STATISTICS 216 Name: ______________________
Review Final Exam - Spring 2015

1. An ongoing study in Montana is investigating the incidence of whirling disease in different trout populations. The data recorded includes the river (Gallatin, Madison, Ruby, etc), species of trout (Rainbow, Brown, Bull, etc), length and weight of the fish, and whether the fish has whirling disease or not. The cases and the variables in the study. For each variable, identify it as categorical or quantitative.

   (a) Cases
   Each trout is a case

   (b) Variables
   River - categorical, Species - categorical, Length of fish - quantitative, Weight of fish - quantitative, Diseased of not - categorical

2. Give the four possibilities for scope of inference and describe the study design for each.
   Causal to the Population for a study with both random assignment (experiment) and a random sample.
   Causal to the Sample for a study with random assignment (experiment) but no random sample.
   Association to the Population for a study with no random assignment (observation study) but that did use a random sample. Association to the Sample for a study with no random assignment (observational study) that did not have a random sample.

3. A college official conducted a survey to estimate proportions of students currently living in dormitories on campus who prefer each of the following room types: single, double, or multiple (more than two people). Five thousand students live in dormitories on campus. A **random sample of 500 students** was selected to receive the survey.
   True or False: The survey results cannot be generalized to the population of students currently living in dormitories because it was sent to only 500 students.

   A. True

   B. False

4. A researcher investigated the impact of a particular herbicide on fish. He randomly assigned 60 healthy fish to either be exposed or not be exposed to the herbicide, with 30 fish in each group. The fish exposed to the herbicide showed higher levels, on average, of an enzyme associated with cancer. Suppose the difference in average enzyme level between the two groups of fish was determined NOT to be statistically significant (i.e., strong evidence of a difference between the groups was not found). Indicate whether the statement provided is valid or invalid.

   (a) The sample size may have been too small to detect a statistically significant difference even if the herbicide did have an impact.

      A. Valid, larger sample size will result in a smaller p-value is the sample statistic remains close to the same.

      B. Invalid

(b) The research results demonstrated strong evidence that the herbicide does not cause higher levels of the enzyme.

A. Valid

B. Invalid (p-value only gives evidence AGAINST the null, not FOR the null)

5. Just how accurate are the weather forecasts we hear every day? The table below compares the daily forecast with a city's actual weather for 365 randomly selected days.

		Actual Weather	
		Rain	No Rain
Forecasted	Rain	15	76
Weather	No Rain	45	229

(a) What is the probability day is forecasted not rain?
274/365

(b) What is the probability that it will rain on a day it was forecasted to rain?
15/91

(c) What is the probability it was forecasted to NOT rain on a day it actually did rain?
45/60

6. You want to see what proportion of MSU students smoke. Write a sampling plan that uses

(a) an unbiased simple random sample.
Use a computer program to take a random sample of student ID numbers.

(b) an unbiased stratified random sample.
Separate the MSU student body by gender, then take a random sample of males by selecting ID numbers at random using software, then do the same for females.

(c) a biased sample.
Tons of options here. Voluntary surveys, convenience surveys (ask everyone that walks by the library or all your friends) would be easiest to use.

7. Researchers reported that unmarried men are more likely to suffer from clinical depression than married men. These findings were based on the marriage histories of 700 men suffering from depression and 600 men not suffering from depression. What type of study is this? Can the researchers conclude that marriage causes increased happiness?
Observational Study. No, because we cannot conclude causation if no random assignment is used.

8. An educational researcher was interested in examining the effect of the teaching method and the effect of the teacher on students scores on a reading test. Suppose there are three different teaching methods (A, B, and C). Students were randomly assigned to a teaching method.

(a) What is the response variable?
Reading test scores

(b) What is the explanatory variable?
Teaching method (3 levels)

(c) Is this an experiment or an observational study?
Experiment since the students are randomly assigned to treatments.

(d) What summary measure should be used to compare methods A and B?
Difference in means since we are comparing a quantitative variable across two groups and this is not a matched pairs study.

9. Insurance companies track life expectancy information to assist in determining the cost of life insurance policies. The insurance company knows that, last year, the life expectancy of its policyholders was 77 years. They want to know if their clients this year have a longer life expectancy, on average, so the company randomly samples some of the recently paid policies to see if the mean life expectancy of policyholders has increased. The insurance company will only change their premium structure if there is evidence that people who buy their policies are living longer than before. The data is listed below.
70, 72, 73, 74, 75, 76, 76, 77, 78, 79, 79, 79, 81, 81, 81, 83, 83, 84, 85, 86

   (a) Find the mean of the data. 78.6

   (b) Find the median of the data. 79

   (c) Is the data right-skewed, left-skewed, or symmetric? How do you know? Symmetric because the mean and median are approximately the same.

