Probability Theory and Introductory Statistics

- ALY6010, Section 81229
- Week 2: May 31, 2022
- Professor Dan Koloski (Professor K)
- d.koloski@northeastern.edu
- Roux Institute at Northeastern University

Agenda: 5/31/2022

Time Slot	
6:00-6:15 (15min)	Recap of last week; Breakouts to Review Bluman Problems
6:15-6:55 (40min)	Section 1: Using Sample Statistics to Estimate Population Parameters (Part 1)
6:55-7:00 (5min)	Break
7:00-7:40 (40min)	Section 2: Confidence Intervals and Estimating Population Parameters (Part 2)
7:40-7:45 (5min)	Break
7:45-8:35 (40min)	Section 3: Final Project Overview and Helpful new R Techniques
8:35-8:40 (5min)	Wrap-Up

Week 1 Objectives and Task List

Learning Objectives

By the end of this module, you should be able to:

- Distinguish between descriptive statistics and inferential statistics
- Calculate expected value and variance
- Use computational tools (such as R) to calculate probability distribution
- Use R to create a descriptive statistics table by sub-group

Task List

- View lessons in Canvas
- Read Elementary Statistics, Chapters 5 & 6
- Review R in Action, Chapters 1-5
- Complete primary Discussion post by Thursday
- Complete Practice Problem Set (not submitted)
- Complete R Practice assignment
- Take quiz
- Review and start Final Project

Observations from Week 1 Assignments

Things I liked seeing

- Stepwise, careful and thorough exploratory data analysis (EDA)
- A variety of plot types for different analytical emphases (not just histograms)
- Precise writeup language
- Clean, well-commented code

Areas to focus on

- Code commenting
- Assigning output you want to use
- Data intake (read.csv, etc.)

Breakouts: Week 1 Problem Sets

- In your Zoom Breakout Rooms
  - Share your Blumen assignments and answers with each other

- Identify any outstanding questions the group has and bring back to the larger group

Week 2 Objectives and Task List

Learning Objectives

By the end of this module, you should be able to:

- Distinguish sample distribution vs. sampling distribution
- Connect the Central Limit Theorem and the Law of Large Numbers to sample statistics
- Understand the role of estimator and standard error
- Calculate a confidence interval around mean and proportion
- Determine minimum sample size

Task List

- View lessons in Canvas
- Read *Elementary Statistics*, Chapter 7
- Read *R in Action*, Chapter 6, all Sections and Chapter 7, Sections 1 & 2
- Complete **primary Discussion post by Thursday (2 secondary by Saturday)**
- Complete Practice Problem Set (not submitted)
- **Complete and SUBMIT R Practice assignment**
- Take quiz
- **Complete and submit Final Project - Milestone 1**

Section 1

Using Sample Statistics to Estimate Population Parameters (Part 1)

Samples and Populations: Vocabulary

- Population parameter
  - hypothesized parameter about a population
- Hypothesis testing
  - test/validate an assumed hypothesis
- **Variables of note: $\mu, \sigma, \sigma^2, p, q, X$ (individual values)**
  - **α = "alpha" (new this week)**

- Sample statistic
  - measured parameter from a sample of the population
- Estimator
  - the sample statistic used to estimate a population parameter
- **Variables of note: $\overline{x}$, s (a.k.a. $\sigma_{\overline{x}}$), s^2, n, X (individual values)**
  - **$\hat{p}$ = "p-hat", $\hat{q}$ = "q-hat", E = "Margin of Error" (new this week)**

Central Limit Theorem

- Central Limit Theorem
  - A distribution of Sample Means ($\bar{x}$) approaches normal distribution when n >=30 and approaches t-distribution when n <30, regardless of the actual distribution of X
  - Distribution of Sample Variance $\mathbf{s^2}$ approaches chi-square distribution when n >=30, regardless of distribution of X
  - N <= 30 is a rule of thumb for using t-distribution vs normal distribution
- Law of Large Numbers
  - ($\bar{x}$) approaches Expected Value of X, EV(X) = μ Population Mean as n approaches infinity

- THEREFORE
  - If we know something about the population statistics, we can use the standard normal distribution to make inferences about hypothetical sample statistics
  - **AND, if we know something about sample statistics, we can also use the standard normal distribution to make inferences about hypothetical population statistics, with various levels of confidence**

Sample Distribution vs Sampling Distribution

- **Sample distribution** finds $\bar{x}$ for a single sample
- **Sampling distribution** finds mean($\bar{x}$) and SD($\bar{x}$) for a bunch of $\bar{x}$ from multiple samples
  - **Mean($\bar{x}$)** = "**mean of means**"

    = 1/n * (($\bar{x}$1)+ ($\bar{x}$2) + ($\bar{x}$3) + ….) = 1/n * Σ($\bar{x}$)
  - $\boldsymbol{\sigma_{\bar{x}}}$ = **"standard deviation of sample means"**

    = ($\bar{x}$ – mean($\bar{x}$))^2= σ(X) / SQRT(n)
  - $\boldsymbol{\sigma^2_{\bar{x}}}$ = **"variance of sample means"** = $\sigma_{\bar{x}}{}^2/n$

