

# Lesson 8: Data visualization of two variables

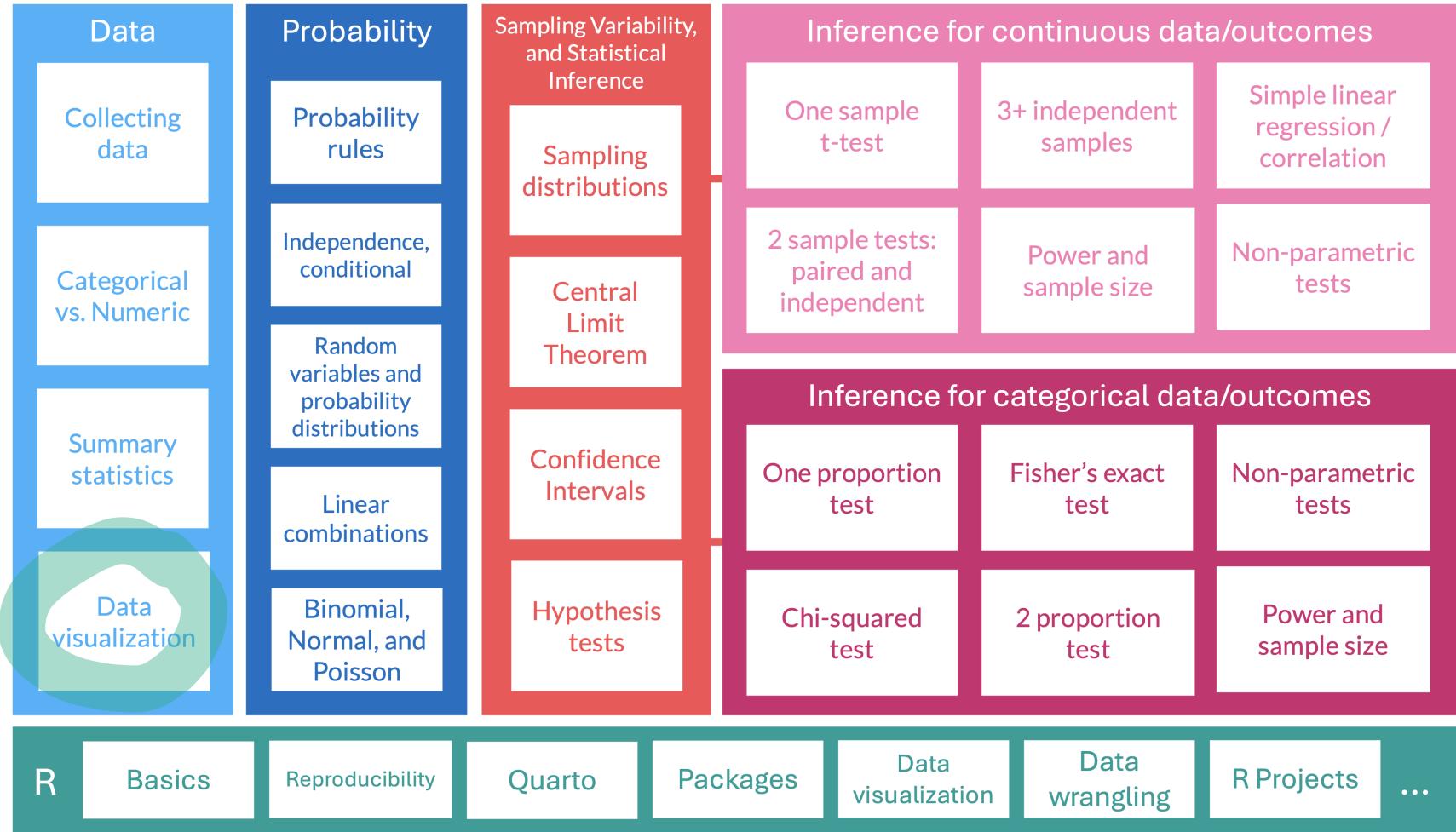
Nicky Wakim

2024-10-23

# Learning Objectives

1. Visualize relationships between two numeric variables using scatterplots and determine their correlation
2. Visualize relationships between two categorical variables using contingency tables and segmented barplots
3. Visualize relationships between a categorical variable and a numeric variable using side-by-side boxplots, density plots, and ridgeline plots

# Where are we?



# Relationships between two variables

- Many studies are motivated by a researcher examining how two or more variables are related
- Example questions about relationships:
  - Do the values of one variable increase as the values of another decrease?
  - Do the values of one variable tend to differ by the levels of another variable?  
*categorical*
- Today we are introducing **summarization and data visualization methods** for exploring and summarizing **relationships between two variables**
- Approaches vary depending on whether the two variables are:
  - Both numerical
  - Both categorical
  - One numerical and one categorical

# We often identify a response variable from our research question

## Response Variable

A **response variable** is defined by the particular research question a study seeks to address

- It measures the outcome of interest in the study

## Explanatory Variable

A study will typically examine whether the values of a response variable differ as values of an **explanatory variable** change, and if so, how the two variables are related.

