Week 1
Introduction & Stats Review

Applied Statistical Analysis/Quantitative Methods I

Jeffrey Ziegler, PhD

Assistant Professor in Political Science & Data Science
Trinity College Dublin

Fall 2024

Road map for today

- About me, and structure of class
- Review of statistics terms
- Quantifying concepts: Types of data
- Making inferences from data
  - Statistic vs. parameter
  - Sampling distribution, C.L.T.
  - Point estimate, confidence interval
- Next time: Hypothesis testing, difference between groups
- By next week, please...
  - Get familiar with problem sets in R, Rstudio, LaTeX, TexStudio

General Info About Course

Instructor	Jeffrey Ziegler, PhD
Email	zieglerj@tcd.ie
In-Person Sessions	09:00 - 11:00 Tuesday
Office Hours	W/Th 15:00-16:00 ?

Data science is social science

- Quantitative social science: Using quantitative data to learn about the social, economic, and political world
    - Industry (Facebook, Google, Netflix, etc.)
    - Nonprofits and governments (Give Directly, OxFam, local services)
- This class will give you the hands-on **tools** and **techniques** of quantitative social scientists a.k.a "data scientists"

Tools: Reasons to learn R, GitHub, & LaTex

R is...

- a tool to perform statistical analysis
- free
- cross-platfrom
- open-source
- can produce high level graphics
- can integrate with document publishing

LaTex is the *word processor* we'll use to format our R code, output clearly

GitHub is how we'll *share our work with each other*

R, GitHub, and LaTex are widely used in academics and industry, put it on your resume!

Techniques: What is Regression Analysis?

- This is your first of two regression analysis courses

Regression analysis examines general relationship between X and Y

$$Y = f(X) + \text{error term}$$

where $f(\cdot)$ may be a linear or non-linear relationship

- In statistics, we use data to fit models that estimate *variability*

Regression Analysis

- A class of statistical regression methods for

  1. Associations: Studying relationships between variables that can be measured

     - Blood pressure and age
     - Income and education (measured in years)

  2. Predictions: Using known values of certain variables to predict outcome for same subjects

     - Given a person's age, cholestorol, and weight, predict blood pressure
     - Given a person's years of education and type of occupation, predict income

Steps in Regression Analysis

1. Research Question
2. Selection of theoretically relevant variables
3. Data collection
4. Model specification
5. Method of fitting
6. Model validation and criticism
7. Model fitting
8. Interpretation

Step 1: Research Question

Formulating research questions is the first and important step in regression analysis! Leads to...

- Selection of relevant variables
- Choice of the statistical methods of analysis

Step 1: Research Question

EX: Suppose we wish to determine whether or not an employer is discriminating against female employees, and we have data from the company on:

- Salary
- Qualifications
- Gender

Potential research questions:

- "On the average, are women paid less than equally qualified men?"
- "On the average, are women more qualified than equally paid men?"

Steps in Regression Analysis

- Step 2: Selection of relevant variables

  Previous research and theory should guide selection of a set of explanatory variables (predictors, X) to explain or predict response variable (Y)

- Step 3: Data collection

  Data collection is from an observational study or an experiment design, determines type of variables and potential inferences we can draw from study

Step 4: Model specification

Classifications of regression analysis

Method	**Response Variable** *(Dependent)*	**Explanatory Variables** *(Independent /predictor)*
Analysis of variance *(ANOVA)*	Continuous response	Categorical predictors
Linear regression	Continuous response	Continuous predictors
Logistic regression	Dichotomous response	Mixture of continuous and categorical predictors
Multionomial regression	Categorial response	Mixture of continuous and categorical predictors
Poisson regression	Discrete response	Mixture of continuous and categorical predictors

Steps in Regression Analysis

- Step 5: Method of fitting

  Most commonly used method of estimation is called least squares

  - We'll briefly also talk about maximum likelihood (MLE)
  - More on this next semester

- Step 6: Model validation and criticism

  We need to determine whether or not required assumptions of model are valid

- Step 7: Model fitting

  Goodness of fit tests

- Step 8: Interpretation

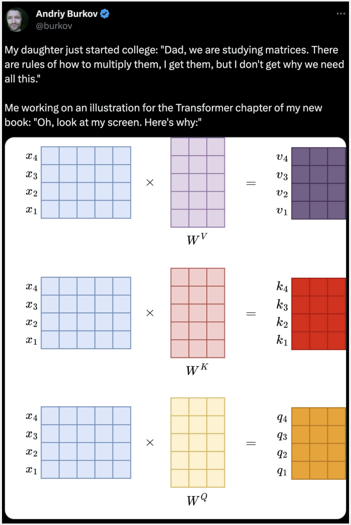
Techniques: What you'll learn

By the end of the course you should be able to:

