Statistics 101

Sonja Winter
Winter Statistics

@winterstat
sonja@winterstatistics.com

Workshop requirements?

- R and Rstudio installed?
- Scripts downloaded from Github?
  - https://github.com/winterstat/Statistics-101

Me, myself, and I

- MSc. Developmental Psychology
- Boutique data analysis company Winter Statistics
- Love: food, kittens, travel, learning

Goals

Goals

Today

- Research methodology
- The nitty gritty
  - Descriptive analyses
  - Exploratory analyses
- Break
- The nitty gritty (con’t)
  - Inferential analyses
  - Predictive analyses

Research methodology

Start with a question

Research methodology

A good question …

1. uses clear terms for the constructs of interest
2. is placed within the right context and application
3. comes with a metric for success, an outcome that you think gives a sufficient answer to your question
4. fits one type of data analytic approach

Research methodology

Data analytic approaches

1. Descriptive
2. Exploratory
3. Inferential
4. Predictive
5. Causal
6. Mechanistic

Source: Leek, J. (2015). *The elements of data analytic style.*

Research methodology

Data analytic approaches

1. Descriptive
2. Exploratory
3. Inferential
4. Predictive
5. Causal
6. Mechanistic

Measure	Average
Temperature	35.99
Lag	5.58
Failure %	51.57

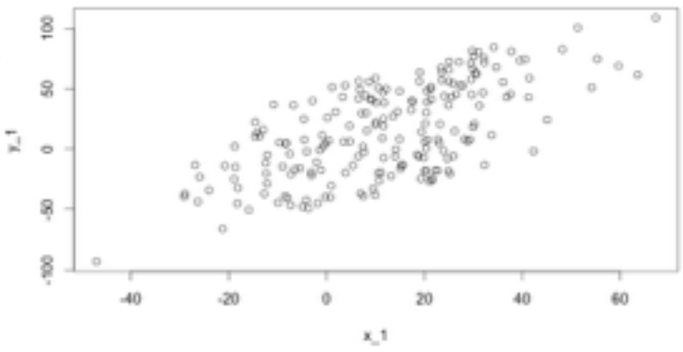
Source: Leek, J. (2015). *The elements of data analytic style.*

Research methodology

Data analytic approaches

1. Descriptive
2. Exploratory
3. Inferential
4. Predictive
5. Causal
6. Mechanistic

Measure	Average
Temperature	35.99
Lag	5.58
Failure %	51.57



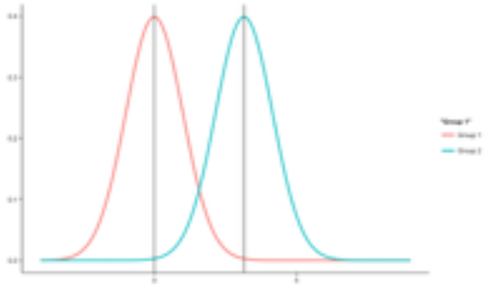
Source: Leek, J. (2015). *The elements of data analytic style.*

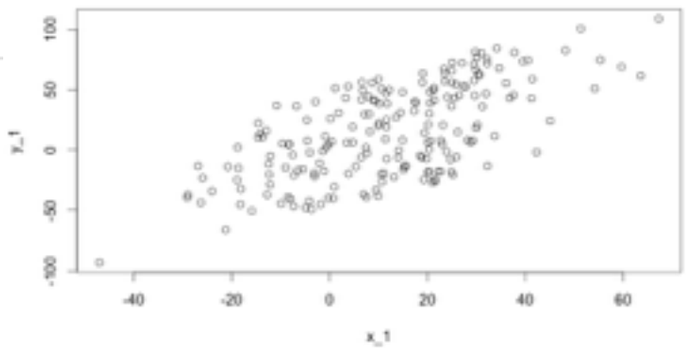
Research methodology

Data analytic approaches

1. Descriptive
2. Exploratory
3. Inferential
4. Predictive
5. Causal
6. Mechanistic

Measure	Average
Temperature	35.99
Lag	5.58
Failure %	51.57





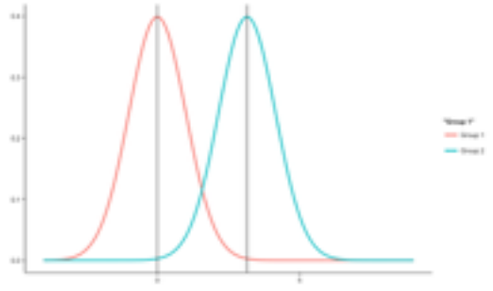
Source: Leek, J. (2015). *The elements of data analytic style.*

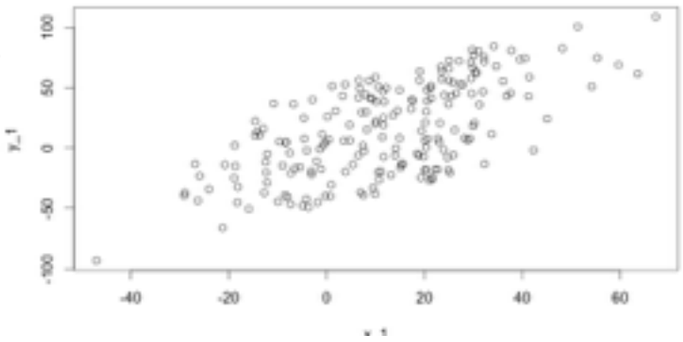
Research methodology

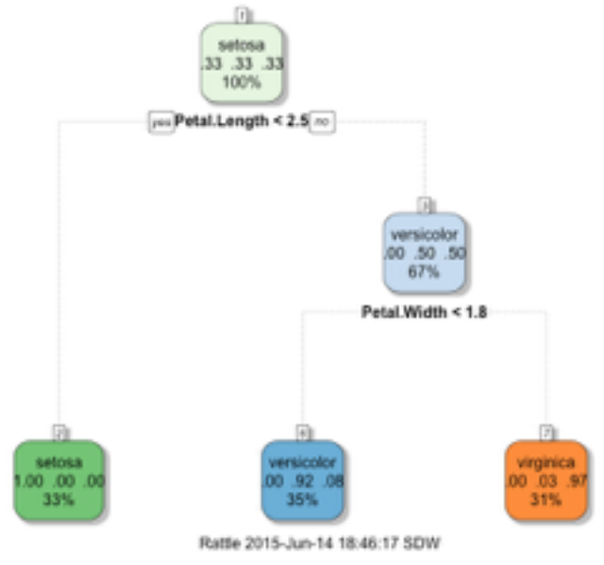
Data analytic approaches

1. Descriptive
2. Exploratory
3. Inferential
4. Predictive
5. Causal
6. Mechanistic

Measure	Average
Temperature	35.99
Lag	5.58
Failure %	51.57







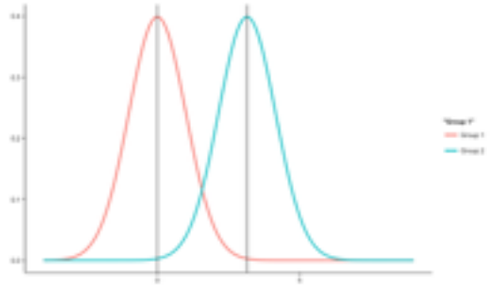
Source: Leek, J. (2015). *The elements of data analytic style.*

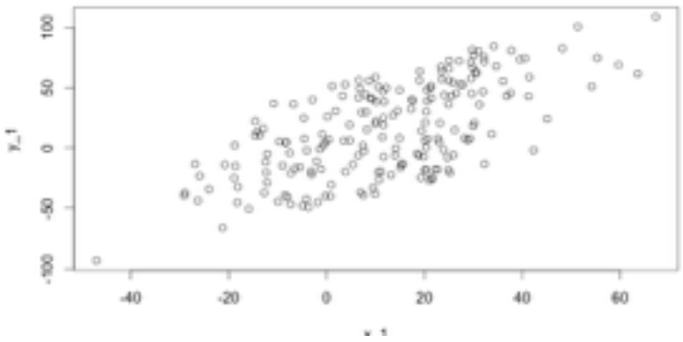
Research methodology

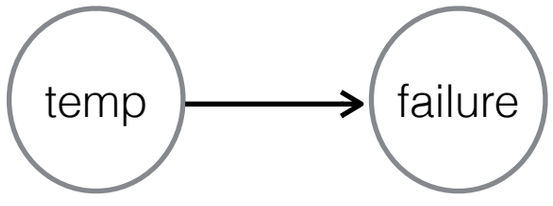
Data analytic approaches

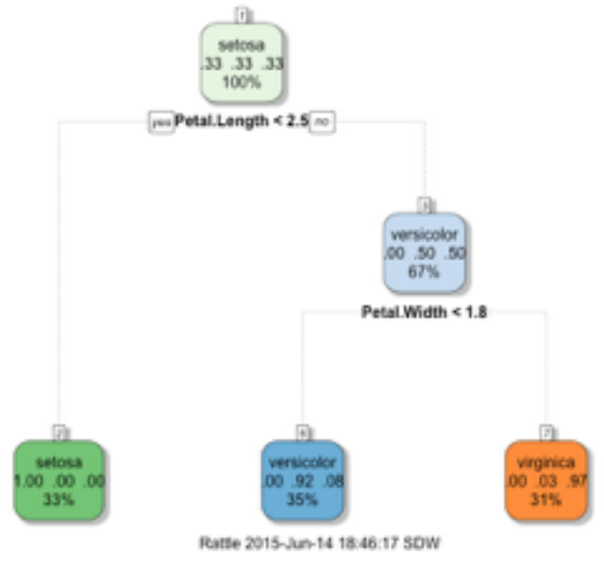
1. Descriptive
2. Exploratory
3. Inferential
4. Predictive
5. Causal
6. Mechanistic

Measure	Average
Temperature	35.99
Lag	5.58
Failure %	51.57





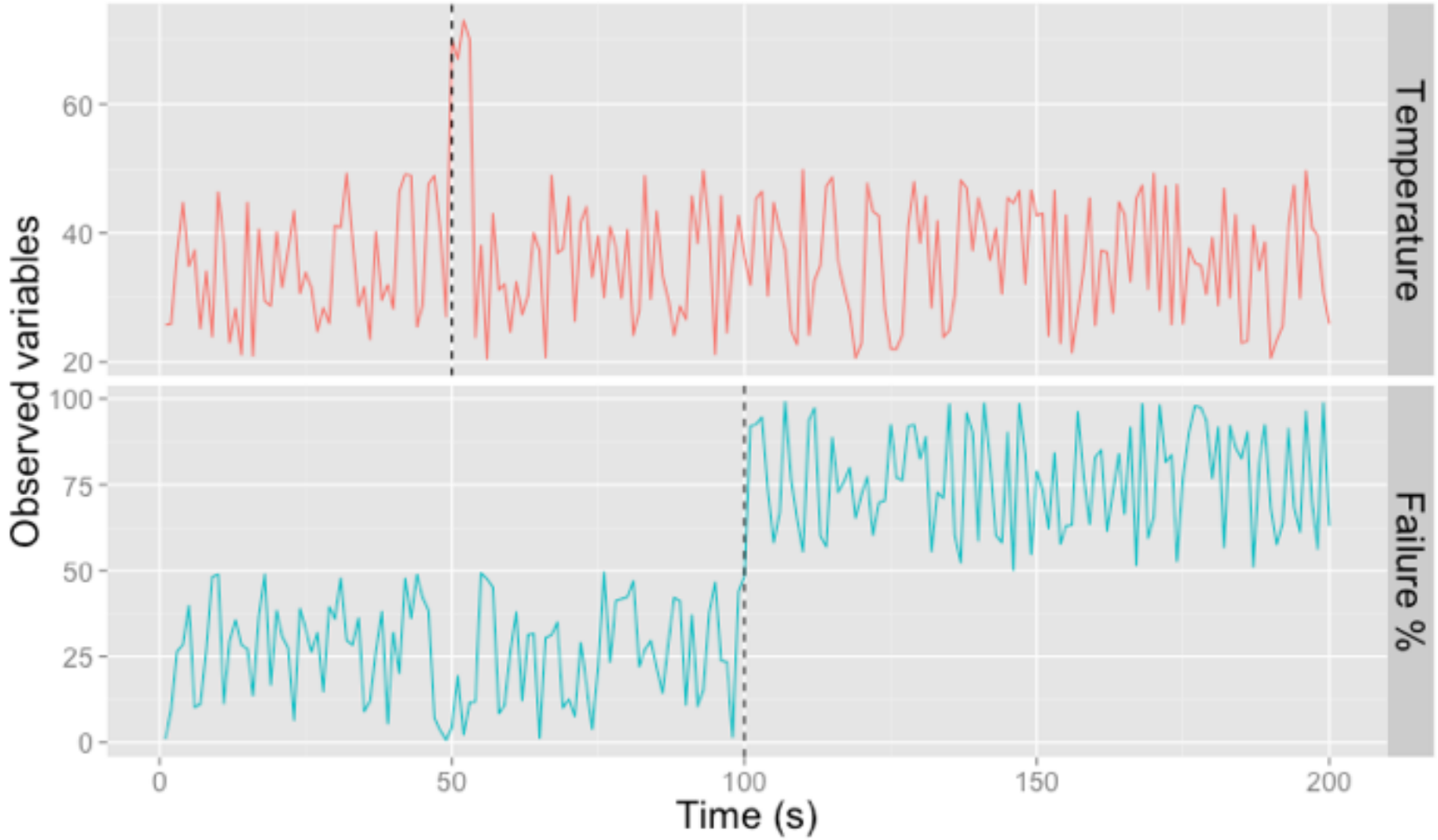




Source: Leek, J. (2015). *The elements of data analytic style.*

Research methodology

- Three criteria of causality (1)
  - Temporal precedence (cause before effect)

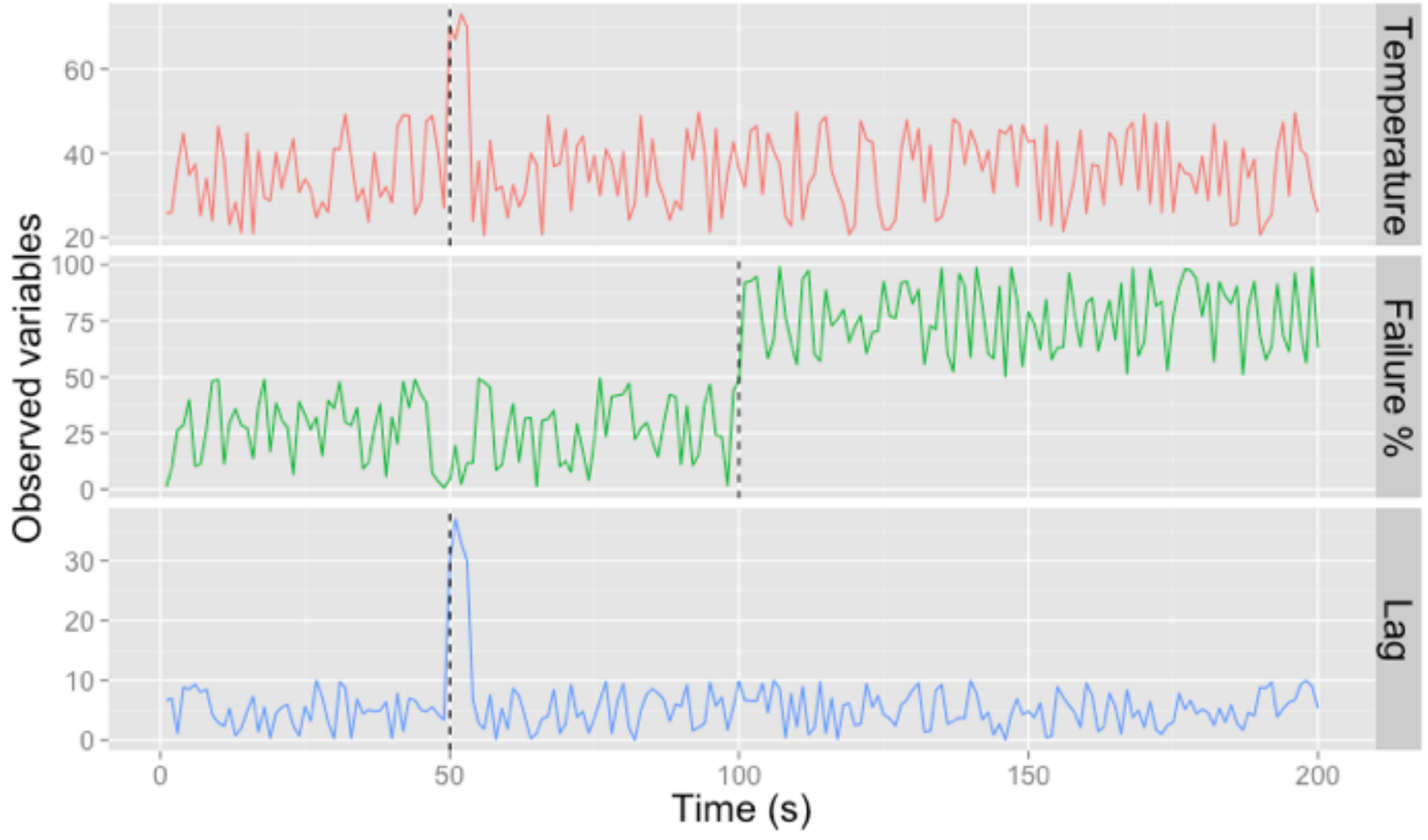


Research methodology

- Three criteria of causality (2)
  - Covariation of cause and effect
    - IF temp spike THEN more failures
    - IF NO temp spike THEN less failures

Research methodology

- Three criteria of causality (3)
  - No plausible alternative explanations (third variable)



Research methodology

- Three criteria of causality (3)
    - No plausible alternative explanations (third variable)
        - **How?** Control groups, experiments, controlling for possible “third variables” in your analysis

- **Important:** No statistical model can prove causality, your research design is the only thing that can

Descriptive data analysis

Descriptive data analysis

- How does one summarise categorical data?
  - Frequencies

Color	Frequency
Blue	14
Yellow	6
Red	18
Green	2

Nominal variable

Movie rating	Frequency
*	3
**	7
***	11
****	14
*****	5

Ordinal variable

Descriptive data analysis

- How does one summarise categorical data?
  - Modes and medians

Color	Frequency
Blue	14
Yellow	6
Red	18
Green	2

Nominal variable

Mode: Red

Movie rating	Frequency
*	3
**	7
***	11
****	14
*****	5

Ordinal variable

Mode: ****
Median: ***

Descriptive data analysis

- How does one summarise categorical data?
-

Definitions

Mode value that occurs most often in the data

Median the middle value when scores are ranked in order of magnitude. To find middle value of variable x, use `(length(x) + 1)/2`

Note: with even number of values, add two middle values and divide by 2 to find median.

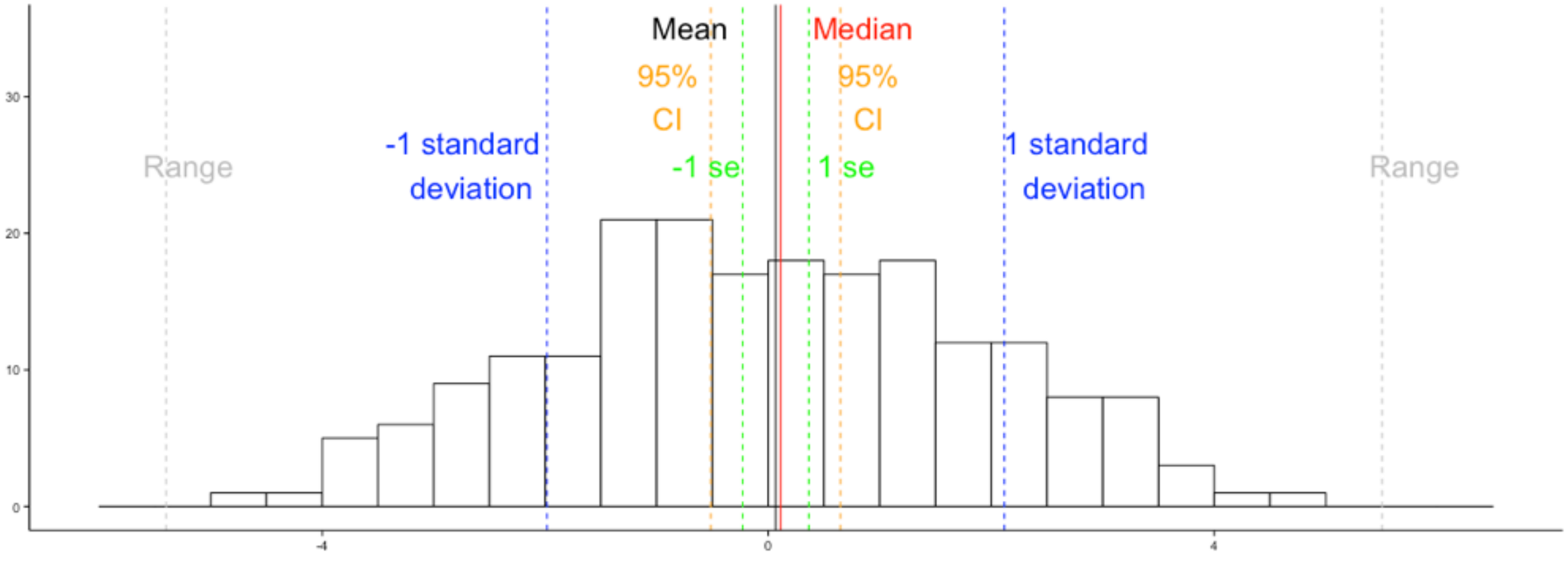
25% and 75% Quartiles the median of the observed data values below and above the overall median.