- A given study may examine several explanatory variables for a single response variable

- Sometimes we're interested in viewing the relationship between our **response variable** and **explanatory variable(s)**
- Sometimes we're just interested in viewing the relationship **between explanatory variables**

# Poll Everywhere Question 1

14:14 Mon Oct 28

Join by Web [PollEv.com/nickywakim275](https://PollEv.com/nickywakim275)

Remember one of the research questions from our first lesson: Is there evidence that introducing peanut products early in life is an effective strategy for reducing risk of developing a peanut allergy?

Which of the following is the response variable?

explanatory

Response Variable	Percentage
Introducing peanut products	9%
Reduced risk	35%
Developing peanut allergy	52%
Predisposition for peanut allergy	4%

Powered by  Poll Everywhere

Reduced risk is  
more statistic  
of the resp  
var

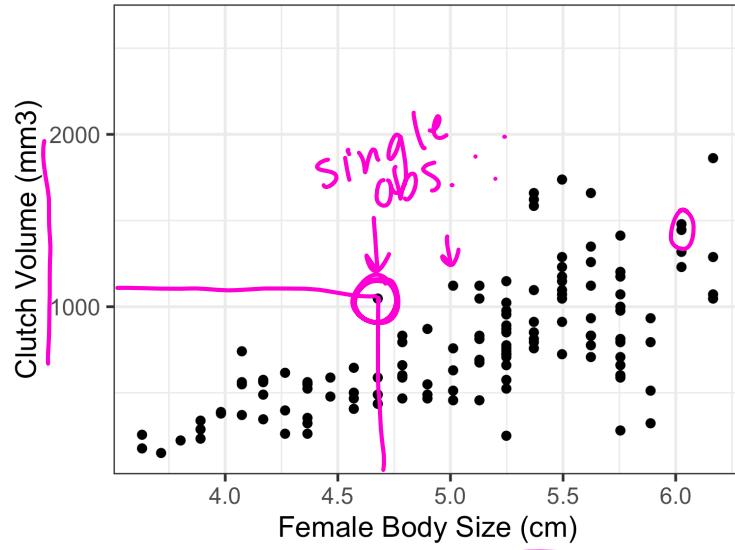
# Learning Objectives

1. Visualize relationships between two numeric variables using scatterplots and determine their correlation
2. Visualize relationships between two categorical variables using contingency tables and segmented barplots
3. Visualize relationships between a categorical variable and a numeric variable using side-by-side boxplots, density plots, and ridgeline plots

# Scatterplots

- Scatterplots provide case-by-case view of the relationship between two numerical variables
- We can make a scatterplot of clutch volume vs. body size, with clutch volume on the y-axis and body size on the x-axis
- Each point represents an observation (egg clutch) with its measurement for clutch volume and body size of parent

```
1 ggplot(data = frog,  
2         aes(x = body.size,  
3                 y = clutch.volume)) +  
4     geom_point(size = 3) +  
5     labs(x = "Female Body Size (cm)",  
6           y = "Clutch Volume (mm3)") ]
```



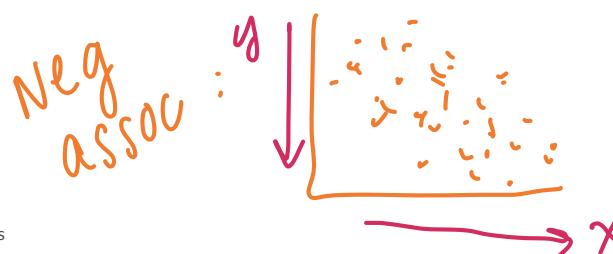
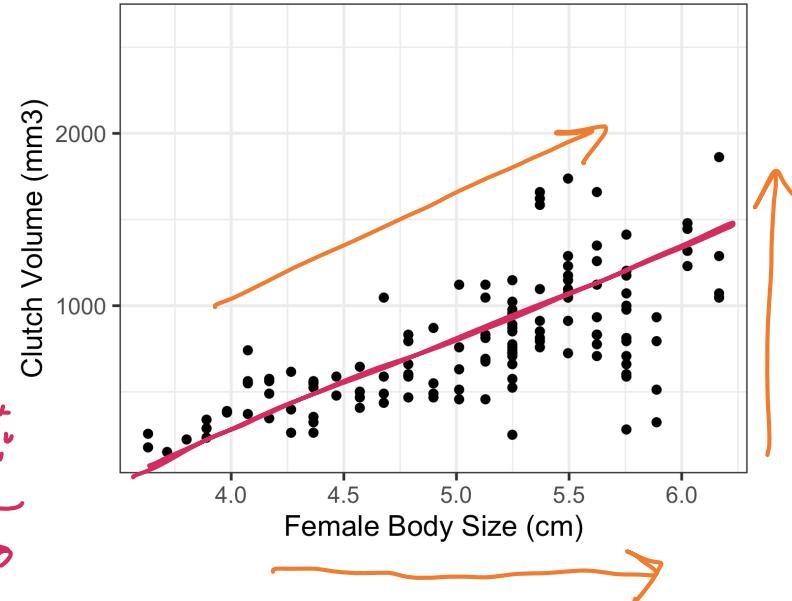
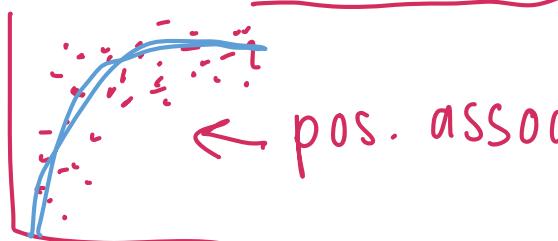
# Describing associations between 2 numerical variables

