Statistics 101 Summer I 2011
Homework #4

Adam Kapelner, Instructor

Due noon, Friday, June 24, 2011 (in my mail slot)

Instructions and Philosophy

Copy the previous first paragraph and tack on the following: reviewing lecture notes from class is really important.

In Stine & Foster, read chapter 15, 16 carefully, then chapter 5 (ignore the section on Cramer's V), then chapter 18. You can ignore the subsections on skewness and kurtosis and the SRS condition with n being 10% of N (since we view $N \approx \infty$ anyway). I also recommend reading Wikipedia's article on Simpson's Paradox.

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 Monday, June 27 at 9AM, 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).

Distribution Theory In this section we will learn about the most important distributions used for testing in the field of Statistics.

Problem 1 We're going to learn more about the CLT and its cousin, the T distribution. We will use the notation T_{df} for a T distribution with df degrees of freedom and the usual $Z = \mathcal{N}(0, 1)$. The notation $\xrightarrow{\mathcal{D}}$ means as $\lim_{n\to\infty}$ the PDF's become equal; the notation $\overset{d}{=}$ means the PDF's are exactly equal right now *i.e.* at any n. S_n will refer to the r.v. that is the sum of $X_1, \ldots, X_n$ as usual, but $\mathbb{S}$ is the r.v. that represents the distribution that s, the sample standard deviation, is drawn from. In other words:

$$\mathbb{S} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

This distribution is immensely complicated and non-standard. It would take PhD's a good few days of computations and algebra to familiarize themselves with its properties.

(a) Show the following by invoking the CLT from class and using shifts and scales:

$$X_1, \ldots, X_n \overset{iid}{\sim} \text{(any PMF or PDF with mean } \mu \text{ and std dev } \sigma\text{)}, \quad \frac{\sqrt{n}\left(\bar{X} - \mu\right)}{\sigma} \xrightarrow{\mathcal{D}} Z$$

(b) Show the following by invoking the CLT from class and using shifts and scales:

$$X_1, \ldots, X_n \overset{iid}{\sim} \text{(any PMF or PDF with mean } \mu \text{ and std dev } \sigma\text{)}, \quad \frac{(S_n - n\mu)}{\sqrt{n}\sigma} \xrightarrow{\mathcal{D}} Z$$

(c) Explain why the following is *not* true:

$$X_1, \ldots, X_n \overset{iid}{\sim} \text{(any PMF or PDF with mean } \mu \text{ and std dev } \sigma\text{)} \quad \Rightarrow \quad \frac{\left(\bar{X} - \mu\right)}{\frac{\sigma}{\sqrt{n}}} \overset{d}{=} Z$$

(d) Explain why the following is *not* true. Be sure to look in your notes; there are two things wrong here.

$$X_1, \ldots, X_n \overset{iid}{\sim} \text{(any PMF or PDF with mean } \mu \text{ and std dev } \sigma\text{)} \quad \Rightarrow \quad \frac{\left(\bar{X} - \mu\right)}{\frac{\sigma}{\sqrt{n}}} \overset{d}{=} T_{n-1}$$

(e) Explain why the following is *not* true. Be sure to look in your notes.

$$X_1, \ldots, X_n \overset{iid}{\sim} \text{(any PMF or PDF with mean } \mu \text{ and std dev } \sigma\text{)} \quad \Rightarrow \quad \frac{\left(\bar{X} - \mu\right)}{\frac{\mathbb{S}}{\sqrt{n}}} \overset{d}{=} T_{n-1}$$

(f) Explain why the following *is* true:

$$X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right) \quad \Rightarrow \quad \frac{\left(\bar{X} - \mu\right)}{\frac{\sigma}{\sqrt{n}}} \overset{d}{=} Z$$

(g) Is the following true?

$$X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right) \quad \Rightarrow \quad \frac{\left(\bar{X} - \mu\right)}{\frac{\mathbb{S}}{\sqrt{n}}} \overset{d}{=} T_{n-1}$$

What is the critical premise most people leave out (even S&F) when claiming their test statistic is T-distributed?

Problem 2 We talk a lot about how the T_{df} converges to a Z as the degrees of freedom (df) gets larger. The ballpark approximation we gave in class was that $df \geq 29$, we call it a day. We're going to see just how close this approximation is.

