ST517 Note Outline 5: Estimation

Lecture 5.1: Introduction and Foundations

Recall:

- Process of Data Analysis
    1. Identify the research question(s)
    2. Conduct background research
    3. Form a hypothesis or make a prediction
    4. Experiment / Collect data
    5. Explore, summarize and analyze data
        - Explore and summarize data = exploratory data analysis (EDA)
        - Analyze data = apply tools of statistical inference
    6. Make conclusions
- *Statistical inference* is the process of using sample information to make conclusions about the population
    - Necessary due to existence of sampling variability
- *Sampling variability* is the variation in the possible values of a sample statistic that results from selecting different random samples
- *Sampling distribution* is the distribution the possible values a statistic could take
    - Describes how likely values are to occur, so we can determine which values would be typical (or surprising) without having to take multiple samples
    - Pattern of sampling variability, which is predictable if sample is random
    - This predictability makes statistical inference possible

Tools of Statistical Inference

- ______________________________ – Provides a range of reasonable estimates (i.e. best guesses) of the true value of a population parameter
    - E.g. We believe that average (fasting) blood sugar for individuals without diabetes is between 70 and 92 mg/dL.

- ______________________________ – Provides evidence of a statistically significant effect in a population
    - E.g. There is evidence that a particular medication can reduce blood sugar for individuals with diabetes.

- ____
    - Explore relationships between variables
    - Used for estimation and prediction
    - Confidence intervals and hypothesis tests play a role in regression as well

This outline: Focus on confidence intervals

Connection between the Sampling Distribution and Confidence Intervals

- Want to know true value of population parameter
- Single best guess:

- This value will be different for different samples
- But we know how the possible values of the statistic are expected to behave

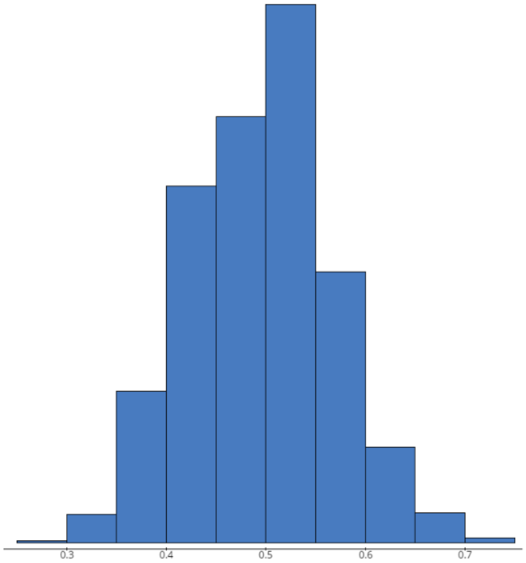
- Recall: Sampling distribution of sample mean $\bar{y}$
  - Conditions:
    - Sample is random
    - Population is normal
      - If sample is large (i.e. $n \geq 30$), don't worry about non-normality in population (thanks to *Central Limit Theorem*)
  - Shape: Well approximated by normal distribution
  - Center: True value of population mean μ
  - Standard deviation of the sample mean: $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$
  - For full details, review Lecture 4.5
- Recall: Sampling distribution of sample proportion $\hat{p}$
  - Conditions:
    - Sample is random
    - Sample is large (i.e. $np \geq 10$ and $n(1-p) \geq 10$)
  - Shape: Well approximated by normal distribution
  - Center: True value of population proportion p
  - Standard deviation of the sample proportion: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
  - For full details, review Lecture 4.6

Lecture 5.2: Margin of Error

Margin of Error

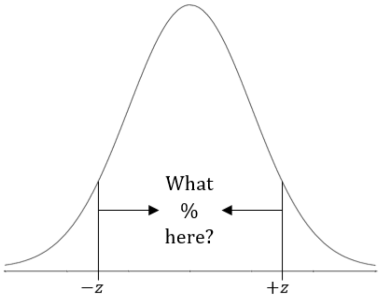
- **Margin of Error (MOE)** is a numeric indication of how far the value of a sample statistic may be from the true value of the population parameter

- Logic behind the margin of error
  - Determine a distance based on the sampling distribution
  - Put that distance around the statistic
  - For most samples, distance will extend far enough to capture true parameter

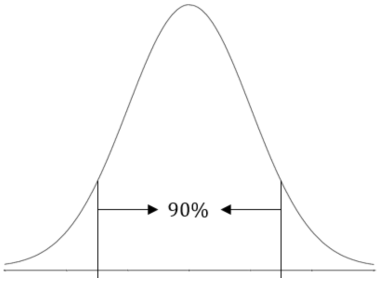


How large should the margin of error be?

