Statistics 101 Summer I 2011
Homework #2

Adam Kapelner, Instructor

Due 1PM, Thursday, June 9, 2011 (in my mail slot)

Instructions and Philosophy

It's easy for me to pretend to teach and you to pretend to learn. It's hard for me to actually teach and for you to actually learn. This requires you to put in a lot of work. This homework is very long but it will really teach you the material. I strongly suggest you start early and *work in teams*; do not do this one solo.

In Stine & Foster, read sections 9.1–9.4, 10.1, 10.3, 10.5 (ignore everything about joint distributions, covariances, and dependent random variables), then 11.1–11.4, then loop back and read ch. 1–4 and stop at p.63 (do not read about the empirical rule). Section 4.2 is about histograms, so please read that carefully.

Once again, green means *easy*, yellow means *intermediate*, red means *difficult*, and purple means *extra credit*. This homework is worth 100 points but the point distribution will not be determined until after the due date. Late homework will be penalized 10 points per day. Beyond Friday June 10 at 5PM, it will receive a zero. 15 points are given as a bonus if the homework is typed using LaTeX (please comment out all extraneous text).

Fundamentals of Random Variables Problems below are related to the fundamentals of r.v.'s which you'll find mostly in ch. 9 of S&F. This section will also cover rules of expectation, variance, and summing r.v.'s found in ch. 10.

Problem 1 Imagine rolling two dice. Let X_1 be the r.v. corresponding to the first die and let X_2 be the r.v. corresponding to the second die. Let the outcomes be \$1 if you roll a 1, \$2 if you roll a 2, ..., and \$6 if you roll a six.

(a) Are these two r.v.'s independent? Identically distributed? Are $X_1, X_2 \overset{iid}{\sim}$?

(b) Find $\mathbb{E}[X_1]$, $\mathbb{V}\text{ar}[X_1]$, $\mathbb{SD}[X_1]$.

(c) Draw the probability mass function (PMF) to scale of X_1 similar to how we did in class (see p.199).

(d) Draw the PMF of $2X_1$.

(e) Imagine a new r.v., call it Y, where you make \$11 if you roll a 1, \$14 if roll a 2, ..., and \$26 if you roll a 6. Express Y as a function of X_1. Using elementary transformation theory, find $\mathbb{E}[Y]$, $\mathbb{V}\text{ar}[Y]$, $\mathbb{SD}[Y]$.

(f) Define a new r.v. $S = X_1 + X_2$. Using the methods in class, find its PMF and then draw it to scale.

(g) What is the support of S?

(h) Calculate $\mathbb{E}[S]$ from first principles and $\mathbb{V}\text{ar}[S], \mathbb{SD}[S]$ from the formulas.

Problem 2 We will investigate some basic facts of mathematical statistics. Consider $X_1, \ldots, X_n \overset{iid}{\sim}$ (some PMF) with nonzero mean μ and nonzero variance σ^2.

(a) Define S_n as we did in class. and compute $\mathbb{E}[S_n]$, $\mathbb{V}\text{ar}[S_n]$, $\mathbb{SD}[S_n]$.

(b) Calculate all three of the following:

$$\lim_{n\to\infty} \mathbb{E}[S_n], \quad \lim_{n\to\infty} \mathbb{V}\text{ar}[S_n], \quad \lim_{n\to\infty} \mathbb{SD}[S_n]$$

Why do your answers make sense?

(c) Define $\bar{X}_n$ as we did in class and compute $\mathbb{E}\left[\bar{X}_n\right]$, $\mathbb{V}\text{ar}\left[\bar{X}_n\right]$, $\mathbb{SD}\left[\bar{X}_n\right]$.

