Summary of Statistical Inference

What is Statistical Inference?
The **population parameter** and the **sample statistic** summarize the same variable. The population parameter summarizes the variable for the population, which is what we want to know, e.g. $\pi$ or $\mu_1 - \mu_2$. However, we usually can’t observe the whole population so we don’t know what the parameter value really is. However, we can measure the variable on a sample or randomized groups and compute a sample statistic, e.g. $\hat{p}$ or $\bar{x}_1 - \bar{x}_2$. The question is what can we **infer** about the parameter based on this statistic? Because these statistics follow a null distribution/regular pattern due to the randomness in the study design, we can estimate or calculate probabilities of different values of the statistic occurring. Different statistics follow different distributions, but once we know which distribution we should use, we can make conclusions about the value of the parameter, e.g., it’s in some interval or we have evidence that it is not a particular value.

Null Distributions
If we specify a value for the population parameter, we can take (or simulate) lots of samples from this population or lots of random assignments and calculate a statistic for each sample/random assignment. This allows us to examine the behavior of the statistic so we can discuss the **shape**, **center**, and spread (i.e. variability) of this “null” distribution. For example, what types of values do we expect the statistic to have, how far away might the statistic stray from the hypothesized value of the parameter?

Simulation vs. “Approximate” (Theory-Based) procedures
In almost every case we have seen two different ways to calculate the p-value: simulation or an approximation based on a mathematical model. We are considering the simulation approaches to always be valid. The mathematical models are only appropriate if the “validity conditions” of the theory-based, approximate approach are met. The advantage of the mathematical models is we can easily get a confidence interval as well. So you may want to consider the mathematical model way first but then if the validity (aka “technical”) conditions aren’t met, use the simulation approach.

Confidence Intervals *Estimate population parameter*
The goal of a confidence interval is to get a range of plausible values that we think the population parameter could be equal to. To do this, we use the sample statistic and a measure of the sampling variability of the sample statistic. This lets us form an interval around the sample statistic that should contain the population parameter. Note, we are trying to contain the population parameter in the interval, not the data and not the sample statistic. In fact, the sample statistic are often the midpoint (center) of the interval.
Tests of Significance  Test claim about population parameter
The goal of a test of significance is to make a decision about the population parameter. Here are the steps we use:

1) Define the parameter(s) of interest. (Should also be able to define the observational units and variable)

2) Specify the hypotheses (e.g. \( H_0: \pi = 1/3, \mu = 50 \), etc.)

   *Always in terms of the population (parameters) because that’s what is unknown and what we are trying to make statements about* (take off the hats!)

   - The null hypothesis is the “dull hypothesis” or the “ho-hum hypothesis”
   - The alternative hypothesis specifies something interesting (“a-ha!”)

   One or two-sided (decide based on wording of research question)

3) Check the validity conditions, and identify the appropriate test procedure by name

   - If the technical conditions are not met, use a simulation method instead

4) Compare the data observed in the sample to what’s “expected” from \( H_0 \).

   Compute the test statistic and do one of the following:

   - Find the p-value = probability of observing a value of the statistic as extreme or more extreme *when \( H_0 \) is true*, using an applet or minitab.
   - Reject \( H_0 \) if \( p\text{-value} \leq \text{level of significance} \) otherwise fail to reject it.
   - Determine the rejection region for the given level of significance using an applet – plug in numbers for “as extreme as” until the probability is equal or just below the desired level of significance.
   - If your test-statistic is inside the rejection region, reject \( H_0 \) otherwise fail to reject it.

5) Draw a conclusion in context

   Make conclusion about research question of interest (back to English)

   - Clearly indicate the population you are generalizing to, taking into account your sampling method (convenience, simple random, etc.)
   - Be careful about inferring a cause-and-effect relationship if you’re comparing two variable based on your study design (was this an observational or experimental study?)

If we repeatedly took different samples or random shuffles and calculated the value of the statistic for each sample, the p-value indicates how often we would expect to see the statistic value that we actually did observe, or one more extreme, when \( H_0 \) is true. If the statistic value is very unlikely (so small p-value) we stop believing \( H_0 \). We can compare to the *significance level* as a benchmark to decide whether the p-value is “too small.”
**T vs Z** With proportions, our observations consist of “yeses” and “nos” for each observational unit in the population. A picture of this population is simply a bar graph. In particular, we don’t worry about its shape or variability. We will always consider approximating the null distribution of the sample proportions with the normal distribution. Thus, we never worry about using the $t$ distribution with proportions. With means, the $t$ distribution is used to take into account the extra variation we will see in the null distribution if we also substitute the sample standard deviation, $s$, into the equation. The key is that both a $z$-statistic and a $t$-statistic have “standardized” our observed statistic onto a comparable scale, e.g., the number of standard deviations from the mean.

**Population vs Variable vs Parameter** A population is a group of objects, a variable is what we measure about the objects, a question (e.g., height). The observational units are the objects we measure (e.g., buildings, volleyball players). You need to be able to decide how many populations you have and how many variables, e.g., are you measuring two different things/answering two different questions about the objects (e.g., height and age); are you measuring the same thing on two different groups (e.g., heights of men and heights of women). Parameters are numbers that describe the population, but we may not know their exact numerical value, (e.g., the average height of all buildings, the average age of all volleyball players).