(a) Run the following code in `R`:

```
x = seq(-6, 6, 0.01);
plot(x, dt(x, 1),
  type = "l",
  col = "chocolate",
  ylim = c(0, 0.39),
  ylab = "prob density",
  xlab = "t (or z)",
  main = "The T distribution as an approximation\nto the Z distribution");
lines(x, dt(x, 5), type = "l", col = "blue");
lines(x, dt(x, 10), type = "l", col = "green");
lines(x, dt(x, 20), type = "l", col = "yellow");
lines(x, dt(x, 30), type = "l", col = "brown");
lines(x, dt(x, 50), type = "l", col = "purple");
lines(x, dt(x, 100), type = "l", col = "gray");
lines(x, dnorm(x), type = "l", col = "red");
#placeholder for last line
```

The code will graph the PDF of T_1 in "chocolate" color, the PDF of T_5 in blue, the PDF of T_{10} green, the PDF of T_{20} in yellow, the PDF of T_{30} in brown, the PDF of T_{50} in purple, the PDF of T_{100} in gray, and the PDF of the $Z = \mathcal{N}(0, 1)$ in red.

What happens as the degrees of freedom gets larger?

(b) Now we're going to run the same code again, but this time we're only going to plot the right tail *i.e.* the portion from $z = 1.96$ (your favorite number) and greater:

```
x = seq(1.96, 3, 0.001);
plot(x, dt(x, 1),
  pch = ".",
  col = "chocolate",
  ylim = c(0, 0.07), xlim = c(1.96, 3),
  ylab = "prob density",
  xlab = "t (or z)",
  main = "The T distribution as an approximation\nto the Z distribution");
lines(x, dt(x, 5), pch = ".", col = "blue");
lines(x, dt(x, 10), pch = ".", col = "green");
lines(x, dt(x, 20), pch = ".", col = "yellow");
lines(x, dt(x, 30), pch = ".", col = "brown");
lines(x, dt(x, 50), pch = ".", col = "purple");
lines(x, dt(x, 100), pch = ".", col = "gray");
lines(x, dnorm(x), pch = ".", col = "red");
#placeholder for last line
```

Maximize the plot window (make the graphic as big as you can). Estimate what percent our approximation is off when we use $z_{0.025}$ instead of $t_{0.025,30}$ by looking at the plot. Is this large enough to care under quotidian circumstances?

(c) Estimate what percent the approximation would be off when you use $z_{0.025}$ instead of $t_{0.025,100}$ by looking at the plot.

(d) In one of the lectures, I wrote on the board that the limit of a T as its *df* (which we will denote ν for this problem) increases is a Z:

$$\lim_{\nu\to\infty} T_\nu = Z$$

For persnikety mathematicians, this is an inexact statement because limits aren't defined for r.v.'s, only for their densities.[1] The following is the density function (the PDF) for the T distribution with ν degrees of freedom:

$$f_{T_\nu}(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}$$

Prove that the density of the T distribution (above) converges to the density of the Z distribution (which is in your notes) for all values of x. Proving this will constitute a more mathematically precise statement and will corroborate the picture you see in parts (a) and (b).

I will give you two hints. One: assume the following is true without proof:

$$\lim_{\nu\to\infty} \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu}\,\Gamma(\frac{\nu}{2})} = \frac{1}{\sqrt{2}}$$

If I didn't give you the above, you would've needed tons of tricks and intimate knowledge of the gamma function (which I explained briefly at the end of one of the lectures) to prove and it's a couple pages of algebra. Two: review the tricks we used during the derivation of the Poisson PMF from the Binomial PMF which is in your notes.

(e) Why is the t-test called "Student's t-test"? Tell the story of the T distribution.

Problem 3 This problem will ask basic questions about the χ^2_{df} distribution in order to familiarize you with it.

(a) Write the definition of the χ^2_{df} r.v. as we did in class.

[1] This is also inexact, but to find out why, you're going to have to enroll in the PhD program.

(b) Use the table from the book (p.743) which I handed out in class to approximate how much density there is in the χ^2_4 PDF when $x \geq 8.2$.

(c) As the degrees of freedom gets larger and larger, what does the χ^2 distribution converge to? Hint: write out your answer to part (a) for *df* large and then look at the theory you invoked for your answers in problem #1.

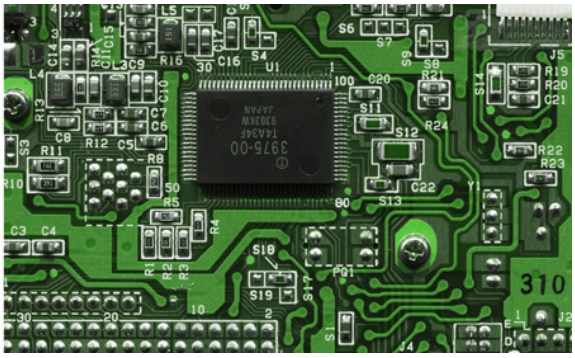
(d) If $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, then what is the distribution of $(n-1)\frac{\mathbb{S}^2}{\sigma^2}$? Can you use this fact for a statistical test? Write a paragraph about this test.