(d) Calculate all three of the following:

$$\lim_{n\to\infty} \mathbb{E}\left[\bar{X}_n\right], \quad \lim_{n\to\infty} \mathbb{V}\text{ar}\left[\bar{X}_n\right], \quad \lim_{n\to\infty} \mathbb{SD}\left[\bar{X}_n\right]$$

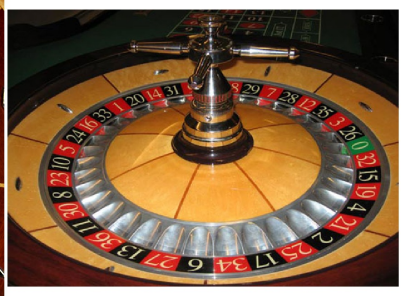
Why do your answers make sense?

(e) Prove from the definition of variance that $\mathbb{V}\text{ar}[X] = \mathbb{E}[X^2] - \mu^2$. This is a very useful formula that you should add to your toolbox.

(f) Let X, Y be discrete r.v.'s just like we are used to in class. Prove $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ from first principles. Super-hard... but if you slog through it, you will have a good command of density functions.

(g) Why does s^2 have the correction factor of $\frac{1}{n-1}$? Read up on Bessel's correction. Define what a "biased estimator" is and prove that s^2 is unbiased. Do not attempt this unless you are *already done* with the assignment.

Problem 3 We are going to learn the casino game of Roulette.

The figure on the left is an American Roulette wheel; the figure on the right is a European Roulette wheel. American Roulette has 38 pockets and European Roulette has 37 pockets. Consider the standard bet of \$5.

(a) Create a r.v. X for betting on the ball landing in black on an American Roulette table. The payoff is 1:1. Use the $\sim$ and piecewise function notation from class. Think carefully about what losing is: how much do you lose? How much do you win? This will enable you to get the outcomes / states of X.

(b) Draw the PMF and calculate $\mathbb{E}[X]$, $\mathbb{V}\text{ar}[X]$, $\mathbb{SD}[X]$ from first principles. Indicate μ on the PMF with the fulcrum symbol like we did in class.

(c) Now, model Y, one bet of \$10,000 on black and an entrance of \$10 to the casino. Calculate $\mathbb{E}[Y]$ and $\mathbb{V}\text{ar}[Y]$ using elementary transformation theory.

(d) Now, model V, one bet of \$10,000 on black with a tip to the cashier of \$200 *if* you win. Calculate $\mathbb{E}[V]$.

(e) Repeat part (b) for a European Roulette wheel.

(f) Would you rather play Roulette in America or Europe? Why?

(g) What is the expected *average* of 100 bets on the number 17? The payoff there is 35:1 and the bet is still \$5 and we are playing in Las Vegas.

Problem 4 In the finance industry, Sharpe ratios are traditionally used to compare mutual funds. In this problem we are going to compare funds JARTX and VFAIX (click the links to search them on Google Finance). Both of these funds invest in large-cap US equities.

(a) Open `two_mutual_funds.JMP` which tabulates the date and the *weekly* return rates for these two mutual funds. Multiply by 100 to get percentages. Now let JMP compute $\bar{x}$ and s for both mutual funds using the "analyze...distribution" function. Calculate the Sharpe ratios for both mutual funds by using the formula from p.208 of F&S. Use $\bar{x}$ to estimate $\mathbb{E}[X]$ and s to estimate σ. Use $r_f = 0.0769\%$ as the "risk free" weekly interest rate.[1]

(b) If you had a choice between these two funds, which would you buy and why?

Bernoullis, Binomials, and Poissons In this section we will be covering the material in Chapter 11.

Problem 5 Derivation of basic facts

(a) Let $S_n = X_1 + \ldots + X_n$ where $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$. How is S_n distributed? Indicate the parameter values clearly and explain them in your own words.

(b) Rederive the PMF of the $\text{Binomial}(n, p)$ just like we did in class from the notes. Explain each step. When you get the final PMF, label what each part of this function means in your own words.

(c) Let $Y \sim \text{Poisson}(\lambda)$. Prove $\mathbb{E}[Y] = \lambda$ from first principles *i.e.* from the definition of expectation. This one is super hard. Do not do this unless you have hours to spare and wish to sharpen your algebra skills. This would be considered an intermediate problem in Stat 430 — the course on probability. Hint: use the Taylor series: $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$

Problem 6 We will be investigating the situation where there are $n = 1500$ students and $p = 2\%$, the probability they enroll in Stat 101 in the summer. Assume each student is ruggedly individualistic about their choice of classes and they each have the same interest level in Statistics.

