Statistics with R Specialization
Course 2: Inferential Statistics

CLT and Sampling

LO 1. Define sample statistic as a point estimate for a population parameter, for example, the sample mean is used to estimate the population mean, and note that point estimate and sample statistic are synonymous.

LO 2. Recognize that point estimates (such as the sample mean) will vary from one sample to another, and define this variability as sampling variability (sometimes also called sampling variation).

LO 3. Calculate the sampling variability of the mean, the standard error, as $SE = \frac{\sigma}{\sqrt{n}}$ where σ is the population standard deviation.

- Note that when the population standard deviation σ is not known (almost always), the standard error SE can be estimated using the sample standard deviation *s*, so that $SE = \frac{s}{\sqrt{n}}$.

LO 4. Distinguish standard deviation (σ or s) and standard error (SE): standard deviation measures the variability in the data, while standard error measures the variability in point estimates from different samples of the same size and from the same population, i.e. measures the sampling variability.

LO 5. Recognize that when the sample size n increases we would expect the sampling variability to decrease.

- Conceptually: Imagine taking many samples from the population. When the size of each sample is large, the sample means will be much more consistent across samples than when the sample sizes are small.
- Mathematically: Remember $SE = \frac{\sigma}{\sqrt{n}}$. Then, when n increases SE will decrease since n is in the denominator.

Confidence Intervals

LO 1. Define a confidence interval as the plausible range of values for a population parameter.

LO 2. Define the confidence level as the percentage of random samples which yield confidence intervals that capture the true population parameter.

LO 3. Recognize that the Central Limit Theorem (CLT) is about the distribution of point estimates, and that given certain conditions, this distribution will be nearly normal.

- In the case of the mean the CLT tells us that if

(1a) the sample size is sufficiently large (n ≥ 30 or larger if the data are considerably skewed), or

(1b) the population is known to have a normal distribution, and

(2) the observations in the sample are independent,

then the distribution of the sample mean will be nearly normal, centered at the true population mean and with a standard error of $\frac{\sigma}{\sqrt{n}}$:

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

When the population distribution is unknown, condition (1a) can be checked using a histogram or some other visualization of the distribution of the observed data in the sample.

The larger the sample size (n), the less important the shape of the distribution becomes, i.e. when n is very large the sampling distribution will be nearly normal regardless of the shape of the population distribution.

LO 4. Recall that independence of observations in a sample is provided by random sampling (in the case of observational studies) or random assignment (in the case of experiments).

In addition, the sample should not be too large compared to the population, or more precisely, should be smaller than 10% of the population, since samples that are too large will likely contain observations that are not independent.

LO 5. Recognize that the nearly normal distribution of the point estimate (as suggested by the CLT) implies that a confidence interval can be calculated as

$$point\ estimate \pm z^\star \times SE,$$

,

where $z^\star$ corresponds to the cutoff points in the standard normal distribution to capture the middle XX% of the data, where XX% is the desired confidence level.

- For means this is: $\bar{x} \pm z^\star \frac{\sigma}{\sqrt{n}}$
- Note that $z^\star$ is always positive.

LO 6. Define margin of error as the distance required to travel in either direction away from the point estimate when constructing a confidence interval, i.e. $z^\star \frac{\sigma}{\sqrt{n}}$.

- Notice that this corresponds to half the width of the confidence interval.

LO 7. Interpret a confidence interval as “We are XX% confident that the true population parameter is in this interval”, where XX% is the desired confidence level.

- Note that your interpretation must always be in context of the data – mention what the population is and what the parameter is (mean or proportion).

Hypothesis Testing

LO 1. Notice that sampling distributions of point estimates coming from samples that don't meet the required conditions for the CLT (about sample size, skew, and independence) will not be normal.