Mode: Red

Mode: ****
Median: ***

Descriptive data analysis

- How does one summarise continuous variables?
  - median, mean, range, variance, standard deviation, standard error, confidence intervals



De sis

- Ho ?
  - s

Definitions

Mean the average score. To find the mean of variable x, use the following formula: `sum(x) / length(x)`

Range the area of variation between the upper and lower limit of a variable. The find the range of x, use: `max(x) - min(x)`

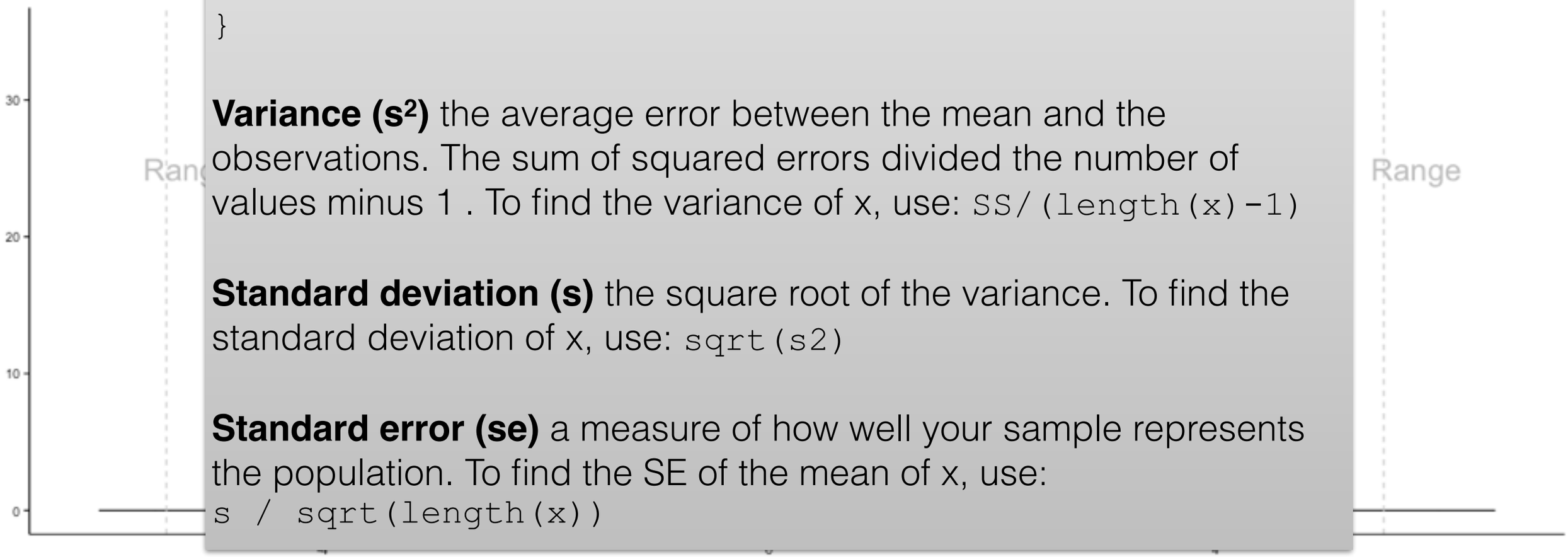
Sum of squared errors (SS) the sum of the squared deviance of each value from the mean. To find the sum of squared errors of variable x, use:

```
for (i in 1:length(x)) {
  value = (x[i] - mean(x))^2
  SS = ss + value
}
```

Variance (s^2) the average error between the mean and the observations. The sum of squared errors divided the number of values minus 1 . To find the variance of x, use: `SS/(length(x)-1)`

Standard deviation (s) the square root of the variance. To find the standard deviation of x, use: `sqrt(s2)`

Standard error (se) a measure of how well your sample represents the population. To find the SE of the mean of x, use:
`s / sqrt(length(x))`



De sis

- Ho
  -

Definitions

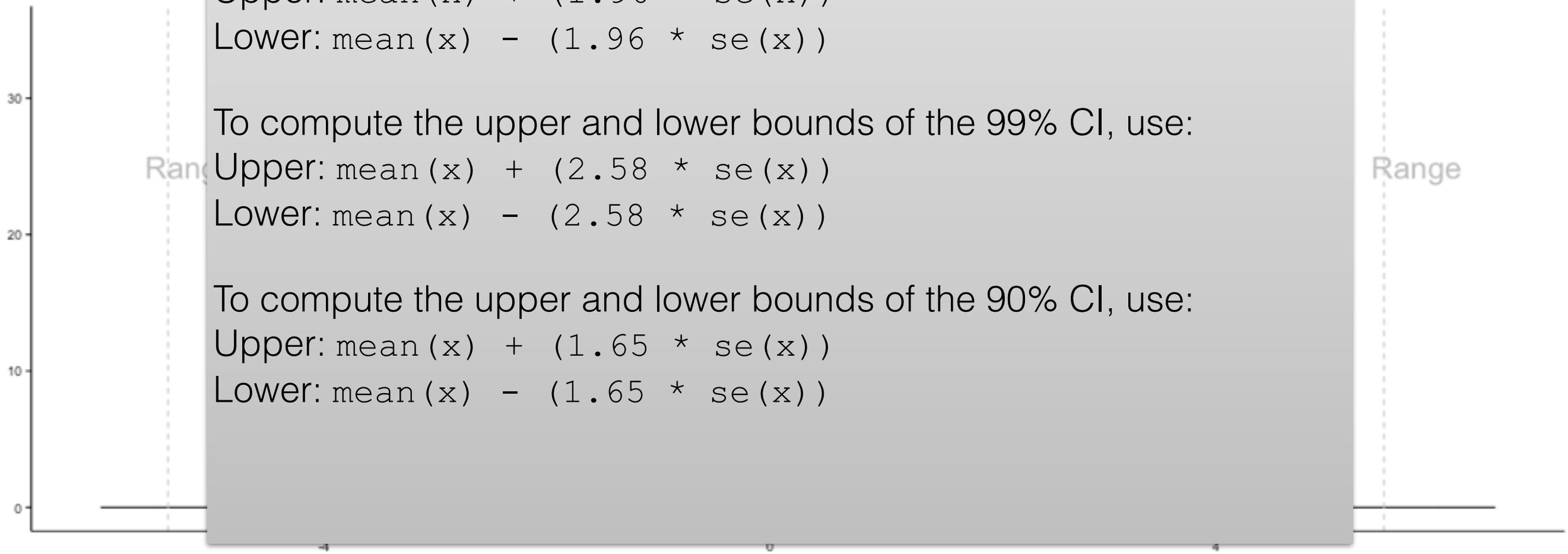
Confidence Interval (CI) a range that indicates that if you were to repeatedly sample from the population of interest, the true population value would be part of the range x % of the time. The higher the percentage, the higher the chance that the true population value is part of the interval. The most commonly used percentage is 95%.

To compute the upper and lower bounds of the 95% CI, we use *z*-scores. The *z*-score that indicates that 2.5% of values will be more positive or negative than that value is **1.96 or -1.96**.

To compute the upper and lower bounds of the 95% CI, use:
Upper: `mean(x) + (1.96 * se(x))`
Lower: `mean(x) - (1.96 * se(x))`

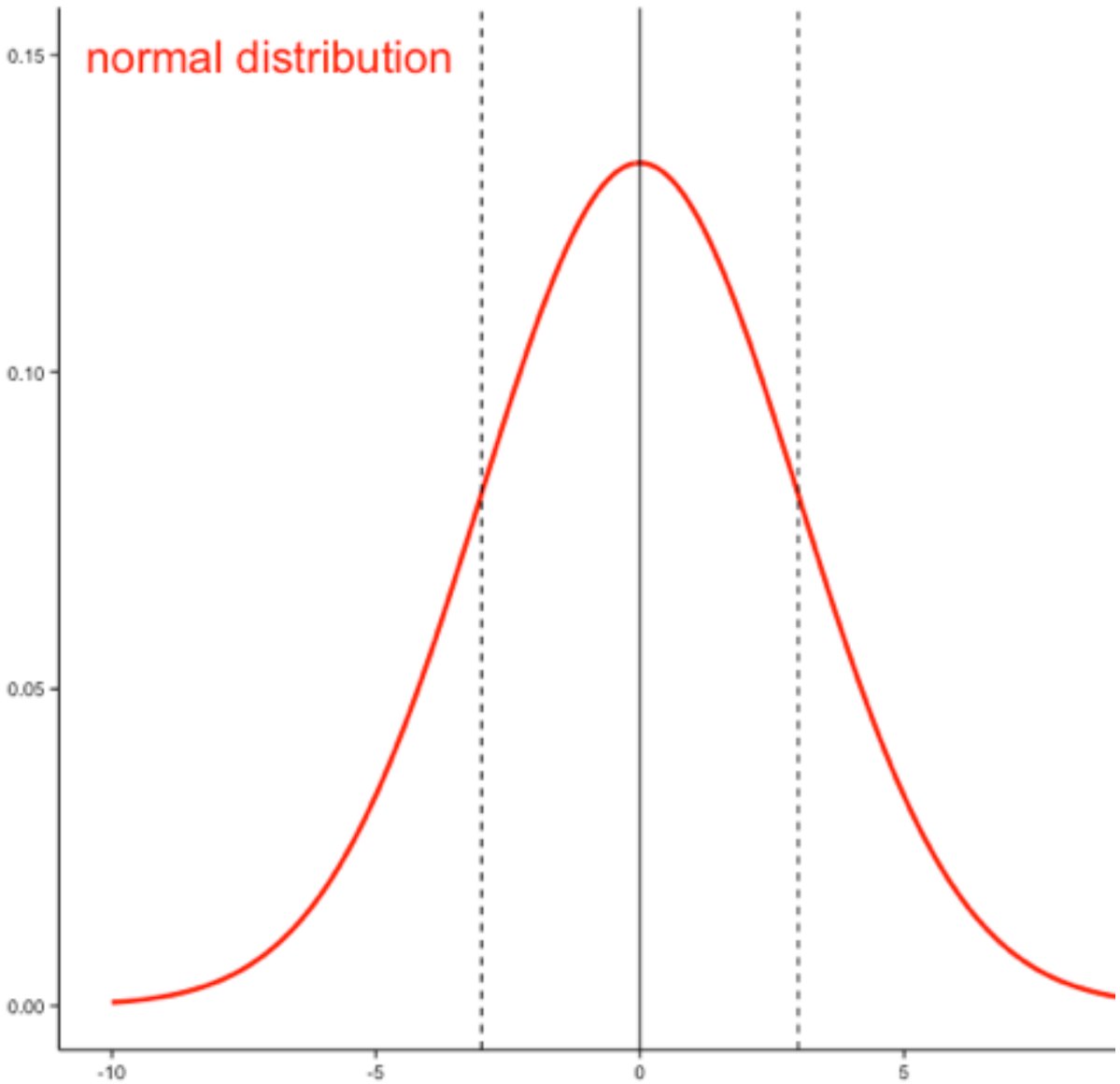
To compute the upper and lower bounds of the 99% CI, use:
Upper: `mean(x) + (2.58 * se(x))`
Lower: `mean(x) - (2.58 * se(x))`

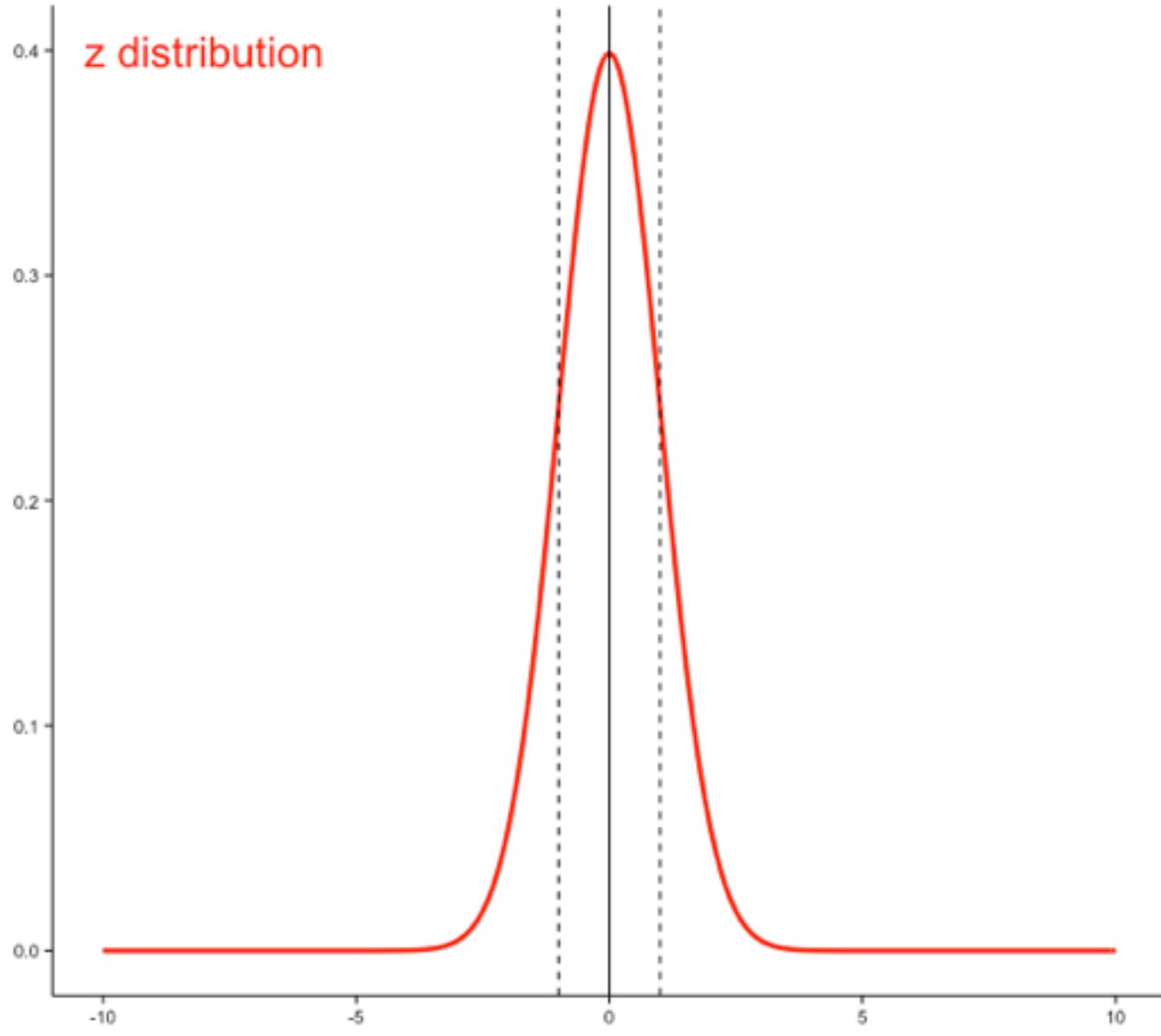
To compute the upper and lower bounds of the 90% CI, use:
Upper: `mean(x) + (1.65 * se(x))`
Lower: `mean(x) - (1.65 * se(x))`



