MATH 3215
Introduction to Probability and Statistics
Fall Semester 2023

John McCuan

November 5, 2023

Contents

Preface

These notes, in more or less the form of a book, were prepared for an undergraduate introduction to probability and statistics. They may be considered to rely heavily on, and constitute a kind of response to, the readings/notes of Orloff and Bloom (MIT Open Courseware course Math 18.05 Introduction to Probability and Statistics Spring 2014).

Prologue

A prologue is often used to cover information preliminary to the main subject. This prologue is a kind of "jump ahead" to some material that is a few sections into the notes below. The idea is to give some foreshadowing of that which is to come and, hopefully, give anyone who is bored with the introductory/review topics of sets, functions, and counting, something (really new) to think about. For those who are not at all bored, you can look back at this prologue after you have figured out the material leading up to it.

0.1 Lecture "-1" Measure

0.1.1 The rough idea

The basic idea of "measure" is that you have some set, say S with a collection of subsets $\mathfrak{M}$ you wish to be able to "measure." This means, you want to associate some kind of **number** with each set $A \subset S$, at least when $A \in \mathfrak{M}$. There are several "players" involved in this idea, but it is really not too complicated. Those players are the following:

the set S

the specified collection of **measurable sets** $\mathfrak{M}$

a real valued function $\mu : \mathfrak{M} \to \mathbb{R}$ which "measures" the sets in $\mathfrak{M}$.

Perhaps the initial question to ask is this:

What does it mean to **measure** sets?

That is to say, "what are reasonable properties to expect/require of the function μ?" We have already said μ is to be a real valued function.[1] It turns

[1] In fact, this is already something of a restriction in some contexts. There are, for example, complex valued measures, but we will have no need for such exotic measuring.

out that this was a rather difficult question to answer, and the formulation of what are recognized by many as the defining properties of a measure only came into focus in the early twentieth century after many years of hard work.

To get some idea of how one might think about this problem, consider for the set S some interval I of the real line, say the unit interval

$$I = [0, 1] = \{x \in \mathbb{R} : 0 \leq x \leq 1\}.$$

The reason we call it the "unit interval," is because the length is 1, and **length** gives a good idea of what one might want for a measure. Length, certainly works well for intervals. Let $\mathcal{I}$ be the collection of all intervals in $I = [0, 1]$. The collection $\mathcal{I}$ contains, for example, the intervals

$$[0, 1/2) = \{x \in \mathbb{R} : 0 \leq x < 1/2\} \quad \text{and} \quad (0, 1/2] = \{x \in \mathbb{R} : 0 < x \leq 1/2\}.$$

Note that both of these intervals have the same length. We can get started with a measure here by setting

$$\mu(J) = b - a$$

where b and a with $a \leq b$ are the endpoints of the interval $J \in \mathcal{I}$. One property you may note about length measure is that the measure of an interval is always nonnegative. This, it turns out, is a reasonable property to require of any measure. Another property, which definitely makes sense but might not be the first thing you think of, is the following:

> The measure of the empty set should be zero.

In fact, when I mention that it was difficult to formulate the definition of a measure, it was really only one property of a measure that was difficult to formulate. The other two are the ones we have just mentioned:

(0) $\mu(A) \geq 0$ for every $A \in \mathfrak{M}$.

(i) $\mu(\phi) = 0$.

Technically, the formulation of property **(i)** requires the consideration of the collection $\mathfrak{M}$ of measurable sets. Specifically, we must say the empty set is among the measurable sets, that is,

$$\phi \in \mathfrak{M}.$$

More generally, it turns out that careful consideration of the natural requirements for the collection of measurable sets $\mathfrak{M}$ is tied up quite intimately with the formulation of the last tricky property of a measure. Collections $\mathfrak{M}$ of subsets of a set S suitable to consider as the domain of a measure have a special name: **sigma algebra**.[2] As with measures, I will give you the first two (of three) properties that define a σ-algebra:

(0) $\phi \in \mathfrak{M}$.

(i) If $A \in \mathfrak{M}$, then $A^c = \{x \in S : x \notin A\} \in \mathfrak{M}$. That is, $\mathfrak{M}$ is closed under complementation.

Here are a few exercises which might give you some ideas about the third properties (both of measures and σ-algebras).

Exercise 0.1.1 Given the two properties of a σ-algebra $\mathfrak{M}$ above, what other set do you know is in $\mathfrak{M}$?

Exercise 0.1.2 Consider the intervals

$$\begin{aligned} I_1 &= (1/2, 1], \\ I_2 &= (1/4, 1/2] \\ I_3 &= (1/8, 1/4] \\ &\vdots \\ I_j &= (1/2^j, 1/2^{j-1}] \\ &\vdots \end{aligned}$$

What is the relation between the length of

$$A = \bigcup_{j=1}^{\infty} I_j$$

and the lengths of the intervals I_j?

[2] We usually write σ-algebra instead of writing out the name "sigma."

Exercise 0.1.3 Consider the intervals

$$\begin{aligned} J_1 &= (1/2, 3/4], \\ J_2 &= (1/4, 3/8] \\ J_3 &= (1/8, 3/16] \\ &\vdots \\ J_j &= (1/2^j, 3/2^{j+1}] \\ &\vdots. \end{aligned}$$

What do you think should be the measure

$$\mu\left(\bigcup_{j=1}^{\infty} J_j\right) \qquad ?$$

Should this value be called "length?"

Exercise 0.1.4 Let $\mathbb{Q}$ denote the rational numbers. That is,

$$\mathbb{Q} = \left\{ \frac{m}{n} : m \in \mathbb{Z} = \{0, \pm 1, \pm 2, \pm 3, \ldots\} \text{ and } n \in \mathbb{N} = \{1, 2, 3, \ldots\} \right\}.$$

(a) Show that for any (positive number) $\epsilon > 0$, there exists a sequence of intervals $I_1, I_2, I_3, \ldots$ such that

$$[0,1] \cap \mathbb{Q} \subset \bigcup_{j=1}^{\infty} I_j$$

and

$$\sum_{j=1}^{\infty} \mu(I_j) < \epsilon.$$

(b) What should this tell you about the measure $\mu([0,1] \cap \mathbb{Q})$?

(c) What should this tell you about the measure of the irrational numbers in $[0,1]$?

(d) Should the numbers considered in parts **(b)** and **(c)** above be called "lengths?"

Rather than get into the details of the third property used to define a σ-algebra and the third property used to define a measure, in the next section I will restrict attention to a much simpler situation.

0.1.2 Some precise details

Let S denote a **set** or what we will call the underlying set or **measure space**.

Let $\wp(S)$ denote the set of subsets of S. This set is called the **power set** of S.

A **measure**[3] $\mu : \wp(S) \to [0, \infty)$ is a nonnegative real valued function satisfying the following:

(i) $\mu(\phi) = 0$ where ϕ denotes the empty set.

(ii) If $A, B \subset S$ with $A \cap B = \phi$, then

$$\mu(A \cup B) = \mu(A) + \mu(B).$$

The property **(ii)** is called **additivity**.

In order to understand (and study) what is written above, you need to know something (at least a little) about sets and functions. These topics will be covered in some detail below as introductory/review material.

Example 1 If $S = \{h, t\}$, then $\#S = 2$. Also,

$$\wp(S) = \{\ \phi,\ \{h\},\ \{t\},\ S\ \},$$

and $\#\wp(S) = 4$. The function $\mu : \wp(S) \to [0, 1]$ satisfying

$$\begin{aligned} \mu(\phi) &= 0, \\ \mu(\{h\}) &= 51/100, \\ \mu(\{t\}) &= 49/100, \text{ and} \\ \mu(S) &= 1 \end{aligned}$$

is a measure.

Exercise 0.1.5 How many other measures are there on the set $S = \{h, t\}$? Classify all of them. Hint(s): Start with $a, b \in \mathbb{R}$ with $0 \leq a \leq b$ and consider the measure with $\mu(\{h\}) = a$.

We can generalize Example 1 to any set S with finitely many elements.

[3] Really what I am defining here is a "baby measure," but this is a good place to start.

Example 2 If $\#S < \infty$, then

$$S = \{x_1, x_2, \ldots, x_n\}$$

for some n and the values of (any) measure $\mu : \wp(S) \to [0, \infty)$ are determined by

$$\mu\big|_{\Omega} \qquad (\mu \text{ restricted to } \Omega)$$

where

$$\Omega = \{\ \{x_1\},\ \{x_2\},\ \ldots,\ \{x_n\}\ \} \subset \wp(S)$$

is the collection of **singleton** subsets of S.

Exercise 0.1.6 Given the base set S of Example 2, how many elements are there in $\wp(S)$, i.e., what is $\#\wp(S)$? For any $A \subset S$ express

$$\mu(A) = \mu\left(\bigcup_{x \in A} \{x\}\right)$$

in terms of the values of the restriction of μ to the singleton set Ω.

0.1.3 Scaling a measure

If $c > 0$ and $\mu : \wp(S) \to [0, \mu(S)]$ is a measure, then $\nu : \wp(S) \to [0, \infty)$ by

$$\nu(A) = c\ \mu(A) \tag{1}$$

is a (new) measure.

Example 3 Remember the measure from Example 1 with $S = \{h, t\}$ and

$$\mu(\{h\}) = 51/100,$$
$$\mu(\{t\}) = 49/100.$$

There is also a measure $\nu : \wp(S) \to [0, \infty)$ with

$$\nu(\{h\}) = 153/100 \text{ and}$$
$$\nu(\{t\}) = 147/100.$$

Exercise 0.1.7 Verify the function $\nu : \wp(S) \to [0, \infty)$ with values determined by (1) is a measure. What is $\nu(S)$ for the new measure in Example 3?

0.1.4 Probability measure

A measure on a set S with $\#S < \infty$ is called a **probability measure** if $\mu : \wp(S) \to [0, \infty)$ is a measure and $\mu(S) = 1$.

In the discussion below, we sometimes use the symbol π to denote a probability measure $\pi : \wp(S) \to [0, 1]$.

Example 4 Let

$$S = \{x_1, x_2, \ldots, x_n\}$$

be a set with n elements. The measure $\pi : \wp(S) \to [0, 1]$ with values determined by

$$\pi(\{x_j\}) = \frac{1}{n} \qquad \text{for } j = 1, 2, \ldots, n$$

is a probability measure. This measure is called the **uniform probability measure** on a finite set.

Exercise 0.1.8 Let $S = \{x_1, x_2, \ldots, x_n\}$ be a set with n elements, and let $\mu : \wp(S) \to [0, \infty)$ be a measure on S. Verify the following:

(a) $\mu(A) \leq \mu(S)$ for every $A \subset S$. This is called **monotonicity** of the measure.

(b) There is a unique measure with $\mu_0(A) = 0$ for every $A \subset S$. This is called the **zero measure**.

(c) If μ is not the zero measure, the scaled measure $\pi : \wp(S) \to [0, 1]$ by

$$\pi(A) = \frac{\mu(A)}{\mu(S)}$$

is a measure and is a probability measure.

0.1.5 Restriction measure

Say $T \subset S$, $\#S < \infty$, and $\mu : \wp(S) \to [0, \infty)$ is a (baby) measure. Consider $r : \wp(T) \to [0, \infty)$ by $r(A) = \mu(A)$.

Theorem 1 The function r is a measure.

Proof:

(0) $r(A) = \mu(A) \geq 0$.

(i) $r(\phi) = \mu(\phi) = 0$.

(ii) If $A, B \subset T$ and $A \cap B = \phi$, then

$$r(A \cup B) = \mu(A \cup B) = \mu(A) + \mu(B) = r(A) + r(B). \qquad \square$$

Example 5 Let

$$S = \{\text{one}, \text{two}, \text{three}, \text{four}, \text{five}, \text{six}\}.$$

Consider the uniform probability measure on S with

$$\pi(\{x\}) = \frac{1}{6} \qquad \text{for all } x \in S.$$

Consider

$$T = \{\text{two}, \text{four}, \text{six}\} \subset S$$

and $r : \wp(T) \to \mathbb{R}$ with

$$r = \pi\big|_{\wp(T)}.$$

Exercise 0.1.9 Let S, T, π, and r be given as in Example 5. Verify the following:

(a) r is a measure.

(b) r is not a probability measure.

(c) There exists a constant c for which the scaled measure $\nu = c\,r$ is a probability measure.

If we wish to emphasize the dependence of a restriction measure r on the subset $T \subset S$, we may write

$$r = r_T.$$

Exercise 0.1.10 What is the difference between the restriction measure r of a measure μ as described in Example 5 and the restriction

$$\mu\big|_{\Omega}$$

considered in Example 2?

0.1.6 Restriction probability measure

If $\pi : \wp(S) \to [0,1]$ is a probability measure and we take (for some $T \subset S$ with $\pi(T) \neq 0$)

$$\rho_T(A) = \frac{\pi(A)}{\pi(T)} \qquad \text{for all } A \subset T,$$

then $\rho_T : \wp(T) \to [0,1]$ is a probability measure. We may call this measure a **restriction probability measure**.

Notice the value of ρ_T is probably not equal the value of π for any nonempty set $A \subset T$.

Example 6 Let π be the uniform measure on

$$S = \{\text{one}, \text{two}, \text{three}, \text{four}, \text{five}, \text{six}\}$$

as in Example 5. The restriction probability measure ρ_T on

$$T = \{\text{two}, \text{four}, \text{six}\} \subset S$$

is the uniform probability measure on T given by

$$\rho_T(\{x\}) = \frac{1}{3} \qquad \text{for all } x \in T$$

because

$$\rho_T(\{x\}) = \frac{\pi(\{x\}}{\pi(T)} = \frac{1/6}{1/2}.$$

In the language of applied probability: Assuming you roll an even number, the probability you (will) roll a four is 1/3.

0.1.7 Expansion of a measure

Say $T \subset S$ and $\sigma : \wp(T) \to [0, \infty)$ is a measure. Consider $\beta : \wp(S) \to [0, \infty)$ by

$$\beta(A) = \sigma(A \cap T).$$

Theorem 2 The function β is a measure (on S).

Proof:

(0) $\beta(A) = \sigma(A \cap T) \geq 0$.

(i) $\beta(\phi) = \sigma(\phi \cap T) = \sigma(\phi) = 0$.

(ii) If $A, B \subset S$ with $A \cap B = \phi$, then

$$(A \cap B) \cap T = (A \cap T) \cap (B \cap T) = \phi$$

and

$$\begin{aligned} \beta(A \cup B) &= \sigma(\ (A \cup B) \cap T\ ) \\ &= \sigma(\ (A \cap T) \cup (B \cap T)\ ) \\ &= \sigma(A \cap T) + \sigma(B \cap T) \\ &= \beta(A) \cap \beta(B). \qquad \square \end{aligned}$$

The measure β (the big measure) in the construction above is called the **expansion** of the (small) measure σ to S.

Exercise 0.1.11 Let $\pi : \wp(S) \to \mathbb{R}$ be a probability measure and let $T \subset S$ be a set with $\pi(T) > 0$. Find a formula for the value(s) of the expansion to S of the probability restriction ρ_T, and show the result gives a probability measure on S.

It is natural to use the notation ρ_T for both the restriction probability measure of a measure π and for the expansion of this measure to S. The values of the latter are said to express the **conditional probability**.

As the application of Example 6 may suggest, the construction of restriction probability measures and the conditional probability values are considered rather important (in applied probability).

Exercise 0.1.12 Express the values of the restricted probability measure of Example 6 in terms of the language of conditional probability:

$$\rho_T(\{\text{four}\}) \text{ is the probability of } \ldots$$
$$\text{conditioned on (i.e., assuming) } \ldots$$

0.1.8 Bayes' rule

The construction of a restricted measure, which may be applied to any subset $B \subset S$, has some relatively interesting consequences. Recall that given a probability measure π on S, we denote, for any set $A \subset S$ with $\pi(A) \neq 0$, the **restriction probability measure on S with respect to A** by ρ_A with

$$\rho_A(B) = \frac{\pi(B \cap A)}{\pi(A)} \qquad \text{for any set } B \subset S.$$

Given also $B \subset S$ (now fixed) with $\pi(B) \neq 0$ as well, we can form a second restriction probability measure ρ_B on S with respect to B. A simple form of Bayes' rule involving two such sets A and B within S is probably about the most sophisticated tool used in statistics we will encounter. Here is an exercise in which you can derive this form of Bayes' rule:

Exercise 0.1.13 Let $\pi : \wp(S) \to [0,1]$ be a probability measure on a set S having two subsets A and B for which $\pi(A) \neq 0$ and $\pi(B) \neq 0$. Complete the following steps:

(a) Find a formula for $\rho_B(A)$ in terms of $\rho_A(B)$ (and the numbers $\pi(A)$ and $\pi(B)$).

(b) Give an example in which $\rho_B(A) \neq \rho_A(B)$.

Exercise 0.1.14 Let

$$S = \{\text{one}, \text{two}, \text{three}, \text{four}, \text{five}, \text{six}\}$$

and let $\pi : \wp(S) \to [0,1]$ be the uniform probability measure on the set S. Let

$$A = \{\text{two}, \text{four}, \text{six}\}$$

and let

$$B = \{\text{one}, \text{four}, \text{five}, \text{six}\}.$$

(a) Compute $\rho_B(A)$.

(b) Compute $\rho_A(B)$.

(c) Describe $\rho_A(B)$ (in words) as a conditional probability.

(d) Describe $\rho_B(A)$ (in words) as a conditional probability.

At this point, we can sharpen slightly the formulation that (mathematical) probability is a "self-contained subject involving calculation" by saying that probability is about starting with an assumed measure and calculating the values of various other derived measures. When all of these measures are probability measures, the calculated values may be called probabilities.

Exercise 0.1.15 How many measures are involved in Bayes' rule? See also Exercise 0.6.8 in Lecture "−2" (section 0.6) below.

0.1.9 Integration

Whenever one has a measure $\mu : \wp(S) \to [0, \infty)$ on a set $S = \{x_1, x_2, \ldots, x_n\}$ with $\#S = n < \infty$, then one can integrate a real valued function[4] on S with respect to the measure μ. More precisely, let $f : S \to \mathbb{R}$ be a real valued function, then the **integral of f with respect to the measure μ** is defined to be the real number

$$\int f = \sum_{j=1}^{n} f(x_j)\ \mu(\{x_j\}). \tag{2}$$

The integral defined in (2) looks a little different from the integrals you may have encountered in calculus, but it shares some of the same properties.

Exercise 0.1.16 If $f, g : S \to \mathbb{R}$ and $a, b \in \mathbb{R}$, then

$$\int (af + bg) = a \int f + b \int g.$$

Thus, we say integration with respect to a measure is **linear**.

There are a couple interesting constructions using the integration we have just defined. The first is that we can integrate on (only) a subset of S instead of all of S: If $A \subset S$, then

$$\int_A f = \sum_{x \in A} f(x)\, \mu(\{x\}).$$

[4]You probably have some idea about functions, but if the notation and concepts here throw you, do not worry, there will be a review later/soon.

Exercise 0.1.17 What properties should a function $f : S \to \mathbb{R}$ (on a measure space S with $\#S < \infty$) have to ensure $\nu : \wp(S) \to \mathbb{R}$ given by

$$\nu(A) = \int_A f$$

defines a measure.

Here is a second construction:

0.1.10 Average value

The **average value of a function** $f : S \to \mathbb{R}$ **with respect to a measure** $\mu : \wp(S) \to \mathbb{R}$ is defined to be the number

$$\frac{1}{\mu(S)} \int f.$$

Similarly, the average value of f over $A \subset S$ is

$$\frac{1}{\mu(A)} \int_A f.$$

Exercise 0.1.18 What is the relation between the average value of a function f over a set S with respect to a probability measure $\pi : \wp(S) \to [0, 1]$ and the integral of f (with respect to the same measure)?

0.1.11 A big daddy measure space: $\mathbb{R}$

We have defined a "baby measure" as a measure on a set S with $\#S < \infty$. We will consider later several generalizations of the idea of measures to sets with infinitely many elements. There are cases where this is relatively easy and situations in which even explaining the definition of what it means to be a measure is somewhat complicated. In fact, the measure we have defined is "restricted" in more ways than simply the size of the set. In the "real" study of measures, there are measures that take on negative values (!), infinite values, and even complex values. One relatively "familiar" measure that takes on the value $+\infty$ is the measure m on $\mathbb{R}$ with the following property: Given an interval I with endpoints $a, b \in \mathbb{R}$ satisfying $a \leq b$,

$$m(I) = b - a.$$

We say: The measure of an interval is its length. It should be no surprise that the entire real line has

$$m(\mathbb{R}) = \infty,$$

so m does not take values in $\mathbb{R}$. We say m takes values in the **extended nonnegative real numbers** $[0, \infty]$. Of course, there are many other subsets of $\mathbb{R}$ which are not intervals. Can you figure out how to find the value of the measure m on more complicated subsets? Here is a hint:

Exercise 0.1.19 Let $\mathbb{Q}$ denote the **rational numbers**:

$$\mathbb{Q} = \left\{ \frac{p}{q} : p \in \mathbb{Z} = \{0, \pm 1, \pm 2, \pm 3, \ldots\} \text{ and } q \in \mathbb{N} = \{1, 2, 3, \ldots\} \right\}.$$

Show that for any positive number $\epsilon > 0$ there exists a sequence of open intervals $I_j = (a_j, b_j)$, $j = 1, 2, 3, \ldots$ with

$$\mathbb{Q} \subset \bigcup_{j=1}^{\infty} I_j \qquad \text{and} \qquad \sum_{j=1}^{\infty} (b_j - a_j) < \epsilon.$$

Conclude that $m(\mathbb{Q}) = 0$. What does this mean about the set of irrational numbers $\mathbb{R} \backslash \mathbb{Q}$ or the set of irrational numbers in the unit interval $[0, 1] \backslash \mathbb{Q}$?

Here is a final exercise.

Exercise 0.1.20 Let $S = \{x_1, x_2, \ldots, x_n\}$ be a set with $\#S = n < \infty$. Consider $\# : \wp(S) \to [0, n]$, i.e., the cardinality of sets given by the number of elements in the set.

(a) Show $\#$ is a measure.

(b) To which familiar probability measure on S is $\#$ related and how?

0.1.12 Integration

It has been mentioned above that when one has a measure $\mu : \wp(S) \to [0, \infty)$ defined on a finite set S, then one can integrate a real valued function $f : S \to \mathbb{R}$. This assertion is also true about the measure m mentioned above for which the measure of an interval is its length. As we did not go into detail concerning the construction of the measure m, we will also only give some suggestive comments about the associated integration.

Note first that the value

$$\sum_{x \in S} f(x)\ \mu(\{x\}) \tag{3}$$

of the integral of $f : S \to \mathbb{R}$ when $S = \{x_1, x_2, \ldots, x_n\}$ is a kind of **weighted sum** of the function values with weights given by the values of the measure (on the singleton set Ω). This is, on the one hand, simply a potentially interesting algebraic quantity. On the other hand, such an integral can have an interesting value in applications.

Example 7 (expectation) Let us say we roll a six sided die with side markings modeled by the set

$$S = \{\text{one}, \text{two}, \text{three}, \text{four}, \text{five}, \text{six}\}$$

as in Example 5. Hopefully, the probabilistic application of the uniform measure on this set is fairly evident by now. Now consider playing with me a game in which you pay me 10 ounces of silver if the die comes up "one," 5 ounces of silver if the die comes up "two." I pay you 2 ounces of silver if the die comes up "three," 4 ounces of silver if the die comes up "four," 6 ounces of silver if the die comes up "five," and 8 ounces of silver if the die comes up "six."

Associated with the game described above, we can define your "payout" function $f : S \to \mathbb{R}$ by

$$\begin{aligned} f(\text{one}) &= -10 \\ f(\text{two}) &= -5 \\ f(\text{three}) &= 2 \\ f(\text{four}) &= 4 \\ f(\text{five}) &= 6 \\ f(\text{six}) &= 8. \end{aligned}$$

The integral of the function f is called the **expectation** associated with this game.

Exercise 0.1.21 What is the meaning of the integral (weighted sum) of the function f in Example 7? Why is this value called the expectation?

Now let us try to make some comparison of the weighted sum associated with a real valued function on a measure space with finitely many elements to an integral with respect to the measure m. For this we restrict to a special case: First of all, we take the measure space to be an interval $I = [a, b] \subset \mathbb{R}$ with $a, b \in \mathbb{R}$ and $a < b$. Second, we assume the function $f : I \to \mathbb{R}$ is **continuous**.

Continuity of the function f in this case is the natural condition under which the integral

$$\int f$$

with respect to the measure m on the interval $I = [a, b]$ takes the same value as the Riemann integral

$$\int_a^b f(x)\, dx$$

which is familiar from elementary calculus. Under these assumptions, furthermore, comparison is relatively easy. The integral here is not quite a weight sum (weighted by the lengths m of certain intervals), but it is a **limit** of such weighted sums, which in this case are called Riemann sums. Precisely,

$$\int f = \lim_{\|\mathcal{P}\| \to 0} \sum_{j=0}^{n-1} f(x_j^*)\ m([x_j, x_{j+1}])$$

where $\mathcal{P}$ is a **partition** of the interval, specifically,

$$\mathcal{P} = \{x_j\}_{j=0}^n \qquad \text{with} \qquad a = x_0 << x_1 < x_2 < \cdots < x_n = b$$

and the **norm of the partition** is

$$\|\mathcal{P}\| = \max\{x_{j+1} - x_j : j = 0, 1, 2, \ldots, n-1\};$$

the points x_j^* for $j = 0, 1, 2, \ldots, n-1$ are called **evaluation points** and satisfy

$$x_j \leq x_j^* \leq x_{j+1}.$$

The sum

$$\sum_{j=0}^{n-1} f(x_j^*)\ m([x_j, x_{j+1}]) \tag{4}$$

is called a **Riemann sum**, and it is clear that it is a sum of function values weighted by measures of associated sets (intervals). The meaning of the limit is that

$$L = \int f$$

is the (unique) number for which the following holds:

> Given any $\epsilon > 0$, there is some $\delta > 0$ such that for any partition $\mathcal{P}$ (and any associated evaluation points) for which $\|\mathcal{P}\| < \delta$, there holds
>
> $$\left| \sum_{j=0}^{n-1} f(x_j^*)\ m([x_j, x_{j+1}]) - L \right| < \epsilon.$$

Exercise 0.1.22 Compare the Riemann sum in (4) to the integral over a finite set given in (3). How are these expressions the same or similar? What are the differences?

Introduction

0.2 Probability and Statistics

A natural division of this course is into two or three sections. The two main sections would be an introduction to probability and an introduction to the application of probability in statistics. I will include a third preliminary section on sets and functions. That is, a course outline might look something like this:

1. Preliminaries (hopefully rather short)
   (a) sets
   (b) functions
2. Probability
   (a) mathematical probability (measures)
   (b) applications of probability
3. Statistics
   (a) elementary statistics (basic quantities associated with data)
   (b) more advanced statistics (inference from data)

This outline is incomplete in several respects. Notably one can identify an additional section involving counting. It is not clear whether the material on counting is to be considered "preliminary" or part of probability. In any case, some material on counting will be covered in the course. Even more notably one can identify an additional section on **simulation** or more properly probabilistic simulation. Again, it is not precisely clear where discussion

of simulation fits in as a self-contained topic. Certainly, it is natural to use simulation to introduce the topic of statistics and create a kind of bridge between probability and statistics via the creation of simulated data sets. It is natural, however, to discuss simulation long before any transition from probability to statistics, and we should do that. Many programs may be used for simulation from Open Office Calc to Mathematica to R (not to mention acutal programming languages like C or Python). There is little doubt that the most marketable option for a student to add to his resume in regard to creating the impression of expertise[5] in statistics is R. More details will be provided on the topics in this outline in the lecture notes below.

I like very much the description of Orloff and Booth [4] asserting roughly that probability is a self-contained subject centering around making calculations. In particular, probability does not have anything inherent to do with data. Statistics, on the other hand, is a subject that begins with data, and in statistics one attempts to apply probabilistic calculation to make conclusions from data. I do not entirely agree with Orloff and Booth, and I will certainly attempt to refine this description in my notes below, but this is probably a good solid starting point (of view) if you have little idea about the subjects themselves, especially statistics: Probability is not about data (sets). Statistics is definitely about data.

0.3 Course materials

I like certain aspects of the notes and course materials of Orloff and Booth published as MIT open courseware in the spring of 2014 [4]. As the authors discuss partially, these materials have certain drawbacks. One of the drawbacks I anticipate for this course is that the problem sets of Orloff and Booth, taken as they are in relation to the provided (MIT) course notes, are (too) challenging. On the other hand, there is significant opportunity for a student who is not overly discouraged to learn something from (those) challenging problem sets. I like also certain aspects of a book by Alexander Farmer [2] portions of which I will make available and have used to compose the lecture notes below. Farmer's book also has certain drawbacks, though I like very

[5]It may be noted, however, that the course is not designed to actually produce anything like expertise in statistics but only to provide an elementary introduction. Specifically, the intention of the instructor is limited to a somewhat modest version of such an elementary introduction.

much his measure theoretic point of view and certain philosophical aspects which I hope the students will find interesting if not valuable.

Passing reference may be made to various other texts on introductory probability and statistics, notably [3] and [1], which I happen to have at hand and the former of which has been used as a text (by other instructors) for MATH 3215 at Georgia Tech. Such references will primarily be to point out contrast with the presentation and point of view taken in my notes. In some sense, the presentation in such books is not so different from that of Orloff and Booth.

0.4 Course activities and procedures

At this point I wish to say something concerning advice for the students in my course (MATH 3215 Introduction to Probability and Statistics, fall semester 2023 at Georgia Tech) and something concerning my philosophy of thinking and learning in particular. Notice I did not say "teaching and learning." First of all, a student may be startled to have me state, up front, that I have **no expectations** concerning your performance in this course. Accordingly, I do not consider it my role to set up routine "hoops" for you to jump through creating the illusion that you are actually learning something when (almost) nothing could be further from the truth. I expect this kind of "jumping through hoops," taking exams and such—doing as little as possible in each course and avoiding thinking at all costs—is probably a pretty good description of most students' experience with "education." I have something very different in mind. I think fundamentally you should be motivated by **your own expectations** (of yourself and what you desire to learn). An indicator that you have some of your own expectations is that you are asking reasonable thoughtful questions (of yourself) and trying to answer those questions. When that happens, you are essentially guaranteed to have started learning. When you are not asking questions independently, you are almost guaranteed to be doing something other than thinking and learning.

My advice is that if you have no idea why you want to learn something about probability and statistics and view this semester as something other than an opportunity to have time set aside to do that thinking and learning, then you should drop the course, sign up for a different course,[6] or simply do

[6]I would say "sign up for a different instructor," but I understand that is administra-

something different. I suggest this because I think you will be wasting your time. Of course, I imagine you are used to wasting your time, but without the busy work of jumping through hoops creating the illusion that something different is happening, the awareness of what is happening is something you will likely find, shall we say, unsettling if not irritating.

There are lots of books, course materials, and problem sets available for probability and statistics. Presumably, you will want to read, look at, and work on some of these. There are three main activities I intend to facilitate, primarily by suggestion but also in other regards, which are hopefully related to learning and definitely related to communicating what you may or may not have learned to others including to me. These three activities are the following:

1. **Assignments**. I will post "assignments" involving various problems throughout the semester. I will compose these assignments myself and provide you an opportunity to upload your work on them on Canvas. You should have at least one week to work on each assignment before the "due" date.[7] A due date is not a deadline. Please do not send me an email with the word "deadline" in it. The only deadline in my course is the end of the semester. After that, we have nothing further to discuss. The due date only means that someone (like a grader) may look at the work you have uploaded and provide you with some feedback (if you turn it in by the due date). You may upload your work at any time until midnight of the scheduled final exam day. I will be able to see it is there at the end of the semester, but you cannot expect the grader will give you feedback on it. There is never any reason to contact me about an "extension" or tell me that you are turning your work in after the due date (or not at all). I recommend you work the problems, write up carefully composed and neatly expressed solutions, and upload them by the due date. This (I hope) will help you make good progress in learning the material. Realistically, I think most of you will need to spend 5 to 8 hours per week (outside of class) in order to learn and master the material in some kind of reasonable fashion.

   As some kind of administrative matter (which I do not view as terribly important because it is not really directly related to learning) some of

tively difficult.

[7] In fact, I will try to keep up so that each assignment is posted two weeks or so before the assignment is due.

the assignments will be designated as "exams." You probably do not need to take much special note of the distinction between "regular assignments" and "exams." There will be no in-class exams. All "exams" will be "take-home" exams. If you want to "cheat" on the exams (and other assignments) by mindlessly copying the work of another student, knock yourself out. I don't think you can learn anything this way. There is little to no motivation for you to do it. If you want to learn something, you will not do it. But as I say, if it makes you feel better, knock yourself out.

Working with other students on problems—if you actually think about what you are doing and develop independent mastery of the material—can be useful in learning. The key is essentially to keep in mind that anything you learn can only be learned by you. You have to do the learning. This means that essentially you have to set the goal concerning what you want to be able to do (and learn) and figure out how to produce that result. If you want to be able to calculate the probability of being dealt "two pair" in poker, then it is essentially up to you to figure out the details of how to do that. If you want the experience to allow you to accomplish other (similar) things, then you'll need to invest the time to think deeply enough about what you are doing in order to make that happen.

In conclusion let me just say: Working with other students on the assignments is generally encouraged if it facilitates thinking and learning.

2. **Papers**. I would like for you to write three papers. The papers do not have to be long or technical. You do not need to make irrefutable arguments, but what you write should give some indication that you have thought carefully about the topic and have something to say. (If you need for me to specify a particular length for you, then I offer "one to ten pages double spaced," though I could imagine writing something very insightful and perfectly acceptable in a half page.)

   (a) In the first paper I would like for you to express what you think about **grades**. You can, of course, discuss anything you like and the various "grades" typically assigned for homework assignments or whatever activities you might undertake in a university course and have "graded" in particular. But what I am really interested in is what you think about **course grades** that appear on your

transcript. How would you like your grade determined? What do you think it means? What should it mean? You may consult what I have written in section 0.5 below.

This paper will have a due date sometime early in the semester. I don't have any idea how much time it will take you, but presumably this is a subject into which you can put some thought and produce something expressive of your beliefs on the topic.

(b) For the second paper, I would like you to address the following question related to the application of probability:

> Does the number 1/2 have anything to do with an individual coin flip?

You watch another person flip a coin (one time) into the air, grab it in the air while it is spinning, and slap it down on a table. You didn't see the outcome (heads or tales) and the coin is hidden beneath the person's hand. Does the number 1/2 give you any information about what has happened?

This paper will be due sometime early in the second half of the semester.

(c) For the third paper, I would like for you to address the following hypothetical situation:

> Assume you experience severe aches and pains along with an apparent reduction in circulatory function resulting in recurrent, nearly debilitating pain, weakness and significant loss of use of your hands, feet, arms and legs. After many medical tests it is determined that you have experienced a rare identifiable viral infection which, though dormant, has left you with a chemical-neural auto-immune disorder. One possibility is that the symptoms will decrease over a long period of time and you will (likely) regain most function. Another possibility is that the viral infection becomes active a second time. It is known that a second infection often results in permanent paralysis or death.
> Your doctor(s) recommend the ingestion of a certain substance and give you the following information:

i. Ingesting this substance causes moderately serious side effects such as reduced kidney and liver function and possible heart damage. The side effects are considered relatively minor compared to the outcome of a second infection.

ii. In multiple studies of people who have experienced the same viral infection and auto-immune response that you have, it has been found that 97% of those who ingest the recommended substance do not experience a second infection (while 3% do expereince a second infection).

iii. In multiple studies of people who have experienced the same viral infection and auto-immune response that you have, it has been found that 90% of those who do not ingest the recommended substance do not expereince a second infection (while 10% do experience a second infection).

You ask the doctor(s) if there is any information on what differentiates the individuals who fall into the 97% from those who fall into the 3% among the individuals ingesting the substance in the studies, and/or if he can discern any particular characteristics of you as an individual that might help you determine if you as an individual fall into one of these categories or the other. Similarly, you ask the doctor(s) if there is any information on what differentiates the individuals falling into the 90% from those who fall into the 10% among the individuals not taking the substance in the studies, and/or if he can discern any particular characteristics of you as an individual that might help you determine if you as an individual fall into one of these categories or the other. The response is uniform: Your doctors have no further information whatsoever.

Here is the question: Have you been given any useful information to help you decide on whether or not you should ingest the recommended substance or how you should proceed?

This paper will be due near the end of the course. Presumably this question is worthy of significant thought and significant in-

vestment of time. Presumably, you will want to know at least something about statistics before you write this paper.

3. **Presentations/projects**. I would like for you to make two twenty minute in-class project presentations and prepare auxiliary materials for each. One presentation should be on a topic in probability, and the other presentation should be on a topic in statistics.

   I will give you some fairly detailed guidelines for this activity I hope are good ones, but everything is flexible.

   (a) The basic idea here is to study a topic in detail so that you can make a twenty minute in-class presentation on the topic—a kind of mini-lecture—from which your classmates can benefit.

   (b) In addition, I would like for you to prepare a set of notes containing significantly more material than you cover in the twenty minute presentation (for distribution to your classmates).

   (c) Similarly, I would like for you to prepare a problem set containing around 5 illustrative problems concerning the material you have covered.

   The two topics you choose should be "balanced" in some way. For example, you should not give one presentation on a very elementary topic in probability and the second on a very elementary topic in statistics. One presentation can be relatively elementary or both may be of moderate difficulty. Whatever the case, you should try to push yourself, think carefully about the topic, and make both presentations of high quality. I think you should anticipate spending up to 40 hours on each project presentation.

   I hope you will use my notes, problem sets, and in-class presentations as examples.

0.5 Grades

It is the case, for better or for worse, that each of you who takes this course and are still on the role (maintained by the registrar) at the end of the semester will be assigned a course grade which will appear on your transcript, assuming you are taking the course under the "graded" option, which should

be the vast majority of you.[8] It is also the case, for better or worse, that I am the one who has the job of assigning those grades.

Aside from a certain distaste concerning this activity of assigning grades, here are some things you may wish to know about how I am thinking about "course grades," "assignment grades," and so on.

1. My basic (abstract) view is that if you want to learn the subject and you put in a reasonable effort to learn the subject, then I want to give you an "A" for your transcript grade, and I will make every effort to do that.

2. I'm also open to the idea that you do not want to learn the subject (or have no idea what it might mean to learn a subject) but are signed up for this course because some administrator decided it was a "requirement" and it is a "requirement" that you have somehow gotten yourself into being "required" of you. This is a bit more tricky for me, but I am happy for you to have an "A" on your transcript too, which is probably what you want as well, and hopefully it will be relatively easy for you to make that happen.

3. I view the "intermediate grades," i.e., grades for assignments, papers, and project presentations (especially the grades for assignments) as primarily for feedback. If you are really not getting something on an assignment, then hopefully when you see "points" taken off, you will go back and redo whatever it was you did not yet figure out. But if you see 50 out of 100 points for an assignment, this does not mean your course grade is not going to be an "A." You can get 50 out of 100 points and still be trying quite honestly and reasonably to learn the material, and you should still get an "A" on your transcript. So that 50 out of 100 on an assignment does not mean "B" or "C" course grade. You can think of assignment grades as, more or less, "marks of completion." When I see 50 out of 100 as an assignment grade, it means to me: "This student did the assignment." Of course, it may mean "This student only did half the assignment," so if all (or the vast majority) of your assignment grades are 50 out of 100, then I will (have to) look more closely to see what you've turned in. Hopefully it won't come to that.

[8] If you are taking the course "pass/fail," perhaps you can talk to me briefly about that.

Getting 50 out of 100 points on a project presentation is also a kind of feedback, but of a different sort. If you see this, it means I really don't think you did a reasonable job on your presentation. You have wasted your time, the time of your classmates, and my time. Otherwise, grades for papers and project presentations should be mostly indications of completion as well. Typical such grades should be between 90 and 100.

Of course, if you do not do some assignments, papers, project presentations at all, then you'll see 0 out of 100 points, and we'll all know what that means, which I come to presently.

4. Practically speaking, if I see at the end of the semester that you've done all (or most all of) the homework assignments, you've written three papers, and you've completed two project presentations, then I will assign you an "A," whether you did it to learn or for some other reason.

   If you do not do some of the suggested activities that make an "A" obvious (which is really not so much) then I'll start to wonder what happened. Maybe I'll ask you: "What do you want me to give you for a grade on your transcript?" Maybe I'll still give you an "A," or maybe I'll suggest a "B" would be more appropriate. If it's clear you didn't do (or learn) much during the semester, then hopefully you really can't expect to get an "A," though amazingly that does happen. If you do all the assignments well but make no project presentation, then you should not expect to get an "A." In that case you will probably be assigned a "B." Similarly, if you show up to class (though I'm not going to take roll or anything like that) and I see you did at least a reasonable amount of work (say many assignments, a paper or two indicating some serious thought, and one project presentation) then you can expect to get a "B."

   If I look at the end of the semester and (to me) it doesn't look like you've done much of anything, then I'll probably assign you a "C." So there's something: I don't have any intention to assign anyone a course grade to go on your transcript that is lower than a "C." You can sign up for the course, do nothing else (maybe pay some money to the registrar) and get a "C" on your transcript. Maybe that's good enough for your "requirement." If you do not want to get a "C," then you should make some kind of effort to make sure I do not assign you

one.

5. I'm pretty flexible on grades. I have a lot of thoughts about them—course grades on a transcript in particular—and I'm open to discuss the matter with you. In fact, I'm interested to see if you can make any sense of the subject (of grades) and that is why I suggest you write a paper on it. Maybe the way you are viewing course grades on your transcript, it *should* be the case that at the end of the semester I have no idea who you are and can see nothing you've done (with the exception of your extremely persuasive paper on grades of course) and I should assign you an "A" instead of a "C." Maybe I'm totally neglecting your insightful perspective on the matter. I'm open to that.[9]

   Of course, maybe you think every time you turn in some silly hoop jumping exercise, you should be able to tally some specific (and meaningless) intermediate grade and sum up everything with a calculator to see your course grade to within one ten-thousandth of a point and/or determine precisely how many minutes you can spend on each of your courses to optiimize such things for the purpose of maintaining your scholarship. That's fine too. Write me a paper about it. Hopefully we can get along anyway.

6. If you have any question about grades, please consider it carefully in light of the cursory comments I have written above first. If you still think there is something to discuss, then I'm happy to hear your question.

0.6 Lecture "−2": The base rate fallacy

In their notes for "class 3" (section 7.1) Orloff and Booth give an example (Example 10) which I think may be an interesting place for some of you to start. It is a little complicated, maybe a little confusing, and maybe somewhat counter-intuitive (which is the point) so I will try to take it slowly. Here is the basic question:

> Say you have a population with a base rate for a certain sickness/disease of 0.5 percent. To be very specific, say this popula-

[9] If you're going to go for this, it's probably a good idea to give me a "heads up," though of course this will pretty rapidly negate the possibilty that I have no idea who you are.

> tion contains 4000 individuals. (See Exercise 0.6.1 below.) There is also a "test" designed to determine if an individual has the disease, but the results are subject to error. You are given/told that the test has a 5 percent false positive rate, and the test has a 10 percent false negative rate.
>
> Now, say you specifically are in this population and the test is administered to you. The result comes back positive, in other words the "test" indicates you have the disease. What is the "probability" you actually have the disease?

Below is a sequence of exercises some of which are, more or less, designed to give you a warm up calculating percentages, some are designed to solidify your understanding of what the question has as given, and eventually I hope they can help you figure out what the question is asking and how to answer it. We can discuss later what the (counterintuitive) answer actually means.

Incidentally, Orloff and Booth characterize the 5 percent false positive rate and 10 percent false negative rate for the "test" as "highly accurate."

Exercise 0.6.1 If there are 4000 individuals in a population, and there is an infection rate of 0.5 percent, how many individuals are infected?

Exercise 0.6.2 When you are given that the "test" has a 5 percent false positive rate, this means something like the following:

> Among individuals who do not have the disease,[10] 5% of them will receive a positive test result.

It is not entirely clear how one can arrive at such a false positive rate. In order to do so, presumably one needs an alternative test (perhaps a more accurate test) to determine a collection of well individuals. In any case, the "given" is probably put forward to be understood as applicable (approximately) to any population. Let us apply the concept directly to the specific population of 4000 individuals mentioned above.

(a) Assuming every individual in the population of 4000 individuals is tested, how many individuals receive a false positive result?

[10] For simplicity, let us refer to people who have the disease as "sick" and those who do not have the disease as "well."

(b) What collection of individual's results represent the **complement** of the false positive results, and how many of these individuals are there? Hint: The answer to the first question is the "true negatives."

Exercise 0.6.3 In view of Exercise 0.6.2 above:

(a) State in words what it means for the test to have a 10% false negative rate.

(b) Assuming every individual in the population of 4000 individuals is tested, how many individuals receive a false negative result?

(c) What collection of individual's results represent the **complement** of the false negative results, and how many of these individuals are there?

Exercise 0.6.4 Were one to assert that this test has 95% **accuracy**, what might this mean? Can you make a calculation to see if the assertion is true?

Exercise 0.6.5 Finally, let's assume you are in the collection of individuals who have received a positive test result. Let us say the "probability" you are actually sick is given by m/n where n is the given number of individuals who received a positive test result like you did, and m is the number of those individuals who are actually sick. What is this number?

Note carefully, the "probability" that, given you are sick, the test result is positive is 90/100. Also, given simply that you take the test (and nothing else), the "probability" the result you get is correct is rather high. (What is it?) But these numbers are very different from the "probability" that, given you have received a positive test result, you are sick.

I like very much the way it is phrased by Orloff and Booth:

> The fact that the test is 95% accurate does not mean that 95% of positive test results are accurate.

This is what Orloff and Booth call the "base rate fallacy." The base rate is the 0.5% of disease in the population.

Exercise 0.6.6 How do the calculations/probabilities change if you change the base rate? How about the other parameters? Can you make a spread sheet allowing each parameter to be changed?

Exercise 0.6.7 As mentioned above Orloff and Booth assert in the statement of their problem that the test is "highly accurate." How would you characterize this statement? For example, is it a statement of fact or a statement of opinion? Is the statement required for finding the answer to the problem? Why is such an assertion included?

Recall the characterization of probability in section 0.1.8 as starting with a given measure and calculating values associated with other (derived) measures.

Exercise 0.6.8 Formulate and derive the solution of the base rate fallacy in terms of Bayes' rule. Explain in detail the various measures and values of those measures involved. Which are given and which are calculated? See also Exercise 0.1.15 in section 0.1.8 above.

0.7 Lecture 0: Events and outcomes

The world is full of events: a solar eclipse, a high school student drawing an ellipse, the computation of a first derivative, the purchase of a first house, a wedding, the fighting of a battle, the reading of a prison sentence. It seems to me that some events are particularly suited to the subjects of probability and statistics. I'm not sure how to define which events those are precisely, but I think I'll start by trying to informally describe some of them and some of their characteristics.

0.7.1 The main event

For me, the main event is a coin flip. Most introductory textbooks start with more complicated events like three consecutive coin flips, rolling multiple dice, and things of this sort. I am pretty sure this is done partially in order to get the students computing things immediately (namely probabilities) before thinking carefully about the meaning of those computations, and thinking carefully about the single coin flip in particular. But this is not most textbooks, most courses, or most anything else, so I'm going to start with the single solitary coin flip.

The first thing I'll point out is that it is natural for me to think about a coin flip in at least three fundamentally different ways. More precisely, I can think of three different kinds of coin flips. First, there are those coin flips

that have happened in the past. The outcome of such a coin flip (heads or tails) has been observed. Among these coin flips is, for example, the coin flip at the beginning of the 2022 Super Bowl[11] a video of which may be viewed on the internet. There are also coin flips that have not taken place yet, and these fall into a fundamentally different category. Examples (at this moment) would be the coin flip at the beginning of the 2030 Super Bowl or the coin flip at the beginning of the 4022 Super Bowl. Another example might be a particular coin flip I have scheduled to make part way through one of my lectures. Moving a bit further in the direction of the abstract, I can consider a future coin flip that is not particularly scheduled but that I expect will take place sometime when I want to choose between having dinner in a restaurant or going home to eat. These future coin flips are much more abstract. Any one of them may never happen. In some sense they are "just in my mind," and one thing that is certainly very different about them from the past coin flips, is that I do not know the outcome.[12] The third kind of coin flips that immediately come to mind are purely abstract coin flips. These have no, and no anticipated, realization as (concrete, physical) past coin flips. They are just totally in my mind, and are going to stay there. I still think of these (totally abstract) coin flips as "real world" coin flips because they share certain characteristics with the other coin flips. They still have possible outcomes heads and tails (and edge). Most importantly, they are "real world" in the sense that they share with future coin flips the possibility that I do not know the outcome.

In summary, I am suggesting the consideration of three kinds of coin flips:

(i) concrete coin flips (that actually took place in the past)

(ii) abstract coin flips (that are imagined to take place in the future)

(iii) abstract coin flips considered purely as an abstraction.

It may be reasonable to consider coin flips in a different way or in multiple different ways. Notably, one may consider, for example, "a coin flip that comes up tails" as an event so that the outcome is included in the event. That would be a different way to think about coin flips. Whatever the case, one should try to be clear.

[11] This is some kind of American football competition. The coin came up tails.

[12] At least it seems possible I do not know the outcome, and more generally, it may be natural to consider the outcome, in some sense, "unknown."

With the distinction I have drawn (between events and outcomes) coin flips are not of direct interest in probability (and statistics) per se. It is rather the **outcomes** of coin flips that are of nominal interest. Having mentioned outcomes several times, let me attempt to describe them informally but with a certain attention to detail. To each of the three kinds of events (specifically kinds of coin flips) discussed above there corresponds a kind of outcome:

(i) concrete event $\rightarrow$ known (concrete) outcome

(ii) future event $\rightarrow$ future outcome

(iii) hypothetical event $\rightarrow$ hypothetical outcome.

Exercise 0.7.1 Is it reasonable to distinguish between events and outcomes as I have done above? Can there be an event without an outcome? Can there be an outcome without an event?

For abstract events, I will now introduce a fourth kind of outcome which generalizes the outcomes considered above in an important way. Unfortunately, the main event (the coin flip) is not the best example to use in illustrating this fourth kind of outcome. Something slightly more complicated is better. Let me suggest the extracting of three balls from a jar containing ten balls, three of which are red, three of which are yellow, and four of which are blue. This is the event. I take as outcome the collection of colors (of balls) drawn.[13] With this basic identification of the outcome, there are $3+3+1 = 7$ possible outcomes. (Three outcomes with one color, three outcomes with two colors, and one outcome with all three colors.) The fourth kind of outcome is distinct in that it may not specify/identify any of the possible outcomes but may allow for several. For example, we can consider

"one of the three extracted balls is red"

as an outcome (in its own right). It will be noted that exactly four of the seven primary outcomes fulfill this new kind of outcome. We may call this fourth kind of outcome a derivative outcome, a **compound outcome**, or a specified outcome. The last term, specified outcome, is slightly paradoxical as these outcomes tend to be less specific than those considered previously.

[13]There are of course other possibilities for the outcome (of interest) of this particular event of "extracting balls." Notably, one might consider the **ordered** triple of colors corresponding to the first, second, and third extraction. This is not an important observation/distinction for our current discussion, but it may be worth noting in general.

Exercise 0.7.2 Can you think of a situation where it is natural to associate a proper compound outcome (meaning one that is not simply a single possible abstract outcome) with a concrete event? What restrictions apply? For example, if three blue balls have been extracted from a jar, is it natural or reasonable to consider the outcome "one of the three balls is red?"

Exercise 0.7.3 What is the only possible proper compound outcome for a single coin flip?

I will suggest below a framework in which to mathematically model the four kinds of outcomes considered above.

For the moment, however, here is the fundamental question:

> Can you say anything about the outcome of an abstract coin flip?

More generally, can you say anything about the outcome of an abstract event? One thing that many people agree on is that an abstract event (at least one suitable for consideration in probability and statistics) can only have one outcome. (And this is saying something.) A coin flip can only come up heads or tails. The outcome may not be known, there may be two (or many more) possible outcomes, and for future or abstract events one may consider compound/specified outcomes, but when an abstract event becomes concrete (or is imagined to have become concrete and to possess a known outcome) there can be only one outcome.