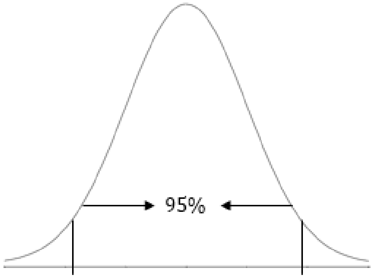
- Depends on the standard deviation of the statistic
  - For means: $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$ For proportions: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
- Depends on the percentage of the sampling distribution we would like to capture



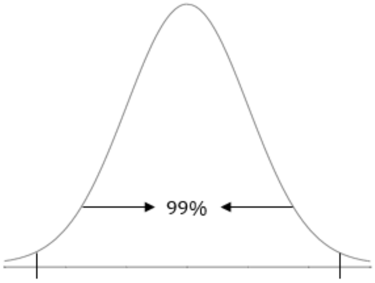
  - Suppose we want to divide the standard normal distribution so that 90% of the area is in the middle. Based on what we've learned, how do we do this?



What about 95%?



What about 99%?



General Form for the MOE:

MOE for a Proportion:

- Multiplier found using

MOE for a Mean:

- When σ is known:
- When σ is <u>not</u> known:
- Explaining distinction between these requires its own lecture!

Example: A simple random sample of 100 undergraduate students from a large university found that 78 of them had used the university library's website to find resources for a class.

a. Calculate and interpret the 90% margin of error.

b. This study also found that students spent an average of 15 minutes, with a MOE of 5 minutes, on the library's website looking for resources. Interpret the margin of error.

Lecture 5.3: The t-distribution

Recall: MOE for a Mean

σ known: $z^*_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$

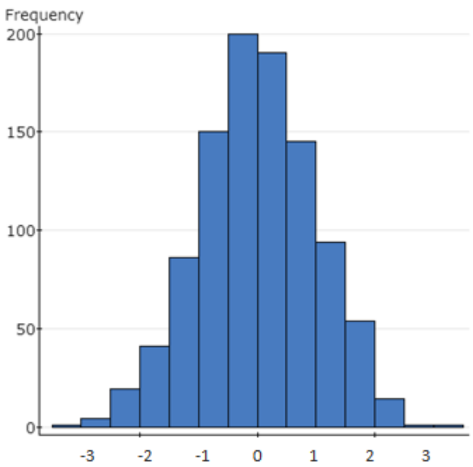
σ unknown: $t^*_{\alpha/2,n-1} \times \frac{s}{\sqrt{n}}$

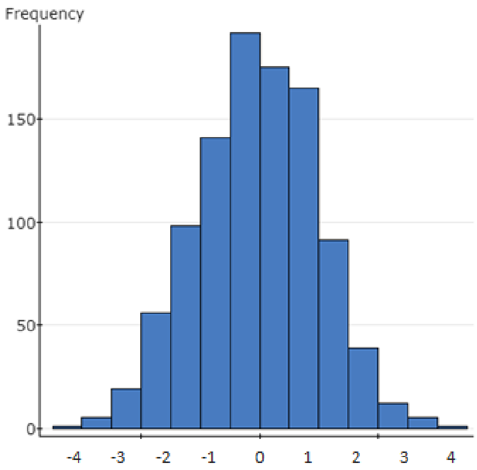
Why the distinction?

- In practice:

- Solution:

- Consequence:





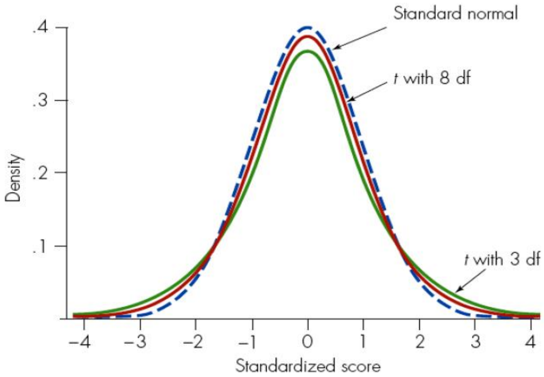
The t-distribution

- Also known as Student's t-distribution
- Most commonly used when conducting inference for a population mean
  - Models additional uncertainly due to population standard deviation being unknown

- PDF: $f(x;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$

  - Notation:

- Bell shaped, unimodal, symmetric
- Depends on one parameter:



  - Ex: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

- As $\nu \to \infty$, the t-distribution converges to the standard normal distribution
- If you know ν, you know everything there is to know about that t-distribution

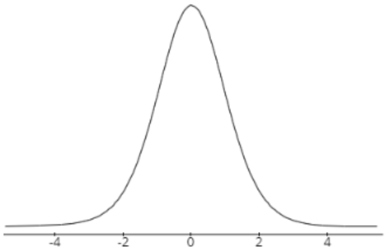
- Mean of t-distribution:

- Variance of t-distribution:

