Math 256/656
December 9, 2012

Final Exam Review 2012

Chapter 1

3. List and define the five elements of an inferential statistical analysis.
   (a) Identify population (Collection of all experimental units)
   (b) Identify variables
   (c) Collect sample data (Subset of population)
   (d) Inference about population based on sample
   (e) Measure of reliability
4. List the three major methods of collecting data and explain their differences.
   (a) Published source (book, newspaper, magazine,etc.)
   (b) Observational (survey, etc.)
   (c) Designed experiment
8. What is a representative sample? What is its value?
   - Exhibits characteristics typical of those possessed by the target population
   - Allows valid inferences to population
23. According to the American Lung Association, lung cancer accounts for 28% of all cancer deaths in the United states. A new type of screening for lung cancer, CT, has been developed. Medical researchers believe that CT scans are more sensitive than regular X-rays in pinpointing small tumors. The Cancer Center is currently conducting a clinical trial of 50,000 smokers nationwide to compare the effectiveness of CT scans with X-rays for detecting cancer. Each participating smoker is randomly assigned to one of the two screening methods and his progress is tracked over time. The age at which the scanning method first detects a tumor is the variable interest.
   (a) Identify the data collection method used by the cancer researchers
      - Designed Experiment
   (b) Identify the experimental units of the study
      - The smokers
   (c) Identify the type (quantitative or qualitative) of the variable measured.
      - quantitative (they are measuring the age)
   (d) Identify the population and the sample

- Population: Smokes of the US
- Sample: 50K Smokers

(e) What is the inference that will ultimately be drawn from the clinical trial?

- The difference in the age at which either method detects a tumor.

Chapter 2

159. Discuss conditions under which the median is preferred to the mean as a measure of central tendency.

- When there are outliers
- Skewness of population or sample

176. $mean = \dfrac{\sum x_i}{n}$
$median = \frac{n}{2} = \frac{14}{2} = 7$. The 7th largest number
$mode =$The number that appears the most in the data

Chapter 3

155. Use the symbols $\cup$, $\cap$, $|$, and ^{c}, to convert the following statements into compound events involving events A and B where:

$A = \{\text{You purchase a computer}\}$
$B = \{\text{You vacation}\}$

(a) You purchase or vacation: $A \cup B$

(b) You will not vacation: B^c

(c) You purchase and vacation: $A \cap B$

(d) Given that you vacation, you will not purchase: $A^c|B$

156. A sample space consists of four sample points where:
$P(S_1) = .2$
$P(S_2) = .1$
$P(S_3) = .3$
$P(S_4) = .4$

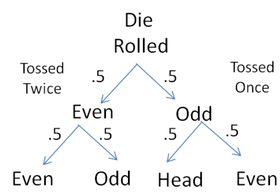
(a) Show that the sample points obey the two probability rules

$.2 + .1 + .3 + .4 = 1$
Each S_i is between 0 and 1

(b) If an event $A = \{S_1, S_4\}$, find $P(A)$

$P(A) = P(S_1) + P(S_2) = 0.2 + 0.4 = 0.6$

161. A fair die is tossed, and the up face is noted. If the number is even, the die is tossed again; if the number is odd, a fair coin is tossed. Consider the following events:
A : {A head appears on the coin}
B : {The die is tossed only one time}



(a) List the sample points in the sample space.

{Odd,Tail},{Odd,Head},{Even,Even},{Even,Odd}

(b) Give the probability for each of the sample points.

$P(\{Odd, Tail\}) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$
$P(\{Odd, Head\}) = \frac{1}{4}$
$P(\{Even, Even\}) = \frac{1}{4}$
$P(\{Even, Odd\}) = \frac{1}{4}$

(c) Find $P(A)$ and $P(B)$

$P(A) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$
$P(B) = \frac{1}{2}$

(d) Identify the sample points in A^c, B^c, $A \cap B$, and $A \cup B$.

$P(A^c) =$ Head does not appear on the coin.

$P(B^c) =$ The die is tossed twice

$P(A \cap B) =$ A head appears and the die is tossed once

$P(A \cup B) =$ A head appears or the die is tossed once

(e) Find $P(A^c), P(B^c), P(A \cap B), P(A \cup B), P(A|B), P(B|A)$.