Beyond this, I think, a choice needs to be made—or a belief, a matter of faith, needs to be asserted or come into play. One choice is to say:

> "No more can be said without further information. Yes, an abstract coin flip can come up heads or tails and only one of those, but nothing meaningful can be said about which outcome will be observed (at this point in time)."

An alternative suggestion is to say something like this:

> "There is a 50% **chance** that the coin will come up heads."

That is to say, the coin will come up heads with **probability** 1/2.

Perhaps some elaboration on the statements above, and especially the first one, will make what I am suggesting clearer. A simple appendix to the first quote should be adequate:

> "No more can be said without further information. Yes, an abstract coin flip can come up heads or tails and only one of those, but nothing meaningful can be said about which outcome will be observed (at this point in time). The outcome of the coin flip will be entirely deterministic, determined by the (admittedly complicated) details of the coin flip process, accelerations, velocities, initial positions, nature of the coins, and other things, but chance is not involved. There is no such thing as chance, and there is no such thing as (applied) probability."

In contrast,

> "The outcome of a coin flip is **random**. There are actual observable **random** processes giving meaning to the ideas of **chance** and **probability**. The assertion that there is a 50% **chance** or **probability** 1/2 that the coin will come up heads conveys something fundamental about an abstract coin flip."

I suggest you take as a first task thinking about which of these statements best represents your belief. I am going to flip a coin later in this lecture. Which of these statements best expresses your understanding, based on your beliefs, about that future abstract coin flip? Is there a 50% chance the coin comes up heads? Will there be a 50% chance the coin comes up heads? I think this is an important first question to consider, and I think the single coin flip is a very good event to focus on when considering it.

Let me also suggest that, if you haven't thought deeply about the question before, you not settle on one belief or the other, at least for a few days. Some comments in the next sections below may also be worth considering.

0.7.2 Other considerations

Even if one embraces the first (deterministic) belief and, make no mistake, **probability** as it is mostly known and practiced, i.e., applied probability, is fundamentally and at a very deep level meaningless to such a person, there is still something to consider. There is **mathematical probability**. Mathematical probability, as we shall see, says nothing directly about chance, random processes, or applications of probability. Even the objectionable word "probability" can, for the most part, be avoided. Another way to describe **mathematical probability** is as "the study of measures with total

value one." We can soon see what this subject (mathematical probability) is about, and let me suggest that it should be (at least a little) fun for everyone.

Kinds and repetition

A second consideration for those who embrace determinism and reject chance[14] is the observation that in certain instances when, not a single coin flip, but a rather "large" number of coin flips (actual concrete coin flips) have their outcomes recorded, and the number of heads is counted, then something at least close to the number 1/2 does appear. Let's make this quite precise: There are instances when one records the number $\mathcal{H}(n)$ of heads appearing among n concrete coin flips for various increasing values of n, and for large values of n the ratio

$$\frac{\mathcal{H}(n)}{n} \tag{5}$$

takes values close to 1/2. You can create such instances for yourself. Thus, it seems quite possible that one can maintain that while, on the one hand, the number 1/2 (or possibly some other number close to 1/2) is utterly irrelevant to every single (deterministic) coin flip, such a number is actually observable with respect to (at least certain classes of) coin flips in general.

Exercise 0.7.4 Note something interesting about the ratios (5):

(a) How different can $\mathcal{H}(101)/101$ be from $\mathcal{H}(100)/100$?

(b) How different can $\mathcal{H}(1001)/1001$ be from $\mathcal{H}(1000)/1000$?

(c) How different can $\mathcal{H}(n+1)/(n+1)$ be from $\mathcal{H}(n)/n$?

(d) What does this tell you about the rate of change of the ratio $\mathcal{H}(n)/n$ as a function of n?

For those who believe in chance (and in infinity or a reference to infinity in this context) it is usually very easy to believe something like the following:

> The quantity $\mathcal{H}(n)/n$ appearing in (5) has a limit as n tends to infinity,
>
> $$\lim_{n\to\infty} \frac{\mathcal{H}(n)}{n} = \alpha,$$
>
> and that limit α is close to 1/2.

[14] And probably for everyone else as well.

Even for those who might have objections to this strong statement, it is difficult to completely ignore some roughly equivalent practical statement like the following:

> In many practical instances of large numbers of concrete event outcomes, quantities like the ratio
>
> $$\frac{\mathcal{H}(n)}{n}$$
>
> of heads to the total number n of coin flips, tend to "stabilize" as n becomes larger, and when n is a large number of (certain) coin flips, the ratio in (5) "stabilizes" somewhere near 1/2.

Our use of the word "stabilizes" here is somewhat vague, and it is certainly not mathematical. Nevertheless, I think it is somewhat difficult to deny that **something** observable is present, and this something has some substance to it. I think, furthermore, that this is really the origin of the numbers associated with probability. Such numbers, for example 1/2 for coin flips, may not be saying anything about a single coin flip, but rather the number 1/2 is saying something about ratios associated with large numbers of coin flips.

Exercise 0.7.5 You can further explore your belief about probability by considering the following: A six sided die is rolled many times, and it is found that the ratio

$$\frac{s(n)}{n}$$

of sixes that are observed to the number n of rolls stabilizes around 1/6. If you roll that die a single time (abstractly), can you say anything about the outcome of that single roll (beyond the fact that there are six possibilities—one, two, three, four, five, and six—for the outcome)? For example, can you say the probability of rolling a six is 1/6? (Does that statement have any meaning or relevance to your roll?)

Exercise 0.7.6 If you can play the game described below with the die described in Exercise 0.7.5 but you can only play it exactly once, would you play?

(i) If you roll something other than a six, you lose/pay 5 ounces of silver.

(ii) If you roll a six, you win 1000 ounces of silver.

Exercise 0.7.7 Describe circumstances in which games similar to the one described in Exercise 0.7.6 might be played repeatedly large numbers of times.

(a) What difference does it make if the payouts for winning and losing are varied?

(b) Assume you play the game with a single other individual human being (as your opponent). Does this make a difference?

Kinds of events

I have mentioned "kinds" of events above. This is another word that is perhaps worth considering somewhat carefully, or with some attention to detail. To say two events are of the same "kind" is to suggest the two events share certain characteristics, so one should perhaps be precise about what those characteristics happen to be. I will not attempt to make the word "kind" (along with "event" and "outcome") entirely precise—perhaps because I do not know how. I will try to suggest some possible shades of meaning by considering "kinds" of events in the spectrum of concrete and abstract events suggested above.

Repetition of events

In a certain sense there is no coin flip "of the same kind" as that at the beginning of the 2022 Super Bowl. Each concrete past event is "one of a kind." On the other hand, it is natural to say the coin flips in previous Super Bowls are of the same kind or even that all coin flips are of the same kind. Thus, we see there is a large degree of imprecision and some kind of assumed context here. For the purposes of studying probability, and consideration of the kinds of events suitable for consideration in probability, being able to at least imagine large numbers of events of the same kind is relatively important. This is illustrated, in part, by the discussion of frequency stabilization above. Specific future coin flips, like that at the beginning of the 2030 Super Bowl, or one that might take place to determine my dinner plans, constitute a collection of coin flips which, on the face of it, may be relatively large but not infinite. As suggested in the discussion of frequency stabilization above, the consideration of collections of infinitely many events of the same kind,

at least abstractly, is relatively important in the study of probability. In particular, the consideration of infinite sequences of events of the same kind, like a sequence of infinitely many coin flips, is important. It is interesting that while humans have no experience[15] whatsoever (as far as I know) of anything comparable to the mathematically infinite sequence (for example an infinite sequence of events) the consideration of an infinite sequence of events appears to be much more intuntively appealing than other infinite collections of events. For example one might abstractly consider a collection of coin flips with one coin flip associated with each real number. Indeed, as we will see below, uncountable collections of events and outcomes are inherently contemplated in the study of probability and statistics (in certain circumstances). In these instances, it may not be immediately obvious how the notion of a limit of ratios discussed above generalizes or the meaning of frequency stabilization.

I will make two other comments about kinds (of events). Among events suitable for consideration in probability, it is natural to designate many classes of events as different (kinds). I simply give some examples:

1. A coin flip is (usually considered) a different kind of event from a roll of a die.

2. The consecutive flipping of three coins is a different kind of event than a single coin flip.

3. Rolling a die with non-standard face values, say a hexahedral (six sided cube) die with two faces marked "2" and four faces marked "6" is a different kind of event than rolling a standard die.

I hope this idea of kinds of events is intuitively clear and you "get it."

Exercise 0.7.8 List some (interesting) events possibly suitable for consideration in the study of probability. Can you list events that might be excluded from consideration in the study of probability?

Finally, I remark that the terminology used in many books on probability (and statistics) strikes me as extremely sloppy and poorly constructed. This is part of the motivation for the discussion above. Some of the words used,

[15]Have you ever flipped a coin infinitely many times? Can you prove you never could? Have you ever watched someone flip a coin infinitely many times? Can you prove you never could?

nevertheless, are quite suggestive (of something). In particular, once the notion of a sequence of events (of the same kind) with well-defined possible outcomes is embraced, then the terms "experiment" and "trial" to refer to one event in the sequence are common.

Here is a short outline of the kinds events discussed above and a short outline of the corresponding outcomes.

Events

1. concrete (past)
2. abstract
   (a) future
   (b) hypothetical

Outcomes

1. concrete (observed)
2. abstract
   (a) future (unknown)
   (b) hypothetical
      i. realistic (single outcome)
      ii. specified/compound (multiple outcomes)

Naturally, within this framework further distinctions among kinds may be made, for example, coin flips and dice rolls may be distinguished as different kinds of events.

Here is an attempt to list some features making a particular kind of event suitable for consideration in probability:

(i) The event should have associated with it a collection of well-defined possible outcomes. Any concrete example (trial) should have resulted in a single one of these possible outcomes.

(ii) The characteristics of the event should be adequately defined so that large numbers (and infinite numbers) of events of the same kind (or trials of the event) may be contemplated (abstractly).

Chapter 1

Lecture 1: Sets, Functions, and Measures

There are many references for sets. I have taken these notes from an appendix of Alex Farmer's book [2] on probability and statistics.

1.1 Sets

1.1.1 The main thing: power set

Given a set S, the **power set** of S is the set of all subsets of S.

Example 8 $S = \{0, 1, 2, 3\}$.

$$\begin{aligned}\wp(S) = \{ \quad & \phi, \{0\}, \{1\}, \{2\}, \{3\}, \\ & \{0,1\}, \{0,2\}, \{0,3\}, \\ & \{1,2\}, \{1,3\}, \{2,3\}, \\ & \{0,1,2\}, \{0,1,3\}, \\ & \{0,2,3\}, \{1,2,3\}, \\ & \{0,1,2,3\} \quad \}.\end{aligned}$$

The symbol ϕ represents the **empty set**.

Exercise 1.1.1 If a set has n elements, how many elements are there in the power set $\wp(S)$?

1.1.2 Other things about sets...

This is mostly notation. Hopefully most of it should be familiar.

1. Writing $x \in A$ means "A is a set and x is **an element of** A."

2. Writing $A \subset S$ means "the set A is a **subset** of the set S." That is, each element of A is an element of S.

3. (**equality** of sets) In order to show $A = B$ for sets A and B means to establish the two implications

$$\begin{array}{lll} x \in A & \Longrightarrow & x \in B, \text{ and} \\ x \in B & \Longrightarrow & x \in A. \end{array}$$

4. Given $A, B \subset S$, the **union** of A and B is the set

$$A \cup B = \{x \in S : x \in A \text{ or } x \in B\}.$$

5. The **intersection** of A and B is the set

$$A \cap B = \{x \in S : x \in A \text{ and } x \in B\}.$$

6. The **complement** of a set $A \subset S$ is

$$S \backslash A = \{x \in S : x \notin A\}.$$

   This complement of A is sometimes denoted A^c when the set S is understood (from the context).

7. De Morgan's laws state

$$\begin{array}{l} (A \cup B)^c = A^c \cap B^c, \text{ and} \\ (A \cap B)^c = A^c \cup B^c. \end{array}$$

8. The **Cartesian product** of two sets A and B is

$$A \times B = \{(x, y) : x \in A \text{ and } y \in B\}.$$

An element (a, b) of the Cartesian product $A \times B$ is called an **ordered pair**.

Exercise 1.1.2 If A has n elements and B has m elements, how many ordered pairs/elements are there in $A \times B$.

9. The **cardinality** of a set is the number of elements in the set when the number of elements in the set is a finite number. If there are n elements in a set A, then one writes $\#A = n$ and/or $\#A < \infty$. If there is a one-to-one correspondence $\gamma : A \to \mathbb{N}$ between the elements in a set A and the natural numbers $\mathbb{N} = \{1, 2, 3, \ldots\}$, then A is said to be **countably infinite** and one writes $\#A = \aleph_0$. If a set A has neither a finite number of elements nor a countably infinite number of elements, then one says the set A is **uncountable**. If there is a one-to-one correspondence $\gamma : A \to \mathbb{R}$ between the elements in a set A and the real numbers $\mathbb{R}$, then A is uncountable and one writes $\#A = \aleph_1$.

10. A generalization of the concept of an ordered pair is a **sequence**. There are finite sequences $(x_1, x_2, \ldots, x_n)$ which are also called n-tuples. There are infinite sequences $(x_1, x_2, x_3, \ldots)$ with countably many entries.

Exercise 1.1.3 Find a one-to-one correspondence between the set

$$A = \left\{ \ (a_1, a_2, a_3, \ldots) : a_j \in \{0, 1\}, \text{ and } \#\{j : a_j = 1\} < \infty \ \right\}$$

and $\mathbb{N}$.

1.2 Functions

1.2.1 The main thing: Definition

Given two sets X and Y, a **function from X to Y** is a rule or correspondence assigning to each $x \in X$ a unique $y \in Y$.

Example 9 $S = \{h, t\}$, $B = \{0, 1\}$.

$$\beta(h) = 1 \qquad \text{and} \qquad \beta(t) = 0. \tag{1.1}$$

In the situation described in Example 9 we can say a **function β from S to B** is determined by (1.1). We can also write

$$\beta : S \to B.$$

Conditions determining/defining the same function may be expressed in the following alternative ways:

mapping values.

$$\begin{aligned} h &\mapsto 1 \qquad\qquad (\text{read } \text{“}h \text{ maps to } 1\text{”}) \\ t &\mapsto 0. \end{aligned}$$

a function as a set.

$$\{\ (h, 1),\ (t, 0)\ \}. \tag{1.2}$$

The set in (1.2) is called the **graph** of the function.

1.2.2 Other things about functions...

If f is a function from X to Y, which as suggested above we will usually express by writing simply $f : X \to Y$, then the set X is called the **domain** of the function, and the set Y is called the **codomain** of the function. **Values of the function**, or the **value** of the function at x, may be denoted by $f(x)$. In this case, we sometimes refer to the domain element x as the **argument** of f, and f is said to be **evaluated** at x to obtain the value $f(x)$. The set of all function values

$$\{f(x) : x \in X\}$$

is called the **range** of the function. Note that the range of $f : X \to Y$ is a subset of the codomain Y. The range is also sometimes denoted by $f(X)$, and more generally if $A \subset X$, then

$$f(A) = \{f(x) : x \in A\} \subset Y.$$

The set $f(A) \subset Y$ is called the **image** of the set A.

If $f(X) = Y$, then the function f is said to be **onto** or **surjective**.

Notice that there is not always a "rule" assigning to a value $y \in Y$ of f an element $x \in X$ for which $f(x) = y$.

Exercise 1.2.1 Give an example of a surjective function $f : X \to Y$ for which there is no rule $g : Y \to X$ such that

$$g(f(x)) = x \qquad \text{for every } x \in X. \tag{1.3}$$

Notice that $f(A)$ is a set, when the argument of f is a set. One can think of this as primarily a kind of set notation. We extend this kind of set notation even when there is no rule $g : Y \to X$ assigning a unique x to each y as in Exercise 1.2.1. Specifically, given a set $E \subset Y$ we write

$$f^{-1}(E) = \{x \in X : f(x) \in E\}. \tag{1.4}$$

This set is called the **preimage** of E. Note carefully that the set $f^{-1}(E)$ always makes perfectly good sense precisely defined by (1.4) even when there is no function $g : E \to X$ for which

$$g(f(x)) = x \qquad \text{for every } x \in f^{-1}(E).$$

There are cases in which $f(X) = E$ and there exists a function $g : E \to X$ satisfying

$$g(f(x)) = x \qquad \text{for every } x \in X. \tag{1.5}$$

Exercise 1.2.2 If $f : X \to Y$ and $f(X) = E \subset Y$, then f is said to **map** X **onto** E. If this is the case, what is

$$f^{-1}(E)?$$

In cases in which $f : X \to Y$ is surjective and there is a function $g : Y \to X$ for which (1.3) holds, we say f is **invertible** and the function g is the **inverse** of f. In this case, we write $g = f^{-1}$ (using the same notation for the function we used in reference to the preimage set).

Exercise 1.2.3 Show that if f is invertible, then the following condition holds:

$$\text{If } x_1, x_2 \in X \text{ with } f(x_1) = f(x_2), \text{ then } x_1 = x_2. \tag{1.6}$$

The condition (1.6) may hold even when $f : X \to Y$ is not surjective. Any time a function $f : X \to Y$ satisfies condition (1.6) the function f is said to be **one-to-one** or **injective**.

Exercise 1.2.4 Show that if f is injective, then the function $\phi : X \to f(X)$ by $\phi(x) = f(x)$ for all $x \in X$ is invertible.

1.2.3 Review and fun fact(s)

Given two sets X and Y, a function $f : X \to Y$ is a rule or correspondence assigning to each $x \in X$ a unique $y \in Y$.

There is some convenient terminology and there are some convenient notations associated with functions. Among the terms that are useful to know and be able to use correctly are domain, codomain, range, image, preimage, one-to-one, onto, inverse, and invertible.

The power set of a set X is sometimes denoted 2^X. More generally, the set of all functions $f : X \to Y$ is denoted Y^X.

Exercise 1.2.5 Can you explain why the notation 2^X makes sense in terms of the notation Y^X? Hint: In set theory $2 = \{0, 1\}$.

1.2.4 Probability and Statistics (preview)

In probability and statistics, and in this course, it may be said that one **uses sets to model the outcomes of events**.[1] Measures which are discussed in the sections below are also important. In all cases, a particular phenomenon of note takes place: Different symbols are (traditionally) used. When discussing sets in the theory of sets and functions in the theory of functions, it is very common to encounter an element x denoted by the symbol "x." When discussing sets in probability, the symbol "x" is almost never used to denote an element of a set. This may be strange, but traditions are sometimes confusing and cannot always be explained. In principle, which symbols are used does not really matter as long as one can keep the concepts/ideas clear in one's mind. It may be anticipated, however, that this "harmless" rearrangement of symbols may tend to result in potentially great difficulty of some sort or another. Among other things, this section provides an opportunity to "warm up" to this potential difficulty.

Exercise 1.2.6 The elements of the base sets under consideration in probability and statistics are often denoted by the symbol "ω" (instead of the symbol "x"). Go through section 1.1.2 above and rewrite/replace each use of the symbol "x" with the symbol "ω."

You may be interested and amused[2] to note that the symbol "x" is usually used to denote a function in probability and statistics. Specifically, you may recall that in calculus, and in many other contexts where functions and their values are discussed, the symbols and notation "f" and "$f(x)$" are used. In probability, these symbols are usually replaced by something like "x" and "$x(\omega)$."

This course is intended to be about two closely related subjects: probability and statistics.

[1] It may be said that the material in this chapter is so important that it will be covered twice (or over and over again). There are other reasons for the repetition as well. This section in particular may be taken as an opportunity to "jump forward" a little bit. That is to say the exact same material will be covered again later, some of it certainly in Lecture 2, so you may wish to skip this section now and move directly to section 1.2.5 and the introduction to measures below (without any reference to probability and statistics).

[2] And you may also be enlightened to the upcoming difficulty anticipated with the traditional notation of probability and statistics.

The subject of probability may be divided into two distinct parts: mathematical probability and aplications of probability.

Mathematical probability is the study of certain measures (for example baby measures $\pi : \wp(S) \to \mathbb{R}$) satisfying the **additional condition**

$$\pi(S) = 1. \tag{1.7}$$

Modeling using sets

An "event" is, for us, very nearly an undefined term, but nevertheless it is good to be able to recognize one when it is observed. The "flipping of a coin" is considered an "event." The "rolling of a die" and the "rolling of a pair of dice" are "events." An "event" is by nature something that happens in the real world and consequently can be expected to involve some degree, and probably a large degree, of complexity and ambiguity. This can be said of the examples we have given. Nevertheless, we associate something relatively simple and definite with each of these example "events," and at least for the time being let us make a restriction suitable to these examples:

> We start by considering "events" for which we believe we can identify a collection of definite and distinct "possible outcomes," and this collection of "possible outcomes" is finite.

For a "coin flip" the usual "outcomes" are "heads" and "tails," and there are two of them. For "rolling a six sided die" the usual outcomes are "one," "two," "three," "four," "five," and "six," and there are six of them. In this case of "rolling a six sided die" there is another obvious alternative: We can take the outcomes "odd" and "even," in which case there will be two "possible outcomes." Thus, we may be told, or get to decide, some details of what is intended by the "possible outcomes" of an "event."

Exercise 1.2.7 Consider the event of "rolling a pair of six sided dice." Identify with words and count the finitely many "possible outcomes" associated with the following descriptions of them.

(a) The sum of the roll.

(b) The two values showing. (Order doesn't count.)

(c) The first value and the second value. (Here one might imagine the dice have different colors, and order does count.)

I am now going to add to our assumption of finitely many (designated and definite) "possible outcomes" described above a somewhat subtle distinction among (real world) "events." This is a distinction between "past events" and "future events." Certain "coin flips" have already taken place in the past and the "outcome" has been observed. Other "coin flips" have not yet taken place, and some may never take place. These "future coin flips" or "abstract coin flips" only exist in the minds of certain humans, and in that sense they are very different from the "concrete coin flips" that have taken place in the past. Nevertheless, we consider both kinds of coin flips "real world events," and I am now going to introduce mathematical models which may be compared to them.

Concrete events

A concrete event has a single well-defined outcome. In principle, we know this outcome and we can associate it with a well-defined (mathematical) symbol. Specifically, given a real world "event" which falls into a family of similar events for which there are exactly n outcomes, we introduce a **base symbol set**

$$S = \{\omega_1, \omega_2, \ldots, \omega_n\}.$$

We associate the specific known outcome of the "concrete event" under consideration to one of the symbols ω_j in the set S. The key property of a "concrete event" is that its outcome can be associated to one and only one element in S, **and** the particular element used to model the outcome of the event is known. For example, I can model the outcomes of "concrete coin flips" using the base set $S = \{h, t\}$. If I have "flipped a coin" five minutes ago, and it came up "heads," then I model this outcome by the symbol $h \in S$. There is nothing particularly tricky about this. We can say the symbol set S models the collection of all **concrete outcomes**.

Abstract events

Usually, less can be said about "abstract events." One thing that can be said is that it is still assumed an "abstract event" can only have a single unique "outcome." It is (perhaps) not known what that outcome wlll be. Nevertheless, I would like to model all the "possible outcomes" not directly associated with a particular abstract event in any way, but associated with the family of similar events like the abstract event under consideration, that is the events of the same "kind." For example, if I am going to "flip a coin" in five minutes (one can say that "abstract event" is "on the schedule") then I can consider similar "events," that is to say "flipping coins." Some of those events are concrete, and I have a symbol set $S = \{h, t\}$ to model their "concrete outcomes." For "abstract events" that are similar, the same symbols can be used to model the "concrete outcomes," i.e., of the same kind, **once they are known** and become concrete. I would like, however, a mathematical model for those "outcomes" **before they are known**. I want a model not for "concrete outcomes" but rather for "abstract outcomes" or "hypothetical outcomes." For this, I use a particular subset of the power set $\wp(S)$ of the symbol set, namely I use

$$\Omega = \{A \in \wp(S) : \#A = 1\}.$$

consisting of all singleton[3] sets in $\wp(S)$. I call this set the (hypothetical) **outcome model**. Obviously, the base symbol set and the outcome model are closely related. If

$$S = \{\omega_1, \omega_2, \omega_3, \ldots, \omega_n\},$$

then

$$\Omega = \{\ \{\omega_1\},\ \{\omega_2\},\ \{\omega_3\},\ \ldots,\ \{\omega_n\}\ \}.$$

In the particular example of a "single coin toss" where $S = \{h, t\}$, we have

$$\Omega = \{\ \{h\},\ \{t\}\ \}.$$

Note carefully that S and Ω contain different kinds of elements and model different kinds of real world outcomes.

[3] A set A is said to be a **singleton** if $\#A = 1$.

Aside from the distinction between "concrete outcomes" and "abstract outcomes" there are other reasons to introduce the set Ω. In particular, it will be quite important for us to consider other subsets of S containing more than one element to model yet other kinds of "abstract outcomes," and the use of Ω will make this extension consistent.

In summary, the (abstract) outcome of flipping a coin (abstractly) and having it come up "heads" is modeled by the set $\{h\}$, and the outcome of flipping a coin and having it come up "tails" is modeled by the set $\{t\}$. The two possible outcomes of an abstract coin flip are modeled using the set Ω.

The essence of **mathematical modeling** is **comparison** of something observed (in the world, in nature, however you wish to express it) or what is often called a **natural phenomenon** with a mathematical construction, ultimately involving sets, which fundamentally only exist in the minds of (certain) people. Mathematical modeling is the comparison of (mathematical) ideas to real world phenomena.

The modeling used in probability is for the most part quite elementary and may be considered rather uninteresting.[4] There are several other hallmarks of mathematical modeling which may be considered difficult to even perceive in the modeling of outcomes by sets as described above. Here are some of those hallmarks:

(a) The comparison of the (mathematical) model to the observed phenomenon is usually **interesting**. This is partially because there is usually an emphasis on **differences** in the comparison.

(b) Mathematical modeling usually results in **prediction**.

(c) Mathematical modeling usually results in **surprises** or unexpected aspects of the observation and comparison.

(d) As a consequence, mathematical models are usually **based on understanding** of the underlying phenomenon and lead to **further or refined understanding**.

(e) The basic standard is agreement of observations with the model **within the tolerance of measurement**. That is to say, if you

[4]This is not a very good excuse for doing it poorly or incorrectly. As we shall see there are other motivations at work for that.

> have the capability to measure differences in the prediction of your model with the observed phenomenon, then your job as a mathematical modeler is not done. There is something else to understand and/or incoroprate into a model.

A more typical example of mathematical modeling is the comparison of solutions of initial value problems for differential equations (using Newton's laws of motion) to the motions of physical masses. In this case, the hallmarks mentioned above are easy to see. One can predict, for example, that with respect to some assumed fixed frame of reference the orbits of planets should be approximately the shape of ellipses. This prediction was compared historically to observations of Kepler. It is something of a surprise that among the mathematical models for orbits of planets there are also parabolic orbits suggesting a "single pass" for some physical orbiting masses. This mathematical conclusion compares/applies relatively accurately to the motions of certain comets. By focusing on the differences between the predictions of Newton's model and the observed motion of planets, the possibility and significance of the relativistic corrections (contained in a different model largely due to Albert Einstein) were/are better appreciated. Many people felt that Newton expressed deep insight and understanding concerning the motion of masses they could also embrace. The work of Einstein represents for many some kind of clear refinement in this understanding. At least some people view these aspects of the mathematical modeling of the motion of masses as altogether quite interesting.

If we look for these hallmarks in the modeling of the outcomes of a coin toss described above, the situation is less clear. Nevertheless, hopefully some basis of comparison (and distinction) is obvious. Perhaps the comparison here is not terribly interesting. One can focus on differences: Perhaps the outcome of a coin toss is that the coin comes to rest on edge—neither heads nor tails. This, we might say, is "not included in the model," and quite honestly, while such a thing may be an apparent physical possibility, it is one that receives (and perhaps merits) little attention in regard to most coin flips. Reflection suggests, however, that there is some potentially relevant and interesting distinction to be made. There is perhaps something we have missed. Namely, there is a natural distinction among **events** we have not taken into account. Let us postpone this refinement until a little later. (You can think about

what it might be.)

The basic prediction of this model (if it can be called a prediction) is that only two outcomes can be observed. Is there much more (interesting) to say about that? The actual physical process (e.g., modeled by Newtonian mechanics) taking place in a specific coin toss and leading to a particular outcome (and the possible prediction of such an outcome) is very specifically not something contemplated/included in our model. There is no real expectation of much real understanding or further refined understanding. We "understand" that there are two possible outcomes. We are (presumably) satisfied with that, and we are going to go in a different direction. From this limited perspective (either "heads" or "tails"), setting aside the basic nature of mathematical modeling to a certain extent, our comparison of outcomes to sets of symbols is exact and precise—completely accurate.

Having explained in some detail the actual nature and content of the mathematical modeling of outcomes using sets, I suggest we give names to the objects (both in the real world and the mathematical world) under consideration.[5] We have already discussed briefly "event" and "outcome." These are "real world" objects/entities. Most textbooks proceed as if an "event," the "outcome" of an event, and the set (or sets) used to model one or both of them are all the same thing. I am suggesting something very different. An "event" is neither the "outcome" of the event nor is it (certainly) a set. I think the terms "procedure," "trial," and "experiment" may be used essentially synonomously with "event." Notice there may also be an outcome of a procedure, an outcome of a trial, or an outcome of an experiment.

Turning to the mathematical objects, let us call the set S the **symbol set** or base set for the modeling. We have already discussed the power set $\wp(S)$ of the base set, which is really the set we wish to use for modeling outcomes. The entire power set $\wp(S)$ will be used in some capacity, but we distinguish a certain subset $\Omega \subset \wp(S)$ called the **base**

[5] In textbooks on probability and statistics, one tends to find a jumble of words including "sample space," "procedure," "experiment," "outcome space," "outcome," "event," and others. The discussion to follow is partially intended to suggest some reasonable uses of some of these (and some other) words.

model set, namely,

$$\Omega = \{A \in \wp(S) : \#A = 1\}.$$

Warning: Many textbooks "identify," that is to say "make little or no distinction among" (or otherwise confuse) the sets S, $\wp(S)$, and Ω. In addition, these sets are usually confused with "events," "outcomes," "samples," "sample spaces," and other nominally real world entities/objects. Again, I am suggesting something very different. Most textbooks "define" an "event" to be a subset of some set—usually something like what I have called the symbol set, though often denoted by Ω and called the "sample set" or "outcome space."

In summary so far, let us say we are considering an "event." This "event" is a real world entity and consequently can be expected to involve a significant level of complexity. Let us assume, however, that this is an "event" for which there are (under any circumstances) finitely many concrete outcomes corresponding to the elements ω in a set S with $\#S = n$. It may be the case that the event has already taken place, has already been observed, and we already know the specific outcome corresponding to one of the symbols in the base set of symbols S. Of greater interest is the situation in which we consider (abstractly) the same "kind" of event with abstract outcome modeled by a set in the base model set Ω.

At this point we introduce a crucial generalization: We consider abstractly any arbitrary set $C \in \wp(S)$, and we consider such a set to model the **abstract possibility an event (will) have outcome corresponding to an element of** C. Do not miss the subtlety of what is being done here. Nomininally, there is a kind of contradiction. An abstract event is required to have a single outcome modeled by a singleton in Ω; when we consider $C \in \wp(S)\backslash\Omega$, we are not modeling a real world outcome but something purely psychological—something purely in the mind of a human being (or at least only conceivably there). We are not modeling the possible outcome of an event but the possibility of a certain kind of outcome of the event.

To better appreciate the distinction between modeling a possible outcome and modeling the possibility of an outcome, let us consider modeling a different more complicated kind of event. A coin is flipped three

times, and the first, second, and third outcomes (heads or tails) are recorded. We model the outcome of the "three flips" by the (symbol) set

$$S = \{\ hhh,\ hht,\ hth,\ htt,\ thh,\ tht,\ tth,\ ttt\ \}.$$

We could use, and it might be more natural to use,

$$S = \{h,t\} \times \{h,t\} \times \{h,t\} = \{h,t\}^3 = \{(a,b,c) : a,b,c \in \{h,t\}\}.$$

Either way, there are eight possible concrete outcomes corresponding to the eight elements ω in the set S. We have then the abstract outcome set

$$\Omega = \big\{\ \{hhh\},\ \{hht\},\ \{hth\},\ \{htt\},\ \{thh\},\ \{tht\},\ \{tth\},\ \{ttt\}\ \big\}.$$

This models, on the one hand, the possible outcomes of a future or otherwise abstract event involving three coin flips.

Here are some examples of specified compound outcomes:

(a) For the specified outcome

"among the three coin flips, at least one comes up heads"

we can use the model set

$$\{hhh, hht, hth, htt, thh, tht, tth\}.$$

Note, we are not taking this set as a different/alternative symbol set. Rather, we consider $A = \{hhh, hht, hth, htt, thh, tht, tth\} \in \wp(S)$ as modeling a different kind of outcome.

(b) For the specified outcome

"among the three coin flips, exactly one comes up heads"

we can use the model set

$$\{htt, tht, tth\}.$$

(c) For the specified outcome

"the first coin flip (of the three) comes up heads"

we can use the model set

$$\{hhh, hht, hth, htt\}.$$

These examples highlight the difference between the outcome of an event and a specified outcome of an event. They also suggest the natural generalization of our modeling mentioned above. It may be noted that with this generalization, the singletons in the model outcome set Ω become somewhat subtly ambiguous. On the one hand, the singleton $\{hhh\}$ models the **possible** abstract outcome that each of the three flips results in "heads." On the other hand, we are now using the same set to model a **hypothetical** abstract outcome, namely, that each of the three flips results in "heads." When we apply these model descriptions to a singleton set like $\{hhh\}$, it is very difficult to perceive the difference. If, however, we consider

$$C = \{hhh, hht, hth, htt\},$$

it is more or less clear that this does not correspond to any known **possible** outcome (abstract or otherwise) of flipping a coin three times. In short, $C \notin \Omega$ the set we use to model abstract outcomes. Changing perspectives, however, one can specify the **hypothetical** outcome that "the first flip comes up heads." If a coin is flipped three times and the first flip comes up heads, and yet the (sub)outcomes of the second and third flips are observed, then the overall outcome will only correspond to one of the symbols in C. The kind of outcome modeled by $C \notin \Omega$ in this case is more properly called a "hypothetical compound outcome of an event" or a "specified outcome of an event."

As a final exercise to solidify the model suggested above, let us consider the very easy case of a single coin flip. A description (in words) of one possible collection of outcomes is

"the outcome of a coin flip is heads or tails." (1.8)

This collection of outcomes may be modeled by the set

$$C = S = \{h, t\} = \{h\} \cup \{t\}. \tag{1.9}$$

Note carefully, that $C \notin \Omega$.

Exercise 1.2.8 What specified collection of outcomes for tossing a coin does the empty set $\phi \in \wp(\{h, t\})$ model?

Exercise 1.2.9 What relevance does the outcome "heads" have in relation to the coin flip at the beginning of the 2022 Super Bowl? See the footnote in Lecture 0 (section 0.7) of the Introduction above.

1.2.5 Review and fun stuff

In closing, the main construction we need concerning sets is that of the **power set** of a set A, which is a set consisting of all the subsets of A including the empty set ϕ and the entire set A. The other subsets of A are called **proper subsets**. The reason the power set is important is because later/soon we will discuss what it means to **measure** the subsets of A.

Both the outcomes of individual events and specified collections of outcomes of events may be modeled by subsets of a **symbol set** S. Very often the singleton sets of a symbol set S correspond to the possible outcomes of a single event, and very often the collection of these singleton sets is called the **outcome set** Ω. More general subsets in $C \in \wp(S)$ are used to model specified collections of outcomes. These sets $C \in \wp(S)$ or $C \subset S$ are called **specification sets**.

Exercise 1.2.10 What is the difference between an outcome and a specified outcome? What is the relation between specified outcomes and outcomes?

Perhaps it will be fun for you to learn the Greek alphabet which is used

to provide symbols for many things in mathematics.

α	A	alpha	
β	B	beta	
γ	Γ	gamma	
δ	Δ	delta	
ϵ	E	epsilon	
ζ	Z	zeta	(pronounced by most people as "z$\bar{a}$tuh")
η	H	eta	("$\bar{a}$tuh" but by most people as "eat-uh")
θ	Θ	theta	
ι	I	iota	("$\bar{e}$-oh-tuh", "yoda," or "$\bar{i}$-oh-tuh")
κ	K	kappa	
λ	Λ	lambda	
μ	M	mu	"m$\tilde{e}$w"
ν	N	nu	"new"
ξ	Ξ	xi	"ksee" or (by some) "ks$\bar{i}$"
o	O	omicron	
π	Π	pi	"pee" or "pie," that is to say "p$\bar{i}$"
ρ	P	rho	"row"
σ	Σ	sigma	
τ	T	tau	
υ	Υ	upsilon	"oops-il-on"
ϕ	Φ	phi	"fee" or (by some) "f$\bar{i}$"
χ	X	chi	"key" or (by some) "k$\bar{i}$"
ψ	Ψ	psi	"psee" or (by some) "ps$\bar{i}$"
ω	Ω	omega	

Exercise 1.2.11 What Greek symbol(s) are used in the discussion of sets above?

1.2.6 (Strange) Mathematical stuff

The simplest set is the empty set ϕ which is the set containing no elements. Sets with a single element are called singletons. Sets with two elements have a special name too; they are called **pairs**, but they are not ordered pairs.

Informally, one may talk about sets or collections of real world objects like cars or people or balls our outcomes of coin tosses, but from the mathematical perspective sets can only contain one kind of object: other sets. Thus, one can consider sets of sets, sets of ordered pairs, sets of numbers, or sets of functions, but when one does so each of the elements under consideration should ultimately be expressed as a set. That is, an ordered pair is some kind of set, a number is some kind of set, and a function is some kind of set. And (mathematically) there is no such thing as a set of events or outcomes. Ultimately, all sets trace their form back to the simplest set containing a set, namely the singleton $\{\phi\}$.

1.3 Measure(s)

Very generally, a measure is a function assigning a number to subsets of a set (and satisfying certain kinds of rules). For us, it will be (mostly) enough to consider measures that assign non-negative real numbers to sets. The assignment of **length** to intervals $I \subset \mathbb{R}$ in the real line for which there are real numbers a and b with

$$a \leq x \leq b \qquad \text{for every } x \in I,$$

i.e., bounded intervals, gives a pretty good idea of what is meant by a measure. You can ask yourself if and how the general properties of a measure described below might apply to the length of bounded intervals.

1.3.1 The main thing: additivity

We initially restrict to measures defined on the power set of a set with finitely many elements.

Definition 1 Given S with $\# S < \infty$, a **baby measure** on S is a function

$$\mu : \wp(S) \to \mathbb{R}$$

(assigning real numbers to subsets of S) satisfying the following properties:

(0) $\mu(A) \geq 0$ for every $A \subset S$.

(i) $\mu(\phi) = 0$.

(ii) If $A, B \subset S$, and $A \cap B = \phi$, i.e., A and B are disjoint, then

$$\mu(A \cup B) = \mu(A) + \mu(B).$$

Property **(ii)** is called **additivity**, and μ is said to **measure the subsets** of S.

Theorem 3 If $\mu : \wp(S) \to \mathbb{R}$ is a baby measure and $A \subset S$, then

$$\mu(A^c) = \mu(S) - \mu(A).$$

Theorem 4 If $\mu : \wp(S) \to \mathbb{R}$ is a baby measure and $A_1, A_2, \ldots, A_k \subset S$ with

$$A_i \cap A_j = \phi \qquad \text{when } i \neq j, \tag{1.10}$$

then

$$\mu\left(\bigcup_{j=1}^{k} A_j\right) = \sum_{j=1}^{k} \mu(A_j). \tag{1.11}$$

Sets $A_1, A_2, \ldots, A_k$ satisfying (1.10) are said to be **pairwise disjoint**.

The property concluded in Theorem 4 is called **finite additivity**.

Whenever we have a measure $\mu : \wp(S) \to [0, \infty)$, the underlying base set of elements is called the **measure space**. Note that the underlying base set S is not the domain of μ. The domain of μ in this case, is called the **algebra** of the measure. Note that there is a kind of "algebra" in $\wp(S)$ with respect to the operations "union" and "intersection."

Exercise 1.3.1 What can you say about the following "algebraic" properties of union and intersection in $\wp(S)$ when S is a set with finitely many elements?

(a) closure

(b) associative (property)

(c) commutative (property)

(d) distributive (property)

Hint(s): Think in terms of the given properties for the arithmetic of addition and multiplication in a set with which you are familiar, for example, the natural numbers

$$\mathbb{N} = \{1, 2, 3, \ldots\}$$

where "$\mathbb{N}$ is closed under addition" means given $n, m \in \mathbb{N}$ the sum $n + m$ is also in $\mathbb{N}$, and the distributive property (of multiplication across addition) says

$$a(n + m) = an + am \qquad \text{for all } a, n, m \in \mathbb{N}.$$

A new measure obtained by scaling

Theorem 5 If $\mu : \wp(S) \to \mathbb{R}$ is a baby measure and $c > 0$ is a (positive) real number, then $\nu : \wp(S) \to \mathbb{R}$ by

$$\nu(A) = c\mu(A) \qquad \text{for every } A \subset S$$

is a (new) baby measure. (The measure ν is really new if $c \neq 1$.)

A new measure obtained by restriction

As above, let S be a set with finitely many elements, say $\#S = n < \infty$, and let $\mu : \wp(S) \to [0, \infty)$ be a baby measure. Now say $T \subset S$. We consider

$$r : \wp(T) \to [0, \infty) \qquad \text{by} \qquad r(A) = \mu(A).$$

The function r is called the **restriction** of μ to $\wp(T)$ and may be denoted by

$$r = \mu\big|_{\wp(T)}.$$

Exercise 1.3.2 Prove the restriction measure r is a measure.

Solution:

(i) $r(A) = \mu(A) \geq 0$ for every $A \subset S$ (because μ is a measure).

(ii) $r(\phi)\mu(\phi) = 0$ (because μ is a measure).

(iii) If $A, B \subset T$, and $A \cap B = \phi$, then $A, B \subset S$ as well. Therefore,

$$\begin{aligned} r(A \cup B) &= \mu(A \cup B) \\ &= \mu(A) + \mu(B) \qquad \text{(because } \mu \text{ is a measure)} \\ &= r(A) + r(B). \end{aligned}$$

Therefore, r is additive and

r is a measure. $\square$

Example 10 Consider

$$S = \{\text{one}, \text{two}, \text{three}, \text{four}, \text{five}, \text{six}\}$$

with $\mu : \wp(S) \to [0,1]$ by $\mu(\ \{x\}\ ) = 1/6$ for all $x \in S$. The function μ is a measure (Exercise 1.3.2).

Consider also

$$T = \{\text{two}, \text{four}, \text{six}\}.$$

Then

$$r = \mu\big|_{\wp(T)} \qquad \text{is a measure,}$$

and is the restriction measure of μ with respect to the subset (measure space) T.

Exercise 1.3.3 Let μ denote the measure dicsussed in Example 10. Find the following:

(a) $\mu(S)$

(b) $r(T)$

(c) $c > 0$ so that the scaled measure $\nu = cr$ has $\nu(T) = 1$.

A new measure obtained by expansion

As above, let S be a set with finitely many elements, say $\#S = n < \infty$, and let $T \subset S$. Now say we have a measure $\sigma : \wp(T) \to [0, \infty)$ on the set T. Consider

$$\beta : \wp(S) \to [0, \infty) \qquad \text{by} \qquad \beta(A) = \sigma(A \cap T).$$

Exercise 1.3.4 Prove β is a measure on S and $\beta(S) = \sigma(T)$.

1.3.2 Integration and average values

Whenever one has a measure μ defined on subsets of a measure space S, it is possible to introduce a notion of **integration** of real valued functions on S. This is somewhat more general than the integration considered in elementary calculus courses. In the special case when S is a set with finitely many elements, however, integration is (at least in some ways) much simpler.

Integration

Here is the definition in this case:

Let S be a set with $\#S = n < \infty$ so that

$$S = \{x_1, x_2, \ldots, x_n\}$$

and let $\mu : \wp(S) \to [0, \infty)$ be a measure on S. Let $f : S \to \mathbb{R}$ be a real valued function with domain S. The integral of f is defined to be the sum

$$\int f = \sum_{j=1}^{n} f(x_j)\mu(\{x_j\}). \tag{1.12}$$

This integral may also be denoted by

$$\int_S f,$$

and this notation can be generalized to allow for integration over any subset $A \in \wp(S)$:

$$\int_A f = \sum_{x \in A} f(x)\mu(\{x_j\}).$$

The idea is the same: We take the value $f(x)$ of f at x and multiply it by (or "weight" it with) the measure of the singleton set $\{x\}$; we add up all the resulting weighted values. Of course, there is a symmetry here: It is perfectly valid to say the measure of the singleton $\mu(\{x_j\})$ is "weighted" by the function value $f(x_j)$, and sometimes one does wish to take this point of view.

Average value

When $f : S \to \mathbb{R}$ is a real valued function on a measure space S with finitely many elements and baby measure μ, the **average value** of f is defined to be

$$\frac{1}{\mu(S)} \int_S f = \frac{1}{\mu(S)} \sum_{x \in S} f(x)\mu(\{x\}). \tag{1.13}$$

More generally, the average value of f over any set $A \subset S$ is defined to be

$$\frac{1}{\mu(A)} \int_A f = \frac{1}{\mu(A)} \sum_{x \in A} f(x)\mu(\{x\}).$$

Notice these are **averages with respect to the measure** μ.

1.3.3 A special measure: Lebesgue measure

We started this section with a comparison of measures to the lengths of intervals of real numbers. In order to fully describe "length" considered as a measure will take a bit too much time right now, though we will address this consideration in some detail before the end of the course. For now, I simply want to point out that there can be measures on sets with infinitely many elements, like an interval in the real line. There is also a very special measure called **Lebesgue measure** for which the measure of every interval is its length. It is also true for measures on measure spaces with infinitely many elements that the presence of a measure leads to a notion of **integration of real valued functions**. In particular, the most natural generalization of the integration considered in (1.12) is called **Lebesge integration**. The integration with which you are probably familiar, however, is **Riemann integration** which is slightly different.

Let us attempt to loosely tie some of these integrations together. Denote Lebesgue measure by length. Then Riemann integration of a function $f : [a, b] \to \mathbb{R}$ with domain a closed interval determined by $a, b \in \mathbb{R}$ with $a < b$, when it is well-defined, can be expressed as

$$\begin{aligned} \int_a^b f(x)\,dx &= \lim_{\|\mathcal{P}\|\to 0} \sum_{j=0}^{k-1} f(x_j^*)(x_{j+1} - x_j) \\ &= \lim_{\|\mathcal{P}\|\to 0} \sum_{j=0}^{k-1} f(x_j^*)\,\text{length}([x_{j+1} - x_j]) \end{aligned} \tag{1.14}$$

where $\mathcal{P} = \{x_0, x_1, \ldots, x_k\}$ is a **partition** of the interval $[a, b]$ with

$$a = x_0 < x_1 < \cdots < x_k = b,$$

the points x_j^* are called **evaluation points** and satisfy $x_j^* \in [x_j, x_{j+1}]$ for $j = 0, 1, 2, \ldots, k-1$, and the **norm of the partition** is defined by

$$\begin{aligned} \|\mathcal{P}\| &= \max\{x_{j+1} - x_j : j = 0, 1, \ldots, k-1\} \\ &= \max\{\text{length}([x_j, x_{j+1}]) : j = 0, 1, \ldots, k-1\}. \end{aligned}$$

The condition that the Riemann integral is well-defined and the meaning of the existence of the limit in (1.14) is as follows: There is some real number

$$L = \int_a^b f(x)\,dx$$

such that for any $\epsilon > 0$, there exists some $\delta > 0$ so that whenever $\mathcal{P}$ is a partition with $\|\mathcal{P}\| < \delta$ and x_j^* for $j = 0, 1, \ldots, k-1$ are (any) evaluation points, then

$$\left| \sum_{j=0}^{k-1} f(x_j^*)(x_{j+1} - x_j) - L \right| < \epsilon.$$

Of course, the Lebesgue integral is not defined quite like this, but the point is that a comparison may be made between the Riemann sum and the integral sum in (1.12). Specifically these two quantities are

$$\sum_{j=0}^{k-1} f(x_j^*)\,\text{length}([x_{j+1} - x_j]) \qquad \text{and} \qquad \sum_{j=1}^{n} f(x_j)\mu(\{x_j\}). \tag{1.15}$$

Starting with the expression on the right of (1.15) for integration on the set $S = \{x_1, x_2, \ldots, x_n\}$ with finitely many elements, one replaces the singleton $\{x_j\}$ with the interval $[x_j, x_{j+1}]$; the evaluation point x_j is replaced with the evaluation point x_j^*, and the measure $\mu : \wp(S) \to [0, \infty)$ is replaced with Lebesgue measure.

You may also recall the notion of the average value of a function $f : [a, b] \to \mathbb{R}$ in calculus given by

$$\frac{1}{b-a} \int_a^b f(x)\,dx$$

which may be generalized to (simply)

$$\frac{1}{\text{length}([a, b])} \int f$$

in terms of Lebesgue measure and Lebesgue integration. This kind of average agrees identically in form to the average determined by the measure $\mu : \wp(S) \to [0, \infty)$ in (1.13).

1.4 Repetition

One way to view what we are going to do now, which is repeat some of the material from the last few sections with different notation and in a slightly restricted context, is that this material is so fundamentally important that

it should be gone over twice (or three times) and understood "forward and backward" so to speak. Changing the notation gives us the opportunity to solidify the ideas. The notation we have used above, especially with elements in sets denoted by $x \in S$ and functions denoted by $f : X \to Y$ with values $f(x) \in Y$ when $x \in X$ is the standard notation associated with the subjects of calculus, set theory, analysis, measure theory, and function theory. Some of that notation should have been at least somewhat familir to you from calculus.

It is traditional, however, in probability and statistics to use **different symbols** for these same objects. Specifically, elements in sets are denoted by $\omega \in S$ and most functions of interest are real valued and typically denoted by something like $x : S \to \mathbb{R}$. A cursory comparison suggests that some confusion may arise if we do not take the time to accustom our thinking to the new notation, and this is the real reason for the repetition. We can also take this an an opportunity to introduce some additional aspects specific to probability and statistics, the definition of a **probability measure** in particular.

1.4.1 Sets

We start again with a set S. Now the elements of S will be typically denoted by the symbol ω. The **power set** $\wp(S)$ of S is, as before, the set of all subsets of S. The empty set is still denoted by ϕ and is one of the elements of the power set so we write

$$\phi \in \wp(S)$$

and more generally $\omega \in A$ to mean A is a set and ω is an **element** of A. Certain subsets of S may play a special role, these are the **singleton** sets

$$\Omega = \{ \{\omega\} : \omega \in S \}.$$

Thus, we write $\{\omega\} \in \Omega$ or $A \in \Omega$ if A is a singleton set, meaning $\#A = 1$.

Writing $A \subset B$ means for each $\omega \in A$ one has $\omega \in B$, that is A is a **subset** of B. Two sets A and B are **equal** if $A \subset B$ and $B \subset A$, that is in order to show $A = B$ one shows that each $\omega \in A$ satisfies $\omega \in B$ and each $\omega \in B$ satisfies also $\omega \in A$.

The set A is said to be a **proper subset** of the set B if $A \subset B$ but $A \neq B$. This is sometimes indicated by writing $A \subsetneq B$. In general, the expressions $A \subset B$ and $A \subseteq B$ mean precisely the same thing.

