




# Week 2 Review Session

GSI team  
July 13th, 2023



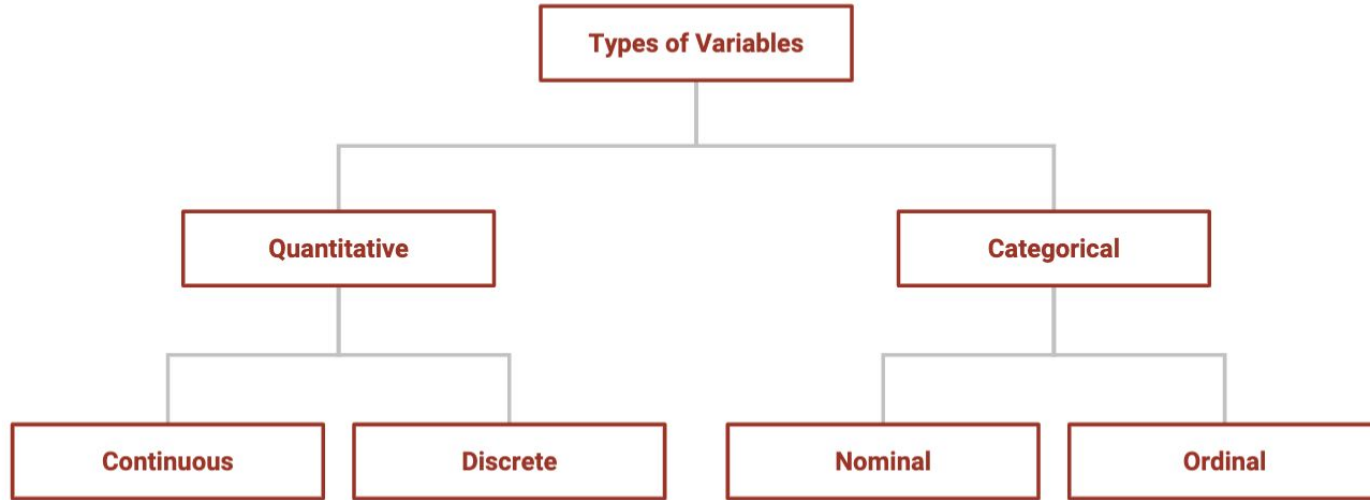
# Announcements

- Midterm #1 begins FRIDAY July 14<sup>th</sup>, 2023
  - Both the timed gradescope portion and the take-home portion are due on Saturday July 15<sup>th</sup> at 12pm (noon).
  - Exam is open notes, but no use of the internet or collaboration with other students is allowed.
  - Public questions on Ed will be disabled starting 10pm Thursday
  - Lab will be held on Friday just for project questions
- Quiz 6 due July 13<sup>th</sup>, 10 pm
- Data Project Part 1 due July 17<sup>th</sup>, 10 pm (Monday)
  - Please meet with your assigned GSI before you submit (required for participation points!)

# Key Points

- PPDAC
  - What does this stand for?
- Types of Problems
  - What is the difference between descriptive, causal, and predictive problem types?
- Describing Data
  - What kind of observations do we have? (e.g. Individual people, lab tests, etc.)
- What kind of variables do we have?
  - Nominal, Ordinal, Continuous, Discrete?

# Types of Variables



Can you give an example for each of these variable types?

# Visualizing and Describing Data

- Bar Charts vs. Histograms
  - What does the height of a bar represent? Frequency (count)
  - What does the x axis represent? Levels of the variable
- What measures of central tendency did we learn?
- What measures of spread did we learn?

# Continuous vs. Categorical Data Visualization

For two **continuous** variables?