1. examine and transform data, diagnose influential data and collinearity
2. identify appropriate statistical regression models when presented with new data
3. perform real data analysis with R and interprete statistical analyses

Reasons you may have doubts about this class:

(1) "What does this class have to do with . . . ? "



Andriy Burkov
@burkov

My daughter just started college: "Dad, we are studying matrices. There are rules of how to multiply them, I get them, but I don't get why we need all this."

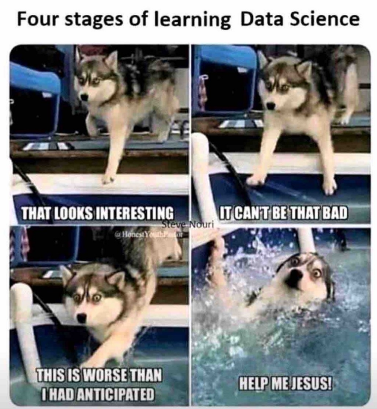
Me working on an illustration for the Transformer chapter of my new book: "Oh, look at my screen. Here's why:"

x_4 x_3 x_2 x_1 × W^V = v_4 v_3 v_2 v_1
x_4 x_3 x_2 x_1 × W^K = k_4 k_3 k_2 k_1
x_4 x_3 x_2 x_1 × W^Q = q_4 q_3 q_2 q_1

- That is an excellent question, please ask it often!
- The short answer is most likely **literacy**
    - To read or produce social science research
    - Examine empirical claims about politics, policy, and the world in general

Reasons you may have doubts about this class:

(2) "I'm not good at math/coding/whatever."



- Everyone feels this way sometimes, and YES YOU ARE!
  - This is not a math class ($+, -, \times, \div, X_{ij}$)
- You can do it!
- Be patient—like learning a foreign language or a new sport
  - Ask questions
  - Practice!
  - **Don't fall behind**

Now, a dramatic reading of the syllabus...

Direct link to syllabus, or find on course website

Required Materials

Texts	All readings are provided, **don’t buy books**!
R and LaTeX	Should be installed
Rstudio and TexStudio	Should be installed

Course Evaluation/Assessment

- Problem Sets ($\bar{x}$ of top 3/4): 50%
- Exams (2): 50%

Problem Sets (50%)

- Typically assigned every other week, and you will generally have two weeks to do the homework (this will vary)
- Problem sets require R and should be written in LaTex (must include figures and code within text)
- All homework will be posted on GitHub
- Evaluated by me
- I will publish correct answers each week, but look at others' GitHubs so we can learn from each other
- The lowest PS grade will be dropped, so "I have been so busy with other classes" is not a legitimate reason to not turn in problem set!

Exams (50%)

Exam 1	In class: October 15
Exam 2	In class: November 26

- All exams are cumulative
- All exams are multiple choice and open-response questions
- All exams are graded by me
- You will be allowed a formula sheet for exams
- Make-up exams are **only** allowed by **written** approval from Students Services

How to do start problem sets: LaTeX and R

Should have:

(1) Downloaded LaTeX

(2) Downloaded word processing interface for LaTeX (TexStudio)

Then:

(1) Download/copy .tex and .R files from class GitHub

(2) Create new files (i.e. "PS1_answers_JZ.tex", "PS1_answers_JZ.R") with template from GitHub

(3) Edit new files and push to GitHub

If you don’t ask, I assume you know how to do this!