Descriptive data analysis

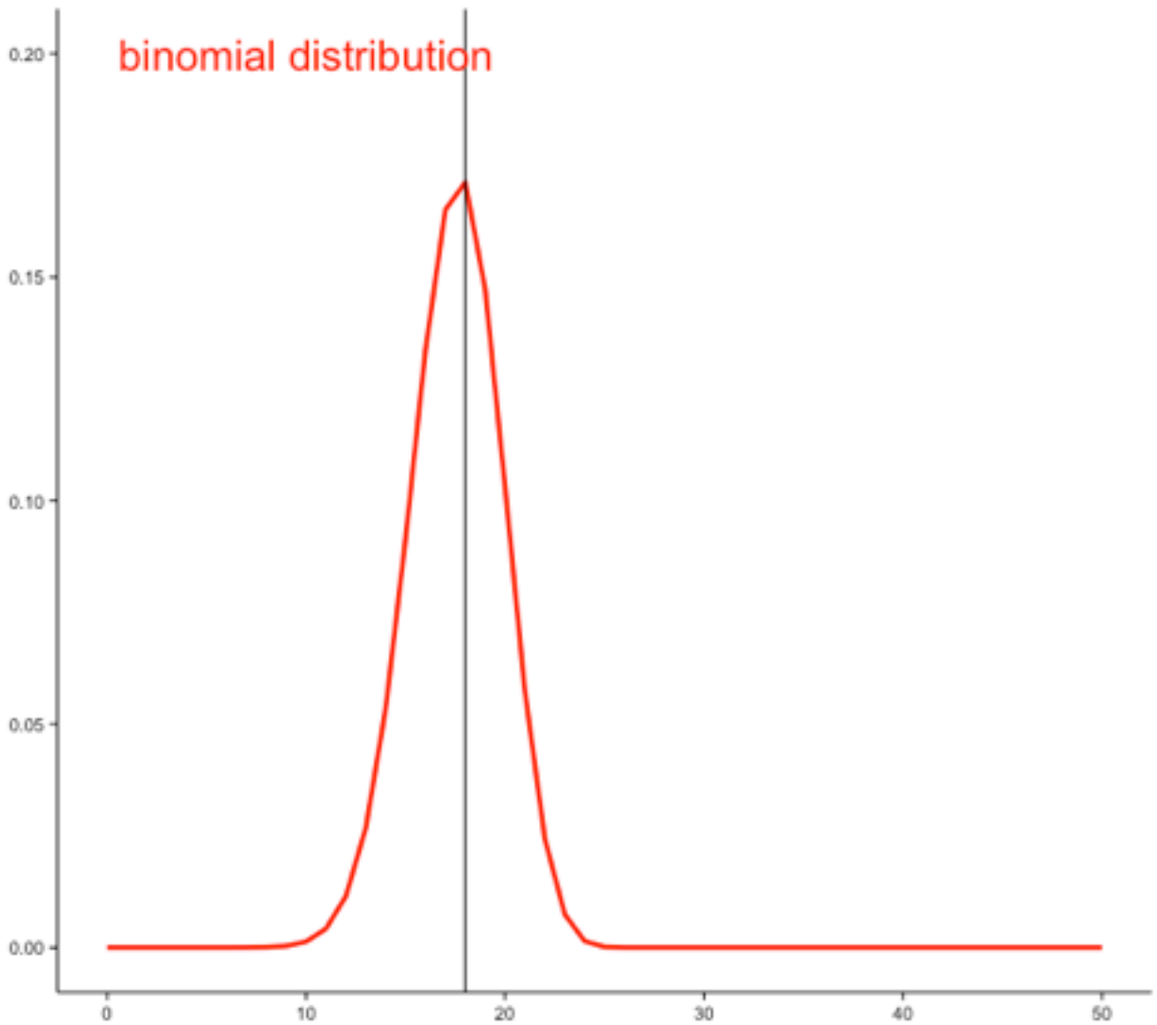
- A note about distributions…

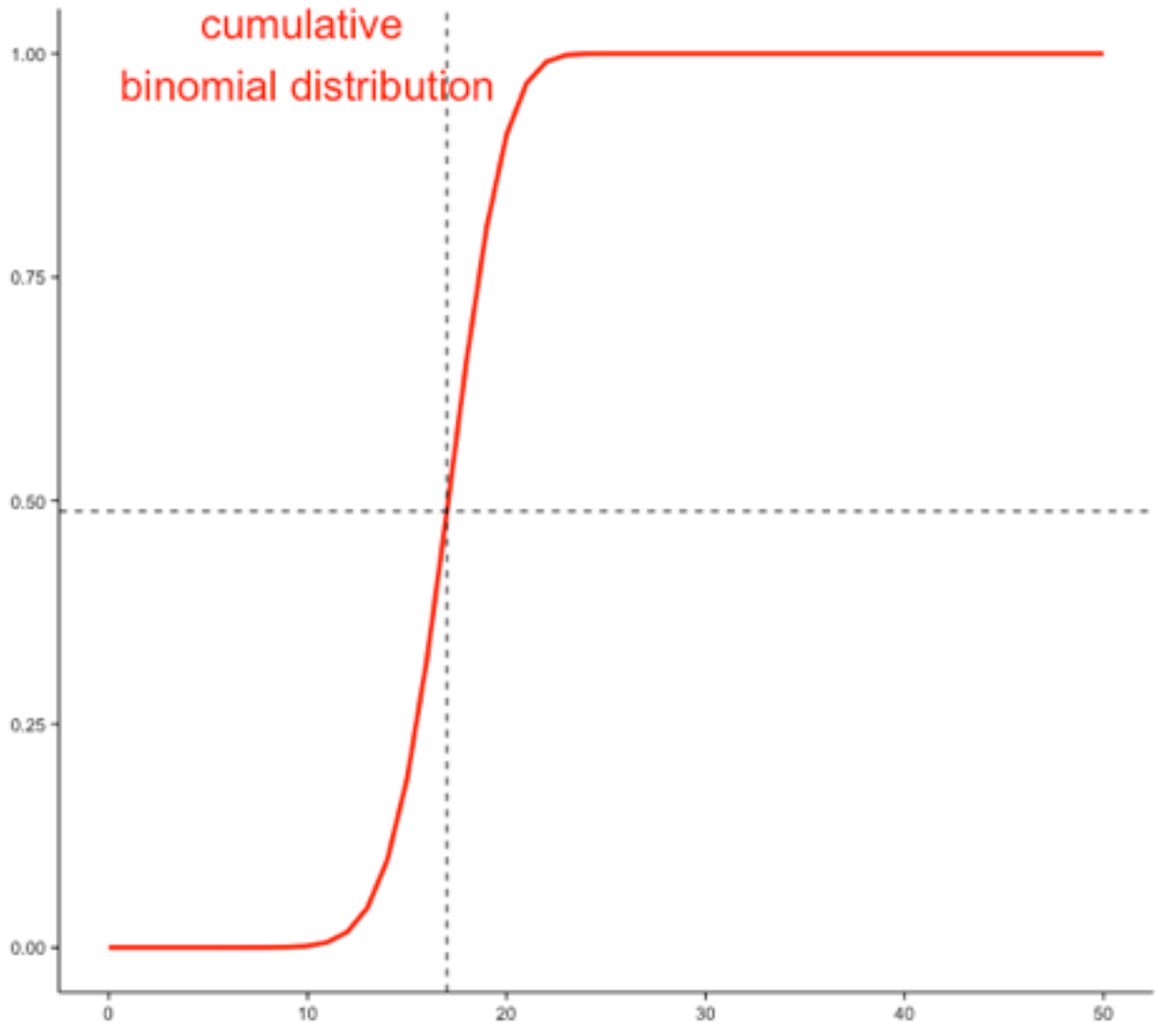




Descriptive data analysis

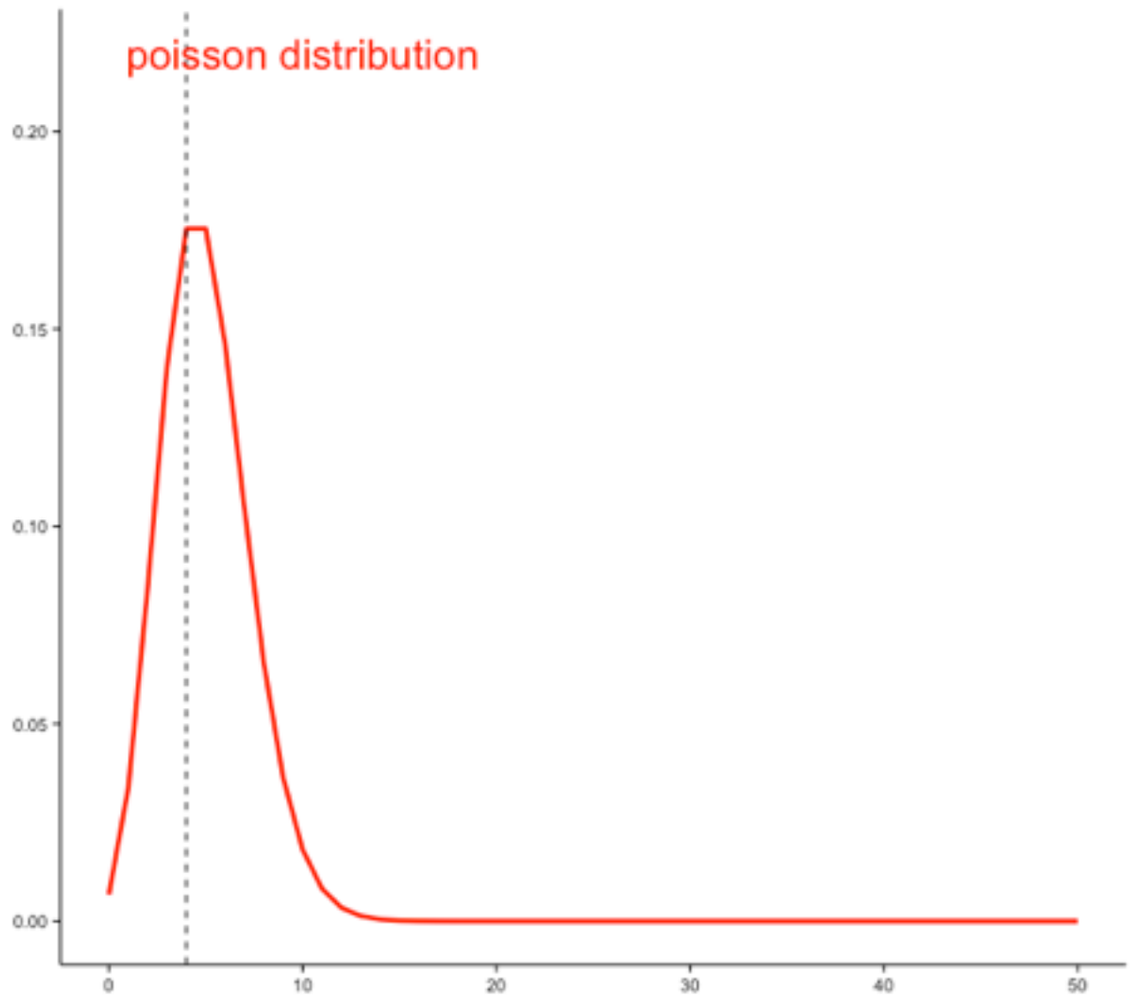
- A note about distributions…





Descriptive data analysis

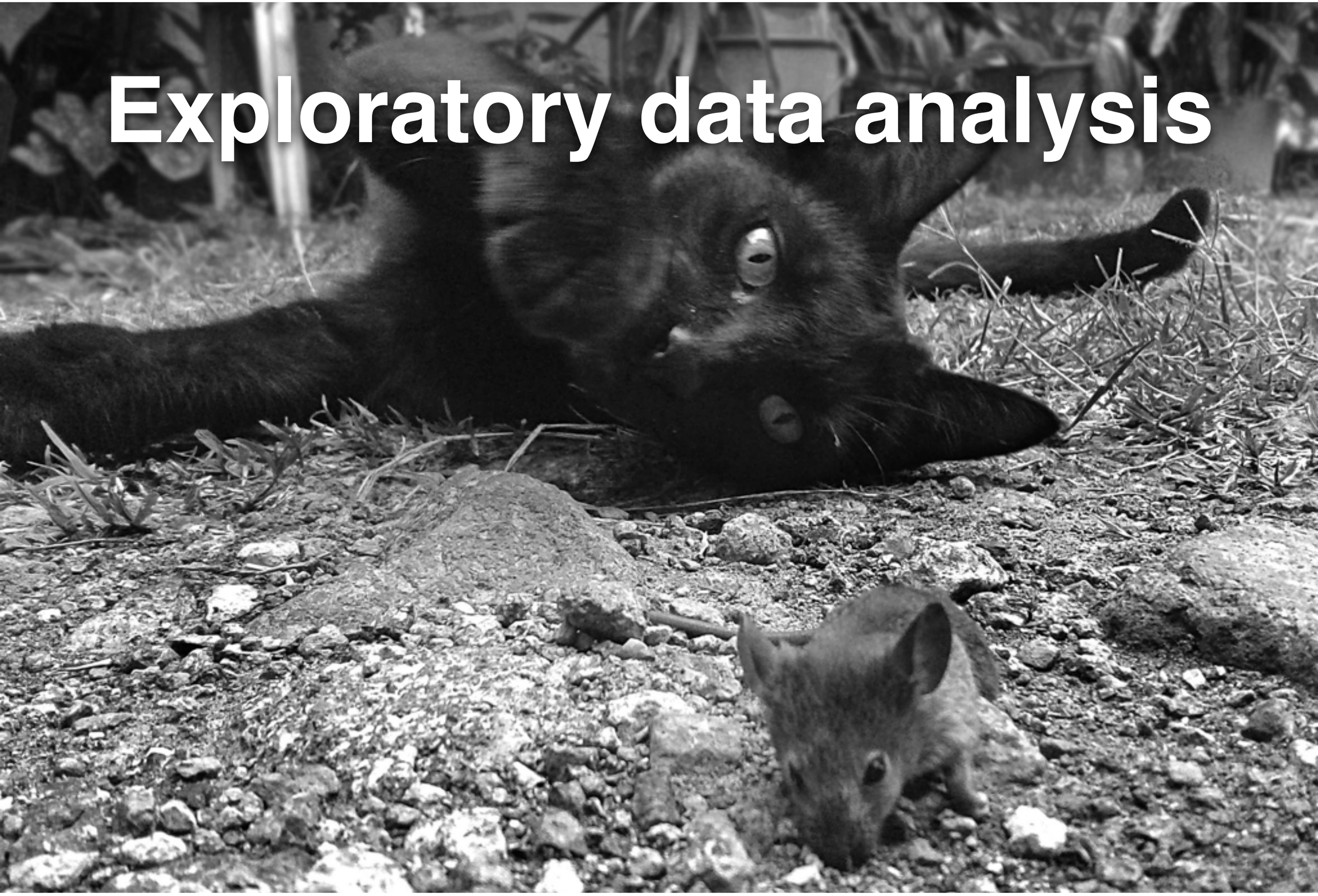
- A note about distributions…



Our data!

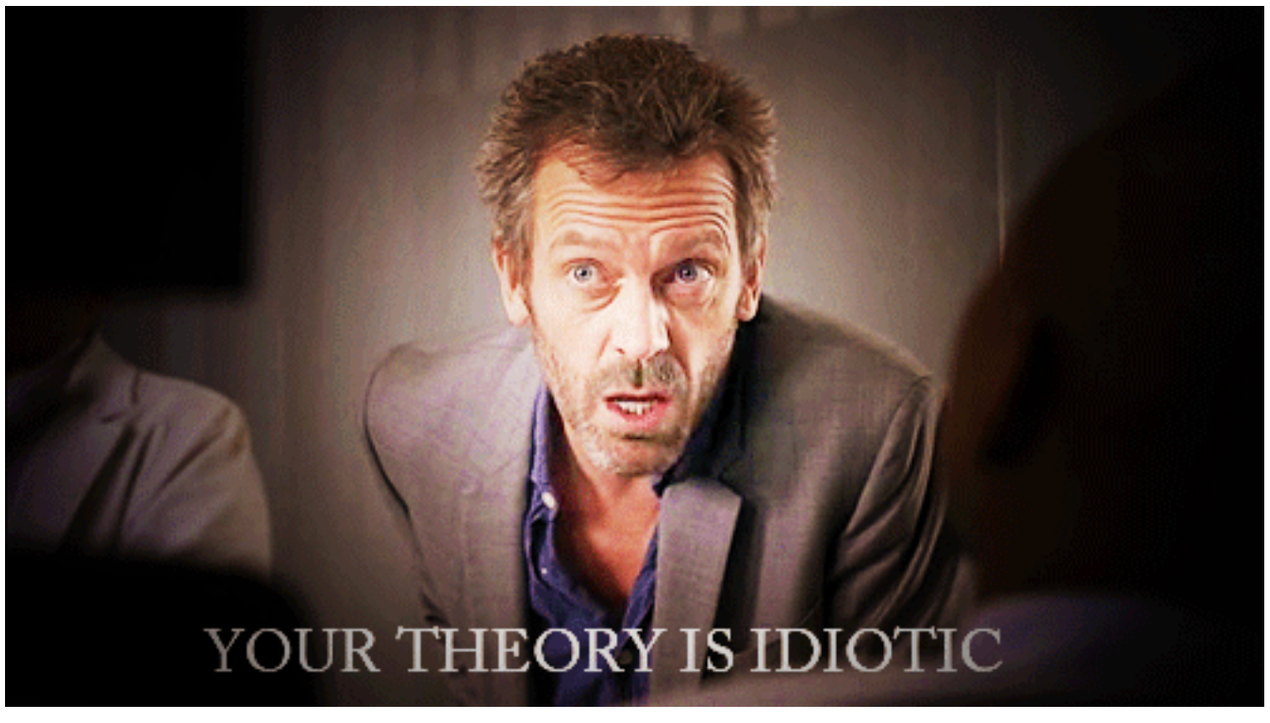
- LAMP stack
  - KittenInc vs DragonInc
  - Users at once, cycle duration, cycles per hour, duration of process A, B, and C

Exploratory data analysis



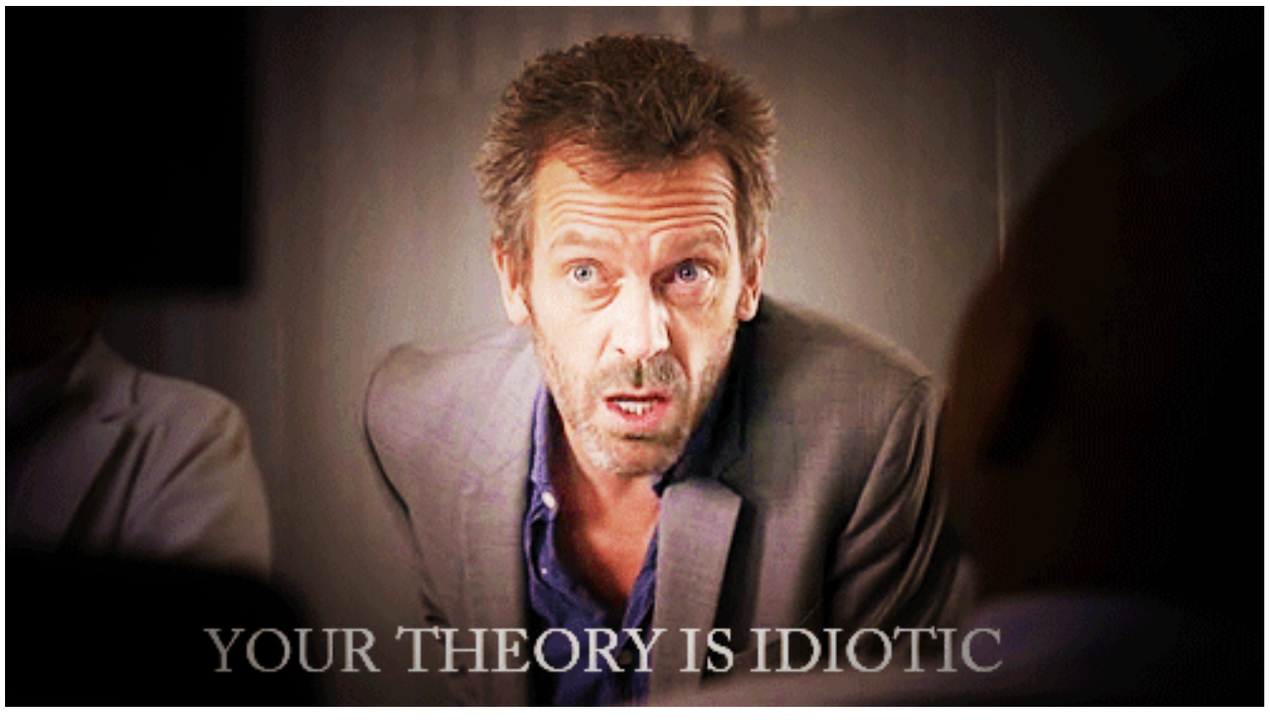
Exploratory data analysis

- Some tips
  - Big data? Random sampling
  - Plots and tables don’t need to be pretty (yet)
  - Split data for exploration and inference/prediction. Don’t re-use data



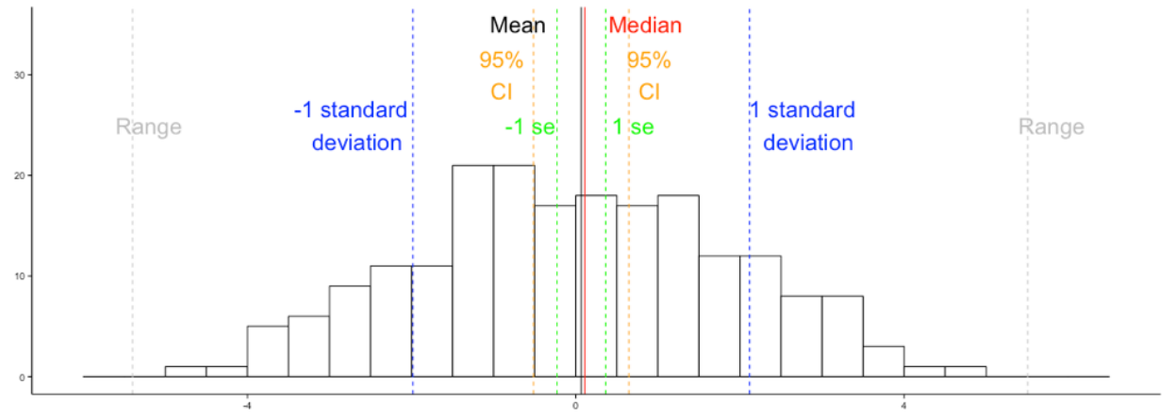
Exploratory data analysis

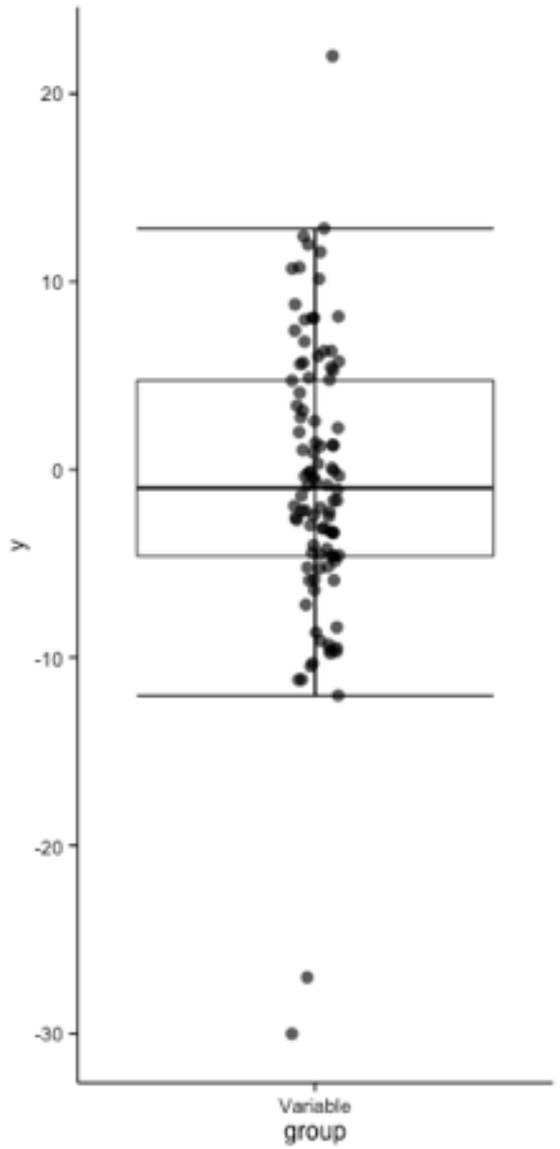
- Some tips
  - Big data? Random sampling
  - Plots and tables don’t need to be pretty (yet)
  - Split data for exploration and inference/prediction. Don’t re-use data

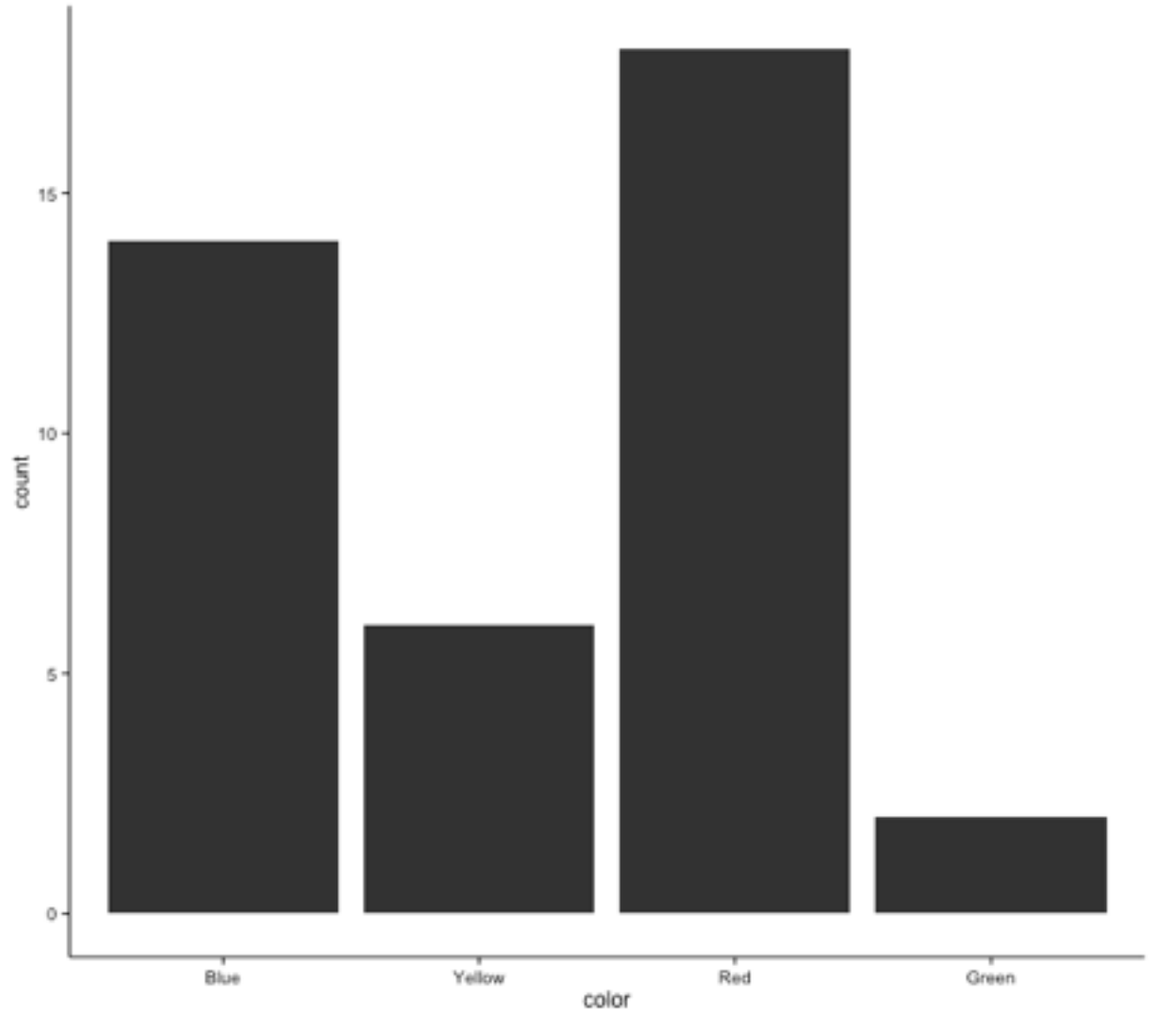


Exploratory data analysis

- **Exploring one variable**







Exploratory data analysis

- **Exploring two or more variables together**
  - Categorical: chi-square test, cross tables

	Red	**Blue**	**Total**
Male	25	15	40
Female	12	28	40
Total	**37**	**43**	**80**

Exploratory data analysis

- **Exploring two or more variables together**
  - Categorical: chi-square test, cross tables

	Red	**Blue**	**Total**
Male	25	15	40
Female	12	28	40
Total	**37**	**43**	**80**

$$\chi^2 = 7.24,\ p = .007,\ df = 1$$

Ex sis

- Ex

Definitions

Pearson's chi-square test a test to see whether there is a relationship between two categorical variables. To compute the chi square of variable x and y use:
`(xy_i - expected_xy_i)`^{2}`/expected_xy_i`, where

`i` you compute this ratio for each cell (`i`) in your cross table. To find the expected cell value, use:

- `(total_x_i * total_y_i) / length(x)`, where

`i` is each level of variable `x` and `y` separately. So, for Sex and Color, this would be:

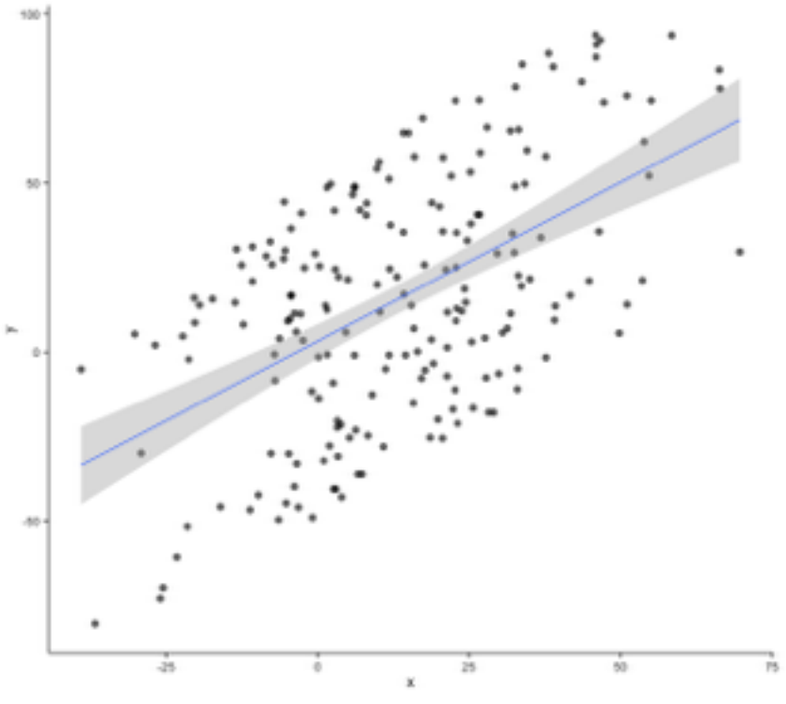
expected_Male_Red = 40 * 37 / 80 = 18.5
expected_Male_Blue = 40 * 43 / 80 = 21.5
expected_Female_Red = 40 * 37 / 80 = 18.5
expected_Female_Blue = 40 * 43 / 80 = 21.5

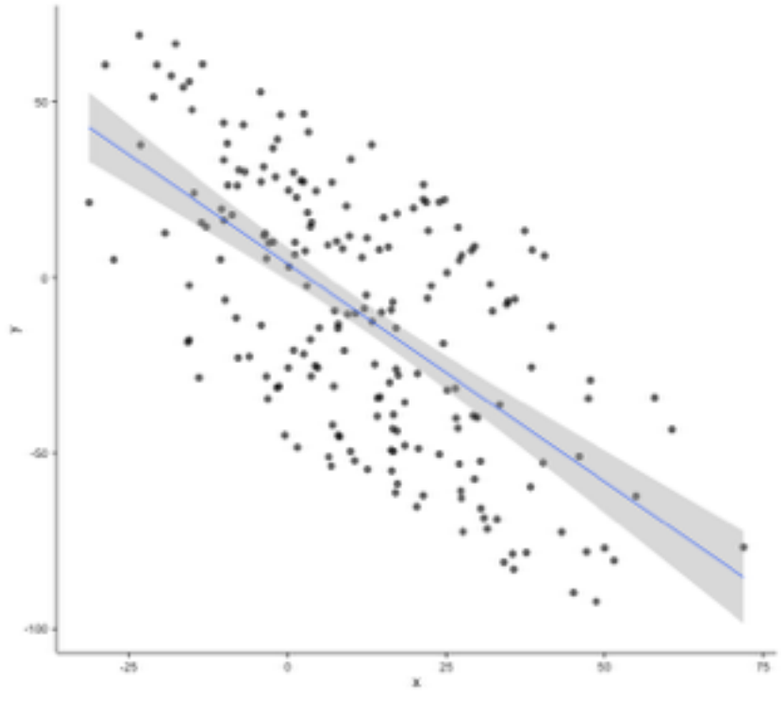
Using the first formula for all 4 cells, we come to the chi-square (χ^2): 7.2407. To know whether this is significant (meaning the dependence between Sex and Color is not zero), we need to know the degrees of freedom. Use this formula: `(levels_x - 1)(levels_y - 1)`

