14.30 Introduction to Statistical Methods in Economics
Spring 2009

Problem Set #7

14.30 - Intro. to Statistical Methods in Economics

Instructor: Konrad Menzel

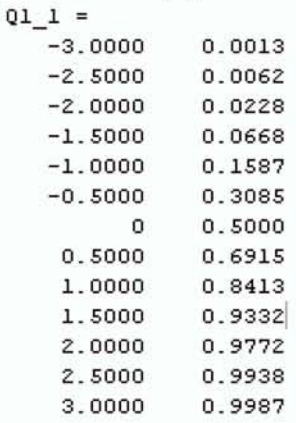
Due: Tuesday, April 14, 2009

Question One

We just learned about the standard Normal distribution with PDF $\phi(z)$ and CDF $\Phi(z)$. Let's familiarize ourselves with it, as we will be using it a lot in the future.

1. Find $u = \Phi(z)$ for $z = \{-3, -2.5, -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2, 2.5, 3\}$.

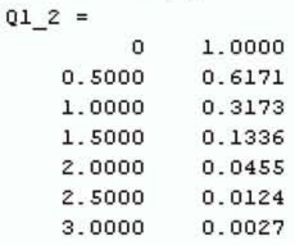
   - Solution to (1):



```
Q1_1 =
   -3.0000    0.0013
   -2.5000    0.0062
   -2.0000    0.0228
   -1.5000    0.0668
   -1.0000    0.1587
   -0.5000    0.3085
         0    0.5000
    0.5000    0.6915
    1.0000    0.8413
    1.5000    0.9332
    2.0000    0.9772
    2.5000    0.9938
    3.0000    0.9987
```

2. Find $u = 1 - Pr(|Z| \leq z)$ for $z = \{0, 0.5, 1, 1.5, 2, 2.5, 3\}$ using your answers from part (1). Explain (via math or words) how you obtained your answers. We call these values of u "p-values" which stands for "probability values" of a result at least that extreme occuring. Memorize these seven values–you will certainly use them in the future.

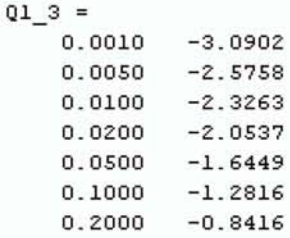
   - Solution to (2):



```
Q1_2 =
         0    1.0000
    0.5000    0.6171
    1.0000    0.3173
    1.5000    0.1336
    2.0000    0.0455
    2.5000    0.0124
    3.0000    0.0027
```

3. Find $z = \Phi^{-1}(u)$ for $u = \{0.001, 0.005, 0.01, 0.02, 0.05, 0.10, 0.20\}$.

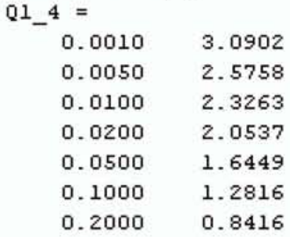
- Solution to (3):



```
Q1_3 =
        0.0010   -3.0902
        0.0050   -2.5758
        0.0100   -2.3263
        0.0200   -2.0537
        0.0500   -1.6449
        0.1000   -1.2816
        0.2000   -0.8416
```

4. Use the results from part (3) to obtain $z = \Phi^{-1}(1-u)$ for each u.

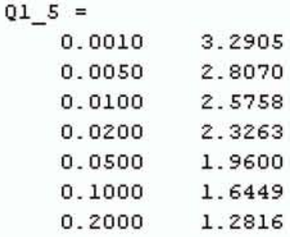
- Solution to (4):



```
Q1_4 =
        0.0010    3.0902
        0.0050    2.5758
        0.0100    2.3263
        0.0200    2.0537
        0.0500    1.6449
        0.1000    1.2816
        0.2000    0.8416
```

5. Use the results and/or methods from part (3) and (4) to obtain z where $1 - Pr(|Z| \leq z) = u$ for each u in part (3). These values of z are called the "two-sided α-level critical values" where $\alpha = u$ in this example. For example, we say, "The (two-sided) 10% critical value of the standard normal distribution is $z = \ldots$" These will be useful in the future. Memorize these seven values as well.