The t-distribution: Finding Probabilities

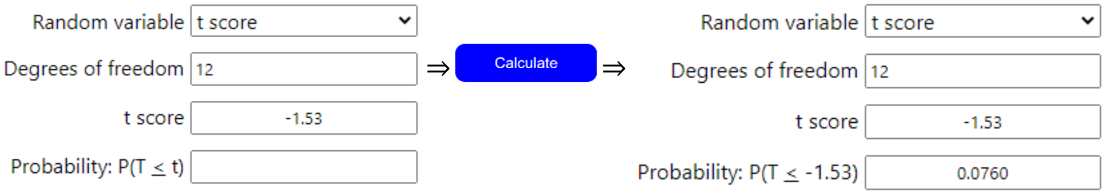
- Finding probabilities under a t-distribution will be particularly important for calculating p-values when we learn about hypothesis testing in the next outline
- As with finding probabilities under a normal distribution, focus on conceptual understanding and use technology to calculate the value
  - Graphing calculator (e.g. TI-84): `tcdf(lower_bound,upper_bound,df)`
  - Software, e.g.
    - SAS: `DATA temp; prob_t=cdf('t',UB,df);`
      `PROC PRINT; var prob_t; run;`
    - Excel: `t.dist(upper_bound,df,TRUE)`
  - Online calculators
    - E.g. stattrek.com/online-calculator/t-distribution.aspx
    - Fill in: *df* and t score; computer provides area to the left of that t score

Example: For a t-distribution with $\nu = 12$, find $P(t \leq -1.53)$.



Using the technology for this example:

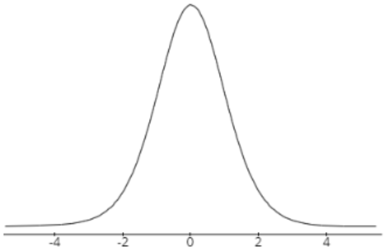
- Graphing calc: `tcdf(-1000,-1.53,12)`
- SAS: `DATA temp; prob_t=cdf('t',-1.53,12);`
  `PROC PRINT; var prob_t; run;`
- Excel: `t.dist(-1.53,12,TRUE)`
- Online calc:


Random variable: t score
Degrees of freedom: 12
t score: -1.53
Probability: P(T ≤ t):

Random variable: t score
Degrees of freedom: 12
t score: -1.53
Probability: P(T ≤ -1.53): 0.0760

- From each of these, the probability is:

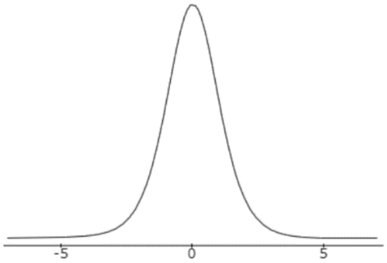
Example: For a t-distribution with $\nu = 19$, find $P(t > 2.01)$.



The t-distribution: Finding Percentiles

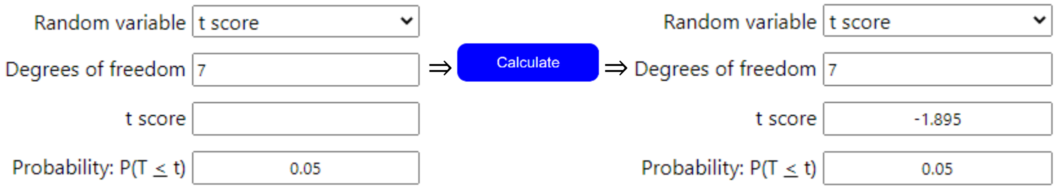
- Confidence coefficients are percentiles of a distribution
- As with finding percentiles under a normal distribution, focus on conceptual understanding and use technology to calculate the value
    - Graphing calculator (e.g. TI-84): `invT(proportion to left, df)`
    - Software, e.g.
        - SAS: `DATA temp; t=tinv(proportion to left, df);`
          `PROC PRINT; var t; run;`
        - Excel: `t.inv(proportion to left, df)`
    - Online calculators
        - E.g. stattrek.com/online-calculator/t-distribution.aspx
        - Fill in: Value of *df* and the proportion to the left; computer will provide the t-score that is the value of the confidence coefficient

Example: For a t-distribution with $\nu = 7$, what is the multiplier for 90% confidence?