The **union** of A and B is

$$A \cup B = \{\omega : \omega \in A \text{ or } \omega \in B\}.$$

The **intersection** of A and B is

$$A \cap B = \{\omega : \omega \in A \text{ and } \omega \in B\}.$$

More generally if $\mathcal{F}$ is a **family of sets**, then the **union** of the sets in $\mathcal{F}$ is

$$\bigcup_{A \in \mathcal{F}} A = \{\omega : \omega \in A \text{ for some } A \in \mathcal{F}\}.$$

The **intersection** of the sets in $\mathcal{F}$ is

$$\bigcap_{A \in \mathcal{F}} A = \{\omega : \omega \in A \text{ for every } A \in \mathcal{F}\}.$$

The **complement** of A with respect to B is

$$B \backslash A = \{\omega \in B : \omega \notin A\}.$$

The complement of A can also be denoted

$$A^c = \{\omega : \omega \notin A\}$$

when the superset B is understood or known from the context.

De Morgan's laws are

$$\left(\bigcap_{A \in \mathcal{F}} A \right)^c = \bigcup_{A \in \mathcal{F}} A^c \qquad \text{and} \qquad \left(\bigcup_{A \in \mathcal{F}} A \right)^c = \bigcap_{A \in \mathcal{F}} A^c.$$

The **Cartesian product** of a family $\mathcal{F}$ of sets is

$$\prod_{A \in \mathcal{F}} A = \{(\omega_A)_{A \in \mathcal{F}} : \omega_A \in A\}. \tag{1.16}$$

An element in this Cartesian product might be called an $\mathcal{F}$-tuple.

The **axiom of choice** states that the Cartesian product of a family of nonempty sets $A \neq \phi$ is nonempty. That is, if it is possible to choose an element $\omega_A \in A$ for each $A \in \mathcal{F}$, then it is possible to choose elements $\omega_A \in A$ for every $A \in \mathcal{F}$ (all at the same time). (Believe it or not!)

The sets A in the product given in (1.16) are called **factors**. When there are only two factors A and B, we write $A \times B$ for the product. More generally, when there are a finite number of factors $A_1, A_2, \ldots, A_n$, we write

$$A_1 \times A_2 \times \cdots \times A_n \qquad \text{or} \qquad \prod_{j=1}^{n} A_j \tag{1.17}$$

for the product. When $A_1 = A_2 = \cdots = A_n$ are all one and the same set A, then we can express the finite Cartesian product described in (1.17) as A^n. In the case of countably many factors $A_1, A_2, A_3, \ldots$ we also write

$$A_1 \times A_2 \times A_3 \times \cdots \qquad \text{or} \qquad \prod_{j=1}^{\infty} A_j.$$

The **cardinality** of a set A is the same as that of a set B if there exists a bijection $\phi : A \to B$. Cardinality is denoted by the symbol "#."

The empty set has **cardinality zero** and is the only such set. All sets with the same cardinality as $1 = \{\phi\}$ are said to have **one element**. In this case we write $\#A = 1$ which is read "the cardinality of A is one." You can guess where this is going:

> Any set A with the same cardinality as $n = \{0, 1, 2, \ldots, n-1\}$ is said to have n **elements**, we write $\#A = n$, and
>
> any set with n elements for some $n \in \mathbb{N} = \{1, 2, 3, \ldots\}$ is said to have **finitely many elements**.

If a set A admits a bijection $\phi : A \to \mathbb{N}$, then A is said to be **countably infinite**. Sets which are countably infinite or have a finite number of elements are said to have a **countable number of elements** or to be simply **countable**.

Sets that do not have a countable number of elements are said to be **uncountable**. The natural numbers $\mathbb{N}$ have $\#\mathbb{N} = \aleph_0$ and are countable. The real numbers $\mathbb{R}$ are uncountable. The **continuum hypothesis** asserts that the smallest set which is uncountable (or that is to say any uncountable set A having the property that A may be injected into any other uncountable set) has the cardinality of $\mathbb{R}$ meaning that if A is uncountable and there exists an injection $\phi : A \to \mathbb{R}$, then $\#A = \#\mathbb{R}$. (Believe it or not!) In accord with the continuum hypothesis, we write $\#\mathbb{R} = \aleph_1$.

If the family $\mathcal{F} = \{A_1, A_2, A_3, \ldots\}$ is countable, then an $\mathcal{F}$-tuple is called a **sequence**. A sequence is equivalent to a certain function

$$a : \mathbb{N} \to \bigcup_{A \in \mathcal{F}} A \qquad \text{with} \qquad a(j) \in A_j \qquad \text{for each } j \in \mathbb{N}.$$

In this case, the value $a(j)$ of the function is often denoted by a_j for $j \in \mathbb{N}$.

1.4.2 Functions

Given sets S and R, a **function** $x : S \to R$ is a rule or correspondence which assigns to each $\omega \in S$ a unique $\xi \in R$.

The value assigned to a particular $\omega \in S$ is denoted $x(\omega)$, and we can also write

$$\omega \mapsto \xi \qquad \text{or} \qquad \omega \mapsto x(\omega).$$

The symbol "$\mapsto$" is read "maps to." The expression "$x : S \to R$" is read "x is a function from S to R." The **graph** of a function $x : S \to R$ is the set

$$\{(\omega, x(\omega)) : \omega \in S\} \subset S \times R.$$

The graph of a function determines the assignment of the function and vice-versa.

When $x : S \to R$, the set S is the **domain** of the function x and the set R is the **codomain**. The **range** or **image** of x is the set

$$x(S) = \{x(\omega) : \omega \in S\}.$$

The **preimage** of a set $B \subset R$ is

$$x^{-1}(B) = \{\omega \in S : x(\omega) \in B\}. \tag{1.18}$$

Note that this **preimage set** is always well-defined for any function $x : S \to R$ even when there are elements $\xi \in R$ for which

$$x^{-1}(\{\xi\}) = \{\omega \in S : x(\omega) = \xi\}$$

consists of more than one point. In the case $x^{-1}(\{\xi\})$ is a set with either no points (cardinality zero) or exactly one point, then the function $x : S \to R$ is said to be **injective** or **one-to-one**. In this case, the function $x : S \to x(S)$ has the property that for each $\xi \in x(S)$, there exists a unique $\omega \in S$ with

$x(\omega) = \xi$. This means there exists a function $x^{-1} : x(S) \to S$ assigning to each $\xi \in x(S)$ the unique $\omega \in S$ with $x(\omega) = \xi$. This function is called the **inverse** of the injective function $x : S \to x(S)$. Please note carefully (again) that the preimage set $x^{-1}(B)$ is well-defined even when the function x^{-1} does not exist.

In general $x(S)$ is a proper subset of R. If $x(S) = R$, then the function x is said to be **surjective** or **onto**. Of course, a surjective function need not be injective, just consider $x : \mathbb{R} \to [0, \infty)$ by $x(\omega) = \omega^2$ which is onto but not one-to-one.

If $x : S \to R$ does happen to be both one-to-one and onto, then the inverse function has domain R and we have $x^{-1} : R \to S$. This is a very nice case, and such a function is said to be **bijective** or is said to be a **one-to-one correspondence**. We will use bijective functions to "rename" the elements of certain base model sets S in our study of mathematical probability. You should be able to find something about that below.

Given any nonempty set S, there is a "trivial" function $\mathrm{id} : S \to S$ with

$$\mathrm{id}(\omega) = \omega \qquad \text{for every } \omega \in S.$$

This function is called the **identity** function on S and is often denoted id_S. Note that id_S has codomain and domain the same set and is always a bijection.

Composition

Given a function $x : S \to R$ and a function $y : R \to T$, there is a function $y \circ x$ called the **composition** of y on x with domain S and codomain T. That is specifically,

$$y \circ x : S \to T \qquad \text{by} \qquad y \circ x(\omega) = y(x(\omega)).$$

The injectivity and surjectivity of a function can be formulated in terms of statements about compositions and identities.

Exercise 1.4.1 Prove the following about a function $x : S \to R$:

(a) x is injective if and only if there exists a function $y : R \to S$ such that $y \circ x = \mathrm{id}_S$.

(b) x is surjective if and only if there exists a function $y : R \to S$ such that $x \circ y = \mathrm{id}_R$.

(c) x is bijective if and only if there exists a function $y : R \to S$ such that $x \circ y = \mathrm{id}_R$ and $y \circ x = \mathrm{id}_S$.

The function y in **(a)** is called a **left inverse**. The function y in **(b)** is called a **right inverse**. In **(c)** the function y is simply called the **inverse** of x and x is said to be **invertible**; in this case we write $y = x^{-1}$.

Incidentally, I believe the axiom of choice is required to correctly complete one of the assertions **(a)** or **(b)** in Exercise 1.4.1. Did you notice this (and notice which one required the axiom of choice)?

Restriction (functions)

We will discuss restriction measures below, but there is also a more general construction which applies to functions and can also be useful to know about. Given any function $x : S \to R$, if $T \subset S$, then the **restriction** of x to T is a function with domain T as follows:

$$x\big|_T : T \to R \qquad \text{by} \qquad x\big|_T(\omega) = x(\omega).$$

1.4.3 Measure(s)

In the notation introduced above, a **baby measure** on a set S with finitely many elements $\omega_1, \omega_2, \ldots, \omega_n$ is a function $\mu : \wp(S) \to [0, \infty)$ satisfying

(i) $\mu(\phi) = 0$.

(ii) If $A, B \subset S$, and $A \cap B = \phi$, i.e., A and B are disjoint, then

$$\mu(A \cup B) = \mu(A) + \mu(B).$$

We will now restrict attention (mostly) to measures satisfying one additional property. These measures are called **probability measures**[6] and (at least for starters) we will denote them using the symbol π. It is still useful to keep in mind the general definition.

[6]Technically, these should be called "baby probability measures" because we are still only really considering the case when S has finitely many elements.

Definition 2 Given a measure $\pi : \wp(S) \to [0, \infty)$ on a set $S = \{\omega_1, \omega_2, \ldots, \omega_n\}$ we say π is a **probability measure** if

$$\pi(S) = 1. \tag{1.19}$$

The condition (1.19) is of course a very simple condition. Note that this condition implies the codomain of a probabiity measure π may be taken to be the interval $[0, 1]$, though of course a baby measure will not be surjective.

Exercise 1.4.2 Let $\pi : \wp(S) \to [0, \infty)$ be a probability measure as above. Show the following:

(a) $\pi(A) \leq 1$ for every $A \subset S$.

(b) There exists a set $A \subset S$ with $\pi(A) = 1$.

When we say π is a (probability) measure on a set S, we are not saying S is the domain of π. The set S is called the **measure space**, but π assigns real numbers not to elements of S but rather to subsets of S. The domain of π is the power set $\wp(S)$. Single elements in the case under consideration do play a special role. Recall that a subset $\{\omega\}$ of S consisting of a single element ω is called a **singleton**. The collection of all singleton sets is traditionally denoted[7] by

$$\begin{aligned} \Omega &= \{\ \{\omega\} : \omega \in S\ \} \\ &= \{\ \{\omega_1\}, \{\omega_2\}, \ldots, \{\omega_n\}\ \}. \end{aligned}$$

Since every set $A \subset S$ is a union of (fintely many) singleton sets, we can write

$$\pi(A) = \sum_{\omega \in A} \pi(\{\omega\})$$

and this can be proved using additivity and induction. Therefore, the measure π, i.e., all the values of π, are determined by the values of the restriction

$$\pi\big|_{\Omega}$$

though this restriction is not a measure. Can you explain why?

The set Ω may be called the **outcome model set**, especially in applications of probability.

[7] At least this is true in the tradition I'm starting now.

Here is the proper way to construct a **restriction measure**: Say you have a (baby probability) measure $\pi : \wp(S) \to [0,1]$ on a (probability) measure space S and you have a subset $T \subset S$. The **restriction measure** on T is given by

$$r = r_T : \wp(T) \to [0,1] \qquad \text{by} \qquad r(A) = \pi(A). \tag{1.20}$$

Exercise 1.4.3 Let $\pi : \wp(S) \to [0,\infty)$ be a probability measure as above, and let $T \subset S$ be a subset of S. Show by example that the restriction measure r defined in (1.20) is **not** always a probability measure.

One trivial example where the restriction measure ρ is a probability measure is when $T = S$. In this case, r and π are the same measure. Can you give an example in which T is a proper subset of S so that the original measure π and the restriction measure r are different measures, but r **is** still a probability measure?

As we might have mentioned or suggested above, mathematical probability is (mostly) the study of probability measures, and the fact that the restriction of a probability measure is usually not a probability measure is the main reason for the following construction(s):

Conditional probability measure

Let $\pi : \wp(S) \to [0,1]$ be a probability measure on S, and let $T \subset S$ be a subset with $\pi(T) > 0$. We define the **restriction probability measure**

$$\rho = \rho_T : \wp(T) \to [0,1] \qquad \text{by} \qquad \rho(A) = \frac{\pi(A)}{\pi(T)}. \tag{1.21}$$

We may note, first of all, that the restriction probability measure is a scaling of the restriction measure $r : \wp(T) \to [0,1]$ considered above. It is the case, furthermore, that ρ is a measure and is a probability measure. The fact that ρ is a probability measure follows because

$$\rho(T) = \frac{\pi(T)}{\pi(T)} = 1.$$

Note carefully the restriction probability measure ρ_T on T and the restriction measure r_T on T are usually different measures.

It turns out that it is easy to consider the restriction probability measure as a probability measure on (the entire measure space) S. We use the same symbol $\rho = \rho_T$ to denote the result of doing so, though in this case, the resulting probability measure is usually called the **conditional probability measure** or the **relative probability measure**. Here are the details:

Definition 3 Given $\pi : \wp(S) \to [0,1]$ a probability measure on S and a set $T \subset S$ with $\pi(T) > 0$, the **conditional probability measure** relative to T is defined by

$$\rho = \rho_T : \wp(S) \to [0,1] \qquad \text{by} \qquad \rho(A) = \frac{\pi(A \cap T)}{\pi(T)}. \tag{1.22}$$

The conditional probability measure is defined on every subset A of the original measure space S. If a set A happens to satisfy $A \subset T$, then $A \cap T = A$, and the value given in (1.22) agrees with the value $\rho(A)$ of the restricted probability measure given by (1.21). In particular, the conditional probability measure also satisfies

$$\rho(S) = \frac{\pi(S \cap T)}{\pi(T)} = 1.$$

Bayes' rule

When A and B are subsets of positive measure in a probability measure space S, the natural relation between the values of the conditional probability measures ρ_A and ρ_B is called **Bayes' rule**. In particular, Bayes' rule relates $\rho_B(A)$, the conditional probability (measure) of A relative to B, to $\rho_A(B)$, the conditional probability measure of B relative to A. The relation naturally centers around the value of $\pi(A \cap B)$. In fact, directly from the definition we have

$$\rho_B(A) = \frac{\pi(A \cap B)}{\pi(B)} \qquad \text{and} \qquad \rho_A(B) = \frac{\pi(A \cap B)}{\pi(A)}.$$

Thus, we can state Bayes' rule as

$$\pi(A \cap B) = \rho_B(A)\ \pi(B) = \rho_A(B)\ \pi(A)$$

or (in the more usual form)

$$\rho_B(A) = \frac{\pi(A)}{\pi(B)}\ \rho_A(B).$$

1.4.4 Integration and Averages

For a probability measure $\pi : \wp(S) \to [0,1]$ on a finite set $S = \{\omega_1, \omega_2, \ldots, \omega_n\}$, the **integral** of a real valued function $x : S \to \mathbb{R}$ is defined by

$$\int_S x = \sum_{j=1}^{n} x(\omega_j)\, \pi(\{\omega_j\}) = \sum_{\omega \in S} x(\omega)\, \pi(\{\omega\}).$$

Because $\pi(S) = 1$, this is the same as the **average value** of x with respect to π, and is also called the **expected value**, though the latter terminology takes us out of mathematical probability and into applications of probability.

Exercise 1.4.4 Let $\pi : \wp(S) \to [0,1]$ be the uniform probability measure on

$$S = \{\text{one}, \text{two}, \text{three}, \text{four}, \text{five}, \text{six}\}$$

with $\pi : \wp(S) \to [0,1]$ by $\pi(\ \{\omega\}\ ) = 1/6$ for all $\omega \in S$. Consider $x : S \to \mathbb{R}$ by $x(\text{one}) = 1$, $x(\text{two}) = 2$, $\ldots$, $x(\text{six}) = 6$ and the subsets

$$A = \{\text{one}, \text{three}, \text{five}\}$$

and $B = S \backslash A$. Calculate the following:

(a) $r(\{\text{one}\})$ where $r = r_A$ is the restriction measure determined by A.

(b) $\rho(\{\text{one}\})$ where $\rho = \rho_A$ is the restriction probability measure determined by A.

(c) $\rho_B(A)$ where ρ_B is the conditional probability measure relative to B.

(d) $\int x$.

Exercise 1.4.5 Given a set $A \subset S$ where S is a measure space with finitely many elements, express $\pi(A)$ as the integral of an appropriate function $x : S \to \mathbb{R}$.

Chapter 2

Lecture 2: Initial Examples and Concepts

We recall that we begin by restricting attention to measure spaces S with $\#S < \infty$. That is, sets with finitely many elements. A probability measure on such a set is a function $\pi : \mathcal{P}(S) \to [0, 1]$ satisfying

(i) $\pi(A) \geq 0$ for every $A \subset S$.

(ii) $\pi(\phi) = 0$.

(iii) If $A, B \subset S$ with $A \cap B = \phi$, then

$$\pi(A \cup B) = \pi(A) + \pi(B).$$

(iv) $\pi(S) = 1$.

2.1 Bernoulli measure

For the Bernoulli measure $\#S = 2$. In some sense, the names of the elements ω_1 and ω_2 for which $S = \{\omega_1, \omega_2\}$ are not important, but very often one starts by considering $S = \{h, t\}$ with the rough interpretation that h represents or corresponds to "success," and t corresponds to "failure."

Actually, to be more precise, there is a one parameter family of Bernoulli measures determined by a parameter p with $0 < p < 1$. Specifically, the

Bernoulli measure $\beta : \mathcal{P}(\{h,t\}) \to [0,1]$ is determined by the single requirement

$$\beta(\{h\}) = p.$$

It follows that $\beta(\{t\}) = 1 - p$.

2.1.1 Renaming and simulation

There can be many functions on a measure space, but very often when we talk about a "function" on a measure space we mean specifically a **real valued** function

$$a : S \to \mathbb{R}.$$

If this function is injective, then $a : S \to a(S)$ is invertible and consideration of such a function amounts to simply renaming the elements of S.

The Bernoulli measure gives a good opportunity to illustrate the use of a **renaming function**. Here, we traditionally take $a : \{h,t\} \to \{0,1\} \subset \mathbb{R}$ with $a(h) = 1$ and $a(t) = 0$. There are a number of advantages in using real numbers and effectively transferring the measure to $\mathbb{R}$. For example, if we wish to generate selections which are supposed to appear "random" using a computer, then the software is usually suited to the selection of numbers (rational numbers or integers usually) rather than symbols like "h" and "t." Generating such selections is called probabilistic simulation or just **simulation**.

Libre Office Calc has a function called `RANDBETWEEN` which can be used[1] to generate such a list for the Bernoulli measure. Specifically, if I enter

`= RANDBETWEEN(0,1)` <return>

in a cell (say cell A1), then I see a number (either 0 or 1, but I do not know which to expect). If I copy (ctrl-c) this cell and paste (ctrl-v) it into the cell below, I'll see another number. I can do this repeatedly, or I can highlight the nine cells below (A2-A10) and paste to get ten selections from $\{0,1\}$

[1]The documentation on `RANDBETWEEN` seems to be somewhat ambiguous. The function is supposed to be "volatile" meaning the value will recompute each time there is an evaluation. I have, on the one hand, made repeated evaluations, and I've never seen any previously computed value change, so the implementation does not seem to be "volatile," and on the other hand, the "non volatile" version `RANDBETWEEN.NV` does not seem to be recognized by my version of Libre Office at all.

which (at least for me) are difficult to predict and (perhaps) can be used to simulate the results of flipping a coin ten times. Here is the result:

1
1
1
0
1
1
1
1
0
0

The same exercise may be completed in several different ways using the statistical software package R. For example, the command

```
sample(0:1, 10, replace = T)      <return>
```

produces

```
[1] 1 1 1 1 0 0 1 1 1 0
```

which is a "random" sample of ten selections from the "vector" `0:1`, i.e., from the set $\{0, 1\}$, with replacement.

If we translate the output of R into selections from the measure space $S = \{h, t\}$ we might obtain the selection

$$h\ h\ h\ h\ t\ t\ h\ h\ h\ t.$$

If we want to know the number of "heads" or the ratio of the number of heads to the total number of selections, then it can be cumbersome to either count (think about how it would be if the size were 100 or 1000 instead of 10) or automate the counting. If, however, we use the renamed set $R = \{0, 1\}$ and the corresponding measure, $\pi : \wp(\{0, 1\}) \to [0, 1]$ with $\pi(\{1\}) = p$, then we can find the sum easily: There is an arithmetic function `sum` in R, and we can simply sum the selections using

```
sum( sample(0:1, 10, replace = T))
```

to determine the number of "heads," namely 7 out of 10. Caution: In practice my output from R looks like this:

```
> sample(0:1,10,replace=T)
    [1] 1 1 1 1 0 0 1 1 1 0
> sum(sample(0:1,10,replace=T))
    [1] 4
>
```

Do you see why I got 4 for the sum instead of 7? If you do not understand why, see the bottom of the next page for a hint.

Exercise 2.1.1 Use a spreadsheet program to generate a list of ten values from the set $\{0, 1\}$ which appear to have been chosen "randomly" with respect to the measure $\pi(\{1\}) = 0.3$.

In most instances when one refers to the Bernoulli measure, one has actually in mind the renamed measure we have just discussed. For the sake of understanding various topics later in the course, it is well worth the time to understand the renaming process and the associated measure in this simple case in detail.

Exercise 2.1.2 Start here with a probability measure $\pi : \wp(\{h, t\}) \to [0, 1]$ with $\pi(\{h\}) = p$. This is the measure called the Bernoulli measure above. Now consider the renaming function $a : \{h, t\} \to R = \{0, 1\}$ with $a(h) = 1$ and $a(t) = 0$. Note it is perfectly fine to also consider the renaming function a to have codomain all of the real numbers so that $a : \{h, t\} \to \mathbb{R}$. Now, define

$$\beta : \wp(\mathbb{R}) \to [0, 1] \qquad \text{by} \qquad \beta(A) = \pi(a^{-1}(A)). \tag{2.1}$$

(a) Show that β is a measure.

(b) Technically, you may not know how to complete part **(a)** above properly because we have not yet discussed the definition of a measure on a measure space like $\mathbb{R}$ with infinitely many elements. You can do two things properly, however:

(b1) Show β is a baby measure.

(b2) Explain precisely how β works and identify a familiar baby measure with which β may be compared.

For the record, when one refers to the Bernoulli measure one really has in mind the measure $\beta : \mathcal{P}(\mathbb{R}) \to [0,1]$ constructed in Exercise 2.1.2, so this exercise and this measure may be worth knowing inside and out backwards and forwards.

Here is the promised hint concerning the output[2] of R:

```
> sum(sample(0:1,10,replace=T))
    [1] 4
> sum(sample(0:1,10,replace=T))
    [1] 5
> sum(sample(0:1,10,replace=T))
    [1] 6
> sum(sample(0:1,10,replace=T))
    [1] 8
> sum(sample(0:1,10,replace=T))
    [1] 5
> sum(sample(0:1,10,replace=T))
    [1] 6
> sum(sample(0:1,10,replace=T))
    [1] 3
> sum(sample(0:1,10,replace=T))
    [1] 5
> sum(sample(0:1,10,replace=T))
    [1] 6
> sum(sample(0:1,10,replace=T))
    [1] 6
> sum(sample(0:1,10,replace=T))
    [1] 5
> sum(sample(0:1,10,replace=T))
    [1] 5
> sum(sample(0:1,10,replace=T))
    [1] 6
> sum(sample(0:1,10,replace=T))
    [1] 5
> sum(sample(0:1,10,replace=T))
    [1] 4
> sum(sample(0:1,10,replace=T))
    [1] 4
> sum(sample(0:1,10,replace=T))
    [1] 5
> sum(sample(0:1,10,replace=T))
    [1] 7
>
```

[2]One may compare this output the the alternative obtained by proceeding each of these commands by `set.seed(1528)`.

2.1.2 Measure induced by renaming

In general if S is a measure space with $\#S < \infty$ and probability measure $\pi : \mathcal{P}(S) \to [0,1]$, we can assume there exists a renaming bijection $a : S \to R \subset \mathbb{R}$ and associate with a the **transferred measure** $\alpha : \mathcal{P}(R) \to [0,1]$ given by

$$\alpha(\{\xi\}) = \pi(\ \{a^{-1}(\xi)\}\ ).$$

The measure α may be easily considered on the entire measure space $\mathbb{R}$ if we take

$$\alpha(\{\xi\}) = \begin{cases} \pi(\ \{a^{-1}(\xi)\}\ ), & \xi \in R = a(S) \\ 0, & \xi \in R^c = \mathbb{R}\backslash a(S). \end{cases} \tag{2.2}$$

You may recall that a measure μ on a measure space S with $\#S < \infty$ is determined by its values on the singleton set

$$\Omega = \{\ \{\omega\} : \omega \in S\ \}$$

by

$$\mu(A) = \sum_{\omega \in A} \mu(\{\omega\}).$$

The measure space $\mathbb{R}$ on which we are considering α above does not have finitely many elements, however, the formula

$$\alpha(A) = \sum_{\xi \in A} \alpha(\{\xi\})$$

still makes sense for any set $A \subset \mathbb{R}$. Can you explain why?

Exercise 2.1.3 Apply the general renaming procedure associated with (2.2) to the pre-Bernoulli measure $\pi : \wp(\{h,t\}) \to [0,1]$ with $\pi(\{h\}) = p$ and show you get the same Bernoulli measure β defined in (2.1).

Solution: In this case, we use the renaming function $a : \{h,t\} \to \{0,1\} \subset \mathbb{R}$ with $a(h) = 1$ corresponding to "success" and $a(t) = 0$. We have then from the definition in (2.2)

$$\alpha(\{\xi\}) = \begin{cases} \pi(\{a^{-1}(\xi)\}), & \xi = 0, 1 \\ 0, & \xi \neq 0, 1. \end{cases}$$

Therefore, the measure of a general set A is obtained in the usual manner for baby measures (see also the discussion of generalized baby measures below) with

$$\alpha(A) = \sum_{\xi \in A} \alpha(\{\xi\}). \tag{2.3}$$

Note that since $\#(\{h,t\}) = 2$, there can be at most two real numbers in $a(\{h,t\})$, and in fact there are exactly two elements in $a(\{h,t\})$ since $a(\{h,t\}) = \{0,1\}$. Consequently, all but possibly two values in the sum appearing on the right in (2.3) are zero; there are at most two nonzero terms. In particular,

$$\alpha(A) = \sum_{\xi \in A \cap \{0,1\}} \alpha(\{\xi\}) = \begin{cases} \sum_{\xi \in A \cap \{0,1\}} \alpha(\{\xi\}), & A \cap \{0,1\} \neq \phi \\ 0, & A \cap \{0,1\} = \phi. \end{cases}$$

Now, consider the two possibilities: If $A \cap \{0,1\} = \phi$, then $a^{-1}(A) = \phi$ and $\pi(a^{-1}(A)) = 0$. On the other hand, if $A \cap \{0,1\} \neq \phi$, then

$$\sum_{\xi \in A \cap \{0,1\}} \alpha(\{\xi\}) = \sum_{\xi \in A \cap \{0,1\}} \pi(\{a^{-1}(\xi)\}).$$

Here we have used (2.2) again. Note however that in this case the set $A \cap \{0,1\}$ contains either one element or two elements. If $A \cap \{0,1\}$ contains exactly one element ξ_1, then $\omega_1 = a^{-1}(\xi_1)$ is a unique well-defined element in $\{h,t\}$, and $a^{-1}(A) = \{\omega_1\}$. Therefore,

$$\sum_{\xi \in A \cap \{0,1\}} \pi(\{a^{-1}(\xi)\}) = \pi(\{a^{-1}(\xi_1)\}) = \pi(\{\omega_1\}) = \pi(a^{-1}(A)).$$

Similarly, if $A \cap \{0,1\} = \{0,1\}$ contains two elements, then $a^{-1}(A) = \{h,t\}$ and

$$\begin{aligned} \sum_{\xi \in A \cap \{0,1\}} \pi(\{a^{-1}(\xi)\}) &= \pi(\{a^{-1}(0)\}) + \pi(\{a^{-1}(1)\}) \\ &= \pi(\{h\}) + \pi(\{t\}) \\ &= 1 \\ &= \pi(a^{-1}(A)). \end{aligned}$$

In all cases then

$$\alpha(A) = \pi(a^{-1}(A)) = \beta(A),$$

and we have shown the renamed measure is the Bernoulli measure defined in (2.1).

2.1.3 Measure induced by a function

The construction of a transferred measure induced by renaming discussed in the previous section has an important generalization. Again, let us start with a measure space S satisfying $\#S < \infty$ and having an associated probability measure $\pi : \mathcal{P}(S) \to [0, 1]$.

This time, however, we consider a general real valued function $x : S \to R \subset \mathbb{R}$ and associate with x the **induced measure** $\alpha_x : \mathcal{P}(R) \to [0, 1]$ given by

$$\alpha_x(\{\xi\}) = \pi(\ \{x^{-1}(\ \{\xi\}\ )\}\ )$$

where we recall $x^{-1}(\ \{\xi\}\ ) = \{\omega \in S : x(\omega) = \xi\}$. Again, the measure α_x may be considered (and will be considered) on the entire measure space $\mathbb{R}$ by taking

$$\alpha_x(\{\xi\}) = \begin{cases} \pi(\ x^{-1}(\ \{\xi\}\ )\ ), & \xi \in R \\ 0, & \xi \notin R. \end{cases} \tag{2.4}$$

You may note that this definition is essentially identical to (2.2) associated with the transferred measure. The only difference is that the function $a : S \to R$ in (2.2) was assumed to be a bijection while the function $x : S \to R$ may or may not be a bijection.

Exercise 2.1.4 Show the formula

$$\alpha_x(A) = \sum_{\xi \in A} \pi(\ x^{-1}(\ \{\xi\}\ )\ )$$

makes sense for any $A \subset \mathbb{R}$ and gives the same value defined in (2.4). Hint: Show an alternative to (2.4) is given by

$$\alpha_x(\{\xi\}) = \begin{cases} \pi(\ x^{-1}(\ \{\xi\}\ )\ ), & \xi \in x(S) \\ 0, & \xi \notin x(S). \end{cases} \tag{2.5}$$

Solution: Notice (2.5) follows because $x(S) \subset R$ and if $\xi \in R \backslash x(S)$, then $x^{-1}(\{\xi\}) = \phi$, so

$$\pi(\ x^{-1}(\ \{\xi\}\ )\ ) = \pi(\phi) = 0.$$

Using either (2.4) or (2.5) we have

$$\alpha_x(A) = \sum_{\xi \in A} \alpha_x(\{\xi\})$$

is a sum with at most finitely many nonzero terms corresponding to elements ξ for which $\xi \in x(S) \subset R$. More precisely,

$$\begin{aligned}\alpha_x(A) &= \sum_{\xi \in A \cap x(S)} \alpha_x(\{\xi\}) \\ &= \sum_{\xi \in A \cap x(S)} \pi(\ x^{-1}(\{\xi\})\ ).\end{aligned}$$

Note that if $\xi \notin x(S)$, we have $x^{-1}(\{\xi\}) = \phi$ and

$$\pi(\ x^{-1}(\{\xi\})\ ) = \pi(\phi) = 0.$$

Therefore we can write

$$\begin{aligned}\alpha_x(A) &= \sum_{\xi \in A \cap x(S)} \pi(\ x^{-1}(\{\xi\})\ ) + \sum_{\xi \in A \setminus x(S)} \pi(\ x^{-1}(\{\xi\})\ ) \\ &= \sum_{\xi \in A} \pi(\ x^{-1}(\{\xi\})\ )\end{aligned}$$

which is the desired formula.

2.1.4 Probability Mass Function (PMF)

Closely related to the induced measure is the **probability mass function** (PMF). Recall that the induced measure is a measure associated with a real valued function $x : S \to \mathbb{R}$. (or very often $x : S \to R$ with some designated codomain $R \subset \mathbb{R}$). Similarly, the probability mass function is a real valued function associated with a function $x : S \to \mathbb{R}$ on a measure space. Specifically, the **probability mass function** $M_x : \mathbb{R} \to [0, 1]$ is defined as follows:

$$M_x(\xi) = \pi(\ x^{-1}(\{\xi\})\ ). \tag{2.6}$$

Exercise 2.1.5 Explain clearly the difference between the probability mass function and the induced measure.

Solution: We have

$$\begin{aligned} \alpha_x(\{\xi\}) &= \pi(\ x^{-1}(\{\xi\})\ ) \qquad \text{and} \\ M_x(\xi) &= \pi(\ x^{-1}(\{\xi\})\ ), \end{aligned}$$

so the values look the "same." The domains, however, are different:

$$\begin{aligned} \alpha_x &: \wp(\mathbb{R}) \to [0, 1] \\ M_x &: \mathbb{R} \to [0, 1]. \end{aligned}$$

In particular, the definition (2.6) gives all the values of the PMF M_x, but the other values of the induced measure are obtained from the formula

$$\alpha_x(A) = \sum_{\xi \in A} \alpha_x(\ \{\xi\}\ ) = \sum_{\xi \in A} \pi(\ x^{-1}(\{\xi\})\ )$$

given in Exercise 2.1.4.

Exercise 2.1.6 Give a formula for the values $M_x(\xi)$ of a probability mass function in terms of the values of the induced measure.

Exercise 2.1.7 Give a formula for the values $\alpha_x(A)$ of an induced measure in terms of the probability mass function.

In the particular case when we consider a renaming bijection $a : S \to R \subset \mathbb{R}$, we find

$$a^{-1}(\{\xi\}) = \{\omega \in S : a(\omega) = \xi\}$$

is either the empty set (if $\xi \notin R$) or the singleton $\{a^{-1}(\xi)\}$ (given in terms of the inverse function a^{-1}). Therefore we can also write

$$M_a(\xi) = \pi(\ \{a^{-1}(\xi)\}\ ) \qquad \text{when} \qquad \xi \in R$$

in terms of the original measure $\pi : \wp(S) \to [0,1]$ and the inverse function $a^{-1} : R \to S$.

Exercise 2.1.8 Let $\pi : S \to [0,1]$ be the uniform measure on the set

$$S = \{\text{one}, \text{two}, \text{three}, \text{four}, \text{five}, \text{six}\}.$$

Consider the function $x : S \to \{3, 6\}$ by $x(\omega) = 3$ for $\omega \neq \text{six}$ and $x(\text{six}) = 6$. Find the induced measure and graph the PMF. For what event do you think the induced measure gives probabilities?

2.2 PMF and CMF of the Bernoulli measure

This gets slightly complicated. When we discuss the Bernoulli measure, we mean the probability measure $\beta : \wp(\mathbb{R}) \to [0,1]$ with $\beta(\{1\}) = p$ and $\beta(\{0\}) = 1 - p$.

Exercise 2.2.1 Prove the conditions

(i) $\beta : \wp(\mathbb{R}) \to [0,1]$ is a probability measure,

(ii) $\beta(\{1\}) = p$, and

(ii) $\beta(\{0\}) = 1 - p$

alone characterize the Bernoulli measure, i.e., show $\beta(\{\xi\}) = 0$ for $\xi \in \mathbb{R}\backslash\{0,1\}$.

Solution:

$$\begin{aligned}
\beta(\{\xi\}) &= \beta(\mathbb{R}) - \beta(\mathbb{R}\backslash\{\xi\}) \\
&\leq 1 - \beta(\{0\} \cup \{1\}) \\
&= 1 - \big[\beta(\{0\}) + \beta(\{1\})\big] \\
&= 1 - [1 - p \ + \ p] \\
&= 0.
\end{aligned}$$

Thus,

$$\beta(A) = \sum_{\xi \in A} \beta(\{\xi\}) = \sum_{\xi \in A \cap \{0,1\}} \beta(\{\xi\}). \qquad \square$$

2.2.1 The PMF

In principle there is no "the" PMF associated with the Bernoulli measure itself, but we need a real valued function $x : \{0,1\} \to \mathbb{R}$ in order to calculate "a" PMF M_x associated with β and x. In practice, however, there is one specific real valued function M designated as the PMF of the Bernoulli measure. The real valued function determining M is the obvious one id : $\{0,1\} \to \{0,1\}$ given by the identity. The induced measure is of course again β. The PMF satisfies

$$M(0) = 1 - p$$
$$M(1) = p,$$

and the graph of M is indicated on the left in Figure 2.1. The PMF shows how

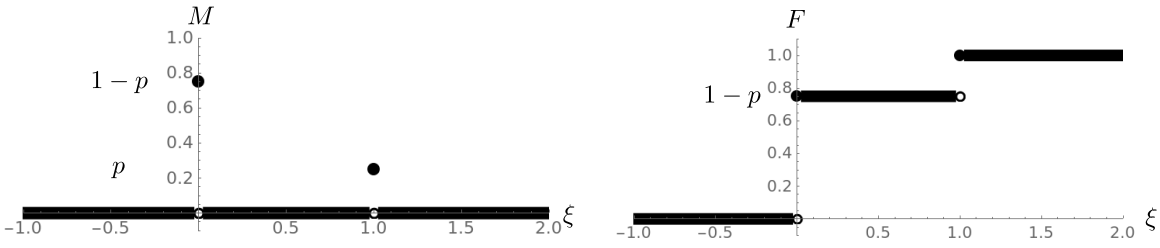

Figure 2.1: The probability mass function (PMF) of the Bernoulli measure (left) and the associated cumulative mass function (CMF)

the probabilities are "distributed" among the real numbers by the function x and is sometimes also called the **mass distribution function** (MDF) or just the distribution function.[3] In the case of Bernoulli measure, the PMF gives a good indication of the underlying measure (Bernoulli measure) itself.

[3]Generally, the concepts of PMF and CMF described in this Lecture are said to describe a "distribution" which mathematically involves a probability measure and a (real valued) function as we have described here. Accordingly, the terms (probability) **mass distribution function** (MDF) and **cumulative** (mass) **distribution function** (CDF) are used more or less interchangeably with PMF and CMF in this context. Many authors

It should be noted however, and we will see examples soon, where this is not the case.

There is another real valued function $F = F_x : \mathbb{R} \to [0,1]$ defined on the entire real line which is essentially equivalent to the PMF. This function has some nice properties, and is sometimes preferable to consider instead of the PMF. The general construction is as follows:

Definition 4 Given a probability measure $\pi : \wp(S) \to [0,1]$ on a measure space S with $\#S < \infty$ and a real valued function $x : S \to R \subset \mathbb{R}$, the **cumulative mass function** (CMF) is the function $F_x : \mathbb{R} \to [0,1]$ by

$$F_x(\xi) = \pi\big(\{\omega : x(\omega) \leq \xi\}\big). \tag{2.7}$$

Theorem 6 The CMF as constructed in Definition 4 is a right continuous nondecreasing step function for which there are values

$$a = \sup\{\xi : F_x(\xi) = 0\}$$

and

$$b = \min\{\xi : F_x(\xi) = 1\}$$

satisfying $a, b \in \mathbb{R}$ and $a \leq b$. The interval $[a,b]$ may be called the **support** of the PMF/CMF, but more often[4] is called the **range** of the PMF/CMF/distribution, and $F_x(a) = 0$.

---

seem to include the idea of "randomness" in the idea of "distribution," so that a kind of hypothetical sampling is also involved at least in the background, and the sampling "distributes" the values of the function x in a manner compatible with the PMF. While this idea can be made at least somewhat precise, there seems to be no mathematical structure suited to incorporate such considerations.

[4]Technically, the terms **support** and **range** signify different things. If you wish to keep them straight and understand the concepts very precisely, then here are some hints: What we have defined here is neither generally the support of the CMF nor the CDF. The support of a function would be the closure of the set of points where the value of the function is nonzero. Thus, if $\{\xi_1, \xi_2, \ldots, \xi_k\} \subset \{x(\omega_1), x(\omega_2), \ldots, x(\omega_n)\}$ is the set of values ξ of x for which $M_x(\xi) = \pi(x^{-1}(\{\xi\})) \neq 0$, then the support of the PMF M_x is $\{\xi_1, \xi_2, \ldots, \xi_k\}$ which is different from the smallest closed interval $[a,b]$ containing these points. That smallest closed interval is traditionally, in probability and statistics, called the range, but this usage differs from the usual usage of the word range in reference to general functions, i.e., the set of image values. Note that according to these technically correct definitions, the support of the CMF is $[a,\infty)$ and the range (in the sense of probability and statistics) of the CMF (or the distribution) is simply the range $[a,b]$ of the PMF (by definition).

Exercise 2.2.2 Write $M_x(\xi)$ in terms of π using set notation for $x^{-1}(\{\xi\})$ so your expression may be compared directly to (2.7).

Solution:

$$M_x(\xi) = \pi\big(\ \{\omega \in S : x(\omega) = \xi\}\ \big) \qquad \text{which compares to}$$
$$M_x(\xi) = \pi\big(\ \{\omega \in S : x(\omega) \leq \xi\}\ \big).$$

Exercise 2.2.3 Calculate the integral

$$\int M$$

of the PMF of the Bernoulli measure **with respect to counting measure** on $\mathbb{R}$.

Exercise 2.2.4 Show the induced measure of section 2.1.3 is a probability measure.

Solution: We can write $x(S) = \{\xi_1, \xi_2, \ldots, \xi_k\} \subset \mathbb{R}$ for some distinct ξ_j for $j = 1, 2, \ldots, k$ with $k \leq n = \#S$. We also write

$$S_j = x^{-1}(\{\xi_j\}) = \{\omega \in S : x(\omega) = \xi_j\}.$$

Then

$$S_i \cap S_j = \phi \quad \text{for } i \neq j \qquad \text{and} \qquad \bigcup_{j=1}^{k} S_j = S.$$

With this in mind, we calculate:

$$\begin{aligned}
\alpha_x(\mathbb{R}) &= \sum_{\xi\in\mathbb{R}} \alpha_x(\{\xi\}) \\
&= \sum_{\xi\in x(S)} \alpha_x(\{\xi\}) \\
&= \sum_{j=1}^{k} \alpha_x(\{\xi_j\}) \\
&= \sum_{j=1}^{k} \pi(x^{-1}(\{\xi_j\})) \\
&= \sum_{j=1}^{k} \pi(\{\omega \in S : x(\omega) = \xi_j\}) \\
&= \sum_{j=1}^{k} \pi(S_j) \\
&= \pi(S) \\
&= 1.
\end{aligned}$$

Exercise 2.2.5 Plot the CMF of the measure and function discussed in Exercise 2.1.8. Show the CMF determines the induced measure uniquely; give a formula in terms of integration. Show the CMF does not determine the underlying measure.

2.2.2 A Detail: Generalized Baby Measure(s)

We have been discussing informally measures $\mu : \wp(\mathbb{R}) \to [0, \infty)$ which essentially satisfy the definition of a baby measure (Definition 1 in section 1.3.1), but have $S = \mathbb{R}$ as a measure space and therefore do not technically satisfy the condition $\#S < \infty$. This is very easy to fix:

Definition 5 Given any set S, a **generalized baby measure** on S is a function

$$\mu : \wp(S) \to [0, \infty)$$

satisfying the following properties:

(0) There is a set $S_0 = \{\omega_1, \omega_2, \ldots, \omega_n\} \subset S$ with $\#S_0 = n < \infty$ such that

$$\mu(S \backslash S_0) = 0.$$

(i) $\mu(\phi) = 0$.

(ii) If $A, B \subset S$, and $A \cap B = \phi$, i.e., A and B are disjoint, then

$$\mu(A \cup B) = \mu(A) + \mu(B).$$

If $\mu(\{\omega_j\}) > 0$ for $j = 1, 2, \ldots, n$, then S_0 is called the **support** of the measure μ.

Again property **(ii)** (additivity) is the main property and is the one that needs to be somehow generalized to get a full-fledged definition of a measure on a measure space with infinitely many elements (and infinitely many subsets of positive measure).

Exercise 2.2.6 Given a generalized baby measure μ with support S_0, show

$$\mu(A) = 0 \qquad \text{for every } A \subset S \backslash S_0$$

and

$$\mu(A) = \sum_{\omega \in A} \mu(\{\omega\}) \qquad \text{for every } A \in \wp(S).$$

2.3 The binomial distribution

Here by "distribution" we mean a combination of an underlying measure and a certain real valued function. Alternatively, we are interested in a particular induced measure. To be precise, we consider the underlying measure space

$$S = \{0,1\}^n = \{(\omega_1, \omega_2, \ldots, \omega_n) \in \mathbb{R}^n : \omega_j \in \{0,1\} \text{ for } j = 1, 2, \ldots, n\}$$

where n is some fixed natural number. It should be pretty clear that $\#S = 2^n < \infty$, so this is a measure space appropriate for a baby measure. In particular, we consider $\beta : \wp(S) \to [0, \infty)$ with

$$\beta(\ \{(\omega_1, \omega_2, \ldots, \omega_n)\}\ ) = p^{\#\{j\ :\ \omega_j = 1\}}(1-p)^{\#\{j\ :\ \omega_j = 0\}}$$

for each $(\omega_1, \omega_2, \ldots, \omega_n) \in S$ where as usual p is some real number in the interval $[0, 1]$.

Exercise 2.3.1 We can give an alternative formulation of the binomial measure as follows: Consider

$$S = \{h,t\}^n = \{(\omega_1, \omega_2, \ldots, \omega_n) : \omega_j \in \{h,t\} \text{ for } j = 1, 2, \ldots, n\}$$

and $\beta : \wp(S) \to [0, \infty)$ by

$$\beta(\ \{(\omega_1, \omega_2, \ldots, \omega_n)\}\ ) = p^{\#\{j\ :\ \omega_j = h\}}(1-p)^{\#\{j\ :\ \omega_j = t\}}.$$

(a) What event does this formulation suggest?

(b) What collection of outcomes is here modeled by the base set S?

(c) Write $R = \{0,1\}^n$ and define an appropriate bijection $a : S \to R$ to obtain the first formulation of the binomial measure above by renaming.

(d) How does this renaming differ from the renaming considered in section 2.1.2?

Exercise 2.3.2 Determine the well-known probability measure the binomial measure becomes in each of the following special cases:

(a) $n = 1$.

(b) $p = 1/2$.

(c) $p = 0$.

(d) $p = 1$.

As suggested by Exercise 2.3.2, the binomial measure depends on two parameters, the number n of components in the measure space (a natural number) and the value $p \in [0, 1]$. Thus, we should really write $\beta = \beta_{n,p}$ or at least $\beta = \beta_n$ with β_1 giving the Bernoulli measure.

In the cases $p \neq 0, 1, 1/2$ it may not be immediately clear that the binomial measure β is a probability measure, though this conclusion may be strongly suggested to you by parts **(a)** and **(b)** of Exercise 2.3.1. This is in fact the case:

Theorem 7 The binomial measure is a probability measure.

Proof: We need to compute $\beta(S)$. Using the notation $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ for elements of S, we have

$$\begin{aligned}\beta(S) &= \sum_{\omega \in S} \beta(\{\omega\}) \\ &= \sum_{\omega \in S} p^{\#\{j \,:\, \omega_j = 1\}} (1-p)^{\#\{j \,:\, \omega_j = 0\}}.\end{aligned}$$

Now, notice there may be several elements $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ for which $\#\{j : \omega_j = 1\}$ is a given natural number k and it is always the case that $\#\{j \ : \ \omega_j = 0\} = n - \#\{j \ : \ \omega_j = 1\} = n - k$. For example, if $n = 3$ and $k = 2$, then for $\omega = (1, 1, 0)$ there holds $\#\{j \ : \ \omega_j = 1\} = 2$, but one also has

$$\#\{j \ : \ \omega_j = 1\} = 2 \qquad \text{for} \qquad \omega = (1, 0, 1) \qquad \text{and} \qquad \omega = (0, 1, 1).$$

In view of this observation, we can set

$$A_k = \big\{\omega \in S : \#\{j \ : \ \omega_j = 1\} = k\big\}$$

for $k = 0, 1, 2, \ldots, n$ and notice that $A_k \cap A_j = \phi$ for $j \neq k$ and

$$S = \bigcup_{k=0}^{n} A_k.$$

That is the sets $A_0, A_1, A_2, \ldots, A_n$ partition the measure space S. This allows us to rewrite $\beta(S)$ as

$$\begin{aligned}\beta(S) &= \sum_{k=0}^{n} \sum_{\omega \in A_k} p^{\#\{j \,:\, \omega_j = 1\}} (1-p)^{\#\{j \,:\, \omega_j = 0\}} \\ &= \sum_{k=0}^{n} \sum_{\omega \in A_k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^{n} p^k (1-p)^{1-k} \sum_{\omega \in A_k} 1 \\ &= \sum_{k=0}^{n} p^k (1-p)^{n-k} \# A_k. \end{aligned} \tag{2.8}$$

Thus we find ourselves faced with a counting problem: How many elements $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ are there in

$$A_k = \big\{\omega \in S : \#\{j \;:\; \omega_j = 1\} = k\big\}?$$

If $\ell = 0$, there is exactly one element in $A_k = A_0$ (the one with all zero components). If $k = 1$, there are n elements in $A_k = A_1$, and these elements have special names:

$$\begin{aligned} \mathbf{e}_1 &= (1, 0, 0, \ldots, 0) \in \mathbb{R}^n \\ \mathbf{e}_2 &= (0, 1, 0, \ldots, 0) \in \mathbb{R}^n \\ &\vdots \\ \mathbf{e}_n &= (0, 0, 0, \ldots, 1) \in \mathbb{R}^n, \end{aligned}$$

These are also called the standard unit basis vectors in $\mathbb{R}^n$. If $k = 2$, we can try to count the elements ω starting with sums of $\mathbf{e}_1 + \mathbf{e}_j$ for $j = 2, \ldots, n$. There are $n-1$ such elements. Then we can take sums $\mathbf{e}_2 + \mathbf{e}_j$ for $j = 3, \ldots, n$, and there are $n - 2$ such elements. Continuing we find the total number of elements is

$$(n-1) + (n-2) + \cdots + 1$$

with the last term 1 corresponding to $\mathbf{e}_{n-1} + \mathbf{e}_n$. The sum of the first $n - 1$ integers is

$$\frac{n(n-1)}{2},$$

and this is the number of elements $\#A_2$ we seek when $k = 2$. Unfortunately, this approach of summing vectors gets more and more complicated.

We consider a different approach. Notice that determining an element $\omega \in A_k$ is equivalent to determining which, among the n components of ω will have the value 1. Put another way, we wish to **choose** among the n component places, which we can imagine as labeled $1, 2, 3, \ldots, n$, which k places will be occupied with the entry 1. This is the same as choosing k numbers from among the natural numbers in $\{1, 2, 3, \ldots, n\}$ (without replacing a particular number once it is chosen). Clearly, there are n choices for the first number, and $n-1$ choices for the second number, and so on until there are $n-(k-1) = n - k + 1$ choices for the last number.

If we take the product

$$n(n-1)\cdots(n-k+1) \tag{2.9}$$

of these numbers of choices, we do not get the number $\#A_k$ we are trying to find. This would be the number of **ordered k-tuples** (with entries from among $1, 2, 3, \ldots, n$ with no repetitions) we might obtain by this procedure. But an element in A_k is determined by these entries without regard to order.

For example, take the case $k = 2$ for which we already know

$$\#A_2 = \frac{n(n-1)}{2}. \tag{2.10}$$

Looking at the set $\{1, 2, 3, \ldots, n\}$, there are n choices for a first place, say we choose place 3. Then there are $n-1$ choices for a second place, say we choose place 5 for the second place. In this way we obtain $(\text{place3}, \text{place5})$ or simply $(3, 5)$ for our choice, and a total of $n(n-1)$ choices altogether. However, this approach counts $(3, 5)$ and $(5, 3)$ as different, but in fact $\omega = \mathbf{e}_3 + \mathbf{e}_5$ and $\omega = \mathbf{e}_5 + \mathbf{e}_3$ are the same element in A_2, so we have overcounted. Specifically, we have counted each element exactly twice. Consequently, we need to divide by $k = 2$ to get the correct answer in (2.10).