Other course policies to consider

- Absences for religious holidays are excused
- Talk to me ASAP if you have any illness or family emergencies
- All students with special accommodations should notify me as soon as possible
  - Documentation from the Trinity Office of Disability Services is required
- The schedule posted on the syllabus is **tentative** and subject to change

Approach Toward Learning

Preparation + synthesis + practice = learning

- Individual preparedness: Reading & slides before class
- In class:
  - Discussion and Q&A on important concepts
- Tutorial:
  - Advanced theoretical problems
  - Group work on problems in R
- Office hours: Review and correct mistakes
- Problem sets: Individual homework assignments

Review: Statistics Terminology

- **Variable**: Characteristic(s) that take on different values for different subjects or experimental units
  - Age, income, partisanship, gender
  - Some variables change more than others (need variation)
- **Response variable** (Outcome variable): What we want to *explain* or *predict*
- **Explanatory variable** (Input variable): Variable used to predict variation in response; can also be called *covariate* or *predictor*

Review: Statistics Terminology

- **Observation**: Single subject/unit, and corresponds to a row in a data matrix or table* such as this...

```
country                         infant.mortality   gdp
---------                       -----------------  ---
Afghanistan                                   154  2848
Albania                                        32   863
Algeria                                        44  1531
American.Samoa                                 11    NA
Andorra                                        NA    NA
Angola                                        124   355
```

(* how you organize your data is very important)

- An observation includes both a response (y_i) and explanatory variable(s) (x_i or $\mathbf{X}_i$)

How to quantify concepts into measures

- *Continuous:* Any value within an interval is possible (income, temperature)
- *Discrete:* Finite (or infinitely countable) number of values are possible (# of children in a family, population)
- *Nominal (Binary/Not Binary):* Color (green, blue, red, etc)
- *Ordinal (Binary/Not Binary):* Height (upper tercile/median tercile/lower tercile)

Numerical summaries for Quantitative variables

1. Measure of center
    - Mean
    - Median
2. Quartiles: (min, Q_1, median, Q_3, max)
3. Measure of variability
    - Standard deviation (variance)
    - Interquartile range
4. Shape of distributions
    - Skewed right (mean $>$ median)
    - Skewed left (mean $<$ median)
    - Symmetric (mean $\approx$ median)
    - Bimodal

How do we estimate an effect?

Start thinking of summary statistics from samples as parameters!

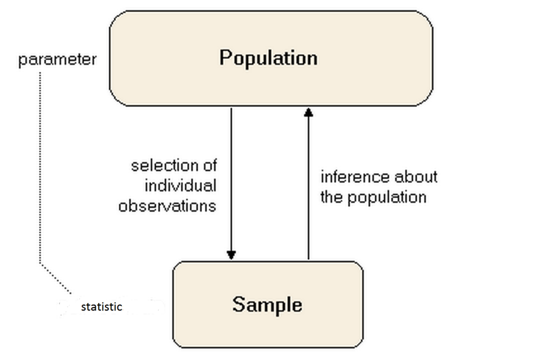
A **parameter** is a numerical summary of the *population*

- We want to make inference from parameters
- The true value of a parameter is unknown
- We denote parameters with Greek letters (i.e., μ, σ, ρ, β)

A **statistic** is a numerical summary of the *sample*

- We calculate statistics from our *sample* data
- **Statistics** are our best estimate of **parameters**
- We denote statistics with lower case letters, bars, and hats (i.e., $\bar{x}$, s, r, $\hat{\beta}$)

What is statistical inference?


parameter
Population
selection of individual observations
inference about the population
statistic
Sample

We make *inferences* about the population based on info in our sample

- Point estimation
- Confidence intervals
- Hypothesis testing (next lecture!)

1^{st} IMPORTANT STATISTIC: SAMPLE MEAN

$$
\begin{aligned}
n &= \text{Sample size} && (1)\\
y_1, y_2, \ldots, y_n &= \text{Observations} && (2)\\
\bar{y} &= \frac{y_1 + y_2 + \ldots + y_n}{n} && (3)\\
&= \frac{1}{n}\sum_{i=1}^{n} y_i && (4)\\
&\equiv \frac{1}{n}\sum y_i \; (\textit{law of large \#s, not C.L.T.}) && (5)
\end{aligned}
$$

Calculation: Standard deviation ($\sqrt{Var}$)

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} \tag{6}$$

$$= \sqrt{\frac{\sum \text{of squared deviations}}{\text{sample size} - 1}} \tag{7}$$

Things to note about Standard deviation

- One standard deviation will be equal to something in original units (e.g., 2 inches, 10 crimes reported per month)
- $S \geq 0$, why?
- $S = 0$ only if y is a constant, why?

