Business Statistics 41000: Homework # 4

Drew Creal

Due date: Homework #4 will be due at the beginning of class in week # 8

Remarks: These questions cover Lectures #5, #6 and part of #7.

Question # 1. The Normal Distribution

Suppose $R \sim N(0.1, 0.0016)$. (HINT: Use functions in Excel. Which one did you use?)

(a) What is $P(R < 0)$?

$NORM.DIST(0,0.1,0.04,\text{TRUE}) = 0.00621$

(b) What is $P(0 < R < 0.1)$?

$NORM.DIST(0.1,0.1,0.04,\text{TRUE}) - NORM.DIST(0,0.1,0.04,\text{TRUE}) = 0.5 - 0.00621 = 0.4938$

(c) What is $P(R > 0.15)$?

$1 - NORM.DIST(0.15,0.1,0.04,\text{TRUE}) = 1 - 0.8944 = 0.1057$
Question # 2. The Normal Distribution

The problem is intended to learn more about the normal distribution.

(a) In Excel, evaluate the following:

1. =NORM.INV(.5,0,1) = -1.39214e^{-16}  
   NOTE: This should be zero! Excel doesn’t have very good numerical accuracy.

2. =NORM.INV(.025,0,1) ≈ -1.96

3. =NORM.INV(.05,0,1) = -1.645

4. =NORM.INV(0.16,0,1) = -0.99445  
   NOTE: This is approximately negative one!

5. =NORM.INV(.84,0,1) = 0.99445   
   NOTE: This is approximately positive one!

6. =NORM.INV(0.95,0,1) = 1.645

(b) What does the function NORM.INV do? How does it relate to the standard normal c.d.f.?

The NORM.INV function is the inverse of the CDF. In other words, you provide the function with a probability and it returns to you the point $x$ on the horizontal access which has that much probability under the pdf (to the left of the point). It does exactly the opposite of what NORM.DIST does.

(c) Without using Excel, what would NORM.INV(0.975,0,1) be?

It should equal (roughly) 2! We know this because for a normal distribution two standard deviations from the mean is 95% of the probability! But, remember that we also need to add to this the additional 2.5% of the probability in the left-hand tail of the pdf.
Question # 3. The Law of Large Numbers

Questions:

(a) Draw \( n \) random numbers from the normal distribution \( x_i \sim N(\mu, \sigma^2) \). You may choose the values of \( \mu \) and \( \sigma \). Calculate the sample mean \( \bar{x} \) and sample variance \( s_x^2 \) for \( n = 100, 200, 500, \) and 1000 draws. Are the sample mean and variance getting closer to the true values of \( \mu \) and \( \sigma^2 \) as \( n \) gets larger?

To generate random numbers from the normal distribution \( N(2, 9) \) in Excel, you would type \( =\text{NORM.INV(RAND(),2,3)} \).

As \( n \) gets larger, the sample mean should converge to \( \mu = 2 \) and the sample variance should converge to \( \sigma^2 = 9 \).

(b) Draw \( n \) random numbers from the binomial distribution \( x_i \sim \text{binomial}(20, 0.75) \). Calculate the sample mean \( \bar{x} \) and sample variance \( s_x^2 \) for \( n = 100, 200, 500, \) and 1000 draws. What values are \( \bar{x} \) and \( s_x^2 \) converging to as \( n \) becomes larger?

To generate random numbers from the binomial distribution \( \text{binomial}(20, 0.75) \) in Excel, you would type \( =\text{BINOM.INV(20,0.75,RAND())} \).

As \( n \) gets larger, the sample mean should converge to \( n * p = 15 \) and the sample variance should converge to \( np(1 - p) = 3.75 \).
Question # 4. The Central Limit Theorem

The goal of this question is to learn what a central limit theorem is. For each part, you may hand in a histogram and a brief description of what you learned.

(a) The solution to this question proceeds in 3 steps:

1. In column A of Excel, draw \( n = 200 \) random numbers \( x_i \sim \text{binomial}(40, 0.35) \). In another excel spreadsheet, calculate the sample mean \( \bar{x} \) of these values. For example, my random numbers are in column A of Sheet1. In cell A1 on sheet 2, I typed: “=AVERAGE(Sheet1!A1:Sheet1!A200)” to calculate the sample mean of the 200 values.

