Part I

Review of Inference

Deciding whether to use a test or a confidence interval

Generally, use a significance test to test a claim. Use a confidence interval if you wish to estimate a population parameter (μ or p) based on statistics from your sample ($\bar{x}$ or $\hat{p}$). Once a confidence interval is constructed, you may use it to test claims (where you fail to reject a claim that falls within the confidence interval, and you reject a claim that falls outside of a confidence interval). The α -level of a **two**-sided hypothesis test is related to the confidence level of an interval by $C = 1 - \alpha$.

Identifying the type of test or interval to use

- Is there a single numerical variable being measured for each subject? Then we will perform a test for means.
- Is there a categorical variable being measured, and we are only concerned with how often a single response (a "success") occurs? Then we will perform a test for a proportion. *For example, if we only care about the proportion of brown-eyed people, then we can do a z-test for the proportion of brown-eyed people.*
- Is there a categorical variable being measured, and we are concerned with how many responses fall into **each** category? Then we will perform a χ^2-test. *For example, if we want to compare the occurrence of brown, blue, green, and grey eyes in two different groups, we will do a χ^2-test because we are looking at multiple categories for the categorical variable "eye color."*
- Are we looking at the relationship between two numerical variables? Then we will perform a *t*-test for the slope of a regression line.

How many samples?

- Be careful to identify the source of each mean or proportion mentioned in a problem. If a mean or proportion does not clearly come from a sample (with an identifiable sample size *n*), then it is probably a claim or a population proportion which should be used in the null hypothesis. A two-sample test should have two clearly identified sample sizes, and each sample should result in an $\bar{x}$ or a $\hat{p}$.
- Some problems have two lists of numerical data that are linked in some way. For example, they could be pre-test and post-test scores for a list of students or temperatures in the sun and temperatures in the shade for a list of days. In these cases, the improvement (post-test score minus pre-test score) or the difference (temperature in sun minus temperature in shade) is the important variable. These are called matched-pair *t* tests. Begin by subtracting the two lists of data to obtain one list of improvements or differences. Then do a one sample *t*-test for a mean. (You will ignore the original two lists after you subtract.) Your null hypothesis will often be that the mean improvement was zero. For example in the pre- and

post-test problem, you might use $H_0 : \mu_{improvement} = 0$ and $H_a : \mu_{improvement} > 0$, where improvement is defined as post-test score minus pre-test score.

- In some χ^2-tests, there is a clear claim. *For example, a company claims that 50% of the prizes in the popcorn boxes are stickers, 20% are rings, and 30% are temporary tattoos.* In this case, we are comparing the data from one sample to a claim, using a χ^2-test for goodness of fit. The data table for this test consists of a single row of data.
- In other χ^2-tests, we are comparing two groups to see if they have the same percentages in each category. *For example, we could compare eye colors of a group of men and a group of women.* In this case, do a χ^2-test for homogeneity. The data table would have two rows (one for male and one for female) and multiple columns (for brown eyes, green eyes, etc.). Rows and columns may be switched.
- The data for some χ^2-tests consists of two categorical questions asked to a single sample of people. *For example, we could ask a group of teachers whether they exercise frequently, often, or never and whether or not they missed any days of school last year due to illness.* We would like to see if their answers to the two questions are independent of each other. If they are independent, then the proportion who missed school due to illness should be the same for all three exercise categories. In this case, do a χ^2-test for independence. The data table would have two rows (one for people who missed school due to illness and one for those who did not) and multiple columns (for the different exercise categories). Rows and columns could be switched. **The mechanics of the χ^2-test for independence and the χ^2-test for homogeneity are exactly the same.**

What to put in your significance test

- A null and an alternative hypothesis (define the parameter of interest in words) *Note: Always hypothesize about the unknown population parameters (μ and p), not the sample statistics ($\bar{x}$ and $\hat{p}$), which are known from the data.*
- Identify the test you are using and check the conditions necessary for doing that test.
- Formula for the test statistic (z or t or χ^2)
- Value for the statistic (can be from calculator if you have written the formula) and a shaded picture of the distribution if you have time to draw it
- The P value (from calculator or table) ***related to*** the α-level, plus *df* for *t*-tests or the expected values for χ^2 tests
- Two conclusions: either reject H_0 or don't reject it based on the relationship of the P-value and the α–level AND write a conclusion about the alternative hypothesis in the context of the problem

Phrasing to put in your conclusions

Example: A seed manufacturer claims that at least 97% of their seeds will germinate. You suspect that the germination rate is less, so you buy a random selection of these seeds to test this claim. You calculate a P-value based on the hypotheses $H_0: p = 0.97$ and $H_a: p < 0.97$, where p is the germination rate of all seeds sold by this company.

- ***When $P < \alpha$, reject H_0.***
  First, draw a mathematical conclusion about H_0: "Since the P-value of 0.017 is less than the α-level of 0.05, reject H_0. A value as extreme as my sample's germination rate should only occur 1.7% of the time by random chance if the company's claim is true." *Then, write a conclusion about H_a in the context of the problem:* "We can conclude that the germination rate of the seeds is significantly lower than the 97% claimed by the company."