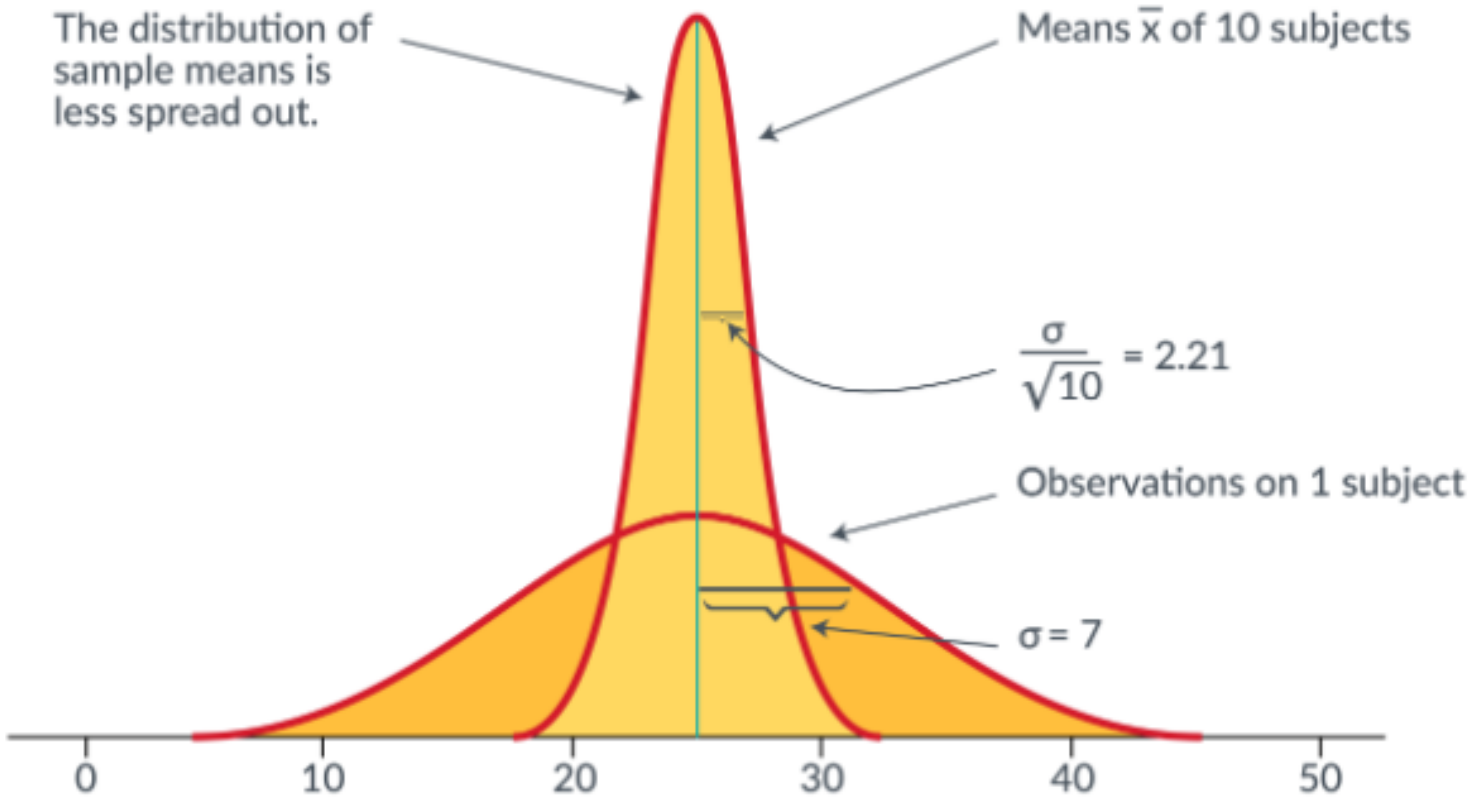
P(X) = probability distribution of actual X values

P($\bar{x}$) = probability distribution of sample means

- P($\bar{x}$) is always normally distributed when n > 30
- P($\bar{x}$) is always t-distributed when n <= 30

Sample Distribution vs Sampling Distribution

Distribution of X and $\overline{X}$ are both treated as random variables as seen in this graph below.



Reminder: Z Score and T-Statistic

Z Score (Standard Normal Dist.)

- Used when sample size **n** >= 30
- Used when you know population mean and population standard deviation
  - $Z = \frac{x-\mu}{\sigma}$
- Z (distance from mean) helps estimate probability (p-value, or area under the curve) of value of X being a certain distance from the mean
- Z(0) = Z(μ) = 0.5

T-Statistic (T-Distribution)

- Used when sample size **n** < 30
- T-Distribution is wider and flatter than Standard Norm. Dist.
  - Less precise estimation as a result
- Used when you don't know population mean and population standard deviation
- T-Statistic (distance from mean) helps estimate probability (p-value, or area under the curve) of value of X being a certain distance from the mean

Z-score Equations for Sample Statistics

Z-Score for Distribution of Sample Means $\bar{x}$

- $z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}}$
- $\sigma_{\bar{x}}$ =

  "(population) standard deviation for the sampling distribution"
  = "standard error of the means"