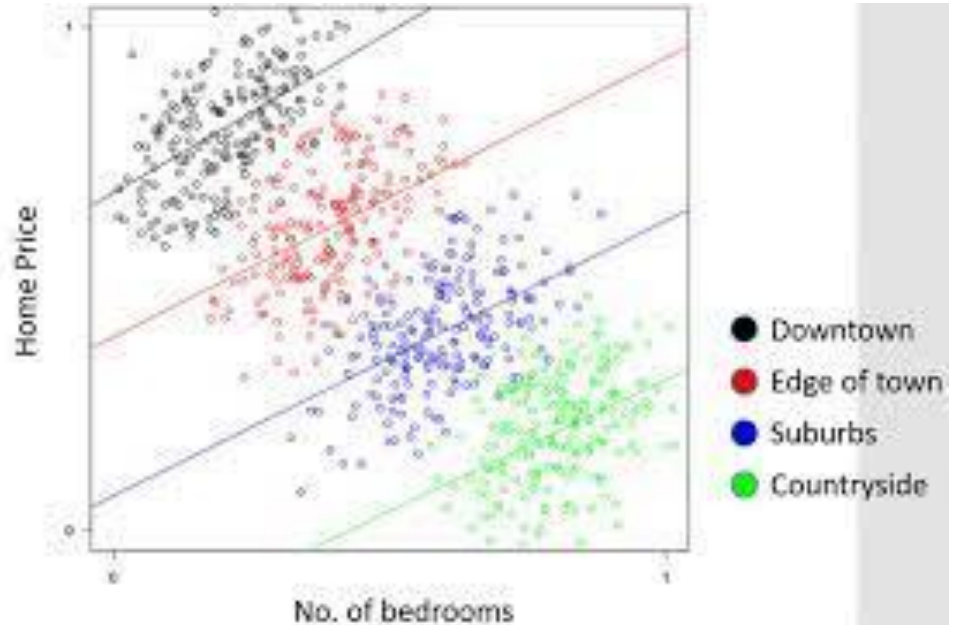
- Scatterplots, correlation coefficients, linear regression

For two **categorical** variables?

- Bar charts, tables, conditional v. marginal distributions, think about Simpson's Paradox

# Simpson's Paradox

Statistical phenomenon where an association between two variables in a population emerges, disappears or reverses when the population is divided into subpopulations



# Visual Summaries in R

- **ggplot** - `new_dataset <- dataset %>% ggplot(aes(x= variable, y=variable))`
- **Histograms** - `geom_histogram`;
  - `ggplot(data = dataset, aes(x = variable)) +  
 geom_histogram(col = "white", binwidth = 10) +  
 labs(x= "", y = "") +  
 theme_minimal(base_size = 15)`
- **Bar plots** - `geom_bar(stat='count')` or `geom_bar(stat='identity')`;
  - `ggplot(dataset, aes( x= variable, y=variable)) +  
 geom_bar(stat = "identity") +  
 labs(y = "", x = "") +  
 theme_minimal(base_size = 15)`
    - Stat = "identity" tells `geom_bar`/histogram that we supplied a y variable this is exactly what we want to plot and we do not need `geom_bar` to calculate the number or percent for us
    - Stat = count will tell `ggplot` to calculate percentages
- **Dodged bar plots** - `geom_bar(aes(x=var, fill=var), stat='identity', position='dodge')`
  - Position = "dodge" → puts bars next to each other
  - Position = "stack" → stacks variables on top of each other in one bar
  - `aes(fill = var)` links the bar's color fill to the x variable
  - `Fct_relevel` = allows you to line up the bars in ascending order



# Quantitative Summaries and IQR

- **Quantitative Summaries**

- **dplyr** - `dataset <- dataset %>% dplyr_function(argument_1 = , argument_2 = )`
- **mutate** - `mutate(new_column = 500 * existing_column)`
- **summarize** - `summarize(summarize_stat_name = 500 * existing_column)`
  - Note: see next slide for more on DPLYR library

- **Interquartile Range (IQR) Q3 - Q1**

- Resistant to outliers
- Quantiles using R
  - `Dataset %>% summarize(Q1 = quantile(var, 0.25) , median = median(var) , Q3 = quantile(var, 0.75))`
  - \*To get exact answer as you would by hand = `quantile(data, 0.25, type = 2)`

# dplyr library

## Functions to know

- Filter
- Mutate
- Summarize / group\_by
- Arrange
- Select

Managing large data frames is hard – dplyr can help! With dplyr, you can organize your data to match your needs! Here are some useful functions to help you transform your data frame from scattered to focused.

### filter

This function allows you to “filter” through your data frame and select observations that meet your set criteria through logical statements (`==`/`>`/`<`/`!=`).

```
filtered_data <- filter(data file, condition)
filtered_data <- filter(pet, state == "CA")
```

Pets owned in USA			
State	Cat	Dog	
CA	0	1	
AZ	0	1	
CA	1	1	
DE	2	0	



Pets owned in USA			
State	Cat	Dog	
CA	0	1	
CA	1	1	

### mutate

This function allows you to change variables or append columns to a data frame. This can be useful to perform new calculations on existing data.

```
mutated_data <- filter(data file, condition)
filtered_data <- mutate(pet, Total_Pets = Cat + Dog)
```

Pets owned in USA			
State	Cat	Dog	
CA	0	1	
AZ	0	1	
CA	1	1	
DE	2	0	



Pets owned in USA				
State	Cat	Dog	Total Pet	
CA	0	1		1
AZ	0	1		1
CA	1	1		2
DE	2	0		2