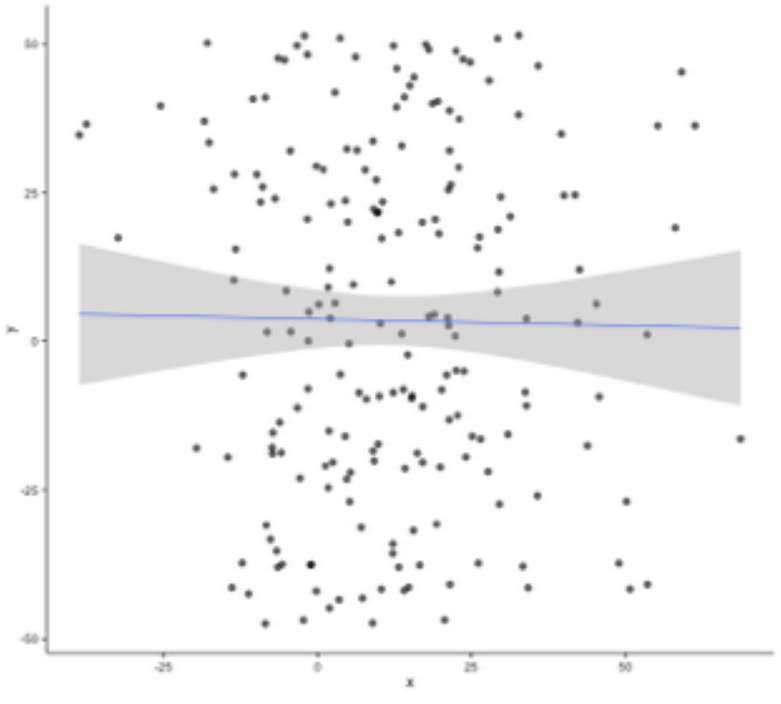
In this case, we have 1 degree of freedom and 7.2407 is a significant value, with a *p*-value of .007

Exploratory data analysis

- **Exploring two or more variables together**
  - Continuous: correlations, scatterplots

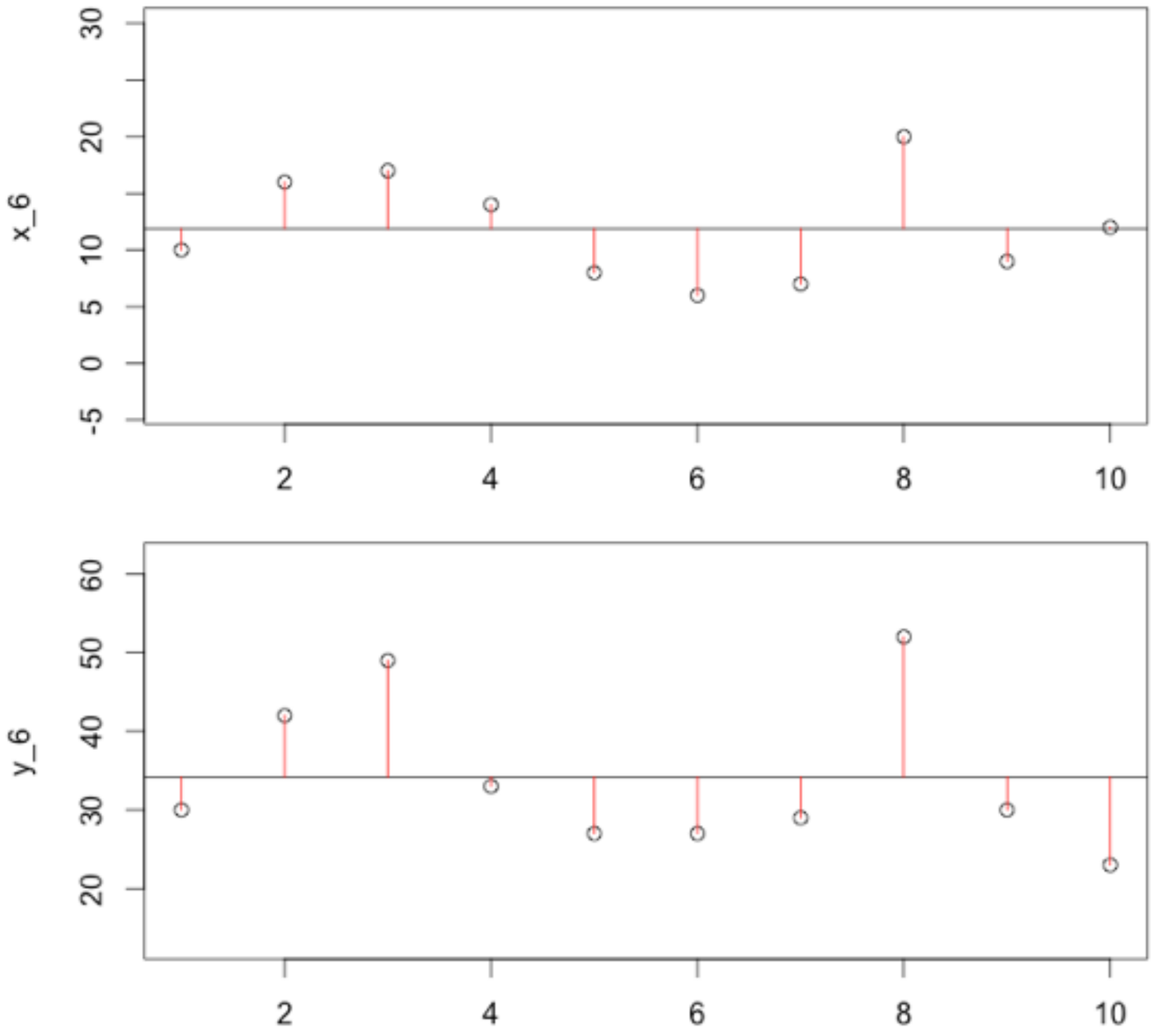






Exploratory data analysis

- Correlations explained



Exploratory data analysis

- Correlations explained
  - We start with covariance

Ex sis

- Co
  -

Definitions

Remember these two concepts?

Sum of squared errors (SS) the sum of the squared deviance of each value from the mean. To find the sum of squared errors of variable x, use:

```
for (i in 1:length(x)) {
  value = (x[i] - mean(x))^2
  SS = SS + value
}
```

Variance (s^2) the average error between the mean and the observations. The sum of squared errors divided the number of values minus 1 . To find the variance of x, use: `SS/(length(x)-1)`

Now for the **covariance (cov(x, y))**…

```
for (i in 1:length(x)) {
  s_x = (x[i] - mean(x))
  s_y = (y[i] - mean(y))
  s_xy = s_xy + (s_x * s_y)
}

s_xy / (length(x)-1)
```

Exploratory data analysis

- Correlations explained
  - We start with covariance
    - Covariance depends on scale of variables
    - To make covariation comparable across variables, we standardise and end up with a correlation!
    - The magnitude of the correlation indicates the strength of the association between two variables:
      - .1 is a weak association,
      - .3 is a medium association,
      - .5 or higher is a strong association

Exploratory data analysis

- Correlations explained
- W[illegible] we

Definitions

Correlation (r) value that indicates the standardised strength of association between two continuous variables. To compute the correlation from the covariance of x and y, use:

```
cov(x, y) / (s(x) * s(y))
```

ngth of the association between two variables:

- .1 is a weak association,
- .3 is a medium association,
- .5 or higher is a strong association

Exploratory data analysis

- Correlations explained
  - Want to test the significance of your correlation? These are the assumptions that you need to meet:
    - Both x and y are continuous
    - x and y are normally distributed (I will show some methods to test for this after the break)

Exploratory data analysis

- Co
  - ese

Definitions

T-statistic to test significance correlation the t-statistic comes from a distribution (much like the z-score comes from the z-distribution). To compute the t-statistic for a certain correlation (*r*), use:

`(r * sqrt(length(x)-2)) / sqrt(1-r²)`

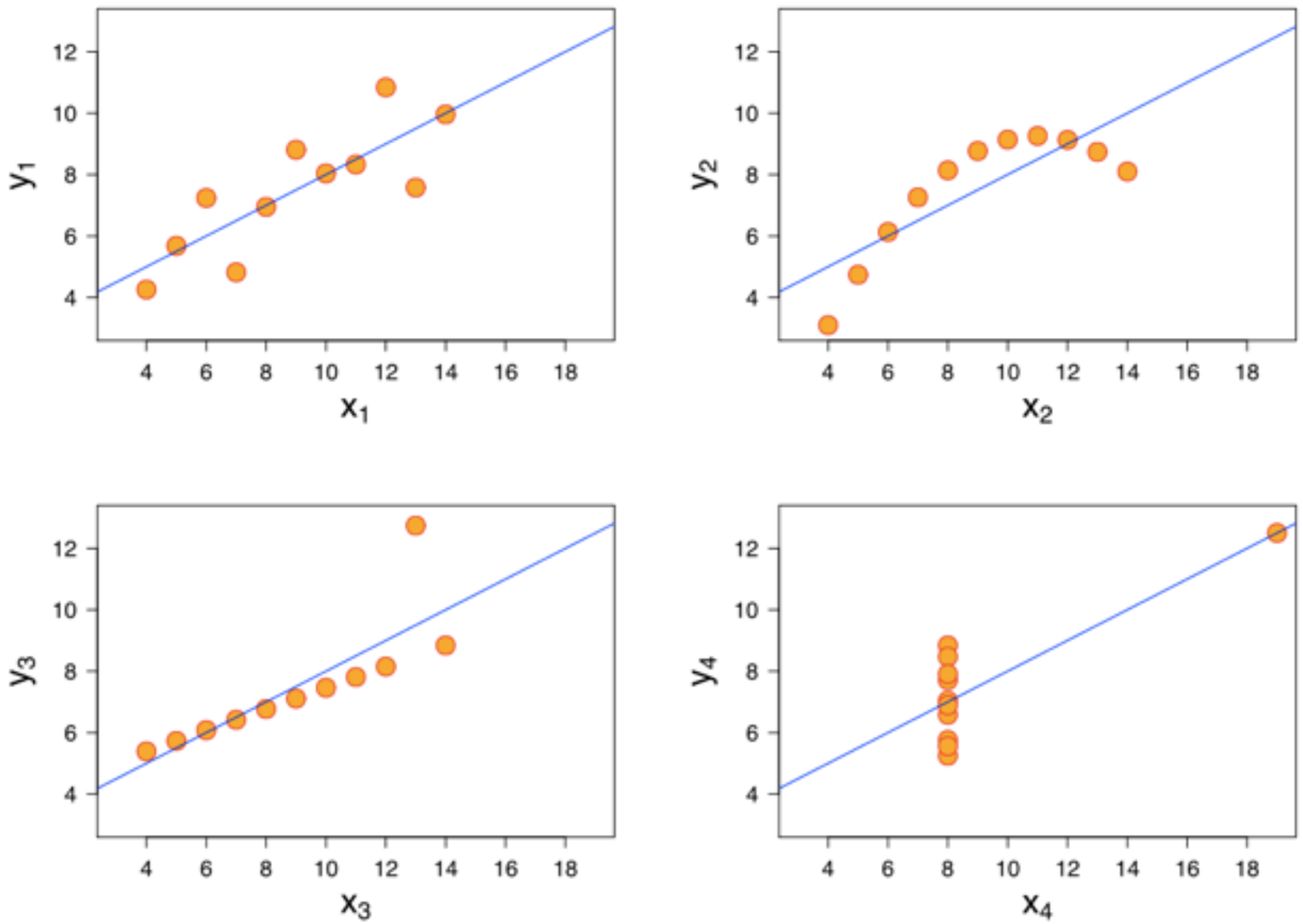
Look the resulting t-value up in a t-statistic table (google!) and find the significance value that goes with `length(x) - 2` degrees of freedom. If it is < .05, the correlation is significantly different from 0.

R² or coefficient of determination another indicator of effect size of a correlation. Indicates the percentage of variation in variable x accounted for by variable y and vice versa. To compute, simply:

`r² * 100`

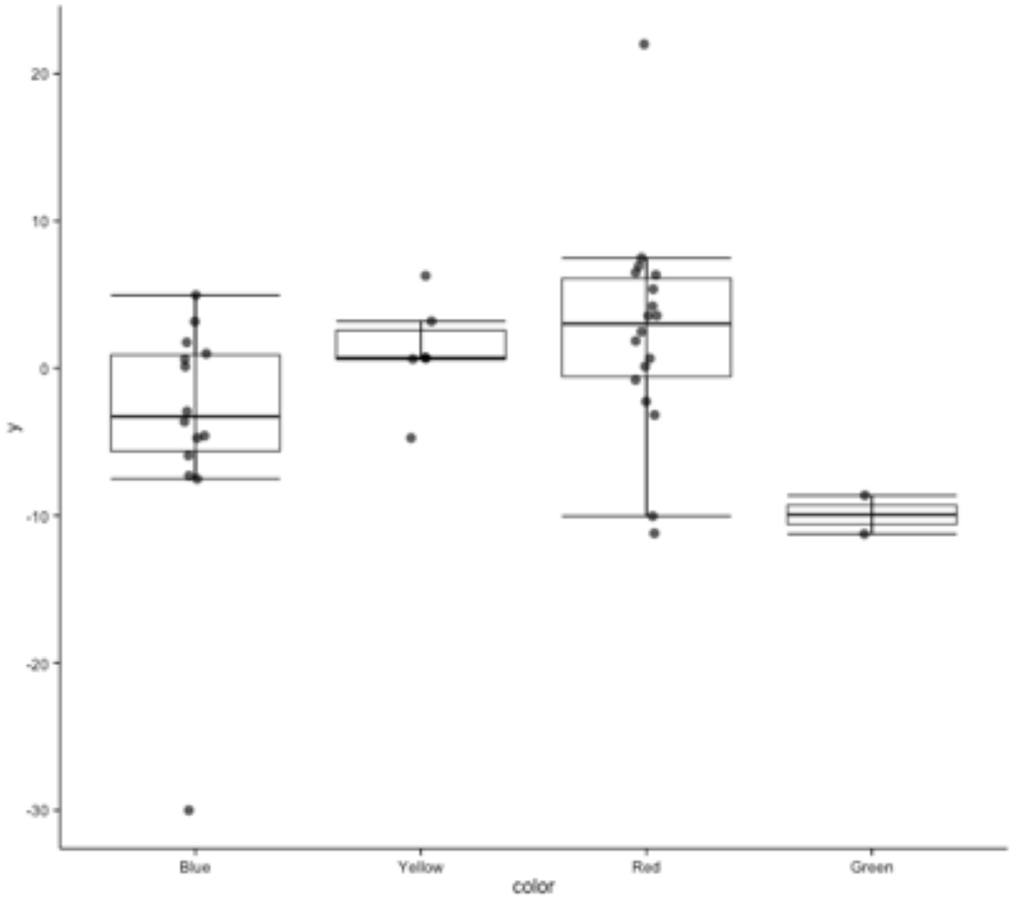
Exploratory data analysis

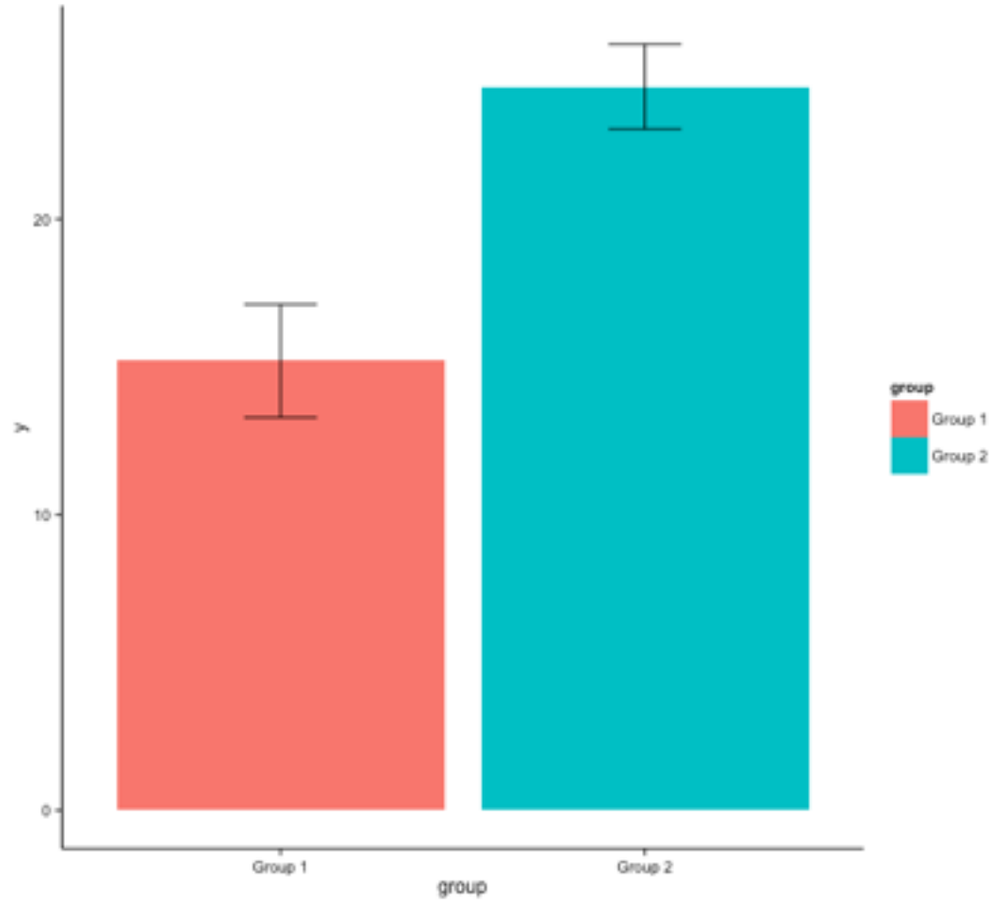
- Exploring two or more variables together
  - **Why are scatterplots important?**



Exploratory data analysis

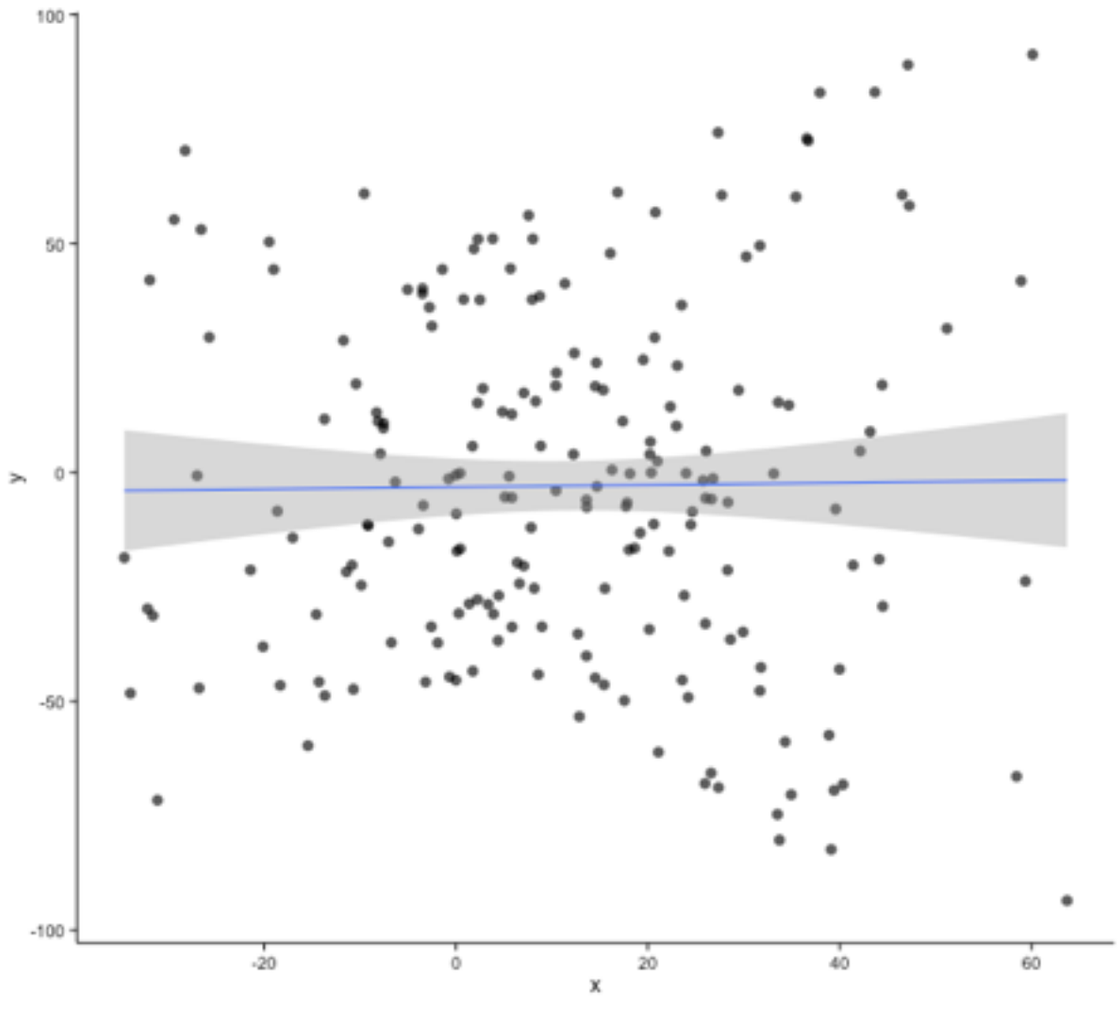
- **Exploring two or more variables together**
  - Categorical and continuous