(e) Write a paragraph about the history of the χ^2_{df} distribution and make sure to tell me how it got its name.

Confidence Intervals (CIs) This section is all about CIs. The goal is to learn how to build CIs and understand their limitations.

Problem 4 This will be a more interactive `R` example (as opposed to the usual passive examples). We return to our discussion of GPS chips from chapter 14 (you may want to reread pp. 325-330). We take a simple random sample (SRS) of 40 GPS chips from the assembly line per day and do a HALT sequence on each to measure their HALT scores. Consider each of the samples as independent and identically distributed samples each distributed like:

$$X_1, \ldots, X_n \overset{iid}{\sim} \text{an unknown PDF with mean } \mu = 7 \text{ and standard deviation } \sigma = 4$$



(a) If we were to build a CI for the true mean HALT score of chips in production that day, should we use the T distribution or the Z distribution to compute the margin of error? Use the tree diagram we did in class for univariate data testing. If we can't do anything, explain why.

(b) Draw the distribution of the average HALT score of the SRS to scale as best you can. Indicate where μ is. Calculate the standard error. Label the x-axis with numbers to the second decimal place. Indicate the name of the r.v. whose density you are sketching and how it's distributed. Leave the entire page below your sketch blank as we will be using it later in the question.

(c) If the chips have a different HALT score than 7, then the plant has a big problem and they should investigate. The cost of not finding a problem is much higher than the cost of shutting down and investigating. Frame this statement in terms of Type I and II errors and explain why α should be picked high or low.

(d) Regardless of the α you picked, let $\alpha = 10\%$ since we want to be conservative. Indicate the cutoff point(s) for $\bar{x}$ and lightly shade in the "rejection" region in the sketch in part (b).

(e) Now, we will be simulating a day's sample. We can do this via the following code:

```
samp = rnorm(40, 7, 4);
samp;
#end placeholder
```

Copy and past this into `R`. It should show you the 50 measurements in the sample on the screen. Now, run the following code to get the average of the sample:

```
mean(samp);
#end placeholder
```

The number on the screen is $\bar{x}$. Find the 90% confidence margin of error. Write this number down separately as it will be useful for later.

(f) What is the $\mathbb{P}(\mu \in \text{CI}_{\mu,90\%})$ where $\text{CI}_{\mu,90\%}$ is the confidence interval from part (e) which you did not have to compute.

(g) Before we do another day's SRS and build another confidence interval, what is the $\mathbb{P}(\mu \in \text{CI}_{\mu,90\%})$? In this case, $\text{CI}_{\mu,90\%}$ refers to the *future* confidence interval from tomorrow's experiment, not the one we built in part (e).

(h) Now, we're going to get samples for many days. Since we only care about inference for μ, we're going to trash the actual data[2] and only look at its average, $\bar{x}$. Run the following code 25 times in a row:[3]

```
mean(rnorm(40, 7, 4));
#end placeholder
```

For each draw (or realization) of $\bar{x}$ from $\bar{X}$, build a 90% CI. Draw each of these 25 CI's using a line segment below the sketch from part (a).

[2] If you're interested in why we can just "trash" the rest of the data, this is because $\bar{x}$ is a *sufficient statistic* for μ, something you'll learn about when you take Stat 434 with Professor Ewens.

[3] use the $\uparrow$ key on your keyboard in the `R` console and press enter to make this go faster.

(i) Now extend the line of μ and count the number of CI's that do *not* include μ. This is a duplication of the graphic that appears on p.361 of S&F. Does the count of these unlucky CI's make sense given our α level?

(j) Calculate the probability you take 25 SRS's and have this many CI's not capture μ.

(k) Consider the more likely case that we do not know the true standard deviation σ. Can we build a confidence interval for the true mean HALT score?

(l) Assume HALT scores are normally distributed. Take another SRS and compute $\bar{x}$ and s by running the following code to get $\bar{x}$ and s:

```
samp = rnorm(40, 7, 4);
mean(samp);
sd(samp);
#end placeholder
```

Build a 99% CI for μ for the sample above and notate it like we do in class.