**Independence/Matched Pairs** We can also assess the “independence” between samples to justify a two-sample procedure. This is not the same as the independence between variables. Instead we are making sure the responses of one group are not influencing or related to the responses in the other group. If they are, then a better analysis is to take that dependence into account (e.g., a “paired $t$” procedure).

**Other notes:**

- A p-value is **not** the probability of a null hypothesis or a conclusion being true.
- Remember your p-value allows you to make a conclusion about whether there is evidence again $H_0$ or not. We can’t say “there is strong evidence of no association” because we assumed no association in the calculations/simulation. So all we can say is “there is not strong evidence on an association.”
- Make the link between your p-value and your decision explicit. Don’t forget to then make a conclusion in context.
<table>
<thead>
<tr>
<th>Question Translations</th>
<th>If the question asks you to</th>
</tr>
</thead>
<tbody>
<tr>
<td>describe/compare the distribution(s) of a categorical variable</td>
<td>Look at (conditional) proportions</td>
</tr>
<tr>
<td>describe/compare the distribution(s) of a quantitative variable</td>
<td>Shape, center, spread Use mean, median, SD, IQR if available</td>
</tr>
<tr>
<td>estimate &quot;how large the difference is&quot; or &quot;plausible values for the parameter&quot;</td>
<td>Consider the confidence interval</td>
</tr>
<tr>
<td>test a parameter</td>
<td>Conduct a test of significance (use your “methods” table to find the appropriate test)</td>
</tr>
<tr>
<td>comment on generalizability</td>
<td>Consider the data collection methods and specify a reasonable population</td>
</tr>
<tr>
<td>comment on causation</td>
<td>Consider whether you have a randomized experiment and statistical significance</td>
</tr>
<tr>
<td>describe a confounding variable</td>
<td>Specify a variable and argue how it might differ between the explanatory variable groups and relate to the response variable</td>
</tr>
<tr>
<td>describe a parameter</td>
<td>Specify the number (e.g., mean or proportion or slope), the variable (e.g., how you are defining success), and the population (don’t worry too much at this point about whether it’s a reasonable population)</td>
</tr>
<tr>
<td>interpret a p-value</td>
<td>Begin the sentence “the probability that…” or “the proportion of …” and put your answer in context of the problem (e.g., what source of randomness are we are modelling? what statistic are you talking about, what value was observed, what the null hypothesis specified in context, what do you mean by “or more extreme”)</td>
</tr>
<tr>
<td>interpret a confidence interval</td>
<td>Begin the sentence “I am 95% confident that &lt;&lt;parameter&gt;&gt; is in (XX, XX)”. Clarify parameter, population, context If an interval about a difference, clarify which population parameter has a higher value.</td>
</tr>
<tr>
<td>interpret the confidence level</td>
<td>Talk about the reliability of the method, if you repeated the process for different samples, what percentage of the resulting intervals would succeed in capturing the parameter</td>
</tr>
<tr>
<td>draw a conclusion from a p-value</td>
<td>Comment on whether the p-value should be considered small, reject or fail to reject the null hypothesis, and restate the conclusion you are going with in the study context</td>
</tr>
<tr>
<td>calculate a confidence interval</td>
<td>Theory-based inference applet</td>
</tr>
<tr>
<td>calculate a p-value</td>
<td>Theory-based inference applet, or simulation in Minitab</td>
</tr>
<tr>
<td>state hypotheses</td>
<td>Probably want both null and alternative. Could ask for you to do this in words and/or in symbols. Make sure you are clearly talking about the population parameter and in context</td>
</tr>
<tr>
<td>identify the procedure</td>
<td>Name the test you would use (e.g., one proportion). Also be prepared to describe a simulation process you could use to estimate a p-value (e.g., flip a coin X times, shuffle X blue and X green cards X times)</td>
</tr>
<tr>
<td>comment on validity conditions</td>
<td>Consider the sample size condition for the relevant procedure as stated on your “methods” table</td>
</tr>
</tbody>
</table>
p-value interpretation template
X% of (randomness) would produce (statistic) at least as extreme (higher, lower, beyond) the observed result (cite the value) if the null is true (context)

CI interpretation template
I’m 95% confident that the (parameter – mean(s) or proportion(s), define population or process) is in the interval, context (variable, measurement units)
With comparison – make sure the “direction” is clear

Test statistic interpretation template (t or z statistic)
The (observed mean or proportion or difference in mean or proportion) is (insert absolute value of test statistic) standard deviations (below if -, above if +) the mean of (your variable).

R² interpretation template
X% of the variability in the (response variable y) is explained by a least squares regression model with (the explanatory variable x)

Regression slope interpretation template
A one (unit) increase in (the explanatory variable x) results in an X (unit) (increase if +, decrease if -) in the predicted value of (the response variable y).

CI level interpretation
This means 95% of intervals constructed this way would succeed in capturing the (parameter)

Power Interpretation
Probability of rejecting the null hypothesis when the alternative hypothesis is true. Calculated for a particular instance of the alternative hypothesis

Misc.
1 – P(type 2 error)
P(type 1 error) = level of significance
p-value = “observed” level of significance