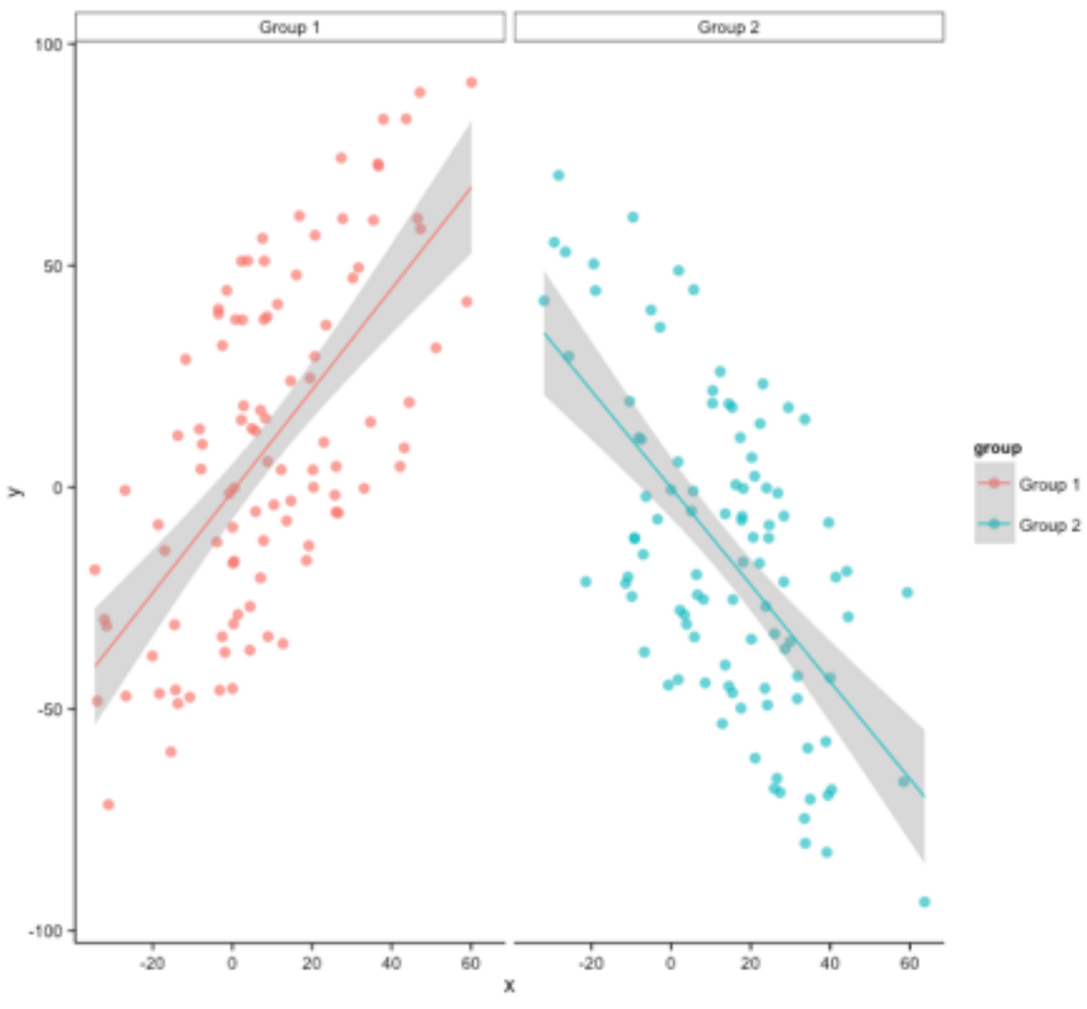




Exploratory data analysis

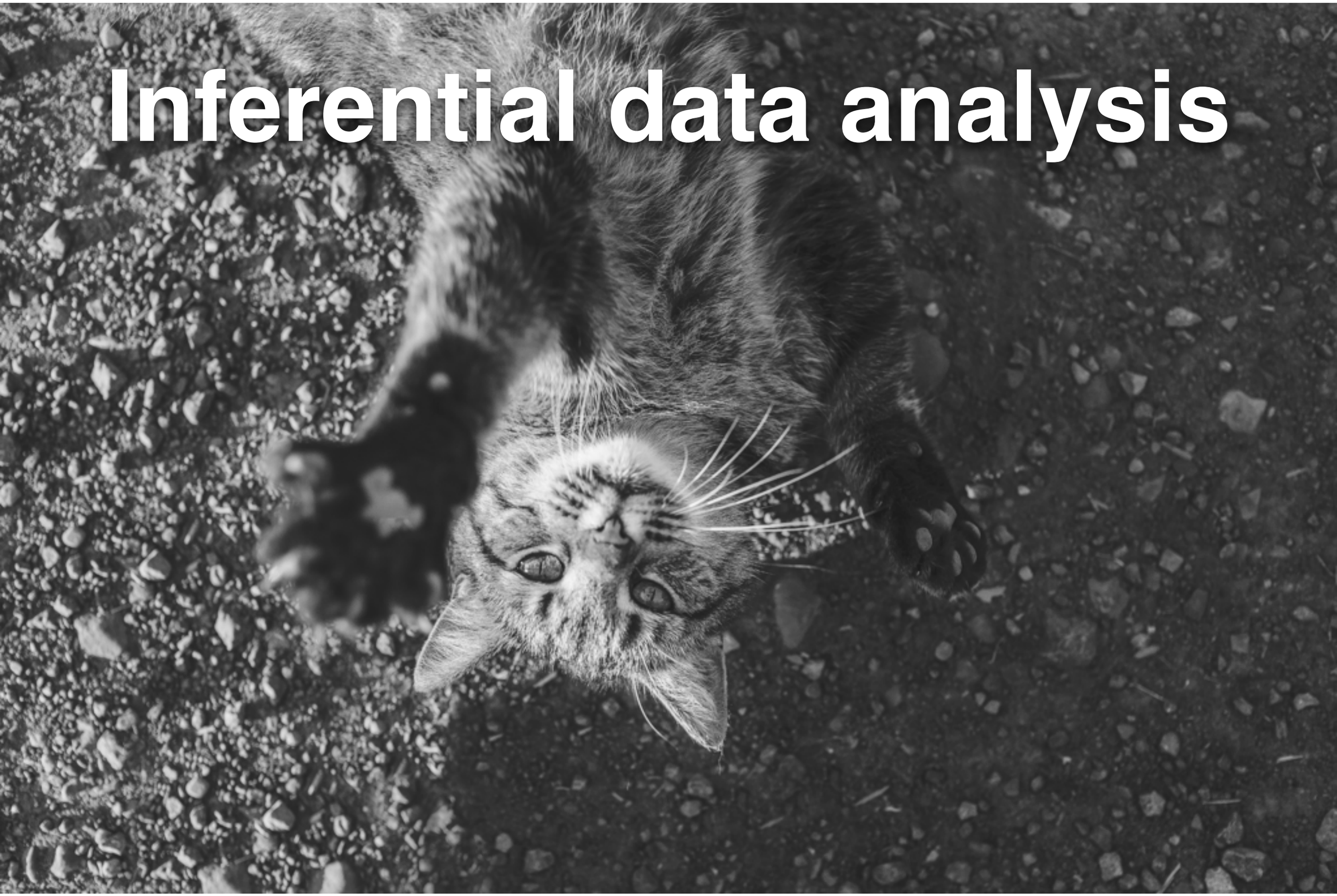
- **Exploring two or more variables together**
  - Categorical and continuous





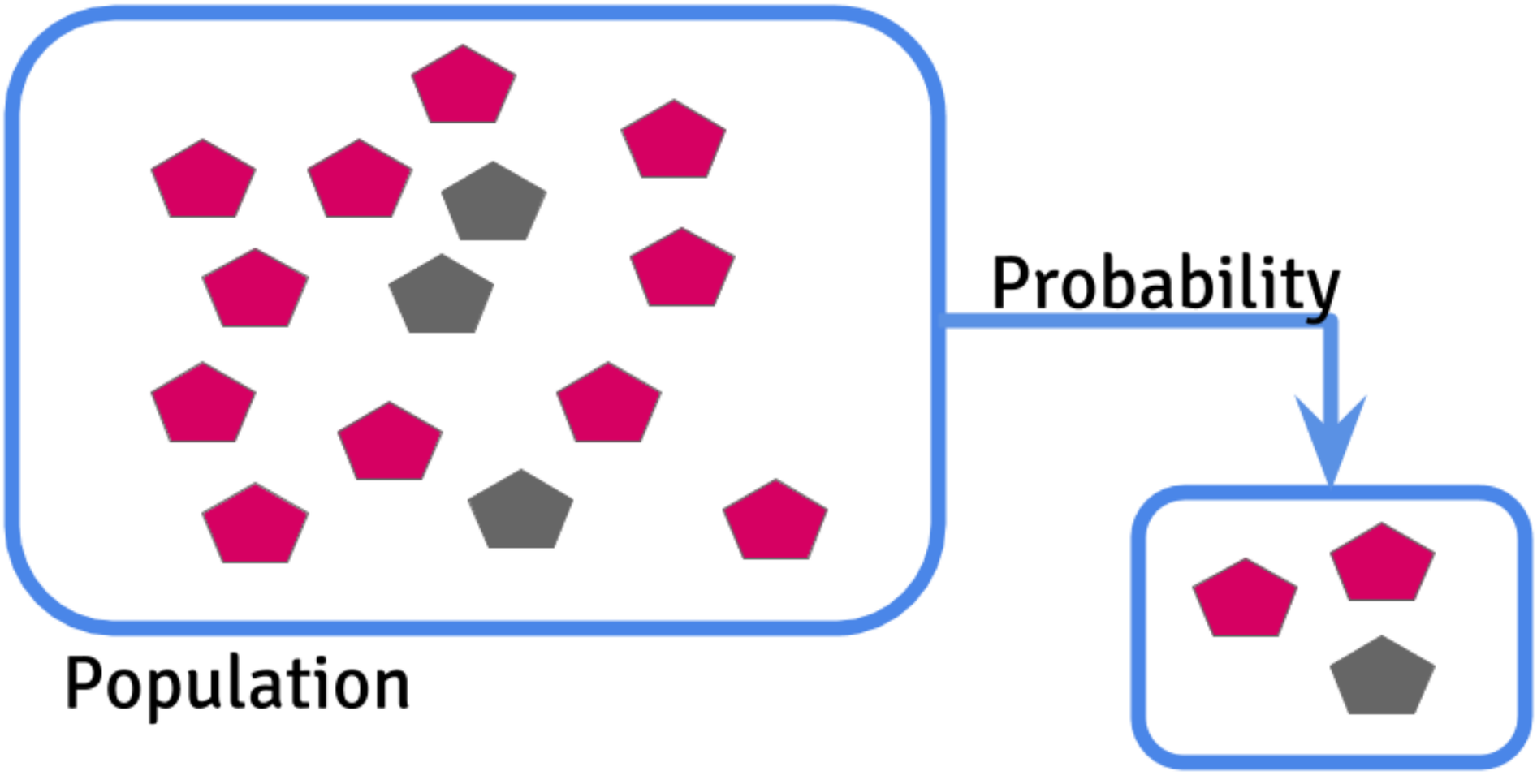

Break…

Inferential data analysis



Inferential data analysis

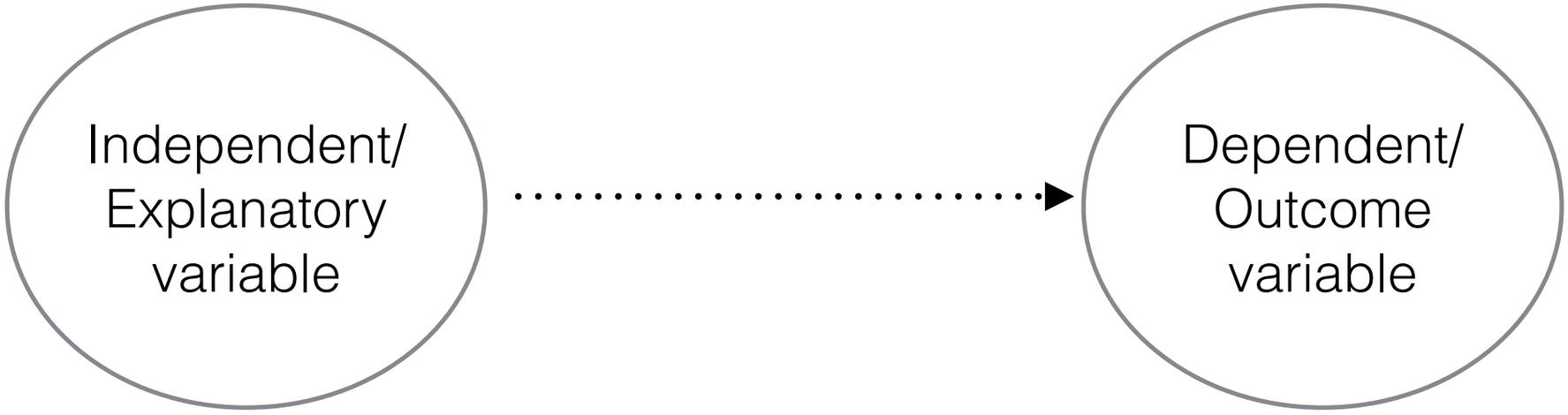
- A note about populations and samples



Source: Leek, J. (2015). *The elements of data analytic style.*

Inferential data analysis

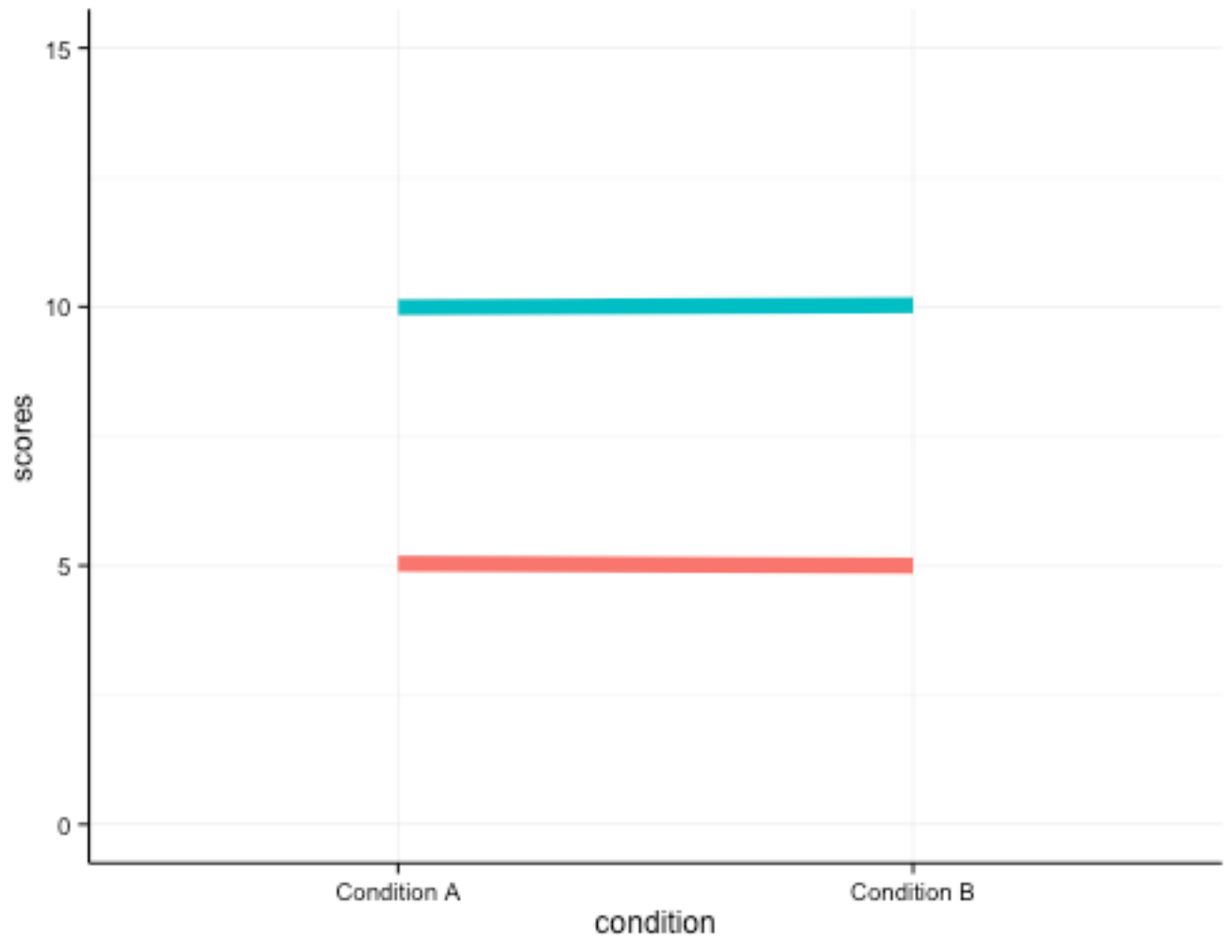
- A note about types of variables in an analysis



Cause **Effect**

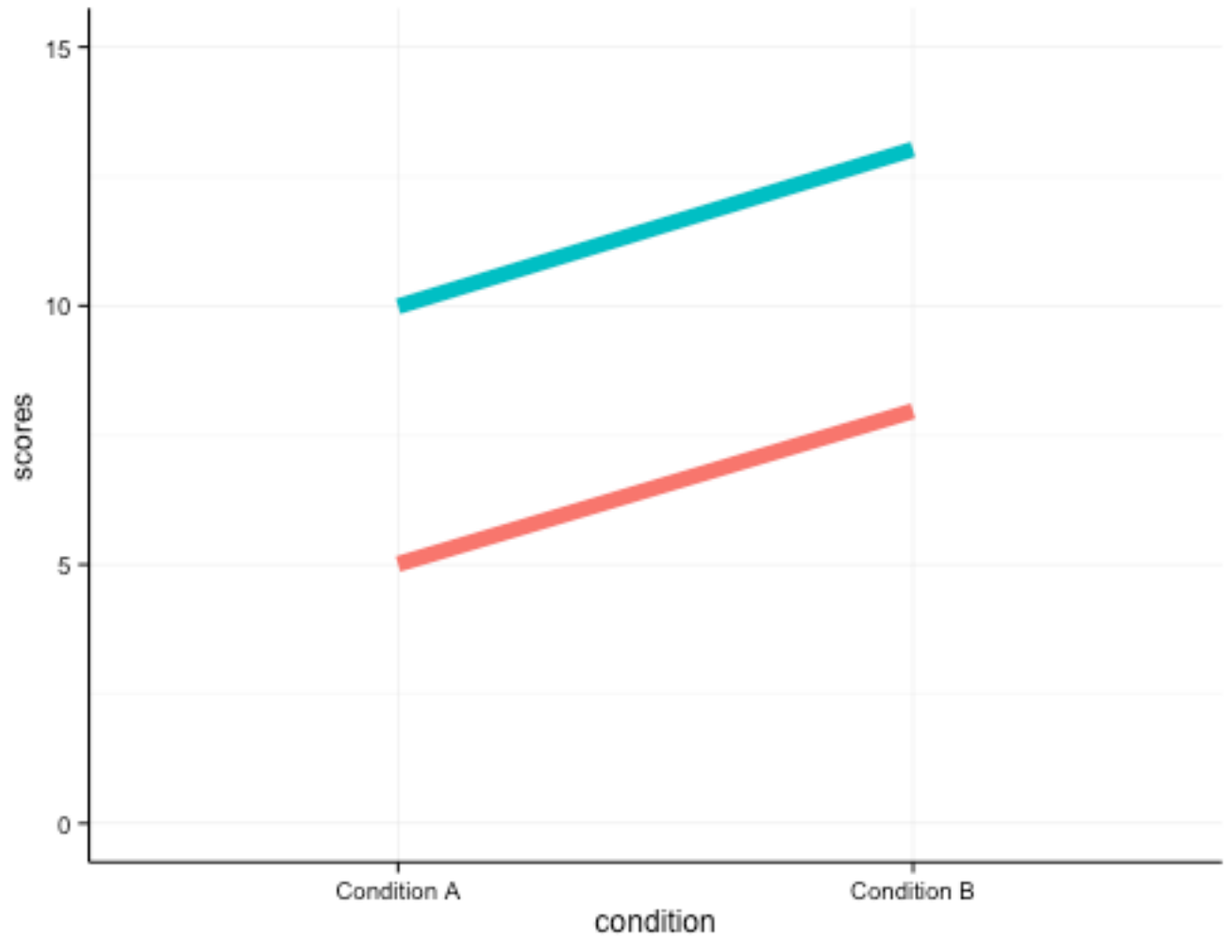
Inferential data analysis

- Main effects, interaction effects



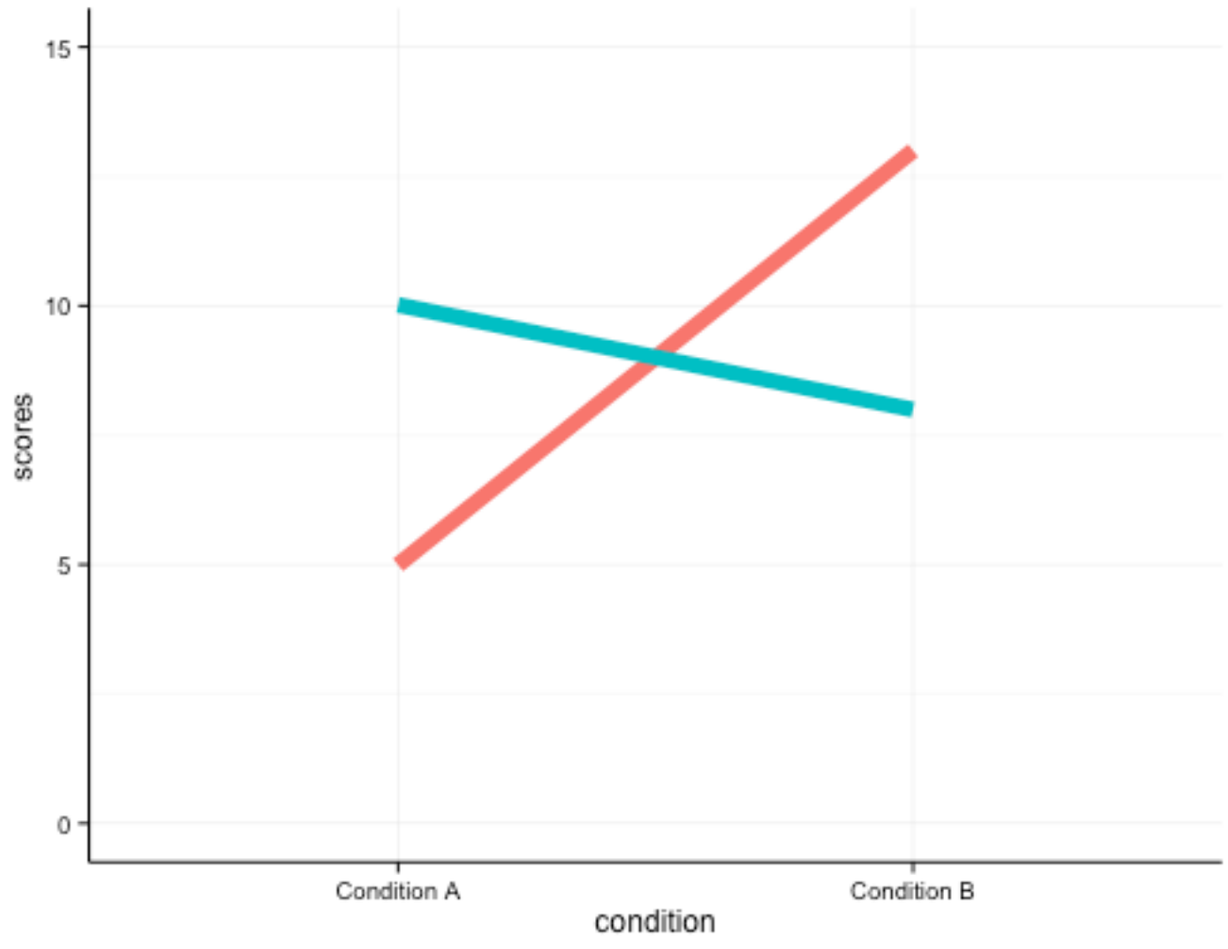
Inferential data analysis

- Main effects, interaction effects



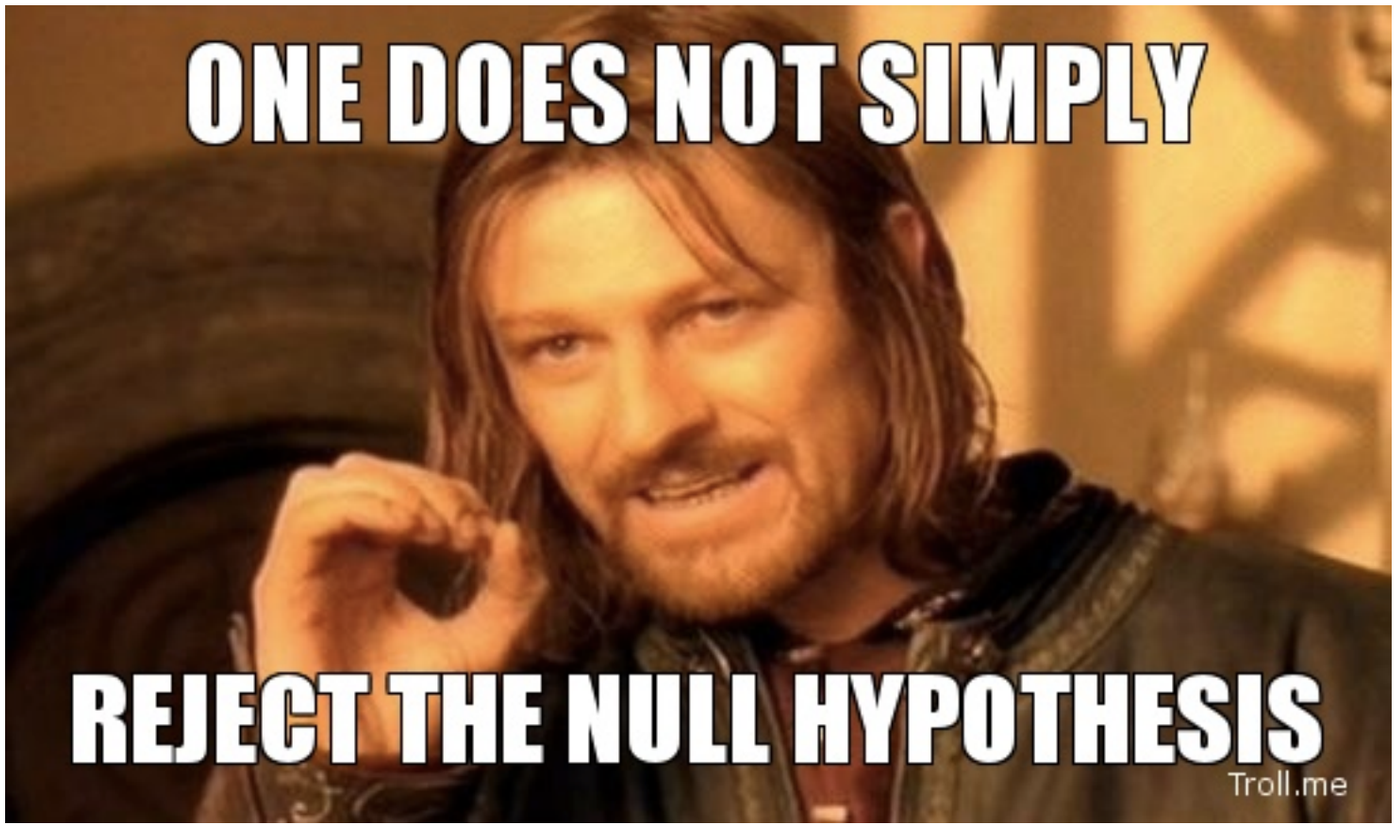
Inferential data analysis

- Main effects, interaction effects



Inferential data analysis

- Hypothesis testing



Inferential data analysis

- Hypothesis testing
  - Null hypothesis: value 1 == value 2 or 0
  - Alternative hypothesis: value 1 ≠ value 2 or 0