Using the technology for this example:

- Graphing calc: `invT(0.05,7)` OR `invT(0.95,7)`
- SAS: `DATA temp; t=tinv(0.05,7)` OR `tinv(0.95,7)`
  `PROC PRINT; var t; run;`
- Excel: `t.inv(0.05,7)` OR `t.inv(0.95,7)`
- Online calc:



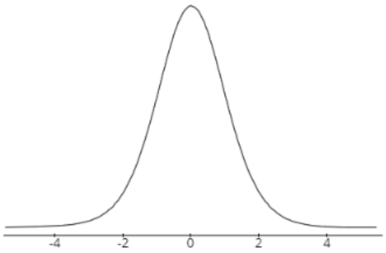
Random variable: t score
Degrees of freedom: 7
t score:
Probability: P(T ≤ t): 0.05

⇒ Calculate ⇒

Random variable: t score
Degrees of freedom: 7
t score: -1.895
Probability: P(T ≤ t): 0.05

- From each of these, the value of the multiplier is:
    - Don't worry about sign! Just report the positive value

Example: For a t-distribution with $\nu = 21$, what is the multiplier for 90% confidence?



Lecture 5.4 Confidence Intervals

Recall

- A *point estimate* is a single guess of the true value of the population parameter
    - Corresponding sample statistic is best point estimate
    - But because of *sampling variability*, we know the value of the estimate will vary from sample to sample
- *Margin of Error (MOE)* is a numeric indication of how far the value of a sample statistic may be from the true value of the population parameter
    - Accounts for amount of sampling variability via the standard deviation of the statistic
    - Accounts for percentage of sampling distribution we would like to capture via the percentiles/multiplier/confidence coefficient

Limitation of Point Estimation

- Want to use point estimate to tell us value of population parameter
- But value of point estimate depends on sample

- How do we use something that changes to estimate something that is fixed?
    - Need an estimator that accounts for sampling variability
    - Rather than only reporting a single guess at the value of the parameter, report a range of guesses, called an **interval estimate**
    - Most common type of interval estimate is a confidence interval

Confidence Intervals

- A **confidence interval (CI)** represents a range of reasonable estimates for the true value of a population parameter

- Interpretation:

General form for a CI:

CI for a Population Proportion:

  - Multiplier found using
  - Conditions required for this interval to be valid:

CI for a Population Mean (when σ is unknown):

  - Multiplier found using
  - Conditions required for this interval to be valid:

How often do we capture the parameter?

- Sometimes the confidence interval will include the true parameter

- Sometimes it will not

- **<u>Confidence level</u>**: Percent of possible samples in which MOE captures parameter
  - Percent of possible samples for which CI includes true value of parameter
  - Interpretation (for 95% confidence):

Example: How often do you laugh in a typical day? A study randomly selects 30 adults and records how often they laugh in a day. The sample average is 21 laughs per day with a standard deviation of 13.7 laughs.

a. What are the response variable and parameter of interest?

b. What is a possible research question of interest?

c. Calculate a 90% CI for the parameter and interpret this interval in context.

d. Calculate a 95% CI for the parameter and interpret this interval in context.

Example: A simple random sample of 100 undergraduate students from a large university found that 78 of them had used the university library's website to find resources for a class.

a. What are the response variable and parameter of interest?

b. What is a possible research question of interest?

c. Calculate a 95% CI for the parameter.

d. Interpret the confidence interval in context.

e. Interpret the confidence level in context.

Lecture 5.5: Important Points about Confidence

- A larger level of confidence produces a larger margin of error.

- A large sample size produces a smaller margin of error.

- News media uses 95% confidence by convention.

- Confidence is arbitrary.

- MOE only accounts for random sampling variability.

- The confidence interval is about the parameter not about the statistic or the individuals.

- Confidence is in the procedure.

Lecture 5.6: Additional Examples

Example 1: Of 3,865 full-time employees surveyed by the Gallup Organization, 42% said that the ability to use their computers and mobile devices to stay connected to the workplace outside of their normal working hours was a 'strongly positive' development.

a. Calculate and interpret a 99% confidence interval for the parameter in this study.

b. Are the conditions necessary for this interval to be valid met?

Example 2: An economist was interested in how recent economic events have influenced home sizes in Maricopa County, Arizona. The researcher used the county data base of homes to randomly select 200 homes and recorded several variables on each. One of the variables was the size of the home (in square feet). The researcher found that the average size of the homes was 1929 sqft and the standard deviation of 833 sqft. Use this sample information to calculate a 95% confidence interval for the mean size of all homes in Maricopa County.

Example 3: Data collected by child development scientists produced the following 90% CI for the average age (in weeks) at which babies begin to crawl: (29.2, 31.8). You can assume that the conditions necessary for inference using this interval were met.

a. What was the sample average age at which babies begin to crawl found by these scientists?

b. What was the margin of error for this interval?

c. Below are three <u>incorrect</u> interpretations for the confidence interval. For each, explain why the interpretation is incorrect.

1. "90% of all babies begin to crawl between 29.2 and 31.8 weeks of age."

2. "90% of all samples will have mean ages between 29.2 and 31.8 weeks."

3. "The mean age at which babies begin to crawl is between 29.2 and 31.8 weeks 90% of the time."

d. Write a correct interpretation of the confidence level.