TEST 3 MATH-1140 (DUPRÉ) SPRING 2013 ANSWERS

FIRST: PRINT YOUR LAST NAME IN LARGE CAPITAL LETTERS ON THE UPPER RIGHT CORNER OF EACH SHEET TURNED IN, INCLUDING THE COVER SHEET, AND MAKE SURE YOUR PRINTED COVER SHEET IS ON TOP.

SECOND: ALL ANSWERS MUST BE CLEARLY WRITTEN AND FINAL ANSWERS GIVEN AS EXACT FRACTIONS OR CORRECT TO AT LEAST THREE SIGNIFICANT DIGITS.

1. Make sure you have followed the above directions exactly to get credit for this first Test Problem.

Suppose that a box contains 50 blocks of which exactly 20 are red. We randomly draw 10 blocks from the box and count the number T of times we get a red block. Calculate:

2. $E(T) = 4$

ANSWER: We have a population of blocks of size $N = 50$, of which we consider that $R = 20$ are *successes*, that is, they are RED. The success rate is then $p = R/N = 20/50 = .4$, or simply $p = 0.4$, is the success rate for drawing red blocks from the box, or equivalently, $p = 0.4$ is the probability of drawing a red block from this block on a single draw. When we randomy draw 10 blocks from the box, we are taking a sample of size $n = 10$ from the population of size $N = 50$. The total T of the number of red blocks in the sample is then

$$E(T) = np = (10)(.4) = 4.$$

3. σ_T(given drawing with replacement)=1.549

ANSWERS: When drawing with replacement, then T is simply the sample total for a sample of size $n = 10$ of an *Independent Random Sample (IRS)* of size $n = 10$, so T has the binomial distribution with $n = 10$ and success rate $p = 0.4$. This means

$$\sigma_T \text{ (given drawing with replacement) } = \sqrt{np(1-p)} = \sqrt{[E(T)][1-p]} = \sqrt{4(.6)} = \sqrt{2.4},$$

which is approximately 1.549193338, or 1.549, to three significant digits.

ANSWERS:

4. σ_T(given drawing without replacement)=1.400

ANSWERS: σ_T(given drawing without replacement)= $\sqrt{np(1-p)\dfrac{N-n}{N-1}}$ which is approximately 1.399708424, or 1.400, to three significant digits. Notice that the distribution of T is hypergeometric when drawing without replacement, so the sample of size $n = 10$ becomes a *Simple Random Sample (SRS)* of size $n = 10$, so we must multiply the IRS standard deviation by the SRS correction factor c_{SRS}, where

$$c_{SRS} = \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{40}{49}}.$$

Therefore, now

$$\sigma_T = c_{SRS}\sqrt{np(1-p)} = \text{approximately, } 1.399708424,$$

or 1.400, to three significant digits.

5. $P(T \leq 4|\text{drawing with replacement}) =$ approximately, 0.633

ANSWERS: When drawing with replacement, the distribution of T is binomial with $n = 10$ and $p = 0.4$, so we use the binomial CDF table for $n = 10$ and find 0.633.

A more accurate calculation with a calculator gives

$$P(T \leq 4|\text{drawing with replacement}) = \text{ approximately, } 0.6331032576,$$

which is 0.633, to three significant digits.

6. $P(T = 4|\text{drawing without replacement}) = \dfrac{C(20,4)C(30,6)}{C(50,10)} =$ approximately, 0.280

ANSWERS: When drawing without replacement, the distribution of T is hypergeometric, so to compute the probability of exactly 4 red blocks when drawing 10, we know we have to get 4 red blocks from the 20 red blocks in the population, and 6 non-red blocks from the 30 non-red blocks in the population of 50 blocks, so the probability is the number of ways to get 4 of the 20 red blocks multiplied by the number of ways to get 6 of the 30 non-red blocks divided by the number of ways to get 10 blocks from the 50 in the box, which is approximately 0.280058603, or simply 0.280, to three significant digits.