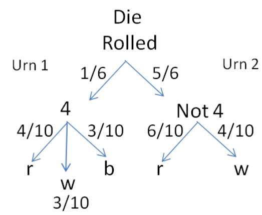
$P(A^c) = 1 - \frac{1}{4} = \frac{3}{4}$
$P(B^c) = 1 - \frac{1}{2} = \frac{1}{2}$
$P(A \cap B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$
$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{4} + \frac{1}{2} - \frac{1}{4} = \frac{1}{2}$
$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/4}{1/2} = \frac{1}{2}$
$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{1/4}{1/4} = 1$

(f) Are A and B mutually exclusive events? Independent events? Why?

$P(A|B) \neq 0$ Thus these are not mutually exclusive

$P(A \cap B) \neq P(A)P(B)$ Thus these are not independent.

162. A balanced die is thrown once. If a 4 appears, a ball is drawn from urn 1; otherwise a ball is drawn from urn 2. urn 1 contains 4 red, 3 white, and 3 black balls. Urn 2 contains 6 red and 4 white balls.



(a) Find the probability that a red ball is drawn.

$P(red) = \frac{1}{6} \cdot \frac{4}{10} + \frac{5}{6} + \frac{6}{10} = 0.567$

(b) Find the probability that urn 1 was used given that a red ball was drawn.

$P(urn\ 1|red) = \frac{P(urn\ 1\ and\ red)}{P(red)} = \frac{1/6 \cdot 4/10}{.567} = 0.118$

174. Researcher studying insects. Several insects are released. If the hormone has an effect, more insects will travel toward it rather than the control. Otherwise, the insects are equally likely to travel in either direction. Suppose the hormone under study has no effect, so it is equally likely that an insect will move toward the hormone or toward the control. Suppose that 5 insects are released.

(a) List or count the number of different ways the insects can travel.

Each insect can go 2 ways so there are 2^5 different ways.

(b) What is the chance that all 5 insects travel toward the hormone.

Each insect has probability (1/2) of going toward the hormone so the chance is $(1/2)^5 = 1/32$

(c) What is the change that exactly 4 travel toward the hormone.

$\binom{5}{4} \cdot (\frac{1}{2})^4(\frac{1}{2})^1 = \frac{5}{32}$

(d) What inference would you make if the event in part (c) actually occurs? Explain?

I would conclude that the hormone has an effect because the chance that all 4 of them would go to the hormone is very low (15% chance).

190. Playing Blackjack.

(a) What is the probability that the dealer will draw blackjack?

P(ace and face card)+P(face card and ace)= $\frac{4}{52} \cdot \frac{12}{51} + \frac{12}{52} \cdot \frac{4}{51}$

Chapter 4

124.

x	10	12	18	20
p(x)	.2	.3	.1	.4

(a) Calculate μ, σ^2, σ

$\mu = \sum xp(x) = 10 \cdot .2 + 12 \cdot .3 + 18 \cdot .1 + 20 \cdot .4 = 15.4$
$\sigma^2 = \sum (x-\mu)^2 p(x) = (10-15.4)^2(.2) + (12-15.4)^2(.3) + (18-15.4)^2(.1) + (20-15.4)^2(.4) = 18.44$
$\sigma = \sqrt{18.44} = 4.29$

(b) $P(x < 15) = P(10) + P(12) = .2 + .3 = .5$

Chapter 5

112. Consider the continuous random variables that follow. Give the probability distribution (uniform, normal, or exponential) that is likely to best approximate the distribution of the random variable:

    (a) Score on an IQ test: Normal

    (b) Time (in minutes) waiting in line: Exponential

    (c) Amount of liquid (in ounces) dispensed: $\boxed{\text{??? Why wouldn't this be uniform?}}$ Normal

127. 8 % of African Americans is known to carry sickle-cell trait. If 1,000 African Americans are sampled at random, what is the approximate probability that:

    $n = 1000$, $p = 0.08$, $\mu = np = 1000 \cdot (0.08) = 80$, $\sigma = \sqrt{npq} = \sqrt{80 \cdot (.08) \cdot .92} = 8.6$

    (a) More than 175 carry the trait?

    $P\left(z > \frac{x-\mu}{\sigma}\right) = P\left(z > \frac{175-80}{8.6}\right) = P(z > 11.04) = 0$

    (b) Fewer than 140 carry the trait?