In the general case the product

$$n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

in (2.9) gives the number of elements in $\{1, 2, 3 \ldots, n\}^k$ with no repeated entries, but what we really want is the number of subsets of $\{1, 2, 3, \ldots, n\}$ with k distinct elements. Thus, we ask another counting question: How

many elements are there in $\{1, 2, 3 \ldots, n\}^k$ having **the same** distinct entries. Thus, we take a fixed subcollection $\{n_1, n_2, \ldots, n_k\} \subset \{1, 2, 3, \ldots, n\}$ consisting of k distinct elements $n_1, n_2, \ldots, n_k$, and ask how many elements of $\{1, 2, 3 \ldots, n\}^k$ have $n_1, n_2, \ldots, n_k$ as entries? Here, of course, order does matter. There are k choices for the first entry, $k-1$ choices for the second entry, and so on until there is only one choice for the last entry. And the number of elements in this case is given by the product

$$k(k-1)\cdots 1 = k!.$$

Thus, the product given in (2.9) counts each choice of entry numbers $k!$ times. The correct number is therefore

$$\#A_k = \frac{n(n-1)\cdots(n-k+1)}{k!} = \frac{n!}{(n-k)!k!}. \tag{2.11}$$

The number in (2.11) is called the **combination of n things taken k at a time** and is denoted by

$$\left(\begin{array}{c} n \\ k \end{array} \right).$$

The combination of n things taken k at a time is also called the **binomial coefficient** for reasons that should become clear shortly. See the **binomial formula** (2.12) below. We will review the counting techniques discussed above and learn some others in the next chapter. For now, we can write the expression for $\beta(S)$ given in (2.8) as

$$\beta(S) = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}.$$

This turns out to be a special case of the binomial expansion formula

$$(a+b)^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) a^k b^{n-k}, \tag{2.12}$$

and it is probably this circumstance from which the names binomial measure and binomial distribution are derived. In any case, the formula (2.12) can be verified by induction. Once (2.12) is known, then we can take $a = p$ and $b = 1 - p$ to find

$$\beta(S) = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} = [p + (1-p)]^n = 1.$$

This establishes that the binomial measure is a probability measure in all cases. □

Exercise 2.3.3 Prove the binomial expansion formula using induction.

Solution: When $n = 0$, we have $(a+b)^0 = 1$ and

$$\sum_{k=0}^{0} \left(\begin{array}{c} 0 \\ k \end{array} \right) a^k b^{-k} = \left(\begin{array}{c} 0 \\ 0 \end{array} \right) a^0 b^0 = (1)(1)(1) = 1.$$

Therefore, the formula holds when $n = 0$. When $n = 1$, $(a+b)^1 = a + b$ and

$$\sum_{k=0}^{1} \left(\begin{array}{c} 1 \\ k \end{array} \right) a^k b^{1-k} = \left(\begin{array}{c} 1 \\ 0 \end{array} \right) a^0 b^1 + \left(\begin{array}{c} 1 \\ 1 \end{array} \right) a^1 b^0 = b + a.$$

Therefore, the formula holds when $n = 1$. We can take this as our base case.[5]

Next, we assume the binomial expansion formula holds for $n = \ell$ (as an inductive hypothesis) and consider the quantities

$$(a+b)^{\ell+1} \qquad \text{and} \qquad \sum_{k=0}^{\ell+1} \left(\begin{array}{c} \ell+1 \\ k \end{array} \right) a^k b^{\ell+1-k}.$$

By the inductive hypothesis

$$\begin{aligned} (a+b)^{\ell+1} &= (a+b) \sum_{k=0}^{\ell} \left(\begin{array}{c} \ell \\ k \end{array} \right) a^k b^{\ell-k} \\ &= \sum_{k=0}^{\ell} \left(\begin{array}{c} \ell \\ k \end{array} \right) a^{k+1} b^{\ell-k} + \sum_{k=0}^{\ell} \left(\begin{array}{c} \ell \\ k \end{array} \right) a^k b^{\ell-k+1}. \end{aligned} \tag{2.13}$$

Shifting indices in the first sum in (2.13) gives

$$\begin{aligned} \sum_{k=0}^{\ell} \left(\begin{array}{c} \ell \\ k \end{array} \right) a^{k+1} b^{\ell-k} &= \sum_{k=1}^{\ell+1} \left(\begin{array}{c} \ell \\ k-1 \end{array} \right) a^k b^{\ell-k+1} \\ &= a^{\ell+1} + \sum_{k=1}^{\ell} \left(\begin{array}{c} \ell \\ k-1 \end{array} \right) a^k b^{\ell-k+1}. \end{aligned}$$

[5]We can probably take either the case $n = 0$ or the case $n = 1$ as the base case, but it is sometimes nice to see the first couple cases in an induction worked out explicitly.

The second sum in (2.13) can be written as

$$\sum_{k=0}^{\ell} \left(\begin{array}{c} \ell \\ k \end{array} \right) a^k b^{\ell-k+1} = \sum_{k=1}^{\ell} \left(\begin{array}{c} \ell \\ k \end{array} \right) a^k b^{\ell-k+1} + b^{\ell+1}.$$

Combining these two expressions we obtain from (2.13)

$$(a+b)^{\ell+1} = a^{\ell+1} + \sum_{k=1}^{\ell} \left[\left(\begin{array}{c} \ell \\ k-1 \end{array} \right) + \left(\begin{array}{c} \ell \\ k \end{array} \right) \right] a^k b^{\ell-k+1} + b^{\ell+1}.$$

The coefficient of the middle sum can be simplified:

$$\begin{aligned} \left(\begin{array}{c} \ell \\ k-1 \end{array} \right) + \left(\begin{array}{c} \ell \\ k \end{array} \right) &= \frac{\ell!}{(\ell-k+1)!(k-1)!} + \frac{\ell!}{(\ell-k)!k!} \\ &= \frac{\ell!}{(\ell-k)!(k-1)!} \left[\frac{1}{\ell-k+1} + \frac{1}{k} \right] \\ &= \frac{\ell!}{(\ell-k)!(k-1)!} \left[\frac{\ell+1}{(\ell-k+1)k} \right] \\ &= \frac{(\ell+1)!}{(\ell-k+1)!k!} \\ &= \left(\begin{array}{c} \ell+1 \\ k \end{array} \right). \end{aligned} \tag{2.14}$$

Thus we have shown

$$\begin{aligned} (a+b)^{\ell+1} &= a^{\ell+1} + \sum_{k=1}^{\ell} \left(\begin{array}{c} \ell+1 \\ k \end{array} \right) a^k b^{\ell-k+1} + b^{\ell+1} \\ &= \sum_{k=0}^{\ell+1} \left(\begin{array}{c} \ell+1 \\ k \end{array} \right) a^k b^{\ell-k+1} \end{aligned}$$

which is the desired binomial formula in the case $n = \ell + 1$. □

Note: If you have not done so, the binomial formula

$$(a+b)^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) a^k b^{n-k}$$

given in (2.12) is worth memorizing and knowing how to use. Also, the identity

$$\left(\begin{array}{c} \ell \\ k-1 \end{array} \right) + \left(\begin{array}{c} \ell \\ k \end{array} \right) = \left(\begin{array}{c} \ell+1 \\ k \end{array} \right)$$

obtained in (2.14) is a well-known identity for combinations which is useful to know and understand.

Exercise 2.3.4 Explain how the identity

$$\left(\begin{array}{c} \ell \\ k-1 \end{array} \right) + \left(\begin{array}{c} \ell \\ k \end{array} \right) = \left(\begin{array}{c} \ell+1 \\ k \end{array} \right)$$

appears in Pascal's triangle.

To summarize our discussion so far, we have introduced for each $n = 1, 2, 3, \ldots$ and each $p \in [0, 1]$ a measure

$$\beta : \wp(S) \to [0, \infty)$$

called the **binomial measure** on the measure space

$$S = \{0, 1\}^n = \{(\omega_1, \omega_2, \ldots, \omega_n) \in \mathbb{R}^n : \omega_j \in \{0, 1\} \text{ for } j = 1, 2, \ldots, n\}$$

and having values determined by

$$\beta(\ \{(\omega_1, \omega_2, \ldots, \omega_n)\}\ ) = p^{\#\{j\ :\ \omega_j = 1\}}(1-p)^{\#\{j\ :\ \omega_j = 0\}}.$$

For $n = 1$, the binomial measure is the same as the Bernoulli measure $\beta = \beta_1$ and for any n and p, the binomial measure $\beta = \beta_{n,p}$ is a probability measure.

When we speak of the **binomial distribution** we couple the binomial measure with a certain real valued function $x : S \to \mathbb{R}$. This function gives the number of "successes" represented by an element $\omega = (\omega_1, \omega_2, \ldots, \omega_n) \in S$. That is,

$$x(\omega) = \sum_{j=1}^{n} \omega_j.$$

Perhaps the first question to ask is:

> What is the (binomial) induced measure $\alpha_x : \wp(\mathbb{R}) \to [0, 1]$ on the measure space $\mathbb{R}$ associated with the function x?

Notice that in the case, $n = 1$, the function x becomes essentially[6] the identity function, so the induced measure is the Bernoulli measure with the associated PMF and CMF illustrated in Figure 2.1. For $n > 1$, we have a new (induced) measure on the measure space $\mathbb{R}$.

In order to understand α_x in the case $n > 1$, let us start by observing the range (as a function) or image of the function x is the set $\{0, 1, \ldots, n\} \subset \mathbb{R}$. This tells us the support or "range" of the measure α_x lies in this set. In particular,

$$\alpha_x(\{\xi\}) = 0 \qquad \text{for} \qquad \xi \in \mathbb{R}\backslash\{0, 1, \ldots, n\}.$$

It is natural to consider first

$$\alpha_x(\{0\}) = \beta(\{\omega \in S : x(\omega) = 0\}) = \beta(\{(0, 0, \ldots, 0)\}) = (1 - p)^n.$$

Exercise 2.3.5 Find

$$\alpha_x(\{\xi\}) = \beta(\{\omega \in S : x(\omega) = \xi\})$$

for $\xi = 1, 2, \ldots, n$ where β is the binomial measure and α_x is the induced measure associated with the binomial distribution.

Solution: The condition $x(\omega) = \xi$ means that ω is associated with ξ "successes," i.e., there are exactly ξ components ω_j of ω that have value $\omega_j = 1$. We know there are

$$\left(\begin{array}{c} n \\ \xi \end{array} \right)$$

such elements $\omega \in S$ and each has

$$\beta(\{\omega\}) = p^\xi (1 - p)^{n-\xi}.$$

Therefore,

$$\beta(\{\omega \in S : x(\omega) = \xi\}) = \sum_{\omega \in x^{-1}(\{\xi\})} \beta(\{\omega\}) = \left(\begin{array}{c} n \\ \xi \end{array} \right) p^\xi (1 - p)^{n-\xi}. \qquad \square$$

[6]Technically in this case $\omega = (\omega_1)$ and $x(\omega) = x((\omega_1)) = \omega_1$, so the function x is the same as the identity function up to the identification of ω with (ω_1) by the removal of one set of parentheses.

2.4 The PMF and CMF of the binomial distribution

The values $\alpha_x(\{\xi\})$ of the binomial induced measure on singleton sets may be visualized as the graph of the PMF $M_x : \mathbb{R} \to [0,1]$ with

$$M_x(\xi) = \alpha_x(\{\xi\}) = \beta(\{\omega \in S : x(\omega) = \xi\}).$$

A formula for the (potentially nonzero) values $M_x(\xi)$ for $\xi = 0, 1, 2, \ldots, n$ was calculated in the solution of Exercise 2.3.5:

$$M_x(\xi) = \left(\begin{array}{c} n \\ \xi \end{array} \right) p^{\xi}(1-p)^{n-\xi}. \tag{2.15}$$

One approach to understanding these values is to simply plot some of them with mathematical software like Mathematica. The special cases $p = 0$, $p = 1/2$ and $p = 1$ are relatively easy to understand directly.

When $p = 0$, we find $M_x(0) = 1$ and $M_x(\xi) = 0$ for $\xi \neq 0$, in particular for $\xi = 1, 2, \ldots, n$. The first value is to a certain extent obtained by convention or definition, for if we substitute $p = 0$ and $\xi = 0$ into (2.15) directly we find

$$M_x(0) = \left(\begin{array}{c} n \\ 0 \end{array} \right) 0^0(1)^n = (1)(0^0)(1).$$

In principle, the factor 0^0 is indeterminate. We take the value $0^0 = 1$ in this case either by convention, or through the limit

$$\lim_{p \searrow 0} M_x(0) = \lim_{p \searrow 0} \left(\begin{array}{c} n \\ 0 \end{array} \right) p^0(1-p)^n = \lim_{p \searrow 0}[(1)(1)(1-p)^n = 1,$$

or heuristically by interpretation of $M_x(0)$ when $p = 0$ as the probability of having "no successes" in n Bernoulli trials when the probability of a "success" in one Bernoulli trial is $p = 0$. Put another way:

$$\text{If you can never win, then you must always lose.} \tag{2.16}$$

The situation with $p = 1$ is (symmetrically) similar with $M_x(\xi) = 0$ for $\xi = 0, 1, 2, \ldots, n-1$ and $M_x(n) = 1$.

Exercise 2.4.1 Find a statement analogous to the statement given in (2.16) summarizing the binomial PMF in the case $p = 1$.

2.4.1 The symmetric case $p = 1/2$

When $p = 1/2$ the values given in (2.15) become

$$M_x(\xi) = \frac{1}{2^n} \left(\begin{array}{c} n \\ \xi \end{array} \right) \qquad \text{for } \xi = 0, 1, 2, \ldots, n.$$

This means the values $M_x(\xi)$ are simply a scaling of the binomial coefficients

$$\left(\begin{array}{c} n \\ 0 \end{array} \right) \qquad \left(\begin{array}{c} n \\ 1 \end{array} \right) \qquad \left(\begin{array}{c} n \\ 2 \end{array} \right) \qquad \cdots \qquad \left(\begin{array}{c} n \\ n \end{array} \right)$$

appearing as the n-th (or technically "n plus first") row of Pascal's triangle:

$$\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & & & & & \vdots & & & & & \\
 & & & & & \vdots & & & & & \\
1 & & n & & \frac{n(n-1)}{2} & & \cdots & & \frac{n(n-1)}{2} & & n & & 1 \\
 & & & & & \vdots & & & & &
\end{array}$$

The values (of the binomial coefficients) may be seen to display an obvious symmetry which may be described in more detail as follows:

If the index of the row is even (starting with the initial entry 1 as the index $n = 0$ row) then there is a middle entry in the row with value

$$\left(\begin{array}{c} n \\ n/2 \end{array} \right).$$

This is the maximum value in the row, and the values

$$\left(\begin{array}{c} n \\ k \end{array} \right)$$

increase from 1 (when $k = 0$) for $k = 0, 1, 2, \ldots, n/2$ and then decrease symmetrically with

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \left(\begin{array}{c} n \\ n-k \end{array} \right) \qquad \text{for } k = n/2, n/2+1, \ldots, n. \tag{2.17}$$

If on the other hand, n is odd, then $n-1$ and $n+1$ are even, and

$$\left(\begin{array}{c} n \\ (n-1)/2 \end{array} \right) \qquad \text{and} \qquad \left(\begin{array}{c} n \\ (n+1)/2 \end{array} \right) \tag{2.18}$$

constitute two equal "middle" terms sharing the common maximum value for the row. Again the values

$$\left(\begin{array}{c} n \\ k \end{array} \right)$$

increase from 1 (when $k = 0$) for $k = 0, 1, 2, \ldots, (n-1)/2$ and then decrease symmetrically with

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \left(\begin{array}{c} n \\ n-k \end{array} \right) \qquad \text{for } k = (n+1)/2, (n+1)/2+1, \ldots, n. \tag{2.19}$$

Exercise 2.4.2 Verify the symmetry relations (2.17) and (2.19) by using the explicit formula for the combinations.

Given the symmetry relations, the monotonicity of the binomial coefficients is perhaps most easily expressed by considering two rows at a time with the first even indexed and the second odd indexed. In this formulation, the monotonicity asserts simply that for n even the expressions

$$\left(\begin{array}{c} n \\ k \end{array} \right) \qquad \text{and} \qquad \left(\begin{array}{c} n+1 \\ k \end{array} \right) \tag{2.20}$$

are increasing in k for $k = 0, 1, 2, \ldots, n/2$. This monotonicity assertion may be verified by induction as follows: The first row of Pascal's triangle (corresponding to index $n = 0$) satisfies the monotonicity vacuously with a single entry 1, a (single) middle entry, and a (single) maximum entry. Similarly, the second row, corresponding to the odd index $n = 1$ has precisely two entries

$$\left(\begin{array}{c} 1 \\ 0 \end{array} \right) \qquad \text{and} \qquad \left(\begin{array}{c} 1 \\ 1 \end{array} \right)$$

both with value 1 constituting two equal, maximum, "middle" values. The third row (index $n = 2$) has a middle element

$$\left(\begin{array}{c} 2 \\ 1 \end{array} \right) = 2$$

giving

$$1 \qquad 2 \qquad 1$$

as the $n = 2$ row with evident monotonicity. The rows for $n = 0, 1, 2$ are adequate for a base case for the induction. The fourth ($n = 3$) row[7] may also be easily checked.

Now, say that inductively we have a row corresponding to even index n and the the expressions in (2.20) increase in k for $k = 0, 1, 2, \ldots, n/2$. The first halves of the initial rows (in Pascal's triangle) with index n and index $n + 1$ look something like this:

$$\begin{array}{ccccccc}
\left(\begin{array}{c} n \\ 0 \end{array} \right) & & \left(\begin{array}{c} n \\ 1 \end{array} \right) & \cdots & & & \\
\left(\begin{array}{c} n+1 \\ 0 \end{array} \right) & & \left(\begin{array}{c} n+1 \\ 1 \end{array} \right) & \cdots & & & \\
\\
\left(\begin{array}{c} n \\ k \end{array} \right) & & \left(\begin{array}{c} n \\ k+1 \end{array} \right) & \cdots & & & \left(\begin{array}{c} n \\ n/2 \end{array} \right) \\
& \left(\begin{array}{c} n+1 \\ k+1 \end{array} \right) & & & \cdots & \left(\begin{array}{c} n+1 \\ n/2 \end{array} \right).
\end{array}$$

It may be worth taking a moment to mentally digest this display. The combination

$$\left(\begin{array}{c} n \\ n/2 \end{array} \right) \qquad \text{is a middle maximum element in the index } n \text{ row,}$$

while

$$\left(\begin{array}{c} n+1 \\ n/2 \end{array} \right) \qquad \text{is one of two maximum elements in the index } n+1 \text{ row,}$$

[7] . . . and another half dozen rows if you like.

in which the next entry (not shown) is

$$\begin{pmatrix} n+1 \\ n/2+1 \end{pmatrix}$$

and

$$\begin{pmatrix} n+1 \\ n/2 \end{pmatrix} = \begin{pmatrix} n+1 \\ n/2+1 \end{pmatrix}$$

by symmetry.

Using the same kind of display, we can consider the first halves of the rows with index $n+1$ and index $n+2$:

$$\begin{pmatrix} n+1 \\ 0 \end{pmatrix} \qquad \begin{pmatrix} n+1 \\ 1 \end{pmatrix} \quad \cdots$$

$$\begin{pmatrix} n+2 \\ 0 \end{pmatrix} \qquad \begin{pmatrix} n+2 \\ 1 \end{pmatrix} \quad \cdots$$

$$\begin{array}{ccccccc} \begin{pmatrix} n+1 \\ k \end{pmatrix} & & \begin{pmatrix} n+1 \\ k+1 \end{pmatrix} & \cdots & \begin{pmatrix} n+1 \\ n/2 \end{pmatrix} & & \\ & \begin{pmatrix} n+2 \\ k+1 \end{pmatrix} & & \cdots & & \begin{pmatrix} n+2 \\ (n+2)/2 \end{pmatrix} \end{array}$$

Note that $(n+2)/2 = n/2+1$ and

$$\begin{pmatrix} n+2 \\ (n+2)/2 \end{pmatrix}$$ is a middle maximum entry in an even indexed row.

The identity (2.14) along with Exercise 2.4.1 indicates how each entry in the index $n+2$ row is the sum of two adjacent entries in the previous index $n+1$ row. In particular, for $0 < k < (n+2)/2$

$$\begin{pmatrix} n+2 \\ k \end{pmatrix} = \begin{pmatrix} n+1 \\ k-1 \end{pmatrix} + \begin{pmatrix} n+1 \\ k \end{pmatrix}$$

and

$$\begin{pmatrix} n+2 \\ k+1 \end{pmatrix} = \begin{pmatrix} n+1 \\ k \end{pmatrix} + \begin{pmatrix} n+1 \\ k+1 \end{pmatrix}.$$

By the strict monotonicity assumption for the odd indexed $n+1$ row we have

$$\begin{pmatrix} n+1 \\ k-1 \end{pmatrix} < \begin{pmatrix} n+1 \\ k \end{pmatrix} \leq \begin{pmatrix} n+1 \\ k+1 \end{pmatrix}$$

with equality holding on the right if and only if $k = n/2$. This allows the comparison

$$\left(\begin{array}{c} n+2 \\ k \end{array} \right) = \left(\begin{array}{c} n+1 \\ k-1 \end{array} \right) + \left(\begin{array}{c} n+1 \\ k \end{array} \right) < \left(\begin{array}{c} n+1 \\ k \end{array} \right) + \left(\begin{array}{c} n+1 \\ k+1 \end{array} \right) = \left(\begin{array}{c} n+2 \\ k+1 \end{array} \right)$$

which establishes the monotonicity of the expression

$$\left(\begin{array}{c} n+2 \\ k \end{array} \right)$$

corresponding to expression on the left (2.20) for the next even indexed $n+2$ row.

Similarly, we can consider the next odd indexed $n+3$ row. Within Pascal's triangle the first half of this row appears something like this:

$$\begin{array}{ccccc} \left(\begin{array}{c} n+2 \\ 0 \end{array} \right) & & \left(\begin{array}{c} n+2 \\ 1 \end{array} \right) & \cdots & \\ \left(\begin{array}{c} n+3 \\ 0 \end{array} \right) & & \left(\begin{array}{c} n+3 \\ 1 \end{array} \right) & \cdots & \end{array}$$

$$\begin{array}{ccccccc} \left(\begin{array}{c} n+2 \\ k \end{array} \right) & & \left(\begin{array}{c} n+2 \\ k+1 \end{array} \right) & \cdots & & & \left(\begin{array}{c} n+2 \\ (n+2)/2 \end{array} \right) \\ & \left(\begin{array}{c} n+3 \\ k+1 \end{array} \right) & & & \cdots & \left(\begin{array}{c} n+3 \\ (n+2)/2 \end{array} \right) & \end{array}$$

Using monotonicity for the $n+2$ row allows the comparison

$$\left(\begin{array}{c} n+3 \\ k \end{array} \right) < \left(\begin{array}{c} n+3 \\ k+1 \end{array} \right)$$

which gives the monotonicity for the next odd indexed $n+3$ row and completes the induction. $\square$

Now that we've established something about the qualitative properties of the PMF M_x of the binomial distribution/measure in the cases $p = 0$, $p = 1/2$, and $p = 1$, we can plot the actual values to confirm what we have shown above. In Figure 2.2 and Figure 2.3 are plots of the cases $p = 0$ and $p = 1/2$ for $n = 1, 2, 3, \ldots, 9$. The monotonicity and symmetry properties should be evident in each plot.

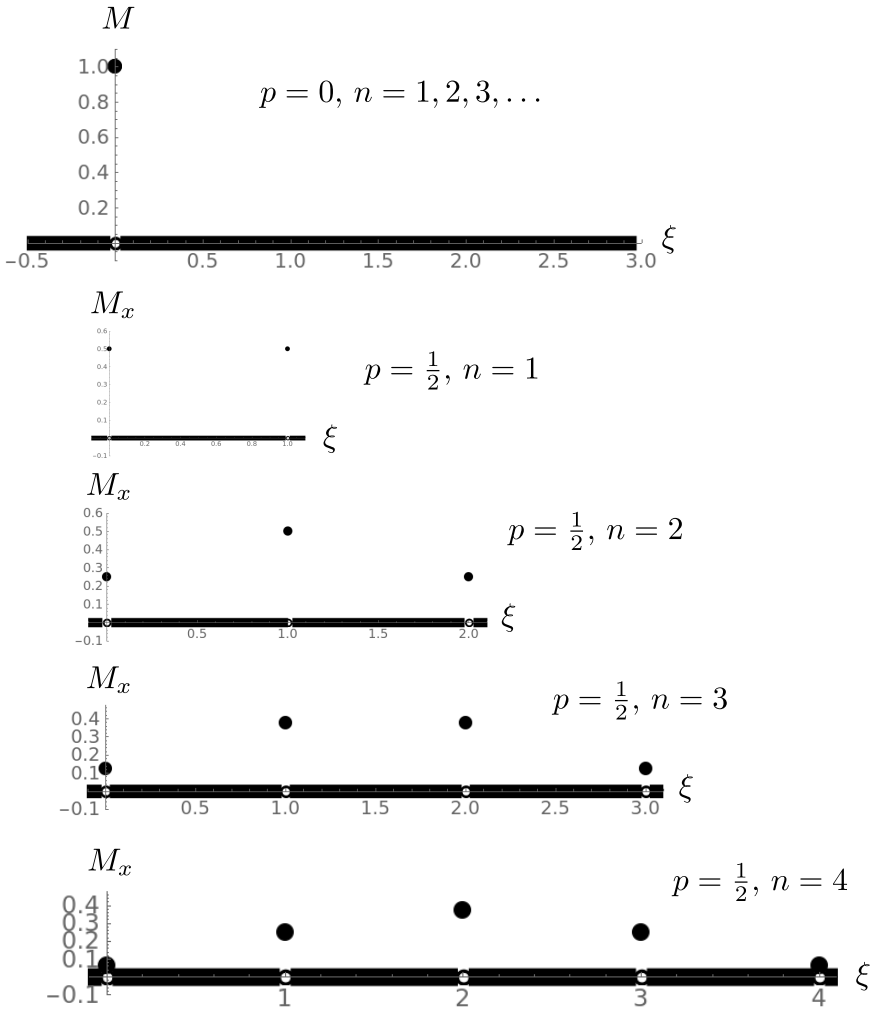

Figure 2.2: The PMF for the binomial distribution when $p = 0$ is the same for every $n = 1, 2, 3, \ldots$ having support $\{0\}$ and corresponds to the probability measure $\alpha_x : \wp(\mathbb{R}) \to [0, 1]$ with $\alpha_x(\{0\}) = 1$ which is the same as the Bernoulli measure with $p = 0$. The PMF of the binomial distribution when $p = 1/2$ and $n = 1$ is the Bernoulli PMF for $p = 1/2$. For $p = 1/2$ and $n = 2$, the probability measure values are distributed among three points of support on the "range" interval $[0, 2]$ with the maximum value at $\xi = 1$. These plots are all shown with no relative scaling between the horizontal and vertical axes.

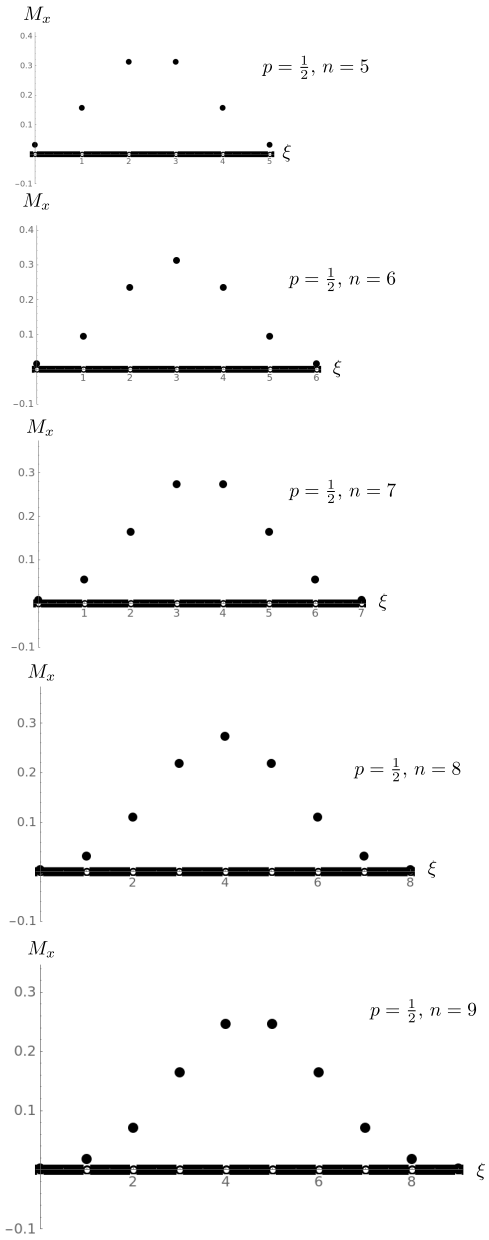

Figure 2.3: The PMF of the binomial distributions with $p = 1/2$ and $n = 5, 6, 7, 8, 9$. Here the vertical axis is scaled relative to the horizontal axis in order to improve the visibility of the plot.

For $n \geq 5$ (and $p = 1/2$) it starts to become evident that the points corresponding to nonzero values on the graph of the PMF have other interesting qualitative properties beyond monotonicity and symmetry. In particular, one can see evident changes in concavity among the points. It is natural to attempt to make this kind of property precise using some continuous function passing through the same points as indicated in the case $p = 1/2$ and $n = 10$ in Figure 2.4. One might be tempted to imagine the value of M_x indicated

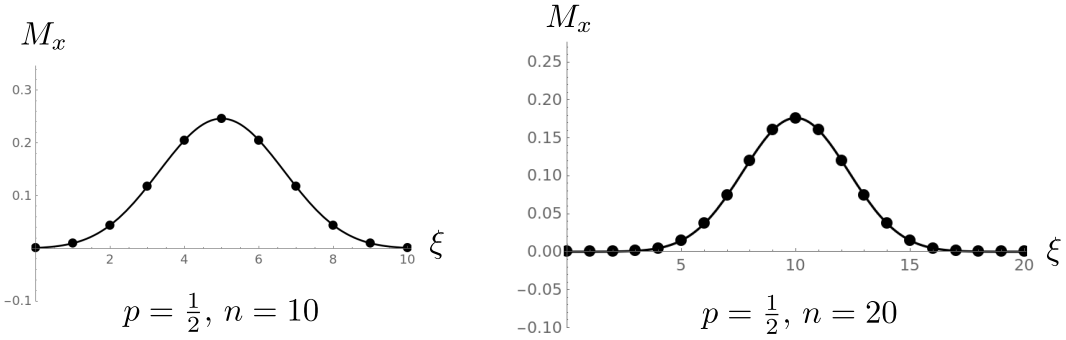

Figure 2.4: The PMF of the binomial distributions with $p = 1/2$ and $n = 10, 20$. Here we have suppressed plotting the zero values $M_x(\xi) = 0$ for to enhance the clarity of the illustration, but it should not be forgotten that these are, in fact, the actual values of the PMF. Also indicated is a smooth function passing through the points $(\xi, M_x(\xi))$ for $\xi = 0, 1, 2, \ldots, 10$ on the left and for $\xi = 0, 1, 2, \ldots, 20$ on the right. In this case, the horizontal axes do not share the same scale.

by these plots is given by an elementary function, e.g., a polynomial function of ξ. That however is not the case, but there is a special (transcendental) function giving the values precisely constructed in terms of the **Gamma function** which is perhaps worth knowing about.

2.4.2 The Gamma function

Let us briefly consider the function $\Gamma : (0, \infty) \to (0, \infty)$ given by

$$\Gamma(\xi) = \int_0^\infty t^{\xi - 1} e^{-t} \, dt$$

Notice that when $\xi = 1$ we get

$$\Gamma(1) = \int_0^\infty e^{-t}\,dt = -e^{-t}\Big|_{t=0}^{\infty} = 1 = 0!.$$

Also, for every $\xi > 0$ we can integrate by parts to see

$$\begin{aligned}\Gamma(\xi+1) &= \int_0^\infty t^\xi e^{-t}\,dt \\ &= -t^\xi e^{-t}\Big|_{t=0}^{\infty} + \int_0^\infty \xi t^{\xi-1} e^{-t}\,dt \\ &= \xi\Gamma(\xi).\end{aligned}$$

It follows from the fact that $\Gamma(1) = 1 = 0!$ and the identity $\Gamma(\xi+1) = \xi\Gamma(\xi)$ that

$$\Gamma(2) = \Gamma(1+1) = 1\Gamma(1) = (1)0! = 1!$$

and inductively that

$$\Gamma(n) = \Gamma(n-1+1) = (n-1)\Gamma(n-1) = (n-1)(n-2)! = (n-1)!$$

for all $n = 3, 4, 5, \ldots$. Consequently, the binomial PMF is given for integer values by

$$M_x(\xi) = \frac{1}{2^n}\left(\begin{array}{c} n \\ \xi \end{array}\right) = \frac{1}{2^n}\,\frac{n!}{(n-\xi)!\xi!} = \frac{1}{2^n}\,\frac{n!}{\Gamma(n-\xi+1)\Gamma(\xi+1)}.$$

This is the formula used to plot the smooth companion graphs in Figure 2.4. You may not be familiar with the Gamma function, but it is a nice special function like the sine function or the logarithm function, and it is available as a standard function in most mathematical software.

Exercise 2.4.3 Plot the PMF for the binomial distribution when $p = 1$ and $n = 1, 2, 3, \ldots$. Hint: Each one is different. What is the range?

Exercise 2.4.4 Use mathematical software to plot the Gamma function on the interval $0 < \xi < \infty$.

Exercise 2.4.5 Plot the CMF of the binomial distribution in the case $p = 1/2$ and $n = 1, 2, 3$.

The PMF of the binomial distribution(s) with $p = 1/2$ and range containing the n points/values $0, 1, 2, \ldots, n$ may be taken as prototypical. Notice that the description of the values $M_x(\xi)$ is somewhat complicated with the maximum value sometimes (for n even to be exact) occurring at a single point $\xi = n/2$ and other times (for n odd to be exact) occurring at two points. And then outside the maximum point(s) certain monotonicity properties have been identified and certain symmetry properties as well. Relaxing the symmetry properties, we can say a PMF M_x with nonzero values $M_x(\xi)$ on $\xi = 0, 1, 2, \ldots, n$ is **roughly bell shaped** if it (roughly) shares the maximal value and monotonicity properties of the binomial $p = 1/2$ PMF. This definition of "roughly bell shaped" is, by admission, not entirely precise, but the binomial PMF with $p \neq 1/2$ discussed in the next section is a specific example that should clarify the basic idea.

2.4.3 The nonsymmetric case $p \neq 0, 1/2, 1$

Having basically understood the situation when $p = 1/2$, let us consider what happens when the parameter p moves away from $1/2$. In particular, let us consider the values

$$M_x(\xi) = \left(\begin{array}{c} n \\ \xi \end{array} \right) \, p^{\xi} \, (1-p)^{n-\xi} \qquad \text{for } \xi = 0, 1, 2, \ldots, n,$$

corresponding to the probability of exactly ξ successes out of n Bernoulli trials at probability p of success, when $0 < p < 1/2$, that is, when the probability of success in each Bernoulli trial is less than $1/2$. For fixed ξ, the changing factor

$$p^{\xi}(1-p)^{n-\xi}$$

may be considered as a function of p. Let us write

$$f(p) = p^{\xi}(1-p)^{n-\xi}.$$

It is not difficult to see that for $\xi > n/2$

$$f'(p) = (\xi - np)p^{\xi-1}(1-p)^{n-\xi-1} > 0 \qquad \text{for } 0 < p \leq 1/2.$$

This means each of the values $M_x(\xi)$ for $\xi > n/2$ **decreases** as p decreases below $1/2$. If n is even and $\xi = n/2$, then $f'(1/2) = 0$, but $f'(p) > 0$ for $0 < p < 1/2$, so the value $M_x(n/2)$ also decreases as p decreases below

1/2. The rate at which $M_x(\xi)$ decreases, at least when p is near 1/2, is an increasing function of ξ for $\xi \geq n/2$, and this means the monotonicity of the values

$$M_x(n/2+1) > M_x(n/2+2) > \cdots > M_x(n)$$

is maintained at least initially when p decreases below 1/2. When n is odd, the values $M_x(\xi)$ for $\lceil n/2 \rceil = (n+1)/2 \leq \xi \leq n$ also all decrease as p decreases from $p = 1/2$ in such a way that the monotonicity

$$M_x((n+1)/2) > M_x((n+1)/2+1) > \cdots > M_x(n)$$

is maintained.

For $\xi < n/2$ we see $f'(1/2) < 0$, so the value $M_x(\xi)$ will initially increase when p decreases from $p = 1/2$. We have not shown it, but for $0 < \xi < n/2$ as p decreases further this initial increase results in a unique maximum value, and then $M_x(\xi)$ decreases monotonically to $M_x(\xi) = 0$ as p approaches $p = 0$. In fact, the limiting value is clear:

$$\lim_{p \searrow 0} M_x(\xi) = \lim_{p \searrow 0} \left(\begin{array}{c} n \\ \xi \end{array} \right) p^{\xi} (1-p)^{n-\xi} = 0 \qquad \text{for } 0 < \xi < n/2,$$

and this (evidently) holds also for $0 < \xi \leq n$ as well.

The case $\xi = 0$ is exceptional. Notice that

$$M_x(0) = (1-p)^n.$$

This value not only increases initially when p decreases from 1/2 but continues to increase monotonically overtaking each of the values $M_x(\xi)$ for $0 < \xi < n$ with limit

$$\lim_{p \searrow 0} M_x(0) = 1.$$

The overall resulting behavior is that as p decreases from $p = 1/2$ to $p = 0$, the nonzero values

$$M_x(0), M_x(1), M_x(2), \ldots, M_x(n)$$

of M_x maintain a form somewhat similar to the binomial PMF with $p = 1/2$; the PMF is roughly bell shaped with maximum value(s) shifting left. An illustration for $n = 10$ should make this description more or less clear. In Figure 2.5 one finds (on the left) an illustration of the nonzero values of the

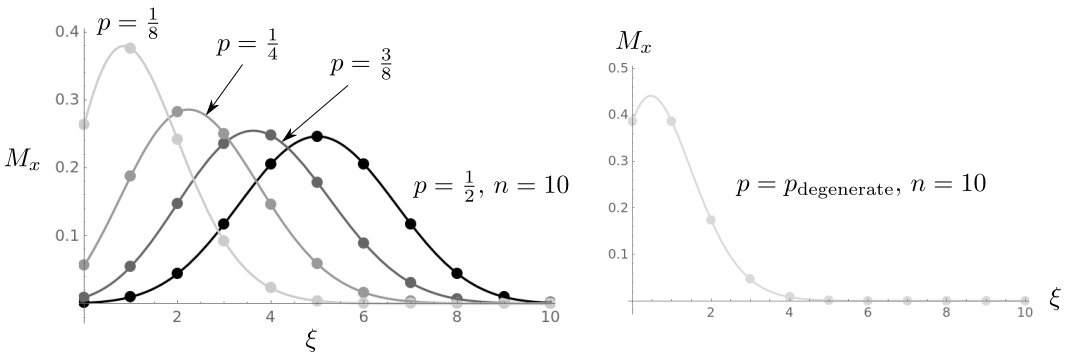

Figure 2.5: The PMF of the binomial distributions with $p = 1/2, 3/8, 1/4, 1/8$ and $n = 10$ (left). On the right is plot for $n = 10$ and the degeneracy value p_* for which the nonzero values have maximum on the left and left monotonicity for the rough bell shape is degenerate.

PMF for $n = 10$ trials in a binomial distribution for $p = 1/8, 1/4, 3/8$ and $p = 1/2$ (the symmetric case). Taking account of the limiting behavior, there is a unique value $p = p_{\text{degenerate}}$ for the probability in the Bernoulli trial for which $M_x(0) = M_x(1)$. Let us call this value $p = p_{\text{degenerate}}$ the **degeneracy value**. For $p < p_*$ the probability of utter failure (i.e., zero successes in 10 trials) is greater than the probability of a single success.

Exercise 2.4.6 What function is used to plot the smooth curve passing through the nonzero graph values of the PMF(s) in Figure 2.5?

Exercise 2.4.7 Find the degeneracy value $p = p_*$ for $n = 10$ corresponding to the PMF illustration on the right in Figure 2.5.

We can finish this chapter/lecture with an easy generalization of the crucial definitions of σ-algebra and measure.

2.5 Countably infinite measure spaces

As far as I know no human has ever had enough time, enough space, or enough resolution to observe anything in the real world corresponding to a (countably) infinite set. Among sets in mathematics[8] and sets of numbers in particular such sets are commonplace and, apparently, quite compelling. The prototypical such set is the set of natural numbers

$$\mathbb{N} = \{1, 2, 3, \ldots\}$$

about which there are immediately a great many questions to which no one knows the answer(s). There are, however, certain of these psychological questions to which it does appear the answers are known. One of these "answers" is the following:

> If $p_1, p_2, p_3, \ldots$ is a **countably infinite sequence of positive real numbers**, then the following hold:
>
> 1. Either the set of sums
>
> $$P = \left\{ \sum_{j \in N} p_j : N \subset \mathbb{N} \right\}$$
>
> is bounded above, i.e., there is some $M \in \mathbb{R}$ for which $s \leq M$ for every $s \in P$, in which case the sum of the sequence of numbers $p_1, p_2, p_3, \ldots$ is unambiguously defined as the least upper bound $L \in \mathbb{R}$ of the set P and we write
>
> $$\sum_{j=1}^{\infty} p_j = L,$$
>
> or the set P is not bounded above and we write
>
> $$\sum_{j=1}^{\infty} p_j = \infty.$$
>
> 2. In the case where
>
> $$\sum_{j=1}^{\infty} p_j < \infty,$$

[8] That is to say, among constructions existing in the minds of certain humans.

all sums of the form

$$\sum_{j\in N} p_j$$

where N is some fixed subset of $\mathbb{N}$ are unambiguously well-defined as real numbers and satisfy the following innocuous condition:

> If $N_1, N_2, N_3, \ldots$ is a sequence of pairwise disjoint subsets of $\mathbb{N}$, then
>
> $$\sum_{k=1}^{\infty}\left(\sum_{j\in N_k} p_j\right) = \sum_{j\in\cup_{k=1}^{\infty} N_k} p_j.$$

These known assertions about countable collections of positive numbers may be called the **fundamental facts of positive sequences** and they readily generalize to sequences of nonnegative real numbers. The last conclusion, which includes the possibility that some (and possibly infinitely many) of the sets N_k are empty, makes the definitions we present below "easy."

Example 11 The sequence

$$\frac{1}{2},\ \frac{1}{4},\ \frac{1}{8},\ \cdots$$

or to be more precise the sequence $\{p_j\}_{j=1}^{\infty}$ with

$$p_j = \frac{1}{2^j}$$

is a sequence with

$$\sum_{j=1}^{\infty}\frac{1}{2^j} = 1.$$

As phrased above, the order of summing over finite subsets of $\mathbb{N}$ does not make any difference in this case. It is usual, however, to adopt the order suggested by an ordering of the sequences and consider **partial sums** so that

$$\sum_{j=1}^{\infty} p_j = \lim_{k\to\infty}\sum_{j=1}^{k} p_j.$$

The sequence $\{h_j\}_{j=1}^{\infty}$ with

$$h_j = \frac{1}{j}$$

is called the **harmonic sequence** and has

$$\sum_{j=1}^{\infty} h_j = \infty.$$

Exercise 2.5.1 Prove the harmonic sequence does not have a finite sum.

We define first a measure on a set $S = \{\omega_1, \omega_2, \omega_3, \ldots\}$ which is countably infinite. Such a measure space shares the following properties with the previously considered measure spaces:

1. The domain of the measure is the entire power set of S.
2. The values can be defined in terms of the values on the singleton set Ω.

Generally, there is associated with each such measure space a sequence determined by the values $\mu(\{\omega_j\})$.

Definition 6 (countably infinite measure space) Given a set S with $\#S = \aleph_0$, an **adolescent measure** on S is a function

$$\mu : \wp(S) \to [0, \infty)$$

satisfying the properties:

(i) $\mu(\phi) = 0$.

(ii) If $A_1, A_2, A_3, \ldots$ is a sequence of (pairwise) disjoint sets in S, then

$$\mu\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mu(A_j).$$

Again property **(ii)** in this definition is called (countable) **additivity**, and it is the most prominent feature of a measure. According to property **(ii)** the measure of every set $A \subset S$ is given by the formula

$$\mu(A) = \sum_{\omega \in A} \mu(\{\omega\}).$$

Thus, the values of μ are all determined by the values of μ restricted to the singleton set

$$\Omega = \{\ \{\omega\} : \omega \in S\ \}.$$

In particular, there is associated with each enumeration[9] $\{\omega_j\}_{j=1}^{\infty}$ the measure space S a sequence determined by

$$p_j = \mu(\{\omega_j\}) \qquad \text{and} \qquad \mu(S) = \sum_{j=1}^{\infty} p_j.$$

Finally, we recall Definition 5 from section 2.2.2 above, and we have a very similar formulation for adolescent measures.

Definition 7 Given any set S, a **generalized adolescent measure** on S is a function

$$\mu : \wp(S) \to [0, \infty)$$

satisfying the following properties:

(0) There is a countably infinite set $S_0 = \{\omega_1, \omega_2, \omega_3, \ldots\} \subset S$ such that

$$\mu(S \backslash S_0) = 0.$$

(i) $\mu(\phi) = 0$.

(ii) If $A_1, A_2, A_3, \ldots$ is a sequence of (pairwise) disjoint sets in $\mathfrak{M}$, then

$$\mu\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mu(A_j).$$

2.5.1 Geometric (series) measure

If we take an infinite sequence $\{p_j\}_{j=1}^{\infty}$ of positive terms with a finite sum, like the sequence,

$$\frac{1}{2},\ \frac{1}{4},\ \frac{1}{8},\ \cdots \tag{2.21}$$

considered above, then we can assign measure values

$$\mu(\{\omega_j\}) = p_j$$

[9] An **enumeration** of a countably infinite set S is a bijection $\nu : \mathbb{N} \to S$, and the values $\nu(j)$ of the bijection are often denoted by ν_j for $j = 1, 2, 3, \ldots$.

on the singleton set

$$\Omega = \{\ \{\omega_1\},\ \{\omega_2\},\ \{\omega_3\},\ \dots\ \}$$

associated with any countably infinite set

$$S = \{\omega_1, \omega_2, \omega_3, \dots\}.$$

Extending μ to $\wp(S)$ by

$$\mu(A) = \sum_{\omega \in A} \mu(\{\omega\}) \tag{2.22}$$

makes S into a countably infinite measure space with measure $\mu : \wp(S) \to [0, \infty)$. When we have a new concept like the concept of a countably infinite measure, we should check to see if the associated concepts and intuition we have developed in simpler contexts still apply. Fortunately, in this case almost everything still works very much like in the case of a measure space with finitely many elements.

For one thing, we can still define a **probability measure on a countably infinite measure space** to be one satisfying the first fundamental deception

$$\mu(S) = 1.$$

This is the case for the measure obtained using the sequence (2.21).

A standard generalization is the following: With

$$S = \mathbb{N}_0 = \{0, 1, 2, 3, \dots\}$$

the set of **natural numbers with zero** and a number $p \in [0, 1]$ we make the singleton assignment

$$\pi(\{j\}) = p(1-p)^j.$$

When $p > 0$, this turns out to be a sequence with sum 1, and consequently the measure $\pi : \wp(\mathbb{N}_0) \to [0, 1]$ constructed using the using the usual extension formula (2.22) is a probability measure. To find the sum of the singleton assignments, it is helpful to know something about geometric series. If you do not know about geometric series, the following may seem mysterious. Starting with a partial sum

$$\sum_{j=0}^{k} p(1-p)^j = p \sum_{j=0}^{k} (1-p)^j$$

we write p as $p = 1 - (1 - p)$. Then the partial sum can be written as

$$\sum_{j=0}^{k} p(1-p)^j = [1-(1-p)]\sum_{j=0}^{k}(1-p)^j = \sum_{j=0}^{k}(1-p)^j - \sum_{j=0}^{k}(1-p)^{j+1}.$$

Separating the first term from the first sum and the last term from the last sum

$$\sum_{j=0}^{k} p(1-p)^j = 1 + \sum_{j=1}^{k}(1-p)^j - \sum_{j=0}^{k-1}(1-p)^{j+1} - (1-p)^{k+1},$$

and the remaining sums are the same number. (Just shift the indices in the second sum up by one and the power down by one.) Therefore, the partial sum takes the nice "closed" form

$$\sum_{j=0}^{k} p(1-p)^j = 1 - (1-p)^{k+1}.$$

This expression always has a limit. If $p = 0$, the limit is 0; in fact, the sequence of partial sums is identically the zero sequence, so we do not get a probability measure. If $p = 1$, then the sequence of partial sums takes identically the value 1, and the limit is 1.

Perhaps the most interesting case is when $0 < p < 1$. In this case, $0 < 1 - p < 1$,

$$\lim_{k\to\infty}(1-p)^{k+1} = 0 \qquad \text{and} \qquad \sum_{j=0}^{k} p(1-p)^j = \lim_{k\to\infty}\sum_{j=0}^{k} p(1-p)^j = 1$$

once again.

Thus, for $0 < p \leq 1$ we have an adolescent probability measure $\pi : \wp\mathcal{O}(\mathbb{N}_0) \to [0,1]$. In the language of the Bernoulli and binomial distributions, the value $\pi(\{n\})$ is said to be the probability of n failures before the first success when undertaking repeated Bernoulli trials with probability p. For example, when flipping a fair coin ($p = 1/2$) this is the probability of getting n "tails" before getting the first "head." Getting a "head" on the first coin flip is assigned probability $p = 1/2$.

We introduce, as usual, a real injection $\mathrm{id}_{\mathbb{N}_0} : \mathbb{N}_0 \to \mathbb{N}_0 \subset \mathbb{R}$ to obtain an induced (generalized adolescent) measure

$$\gamma : \wp\mathcal{O}(\mathbb{R}) \to [0,1] \qquad \text{by} \qquad \gamma(A) = \pi(A \cap \mathbb{N}_0).$$

The measure γ is called the **geometric (series) measure**. The concept of a PMF and CMF carries over readily to this situation with the PMF $M : \mathbb{R} \to [0, 1]$ of the geometric series measure given by

$$M(\xi) = \begin{cases} p(1-p)^{\xi}, & \xi \in \mathbb{N}_0 \\ 0, & \xi \in \mathbb{R} \backslash \mathbb{N}_0. \end{cases}$$

Exercise 2.5.2 Show the CMF $F : \mathbb{R} \to [0, 1]$ of the geometric distribution is given by

$$F(\xi) = \begin{cases} \sum_{n=0}^{\lfloor \xi \rfloor} p(1-p)^n, & \xi \geq 0 \\ 0, & \xi < 0 \end{cases}$$

where $\lfloor \xi \rfloor$ denotes the "floor" function of ξ giving the greatest integer less than or equal to ξ.

Exercise 2.5.3 Use a renaming of the geometric probability distribution to give a probability measure on the set $\mathbb{N}$ giving the number of coin flips required to obtain a first "head," in other words, the total number of attempts instead of just the number of failures.

The PMF and CMF of the geometric distribution are illustrated in Figure 2.6.

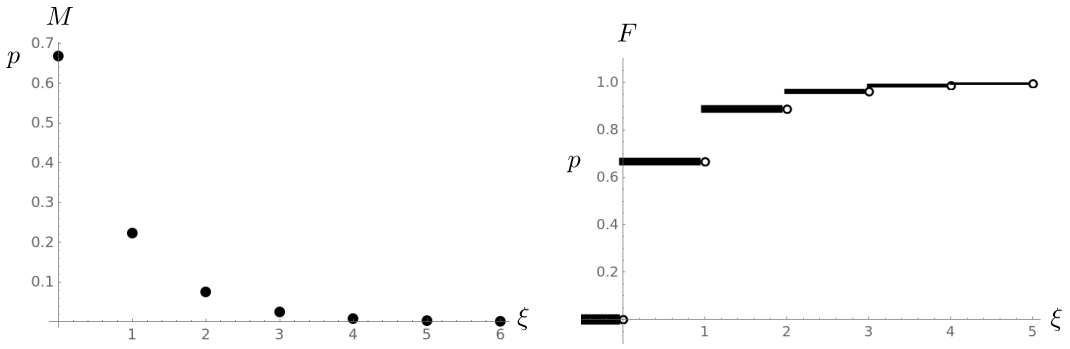

Figure 2.6: The PMF of the gometric distribution with $p = 2/3$ (left) and the corresponding CMF (right). The bell shape in this distribution is degenerate.