Suppose that we are studying the length of fish in Lake Wobegon. An independent random sample of 4 fish from Lake Wobegon has a sample mean length of 16 inches with a sample standard deviation of 7 inches. We assume that fish length is normally distributed for fish in Lake Wobegon.

7. What is the MARGIN OF ERROR in the 95 percent confidence interval for the true mean length of fish in Lake Wobegon if we know that the POPULATION standard deviation for fish length in Lake Wobegon is 6 inches?

ANSWERS:

$$\text{MARGIN OF ERROR} = ME = \text{ approximately } 5.880.$$

The **Margin of Error**, denoted ME, is given here, using the known value of the standard deviation of fish length, $\sigma = 6$, by the margin of error of error formula

$$ME = \frac{z_C \sigma}{\sqrt{n}} = \frac{(1.960)(6)}{2} = 5.880.$$

A comment on calculation accuracy is appropriate here. If we use a calculator to get a more precise value of z_C, for $C = 95$ percent confidence, of course we need the the score which cuts off the right tail of area .025, and find that to be 1.959963986, leading to the more accurate value of 5.79891958, as the margin of error, which is 0.5880, to three significant digits. But what sense does it make to have such accuracy (nine decimal points) when there are only 4 fish in the sample, so we know the margin of error is big. Think about it.

8. If we know that the POPULATION standard deviation for fish length in Lake Wobegon is 6 inches, does our sample data establish that the true mean length of the fish exceeds 10 inches at the .05 significance level? Give the value of the standardized test statistic for the sample data and give the P-Value of the data.

ANSWERS:

$$\text{P-Value} \ = \ \text{approximately } 0.0228.$$

This is obviously a hypothesis test because the significance level is asked for. Let X be fish length. We are asked if the data establishes $\mu_X > 10$, and you only have a chance of proving the alternate hypothesis in a hypothesis test, so the alternate hypothesis is

$$H_{alt} : \mu > 10.$$

This means that the null hypothesis is

$$H_0 : \mu = 10.$$

The P-Value of the data is then the probability that another sample obtained in the same manner would have a sample mean as or more contradictory, that is as much or more than our sample mean of $\bar{x}_{data} = 16$. That is,

$$\text{P-Value} = P(\bar{X} \geq \bar{x}_{data} | X \text{ normal }, \mu = 10, \sigma_X = 6).$$

To compute the probability here we need to standardize the inequality $\bar{X} \geq \bar{x}_{data}$, using

$$Z_{\bar{X}} = \frac{\bar{X} - \mu_X}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu_X}{(\sigma_X/\sqrt{n})}, \text{ so } z_{data} = \frac{16 - 10}{(6/\sqrt{4}} = \sqrt{4} = 2.$$

Therefore, the value of the standardized test statistic is $z_{data} = 2$, and

$$\text{P-Value} = P(\bar{X} \geq \bar{x}_{data} | X \text{ normal }, \mu = 10, \sigma_X = 6) = P(Z \geq 2) = .5 - P(0 \leq Z \leq 2).$$

The probability that standard normal Z lies between 0 and 2 can be found in the Right Body Table for the Normal Distribution and is 0.4772, so the P-Value is $.5 - .4772 = 0.0228$, to three significant digits. The criteria for rejecting the null hypothesis at level of significance α is that the P-value should be less than or equal to α, and here $\alpha = 0.05$. Thus here, the criteria to establish the alternate hypothesis is that the P-Value of the is data less than or equal to 0.05. Since 0.0228 is less than 0.05, we reject the null hypothesis and regard the alternate hypothesis $\mu > 10$ as having established by our data at the 0.05 level of significance.

The more accurate calculation of the P-Value using a calculator instead of the normal distribution table gives the P-Value as

$$P(Z \geq 2) = \text{ approximately } 0.0227500625,$$

which is 0.0228, to three significant digits. Again, the comments on calculation accuracy apply, it is sort of a waste of time to worry about getting 9 decimal place accuracy when trying to draw conclusions based on a sample of size 4.