Problem 5 We take a survey to see if people will buy a piece of software. The product costs \$10. We do many different surveys. They all consist of asking n users who come across the homepage and ask them if they would buy the service:

Survey 1 :	$n = 200$,	36 people say yes
Survey 2 :	$n = 400$,	67 people say yes
Survey 3 :	$n = 800$,	121 people say yes
Survey 4 :	$n = 1600$,	265 people say yes
Survey 5 :	$n = 3200$,	501 people say yes

(a) Are any of the above good surveys if our goal is to make inference about the proportion of people who will buy the product online?

(b) Build a 95% CI for the proportion of people who will buy this product when it is released using each of the above surveys.

(c) If we're comfortable with this margin of error on survey 1, allow the margin of error to remain constant for surveys 2, 3, 4, and 5. What will the coverage probabilities be for each survey?

(d) Can we build a 95% CI for the number of people that will buy the product when it is released? Explain why or why not.

(e) Can we build a 95% CI for the amount of revenue the company will make when the product is released? Explain why or why not.

(f) Using survey 3, provide a confidence interval for the *number of people* who subscribe for one month if 100,000 people visit the homepage in one year.

(g) Using survey 3 and the previous question, provide a confidence interval for the year's revenue in euros. Assume everyone subscribes and pays up front for one month. Use the exchange rate \$1 = €0.70.

(h) Last year the company had revenues of \$120,000. Create a 95% CI for the *percentage change* in revenue. You may want to reread pp. 362-364.

One-proportion and one-sample hypothesis testing These questions are going to familiarize you with the methodology of univariate hypothesis testing and introduce you to a basic power calculation.

Problem 6 You are running the craps tables at the casino.

There is someone suspiciously winning. He seems to hardly ever "crap out" *i.e.* roll a seven on two dice (the sum of two dice is 7). From table records you see he has rolled 113 rounds (rolls of both dice simultaneously) and only crapped out 13 times. He then left the casino with a large bundle of winnings and you face the decision whether or not to let him back in next time.

(a) Frame this as a hypothesis test. Write down H_0 and H_a. Write it in English first then write the mathematical statements for the hypotheses.

(b) Talk about what a Type I error is and what a type II error is in this context.

(c) What would the costs of each error be?

(d) You decide to take a middle-of-the-road approach and take an $\alpha = 5\%$. Run the hypothesis test and calculate the p_{val}.

(e) What is your conclusion? Is the result statistically significant?

(f) What is your strategy for the future when you see him playing at the craps tables again?

(g) If this guy really knows how to set the dice and actually can set so that the proportion of 7's is 11% which is for an expert dice setter, he can really cheat the casino out of a lot of money at his gambling rate. Set this scenario up as a hypothesis test but do not do any computations.

(h) Calculate the power of this test *i.e.* the probability of detecting whether this guy is an expert dice setter.

(i) How many rounds do you have to observe this guy before you can detect that he's an expert dice setter with probability 80%?[4]

Problem 7 According to the official Collegeboard 2010 report, the mean math score is 516 with SD 112 (see table 1, page 1). Download `SAT.jmp` which is all the SAT data for the incoming freshmen class of 2008 anonymized.

(a) Analyze the distribution of the math scores. Are they normally distributed? Print a Q-Q plot and attach it to your homework.

(b) It looks like there's a black vertical line in the Q-Q plot. Why is it there? What does it correspond to?

(c) JMP displays a diamond in the box and whisker plot to indicate a 95% CI for the mean. Is JMP's CI exact in this case? Explain.

[4]For some reason 80% is the standard power-level statisticians use; it is analagous to the standard $\alpha = 5\%$.

(d) SAT math scores are known to be normally distributed. Are the incoming Freshmen class of 2008 an SRS of all students that took the SAT's? Base your answer on the analysis of distribution on the screen in JMP.

(e) Test that the incoming students have a higher than average math score. Compute a p_{val}. Is this lower than you expected? Think about what the p_{val} means: is the null hypothesis ridiculous in this context?

(f) We now examine the $n = 43$ homeschooled students of that incoming class. Their SAT math scores are below:

```
760 560 690 800 510 800 670 540 730 680 550 800 640 770 420 680 530
730 560 430 630 560 670 600 800 800 710 530 710 390 660 750 440 610
660 400 560 800 770 790 330 720 380
```

We have reason to suggest that they come from a normal distribution but *not* the same distribution as what is published in the SAT 2010 report above. We are interested in testing whether or not they have the same math skills as the average incoming student (as measured by the SAT math examination). Set up a hypothesis test.

(g) Use JMP to compute a 95% CI for the mean of Penn-admitted homeschooled student math scores.

(h) Carry out the hypothesis test at $\alpha = 5\%$. Use the p_{val} criterion for acceptance / rejection. Indicate whether or not your result is statistically significant.

