Homework 2

Due Monday July 16, 2012 *(before class starts!)*

This homework looks long, but question 1 and 2 shouldn't take you too long. They are conceptual questions, designed to get you to think about CIs and hypothesis testing. Question 4 is a "mechanical" question, designed to help you to derive CIs and obtain p-values. Overall, I don't expect you to take more than four hours on this homework.

1. These questions are designed to test your conceptual knowledge of CIs. No math required, *seriously.*

   (a) Fill in the following table. Here, $S = \hat{\sigma}$ and $S^2 = \hat{\sigma}^2$

CI	Confidence Level	Parameter(s) Covered	Population normal?	Variance known?
$\bar{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$	$1-\alpha$	μ, popu. mean	Yes and No (for large n)	Yes
$\bar{X} \pm z_{(0.95)}\frac{\sigma}{\sqrt{n}}$				
$\bar{X} \pm z_{(0.80)}\frac{S}{\sqrt{n}}$				
$\bar{X} \pm t_{(0.975),n-1}\frac{S}{\sqrt{n}}$				
$\left(\frac{(n-1)S^2}{\chi^2_{(0.01),n-1}}, \frac{(n-1)S^2}{\chi^2_{(0.99),n-1}}\right)$				
$(-\infty, \infty)$				
$(0, \infty)$		σ^2, popu. variance		

   (b) Suppose you already computed the 95% CI for a parameter. Can we determine whether the parameter is in the CI? Briefly explain your response.

   (c) Suppose we wanted to study the average balance, μ, of checking accounts for PNC customers. We sample 100 accounts randomly and find the 95% confidence interval to be [$1,000,$2,500]. Which of the following statements about this 95% CI is correct? Briefly explain your reasoning

      i. The average balance, μ, of checking accounts is in between $1,000 and $2,500

      ii. 95% of all PNC customers keep a balance between $1,000 and $2,500 in their checking accounts

      iii. We have 95% confidence that μ lies within $1,000 and $2,500

      iv. The sample mean of 95% of the 100 individuals we sampled has a balance between between $1,000 and $2,500 in their checking accounts

      v. Every new sample of 100 accounts and the procedure to obtain the CI from the sample will generate a CI that will cover μ 95% of the time.

(d) Suppose we have $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, $i = 1, ..., n$, and we have two α-CIs,

$$\text{CI 1: } \bar{X} \pm Z_{(1-\alpha/2)} \frac{\sigma}{\sqrt{n}}$$
$$\text{CI 2: } \bar{X} \pm t_{(1-\alpha/2),n-1} \frac{\sigma}{\sqrt{n}}$$

Which CI has the largest interval? Which interval is preferred over the other? Briefly explain your response.

2. These questions are conceptual questions for hypothesis testing. No math required, *seriously.*

(a) Suppose our testing procedure decides to reject the null. Does that mean that the null is actually false?

(b) (TRUE/FALSE) P-value is the probability of H_0 being incorrect. Explain your reasoning.

(c) (True/False) P-value is a random variable. Explain your reasoning
Hint: This is a very tricky question and you must explain very carefully. Here are a couple of questions to get you to think about this question. Is there some sort of randomness involved in our procedure? Does this impact how to accept the null or the alternative (i.e. p-value $\leq \alpha$)? Can you assign a probability to $P($ p-value $\leq \alpha)$?

(d) Suppose I collected data $(X_1, ..., X_n)$ where $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ and I want to test

$$H_0 : \mu = 0 \ vs. \ H_a : \mu \neq 0$$

Let $\bar{X}_{obs} = 3$ and $\hat{\sigma} = 1.2$. Suppose I knew the actual variance for this population, say $\sigma^2 = 1$. But, I decided to use $\frac{\bar{X}-\mu_0}{\hat{\sigma}/\sqrt{n}}$ instead of $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$ as my sampling distribution. How would the p-values from both sampling distributions compare?

(e) Suppose I had a testing procedure where I always rejected the null, regardless of what my sample said. What are the Type I, Type II, and power for this procedure? What α achieves this procedure ?

(f) Suppose I had a testing procedure where I always retained the null, regardless of what my sample said. What are the Type I, Type II, and power for this procedure? What α achieves this procedure?

3. (Bonus question) Find the probability of the following for $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, $i = 1, ..., n$.

$$P(\text{ p-value } \leq \alpha) \tag{1}$$

where α is the significance level.

4. These are "mechanical" questions on CIs and hypothesis testing. That is, they are designed to drill you on how to derive confidence intervals for a variety of situations. By the end of these series of (tedious) questions, I hope you can construct CIs as easily as adding two binary numbers. FYI: All these questions require mathematical justification! Simply writing down the formula will get you no points!

(a) Please download the "PennSexSurvey.txt" dataset and import it into R. Read the description about the Penn Sex Survey that was done a couple years ago on the course website.

(b) Suppose we are interested in the average BMI at Penn, denoted as μ. We'll assume, for a moment, that $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ where X_i is the height of the ith male in the data set.

i. How many males are in the data? How many females are in the data?

ii. What is the sample average BMI of Penn students? What is the sample average BMI of males at Penn? How about females?

iii. Derive the $1 - \alpha$ confidence interval for the population mean. Assume that the variance, σ^2, is *known*.

iv. Derive the $1 - \alpha$ confidence interval for the population mean. Assume that the variance, σ^2, is *unknown*. What is the 99% CI for mean BMI amongst males? Amongst females?

v. Derive the $1 - \alpha$ confidence interval for the population variance. Assume that we don't know what μ is. What is the 90% CI for variance of BMI amongst males? Amongst females?

(c) Now assume that X_i comes from any arbitrary population F with mean μ and variance σ^2. As before, we are interested in the average BMI at Penn. Derive the $1-\alpha$ confidence interval for the population mean. Assume that the variance, σ^2, is known. What is the 95% CI for mean BMI amongst all Penn students?

Note: You must use the derivation procedure outlined in homework 1.5. Simply stating the formula will get you no points.

(d) Suppose you want to test that the average BMI of Penn students is less than 25 (i.e. Penn students are not obese). Set up the hypotheses and test your hypotheses against an $\alpha = 0.05$. In particular, do it for the case where

i. The population is not Normal and the variance is unknown

ii. The population is Normal and the variance is unknown

(e) Suppose you want to test the hypothesis that female's BMI is lower than male's BMI. Set up the hypotheses and test your hypothesis against an $\alpha = 0.05$. In particular, do it for the case where...

i. the population is Normal and the variance of female's BMI is the same as the variance of male's BMI, but unknown.

ii. the population is Normal and the variance of female's BMI is different than the variance of male's BMI. Assume both variances are unknown

Note: Here, simply writing down the formula is sufficient

iii. the population is not Normal and the variance of female's BMI is the same as the variance of male's BMI, but unknown.

iv. the population is not Normal and the variance of female's BMI is different than the variance of male's BMI. Assume both variances are unknown