# Ch. 10: Two Quantitative Variables



# Navigation

By Section


- 10.1: start - end
- 10.2: start - end
- 10.3: start - end
- 10.4: start - end
- 10.5: start - end




# 10.1: Two Quantitative Variables

## Scatterplots and Correlation

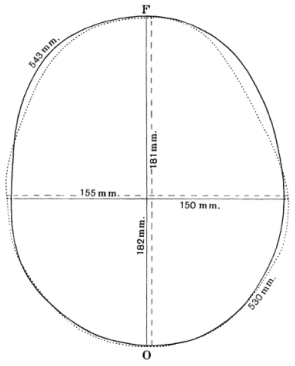
# Graphical Summaries of Quantitative Variables

R.J. Gladstone (1905). "A Study of the Relations of the Brain to the Size of the Head", Biometrika, Vol. 4, p 105-123.

Data collected during 237 autopsies at Middlesex Hospital in London, excluding cases "in which the brain showed a distinctly pathological condition which would have obviously affected its weight"

Variables:

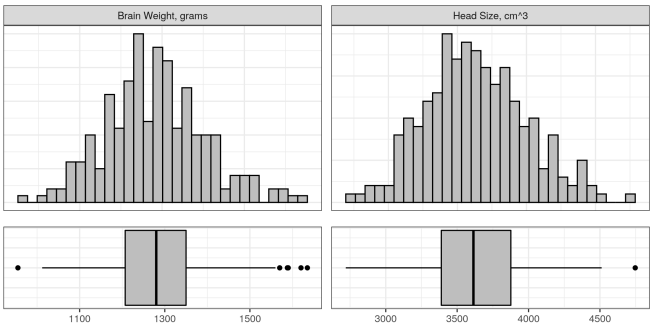
- Gender
- Age (20 - 45 or 46+)
- Brain Weight (g)
- Head Size (cubic cm) the smallest rectangular block which could contain the head






# Graphical Summaries of Quantitative Variables

A single quantitative variable can be summarized visually using a histogram or a bar chart:

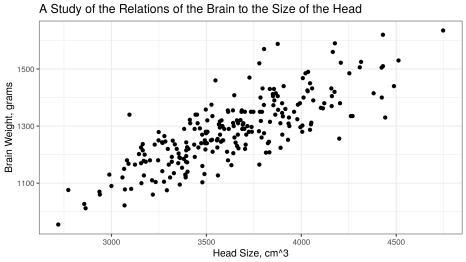





# Graphical Summaries of Quantitative Variables

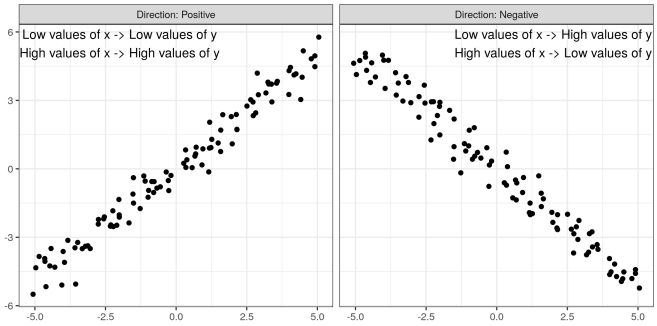
But, summarizing each variable separately doesn't tell us how the two variables might be related.

Is there a relationship between brain weight and head size? How do you know?




A **scatterplot** is a plot with the explanatory variable on the x-axis, and the response variable on the y-axis. Observations are shown as points corresponding to a set of quantitative measurements.

# Describing Variable Relationships: Direction

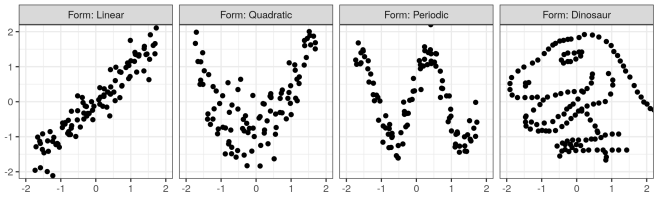



Positive slope: as $x$ increases, $y$ increases too.