   (d) In this data set, Q1 = 75.5 and Q3 = 82. Are there any outliers in this dataset? Explain.
   IQR = 82 - 75.5 = 6.5. 1.5IQR = 9.75, Lower fence: 65.75, No low outliers since all values are greater than the lower fence. Upper fence: 91.75, No high outliers since all values are less than the upper fence.

10. List 4 properties of correlation.
No units. Must be between -1 and 1. Farther from 0 means stronger correlation. Value of 0 means random scatter (or non-linear relationship). Only measures strength of a linear relationship between two quantitative variables. Sign gives direction of scatterplot. Non-resistant, meaning it is influenced my outliers. Unaffected by changes in scale (so measuring in feet vs meters will not change correlation).

11. The regression line for predicting body fat percentage from neck circumference is

$$\widehat{BodyFat} = -47.9 + 1.75 \times Neck$$

   (a) What body fat percent is predicted for a person with a neck circumference of 35?
   $\widehat{BodyFat} = -47.9 + 1.75 \times 35 = 13.35$

   (b) Interpret the slope in the context of the problem.
   (Assuming neck circumference in measured in centimeters) For every 1 centimeter increase in neck circumference, we expect body fat percentage to increase by 1.75%.

   (c) One of the men in the study had a neck circumference of 38.7 cm and a body fat percentage of 11.3. Find the residual for this man.
   $\widehat{BodyFat} = -47.9 + 1.75 \times 38.7 = 19.825$, Residual = Observed - Predicted = 11.3 - 19.825 = -8.527.

12. Every year favorite songs compete to be on a Top 200 list based upon sales and rankings by the experts in the music industry. These songs have many characteristics, such as song length and beats per minute, which vary from category to category in the music industry. A disc jockey wondered if the number of beats per minute in songs classified as dance music were lower than the beats per minute in the songs that are ranked on a Top 200 list from 2001. A

random sample of songs from each group was selected and the data are summarized below.

Group	Mean	SD	n
Dance	119.7	1.74	18
Top 200	120.6	1.72	15

(a) What type of test should be used?

A. Two-sample Z-test

B. One-sample Z-test

C. Two-sample T-test - independent groups and working with means

D. One-sample T-test

(b) What are the null and alternative hypotheses?
$H_0 : \mu_d - \mu_t = 0$ versus $H_a : \mu_d - \mu_t < 0$

(c) Find the test statistic. Give the degrees of freedom.
Degrees of freedom = 14, T = $(119.7 - 120.6)/\sqrt{1.74^2/18 + 1.72^2/15} = -1.489$

(d) Estimate the p-value using your test statistic.

A. Below 0.025

B. Around 0.025

C. Above 0.025 because the test statistic is less than 2 so there will be more than 0.025 in the left tail.

(e) Write a decision and conclusion at the $\alpha = 0.01$ level.
With a p-value above 0.025 (the actual p-value is 0.079) which is higher than the significance level, there is insufficient evidence to conclude the beats per minute in the dance group are lower than the beats per minute in the top 200 group. We would fail to reject the null hypothesis that the two types of music have the same mean beats per minute.

(f) Interpret the p-value (which was found to be 0.079 using technology) in the context of the problem.
There is a 7.9% chance of seeing a random sample of 18 dance songs have at least 0.9 fewer beats per minute than a random sample of 15 top 200 songs is there really is no difference in the average beats per minute between the two groups.

13. Most people are definitely dominant on one side of their body - either right or left. For some sports being able to use both sides is an advantage, such as batting in baseball or softball. In order to determine if there is a difference in strength between the dominant and non-dominant sides, 19 switch-hitting members of some school baseball and softball teams were asked to hit from both sides of the plate during batting practice. The longest hit (in feet) from each side was recorded for each player. The plot of the difference in hits is roughly symmetric with no outliers. The data are summarized below.

Group	Mean	SD
Dominant	154.8	9.93
Non-dominant	129.7	9.60
Differences	25.1	2.31

(a) What type of test should be used? Explain.

