Assignment 7: Statistics
Due November 14, 2023

John McCuan

Problems 1-5 below concern a “colorful” set of statistical data and are related to the first fundamental question of statistics: What can you do with data?

Measurements in inches for the length of Milkyway minis are found to be the following:

```
0.88 1.04 0.77 1.11 0.88 1.04 0.89 0.90 0.97 0.90
1.14 0.96 1.10 1.02 0.94 0.75 0.82 0.83 0.94 1.02
0.90 0.62 0.89 0.88 0.81 1.01 0.82 0.67 0.96 0.89
0.90 0.92 1.02 0.79 0.96 0.80 0.91 0.99 0.72 1.02
0.83 0.90 0.95 1.03 0.78 0.79 0.94 0.85 0.86 0.98
```

Problem 1 (some of the easiest statistical data values) Let’s make a **stem-and-leaf display** with three columns: one for stems, one for leaves, and one for frequencies. I’ll start with the first data point 0.88 and take the last digit as the “leaf.” The rest is the “stem.” I can’t count the frequencies until I’m done. So I start with the following:

stems	leaves	frequencies
0.8	8	

With each data point your display should grow. Order the stems vertically and the leaves horizontally as you add data to the display. Here is what I get after the first ten data points:

stems	leaves	frequencies
0.7	7	
0.8	8 8 9	
0.9	0 0 7	
1.0	4 4	
1.1	1	

You can see there should be ten leaves for the ten points. The frequencies associated with the first ten data points are $1, 3, 3, 2, 1$ which would appear in the frequencies column if there were only ten lengths.

(a) Complete the stem-and-leaf display for the lengths of Milkyway minis and fill in the frequencies.

(b) Reconstruct the data in an ordered list from your stem-and-leaf display. These numbers (in this order) are called the **order statistics** associated with the sample (data).

(c) If someone asks for the **order statistic of rank** n in a sample, he means the n-th entry in your list from part **(b)** above. What is the rank ten order statistic for the Milkyway mini data?

(d) One thing that is very easy to do using a stem-and-leaf display is to find the sample range of the data. The **sample range** of the data is the interval between the largest sample value and the smallest sample value. What is the sample range for the Milkyway mini data?

Problem 2 (order of observations) A different kind of stem-and-leaf display is possible if the stems are ordered according to value as before, but the leaves are ordered according to their appearance in the rows of the raw data set.

(a) Make a second stem-and-leaf display containing the Milkyway mini data but using the order of observations for the leaves as described above.

(b) Assuming the Milkyway minis were measured in the order of the raw data from top to bottom and left to right along rows, find information about the raw data that is lost in the stem-and-leaf display from Problem 1 but is preserved in the stem-and-leaf display from part **(a)** of this problem.

(c) It is typical that information about the raw data is lost when sample statistics, like the order statistics or histogram statistics, are extracted from a data set. Find something about the order of the Milkyway mini observations that is lost in the stem-and-leaf display from part **(a)** of this problem under the same assumption in part **(b)**.

Problem 3 (histogram statistics) The actual sample values are preserved in a stem-and-leaf display, though perhaps not the order of observation, and associating with each stem the corresponding frequency gives an example of **histogram statistics**.

(a) Arrange the stems from the stem-and-leaf display obtained in Problem 1 along a horizontal axis and plot the values of the associated frequencies as a function of the stem. The result is called a **frequency histogram**.

(b) Display the frequency histogram from part **(a)** by making a bar graph as follows: Let a_j for $j = 1, 2, \ldots, k$ denote the stems. Observe that $L = a_{j+1} - a_j$ is a constant for $j = 1, \ldots, k-1$. Draw the rectangles $[a_j - L/2, a_j + L/2] \times [0, f_j]$ where f_j is the frequency associated with the stem a_j. Sometimes such a bar graph is called a **histogram**.

(c) If you have a stem-and-leaf display for data and you observe that the stems a_j, $j = 1, 2, \ldots, k$ do not satisfy the condition

$$a_{j+1} - a_j \text{ is constant for } j = 1, \ldots, k-1,$$

what does this tell you about the raw data?

(d) Can you recover the length measurements of the raw Milkyway mini data from the histogram?

Problem 4 (relative frequency) If the frequencies $f_1, f_2, \ldots, f_k$ in a stem-and-leaf display are divided by the total number of sample values (in this case 50), one obtains the **relative frequency data** for the sample.

(a) Arrange the stems from the stem-and-leaf display obtained in Problem 1 along a horizontal axis and plot the values of the associated relative frequencies as a function of the stem. The result is called a **relative frequency histogram** or a **density histogram**.

(b) What probabilistic construction does the relative frequency histogram resemble? Hint: What is the sum of the relative frequencies?

(c) Make a bar graph histogram for the relative frequency data using the procedure described for the frequency data in part **(b)** of Problem 3 above.

(d) What is the area covered by the rectangles in your bar graph histogram from part **(c)** above?

Problem 5 (relative frequency polygon) Make a second copy of your relative frequency histogram and relative frequency bar graph histogram (on the same coordinate plane) from parts **(a)** and **(c)** of Problem 4 above.

(a) Consider the stem values $a_1, a_2, \ldots, a_k$ and the associated relative frequency values $r_1, r_2, \ldots, r_k$ mentioned in part **(b)** of Problem 3. Starting from the left on your second copy of the relative frequency histogram, connect the following points with a polygonal path:

$$(a_1 - L, 0), (a_1, r_1), (a_2, r_2), \ldots, (a_k, r_k), (a_k + L, 0).$$

The result is called the **relative frequency polygon**.

(b) What is the area under the relative frequency polygon?