Ex: Calculating mean, variance, and SD

y_i	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$
0		
4		
4		
5		
7		
10		

Doing it in R: Find the sum

```r
# create vector y
```

Doing it in R: Find the sum

```r
# create vector y
y <- c(0, 4, 4, 5, 7, 10)
```

Doing it in R: Find the sum

```r
# create vector y
y <- c(0, 4, 4, 5, 7, 10)
# (1) find sum of y
```

Doing it in R: Find the sum

```r
# create vector y
y <- c(0, 4, 4, 5, 7, 10)
# (1) find sum of y
sum(y)
```

$\sum y_i = 30$

Doing it in R: Find the mean

The old fashioned way:

$$\frac{\sum y_i}{n} \tag{8}$$

Doing it in R: Find the mean

```r
# (2) find mean of y
sum(y)/length(y)
```

Doing it in R: Find the mean

```r
# (2) find mean of y
sum(y)/length(y)
# same thing, but faster
mean(y)
```

$\bar{y} = 5$

Doing it in R: Find the sum of errors

Let’s try to construct this ourselves:

$$\sum (y_i - \bar{y})$$

Doing it in R: Find the sum of errors

```r
# (3) find sum of demeaned values
# substract mean from each y_i
# first, I’ll create a vector to fill
# there is nothing in there when I create an object
# like x <- NULL
demeanedSum <- NULL
# remember that [i] is indexing each element of y
# I’m basically saying, for each element in all of the elements in y
# take that y_i and subtract the mean of y
# and store that value in my vector of demeaned values for each y_i
for(i in 1:length(y)){
  demeanedSum[i] <- y[i] - mean(y)
}
```

```
[1] -5 -1 -1 0 2 5
```

Filling in the table with our demeaned values

y_i	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$
0	-5	
4	-1	
4	-1	
5	0	
7	2	
10	5	

Doing it in R: Find the sum of errors

```r
# the simple way (R figures it out)
demeanedSumSimple <- y - mean(y)
# now we take the sum of all those demeaned values
sum(demeanedSumSimple)
```

$$\sum(y_i - \bar{y}) = 0$$

Finishing sum of squared errors in R

```r
# (4) sum of squared error
squaredError <- demeanedSum^2
sum(squaredError)
```

$$\sum(y_i - \bar{y})^2 = 56$$

Finishing our table with the squared errors

y_i	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$
0	-5	25
4	-1	1
4	-1	1
5	0	0
7	2	4
10	5	25

Answer: Calculating variance

$$S^2 = \frac{\sum(y_i - \bar{y})^2}{n-1}$$

Answer: Calculating variance

$$S^2 = \frac{\sum(y_i - \bar{y})^2}{n - 1}$$

$$S^2 = \frac{56}{5} = 11.2$$

Answer: Calculating standard deviation

$$S^2 = \frac{\sum(y_i - \bar{y})^2}{n-1}$$

$$S^2 = \frac{56}{5} = 11.2$$

$$S = \sqrt{11.2} = 3.3$$

WE USE *SAMPLE STATISTICS* TO ESTIMATE *POPULATION PARAMETERS*

Now that we have our *sample statistics*, what do we do?

- Use probability theory and sampling distributions
- Some assumptions (i.e. $\bar{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$)
- This will be our first true statistical inference (it's getting exciting 😄)

Definition: Sampling Distributions

> *A* ***sampling distribution*** *is the distribution of a* ***statistic*** *given repeated sampling*

We use probability theory to derive a **distribution for a statistic**, which allows us (eventually) to make inferences about **population parameters**

Similar to, but different from sampling distributions...

We also need **probability distributions**!

- Binomial distribution: $Bin(n, p)$ (# of trials, Pr(success))
- Normal distribution: $N(\mu, \sigma^2)$ (mean, variance)
- Standard Normal distribution: $N(0, 1)$ (mean=0, variance=1)
- t-distribution: t_{df}
  - Centered at zero
  - Bell shaped like normal
  - Fatter tails than the normal
  - Degrees of freedom (df) define 'fatness'
  - As DF ↑, t distribution gets close to standard normal

