Homework #4: Intro to inference
Due: In lab, Wednesday, Feb 25-28

DIRECTIONS: For this homework, please turn in hard copies in class on Wed, Feb 13. This homework counts double. Write answers out in a separate Microsoft Word document (note that some questions may not require an answer – so you can skip these or just say "completed" – but we'll try to put what the actual question to be answered, or what you need to do to get credit, by placing that in **bold** in the question). We encourage you to work in groups for R-related material, but make sure to do the work yourself too and *always* come up with independent answers. Homeworks that have identical answers will receive F's.

Question 1:
You are interested in whether people smoke more marijuana in communities where medical marijuana is legally sold. The population mean number of marijuana cigarettes ("joints") or equivalent smoked per person per year is 5 in the nation, but this number is highly skewed (most people smoke 0, but some smoke many hundreds of joints). The population standard deviation of joints smoked per year per person is 8.
A) What percentage of individuals in the US smoke more than 21 joints or equivalent per year? If you cannot answer this, explain why.

B) The distribution of number of marijuana cigarettes smokes is highly skewed. Let us say we took "a lot" of samples of size 400 from the US population, collected data on how much marijuana each person smoked, and figured the mean of number of joints smoked per year. **Describe what you would call this distribution, what its shape would be, what its mean would be, and what its standard deviation would be. What is the name of this standard deviation?**

C) Given your answer in (B), and assuming that we did indeed draw from the entire US population, **what is the probability of observing a mean (figured from 400 individuals) as far or further than .4 away from the population mean of 5 (e.g., less than 4.6 or more than 5.4)?**

D) What about as far or further than .8 away from the population mean?

E) We collect data on marijuana smoking among 400 individuals who live in communities across the US where medical marijuana is legal. We find that the mean number of joints per year in this sample is 5.2. **Describe the null and alternative hypothesis here, compute the inferential z-statistic for this finding, report a p-value, and decide whether to reject or not reject the null hypothesis.** Remember, the inferential z-statistic is:

$$z = \frac{\bar{X} - \mu}{SEM} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

F) The "Cohen's *d*" is a measure of "effect size" – it tells us how far above or below the population mean an observed mean is in terms of the ORIGINAL standard deviation (not in terms of the standard error of the mean). An effect size is an important concept in this class – a sample mean can be highly significantly different from a population mean (have a really small p-

value) even if the difference between the means is trivial in practical terms. For example, a huge (n=10,000) sample might allow us to know that the average extraversion scores among college-students is higher than non-college students of the same age, but the difference in real terms might be tiny (.001). All a Cohen's *d* does is standardize that .001 difference in terms of the original standard deviation (much as we've been doing with z-scores all along). On the other hand, we might have a non-significant difference in a small sample (n=10) that nevertheless has a very large effect size (say a difference of .2). Cohen's d allows us to disambiguate the SIZE of an effect from its STATISTICAL SIGNIFICANCE – two very different things! **What is the Cohen's *d* for the result in (E) above?** Remember, the Cohen's *d* is:

Cohen's $d = \frac{\bar{X} - \mu}{\sigma}$

(Note the difference between this formula and the one in (E) above!)

G) Your research assistant made a mistake in reporting the study to you. Instead of collecting information on 400 people, you actually collected information on 10,000 people who live in communities where medical marijuana is legal (everything else is identical). **Refigure (E) and (F) above using the new sample size of 10,000.**

H) For your answers above, explain intuitively why your inferential z-value, p-value, and decision to reject the null hypothesis changed between (E) and (G). Explain intuitively why your effect size estimate (Cohen's *d* statistic) did not change.

I) Why do we use a "Cohen's *d*" effect size estimate rather than simply reporting inferential z-statistic for a finding? What does a Cohen's *d* tell you that is different than what an inferential z-statistic does?

2) Describe what the statistical power for a test is. Name three factors that increase the statistical power for a test, and explain in words grandma can understand why they have this effect.

3) Load the datasets that I have stored on an external server. To do that, use the command below:
load(url("http://www.matthewckeller.com/Stats3101/Stat3101.Datasets-2013.RData"))

In the first lab, you completed a survey on neuroticism. It is a tendency to experience negative emotional states. Individuals who score high on neuroticism tend to experience feelings such as anxiety, anger, guilt, and depressed mood. **We are interested in understanding whether the students who take this course (PSY 3101) tend to have higher or lower neuroticism scores than the national average. The population mean level of neuroticism on this scale is 2.7 and the population SD is .6. Use R to answer this question, and report your results in a 4-sentence summary below.** A 4-sentence summary should have an *Intro sentence*, a *Methods sentence*, a *Results sentence* (with relevant descriptive and inferential statistics reported in parentheses), and a *Discussion/Conclusion sentence*.