LO 2. Formulate the framework for statistical inference using hypothesis testing and nearly normal point estimates:

1. Set up the hypotheses first in plain language and then using appropriate notation.
2. Identify the appropriate sample statistic that can be used as a point estimate for the parameter of interest.
3. Verify that the conditions for the CLT hold.
4. Compute the SE, sketch the sampling distribution, and shade area(s) representing the p-value.
5. Using the sketch and the normal model, calculate the p-value and determine if the null hypothesis should be rejected or not, and state your conclusion in context of the data and the research question.

LO 3. If the conditions necessary for the CLT to hold are not met, note this and do not go forward with the analysis. (We will later learn about methods to use in these situations.)

LO 4. Calculate the required sample size to obtain a given margin of error at a given confidence level by working backwards from the given margin of error.

LO 5. Distinguish between statistical significance vs. practical significance.

LO 6. Define power as the probability of correctly rejecting the null hypothesis (complement of Type 2 error).

t-distribution and comparing two means

LO 1. Use the t-distribution for inference on a single mean, difference of paired (dependent) means, and difference of independent means.

LO 2. Explain why the t-distribution helps make up for the additional variability introduced by using s (sample standard deviation) in calculation of the standard error, in place of σ (population standard deviation).

LO 3. Describe how the t-distribution is different from the normal distribution, and what "heavy tail" means in this context.

LO 4. Note that the t-distribution has a single parameter, degrees of freedom, and as the degrees of freedom increases this distribution approaches the normal distribution.

LO 5. Use a t-statistic, with degrees of freedom df=n−1 for inference for a population mean:

$$\text{CI: } \bar{x} \pm t^\star_{df} SE \qquad\qquad \text{HT: } T_{df} = \frac{\bar{x} - \mu}{SE}$$

where $SE = \frac{s}{\sqrt{n}}$.

LO 6. Describe how to obtain a p-value for a t-test and a critical t-score (t⋆df) for a confidence interval.

LO 7. Define observations as paired if each observation in one dataset has a special correspondence or connection with exactly one observation in the other data set.

LO 8. Carry out inference for paired data by first subtracting the paired observations from each other, and then treating the set of differences as a new numerical variable on which to do inference (such as a confidence interval or hypothesis test for the average difference).

LO 9. Calculate the standard error of the difference between means of two paired (dependent) samples as $SE = \frac{s_{diff}}{\sqrt{n_{diff}}}$ and use this standard error in hypothesis testing and confidence intervals comparing means of paired (dependent) groups.

LO 10. Use a t-statistic, with degrees of freedom $df = n_{diff} - 1$ for inference for the difference in two paired (dependent) means:

$$\text{CI: } \bar{x}_{diff} \pm t^\star_{df} SE \qquad\qquad \text{HT: } T_{df} = \frac{\bar{x}_{diff} - \mu_{diff}}{SE}$$

where $SE = \frac{s}{\sqrt{n}}$. Note that μdiff is often 0, since often $H_0 : \mu_{diff} = 0$

LO 11. Recognize that a good interpretation of a confidence interval for the difference between two parameters includes a comparative statement (mentioning which group has the larger parameter).

LO 12. Recognize that a confidence interval for the difference between two parameters that doesn't include 0 is in agreement with a hypothesis test where the null hypothesis that sets the two parameters equal to each other is rejected.

LO 13. Calculate the standard error of the difference between means of two independent samples as $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ and use this standard error in hypothesis testing and confidence intervals comparing means of independent groups.

LO 14. Use a t-statistic, with degrees of freedom $df = min(n_1 - 1, n_2 - 1)$ for inference for the difference in two independent means:

$$\text{CI: } (\bar{x}_1 - \bar{x}_2) \pm t^\star_{df} SE \qquad\qquad \text{HT: } T_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE}$$

where $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ Note that (μ1−μ2) is often 0, since often $H_0 : \mu_1 - \mu_2 = 0$.

LO 15. Calculate the power of a test for a given effect size and significance level in two steps: (1) Find the cutoff for the sample statistic that will allow the null hypothesis to be rejected at the given significance level, (2) Calculate the probability of obtaining that sample statistic given the effect size.

