Module 7 Review

Overview of Statistical Inference

- A **parameter** is a number that ***describes*** a **population characteristic**.
- A **statistic** is a number that ***describes*** a **sample characteristic**.
    - For a categorical variable, the parameter and the statistics are **proportions**.
    - For a quantitative variable, the parameter and statistics are **means**.
- **Inference** is based on probability.
- We use a statistic to draw a conclusion about a parameter. These conclusions include a probability statement that describes the strength of the evidence or our certainty.
- For a given situation, we assume that the parameter is fixed. It does not change. We create simulations and mathematical models to describe the variability we expect to see in sample statistics.

Sampling Distribution for a Sample Proportion

- Larger samples have less variability.
- For a categorical variable we assume that population has a proportion *p* of successes. And can create a mathematical model for the distribution which will have the following center, spread and shape:

  Center: Mean of the sample proportions, $\mu_{\hat{p}}$, is the population proportion, p.Therefore $\boldsymbol{\mu_{\hat{p}} = p}$.

  Spread: Standard deviation (**standard error**) of the sample proportions is $\boldsymbol{\sigma_{\hat{p}} = \sqrt{\frac{p\cdot(1-p)}{n}}}$.

  Shape: A normal model is a good fit if the expected number of successes and failures is at least 10: $(np \geq 10 \; and \; n(1-p) \geq 10)$.

- When a normal distribution is a good fit for the sampling distribution, we can calculate a Z-score $\left(\boldsymbol{z = \frac{\hat{p}-p}{\sigma_{\hat{p}}}}\right)$for a given sample proportion, allowing us to use the normal distribution to find probabilities.
- Z-scores that are >2 or<-2 are unusual.
- Probabilities that are< 5% or <.05 are unusual.
- When looking for Area or Probability, we need a Z-score so we can use **DISTR>normalcdf**
- When given a probability and asked to find a value, we need to first find a z-score using **DISTR>invnorm**, then reverse the Z-score formula to find the value $(x = \mu + z\sigma)$.

Confidence Intervals

- The goal of a confidence interval is to estimate a parameter value.
  - Sample statistics are used to create a confidence interval.
  - Sample statistics vary, so there is always error in our estimate. So we use the standard error, which is the average error in our sample estimates, to create a margin of error.
  - The margin of error is related to our confidence that the interval contains the population parameter.
  - The **estimated standard error** is calculated as $s_{\hat{p}} = \sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}$
  - To use the Normal distribution, check there are at least 10 yes and 10 no responses.
  - Confidence intervals can be calculated for any level of confidence, but are in practice limited to 90%, 95% and 99%
  - Critical value $\mathbf{Z_c}$ is the determined from using Invnorm. Here are the values for the practical confidence intervals:

Confidence Level C	90%	95%	99%
Critical value $\mathbf{Z_c}$	1.645	1.96	2.576

  - For an estimated **<u>confidence interval</u>**, to find the margin of error, E, we multiply the standard error, $s_{\hat{p}}$, by the critical value $\mathbf{Z_c}$ so we get:

$$E = Z_c s_{\hat{p}}$$

- To calculate the confidence interval, we use the sample statistic as the center of the interval. The endpoints of the interval will be $\hat{p} \pm E$.

- TI-calculator command for confidence interval: **STAT>TESTS>1-PropZInt**

- Interpretation of the interval:
  - We say we are C% confident that the <proportion in context> is between <lower bound> and <upper bound>.
  - C% of the time these intervals will actually contain the true population proportion and we will be right.
  - (100-C)% of the time, we will be wrong.

Hypothesis Tests in General:

Hypothesis Tests consist of 4 steps. These steps apply to all the hypothesis tests we will do in this course.

Step 1: Determine the hypotheses.

The hypotheses are statements about the parameter(s) in question. The null hypothesis, H_0 is always a statement of equality and usually means no change or difference. The alternative hypothesis, H_a, is always an inequality, either <, >, or ≠ and is based on the research question.

Step 2: Collect the data.

The data must come from a random sample that is representative of the population in question.

Step 3: Assess the evidence.

The p-value is the evidence. The p-value is the probability that we would get sample results at least as extreme as those observed if the null hypothesis is true. If the p-value is smaller than the significance level, the results are unusual enough for us to reject the null hypothesis. Otherwise, we "fail to reject" the null hypothesis.

Step 4: Give the conclusion.