A. Two-sample Z-test

B. One-sample Z-test

C. Two-sample T-test

D. One-sample T-test - this is paired data so we will analyze the data as a single mean (of the differences).

(b) Are the appropriate assumptions and conditions met to create a 90% confidence interval? Be sure to check each.
This is paired data. It seems reasonable that one hitter should not effect another so the independence assumption appears valid. The Large Enough Sample assumption is valid since we are told the data is roughly symmetric with no outliers and our sample size is 19. The randomization assumption is likely violated as it does not say the hitters were randomly assigned which side to hit with first nor does it say they were randomly selected.

(c) Create and interpret a 90% confidence interval. Use $t^{\star} = 1.734$
$25.1 \pm 1.734(2.31/\sqrt{19}) = (24.2, 26.0)$ We are 90% confident the true increase in length of hit for dominant versus non-dominant hits is between 24.2 and 26.0 feet.

(d) What does it mean to be "90% confident" in the interval? Approximately 90% of random samples of 19 hitters will create confidence intervals that contain the true average difference in hit lengths between dominant and non-dominant hands.

(e) Does this provide evidence there is a difference in the maximum length of a hit between dominant and non-dominant hands?
Yes because 0 is not in the interval.

14. Imagine you have a barrel that contains thousands of candies that are several different colors. A student takes a random sample of 20 candies and finds that 35% of candies are yellow.

(a) Which of the following provides an plausible range for the percentage of yellow candies that you can expect with 95% confidence?

A. About 0% to 100%

B. About 15% to 55% ($SE = \sqrt{0.35 \times (1 - 0.35)/20} = 0.106654$, $\text{CI} = 0.35 \pm 1.96 \times 0.106654 = (0.14, 0.56)$

C. About 30% to 40%

D. About 35% to 65%

(b) Explain what 95% confidence means.
In the long run, 95% of random samples of 20 candies will create a 95% confidence interval that contains the true proportion of yellow candies produced by the manufacturer.

(c) How large a sample would be needed to give a 99% confidence interval with the same margin of error as (a)?
From (a), $\text{ME} = 1.96 \times 0.106654 = 0.209041$. $n = (z^{\star}/ME)^2 \times \hat{p} \times (1 - \hat{p}) = (2.576/0.209041)^2 \times 0.35 \times 0.65 = 34.55$ or 35 candies.

15. In past years, adult bass (a species of fish) in Silver Lake were known to have an average length of 12.3 inches with a standard deviation of 3 inches. People who fish often in Silver Lake claim that this year they are catching smaller adult bass than usual. A research group randomly sampled 100 adult bass from Silver Lake and found the mean of this sample to be 11.2 inches, and they estimated the **standard error** of the mean as 0.3 inches.

(a) If the fish are currently no different in size from past years, how unusual is the sample data?

A. Very unusual (T = (11.2 - 12.3)/0.3 = -3.667. This is well beyond the cutoff of 2 SE from the mean so it should be considered unusual.)

B. Not unusual

(b) Which of the following best matches the reason for your answer to the question above?

A. Because this years mean was more than three standard errors below the mean length of adult bass in past years.

B. Because the researchers did not test more than one sample, which is not accurate enough to make a generalization.

C. Because the average length of fish found this year still falls within the typical range of fish from the previous years.

D. None of the above.

16. A newspaper article claims that the average age for people who receive food stamps in a community is 40 years. A local researcher believes that the average age is less than that. The researcher takes a random sample of 100 people in the community who receive food stamps, and finds their average age to be 39.2 years, which is statistically significantly lower than the age of 40 stated in the article (p-value $< .05$). Indicate for each of the following interpretations whether they are valid or invalid.

(a) The statistically significant result indicates that the majority of people who receive food stamps is younger than 40.

A. Valid

B. Invalid

(b) An error must have been made. This difference in means (39.2 vs. 40 years) is too small to be statistically significant.

A. Valid

B. Invalid

17. Explain the difference between a sample and a bootstrap resample.
A sample comes from the population. A bootstrap resample is the same size as the sample but solely comes from the sample, taken with replacement, so some values are not included and some values are repeated.

18. Explain the difference between $t^\star$ and the T test statistic (or the difference between $z^\star$ and the Z test statistic).
$t^\star$ and $z^\star$ are multipliers and are used in confidence intervals from theoretical distributions. In the standard Normal and t distributions, they tell where the center __% of data lies. With the exception of the sample size for t distributions, $t^\star$ and $z^\star$ do not depend on the data at all. (Be sure to write down the $z^\star$ values for 80, 90, 95, 98 and 99% confidence!) T and Z are test statistics and only used in hypothesis testing. These are calculated using the data and the null hypothesis. Once we calculate these values, we look them up on the standard normal or t distributions to calculate a p-value.