R help for question 3:

After loading the data (above), make "survey2013" the active dataset. Then find the sample mean for the variable "neuroticism" by going through the Menu (Statistics -> Summaries -> Numerical summaries, choose "neuroticism") or by using syntax:

```
mean(survey2013$neuroticism, na.rm=TRUE)
```

Once you have the sample mean, you should be able to use the formula from Question 1E above to find the inferential z-statistic, given that you know the population ("null") mean and the population standard deviation. Once you have your inferential z-statistic, you need to know the proportion of scores above/below a score that extreme assuming normality.You can look up the proportion of scores below or above a given z-score – assuming the distribution is normal – using R. You can assume the distribution of sample means of sample size *n=48* would be normally distributed (right?). So you can look up whatever z-score you find in R using the Menu (Distributions -> Continuous distributions -> Normal distribution -> Normal probabilities -> then enter your z-score and make sure "lower tail" is highlighted if your z-score is negative). Or using the syntax below (assuming you get a negative z-score, which you should!):

```
pnorm(<your z-score>,lower.tail=TRUE)
```

Just a bit of digression on using R to find normal probabilities. It's actually a lot easier to use R than the book to find the upper or lower tail of the distribution. If, for example, we want the proportion of scores above a z-score of 1 for normally distributed data:

```
pnorm(1,lower.tail=FALSE)
```

If we want the proportion of scores below a z-score of 1 for normally distributed data:

```
pnorm(1,lower.tail=TRUE)
```

(Note that pnorm(1,lower.tail=FALSE) + pnorm(1,lower.tail=TRUE) = 1 always)

Similarly, if we want the proportion of scores above a z-score of -1 and a z-score below a score of -1 for normally distributed data

```
pnorm(-1,lower.tail=FALSE)
pnorm(-1,lower.tail=TRUE)
```

The "p-value" for a normally distributed variable is asking the question, "what is the probability of getting a z-score this extreme or more extreme if the population mean is really equal to 2.8"? THINK before you act! You really need to be careful about whether you ask R for the lower tail or upper tail to answer this question. DRAW the normal distribution to help you. Thus, the proportion of scores further than a z-score of 1 away from the mean for normally distributed data is:

```
pnorm(1,lower.tail=FALSE) + pnorm(-1,lower.tail=TRUE)
```

The first one gives you the proportion of scores above 1, and the last command the proportion of scores below -1. It is important that you realize that the syntax above also gives you the "p-value" for a z-score of 1 – i.e., 31.7% of scores are as far or further than a z-score of 1 away from

the mean (of 0). You knew this already (.16*2 = .32), but we need to use R to find p-values for any possible z-score you might get.

THUS, to find YOUR p-value, you would do this
pnorm(<positive absolute z-score you observed>,lower.tail=FALSE) + pnorm(<negative absolute z-score you observed>,lower.tail=TRUE)

For example, assuming you get an observed z-score of -1.5:
pnorm(1.5,lower.tail=FALSE) + pnorm(-1.5,lower.tail=TRUE)
is your "p-value" – i.e., the proportion of times you would get a mean THIS FAR OR FURTHER from the population mean if we really are drawing samples from that population. Make sure you understand why the above works by drawing out the normal distribution. P-values for a normal distribution are trying to get the sum of BOTH TAILS of the distribution.

Note that the "magical" p-value cut-off in science is typically .05. P-values lower than this are deemed "significant" – i.e., unlikely to have arisen by chance sampling from the null distribution. In such cases, we "reject" the null hypothesis and conclude that the null is unlikely to be true. Use this cut-off in your answer to determine if you have "significant" findings.

4-sentence summary help for question 3:
We need to be able to quickly convey our findings to the scientific community. This is typically done in scientific abstracts. From here on in this class, including on almost all HW's and tests, we'll get practice in writing "4-sentence summaries" of our findings, which is really the framework for all scientific abstracts.

SENTENCE 1 (INTRO): This is an introductory sentence, explaining what it is that we're looking at and (potentially) why. E.g., "Several studies have found that males have lower average conscientiousness scores than females." OR "We were interested in understanding whether males or females had higher average consciousness scores in college-aged populations".

SENTENCE 2 (METHODS): This tells the reader HOW you went about trying to answer the question you introduced in the first sentence. E.g., "To understand this, we asked 48 students at CU to take a 5-item scale measuring consciousness" OR "To investigate this, we randomly selected 50 males and 50 females from the general population and had them take a 10-item questionnaire on consciousness".

SENTENCE 3 (RESULTS): This is the meat of your summary! Here, you tell the reader what you found. We want this to be able to be read by grandma and still understand it, so put it in plain English, placing the statistics in parentheses so that people who understand the stats can see exactly what you found. E.g., "We found that the average score for male conscientiousness (mean=3.2, sd=.5) was lower than female conscientiousness (mean=3.8, sd=.6), and this difference was very unlikely to be due to chance (z=3.23, p=.0012)."

SENTENCE 4 (DISCUSSION): Here, we give a "so what" sentence, telling the reader our conclusions. E.g., "We conclude that males from this population have lower conscientiousness scores, although the reason this may be remains unclear." or something along those lines.