9. What is the MARGIN OF ERROR in the 95 percent confidence interval for the true mean length of fish in Lake Wobegon if we DO NOT know that the population standard deviation for fish length in Lake Wobegon is 6 inches but instead use our sample standard deviation of 7 inches?

ANSWERS:

$$\text{MARGIN OF ERROR } = ME = \text{ approximately, } 11.14.$$

If we do not know that $\sigma_X = 6$, then we only have our data to go on, it it gives the sample standard deviation as $s = 7$. We then standardize using the sample mean random variable $\bar{X}$ and the sample standard deviation random variable S_X resulting in the student $t-$distribution for df degrees of freedom and here $df = n - 1 = 4 - 1 = 3$, and

$$t = \frac{\bar{X} - \mu_X}{(S/\sqrt{n})}, \text{ so } t_{data} = \frac{\bar{x}_{data} - 10}{(s/\sqrt{4})} = \frac{16 - \mu}{(7/2)}.$$

Notice the standardization procedure is the same as when σ is known, except we use s in place of σ and the resulting standard statistic is t instead of z. However, the $t-$ distribution is really a whole family of distributions, parametrized by *degrees of freedom*, denoted df, and here we have $df = n - 1$. Thus, the margin of error formula replaces the z upper centile with the t upper centile for $n - 1$ degrees of freedom. Here $n = 4$, so we have only 3 degrees of freedom. For 95 percent confidence, we need the value of t which cuts off a right tail of area 0.025, which we find from the $t-$ table using the line for 3 degrees of freedom and we find it to have the value $t_C = 3.182$. Thus, t_C replaces z_C when $\sigma =$ is unknown and we use s in place of σ. Therefore the margin of error is now

$$ME = \frac{t_C s}{\sqrt{n}} = \frac{(3.182)(7)}{\sqrt{4}} = 11.137.$$

The main thing to notice here, is how much the margin of error is increased by not knowing the population standard deviation in case of a small sample. If you look at the table of right tail criitcal values, you see that as the number of degress of freedom approaches 30, there is not a very big difference between Z and t, but for a small number of degrees of freedom, the difference is substantial.

The calculation using the calculator gives the value $t_C = 3.182446305$ which then results in a more accurate value of

$$ME = \text{ approximately, } 11.13856207,$$

which does equal 11.14 to three significant digits, and of course the comments made in the previous problems about accuracy of calculation apply even moreso here where the population standard deviation is unknown and the sample standard deviation of a very small sample is being used to estimate the population standard deviation.

10. If we DO NOT know that the POPULATION standard deviation for fish length in Lake Wobegon is 6 inches, and instead use the sample standard deviation of 7 inches, does our sample establish that the true mean length of the fish exceeds 10 inches at the .05 significance level? Give the value of the standardized test statistic for the sample data and the REJECTION REGION.

ANSWERS:

Data Standardized Test Statistic $= t_{data} = 1.714$, REJECTION REGION: $t_{data} > t_{0.05} = 2.353$.

We are working the same hypothesis test as in problem 8, but we do not have the same information. If we do not know that $\sigma_X = 6$, then we only have our data to go on, and it gives the sample standard deviation as $s = 7$. We then standardize using the sample mean random variable $\bar{X}$ and the sample standard deviation random variable S_X resulting in the student $t-$distribution for df degrees of freedom and here $df = n - 1 = 4 - 1 = 3$, and

$$t = \frac{\bar{X} - \mu_X}{(S/\sqrt{n})}, \text{ so } t_{data} = \frac{\bar{x}_{data} - 10}{(s/\sqrt{4}} = \frac{16 - 10}{(7/2)} = \frac{12}{7} = \text{ approximately, } 1.714285714,$$

