Public Policy 529
Midterm Exam

Student ID number (8-digits): ____________________

List of Formulas

$$\bar{y} = \frac{\sum y_i}{n}$$

$$IQR = Q_3 - Q_1$$

$$SS = \sum (y_i - \bar{y})^2$$

$$s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

$$s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$$

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

$$P(\sim A) = 1 - P(A)$$

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

$$P(A \text{ and } B) = P(A) \times P(B|A)$$

$$P(A \text{ and } B) = P(A) \times P(B)$$

$$Z = \frac{y - \mu_y}{\sigma}$$

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

$$\hat{\sigma}_{\bar{y}} = \frac{s}{\sqrt{n}}$$

$$\hat{\sigma}_{\hat{\pi}} = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}$$

$$\hat{\sigma}_{\pi_0} = \sqrt{\frac{\pi_0(1 - \pi_0)}{n}}$$

$$\text{c.i.} = \bar{y} \pm t \cdot \hat{\sigma}_{\bar{y}}$$

$$\text{c.i.} = \bar{y} \pm Z \cdot \hat{\sigma}_{\bar{y}}$$

$$\text{c.i.} = \hat{\pi} \pm Z \cdot \hat{\sigma}_{\hat{\pi}}$$

$$t = \frac{\bar{y} - \mu_0}{\hat{\sigma}_{\bar{y}}}$$

$$Z = \frac{\hat{\pi} - \pi_0}{\hat{\sigma}_{\pi_0}}$$

1. Suppose that, in the population, the number of hours people spend each month listening to news has a normal distribution with a mean of 10 and a standard deviation of 4.

   (a) What percentage of people listen to 12 or more hours? (5 points)

   $$Z = \frac{12-10}{4} = .50$$

   The area above Z=0.50 on the standard normal is .3085, so about 30.9%. Rounding is okay.

   (b) What percentage of people listen to 11 or fewer hours? (5 points)

   $$Z = \frac{11-10}{4} = .25$$

   The area above 0.25 on the standard normal is .4013, so an area of 1-.4013=.5987 lies below. This corresponds to about 59.9%. Rounding is okay.

   (c) What percentage of people are in the range of 6 to 12 hours ? (5 points)

   There are multiple ways to solve this problem. First, one can find the area under the above 6 and subtract off the area above 12. Second, one can find area below 6 and the area above 12, subtracting both from 100% (or 1.00). Note that we already found the area above 12 in part (a), so we only need to worry about 6.

   $$Z = \frac{6-10}{4} = -1.00$$

   By symmetry, we know that a proportion of .1587 lies below Z=-1.00, so .8413 lies above Z=-1.00.

   - By method 1, .8413 - .3085 = .5328. In other words, 53.3%.
   - By method 2, 1.000 - .1587 - .3085 = .5328. In other words, 53.3%.

2. According to the 2012 American National Election Study, 43% of respondents report that they performed volunteer work during the past 12 months. The remaining 57% say that they did not perform volunteer work. The sample size of this survey is 5,507. Construct a 95% confidence interval around the estimated proportion of Americans who did volunteer work *and* interpret what this confidence interval tells us about the "true" population proportion. (10 points)

   $$.43 \pm 1.96 \cdot \sqrt{\frac{.43(1-.43)}{5,507}} = .43 \pm 1.96 \cdot 0.0067 = .43 \pm 0.013$$

   The range goes from .417 to .443. Rounding is okay, so if they say .42 to .44, that's fine.

   A good interpretation is: in 95% of random samples, confidence intervals constructed in this way will include the population proportion. Alternatively, one could say that there is a 95%

chance that using these procedures will produce confidence interval that includes the population proportion. The interpretation is worth 4 points.

A good answer will *not* say that we are 95% confident that the true population proportion is between .42 and .44.

3. In your own words, explain why the Central Limit Theorem is so important for significance testing. (5 points)

   Two key points from the CLT are crucial. First, that the sampling distribution of statistics from random samples is normal and centered upon their population counterparts, so our estimates are unbiased. Second, that the CLT gives us the size of sampling error, so we know by how far our estimates deviate from the truth on average.

   We use these insights to estimate how plausible it would be to obtain the sample statistics we observe under the scenario that the null hypothesis is true. If the sample statistics are highly implausible under this scenario – that is, they are unlikely to result from random sampling error given the estimated size of those errors – we reject the null hypothesis as true.

4. A survey asked respondents whether they favor or oppose limits on imports of goods from other countries. The survey also collected information on political party identification. The joint probability distribution of these two variables is presented in the table below.

Favor/Oppose	Democratic	Independent	Republican	Total
Favor	.24	.21	.15	.60
Oppose	.16	.14	.10	.40
Total	.40	.35	.25	1.00

   (a) What is P(Favor and Independent)? (5 points)

   This is simply .21.

   (b) What is P(Oppose | Independent)? (5 points)

   This is from the fraction $\frac{.14}{.35} = .40$

   (c) Are party identification and opinions about limits on imports independent? Demonstrate mathematically. (5 points)

   Yes, they are independent, since the conditional probabilities match the unconditional probabilities. For example:

$$\begin{aligned} P(\text{Favor}) &= .60 \\ P(\text{Favor}|\text{Democratic}) &= \frac{.24}{.40} = .60 \\ P(\text{Favor}|\text{Independent}) &= \frac{.21}{.35} = .60 \\ P(\text{Favor}|\text{Republican}) &= \frac{.15}{.25} = .60 \end{aligned}$$

We can find the same thing by looking at the probabilities the other way: $P(\text{Democratic}) = P(\text{Democratic}|\text{Favor}) = .40$.