Z-Score for Distribution of Sample Means $\bar{x}$

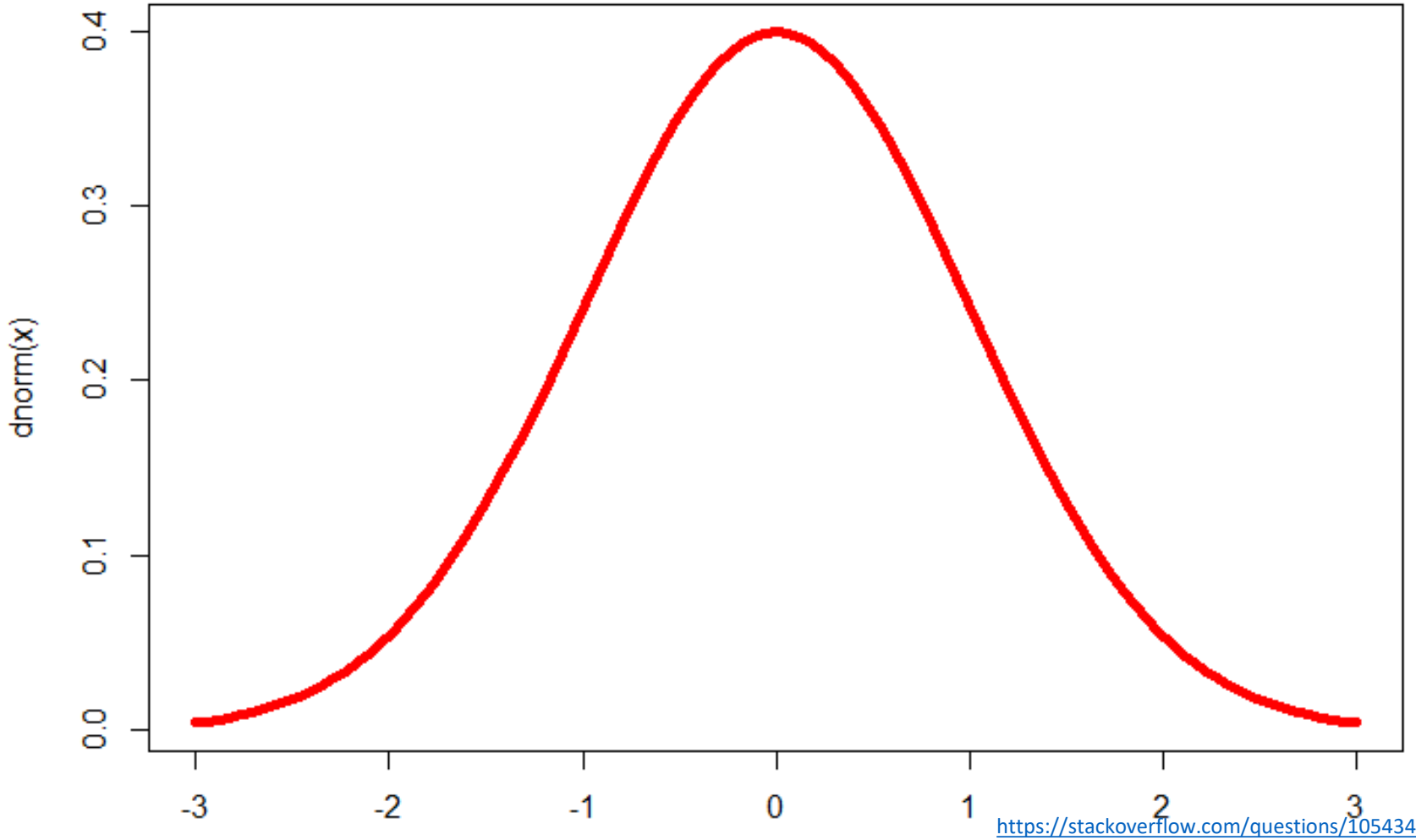
- $z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}}$
- $\sigma_{\bar{x}} = \frac{\sigma}{\mathrm{SQRT(n)}}$, therefore
- $\mathrm{z} = \frac{\bar{x}-\mu}{\frac{\sigma}{\mathrm{SQRT(n)}}}$

What if we don't know the population mean?

Estimating Pop. Mean μ from Samp. Mean $\bar{X}$

- Can estimate μ from Mean($\bar{x}$) thanks to Central Limit Theorem + Law of Large Numbers, *by providing a level of confidence in our estimate*
- Levels of confidence: expressed as percentages (ex, 90%, 95%, 99%)
- α = "alpha" = Significance Level = the likelihood that the true population parameter lies outside the confidence interval
  - Think of it as the point on the distribution curve outside of which P-value would be below the desired level of confidence
  - Or think of it as 1-(confidence level)…Ex: for 95% confidence, α = 5%
- Critical Value = Z-score or T-statistic of α or α/2, depending on the question you are asking (one-tailed or two-tailed test)

Alpha, Critical Value & P-Value, Graphically



Alpha: point(s) on the curve corresponding to level of confidence

Critical Value: Z or T of alpha

P-value: Area under the curve for C.V.

```
# R code used to plot a sample normal distribution curve
# option 2 -- use the curve function to plot a smooth curve from -3 to 3
curve(dnorm, -3,3, col="red", lwd=5)
```

https://stackoverflow.com/questions/10543443/how-to-draw-a-standard-normal-distribution-in-r/10543555