- Alpha: the probability of rejecting the null hypothesis when the null hypothesis (i.e. there is no effect or difference) is true
  - Alpha = .05 means that there is a 5% chance of thinking you found a result, when in fact there is none in the population

Inferential data analysis

- **Power**
  - The probability of rejecting the null hypothesis when it is false
  - Power is higher when…
    - Sample is larger
    - Effect is bigger

		Truth	
		Null hypothesis is true	**Null hypothesis is false**
Your data	**Null hypothesis is true**	GOOD (1 - α)	Type II Error (β)
	Null hypothesis is false	Type I Error (α)	GOOD (1 - β)

Inferential data analysis

- **Bootstrapping**
  - Helps you estimate SEs and 95% CIs for statistics that do not have a clear distribution
  - Use bootstrapped confidence intervals for a statistic when your data does not meet the distribution assumptions of your statistical test.

In is

- Bo
  -
  -

st.

Definitions

Bootstrapping technique used to created distributions of population statistics for which you don't know the distribution. Follow these steps:

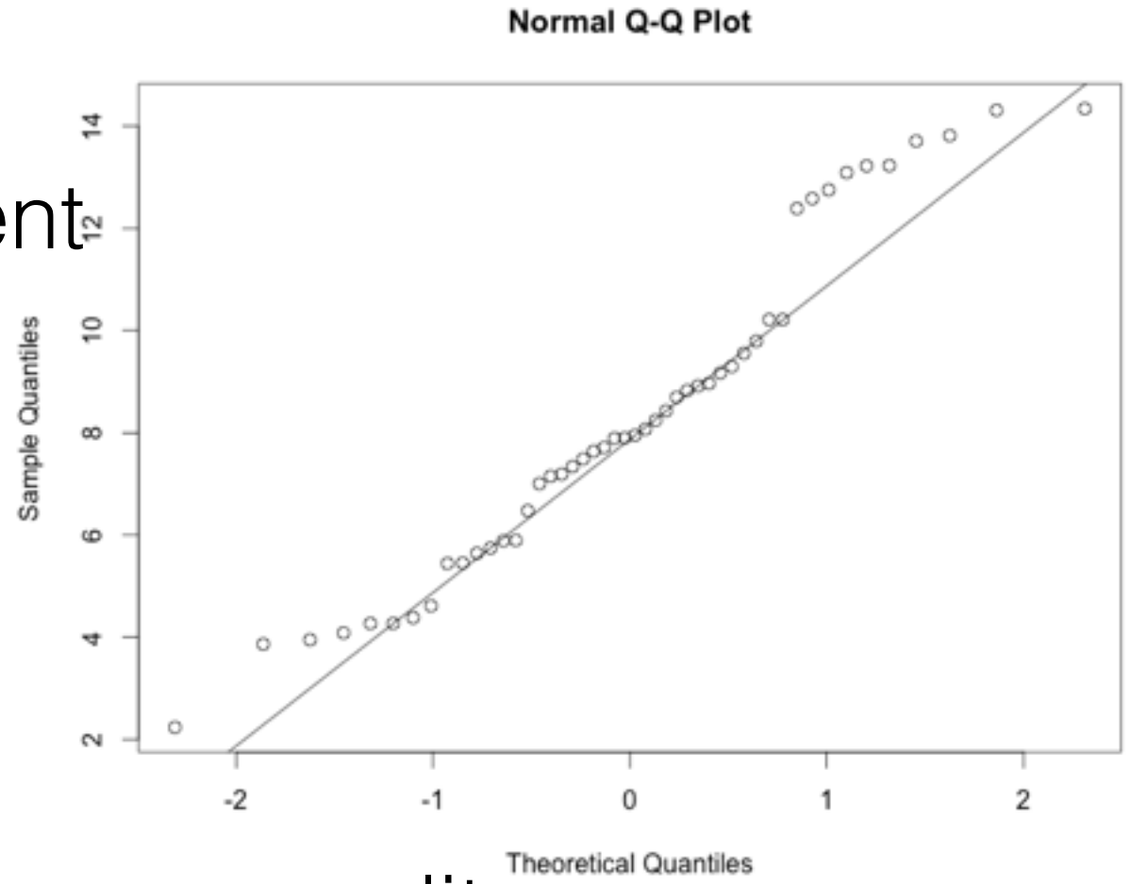
1. Sample from your variable x a sample with length(x). To make sure you don't end up with the exact same x, use REPLACEMENT
2. Compute the statistic of interest (for example the median) for the new sample
3. Repeat step 1 and 2 B times, resulting in B statistics (e.g. medians). Make sure B is very large (e.g. 10,000).

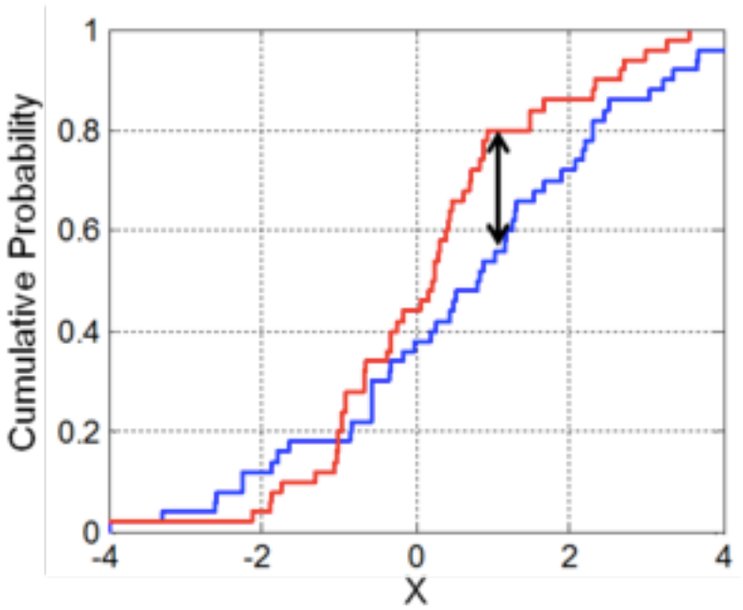
To use bootstrapping to find the population distribution of the median of variable x, use:

```
B <- 10000
resamples <- matrix(sample(x, n*B, replace = TRUE), B,
n)
medians <- apply(resamples, 1, median)
```

Inferential data analysis

- **Common assumptions**
  - scores are independent
  - normality
    - Q-Q plot
    - Kolmogorov-Smirnov test

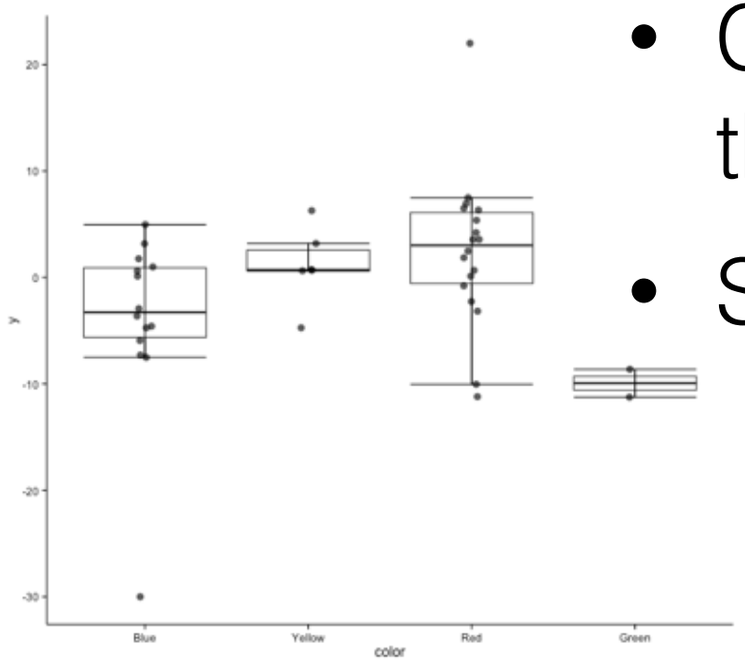




Significant == non-normality

Inferential data analysis

- **Common assumptions**
    - variances are roughly equal across groups/ levels (homogeneity of variances)
    - Boxplots
    - Levene's test
        - Compares the absolute deviations from the mean per group
        - Significant == heterogeneity



Inferential data analysis

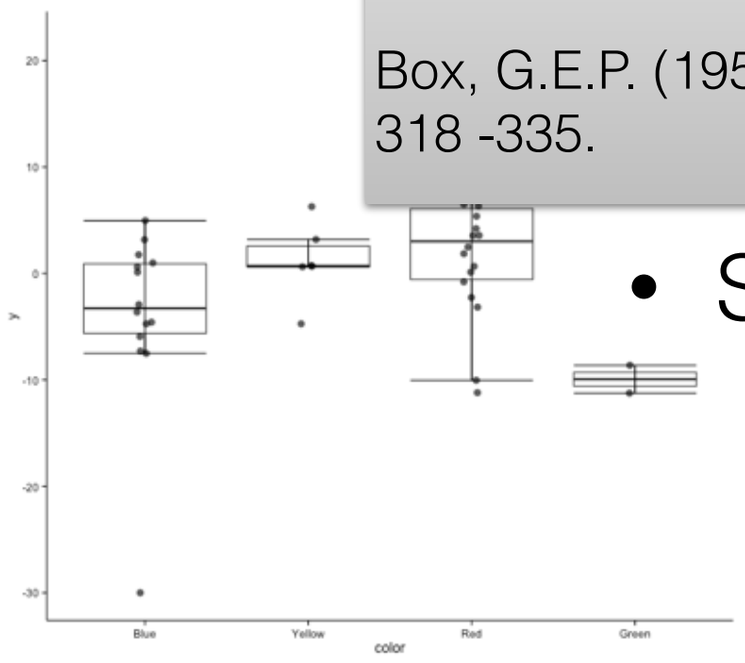
- **Common assumptions**
    - variances are roughly equal across groups/ levels (homogeneity of variances)

George Box

To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!

Box, G.E.P. (1953), "Non-normality and Tests on Variance," Biometrika, 40, 318 -335.

om



- Significant == heterogeneity

Inferential data analysis

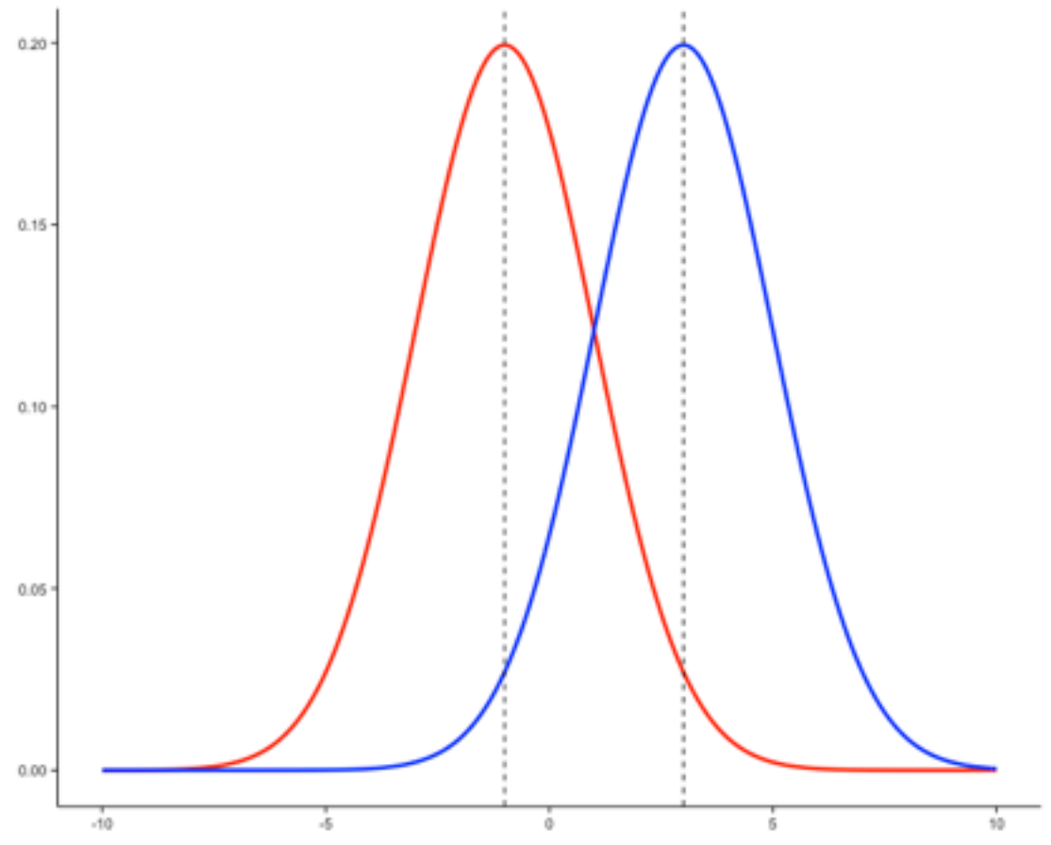
- **Common assumptions**
    - What if assumptions aren't met?
        - Some tests are robust against deviations from normality and homogeneity
        - Pick an alternative test that is robust
        - Use bootstrapping
        - Transform your data (makes interpretation more difficult)

Inferential data analysis

- **Comparing groups**
  - t-test (comparing two groups)
  - William Sealy Gosset (1876-1937)
    - Worked for Guinness
- Assumptions:
  - normality
  - independence
  - continuous outcome variable
  - homogeneity of variances

Inferential data analysis

- **Comparing groups**
  - t-test (comparing two means)
    - One sample t-test
    - Independent samples t-test
    - Paired samples t-test



In is

- C

Definitions

One sample t-test test to compare the mean of one group to a hypothetical population mean (often 0). To perform a one sample t-test on the values of group x compared to 0, use:

```
t_value = (mean(x) - 0) / (s(x) / sqrt(length(x))
```

- **Independent samples t-test** test to compare the means of two samples, to see if they come from different populations. To perform a t-test on the values of group x and y, use:

```
df = length(x) + length(y) - 2
pooled_var = ((s(x) * length(x) - 1) + (s(y) * length(y) - 1)) / df
se_diff = sqrt(pooled_var/length(x) + pooled_var/length(y))
t_value = (mean(x) - mean(y)) / se_diff
```

Paired samples t-test test to compare the means of two measurements of the same sample. To perform a t-test on the values of measurement x_1 and x_2, use:

```
diff = mean(x_1) - mean(x_2)
se = s(diff) / sqrt(length(diff))
t_value = (diff - 0) / se
```

Inferential data analysis

- **Comparing groups**
  - t-test (comparing two means)
    - Example output

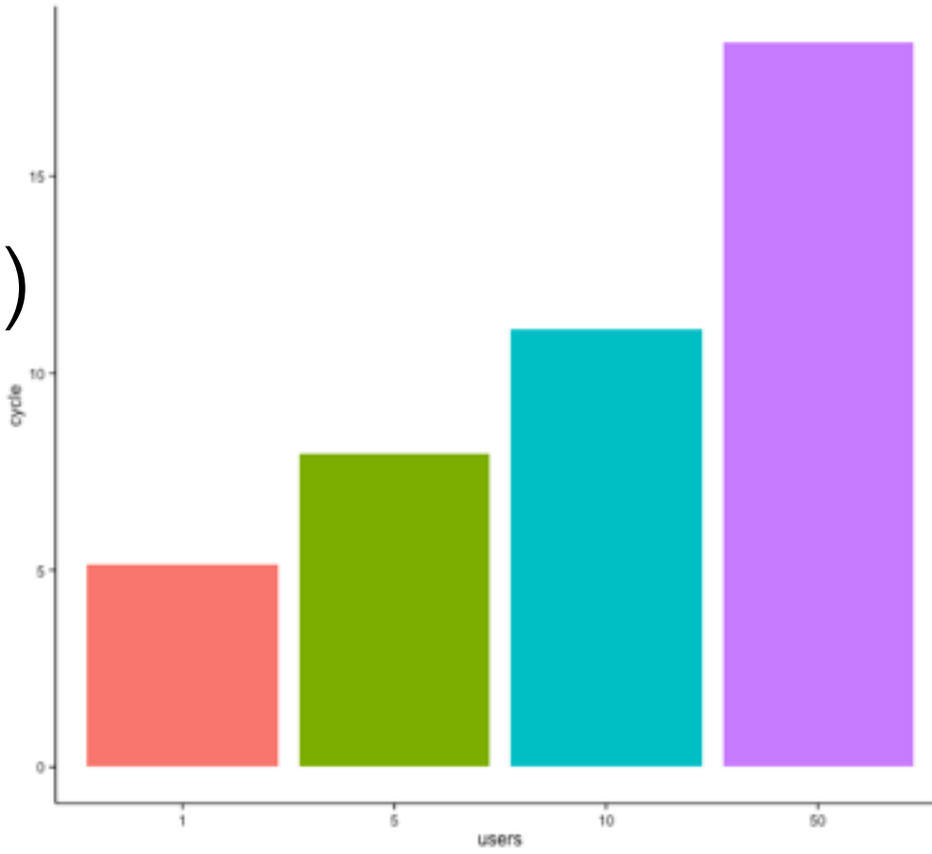
```
        Welch Two Sample t-test

data:  cycle by system
t = 5.0895, df = 75.028, p-value = 2.579e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.168084 7.243232
sample estimates:
      mean in group BazingaShopping mean in group Rainbow Dash Shopping
                           13.522449                           8.316791
```

Inferential data analysis

- **Comparing groups**
  - ANOVA (comparing > 2 groups)
  - Ronald Fisher (1890-1962)
  - Assumptions
    - normality (within groups)
    - independence
    - continuous outcome variable
    - homogeneity of variances

Inferential data analysis

- **Comparing groups**
  - ANOVA (comparing > 2 groups)
    - Basic idea
      - Comparing variability within groups to variability between groups



Definitions

One-way ANOVA test to compare the means of more than two groups.
To perform a one-way ANOVA on the values of group x, y, and z, use:

```
sum_value = sum(x,y,z)
n = length(x) + length(y) + length(z)
mean_overall = sum_value / n
mean_x = mean(x)
mean_y = mean(y)
mean_z = mean(z)

sst = sum((c(x,y,z) - mean_overall)^2)
ssb = length(x)*(mean_x - mean_overall)^2 +
      length(y)*(mean_y - mean_overall)^2 +
      length(z)*(mean_z - mean_overall)^2
ssr = sum(c((x - mean(x))^2 +
            (y - mean(y))^2 +
            (z - mean(z))^2))

df_total = n - 1
df_model = groups - 1
df_residual = df_total - df_model

MS_m = ssb / df_model
MS_r = ssr / df_residual

F = MS_m / MS_r
```

Inferential data analysis

- **Comparing groups**
  - ANOVA (comparing > 2 groups)
    - Other forms
      - Factorial ANOVA
      - ANCOVA
      - Repeated measures ANOVA
      - MAN(C)OVA

