DATA 8 FINAL STATS REVIEW

I. Hypothesis Testing

Purpose: To answer a question about a process or the world by testing two hypotheses, a null and an alternative. Usually the null hypothesis makes a statement that "the world/process works this way", and the alternative hypothesis says "the world/process does not work that way".

Examples:
Null: "The customer was not cheating-his chances of winning and losing were like random tosses of a fair coin-50% chance of winning, 50% of losing. Any variation from what we expect is due to chance variation."
Alternative: "The customer was cheating-his chances of winning were something other than 50%".

Pro tip: You must be very precise about chances in your hypotheses. Hypotheses such as "the customer cheated" or "Their chances of winning were normal" are vague and might be considered incorrect, because you don't state the exact chances associated with the events.

Pro tip: Null hypothesis should also explain differences in the data. For example, if your hypothesis stated that the coin was fair, then why did you get 70 heads out of 100 flips? Since it's possible to get that many (though not expected), your null hypothesis should also contain a statement along the lines of "Any difference in outcome from what we expect is due to chance variation".

Steps:
1) Precisely state your null and alternative hypotheses.
2) Decide on a test statistic (think of it as a general formula) to help you either reject or fail to reject the null hypothesis.
    - If your data is categorical, a good test statistic might be the Total Variation Distance (TVD) between your sample and the distribution it was drawn from. An example of this case would be the Alameda jurors example we went over in class.
    - If your data is numerical, you have a lot of options. A common example is when you're testing whether a coin is fair, a good test stat might be |*observed proportion of heads -0.50* | (i.e. the absolute difference from what you get and what you expect).
3) Calculate the observed value of the test statistic, using your original sample of data. Remember this value, because it is used to calculate the p-value later.
4) Simulate all possible values of the test statistic under the null hypothesis.
    - Since our sample could have come out differently (since it was drawn at random), the observed test statistic could have been different. So we want to find all possible values.

- Steps to simulate:

i) Create empty array, that will contain all values of the test stat at the end.

ii) Set up a *for loop*, either one that runs the given number of repetitions, or if no number is specified, 10,000 times (the convention).

iii) Within the body of the for loop (i.e. the code that will be run on each loop), simulate under the null hypothesis. For example, if you were testing whether a coin was fair, you would use *np.random.choice* on *coin = make_array ('Heads', 'Tails')*, which would simulate flipping a fair coin.

iv) Still in the body of the loop, calculate the value of the test statistic for that resample and append (using np.append) that value to your results array outside the loop.

5) Calculate the p-value.
   - Remember, the p-value is the *chance, under the null hypothesis, of getting a test statistic equal to the observed test statistic or more extreme in the direction of the alternative.*
   - In other words, we want to see if the observed test statistic is likely to come from the null distribution. If the observed test statistic is inconsistent (i.e. looks nothing like the others) with the rest of the test statistics generated under the null, then it starts to look like the null hypothesis is not true.

   i) Either calculate *np.count_nonzero(results_array >= observed_test_stat)/len(results_array)* or *np.count_nonzero(results_array <= observed_test_stat)/len(results_array)* depending on what the alternative hypothesis says.

   ii) If your p-value cut-off (which is either given, or chosen to be 0.05, the convention) is 0.05, and your p-value is 0.023 (less than cutoff), then you reject the null. That means that the observed test statistic is so unlikely to occur under the null hypothesis that the alternative hypothesis makes more sense.

   iii) If your p-value is greater than the cut-off, then you fail to reject the null hypothesis. The observed test statistic was consistent enough with the null distribution that we cannot conclude the alternative.

Pro tip: Do not say we "accept the null". Because of the randomness of this process, a p-value greater than the cut-off is not enough evidence to prove that the null hypothesis is true-it only proves that we do not have enough evidence to reject it. Thus, it is better to say "fail to reject the null" rather than "accept the null". (This is a subtle difference, but important).

For testing hypotheses using confidence intervals, please see the next section.

II. Confidence Intervals

Purpose: To provide an interval of estimates for a population parameter. For example, let's say we wanted to estimate the median annual household income in the United States. We collect a large random sample from the population. However, the median of our sample isn't a good estimate by itself. Due to random chance, our sample could have come out differently, and then the sample median would have been different. Thus, we need to take many samples, and provide

an interval of estimates from the sample medians. However, we do not have the resources to physically take more samples, so we bootstrap (or "resample") from our original sample.

Examples: Construct a 95% confidence interval for the average height of the entire U.S. adult population.