2. Repeat the first step in columns B, C, D, E, etc. a large number of times, e.g. 500 columns. In other words, you are generating 500 data sets (one data set per column) each of size \( n = 200 \). And, for each of these 500 data sets, you are calculating the sample mean \( \bar{x} \).

3. Draw a histogram of the 500 values of \( \bar{x} \). What does the histogram look like? What is the variance of this distribution? What happens if \( n \) becomes large?

Here is my histogram from the binomial distribution.
(b) Repeat question (a) above but draw the random variables from a uniform distribution $x_i \sim U(0, 1)$.

Here is my histogram from the uniform distribution.

(c) Repeat question (a) above but draw the random variables from another distribution of your choice other than the uniform or binomial distribution.

Here is my histogram for a Poisson distribution with parameter $\lambda = 5$. 
Question # 5. Confidence Intervals for voting

It is election night and you are working for a TV news service in Florida. You are covering a race between two candidates, George and Al. Assume that all voters in the state of Florida vote for one of these two candidates and let $p$ be equal to the proportion of people who vote for Al. Votes are now being counted, and the network executives want you to make a projection for which candidate will win the state. Though of course they want this projection as soon as possible, they are concerned with the network’s credibility and are willing to accept at most a 5% chance that the projection is wrong.

(a) Suppose that 500 votes from the state have been counted. The results are 263 votes for Al (the other 237 voted for George). Construct a 95% confidence interval for $p$. Can you project a winner in the election?

Here we have $\hat{p} = 263/500 = 0.526$, and $n = 500$. So $S.E.(\hat{p}) = \sqrt{\frac{0.526(1-0.526)}{500}} = 0.02233$. The 95% CI is $\hat{p} \pm 2 \times S.E.(\hat{p}) = 0.526 \pm 0.04466 = (0.4813, 0.5707)$. This confidence interval contains the value $p = 0.5$. In other words, we can’t tell based on this data whether more than 50% of all voters voted for Al. (We could also have done this by testing $H_0 : p = 0.5$ directly; we’d fail to reject.)

(b) Now suppose that 10,000 votes have been counted and 5,150 of them were votes for Al. Can you project a winner in the election?

This time $\hat{p} = 5150/10000 = 0.515$, and $n = 10,000$. Now $S.E.(\hat{p}) = \sqrt{\frac{0.515(1-0.515)}{10,000}} = 0.005$. The 95% CI is $\hat{p} \pm 2 \times S.E.(\hat{p}) = 0.515 \pm 0.005 = (0.505, 0.525)$. In particular, this CI gives us enough information (at the 95% level) to say that $p$ is bigger than 0.5, or that more than 50% of people voted for Al.

(c) Now suppose you learn that the 10,000 voters in part (b) came from parts of the state known to support Al’s party, and that some parts of the state known to support George have not yet reported their results. Are you still willing to make a projection based on your confidence interval from (b)? Explain.

This changes things completely. Remember, to build our confidence interval, we assumed that the observations $X_i$ in our data were $X_i \sim Bernoulli(p)$ i.i.d. In
particular, all of the observations in our sample are independent of each other and have the same probability, $p$, of being a one. Remember, $p$, (the probability a randomly chosen voter in the state voted for Al) is the probability we are trying to learn about. If the voters in our sample were drawn from counties known a priori to support Al, they may still be i.i.d. Bernoulli, but they don’t have the same probability of being one as if we took a random sample of voters from the entire state. In other words, our estimate, $\hat{p}$, will be biased. So no, you would not want to project the election results based on this data.
Question # 6. The Normal Distribution

Suppose you are in charge of the “machine that fills cereal boxes” from the notes. You have lots of data on your process and are sure that the amount (in grams) of cereal put in each box can be described as i.i.d. draws from the normal distribution with mean $\mu = 345$ and standard deviation $\sigma = 25$. You are about to be audited by an inspector who will take a sample of 25 boxes and compute the average amount put in the 25 boxes. If that average is greater than 350 or less than 340, you will fail the audit and be reprimanded by your superiors.

(a) The inspector is about to look at the average amount in 25 boxes. Before the inspector begins looking at boxes, what is the probability distribution of this average? (This is the “sampling distribution” we talked about in class.)