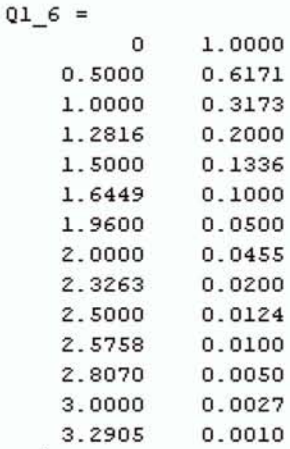
- Solution to (5):



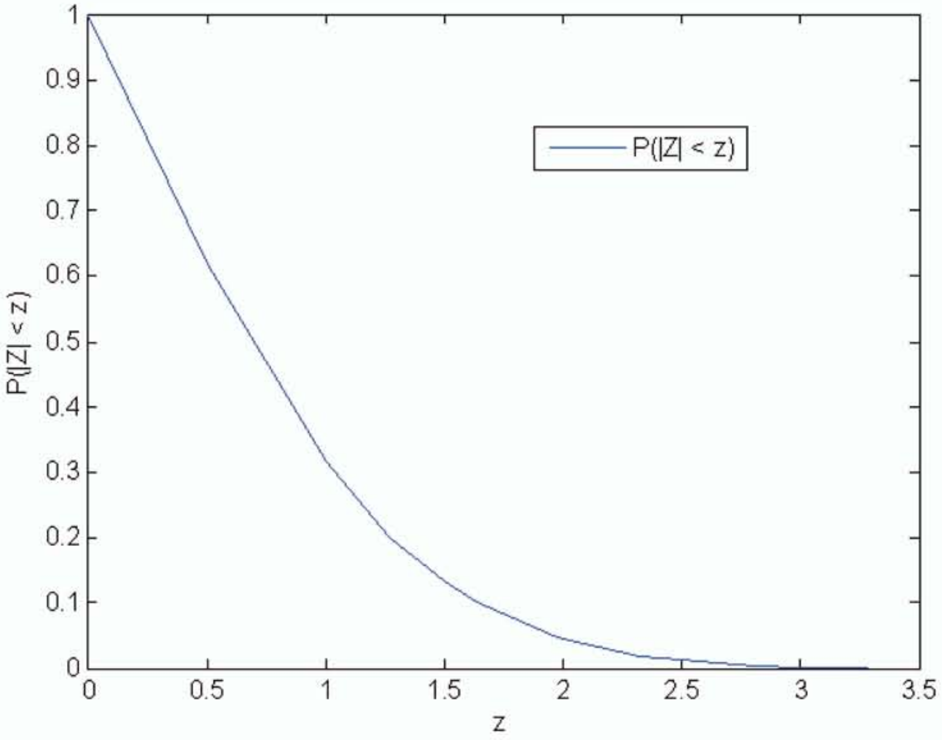
```
Q1_5 =
        0.0010    3.2905
        0.0050    2.8070
        0.0100    2.5758
        0.0200    2.3263
        0.0500    1.9600
        0.1000    1.6449
        0.2000    1.2816
```

6. Finally, interlace (i.e. sort) the values of z and u in a short table with the 14 values of z and u you obtained in parts (2) and (5). Start with $z = 0$ and $p = 1$ and put the z's in the right order with their corresponding probabilities.

- Solution to (6):



```
Q1_6 =
         0    1.0000
    0.5000    0.6171
    1.0000    0.3173
    1.2816    0.2000
    1.5000    0.1336
    1.6449    0.1000
    1.9600    0.0500
    2.0000    0.0455
    2.3263    0.0200
    2.5000    0.0124
    2.5758    0.0100
    2.8070    0.0050
    3.0000    0.0027
    3.2905    0.0010
```



Question Two

1. Given a random variable X, define the standardization Z of X and derive its variance.

    - Solution to (1): The standardization Z of X is the affine transformation where we subtract off the mean and divide by the standard deviation (square-root of the variance):

$$Z = \frac{X - \mathbb{E}[X]}{\sqrt{Var(X)}} = \frac{X - \mu_x}{\sigma}.$$

Its variance is 1: $Var(Z) = Var\left(\frac{X-\mu_x}{\sigma}\right) = \frac{1}{\sigma^2}Var(X-\mu_x) = \frac{1}{\sigma^2}Var(X) = \frac{1}{\sigma^2}\sigma^2 = 1$.