(c) The relative frequency polygon determines (and is equivalent to) a piecewise affine function $\rho : [a_0, a_{k+1}] \to [0, \infty)$ where $a_0 = a_1 - L$ and $a_{k+1} = a_k + L$. Find the formula for the value

$$\rho(\omega) \qquad \text{for } a_j \le \omega \le a_{j+1},\ j = 0, 1, \ldots, k.$$

(Just give the formula in terms of the parameters

$$a_0, \ldots, a_{k+1}, \quad 0 = r_0, r_1, \ldots, r_k, r_{k+1} = 0, \quad \text{and} \quad L$$

instead of the actual numerical values associated with Milkyway minis.)

Let's call the function ρ a **piecewise affine density**.

(d) Under what condition(s) is $\delta : \mathbb{R} \to [0, \infty)$ by

$$\delta(\omega) = \rho(\omega)\, \chi_{[a_0, a_{k+1}]}(\omega)$$

the MDF of an integral[1] probability measure?

[1] i.e. continuous

Problems 6-9 are more or less designed to outline a way for you to produce some "interesting" data related to Problem 10 of Assignment 6. Broadly speaking, there are probably three approaches by which we (meaning you) can develop some understanding of the basics of statistics. One is to consider various "colorful" data sets and analyze them. From some point of view, each such data set is not too different from the one considered in Problems 1-5 above. We could, for example, do one that is a little more "Singapore math" style and consider measured weights of Bánh Pía durian hopia cakes which weigh about 40 grams each. Maybe this is worthwhile.

A second (also not so different) approach might be to consider some basic problems that make headlines. The two obvious ones that come to mind are (1) How do statisticians deal with polling data to predict election results? and (2) How do statisticians deal with data from medical trials to determine if drugs and/or treatments are effective? There could be other variations of "popular" data to consider.

The problems below take up a third approach which in some ways is simpler. In some ways, however, it is a little complicated. It certainly comes across as a little more abstract. We generate data using simulation in some well-defined way, and then ask the question: Can you tell from the data the process by which the data was generated? What can you tell? Under what circumstances can you tell it and how? And what can you not tell?

Problem 6 (MDF on a finite interval) The construction of the relative frequency polygon and its associated piecewise affine density ρ in Problem 5 above may be applied to any PMF $M : \mathbb{R} \to [0,1]$ associated to some probability measure $\pi : \wp(R) \to [0,1]$ where $R \subset \mathbb{R}$ is a measure space with $\#R < \infty$. For example, the PMF M and π may be taken to be

$$M(\omega) = \begin{pmatrix} n \\ \omega \end{pmatrix} p^\omega (1-p)^{n-\omega} \qquad \text{for } \omega = 0, 1, 2 \ldots, n \tag{1}$$

and $\pi = \beta$ the binomial measure.

(a) Let $a_1, a_2, \ldots, a_k \in \mathbb{R}$ with

$$a_{j+1} - a_j = L \quad \text{(constant)} \quad \text{for } j = 1, 2, \ldots, k-1$$

with associated PMF values $M(a_j) = \pi(\{a_j\})$ for a probability measure π. Define values a_0 and a_{k+1} and construct a piecewise affine density ρ by appropriately adapting the construction of Problem 5.

(b) Show $\delta : \mathbb{R} \to [0, \infty)$ by

$$\delta(\omega) = \frac{1}{L}\, \rho(a_0 + (k+1)L\omega)\ \chi_{[0,1]}(\omega) \tag{2}$$

is an MDF.

(c) Find the mean and variance of the MDF defined in (2).

Problem 7 (PMF on $[0,1]$) Something like the scaling used in part **(b)** of Problem 6 above may also be applied directly to the PMF of the binomial distribution.

(a) Let $a_j = j/n$ for $j = 0, 1, \ldots, n$. Show $N : \{a_0, a_1, \ldots, a_n\} \to [0, 1]$ by

$$N(\omega) = M(n\omega)$$

is a PMF.

(b) Use R to plot the PMF N for $n = 100$ and $p = 0.6$.

Problem 8 (interesting Bernoulli data) Say we wish to conduct 15 experiments and each experiment consists of three Bernoulli trials simulated using `rbinom(3,1,p)` where p is also a simulated value following the distribution associated with $n = 100$, $p = 0.6$ and N from Problem 7.

(a) Execute the following `R` commands to understand N:

```
> x <- 0:100
> y <- dbinom(x, 100, 0.6)
> plot(x/100,y)
```

(b) Execute the commands

```
> probsbase <- rbinom(15,100,0.6)/100
> probs <- replicate(n=3,probsbase)
> apply(probs,1,function(z)rbinom(3,1,z))
```

(c) Explain the meaning of each command in part **(b)** above, what is being generated, and the final output.

Problem 9 (statistical analysis) Generate a data set using the following R commands

```
> set.seed(161)
> probsbase <- rbinom(10000,100,0.6)/100
> probs <- replicate(n=100,probsbase)
> set.seed(361)
> dataset <- apply(probs,1,function(z)rbinom(100,1,z))
```

(a) Viewing each column of `dataset` as the result of the execution of 100 Bernoulli trials with the same probability p, estimate the value p used in those trials. You should obtain a vector `eprobs` with 10000 estimated values of p, one for each trial.

(b) Produce a relative frequency histogram for the data `eprobs`.

(c) What does your picture from part **(b)** resemble and why?

(d) How would the code and results change if instead of 100 trials in each of the 10000 experiments, there were 1000 trials?

Problem 10 (David Chan's probability problem) Give a complete and detailed solution with a full explanation for the following problem:

> What is the probability that team A wins four games in a seven game series against team B if the probability that team A beats team B each time the two teams play is p?