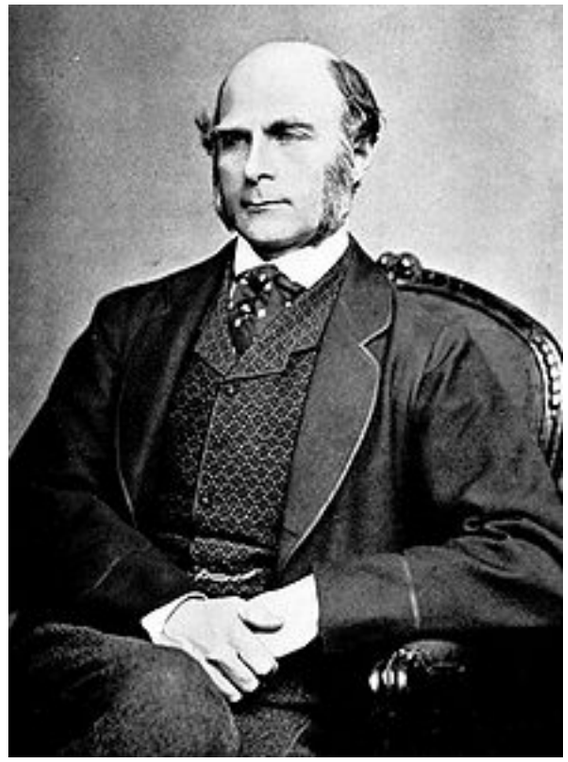
Inferential data analysis

- **Comparing groups**
  - ANOVA (comparing > 2 groups)
    - Example output

```
            Df Sum Sq Mean Sq F value Pr(>F)
users        3 2781.1   927.0     152 <2e-16 ***
Residuals   96  585.6     6.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

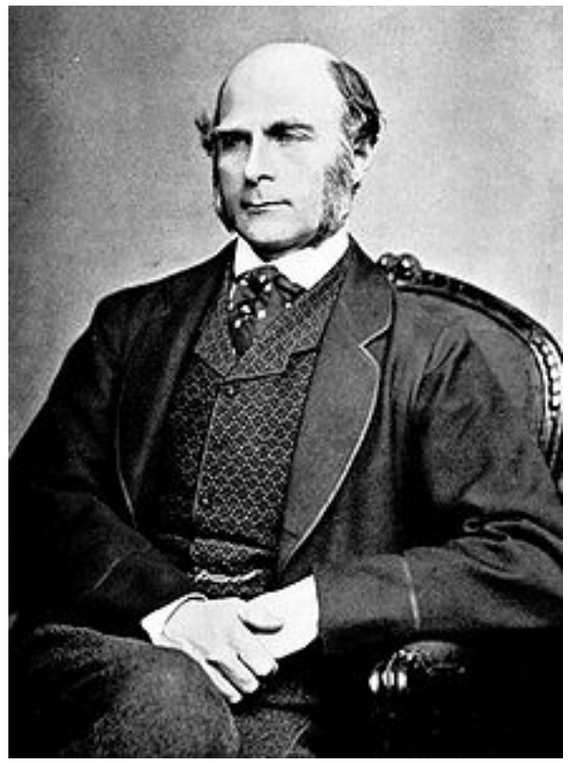
Inferential data analysis

- **Continuous and categorical explanatory variables**
  - Regression (Francis Galton, 1822-1911)
  - Assumptions
    - Continuous or categorical (with 2 categories) predictors
    - Each variable has to have some variance
    - No multicollinearity (perfect linear relationship between two predictors)
    - homoscedasticity (same as homogeneity, but for continuous variables



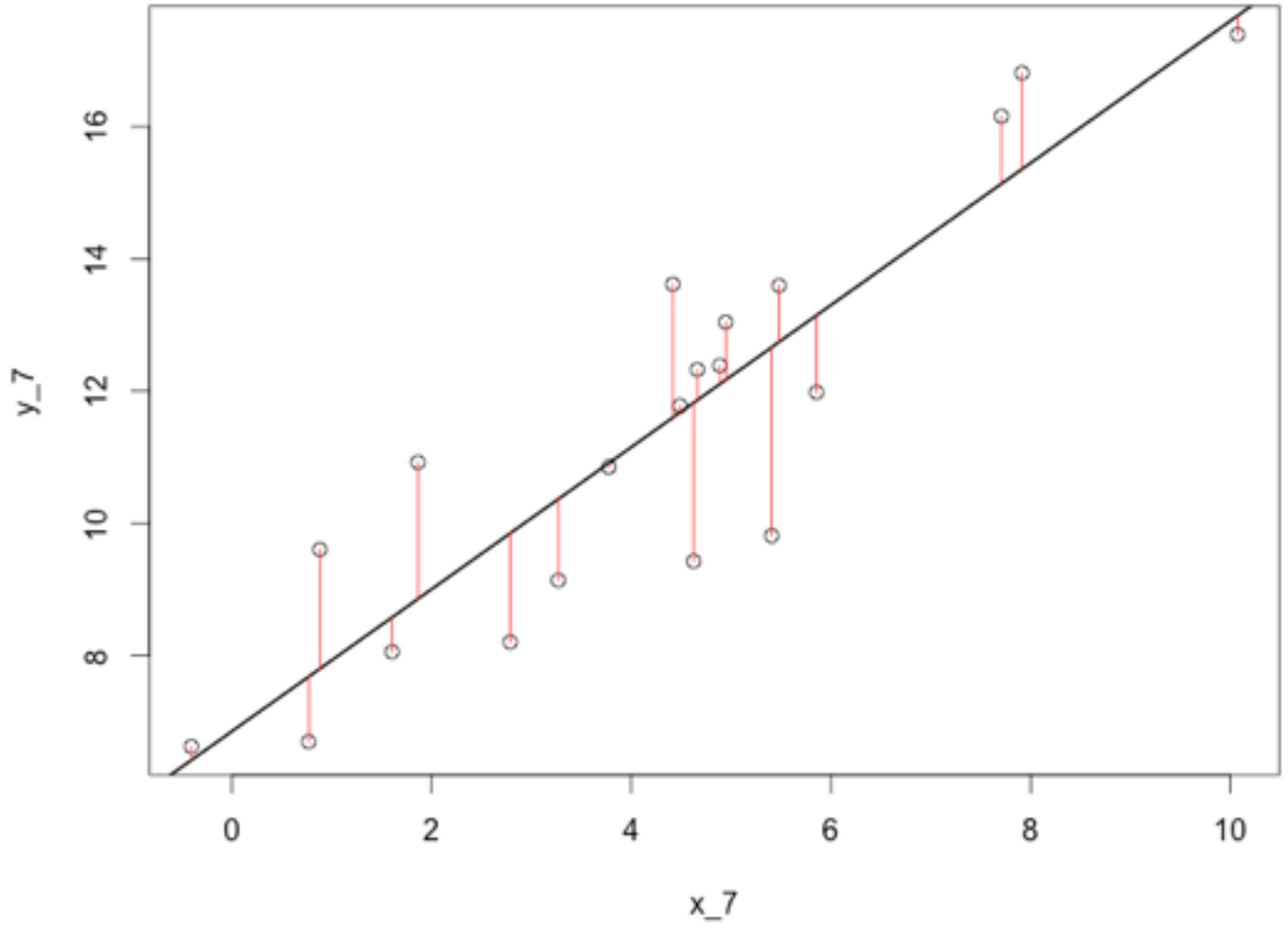
Inferential data analysis

- **Continuous and categorical explanatory variables**
  - Regression (Francis Galton, 1822-1911)
  - Assumptions (con't)
    - the residual variances of any two observations are uncorrelated
    - the residual variance should be normally distributed with a mean of 0
    - independence of observations
    - linearity, the relationship between the predictor and the outcome is linear (i.e. 1 increase in predictor == 1 increase in outcome).



Inferential data analysis

- **Continuous and categorical explanatory variables**
  - Least squares estimation

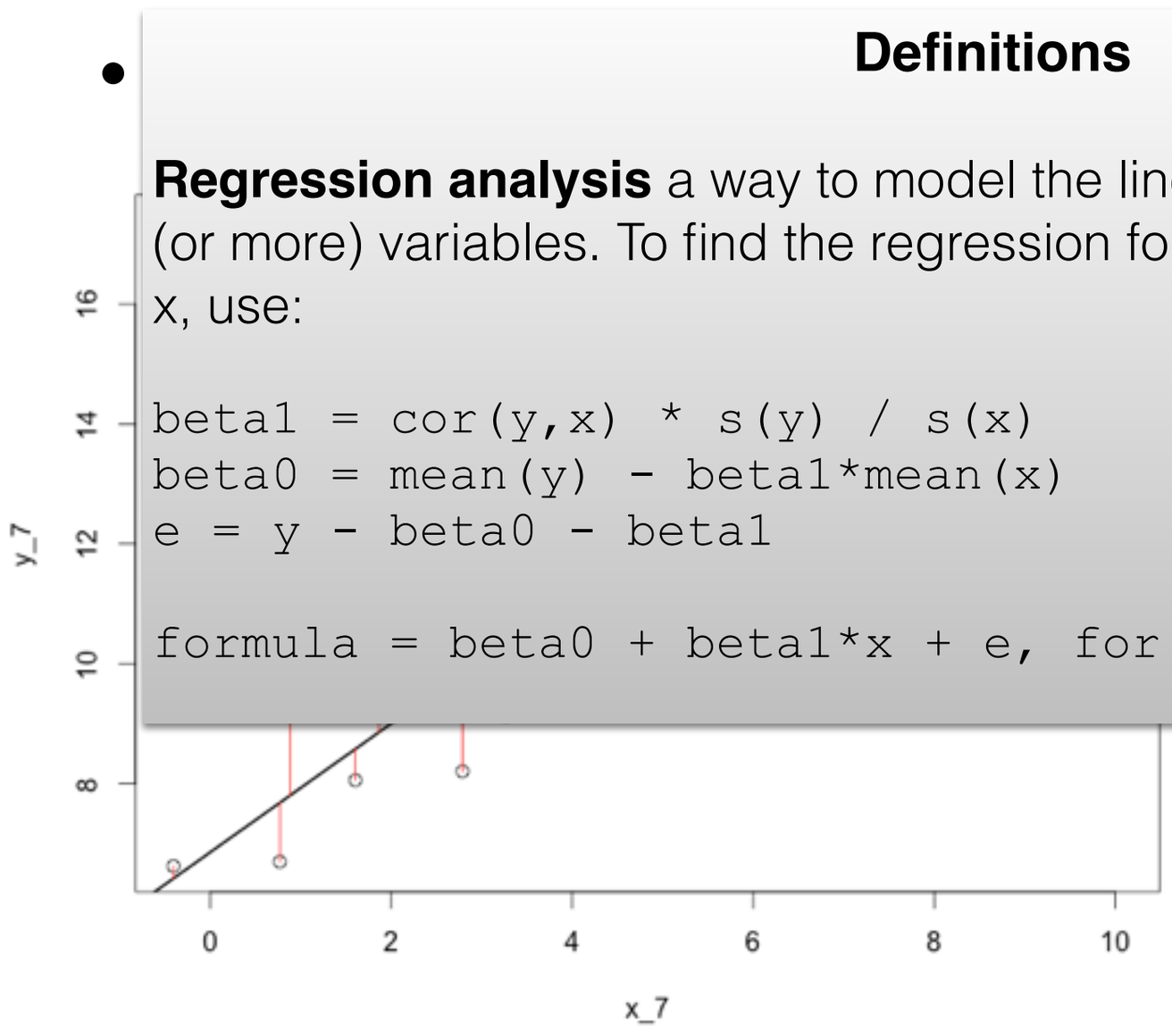


Inferential data analysis

- **Continuous and categorical explanatory variables**
  - **Definitions**

    Regression analysis a way to model the linear relationship between two (or more) variables. To find the regression formula for predicting y from x, use:

    ```
    beta1 = cor(y,x) * s(y) / s(x)
    beta0 = mean(y) - beta1*mean(x)
    e = y - beta0 - beta1

    formula = beta0 + beta1*x + e, for each observation
    ```

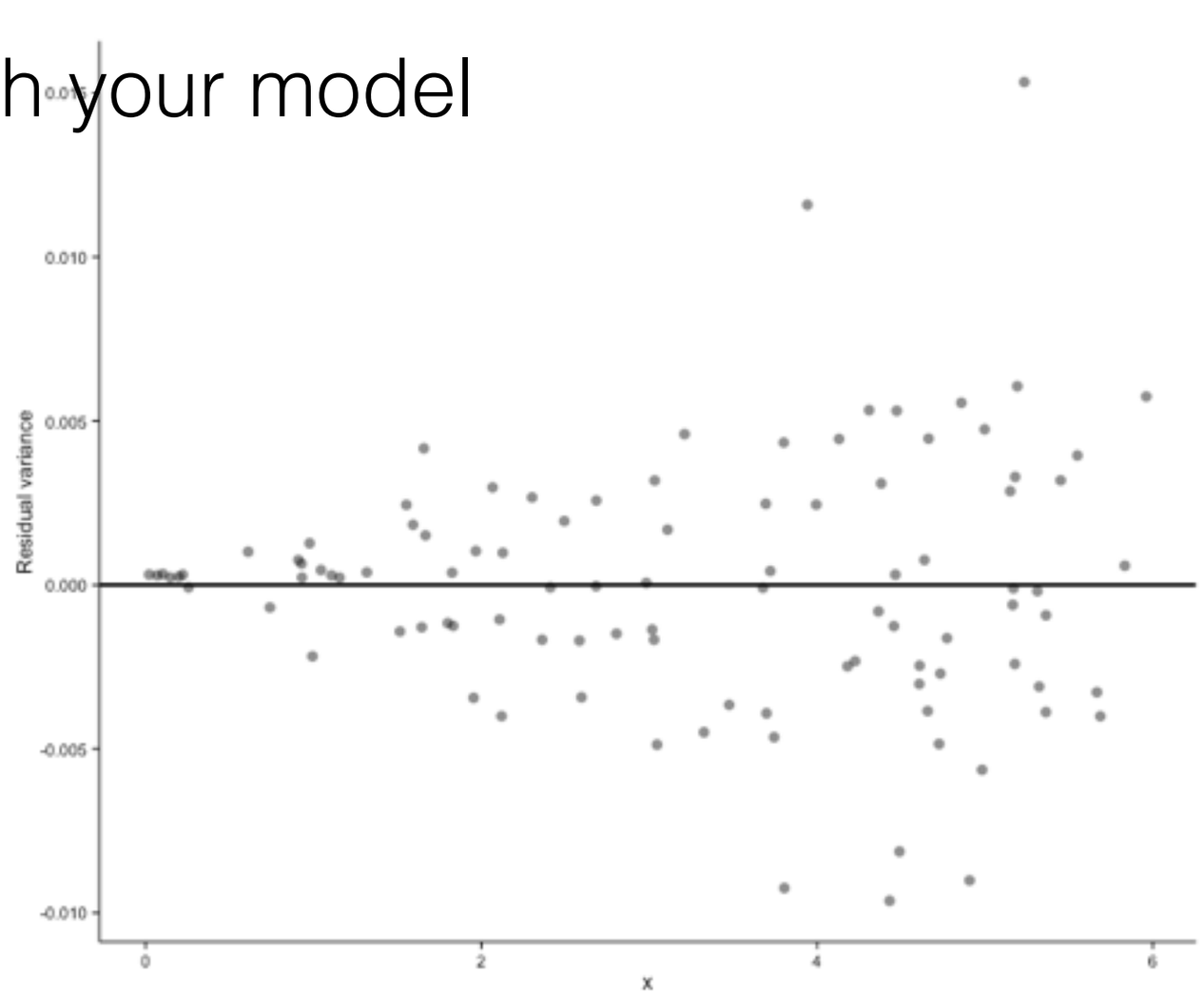


Inferential data analysis

- **Continuous and categorical explanatory variables**
  - Residual variation
    - The variation in the outcome variable that is left after adding predictors to the model
    - The outcome with the linear association with the predictor removed

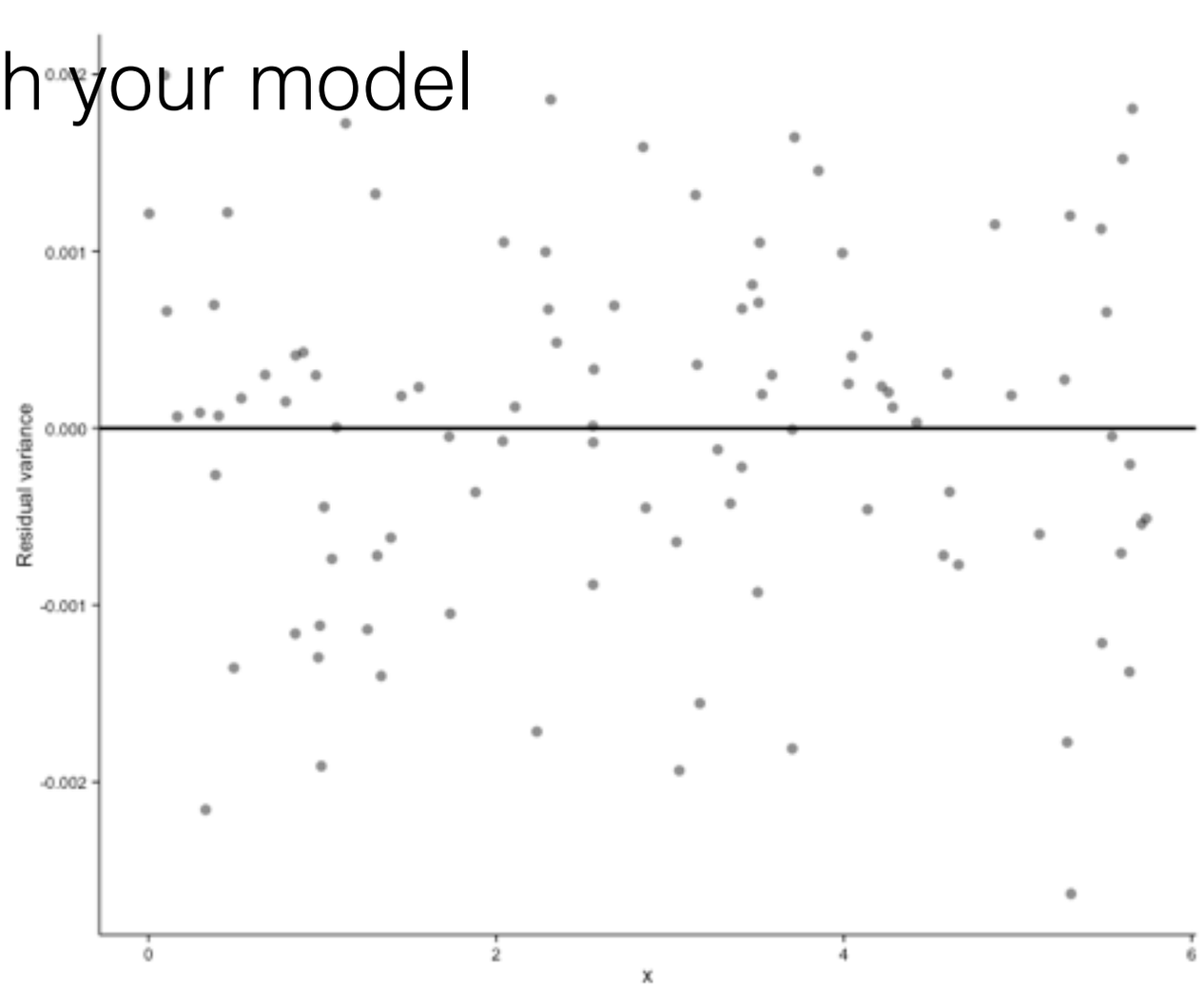
Inferential data analysis

- **Continuous and categorical explanatory variables**
  - Residual variation
    - Can show issues with your model
      - Heteroscedasticity
      - Normal distribution
      - Uncorrelated residuals



Inferential data analysis

- **Continuous and categorical explanatory variables**
  - Residual variation
    - Can show issues with your model
      - Heteroscedasticity
      - Normal distribution
      - Uncorrelated residuals



Inferential data analysis

- **Continuous and categorical explanatory variables**
  - Is my predictor significant?
    - To figure out if a predictor is significant, we compute a t-value for this estimate and find its p-value
  - How do I interpret the estimate of my predictor?
    - It shows you how much the outcome will change when your predictor is increased by one unit

Definitions

Regression analysis a way to model the linear relationship between two (or more) variables. To find the regression formula for predicting y from x, use:

```
beta1 = cor(y,x) * s(y) / s(x)
beta0 = mean(y) - beta1*mean(x)
e = y - beta0 - beta1

formula = beta0 + beta1*x + e, for each observation
```

Now, to find out whether the estimate of beta 0 and beta1 are significant, we use:

```
sigma = sqrt(sum(e^2) / (n - 2))
ssx = sum((x - mean(x))^2)

seBeta0 = (1/n + mean(x)^2 / ssx) ^.5 * sigma
tBeta0 = beta0 / seBeta0

seBeta1 = sigma / sqrt(ssx)
tBeta1 = beta1 / seBeta1
```

Inferential data analysis

- **Continuous and categorical explanatory variables**
  - More than one predictor?
  - For each extra predictor, the analysis looks whether the predictor can explain a significant amount of remaining residual variance
  - How to interpret the estimate?
    - The expected change in the outcome variable when the predictor is increased by one unit, keeping all other predictors fixed

Inferential data analysis

- **Continuous and categorical explanatory variables**
  - Categorical explanatory variables?
  - Two categories? Code as 0 and 1

Inferential data analysis

- **Continuous and categorical explanatory v**
  - **Dummy variable** a 0/1 variable that denotes whether an observation is or is not part of a group. A popular example is gender. Males are part of the male group and coded as 1, females are not part of the male group, so are coded as 0. In the regression formula, this looks like:

Definitions

```
formula = beta0 + beta1*gender + e, for each observation
males = beta0 + beta1*1 + e
females = beta0 + beta1*0 + e
```

So, for this model, the value of beta0 will be the mean of the group of females, while the sum of beta0 + beta1 will be the mean of the group of males. Beta1 is then the difference between males and females on your outcome variable.

Inferential data analysis

- **Continuous and categorical explanatory variables**
  - Categorical explanatory variables?
  - Two categories? Code as 0 and 1
  - More than two categories? Create dummy variables

Reference category

	D1	D2
Small	1	0
Medium	0	1
Large	0	0

Inferential data analysis

- **Continuous and categorical explanatory v**
  - **Dummy variable** a 0/1 variable that denotes whether an observation is or is not part of a group. When a variable has more than 2 categories, you need to create categories - 1 dummy variables to represent all the categories:

