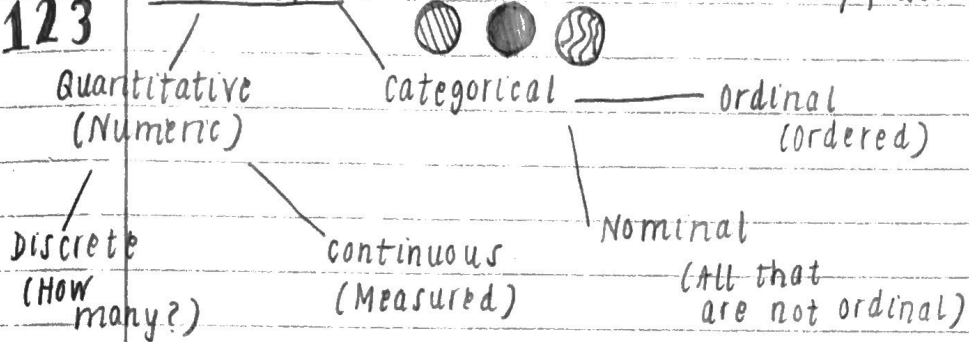


# Midterm 1 Review

123

## Data Types



\* Whole numbers  $\rightarrow$  Discrete

\* Decimals/fractions  $\rightarrow$  Continuous

## Problem Types

1. Descriptive
2. Predictive
3. Causative

\* The JAMA article (discussion 1)

looks at how gun mortality changes over time BUT DOES NOT ASK:

- How will mortality be in the future?

- What causes these mortalities?

\* Can we take these insurance charges and ages, then predict charges for someone new who gives us their age?

In order to deal with data on the computer we use

**R**

ggplot2: data visualization  
dplyr: data manipulation

ggplot2

```
ggplot(dataset, aes( ))  
+ geom_something()
```

\* Arguments go into  
- aes() vary depending  
- on which geom you  
- are using

dplyr

dataset %>% function()

mutate(new\_col = 100 \* old\_col)

This is an example! Many more ways to mutate, like log()!

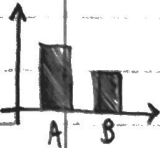

MORE summarize() ← group-by()  
filter() ← rename()  
select() ←  
arrange()

x/y/fill

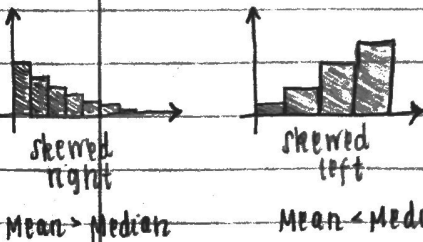
```
ggplot(your_data, aes( )) +  
geom_something()
```



### All the Plots

Plot	Variable	code
Bar plot	1. Categorical variable on the x	x = categorical_col geom_bar()
	2. Categorical x Numeric y (for bar heights)	x = categorical_col y = numeric_col geom_bar(stat = "identity")
	3. 2 categorical variables	x = cat-1_col fill = cat-2_col geom_bar(position = "dodge")

### Histogram



Numeric Quantitative on x that can be grouped into intervals (bins)

```
x = numeric_col  
geom_histogram()  
binwidth = ?
```

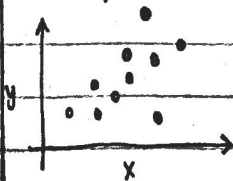
### Boxplot



Numeric on y

```
y = numeric_col  
geom_boxplot()
```

### Scatterplot



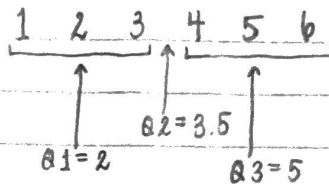
Numeric on both x and y

```
x = num_col-1  
y = num_col-2  
geom_point()
```

# The IQR Visual: Boxplot

## Calculations

Example 1



$$IQR = Q3 - Q1 = 3$$

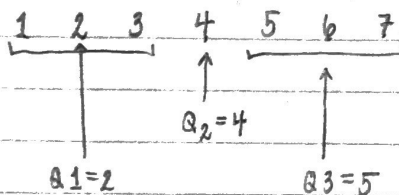
$$\text{Up Whisker} = Q3 + 1.5(IQR)$$

$$= 5 + 4.5 = 9$$

$$\text{Low Whisker} = Q1 - 1.5(IQR)$$

$$= 2 - 4.5 = -2.5$$

Example 2



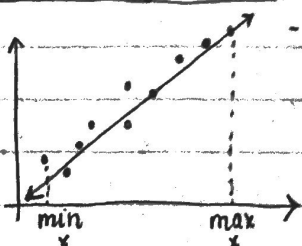
\* Outliers are data points outside the range  $[-2.5, 9]$

\* TO OUTLIERS...

ROBUST!  
Median

NOT!  
Mean

## Linear Regression



$$-1 \leq \text{Correlation} \leq 1$$

$$0 \leq R^2 \leq 1$$

Correlation

↓  
Causation

`your_model <- lm(y ~ x, data = your_data)`

`{ glance(your_model) }`  
`{ tidy(your_model) }`

From the output, you can write

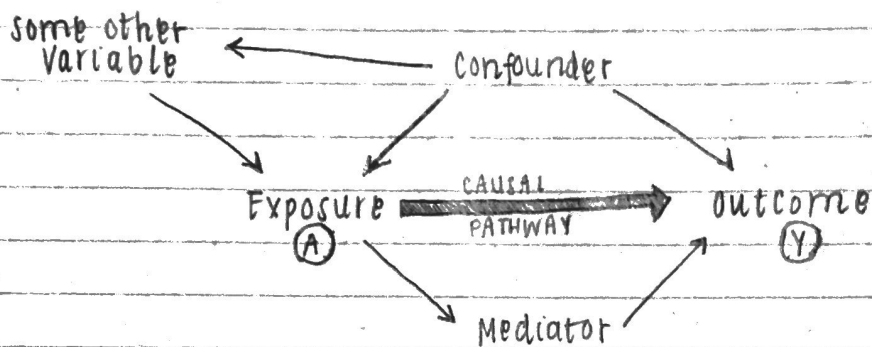
$$y = b_0 + b_1 x$$

↑                    ↑  
intercept          slope

\* Know the interpretations of slope/intercept (see review slides!)

\* To predict, set  $x$  equal to a value and evaluate  $y = mx + b$

## Causal Graphs



## Backdoor Path

A path that does not start at  $(A)$  but goes through  $(A)$  and ends at  $(Y)$