Statistics 101 Summer I 2011
Homework #3

Adam Kapelner, Instructor

Due noon, Friday, June 17, 2011 (in my mail slot)

Instructions and Philosophy

Copy paragraph one from last week's introduction about working hard on problems and working in teams.

In Stine & Foster, read chapter 4 over including the last section on the empirical rule, read chapter 13 on surveys and sampling, read chapter 12 on continuous r.v.'s and the central limit theorem, then chapter 14 on basic applications including control limits. You can ignore the subsection on skewness and kurtosis (pp 275-276) and the subsection on "sample size condition" to test for normality (pp 328-329). Also, read the first few pages of chapter 15.

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 17 at 5PM, it will receive a zero. 15 points are given as a bonus if the homework is typed using LaTeX (please comment out or delete all extraneous text).

Basic Data Analysis and Sampling This section will cover the nitty-gritty of data tables, summary statistics, and sampling / surveys.

Problem 1 This is a simple data problem. The ages of 25 billionaires living in NY today are 66, 92, 54, 86, 59, 79, 64, 78, 60, 73, 34, 42, 85, 77, 90, 64, 64, 64, 78, 58, 48, 78, 81, 72, 66.

(a) Find the average and sample standard deviation with JMP.[1] Indicate units.

(b) Estimate the mean and standard deviation for all billionaires in NY.

(c) In the previous question, are we allowed to do that? Explain.

[1] This part was originally assigned without JMP, but now that we've had the midterm, you can use JMP only if you know the formulas well, otherwise I recommend doing it from scratch so you learn how to to do it well.

(d) Find the five-number summary, IQR, and range.

(e) Without doing the next parts of the question, evaluate whether or not there is a skew.

(f) If there is a skew, what methods would you use to fix the skew? Does the transformation make sense?

(g) Find the median of the transformed data set. Indicate how you would find the mean but do not calculate it.

(h) How will the median and mean change if the youngest and oldest billionaires in the sample moved out of NY?

(i) Construct a histogram with bin width = 5 by hand

(j) Construct a box and whisker plot by hand above the histogram. Add dots for outliers if they exist. For the half-width of the diamond, use 7. We will be learning how to calculate this half-width next week.

(k) Could we use the empirical rule to estimate the amount of data between sd's?

Problem 2 This problem is to be done without the use of JMP. Table 1 is an excerpt of real data from a study I did with a collaborator at MIT about online survey presentation.[2]

(a) We spoke about different types of variables in class: numeric (either interval or taio) and categorical (either nominal or ordinal). What kind of variable is the "treatment" column?

(b) Create a pie chart for the data in the "treatment" column.

(c) What kind of variable is the "id" column? This question is designated intermediate for a reason.

(d) The "created at date" column is sort of useless. Explain how you would combine the "created at date" column and another column to make a more useful variable.

(e) What type of data is the "amount qu 1" column?

(f) What type of data is the "# switches" column?

(g) The "# switches" column looks like it's trimodal (*i.e.* has three clusters). This is an example where it may be easier to analyze the data if we convert it to a categorical variable. Let's do this. Name the variable "switch behavior" and convert each observation into your new variable.

(h) What type of variable is the new "switch behavior" column?

[2]If you care more, click here to watch us embarass ourselves in San Francisco (minute 7:35).

id	treatment	created at date	created at time	amount qu 1	# switches
1147	kapcha	9/1/2010	13:33:42	1.75	2
1150	timing	9/1/2010	13:33:43	3	194
1152	timing	9/1/2010	14:00:08	1.5	8
1153	kapcha	9/1/2010	14:00:08	5	200
1154	kapcha	9/1/2010	14:00:08	1.5	31
1157	control	9/1/2010	14:00:10	1	32
1158	exhortation	9/1/2010	14:00:10	2	32
1159	kapcha	9/1/2010	14:00:10	1.29	40
1160	exhortation	9/1/2010	14:00:11	2	196
1161	timing	9/1/2010	14:00:11	1.5	214
1162	exhortation	9/1/2010	14:00:12	1	200
1163	kapcha	9/1/2010	14:00:12	1.5	208
1170	control	9/1/2010	14:00:15	1.5	32
1171	kapcha	9/1/2010	14:00:15	2	36
1173	control	9/1/2010	14:00:16	1.25	196
1178	control	9/1/2010	14:00:18	2.5	195
1184	timing	9/1/2010	14:00:20	1.3	223
1186	kapcha	9/1/2010	14:00:21	2	203
1187	exhortation	9/1/2010	14:00:22	2	199
1188	kapcha	9/1/2010	14:00:22	1	36
1191	kapcha	9/1/2010	14:00:23	2	200
1192	timing	9/1/2010	14:00:24	2.5	190
1197	kapcha	9/1/2010	14:00:26	1	0
1198	exhortation	9/1/2010	14:00:26	5	32

Table 1: Data excerpted from the Kapcha study

(i) Now, create a bar chart for the "switch behavior" data. Does the order of the categories on the x-axis matter?

