Midterm 1 Review **Math 222**

Note: If you need to calculate a percentile or a z-value for either a normal or t-distribution, you may write your answer using the R commands `pnorm()`, `qnorm()`, `pt()`, and `qt()`. For example `pnorm(1.5)` is the percentile of the z-value 1.5 on a normal distribution, and `qt(0.95,7)` gives the t-value at the 95th percentile in a t-distribution with 7 degrees of freedom.

1. Attitudes toward drinking and behavior studies. Some of the methods we have covered are approximations rather than exact probability results. We have given rules of thumb for safe use of these approximations.
   (a) You are interested in attitudes toward drinking among the 75 members of a fraternity. You choose 30 members at random to interview. One question is "Have you had five or more drinks at one time during the last week?" Suppose that in fact 30% of the 75 members would say "Yes." Explain why you cannot safely use the $B(30, 0.3)$ distribution for the count X in your sample who say "Yes."

   (b) The National AIDS Behavioral Surveys found that 0.2% (thats 0.002 as a decimal fraction) of adult heterosexuals had both received a blood transfusion and had a sexual partner from a group at high risk of AIDS. Suppose that this national proportion holds for your region. Explain why you cannot safely use the Normal approximation for the sample proportion who fall in this group when you interview an SRS of 1000 adults.

2. The Harvard College Alcohol Study finds that 67% of college students around the country support efforts to "crack down on underage drinking." The study took a sample of almost 15,000 students, so the population proportion who support a crackdown is very close to p = 0.67. Suppose that the administration at HSC wants to know if the opinions of HSC are similar to those of college student's elsewhere. They survey an SRS of 200 HSC students and find that 140 support a crackdown on underage drinking.
   (a) What is the sample proportion who support a crackdown on underage drinking?

   (b) If in fact the proportion of all students at HSC who support a crackdown is the same as the national 67%, what is the probability that the proportion in an SRS of 200 students is as large or larger than the result of the administration's sample?

   (c) A writer in the student paper says that support for a crackdown is higher on your campus than nationally. Write a short letter to the editor explaining why the survey does not support this conclusion.

3. A study of the health of teenagers plans to measure the blood cholesterol level of an SRS of 13- to 16-year olds. The researchers will report the mean $\bar{x}$ from their sample as an estimate of the mean cholesterol level μ in this population.

   (a) Explain to someone who knows no statistics what it means to say that $\bar{x}$ is an "unbiased" estimator of μ.

   (b) The sample result $\bar{x}$ is an unbiased estimator of the population truth μ no matter what size SRS the study chooses. Explain to someone who knows no statistics why a large sample gives more trustworthy results than a small sample.

4. The scores of high school seniors on the ACT college entrance examination in 2003 had mean $\mu = 20.8$ and standard deviation $\sigma = 4.8$. The distribution of scores is only roughly Normal.

   (a) What is the approximate probability that a single student randomly chosen from all those taking the test scores 23 or higher?

   (b) Now take an SRS of 25 students who took the test. What are the mean and standard deviation of the sample mean score $\bar{x}$ of these 25 students?

   (c) What is the approximate probability that the mean score $\bar{x}$ of these students is 23 or higher?

   (d) Which of your two Normal probability calculations in (a) and (c) is more accurate? Why?

5. Suppose we are studying a group of students ($N = 80$) who are enrolled in an SAT test prep course. We want to see if their SAT math scores are significantly higher than the average for students across the country. For simplicity, assume that the national average SAT math score is 500 with a population standard deviation of $\sigma = 100$ (assume that our students have the same standard deviation too). Since we are assuming that we know σ, we can use normal distributions instead of t-distributions. We are testing

$$H_0 : \mu_{\text{prep students}} = 500$$
$$H_A : \mu_{\text{prep students}} > 500$$

Instead of aiming for a 5% significance level, what if we just decided that a sample mean of 520 or higher counts as significant enough to reject the null hypothesis.

(a) Find the probability α of a Type I error if we use this rejection region.

(b) Find the power of this test if $\mu_{\text{prep students}} = 525$.

(c) How large would N need to be in order to increase the power in part (b) to 80%?

6. Healthy bones are continually being renewed by two processes. Through bone formation, new bone is built; through bone resorption, old bone is removed. If one or both of these processes are disturbed, by disease, aging, or space travel, for example, bone loss can be the result. Osteocalcin (OC) is a biochemical marker for bone formation: higher levels of bone formation are associated with higher levels of OC. A blood sample is used to measure OC, and it is much less expensive to obtain than direct measures of bone formation. The units are milligrams of OC per milliliter of blood (mg/ml). One study examined various biomarkers of bone turnover. Here is a summary of the OC measurements on 31 healthy females aged 11 to 32 years who participated in this study:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
8.10	18.60	30.20	33.42	46.05	77.90

The sample standard deviation was 19.61.

(a) Display the data with a boxplot. Are there any outliers? Is the data skewed?

(b) Find a 95% confidence interval for the mean OC.

7. Suppose that an insurance company estimates that in the entire population of homeowners, the mean loss from fire is $\mu = \$300$ and the standard deviation of the loss is $\sigma = \$400$. What are the mean and standard deviation of the average loss for 10 policies? (Losses on separate policies are independent.)

8. A study that evaluated the effects of a reduction in exposure to traffic-related air pollutants compared respiratory symptoms of 283 residents of an area with congested streets with 165 residents in a similar area where the congestion was removed because a bypass was constructed. The symptoms of the residents of both areas were evaluated at baseline and again a year after the bypass was completed. 24 For the residents of the congested streets, 17 reported that their symptoms of wheezing improved between baseline and one year later, while 35 of the residents of the bypass streets reported improvement.

   (a) Find the two sample proportions.

   (b) Report the difference in the proportions and the standard error of the difference.

   (c) What are the appropriate null and alternative hypotheses for examining the question of interest? Be sure to explain your choice of the alternative hypothesis.

   (d) Find the test statistic. Construct a sketch of the distribution of the test statistic under the assumption that the null hypothesis is true. Find the P-value and use your sketch to explain its meaning.

   (e) Is no evidence of an effect the same as evidence that there is no effect? Use a plus-4 method 95% confidence interval to answer this question. Summarize your ideas in a way that could be understood by someone who has very little experience with statistics.

General Form of a Confidence Interval *estimate* $\pm$ *margin of error*, where

$$\textit{margin of error} = (\textit{critical value}) \times SE_{\textit{estimate}}$$

General form of a test statistic

$$\frac{\textit{estimate} - \textit{hypothesized valued}}{SE_{\textit{estimate}}}$$

Standard errors*

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}} \qquad SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$*$*When testing a null hypothesis about proportions, it is better to use the hypothesized population proportion* p_0 *rather than* $\hat{p}$ *in the formula for standard error for one-sample tests, and it is better to used the pooled proportion* $\hat{p}$ *rather than either individual sample proportion* $\hat{p}_1$ *or* $\hat{p}_2$ *in the formula for standard error in two-sample tests.*