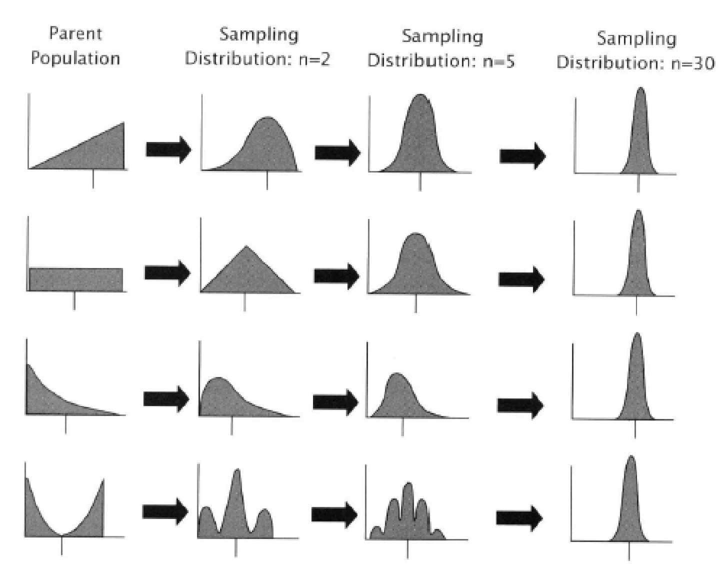
Central limit theorem (not law of large #s)

For random sampling with a **large** sample size n, the sampling distribution of the sample mean $\bar{y}$ is approximately normal, where $\bar{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

Central limit theorem

- $\sigma/\sqrt{n}$ is called the standard error
  - **Standard deviation** of the **sampling distribution**
  - Note: formula includes population standard deviation σ
  - Pay attention or you will get them mixed up!
- As $n \rightarrow \infty$, standard error gets smaller and smaller
- Unbiased distribution of $\bar{y}$
- **This** is why the normal distribution is very important
- Usually $n = 30$ is "good enough", but it will depend on the distribution... (not a stead-fast rule)

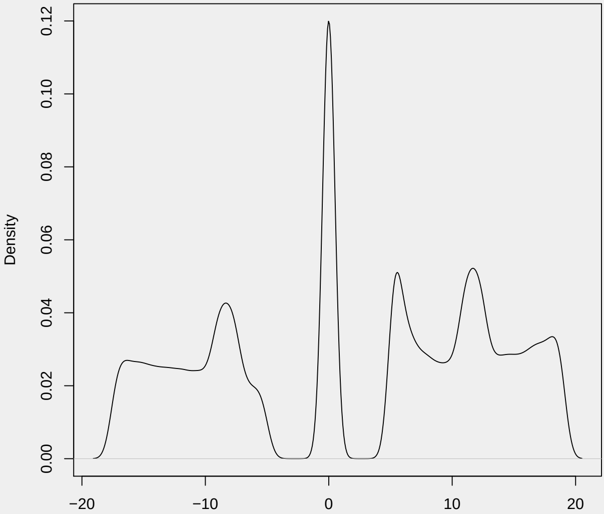
IT WORKS FOR EVERYTHING (FOR RIGHT NOW…)


Parent Population
Sampling Distribution: n=2
Sampling Distribution: n=5
Sampling Distribution: n=30

Let's do our own experiment in R

```r
# we want to create "abnormal" distribution to make our point
# generate an odd transformation of the data
x <- runif(100000, min = -1, max=1)
x <- sqrt(1 + x) + 2*x^3 - 23*x + abs(log(abs(x))) + 2*(x > .5) + -2*(x < -.5)
# add a large mass of data in the middle at zero
# do so by assigning all values between -5 and 5
# to zero
x[x > -5 & x < 5] <- 0
# plot and save as pdf
pdf("zeroCenteredDist.pdf")
plot(density(x, bw = .5), main="", xlab = "")
dev.off()
```

Our distribution centered at zero



Create sampling distribution of $\bar{x}$

Step 1: Create "random sample" parameters to estimate

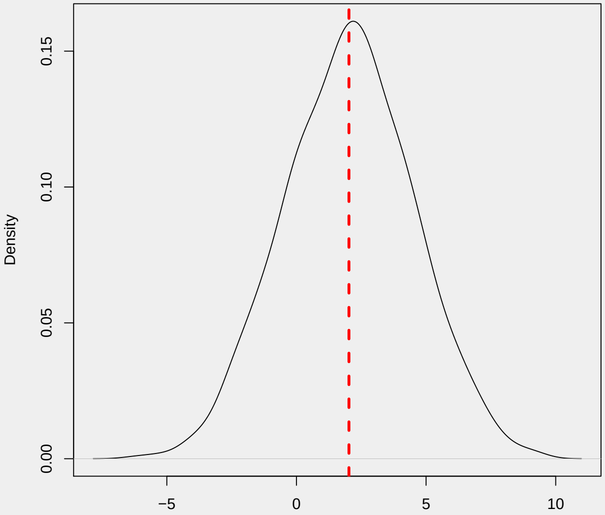
```r
# don't forget to set your seed so that you can reproduce your example!
set.seed(123)
# we'll take 1000 random samples
n.samples <- 1000
# with sample size = 20
sample.size <- 20
# use a "for loop"
# (1) make an empty vector to store all of our sample statistics
# fill the empty vector with NAs (missing data)
# remember, make it length 1000
x.bars <- rep(NA, n.samples)
```