Commonly-Used Critical Values

- For a 90% confidence level, where α=.10
  - Z(α/2) = Z(0.05, 0.95) = ±1.65
    or ± 1.65 * σ of μ(hypothesized)
- For a 95% confidence level, where α=.05
  - Z(α/2) = Z(0.025, 0.975) = ±1.96
    or ± 1.96 * σ of μ(hypothesized)
- For a 99% confidence level, , where α=.01
  - Z(α/2) = Z(0.005, 0.995) = ±2.58
    or ± 2.58 * σ of μ(hypothesized)

To find Critical Value:

STEP 1: Calculate α/2

STEP 2: Use Bluman Table E (find nearest value) or Excel/R/Calc to get Z-Score

Confidence Intervals and Sample Sizes

- **Confidence Interval (CI):** lower & upper estimate range within which X is likely to occur at a certain confidence level
  - ex: "The 95% CI for μ is (a, b)." =
    "We estimate that a <= μ <= b with a 95% confidence level." =
    "There is a 95% chance that μ falls between a and b.

- **Calculating minimum sample size** n required to get a certain confidence interval level of population mean μ
  - $\boldsymbol{n = (z_{\alpha/2} * \sigma \text{ / E})^2}$, where $\boldsymbol{\sigma}$ = pop. Std. dev and E = margin of error

Conf. Interval Formula (known σ, n>30)

- For a given sample mean $\bar{X}$, a given alpha α, a given sample size *n* and a known population standard deviation sigma σ, the confidence interval CI of population mean μ is:

$$\bar{X} - E < \mu < \bar{X} + E \quad , \quad \text{where } E = Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \quad \textbf{, or}$$

$$\bar{X} - \left(Z_{\alpha/2} * \left(\frac{\sigma}{\sqrt{n}}\right)\right) < \mu < \bar{X} + \left(Z_{\alpha/2} * \left(\frac{\sigma}{\sqrt{n}}\right)\right)$$

See Bluman p. 378 for algebraic derivation

Breakouts: Bluman Examples

- 7.1.19
- 7.1.24

[Break]

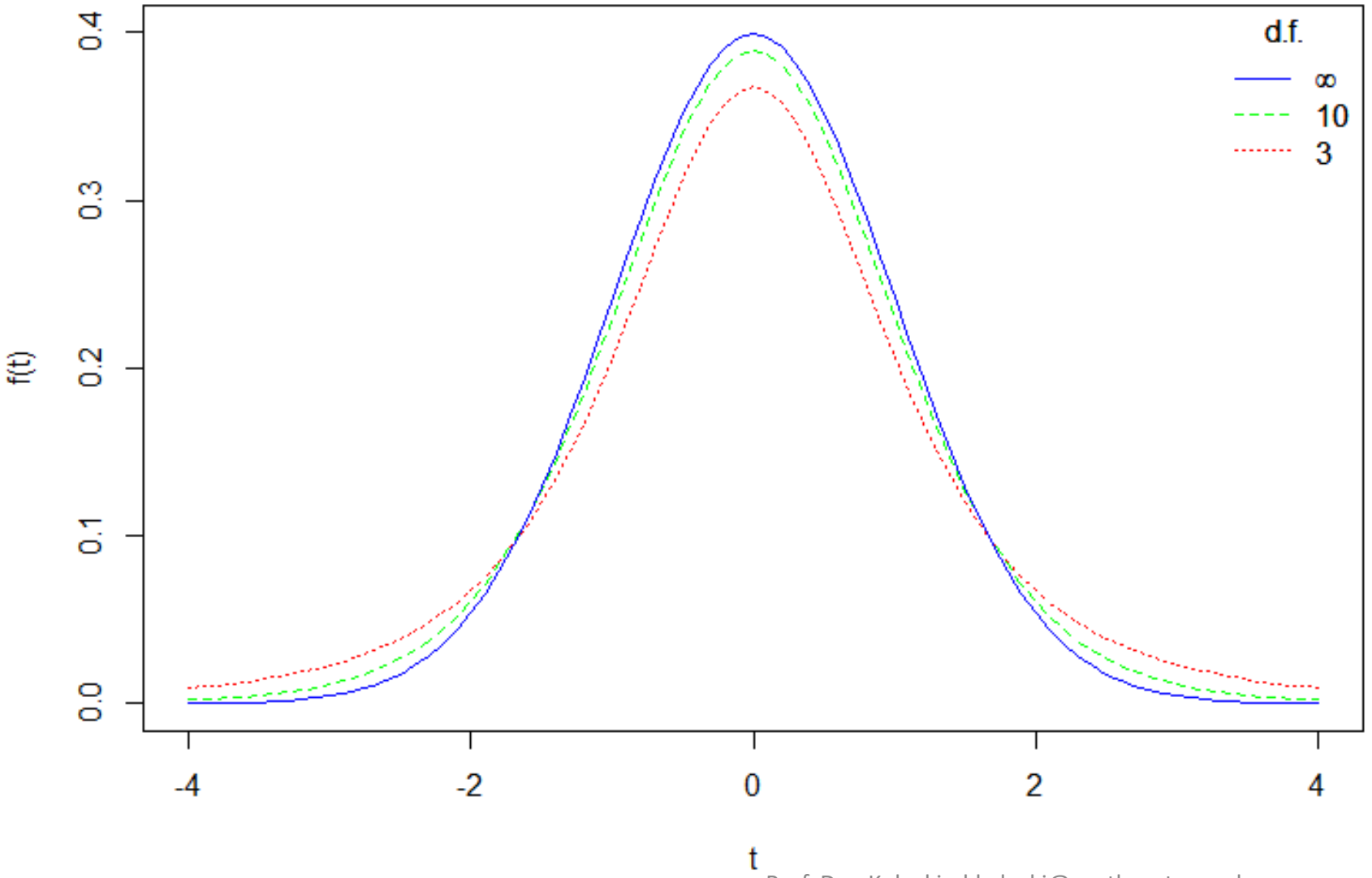
Section 2

Confidence Intervals and Estimating Population Parameters (Part 2)