- Two variables  $x$  and  $y$  are
  - **Positively associated** if  $y$  increases as  $x$  increases
  - **Negatively associated** if  $y$  decreases as  $x$  increases

- If there is no association between the variables, then we say they are **uncorrelated** or **independent**

*regardless of  $x$ ,  $y$  generally stays same*

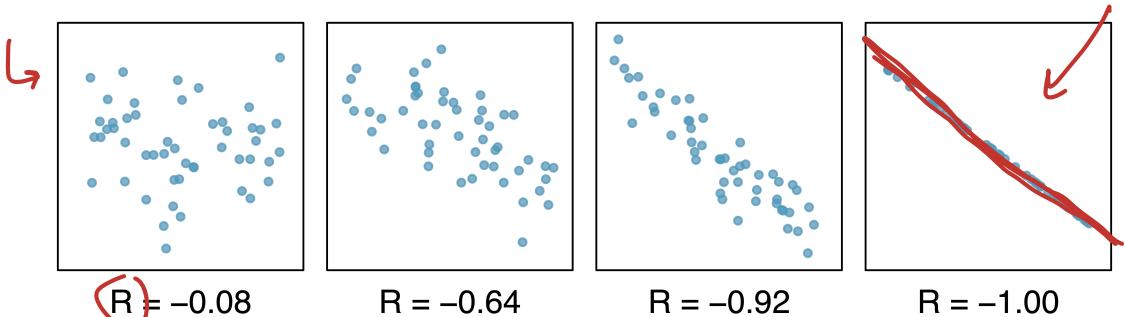
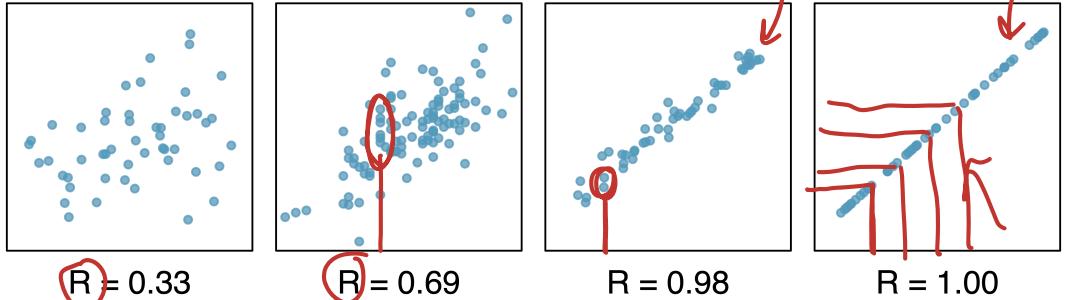
- The term “association” is a very general term
  - Can be used for numerical or categorical variables
  - Not specifically referring to linear associations



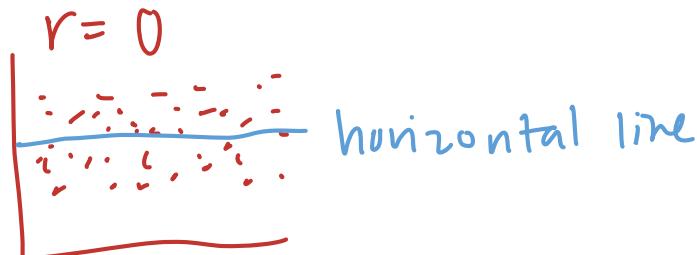
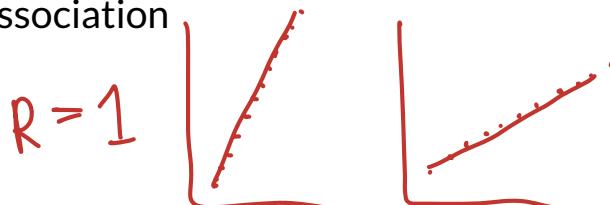
# (Pearson) Correlation coefficient ( $r$ )

$$-1 \leq r \leq 1$$

- $r = -1$  indicates a **perfect negative linear relationship**: As one variable increases, the value of the other variable tends to go down, following a *straight line*
- $r = 0$  indicates **no linear relationship**: The values of both variables go up/down independently of each other
- $r = 1$  indicates a **perfect positive linear relationship**: As the value of one variable goes up, the value of the other variable tends to go up as well in a linear fashion



- The closer  $r$  is to  $\pm 1$ , the stronger the linear association



# Poll Everywhere Question 2

14:27 Mon Oct 28



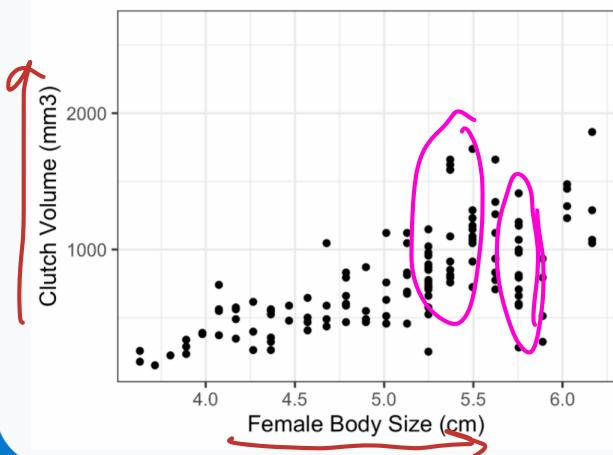
66%



Join by Web [PollEv.com/nickywakim275](https://PollEv.com/nickywakim275)



How would you describe the correlation between female body size and clutch volume?