5. Suppose $\alpha = .01$ and you have 50 degrees of freedom.

   (a) What is the critical value of the t-statistic needed to reject H_0 in a two-sided test? (5 points)

   This comes from looking at the t-table at the $t_{.005}$ level: 2.678.

   (b) What is the corresponding critical value of Z? (5 points)

   We need to find the Z-statistic associated with have an area .005 in the right-hand tail. That is a Z-statistic somewhere in the range of $2.57 < Z \leq 2.58$. Accept any answer in this range. Include 2.58 as a full-credit answer, since it is the more lowest value of Z on the table that we know does not have more than .005 area in the tail.

   (c) If your t-statistic were 2.403, what would be the associated p-value? (5 points)

   According to the t-table, .01 of the area is in the right-hand tail for a t-statistic of 2.403. $p = 2 \times .01 = .02$. Take off 2 points if they forget to get the two-sided p-value.

6. Use the Stata output below to answer the following questions. The data are about countries.

```
. sum oilprod, detail

        oil production, thousands of barrels per day
-------------------------------------------------------------

      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs                 207
25%            0              0       Sum of Wgt.         207

50%        .9791                      Mean           409.7014
                        Largest       Std. Dev.      1333.729
75%        96.27           4172
90%         1023           9056       Variance        1778833
95%         2472           9764       Skewness       5.309319
99%         9056          10120       Kurtosis       35.10556
```

   (a) What is the measurement level of this variable? Which measures of central tendency and dispersion are appropriate? (5 points) This is an interval-level variable. All three measures

   of central tendency are appropriate. For measures of dispersion, we can use any of the four mentioned in lecture: variance, standard deviation, range and interquartile range.

   (b) Find the interquartile range. Explain the result in sentence form.(5 points)

This is easy since the relevant percentiles are provided. The IQR is 96.27, which means that the middle 50% of the countries in terms of oil production falls into the range of 0 to 96.27 thousands of barrels per day.

(c) After inspecting the summary output, including the measures of central tendency and dispersion, how would you describe the distribution of this variable? (5 points)

The variable has a strong positive skew, with very high relative frequency on 0 and other very low values. Half the cases are less than 1. The top 5% of the data range from 2472 to 10120, which is far away from the median of the data at .9791. The mean is not only higher than the median, but it is higher than the 75th percentile.

(d) Suppose you were going to perform a significance test in which the null hypothesis is that the population mean is 502.5. Choose an appropriate test statistic and write out the formula to calculate it, plugging in the appropriate numbers from the table above. You do not need to calculate the statistic. (5 points)

$$t = \frac{409.7 - 502.5}{\frac{1333.7}{\sqrt{207}}} = -1.00$$

The t-statistic is preferred, but the difference between the Z and t is probably not enough to make a difference.

7. A statistics professor has taught a particular course 20 times. The mean number of students is 25 with a standard deviation of 7.5. When asked about the "typical enrollment" in this course, the professor naturally answers in the form of a 90% confidence interval.

   (a) In thinking about this scenario, is the t-distribution or the Z-distribution the better approximation for the sampling distribution of the sample mean? Explain. (5 points)

   The answer is the *t*-distribution given the small sample size. We have to make a wider interval in order to depict a given level of confidence compared to large-*n* scenarios.

   (b) Find the 90% confidence interval for the mean.(5 points)

$$\begin{aligned} ci &= \bar{y} \pm t_{.05} \cdot se \\ &= 25 \pm 1.729\left(\frac{7.5}{\sqrt{20}}\right) \\ &= 25 \pm 1.729(1.677) \\ &= 25 \pm 2.899 \end{aligned}$$

   This makes the 90% confidence interval range from 22.10 to 27.90. Allow for rounding.

8. A math teacher decides to experiment with a new approach. Her class has 16 students. At the end of the year, the mean standardized test score for her students was 328.4 and the standard deviation was 8. This compares to the baseline expected mean score of 324. Perform a test of statistical significance on whether the new approach made a difference. How does the *p*-value compare to α=.05? (10 points)

H_0: $\bar{y} = 324$
H_A: $\bar{y} \neq 324$

Using a *t*-statistic is necessary due to the small sample size.

$$t = \frac{328.4 - 324}{\frac{8}{\sqrt{16}}} = \frac{4.4}{2} = 2.2$$

For $\alpha = .05$, the critical value of t needed to reject the null hypothesis is given on the *t*-table where the $t_{.025}$ column intersects the row for 15 degrees of freedom. The critical value is thus 2.131, which is less than the *t*-statistic we just calculated. We can reject the null hypothesis.

Based upon the table, the *p*-value associated with a *t*-statistic of 2.2 at 15 degrees of freedom is in the range: $2 \times .01 < p < 2 \times .025$, which is $.02 < p < .05$. Since $p <.05$, we can reject the null hypothesis.