Estimation when less is known up-front

- Confidence Interval Formula for μ for a given α **when σ is known AND n>30** uses Z score
- **If we don't know σ, OR n< 30** we have a few options
  - Option one (less good) – "swag it": Use sample standard deviation *s* as a proxy for σ and proceed as before (can be risky with Z since *s* varies more)
  - Option 2 (better!) – "use a more variable estimator": replace Z score with T-statistic, which has a wider amount of variability built-in
- T-Distribution (a.k.a. "Student T-Distribution")
  - Like norm. dist in that it is bell-shaped, symmetric with mean = 0
  - Different than norm dist. in that it has more variability and is actually an entire family of curves based on degrees of freedom; as n gets larger, it approaches norm. dist.
  - Degrees of freedom (d.f.) = "how many variables could vary in a given sample size that has a certain mean" = one less than the sample size = **d.f. = n-1**

T-Distribution, Graphically



Critical Value: T of alpha for a certain number of degrees of freedom

P-value: Area under the curve for C.V.

```
# R code used to plot a sample t-distribution family of curves
#https://www.dummies.com/education/math/statistics/plotting-t-base-r-graphics/ for this example
# add infinity symbol to the fonts used in legends
#install.packages("grDevices")
library(grDevices)

# set some sample t-values from -4 to 4, stepping by 0.1
t.values <- seq(-4,4,.1)

#plot a curve for the t-values against their frequencies using dt() function for 3 degrees of freedom
plot(x = t.values,y = dt(t.values,3), type = "l", lty = "dotted", ylim = c(0,.4), xlab = "t", ylab = "f(t)", col="red")

#plot a curve for the t-values against their frequencies using dt() function for 10 degrees of freedom
lines(t.values,dt(t.values,10),lty = "dashed", col="green")

#plot a curve for the t-values against their frequencies using dnorm() function for ifinity degrees of freedom
lines(t.values,dnorm(t.values),lty = "solid", col="blue")

# add legend
legend("topright",
       title = "d.f.",
       legend = c(expression(infinity),"10","3"),
       lty = c("solid","dashed","dotted"),
       text,col = c("blue","green","red"),
       bty = "n")
```

Calculating T-statistics

- Collect or calculate source data
  - Confidence Interval CI
  - α/2 = (1-CI)/2
  - Degrees of Freedom d.f. = (*n*-1)
- Look up value in table F in Bluman
- In this example:
  - CI = 95%
  - Alpha/2 = .025
  - d.f. = 21 (means *n* was 22)
  - T=2.080

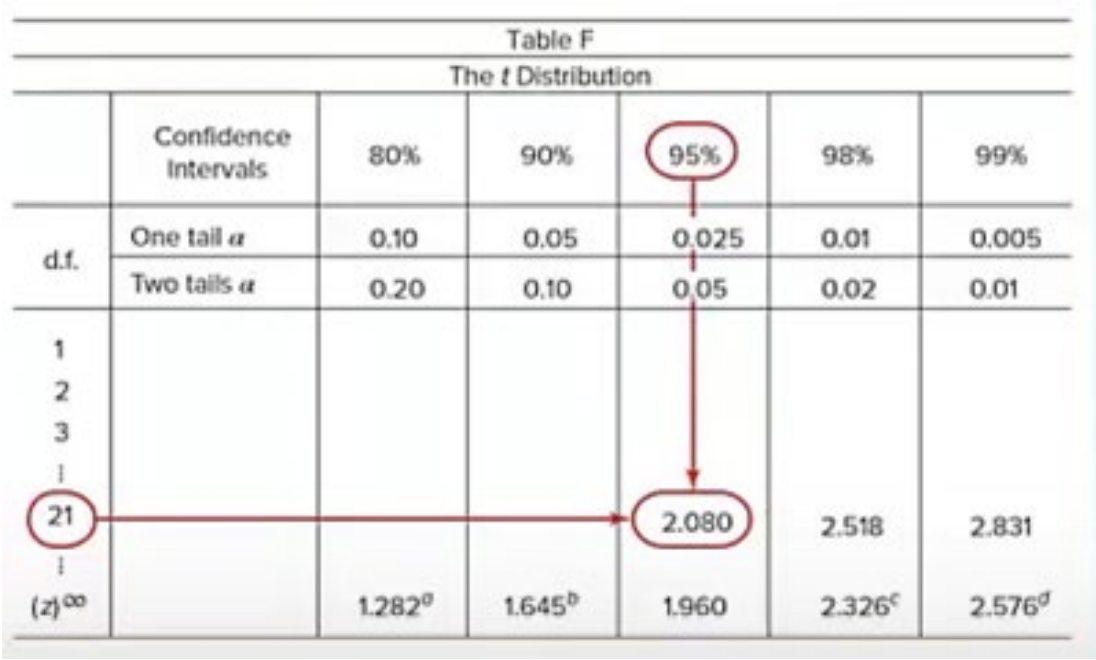


Table F

The *t* Distribution

d.f.	Confidence Intervals	80%	90%	95%	98%	99%
	One tail α	0.10	0.05	0.025	0.01	0.005
	Two tails α	0.20	0.10	0.05	0.02	0.01
1						
2						
3						
⋮						
21				2.080	2.518	2.831
⋮						
$(z)^{\infty}$		1.282[a]	1.645[b]	1.960	2.326[c]	2.576[d]

Helpful Functions in Excel and R: T-Dist.