Steps:

1) Calculate your original sample statistic (which is the sample version of the population parameter). An example would be the sample mean or median.
2) Create an empty array to collect all of your sample statistics.
3) Create a *for loop* that will run either the given number of repetitions or 10,000 times.
4) Within the body of the for loop (the code that will run each loop), perform one bootstrap ("resample") replication of your original sample. The bootstrap must be drawn at random *with* replacement, and the same size as your original sample to be valid. This can be done with *original_table.sample()*
5) Still in the body, calculate the sample statistic of your bootstrap resample and append it to your results array using *np.append*.
6) Once you have all 10,000 sample statistics, take the middle A% of them, where A is the confidence level of your confidence interval.
    - For a 95% confidence interval, take the middle 95% of statistics.
    - This means that there is 5% of data left over, 2.5% on each end. Thus, the left end of the middle 95% (the left end of the interval) is at the 2.5^{th} percentile, and the right end is at the 97.5^{th} percentile.
    - Find the left and right end by using *percentile(2.5, results_array)* and *percentile(97.5, results_array)*

What if you don't have a computer?
Mathematically, a confidence interval is defined as:

$$Sample\ Mean \pm\ 2{\times}Sample\ SD$$

But what if you don't have the sample standard deviation? Then you can use this formula:

$$Sample\ SD = \frac{population\ SD}{\sqrt{Sample\ Size}}$$

A common problem where this formula is useful is worked out at the end.

III. Testing Hypotheses Using Confidence Intervals:

Purpose: For certain types of hypothesis tests, it's easier and more beneficial to make a conclusion about the hypotheses with a confidence interval. If your hypotheses are something like: *Null:* "The population parameter is some number *X*" and *Alternative:* "The population parameter is not that number *X*", then you can construct a confidence interval to answer this question.

Examples:
Null: The true slope of the regression line between X and Y is 0.

Alternative: The true slope of the regression line between X and Y is not 0.

Pro tip: In order to use this method, the hypotheses must be in the format of "this parameter is this number x". A hypothesis test about whether a coin was fair does not work for this format, because there is no number to construct a confidence interval around.

Steps:

1) Carefully define your null and alternative hypotheses. For example: "The average age is 30".
2) Construct a confidence interval for the parameter in question (age in this example), by following the steps outlined in the previous section.
3) If your constructed confidence interval does not contain the number in question, reject the null hypothesis. In our example, if the confidence interval for average age was (31.2, 35.8), that does not contain 30, as specified by the null.
4) If your constructed confidence interval does contain the number in question, then fail to reject the null hypothesis.
    - This method is valid because of the deep connection between a confidence level and a p-value cut-off.
    - If a 95% confidence interval does not contain the number in question, that's like a p-value being less than a 0.05 cutoff value.
    - Likewise, a 99% confidence interval that does not contain the number in question is like a p-value being less than a 0.01 cutoff value.

IV. Correlation

Purpose: Measures how tightly clustered a scatter diagram is about a straight line, which indicates the strength of the linear relationship ("association") between two variables X and Y. Also used to calculate the regression line between X and Y.

Properties of Correlation Coefficient r:

- r is a number between -1 and 1. 1 and -1 indicate the strongest relationship, while r = 0 indicates no relationship between X and Y.
- r has no units.
- r is not affected by changing the units of X and Y. For example, if X was originally measured in inches, and then X was converted to centimeters, r would not change.

Formula:

- r is the average of the product of X and Y in standard units

Word of Caution:

- Correlation does not mean Causation!!! Just because X has a strong correlation with Y, does not mean that X *causes* Y. For example, shoe size and reading ability are strongly correlated. But that does not mean that the bigger feet a person has, the better reader they are. (What's the confounding variable? Age).
- Correlation measures *linear* association! If your scatter plot looks like it fits a curve (like a parabola) better than a straight line, you should not use correlation or regression on the data! Your results will be very inaccurate.

V. Linear Regression

Purpose: If we have two variables X and Y, we want to be able quantify the relationship between them, so that given a new value of X, we can predict its Y value. How do we quantify the relationship? We find the least squares regression line of the scatterplot of X and Y.

Example: Heights and weights in a population are related. Once we find the equation of the regression line between heights and weights, we can predict the weight of a new person, given their height.