Negative linear relationship 5%

No linear relationship 0%

Positive linear relationship  95%

# (Pearson) Correlation coefficient ( $r$ ): formula

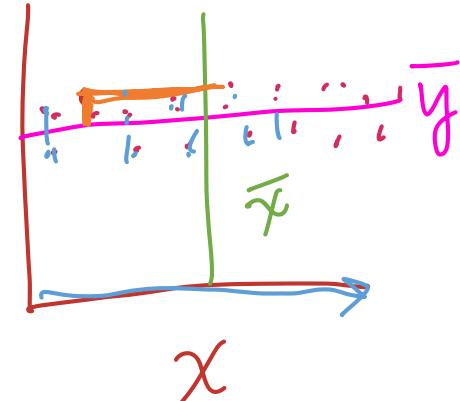
The (Pearson) correlation coefficient of variables  $x$  and  $y$  can be computed using the formula

$$\rightarrow r = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

avg are 0.

where

- $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  are the  $n$  paired values of the variables  $x$  and  $y$
- $s_x$  and  $s_y$  are the sample standard deviations of the variables  $x$  and  $y$ , respectively



- We can use `cor()` in R to calculate this! **two numeric variables!**

```
1 cor(frog$body.size, frog$clutch.volume, use = "pairwise.complete.obs")  
[1] 0.6755435
```

? cor

if no missing data, don't need to use "use"

# Guess the correlation game!

Rossmann & Chance's applet

## Guess the Correlation

Number of points: 25

New Sample

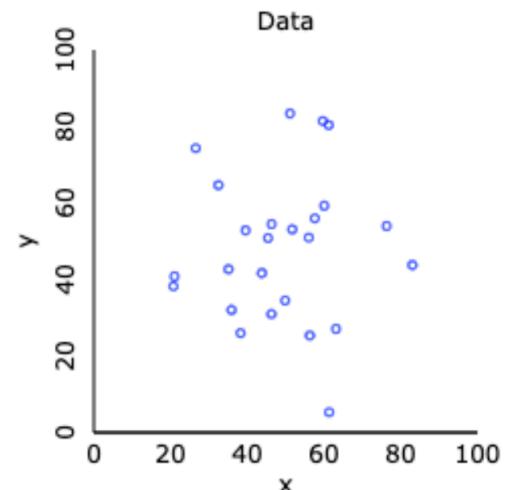
Edit/Paste Data

Correlation guess: -.01

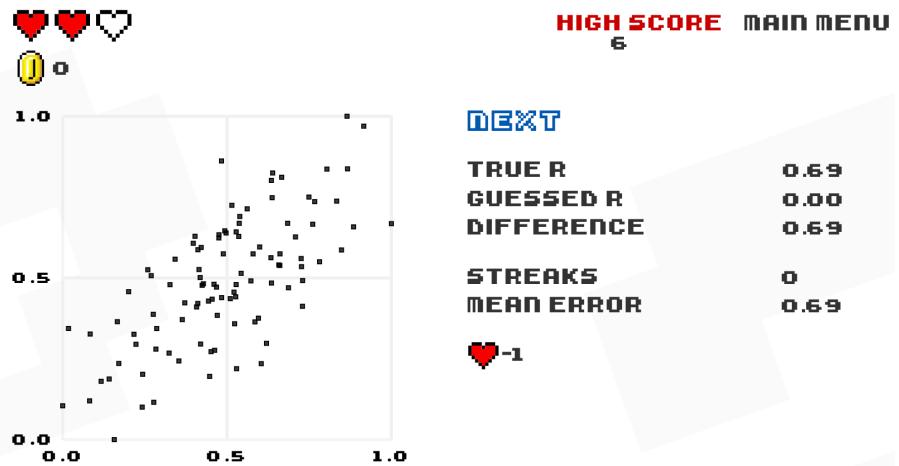
Check Guess

$r = 0.030$

Reset



Or, for the Atari-like experience



<http://guessthecorrelation.com/>

Tracks performance of guess vs. actual, error vs. actual, and error vs. trial

<http://www.rossmanchance.com/applets/GuessCorrelation.html>

# Learning Objectives

1. Visualize relationships between two numeric variables using scatterplots and determine their correlation
2. Visualize relationships between two categorical variables using contingency tables and segmented barplots
3. Visualize relationships between a categorical variable and a numeric variable using side-by-side boxplots, density plots, and ridgeline plots

## From Lesson 4: Contingency tables

- We can start looking at the **contingency table** for hypertension for different age groups
  - **Contingency table:** type of data table that displays the frequency distribution of two or more categorical variables

Table: Contingency table showing hypertension status and age group, in thousands.

Age Group	Hypertension	No Hypertension	Total
18-39 years	8836	112206	121042
40-59 years	42109	88663	130772
60+ years	39917	21589	61506
Total	90862	222458	313320

8,836,000 people are both 18-39 yo & have hyp

## From Lesson 4: Probability tables

Table: Probability table summarizing hypertension status and age group.