19. Consider an experiment where a researcher wants to study the effects of two different exam preparation strategies on exam scores. Twenty students volunteered to be in the study, and were randomly assigned to one of two different exam preparation strategies, 10 students per strategy. After the preparation, all students were given the same exam (which is scored from

0 to 100). The researcher calculated the mean exam score for each group of students. The mean exam score for the students assigned to preparation strategy A was 5 points higher than the mean exam score for the students assigned to preparation strategy B. What can the researcher do to estimate the true effect (size of the difference between groups) for students who use the two exam preparation strategies?

A. The researcher can produce a p-value based on the difference in mean exam scores between the two strategies.

B. The researcher can produce a confidence interval for the mean exam score for the strategy that produced better results.

C. The researcher can produce a confidence interval for the difference in mean exam scores between the two strategies because these are independent samples.

D. The researcher can produce a confidence interval for the mean of the difference in exam scores between the two strategies because the data is a paired sample.

20. Suppose that in a random sample of 500 Bozeman residents, 61% indicated they own a dog, while in a random sample of 500 Missoula residents, 53% indicated they owned a dog.

(a) What are the null and alternative hypotheses?
Without more information we must assume a two-tailed test. Let 1 = Bozeman, 2 = Missoula. $H_0 : p_1 = p_2, H_A : p_1 \neq p_2$

(b) How many of those surveyed from Bozeman owned a dog? How many from Missoula?
Bozeman: $0.61 \times 500 = 305$
Missoula: $0.53 \times 500 = 265$

(c) What is the combined proportion of people who owned a dog?
$\widehat{p_m} = \frac{305+265}{500+500} = 0.57$

(d) What is the standard error of the difference in the sample proportions under H_0?
$SE = \sqrt{0.57 \times 0.43 \times (1/500 + 1/500)} = 0.0313113$

(e) Find the test statistic.
$Z = \frac{0.61-0.53}{0.0313113} = 2.555$

(f) What distribution should be used to calculate the p-value?
Standard normal since we are working with proportions.

(g) Based on the test statistic, what can you say about the p-value?

A. It is less than 0.025

B. It is less than 0.05, because the test statistic is greater than 2 so there is less than 0.025 above the test statistic and less than 0.025 below the negative of the test statistic, so there has to be less than 0.05 in both tails combined (and we care about both tails because HA is a $\neq$ sign).

C. It is between 0.025 and 0.05

D. It is greater than 0.05

E. It is greater than 0.025

Note: p-value (from app using N(0,1)) = 2(0.0054) = 0.0108, but based on the Empirical rule we only know it is less than 0.05.

(h) Give and interpret a 99% confidence interval for the true difference in proportion of dog owners between Missoula and Bozeman.
CI $= (0.61-0.53) \pm 2.576 \times \sqrt{0.61 \times (1-0.61)/500 + 0.53 \times (1-0.53)/500} = (-0.0004, 0.1604)$
We are 99% confident the true difference in proportion of dog owners between Bozeman and Missoula (Bozeman - Missoula) is between -0.0004 and 0.1604.

(i) Can we say with 99% confidence that one city has higher dog ownership rates? Explain.
No. No difference between the cities proportion of dog owners (or a difference of 0) is included in the confidence interval.

(j) Does your confidence interval agree with your test decision? Explain.
Yes. At the 1% significance level, we would fail to reject a difference in the dog ownership rates between the two cities because the p-value is greater than 0.01.

(k) Interpret Type I and Type II error in the context of the problem.
Type I: Concluding the dog ownership rates are different when in fact there is no difference in the proportion of residents who own dogs between Missoula and Bozeman.
Type II: Concluding the dog ownership rates are the same (or failing to conclude they are different), when in fact there is a difference in the proportion of residents who own dogs between Missoula and Bozeman.

21. Explain how to get a Bootstrap confidence interval using the

(a) 2SE Method
Use the bootstrap distribution to estimate SE (on the big plot), then use CI = statistic $\pm$ 2SE.

(b) Percentile Method
Find the values that cut-off or bound the middle __% of the bootstrap resample statistics. These values are the upper and lower bounds of the confidence interval.