### summarise/group\_by

This function allows you to complete single number output calculations on your entire dataset (i.e. means, total sums, etc.). You can use the `group_by` function to arrange summary functions by category. The `group_by` function does not change the data frame.

```
summary_data <- summarise(data file, condition)
summary_data <- summarise(pet, mean_cats =mean(Cat), mean_dogs=mean(Dog))
group_data <- summarise(data file, var)
group_data <- group_by(pet, State)
summarise(group_data, mean_cats =mean(Cat),
```

Pets owned in USA			
State	Cat	Dog	
CA	0	1	
AZ	0	1	
CA	1	1	
DE	2	0	



mean_cats	mean_dogs
<dbl>	<dbl>
0.75	0.75
1 row	



State	mean_cats	mean_dogs
<chr>	<dbl>	<dbl>
AZ	0.0	1
CA	0.5	1
DE	2.0	0

### arrange

This function allows you to arrange data numerically or alphabetically by variable.

```
arrange_data <- arrange(data file, var)
arrange_data <- arrange(pet, State)
```

Pets owned in USA			
State	Cat	Dog	
CA	0	1	
AZ	0	1	
CA	1	1	
DE	2	0	



Pets owned in USA			
State	Cat	Dog	
AZ	0	1	
CA	0	1	
CA	1	1	
DE	2	0	

### select

This function allows you to select variables to include or exclude in your data frame. Use the `“.”` sign to exclude.

```
select_data <- select(data file, var)
select_data <- select(pet, -Dog)
```

Pets owned in USA			
State	Cat	Dog	
CA	0	1	
AZ	0	1	
CA	1	1	
DE	2	0	



Pets owned in USA		
State	Cat	
AZ	0	
CA	0	
CA	1	
DE	2	

**Useful Tip:** The dplyr library includes a `%>%` (pronounced “pipe”) operator function. You can use it to “chain” the functions above to succinctly complete more than one function on one dataset at the same time. This makes it easy to add steps and minimize local variables stored in the environment.

# Variance, SD, and 5 Number Summary

## Sample Variance

- $s^2$  represents sample variance
- $s^2 = 1/(n-1) \sum (X_i - \bar{X})^2$
- Use `summarize()` dataset %>% `summarize(name_var = var(variable))`

## Standard Deviation

- $s$  represents standard deviation
- Square root of variance
- Use `summarize()` dataset %>% `summarize(name_sd = sd(variable))`

## Five Number Summary = Min, Q1, median, Q3, max

- To visualize, use a box plot
  - Center line = median
  - Top box = Q3
  - Bottom box = Q1
  - Top whisker = Max or  $Q3 + 1.5 \cdot IQR$
  - Bottom whisker = Min or  $Q1 - 1.5 \cdot IQR$
- Boxplots in R = Use `ggplot()`'s `geom_boxplot()`
- Use `dplyr`'s `summarize()` function to calculate 5 number summary
- Ex. dataset %>% `summarize(min = min(var) , Q1 = quantile(var, 0.25) , median = median(var) , Q3 = quantile(var, 0.75), max = max(var))`

# Scatterplots

**Scatterplot** - geom\_point is used for scatterplot

```
Ex. name of plot <- ggplot(data = dataset, aes(x=var, y=var)) +  
  geom_point(na.rm = TRUE) + theme_minimal(base_size  
    = 15) +  
  labs (x= "", y= "", title= "")
```

- To color the points by gender include col=gender

```
Ex. name of plot <- ggplot(data = dataset, aes(x=var, y=var)) +  
  geom_point(aes(col=gender))
```

- To create separate plots for combinations of levels of 2 vars i.e. (gender) use facet\_wrap ex. facet\_wrap ( ~ gender)

# Regression

**Regression:** Straight line fitted to data to minimize distance b/w data and fitted line

- "Line of best fit" =  $a + bx$
- $a$  = **intercept** (Predictive Value of  $y$  when  $x=0$ )
- $b$  = **slope**  $r^*(s_y/s_x)$
- **Interpretation:** an increase from  $x$  to  $x+1$  is associated w/ an increase in  $y$  by the amount  $B$  (know the units of  $x$  to know the slope) (level-level)
- If you change units of  $x$  slope will change but  $r^2$  and correlation coefficient will not change
- Correlation coefficient and slope have same sign
- Used to describe relationship b/w explanatory and response variables
- Correlation coefficient is just for one pair of variables vs.  $r^2$  can be used for multiple variables