    $P\left(z < \frac{140-80}{8.6}\right) = P(z < 6.98) = 1$

Chapter 6

48. Describe how you would obtain the simulated sampling distribution of a sample statistic.

    Repeat a sampling experiment a very large number of times and get resulting distribution of sample statistic.

50. **True or False.** The sampling distribution of $\bar{x}$ is normally distributed, regardless of the size of the sample n.

    True

52. The standard deviation (standard error) of the sampling distribution for the sample mean, $\bar{x}$, is equal to the standard deviation of the population from which the sample was selected, divided by the square root of the sample size. That is: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

    (a) As the sample size is increased, what happens to the standard error of $\bar{x}$? Why is this property important?

The standard error decreases by a factor of $\frac{1}{\sqrt{n}}$; this property tells us that the more data we have the less error we have.

(b) Suppose a sample statistic has a standard error that is not a function of the sample size. In other words, the standard error remains constant as n changes. What would this imply about the statistic as an estimator of a population parameter.

It is a poor estimate.

(c) Suppose another unbiased estimator (call it A) of the population is a sample statistic with a standard error equal to $\sigma_A = \frac{\sigma}{\sqrt[3]{n}}$
Which of these sample statistics, $\bar{x}$ or A is preferable as an estimator of the population mean? Why?

$\bar{x}$ is preferred since $\sqrt{n} > \sqrt[3]{n}$, thus $\sqrt{n}$ will give us a smaller error.

(d) Suppose that the population standard deviation σ is equal to 10 and that the sample size is 64. Calculate the standard errors of $\bar{x}$ and A. Assuming that the sampling distribution of A is approximately normal, interpret the standard errors. Why is the assumption of normality unnecessary for the sampling distribution of $\bar{x}$?

$\sigma_{\bar{x}} = \frac{10}{\sqrt{64}} = 1.25$ and
$\sigma_A = \frac{10}{\sqrt[3]{64}} = 2.5$, We have a larger error with A. Because of the Central Limit Theorem, we do not have to assume normality.

58. Suppose x equals the number of heads observed when a single coin is tossed; that is, $x = 0$ or $x = 1$. The population corresponding to x is the set of 0's and 1's generated when the coin is tossed repeatedly a large number of times. Suppose we select $n = 2$ observations from this population. (that is, we toss the coin twice and observe two values of x)

(a) List the three different samples (combinations of 0's and 1's) that could be obtained.

$\{0, 0\}, \{0, 1\}, \{1, 1\}$

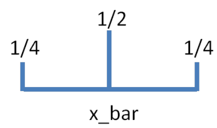
(b) Calculate the value of $\bar{x}$ for each of the samples.

$\{0, 0\} = \frac{0+0}{2} = 0$
$\{0, 1\} = \frac{0+1}{2} = \frac{1}{2}$
$\{1, 1\} = \frac{1+1}{2} = 1$

(c) List the values that $\bar{x}$ can assume, and find the probabilities of observing these values.

$\bar{x} = 0 \Rightarrow \frac{1}{4}$
$\bar{x} = \frac{1}{2} \Rightarrow \frac{2}{4}$
$\bar{x} = 1 \Rightarrow \frac{1}{4}$

(d) Construct a graph of the sampling distribution of $\bar{x}$



Chapter 7

110. Interpret the phrase "95% confident" in the following statement: "We are 95% confident that the proportion of all PCs with a computer virus falls between .12 and .18."

    We have used a procedure that would work 95% of the time in the long run

111. In each of the following instances, determine whether you would use a $z-$statistic or a $t-$statistic (or neither) to form a 95% confidence interval for μ; then look up the appropriate z or t value.

    (a) Random sample size of $n = 21$ from a normal distribution with unknown mean μ and standard deviation σ.

        $t-$statistic, 2.0896

    (b) Random sample size of $n = 175$ from a normal distribution with unknown mean μ and standard deviation σ.

        $z-$statistic $\Rightarrow \alpha/2 = .05/2 = 0.025 \Rightarrow .5 - .025 = .475 \Rightarrow z = 1.96$

    (c) Random sample size of $n = 12$ from a normal distribution with unknown mean μ and standard deviation $\sigma = 5$.

        $t-$statistic, 2.228

    (d) Random sample size of $n = 65$ from a normal distribution about which nothing is known.

        $z-$statistic, 1.96

    (e) Random sample size of $n = 8$ from a normal distribution about which nothing is known.

        You cannot use either one of these options. Sample size is too small and we know nothing about it.