Problem 3 This problem will ask questions about sampling theory. This will be the fuzziest section of any homework I will assign. You may want to have the book and your notes handy.

(a) Define "population" and "sample" and their respective sizes using the notation from class.

(b) Explain why a non-representative sample can be biased and why that would be bad.

(c) We can use $\bar{x}$ to estimate μ and s^2 to estimate σ^2 only if the data was sampled using a _____. Fill in the blank and explain.

(d) If your sampling method is non-representative, explain why simply sampling more data won't help you.

(e) Download the `cars_mpg.JMP` file. Use these cars as a sampling frame and sample 30 cars using the method taught in class (p.309 of S&F). Print out your resulting table and circle the 30 you're choosing.

(f) Do the same sampling five times and list the five $\bar{x}$'s for the HWYMPG column for each of the 5 samplings. Why are each of these $\bar{x}$'s different?

(g) I am running a survey for my arts and crafts store and I sample 30 customers between the ages of 5-15, 30 customers between the ages of 15-25, 30 customers between the ages of 25-35, 30 customers between the ages of 35-45, 30 customers between the ages of 45-55, 30 customers between the ages of 55-65, and 30 customers between the ages of 65-75. What kind of survey is this? Is it biased? Can the biased be fixed?

(h) I want to find out about people who use public transportation in Philadelphia so I talk to people who will talk to me on the 42 bus[3]. Is this a simple random sample? If not, what did I fail to account for?

(i) When the United States runs their census every 10 years, what is their sampling goal?

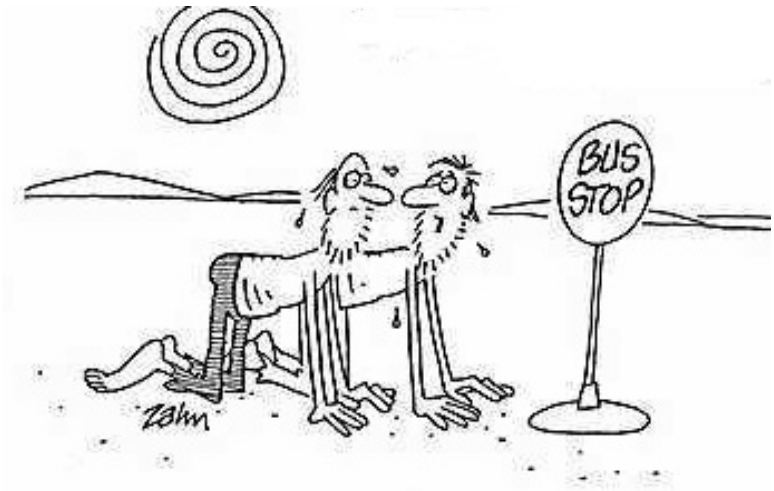
Continuous r.v.'s This section will cover some ideas about continuous r.v.'s as well as the basics of the normal curve. Since we are not focusing on computation, I will expect you to use tools that make the computation trivial, so you can learn the concepts.

Problem 4

In this problem we will look at the exponential r.v. below:

$$X \sim \text{Exp}(\lambda) = f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

Notice we have one parameter, λ. Exponential random variables are used to model waiting times. Let's pretend X models the time to wait for a bus in minutes.



"It wouldn't hurt to wait around for a little while."

(a) To make the case simple, let $\lambda = 0.1$, something we'll relax later. Write the PDF (not PMF, this is a continuous r.v.). Use the $f_X(x)$ notation. Be sure to indicate where x is defined.

[3]service to Upper Darby via Walnut Street and Spruce Street

(b) Plot the PDF by sketching it with pencil. Since this is not a calculus class, I recommend using the website Wolfram Alpha (WA). Type in "`Plot[0.1 e^{-0.1 x}, {x,0,50}]`" into WA and press enter. Copy the plot onto your paper.

(c) Find the probability that you wait for the bus more than 10 minutes *i.e.* $\mathbb{P}(X > 10)$. On WA, type in "`Integrate[0.1 e^{-0.1 x}, {x, 10, Inf}]`". Make sure you write the integral you use, then answer as a number.

(d) Calculate $\mathbb{E}[X]$ by using "`Integrate[x 0.1 e^{-0.1 x}, {x,0,Inf}]`".