It will be noted that compared to the rough bell shaped binomial distributions, the geometric distribution may be said to have degenerate form with

a maximum at the left endpoint of the range and monotonically decreasing nonzero values to the right. There are certain situations in which a more standard form bell shaped distribution with nonzero values increasing to a maximum and decreasing to the right is desirable. We will explain this more fully in the next section and give an example on a countably infinite measure space.

2.6 The Poisson distribution

Naturally, the geometric probability p with $0 < p < 1$ of initial success is greater than the probability $p(1-p)^n$ of any particular number n of failures before the first success.

Exercise 2.6.1 Characterize the situations under which the probability p of initial success assigned by the geometric probability measure is greater than the probability of any positive number of failures before a/the first success.

If a specific time window for observation and a specific observation area are specified and the observed entities are moving in and out of a the specified observation area over time, then one may wish to assign a smaller probability to the observation of a smaller number of entities than some larger one. This may be accomplished using a **Poisson measure**.

As with other measures on countably infinite measure spaces, we start with a sequence of values

$$\pi(\{n\}) = e^{-\lambda}\ \frac{\lambda^n}{n!} \qquad \text{for} \qquad n = 0, 1, 2, 3, \ldots .$$

From this we determine a one parameter family of probability measures $\pi : \wp(\mathbb{N}_0) \to [0, 1]$ where the parameter λ is any positive real number. Each is, of course, an adolescent Poisson measure. To see $\pi(\mathbb{N}_0) = 1$, we recall the power series for the exponential function is

$$e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}.$$

This series converges to the value of the exponential for any real number z, and in fact, for any complex number z. In particular, for $\lambda > 0$,

$$\sum_{n=0}^{\infty} \pi(\{n\}) = \sum_{n=0}^{\infty} e^{-\lambda}\frac{\lambda^n}{n!} = e^{-\lambda}\sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = e^{-\lambda}e^{\lambda} = 1.$$

Using the same **real injection** giving the geometric series measure, namely,

$$\mathrm{id}_{\mathbb{N}_0} : \mathbb{N}_0 \to \mathbb{R} \qquad \text{by} \qquad \mathrm{id}_{\mathbb{N}_0}(n) = n,$$

we obtain the induced measure $\alpha : \wp(\mathbb{R}) \to [0,1]$ which is a generalized adolescent measure on $\mathbb{R}$ and we can calculate the PMF $M : \mathbb{R} \to [0,1]$:

$$M(\xi) = \begin{cases} e^{-\lambda}\frac{\lambda^{\xi}}{\xi!}, & \xi \in \mathbb{N}_0 \\ 0, & \xi\mathbb{R}\backslash\mathbb{N}_0. \end{cases}$$

The CMF of the Poisson measure is given by

$$F(\xi) = \begin{cases} e^{-\lambda}\sum_{n=0}^{\lfloor \xi \rfloor}\frac{\lambda^n}{n!}, & \xi \geq 0 \\ 0, & \xi < 0. \end{cases}$$

Exercise 2.6.2 Plot the PMF and CMF of the Poisson distribution.

Hopefully the example on the next page will give you some kind of idea how the Poisson distribution might be used.

Example 12 Say a certain kind of bird lives in relatively large numbers in a particular forest. A person sits each day for ten days in one location in the forest and records each day how many of this kind of bird he sees. The numbers[10] he records are indicated in Table 2.1 below. This is a **data set**.

Table 2.1: number of hummingbirds observed per day July 5-14, 2016

day	1	2	3	4	5	6	7	8	9	10
no. birds	96	96	85	104	105	82	99	87	83	102

For whatever reason, a statistician wishes to find a data set using the statistics program R giving numbers of birds someone might think are actually observed in the same forest, e.g., these numbers might be the numbers of birds observed by the same person over the next seven days July 15-19, 2016. The statistician uses the command `rpois(7,`λ`)` which provides numbers that look like "random" selections chosen according to a Poisson distribution with parameter λ and obtains[11] the values recorded in Table 2.2.

Table 2.2: simulated number of hummingbirds observed per day July 15-19, 2016

day	1	2	3	4	5	6	7
no. birds	91	101	93	81	96	85	88

The statistician uses the commands

```
> set.seed(120)
> rpois(7,λ)
```

but he needs to determine the value λ from the initial data set in Table 2.1. Can you figure out the value of λ he used? Hint: Calculate

$$\int_{\xi\in\mathbb{R}} \xi$$

with respect to the Poisson measure.

[10] `set.seed(126)`

[11] `set.seed(120)`

2.7 Summary

We have now introduced several standard measures and probability measures:

1. uniform measure and uniform probability measure
2. counting measure
3. Bernoulli measure
4. binomial measure and binomial induced measure
5. geometric (series) measure
6. Poisson measure

We have also introduced a general framework: Given a probability measure $\pi : S \to [0, 1]$ on a measure space S (either with $\#S < \infty$ or $\#S = \aleph_0$) we may encounter the following:

1. a renaming function $a : S \to T$,
2. an inducing function $x : S \to \mathbb{R}$, and
3. an induced (probability) measure $\alpha_x : \wp(\mathbb{R}) \to [0, 1]$.

Associated with the induced measure is a probability mass function (PMF)

$$M_x : \mathbb{R} \to [0, 1]$$

and a cumulative mass function (CMF)

$$F_x : \mathbb{R} \to [0, 1].$$

Together this construction constitutes a "distribution."

We have not yet formulated what is meant by a general measure on $\mathbb{R}$ or an interval in $\mathbb{R}$, but our discussion is restricted to baby and adolescent measures.

2.8 General expansion principle

We have introduced two restriction measure above associated with a subset T of a measure space S, a basic restriction measure and a probability restriction measure. The probability restriction measure associated with a subset T of a measure space S with measuare $\mu : \wp(S) \to [0, \infty)$ is of particular interest. Specifically, if $\mu(T) > 0$, then we take

$$\rho_T(A) = \frac{\mu(A)}{\mu(T)} \qquad \text{for any } A \subset T$$

and more generally, we can consider ρ_T as a measure on the larger space S by the **expansion formula**

$$\rho_T(A) = \frac{\mu(A \cap T)}{\mu(T)} \qquad \text{for any } A \subset S.$$

This expansion formula works more generally:

Theorem 8 (general expansion formula) Given any measure $\alpha : \wp(T) \to [0, \infty)$ and a superset $S \supset T$, the measure $\beta : \wp(S) \to [0, \infty)$ by

$$\beta(A) = \alpha(A \cap T) \qquad \text{for } A \subset S$$

is called the **expansion** of α. If α is a probability measure, then so is any expansion of α.

Exercise 2.8.1 If β is an expansion of a measure α from the measure space T to the superset S, then show

$$\beta(A) = 0 \qquad \text{for any } A \subset S \text{ with } A \cap T = \phi.$$

Chapter 3

Lecture 3: Counting and Probability

We have discussed countably infinite sets above, that is sets with the same cardinality as the natural numbers $\mathbb{N}$. Though countably infinite sets are called "countable," this does not mean one can determine the number (in the sense of a single natural number) of elements in such a set. In fact, some people prefer to call countably infinite sets **denumerable** meaning simply that a natural number may be assigned to each element rather than to suggest that all the elements can be counted. We return to sets with finite cardinality in this chapter/lecture and discuss some techniques for counting the number of elements in various sets. We have already considered some preliminary counting techniques in our discussion of the binomial distribution in section 2.3 above. We will also review and expand the discussion of counting related to the binomial distribution.

Products are perhaps the simplest sets to count: If $\#A = n \in \mathbb{N}$ and $\#B = m \in \mathbb{N}$, then the set of ordered pairs $A \times B = \{(a, b) : a \in A, b \in B\}$ has finite cardinality and

$$\#(A \times B) = (\#A)(\#B) = nm. \tag{3.1}$$

If A and B are measure spaces with $\pi_1 : \wp(A) \to [0, 1]$ and $\pi_2 : \wp(B) \to [0, 1]$ probability measures on A and B respectively, then there is a measure on the Cartesian product $A \times B$ given by a formula you might expect:

$$\pi : \wp(A \times B) \to [0, 1] \qquad \text{by} \qquad \pi(\{(a, b)\}) = \pi_1(\{a\})\ \pi_2(\{b\}). \tag{3.2}$$

At this point, and before addressing other questions about (3.2) which are taken up below, it is perhaps natural to ask the following question(s):

1. What is the cardinality of $\wp(A \times B)$?

2. More generally, if the power set of a set with finitely many elements always a set with finitely many elements, and if so, given S with $\#S < \infty$ what is $\#\wp(S)$?

We will answer these questions very soon, but it turns out there is a simpler counting technique it will be convenient to address first.

3.1 Permutations

The multiplication principle of counting expressed by (3.1) may be extended by induction as follows: If $\#S_j = n_j \in \mathbb{N}$ for $j = 1, 2, \ldots, k$, then

$$\# \left(\prod_{j=1}^{k} S_j \right) = n_1 n_2 \cdots n_k. \tag{3.3}$$

This may be heuristically interpreted as follows: Each point $(\tau_1, \tau_2, \ldots, \tau_k) \in \prod_{j=1}^{k} S_j$ admits n_j "choices" for the entry τ_j. For example, there are n_1 choices for τ_1, and each of these choices may be "paired" with any point $(\tau_2, \ldots, \tau_k) \in \prod_{j=2}^{k} S_j$. Consequently, since by (heuristic) induction

$$\# \left(\prod_{j=1}^{k-1} S_j \right) = n_2 \cdots n_{k-1},$$

we see how the product in (3.3) arises or is constructed.

We can use this same construction to count the number of elements in a certain subset of $\prod_{j=1}^{k} S = S^k$ for any single set S with $\#S = n$. The set we have in mind is

$$D = \left\{ (\tau_1, \tau_2, \ldots, \tau_k) \in \prod_{j=1}^{k} S : \tau_i \neq \tau_j \text{ when } i \neq j \right\}. \tag{3.4}$$

This is the subset of S^k containing points with distinct entries from the set S. Again, there are n choices for the entry τ_1. For each such choice, however,

there are only $n-1$ choices for τ_2. Executing the more or less obvious heuristic induction we obtain

$$\#D = n(n-1)\cdots(n-k+1). \tag{3.5}$$

This number is called the **permutation of n elements taken k at a time** or simply the **permutation of n taken k** or $P(n,k)$ for short. Writing

$$n(n-1)\cdots(2)(1) = n! \qquad \text{and} \qquad k(k-1)\cdots(2)(1) = k!$$

as usual,

$$P(n,k) = \#D = \frac{n!}{k!}.$$

It may be noted that this is also the number of ways to arrange k (distinct) elements in order from among the elements of a set having n elements. This is essentially equivalent to specifying a point in the set D given in (3.4).

While counting using permutations is heuristically simpler than some other counting techniques we will encounter, one should not get the impression that the permutation counting formula $\#D = P(n,k)$ is simple. In fact, what we have described above is somewhat subtle and interesting both analytically (or algebraically) and geometrically. First of all, one may make a distinction between mathematical counting, in which one obtains a bijection between a given set S and a second set of the form $\{1,2,\ldots,\#S\}$ and heuristic counting in which one asserts the existence of such a bijection without actually obtaining one.

It may be suspected in fact that we have not obtained such a bijection even in the case of a Cartesian product (3.1). This is not too difficult, and to illustrate this I will now outline the details. The process is inductive: As a base case we consider $\#B = 1$ with $B = \{b\}$. In this case we obtain a bijection

$$\psi : A \times B \to \{1,2,\ldots,\#A\} \qquad \text{by} \qquad f(a,b) = a$$

where $f : A \to \{1,2,\ldots,\#A\}$ is a bijection enumerating A. For the inductive step, we begin with bijections $f : A \to \{1,2,\ldots,\#A\}$ and $g : B \to \{1,2,\ldots,\#B\}$ where $\#B = n+1$. Then $B\backslash\{g^{-1}(n+1)\}$ is a set with n elements. Thus, we have by induction a bijection

$$\psi_n : A \times [B\backslash\{g^{-1}(n+1)\} \to \{1,2,\ldots \#A\,(n)\}.$$

The function $\psi : A \times B \to \{1, 2, \ldots, \#A\#B\}$ by

$$\psi(a,b) = \begin{cases} \psi_n(a,b), & \text{if } b \neq g^{-1}(n+1) \\ \#A\#B + f(a), & \text{if } b = g^{-1}(n+1) \end{cases} \tag{3.6}$$

is relatively easily seen to be a bijection.

Exercise 3.1.1 Consider the function ψ defined inductively in (3.6).

(i) Show ψ is a bijection.

(ii) Draw a picture illustrating the definition (3.6).

Exercise 3.1.2 Use the counting formula (3.1) and induction to obtain a bijection

$$\psi : \prod_{j=1}^{n} S_j \to \{1, 2, \ldots, (\#S_1) \cdots (\#S_n)\}.$$

In contrast to the bijections obtained above for various product counting formulas, justification of the heuristically obtained permutation counting formula $\#D = P(n,k)$ as a mathematical counting formula is not so straightforward.

A potential base step for induction on the dimension k starting with $k = 2$ seems relatively straightforward: Given $n \geq 2$

$$\psi : \{(a_1, a_2) \in \{1, 2, \ldots, n\}^2 : a_1 \neq a_2\} \to \{1, 2, \ldots, n\} \times \{1, 2, \ldots, n-1\}$$

by

$$\psi(a_1, a_2) = \begin{cases} (a_1, a_2), & \text{if } a_2 < a_1 \\ (a_1, a_2 - 1), & \text{if } a_1 < a_2 \end{cases}$$

Is a well-defined bijection, and by the counting principle for products the image set has cardinality $n(n-1)$.

Given $n \geq k = 3$, I invite you to write down a bijection from

$$D = \{(a_1, a_2, a_3) \in \{1, 2, \ldots, n\}^3 : a_i \neq a_j, i \neq j\}$$

to

$$\{1, 2, \ldots, n\} \times \{1, 2, \ldots, n-1\} \times \{1, 2, \ldots, n-2\}.$$

Even in the case $n = 3$, the images of the six resulting points are relatively complicated to write down. Geometrically, these six points $(1, 2, 3)$, $(1, 3, 2)$,

$(2,1,3)$, $(2,3,1)$, $(3,1,2)$, $(3,2,1)$ lie on a circle in the plane $L = \{(a_1, a_2, a_3) : a_1 + a_2 + a_3 = 6\}$ with center $(2,2,2)$ and radius $\sqrt{2}$. As one might expect, these six points are the vertices of a regular hexagon in the plane L.

As the dimension increases, the situation becomes even more interesting. Taking $n = k = 4$ there are 24 points with distinct coordinates in $\{1,2,3,4\}^4 \subset \mathbb{R}^4$ which may be put into one-to-one correspondence with the (product enumerable) set

$$\{1,2,\ldots,4\} \times \{1,2,3\} \times \{1,2\} \times \{1\}$$

or simply $\{1,2,\ldots,4\} \times \{1,2,3\} \times \{1,2\}$. The original 24 points lie on a 2-sphere in the hyperplane

$$L = \{(a_1, a_2, a_3) : a_1 + a_2 + a_3 + a_4 = 10\}$$

with center $(5/2, 5/2, 5/2, 5/2)$ and radius $\sqrt{5}$. The arrangement of these points in the sphere is quite interesting; note there is no platonic solid with 24 vertices.

One conclusion/observation of this section is that one can "count" or "heuristically count" the number of elements in sets that are (too) complicated to mathematically count, in the sense that one can write down the cardinality but it is inconvenient to write down the bijection. Most of the counting considered below is heuristic rather than mathematical.

3.2 Combinations

It doesn't really make sense to consider **ordered** k-tuples "without order," but the equivalent heuristic formulation given in the last section concerning arrangements of elements does admit a natural modification. That is, one can consider the number of ways to arrange k (distinct) elements from among the elements of a set having n elements (where the order of the arrangement does not matter). Again, this is not essentially equivalent to specifying some set, or subset, of ordered k-tuples, but it is equivalent to counting some collection of (sub)sets in $\wp(S)$ the power set of a given set S. Specifically, an unordered collection of k (distinct) elements from a set S of cardinality $n > k$ is the same thing as the subset $A \subset S$ containing those k elements and having $\#A = k$. That is, we are now going to count the number of elements in

$$K = \{A \subset S : \#S = k\}. \tag{3.7}$$

In the case $k = 1$, these are the singleton sets, and there are (clearly) n of them, namely, the sets in

$$\Omega = \{\ \{\omega_1\},\ \{\omega_2\},\ \ldots,\ \{\omega_n\}\ \}$$

where

$$S = \{\omega_1, \omega_2, \ldots, \omega_n\}.$$

For other values of k, the cardinality of the set K defined in (3.7) is called the **combination of n elements taken k at a time** or the **combination of n taken k** or simply $C(n, k)$. Notice the set D appearing in (3.4) is very different from the set K under consideration now. Any element of the set K takes the form

$$A = \{\tau_1, \tau_2, \ldots, \tau_k\}$$

where $\tau_1, \tau_2, \ldots, \tau_k \in S$ and $\tau_i \neq \tau_j$ for $i \neq j$. Given such a set A, we can find an element $(\tau_1, \tau_2, \ldots, \tau_k) \in D \subset S^k$ with entries the elements in A. But this assignment is not unique. For example, the element $(\tau_2, \tau_1, \ldots, \tau_k) \in D$ also has entries the elements of A. On the other hand, every element of S^k with distinct entries does determine a set $A \in K$. They key questions is this: How many elements in $D \subset S^k$ give the same element $A = \{\tau_1, \tau_2, \ldots, \tau_k\} \in K$? The answer is the number of different ways the elements $\tau_1, \tau_2, \ldots, \tau_k$ may be arranged in the k entries of a point in S^k or the permutation of k things taken k at a time. According to (3.5) with $n = k$ this number is simply $k!$. Thus, each element in K corresponds to $k!$ elements in the general set D of (3.4). In order to count the number of elements in K, we we can take the number of elements in D and divide by $k!$. That is,

$$\#K = \frac{P(n, k)}{k!} = \frac{n!}{(n-k)!k!}.$$

This is $C(n, k)$ the combination of n taken k. In section 2.3 we introduced the notation

$$\left(\begin{array}{c} n \\ k \end{array} \right)$$

for the combination $C(n, k)$ and used both permutations and combinations to count the elements in another set in giving a detailed proof of the binomial formula; see especially (2.11) and (2.12). There is no comparable notation for the permutation of n taken k, but the alternative notations

$$_nC_r \qquad \text{and} \qquad _nP_r$$

for combinations and permutations respectively are sometimes used.

The formula we have obtained for the number of subsets in the power set of a given set S containing a specified number of elements can be used to find the cardinality of $\wp(S)$ when $\#S = n$ in general:

$$\#\wp(S) = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right).$$

Notice that we have already proved the binomial formula (2.12) in detail and taking $a = b = 1$ in $(a+b)^n$ we get

$$\sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) \; 1^k \; 1^{n-k} = (1+1)^n.$$

Therefore

$$\#\wp(S) = 2^{\#S}.$$

Before expanding our discussion to nominally address the question of counting real world collections of objects, I will try to repeat the counting discussion above in a seemingly different manner. We consider the two sets

$$D = \left\{ (\tau_1, \tau_2, \ldots, \tau_k) \in \prod_{j=1}^{k} S : \tau_i \neq \tau_j \text{ when } i \neq j \right\}.$$

and

$$K = \left\{ \; \{\tau_1, \tau_2, \ldots, \tau_k\} \in \wp(S) : \tau_i \neq \tau_j \text{ when } i \neq j \; \right\}.$$

The set D is the set of ordered k-tuples with distinct entries from S considered in (3.4) above. The set K is the collection of subsets of S having k distinct elements considered in (3.7) above. In both cases we may assume

$$S = \{\omega_1, \omega_2, \ldots, \omega_n\}$$

with $\#S = n$. One thing I note is that given a subset

$$A = \{\tau_1, \tau_2, \ldots, \tau_k\} \in K$$

with $\#A = k$ there are a number of k-tuples $p \in D$ I can construct with the components $\tau_1, \tau_2, \ldots, \tau_k$. In fact, the number of such k-tuples I can construct using $\tau_1, \tau_2, \ldots, \tau_k$ is

$${}_kP_k = k!.$$

(This is the number of distinct k-tuples in $\{\tau_1, \tau_2, \ldots, \tau_k\}^k$, and counting them is the special case of our discussion of (3.4) when $\#S = k$.)

If the number of such subsets $A \in K$, namely $\#K$ is denoted by C, then the cardinality of D is given by the product

$$C\ k!.$$

That is, $\#D = C\ k! = k!\ \#K$ or

$$C = \#K = \frac{\#D}{k!}.$$

On the other hand, $\#D$ in the general case of (3.4) above was found to be the permutation of n taken k, so

$$C = \#K = \frac{P(n,k)}{k!},$$

and this is the number we take as the definition of the combination of n taken k.

3.3 Inclusion/exclusion principle of counting

Sometimes the following simple principle of counting is useful:

If A and B are any sets, then

$$\#(A \cup B) = \#A + \#B - \#(A \cap B).$$

This may be derived by an even simpler principle, the counting **principle of addition**:

If A and B are **disjoint** sets, then

$$\#(A \cup B) = \#A + \#B.$$

Thus, in general since $A \backslash B = A \backslash (A \cap B)$ and $A \cap B$ are disjoint with $(A \backslash B) \cup (A \cap B) = A$, we have

$$\#A = \#(A \backslash B) + \#(A \cap B).$$

Similarly,

$$\#B = \#(B\backslash A) + \#(A \cap B).$$

Therefore,

$$\#A + \#B = \#(A\backslash B) + \#(B\backslash A) + 2\#(A \cap B).$$

On the other hand the three sets $A\backslash B$, $B\backslash A$, and $A \cap B$ are also disjoint with $(A\backslash B) + (B\backslash A) + (A \cap B) = A \cup B$, so

$$\#(A \cup B) = \#(A\backslash B) + \#(B\backslash A) + \#(A \cap B) = \#A + \#B - \#(A \cap B)$$

which is the **inclusion/exclusion principle** of counting.

On the other hand, the counting principle of addition follows as a special case of the inclusion/exclusion principle when $A \cap B = \phi$.

The inclusion/exclusion principle is especially useful when one is given precisely the information appearing within it.

Example 13 In a drum and tuba marching band 15 people play drums, 25 people play tuba, and 7 people play both (at the same time). In this case we know there are

$$15 + 25 - 7 = 23$$

people in the band.

3.4 Modeling with sets

There may seem to be no ambiguity in the previous example, but technically the counting principles above apply only to sets. While one may informally refer to sets of people, and sets of band members, and sets of tubas and such things, I suggest a strict distinction be made between sets as mathematical objects and real world "collections" which often entail a greater degree of ambiguity. The idea here is probably rather new to you and is a little bit subtle.

3.4.1 Models in counting

The suggestion is that in the restricted sphere of human psychological constructions which are considered mathematical, starting especially with the

notion of set and number—where numbers are particular kinds of sets also—there appears to be less ambiguity than in the consideration of "real world" phenomena, objects, events, and so on. Thus, **mathematics** of which subjects like calculus, set theory, algebra, etc. are examples is comprised of "imaginary pictures" or abstractions. The discipline of **applied mathematics** involves primarily the comparison of these abstractions with real world phenomena. It is true that one can be sloppy and say that a "hand of cards" (in blackjack) is a "set" of two cards from a deck of 52 cards and "count" the number of such "hands" to obtain

$$ {}_{52}C_2 = \frac{52(51)}{2} = 1326 \text{ hands.} \tag{3.8} $$

Technically, however, cards are cards which are real world objects and counting using combinations applies to calculating the cardinalities of certain sets which are mathematical objects. In this case, one can say the calculation given in (3.8) is not a mathematical calculation or that some details are left out at the very least. The proper way to proceed would be something like this: We model a deck of cards using the set

$$ \begin{aligned} \mathcal{C} &= \{1, 2, 3, \ldots, 13\} \times \{1, 2, 3, 4\} \\ &= \left\{ \ (r, s) : r \in \{1, 2, 3, \ldots, 13\} \text{ and } s \in \{1, 2, 3, 4\} \ \right\} \end{aligned} $$

where the first coordinate corresponds to the **rank** of a card (1 corresponds to ace and 11 corresponds to jack for example) and the second coordinate corresponds to the **suit** of the card (1 corresponds to spades, 2 to hearts, 3 to diamonds, and 4 to clubs). The set $\mathcal{C}$ is a mathematical object. The calculations

$$ \#\mathcal{C} = 52 \qquad \text{and} \qquad \#\{h \in \wp(\mathcal{C}) : \#h = 2\} =_{52} C_2 = 1326 $$

are mathematical calculations. Modeling the cards and the blackjack hands using the sets $\mathcal{C}$ and

$$ D_2 = \{h \in \wp(\mathcal{C}) : \#h = 2\} $$

is applied mathematics, especially when a comparison is made like (saying) there are 1326 possible hands in blackjack because the model set

$$ \{h \in \wp(\mathcal{C}) : \#h = 2\} \qquad \text{satisfies} \qquad \#\{h \in \wp(\mathcal{C}) : \#h = 2\} = 1326. $$

Naturally, this seems like a bit of overkill. This is the case partially because the mathematical modeling here is relatively elementary, but it is also the case because (many) people are used to being sloppy in this regard and ignoring the important distinction between a mathematical model and a real world object.

Let me suggest one possibility of why filling in the details can be worthwhile. Orloff and Booth (and many other texts) offer a counting principle for real world objects along the following lines:

> (rule of product) If an object can be constructed in a two step process in which the first step has n possible results and the second step has m possible results, then one can construct nm objects using this two step process.

Thus, we can construct a blackjack hand by dealing a first card from the deck (a step which has 52 possible results) and then dealing a second card from the same deck (a second step which always has 51 possibilities). In this way, we see

$$\text{There are } 52(51) = 2652 \text{ possible blackjack hands.} \qquad \text{(wrong)}$$

In effect what we have done is model the blackjack hands with the set

$$K_2 = \{(\omega_1, \omega_2) \in \mathcal{C}^2 : \omega_1 \neq \omega_2\}.$$

Indeed $\#K_2 = 2652$, but this set does not provide a very good model for blackjack hands. If we've done the mathematical modeling, then at least we have the opportunity to go back and make a comparison:

In K_2 the elements $((1,1),(1,2))$ and $((1,2),(1,1))$ are considered as distinct elements. These correspond to the "hand" consisting of the ace of spades and the ace of hearts, but in the "real world" the "hand" consisting of the ace of hearts (followed by) the ace of spades is considered the same blackjack hand. The model using (sets in) D_2 is a better model than the one using (ordered pairs in) K_2.

A famous example of failing to make such a distinction is in the modeling of populations, i.e., observed numbers of individual people, using the solution of an ordinary differential equation. By being sloppy and identifying the population (a real world quantity/object) with the (value of the) solution of the differential equation (a mathematical object, i.e., an imaginary picture) many

people made the mistake of predicting a "population explosion" or "doomsday scenario" which was not, as a matter of fact, realized. If the distinction between the mathematical model and the real world phenomenon had been maintained, and especially if the emphasis on comparison (and specifically the identification of **differences** between the two rather than identification) had been the focus, then certain individuals might have been saved from[1] participating in the the propagation of manifestly false predictions.

One consequence of the point of view presented above, when carried to its natural conclusion is that only mathematical probability (the theory of measures with total value one) involves mathematics while applied probability is based on the strict identification of a real world outcome with the value of a certain measure in such a manner that no comparison is possible. Thus, applied probability does not involve mathematics; it is not in particular applied mathematics—it is something else. Similarly, essentially all aspects of statistics are fundamentally disjoint from mathematics, though some calculations are involved—the discipline of statistics is also very much something else.

3.4.2 Models in probability

In applied probability, first of all, one does not model events. Events are real world phenomena which could be modeled, but the objective of probability is not (at all) to model them. For example, an abstract coin flip could be modeled. This would involve understanding what happens during the process of the coin flip in some way, but nothing of the sort is contemplated when one considers a coin flip as an event in probability. What is properly modeled, and mathematically modeled, in mathematical probability is the collection of **outcomes** of an event. The outcome of a single coin flip can be modeled by the set $S = \{0, 1\}$ with 1 (mathematical) corresponding to an outcome of "heads" (real world) and the (mathematical) element 0 corresponding to the outcome "tails." The outcome of ten coin flips taken in order may be modeled by $\{0, 1\}^{10}$.

I have suggested above, following A. Farmer, and I suggest here more

[1]Of course, one might entertain philosophical motivations for propagating manifestly false predictions, and mathematical modeling can be and certainly is largely used in this way. This, however, is generally not considered a proper use of mathematical modeling. Perhaps that assertion needs modification: At least a few people like myself do not consider such use proper, justified, or legitimate.

properly a refinement of the modeling of outcomes using sets. The actual known totality of possible concrete outcomes for an event is modeled by some set S, the **base set**. The abstract outcomes associated with a single event of the same kind are modeled by the **singleton set**

$$\Omega = K_1 = \{\ \{\omega\} : \omega \in S\ \}.$$

More generally, a variety of hypothetical (compound) outcomes may be modeled by the elements of the entire power set $\wp(S)$, and such sets are often described with the aid of particular real valued functions $x : S \to \mathbb{R}$, or more generally other functions on S.

The introduction of a measure $\mu : \wp(S) \to [0, \infty)$ as considered above is an element of the mathematical subject of measure theory. The restriction to measures $\pi : \wp(S) \to [0, 1]$ with $\pi(S) = 1$, called "probability measures," leads to the subject of mathematical probability. To be very precise, one can recognize the beginning of applied probability with the blurring of all distinctions between real world events, their outcomes, and various sets which might model them. Most often, one takes something like the the base set S (called a sample space or event space) and introduces something that looks mathematical called a **probability function** $P : S \to [0, 1]$. One then identifies the value/number

$$P(\omega) = \pi(\{\omega\})$$

as something conveying information about the occurrence of a **single event**. The problem is not that P is not a function, but the problem is a psychological one in that one is not offering something that is available for comparison, one is not participating in mathematical modeling, one is doing something else. More specifically, when one says

"The probability of a coin flip coming up heads is $P(h) = 1/2$,"

one is not offering something that can be compared to anything about the event of flipping a coin (once) or to the outcome of such an event. This is not mathematical modeling. It is something different. One benign, and perhaps the most common, explanation of what one is doing here is offering a psychological "stand-in" for actual understanding of the event and its outcome. One does not understand what is happening in the event in any particular way, and one cannot really say anything (especially make a prediction) of the

outcome of the event, but somehow attaching a numerical measure of this complete ignorance to the event is intended to "stand-in" for understanding. The particular value 1/2 should not present itself as a red herring in this discussion.

If I roll a six sided "fair" die, the contribution or position of applied probability is the following:

> I do not know what the outcome will be. In particular, I do not know if the outcome will be a one. I can model the outcomes using $S = \{1, 2, 3, 4, 5, 6\}$, I can introduce a uniform probability measure on this set, and I can say
>
> > "The probability the outcome will be one is
> >
> > $$P(1) = 1/6,"$$
>
> but I can make no claim that this number will compare in any particular way with the actual outcome.

Prediction is a hallmark of mathematical modeling, applied mathematics and science, and if the prediction disagrees with the physical phenomenon being modeled, then the model is deemed incorrect or innacurate. This does not apply in applied probability because applied probability is not any of these things. In short the statement

> "The probability the outcome will be one is $P(1) = 1/6$,"

is offered to give you (or someone else) the **perception** that something about the event and its outcome is understood when in fact that is not the case (because in fact, by any objective standard, there is total ignorance—and at length avowed total ignorance—concerning the process of the event and its outcome). It can be said, however, that there is a "market" for the calculation of such numbers.

Returning to the topic of counting, we have mathematical counting (the study of bijections) and informal counting with which, on the one hand, one must exercise some care in order to avoid "non-marketable" answers, i.e., incorrect answers, and on the other hand can sometimes be used to obtain cardinalities even when strict mathematical enumeration is quite difficult. We have seen some examples of this above, and most of the examples below fall into this category. The techniques of counting are of particular interest

when a uniform measure is introduced on $S = \{\omega_1, \omega_2, \ldots, \omega_n\}$. This will be illustrated below in the sections on dice, cards, and urns. Though things will become even more complicated, it is perhaps worth keeping in mind several informal counting rules

3.5 Informal Counting Rules

Here are some informal rules, like the one of Orlof and Booth concerning "constructions" above, that may be worth keeping in mind. Some care is required in application of these rules. They are phrased in terms of tasks, choices, arrangements, and objects, none of which is a precise mathematical object.

1. If two tasks (must) follow one another and there are n ways to perform the first task and m ways to perform the second task, once the first task has been performed, then there are mn ways to perform the two tasks (or the compound task of performing the first task (first) and then the second task).

2. If there are n choices for one thing and m choices for another, then there are mn choices for both things. Example: Three crusts and seven toppings makes 21 different kinds of pizzas.

3. If you want to choose k objects from among n objects (assuming $n \geq k$) and arrange the k objects you have chosen, then there are $P(n, k)$ possible arrangements.

4. If you want to choose k objects from among n distinct objects, then there are $C(n, k)$ possible choices.

3.6 Product measure (two factors)

We mentioned the product measure on the Cartesian product $A \times B$ above. It is clear that the **product measure** with values

$$\pi(\{(a,b)\}) = \pi_1(\{a\})\pi_2(\{b\})$$

takes values in $[0,1]$ and extends to a measure on $A \times B$. In fact, we have now established that $\wp(A \times B)$ is a set with finite cardinality and

$$\#\wp(A \times B) = 2^{\#A\#B},$$

so that we obtain the measure π by the usual extension formula

$$\pi(W) = \sum_{\omega \in W} \pi(\{\omega\}).$$

There are of course, many different kinds of sets in $\wp(A \times B)$. Among those sets are "product sets" having the form $U \times V$ for some $U \subset A$ and $V \subset B$. We can make the general observation that for a product set $U \times V \in \wp(A \times B)$ there holds

$$\pi(U \times V) = \pi_1(U) \times \pi_2(V).$$

To see this, we use the extension formula:

$$\begin{aligned}
\pi(U \times V) &= \sum_{(a,b) \in U \times V} \pi(\{(a,b)\}) \\
&= \sum_{(a,b) \in U \times V} \pi_1(\{a\})\ \pi_2(\{b\}) \\
&= \sum_{a \in U, b \in V} \pi_1(\{a\})\ \pi_2(\{b\}) \\
&= \sum_{a \in U} \sum_{b \in V} \pi_1(\{a\})\ \pi_2(\{b\}) \\
&= \sum_{a \in U} \pi_1(\{a\})\ \sum_{b \in V} \pi_2(\{b\}) \\
&= \pi_1(U)\ \pi_2(V).
\end{aligned}$$

In particular, it follows that

$$\pi(A \times B) = \pi_1(A)\ \pi_2(B) = 1,$$

so the product measure is a probability measure.

There can be many other measures on the product $A \times B$ and many other probability measures in particular. Let's look at some examples.

Example 14 Let $A = B = \{0, 1\}$. Then

$$S = A \times B = \{0, 1\}^2.$$

This is the natural measure space to model the outcome of two Bernoulli trials, i.e., coin flips. We can associate with the first factor the measure π_1 with $\pi_1(\{1\}) = p$ and with the second factor the measure π_2 with $\pi_2(\{1\}) = q$ where p and q satisfy $0 < p, q < 1$. The product measure is probably the one you would naturally associate with a single ordered outcome of the two flips. This is associated with the idea that the two flips are "independent."

It is possible however to make the outcome of the second flip dependent on the first flip. For example, if the first flip is a failure, then perhaps the ordered pair must also be considered a failure, or that is to say the second flip is irrelevant. There are four elements in the product giving sixteen elements in $\wp(A \times B)$. We can take a measure $\alpha : \wp(A \times B) \to [0, 1]$ with

$$\begin{aligned} \alpha(\{(0,0)\}) &= 1 - p \\ \alpha(\{(0,1)\}) &= 0 \\ \alpha(\{(1,0)\}) &= p(1-q) \\ \alpha(\{(1,1)\}) &= pq. \end{aligned}$$

This may be compared to

$$\begin{aligned} \pi(\{(0,0)\}) &= (1-p)(1-q) \\ \pi(\{(0,1)\}) &= (1-p)q \\ \pi(\{(1,0)\}) &= p(1-q) \\ \pi(\{(1,1)\}) &= pq. \end{aligned}$$

Exercise 3.6.1 Let A and B be measure spaces with probability measures π_1 and π_2 respectively. If $\alpha : \wp(A \times B) \to [0, 1]$ is a probability measure and

(i) $U \subset A$,

(ii) $V \subset B$,

(iii) $\alpha(U \times V) = \pi(U \times V) = \pi_1(U)\ \pi_2(V)$ where π is the product measure, then is it necessarily true that $\alpha(\{(a, b)\}) = \pi_1(\{a\})\pi_2(\{b\})$ for $a \in U$ and $b \in V$?

3.7 Dice

In this section we consider the category of events based on "rolling a die." As with all real world events, there is a certain ambiguity in the nature of the event itself, but the basic assumption of this event is that a "number of sides" is involved, and this number of sides is the same as the number of possible outcomes modeled by a base set

$$S = \{\omega_1, \omega_2, \ldots, \omega_m\}$$

with $\#S = m$ on which one considers a probability measure $\pi : \wp(S) \to [0, 1]$ with

$$\pi(\{\omega_j\}) = p_j \quad \text{for some } p_1, p_2, \ldots, p_m > 0 \text{ with } \quad \sum_{j-1}^{m} p_j = 1. \tag{3.9}$$

As mentioned above, the numbers $p_1, p_2, \ldots, p_m$ may be imagined to be obtained as certain frequency stabilization limits for the actual occurrence of concrete outcomes, though the actual verification of those limits is impossible. The die (or more properly the probabilistic scheme) is said to be "fair" if $p_1 = p_2 = \cdots = p_m$.

Exercise 3.7.1 How many numbers does it take to determine the base measure (3.9) associated with "rolling a die" with m sides?

A "coin flip" may be considered the "rolling of a die" with two sides, and with a single coin flip (in the consideration of applied probability) one associates the Bernoulli measure. Again from the perspective of applied probability, if one considers multiple coin flips, one possibility is to associate with n coin flips the binomial measure/distribution.

Exercise 3.7.2 Give two other possible derived measures one may consider probabilistically in reference to multiple coin flips (Bernoulli trials) with probability p.

Exercise 3.7.3 Given p with $0 < p < 1$, consider the adolescent measure $\gamma : \wp(\mathbb{N}_0) \to [0, 1]$ by

$$\gamma(\{n\}) = (1 - p)^n p.$$

(a) Show γ is a probability measure.

(b) To what well-know measure is γ related and how?

(c) What is the outcome modeled by $\{n\}$?

There are even more possibilities if one considers dice with three or more sides. Let us say for the sake of argument that a given tetrahedral die with sides identifed with the symbols "one," "two," "three," and "four" displays frequency stabilization corresponding to the measure τ with

$$\begin{aligned}\tau(\{\text{one}\}) &= p_1\\ \tau(\{\text{two}\}) &= p_2\\ \tau(\{\text{three}\}) &= p_3.\end{aligned}$$

Alternatively, we can start with a base set $S = \{1, 2, 3, 4\}$ to model the outcomes and a (renamed) probability measure $\tau : \wp(S) \to [0, 1]$. It will be noted that we have here a (continuous) three parameter family of measures under the necessary condition(s)

$$0 < p_1, p_2, p_3 < \sum_{j=1}^{3} p_j < 1.$$

Drawing inspiration from the construction of the binomial measure, we can also consider the generalization $\tau : \wp(S^n) \to [0, 1]$ with

$$\tau(\{(\omega_1, \omega_2, \ldots, \omega_n)\}) = [1 - (p_1 + p_2 + p_3)]^{\#\{j \,:\, \omega_j = 4\}} \prod_{k=1}^{3} p_k^{\#\{j \,:\, \omega_j = k\}}. \quad (3.10)$$

We can call this the **tetranomial measure**. Naturally, there are counting problems involved in verifying that $\tau = \tau_n = \tau_{n,p_1,p_2,p_3}$ is a probability measure. There are 4^n elements in S^n and the associated singleton set Ω, so nominally the measure of the entire space is the sum of the 4^n values given in (3.10). As with the binomial measure, however, we can group these according to compound outcomes with the same number of each outcome $\{j\}$ for $j = 1, 2, 3, 4$. For example, we can consider the compound outcome that all n rolls result in "one" corresponding to the single element

$$\sum_{j=1}^{n} \mathbf{e}_j \in S^n \subset \mathbb{R}^n.$$

Next might be the compound outcome that $n-1$ rolls result in "one" and one roll results in "two." There are n elements in S^n, namely,

$$2\mathbf{e}_k + \sum_{j\neq k} \mathbf{e}_j, \qquad k = 1, 2, 3, \ldots, n$$

corresponding to this compound outcome. In general, if there are ν_k entries/components ω_j in $(\omega_1, \ldots, \omega_n)$ with $\omega_j = k$, that is $\#\{j \ : \ \omega_j = k\} = \nu_k$ for $k = 1, 2, 3, 4$, then we would like to know (or count) the number of elements in S^n for which this condition holds.

Starting with $k = 1$, there are n components of $(\omega_1, \omega_2, \ldots, \omega_n)$ which might be assigned (or take the value) 1. Order of the assignment does not matter, so this constitutes an initial factor in counting of

$${}_nC_{\nu_1} = \left(\begin{array}{c} n \\ \nu_1 \end{array} \right) = \frac{n!}{(n-\nu_1)!\nu_1!}.$$

Once the components assigned to 1 are assigned, and for each such assignment, we can consider the possible assignment of $n-\nu_2$ remaining components to 2. There are

$${}_{n-\nu_1}C_{\nu_2} = \left(\begin{array}{c} n-\nu_1 \\ \nu_2 \end{array} \right) = \frac{(n-\nu_1)!}{[n-(\nu_1+\nu_2)]!\nu_2!}$$

ways to assign the ν_2 "twos." Continuing until we assign the "threes," for which there are

$${}_{n-(\nu_1+\nu_2)}C_{\nu_3} = \left(\begin{array}{c} n-(\nu_1+\nu_2) \\ \nu_3 \end{array} \right) = \frac{[n-(\nu_1+\nu_2)]!}{[n-(\nu_1+\nu_2+\nu_3)]!\nu_3!}$$

possible assignments, we have

$$\left(\begin{array}{c} n \\ \nu_1 \end{array} \right) \left(\begin{array}{c} n-\nu_1 \\ \nu_2 \end{array} \right) \left(\begin{array}{c} n-(\nu_1+\nu_2) \\ \nu_3 \end{array} \right) \tag{3.11}$$

for the desired total number of elements in

$$\begin{aligned} A &= A(\nu_1, \nu_2, \nu_3, \nu_4) \\ &= \{(\omega_1, \omega_2, \ldots, \omega_n) \in S^n \ : \ \#\{j \ : \ \omega_j = k\} = \nu_k, \ k = 1, 2, 3, 4\}. \end{aligned}$$

Note: There is no need to consider the assignment of the "fours" because after the "threes" are assigned all remaining components must be assigned "four." In particular, we must have the necessary relation

$$\sum_{k=1}^{4} \nu_k = n$$

so that

$$\nu_4 = n - (\nu_1 + \nu_2 + \nu_3)$$

in particular. Simplifying the product in (3.11) gives

$$\frac{n!}{\nu_1!\nu_2!\nu_3!\nu_4!}$$

which is an example of a **multinomial coefficient**. In particular,

$$1 = (p_1 + p_2 + p_3 + p_4)^n = \sum_{\nu} \frac{n!}{\nu_1!\nu_2!\nu_3!\nu_4!}\, p_1^{\nu_1}\, p_2^{\nu_2}\, p_3^{\nu_3}\, p_4^{\nu_4},$$

where the sum is taken over 4-tuples $\nu = (\nu_1, \nu_2, \nu_3, \nu_4) \in \mathbb{N}_0^4$ with $\nu_1 + \nu_2 + \nu_3 + \nu_4 = n$. With this in mind, we can rewrite the tetranomial measure of the entire space to see

$$\begin{aligned}
\tau(S^n) &= \sum_{\omega=(\omega_1,\ldots,\omega_n)\in S^n} \tau(\{\omega\}) \\
&= \sum_{\omega\in S^n} \prod_{k=1}^{4} p_k^{\#\{j\,:\,\omega_j=k\}} \\
&= \sum_{\nu} \sum_{\omega\in A(\nu_1,\nu_2,\nu_3,\nu_4)} \prod_{k=1}^{4} p_k^{\nu_k} \\
&= \sum_{\nu} \left(\frac{n!}{\nu_1!\nu_2!\nu_3!\nu_4!}\right) \prod_{k=1}^{4} p_k^{\nu_k} \\
&= (p_1 + p_2 + p_3 + p_4)^n \\
&= 1.
\end{aligned}$$

We have counted above the number of elements in S^n corresponding to a particular collection A of outcomes $(\omega_1, \omega_2, \ldots, \omega_n)$ with prescribed numbers

ν_1, ν_2, ν_3, and ν_4 of "ones," "twos," "threes," and "fours" respectively (on a tetrahedral die). These sets $A = A(\nu_1, \nu_2, \nu_3, \nu_4)$ partition the measure space S^n, and a different kind of counting question arises if we ask how many of these sets are there? That is to say:

> If we model the outcome of n rolls of a tetrahedral die as a 4-tuple
>
> $$\left(\sum_{j=1}^{\nu_1} \mathbf{e}_j, 2\sum_{j=1}^{\nu_2} \mathbf{e}_j, 3\sum_{j=1}^{\nu_3} \mathbf{e}_j, 4\sum_{j=1}^{\nu_4} \mathbf{e}_j \right) \in \{1\}^{\nu_1} \times \{2\}^{\nu_2} \times \{3\}^{\nu_3} \times \{4\}^{\nu_4} \quad (3.12)$$
>
> where $\nu_1 + \nu_2 + \nu_3 + \nu_4 = n$, then how many distinct outcomes are there?

Whatever the perception of the precision and consensus of mathematics, it is very often important, or at least convenient, to be guided by the perception of and comparison to real world phenomena and objects. The set of all elements indicated in (3.12) is mathematically precise, but the method of counting their number may be aided by the introduction of elements not obvious in this mathematical set itself. In particular, one may note that the number we seek would be the same, at least heuristically, as the number of ways to sort n baseballs (none of which can be distinguished from the other) into four boxes, labeled perhaps 1, 2, 3, and 4, from left to right. It is of course possible to put no baseballs into one box or, in fact, to put all the baseballs into a single box. Letting 0 correspond to (i.e., model) a baseball and 1 model the division between two boxes (of which there are three) a sorting may be compared to a point in the set

$$B = \left\{ q = (q_1, q_2, \ldots, q_{n+3}) \in \{0,1\}^{n+3} \ : \ \#\{j \ : q_j = 1\} = 3 \right\}. \quad (3.13)$$

For $n = 6$ baseballs, the point $(1,0,0,0,1,1,0,0,0)$ corresponds to three baseballs in box 2 and three baseballs in box 4; the point $(0,0,1,0,0,1,0,0,1)$ corresponds to two baseballs each in boxes 1, 2 and 3 (and no baseballs in box 4).

Exercise 3.7.4 (baseballs in boxes and rolls of tetrahedral dice)

(a) Write down carefully the entire set of elements T described in (3.12).

(b) Find a bijection between the set T and the set B given in (3.13).

The cardinality of B may be counted rather easily: We just need to determine the locations of the dividing 1's. This is choosing 3 coordinate locations from among the $n+3$ coordinates:

$$\#B = \left(\begin{array}{c} n+3 \\ 3 \end{array} \right) = \frac{(n+3)!}{n!\ 3!}.$$

As a bit of an aside, it may be noted that we have considered sets above consisting of symbols, for example the set $\{h, t\}$ to model concrete "heads" and "tails" outcomes. The elements of B corresponding to baseballs in boxes have also a nice representation in terms of the symbols "$\bullet$" for a baseball and "|" for a divider. The examples given above with six baseballs in four boxes then become

$$| \bullet \bullet \bullet | | \bullet \bullet \bullet \qquad \text{and} \qquad \bullet \bullet | \bullet \bullet | \bullet \bullet |. \tag{3.14}$$

In principle, these symbols or sets of symbols lack some precision and introduce a (very) small element of ambiguity, and mathematicians have opted for a very limited collection of symbols and/or concepts with which to communicate, specifically, aside from the general category of sets as elements and definition of all mathematical objects in terms of sets, there is only one independent symbolic element, namely ϕ for the empty set. Nevertheless, the use of heuristics through the informal consideration of real world phenomenona and objects (baseballs and boxes) and/or a wide variety of symbols as in (3.14) can be extremely suggestive, is very popular, and as hopefully suggested above can assist greatly in answering certain counting questions.

Exercise 3.7.5 Find a formula for the number of possible outcomes of rolling n dice each of which has m sides if only the recorded (and not the specific dice involved) are counted.

Exercise 3.7.6 Two sided dice (coins) and six sided (hexahedral) dice used in games are perhaps the most common dice. Tetrahedral, octahedral, dodecahedral, and icosahedral dice are used in certain games.

(a) Can you think of a commonly occuring every day event involving a three sided die? How would you construct a physical three sided die?

(b) Most tetrahedral dice used in games have the shape of a regular tetrahedron. If one imagines the following two things:

(i) A non-regular tetrahedral die determines frequency stabilization limits (probabilities) p_1, p_2, p_3, p_4 depending on the geometric shape of the die, and

(ii) These numbers p_1, p_2, p_3, p_4 depend **only** on the shape of the die, and not one other factors like friction, e.g., the nature of the materials of which the die is constructed or the substrate onto which it is "tossed,"

then one can formulate a nice problem in mathematical modeling/applied mathematics/science:

> What is the dependence of the numbers p_1, p_2, p_3, p_4 on the shape of the tetrahedral die?

(c) A first step in consideration of part **(b)** and related questions might be to consider (and construct physical) tetrahedral dice with one side an equilateral triangle and the other three sides congruent isosceles triangles so that geometrically only one parameter, the height, determines the shape up to scaling.

(i) By constructing geometrically congruent dice with different material properties (e.g., with different frictional characteristics) and/or geometrically similar physical dice of the same materials one might explore the validity of assumption **(ii)** in part **(b)** above.

(ii) By constructing dice of the suggested form with the same size equlateral face and differing heights, one might explore the apparent dependence of p_1, p_2, p_3, p_4 on height.[2]

[2]Something like this was done with dice of the form suggested in part **(c)** in the case when the isosceles sides were right triangles. These are called **trirectangular tetrahedral dice**. You can read about what has been done in the paper *Probabilities involving standard trirectangular tetrahedral dice rolls,* Rose-Hullman Undergraduate Mathematics Journal, Vol. 19 Issue 1.

3.8 Cards

3.8.1 Standard deck of cards

Let's model a standard deck of 52 cards using a cross product of rank and suit, so a card of rank $r \in \{1, 2, \ldots, 13\}$ and suit $s \in \{1, 2, 3, 4\}$ corresponds to the ordered pair (r, s). By the multiplication principle, there are $(4)(13) = 52$ such points, which is good in comparison to a deck of cards since we said it had 52 cards in it. For the record: rank 11 is a "jack," rank 12 a "queen," rank 13 a "king," and rank 1 an "ace." Suit 1 is spades, suit 2 is hearts, suit 3 is diamonds, and suit 4 is clubs.

The natural probaility measure on $C = \{(r, s) : r \in \{1, 2, \ldots, 13\}, s \in \{1, 2, 3, 4\}\}$ is determined by

$$\pi(\{(r, s)\}) = \frac{1}{52}.$$

This is of course a uniform measure, so probabilities can almost always be determined by counting. Since the cards have some nuance, and "value" both in rank and suit is often of interest, the counting can get complicated. Also, it's a pretty large measure space with 52 elements in the base set and $2^{52} = 4503599627370496$ measurable sets.

