Assignment 6: Bayes' theorem, simulation
Due November 7, 2023

John McCuan

Problem 1 (Bayes' theorem and total probability) Consider an urn containing two red balls and four green balls. If two balls are drawn in order without replacement, model the outcomes with the set $S^2 = \{1,2,3,4,5,6\}^2$ where $S = \{1,2,3,4,5,6\}$. Consider also the function $x : S \to \{r, g\}$ with $x(j) = r$ for $j = 1, 2$ and $x(j) = g$ for $j = 3, 4, 5, 6$.

(a) Introduce appropriate probability measures on S and construct from them a generalized product measure π on S^2 so that $\pi(\{(i,j)\})$ is the probability the i-th ball is drawn first and the j-th ball is drawn second.

(b) Use the function x to obtain subsets of S^2 modeling the following compound outcomes:

(i) The first ball drawn is red.

(ii) The first ball drawn is green.

(iii) The second ball drawn is red.

(c) Let R be the set you found in part **(b)(i)** above. Let G be the set you found in part **(b)(ii)** above. Let R_2 be the set you found in part **(b)(iii)** above. Verify the following law of total probability:

$$\pi(R_2) = \rho_R(R_2)\ \pi(R) + \rho_G(R_2)\ \pi(G)$$

(d) Calculate $\pi(R_2)$ directly from the values of π and part **(b)(iii)**.

(e) Calculate $\pi(R_2)$ using the law of total probability.

Problem 2 (Bayes' rule) The following forms of Bayes' rule appear in Matthew Buchan's project:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \qquad \text{and} \qquad P(H|E) = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|\overline{H})P(\overline{H})}.$$

Assume the symbols E and H represent sets modeling outcomes in some probability measure space S.

(a) Reexpress each of the probability rules above in terms of the measure

$$\pi : \mathfrak{M} \to [0,1] \tag{1}$$

and the associated probability restriction measures ρ_E and ρ_H.

(b) Use (only) the first form of Bayes' rule to justify the equality

$$P(E) = P(E|H)P(H) + P(E|\overline{H})P(\overline{H}). \tag{2}$$

(c) What would you (or Orloff and Booth) call the relation (2)?

(d) What is $\mathfrak{M}$ in (1) and why would it (possibly) be incorrect to replace (1) with

$$\pi : \wp(S) \to [0,1] \text{ ?}$$

In the following problems involving **simulation** using the statistics package `R`, I will use ">" to represent the commandline prompt.

Problem 3 (`rbinom`, `sample`, and frequency)

(a) What is the difference between

`sample(0:1,5,prob=c(0.4,0.6))` and `rbinom(5,1,0.6)` ?

(b) Find an implementation of `sample` equivalent to `rbinom(5,1,0.6)`.

(c) Evaluate the following commands:

```
> set.seed(150)
> x <- sample(0:1,10,replace=TRUE,prob=c(0.4,0.6))
> set.seed(150)
> y <- rbinom(10,1,0.6)
```

Problem 4 (frequency) Let `x` and `y` denote the vectors calculated in part **(c)** of Problem 3 above.

The **frequency of an entry** (or element) **in a vector** is the number of times that entry appears in the vector. For example, if the fourth entry in `x` is `0`, then the frequency of the fourth entry is the number of zeros in `x`.

(a) Does it make sense to talk about the **frequency of the number** `1` in the vector `x`?

(b) How would you define the frequency of `4` in the vector `z` if

```
> pv <- c(0.45,0.045,0.0045,0.00045,0.0001,0.00045,0.0045,0.045,0.45)

> z <- sample(0:8,5,prob=pv) ?
```

(c) What is the frequency of `9` in the vector `z` from part **(b)** above?

(d) Give an easy way to calculate the frequencies of the numbers `0` and `1` in the vector `y`.

Problem 5 (binomial experiments) Consider

```
> set.seed(15)
> tottrials <- rbinom(100,1,0.6)
> experiments <- matrix(tottrials,nrow=20,ncol=5)
```

(a) The code above may be interpreted to simulate twenty experiments in which each experiment is a simulation of 5 Bernoulli trials. What is the probability associated with each Bernoulli trial?

(b) What are the outcomes in the twelfth experiment?

(c) What R code/command could you use (or would you use) to display the outcome of the twelfth experiment?