# Regression Cont.

- `lm()` is the function for a linear model
  - The 1st argument that `lm()` wants is a formula  $y \sim x$  (Y = response X = explanatory)
  - The 2nd argument is the data set
  - `lm(formula = y ~ x, data = your_dataset)`
  - Add regression line to a scatterplot using `geom_abline(slope=, intercept=)`
  - `glance(data_lm)` will give r squared
    - Ex. `glance(seed_mod) %>% pull(r.squared)`
- **Interpretation of r-squared:** the fraction of the variation in the values of y that is explained by the line of best fit (the regression of y on x)
- Find **correlation** in data using `summarize(corr_variable = cor(var1, var2))`

# Regression Cont.

- **Correlation ( $\rho$ )**

- Degree to which two variables move in coordination with one another
- $-1 < \rho < 1$

$$\frac{\text{Cov}(x,y)}{\text{SD}(x) \text{SD}(y)}$$

- **Regression ( $\beta$ )**

- Change in the outcome variable for every 1-unit of change in the predictor variable

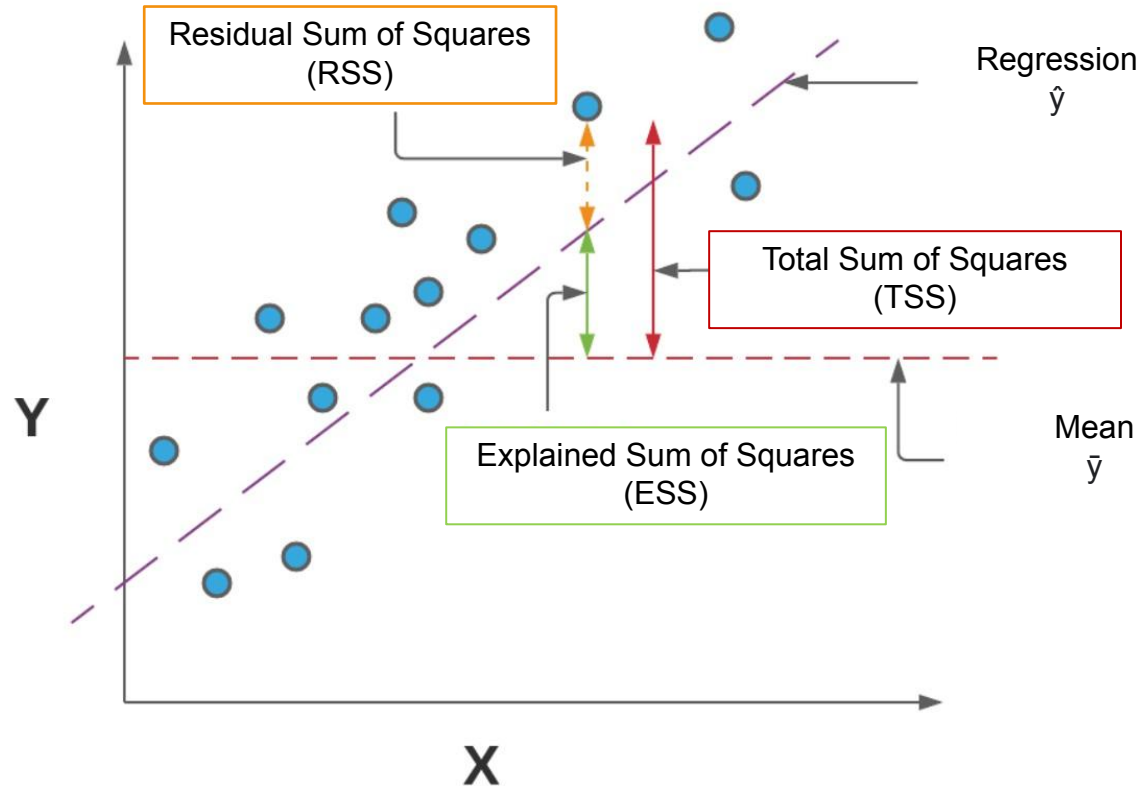
$$\frac{\text{Cov}(x,y)}{\text{Var}(x)}$$

# Regression Cont.

- **$R^2$  (of a linear model)**
- $0 < R^2 < 1$ , also  $\rho^2$
- Goodness of fit of the fitted regression line to a set of data
- $R^2 = \text{ESS} / \text{TSS} = 1 - \text{RSS} / \text{TSS}$
- **Total** Sum of Squares = **Explained** SS + **Residual** SS
$$\text{TSS} = \text{ESS} + \text{RSS}$$
$$\sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y - \hat{y})^2$$



# Regression Cont.



# Key Terms

**Counterfactual** - (how long would a person have lived without treatment X? Aka The “Ideal Experiment”); Observed: Person A  $\rightarrow$  exposure to x; Unobserved: Person A  $\rightarrow$  no exposure to x \*\*the counterfactual