IN EXCEL

```
=T.DIST(x, deg_freedom, cumulative)
(note we will cover =T.DIST.RT and =T.DIST.2T
later)
```

T-Statistic -> Cumulative P-Value

- **X**: Required. The numeric value at which to evaluate the distribution
- **Deg_freedom:** Required. An integer indicating the number of degrees of freedom.
- **Cumulative:** Required. A logical value that determines the form of the function. If cumulative is TRUE, T.DIST returns the cumulative distribution function; if FALSE, it returns the probability density function

```
=T.INV(probability,deg_freedom)
=T.INV.2t(probability,deg_freedom)
```

Cumulative P-Value -> T-statistic

- **probability**: Required. The P-value at which to evaluate the distribution
- **Deg_freedom:** Required. An integer indicating the number of degrees of freedom.

IN R

```
pt(q, df, lower.tail = TRUE)
```

T-Statistic -> Cumulative P-Value

- **q:** Vector of T-scores (a.k.a. “quantiles”)
- **df:** degrees of freedom (n-1).
- **lower.tail:** If TRUE, probabilities are P[X ≤ x] (cumulative P-value), otherwise, P[X > x] (1-cum. P-val). Default is TRUE.

```
qt(p, df, lower.tail = TRUE)
```

Cumulative P-Value -> T-statistic

- **p:** Vector of P-values . Rest of variables are same as `pt`

Conf. Interval Formula (unknown σ, n<30)

- For a given sample mean $\bar{X}$, a given alpha α, a given sample size *n* and a given sample standard deviation s, the confidence interval CI of population mean μ is:

$$\bar{X} - E < \mu < \bar{X} + E \quad , \quad \text{where } E = t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \quad \textbf{, or}$$

$$\bar{X} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) < \mu < \bar{X} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

Solving T-Dist. Problems for Conf. Intervals

STEP 1:
State the question & data you have

- Ex. question: "Given a certain sample mean and standard deviation, estimate the 95% confidence interval for the population mean $\boldsymbol{\mu}$. "
- Ex. data: α, $\bar{x}$, *s*, *n, d.f.* (maybe *X*'s)

STEP 2: DRAW the curve & area

- This is optional since T-distribution is symmetrical, but it may help you figure out which Ts to use and will be done when we do hypothesis testing

STEP 3: Calculate relevant T-statistics

- Use Table F or Excel/R to find the relevant T-statistics for the relevant $\alpha/2$ and degrees of freedom

STEP 4: Use the formula to calculate and state the lower and upper values of the range

- Lower < $\boldsymbol{\mu}$ < Higher

Conf. Intervals for Population Proportions

- Used when estimating proportion parameter of population (ex., 12% of boats are named Serenity)
- $\hat{p}$ = "p hat" = X/n is the sample proportion
  - X= # of sample units with the characteristic
  - n = sample size
- $\hat{q}$ = "q hat" = $1-\hat{p}$ = (n−X) /n
- As with Z scores from last week, both $n\hat{p}$ * and $n\hat{q}$ must be >=5 for this to work

$$\hat{p} - \boldsymbol{E} < \quad \boldsymbol{P} \quad < \hat{p} + \boldsymbol{E} \ , \quad \text{where } \boldsymbol{E} = \boldsymbol{Z_{\alpha/2}} \left(\sqrt{\frac{\hat{p}\hat{q}}{\boldsymbol{n}}} \right), \text{ or}$$

$$\hat{p} - \boldsymbol{Z_{\frac{\alpha}{2}}} \left(\sqrt{\frac{\hat{p}\hat{q}}{\boldsymbol{n}}} \right) < \quad \boldsymbol{P} \quad < \quad \hat{p} + \boldsymbol{Z_{\alpha/2}} \left(\sqrt{\frac{\hat{p}\hat{q}}{\boldsymbol{n}}} \right)$$