or, $t_{data} = 1.714$, to three significant digits. Thus, the P-value is now

$$\text{P-Value} = P(t \geq t_{data}|df = 3).$$

The problem now is that you only have a simple table of common upper centiles for the $t-$distribution, you do not have the distribution itself. Now, at the level of significance $\alpha = 0.05$, we reject the null hypothesis if the P-Value of the data is $\leq \alpha$. But, this will be the case, if and only if the right tail cutoff for the right tail of area α is no more than t_{data}. That is, we can safely say that the P-Value is $\leq \alpha$ if and only if $t_{data} \geq t_\alpha$. We can look up $t_\alpha = t_{0.05}$ for 3 degrees of freedom in our $t-$table of Right Tail Critical Values for the $t-$distribution and find that $t_{0.05} = 2.353$. Thus, our rejection (of H_0) criteria which establishes our alternate hypothesis, $H_{alt} : \mu > 10$ is that $t_{data} > 2.353$. However, our t_{data} is only 1.714, far too small to be significant, and now the result of the hypothesis test is that **the data is inconclusive**. That is, in this case, the data fails to establish that $\mu > 10$ at the 0.05 level of significance.

Again, we can comment on levels of accuracy of computing here. If we use a calculator instead of the table, we can find that

$$t_\alpha = 2.35336342,$$

and as well, we can calculate the P-Value,

$$\text{P-Value } = P(t \geq t_{data}|df = 3) = 0.0924943145,$$

and this is certainly too big, much bigger than $\alpha = 0.05$, so the data is inconclusive. Thus, we cannot reject the null hypothesis nor establish an alternate hypothesis at level of significance $\alpha = 0.05$ using our data, as the P-Value of the data is way too big. Either way, using the calculator or the table, this is not a close call. However, if you do have a close call with a very small sample, you should be very wary of drawing a conclusion in a real situation.

Finally, we should notice here that our data FAILS to establish any valid conclusion about the mean here where we do not know the true population standard deviation, whereas it did establish the conclusion that $\mu > 10$ when we knew the population standard deviation was $\sigma = 6$. In fact, if we had $s = 5.5$ instead of $s = 7$, so that we thought the population standard deviation might be less than 6, then the P-Value of the data would have been $P(t \geq [12/(5.5)])|df = 3) = 0.0585745499$, which still would mean the data is not significant at the $\alpha = 0.05$ level of significance. Not knowing the population standard deviation represents a big lack of information when dealing with small samples naturally leading to less conclusiveness for the data in hypothesis testing and larger margins of error in confidence intervals.

Suppose, in a SMALL SAMPLE, that we ask 10 ducks in Duckburg if they will vote for Donald for Mayor of Duckburg in the upcoming election. Suppose that only 3 say yes. In a LARGE SAMPLE we ask the same question of 1000 ducks and 300 say yes.

11. What is the P-VALUE or SIGNIFICANCE of the SMALL SAMPLE data as evidence that the true percentage of ducks in Duckburg who currently say they will vote for Donald for Mayor of Duckburg is less than 40 percent?

ANSWERS:

$$\text{P-Value of Data} \;=\; \text{Significance of Data} \;=\; \text{approximately, } 0.382.$$

If we had a large enough sample here, we could use the normal approximation to the binomial and work in a similar fashion to problem 8, using $\sigma = 0.5$, because we know that for any indicator unknown, the standard deviation is at most one half. However, our criteria for a binomial to be approximately normal is that there should be at least 10 expected success and 10 expected failures. Since $\mu = np$ for the binomial, clearly if $n = 10$, then we cannot expect at least 10 successes AND 10 failures. Therefore, we must use the binomial distribution directly. We are asking if the data can prove that $p < 0.4$, so this is the alternate hypothesis, and therefore the null hypothesis is that $p = 0.4$. Thus our hypothesis test can be summarized as

$$H_0 : p = 0.4 \text{ versus } H_{alt} : p < 0.4.$$

Therefore, the approximate P-Value is simply computed using the binomial CDF directly,

$$\text{P-Value } = P(\text{number who say yes } \leq 3 \mid \text{ binomial }, n = 10,\; p = 0.4) = 0.382.$$

A more accurate calculation using a calculator gives

$$P(\text{number who say yes } \leq 3 \mid \text{ binomial }, n = 10,\; p = 0.4) = 0.3822806016,$$

which is 0.382, to three significant digits.