Formulas:
Quick reminder of standard units (which we use to compare variables with different scales):

$$X \text{ in Standard Units} = \frac{Value\ of\ X - Average\ of\ X}{Standard\ Deviation\ of\ X}$$

If X and Y are measured in Standard Units, and r is the correlation coefficient, the line is:

$$Estimate\ of\ Y = r \cdot x$$

If X and Y are not measured in Standard Units, then the line is:

$$Slope\ of\ line: r \cdot \frac{SD\ of\ Y}{SD\ of\ X}$$

$$Intercept\ of\ line: average\ of\ Y - slope \cdot average\ of\ X$$

With these formulas, you can predict Y, given a value of X, by:

$$Estimate\ of\ Y: slope \cdot given\ value\ of\ X + intercept$$

Caution:
There are some situations where you should not use linear regression (the above formulas), because they would be a horribly inaccurate form of prediction.

1) The correlation coefficient and the above linear regression formulas measure LINEAR association only. So if you look at the scatterplot of X and Y and the points seem to be clustered around a curve (like a parabola), you should not use linear regression on that data.
2) Sometimes non-linearity is hard to see in the original scatterplot, so we look at the residual plot as well.

   $$Residual = observed\ value\ of\ Y - regression\ estimate\ of\ Y$$

   If we calculate the residual for every point (X, Y), and plot them, the residual plot should *show no trend or pattern* (i.e. it should look like a blob). If you notice a pattern, for example all the residuals are positive, that means the data may not be linear, and you should not use linear regression on the data.

So if a question on the final asks, "Can you use linear regression on this data?" Check the above conditions.

The next two sections will go over some common problems in regression

VI. Inference for the True Slope

Purpose: We have some sample, and we can calculate a regression line between X and Y in our sample. But since our sample was drawn at random, it could have come out differently (i.e. we could have different data). Then our regression line would be different. Thus, we want to estimate the true slope of the line between X and Y in the population (i.e. the true relationship between X and Y). How do we estimate the true slope? With a bootstrapped confidence interval!

Steps:

1) Follow the steps of constructing a confidence interval in the earlier section, where your sample statistic is the slope of the regression line for each sample. (Use the regression line formulas to calculate slope each time).

Question: A thought strikes us: what if there is no relationship between X and Y? What if we got a non-zero slope in our sample by chance? What if the true slope of the line between X and Y is 0? These questions lead very nicely to a set of hypotheses:

Null: The true slope of the line between X and Y is 0.

Alternative: The true slope of the line between X and Y is not 0.

How do we test these hypotheses? With a bootstrapped confidence interval for the true slope!

Steps:

1) Follow the steps for testing hypotheses with confidence intervals defined in the previous sections, where you are constructing a confidence interval for the true slope.
2) If your final confidence interval does NOT contain 0, then you reject your null hypothesis. That means your sample wasn't a fluke, there is a relationship between X and Y.
3) If your final confidence interval DOES contain 0, then you fail to reject your null hypothesis. That means there is no relationship between X and Y in the population.

VII. Prediction Intervals

Purpose: We calculate a regression line so that given a new value of X, we can predict its Y value. But if our sample was drawn at random, then the sample could have been different, so the equation of our regression line could have been different, which would give us a different predicted value of Y for that given value of X. How do we remedy this situation? We create a confidence interval for the predicted value of Y given a value of X!

Steps:

1) Follow the steps for constructing a confidence interval outlined in previous sections.
2) In each bootstrap replication, recalculate the equation of the regression line for that sample and use it to predict a value of Y for the given X.
3) At the end of the process, you will have a confidence interval for predicted values of Y.

Note:
The value of X that you plug into the regression equation never changes. You are creating an interval for predicted Y's, *based off of that one given value of X.* The slope and intercepts of your regression line will change (because of bootstrapping), but the value of X that you plug in never changes.

VIII. Classification

Purpose: We want to develop an algorithm so that when given a new person/subject, we can predict their class based on values of their features/attributes (i.e. using multiple attributes to predict a categorical variable).

Examples: Using cell measurements and features to classify the cells as cancerous or non-cancerous. Using the frequency of certain words to classify a song as hip-hop or country.

Steps:
Note: for the exact code of these steps, please see the textbook section on classification:
https://www.inferentialthinking.com/chapters/15/4/implementing-the-classifier.html

1) Make sure you have a data set of individuals with their attribute measurements and known class.
2) To classify a new individual, find the *k* nearest neighbors to your new point in your dataset, by computing the Euclidian distance between them.
3) Among the *k* nearest neighbors, find the majority class among them, and assign that class to the new individual.

IX. A/B Testing (Permutation Tests)

Purpose: To determine if two samples come from the same underlying distribution. Or, to see if the distribution of some feature/attribute for one class is the same as the distribution of the feature/attribute for another class (in the population). These questions are answered by permutation tests, which are similar to hypothesis tests with one key difference. Another use is to help us decide whether we should use a certain attribute during classification.