2. Suppose $\bar{X}_{20}$ is the mean of a sample of $n = 20$ i.i.d. observations with $X_i \sim N(1,2)$. What is the expected mean and variance of this average? What is its distribution (Hint: see lecture notes, around Proposition 24)? What is the probability that the sample mean $\bar{X}_{20}$ is between 0.75 and 1.25?

   - Solution to (2): The expected mean of the sum of k i.i.d. random variables is just the expectation of a single draw:

   $$\mathbb{E}[\frac{1}{k}(X_1 + ... + X_k)] = \mathbb{E}[X]$$

   which means that we should expect the mean to be equal to 1, and the variance follows a similar formula, which we haven't derived up until now:

   $$\begin{aligned} Var\left[\frac{1}{k}(X_1 + ... + X_k)\right] &= \frac{1}{k^2}Var(X_1 + ... + X_k) \\ &= \frac{1}{k^2}(Var(X_1) + ... + Var(X_k)) \\ &= \frac{1}{k^2}k \cdot Var(X) \\ Var\left[\frac{1}{k}(X_1 + ... + X_k)\right] &= \frac{1}{k}Var(X) \end{aligned}$$

   where we make use of the independence of the k random variables in the second line and identical distribution in the third line. The variance would thus be, for this particular problem:

   $$Var(\bar{X}_{20}) = \frac{1}{20}4 = \frac{1}{5}.$$

   Its distribution is Normal (by Proposition 24) with mean 1 and variance $\frac{1}{5}$. The probability that the sample mean is between 0.75 and 1.25 is $P(0.75 \leq \bar{X}_{20} \leq 1.25)$ which we can standardized to the random variable $Z = \frac{\bar{X}_{20}-1}{\sqrt{\frac{1}{5}}}$:

   $$\begin{aligned} P(0.75 \leq \bar{X}_{20} \leq 1.25) &= P\left(\frac{-0.25}{\sqrt{\frac{1}{5}}} \leq \frac{\bar{X}_{20}-1}{\sqrt{\frac{1}{5}}} \leq \frac{0.25}{\sqrt{\frac{1}{5}}}\right) \\ &= P\left(-0.25\sqrt{5} \leq z \leq 0.25\sqrt{5}\right) \\ &= 1 - 2(0.2881) \\ &= 0.4238 \end{aligned}$$

3. In Pset #5, we used the convolution formula to determine the finite distribution of the average (and sum) of k i.i.d. exponential random variables, noting that the mean did not depend on the distribution. What is the variance of the sum of k i.i.d. random

variables? What's the variance of their average? We noticed (in the solution set to Pset #5) that the average over the exponential RVs approaches the normal distribution. Do you think that this property was specific to the exponential distribution?

- Solution to (3): We derived the answers to this question in part (1):

$$\begin{aligned} Var\left[X_1 + ... + X_k\right] &= kVar(X) \\ Var\left[\frac{1}{k}(X_1 + ... + X_k)\right] &= \frac{1}{k}Var(X) \end{aligned}$$

  The average over exponential RVs approaching the Normal distribution was not specific to the exponential distribution. It is the property of random variables whose PDF satisfies general requirements for the Central Limit Theorem to hold.

Question Three

Suppose that a random sample of size n is taken from a normal distribution with mean μ and variance 3, and that the sample mean, $\bar{X}_n$, is calculated.

1. What does n need to be so that the probability of $\bar{X}_n$ being within 0.1 of μ is at least 90%?[1]

   - Solution to (1): This power calculation follows:

$$\begin{aligned} P\left(\left|\bar{X}_n - \mu\right| \leq 0.1\right) &\geq 0.90 \\ P\left(\left|\frac{\bar{X}_n - \mu}{\sqrt{3/n}}\right| \leq \frac{0.1}{\sqrt{3/n}}\right) &\geq 0.90 \\ P\left(|Z| \leq \frac{0.1}{\sqrt{3/n}}\right) &\geq 0.90 \end{aligned}$$

     which we can now solve for equality to find the minimum n for which this will hold (you can check to make sure it actually is a minimum n by checking whether the probability increases or decreases for $n+1$). We first need the critical value for Z such that the probability statement will hold:

$$\begin{aligned} P\left(|Z| \leq z\right) &= 0.90 \\ z &= 1.6449 \end{aligned}$$