115. A random sample of 225 measurements is selected from a population, and the sample mean and standard deviation are $\bar{x} = 32.5$ and $s = 30.0$, respectively.

    (a) Use a 99% confidence interval to estimate the mean of the population μ.

        $z_{\alpha/2} = z_{0.01/2} = z_{0.005} = 2.554$,

        CI$= \bar{x} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{2}}\right)$

$$\text{CI}= 32.5 \pm 2.554 \left(\frac{30.0}{\sqrt{225}}\right) \Rightarrow 32.5 \pm 5.108$$

Chapter 8

129. *Complete the following statement:* The smaller the $p-$value associated with a test of hypothesis, the strong the support for the **alternative** hypothesis.

130. Specify the differences between a large-sample and small-sample test of hypothesis about a population mean μ. Focus on the assumptions and test statistics.

    For n large: use $z-$test, only assume random sample is selected from the target population

    For n small: use $t-$test, assume the sample has a normal distribution

137. A random sample of 41 observations from a normal population possessed a mean of $\bar{x} = 88$ and a standard deviation $s = 6.9$

    (a) Test $H_0 : \sigma^2 = 30$ against $\sigma^2 > 30$. Use $\alpha = 0.05$.

    Test statistic: $\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{40 \cdot 6.9^2}{30} = 63.48$
    Rejection Region: Reject if $\chi^2 > \chi^2_{0.05}$
    $\chi^2_\alpha = \chi^2_{0.05} = 55.75$
    Since $\chi^2 > \chi^2_{0.05}$ we reject our null hypothesis

    (b) Test $H_0 : \sigma^2 = 30$ against $\sigma^2 \neq 30$. Use $\alpha = 0.05$.

    Test statistic: $\chi^2 = \frac{(n-1)s^2}{\sigma^2} = 63.48$
    Rejection region: Reject if $\chi^2 < \chi_{1-\frac{\alpha}{2}}$ or if $\chi^2 > \chi^2_{0.05/2}$
    $\chi^2_{0.975} = 24.4331$ and $\chi^2_{0.025} = 59.3417$ Since $\chi^2 > \chi^2_{\alpha/2}$ we reject the null hypothesis.

141. In a study, 12 healthy college students deprived of one night's sleep received an array of tests intended to measure their thinking time. The overall test scores of the sleep-deprived students were compared with the average score expected from students who received their accustomed sleep. Suppose the overall scores of the 12 sleep-deprived students had a mean of $\bar{x} = 63$ and a standard deviation of 17.

    (a) Test the hypothesis that the true mean score of sleep-deprived subjects is less than 80, the mean score of subjects who received sleep prior to taking the test. Use $\alpha = 0.05$.

    $H_0 : \mu = 80, \;\; H_a : \mu < 80$
    Test Statistic: $t = \frac{\mu - \mu_0}{s/\sqrt{n}} = \frac{63-80}{17/\sqrt{12}} = -3.46$
    Rejection Region: Reject if $t < -t_\alpha$ with $(12 - 1)df$ ie. Reject if $t < -1.812$.

    Thus we reject null and conclude H_a.

    (b) What assumption is required for the hypothesis test of part **a** to be valid?

We must assume that the population is normal.

Chapter 9

106. List the assumptions necessary for each of the following inferential techniques:

(a) Large-sample inferences about the difference $(\mu_1 - \mu_2)$ between population means, using a two-sample $z-$statistic.

- 2 good independent random samples

(b) Small-sample inferences about $(\mu_1 - \mu_2)$, using an independent samples design and a two-sample $t-$statistic

- 2 good independent random samples
- Each population is normally distributed with a common variance

(c) Small-sample inferences about $(\mu_1 - \mu_2)$, using a paired difference design and a single-sample $t-$ statistic to analyze the differences.

- 2 good random pair selection
- normal population differences

(d) Large-sample inferences about the differences $(p_1 - p_2)$ between binomial proportions, using a two-sample $z-$statistic

- n and p large for CLT

(e) Inferences about the ratio $\frac{\sigma_1^2}{\sigma_2^2}$ of two population variances, using an $F-$test.

- 2 good independent samples
- normal population

107. For each of the following, identify the target parameter as $\mu_1 - \mu_2$, $p_1 - p_2$ or $\frac{\sigma_1^2}{\sigma_2^2}$.