**Confounding** - a relationship between your variable of interest and your outcome of interest is confounded when there is a variable that is associated with both the exposure and the outcome, and is not on the causal pathway between the two

**Conditional distribution** - distribution of variable within or conditional on the level of a second variable

**Marginal distribution** - “in the margin” of the table; row total or column total are the two margins of a two-way table  $\rightarrow$  answers ?’s about the overall distribution of one variable

**Simpson’s Paradox** - An association or comparison that holds for all of several groups can reverse direction when the data are combined to form a single group. This reversal is called Simpson's Paradox

# Observation vs. Experiment

**Observation** - No control over treatment or exposure; Doesn't control for confounding

**Experimentation** - The investigator is experimenting by controlling who is getting the exposure (treatment) and who is not; Ways exposure is assigned = Pre and post designs ; Randomized roll out (stepped wedge) -- Random assignment by time rather than overall

# Population, Sampling, and Conditional Selection

## Population of Interest

- Target Population - entire group of individuals about which we want estimates to apply  
\*problem in PPDAC\*
- Study Population - part of the population which we can select individuals & collect information to draw conclusions about the entire population

## Sampling

- Simple Random Sample - sample chosen by chance, where each individual had the same chance of being selected
- In R an SRS of size 100 can be generated by: `name_1 <- dataframe %>% sample_n(100)`

## Conditional selection

- Conditional on exposure or outcome
- If we are choosing people to participate in our study based on their exposure status this is generally a **cohort design**
- If we are choosing people to participate in our study based on outcome status this is generally a **case control design**
- If we select participants conditional on exposure OR outcome then marginal distribution of exposure is not meaningful

# Sources of Bias

- **Sampling bias** if sampling frame does not cover target population aka “undercoverage bias”
- Participation once sampled an individual may not agree to participate -- referred to as “**response**” or “**nonresponse**” **bias** participants lost to follow-up
- **Contamination** if people who were randomized to control receive the medication or exposure that was intended to be given only to treatment arm participants
- **Adherence** once a participant is randomized to a given treatment, they do not always adhere to that treatment
- Measurement Error
- **Self-report bias** associated w/ willingness or ability to report info = social desirability bias + recall

# Practice Problems

## Question 1

Adapted from “Effects of water quality, sanitation, hand washing, and nutritional interventions on diarrhea and child growth in rural Bangladesh: a cluster randomized controlled trial”

Diarrhea and growth faltering in early childhood are associated with subsequent adverse outcomes. The authors aimed to assess whether water quality, sanitation, and hand washing interventions alone or combined with nutrition interventions reduced diarrhea or growth faltering. There were multiple intervention condition, one of which was a nutritional intervention only. One of the outcomes was childhood stunting.

The authors present the following information:

Group	Stunting	No Stunting	Total
Nutrition	186	381	567
Control	451	652	1103
Total	637	1033	1670

1a.) What is the marginal probability of stunting in this table?

1b.) What are the probabilities of stunting conditional on study group?

1c.) What kind of variable is stunting?

# Question 2

Look at the title of this article from New England Journal of Medicine 2019; 380:415-424

"Partial Oral versus Intravenous Antibiotic Treatment of Endocarditis a randomized, noninferiority, multicenter trial"

Which of the following study design terms apply to this study:

- A. Experimental
- B. Observational
- C. Simple Random Sample
- D. Case-Control

# Question 3

We have a dataset that looks like this:

```
## # A tibble: 4 x 4
```

```
##   smoking   lung_cancer percent number
```

```
##   <chr>      <chr>          <dbl> <dbl>
```

```
## 1 smoker    lung cancer      4.8    12
```

```
## 2 smoker    no lung cancer  95.2   238
```

```
## 3 non-smoker lung cancer    0.9     7
```

```
## 4 non-smoker no lung cancer 99.1  743
```

Which of these statements would produce our preferred bar graph for comparing lung cancer by group

- A. `geom_bar(aes(fill = lung_cancer), stat = "identity", position = "stack")`
- B. `geom_bar(aes(fill = lung_cancer), stat = "identity", position = "dodge")`
- C. `geom_bar(aes(fill = lung_cancer), stat = "count", position = "stack")`
- D. `geom_bar(aes(fill = lung_cancer), stat = "count", position = "dodge")`



# Question 3

```
ggplot(data=data, aes(x=smoking, y=percent))+  
  geom_bar(aes(fill = lung_cancer), stat = "identity", position = "dodge")
```

Percent of given  
smoking status, if they  
have lung cancer or not