LO 16. Explain how power changes for changes in effect size, sample size, significance level, and standard error.

LO 17. Use bootstrap methods for confidence intervals for categorical variables with at most two levels.

ANOVA and bootstrapping

LO 1. Define analysis of variance (ANOVA) as a statistical inference method that is used to determine - by simultaneously considering many groups at once - if the variability in the sample means is so large that it seems unlikely to be from chance alone.

LO 2. Recognize that the null hypothesis in ANOVA sets all means equal to each other, and the alternative hypothesis suggest that at least one mean is different.

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$

H_A: At least one mean is different

LO 3. List the conditions necessary for performing ANOVA

1. the observations should be independent within and across groups
2. the data within each group are nearly normal
3. the variability across the groups is about equal and use graphical diagnostics to check if these conditions are met.

LO 4. Recognize that the test statistic for ANOVA, the F statistic, is calculated as the ratio of the mean square between groups (MSG, variability between groups) and mean square error (MSE, variability within errors). Also recognize that the F statistic has a right skewed distribution with two different measures of degrees of freedom: one for the numerator ($df_G = k - 1$, where k is the number of groups) and one for the denominator ($df_E = n - k$, where n is the total sample size).

- Note that you won't be expected to calculate MSG or MSE from the raw data, but you should have a conceptual understanding of how they're calculated and what they measure.

LO 5. Describe why calculation of the p-value for ANOVA is always “one sided”.

LO 6. Describe why conducting many t-tests for differences between each pair of means leads to an increased Type 1 Error rate, and we use a corrected significance level (Bonferroni correction, $\alpha^\star = \alpha/K$, where K is the number of comparisons being considered) to combat inflating this error rate.

- Note that $K = \frac{k(k-1)}{2}$, where k is the number of groups.

LO 7. Describe why it is possible to reject the null hypothesis in ANOVA but not find significant differences between groups when doing pairwise comparisons.

LO 8. Describe how bootstrap distributions are constructed, and recognize how they are different from sampling distributions.

LO 9. Construct bootstrap confidence intervals using one of the following methods:

- Percentile method: XX% confidence level is the middle XX% of the bootstrap distribution.
- Standard error method: If the standard error of the bootstrap distribution is known, and the distribution is nearly normal, the bootstrap interval can also be calculated as $point\ estimate \pm t^\star SE_{boot}$.

LO 10. Recognize that when the bootstrap distribution is extremely skewed and sparse, the bootstrap confidence interval may not be reliable.

Inference for proportions

LO 1. Define population proportion p (parameter) and sample proportion $\hat{p}$.

LO 2. Calculate the sampling variability of the proportion, the standard error, as $SE = \sqrt{\frac{p(1-p)}{n}}$, where p is the population proportion.

- Note that when the population proportion p is not known (almost always), this can be estimated using the sample proportion, $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

LO 3. Recognize that the Central Limit Theorem (CLT) is about the distribution of point estimates, and that given certain conditions, this distribution will be nearly normal. In the case of the proportion the CLT tells us that if

- the observations in the sample are independent,
- the sample size is sufficiently large (checked using the success/failure condition: $np \geq 10$ and $n(1-p) \geq 10$,

then the distribution of the sample proportion will be nearly normal, centered at the true population proportion and with a standard error of $SE = \sqrt{\frac{p(1-p)}{n}}$.

$$\hat{p} \sim N\left(mean = p, SE = \sqrt{\frac{p(1-p)}{n}}\right)$$

LO 4. Note that if the CLT doesn't apply and the sample proportion is low (close to 0) the sampling distribution will likely be right skewed, if the sample proportion is high (close to 1) the sampling distribution will likely be left skewed.

LO 5. Remember that confidence intervals are calculated as

$$\text{point estimate} \pm \text{margin of error}$$

and test statistics are calculated as

$$\text{test statistic} = \frac{\text{point estimate - null value}}{\text{standard error}}$$

LO 6. Note that the standard error calculation for the confidence interval and the hypothesis test are different when dealing with proportions, since in the hypothesis test we need to assume that the null hypothesis is true – remember: p-value = P(observed or more extreme test statistic | H0 true).