# Describing Variable Relationships: Form

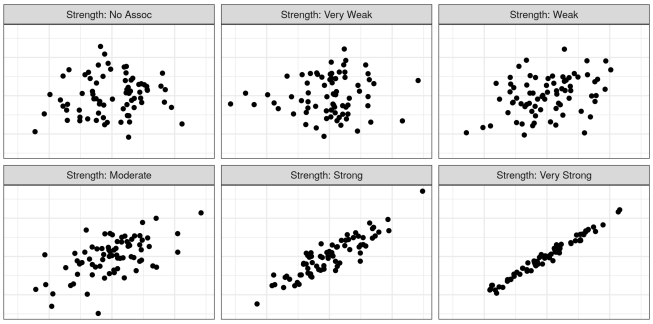
The **form** of an association is whether it follows a linear pattern, or some sort of more complicated pattern - periodic, polynomial (quadratic, cubic, etc.)






# Describing Variable Relationships: Strength

The **strength** of an association indicates how well the value of one variable can be predicted if you know the value of the other variable.

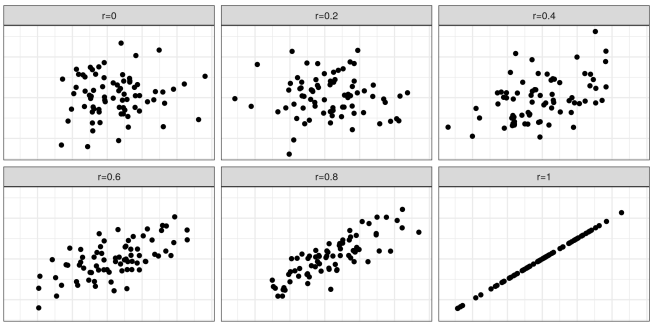



We can describe the strength and direction of a *linear* relationship using the **correlation coefficient**

# Correlation Coefficient

The **correlation coefficient**, $r$, is always between -1 and 1.

| -1 | -0.5 | 0 | 0.5 | 1 |
|---|---|---|---|---|
| perfect negative correlation | moderate negative correlation | no correlation | moderate positive correlation | perfect positive correlation |

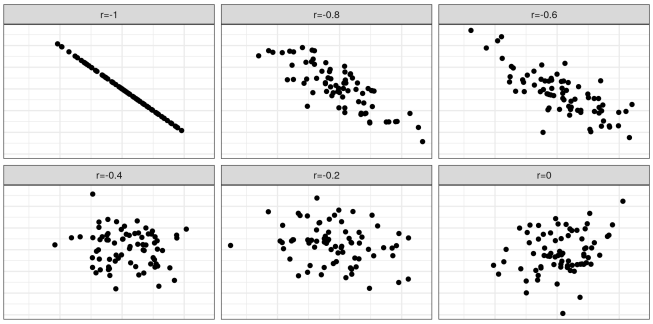





# Correlation Coefficient

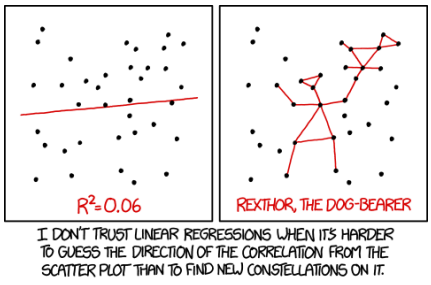
The **correlation coefficient**, $r$, is always between -1 and 1.

| -1 | -0.5 | 0 | 0.5 | 1 |
|---|---|---|---|---|
| perfect negative correlation | moderate negative correlation | no correlation | moderate positive correlation | perfect positive correlation |