Age Group	Hypertension	No Hypertension	Total
18-39 years	0.0282	0.3581	0.3863
40-59 years	0.1344	0.2830	0.4174
60+ years	0.1274	0.0689	0.1963
Total	0.2900	0.7100	1.0000

marginal for  
age grp

• **Joint probability**: intersection of row and column

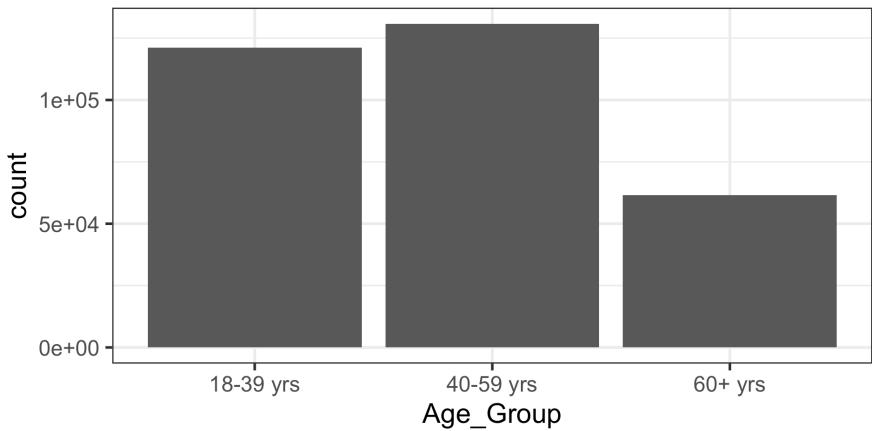
• **Marginal probability**: row or column total

We can work towards visualizing the data in contingency and probability tables

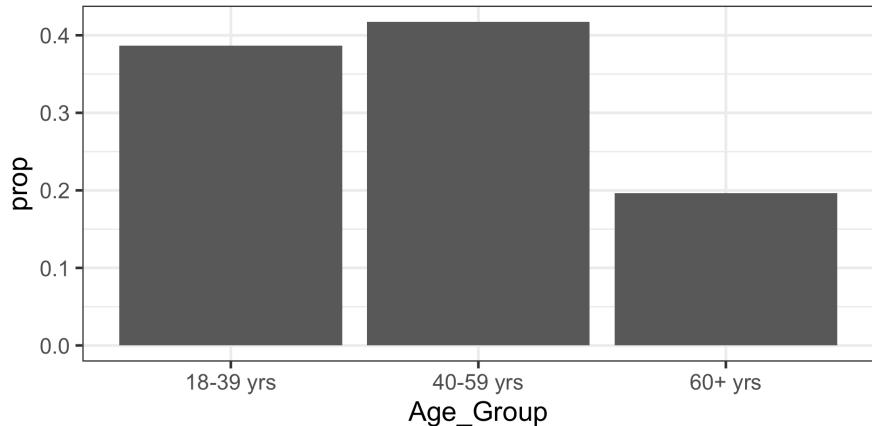
# Last time: Barplots

Counts (below) vs. percentages (right)

```
1 ggplot(data = hyp_data,  
2         aes(x = Age_Group)) +  
3     geom_bar()
```



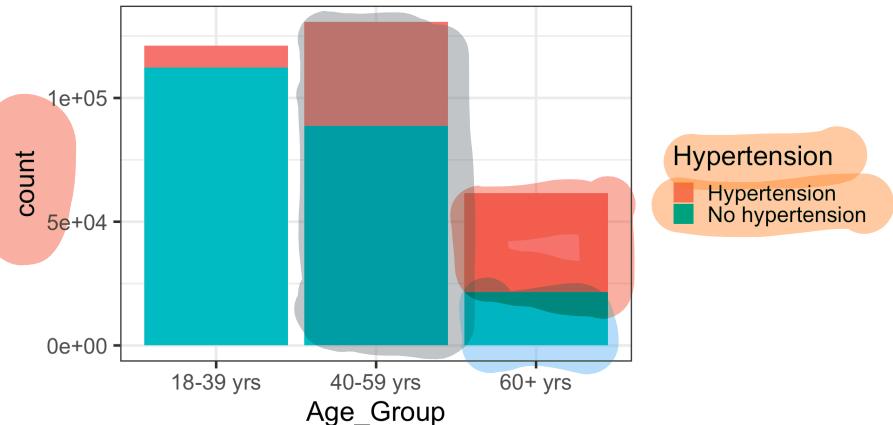
```
1 ggplot(data = hyp_data,  
2         aes(x = Age_Group)) +  
3     geom_bar(aes(y = stat(prop)),  
4                group = 1))
```