Exercise 3.8.1 Write out the number of elements in $\wp(C)$ in words.

3.8.2 Blackjack

A **blackjack** hand consists of two cards and corresponds to a two element subset of C, or a **doubleton**. The doubletons in $\wp(C)$ can be counted using the combination

$$\left(\begin{array}{c} 52 \\ 2 \end{array} \right) = \frac{52(51)}{2} = 1326$$

because a combination $C(n, m)$ counts the number of ways m distinct (but unordered) selections may be made from among n elements/objects/sets. That is, we calculate 52 taken 2 or "52 choose 2."

Exercise 3.8.2 If you ask someone who plays a lot of blackjack[3] how many (starting) hands are possible, he will tell you 34. Why?

[3]Reportedly at the current time, blackjack is the most played casino card game on earth.

There is another way to calculate the number of hands in a card game. While it is not easier, this way allows us to distinguish among the hands by rank and suit (or at least by suit). Consider first the suits represented in a blackjack hand: There are four possibilities that only one suit is represented (both cards of the same suit). For each suit, there are

$$\left(\begin{array}{c} 13 \\ 2 \end{array} \right) = \frac{13(12)}{2} = 78$$

hands consisting of cards from only that suit. These hands account for $4(78) = 312$ of the 1326 blackjack hands. The other hands consist of two cards having different suits, and there are $C(4,2) = 6$ ways for this to happen. It's easy to list these: (spade and heart) (spade and diamond) (spade and club) (heart and diamond) (heart and club) (diamond and club). For each choice of the two suits, there are $13^2 = 169$ possible hands. This means $6(169) + 312$ should be 1326, and it is.

Exercise 3.8.3 Consider $v : C \to \mathbb{N}$ by

$$v(r,s) = \begin{cases} 11 & \text{if } r = 1 \\ r & \text{if } 1 < r < 10 \\ 10 & \text{if } r \geq 10 \end{cases}$$

and

$$B = \big\{ \{\omega_1, \omega_2\} \in \wp(C) \ : \ \#\{\omega_1, \omega_2\} = 2 \big\}.$$

Calculate and translate the probabilities:

(a) Express

$$\mu(\{\omega = \{\omega_1, \omega_2\} \in B \ : v(\omega_1) + v(\omega_2) = 21\})$$

in words where $\mu : \wp(B) \to [0,1]$ is the uniform probability measure on B: "This is the probability that...," and calculate the value.

(b) Find a set $A \in \wp(B)$ for which $\mu(B)$ is the probability that an initially dealt blackjack hand counts for 20 points or more in value, and calculate the probability.

3.8.3 Poker

For more practice counting we can consider poker hands consisting of five cards. There are

$$\begin{aligned}\left(\begin{array}{c} 52 \\ 5 \end{array} \right) &= \frac{52(51)(50)(49)(48)}{5!} \\ &= 52(51)(10)(49)(2) \\ &= 17(13)(7)^2(5)(3)(2)^4 \\ &= 2\,598\,960\end{aligned}$$

possible poker hands. We can model these hands by

$$H = \{h \in \wp(C) : \#h = 5\}$$

and introduce a second uniform probability measure η on H with $\eta(\{h\}) = 1/2598960$.

Exercise 3.8.4 What are the domain and codomain of the probability measure η?

There are some good counting questions associated with hands of cards and poker hands in particular. These illustrate that it is often easier to count things than it is to count things using precise mathematical rules of counting. For example, a poker hand is considered to be a "one pair" hand if it contains two cards with the same rank and no two other cards of the same rank. Thus, "one pair" hands may be (informally) counted as follows: There are 13 choices for the rank of the pair and $C(4, 2) = 6$ choices for the suits of the two cards in the pair for a total of

$$(13)(6) = 78 \qquad \text{pairs.}$$

To finish the hand, we need three cards each of which has rank distinct from the rank chosen for the pair. There are $C(12, 3) = 12(11)(10)/6 = 220$ choices for the ranks of the three remaining cards and $4^3 = 64$ choices for the suits of the remaining three cards giving a total of

$$220(64) = 14080 \qquad \text{complementary three card "hands"}$$

to go with a pair to make a "one pair" poker hand. This gives a total of

$$78(14080) = 1098240 \qquad \text{one pair poker hands.}$$

With this number we can also calculate what people mean by the probability of getting a "one pair" hand dealt to you in a game of poker. That would be

$$P(\text{two pair}) = \frac{1098240}{2598960} = \frac{352}{833} \doteq 0.422569$$

or about 42.2569%. Remember this number is

$$\eta(A_2) = \sum_{c \in A_2} \mu(\{c\})$$

where $A_2 \subset H \subset \wp(C)$ is the subset of H (the model set for poker hands) corresponding to "two pair" hands and η is the probability measure of Exercise 3.8.4.

Exercise 3.8.5 If one wishes to make no distinction among "two pair" poker hands having the same rank for the pair and the same ranks for the remaining three cards (without reference to suits) how many distinct "two pair" poker hands are there? Counting hands this way is sometimes called counting "distinct hands." For example, a pair of sixes beats a pair of fives, and a pair of sixes (with remaining cards having) jack high beats a pair of sixes having a ten high.

A good counting exercise is to calculate the number of each kind of poker hand and each kind of distinct hand.

As a final comment on cards, we might attempt to model "two pair" hands as a set which is nominally easier to count than H_2. One way to do this is as ordered pairs with the first entry modeling the pair and having the form $(r_*, \{s_{*1}, s_{*2}\})$ and second entry modeling the remaining three cards and having the form $\{c_1, c_2, c_3\}$ where $c_j = (r_j, s_j) \in C$. The pair is then given by $\{(r_*, s_{*1}), (r_*, s_{*2})\} \in \wp(C)$, and the overall set to enumerate is

$$\Big\{\big((r_*, \{s_{*1}, s_{*2}\}), \{c_1, c_2, c_3\}\big) : \\ c_j = (r_j, s_j) \in C,\ j = 1, 2, 3, \text{ and} \\ s_{*1} \neq s_{*2} \text{ and} \\ r_i \notin \{r_*, r_j\},\ i \neq j\Big\}.$$

Writing down a bijection between this set and the usual model A_2 of the "two pair" hands in

$$H = \{h \in \wp(C) : \#h = 5\}$$

is complicated.

Exercise 3.8.6 Write down the set A_2 that directly models "two pair" poker hands in H.

3.8.4 Example 2 Class 3 (Orloff and Booth)

We finish this section with a discussion of Example 2 from the MIT 18.05 class 3 notes of Orloff and Booth. Here is what they have written:

> **Example 2.** Draw two cards from a deck. Define the events: S_1 = 'first card is a spade' and S_2 = 'second card is a spade'. What is the $P(S_2|S_1)$?

Aside from reinterpretation, this example gives us the opportunity to introduce some new kinds of measures.

Orloff and Booth first assert that the desired probability in their problem may be obtained "directly by counting:"

> If the first card is a spade then of the 51 cards remaining 12 are spades.
>
> $$P(S_2|S_1) = \frac{12}{51}.$$

A measure based approach approach to this counting is the following: The first card drawn corresponds to some model outcome $c_* \in C$. The remaining possible outcomes (for a second draw) are naturally modeled by $S = C \backslash \{c_*\}$, and the uniform probability measure π_* on S has

$$\pi_*(\{c\}) = \frac{1}{51}$$

because there are 51 elements in S. Given that $c_* = (r_*, 1)$, i.e., the first card drawn is a spade, there are 12 elements $c = (r, s)$ left in S with $s = 1$. This suggests modeling the outcome "the second card drawn is a spade" by

$$S_* = \{c = (r, 1) \in C : r \neq r_*\} \subset S.$$

In fact,

$$\pi_*(S_*) = \sum_{c \in S_*} \frac{1}{51} = \frac{\#S_*}{51} = \frac{12}{51} = \frac{4}{17} \tag{3.15}$$

is the value of a measure on a set giving the claimed/desired probability.

There are a few unsettling features of this computation. The primary problem is that we haven't really properly modeled any specific outcome of the event in question nor obtained the probability as the value of any specific measure. There are thirteen measures π_* involved and thirteen different sets S_*. It happens that $\pi_*(S_*)$ has the same value in all thirteen cases, so this works to make the computation "feel" right. It would be better, however, if we could achieve the following:

1. Obtain the value of the probability as a specific value of a measure on a measure space appropriate to model the actual possible outcome(s) of the event, and

2. do so in such a way that a/the relation to the original base measure π_0 on C determined by
$$\pi_0(\{c\}) = \frac{1}{52}$$
comes into play.

These objectives are somehow what is essentially behind the effort of Orloff and Booth to "recompute" the value using their "formal definition of conditional probability."

From our point of view, i.e., my point of view, a "conditional probability" is the value of some restriction probability measure. This is a good time to make a comparison between my formula for a restriction probability measure

$$\rho_B(A) = \frac{\pi(A \cap B)}{\pi(B)} \tag{3.16}$$

and the "formal definition of conditional probability" of Orloff and Booth, namely

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{3.17}$$

The two "formulas" look superficially similar. The big difference is that my formula involves two distinct measures which are well-defined **functions** with specified domains, codomains, and values. As the symbol P is used for "probability" by Orloff and Booth, however, essentially nothing is formally correct. The values appearing in (3.17) are not actually the values of a function with a specified domain and range, but the symbol "P" is being used in some vague non-mathematical way to stand in for "whatever is the correct thing."

While it may be granted that it is more "complicated" to properly distinguish the measures involved and avoid the use of the universal "probability" symbol "P," I would suggest that the fundamental level of complication is inherent in the calculation, and it is important to understand precisely what is being computed, or at least can be. I'm not sure if the following development convincingly illustrates this point, but at the very least a fixed model outcome space is introduced and a way to remove some of the ambiguity associated with the thirteen values in (3.15) is provided.

Consider the event described by

> "drawing two cards from a deck of cards[4] and keeping track of the ordered pair of cards drawn."

From this point of view, what are described as "events" by Orloff and Booth are actually (real world) compound outcomes, so we need, first of all, a set to model the outcomes. The elements of such a set would be (or at least might be) ordered pairs with entries from our set C modeling cards. More precisely, we can consider

$$D = \{(c_1, c_2) \in C^2 : c_1 \neq c_2\}.$$

In order to specify the desired specified/compound outcomes, we can use a function $x : D \to \mathbb{R}$ by $x(c_1, c_2) = s_1$ where $c_1 = (r_1, s_1) \in C$ and then

$$S_1 = \{(c_1, c_2) \in D : x(c_1, c_2) = 1\}.$$

Remember 1 corresponds to "spades." Similarly,

$$S_2 = \{(c_1, c_2) \in D : y(c_1, c_2) = 1\}$$

where $y : D \to \mathbb{R}$ by $y(c_1, c_2) = s_2$ and $c_2 = (r_2, s_2) \in C$ as usual.

But what is the "natural" measure on D in this case? Recall there are $52(51)$ ordered pairs (with distinct entries) in D. That is, $\#D = 2652$. Thus, the uniform probability measure π on D has

$$\pi(\{(c_1, c_2)\}) = \frac{1}{52(51)}.$$

The set D is not a product, and π is not a simple product measure, and it may not be immediately obvious how π arises in relation to the original uniform measure π_0 on C. The measure π does arise, however, as a more general product:

[4] specifically from a standard deck of 52 cards

Definition 8 (restriction product measure for two factors) Given a measure space

$$S = \{\omega_1, \omega_2, \ldots, \omega_n\}$$

with a probability measure $\pi_0 : \wp(S) \to [0,1]$, for each element ω_j of S, i.e., for each $j = 1, 2, \ldots, n$ let π_j be the restriction probability measure on $S \backslash \{\omega_j\}$. Recall that this means

$$\pi_j(\{\omega\}) = \rho_{S\backslash\{\omega_j\}}(\{\omega\}) = \frac{\pi_0(\{\omega\}\backslash\{\omega_j\})}{\pi_0(S\backslash\{\omega_j\})}.$$

The measure $\pi : \wp(S^2) \to [0,1]$ satisfying

$$\pi(\{(\omega_i, \omega_j)\}) = \pi_0(\{\omega_i\})\ \pi_i(\{\omega_j\})$$

is a (probability) **restriction product measure** on S^2.

The restriction product measure has some interesting properties. First, π given in Definition 8 is a probability measure. To see this, we need to show

$$\sum_{(a,b)\in S^2} \pi(\{(a,b)\}) = 1.$$

In fact,

$$\begin{aligned}
\sum_{(a,b)\in S^2} \pi(\{(a,b)\}) &= \sum_{i=1}^{n}\sum_{j=1}^{n} \pi_0(\{\omega_i\})\ \pi_i(\{\omega_j\}) \\
&= \sum_{i=1}^{n} \pi_0(\{\omega_i\}) \sum_{j=1}^{n} \pi_i(\{\omega_j\}) \\
&= \sum_{i=1}^{n} \pi_0(\{\omega_i\}) \qquad \text{because } \pi_i \text{ is a probability measure on } S \\
&= 1.
\end{aligned}$$

Even more interesting is the fact that the *regular* restriction of the restriction product measure π to

$$D = \{(a,b) \in S^2 : a \neq b\},$$

i.e., to $\wp(D)$, is the same as the probability restriction of the restriction product measure to D. In particular, π is a probability measure on D. To see this, it is enough to show $\pi(\{(\omega_j, \omega_j)\}) = 0$ for every $\omega_j \in S$. In fact,

$$\pi(\{(\omega_j, \omega_j)\}) = \pi_0(\{\omega_j\})\ \pi_j(\{\omega_j\}),$$

but

$$\pi_j(\{\omega_j\}) = \frac{\pi_0(\{\omega_j\}\backslash\{\omega_j\})}{\pi_0(S\backslash\{\omega_j\})} = \frac{\pi_0(\phi)}{\pi_0(S\backslash\{\omega_j\})} = 0.$$

Returning to our discussion of drawing two cards (and keeping them in order) if we use the uniform probability measure $\pi_0 : \wp(C) \to [0,1]$ associated with a deck of cards, or with drawing a single card from a deck of cards, as the base measure in Definition 8, then the restriction probability measure π we obtain restricted to

$$D = \{(c_1, c_2) \in C^2 : c_1 \neq c_2\}$$

is the natural measure to consider for this problem. What is special here is that the base measure is a uniform probability measure. This is not required by the definition. In this case, the restriction probability measures π_j are not technically uniform probability measures because

$$\pi_j(\{c\}) = \begin{cases} \frac{1/52}{\pi_0(C\backslash\{\omega_j\})} = \frac{1}{51}, & c \neq \omega_j \\ 0, & c = \omega_j, \end{cases}$$

but these measures are "virtually" uniform probability measures and actually uniform probability measures when restricted to $C\backslash\{\omega_j\}$ since

$$\#(C\backslash\{\omega_j\}) = 51$$

where $C = \{\omega_1, \omega_2, \ldots, \omega_{52}\}$. It follows that the restriction product measure π has nonzero values

$$\pi(\{(c_1, c_2)\}) = \frac{1}{52(51)} \qquad \text{for each singleton with } (c_1, c_2) \in D.$$

Recalling also that $\#D = P(52, 2) = 52(51)$ we see π is also the uniform probability measure on D. Thus, the measure π is a natural measure to consider on D simply because it is a uniform probability measure, but we

can see now how it naturally arises as a kind of product measure. More general product measures are considered below.

The desired probability is the value of the measure of the set S_2 with respect to the measure ρ_{S_1} where ρ_{S_1} is the restriction probability measure of the restriction product measure π on D. Note that $\#S_1 = \#S_2 = 13(51)$ and $\#(S_1 \cap S_2) = 13(12)$. The restriction probability measure on S_1 is determined by

$$\rho_{S_1}(A) = \frac{\pi(A \cap S_1)}{\pi(S_1)}. \tag{3.18}$$

By the counting above and the definition of π we find

$$\pi(S_2 \cap S_1) = \frac{\#(S_2 \cap S_1)}{52(51)} = \frac{13(12)}{52(51)} = \frac{1}{17}$$

and

$$\pi(S_1) = \frac{\#S_1}{52(51)} = \frac{13(51)}{52(51)} = \frac{1}{4}. \tag{3.19}$$

Therefore, taking $A = S_1$ in (3.18 gives

$$\rho_{S_1}(S_2) = \frac{1/17}{1/4} = \frac{4}{17}.$$

This is the number Orloff and Booth call $P(S_2|S_1)$.

Exercise 3.8.7 The formula (3.16) looks quite complicated in comparison to the formula (3.17) because it involves two measures ρ_{S_1} and π and, in fact, the measure π itself is quite complicated being given in terms of product measure constructed from 53 or so other measures. Orloff and Booth use (3.17) in the form

$$P(S_2|S_1) = \frac{P(S_2 \cap S_1)}{P(S_1)}.$$

Their justification for (taking) the value $P(S_1) = 1/4$ is as follows:

> "...there are 52 equally likely ways to draw the first card and 13 of them are spades."

What measure space M and what measure μ does this description suggest? For what set $A \subset M$ is $\mu(A) = 1/4$?

Exercise 3.8.8 In order to apply the formula

$$P(S_2|S_1) = \frac{P(S_2 \cap S_1)}{P(s_1)}$$

to find $P(S_2|S_1)$ Orloff and Booth say they need to compute $P(S_2)$. It will be noted that they do not actually need to compute $P(S_2)$ whatever that value might be, but probably they are just giving this computation as a separate interesting and as they say "surprising" example. Their extended justification for (taking) the value $P(S_2) = 1/4$ is more complicated:

> "...the value of the first card certainly affects the probabilities for the second card. However, if we look at *all* possible two card sequences we will see that every card in the deck has equal probability of being the second card.... $P(S_2) = 1/4$."

Considering all two card sequences as suggested by Orloff and Booth corresponds well to our consideration of the set $D = \{(c_1, c_2) \in C^2 : c_1 \neq c_2\}$. What probability measure μ might one have in mind on D to conclude $\mu(S_2) = 1/4$ for $S_2 \subset D$?

I mentioned above that my attempt to justify some additional complication in modeling the outcome of the actual event of drawing two cards in Orloff and Booth's Example 3 above might not be convincing. There is one point at which I will admit the above presentation is particularly "weak" or vulnerable to criticism (at least on the face of it). This is in the computation of $\pi(S_1)$ given in (3.19). I have introduced a more complicated restriction product measure π and seemingly the computation of it's value on the set S_1 corresponding to "the first card is a spade" is unnecessarily complicated. The argument of Orloff and Booth that this number should be simply

$$P(S_1) = \pi_0(A_1) = \frac{1}{4}$$

(see Exercise 3.8.7) where $A_1 = \{c = (r, s) \in C : s = 1\}$ is, at least on the face of it, very "strong." Of course, $P(S_1)$ is nonsense mathematically, but the contention that $\pi_0(A_1)$ is "the probability that the first card is a spade" does make good sense and is quite persuasive (if one believes in applied probability of course). I will make an effort to address this "situation" below in a setting where the discussion is more amenable to illustration. I will even try to prove a theorem relating the derived restriction product measure π to the base measure π_0 so that one can conclude $\pi(S_1) = \pi_0(A_1)$ properly and in general.

3.9 General product measures

Definition 8 may be generalized in a couple different directions. One possibility is to allow the measures π_j associated to each element to be simply any measures on the second factor. Another more complicated direction is to allow the factors to be different and/or to allow more than two factors.

Let's start with the definition of product measures for more than two factors.

Definition 9 (product measure) Given measure spaces S_1, S_2, ..., S_k each with finitely many elements $\#S_j = n_j$ for $j = 1, 2, \ldots, k$ and with measures μ_1, μ_2, ..., μ_k respectively, the **product measure** μ on the Cartesian product

$$\prod_{j=1}^{k} S_j$$

is determined by

$$\mu(\{(\omega_1, \omega_2, \ldots, \omega_k)\}) = \prod_{j=1}^{k} \mu_j(\{\omega_j\}).$$

If each of the measures μ_1, μ_2, ..., μ_k is a probability measure, then the product measure is a probability measure as well.

Notice that the measures in this definition can be any measures, so this allows the option to choose different measures on a single measure space S to obtain a product measure on S^n different from the "standard" product measure with

$$\mu(\{(\omega_1, \omega_2, \ldots, \omega_k)\}) = \prod_{j=1}^{k} \mu_0(\{\omega_j\})$$

where $\mu_0 : \wp(S) \to [0, \infty)$ is one fixed measure on S. This however is not the same as the construction leading to the restriction probability measure where even more measures are needed.

Definition 10 (generalized product measure) Given measure spaces S_1, S_2, ..., S_k each with finitely many elements so that

$$S_i = \{\omega_{i1}, \omega_{i2}, \ldots, \omega_{in_i}\} \qquad \text{for } i = 1, 2, \ldots, k$$

and measures $\mu_{\ell a}$ on S_ℓ, one for each $\ell = 1, 2, \ldots, k$ and each

$$a = (a_1, a_2, \ldots, a_k) \in \prod_{j=1}^{k} S_j,$$

the **generalized product measure** μ on the Cartesian product

$$\prod_{j=1}^{k} S_j$$

is determined by

$$\mu(\{a = (a_1, a_2, \ldots, a_k)\}) = \prod_{\ell=1}^{k} \mu_{\ell a}(\{a_\ell\}). \tag{3.20}$$

It is clear that the values (3.20) determine a measure on $\prod_{j=1}^{k} S_j$ simply because assigning a non-negative real number to each singleton set determines a measure on any set with finitely many elements by the extension formula. It is not so clear that this generalized product measure is a very interesting measure, but the definition does serve to introduce the use of a rather large number of possibly distinct measures in the construction of a product measure. Notice there are more measures involved in this definition than there are elements in the union of all the spaces S_j for $j = 1, 2, \ldots, k$ or even all the elements in the Cartesian product $\prod_{j=1}^{k} S_j$ itself.

Let us specialize the generalized product measure to the case of two factors S_1 and S_2 in order to see how the restriction product measure of Defnition 8 arises as a special case of a generalized product measure.

Definition 11 (generalized product measure for two factors) Given measure spaces S_1, and S_2 each with finitely many elements so that

$$S_1 = \{\omega_{11}, \omega_{12}, \ldots, \omega_{1n_1}\} \qquad \text{and} \qquad S_2 = \{\omega_{21}, \omega_{22}, \ldots, \omega_{2n_2}\}$$

and measures

$$\mu_{1(a_1,a_2)} : \wp(S_1) \to [0, \infty) \qquad \text{for } (a_1, a_2) \in S_1 \times S_2$$

and

$$\mu_{2(a_1,a_2)} : \wp(S_2) \to [0, \infty) \qquad \text{for } (a_1, a_2) \in S_1 \times S_2,$$

the **generalized product measure** μ on the Cartesian product $S_1 \times S_2$ is determined by

$$\mu(\{(a_1, a_2)\}) = \mu_{1(a_1,a_2)}(\{a_1\})\ \mu_{2(a_1,a_2)}(\{a_2\}). \tag{3.21}$$

Notice that in this case there are $2n^2$ measures, while the value of the restriction product measure on S^2

$$\pi(\{(\omega_i, \omega_j)\}) = \pi_0(\{\omega_i\})\ \pi_i(\{\omega_j\})$$

is given in terms of the $n+1$ measures π_0, π_1, π_2, ..., π_n on S. Specifically, if we take $S_1 = S_2$ in the definition of the generalized product measure for two factors and the n^2 measures $\mu_{1(\omega_i,\omega_j)} = \pi_0$ for $i, j = 1, 2, \ldots, n$, then we have the n^2 measures $\mu_{2(\omega_i,\omega_j)}$ to specify. For $j = 1, 2, \ldots, n$ we take

$$\mu_{2(\omega_i,\omega_j)} = \pi_i = \rho_{S\setminus\{\omega_i\}}.$$

This gives the remaining n^2 measures for which the generalized product measure for two (identical) factors yields the restriction product measure. Notice the asymmetric role played by the original probability measure π_0.

It remains to define the restriction product measure for more than two factor spaces.

Definition 12 (probability restriction product measure) Given a measure space S with

$$S = \{\omega_1, \omega_2, \ldots, \omega_n\}$$

and a probability measure μ_0 on S the **restriction product measure** π on the k-fold Cartesian product S^k is determined by

$$\pi(\{(a_1, a_2, \ldots, a_k)\}) = \pi_0(\{a_1\}) \prod_{j=2}^{k} \rho_{S\setminus\cup_{\ell=1}^{j-1}\{a_\ell\}}(\{a_j\}) \tag{3.22}$$

where the restriction measure

$$\rho_{S\setminus\cup_{\ell=1}^{j-1}\{a_\ell\}}$$

is obtained from the base probability measure π_0 for $j = 2, 3, \ldots, k$.

Exercise 3.9.1 Explain how the restriction product measure on S^k is obtained as a generalized product measure on k factors.

Exercise 3.9.2 Let π be a restriction product measure on S^k determined by a base probability measure π_0 as in Definition 12. Show that if

$$(a_1, a_2, \ldots, a_k) \in S^k \backslash \{(\tau_1, \tau_2, \ldots, \tau_k) : \tau_i \neq \tau_j,\ i \neq j\},$$

then $\pi(\{(a_1, a_2, \ldots, a_k)\}) = 0$.

Exercise 3.9.3 If the base measure π_0 on S is the uniform probability measure what is the value of $\pi(\{(a_1, a_2, \ldots, a_k)\})$ given in (3.22)?

3.10 Urns

The problems involving cards tend to be a little more complicated than those involving dice, at least sometimes, because of the large number and diversity of cards in a deck. Problems involving balls in urns can also be quite complicated but for slightly different reasons.

3.10.1 Example 3 Class 3 (Orloff and Booth)

As we ended the section 3.8 with a discussion of Example 2 from class 3 of Orloff and Booth we begin this section with a discussion of Example 3 from the same class 3 notes of Orloff and Booth. Here is the statement:

> **Example 3.** An urn contains five red balls and two green balls. Two balls are drawn one after the other. What is the probability that the second ball is red?

Orloff and Booth use this example to illustrate the use of what they call the **law of total probability**. Phrased in terms of measures, the relation they have in mind becomes

$$\pi(A) = \rho_{A_1}(A)\ \pi(A_1) + \rho_{A_2}(A)\ \pi(A_2)$$

where A_1 and A_2 are disjoint sets that partition a measure space S, that is $A_1 \cup A_2 = S$, and the restriction measures are derived from a base probability measure π.

Exercise 3.10.1 Prove the law of total probability as stated above in terms of measures.

As with our discussion of Orloff and Booth's Example 2, a large part of this problem will be explaining precisely how the modeling (with sets) works. Fortunately, we've already gone over most of the relevant details. In particular, the basic answer to this question is very much like the extraneous calculation of $P(S_2)$ given by Orloff and Booth in their discussion of Example 2 and which is the subject of Exercise 3.8.8 above. Also, the basic concepts in this example are pretty much the same as those in our discussion of Example 2 but there are only 7 elements in our base modeling set instead of 52 corresponding to cards, so this should give a good example to use to illustrate,

and eventually prove, the theorem promised at the end of the discussion of Example 2.

In practice, one may not be able to distinguish which of the five red balls has been (with)drawn when a red ball is drawn out of the urn. In principle, however, each of the five balls is a distinct ball. Let us imagine like cards, say diamonds or hearts, these five red balls are marked with a number or rank, so the concrete outcome of drawing a red ball corresponds to one of the elements 1, 2, 3, 4, or 5 in the set

$$S = \{1, 2, 3, 4, 5, 6, 7\}. \tag{3.23}$$

We could alternatively take elements $(1, 1)$, $(2, 1)$, $(3, 1)$, $(4, 1)$, and $(5, 1)$ representing red balls with the second coordinate corresponding to "red" and then introduce the elements $(1, 2)$ and $(2, 2)$ to correspond to (the drawing of) green balls. Since the resulting set is not a cross product, I've deemed it simpler to just use the set S in (3.23) in conjunction with the function $x : S \to \{1, 2\}$ by $x(j) = 1$ for $j = 1, 2, \ldots, 5$ and $x(6) = x(7) = 2$ to delineate color. On the set S I introduce the uniform probability measure π_0 with $\pi_0(\{j\}) = 1/7$ for each $j \in S$. I can also form the Cartesian product set $S^2 = S \times S$ and the more relevant collection

$$D = \{(\omega_1, \omega_2) \in S^2 : \omega_1 \neq \omega_2\}$$

of distinct ordered pairs to model the concrete outcomes of "drawing two balls without replacement and keeping track of the ordered result."

With this model, Orloff and Booth's "the first ball is red" corresponds to

$$R_1 = \{(\omega_1, \omega_2) \in D : x(\omega_1) \leq 5\}.$$

Similarly, the compound outcome "the first ball is green" is modeled by

$$G_1 = \{(\omega_1, \omega_2) \in D : x(\omega_1) \geq 6\},$$

"the second ball is red" corresponds to

$$R_2 = \{(\omega_1, \omega_2) \in D : x(\omega_2) \leq 5\},$$

and "the second ball is green" to

$$G_2 = \{(\omega_1, \omega_2) \in D : x(\omega_2) \geq 6\}.$$

Ignoring momentarily the suggestion of Orloff and Booth to consider $\pi_0(\{j : j \leq 5\}) = 5/7$ as $P(R_2)$, though that is the correct value. We can observe, however, that

$$\begin{aligned} R_2 = \{&(1,2),(1,3),(1,4),(1,5),\ (2,1),(2,3),\ldots,(5,4)\} \\ \cup\quad &\{(6,1),(6,2),\ldots,(7,5)\} \end{aligned}$$

which is a set with $P(5,2) + 5C(2,1) = 30$ elements. If we introduce the restriction product measure π, or equivalently the uniform measure on D, then

$$\pi(R_2) = \frac{30}{7(6)} = \frac{5}{7}.$$

This is one way to get the informal probability assertion $P(R_2) = 5/7$.

Next, Orloff and Booth wish to recompute this value $\pi(R_2)$ using the restriction measures determined by the disjoint sets R_1 and G_1:

$$\pi(R_2) = \rho_{G_1}(R_2)\ \pi(G_1) + \rho_{R_1}(R_2)\ \pi(R_1).$$

The set G_2 seems to be basically irrelevant. The idea is that, given the first ball is green, the probability that the second ball is red, i.e., the value of the restricted probability measure $\rho_{G_1}(R_2)$, is easy to calculate or obvious: There are five red balls left and 6 balls total. That is,

$$\rho_{G_1}(R_2) = \frac{10/[7(6)]}{12/[7(6)]} = \frac{5}{6}.$$

This is where the "obvious" value involves recourse to the measure(s) π_* on $S\backslash\{j_*\}$ where j_* is 6 or 7 and $\pi_*(\{j\}_) = 1/6$. Similarly,

$$\rho_{R_1}(R_2) = \frac{20/[7(6)]}{30/[7(6)]} = \frac{4}{6} = \frac{2}{3}.$$

It remains to compute $\pi(G_1)$ and $\pi(R_1)$. It may be noted that we have (just) calculated these values:

$$\pi(G_1) = \frac{12}{7(6)} \qquad \text{and} \qquad \pi(R_1) = \frac{30}{7(6)}.$$

Thus as Orloff and Booth conclude:

$$\pi(R_2) = \frac{5}{6} \cdot \frac{2}{7} + \frac{2}{3} \cdot \frac{5}{7} = \frac{5}{7}.$$

I am now going to reconsider each of the calculations above with reference to the illustration of Figure 3.1.

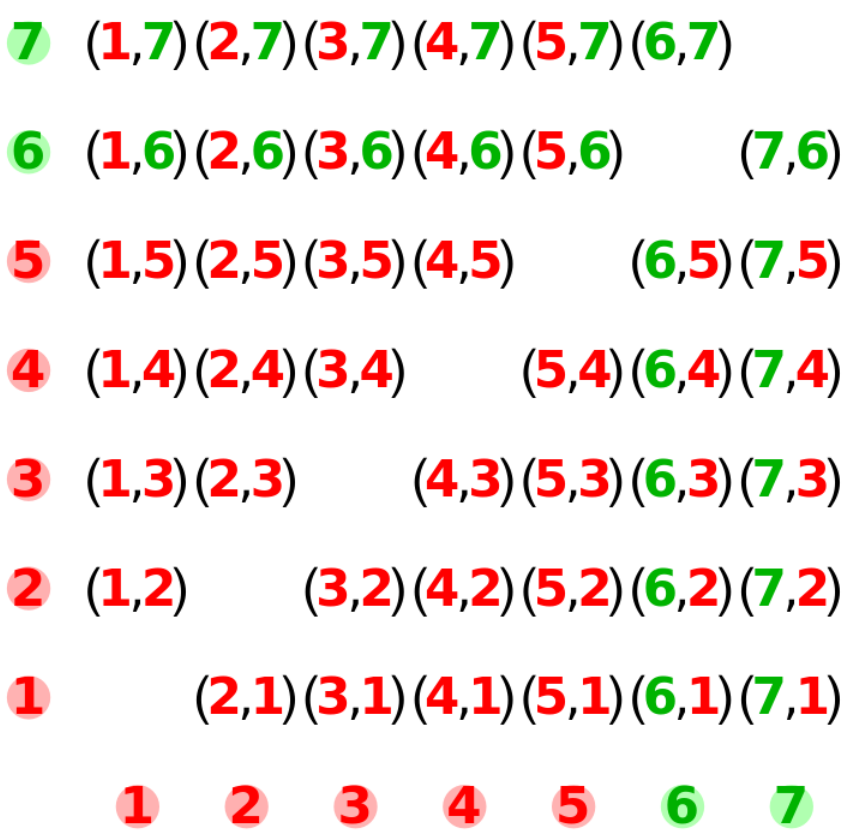

Figure 3.1: A figure indicating the elements in $D \subset S^2 = \{1, 2, \ldots, 7\}^2$ relevant to the restriction product measure and and the base probability space $S = \{1, 2, \ldots, 7\}$ with color coding.

Consider first $\pi(R_1)$. The elements corresponding to

$$R_1 = \{(\omega_1, \omega_2) \in S^2 : \omega_1 \leq 5\}$$

are in the first five columns in Figure 3.1. There are six elements in each of these five columns giving a total of 30 elements. The corresponding singletons each have measure $1/[7(6)]$ giving the total measure $\pi(R_1) = 5/7$. In order to realize this measure as the value of the base measure π_0 on $S = \{1, 2, \ldots, 7\}$ we need to identify an appropriate subset $A_1 \subset S$ and show it has the same measure.

Theorem 9 (basic projection theorem) If the restriction product measure is defined on a product S^2 by a base measure π_0 on $S = \{\omega_1, \omega_2, \ldots, \omega_n\}$, then recall

$$\pi(\{\omega_i, \omega_j\}) = \pi_0(\{\omega_i\})\ \pi_i(\{\omega_j\})$$

where $\pi_i = \rho_{S\setminus\{\omega_i\}}$.

(i) If a set $R_1 \subset S^2$ has the form

$$R_1 = \{(a_1, a_2) \in S^2 : a_1 \in A_1\},$$

for some $A_1 \subset S$, then $\pi(R_1) = \pi_0(A_1)$.

(ii) If π_0 is a uniform probability measure and $R_2 \subset S^2$ has the form

$$R_2 = \{(a_1, a_2) \in S^2 : a_2 \in A_2\},$$

for some $A_2 \subset S$, then $\pi(R_2) = \pi_0(A_2)$.

Proof: In case **(i)** note that for each $a \in S$, there is some unique ω_ℓ for which $a = \omega_\ell$. In particular, the (probability) restriction product measure of the singleton $\{(a, b)\} \subset S^2$ is given by

$$\pi(\{(a, b)\}) = \pi_0(\{a\})\pi_\ell(\{b\}).$$

Formally, there is a function $\ell : S \to \{1, 2, \ldots, n\}$ such that $\omega_{\ell(a)} = a$. Consequently, we can write

$$\begin{aligned}
\pi(R_1) &= \sum_{a_1 \in A_1} \sum_{a_2 \in S} \pi(\{(a_1, a_2)\}) \\
&= \sum_{a_1 \in A_1} \sum_{\omega_2 \in S} \pi_0(\{a_1\})\ \pi_{\ell(a_1)}(\{a_2\}) \\
&= \sum_{a_1 \in A_1} \pi_0(\{a_1\}) \sum_{\omega_2 \in S} \pi_{\ell(a_1)}(\{a_2\}) \\
&= \sum_{a_1 \in A_1} \pi_0(\{a_1\}) \qquad \text{because } \pi_{\ell(a_1)} \text{ is a probability measure} \\
&= \pi_0(A_1).
\end{aligned}$$

This establishes assertion **(i)** in the theorem.

Assertion **(ii)** is a little more tricky.

$$\begin{aligned}\pi(R_2) &= \sum_{a_2\in A_2}\sum_{a_1\in S}\pi(\{(a_1,a_2)\}) \\ &= \sum_{a_2\in A_2}\sum_{a_1\in S}\pi_0(\{a_1\})\ \pi_{\ell(a_1)}(\{a_2\}).\end{aligned} \tag{3.24}$$

On the face of it, if ℓ and m are different integers, there is no real reason to believe $\pi_\ell(A_2)$ and $\pi_m(A_2)$ are the same numbers. We recall, however, that in the case of a restriction product measure under consideration we have

$$\pi_{\ell(a_1)} = \rho_{S\backslash\{\omega_{\ell(a_1)}\}} = \rho_{S\backslash\{a_1\}}.$$

Furthermore, by the definition of probability restriction measure, we have

$$\rho_{S\backslash\{a_1\}}(\{a_2\}) = \frac{\pi_0(\{a_2\}\backslash\{a_1\})}{\pi_0(S\backslash\{a_1\})}$$

Nothing so far promises to mitigate the dependence on the element a_1 in the second factors of (3.24). For a uniform base measure π_0 however we can proceed one step further to find

$$\pi_{\ell(a_1)}(\{a_2\}) = \begin{cases} 0, & a_2 = a_1 \\ n\pi_0(\{a_2\})/(n-1), & a_2 \neq a_1, \end{cases}$$

since

$$\pi_0(S\backslash\{a_1\}) = \frac{\#(S\backslash\{a_1\})}{\#S} = \frac{n-1}{n}.$$

Furthermore, $\pi_0(\{a_2\}) = 1/n$, so the nonzero values of $\pi_{\ell(a_1)}(\{a_2\})$ are given

simply by $1/(n-1)$. Continuing from (3.24) then

$$\begin{aligned}
\pi(R_2) &= \sum_{a_2 \in A_2} \sum_{a_1 \in S \setminus \{a_2\}} \pi_0(\{a_1\}) \, \frac{1}{n-1} \\
&= \frac{1}{n-1} \sum_{a_2 \in A_2} \sum_{a_1 \in S \setminus \{a_2\}} \pi_0(\{a_1\}) \\
&= \frac{1}{n-1} \sum_{a_2 \in A_2} \sum_{a_1 \in S \setminus \{a_2\}} \frac{1}{n} \\
&= \frac{1}{n-1} \sum_{a_2 \in A_2} \frac{n-1}{n} \\
&= \sum_{a_2 \in A_2} \frac{1}{n} \\
&= \sum_{a_2 \in A_2} \pi_0(\{a_2\}) \\
&= \pi_0(A_2). \qquad \square
\end{aligned}$$

There are weaker conditions under which the basic conclusions of Theorem 9 hold. We will try to consider those later.

This theorem as it stands, however, justifies the simple calculation

$$\pi(R_1) = \pi_0(A_1) = \frac{5}{7}$$

as the measure of the set modeling the outcome "the first ball is red." It also justifies

$$\pi(R_2) = \pi_0(A_1) = \frac{5}{7}$$

since

$$\begin{aligned}
R_2 &= \{(\omega_1, \omega_2) \in \{1, 2, \ldots, 7\}^2 : \omega_2 \leq 5\} \\
&= \big\{(\omega_1, \omega_2) \in \{1, 2, \ldots, 7\}^2 : \omega_2 \in A_1 = \{1, 2, 3, 4, 5\}\big\}.
\end{aligned}$$

3.10.2 Example 4 Class 3 (Orloff and Booth)

I have worked out a mathematical solution for this example, and it is somewhat interesting. I will try to come back and type up that solution and also a generalized version of the projection theorem (Theorem 9 which goes along with it. In the mean time, you are welcome to undertake the modeling of the outcomes (a bit more complicated than what we have done above) and determination of the probability (as usual a bit more difficult than the proper mathematical modeling) for this problem:

> An urn contains 5 red balls and 2 green balls. A ball is drawn. If it is green a red ball is added to the urn, and if it is red a green ball is added to the urn. (The original ball is not returned to the urn.) Then a second ball is drawn. What is the probability the second ball is red?

3.10.3 Other Urn Problems

Let us warm up with a relatively simple problem: Two balls are withdrawn from an urn containing two red balls, two green balls, and two blue balls. The first ball is not returned to the urn before the second ball is withdrawn. This is called (with)drawing **without replacement**. What is the probability that the second ball withdrawn is red?

One way to find the answer to this question is to imagine the balls labeled with a number in addition to a color so we can keep track of exactly what is happening (or more properly what might happen). Thus we can model the balls in the urn by the set of symbols $S = \{R1, R2, G1, G2, B1, B2\}$. There are six possibilities for the outcome of the first draw and five possible outcomes for the second draw. Thus, modeling the compound outcome as an ordered pair in S^2, there are $5(6) = P(6, 2) = 30$ possible outcomes. In order to have a red ball second, we might have a red ball first, and there are two ways this can happen: $(R1, R2)$ and $(R2, R1)$. This accounts for $P(2, 2) = 2$ possibilities. We could also have a non-red ball first and a red ball second. There are then 4 choices for the first ball and 2 choices for the second (red) ball accounting for 8 possibilities. Together, there are 10 ways to have a red ball second giving

$$P(\text{second ball red}) = \frac{10}{30} = \frac{1}{3}. \tag{3.25}$$

Notice that while we have said it is possible to interpret a "probability function" as a proper function, this is not quite what is happening in (3.25). The argument, for example, is not a proper mathematical element of a set; "second ball red" is probably best described as a real world outcome. Nevertheless, this kind of informal usage of a "probability function" is very common.

Exercise 3.10.2 Introduce an appropriate base set S_2 modeling the possible compound outcomes of withdrawing two balls (in order without replacement) from an urn with two red balls, two green balls, and two blue balls. Find the set in $\wp(S_2)$ modeling the outcome "second ball red."

It may be noticed that no measure (was harmed or even) mentioned in the calculation leading to (3.25). We have claimed, on the other hand, that the values of (all) probabilities in the practice of applied probability are obtained as the values of measures. The reason attention to measures was easily avoided is largely becuase the measueres in question were uniform measures. This circumstance essentially reduces the calculation to counting.

Exercise 3.10.3 Prove that given a measure space with infinitely many points, there is no uniform probability measure for which the measure of every singleton set is positive.

Here is a way to see the measure which serves as the mathematical foundation of (3.25):

We start with the uniform measure on $S = \{1, 2, 3, 4, 5, 6\}$ and introduce the color function $x : S \to \{1, 2, 3\}$ with

$$x(j) = \begin{cases} 1, & j = 1, 2 \\ 2, & j = 3, 4 \\ 3, & j = 5, 6. \end{cases}$$

Here S models the outcome of withdrawing a single ball from among the six balls in the urn, and the elements in the codomain of x model the three colors with 1 corresponding to "red," 2 corresponding to "green," and 3 corresponding to "blue."

The set $S^2 \backslash \Delta$ where

$$\Delta = \{(a, a) : a \in S\}$$

is the diagonal of S^2 may be used to model the possible compound outcomes of withdrawing two balls in order. Notice there are 36 elements in S^2 and 6

elements in Δ so that $\#(S^2 \backslash \Delta) = 30$. Alternatively,

$$\#(S^2 \backslash \Delta) = \#D = P(6,2) = 30$$

where D is the subset of ordered pairs in S^2 having distinct entries as discussed in section 3.1 above.

The uniform probability measure π on $S^2 \backslash \Delta$ has singleton value 1/30. The set corresponding to "second ball red" is

$$\begin{aligned} A &= \{(a,b) \in S^2 : x(b) = 1\} \\ &= \{(2,1),(3,1),(4,1),(5,1),(6,1),(1,2),(3,2),(4,2),(5,2),(6,2)\}. \end{aligned}$$

Since $\#A = 10$ we see

$$P(\text{second ball red}) = \pi(A) = \frac{1}{3}.$$

Exercise 3.10.4 Two problems similar to the problem considered above and in Exercise 3.10.2 in particular are considered in the Class 3 notes of Orloff and Booth. See Orloff and Booth Example 2 and Example 3. These problems may be worked in a variety of ways. Review the discussion of Orloff and Booth and work the problem above involving two red balls, two green balls, and two blue balls in an urn in the ways suggested in that discussion.

It is clear that Orloff and Booth would argue that the answer can be obtained immediately simply because 2 out of 6 or 1/3 of the balls are red. Is this reasoning justified?

Exercise 3.10.5 In the solution above of the problem considered in Exercise 3.10.2 the uniform measure on the space $S^2 \backslash \Delta$ modeling the outcome of two draws without replacement was used to justify the probability measure value 1/30 on singletons. Can you define measure spaces and a construction so that the value 1/30 arises as a product

$$\frac{1}{30} = \frac{1}{6}\,\frac{1}{5}$$

where the factor 1/6 is the singleton value of the uniform measure on $S = \{1,2,3,4,5,6\}$?

Here is a more complicated counting problem: If an urn contains 5 red balls, 5 green balls, and 5 blue balls, what is the probability that drawing

three balls from the urn without replacement results in three balls of different colors (and representing all colors, red, green, and blue, contained in the urn)?

This question would work equally well with the previous urn and its six balls. The big differences here are that the outcome is unordered and that the question inherently involves all three draws together.

3.11 Midword: Philosophy

It is clear that some philosophical issues could be addressed, and this seems like a somewhat reasonable place to do that. One student Jeremy Mahoney mentioned that he feels it is necessary to discuss philosophical issues in relation to the subject(s). I am rather sympathetic toward this point of view. On the other hand, I would assert that the suggested "neceessity" is not strictly speaking correct. To wit, if one picks up almost any elementary text on probability and statistics (or in fact almost any text of any sort on probability and statistics) one finds only the most superficial and sophomoric discussion of philosophy, if one finds anything addressing philosophical questions at all. I mentioned in response some personal inclination to avoid philosophical questions connected to probability and statistics because to do so would likely be dangerous for me. Perhaps I will explain some aspects why that might be the case later. Perhaps it is the case that this fear is shared, and more studiously acted upon by writers of textbooks and (other) instructors of courses, but I suspect the real reason is something altogether entirely different.

On the third hand, we've perhaps already touched on some philosophical issues when I asked you, actually, what you imagine the subjects of probability and statistics to be about. Also, I've asked the fundamentally philosophical question:

> When a single coin is flipped, does the information that "the probability the coin comes up heads is 1/2" (or some other number) give you any useful information about the outcome of that event?

This question may be modified and/or rephrased in various ways. One might ask: Can you derive from the information about the probability if the outcome is heads? Another related question is

> When a single coin is flipped, has anything happened which may be classified as "random" or having happened "by chance?"

Several of you mentioned the words "random" and "chance" in your answers concerning the subject(s) of probability and statistics, and indeed most if not all textbooks on the subject(s) feature these words prominently. None of them give much of an explanation of what they might mean, and you may have figured out by now that there is not a mathematical explanation for what they might mean. The subject of measure theory, and in particular, the study of probability measures has nothing to do with "randomness" or "chance."

It is safer for me to avoid giving answers to philosophical questions, but hopefully it won't be too offensive to ask a few (and point out some other curious observations). In most texts on probability and statistics something called a "random variable" is introduced. Honest authors sometimes also mention that a "random variable" is neither random nor a variable. why do you think this kind of language is used?

Speaking of calling things what they are not, I mentioned the connection, at least etymologically, between the subject of statistics and the state.[5] In specific instances a state is often referred to as a "union." In attempting to determine a precise definition of the word "state," I might suggest you think rather in the direction of a "division." You might also ask yourself why the euphemism of "union" is so common, and what is a "euphemism?"

3.11.1 Excerpt from a probability and statistics text

Sheldon Ross has a section called *A brief history of statistics* which might be about as close as any text I have seen comes to addressing philosophical issues. I'm going to reproduce some excerpts from that section here with some commentary.

> A systematic collection of data on the population and the economy was begun in the Italian city-states of Venice and Florence during the Renaissance. The term statistics, derived from the word state, was used to refer to a collection of facts of interest to the state.

Immediately, having a definiiton for the term "state" would be useful. Note carefully the phrase "of interest to the state." What does "of interest"

[5] See for example *Introduction to Probability and Statistics for Engineers and Scientists* by Sheldon Ross.

mean here? It seems to me there are (at least) two distinct possibilities, and maybe both of them are intended. One is "to be interesting," "to be the object of curiousity." It is probably unlikely that you identified the "state" as some particular individual human being, but it doesn't seem too far fetched to think that only a particular individual human beging can properly find something "to be interesting" or "to be an object of curiosity." Thus, if this is the meaning of "of interest," then we are immediately faced with, at the very least, some kind of abstraction to think that something could be "of interest to the state." When I refer to an "abstraction" here, I mean the attempt to express something differently from what it really is or, more precisely in this instance, to attribute something to an entity to which it cannot directly be attributed. You may note that the idea of an "abstraction" is closely related to the idea of a "euphemism" or perhaps more broadly to the practice of deception.

There is another possibility for the meaning of "of interest." This might mean "for the benefit of." This may fit better with your definition of "state." At this point, I would like to suggest something which might not be completely obvious. If the meaning of "of interest" is taken to mean "for the benefit of," then there is an underlying assumption that the "state" is something desirable. And this goes back to particular individual human beings. It seems to me at least that the assertion "the state is desirable or inevitable" can only be attributed to a particular individual human being. More to the point, if a particular individual human being perceives the state as not inevitable and/or not desirable, then such a person might very well question a subject having its origin in that which is "for the benefit of" the state. In fact, such a person might properly interpret the phrase "for the benefit of the state" to be a deceptive way to say "for the benefit of some individuals other than you."

Continuing from the first paragraph in the "brief history" of Ross:

> The idea of collecting data spread from Italy to the other countries of Western Europe. Indeed by the first half of the 16th century it was common for European governments...

I almost got to the second sentence (fourth sentence of the paragraph). Ross has used the word "government." This is a word that is in common use, but it is a word to the use of which I strongly object. I have thought quite carefully about this word and its use. My (startling) conclusion, after long and careful consideration, is that if someone uses this word that person almost certainly

has no idea what he is talking about. The most benevolent interpretation I can give is that this person is mindlessly parroting harmful propaganda. If I am to extract anything usefull at all from what Ross is trying to say here, some modification must be made. I will start by reproducing and focusing on the entire sentence:

> . . . by the first half of the 16th century it was common for European to require parishes to register births, marriages, and deaths.

Clearly the nonsense continues.[6] A **parish** is a kind of geographical political designation like a county (or perhaps a city). I could go into the sordid definition of the word "political," but I will leave that aside for the moment. A fundamentally geographically designated area cannot (as far as I know) register or record anything. Registering or recording something is an activity undertaken sometimes by certain human individuals. It seems Ross would like us to know some individuals in the sixteenth century were "required" to register births, marriages, and deaths. Perhaps it is a natural assumption that such a person, who we might call a "registrar," experienced this "requirement" (and acted upon it) as the result of the activity of some other individual human or individual humans. An alternative might be that some kind of "requirement" arose as the result of some naturally occurring or environmental events. For example, one might say certain North American or North Asian nomadic tribes were "required" to range further to the south in order to survive during an ice age. I will assume, however, that Ross is pointing to the involvment of some other humans. Such a human is, of course, not "government." But as the person who registers the births, marriages, and deaths in some human population might naturally be referred to by a name like a "registrar," there is perhaps a natural name for someone who "requires" another human to record such data. I suggest the name "hu-

[6] Being acutely aware that some significant effort at deception is on offer here and that words like "indeed" are added in such contexts to stifle critical evaluation—as if to say: "The lies I am telling you here are so heavily reliant on your previous indoctrination and the assumption that you have been thoroughly propagandized that you should assume you understand what I'm saying perefectly." The use of the word "indeed" here is an indication of a kind of "mass formation" in the psychological sense, as one guard in Auschwitz might say to another "Indeed, the Final Solution makes perfect sense," or like one Iranian might say to another in the late 1970s "Indeed the face of the Ayatollah is visible on the surface of the moon." Thus, I have deleted the word "indeed."

man livestock manager." If the terminology I suggest is reasonable, then the description of Ross might be reworded thus:

> The idea of collecting data spread from Italy to the other countries of Western Europe. Indeed by the first half of the 16th century it was common for human livestock managers all over Europe to require their registrars (local farm hands?) to register the births, marriages, and deaths (of the local human livestock).