22. Explain how to get a p-value using a Randomization test.
A p-value is the probability of an observed result or something more extreme if the null hypothesis is true. In a randomization distribution, where each dot represent a sample statistic that could happen if the null is true, the p-value is found by find the proportion of randomization samples (dots) that are at least as extreme (as defined by the alternative hypothesis) as the observed result. (# of randomization samples at least as extreme as the observed result / total number of randomization samples).

23. For each situation below (which is a repeat of a previous problem), describe (i) how a randomization sample would be created, what value would be plotted from each, and where the randomization distribution would be centered and (ii) how a bootstrap resample would be created, what value would be plotted from each, and where the bootstrap distribution would be centered.

(a) 12 Analyzing the difference in average beats per minute between Dance songs and Top 200 songs. Independent groups with a quantitative response, so this is a difference in means situation.
(i)Randomization sample: Combine all original responses (beats per minute in the song), shuffle, and deal into two new groups the sample sizes as the original study (18 songs into the Dance category and 15 into the Top 200 group). Record the average beats per minute in each shuffled song category, then plot the difference. The null distribution will

be centered at 0 because we assume $H0 : \mu_1 - \mu_2 = 0$ is true.
(ii) Bootstrap sample: Randomly draw with replacement 18 times from the original list of Dance songs and 15 times from the original list of Top 200. Record the average beats per minute in each resampled song category, then plot the difference. The bootstrap distribution will be centered at 0.9 because that the is observed result, or observed difference in mean beats per minute between Dance and Top 200 songs in the original sample ($\overline{x_1} - \overline{x_2} = 120.6 - 119.7 = 0.9$).

(b) 13 Analyzing the average difference in hit distance between batting with a dominant and non-dominant hand. This is paired data, so use a single mean type analysis on the differences.
(i) Randomization sample: Shift the original data so that the mean of the shifted data is the null value 0 by subtracting 25.1 from each hit distance. Then randomly draw with replacement from the shifted data 19 times. Plot the average difference in hit distance between dominant and non-dominant handed swings of the shifted resample. The null distribution will be centered on the null value 0 because we assume $H_0 : \mu_d = 0$ is true.
(ii) Bootstrap sample: Randomly draw with replacement 19 times from the original list of differences in hit distance between dominant and non-dominant swings. Then plot the mean difference in hit distance of the resample. The bootstrap distribution is centered on the observed mean of the differences of $\overline{x_d} = 25.1$.

(c) 14 Analyzing the proportion of yellow candies. One categorical variable so this is a single proportion situation.
(i)Randomization sample: Spin a spinner with probability of success equal to the null value (p_0) the sample size (20) times. Record the proportion of successes (in this case, yellow candies). The null distribution would be centered at the null value, p_0 because we assume the $H_0 : p = p_0$ is true.
(ii)Bootstrap sample: Randomly draw with replacement 20 times from the original list of candy colors. Record and plot the proportion of successes (yellow candies). The bootstrap distribution would be centered at the observed result ($\hat{p} = 0.35$)

(d) 15 Analyzing the average weight of bass. One quantitative variable so this is a single mean situation.
(i) Randomization sample: Shift the original data so that the mean of the shifted data is the null value 12.3 by adding 1.1 to each fish weight. Then randomly draw with replacement from the shifted data 100 times. Plot the average fish weight of the shifted resample. The null distribution will be centered on the null value 12.3 because we assume $H_0 : \mu = 12.3$ is true.
(ii) Bootstrap sample: Randomly draw with replacement 100 times from the original list of fish weights. Then plot the mean fish weight of the resample. The bootstrap distribution is centered on the observed mean of $\overline{x} = 11.2$.

(e) 20 Analyzing the difference in proportion of dog owners between Bozeman and Missoula. Two categorical variables so this is a difference in proportions situation.
(i)Randomization sample: Combine all original responses (whether the person owned a dog or not), shuffle, and deal into two new groups the sample sizes as the original study (500 in Missoula and 500 in Bozeman). Record the proportion of successes (owning a dog) in each shuffled city, then plot the difference. The null distribution will be centered at 0 because we assume $H0 : p_1 - p_2 = 0$ is true.
(ii) Bootstrap sample: Randomly draw with replacement 500 times from the original list of Missoula residents and 500 times from the original list of Bozeman residents. Record

the proportion of successes (owning a dog) in each resampled city, then plot the difference. The bootstrap distribution will be centered at 0.08 because that the is observed result, or observed difference in proportions of residents who owned a dog between Bozeman and Missoula in the original sample ($\hat{p}_1 - \hat{p}_2 = 0.61 - 0.53 = 0.08$).