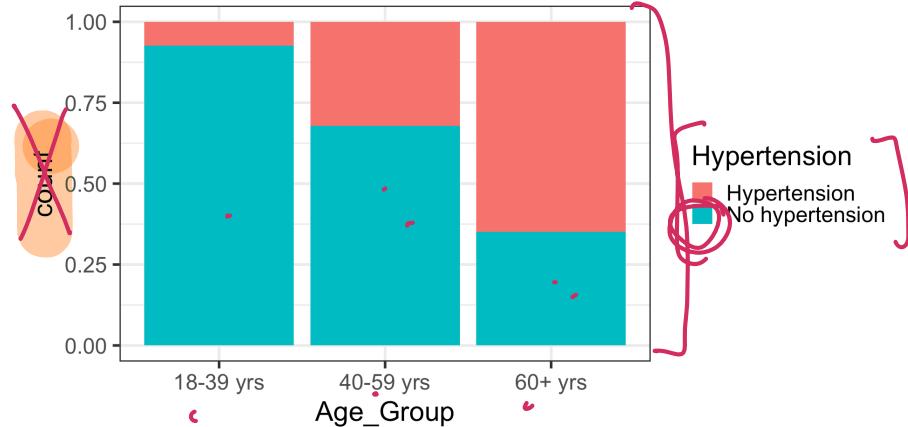
# Barplots with 2 variables: segmented bar plots

- Way of visualizing the information from a contingency table

```
1 ggplot(data = hyp_data,  
2         aes(x = Age_Group,  
3                 fill = Hypertension)) +  
4     geom_bar()
```



```
1 ggplot(data = hyp_data,  
2         aes(x = Age_Group,  
3                 fill = Hypertension)) +  
4     geom_bar(position = "fill")
```



higher age grp,  
more hyp

# Poll Everywhere Question 3

14:42 Mon Oct 28

Join by Web [PollEv.com/nickywakim275](https://PollEv.com/nickywakim275)

How would you describe the relationship between age group and hypertension?

prop hyp ↑ as age grp ↑

Age Group	Hypertension (Red)	No Hypertension (Teal)	Total Count
18-39 yrs	~0.05	~0.95	1.00
40-59 yrs	~0.35	~0.65	1.00
60+ yrs	~0.75	~0.25	1.00

Negatively associated ~~10%~~ 10%

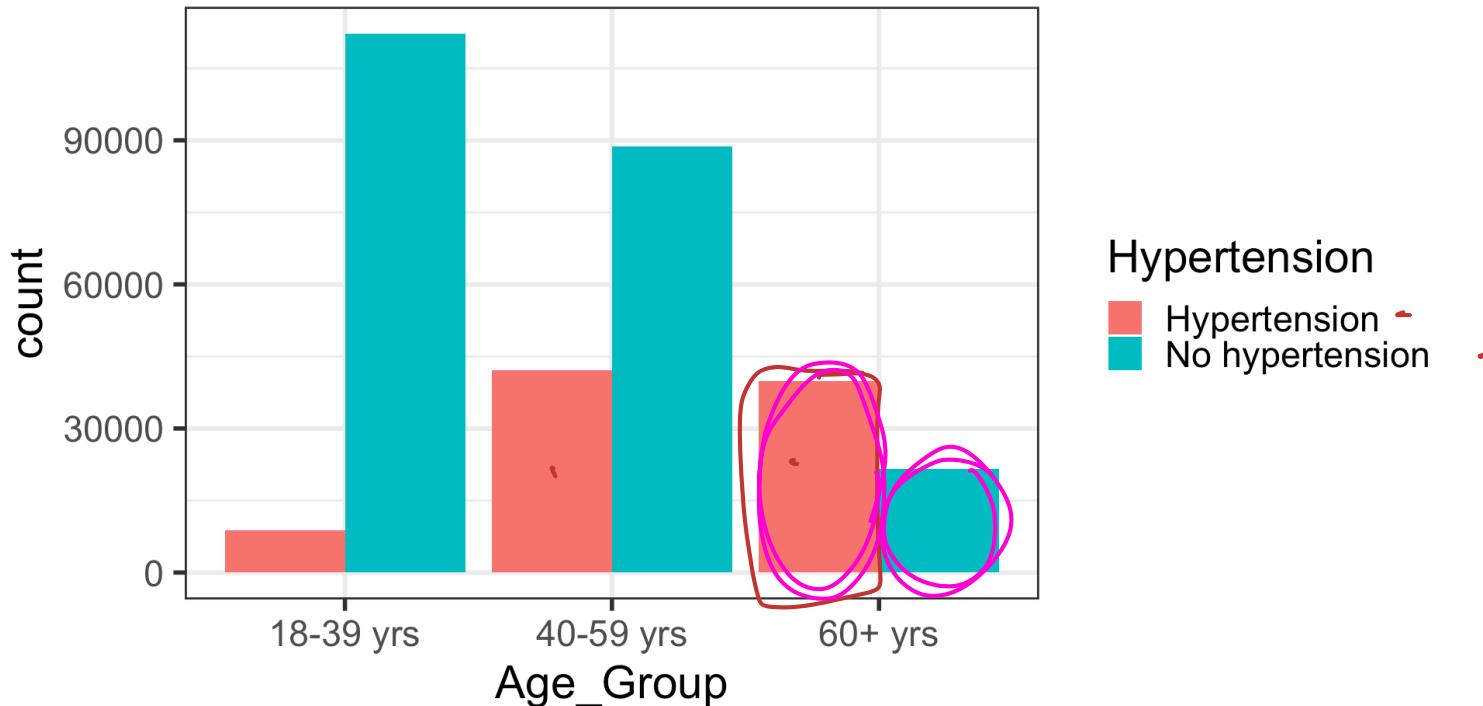
Not associated 0%

Positively associated 90%

Powered by  Poll Everywhere

# Barplots with 2 variables: side-by-side bar plots

```
1 ggplot(data = hyp_data,  
2         aes(x = Age_Group,  
3                 fill = Hypertension)) +  
4     geom_bar(position = "dodge")
```