Examples: Trying to see if the distribution of birth weights is the same for smoking mothers compared to non-smoking mothers. Another example is seeing if the distribution of mitosis levels is the same for cancerous cells as non-cancerous cells.

Steps:
1) Set up your null and alternative hypotheses
    - *Null*: The distribution of the feature/attribute is the same for Class 0 and for Class 1 in the population. (i.e. the feature/attribute is not related to class).

- *Alternative*: The two distributions are different in the population. (i.e. the feature/attribute is related to class).

2) Decide on your test statistic.
    - For the baby weight example, it could be the absolute difference between the average baby weight of each class.
    - For the mitosis example, it could be the Total Variation Distance (TVD) between the two distributions
3) Calculate the observed value of your test statistic (from your original sample). This, like in hypothesis testing, will help you decide between the hypotheses in the end.
4) Simulate your test statistic under the null hypothesis (this is where a permutation test is different from a hypothesis test).
    - The null hypothesis says that the feature/attribute and class are not related. That means it doesn't matter what order the attribute values are in- that wouldn't change the test statistic or distribution.
    - If it doesn't matter what order the attribute values are in, then you can shuffle (randomly permute) the attribute values and nothing would change.
    - This is how we simulate the statistics under the null:
        i) Shuffle (i.e. draw at random *without* replacement) your attribute values (Not the whole table! Just attribute values)
        ii) You can do this with *original_table.column("Attribute").sample(with_replacement = False)*
        iii) Recalculate your test statistic, using the shuffled values, and append it to your results array.
        iv) Repeat this process 10000 times
    - This is how a hypothesis and permutation test are different. To simulate the test statistic under the null hypothesis, a hypothesis test draws at random *with replacement* from the population/distribution (see coin example in hypothesis section). To simulate a test statistic under the null hypothesis, a permutation test draws at random *without replacement* from the population/distribution (i.e. shuffling)
5) Calculate a p-value using your results array and observed test statistic to help you decide between the hypotheses. (See hypothesis test section for more detailed version of this step).

X. Updating Probabilities (Bayes Rule)

Purpose: We have some *prior* probabilities of some event A, and given some new information (*likelihoods)*, we want to *update* our chances of event A (called *posterior probability*).

Context: I have a class, made up of 60% sophomores (second years) and 40% juniors (third years). Half of sophomores have declared a major, while 80% of juniors have declared their major. I select one student at random.

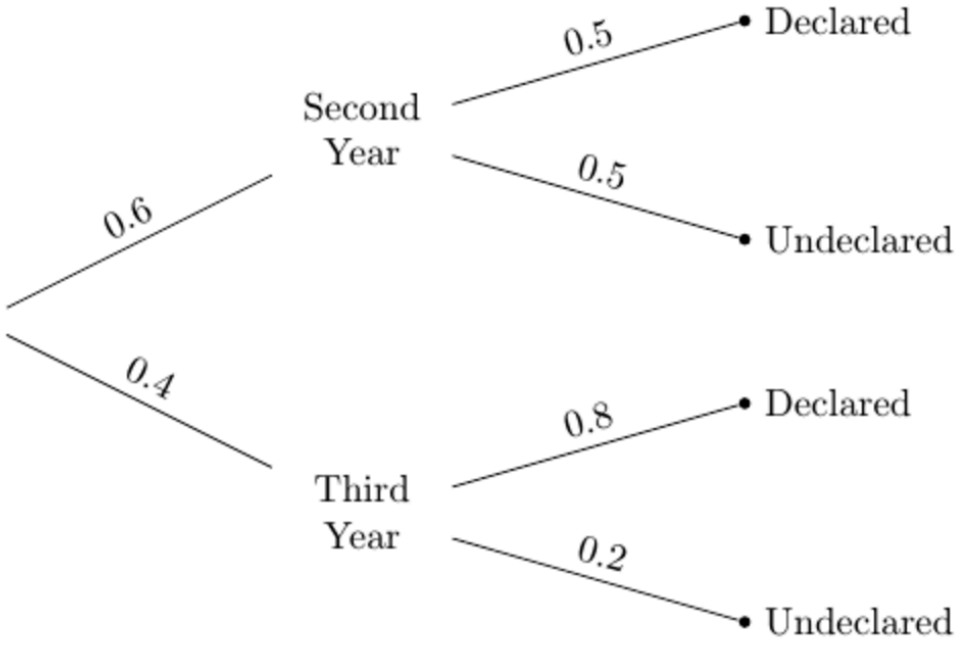
Question: Which year is the student most likely to be in?