(i) Can you reach the same conclusion you did in part (h) from looking at the CI from part (g)? Why or why not?

Contingency tables This section will introduce you to contingency tables, testing using the χ^2 test for independence, and the concept of lurking variables and Simpson's paradox.

Problem 8 We talked about the Berkeley data graduate school admissions data in class. Download the file `berkeley.JMP` and open it. The data I presented in class was from Wikipedia, this data is from another source, so the answers will be a bit different.

(a) Define "observational study".

(b) Define "experiment".

(c) Is this Berkeley data the result of an observational study or an experiment and why?

(d) Use JMP to build a mosaic plot of admission on gender as well as a contingency table (use the "Fit Y by X" analysis option). Print both the mosaic plot and the contingency plot and include it with your homework.

(e) Estimate the probability of being admitted if you're a female. Notate A for "admitted" and F for "female". Verify your answer with the correct percentage on the contingency table and circle the number(s) on the printout.

(f) Let X be the r.v. that someone is admitted regardless of gender. How do you model X? Which special r.v. do you use? Draw a box around the number(s) on the table you're using to come up with the parameter estimate for such a model.

(g) We are interested in whether or not gender and the admissions decision are associated. Frame this as a hypothesis test.

(h) Execute this test for independence at an $\alpha = 1\%$ confidence level. This will involve building a table of expected counts, calculating a χ^2 test statistic, looking up the correct critical value, making a decision, and writing a sentence of discussion. Your χ^2 statistic should match the "Pearson" "ChiSquare" in JMP's "Tests" panel (don't worry about rounding errors).

(i) Is it fair to conclude from the previous question that gender and admission decision are associated? Is it fair to conclude that being female or male actually *causes* a prospective candidate to have a larger or smaller chance of being accepted to graduate school at Berkeley?

(j) Now build a table whose columns are gender and rows are choice of major and whose cells are percentage of applicants admitted. I don't believe that this is an option in JMP (or at least I couldn't figure it out in a reasonable amount of time). I recommend doing a "Fit Y by X" where Y is `admitted?` and X is the interaction `gender*major` and using the percentages in the contingency table (you will have to figure out which of the three percentages to use).

(k) Write down the definition of "lurking variable". Does major appear to be a lurking variable?

(l) Write down the definition of "Simpson's Paradox". Is the direction of association changing here?

(m) Write down the definition of "confounding".

(n) Is it possible that there's *another* lurking variable that we haven't controlled for that would flip the direction of association after you control for it in addition to the previous controls?

Two-proportion and two-sample hypothesis testing These questions are going to familiarize you with the methodology of bivariate hypothesis testing: testing interval samples crossed with a nominal variable of two categories.

Problem 9 We return to the craps table for another round of gambling. This time, two dice setters are at odds with each other both claiming that they are better than the other. We do an experiment. The first dice setter rolls 224 times and gets 24 sevens; the second dice setter rolls 214 times gets 28 sevens.

(a) As a disinterested observer, we want to investigate if there is any statistically significant difference between the two dice setters. Frame this as a hypothesis test.

(b) Run the test at $\alpha = 5\%$, calculate a p_{val}, make a decision, and interpret your results.

Problem 10 We return to the SAT data from the collegeboard report as well as the SAT data from UPenn admissions.

(a) We want to test whether or not men perform better on the critical reading section of the SAT. Frame this as a hypothesis test.

(b) Run the test at $\alpha = 5\%$. Use the data from p.1 of the report (*i.e.* the page after the table of contents) and assume SAT scores are normally distributed.

(c) Open `sat.JMP` again. This time, we want to look at the difference between admitted students' verbal and math scores. In order to do this, we need to combine the `math` and `verbal` variables into one variable. Use the `Tables...Stack` option in the menu. Stack both the `math` and `verbal` columns and call the "Stacked Data Column" the name "score" and the "Source Label Column" the name "mathorverbal". Upon executing the stack, this should spring up a new data table. Now we can analyze the math vs. verbal side-by-side. Go to "Fit Y by X". Enter for Y the variable `score` and for X the variable `mathorverbal`.

You should now see the bivariate plot with the nominal data on the x-axis and interval data on the y-axis. Click on the red arrow and click "Quantiles" to see the box and whisker plot for both the `math` and `verbal` datasets. Print this out and attach it to your homework.

(d) Consider both the `math` and `verbal` scores as coming from a normal distribution (even though this seems to be incorrect). Test the difference of these two scores at $\alpha = 1\%$.

(e) Click the red arrow and click "t Test". Does your answer from the previous part match JMP's answer?