(a) Calculate the probability of 30 students enrolling in Stat 101. This is the expected number of students to enroll. Does the probability you calculated feel low to you? Explain.

(b) This is a case where n is high and p is low. Let U be the r.v. that models the approximation we learned about in class. How is U distributed? Indicate your parameters clearly.

[1] This was calculated using 3-month T-bills, but that's not important for the scope of this class. In your finance classes you will cover this in more depth.

(c) The R code below will graph the PMF of $Y \sim \text{Binomial}(1500, 2\%)$ in red and the PMF of U in green. Copy and paste the code into the console. If this doesn't work, no excuses, copy and paste it to a text editor, format it exactly as you see below and try again.

```
n = 1500;
p = 0.02;
support_max = 300;
bin = array(NA, support_max);
pois = array(NA, support_max);
for (i in 1 : support_max){
  bin[i] = dbinom(i, n, p);
  pois[i] = dpois(i, n * p);
}
plot(bin[1:60],
  pch = 16,
  col = "red",
  xlab = "Number of students enrolled in Stat 101",
  ylab = "probability",
  main = "Class Enrollment Probabilities for Stat 101");
points(pois[1:60], col = "green", pch = 16);
#placeholder for last line
```

Is U a good approximation of Y? Why or why not? Print out the plot and attach it to your homework.

Data We will be reviewing the concepts of data as realizations of r.v.'s and do some modeling in this section.

Problem 7 On p197-198, F&S investigate IBM stock returns by using a mock model. We're going to build a more believable statistical model by using real data for the past 100 trading days between Jan 7, 2011 and June 1, 2011. This will *still* be a mock model. Building more believable models is what hedge fund dudes get paid the big bucks for.

(a) Go to the Yahoo Finance page then click on "historical prices" on the left (you can click on the link above to go straight there). Query for the data between Jan 7, 2011 and June 1, 2011. Click the "Download to Spreadsheet" link. You now have the file `table.csv`. Open this up in JMP and sort so that the date Jan 7, 2011 is first. Now create a new column that calculates the daily returns *as a percentage*. Use the *adjusted closing* price[2]. Now create a histogram of the returns. Print this out and attach it to your homework.

(b) Are daily returns an $\overset{iid}{\sim}$ model? Discuss both conditions and if you think they hold.

(c) Find $\bar{x}$ and s. What do these *statistics* estimate?

(d) What is the median and what is the IQR? Do these statistics estimate a parameter we spoke about in class?

(e) Pretend $\bar{x}$ is the expectation[3] and s is the standard deviation. Estimate the probability of making more than 1% in one day trading IBM stock. The graphical features of JMP make this easy.

(f) Look at the Box and Whisker plot above the histogram. Is the line in the center of that diamond? Now expand the box and whisker plot so it's full screen. Is the line still in the center of the diamond? Explain why the line is slightly off center and interpret the difference in distance in the context of the summary statistics.

(g) Use JMP to locate the row that is the outlier of highest negative return. What day is it? Can you find a story online about IBM that day?

(h) In the next 100 days in the future, estimate the probability you have a positive daily return for *half* the days.

(i) In the next 100 days in the future, estimate the probability you *lose* more than 2% on *at most* 5 days.

(j) Make a histogram of the volumes. What is the sample average and the sample median? Is this symmetric? Does it have a skew? What does this skew mean?

[2] you will learn why this is different from the regular closing price in your finance class

[3] we are pretending that 100 days is enough of a "long term" to satisfy the law of large numbers (we'll discuss this in class in the future)

(k) Figure out how to visualize the distribution of returns *by month*.[4] Print out histograms of the returns by month and attach them to your homework. Is there a month effect on returns?

[4]Hint: Convert the date column to character type and use the substr function on a new column formula.