Answer: Sophomore, because there are more sophomores than juniors, so a student picked at random is more likely to be a sophomore.

New Information: I tell you that the student I picked at random had declared their major. Now which year are they more likely to be in?
Answer: This question is harder to answer. Before, the student was most likely to be the largest year level. But now given some new information, our probabilities of which year the student is changed. This calls for Bayes rule!

Steps:

1) DRAW A TREE DIAGRAM. This problem is so much easier to think about with this diagram, so always draw one.
2) Here is the tree diagram for this situation:



   You have the years on separate branches, and then the possibilities of declared or undeclared for each year. Note the PORPORTIONS along each branch. It's important you use proportions instead of counts for the following calculations.
3) Let's calculate the *posterior* probability the student is a third year given that the student declared their major.
   - Let's denote this P(Third year | Declared). To calculate this, first restrict your mind to the two branches of declared students. So restrict yourself to the second year and declared major path, and then the third year and declared major path. Those will make up the denominator of your probability (because we know the student has declared their major, so we don't consider the undeclared probabilities.
   - And since we are calculating the chance the student is a third year, the numerator of the probability will be the chance the student is a third year and declared.
   - In more concise terms:

$$P(\text{Third Year} \mid \text{Declared}) = \frac{0.4 \times 0.8}{0.6 \times 0.5 + 0.4 \times 0.8}$$

$$= \frac{(\text{prior probability of Third Year}) \times (\text{likelihood of Declared given Third Year})}{\text{total probability of Declared}}$$

The other posterior probability is

$$P(\text{Second Year} \mid \text{Declared}) = \frac{0.6 \times 0.5}{0.6 \times 0.5 + 0.4 \times 0.8}$$

$$= \frac{(\text{prior probability of Second Year}) \times (\text{likelihood of Declared given Second Year})}{\text{total probability of Declared}}$$

- Since the posterior probability that the student is a third year, given that they declared their major, is greater than the posterior probability that the student is a second year, we change (*update*) our classification of the randomly selected student to a third year.

XI. Central Limit Theorem

Purpose: If the conditions of the CLT are satisfied, then we know our distribution/histogram of our sample is normal, which allows us to use the normal distribution's many useful properties.

Conditions: These 2 conditions MUST be satisfied in order for you to use the CLT and say that your sample is normally distributed.

1) Your sample needs to be *large,* and drawn *at random with replacement*. Small sample sizes and drawing without replacement don't work.
2) You must be trying to find the distribution of your sample *sum* or *average.* (Proportions of yes-no populations are also considered averages-see homework 6)

Theorem:
If the above conditions are satisfied, namely you have the probability distribution of sum or average of a large random sample drawn with replacement, then the distribution will be roughly normal, regardless of the distribution of the original population.

Examples:
The distribution of mother's ages when they give birth is NOT normal-it has a long right tail, with a peak around 25. But if you bootstrap and create an empirical distribution (generated by many resamples) for the *average age* of mothers at birth, then that distribution WILL be normal. See textbook section: https://www.inferentialthinking.com/chapters/12/4/central-limit-theorem.html for more examples.

Caution: When grading HW6, I noticed that many students said we could use the CLT because "The sample size was large". But that is not the only condition!! Make sure both conditions are satisfied before assuming we can use CLT and that the distribution is normal.

XII. Miscellaneous

Calculating Sample Size:
A common problem: What sample size should I take such that a 95% confidence interval has a width of no more than 0.04?

Steps:
I will work out the problem and add some comments along the way

$$(Sample\ mean + 2 \cdot Sample\ Sd) - (Sample\ mean - 2 \cdot Sample\ Sd) \leq 0.04$$

This line says the right end minus the left end should have a width of no more than 0.04. And I multiply the SD by 2 because 95% of a normal distribution is within 2 SDs of the mean

$$4 \cdot Sample\ SD \leq 0.04$$

Now, we don't know the sample SD, and we need to somehow get sample size in the equation so we can solve for it. So we remember the formula for sample SD

$$\frac{population\ SD}{\sqrt{Sample\ Size}} \leq \frac{0.04}{4}$$

Next plug in your value for the population SD (given), which I will say is 1, and cross-multiply

$$0.04 \cdot \sqrt{Sample\ Size} \geq 1 \cdot 4$$

$$\sqrt{Sample\ Size} \geq \frac{1 \cdot 4}{0.04}$$

$$Sample\ Size \geq 10{,}000$$

You need a very large sample size to get a confidence interval that small and accurate!