(e) Calculate $\mathbb{E}[X]$ with λ as an indeterminate parameter. This will require you to use WA a little more creatively.[4]

(f) Let's say you've been at the busstop for 10min already. What's the probability that you wait more than 10 minutes more? Compare with part (c). Does that make sense? Is the bus tricking you? Explain why the exponential r.v. is called the "memoryless" r.v.

Problem 5 We will practice using the normal table in this example. Please *draw a picture* and shade in the relevant areas for parts a—f. We assume the standard notation that $Z \sim \mathcal{N}(0, 1)$. Find...

(a) $\mathbb{P}(Z < 1.8)$

(b) $\mathbb{P}(Z \leq 1.8)$

(c) $\mathbb{P}(Z \leq -1.8)$

(d) $\mathbb{P}(|Z| \leq 1.8)$

(e) $\mathbb{P}(|Z| \geq 1.8)$

(f) $\mathbb{P}(Z \in [-1.5, -1.2] \cup Z \in [1.2, 1.5])$

(g) the z such that $\mathbb{P}(Z < z) = 0.99$

(h) the z such that $\mathbb{P}(Z < z) = 0.95$

(i) the z such that $\mathbb{P}(Z < z) = 0.68$

(j) $z_{0.0015}$ *i.e.* find the z such that $\mathbb{P}(|Z| < z) = 0.997$

(k) $z_{0.025}$ *i.e.* find the z such that $\mathbb{P}(|Z| < z) = 0.95$

(l) $z_{0.16}$ *i.e.* find the z such that $\mathbb{P}(|Z| < z) = 0.68$

(m) What are the z values to keep in mind for the *empirical rule* (p.65)? Refer to the last six subparts and the find the three we are looking for.

[4] Don't be confused by the output "Re(·)". This is WA making sure whatever your parameter variable you chose is a real number, which it is for the real case of waiting for a bus.

Problem 6 We will practice shifts and scales. Consider the following:

$$X \sim \mathcal{N}\left(\mu,\, \sigma^2\right) \quad \text{and now we create a transformation:} \quad Z = \frac{X - \mu}{\sigma}$$

(a) What is $\mathbb{E}[X]$ and $\mathbb{SD}[X]$? This is training you to understand the notation.

(b) What is the transformation into Z called? Why do we use that name?

(c) Find $\mathbb{E}[Z]$ using the formulas from chapter 9.

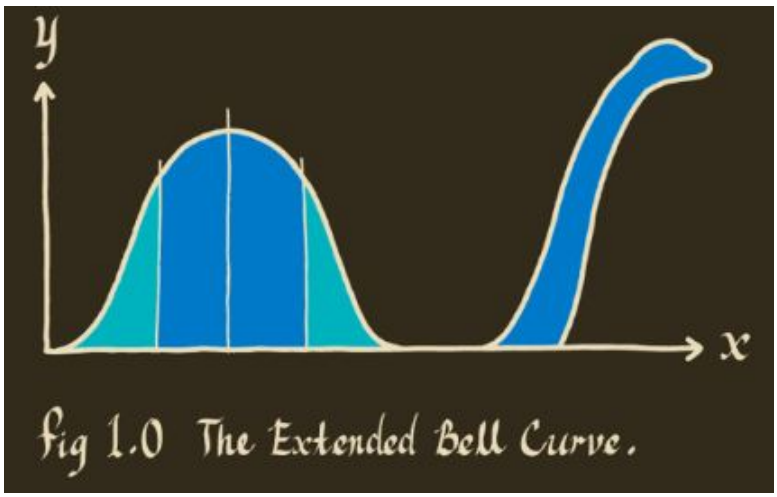
(d) Find $\mathbb{SD}[Z]$ using the formulas from chapter 9.

(e) Graph all three of the following r.v.'s to scale on the same axis and label $f_{X_1}(x)$, $f_{X_2}(x)$, and $f_{X_3}(x)$:

$$X_1 \sim \mathcal{N}\left(-3,\, 1^2\right), \quad X_2 \sim \mathcal{N}\left(4,\, 2^2\right), \quad X_3 \sim \mathcal{N}\left(1,\, \left(\frac{1}{10}\right)^2\right)$$

Rule of thumb: by ± 3 sd's away from the mean, the PDF (the bell shape) should basically be zero. Skinny bells have to be taller than fatter bells and vice versa.

(f) Let X_3 model the number of quarts of milk in a bottle delivered by a home delivery milk service. Find the amounts of milk you get only 10% of the time as extreme values (you need both the low and the high).



Problem 7 We will look at applications of the Central Limit Theorem (CLT).

(a) State the CLT for $\bar{X}$ using the μ, σ notation. Be as complete as you can.