I have had just about as much of Ross' "brief history" as I care to try to parse in a sensible manner if such is possible. I will attempt one more sentence to finish the paragraph. I may return to the arduous task of trying to understand this "brief history" later, and you may also wish to undertake the translation and explanation of what Ross is trying to say here (if you find it interesting).

> Because of poor public health conditions this last statistic was of particular interest.

The notion of "public health" and more broadly the word "public" are things I have also considered. My basic conclusion is that these words fall into the same objectionable category as "government." These terms cannot be used without contradiction or deception. They should not be used by an honest person who has thought about what they mean. The word "public" in particular makes sense only in a context in which some humans are owned and managed by others. Like the "record keeping of a parish," the ownership involved is often euphemistically transferred to some entity quite incapable of exercising ownership. For example, there might be an attempted euphemistic transfer of ownership to some designated geographic location (the American "public") or to some abstract entity (society's "public" spaces). The abstract approach (i.e., lie) was well-established among the classical Greeks and persisted among the Romans: That which is "public" belongs to the city, or state, or society, or nation, or empire. But the underlying context is very much the same as that considered by Ross: The "farmers" or human livestock managers are the ones quite aware of their exercise of ownership over (when referring to human herds) "the public" and when referring to places in which the human herds may more or less mingle without certain specific restraints "public places" or "public spaces" as opposed to the spaces and places specifically designated by the human livestock managers for the

more or less exclusive use by certain members of the human herd and/or the human livestock management team.

So then, stripping away all the euphemistic language, perhaps it becomes clear what Ross is trying to get across in this last sentence:

> Human livestock managers tend to show a particular interest when there is a high mortality rate among the human livestock, in other words when it seems like the human herd might be dying off.

Nevertheless, there is more to note about the deceptive usage of the term "public health" in particular. Of course, "health" is a perfectly good word. It is only corrupted by the use of the objectionable adjective "public." Health, however, it may be noted is something that only makes sense for individuals, except by deceptive abstraction. For a human livestock manager "public health" indicates nothing about the actual health of any particular individual. To an individual, his own personal health may be of interest, and the health of one individual may be of interest to another. But for the human livestock manager, as for many managers of many kinds of livestock, it is only what is called abstractly the "health of the herd or flock" that is of interest. In particular, it is the total output at harvest, i.e., what the herd can produce, that is of primary intersest. If a management decision can be made resulting in the diminishing of the health of one individual, or resulting in the perishing of that individual, this is very often acceptable to livestock managers both of humans and of other livestock.

Setting aside the morally objectionable context of human livestock management, the question of whether or not a farmer considers the individual health of his, for example, chickens or only the overall output of the flock, or something in between is a matter of management or farming style. It is rare that a chicken farmer will elevate the health of a single individual chicken above overall production, though this can be a livestock management ideal. It is especially the case with respect to disease, and especially disease putting the entire herd or flock at risk, that a farmer will still elevate the health of a single individual (where the concept of health actually makes sense) above the **risk** to the health of the other individuals in the herd or flock. The following question might, or might not, be worthy of consideration:

> In the context of livestock management of animals by humans, might it be worthwhile to consider the modification of livestock management style based on statistical techniques?

Even here, the evaluation of what is "worthwhile" and the overall objective and values of the individual farmer deserve careful consideration. What Ross has made clear so far, I think, is that this is very much not the context in which the development of the use of statistics arose.

3.11.2 A side note on philosophy

Here I will step aside from the questions and offer briefly my perception of the meaning of philosophy itself. In very broad strokes philosophy from Socrates to Nietzsche was the attempt to isolate that which all humans **must** embrace. The efforts varied from wording things in clever ways to spending lots of time giving things names and categorizing observations. There was often an emphasis on the inclusion of the enslavement of some humans by others as something that all humans **must** embrace. These efforts were broadly placed under the heading of "beauty and truth" where "beauty" roughly speaking corresponds to "that which is good" and "truth" corresponds to that which "must be" (good or evil).

Of particular interest for me is the effort of Immanuel Kant who, after seeing some general weakness in the contention that some humans must be enslaved by others, looked for at least some specific thing all humans **must** embrace, or that is to say, be enslaved by. Kant's summary attempt, the "categorical imperative," has been embraced by many (pro enslavement individuals), but my overall evaluation is that Kant was unable to give any justification for anything whatsoever that all humans **must** embrace. In more poetic language: Kant couldn't.

So then philosophy has more or less stagnated with current philosophers in one way or another circling back to the more optimistic approach, and more naive approach, of primitive philosophers like Plato and Aristotle. This is the case, it seems to me, because the objective was poorly chosen. I suggest then a different definition of philosophy itself. Embracing the position (apparently not an inevitable one) that the identification of something that all humans **must** embrace is itself embracing a kind of enslavement and taking the very novel position that enslavement itself is undesirable, I suggest philosophy can reemerge as a **conversation based on common perceptions**. No one is required to participate. Nothing must be embraced by anyone in particular. But for those with common perception, something may be possible. With this in mind, I state here some of my perceptions (organized mostly as axioms and conjectures) for those who might wish to pursue philosophy (the conversation

of philosophy) with me. In further contrast to some others, rather than phrase my perceptions in convincing and clever ways, I will attempt to phrase them in the simplest and, for the most part, most unlikely of forms.

3.11.3 Axiom of perception and reality

There is that which an individual human perceives and, in contrast, there is something fundamentally outside and separate from human perception called **reality**. Each individual human uses language, or something like language to express his perception of reality both inwardly to himself and outwardly to another individual human. This does not rule out perception of reality separate from language, but isolates the expression of perception in language as of particular interest in the conversation of philosophy.

The most important point is the separation: Individual human perception of reality is imperfect and even perhaps viewed as tenuous at best. Perception is not reality. It is to be expected that an expression of the perception of reality in language is much more nonsense than anything else.

3.11.4 Axiom of good and evil

There is that which a person perceives to be "good" or desirable and that which a person perceives to be "evil" or undesirable. Determining common descriptions and perceptions of that which is good and evil is a primary activity of the conversation of philosophy. Determination of disparity in these categories is tantamount to the end of the discussion.

Another primary activity of philosophy is the discussion and determination of definitions.

3.11.5 Definition of freedom

Freedom is a state of existence in which no other human opposes that which the free human desires. That is to say, a human being is free if there does not exist something he desires to do (or think for example) which is opposed by another human. Alternatively, a human being is free if there does not exist another human being who opposes that which he desires.

3.11.6 Axiom/conjecture of freedom

While freedom, like the conversation of philosophy itself, is exceedingly difficult and rare, it is neither impossible nor undesirable. Freedom is good.

3.12 The big questions

Some of you have asked me questions, and some of those are good questions. The "big questions" list below contains questions I asked myself when I started preparing to teach this course, and I am delighted that some of you are starting to ask them. I think they are natural questions for someone taking a course in probability and statistics. I have not answered any of these questions so far, at least explicitly, though the answers of some of them are above expressed in different terminology. When considering these questions, it may be helpful for you to recall something I wrote above in my introduction to the course. I will repeat it here:

> My advice is that if you have no idea why you want to learn something about probability and statistics and view this semester as something other than an opportunity to have time set aside to do that thinking and learning, then you should drop the course, sign up for a different course, or simply do something different.

In view of the fact that some of you are starting to ask these questions (which makes me very happy) this seems like a good time to phrase this sentiment in a more positive way:

> My advice is that you view this semester as an opportunity to set time aside to think about the things **you** want to understand in regard to the subject of probability and statistics and try to learn those things.

I can probably serve as something of a guide, and I might be especially useful as someone to discuss your ideas with you. Above all, I suggest that you are your best guide. Certainly I have thrown some questions at you which are a bit nonstandard and which you might not have anticipated:

- How can I read, write, and speak the language of sets/mathematics?
- What is a function?
- What is a measure?
- What is a probability?
- What is a statistic?
- Is anything random?
- Does "chance" exist?

This may be considered a list of "little questions." They are questions designed to help you learn what **you want to learn** or what I imagine you might want to learn. But I might be very wrong about what you want to learn. If you don't find my qeustions helpful for learning what you want to learn, then they are neither helpful nor important. Don't worry about them.

In contrast if the list below contains questions you have asked, then these are the "big questions," and you should try to answer them.[7]

1. What is a random variable?
2. What is a p-value?
3. What is dependence/independence?

Of course I will try to help you any way I can. That is indeed what I am trying to do, though I may not be very good at it.

[7] And "yes," if you come to what you think is a comprehensive understanding concerning the answer of one of these questions, this is precisely the kind of topic you should choose for a presentation to the class. The possibility that you cannot learn anything that you want to know about probability and statistics (by yourself, independent of me or Georiga Tech or grades or anything else) during an entire four month period (a semester) is just too ghastly to even consider. Of course, if that is the case, and you must confront such a ghastly reality, then there are probably much more important questions to ask yourself rather than questions about probability and statistics, like "How on earth did I end up in such a helpless state?"

Chapter 4

Lecture 4: Uncountable Measure Spaces

The main objective of this lecture is to introduce probability measures on intervals in the real line. When these measures are introduced, we will want to try to check that all the properties of baby measures and adolescent measures with which we are familiar carry over to these new measures. Most constructions[1] will carry over to these new measures, but others like PMF and the singleton extension formula will not. In addition, we will introduce some new constructions, most notably the **variance**, associated with measures and we will want to go back and see how these apply to baby and adolescent measures.

When it comes to measures on intervals in $\mathbb{R}$, there are not many (actually not really any) sets analogous to sets of symbols like $\{h, t\}$ or $\{\omega_1, \ldots, \omega_n\}$ which are not themsleves sets of numbers. For this reason, we can (for the most part, at least initially) dispense with sets other than sets of numbers. Thus, when we want to measure outcomes with an interval in the real line, the outcomes we are measuring are typically already numbers in the real line.

For example, we might model the time in minutes a particular student arrives in his 3:30 class before the scheduled time. In principle, this could be any real number of minutes, and we can take negative numbers to mean that the student is that number of minutes late. In this way, we might want to consider the real line as a measure space. In practice, the time can

[1] For example, monotonicity, restriction, integration, CMF, and measures induced by real valued functions.

only be measured with a certain accuracy, so maybe there are only finitely many possible times before and after the class time that need be considered, but still time can be measured with quite good accuracy, so there can be many possible values, even if there are only finitely many. Furthermore, as you might have guessed from our discussion above of measure spaces with finitely many elements, the analysis of such measure spaces becomes more and more complicated when there are more and more elements. Thus, sometimes, even if there are only finitely many real world outcomes under consideration (which in practice is always the case for one reason or another) it is often convenient to do the modeling of the outcomes using an interval.

Perhaps the key complication arising when a measure space is an interval of real numbers like $[-20, 20]$ or $[0, 1]$ or $(0, 1)$ or $\mathbb{R}$ is that the measure of a singleton should almost always be considered to be zero:

$$\mu(\{\xi\}) = 0 \qquad \text{or for a probability measure} \qquad \pi(\{\xi\}) = 0.$$

This means we definitely cannot have some kind of extension formula where we simply add up the values of the measure on singletons. Conversely, if we imagined singletons to have positive measure in a measure space in which there are uncountably many singletons, like an interval in the real line, then it turns out the measure of any reasonably large sets is always infinity, which of course is no good. For example, if every time between -1 minute (a minute late) and 1 minute (a minute early) in our example above has corresponding singleton $\{t\}$ with positive measure, then additivity would certainly give the measure (or the probability if you like) of the set corresponding to times between $t = -1$ and $t = 1$ would be infinite, which is not what we want.

First note that length measure on the real line has this same "problem." Each singleton $\{\xi\} \subset \mathbb{R}$ is itself a closed interval

$$\{\xi\} = [\xi, \xi] = \{t \in \mathbb{R} : \xi \leq t \leq \xi\}$$

and the measure of this interval is zero. Furthermore, we really can't get the length of larger intervals by adding up the lengths of the singletons. Fortunately, we have a formula for the length of an interval I with endpoints $a < b$, namely $\mu(I) = b - a$. Of course, there are many more complicated subsets of $\mathbb{R}$ one would like to measure.

Historically, this question about how to obtain a measure on $\mathbb{R}$ for which the measure of an interval is its length was a very difficult one and was the subject of intense investigation. One perhaps very surprising result of this investigation was the following:

One cannot expect to measure all the sets in $\wp(\mathbb{R})$.

Of course, we have discussed measures that do measure every subset of $\mathbb{R}$, but these were only (generalized) baby measures and (generalized) adolescent measures. They are very different from length measure and very different from the measures of interest as probability measures we want to talk about here on intervals of $\mathbb{R}$. Another result of the same investigations leading to the possibly disturbing conclusion above led to the construction of a specific measure, now called **Lebesgue measure**, for which the measure of an interval is its length. This measure is also foundational for all probability measures on intervals in the real line, so the general definitions and properties we now present are probably worth knowing.

The first key discovery was the identification of **closure under countable unions** and **countable additivity** as the correct properties[2] to define a measure and a **σ-algebra** the proper domain of a measure.

Definition 13 (σ-algebra) A collection $\mathfrak{M}$ of subsets of a given set S is a **σ-algebra** if the following properties hold:

(0) $\phi \in \mathfrak{M}$.

(i) If $A \in \mathfrak{M}$, then
$$A^c = \{x \in S : x \notin A\} \in \mathfrak{M}.$$

(ii) If $A_1, A_2, \ldots$ is any sequence of subsets in $\mathfrak{M}$, then the countable union
$$\bigcup_{j=1}^{\infty} A_j \in \mathfrak{M}.$$

Property **(i)** is called closure under complements and property **(ii)** is called closure under countable unions.

[2]You may wish to review the preliminary discussion of σ-algebras in Lecture "-1." Also, it may not be strictly speaking correct to say these conditions are "correct," they are simply the "best known" or "most popular" constructions leading to a reasonable notion of measure on uncountable sets.

Definition 14 (measure) Given a σ-algebra $\mathfrak{M}$ of subsets of a set S, a function $\mu : \mathfrak{M} \to [0, \infty]$ is a measure if

(i) $\mu(\phi) = 0$ and

(ii) If $A_1, A_2, \ldots$ is any sequence of **disjoint** subsets in $\mathfrak{M}$, then

$$\mu\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mu(A_j).$$

Property **(ii)** is called **countable additivity**, and in this context the σ-algebra is called the collection of **measurable sets**. The condition $A \in \mathfrak{M}$ is often expressed by saying simply A is **measurable**.

Generally, a measure $\pi : \mathfrak{M} \to [0, 1]$ with $\pi(S) = 1$ is called a **probability measure**, but usually (and certainly in this course) most measures outside of baby and adolescent measures on intervals in $\mathbb{R}$ are constructed using one very special measure:

Theorem 10 (Lebesgue-Carathèodory theorem) There exists a unique measure $m_{\mathbb{R}} : \mathcal{M}_{\mathbb{R}} \to [0, \infty]$ where $\mathcal{M}_{\mathbb{R}}$ is the largest possible σ-algebra of subsets of $\mathbb{R}$ and $m_{\mathbb{R}}(J)$ is the length of J for every interval J. The (regular) restriction of this measure to any subinterval $I \subset \mathbb{R}$ has the same property: The restriction of $m_{\mathbb{R}}$ to I is the unique measure on the interval I defined on the largest possible σ-algebra and having the measure of every (sub)interval J given by the length of J. We may denote restrictions of the measure $m_{\mathbb{R}}$ by

$$m = m_{\mathbb{R}}\big|_{\mathcal{M}}.$$

This measure, whenever it is used on any interval I, is called **Lebesgue measure** and of course the σ-algebra $\mathcal{M}$ sometimes denoted by $\mathcal{L}$, called the collection of **Lebesgue measurable sets**, is a proper subset of $\wp(I)$:

$$\mathcal{M} = \{A \cap I : A \in \mathcal{M}_{\mathbb{R}}\} \subsetneqq \wp(\mathbb{R}).$$

Note: The symbol m (along with symbols like i, j, k, ℓ, and n) is often used to represent a natural number. If one is worried about confusing a natural number m with the Lebesgue measure m, then one can use m_I for the Lebesgue measure, but usually the context and usage make the meaning clear.

Note: The phrasing above "...where $\mathcal{M}_{\mathbb{R}}$ is the largest possible σ-algebra of subsets of $\mathbb{R}$..." is not the best. Of course, $\wp(\mathbb{R})$ is the largest possible σ-algebra of subsets of $\mathbb{R}$. This phrasing is used because I wanted to give a quick introduction emphasizing the existence of the Lebesgue-Carathèodory measure. The actual meaning is something like this: First of all, there exists a **smallest** σ-algebra $\mathcal{B} \subset \wp(\mathbb{R})$ containing all intervals, and there exists a unique measure $m_0 : \mathcal{B} \to [0, \infty]$ for which the measure of every interval is its length. The σ-algebra $\mathcal{B}$ is called the **Borel σ-algebra**. Technically, one cannot extend m_0 to a translation invariant measure m on $\wp(\mathbb{R})$, and there turns out to be a largest σ-albegra $\mathcal{M}_{\mathbb{R}}$ for which such an extension is possible. So this is the basic meaning. I'm not entirely sure if translation invariance is a necessary restriction. That is, I do not know if there exists a non-translation invariant extension of m_0 to a σ-algbra larger than $\mathcal{M}_{\mathbb{R}}$ (or perhaps even to $\wp(\mathbb{R})$), but translation invariance seems to be required in every reference I know or can find.

One can note immediately that Lebesgue measure on an interval I is a probability measure precisely when the length of I is $m(I) = 1$. Given **an interval I of finite positive length** a probability measure can always be obtained by scaling Lebesgue measure: $\pi : \mathcal{M} \to [0, 1]$ by

$$\pi(A) = \frac{m(A)}{m(I)}$$

is called the **uniform probability measure on the interval I**.

Important note: From now on, whenever we mention a probability measure π on an interval in the real line (unless it is explicitly a generalized baby or adolescent measure) we will assume the domain of π is the collection $\mathcal{M}$ of Lebesgue measurable sets. This is the first indication of the importance of the Lebesgue-Carathèodory theorem.

There is no uniform probability measure on an unbounded interval.

Exercise 4.0.1 Show the uniform measure on a bounded interval I is a measure and is a probability measure.

Exercise 4.0.2 Let π be a uniform probability measure on an interval I. Show that given any $\ell > 0$ the measure π is constant on every collection

$$\{J \subset I : \ J \text{ is an interval with } m(J) = \ell\}.$$

Exercise 4.0.3 Let us make the following definition: A probability measure π on an interval $I \subset \mathbb{R}$ is said to be a **uniform probability measure** if given any $\ell > 0$ the measure π is constant on every collection

$$\{J \subset I : \ J \text{ is an interval with } m(J) = \ell\}.$$

(a) Show there does not exist a uniform probability measure on any unbounded interval.

(b) Show a uniform probability measure on a bounded interval is given by a scaling of Lebesgue measure.

Exercise 4.0.4 Show any measure μ satisfies the monotonicity property:

If A and B are measurable sets and $A \subset B$, then $\mu(A) \leq \mu(B)$.

4.1 Integration

Whenever we have a measure space, say an interval I, we can integrate (some) real valued functions $x : I \to \mathbb{R}$. The following is a little bit of a subtlty which is not of central interest to us but is good to know: Just like we can't measure all sets, we cannot expect to integrate all functions. The functions we can integrate are called **measurable functions**. I won't get into the technical details of what it takes to make a function measurable, but I'll simply state two facts about measurable functions:

1. Not all functions are measurable, and hence not all functions can be integrated.

2. Just about every function you can think of is measurable.

   - Every continuous function is measurable.
   - Every function which is continuous except on a countable subset of I is measurable.
   - Many other functions, discontinuous on much larger sets, like

     $$x(\xi) = \begin{cases} 0, & \xi \in I \cap \mathbb{Q} \\ 1, & \xi \in I \backslash \mathbb{Q}, \end{cases}$$

     are measurable.

4.1.1 Lebesgue integration (technical details)

One question you might be asking is: How does integration with respect to Lebesgue measure work? That is, we know how to integrate a function $x : S \to \mathbb{R}$ with respect to a baby measure or an adolescent measure using Riemann sums, you may also know the definition of a Riemann integral

$$\int_a^b x(\xi)\, d\xi = \lim_{\|\mathcal{P}\|\to 0} \sum_{j=0}^{n-1} x(\xi_j)\ (a_{j+1} - a_j) \tag{4.1}$$

using Riemann sums, but the kind of integral we have here

$$\int_I x$$

might be something different (and in principle it is), so what is the definition of the **Lebesgue integral**?

I will give you a short answer to this question, but first let me briefly review the definition of the Riemann integral. Actually, this discussion already appears on page 24 of these notes in Lecture "-1." I will just repeat it here with a change of notation: Specifically, the expression and the limit in (4.1) are somewhat useful to review and understand (or to learn for the first time if you haven't learned them before). In the expression (4.1) one starts with an interval $I = [a, b]$ and considers a **partition** $\mathcal{P}$ which is a collection of points

$$\mathcal{P} = \{a_j\}_{j=0}^n \qquad \text{with} \qquad a = a_0 < a_1 < a_2 < \cdots < a_n = b.$$

The **norm of the partition** is

$$\|\mathcal{P}\| = \max\{a_{j+1} - a_j : j = 0, 1, 2, \ldots, n-1\};$$

the points ξ_j for $j = 0, 1, 2, \ldots, n-1$ are called **evaluation points** and satisfy

$$a_j \le \xi_j \le a_{j+1}.$$

The sum

$$\sum_{j=0}^{n-1} x(\xi_j)\ m([a_j, a_{j+1}]) \tag{4.2}$$

is called a **Riemann sum**. Note that we have introduced the Lebesgue measure here with $m([a_j, a_{j+1}]) = a_{j+1} - a_j$, but we are only measuring the lengths of intervals so nothing like the full power of the Lebesgue measure is needed. It is clear that (4.2) is a sum of function values weighted by measures of associated sets (intervals). The meaning of the limit is that

$$L = \int_a^b x(\xi)\, d\xi$$

is the (unique) number for which the following holds:

> Given any $\epsilon > 0$, there is some $\delta > 0$ such that for any partition $\mathcal{P}$ (and any associated evaluation points) for which $\|\mathcal{P}\| < \delta$, there holds
> $$\left| \sum_{j=0}^{n-1} x(\xi_j)\ m([a_j, a_{j+1}]) - L \right| < \epsilon.$$

A continuous function on a closed interval always has a well-defined Riemann integral.

Exercise 4.1.1 Use the definition of the Riemann integral to calculate the Riemann integral of the function $x : [0, 1] \to \{0, 1\}$ by

$$x(\xi) = \begin{cases} 0, & \xi \in [0, 1/2) \\ 1, & \xi \in (1/2, 1]. \end{cases}$$

Exercise 4.1.2 Show the Riemann integral of the function $x : [0, 1] \to \{0, 1\}$ by

$$x(\xi) = \begin{cases} 0, & \xi \in \mathbb{Q} \\ 1, & \xi \notin \mathbb{Q} \end{cases}$$

does not exist.

For Lebesgue measure we proceed somewhat differently. First let me restrict attention to functions $p : I \to [0, \infty)$ with **non-negative values**. This will be the case of primary interest in the next section anyway. Then I will proceed in two steps:

1. For **simple functions** we can use the same Riemann sum approach we used when integrating with respect to baby measures.

2. For general non-negative functions we approximate using simple functions.

Simple functions are built using even simpler functions called **characteristic functions** which are useful to know about by themselves. Given a measurable set $A \in \mathcal{M}$, the characteristic function $\chi_A : I \to \{0,1\} \subset \mathbb{R}$ is given by

$$\chi_A(\xi) = \begin{cases} 1, & \xi \in A \\ 0, & \xi \notin A. \end{cases}$$

A **simple function** $s : I \to \mathbb{R}$ is a linear combination of simple functions, that is, there are (finitely many) real constants $c_1, c_2, \ldots, c_k \in \mathbb{R}$ and corresponding measurable sets $A_1, A_2, \ldots, A_k \in \mathcal{M}$ such that

$$s(\xi) = \sum_{j=1}^{k} c_j \chi_{A_j}(\xi).$$

A simple function is a **non-negative simple function** if the constants $c_1, c_2, \ldots, c_k$ are all non-negative.

Exercise 4.1.3 Give an example of a simple function

$$s = \sum_{j=1}^{k} c_j \chi_{A_j}$$

which is non-negative but not all of the constants $c_1, c_2, \ldots, c_k$ are non-negative.

Exercise 4.1.4 Show any non-negative, nonzero simple function $s : I \to \mathbb{R}$ has a unique form

$$s = \sum_{j=1}^{k} c_j \chi_{A_j}$$

in which all of the constants $c_1, c_2, \ldots, c_k$ are positive and distinct, and the sets $A_1, A_2, \ldots, A_k$ are (pairwise) disjoint.

Here is our first main definition: The Lebesgue integral of a simple function $s : I \to \mathbb{R}$ by

$$s(\xi) = \sum_{j=1}^{k} c_j \chi_{A_j}(\xi), \tag{4.3}$$

is

$$\int_I s = \sum_{j=1}^{k} c_j m(A_j). \tag{4.4}$$

Exercise 4.1.5 Show the value of the integral of a simple function $s : I \to \mathbb{R}$ given in (4.4) is independent of the constants $c_1, c_2, \ldots, c_k$ and the measurable sets $A_1, A_2, \ldots, A_k$ in the form (4.3).

Here is the second step defining the Lebesgue integral of any non-negative Lebesgue measurable function $x : I \to [0, \infty)$:

$$\int_I x = \sup \left\{ \int_I s : s \text{ is a non-negative simple function with } s \leq x \right\}.$$

At least we have defined the Lebesgue integral for non-negative functions. In particular, we may also defined integrals

$$\int_A x = \int_I x\chi_A$$

over measurable subsets A of I and averages

$$\frac{1}{m(A)} \int_A x.$$

These last two formulas are now well defined whenever $x : I \to \mathbb{R}$ takes non-negative values. There are a few more complications one encounters if a (measurable) function $x : I \to \mathbb{R}$ changes sign. We will not address these complications, but we will use the notation above which carries over to all functions under consideration.

4.1.2 Probability measures and MDF

We have already classified all the uniform probability measures on intervals above. Note that these measures $\pi : \mathcal{M} \to [0, 1]$ have values given by

$$\pi(A) = \frac{1}{m(I)} \int_A 1 = \frac{1}{m(I)} \int_I \chi_A. \tag{4.5}$$

Consequently, these uniform probability measures may be called **uniform integral measures**.

We now generalize the integral form of the uniform probability measures on intervals. A non-negative measurable function $\delta : I \to [0, \infty)$ is called a **mass density function** (MDF) if

$$\int_I \delta = 1.$$

Given any mass density function, the function $\pi : \mathcal{M} \to [0, 1]$ by

$$\pi(A) = \int_A \delta \tag{4.6}$$

is a probability measure on I. Here, the interval I does not need to be bounded.

Exercise 4.1.6 Show the function $\pi : \mathcal{M} \to [0, 1]$ with values given in (4.6) is a measure and is a probability measure.

Not every measure falls into the two categories of generaliezed baby and adolescent measures and the measures defined by integration of a MDF as in (4.6), but we are only going to consider measures of these two types in this course. In most books on probability and statistics, the first kind of measures (generalized baby and adolescent) are called "discrete" measures, which is not really a very good name for them, but I'm not going to explain why. The measures under consideration in this chapter/lecture given by Lebesgue integration of a MDF are called **continuous** measures or more properly **absolutely continuous** measures. See the next section.

4.1.3 Uniqueness of the MDF

Given an absolutely continuous measure π, the mass density function can be recovered from the values of the measure by the formula

$$\delta(\xi) = \lim_{\epsilon \searrow 0} \frac{1}{2\epsilon} \pi((\xi - \epsilon, \xi + \epsilon)). = \lim_{\epsilon \searrow 0} \frac{\pi((\xi - \epsilon, \xi + \epsilon))}{m((\xi - \epsilon, \xi + \epsilon))}.$$

This is called the Radon-Nikodym derivative of the measure π with respect to Lebesgue measure m. We will later express the MDF as the derivative of a different function, namely the CMF (re)introduced in section 4.7 below.

4.1.4 Comparison of PMF and MDF

One may compare the mass density function (MDF) to the probability mass function (PMF) for a baby measure (or an induced generalized baby measure) on the real line. The PMF gives the values of the measures of singletons:

$$M(\xi) = \alpha(\{\xi\}).$$

These values are always in the interval $[0, 1]$. A mass density function (MDF) on the other hand, also takes non-negative values, but these are not the measures of singletons nor is there any restriction concerning how large the values may be. The only restriction is the integral restriction

$$\int_I \delta = 1.$$

Also, since singletons have Lebesgue measure zero, the measure $\pi : \mathcal{M} \to [0, 1]$ determined by an MDF always satisfies

$$\pi(\{\xi\}) = \int_{\{\xi\}} \delta = 0$$

for every singleton $\{\xi\} \subset \mathbb{R}$. The measure of singletons is not of interest for a measure on an interval. Of primary interest is the value of such a measure on subintervals:

$$\pi(J) = \int_J \delta$$

where this is an integral with respect to Lebesgue measure.

We didn't mention it explicitly, but there is a special measure analogous to Lebesgue measure which is used (or at least may be used) to express the values of any generalized baby measure or generalized adolescent measure on $\mathbb{R}$.

Exercise 4.1.7 Identify the measure $\mu : \wp(\mathbb{R}) \to [0, \infty]$ for which any generalized baby or adolescent probability measure $\alpha : \wp(\mathbb{R}) \to [0, 1]$ has values given by

$$\pi(A) = \int_A M$$

where M is the PMF associated directly with α, i.e., induced by the identity function $\mathrm{id}_{\mathbb{R}} : \mathbb{R} \to \mathbb{R}$.

4.1.5 Integration with respect to an integral measure

Remember that whenever we have a measure on a measure space, there is (or at least should be) a way to integrate real valued functions with respect to the measure. We've talked about Lebesgue integration, but we can also talk about integration with respect to an integral measure defined in terms of Lebesgue integration. Specifically, say I is an interval in the real line, and $\delta : \mathbb{R} \to [0, \infty)$ is a (probability) MDF with support or statistical range in I. Then the measure $\pi : \mathcal{M} \to [0, 1]$ determined by δ has values given by

$$\pi(A) = \int_A \delta.$$

where the integral on the right is with respect to Lebesgue measure, i.e. length measure.

With this in the background, we can take a real valued (measurable) function $x : I \to \mathbb{R}$ or even $x : \mathbb{R} \to \mathbb{R}$ and integrate x with respect to the measure π. The value of such an integral is given by

$$\int_{\mathbb{R}} x = \int_{\mathbb{R}} x\delta \tag{4.7}$$

where the integral on the left is with respect to π, and the **definition of that integral** is the integral on the right which is an integral with respect to Lebesgue measure. Put another way, integration of the function $x : \mathbb{R} \to \mathbb{R}$ with respect to the integral measure π **means** integration of the product of x with the MDF of π with respect to Lebesgue measure.

The notation of (4.7) as in our discussion of integration with respect to baby measures has some evident level of confusion associated with it. On the one hand, one can simply say that when integrating, **one needs to know** the measure with respect to which one is integrating. On the other hand, if some notational distinction is needed, I would suggest

$$\int_{(\mathbb{R},\pi)} x = \int_{(\mathbb{R},m)} x\delta$$

where m denotes Lebesgue measure.

4.2 Initial Examples

As pointed out by Orloff and Booth every measure on an interval I is naturally considered as a measure on the entire real line. This is accomplished

by the use of a characteristic function. For example if we take the uniform probability measure $\pi : \mathcal{M} \to [0,1]$ on $I = [0, 1/3]$ considered as Example 2 in the Class 5 notes of Orloff and Booth, then the values of π are given by

$$\pi(A) = \int_A \delta$$

where $\delta(\xi) \equiv 3$. This works for any $A \subset I = [0, 1/3]$. However, if we take $\delta(\xi) = 3\chi_I$, then we can consider the expanded measure with values

$$\pi(A) = \int_A \delta = \int_{\mathbb{R}} 3\chi_{A\cap I}.$$

This is, of course, no longer a uniform measure, as we know there are no uniform integral measures on the entire real line. Nevertheless, this is common practice and, as mentioned by Orloff and Booth, computations like

$$\begin{aligned}\pi([0.1, 0.2]) &= \pi(\{\omega : 0.1 \le \omega \le 0.2\}) \\ &= \int_{0.1}^{0.2} 3\, d\omega \\ &= \frac{3}{10}\end{aligned}$$

and

$$\begin{aligned}\pi([0.1, 1]) &= \pi(\{\omega : 0.1 \le \omega \le 1\}) \\ &= \pi(\{\omega : 0.1 \le \omega \le 1/3\}) \\ &= \int_{0.1}^{1/3} 3\, d\omega \\ &= 3\left(\frac{1}{3} - \frac{1}{10}\right) \\ &= \frac{7}{10}\end{aligned}$$

pass for the values of the "probabilities"

$$P(0.1 \le \omega \le 0.2) \qquad \text{and} \qquad P(0.1 \le \omega \le 1).$$

Figure 4.1 illustrates the MDF of the uniform integral measure on the interval $I = [0, 1/3]$ and areas associated with the value of the measure on two

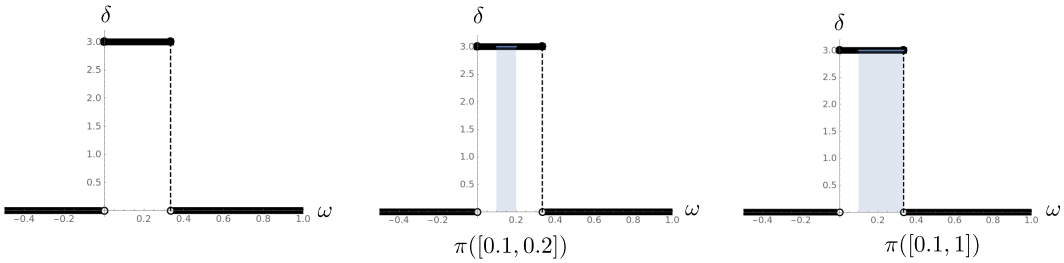

Figure 4.1: The MDF associated with a uniform probability measure on the interval $I = [0, 1/3]$. See Orloff and Booth page 4, Class 5 notes (second installment) for a nicer version of this illustration.

intervals mentioned (and calculated) above. Notice the value of the MDF does not give the values of measures (like a PMF), but it is still the case that the MDF indicates the location of sets with larger measure. In particular, for this meausre $\pi([a, b]) = 0$ whenever $a \geq 1/3$.

As a second example[3] of an integral measure, we can take

$$\delta(\omega) = 2\omega\, \chi_{[0,1]} \qquad \text{or} \qquad \delta(\omega) = \frac{\omega}{2}\, \chi_{[0,2]}. \tag{4.8}$$

Exercise 4.2.1 Identify the probability measure $\pi : \mathcal{M} \to [0, 1]$ determined by each of the probability densities given in (4.8). Plot the density functions and illustrate the measures

$$\pi([0.1, 0.2]) \qquad \text{and} \qquad \pi([0.1, 1])$$

for each measure.

[3]Orloff and Booth mention $\delta(\omega) = (\omega/2)\, \chi_{[0,1]}(\omega)$ at the bottom of page 4 of their Class 5 notes (second installment), but this seems to be a typo. Incidentally, Orloff and Booth offer the Class 5 notes in four installments as follows:

1. Variance of Discrete Random Variables
2. Continuous Random Variables (second installment)
3. Gallery of Continuous Random Variables
4. Manipulating Continuous Random Variables

The present section of my notes is based on the first three installments of these Class 5 notes, and much of the material is presented in different language there.

The measures associated with the MDFs given in (4.8) are not uniform on their support (or range) intervals. Recall that previously we had defined the **statistical range** of a (baby or adolescent) probability measure π on $\mathbb{R}$ to be the smallest interval containing all the singletons $\xi \in \mathbb{R}$ for which $\pi(\{\xi\}) > 0$. We have a similar definition for integral measures: The **statistical range** of an integral measure with MDF δ is the smallest closed interval containing the set $\{\omega \in \mathbb{R} : \delta(\omega) > 0\}$.

4.3 Two important examples

There are many integral measures. Aside from the uniform integral measures introduced above and discussed further below, we will focus only on two or three others. Here we introduce the **exponential probability measure** and the **normal probability measure**. As usual for integral measures, the identity on $\mathbb{R}$ may be considered to induce these measures and they may be (and often are) referred to as the exponential distribution and the normal distribution.

4.3.1 Exponential distribution

This is a one parameter family of measures $\gamma = \gamma_\lambda$ determined by $\lambda > 0$ with MDF

$$\delta(\omega) = \lambda e^{-\lambda\omega}\ \chi_{[0,\infty)}(\omega).$$

The measure itself is given (on subintervals $[a, b]$ with $0 \leq a < b$) by

$$\gamma([a, b]) = \lambda \int_a^b e^{-\lambda\omega}\, d\omega = e^{-a\lambda} - e^{-b\lambda}. \tag{4.9}$$

You may recall that we used the same symbol γ for the geometric probability measure, and there is some nominal relation between the two. To see this, note first that the PMF of the geomtric probability measure

$$M(\omega) = \begin{cases} (1-p)^k p, & \omega = k \in \mathbb{N}_0 \\ 0, & \omega \in \mathbb{R}\backslash\mathbb{N}_0 \end{cases}$$

naturally gives rise to a MDF for a (Lebesgue) integral measure:

$$\delta = \sum_{j=0}^{\infty} (1-p)^j p\ \chi_{[j,j+1)}(\omega).$$

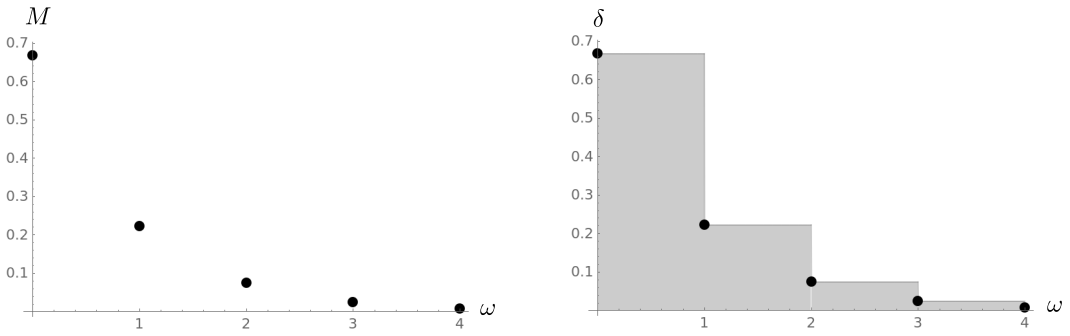

Figure 4.2: The probability mass function (PMF) of the geometric probability measure (an adolescent measure) and an associated mass density function (MDF) δ for an integral probability measure.

These functions are illustrated in Figure 4.2. You may recall that we used the gamma function to find a function of a continuous variable matching the values of the PMF of the binomial distribution. Clearly, we can find a similar function for the geometric probability measure as well. That function is given by

$$m(\omega) = (1-p)^{\omega} p \ \chi_{[0,\infty)}(\omega)$$

and can be expressed in the (striking) form

$$m(\omega) = p e^{-\omega \ln\left(\frac{1}{1-p}\right)} \ \chi_{[0,\infty)}(\omega).$$

It should be clear that while we may define an integral measure

$$\mu(A) = \int_A m$$

using the mass density m, this will **not** be a probability measure. Clearly, a scaling of this measure (and a scaling of the associated MDF) does give a probability measure, and this measure is an exponential probability measure:

$$\delta(\omega) = \ln\left(\frac{1}{1-p}\right) e^{-\omega \ln\left(\frac{1}{1-p}\right)} \ \chi_{[0,\infty)},$$

$$\gamma(A) = \ln\left(\frac{1}{1-p}\right) \int_{\omega \in A \cap [0,\infty)} e^{-\omega \ln\left(\frac{1}{1-p}\right)}.$$

Notice the parameter λ is given in this case by

$$\lambda = \ln\left(\frac{1}{1-p}\right).$$

Exercise 4.3.1 Is it clear that

$$p < \ln\left(\frac{1}{1-p}\right)$$

for $0 < p < 1$?

Alternatively, we can change the scaling in the exponent to obtain a different exponential probability measure corresponding to the parameter $\lambda = p$:

$$\delta(\omega) = pe^{-p\omega}\ \chi_{[0,\infty)},$$

$$\gamma(A) = p\int_{\omega\in A\cap[0,\infty)} e^{-p\omega}.$$

Notice these latter exponential measures only correspond to $\lambda = \delta(0) = p$ with $0 < p < 1$ while in general the exponential probability measure may be considered for any $\lambda > 0$. The MDFs of these last two measures are illustrated in Figure 4.3.

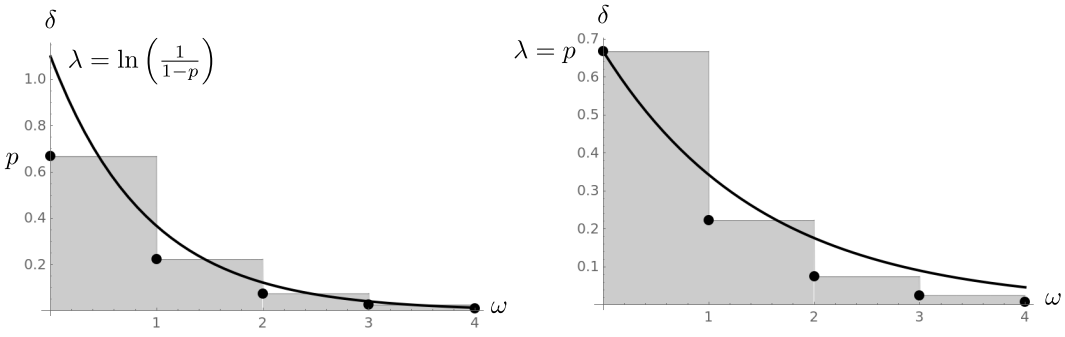

Figure 4.3: The MDF of the exponential probability measure with $\lambda = \ln(1/(1-p))$ (left) adn the MDF of the exponential probability measure with $\lambda = p$ (right).

Exercise 4.3.2 Calculate

$$F(\omega) = \int_{(-\infty,\omega]} \lambda e^{-\lambda\omega}\ \chi_{[0,\infty)}(\omega).$$

The function $F : \mathbb{R} \to [0, 1)$ is called the **cumulative distribution function** of the exponential measure. It is essentially the same as the cumulative mass function (CMF) encountered in connection with generalized baby and adolescent measures.

Finally, one may note from the expression (4.9) that

$$\frac{d}{d\lambda}\gamma([a, b]) = -ae^{-a\lambda} + be^{-b\lambda} = e^{-b\lambda}\left(b - ae^{(b-a)\lambda}\right).$$

This quantity is negative for (a and b fixed and) λ large enough. This is one indication that the mass density shifts toward the center as λ increases. We will attempt to quantify this kind of observation more precisely below.

The parameter λ is called the **rate constant**. The range $[0, \infty)$ of the exponential measure is often used to model **waiting times**. If something happens repeatedly, say a clock ticks, on average $\lambda = 60$ times per minute, then on average the time between occurrences (that is the average time one must wait between occurrences) for example ticks of the clock, is $1/\lambda = 1/60$ minutes or one second.

Simulation using the exponential measure gives quite a bit more variability than one would expect from the ticking of a clock. In particular, the R commands

```
set.seed(24)
sum(rexp(60,rate=60))
```

which are supposed to return the sum of 60 simulated wait times with rate constant $\lambda = 60$, produce the output 0.8218647. I guess that is a New York City minute. In contrast, the analogous commands

```
SeedRandom[24];
waittimes = RandomVariate[ExponentialDistribution[60],
60]; Sum[waittimes[[j]], j, 1, 60]
```

produce output 1.24851 in Mathematica.

4.3.2 Normal distribution

By all accounts, this is the most important integral probability measure on $\mathbb{R}$. The normal measure(s) are given by a two parameter family with parameters μ and σ. The parameter μ called the mean may be any real number and the parameter σ called the standard deviation should be positive. Here are the formulas:

$$\delta(\omega) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(\omega-\mu)^2}{2\sigma^2}}$$

and

$$\pi(A) = \int_A \delta.$$

Notice first that the MDF δ is positive on the entire real line as indicated in Figure 4.4 where we have taken $\sigma = 1$ and $\mu = 0$. This MDF is also called a **Gaussian** after Carl Friedrich Gauss who was one of the first people to consider it carefully, and with this in mind we may sometimes denote the normal MDF by $g : \mathbb{R} \to [0, \infty)$ instead of $\delta : \mathbb{R} \to [0, \infty)$ and the normal measure by $G : \mathcal{M} \to [0, 1]$ instead of $\pi : \mathcal{M} \to [0, 1]$. The MDF associated

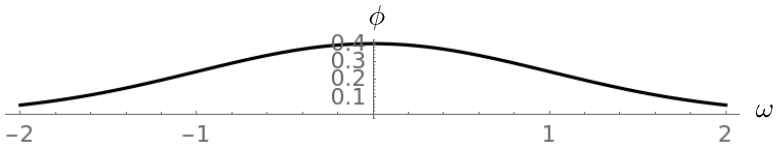

Figure 4.4: The MDF associated with a centrally symmetric normal probability measure with $\sigma = 1$.

with these particular values $\mu = 0$ and $\sigma = 1$ is denoted by Orloff and Booth by ϕ. All other normal mass density functions are obtained from this one by scaling and translation:

$$\delta(\omega) = \frac{1}{\sigma}\, \phi\left(\frac{1}{\sigma^2}(\omega - \mu)\right).$$

The fact that the normal probability measure is a probability measure is a famous exercise in elementary mathematics. It may be, however, a famous exercise in elementary mathematics with which you are not familiar. This can be a time to remedy that. The careful derivation, as with many things in

measure theory, involves a new kind of measure. You should be familiar with this one in some form from elementary calculus. Namely, one traditionally uses here the **product measure on** $\mathbb{R}^2$. Lebesgue product measures are more complicated in principle than the product measures considered above for measure spaces with finitely many elements. Recall or note that we have not actually gone through the details of constructing Lebesgue measure on $\mathbb{R}$ (the measure for which an interval is its length). This measure comes from the Lebesgue-Carathèodory theorem which we just stated. There does exist a nice product measure on $\mathbb{R}^2$: There is an appropriate largest σ-algebra $\mathcal{M}$ of subsets of $\mathbb{R}^2$ and a corresponding measure μ for which the measure of a rectangle (for example, or a disk or any other elementary region in the plane) is its **area**. Associated with this measure is an integration, i.e., with the Lebesgue area measure on $\mathbb{R}^2$ one can integrate real valued functions: If $A \subset \mathbb{R}^2$ is a measureable set and $x : A \to \mathbb{R}$ is a measurable function, then the integral

$$\int_A x \tag{4.10}$$

with respect to the Lebesgue area measure is well-defined. The procedure for defining such an integral is not so different from the one-dimensional case we discussed a little bit above. In any case, these kinds of integrals are considered in elementary calculus mostly in the form of what are called "double integrals." The relation comes through a result called Fubini's theorem which says that in certain cases a two-dimensional Lebesgue integral of the form (4.10) can be expressed (or computed) as two iterated one-dimensional integrals. For example, if $A = [a, b] \times [c, d]$ is a rectangular domain in $\mathbb{R}^2$ and the function $x : A \to \mathbb{R}$ happens to be continuous so that it is also Riemann integrable, then

$$\int_A x = \int_{[a,b]} \left(\int_{[c,d]} x \right) = \int_c^d \left(\int_a^b x(\omega_1, \omega_2)\, d\omega_1 \right)\, d\omega_2.$$

The last expression is one you should find (at least a little) famiilar. There are various manifestations of this kind of "double integration" one of which is the change of variables leading to "integration with respect to polar coordinates." In this case, the region A has a form something like

$$A = \{(r\cos\theta, r\sin\theta) \in \mathbb{R}^2 : r_1 \le r \le r_2,\ \theta_1 \le \theta \le \theta_2\}$$

corresponding to a rectangular domain $[r_1, r_2] \times [\theta_1, \theta_2]$ in $\mathbb{R}^2$, and

$$\int_A x = \int_{r_1}^{r_2} r \left(\int_{\theta_1}^{\theta_2} x(r\cos\theta, r\sin\theta)\, d\theta \right) dr.$$

With these preliminaries, let us start with the basic form of the function appearing in the Gaussian ϕ namely $g(\omega) = e^{-\omega^2}$. You may note that this function g is not quite a normal mass density function, but if you take $\sigma = 1/\sqrt{2}$ and $\mu = 1$, you almost get this function. In any case, we'll use two of these. Note first that since g is clearly even

$$\int_{(-\infty,\infty)} g = 2 \int_{(0,\infty)} g \geq 0.$$

Furthermore, g is continuous, so this (half) value can be expressed as a Riemann integral:

$$\int_0^\infty g(\omega)\, d\omega.$$

More precisely, this is what is called an improper Riemann integral, but that is only a minor complication. Now, as mentioned, we take two of them:

$$\frac{G(\mathbb{R})}{2} = \int_0^\infty g(\omega_1)\, d\omega_1 = \int_0^\infty g(\omega_2)\, d\omega_2.$$

Consequently,

$$\begin{aligned}
\left(\frac{G(\mathbb{R})}{2}\right)^2 &= \left(\int_0^\infty g(\omega_1)\, d\omega_1 \right) \left(\int_0^\infty g(\omega_2)\, d\omega_2 \right) \\
&= \int_0^\infty \left(\int_0^\infty g(\omega_1)\, d\omega_1 \right) g(\omega_2)\, d\omega_2 \\
&= \int_0^\infty \left(\int_0^\infty g(\omega_1) g(\omega_2)\, d\omega_1 \right) d\omega_2 \\
&= \int_0^\infty \left(\int_0^\infty e^{-(\omega_1^2+\omega_2^2)}\, d\omega_1 \right) d\omega_2
\end{aligned}$$

The last expression is an iterated integral, and thus a two-dimensional Lebesgue integral (or an area integral) by Fubini's theorem as discussed above. The planar domain of integration here is the first quadrant

$$A = \{(\omega_1, \omega_2) \in \mathbb{R}^2 : \omega_1, \omega_2 \geq 0\} = \{(r\cos\theta, r\sin\theta) : 0 \leq r < \infty,\ 0 \leq \theta \leq \pi/2\}.$$

This is essentially the form suited to a change of variables to polar coordinates:

$$\int_{(\omega_1,\omega_2)\in A} e^{-(\omega_1^2+\omega_2^2)} = \int_0^\infty r \left(\int_0^{\pi/2} e^{-r^2}\, d\theta \right) dr.$$

It will be noted that the θ dependence in the integrand has been eliminated, and calculation of these integrals is now elementary:

$$\begin{aligned} \left(\frac{G(\mathbb{R})}{2}\right)^2 &= \int_0^\infty r \left(\frac{\pi}{2}\right) e^{-r^2}\, dr \\ &= \frac{\pi}{2} \int_0^\infty r\, e^{-r^2}\, dr \\ &= -\frac{\pi}{4}\, e^{-r^2}\Big|_{r=0}^{\infty} \\ &= \frac{\pi}{4}. \end{aligned}$$

It follows that

$$G(\mathbb{R}) = \int_{-\infty}^{\infty} e^{-\omega^2}\, d\omega = \sqrt{\pi}.$$

This is, of course, not 1, and the measure we have denoted G here is not a probability measure. It is not a normal (or Gaussian) measure either, but it is only off by a factor:

If $g : \mathbb{R} \to [0, \infty)$ by

$$g(\omega) = \frac{1}{\sqrt{\pi}} e^{-\omega^2},$$

then g is the MDF of the normal/Gaussian measure with $\mu = 0$ and $\sigma = \sqrt{2}$. Furthermore, we have shown the corresponding measure $G : \mathcal{M} \to [0, \infty]$ with

$$G(A) = \int_{\omega\in A} \frac{1}{\sqrt{\pi}} e^{-\omega^2}$$

satisfies $G(\mathbb{R}) = 1$ and is therefore a probability measure.

