How to write interpretations, such as “we reject the null hypothesis, so …”

- It’s always a good idea to start by saying either “we reject the null hypothesis …” or “we cannot reject the null hypothesis …” After that, I’m looking for you to say something related to the practical context of the application. For example, if it was the infant heart study comparing PDI/MDI outcomes across surgeries you’d share the surgery you believe to be more effective (if you rejected H0), or you’d say there wasn’t enough evidence in the data to statistically establish a difference. You should recognize that the entire purpose of hypothesis testing is to use the sample data to make a conclusion about a broader population.

I would really appreciate bootstrapping and the difference between standard deviation and standard error conceptually

- Standard deviation is a way of measuring the average distance between a collection of numbers and the mean value of those numbers. For example, if we have the heights of every adult male in the US, the standard deviation of those heights tells us how much, on average, heights vary from the mean height. This is important because it helps us understand what sort of height are typical in the data. If the mean height is 60 inches and the standard deviation is 3 inches, we know that an height of 63 inches (6’1”) is 1 SD above the mean and relatively common, while an height of 69 inches (6’7”) is 3 SD above the mean and relatively uncommon but still possible to encounter if we observe enough people – however, we would be very surprised if we sampled a single person and their height was 69 inches.
- Standard error is just a standard deviation, but of a descriptive statistic rather than a collection of individual cases/data-points. If we take a sample of 100 cases, we can calculate the sample standard deviation of their heights, which represents how much each individual height deviations from the sample average. We could also calculate the mean height of the sample, which we know will be subject to sampling variability. Standard error describes this sampling variability, and it is defined as the standard deviation calculated using a large number of different sample means (each for a different sample of 100 cases).
- It’s not practical to collect many different samples of size 100, calculate a sample mean within each, then take the standard deviation of these means to get the standard error. This is where bootstrapping comes into play. Bootstrapping is a way to simulate or artificially replicate the variability introduced by repeatedly sampling from the population. It allows us to the sample the original sample with replacement in order to get the large number of sample means (bootstrap statistics) necessary to find the standard error, which is simply the standard deviation of the bootstrap statistics.

Know the relationship between confidence intervals and hypothesis testing and how they provide complimentary information.

- Hypothesis testing and confidence intervals are tools for using the data in a sample to reach a conclusion about the cases in a broader population. Confidence intervals provide a range of plausible values for an unknown characteristic of the population with a certain degree of confidence that depends upon the standard error of the characteristic being studied (ie: SE of the sample mean, or sample proportion, etc.) and a probability model (ie: Normal curve, t-distribution, etc.) Hypothesis tests assume a certain value of the population characteristic and calculate how compatible the sample data are with that assumed value using the exact same components: a standard error and a probability model. Because both methods involve these same components, they should yield conclusions that are consistent when applied to the same sample data. So if you do a test of H0: p = 0.5 and your test doesn't provide you enough evidence to reject this null hypothesis at the 0.05 significance threshold, then a 95% confidence interval estimate of p using the same data should suggest that p=0.5 is a plausible value for the population you're studying.

when to use t test and when to use z test, and calculating two sided p values by hand

- As a simple rule, we'll use Z-tests for proportions and T-tests for means. The more nuanced rule is that the T-distribution is necessary whenever we're calculating the SE using another parameter that is estimated from the sample data and independent of the point estimate. Thus, we need the T-distribution for means, but also things like coefficients in regression (as we'll see later on); whereas for proportions the SE depends only upon the proportion itself and the sample size.
- I will not ask you to compute a two-sided p-value by hand for Z or T test. However, you should be able to sketch the area represented by the p-value on a Normal or T curve, and you should be able to estimate the p-value for a bootstrapping/simulation-based hypothesis test (see the question in the Practice exam).

The different ways to get standard error. When to use different methods. How to get Standard deviation (I think just stat key)

- Standard error is the standard deviation of a sample statistic (ie: point estimate for a population characteristic of interest). The Central Limit theorem gives us a mathematical expression for the SE of different descriptive statistics. Otherwise, we can use bootstrapping (StatKey) to repeatedly generate the sample statistic to estimate it's variability. The previous response on Std Dev vs. SE hopefully explains this.

Test statistics formulas, relationship between decision threshold and Type I and II errors

- A test statistic is a standardized way to reliably quantify how far the observed point estimate is from the value conjectured in the null hypothesis. The idea is to come up with a standardized measure of evidence against the null hypothesis. For example, if you observe a sample mean of 10 and the null hypothesis states the mean is 0, you do not know if 10 is far or close to zero because that depends upon the units of the variable involved in the analysis. For example, if you're estimating the average length of piece of building material, the average differing from H0 by 10 millimeters might be a very small deviation from what the hypothesis suggests, while a difference of 10 meters might be enormous, so the test statistic must be standardized by an appropriate measure of variability, the standard error. Thus, a test statistic looks like: (Point Estimate – Null)/SE; This turns out to simply be an algebraic re-arrangement of the Normal approximations that arise from the Central Limit theorem, which is why the test statistic follows a Standard Normal curve.
- The decision threshold reflects the largest p-value that would lead you to reject H0 in favor of an alternative. If H0 is true and you find a p-value smaller than your decision threshold you've made a Type 1 error by rejecting a Null hypothesis that shouldn't have been rejected. The likelihood of this occurring is the value of the threshold you set. So, if you determine that you'll

use alpha=0.05 as an evidence threshold you're giving yourself a 5% chance of making a Type 1 error if the null hypothesis is true. This is because it's possible to observe a test statistic in the most extreme 5% of the Normal curve by chance, but only with a probability of 5%.  A Type 2 occurs when a false null hypothesis is not rejected. You cannot calculate the likelihood of this occurring, but you do know that reducing your chances of rejecting the null hypothesis must increase the likelihood of a Type 2 error.