Breakouts: Bluman Examples

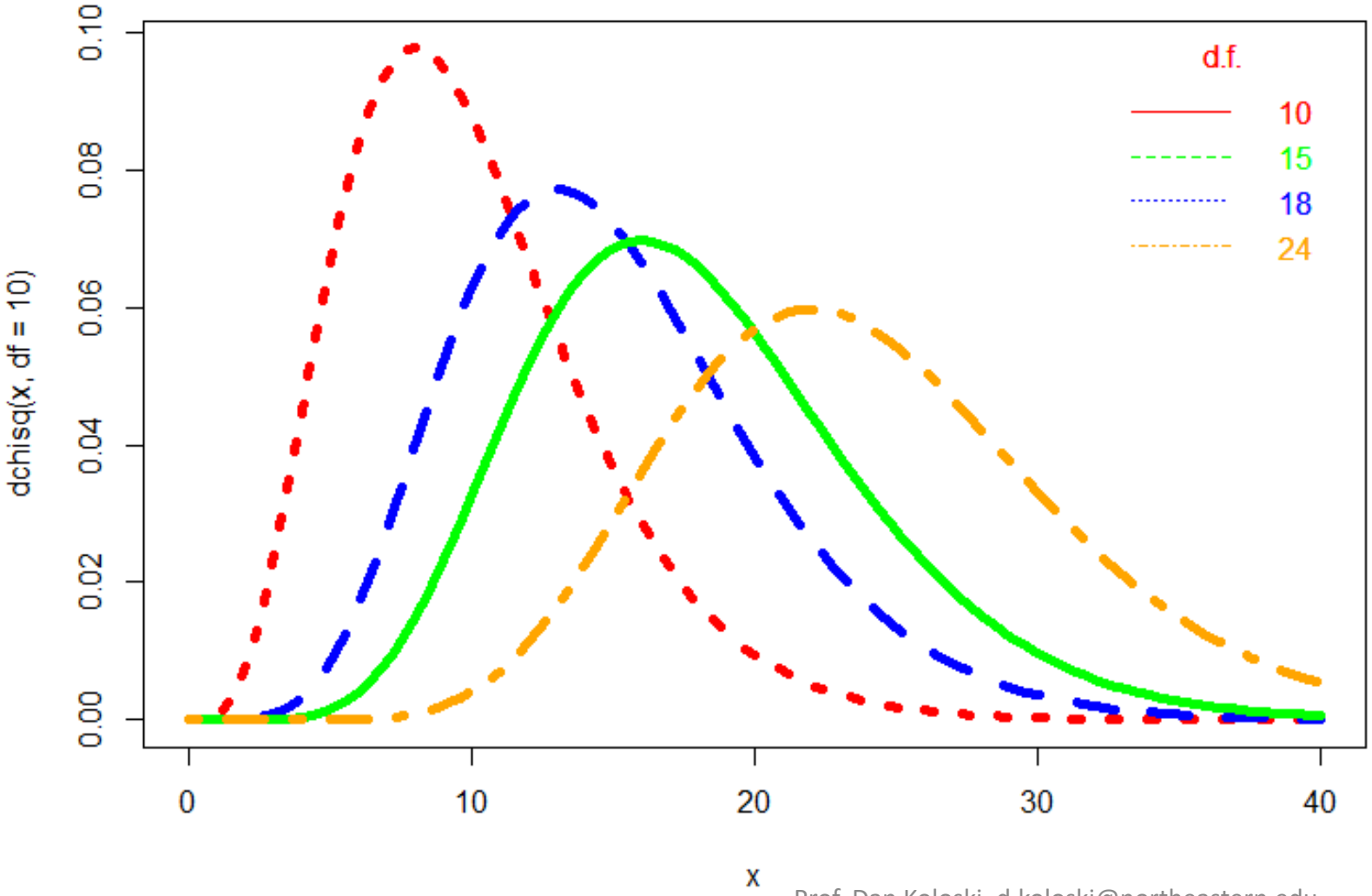
- 7.2.11
- 7.3.12

Conf. Intervals for Population $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma^2}$

- For estimating population σ and σ^2 given sample mean $\bar{X}$ and sample standard deviation s (or sample variance s^2), we must use yet another family of distribution curves (not Std. Norm. Dist. or T-Dist.)

- Chi-square (χ2) family of curves is different in several ways
  - All χ2 values are > 0
  - Area under the chi-square curve = 1
  - All χ2 curves are positively/right skewed (NOT symmetrical)
  - Family of curves are based on degrees of freedom (like T-distribution)

Chi-Square (χ2), Graphically



Critical Values: Right/Upper and Left(Lower) χ2 values – different since curve is not symmetrical!

P-value: Area under the curve between C.V.'s

```
###### R-code to plot a chi-square distribution ########
curve(dchisq(x, df = 10), from = 0, to = 40, col="red", lwd=5, lty = "dotted")
curve(dchisq(x, df = 15), from = 0, to = 40, col="blue", lwd=5, add=TRUE, lty = "dashed")
curve(dchisq(x, df = 18), from = 0, to = 40, col="green", lwd=5, add=TRUE, lty = "solid")
curve(dchisq(x, df = 24), from = 0, to = 40, col="orange", lwd=5, add=TRUE, lty = "dotdash")
legend("topright",
       title = "d.f.",
       legend = c("10","15","18","24"),
       lty = c("solid","dashed","dotted", "dotdash"),
       text.col = c("red","green","blue","orange"),
       col = c("red","green","blue","orange"),
       bty = "n")
```

Conf. Intervals for Population $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}^2$

- For a given sample standard deviation s (or sample variance s^2), a given alpha α, a given sample size *n (and d.f. n-1)*, the confidence interval CI of Population Variance Standard $\boldsymbol{\sigma}^2$ is:

$$\frac{(\boldsymbol{n-1})\boldsymbol{s^2}}{\chi 2_{right}} \quad < \quad \boldsymbol{\sigma}^2 \quad < \quad \frac{(\boldsymbol{n-1})\boldsymbol{s^2}}{\chi 2_{left}}$$