Create sampling distribution of $\bar{x}$

Step 2: Generate "random sample" and estimate value

```r
# (2) "loop" over the vector 1, 2, ..., 1000
# in each iteration the variable "i" will increment up to the next value
for(i in 1:n.samples){
  # draw a random sample
  this.sample <- sample(x, size = sample.size, replace=F)
  # calculate mean and add it to vector of NAs
  x.bars[i] <- mean(this.sample)
}
# plot and save sampling distribution of mean of Xs
pdf("exampleSamplingDist.pdf")
plot(density(x.bars),main="", xlab = "")
abline(v = mean(x.bars), col="red", lwd=3, lty=2)
dev.off()
```

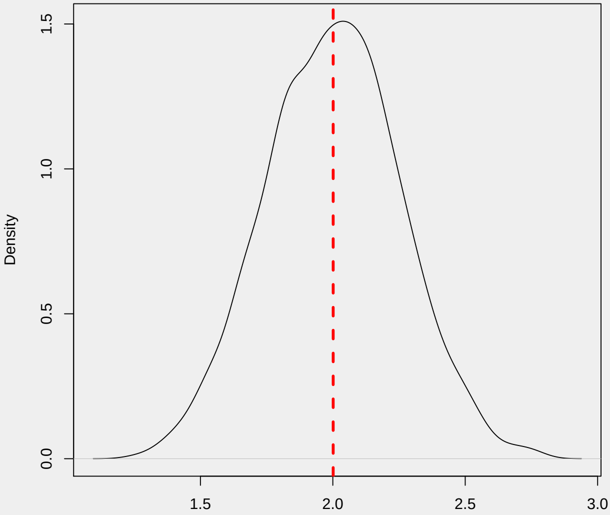
Sampling distribution of $\bar{x}$ with $n = 20$



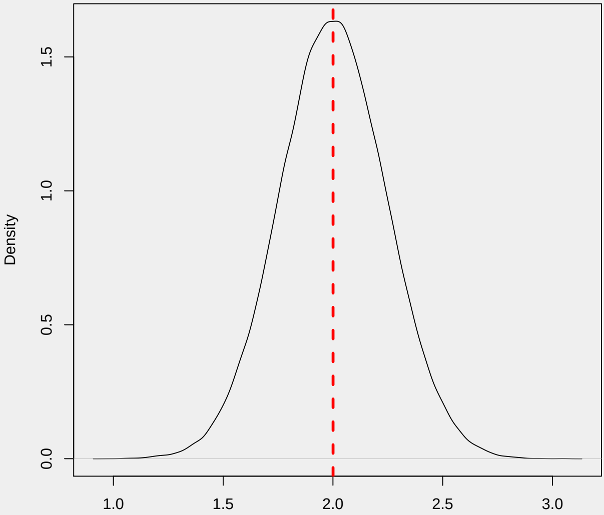
Now with a larger sample size

```r
# update the sample size to 2000 instead of 20
sample.size <- 2000
# create new vector of x bars (still length 100)
new.x.bars <- rep(NA, n.samples)
# "loop" again (create random sample of 2000
# and get the mean, repeat 1000x
for(i in 1:sample.size){
  this.sample <- sample(x, size = sample.size, replace=F)
  new.x.bars[i] <- mean(this.sample)
}
```

Sampling distribution of $\bar{x}$ with $n = 2000$



Sampling distribution of $\bar{x}$ with $n = 100,000$



Sampling Distribution of $\bar{y}$

- How is $\bar{y} = \frac{1}{n} \sum y_i$ distributed?
- The mean of the distribution is known to be μ (population mean)
- What about the spread?

Standard error

The standard deviation of the sampling distribution of $\bar{y}$, denoted $\sigma_{\bar{y}}$, is called the standard error of $\bar{y}$, and is equal to $\frac{\sigma}{\sqrt{n}}$

Under certain circumstances we can safely assume that $\bar{y} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

The story so far ...