(b) Assume $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}\,(p)$. State the CLT for $\hat{P}$ using the notation from class. Explain why we use the $\hat{P}$ notation and not the usual $\bar{X}$ notation.

(c) You randomly sample 100 people in an apartment building. Assume the population as a whole has equal distribution of males and females. What is the probability more than 60 people are male? Assume this sample is an SRS with respect to gender.

(d) In homework #2 problem 3 (a), you created a r.v. for the outcome of a \$5 American Roulette bet on black. Find the probability after 30 bets you come out ahead.

(e) Assume the same configuration in the previous problem. Find the probability after 40 bets you lose more than 30 cents *on average*.

(f) Find a quote of beauty about the CLT that is different from the one in class. Write down the quotation, where you got it from, who said it (whether it's apocryphal or not), and the year.

Problem 8 We will now examine how quickly the CLT converges for the binomial r.v. with different n, p. Download the file `binomial_clt.txt` from the webcafe portal and open it with a text editor. Copy all the code into `R` and press enter. This may take your computer 5 full minutes to compute, please be patient. You will now see a large plot. Each row of the plot represents the following random variables indexed by row number:

$$\begin{aligned} X_{\text{row 1}} &\sim \text{Bernoulli}\,(0.5) \\ X_{\text{row 2}} &\sim \text{Bernoulli}\,(0.1) \\ X_{\text{row 3}} &\sim \text{Bernoulli}\,(0.01) \\ X_{\text{row 4}} &\sim \text{Bernoulli}\,(0.001) \end{aligned}$$

Therefore, the row varies p. The six columns vary n, representing different sample sizes: $n = 5, 10, 100, 1000, 10000, 100000$. The plots are the histogram estimates of the PMF's S_n (the sum of n for each of the different r.v.'s).

(a) Maximize the plot window so you can examine it as closely as possible. At what n for each of these Bernoulli's does its S_n distribution approach the normal density *i.e.* appear bell-like in the plot window? Phrased equivalently: at what n does the binomial "look" normal?

(b) We learned about the two sample size conditions for the normal approximation to the binomial in class as Rule 1: $np > 10$ and Rule 2: $n(1-p) > 10$. Below are the values of each of these calculations:

$p\downarrow, n\rightarrow$	5	10	100	1000	10000	1000000
0.5	2.5	5	50	500	5000	500000
0.1	0.5	1	10	100	1000	100000
0.01	0.05	0.1	1	10	100	10000
0.001	0.005	0.01	0.1	1	10	1000

Table 2: Calculation of np

$p\downarrow, n\rightarrow$	5	10	100	1000	10000	1000000
0.5	2.5	5	50	500	5000	500000
0.1	4.5	9	90	900	9000	900000
0.01	4.95	9.9	99	990	9900	990000
0.001	4.995	9.99	99.9	999	9990	999000

Table 3: Calculation of $n(1-p)$

For each Bernoulli, what is the minimum n that satisfies *both* conditions? Is your answer the same as part (a) when you visually inspected the plots?

(c) Sum up what you've learned: when can we use the normal table to approximate the binomial?

(d) Let's say $X_1, \ldots, X_{100} \overset{iid}{\sim} \text{Bernoulli}\,(40\%)$. Calculate the probability of 39 successes out of 100 using the normal approximation. Compare it to the exact probability computed via the binomial PMF. This is *really* hard and it's not extra credit because I want you all to attempt it. Do this one at the end of the assignment.

Control Theory These questions will get your feet wet with the concepts in chapter 14.

Problem 9 We will learn about the two types of errors.

(a) What is the symbol we use for Type I errors? Make a probability statement using this symbol.

(b) Let's say you have a fire alarm in the building. The fire alarm can be "off" or "on" and there's either a "fire" or "no fire". Draw a 2×2 matrix like the one in class (table 14.2 on p.332) and indicate clearly what the row axis represents, what the column axis represents, and the labels for the rows and columns.

(c) Explain what a Type I error would be in the building / fire example without using technical jargon.

(d) What is the cost of a Type I error?

(e) Explain what a Type II error would be in the building / fire example without using technical jargon.

(f) What is the cost of a Type II error?

(g) You are the casino manager at Harrah's in Las Vegas. You are concerned about your Roulette wheels landing on black too often. You want to use the tools of ch 14 to track this potential problem. "Shutdown" would be shutting down the roulette tables. In fact, you want to duplicate exactly the statistical methodology that the book uses to look at computer chip HALT scores. Draw a mock control chart. Indicates axes and make up control limits. Do not do any calculations. This is purely a conceptual and visual problem.

(h) How can the casino manager make a Type I error?

(i) How can the casino manager make a Type II error?