Let $X_i$ be the amount of cereal in the $i$-th box. We know $E[X_i] = 345$ and $V[X_i] = 25^2$, and we are looking at $n = 25$ observations. We know that the mean and variance of the sampling distribution of $\overline{X}$ are

$$E[\overline{X}] = 345$$
$$S.E.(\overline{X}) = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{25^2}{25}} = 5$$

Finally, the central limit theorem (CLT) tells us when $n$ is large, the sample mean is normal

$$\overline{X} \sim N(345, 5^2)$$

(b) What is the probability that you will pass the audit?

$$P(340 < \overline{X} < 350) = 0.68$$

For a normal random variable like $\overline{X}$, there is a 68% probability of being within one standard deviation of the mean.
(c) How would your answer to part (b) change if the inspector took a sample of 100 boxes?

Now that the sample size has increased to \( n = 100 \), we need to change the standard error (i.e. the standard deviation of our sampling distribution).

\[
S.E.(\bar{X}) = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{25^2}{100}} = 2.5
\]

The probability of passing the audit just increased to 95% because the range 340 to 350 is now two standard deviations away from the mean.

(d) If you are right about \( \mu \) and \( \sigma \), do you want the inspector to look at more boxes or fewer? Explain intuitively.

As long as you are sure that \( \mu \) is actually 345, you want the inspector to look at MORE boxes. The reason is that the sample mean will differ from \( \mu \) due to randomness in the data, but the more observations we have, the closer we expect \( \bar{X} \) to be to \( \mu \).
Question # 7. Confidence Intervals for the Normal Distribution

In this question, we compute confidence intervals using the Australian returns data.

<table>
<thead>
<tr>
<th></th>
<th>Australia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>107.00</td>
</tr>
<tr>
<td>Mean</td>
<td>0.012</td>
</tr>
<tr>
<td>Stand. Dev.</td>
<td>0.054</td>
</tr>
</tbody>
</table>

(a) Above are summary statistics for the Australian returns (see conret.xls). Construct the 95% confidence interval for the mean of the Australian returns.

The 95% confidence interval is:

\[ \bar{X} \pm 2 \times S.E.(\bar{X}) = \bar{X} \pm 2 \times \frac{s_x}{\sqrt{n}} \]

\[ = 0.012 \pm 2 \times \frac{0.054}{\sqrt{107}} \]

\[ = (0.0016, 0.022) \]

(b) Now open the country returns data in Excel. Select StatPro -> Statistical Inference -> One Sample Analysis... and from the menu, select “confidence interval for the mean” and enter “95” as the confidence level. Verify that Excel’s answer matches yours from part (b).
\[
2 \times \frac{0.054}{\sqrt{107}} = 0.0104407 \approx 0.01037 \text{ Excel is using the critical value from a Student's t distribution.}
\]

(c) Using the numbers from part (a), construct a 90% confidence interval and a 99% confidence interval for the mean of the Australian returns (use Excel's NORM.S.INV function to get the critical values as we discussed in class).

Which interval is widest (90, 95, or 99%)? Give a brief intuitive explanation. In particular, does a wider interval mean we know less about the actual mean?

**We need to compute two new critical values. We can use the NORM.S.INV function which is the inverse CDF for the standard normal distribution.**

\[
\text{NORM.S.INV}(0.05) = -1.64485
\]

\[
\text{NORM.S.INV}(0.005) = -2.5758
\]

<table>
<thead>
<tr>
<th>Our two new confidence intervals can be constructed as:</th>
</tr>
</thead>
<tbody>
<tr>
<td>90% interval ( = \bar{X} \pm 1.64485 \times S.E.(\bar{X}) )</td>
</tr>
<tr>
<td>( = 0.012 \pm 1.64485 \times \frac{0.054}{\sqrt{107}} )</td>
</tr>
<tr>
<td>( = (0.0035, 0.0206) )</td>
</tr>
<tr>
<td>and</td>
</tr>
<tr>
<td>99% interval ( = \bar{X} \pm 2.5758 \times S.E.(\bar{X}) )</td>
</tr>
<tr>
<td>( = 0.012 \pm 2.5758 \times \frac{0.054}{\sqrt{107}} )</td>
</tr>
<tr>
<td>( = (-0.00145, 0.0255) )</td>
</tr>
</tbody>
</table>

The 99% CI is widest, because to have a higher probability of containing the unknown “true” mean, we need a wider interval.