- We are hunting for population parameters $[\mu, \sigma]$
  - What percentage of Americans approve of President Biden?
  - How much carbon is emitted from cars?
  - What is the wage gap for women in Ireland?
- We sample population & calculate sample statistics $[\bar{y}, s]$

Estimation basics: Point estimate

*A **point estimate** is a sample statistic that gives a good guess about a population parameter*

- **Remember**: Point estimation for population mean ($\hat{\mu}$)
  - $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- **Remember**: Point estimate for population standard deviation ($\hat{\sigma}$)
  - $S = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}$

Estimation basics: Estimators

- How do we choose among possible estimators?
- We want our estimators to be:
  - Unbiased (i.e., accurate): $E(\hat{\mu}) = \mu$ with repeated sampling
  - Efficient (i.e, precise): $\sigma_{\hat{\mu}}$ is small(er)
  - Consistent: As $n \to \infty$ then $E(\hat{\mu}) \to \mu$

Estimation basics: Summary

The point estimates for populations parameters μ and σ are:

- Denoted as $\hat{\mu}$ and $\hat{\sigma}$
- “Best” estimated by $\bar{y}$ and S
- They are “best” in terms of bias and efficiency

Note: We are **not assuming** that the population is normal

- We are just assuming that our real goal is to find a good estimate of μ and that n is large

Some terms: Confidence intervals

A point estimate is OK, but it is not very useful without knowing how much confidence to have it

Solution – *interval estimation*!

A CONFIDENCE INTERVAL

... for a population parameter, a CI is a range of numbers within which a parameter is believed to fall

The ***confidence coefficient*** *is the probability that an interval would contain the parameter with repeated sampling*

- 0.95 → 95% confidence interval
- 0.70 → 70% confidence interval

Confidence interval for population means

Assuming a large sample...

We can use the sampling distribution of $\bar{y}$ to calculate a confidence interval for the population mean

- Parameter: μ
- Estimator: $\hat{\mu} =$ assuming that $\bar{y} \sim N(\mu_{\bar{y}}, \sigma_{\bar{y}})$
- Remember that $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$ and $\hat{\sigma}_{\bar{y}} = \frac{S}{\sqrt{n}}$

Basic idea:

- We plug in the **estimated** value of σ to get $\hat{\sigma}_{\bar{y}}$
- We use $\bar{y}$ to **estimate** μ, which we'll denote as $\hat{\mu}$
- Now we have an **estimated** sampling distribution, $N(\bar{y}, \hat{\sigma}_{\bar{y}})$
  - We use our knowledge of the normal distribution to find a *confidence interval*
  - E.g., we want 2.5% of the probability to be outside of our interval on each side

Steps for calculating CIs:

1. Calculate $\bar{y}$

2. Calculate S and then $\hat{\sigma}_{\bar{y}} = \dfrac{S}{\sqrt{n}}$

3. How much area do we need under the curve to the right?
   - (1-Confidence Coefficient)/2

   How much area do we need under the curve to the left?
   - (Confidence Coefficient)/2

4. Find the z-score associated with that number

5. Use these values to calculate $\bar{y} \pm Z \times \hat{\sigma}_{\bar{y}}$

Let’s calculate some CIs in R

```r
## Let's say our confidence coefficient = .95
## Calculate the appropriate confidence interval for the
## mean level of support for how George W. Bush was
## handling the war in Iraq (bushIraq)
#using qnorm since n>30
z95 <- qnorm((1 - .95)/2, lower.tail = FALSE)## (1-
    confidence coefficient)/2
n <- length(na.omit(anes$bushiraq))
sample_mean <- mean(anes$bushiraq, na.rm = TRUE)
sample_sd <- sd(anes$bushiraq, na.rm = TRUE)
lower_95 <- sample_mean - (z95 * (sample_sd/sqrt(n)))
upper_95 <- sample_mean + (z95 * (sample_sd/sqrt(n)))
confint95 <- c(lower_95, upper_95)
```

Results = [0.358, 0.443]

What does this mean?

Class business

- Respond to the pre-term survey
- Read the required and suggested online materials
- Should already: install R, set up your `LaTex`, download `GitHub` desktop
- Problem set # 1 is up on GitHub
- These slides are available on the course website
- Next time, we'll pull it all together to make our first true inference (hypothesis testing!)