  Instead! There is convincing evidence that the germination rate of the seeds is lower than 97%.

OK, but I would like the format from our class. [updated!]

- ***When $P > \alpha$, do not reject H_0.***
  First, draw a mathematical conclusion about H_0: "Since the P-value of 0.209 is greater than the α-level of 0.05, there is not sufficient evidence to reject H_0. A value as extreme as my sample's germination rate would occur 20.9% of the time by random chance if the company's claim is true." *Then, write a conclusion about H_a in the context of the problem:* "We cannot conclude that the germination rate of the seeds is significantly lower than the 97% claimed by the company."

  Instead There is Not convincing evidence that....

- Note that we never accept H_0.

How to check the conditions necessary to do the tests for means and proportions

Always check to see if the sample is independent: randomly taken **from the population of interest** and that the population is at least 10 times the sample size.

One Sample

What you are studying	What you know	Use this	Why?
Mean	s (standard deviation of sample) **Graph the sample if $n < 40$ to see if it is approx. normal**	$t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$	properties of t distribution

Proportion	$np_o > 10$ $n(1-p_o) > 10$ *Use the claimed proportion here.*	$z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$	you can approximate the binomial distribution by the normal distribution
Mean *(Note: It is unusual to know σ of the population.)*	σ population is normal	$z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$	because you can always use z scores for normal distributions
Mean *(Note: It is unusual to know σ of the population.)*	σ population may not be normal, but $n > 30$	$z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$	Central Limit Theorem says that distributions get more normal as n increases

Two Samples

What you are studying	**What you know**	**Use this**	**Why?**
Difference of means (two independent samples)	s_1 and s_2 (standard deviation of samples), graph each one to see if it is approx. normal if $n < 40$	$t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$	properties of t distribution
Difference of two proportions	$n_1\hat{p} > 5$ $n_1(1-\hat{p}) > 5$ $n_2\hat{p} > 5$ $n_2(1-\hat{p}) > 5$ *Use the pooled $\hat{p}$ here.*	$z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ where $\hat{p} = \dfrac{X_1 + X_2}{n_1 + n_2}$	you can approximate the binomial distribution by the normal distribution.
Difference of two dependent means	THIS IS MATCHED PAIRS!!!	Find the difference between each pair and do a one sample t test.	We are only interested in one piece of data for each subject; usually the improvement or difference (after-before).
Difference of means (two ind. samples) *(Note: It is unusual to know σ.)*	σ_1 and σ_2 population is normal	$z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	because you can always use z scores for normal distributions

Difference of means (two ind. samples) *(Note: It is unusual to know σ.)*	σ_1 and σ_2 population may not be normal, but $n_1 > 30$ and $n_2 > 30$	$z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	Central Limit Theorem says that distributions get normal as n increases

How to check the conditions for the chi-squared tests

- Data must be counts (not averages or proportions).
- Data in sample are independent (chosen randomly and n < 10% of the population)
- Groups are large enough that all expected values ≥ 5.

How to check the conditions for the t- test for the slope of a regression line

- The scatterplot must look linear.
- There must be no pattern in the residual plot (errors are independent).
- The residual plot has a constant spread (errors have constant variability).
- Histogram of residuals is approximately normal.

Inference using confidence intervals

Keep doing this

1 State
PLAN
DO
conclude

- In general, you must put the following three things in a confidence interval problem:
  1. Identify the interval you will use and check the conditions necessary to use the interval.
  2. Calculate the interval.
  3. Interpret the interval in the context of the problem.
- The conditions we must check are the same as for the associated significance tests (as shown in the table above), with one exception. When performing significance tests for one or two proportions, you check to see that n is large enough by examining np_o and $n(1-p_o)$ where p_o is the claimed proportion which appears in the null hypothesis. Since there is no claim in a confidence interval problem, use the sample proportion $\hat{p}$ in these checks. For one-sample intervals, we require that $n\hat{p}$ and $n(1-\hat{p})$ be over 10. In two-sample intervals, we require that $n_1\hat{p}_1$, $n_1(1-\hat{p}_1)$, $n_2\hat{p}_2$, and $n_2(1-\hat{p}_2)$ be over 5.
- Also, in proportion confidence intervals, we must use the sample proportion $\hat{p}$ to calculate the standard error.

Review of Inference

- Formulas:

One mean (σ unknown)	$\bar{x} \pm t^* \dfrac{s}{\sqrt{n}}$
Difference in two means (σ unknown)	$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$
One mean *(σ known; unusual case)*	$\bar{x} \pm z^* \dfrac{\sigma}{\sqrt{n}}$
Difference in two means *(σ known; unusual case)*	$(\bar{x}_1 - \bar{x}_2) \pm z^* \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$
One proportion	$\hat{p} \pm z^* \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$
Difference in two proportions	$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Good confidence interval conclusions

Make sure to state these in the context of the question.

- I am C% confident that my interval captures the population value μ or p.
- C out of 100 intervals constructed using this method would capture the population value μ or p.

Bad confidence interval conclusions

Avoid making these statements:

- C% of the $\bar{x}$ values or $\hat{p}$ values would fall in my interval.
- C% of the data is in my interval.
- There is a C% chance that the population value μ or p is in my interval.