- For confidence intervals use $\hat{p}$ (observed sample proportion) when calculating the standard error and checking the success/failure condition:

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- For hypothesis tests use p0 (null value) when calculating the standard error and checking the success/failure condition:

$$SE_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}}$$

- Such a discrepancy doesn't exist when conducting inference for means, since the mean doesn't factor into the calculation of the standard error, while the proportion does.

LO 7. Explain why when calculating the required minimum sample size for a given margin of error at a given confidence level, we use p^=0.5 if there are no previous studies suggesting a more accurate estimate.

- Conceptually: When there is no additional information, 50% chance of success is a good guess for events with only two outcomes (success or failure).

- Mathematically: Using $\hat{p} = 0.5$ yields the most conservative (highest) estimate for the required sample size.

LO 8. Note that the calculation of the standard error of the distribution of the difference in two independent sample proportions is different for a confidence interval and a hypothesis test.

- confidence interval and hypothesis test when $H_0 : p_1 - p_2$ = some value other than 0:

$$SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- hypothesis test when $H_0 : p_1 - p_2 = 0$:

$$SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_1} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_2}},$$

,

where $\hat{p}_{pool}$ is the overall rate of success:

$\hat{p}_{pool}$=number of successes in group 1 + number of successes in group 2 (n1+n2)

LO 9. Note that the reason for the difference in calculations of standard error is the same as in the case of the single proportion: when the null hypothesis claims that the two population proportions are equal, we need to take that into consideration when calculating the standard error for the hypothesis test, and use a common proportion for both samples.

Simulation based inference for proportions and chi-square testing

LO 1. Use a chi-square test of goodness of fit to evaluate if the distribution of levels of a single categorical variable follows a hypothesized distribution.

H_0: The distribution of observed counts follows the hypothesized distribution, and any ob- served differences are due to chance.

H_A: The distribution of observed counts does not follow the hypothesized distribution.

LO 2. Calculate the expected counts for a given level (cell) in a one-way table as the sample size times the hypothesized proportion for that level.

LO 3. Calculate the chi-square test statistic as

$$\chi^2 = \sum_{i=1}^{k} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}},$$

, where k is the number of cells.

LO 4. Note that the chi-square statistic is always positive, and follows a right skewed distribution with one parameter: degrees of freedom.

LO 5. Note that the degrees of freedom for the chi-square statistic for the goodness of fit test is $df = k-1$.

LO 6. List the conditions necessary for performing a chi-square test (goodness of fit or independence)

1. the observations should be independent
2. expected counts for each cell should be at least 5
3. degrees of freedom should be at least 2 (if not, use methods for evaluating proportions)

LO 7. Describe how to use the chi-square table to obtain a p-value.

LO 8. When evaluating the independence of two categorical variables where at least one has more than two levels, use a chi-square test of independence.

H_0: The two variables are independent.

H_A: The two variables are dependent.

LO 9. Calculate expected counts in two-way tables as

$$E = \frac{\text{row total} \times \text{column total}}{\text{grand total}}$$

LO 10. Calculate the degrees of freedom for chi-square test of independence as $df = (R-1) \times (C-1)$, where R is the number of rows in a two-way table, and C is the number of columns.

LO 11. Note that there is no such thing as a chi-square confidence interval for proportions, since in the case of a categorical variables with many levels, there isn't one parameter to estimate.

LO 12. Use simulation methods when sample size conditions aren't met for inference for categorical variables.

- Note that the t-distribution is only appropriate to use for means. When sample size isn't sufficiently large, and the parameter of interest is a proportion or a difference between two proportions, we need to use simulation.

LO 13. In hypothesis testing

- for one categorical variable, generate simulated samples based on the null hypothesis, and then calculate the number of samples that are at least as extreme as the observed data.
- for two categorical variables, use a randomization test.