Our conclusion is stated in terms of the alternative hypothesis. Either "there is" or "there is not" enough evidence to say that the alternative hypothesis is true. We always use the context of the problem in the conclusion and always include the p-value. Lastly, we never say that the null hypothesis is true, only that we reject or fail to reject it.

Hypothesis Test for a Population Proportion:

For the 4 steps, the following are specific to hypothesis testing for a population proportion.

Step 1: Determine the hypotheses.

The hypotheses for a test about a population proportion are stated in terms of the p. Here p_o is a number to which we compare the population proportion to.

First, read the problem and write **Ha in words.** This will help you find the appropriate symbol for Ha.Then write Ho and Ha using population proportion p:

$$H_{0:}\, p = p_0$$

$$H_{a:}\, p > p_0 \quad \text{or} \quad Ha{:}\, p < p_0 \quad \text{or} \quad Ha{:}\, p \neq p_0$$

Step 2: Collect the data.

We also check at this point that $np \geq 10$ and $n(1-p) \geq 10$ where p is the value from the null hypothesis,. If these conditions are true, a normal model is a good fit for the sampling distribution of sample proportions. We need this model to do the remaining steps in the hypothesis test.

$$\hat{P} = \frac{X}{n} = \frac{Number\ of\ \ Successes}{Sample\ Size}$$

Step 3: Assess the evidence.

We calculate the test statistic (the *Z*-score) for our sample proportion. We use the test statistic to determine the P-value, using a standard normal curve. We can do this using the *formulas* or technology.

As always, if the p-value is smaller than the significance level, the results are unusual enough for us to reject the null hypothesis. Otherwise, we "fail to reject" the null hypothesis.

To get Z and the p-value on TI calculator: **STAT>TESTS>1-PropZTest**

To get p-value using formulas:

$Z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$	If Ha: $p < \#$, pvalue = NORMCDF (-99999, Z) If Ha: $p > \#$, pvalue = NORMCDF (Z, 99999) If Ha: $p \neq \#$, pvalue = 2 x NORMCDF (Z, 99999) if Z is positive = 2 x NORMCDF (-99999, Z) if Z is negative

- **If the p-value < significance level (α), reject Ho**
- **If the p-value ≥ significance level (α), fail to reject Ho**

Step 4: Give the conclusion.

- **If Ho is rejected:** <Ha in words.> The results are statistically significant.
- **If Ho is not rejected:** "There is insufficient evidence to conclude <Ha in words>.

Important definitions

Statistical Inference
The process of estimating or testing hypotheses of population parameters using statistics from a random sample.

Parameter
A fixed numerical value that describes a characteristic of a population.

Estimation
An inference process that attempts to predict the values of population parameters based on sample statistics.

Point Estimate
A single sample statistic that is used to estimate a population parameter. For example, $\bar{X}$ is a point estimator for μ.

Confidence Interval
An Interval estimate that estimates a population parameter from a random sample using a predetermined probability called the level of confidence.

Level of Confidence
The probability, usually expressed as a percentage, that a Confidence Interval will contain the true population parameter that is being estimated.

Margin of Error
The distance in a symmetric Confidence Interval between the Point Estimator and an endpoint of the interval. For example a confidence interval for p may be expressed as $\hat{p} \pm$ Margin of Error.

Model Assumptions

Criteria which must be satisfied to appropriately use a chosen statistical model. For example, to use the normal distribution for inference about a proportion, show that np>10 and n(1-p)>10.

Hypothesis
A statement about the value of a population parameter developed for the purpose of testing.

Hypothesis Testing
A procedure, based on sample evidence and probability theory, used to determine whether the hypothesis is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.

Null Hypothesis (H_0)
A statement about the value of a population parameter that is assumed to be true for the purpose of testing.

Alternative Hypothesis (H_a)
A statement about the value of a population parameter that is assumed to be true if the Null Hypothesis is rejected during testing.

Level of Significance (α)
The maximum probability, set by design, of rejecting the Null Hypothesis when it is actually true (conditional probability of making Type I error **given Ho is true.**)

p-value
The probability, assuming that the Null Hypothesis is true, of getting a value of the test statistic at least as extreme as the computed value for the test. If the p-value is less than α , Ho is rejected.

Type I Error
Rejecting the Null Hypothesis when it is actually true.

Type II Error
Failing to reject the Null Hypothesis when it is actually false.