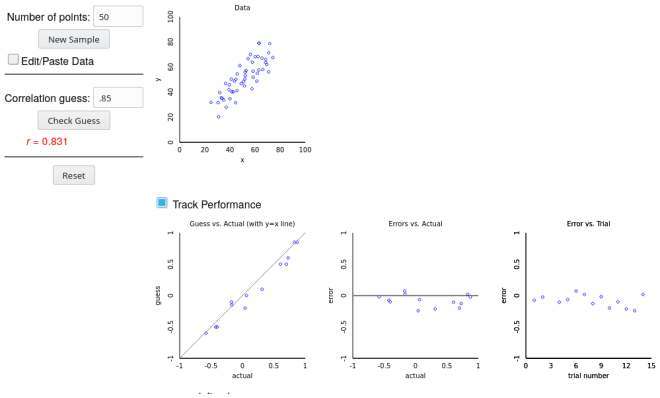



# Correlation Coefficient



# Correlation Coefficient

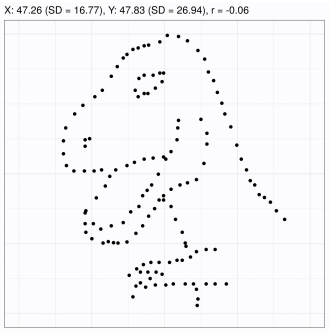
Get a feel for it by playing the correlation guessing game!






# Correlation Coefficient

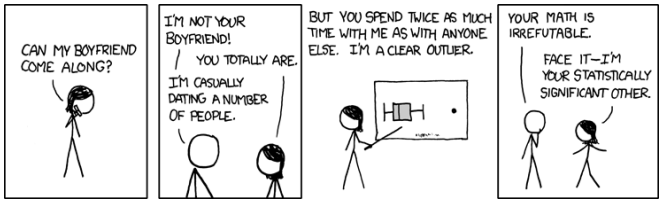
The correlation coefficient is only useful for showing the strength of linear relationships.




All of these plots have essentially the same correlation coefficient, but in some cases there are very clear associations between $x$ and $y$



# Outliers and Influential Observations

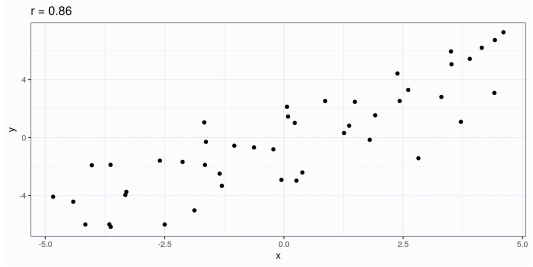



With one variable, outliers are fairly easy to spot

When there are two variables, we don't just have to worry about outliers in one dimension; we also have to worry about **influential observations**

# Outliers and Influential Observations

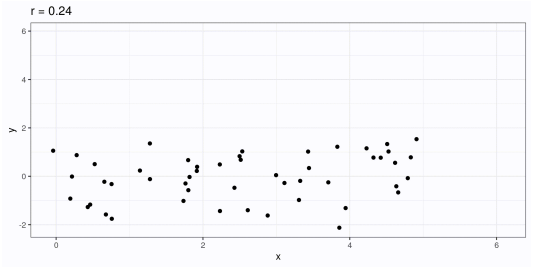
Influential observations are observations which, if included, change our understanding of the relationship between two variables.






# Outliers and Influential Observations

Influential observations are observations which, if included, change our understanding of the relationship between two variables.






# Exploration 10.1

Work through Exploration 10.1 to get a chance to put the material in this section into practice. You can turn it in for 10 points of extra credit in the "Assignment" category.

