Caltech Department of Mathematics

Ma 3/103 KC Border
Introduction to Probability and Statistics Winter 2017

Assignment 3

Due Monday, January 23 by 4:00 p.m. at 253 Sloan

Instructions:

When asked for a probability or an expectation, give both a formula and an explanation for why you used that formula, and also give a numerical value when available.

When asked to plot something, use informative labels (even if handwritten), so the TA knows what you are plotting, attach a copy of the plot, and, if appropriate, the commands that produced it.

Exercise 1 (30 pts) Is it possible to have three random variables X, Y, and Z, where X and Y are stochastically independent, Y and Z are stochastically independent, and X and Z are stochastically independent; but the set $\{X, Y, Z\}$ of random variables is *not* stochastically independent? Explain why your answer is correct. □

Exercise 2 (Problem 3.3.15 in Pitman) (20 pts)
Let X and Y be independent random variables. Show that

$$\boldsymbol{Var}(X - Y) = \boldsymbol{Var}(X + Y).$$

□

Exercise 3 (Cf. problem 3.3.26 in Pitman) (25 pts) Use Jensen's Inequality (Lecture 6) to show that for a random variable X with finite mean μ,

$$\text{std. dev. } X \geqslant \boldsymbol{E}|X - \mu|,$$

with equality if and only if $|X - \mu|$ is degenerate. □

Exercise 4 (25 pts) There are n balls numbered $1, \dots, n$ and n bins numbered $1, \dots, n$. The balls are put into the bins at random, one per bin. What is the expected number of balls put in the matching bin? Explain your reasoning. (Hint: Let E_i be the event that ball i is in bin i. Use indicator functions.) □

Exercise 5 (Exploring some data) (40 pts)

The empirical distribution function is a good way to look at data. It just charts for each number t the fraction of the data that is less than or equal to t. I believe that in many ways, it is superior to histograms for "eyeballing" the distribution of data. Let's see what you think.

At the beginning of the term you flipped coins. This generated a long string of 0s and 1s. A segment of this string can be interpreted as a binary number, and by dividing this by the appropriate power of two, it can be interpreted as a number between 0 and 1. Moreover, if the coin tosses are independent and Heads and Tails are equally likely, then these numbers should be i.i.d. with an approximately uniform distribution. We are going to subject this to an "eyeball test," which is one of the first things you should always do with data.

I have taken the liberty of chopping the coin toss data from this year into 776 strings of length 32, and converting them into numbers between 0 and 1. You can download these results from http://www.math.caltech.edu/~2016-17/2term/ma003/Data/Random32.txt. Or you can do it yourself from the raw data at http://www.math.caltech.edu/~2016-17/2term/ma003/Data/Flips.txt

Using the program/language of your choice do the following. (I give hints for R and Mathematica below.)

1. What is the expected value of a Uniform[0,1] random variable? What is its standard deviation?

2. What is the average of the numbers in your samples? What is the sample standard deviation of each sample? (The sample standard deviation is gotten by squaring the deviation of each sample value from the sample mean, summing them, dividing by (sample size − 1), and then taking the square root.

3. Plot a histogram of these numbers, using the default. Then plot a histogram using bins of length 0.05.

4. Now plot a cumulative histogram or the empirical cumulative distribution function. (In Mathematica, this is just an option of the `Histogram` command, and in R use the `ecdf` command.)

5. Which method makes it easier to check by eye if the data appear to be uniform?

In an appendix, I give some badly documented sample code that you might find useful.

□

Exercise 6 (10 pts) How much time did you spend on the previous exercises? □

Exercise 7 (Optional Exercise) (50 pts) There are n balls numbered $1, \dots, n$ and n bins numbered $1, \dots, n$. The balls are put into the bins at random, one per bin. For each $k = 0, \dots, n$, what is the probability that exactly k balls are put in the matching bin? Explain your reasoning. (Hint: Let E_i be the event that ball i is in bin i. Use indicator functions.) □

Appendix: Sample code

If you don't have a preference, there is a lot to be said for learning the R statistical programming language. It is used widely on campus, and it looks like it will be around for a while. It is also **free** and runs on the major operating systems. You can get it at http://www.r-project.org. But if you are familiar with something else, go ahead. Even Excel can probably handle this assignment, but future ones may be trickier.

Hint: Badly documented sample R code:

Warning: I am not an R programer, and I am sure there are probably better ways to do things. Most of what I know I got by Googling various questions. Also—typing `?command` will bring up help on the command `command`.

First, use `setwd("your_data_pathname")` to change your working directory to the folder where the data file is. (Or be prepared to use a full path name.) You can use `getwd()` and `list.files()` check that you are in the right place.

Read the data from the file into an array. Check the length, it should be 776 for the file Random32.txt. (# is a comment character.)

```
a = as.matrix(read.table("Random32.txt"))  # the as.matrix is important!
length(a)
```

Now try a default histogram:

```
hist(a)
```

Now try a histogram with bins of size 0.05. Also instead of actual counts, use relative frequencies (density):

```
bins=seq(0.0,1.0,by=0.05)
hist(a, breaks=bins, freq=FALSE)   # freq=FALSE uses relative frequencies ?!
```

Now let's examine the empirical cdf.

```
c=ecdf(a)
plot(c)
```

How do you save these plots? Well on my Mac, I just click on the graphic's window and hit Save, and it saves the graphic as a pdf. But here is a better way. Say you want to save the plot above to a `png` file named `Hist.png`. Here you go:

```
png("Hist.png")     # open the file for writing
plot(c)             # plot to the file
dev.off()           # close the file.  This is crucial.
```

To save to a `pdf` file use `pdf("Hist.pdf")` for the first line. I found this at http://wiki.stdout.org/rcookbook/Graphs/Output%20to%20a%20file/.

Hint: Undocumented sample Mathematica code:

```
SetDirectory["Your path goes here"]
a = Flatten[ Import["Random32.txt", "Table"] ];
g = Histogram[a]
Export["File name 1.pdf",g]
g = Histogram[a, {0, 1, 0.05}]
Export["File name 2.pdf",g]
```

References

[1] R. J. Larsen and M. L. Marx. 2012. *An introduction to mathematical statistics and its applications*, fifth ed. Boston: Prentice Hall.

[2] J. Pitman. 1993. *Probability*. New York, Berlin, and Heidelberg: Springer.