Notice this means that the small sample data fails to establish $p < 0.4$ at the popular level of significance $\alpha = 0.05$, for instance. The small sample data fails to establish $p < 0.4$ at the fairly sloppy level of significance $\alpha = 0.1$. However, if we were to ask if the data establishes that Donald will lose the election, then we are asking if the data establishes the alternate hypothesis $p < 0.5$, so now would would look up the probability that the number of yes responses is less than or equal to 3 using $p = 0.5$ in the binomial table, and that result is that the P-Value is only 0.172, a P-Value that still fails to establish anything at any reasonable level of significance, but just might make Donald a little nervous and seek a larger sample to clarify the situation.

In general, in application of statistical hypothesis testing, if a small sample makes you suspicious but is inconclusive, you should look at more data. However, this must be done carefully, as one can "prove" a lot of things by the method of looking until you find what you like. I can keep going to one doctor after another until I find the one that tells me what I want to hear, but that may not be the healthiest method for health care. Thus, if a larger study is warranted, one should choose a sample of substantial size and stick with the result, instead of gradually increasing the size of the sample until you find what you want. Unfortunately, this simple rule has been ignored in many experiments, leading to large numbers of erroneous conclusions in scientific literature. One reported computer simulation concluded in the analysis of a random sample of over 800 scientific articles which use significance testing, that about 70 percent of conclusions thought to be established by data were in error.

12. Using the LARGE SAMPLE data, what is the MARGIN OF ERROR in the 95 percent confidence interval for the true proportion of ducks who say they will vote for Donald for Mayor of Duckburg?

ANSWERS: The standard deviation of the indicator of yes for a single response is

$$\sigma = \sqrt{p(1-p)},$$

but p is unknown to us. However, we know here that

$$\sqrt{p(1-p)} \leq \frac{1}{2}, \text{ since } 0 \leq p \leq 1.$$

So we can proceed as if we know the standard deviation and simply use 0.5 in its place. Since the sample is very large, we can assume the binomial distribution and therefore the sample proportion are both approximately normal to a high degree, and use the margin of error formula for the case of a normal random variable as in the case of the fish length. Therefore,

$$ME = \frac{z_C(0.5)}{\sqrt{n}} = \frac{(1.960)(1/2)}{\sqrt{1000}} = \text{approximately, } 0.0309903211,$$

or $ME = 0.0310$, to three significant digits.

This means that we can be 95 percent confident that the true percentage of ducks who say they will vote for Donald is in the interval

$$0.3 - 0.0320 \leq p \leq 0.3 + 0.0310,$$

or

$$0.2690 \leq p \leq 0.3310.$$

We are therefore at least 95 percent certain that Donald would lose the election if held the day the sample data was taken.

Using a caclulator to calculate the margin of error more accurately using a more accurate value for z_C gives

$$z_C = \text{ approximately, } 1.959963986$$

and

$$ME = \text{ approximately, } 0.0309897516,$$

which is 0.0310, to three significant digits, again.

Even though this was not asked for in the problem, if we ask for the P-Value of this large sample data as evidence that the true proportion of Donald's supporters is less than 40 percent, then we would use the normal approximation assuming that the true proportion has the null hypothesis value of $p_0 = 0.4$ which then makes the standard deviation of yes indicator exactly equal to $\sqrt{p_0(1-p_0)} = \sqrt{.24}$, and therefore the $z-$score of the data is in the normal approximation given by

$$z_{data} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{.3 - .4}{\sqrt{(.24)/(1000)}} = \text{ approximately, } -6.454972244.$$

Thus, the standardized test statistic of the large sample is over 6 standard deviations below zero, so the large sample is highly significant as evidence that the true proportion is less than 0.4. In fact, the P-Value is only P-Value = approximately, 0.0000000005719.