Now consider the "standard Gaussian" given by

$$\phi(\omega) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{\omega^2}{2}}.$$

In this case we can use the change of variables $t = \omega/\sqrt{2}$ so that $d\omega = \sqrt{2}\,dt$ and

$$\int_{\mathbb{R}} \phi = \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} e^{-t^2}\,dt = 1.$$

In general, we can change variables with

$$t = \frac{\omega - \mu}{\sigma\sqrt{2}}$$

to conclude the normal measure is a probability measure for every choice of parameters $\mu \in \mathbb{R}$ and $\sigma > 0$.

4.3.3 Basic interpretation of normal distribution

The meaure space $\mathbb{R}$ is very often used to model "random" measurements of physical characteristics like height and weight for living organisms. Such measurements are often simulated using a normal or Gaussian distribution. To accomplish this, one needs to decide on values for the mean μ and the standard deviation σ. The meaning of "mean" here is relatively straightforward: One wishes to produce simulated data with mean μ. We will discuss the meaning of the variance σ^2 and the standard deviation σ below.

Physical measurements are not the only kinds of "measurements" simulated using the normal distribution. Intelligence quotients (IQ) and simple errors in physical measurement, among many other things, are often simulated using the normal distribution.

4.4 Induced integral measures

We have introduced integral measures directly as probability measures on intervals in the real line. The construction of an induced probability measure in this context is also worth understanding: Say we have an integral probability measure $\pi_0 : \mathcal{M} \to [0,1]$ on an interval $I_0 \subset \mathbb{R}$ and a real valued function $x : I_0 \to \mathbb{R}$. Then the **induced measure** $\pi : \mathcal{M} \to [0,1]$ is given as before by

$$\pi(A) = \pi_0(x^{-1}(A)). \tag{4.11}$$

If we have a MDF $\delta_0 : I_0 \to \mathbb{R}$ for π_0 and the inducing function x is one-to-one (essentially a renaming) and has a positive derivative

$$\frac{dx}{d\omega} > 0,$$

then the image $I = x(I_0)$ is also an interval, and the MDF $\delta : I \to [0, \infty)$ of the induced measure is obtained by a change of variables:

$$\pi(A) = \int_{x^{-1}(A)} \delta_0 = \int_A \delta_0 \circ x^{-1} \Big/ \frac{dx}{d\omega} \circ x^{-1}.$$

That is, the MDF $\delta : I \to [0, \infty)$ of the induced measure is given by

$$\delta(\xi) = \delta_0 \circ x^{-1}(\xi) \Big/ \frac{dx}{d\omega}(x^{-1}(\xi)). \tag{4.12}$$

Technically, since x may not be surjective the natural restriction of the codomain to $I = x(I_0)$ in order to obtain an inverse may be necessary. More importantly if x is not one-to-one and/or not differentiable the situation may be substantially more complicated and the relation between the base MDF δ_0 and the induced MDF δ may be unclear. In particular, the formula (4.12) may not always apply, but the formula (4.11) is always valid.

Consider the centrally symmetric integral probability measure determined by $\delta_0(\omega) = (3/4)(1 - \omega^2)\ \chi_{[-1,1]}$. The measure π induced on $\mathbb{R}$ by $x(\omega) = \omega^2$ has

$$\begin{aligned}
\pi([a, b]) &= \int_{-\sqrt{b}}^{-\sqrt{a}} \delta_0(\omega)\, d\omega + \int_{\sqrt{a}}^{\sqrt{b}} \delta_0(\omega)\, d\omega \\
&= \frac{3}{2} \int_{\sqrt{a}}^{\sqrt{b}} (1 - \omega^2)\, d\omega \\
&= \frac{3}{4} \int_a^b \left(\frac{1}{\sqrt{\xi}} - \sqrt{\xi} \right)\, d\xi
\end{aligned}$$

for $0 \le a < b \le 1$. We see then that

$$\delta(\xi) = \frac{3}{4} \left(\frac{1}{\sqrt{\xi}} - \sqrt{\xi} \right)\ \chi_{[0,1]}.$$

Plots of the MDF δ_0 and the induced MDF δ are illustrated in Figure 4.5. It is essentially impossible to recover the base measure π_0 if one only knows the induced measure π and/or the induced MDF δ.

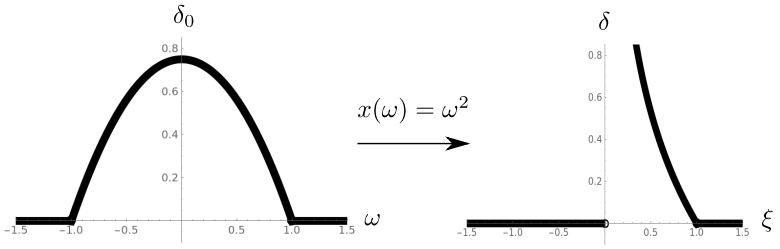

Figure 4.5: The MDF $\delta_0 = \delta_0(\omega)$ associated with a centrally symmetric probability measure on the interval $I = [-1, 1]$ and the induced MDF $\delta = \delta(\xi)$ associated with a probability measure with statistical range $[0, 1]$.

4.5 statistical values

If you look up[4] any of the probability measures (or "distributions") we have described, you should find a table of values. These may be called "statistical values" for reasons we will discuss later. Among these are the PMF (baby and adolescent measures) or MDF (integral measures) and the CMF or CDF. The CMF $F : \mathbb{R} \to [0, 1]$ for an integral probability measure is often called the **cumulative distribution function** (CDF) and is given by the same formula as in the case of (generalized) baby and adolescent measures:

$$F(\omega) = \pi((-\infty, \omega]) = \int_{(-\infty,\omega]} \delta.$$

We will discuss this quantity further below. (See also sections 4.3 and 4.4 of the Class 5 notes (second installment) of Orloff and Booth.) For now, simply note that this is the measure of a particular set. In regard to terminology, because cumulative distribution function (CDF) is the terminology most commonly used (especially in tables) but I think cumulative mass function (CMF) is more descriptive, I will (try to) write CMF/CDF. Similarly, probability density function (PDF) is used for MDF, so I will write MDF/PDF.

[4] using Wikipedia for example

4.5.1 Mean or expectation

On the table should also be the **mean** or **expectation**. If π is a baby or adolescent probability measure on a measure on $\mathbb{R}$, then the **mean** is given by

$$\mu = \sum_{\omega \in \mathbb{R}} \omega \ \pi(\{\omega\}).$$

If π is an integral measure on $\mathbb{R}$ with MDF δ, then the **mean** is given by

$$\mu = \int_{\omega \in \mathbb{R}} \omega \ \delta(\omega). \tag{4.13}$$

Of course, we can take the average value or mean of any real valued function on a measure space, and we've seen such averages before. These values may also be called the **expectation** and we've seen how they can be expressed in terms of the induced measure:

$$\int_S x = \int_{\mathbb{R}} \mathrm{id}_{\mathbb{R}} \tag{4.14}$$

where the integral on the left is with respect to the measure on S and the integral on the right is with respect to the measure induced by $x : S \to \mathbb{R}$ on $\mathbb{R}$. When we speak of the mean associated with a measure on $\mathbb{R}$ (by itself), we usually mean the average value of the identity function on $\mathbb{R}$ as defined here. Abstractly, the same thing is expressed in (4.13): One must always indicate the measure with respect to which one integrates on a measure space. The integral on the right in (4.13) is an integral with respect to Lebesgue measure, and this is the usual formula for the mean. But it is also perfectly correct to write down a version of (4.14) in this case:

$$\mu = \int_{\mathbb{R}} \mathrm{id}_{\mathbb{R}}. \tag{4.15}$$

In this expression the measure space $S = \mathbb{R}$ has two measures (Lebesgue measure and the probability measure π with MDF δ). The integral on the right in (4.15) is with respect to π.

Here are some examples of how to calculate the mean:

1. uniform measure on n points: In order to calculate a mean here we need[5] an injection $x : \{\omega_1, \omega_2, \ldots, \omega_n\} \to \mathbb{R}$. Say $x(\omega_j) = \xi_j$ for $j =$

[5] Actually, you are welcome to also consider what happens if $x : \{\omega_1, \omega_2, \ldots, \omega_n\} \to \mathbb{R}$ is any other inducing function.

$1, 2, \ldots, n$ with $\xi_i \neq \xi_j$ for $i \neq j$. Then

$$\mu = \sum_{j=1}^{n} \xi_j \alpha(\{j\}) = \frac{1}{n} \sum_{j=1}^{n} \xi_j$$

which is the arithmetic mean of the numbers $\xi_1, \xi_2, \ldots, \xi_j$.

2. Bernoulli measure:

$$\mu = 0 \ \beta(\{0\}) + 1 \ \beta(\{1\}) = p.$$

3. induced binomial measure/binomial distribution:

Recall that setting $S^n = \{0, 1\}^n$ there are functions $x_\ell : S^n \to \mathbb{R}$ for $\ell = 1, 2, \ldots, n$ given by $x_\ell(\omega) = \omega_\ell$ and

$$\begin{aligned} \int_{S^n} x_\ell &= \sum_{\omega \in S^n} \omega_\ell \ \beta(\{\omega\}) \\ &= p \int_{S^{n-1}} 1 \\ &= p. \end{aligned}$$

Also,

$$\mu = \sum_{j=0}^{n} j \ \alpha(\{j\}) = \int_{S^n} x$$

where $x : \{0, 1\}^n \to \mathbb{R}$ by

$$x = \sum_{\ell=1}^{n} x_\ell.$$

Therefore,

$$\mu = \sum_{\ell=1}^{n} \int_{S^n} x_\ell = np.$$

See also Assignment 2 Problem 4.

4. geometric probabilty measure:

$$\begin{aligned}
\mu &= \sum_{j=0}^{\infty} j(1-p)^j p \\
&= \sum_{j=1}^{\infty} j(1-p)^j p \\
&= p \sum_{\ell=1}^{\infty} \sum_{j=\ell}^{\infty} (1-p)^j \\
&= p \sum_{\ell=1}^{\infty} (1-p)^\ell \sum_{j=0}^{\infty} (1-p)^j \\
&= \sum_{\ell=1}^{\infty} (1-p)^\ell \\
&= \frac{1}{p} - 1 \\
&= \frac{1-p}{p}.
\end{aligned}$$

This is if one is counting failures (until success). Alternatively, if one is counting trials:

$$\begin{aligned}
\mu &= \sum_{j=1}^{\infty} j(1-p)^{j-1} p \\
&= p \sum_{\ell=1}^{\infty} \sum_{j=\ell}^{\infty} (1-p)^{j-1} \\
&= p \sum_{\ell=1}^{\infty} (1-p)^{\ell-1} \sum_{j=0}^{\infty} (1-p)^j \\
&= \sum_{\ell=1}^{\infty} (1-p)^{\ell-1} \\
&= \frac{1}{p}.
\end{aligned}$$

5. Poisson measure:

$$\begin{aligned}\mu &= \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} \\ &= e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!} \\ &= \lambda.\end{aligned}$$

   This is the meaning of the parameter λ.

6. exponential measure:

$$\begin{aligned}\mu &= \int_0^{\infty} \omega\lambda e^{-\lambda\omega}\,d\omega \\ &= -\omega e^{-\lambda\omega}\Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda\omega}\,d\omega \\ &= -\frac{1}{\lambda}e^{-\lambda\omega}\Big|_0^{\infty} \\ &= \frac{1}{\lambda}.\end{aligned}$$

   Here the parameter λ does not denote the mean, but rather the reciprocal of the mean. It is also called the **rate parameter**.

7. Gaussian measure: Here is a surprising situation in which the mean is much easier to compute than the measure of the entire space. Of course, we will also use that the measure of the space $G(\mathbb{R}) = 1$. We will also use a couple change of variables starting with the translation $\xi = \omega - \mu$. Also, in the following we start with μ on the left as a symbol to represent the mean and the same symbol on the right to denote the parameter μ in the Gaussian mass density.

$$\begin{aligned}\mu &= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} \omega e^{-\frac{(\omega-\mu)^2}{2\sigma^2}}\,d\omega \\ &= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} (\xi+\mu) e^{-\frac{\xi^2}{2\sigma^2}}\,d\xi \\ &= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} \xi e^{-\frac{\xi^2}{2\sigma^2}}\,d\xi + \frac{\mu}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{\xi^2}{2\sigma^2}}\,d\xi.\end{aligned}$$

The second term/integral is just $\mu G(\mathbb{R}) = \mu$ since $G : \mathcal{M} \to [0,1]$ by

$$G(A) = \frac{1}{\sigma\sqrt{2\pi}} \int_{\omega \in A} e^{-\frac{(\omega-\mu)^2}{2\sigma^2}}$$

is a probability measure.

The first term/integral is zero on the one hand because the integrand is odd. On the other hand, we can write

$$\int_{-\infty}^{\infty} \xi e^{-\frac{\xi^2}{2\sigma^2}}\, d\xi = \int_{-\infty}^{0} \xi e^{-\frac{\xi^2}{2\sigma^2}}\, d\xi + \int_{0}^{\infty} \xi e^{-\frac{\xi^2}{2\sigma^2}}\, d\xi.$$

Using the change of variables $u = \xi^2/(2\sigma^2)$ in the first integral where $\xi = -\sigma\sqrt{2u}$, we see

$$\begin{aligned} \int_{-\infty}^{0} \xi e^{-\frac{\xi^2}{2\sigma^2}}\, d\xi &= \int_{\infty}^{0} (-\sigma\sqrt{2u}) e^{-u} \left(\frac{-\sigma\sqrt{2}}{2\sqrt{u}} \right) du \\ &= -\sigma^2 \int_{0}^{\infty} e^{-u}\, du \\ &= -\sigma^2. \end{aligned}$$

Using the change of variables $u = \xi^2/(2\sigma^2)$ in the second integral where $\xi = \sigma\sqrt{2u}$, we see

$$\begin{aligned} \int_{0}^{\infty} \xi e^{-\frac{\xi^2}{2\sigma^2}}\, d\xi &= \int_{0}^{\infty} (\sigma\sqrt{2u}) e^{-u} \left(\frac{-\sigma\sqrt{2}}{2\sqrt{u}} \right) du \\ &= \sigma^2 \int_{0}^{\infty} e^{-u}\, du \\ &= \sigma^2. \end{aligned}$$

This gives that the sum is zero (as we knew). And a third way is to observe directly that

$$\frac{d}{d\xi} \left[-\sigma^2\, e^{-\frac{\xi^2}{2\sigma^2}} \right] = \xi\, e^{-\frac{\xi^2}{2\sigma^2}} \tag{4.16}$$

so by the fundamental theorem of calculus

$$\int_{-\infty}^{\infty} \xi e^{-\frac{\xi^2}{2\sigma^2}}\, d\xi = -\sigma^2\, e^{-\frac{\xi^2}{2\sigma^2}} \Big|_{\xi=-\infty}^{\infty} = 0. \tag{4.17}$$

We conclude the value of μ defined by

$$\mu = \int_{\omega \in \mathbb{R}} \omega \, \delta(\omega)$$

agrees with the parameter μ appearing in the definition of the MDF/PDF

$$\delta(\omega) = \frac{1}{\sigma\sqrt{2\pi}} \, e^{-\frac{(\omega-\mu)^2}{2\sigma^2}}.$$

It is interesting to note that the parameter σ does not effect the value of the mean in any way.

First mathematical principle of distinct probability measures: If two measures α and β on $\mathbb{R}$ have different means, then α and β are not the same measure.

4.5.2 Variance and spread

Say we have a uniform measure α on the $2k+1$ points $0, \pm\omega, \pm 2\omega, \ldots, \pm k\omega$ for some $\omega \in (0, \infty)$. The resulting possibilities (for the measures α) cannot be distinguished by mean. However, the quantity

$$\sigma^2 = \int_{\mathbb{R}} \mathrm{id}\,\mathrm{id} = \sum_{j=1}^{k} (-j\omega)(-j\omega)\alpha(\{-j\}) + \sum_{j=1}^{k} (j\omega)(j\omega)\alpha(\{j\}) = \frac{2\omega^2}{2k+1} \sum_{j=1}^{k} j^2$$

is monotonically dependent on the value of ω. Consequently, each of these measures has a distinct **variance** determined by how far ω is separated from $\xi = 0$. Equivalently, we can phrase the distinguishing characteristic in terms of the distance between any one of the pairs of points $-j\omega$ and $j\omega$ or in terms of the overall diameter (of the range) $2k\omega$. In fact,

$$\sigma^2 = \frac{k(k+1)}{3} \, \omega^2 = \frac{k+1}{12k} \, (2k)^2.$$

We generalize these statistical values as follows:

Definition 15 Given a probability measure α on $\mathbb{R}$ (either an adolescent or integral measure) with mean μ, we consider the translated measure β on $\mathbb{R}$ given by

$$\beta(A) = \alpha(\{\omega + \mu : \omega \in A\}).$$

The **variance** of α is given by

$$\sigma^2 = \int_{\mathbb{R}} (\mathrm{id})^2$$

where the integral is with respect to the measure β.

The **spread** of the measure α is given by

$$J = 2\sqrt{3}\ \sigma.$$

4.6 Translation and mean

Translation gives a particularly simple globally bijective function for renaming/inducing measures. One is often especially interested in translating to a measure with a **central mean** located at $\xi = 0$.

Consider first some measures on measure spaces with finitely many elements.

4.6.1 Bernoulli measure

Recall that β has nonzero values only on the singletons $\{0\}$ and $\{1\}$ with $\beta(\{1\}) = p$. The mean is

$$\mu = p.$$

Thus, the normalized measure π_x induced by the translation $x(\omega) = \omega - p$ has PMF

$$M(\xi) = \begin{cases} p, & \xi = 1 - p \\ 1 - p, & \xi = -p \\ 0, & \xi \neq -p, 1 - p. \end{cases}$$

The variance of the Bernoulli measure is the integral of ξ^2 with respect to the induced measure:

$$\begin{aligned} \sigma^2 &= (1-p)^2 \beta(\{1-p\}) + (-p)^2 \beta(\{-p\}) \\ &= (1-p)^2 p + p^2 (1-p) \\ &= p(1-p). \end{aligned} \tag{4.18}$$

Exercise 4.6.1 The quantity given in (4.18) may be considered the standard variance of the Bernoulli distribution. It is the value given in tables. For $b > 0$ consider the function $y : \mathbb{R} \to \mathbb{R}$ given by

$$y(\xi) = b\xi.$$

Find the variance of the measure induced by y from the normalized base measure π_x.

4.6.2 binomial measure

Taking $\mu = np$,

$$\begin{aligned}\sigma^2 &= \sum_{k=0}^{n}(k-np)^2 \left(\begin{array}{c} n \\ k \end{array} \right) p^k(1-p)^{n-k} \\ &= \sum_{k=0}^{n} k^2 \left(\begin{array}{c} n \\ k \end{array} \right) p^k(1-p)^{n-k} \\ &\qquad - 2np\sum_{k=0}^{n} k \left(\begin{array}{c} n \\ k \end{array} \right) p^k(1-p)^{n-k} \\ &\qquad\qquad + (np)^2 \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) p^k(1-p)^{n-k}.\end{aligned}$$

Consider first

$$\sum_{k=0}^{n} k^2 \left(\begin{array}{c} n \\ k \end{array} \right) p^k(1-p)^{n-k}.$$

Note that

$$\begin{aligned}\sum_{k=0}^{n} k^2 \left(\begin{array}{c} n \\ k \end{array} \right) p^k(1-p)^{n-k} &= \sum_{k=1}^{n} k \frac{n!}{(n-k)!(k-1)!} p^k(1-p)^{n-k} \\ &= n\sum_{k=1}^{n} k \frac{(n-1)!}{[n-1-(k-1)]!(k-1)!} p^k(1-p)^{n-k} \\ &= np\sum_{k=1}^{n} k \left(\begin{array}{c} n-1 \\ k-1 \end{array} \right) p^{k-1}(1-p)^{n-1-(k-1)} \\ &= np\sum_{k=0}^{n-1} (k+1) \left(\begin{array}{c} n-1 \\ k \end{array} \right) p^k(1-p)^{n-1-k} \\ &= np\int_{\mathbb{R}} \mathrm{id} + np\int_{\mathbb{R}} 1\end{aligned}$$

where for $n > 2$ the integrals are with respect to β_{n-1} the induced binomial measure associated with $n-1$ trials. For $n = 1$, we note the original sum

satisfies

$$\sigma^2 = \sum_{k=0}^{n}(k-np)^2 \left(\begin{array}{c} n \\ k \end{array} \right) p^k(1-p)^{n-k} = p^2(1-p) + (1-p)^2 p = np(1-p).$$

Recalling that the expectation of β_{n-1} (for $n \geq 2$) is

$$\int_{\mathbb{R}} \mathrm{id} = (n-1)p \qquad \text{and} \qquad \int_{\mathbb{R}} 1 = \beta_{n-1}(\mathbb{R}) = 1,$$

we have

$$\sum_{k=0}^{n}(k-np)^2 \left(\begin{array}{c} n \\ k \end{array} \right) p^k(1-p)^{n-k} = np^2(n-1) + np.$$

The value of

$$\sum_{k=0}^{n} k \left(\begin{array}{c} n \\ k \end{array} \right) p^k(1-p)^{n-k}$$

is precisely the mean of β_n or np. And finally,

$$\sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) p^k(1-p)^{n-k} = \beta_n(\mathbb{R}) = 1.$$

Therefore, when $n \geq 2$

$$\sigma^2 = \sum_{k=0}^{n}(k-np)^2 \left(\begin{array}{c} n \\ k \end{array} \right) p^k(1-p)^{n-k} = np^2(n-1)+np-2n^2p^2+n^2p^2 = np(1-p)$$

as well.

4.6.3 geometric probability measure

4.6.4 Poisson measure

4.6.5 Exponential measure

4.6.6 Normal/Gaussian measure

The variance of the measure G with density

$$\delta(\omega) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\omega^2}{2\sigma^2}}$$

as might be expected from the notation is σ^2. This can be seen using integration by parts. Setting first

$$u = \omega \qquad \text{and} \qquad v = -\sigma^2 \, e^{-\frac{\omega^2}{2\sigma^2}}$$

and recalling (4.16) we have

$$\begin{aligned}
\sigma^2 &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \omega^2 e^{-\frac{\omega^2}{2\sigma^2}} \, d\omega \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} u \frac{dv}{d\omega} \, d\omega \\
&= \frac{1}{\sigma\sqrt{2\pi}} \left[uv|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} v \, d\omega \right] \\
&= \frac{1}{\sigma\sqrt{2\pi}} \left[-\sigma^2 \int_{-\infty}^{\infty} \omega e^{-\frac{\omega^2}{2\sigma^2}} \, d\omega + \sigma^2 \int_{-\infty}^{\infty} e^{-\frac{\omega^2}{2\sigma^2}} \, d\omega \right] \\
&= \sigma^2 G(\mathbb{R}) \\
&= \sigma^2.
\end{aligned}$$

We have used here the observation (4.17) that

$$\int_{-\infty}^{\infty} \omega e^{-\frac{\omega^2}{2\sigma^2}} \, d\omega = 0,$$

and we have also used that G is a probability measure.

4.7 Cumulative mass function

Recall that the CMF (cumulative mass function) of an induced (generalized) baby measure is given by

$$F(\xi) = \sum_{t \leq \xi} \alpha(\{t\}) = \sum_{t \leq \xi} M(t)$$

where $\alpha = \alpha_x$ is the meausure induced by a real valued function $x : S \to \mathbb{R}$ and M is the associated PMF (probability mass function).

For integral measures or integral distrubutions we have

$$F(\omega) = \int_{(-\infty, \omega]} \delta$$

where the integration is with respect to Lebesgue measure. When the MDF δ is Riemann integrable this may be written in the more familiar form

$$F(\omega) = \int_{-\infty}^{\omega} \delta(t)\, dt.$$

The CMF is generally not so important for adolescent measures due to the extension formula involving the measure of singletons; the measures of singletons are obtained directly: $\alpha(\{t\}) = M(t)$. For integral measures on the other hand, there is no analogous extension formula, and the measure of general **intervals** $J \subset \mathbb{R}$ is usually of primary interest. Thus, the CMF is used much more often because if J has endpoints $a, b \in \mathbb{R}$ with $a < b$, then

$$\alpha(J) = \alpha((-\infty, b]) - \alpha((-\infty, b]) = F(b) - F(a).$$

4.8 Normalization of measures

With all probability measures on $\mathbb{R}$, but especially with integral measures, there is often an interest in **symmetry** with respect to the **mean**. We have seen above how to translate a measure to be **balanced** and how to compute the **variance**. Here we consider families of scalings of a balanced measure and the dependence of the variance on the scaling.

4.8.1 The meaning of variance

We have taken as our basis for the definition of **spread** the uniform integral measure. Recall this measure, if it is balanced, has range an interval $I = [-a, a]$ for some $a > 0$ and MDF/PDF given by $\delta(\omega) = \chi_I(\omega)/(2a)$. Now we induce a measure on $\mathbb{R}$ by the scaling $x : \mathbb{R} \to \mathbb{R}$ with $x(\omega) = b\omega$ where $b > 0$ is fixed. All of the resulting uniform probability measures, each with range $[-ab, ab]$ and density $1/(2ab)$, is balanced, i.e., has mean $\mu = 0$.

Each of these measures, however, can be distinguished from the others by variance. Let $\beta = \beta_b$ denote the measure induced by x. Then the variance of β is

$$\sigma_b^2 = \int_{\mathbb{R}} \xi^2 = \int_{-ab}^{ab} \xi^2 \, \frac{1}{2ab}\, d\xi = \frac{a^2 b^2}{3}.$$

Thus, the **spread** of β_b as we have defined it is precisely

$$2\sqrt{3}\sigma_b = 2ab.$$

The square root of the variance σ is called the **standard deviation**. Notice the value

$$\beta_b([-\sigma_b, \sigma_b]) = \beta_b\left(\left[-\frac{ab}{\sqrt{3}}, \frac{ab}{\sqrt{3}}\right]\right) = \frac{1}{\sqrt{3}} \doteq 0.57735.$$

The main conclusion here is that one can specify a unique measure from among the scaled measures β_b by specifying the variance, the standard deviation, or the spread. Notice that the parameter a plays no real role in this determination. Thus, we can say **there exists a unique uniform integral measure with a specified variance**. Figure 4.6 shows the MDF/PDF of a uniform integral measure with range $[-1, 1]$.

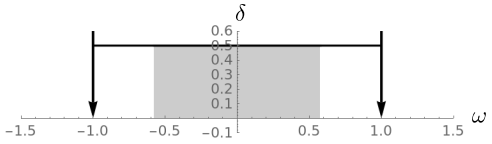

Figure 4.6: The MDF/PDF of a balanced uniform integral measure with range $[-1, 1]$. The spread $J = 2$ is the distance between the arrows. The shaded region corresponds to the measure of the set of points within one standard deviation of the mean.

As we have already seen, the variance of the Gaussian measure is the parameter σ^2, or equivalently the parameter σ is the standard deviation. In this case,

$$G([-\sigma, \sigma]) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\sigma}^{\sigma} e^{-\frac{\omega^2}{2\sigma^2}}\, d\omega \doteq 0.682689.$$

The family of (balanced) Gaussian measures are themselves obtained from any one of them by the same scaling $x(\omega) = b\omega$ with

$$\beta(A) = G(x^{-1}(A)) = \frac{1}{\sigma\sqrt{2\pi}} \int_{\omega \in x^{-1}(A)} e^{-\frac{\omega^2}{2\sigma^2}} = \frac{1}{b\sigma\sqrt{2\pi}} \int_A e^{-\frac{\xi^2}{2b^2\sigma^2}}.$$

Thus, when a specific Gaussian measure is determined by specifying the parameter σ^2 one is also specifying a specific scaling of the domain for the MDF/PDF. This specification is often made with respect to a particular choice of parameter σ, namely $\sigma = 1$ (and $\mu = 0$) which results in what is

known as the **standard normal distribution** with MDF/PDF

$$\phi(\omega) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{\omega^2}{2}}.$$

The spread of the standard normal distribution is

$$J = 2\sqrt{3} \doteq 3.4641.$$

Figure 4.7 shows the MDF/PDF of a standard normal/Gaussian measure.

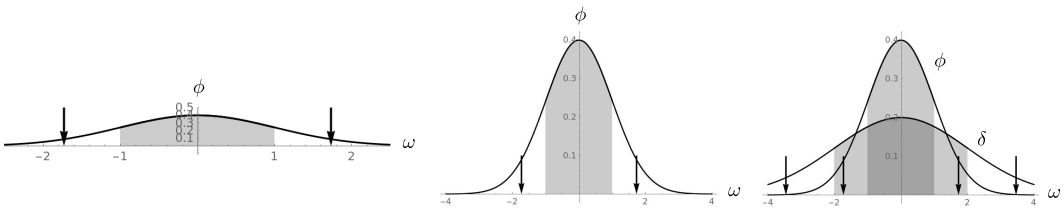

Figure 4.7: The MDF/PDF of a standard normal/Gaussian measure (left and center). The spread $J = 2\sqrt{3}$ is the distance between the arrows. The shaded region corresponds to the measure of the set of points within one standard deviation of the mean. On the right I have superimposed the Gaussian measure obtained by a scaling with $b = 2$.

4.8.2 Standard deviation

Within a family of scaled (balanced) probability measures as considered above, the measure of the set of points within one standard deviation of the mean will be constant (independent of the scaling). Thus, if this measure is taken as the probability of a certain outcome, then the probability of falling within one standard deviation from the mean will be the same for all these probability measures.

On the other hand, if one takes the measure as indicative of the number of outcomes (for example measurement values of individuals) in an overall population falling within one standard deviation of the mean, then this measure may be interpreted as a **percentage of the population**. For example, if height measurements in a population are consistient with a normal distribution, then about 68% of individual heights in that population should

fall within one standard deviation of the mean. Notice the percentage is independent of the variance/standard deviation, but the interval of interest determining the number of individuals itself has length twice the standard deviation.

4.9 Simulation with the exponential distribution

4.10 Simulation with the normal distribution

4.11 The first principle of statistics

The first principle of statistics is this:

> Data is imagined to be the result of "random sampling from some probabilistic distribution."

If one wishes to imagine a particular set of data is the result of random sampling from a probabilistic distribution, then the mean and the variance of the distribution should be reflected in the data. The mean of real data, i.e., a finite set of real numbers, is relatively easy to calculate, and the application is straightforward:

> Given enough data if the mean of the data is significantly different from the mean of a particular probabilistic distribution, then it is difficult to imagine the data is the result of random sampling from that distribution.

Note the superficial resemblance to the first mathematical principle of distinct measures.

Definition 16 A human being who must play psychological games with himself in order to hurt another human being is called **humane**. Note the linguistic root "human" in the word "humane;" to be humane is considered by many human beings to be "normal" for human beings.[6]

A human being who does not need to play psychological games with himself in order to hurt another human being is called a **psychopath**.

The statistical situation with variance is more complicated.

[6] . . . at least superficially as partially indicated by linguistic usage at the current time

Chapter 5

Lecture 5: Restriction Measures and Bayes' Theorem

We have seen restriction measures before. For example, given a six sided die and the associated uniform measure on the set $S = \{1, 2, 3, 4, 5, 6\}$ modeling the outcome of a roll, we may consider the set $T = \{2, 4, 6\}$ and ask the question:

> Given the roll is even, what is the probability the outcome is greater than or equal to 4?

This is equivalent to asking what is the measure $\rho_T(\{4, 5, 6\})$ where ρ_T is the restriction probability measure determined by the set T. Recall the definition giving the values of this measure is

$$\rho_T(A) = \frac{\pi(A \cap T)}{\pi(T)}$$

where π is some base measure. In this case π is the uniform measure mentioned above with $\pi(\{j\}) = 1/6$. We have now covered some additional material on product measures, and we will cover several problems involving both restriction measures and product measures. In particular, I would like to start by revisiting the problem posed by Orloff and Booth which I took as the subject of Lecture "-2" in these notes, namely, the problem these authors chose to exemplify both Bayes' theorem and what they called the **base rate fallacy**.

You may recall this example/problem involved a test for sickness. The false positive rate was said to be 0.05, the false negative rate 0.1, and the

"base rate" for illness 0.005 or half a percent. I gave a solution of this problem by applying these rates precisely to a hypothetical population each individual of which was tested and in which the health of every individual was known and assumed to follow the "distribution" described in the problem exactly. I am now going to solve the problem in a drastically different way. I consider only the single individual in question. In the statement of the problem, this was "you." So be it: I consider myself. I introduce a probability space on two points, essentially a Bernoulli measure. One element of this space, say "s" models the outcome that I am sick. The other "w" models the outcome that I am not sick, that is to say, well. Thus, $H = \{\mathrm{s}, \mathrm{w}\}$, and I introduce a probability measure π_1, or more properly, such a measure is nominally already introduced in the problem, for I am supposed to believe the "probability" I am sick is the base rate. That is,

$$\pi_1(\{\mathrm{s}\}) = 0.005 = \frac{1}{200} \qquad \text{and} \qquad \pi_1(\{\mathrm{w}\}) = 1 - 0.005 = 0.995 = \frac{199}{200}.$$

So this probability measure is given, or at least these probabilities are given, and I have interpreted them as the values of a measure.

I introduce a second probability space $T = \{\mathrm{p}, \mathrm{n}\}$ to model the possible outcomes that I am tested and the test result is positive "p" or negative "n." The values of the Bernoulli measure on this set, **I do not know** at least initially. Nevertheless, I assume there is such a measure π_2.

The produce space is the Cartesian product $S = H \times T$ and is illustrated in Figure 5.1. Having constructed this base product space S on which I also

n	(s, n)	(w, n)
p	(s, p)	(w, p)
	s	w

Figure 5.1: The product space for the possible outcomes of my health (horizontal modeled by the set H) and my testing results (vertical modeled by the set T).

assume there is a base measure π, I turn to the meaning and values of the "false positive" and "false negative" probabilities. The interpretation is that the values of the following measures of sets are given:

$$\rho_W(P) = 0.05 = \frac{1}{20} \qquad \text{where } W = \{(\mathrm{w}, \mathrm{p}), (\mathrm{w}, \mathrm{p})\} \text{ and } P = \{(\mathrm{s}, \mathrm{p}), (\mathrm{w}, \mathrm{p})\}.$$

$$\rho_S(N) = 0.1 = \frac{1}{10} \qquad \text{where } S = \{(\mathrm{s},\mathrm{p}),(\mathrm{s},\mathrm{n})\} \text{ and } N = \{(\mathrm{s},\mathrm{n}),(\mathrm{w},\mathrm{n})\}.$$

These are the values of probability restriction measures both determined by the base measure π which, at the moment, is unknown. Notice we have introduced four doubleton sets corresponding to the outcomes that "I am well" (W), "sick" (S), that "I test positive" (P), and that "I test negative" (N). Of course, these outcomes are modeled separately, and individually in the respective factor spaces. They are modeled as compound outcomes by the doubletons.

Recall that we have introduced various measures on Cartesian products with two factors. The simplest kind of "product measure" is one for which the measure of a singleton pair is the product of specified factor measures of the factor singletons:

$$\pi(\{(\omega_1,\omega_2)\}) = \pi_1(\{\omega_1\})\ \pi_2(\{\omega_2\}).$$

This will always produce a "product measure" and is what we actually defined as a product measure. We also had, however, genealized product measures, in which the measure used for the product was not always the same factor measure.

Exercise 5.0.1 Consider the product measure for an urn problem containing one red ball and one green ball without replacement. For this we model the outcome of the first draw using a base probability measure π_1 on, for example, $\{\mathrm{r},\mathrm{g}\}$. Assigning probabilities to the possible outcomes of the second draws, which we can model using the same set, requires the use of two measures π_{21} and π_{22}. Specifically, we have

$$\pi_1(\{\mathrm{r}\}) = \pi_1(\{\mathrm{g}\}) = \frac{1}{2},$$

$$\pi_{21}(\{\mathrm{r}\}) = \pi_{22}(\{\mathrm{g}\}) = 0, \qquad \text{and} \qquad \pi_{21}(\{\mathrm{g}\}) = \pi_{22}(\{\mathrm{r}\}) = 1,$$

so that

$$\pi(\{(\mathrm{r},\mathrm{r})\} = \pi_1(\{\mathrm{r}\})\,\pi_{21}(\{\mathrm{r}\}) = 0, \quad \pi(\{(\mathrm{r},\mathrm{g})\} = \frac{1}{2}, \quad \pi(\{(\mathrm{g},\mathrm{r})\} = \frac{1}{2}, \quad \text{and} \quad \pi(\{(\mathrm{g},\mathrm{g})\} = 0.$$

Note carefully, the modeling in the vertical factor space: The sums of the values across the rows do not give the value of either measure π_{21} nor π_{22} on the corresponding second factor singleton.

Assume $S = S_1 \times S_2$ where $S_1 = \{\omega_1,\omega_2,\ldots,\omega_k\}$ with $\#S_1 = k$ and $S_2 = \{\tau_1,\tau_2,\ldots,\tau_\ell\}$ wtih $\#S_2 = \ell$. Assume π is a measure on S.

(a) Show the sums of the measures

$$v_i = \sum_{j=1}^{\ell} \pi(\{(\omega_i, \tau_j)\}$$

of the values of π along each column for $i = 1, 2, \ldots, k$ satisfy

$$\sum_{i=1}^{k} v_i = 1$$

Thus, $\mu_1(\{\omega_i\}) = v_i$ for $i = 1, 2, \ldots, k$ defines a probability measure on S_1. The values $v_1, v_2, \ldots, v_k$ are called the **marginals** or the first factor marginals of the measure π.

(b) Show the sums of the measures

$$w_j = \sum_{i=1}^{k} \pi(\{(\omega_i, \tau_j)\}$$

of the values of π along each row for $j = 1, 2, \ldots, \ell$ satisfy

$$\sum_{j=1}^{k} w_j = 1$$

Thus, $\mu_2(\{\tau_j\}) = w_j$ for $j = 1, 2, \ldots, \ell$ defines a probability measure on S_2. The values $w_1, w_2, \ldots, w_\ell$ are also called **marginals** or the second factor marginals of the measure π.

(c)

We will assume here that the marginals satisfy the following:

$$\pi(S) = \pi_1(\{\mathrm{s}\}) \qquad \text{and} \qquad \pi(W) = \pi_1(\{\mathrm{w}\}).$$

By the definition of probability restriction measure we know

$$\rho_W(P) = \frac{\pi(P \cap W)}{\pi(W)} \qquad \text{and} \qquad \rho_S(N) = \frac{\pi(N \cap S)}{\pi(S)}.$$

Notice these values determine $\pi(P \cap W)$ and $\pi(N \cap S)$.

We are asked to find the value

$$\rho_P(T) = \frac{\pi(T \cap P)}{\pi(P)}.$$

Chapter 6

Lecture 6: Summary of Probability

Lectures 1-4 might be characterized as building new measure spaces from old. Generally, we started with simpler or smaller measure spaces and measures and built larger and more complicated measures and measure spaces. This is an interesting topic and approach, and I think it has been fun and worked well for many of you. The transition from certain baby measures to certain adolescent measures, in particular from the induced binomial measure to the Poisson measure, and from adolescent measures to integral measures was, I think, quite interesting and inspiring.

It may be remarked that this approach is not entirely standard for an introduction to probability (especially in the context of introductory statistics). The standard approach is to simply give formulas for all these measures we have thought about rather more carefully and, in many case, attempted to "build" from simpler measure spaces by various measure theoretic constructions.

In the end, I hope you have accumulated a respectable collection of probability measures with which you are reasonably familiar. Among these, you should understand uniform measures both on a set of finitely many points and for an integral measure on an interval, Bernoulli, binomial, geometric, Poisson, exponential, and Gaussian measures. This is not to mention learning the meaning of various words and distinctions, like what it means to be a baby or adolescent (discrete) or integral (continuous) measure, PMF, CMF/CDF, MDF/PDF, mean, variance, and others.

Lecture 5 marks a departure from building measure spaces with greater

size and complication, though the topic of restriction, as we have seen, has its own complications. In terms of applications the consideration of restriction measures, or equivalently and more commonly conditional probabilities, is the main and most powerful technique usually considered in an introductory course.

Chapter 7

Data

7.1 Simulation of data

7.2 Analysis of data

This is called **descriptive statistics**.

Chapter 8

Statistics

Hopefully it has been made clear above that **one view** of probability is the following:

> The only meaning that can be attached to a probability (or the concept of probability) is the value of a measure.

An alternative is the following:

> Some other meaning can be attached to a probability involving "chance" or "random" occurrence.

The alternative must be embraced for the subject of statistics (beyond descriptive statistics) to make any sense, because the basic objective of the subject of statistics[1] is to attach a probabilistic explanation to collections of real world data. It should be emphasized from the outset, as it is in many elementary texts and also by Orloff and Booth, that there is no well-defined procedure for "attaching" such meaning. There are many attempts and approaches to reaching the main objective. These very often depend in extremely important ways on an underlying knowledge of how the data is collected, or at least some intuition concerning the orign of the data and how it was collected. As Orloff and Booth and many other authors put it: Statistics is more an art than a science. With a certain expectation of imprecision then, we can roughly break the subject of statistics into various statistical activities and offer the following which can be taken to embody **one view** of statistics.[2] These are often cast as the "jobs of a statistician."

[1] . . . beyond providing a framework for humans to think in a pathological manner

[2] Naturally a much greater diversity of alternative views may be expected in regard to statistics, and I make no attempt to represent other possibilities.

1. Obtain or collect **data**. (This may be broadly termed "simulation.")

   One "job" of a statistician is to explain how data should be collected. (This is sometimes called "experiment design.")

2. Organize and/or interpret data. (For example, sort data, make histograms, and record frequencies.)

   Determine **statistical quantities**, i.e., statistics. (This is descriptive statistics.)

3. Believe—that not only "chance" and "random" occurrence give meaning to probability, but that all data arises as the result of some "random" sampling or occurrence.

   (i) Identify or propose a hypothetical/theoretical probability distribution as the source of the data.

   (ii) Check for consistency. (For example, does the theoretical statistical mean match the sample mean? Does the theoretical statistical variance match the sample variance?)

4. Determine a second layer of probability/probabilistic explanation.

   This can take various forms but roughly speaking falls into two (sometimes overlapping) categories: internal (or direct comparison of the data to the proposed probabilistic explanation) and external (or extrapolation to larger population(s) assumed to be "represented" by the data). Among the tools and "jobs" associated with this second layer of probabilistic explanation are the following:

   (i) maximum liklihood determination

   (ii) determination of **confidence intervals** (essentially internal but often with application to extrapolation)

   (iii) determination of p**-values** (This is especially associated with "hypothesis testing." A p-value is considered a measure of "statistical significance.")

   (iv) weighting (essentially external and depending on data outside[3] the sample being analyzed/explained)

[3]For example, say one has online polling data but is convinced, by other data, that proportionally few members of a certain demographic in the larger population participated

8.1 One Main Point of Statistics (Part A)

If one has a numerical data set

$$\mathtt{S} = \{\mathtt{a}_1, \mathtt{a}_2, \ldots, \mathtt{a}_N\}$$

of "discrete type" taking values in the sample range

$$\{\omega_1, \omega_2, \ldots, \omega_n\} \subset \mathbb{R},$$

then each element ω_j for $j = 1, 2, \ldots, n$ is considered a "bin" and one may sort through the data and distribute[4] it among the bins.

Recall that the objective is to associate a PMF with the data.

Under normal circumstances one should like to have $N >> n$. Let us consider, however, starting with the first datum $\mathtt{a}_1$. If $\mathtt{a}_1 = \omega_k$, we can consider a PMF with range $\{\omega_k\}$ given simply by

$$F_1(\omega) = \begin{cases} 1, & \omega = \omega_k \\ 0, & \omega \neq \omega_k. \end{cases}$$

Note that the PMF F_1 is a probability mass function (PMF) and does not involve the data explicitly, though it is certainly determined completely by the first datum $\mathtt{a}_1$. There are two possibilities when we come to sort the second datum $\mathtt{a}_2$. Either $\mathtt{a}_2 = \omega_k$ or $\mathtt{a}_2 = \omega_\ell$ for some $\ell \neq k$. Simply keeping track of the frequencies (so far in the sorting process) we will be confronted with one of two functions

$$f_1(\omega) = \begin{cases} 2, & \omega = \omega_k \\ 0, & \omega \neq \omega_k. \end{cases} \qquad \text{or} \qquad f_1(\omega) = \begin{cases} 1, & \omega = \omega_k \\ 1, & \omega = \omega_\ell \\ 0, & \omega \notin \{\omega_k, \omega_\ell\}. \end{cases}$$

in the online poll. For example, Say it is known that 60% of people over sixty years of age do not participate in online polls but comprise 50% of some larger population to which one wishes to extrapolate sample polling data. If 5% of those who did participate in the online poll and provided responses including identifying themselves as over sixty years of age, then a better representation of the "responses" of those in the larger population can be imagined to be obtained by puting some additional weight on the responses of the particular "over sixty" participants. Similar "weighting" may be used if certain demographics are suspected of lying in response to a poll, being "insincere," or simply "unimportant."

[4]It is of note that this use of the word **distribute** is the origin of the term **distribution** in "probability distribution" and "cumulative distribution function."

Neither of these functions is a probability mass function, but dividing the values by the number of samples (in this case two samples) we obtain two PMFs:

$$F_1(\omega) = \begin{cases} 1, & \omega = \omega_k \\ 0, & \omega \neq \omega_k. \end{cases} \qquad \text{and/or} \qquad F_1(\omega) = \begin{cases} 1/2, & \omega = \omega_k \\ 1/2, & \omega = \omega_\ell \\ 0, & \omega \notin \{\omega_k, \omega_\ell\}. \end{cases}$$

Continuing this process we obtain the **sample PMF** $F : \{\omega_1, \ldots, \omega_n\} \to \mathbb{R}$ by

$$F(\omega) = \#\big\{j \in \{1, 2, \ldots, N\} : \mathtt{a}_j = \omega\big\}/N.$$

The **one main point of statistics** referenced in the section title above is, roughly speaking, that as N tends to ∞, the sample PMF tends to the desired PMF.

Of course, there are various qualifications concerning this statement, but like the frequency stabilization limit in probability, if you believe the sampling is indeed "random sampling" according to some PMF associated with a probability measure, then the assertion is more or less straightforward to "understand" and easy to believe. In fact, it is essentially the same statement, though in this context (of statistics) the imagined convergence is more closely associated with the **law of large numbers** which usually refers only to the convergence of the sample mean(s) or more properly the relation of the sample mean to a hypothetical/theoretical mean.

A technical difficulty is that usually N is simply a fixed integer and is not "tending to infinity" in any sense. One may consider (or imagine) a sequence of "experiments" or samplings however with sizes $N_1, N_2, N_3, \ldots$ tending to ∞. Changing notation from that above we may denote by F_j the PMF associated with N_j for $j = 1, 2, 3, \ldots$. Then a mathematical convergence

$$\lim_{j \nearrow \infty} F_j = F$$

may be considered. Of course, no one has ever seen infinitely many samplings of ever increasing size, so such a consideration is practically vaccuous. There are instances, however, in which finitely many different, though not necessarily entirely distinct, sample data sets

$$\mathtt{S}_1, \mathtt{S}_2, \ldots, \mathtt{S}_m$$

may be considered. Sometimes such data sets may be identified by size and ordered to correspond to a finite sequence

$$N_1 < N_2 < \cdots < N_m$$

In such cases, there can be interesting questions about the overlap or "correlation" among the data sets $\mathtt{S}_i$ and $\mathtt{S}_j$ for $i \neq j$. One can ask, for example, should the union

$$\mathtt{S} = \bigcup_{j=1}^{m} \mathtt{S}_j$$

be considered a single data set associated with the (assumed underlying) distribution?

8.2 One Main Point of Statistics (Part B)

Let us imagine for a moment that a data set is apparently of a continuous type, but sorting suggests some relation to discrete type data. For example, imagine there exist sample range values

$$\{\omega_1, \omega_2, \ldots, \omega_n\} \subset \mathbb{R}$$

as in the previous section but in this case one has in fact

$$\{\omega_1, \omega_2, \ldots, \omega_n\} \subset \mathbb{Z} = \{0, \pm 1, \pm 2, \pm 3, \ldots\}$$

and all sample values $\mathtt{a}_j$ for $j = 1, 2, 3, \ldots, N$ fall into the union of intervals

$$\bigcup_{k=1}^{n} (\omega_k - 0.1, \omega_k + 0.1).$$

It makes sense to treat the intervals $I_k = (\omega_k - 0.1, \omega_k + 0.1)$ for $k = 1, 2, \ldots, n$ as bins and sort the data to obtain a relative frequency function or sample PMF F as in the previous section. For example, maybe we have $\omega_k = 1$ and data points

$$\mathtt{S} \cap I_k = \{1.02, 1.027, 0.95, 1.05, 0.993, \ldots\}.$$

In this case, it may be desirable to propose sampling according to a discrete distribution (baby measure) as the underlying probabilistic explanation for

the data even though there is some apparrent pointwise variance around the integers at a finer scale.

When data has a more clearly "continuous" character one can attempt a similar sorting into "bins" but one must determine the bins in some non-obvious manner. Generally, equal length intervals

$$(a_0, a_2], (a_2, a_3], \ldots, (a_n, a_{n+1}]$$

determined by **class boundaries** $a_0, a_1, a_2, \ldots, a_{n+1}$ are used in this context. The law of large numbers in this case, roughly speaking, says that the relative frequency functions should converge to a function proportional to the underlying MDF. Of course there is a technical problem in that for a fixed number n of bins, one obtains (at best) a fundamentally discrete distribution as N tends to $+\infty$. However, one can go further and imagine a sequence $n_\ell \nearrow \infty$ of bin numbers for which (perhaps) $a_{n+1}(\ell) \nearrow \infty$, $a_0(\ell) \searrow -\infty$, and $a_{k+1}(\ell) - a_k(\ell) \searrow 0$ (all as $\ell \nearrow +\infty$) and for each a limiting discrete PMF F_ℓ (for ℓ fixed and $N \nearrow \infty$) represented as a step function over the bin intervals one has

$$\lim_{\ell \nearrow \infty} F_\ell = \alpha\delta$$

where δ is the underlying MDF and α is a constant arising as a result of the fact that F_ℓ arises as (or from) a PMF and cannot be expected to have integral 1. This basic assertion goes under the name GlivenkoCantelli theorem while the corresponding convergence for the mean is the continuous version of the law of large numbers. Some additional assumptions and technicalities are required. In particular, the convergence is GlivenkoCantelli theorem is (only) asserted with "probability 1," and the convergence of the means requires that the distribution δ admit a well-defined mean.

8.3 Interpretation of data

Here we consider more fully the relation of the statistical values of data defined in the previous chapter/lecture with the statistical values associated with various probability measures.

8.4 Inference

Bayesianism, definition 1: Two lies inspire more confidence than one.

Bayesianism, definition 2: Multiple quantification of ignorance represents credible knowledge.

8.4.1 Extrapolation

8.4.2 Hypothesis testing

Bibliography

[1] William Bolstad. *Introduction to Bayesian Statistics*. Wiley, Hoboken, NJ, second edition, 2007.

[2] Alexander Farmer. *The Imposition of Dystopia: Probability and Statistics*. Wonderground Press, Makoti, ND, 2020.

[3] R Hogg, E. Tanis, and D Zimmerman. *Probability and Statistical Inference*. Pearson, New York, ninth edition, 2015.

[4] J. Orloff and J. Bloom. Introduction to probability and statistics. In *Spring Semester—MIT Course No. 18.05J*. MIT, Cambridge MA, 2014. MIT OpenCourseWare.