# Learning Objectives

1. Visualize relationships between two numeric variables using scatterplots and determine their correlation
2. Visualize relationships between two categorical variables using contingency tables and segmented barplots
3. Visualize relationships between a categorical variable and a numeric variable using side-by-side boxplots, density plots, and ridgeline plots

# Visualizing relationships between numerical and categorical variables

- Useful visualizations for directly comparing how the **distribution of a numerical variable differs by category**:
  - Side-by-side boxplots
  - Side-by-side boxplots with data points
  - Density plots by group
  - Ridgeline plot

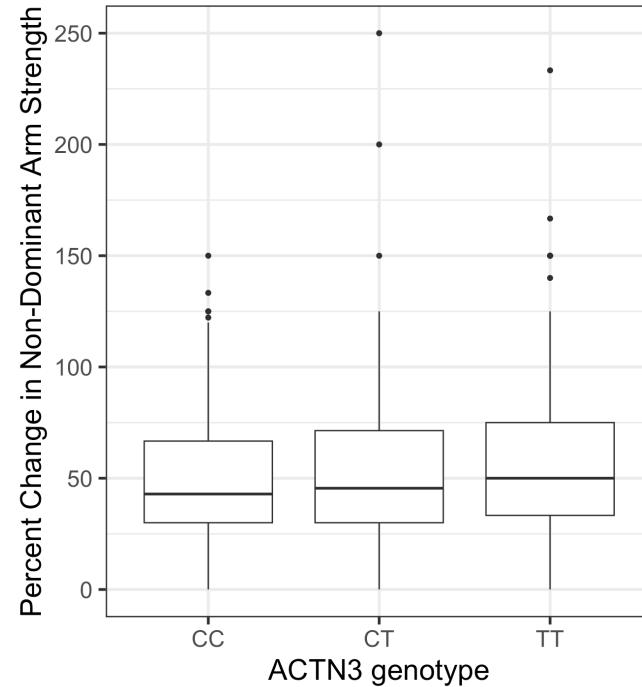
## We need to introduce a new dataset for this

- Study investigating whether ACTN3 genotype at a particular location (residue 577) is associated with change in muscle function
- **Categorical variable:** genotypes (CC, TT, CT)
- **Numeric variable:** Muscle function, measured as percent change in non-dominant arm strength
- We can start the investigation by plotting the relationship

# Side-by-side boxplots

- We can look at the boxplot of percent change for each genotype

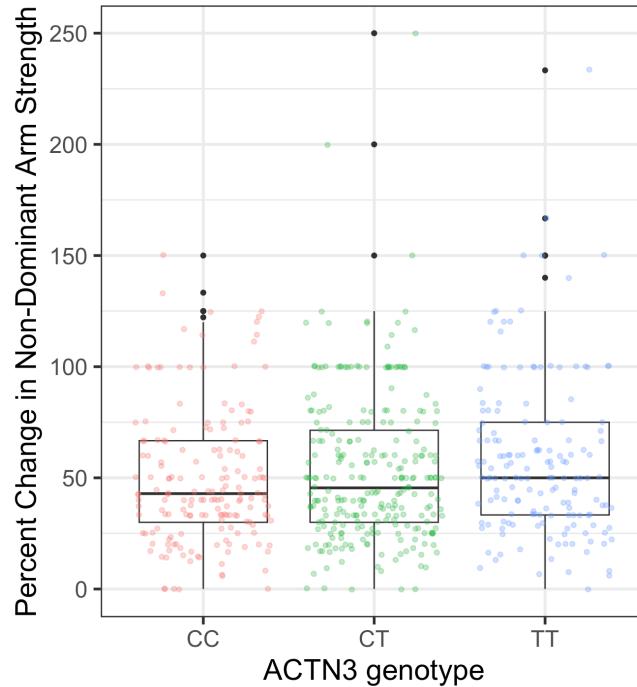
```
1 ggplot(data = famuss,
2         aes(x = actn3.r577x,
3               y = ndrm.ch)) +
4   geom_boxplot() +
5   labs(x = "ACTN3 genotype",
6        y = "Percent Change in Non-Dominant Arm Strength")
```



# Side-by-side boxplots with data points

- We can look at the boxplot of percent change for each genotype **with points shown so we can see the distribution of observations better**