- For a given sample standard deviation s (or sample variance s^2), a given alpha α, a given sample size *n (and d.f. n-1)*, the confidence interval CI of population Standard Deviation $\boldsymbol{\sigma}$ is:

$$\sqrt{\frac{(\boldsymbol{n-1})\boldsymbol{s^2}}{\chi 2_{right}}} \quad < \quad \boldsymbol{\sigma} \quad < \quad \sqrt{\frac{(\boldsymbol{n-1})\boldsymbol{s^2}}{\chi 2_{left}}}$$

Solving χ2 Problems for Conf. Intervals

STEP 1:
State the question & data you have

- Ex. question: "Given a certain sample standard deviation, estimate the 99% confidence interval for the population standard deviation. "
- Ex. data: α, $\bar{x}$, *s*, *n, d.f.* (maybe *X*'s)

STEP 2: DRAW the curve & area

- Best practice since chi-square curves are not symmetrical

STEP 3: Calculate relevant χ2

- Use Table G or Excel/R to find the relevant χ2 for the relevant $\alpha/2$ and degrees of freedom

STEP 4: Use the formula to calculate and state the CI based on the lower and upper values of the range

- Lower < $\boldsymbol{\sigma}$ < Higher

or

- Lower < $\boldsymbol{\sigma^2}$ < Higher

Bluman Examples

- 7.4.9

[Break]

Section 3

Final Project Overview and Helpful new R Techniques

Final Project & Milestone Assignments

Bi-weekly - (all due by 11:59 PM EST Sunday)

- ~~**Data selection decision** - Due this **Sunday 5/29**~~
- Milestone 1 - Exploratory Data Analysis - Due **Sunday 6/5**
- Milestone 2 - Basic Hypothesis Testing - Due **Sunday 6/19**
- Final Project - Multi-variable Relationships & Report - Due **Friday 7/1**

Purpose - End-to-end analysis including data cleanup, exploration, hypothesis testing, presentation to stakeholders

Data source -

1. Use provided data - Study of Women's Health Across the Nation (SWAN): Baseline Dataset, [United States], 1996-1997 (ICPSR 28762)
    - Do not download the original. I have cleaned this data and posted information about it HERE.

Let's talk about Final Project M1 (1 of 2)

- Using the dataset provided by your instructor ~~or a dataset that you have personally obtained (*if you obtain your own dataset, you must have it approved by your instructor; the dataset must have at least 6,000 rows of data*),~~ complete an **exploratory data analysis.**
- **Specifically, pick a subset of the SWAN data that includes at least columns representing AT LEAST** 3 categorical variables and 2 continuous or discrete variables. More is fine. You will reuse this throughout the 3 assignments.

This means:

- getting a sense of the data
- creating visualizations to represent the data; include histograms, boxplots, scatterplots etc. as needed
- obtain descriptive statistic of the data
- finding subsets of data and getting descriptive statistics for each subset
- create visualizations for subset data

Let’s talk about Final Project M1 (2 of 2)

In your report be sure to:

- Describe your dataset.
- What is the purpose of the dataset? What is your data source?
- What kind of data is included? Is it all text data, is it numerical?
- Describe the data fields including the title, the data type, the data description, etc.
- How many rows of data are there? how many fields?
- How many rows of data are there? how many fields?
- Describe any data cleaning you did
- Provide visualizations of the key data and subset data of interest. This should be done for categorical data, discrete data and continuous data.
- Provide descriptive statistical tables for key data fields of interest.
- Provide analysis above and beyond the graphs and tables. Explain what the tables and visualizations tell you about the data.

Present your Exploratory Data Analysis in a report. Incorporate visualizations and tables into the textual analysis of your report. If appropriate, add an appendix of additional data tables and graphs.

- The report should follow this flow:
- **Introduction:** introduce the data, its purpose, the sources, the reason for choosing the data and what you hope to learn from the data. Incorporate a discussion of the data cleaning methods used.
- **Data Analysis:** This is the body of the report where you provide descriptions of the data, basis statistical measures, graphs, tables and analysis.
- **Summary:** Summarize the report. Identify the key take-aways from your analysis. Describe what you want to explore further about your data. Identify questions you want to answer with the data.
- Your report must be 3-5 pages in length, including all graphs but excluding the appendix.

What to Submit: You must submit 3 files:

- The Exploratory Data Analysis report
- Your dataset (if chosen and not provided to you)
- The R code used to analyze the data
- **This assignment is due at the end of Module 2**

In-Class Exercise

- Cool Excel and R techniques for this week's assignment

Wrap Up

This Week's Learning Objectives and Task List

Week 2 Objectives and Task List

Learning Objectives

By the end of this module, you should be able to:

- Distinguish sample distribution vs. sampling distribution
- Connect the Central Limit Theorem and the Law of Large Numbers to sample statistics
- Understand the role of estimator and standard error
- Calculate a confidence interval around mean and proportion
- Determine minimum sample size

Task List

- View lessons in Canvas
- Read *Elementary Statistics*, Chapter 7
- Read *R in Action*, Chapter 6, all Sections and Chapter 7, Sections 1 & 2
- Complete **primary Discussion post by Thursday**
- Complete Practice Problem Set (not submitted)
- Complete **R Practice assignment by Sunday**
- Take quiz
- Complete and submit **Final Project - Milestone 1 by Sunday**

Thank you! See you next week!