(d) Exclude the first line of code above (specifying the seed), and execute the remaining three commands to obtain a different `experiments` vector; compose and execute a command that outputs a vector with the total number of successes in each experiment.

Problem 6 (binomial experiments; Problem 5 above) Consider appending the following R commands to your commands from part **(d)** of Problem 5 above:

```
> xvalues <- rowSums(experiments)
> cepmf=replicate(n=6,xvalues)
> ccepmf <- t(replicate(n=20,0:5))
> calcepmf <- cepmf-ccepmf == 0
> epmf <- colSums(calcepmf)/20
```

The first line/command above may be considered to give the value of the binomial inducing function $x : \{0,1\}^5 \to \mathbb{R}$ for each of the twenty experiments. The remaining lines/commands determine a vector `epmf` giving what is called the **emperical probability mass function** or **sample mass function** induced by x on the data `experiments`.

(a) Study and understand each of the lines/commands above. Can you obtain the same `epmf` in a simpler way?

(b) Execute the following lines/commands to plot the data `epmf`:

```
> omega=0:5

> plot(omega,epmf)
```

(c) Increase the number of experiments, and plot the resulting `epmf` each time. What do you observe?

Problem 7 (frequency and the binomial distribution)

(a) Express the vector you produced in part **(d)** of Problem 5 in terms of frequency and in terms of the quantities/variables introduced in Problem 6.

(b) What frequency is measured by `epmf` in Problem 6?

(c) Use `dbinom(0:5, size = 5, prob = 0.6)` to plot the PMF for the binomial induced measure.

(d) Compare the plots of `epmf` and the PMF of the binomial induced measure for various numbers of experiments. How many experiments do you need to produce (apparent) frequency stabilization?

The following (hint) code may be of interest for part **(d)** of Problem 7:

```
> plot(omega,epmf,pch=20)
> par(new=TRUE)
> plot(omega,dbinom(omega,5,0.6),pch=1).
```

Problem 8 (simulated means) Recall that we calculated the mean of the induced binomial distribution on n (Bernoulli) trials with probability p and found that mean to be $\mu = np$.

(a) Use `R` to calculate `mean(rbinom(100,10,0.3))`.

(b) Increase the number of trials/experiments by factors of 10 each time until you consider a million trials. Record the calculated means.

(c) Calculate `mean(rexp( n, 0.5))` for $n = 10, 100, \ldots, 1\,000\,000$ similarly, in order to simulate the mean of the exponential distribution. Recall that here $\lambda = 0.5$ is called the **rate constant**.

Problem 9 (simulation and variance) Recall that for an integral probability measure with MDF $\delta : \mathbb{R} \to [0, \infty)$, the **variance** of the measure is given by

$$\sigma^2 = \int_{\omega \in \mathbb{R}} (\omega - \mu)^2 \delta(\omega) \qquad \text{where} \qquad \mu = \int_{\omega \in \mathbb{R}} \omega \delta(\omega) \quad \text{is the mean.}$$

(a) Find the formula for the variance of the measure induced by a function $x : \mathbb{R} \to \mathbb{R}$.

(b) Show

$$\sigma^2 = -\mu^2 + \int_{\omega \in \mathbb{R}} \omega^2 \delta(\omega). \tag{3}$$

(c) Use `rnorm(` n`, mean = -2, sd = 3)` to find the variance of the normal distribution with mean -2 and standard deviation 3. Hint:

> mean((rnorm(n, mean = -2, sd = 3) -μ)^2)

(d) Obtain a different simulation to calculate the same number using the formula (3).

Problem 10 (conditional probabilities; Problems 5-7 above) Consider replacing the middle line of `R` code

```
tottrials <- rbinom(100,1,0.6)
```

in Problem 5 with

```
tottrials <- rbinom(100,1,rbinom(100,10,0.6)/10).
```

How does this effect the subsequent discussion of Problems 5-7? In particular,

(a) Can you still obtain `dbinomial(0:5, size = 5, prob = 0.6)` for large numbers of experiments as in parts **(c)** and **(d)** of Problem 7?

(b) Does (and if so "how does") changing the parameter `size`, i.e., the number of trials in each experiment effect the result(s)?

(c) Making `size` smaller in part **(b)** above might be particularly interesting. For example, if `size` is taken to be 2, so that for the twenty experiments considered in the code above related to Problem 5 the number 100 becomes 40, is it possible to make an explicit calculation (without simulation)?