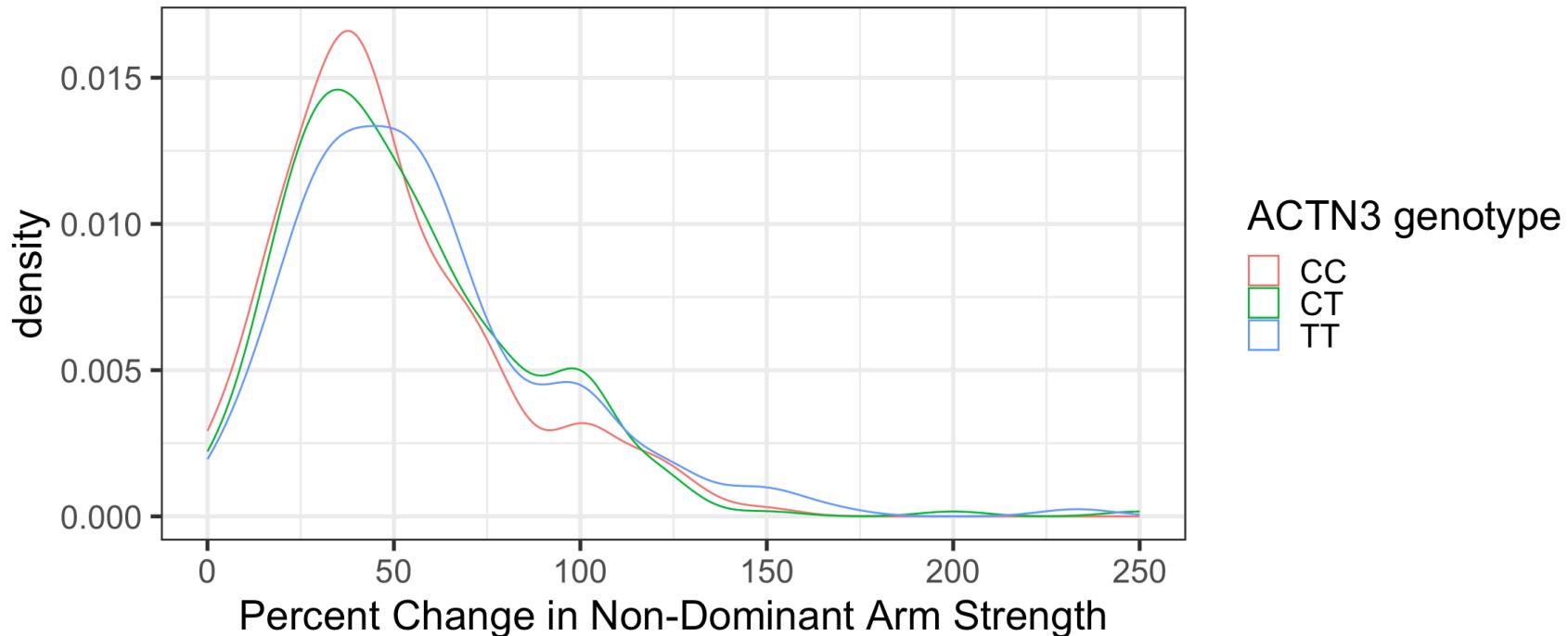
```
1 ggplot(data = famuss,
2         aes(x = actn3.r577x,
3               y = ndrm.ch)) +
4   geom_boxplot() +
5   labs(x = "ACTN3 genotype",
6        y = "Percent Change in Non-Dominant Arm Strength"),
7   geom_jitter(aes(color = actn3.r577x),
8               alpha = 0.3,
9               show.legend = FALSE,
10              position = position_jitter(
11                height = 0.4))
```



# Density plots by group

- Allows us to see the densities of percent change for each genotype

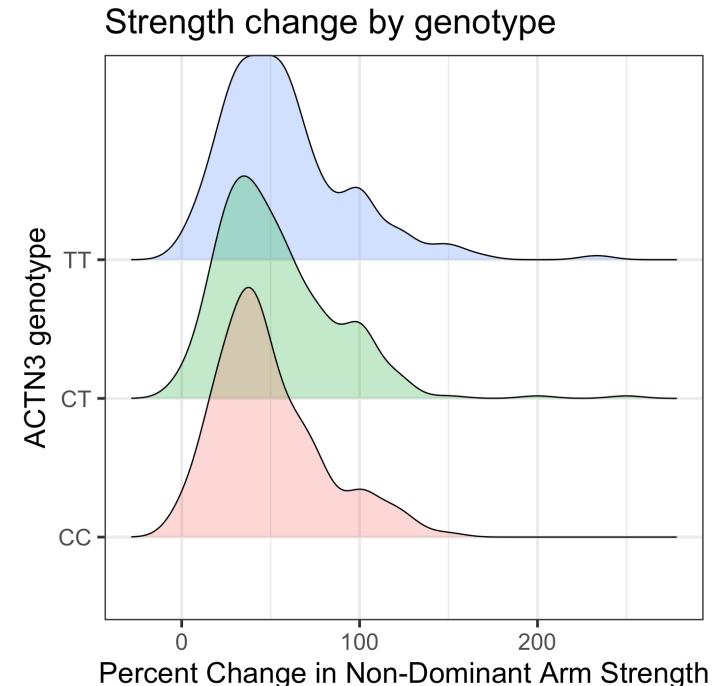
```
1 ggplot(data = famuss, aes(color = actn3.r577x, x = ndrm.ch)) +  
2   geom_density() +  
3   labs(x = "Percent Change in Non-Dominant Arm Strength", color = "ACTN3 genotype")
```



# Ridgeline plot

- Overlapped densities were easy enough to see with 3 genotypes
- If you have **many categories**, a ridgeline plot might make it easier to see

```
1 library(ggridges)
2 ggplot(data = famuss,
3         aes(y = actn3.r577x,
4               x = ndrm.ch,
5               fill = actn3.r577x)) +
6   geom_density_ridges(alpha = 0.3,
7                       show.legend = FALSE) +
8   labs(x = "Percent Change in Non-Dominant Arm Strength",
9        y = "ACTN3 genotype",
10       title = "Strength change by genotype")
```



# Poll Everywhere Question 4