[1] These are what we call power calculations and help us craft surveys and experiments to guarantee before spending lots of money that we'll be able to detect economically meaningful effects, if they are present. For example, we usually think of log wages as being normally distributed. Often times we need to know how many observations we need in a survey for the average log wage to be close enough to the true mean. Further, as we will learn about the Central Limit Theorem, we'll realize that for any average that we might be interested, for $n \geq 30$, the distribution we sample from will generally not affect our calculation of n.

which we now set equal to the critical value for our mean:

$$\begin{aligned} 1.6449 &= z = \frac{0.1}{\sqrt{3/n}} \\ n &= \left\lceil \left(\frac{1.6449}{0.1} \sqrt{3} \right)^2 \right\rceil \end{aligned}$$

which gives us $n = 812$.

2. What does n need to be so that the probability of $\bar{X}_n$ being within 0.01 of μ is at least 90%?

   - Solution to (2): We just insert the different value for the distance to obtain the following expression:

     $$n = \left\lceil \left(\frac{1.6449}{0.01} \sqrt{3} \right)^2 \right\rceil$$

     and the following $n = 81,171$. Thus, a 10-fold, $O(\delta)$, improvement in precision requires a 100-fold, $O(\delta^2)$, increase in sample size.

3. In my experiments at Yahoo!, I have been looking at treatment and control differences to determine the effectiveness of online display advertising. The standard deviation of (scaled) weekly sales is R\$15.00 and average weekly purchases are $R\$1.00$ for the control group and $R\$1.00 + \delta$ for the treatment group. If I am constrained to have only 25% of my observations in the control group and the remaining 75% in the treatment group, what is the variance of the difference between $\bar{X}_T - \bar{X}_C$ where $\bar{X}_T$ is the sample average over all treatment individuals and $\bar{X}_C$ is the average over all control individuals? How large does N have to be in order for the probability of $\bar{X}_T - \bar{X}_C > 0$ is at least 95%? Assume the sample averages for the treatment and control groups are normally distributed. Do this for general δ and then evaluate it for $\delta = R\$0.05$ and $\delta = R\$0.10$. Are these values of N large? Comment on the results.

   - Solution to (3): We are interested in a random variable, $\Delta = \bar{X}_T - \bar{X}_C$ which has expectation $\mathbb{E}[\Delta] = \mathbb{E}[\bar{X}_T - \bar{X}_C] = \mathbb{E}[\bar{X}_T] - \mathbb{E}[\bar{X}_C] = \delta$ and variance $Var(\Delta) = \frac{1}{\frac{3}{4}N} \cdot Var(X_T) + \frac{1}{\frac{1}{4}N} \cdot Var(X_C) = \frac{4+12}{3N} Var(X) = \frac{16}{3N} \cdot 225$. We want to know how large N has to be in order for the probability that $\bar{X}_T - \bar{X}_C > 0$ is at least 95%:

     $$\begin{aligned} P\left(\Delta \leq \delta\right) &\geq 0.95 \\ P\left(\frac{\Delta}{\sqrt{\frac{16}{3N} \cdot 225}} \leq \frac{\delta}{\sqrt{\frac{16}{3N} \cdot 225}} \right) &\geq 0.95 \\ P\left(Z \leq z^*\right) &\geq 0.95 \end{aligned}$$

     for $z^* = \frac{\delta}{\sqrt{\frac{16}{3N} \cdot 225}}$. So, we just need to find the 0.95% critical value: 1.6449 (it's the same as for the last problem since we have a one-sided probability that is one-half (in terms difference from 100%) the 90% probability in the previous parts. So, we find:

$$\begin{aligned} 1.6449 &= z^* = \frac{\delta}{\sqrt{\frac{16}{3N} \cdot 225}} \\ N &= \left\lceil \left(\frac{1.6449}{\delta} \sqrt{\frac{16}{3} \cdot 225} \right)^2 \right\rceil . \end{aligned}$$

