Biometry, Fall 2020 *Your Name: _Danielle Barnas___*

Midterm 1, Part A (100 pts)

There are 20 questions on this portion of the exam. I would budget about an hour for this portion, but no more than 1.5 hours because you want to leave time for the second portion. Most questions can be answered in 1-3 sentences, so don't spend too much time writing more than that. If you finish Part B quickly, you can always come back to Part A to answer questions that you weren't sure about.

1. (5 pts) Descriptive statistics are of two kinds: statistics of location, and statistics of dispersion. (a) Describe what "location" and "dispersion" refer to and (b) provide the 3 most commonly used measures of location and (c) 2 measures of dispersion statistics.

a)

Location refers to the position of the curve, graphically, based on the mean

Dispersion refers to the width of the curve, based on the standard deviation, or how scattered the data is

b)

mean, median, mode

c)

range = max value – min value

sum of squares

2. (5 pts) Imagine you had two samples with the same mean, but different variance. Which sample will have the higher Coefficient of Variation (CV)? Why?

The sample with the higher variance will have a higher coefficient of variation because CV = standard deviation / mean and greater variance will yield a greater standard deviation (= square root of variance), resulting in a higher CV value.

3. (5 pts) In what way is the formula for variance of a sample different from variance of a statistical population? Why do we use different formulas for estimating the variance of a population versus calculating the variance of a sample?

Variance of a population = sum of squares divided by the number of individuals or observations in the population, while a sample variance is the sum of squares divided by the number of individuals or observation in the sample minus 1. Using "n-1" (degree of freedom) eliminates a bias when calculating our sample variance because if we know all but one of our sample values, we know what that last value must be based on the mean or another statistic evaluating the entire sample.

4. (5 pts) Explain the logic behind attempting to falsify the H_0 rather than attempting to prove the H_a true.

In trying to prove the H_a there will always be some chance of confounding effects which may skew our interpretations in attempting to prove our alternative hypothesis. We can falsify a null hypothesis by showing that there was at least some effect, which points to an alternative hypothesis and may provide support of our H_a.

5. (5 pts) How does the width of the confidence interval change as sample size increases (if all other statistics remain the same)?

s.d. may or may not decrease

As n increases, standard deviation decreases, and if the mean of the sample remains the same, the confidence interval width would also decrease (CI = mean +/- SD*1.96).

6. (4 pts) Calculate the Z score (standard deviate) for the value 16.0 if the mean of the sample is 8.0 and the standard deviation is 4.0.

z score = (16 – 8) / 4 = 2

7. (6 pts) Describe an H_0 for each type of t-test.

One-sample t-test: H0: The sample group does not differ from a known constant

Two-sample t-test: H0: The two sample groups do not differ from each other (mean A = mean B)

Paired t-test: H0: Within each pairing, there is no difference between groups

8. (5 pts) Define *pseudoreplication* and make up an example of a pseudoreplicated experiment. What hypothesis is truly being tested by a pseudoreplicated design? How does this differ from the hypothesis that researchers likely *want* to test?

Pseudoreplication is inappropriate or ineffective replication.

Ex. Testing growth of a species of plant over time. Growth measurements are collected from the same plant over time, which is kept in a controlled environment. The hypothesis being tested here is growth of this particular plant over time, while the researchers likely wanted to test the average growth of the population of this species over time. They would have needed to have multiple individuals to test over time for a better population estimate.

9. (4 pts) What is the value of a procedural control? Give an example of why we might want a procedural control rather than just a control.

A procedural control accounts for unwanted variation due to our manipulation during our experiment. A procedural control is useful over simple a control because if we manipulate our samples in some way that has possible confounding effects, but we do not attempt to eliminate the bias of these confounding effects, we will not know how to interpret our results or support our alternate hypothesis. Ex. Giving a placebo pill to a patient who doesn't know if they are getting the real drug or not eliminates bias of a patient simply reacting to begin given a pill to take.

10. (5 pts) In what way is a two-tailed test different from a one-tailed test? What is the logic behind choosing one over the other? Which has higher statistical power?

A two-tailed test assumes that the type I error could be under either tail (in either the upper or lower 2.5%), while a one-tailed test assumes the type I error can only be under either the upper or lower tail (upper or lower 5%), and should only be assumed with a priori knowledge of our population.

Power = 1-Beta

A two-tailed test has greater power.

One tail has greater power, which is why people are suspicious if you

11. (5 pts) What is the difference between a dependent and an independent variable?

A dependent variable is the response variable, which is responding to or predicted by our independent variable (the predictor variable). Dependent variables are represented on the y-axis, while the independent variables are represented on the x-axis of regression plots

12. (5 pts) Why bother to measure effect size, when the p-value has already accepted or rejected our hypothesis?

The p-value tells us that there is a relationship, but not how strong that relationship is. Effect sizes tell us the strength of a relationship, whether there is any real ecological value to it.

13. (5 pts) You measure the concentration of two enzymes expressed by a bacterial colony and test for a correlation between the two variables. You find that r = -0.15 and p=0.032. What does this tell you about the relationship between these enzyme concentrations?

> The two enzymes are correlated, meaning they do co-vary, based on the p-value less than 0.05, however the strength of that negative correlation (the effect size) is too low to have any real ecological value.

14. (5 pts) What are the assumptions of a Pearson's correlation? In what cases might you be better off using Spearman's correlation?

> Assumes linearity and normality of data
>
> Use spearman's rho for nonlinear relationships

15. (5 pts) What is the standard error of the mean and what does it represent? Explain how it is different from the standard deviation.

> Standard deviation represents the average deviation from the mean, and equals the square root of the sum of squares divided by the degrees of freedom (sample size minus one), while the SEM is standard deviation divided by the square root of the sample size. SEM is a measure of dispersion and it can be used to estimate confidence in the mean (precision of our sample). The more we sample (as n increases in the denominator) the more precise our estimate of the mean is (SEM decreases).

16. (6 pts) For each of the following questions, give the appropriate sampling unit and the statistical population.

SE used as precision of the mean

How many mutations per generation in a species of bacteria?

Sampling unit: one generation of a single bacteria in the species
Statistical population: all generations of all bacteria in the species

How many meteorite craters on moons of planets in our solar system?

Sampling unit: one moon
Statistical population: all moons of all planets in our solar system

How many stars per galaxy?

Sampling unit: one galaxy
Statistical population: all galaxies

17. (5 pts) How does the Central Limit Theorem help us to determine how well our sample parameter (e.g. mean/variance/slope/etc.) estimates the population parameter?

As our sample size increases, the means of samples will always approach a normal distribution, even if our sample was not normally distributed, which we can use to make inferences on the population now that we have achieved normality.

18. (5 pts) What is the difference between correlation and regression?

Correlation shows a relationship between two variables which may not be causal (one doesn't cause or drive the other), while regression shows a causal relationship (where x drives y, and we can make assumptions or predictions for y based on x).

19. (5 pts) In the context of simple linear regression, what is a residual?

A residual is how much each y value differs from its expected value based on x. Essentially, residuals tell us unexplained variation, or the error, in our model.

20. (5 pts) What is the coefficient of determination? What does it tell you about the relationship between x and y in a linear regression?

The coefficient of determination is the r-squared value (correlation coefficient squared), which measures explained variation, and it tells us the strength of the relationship between x and y, how much x drives or predicts y. if = 0, x doesn't explain y at all. If = 1, x perfectly explains y.