# 10.2: Inference for the Correlation Coefficient

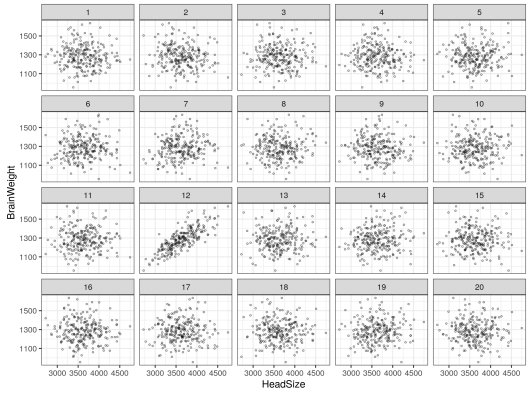
## Simulation Based Approach



## One of these things is not like the others



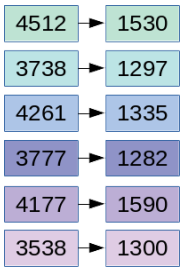
## Which one of these things is not like the others?



## Simulation-based Inference for Correlation Coefficient

Our null hypothesis is $H_0$ : No relationship between $x$ and $y$

How can we simulate this?

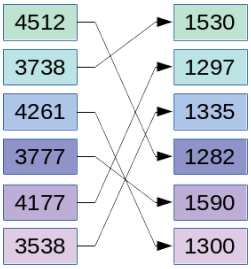





## Simulation-based Inference for Correlation Coefficient

Our null hypothesis is $H_0$ : No relationship between $x$ and $y$

How can we simulate this?




If there's no relationship between $x$ and $y$, then it doesn't really matter what $x$ value is paired with a given $y$ value... so we can just change which values are paired together.



## Summary of Simulation Model

- Null hypothesis: No association between $x$ and $y$ variables
- One repetition: Re randomizing the response outcomes to the explanatory variable values (randomize $y$ values)
- Statistic: Correlation coefficient, $r$

# Simulation-based Inference for Correlation Coefficient

Our null hypothesis is $H_0$ : No relationship between $x$ and $y$

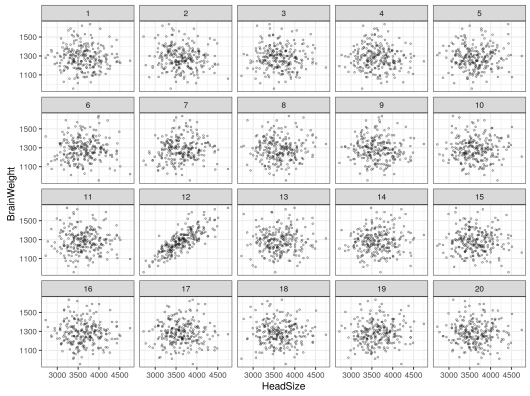
How can we simulate this?

| | |
|---|---|
| 4512 | 1282 |
| 3738 | 1530 |
| 4261 | 1300 |
| 3777 | 1590 |
| 4177 | 1297 |
| 3538 | 1335 |

This is equivalent to shuffling the order of $y$ and creating a new regression



# Which one of these things is not like the others?

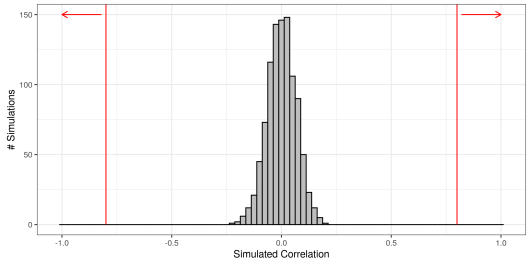



Here, our $y$ values have been shuffled for each sub-plot that isn't plot 12.
Plot 12 contains the original data.



# 3 S strategy

1. **Statistic** - correlation coefficient from the sample
2. **Simulate**
   - Assume there is no relationship between $x$ and $y$
   - Shuffle the values of $y$
   - Calculate the correlation coefficient from the simulated date
3. **Strength of evidence** - in how many simulated samples did we get a correlation coefficient with magnitude greater than our sample correlation coefficient?



## Hypotheses for Correlation Coefficient

$H_0$ : There is no relationship between $x$ and $y$ ($r = 0$)

$H_A$ : There is a relationship between $x$ and $y$ ($r \neq 0$)



# 10.3: Least Squares Regression