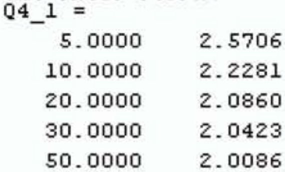
For $\delta = R\$0.05$, we find that $N = 1,298,735$ will yield a 95% probability of discovery, while for $\delta = R\$0.10$, we only need one-fourth as many total observations, or $N = 324,684$, to obtain a 95% probability of discovery. Interestingly enough, if we could balance the treatment and control groups such that they have 50% in each, we would actually need a lower N by a factor of $\frac{4}{16/3} = 0.75$. But, in this case, it would mean that the advertiser would not be able to show the ads to as many individuals–so, economic considerations force us to use a lower powered experiment as we trade off learning with earnings. ;) Further, if there was a way to reduce the variance from 15^2 by controlling for observable characteristics, we could use just the residual variance in our calculation and perhaps further reduce the number of observations (or reduce our desired detection level, δ, to discover smaller effects with a high probability).

Question Four

Use a table, calculator, internet, simulation, or any other method (besides cheating) to determine the following critical values:

1. $T \sim t$ distribution: $1 - P(|T| \leq t; dof) = 0.05$ for $dof = 5$, 10, 20, 30, 50. How do these compare to the 0.05 critical values of the Normal?

   - Solution to (1): The degrees of freedom and t-distribution critical values are in the table below:



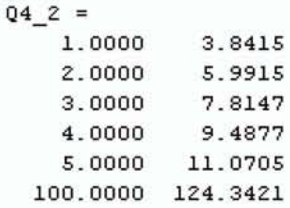
```
Q4_1 =

    5.0000    2.5706
   10.0000    2.2281
   20.0000    2.0860
   30.0000    2.0423
   50.0000    2.0086
```

   These are all larger than the 0.05 critical value of the Normal, 1.96. This will be important to remember when computing averages with small numbers of observations (like $N \leq 100$), as it may influence your inference about the research question you're computing the statistic for.

2. $X \sim \chi^2_k$ distribution: $P(X \geq x; k) = 0.05$ for $k = 1, 2, 3, 4, 5, 100$. How do these compare to the 0.05 critical values of the Normal? (Hint: Divide by k and take the square root–kind of like an average.) A few of these may be worth memorizing as well.

- Solution to (2):



```
Q4_2 =
     1.0000     3.8415
     2.0000     5.9915
     3.0000     7.8147
     4.0000     9.4877
     5.0000    11.0705
   100.0000   124.3421
```

For $k = 1$, we obtain that the critical value is just the squared Normal critical value: $1.96^2 = 3.84$. For greater k, we find that the square root of the average is less. This is just another example of Laws of Large Numbers causing the average to become more concentrated.

Question Five

What is the distribution of $\frac{Y_1/k_1}{Y_2/k_2}$ where $Y_1 \sim \chi^2_{k_1}$ and $Y_2 \sim \chi^2_{k_2}$? What happens to the value and variance of the denominator when when $k_2 \to \infty$? (Hint: What is the sum of two i.i.d. χ^2_1 random variables? Then, what's the variance of their average? And how about the sum and average of k_2 of them?) What distribution do you think that means that the ratio converges to (holding k_1 fixed)?

- Solution: The distribution of $\frac{Y_1/k_1}{Y_2/k_2}$ is $F(k_1, k_2)$. The value of the denominator converges to the average of k_2 χ^2_1 random variables, each of which has expectation 1. Thus, the denominator converges to 1 and its variance, $\frac{1}{k_2}Var(\chi^2_1) = \frac{1}{k_2} \cdot 2 \to 0$ as $k_2 \to \infty$. This means that $\frac{Y_1/k_1}{Y_2/k_2} \to \frac{Y_1}{k_1}$ where $Y_1 \sim \chi^2_{k_1}$. So, the ratio converges to a scaled version of the $\chi^2_{k_1}$, which, for large k_1, is approximately $N(1, \frac{2}{k_1})$. Further, if we now think about Central Limit Theorems, that the sum of k_1 χ^2_1 random variables is going to converge to a Normal with mean equal to k_1 and variance equal to $k_1 Var(\chi^2_1) = k_1 \cdot 2$.