(a) Comparison of average SAT scores of males and females

$\mu_1 - \mu_2$

(b) Difference between mean waiting times at supermarket checkout lanes

$\mu_1 - \mu_2$

(c) Comparison of proportions of Democrats and Republicans who favor a law

$p_1 - p_2$

(d) Comparison of variation in salaries of NBA players

$$\frac{\sigma_1^2}{\sigma_2^2}$$

(e) Difference in dropout rates of college student athletes and regular students.

$p_1 - p_2$

111. Two independent random samples are taken from two populations. The results are summarized in the following table:

Sample 1	Sample 2
$n_1 = 135$	$n_2 = 148$
$\bar{x}_1 = 12.2$	$\bar{x}_2 = 8.3$
$s_1^2 = 2.1$	$s_2^2 = 3.0$

(a) Form a 90% Confidence Interval for $(\mu_1 - \mu_2)$

$\alpha = 0.10 \Rightarrow \alpha/2 = 0.05.$

CI$= (\mu_1 - \mu_2) \pm z_{\alpha/2}\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$

CI$= (12.2 - 8.3) \pm 1.645\sqrt{\dfrac{2.1}{135} + \dfrac{3}{148}}$

CI$=3.9 \pm 0.31$

(b) $H_0 : (\mu_1 - \mu_2) = 0, H_a : (\mu_1 - \mu_2) \neq 0$ and $\alpha = 0.01$

Test Statistic: $\dfrac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = 20.6$

Rejection region: Reject if $z < -z_{\alpha/2}$ or if $z > z_{\alpha/2}$

Since $z_{0.01/2} = 2.53$ Reject null hypothesis in favor of alternative hypothesis

(c) What sample size would be required if you wish to estimate $(\mu_1 - \mu_2)$ within .2 with 90% confidence, assume that $n_1 = n_2$.

We want the second part of our confidence interval to equal 0.2 so we set it equal to 0.2 and solve for $n_1 = n_2 = n$.

$z_{\alpha/2}\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}} = 0.2$. Then solving for n we get that $n = 345$.

Chapter 11

110. Explain the difference between a probablistic model and a deterministic model.

There is no error term in the deterministic model where as there is an error term in the probablistic model.

111. Give the general form of a straight-line model. deterministic model.

$y = \beta_0 + \beta_1 x + \epsilon$

112. Outline the five steps in a simple linear regression analysis:

1. Hypothesize the deterministic component of the model that relates the mean $E(y)$ to the independent variable x.

2. Use the sample data to estimate unknown parameters in the model.
3. Specify the probability distribution of the random-error term and estimate the standard deviation of this distribution
4. Statistically evaluate the usefulness of the model
5. When satisfied that the model is useful, use it for prediction, estimation and other purposes.

113. *True or False.* In simple linear regression, about 95% of the $y-$values fall within $2s$ of their respective predicted values.

    True.

119. Data is given to find out if the number of games won by a baseball team is linearly related to the team's batting average. There were 14 teams.

    (a) If you were to model the relationship between mean (or expected) number of games won by a major league team and the team's batting average x, using a straight line, would you expect the slope of the line to be positive or negative? Explain.

    Positive, as the batting average increases, the number o games won should increase.

    (c) An SAS printout of the simple linear regression is shown below. Find the estimates of the β's on the printout and write the equation of the least squares line.

    Parameter estimate, intercept: -12.62556

    Parameter estimate, BATAVG: 0.36269

    The equation would be $y = -12.63 + 0.36269x$

    (e) Interpret the estimates of β_0 and β_1 in the words of the problem.

    Slope(β_1): For each additional hit per 1000 at bats, we estimate the number of games won will increase by 0.36269.

    (f) Conduct a test (at $\alpha = 0.05$) to determine whether the mean (or expected) number of games won by a major league baseball team is positively linearly related to the team's batting average.

    $H_0 : \beta_1 = 0,\ H_a : \beta_1 > 0$

    Test statistic: $\dfrac{\hat{\beta}_1 - 0}{S_{\hat{\beta}_1}} = \dfrac{0.36269}{0.24591} = 1.47$

    Rejection Region: Reject if $t < -t_{\alpha/2}$ or if $t > t_{\alpha/2}$ (with $n - 2 = 14 - 2 = 12$ df). $t_{0.025} = 2.179$

    Thus we fail to reject H_0 and conclude that the number of games won is not positively related to the team's batting average.