```
formula = beta0 + beta1*d1 + beta2*d2 + e
small = beta0 + beta1*1 + beta2*0 + e
medium = beta0 + beta1*0 + beta2*1 + e
large = beta0 + beta1*0 + beta2*0 + e

beta1 compares Small to Large
beta2 compares Medium to Large
beta1 - beta2 compares Small to Medium
```

Definitions

Reference category

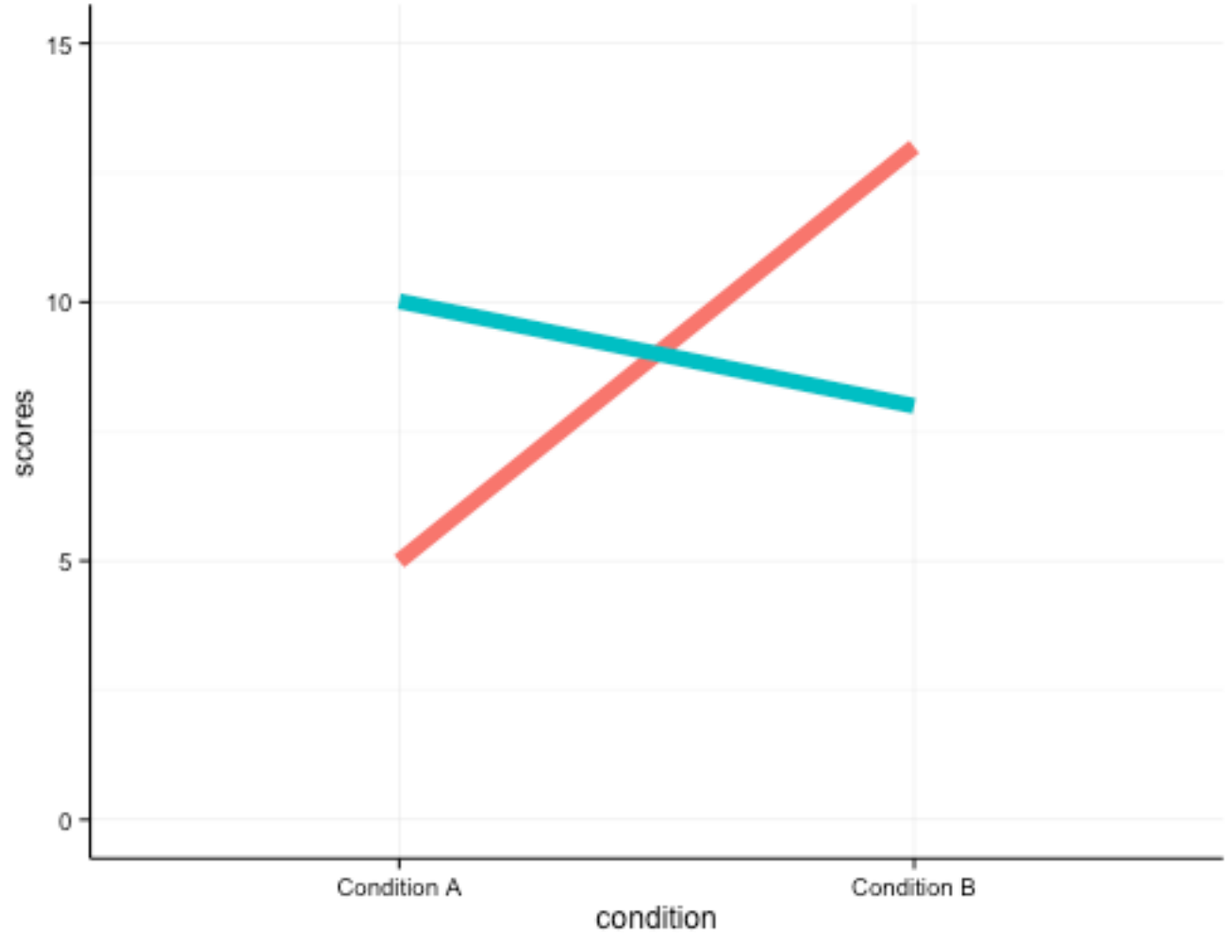
Large	0	0

Inferential data analysis

- **Continuous and categorical explanatory variables**
  - Categorical explanatory variables?
  - Interaction effect
    - The expected change in Y1 per unit change in X1, per unit change in X2

Inferential data analysis

- **Continuous and categorical explanatory variables**
  - Categorical explanatory variables?
  - Interaction effect
    - The expected change in Y1 per unit change in X1, per unit change in X2



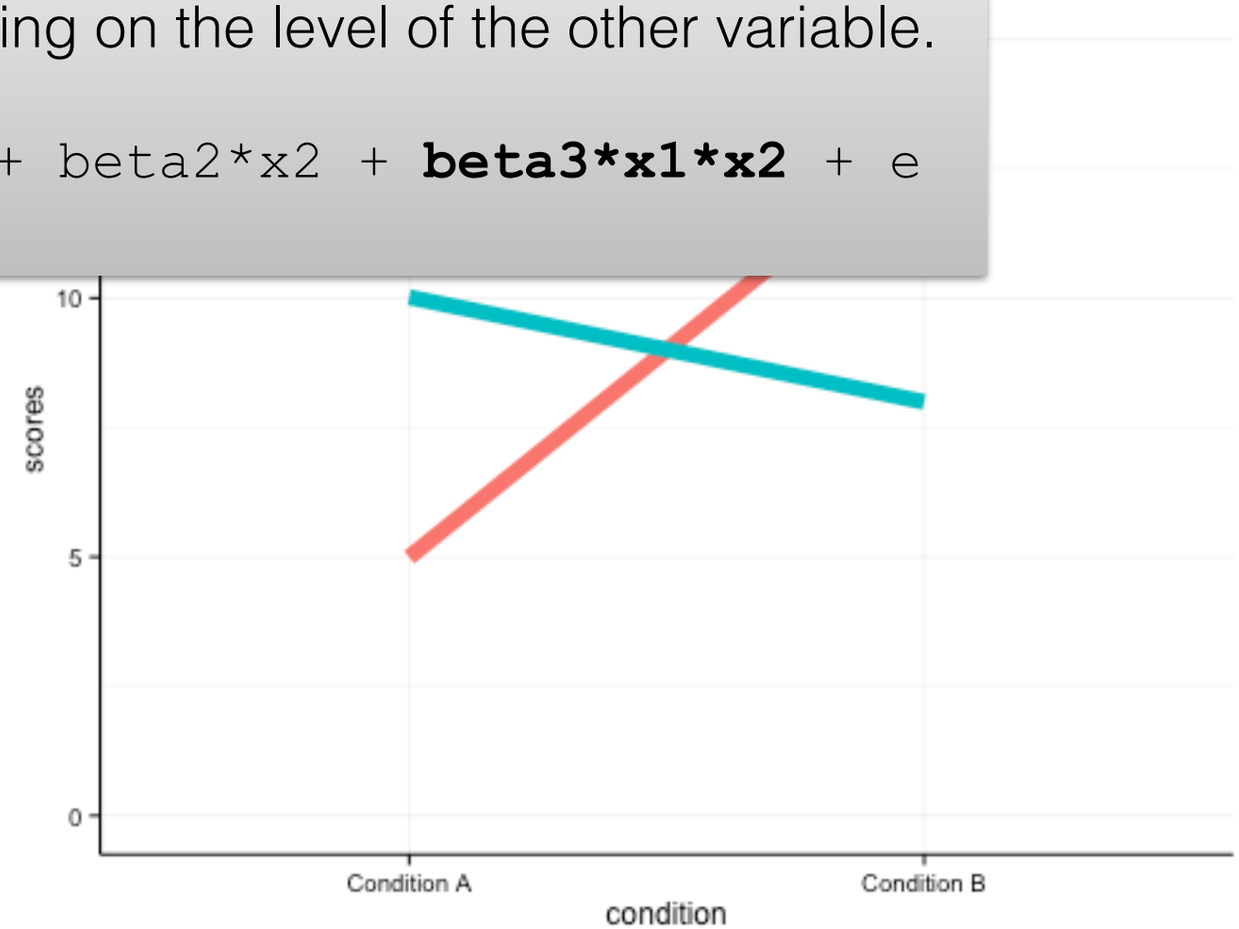
Inferential data analysis

- **Continuous and categorical explanatory variables**
  - change in Y1 per unit change in X1, per unit change in X2

Definitions

Interaction effect two predictors interact with each other if the effect of one of the variables differs depending on the level of the other variable.

`formula = beta0 + beta1*x1 + beta2*x2 + beta3*x1*x2 + e`



Inferential data analysis

- **Continuous and categorical explanatory variables**
  - Example output

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.45781 -0.51600 -0.02595  0.49362  1.49759

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.7652     0.4511   3.913 0.000175 ***
systemDragonInc    -0.7734     0.2293  -3.373 0.001094 **
users5              2.8170     0.8069   3.491 0.000744 ***
users10             4.9037     0.7761   6.318 9.59e-09 ***
users50             5.2597     0.8072   6.516 3.92e-09 ***
process_B           2.0306     0.2566   7.913 5.78e-12 ***
users5:process_B   -1.0698     0.3259  -3.282 0.001461 **
users10:process_B  -1.2191     0.2868  -4.250 5.15e-05 ***
users50:process_B  -1.0649     0.2738  -3.889 0.000191 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6711 on 91 degrees of freedom
Multiple R-squared:  0.9852,    Adjusted R-squared:  0.9839
F-statistic: 759.4 on 8 and 91 DF,  p-value: < 2.2e-16
```

Inferential data analysis

- Other types of regression analysis
  - Binary Logistic Regression
    - If you have a yes/no outcome you'd like to predict
  - Multinomial Logistic Regression
    - If you have a nominal categorical outcome variable with more than 2 categories
  - Poisson Regression
    - If you have a count outcome you'd like to predict

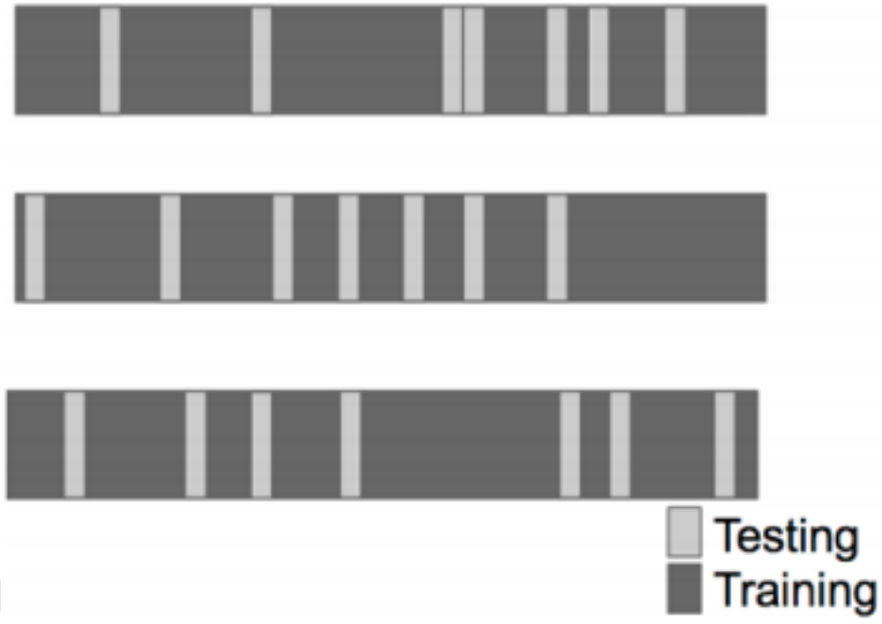
Predictive data analysis

Predictive data analysis

- **Preparation**
  - Splitting data
    - Training data versus testing data
    - 60/40 split is common

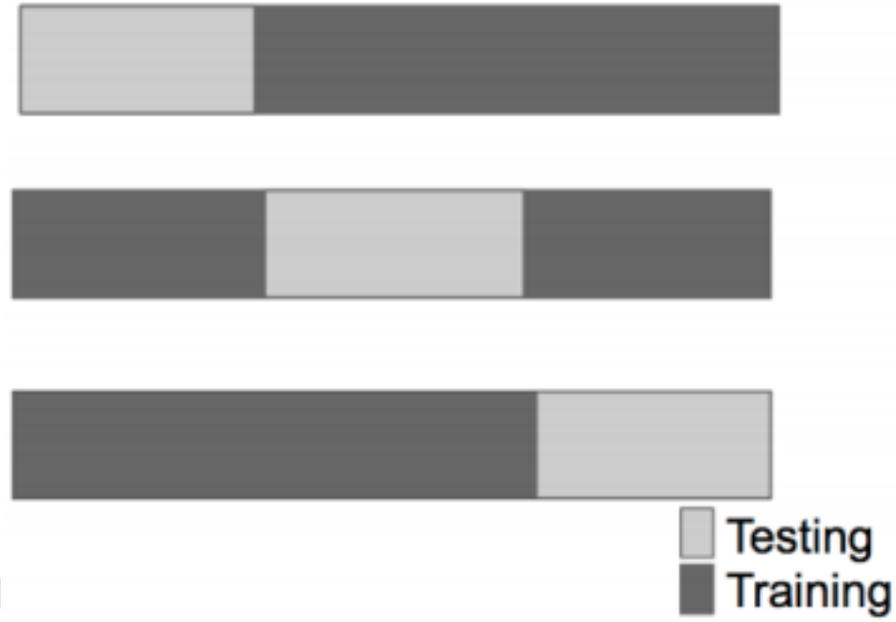
Predictive data analysis

- **Preparation**
  - Cross validation
    - Helps guard against overfitting
    - Within your training data
      - Split again (20/20/20)
      - Random subsampling
      - K-fold cross validation
      - Leave one out cross validation



Predictive data analysis

- **Preparation**
  - Cross validation
    - Helps guard against overfitting
    - Within your training data
      - Split again (20/20/20)
      - Random subsampling
      - K-fold cross validation
      - Leave one out cross validation



Predictive data analysis

- **Preparation**
  - Cross validation
    - Helps guard against overfitting
    - Within your training data
      - Split again (20/20/20)
      - Random subsampling
      - K-fold cross validation
      - Leave one out cross validation

Predictive data analysis

- **Accuracy of your predictions (categorical outcomes with 2 options)**
    - True positive = correctly identified
    - True negative = correctly rejected
    - False positive = incorrectly identified
    - False negative = incorrectly rejected

- Sensitivity = TP / (TP+FN)
- Specificity = TN / (FP + TN)
- Pos Predictive value = TP / (TP + FP)
- Neg Predictive value = TN (FN + TN)
- Accuracy = (TP + TN) / (TP + FP + FN + TN)

		Truth	
		+	**-**
Test	**+**	TP	FP
	-	FN	TN

Predictive data analysis

- **Accuracy of your predictions (categorical outcomes with more than 2 options)**
    - Misclassification Error: misclassified / all options
        - 0 = perfect purity; 0.5 = no purity

    - Gini index: 1 - ((misclassified / all options)^2 + (correctly classified / all options)^2)
        - 0 = perfect purity; 0.5 = no purity

Predictive data analysis

- **Accuracy of your predictions (categorical outcomes)**
    - Information gain: - (misclassified / all options * log2(misclassified / all options) + correctly classified / all options * log2(correctly classified / all options))
    - 0 = perfect impurity; 1 = no purity

Predictive data analysis

- **Accuracy of your predictions (continuous outcomes)**
  - Mean Square Error
  - 1 / length(test) * sum((pred - test)^2)

- Root Mean Square Error
  - sqrt(MSE)

Predictive data analysis

- **Machine Learning Algorithms**

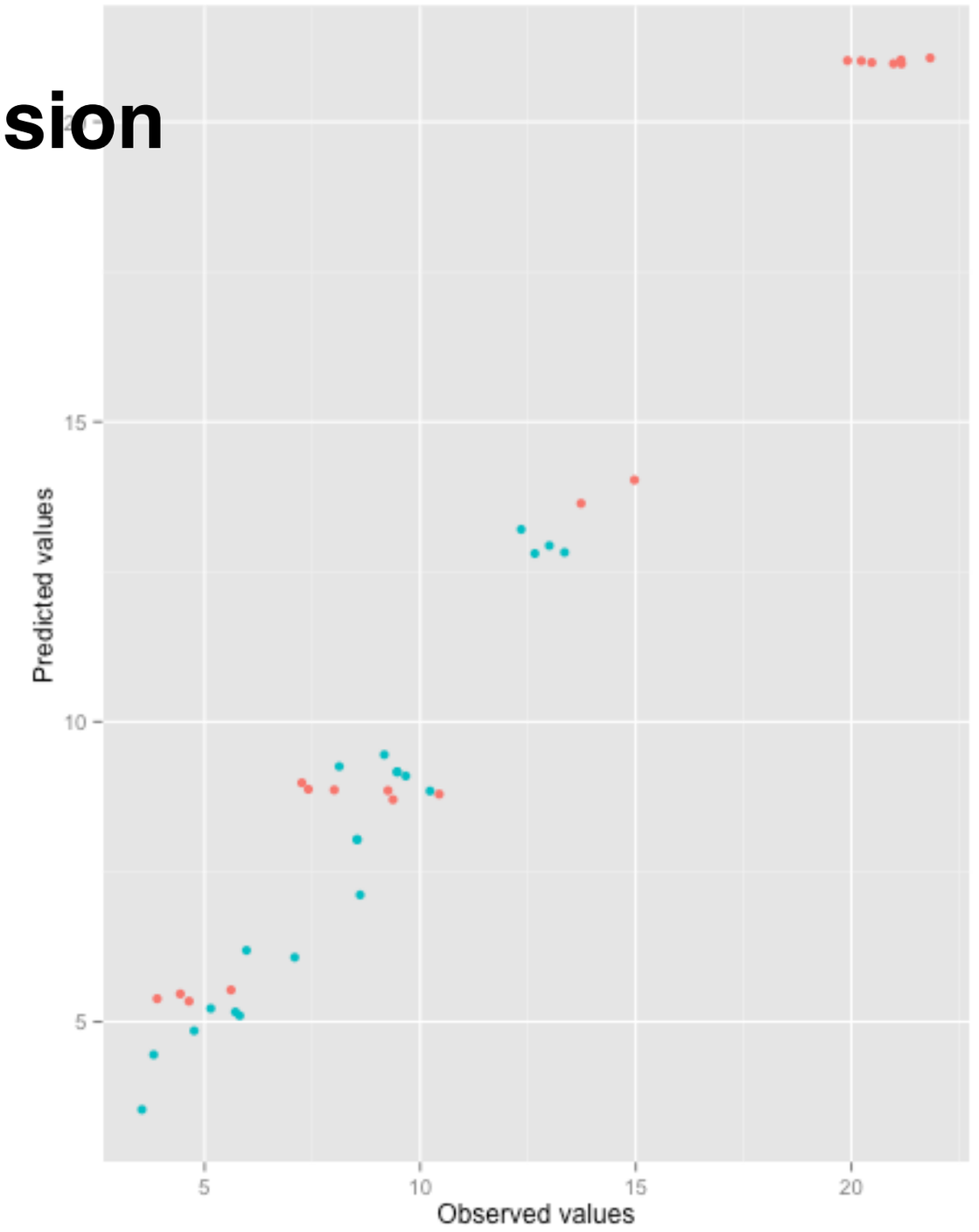
```
> names(getModelInfo())
  [1] "ada"              "AdaBag"           "AdaBoost.M1"      "amdai"            "ANFIS"
  [6] "avNNet"           "bag"              "bagEarth"         "bagEarthGCV"      "bagFDA"
 [11] "bagFDAGCV"        "bayesglm"         "bdk"              "binda"            "blackboost"
 [16] "Boruta"           "brnn"             "bstLs"            "bstSm"            "bstTree"
 [21] "C5.0"             "C5.0Cost"         "C5.0Rules"        "C5.0Tree"         "cforest"
 [26] "chaid"            "CSimca"           "ctree"            "ctree2"           "cubist"
 [31] "DENFIS"           "dnn"              "earth"            "elm"              "enet"
 [36] "enpls.fs"         "enpls"            "evtree"           "extraTrees"       "fda"
 [41] "FH.GBML"          "FIR.DM"           "foba"             "FRBCS.CHI"        "FRBCS.W"
 [46] "FS.HGD"           "gam"              "gamboost"         "gamLoess"         "gamSpline"
 [51] "gaussprLinear"    "gaussprPoly"      "gaussprRadial"    "gbm"              "gcvEarth"
 [56] "GFS.FR.MOGAL"     "GFS.GCCL"         "GFS.LT.RS"        "GFS.THRIFT"       "glm"
 [61] "glmboost"         "glmnet"           "glmStepAIC"       "gpls"             "hda"
 [66] "hdda"             "HYFIS"            "icr"              "J48"              "JRip"
 [71] "kernelpls"        "kknn"             "knn"              "krlsPoly"         "krlsRadial"
 [76] "lars"             "lars2"            "lasso"            "lda"              "lda2"
 [81] "leapBackward"     "leapForward"      "leapSeq"          "Linda"            "lm"
 [86] "lmStepAIC"        "LMT"              "logicBag"         "LogitBoost"       "logreg"
 [91] "lssvmLinear"      "lssvmPoly"        "lssvmRadial"      "lvq"              "M5"
 [96] "M5Rules"          "mda"              "Mlda"             "mlp"              "mlpWeightDecay"
[101] "multinom"         "nb"               "neuralnet"        "nnet"             "nodeHarvest"
[106] "oblique.tree"     "OneR"             "ORFlog"           "ORFpls"           "ORFridge"
[111] "ORFsvm"           "pam"              "parRF"            "PART"             "partDSA"
[116] "pcaNNet"          "pcr"              "pda"              "pda2"             "penalized"
[121] "PenalizedLDA"     "plr"              "pls"              "plsRglm"          "polr"
[126] "ppr"              "protoclass"       "qda"              "QdaCov"           "qrf"
[131] "qrnn"             "rbf"              "rbfDDA"           "rda"              "relaxo"
[136] "rf"               "rFerns"           "RFlda"            "ridge"            "rknn"
[141] "rknnBel"          "rlm"              "rmda"             "rocc"             "rpart"
[146] "rpart2"           "rpartCost"        "RRF"              "RRFglobal"        "rrlda"
[151] "RSimca"           "rvmLinear"        "rvmPoly"          "rvmRadial"        "SBC"
[156] "sda"              "sddaLDA"          "sddaQDA"          "simpls"           "SLAVE"
[161] "slda"             "smda"             "sparseLDA"        "spls"             "stepLDA"
[166] "stepQDA"          "superpc"          "svmBoundrangeString" "svmExpoString" "svmLinear"
[171] "svmPoly"          "svmRadial"        "svmRadialCost"    "svmRadialWeights" "svmSpectrumString"
[176] "treebag"          "vbmpRadial"       "widekernelpls"    "WM"               "wsrf"
[181] "xgbLinear"        "xgbTree"          "xyf"
```

Predictive data analysis

- **Predicting with regression**
  - Use the techniques we learned for inferential statistics, but now predict new values with the model you end up with
  - Pros
    - Easy to implement and understand
  - Cons
    - Poor performance when associations aren’t linear

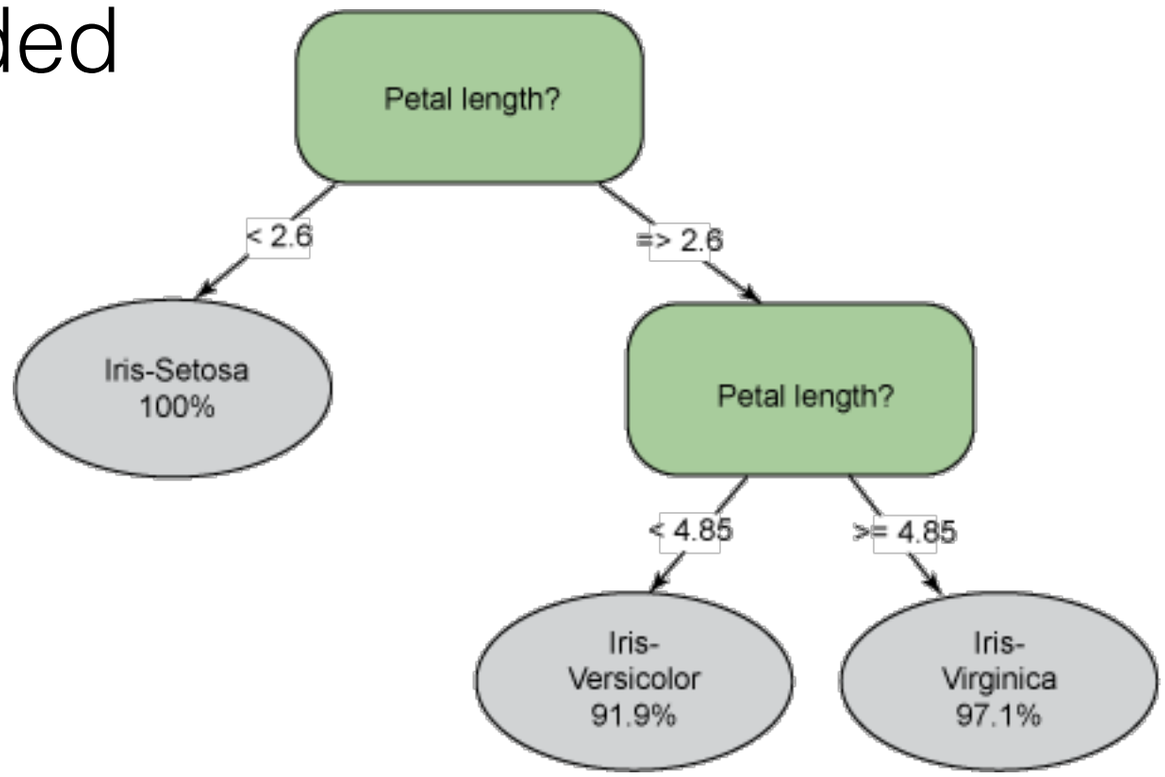
Predictive data analysis

- **Predicting with regression**
  - Basic algorithm
    - Use the linear regression formula that fits the training data to predict the testing data



Predictive data analysis

- **Predicting with trees** (often used for categorical outcomes)
  - Iteratively split variables into groups
  - Evaluate homogeneity within each group
  - Split again if needed



Predictive data analysis

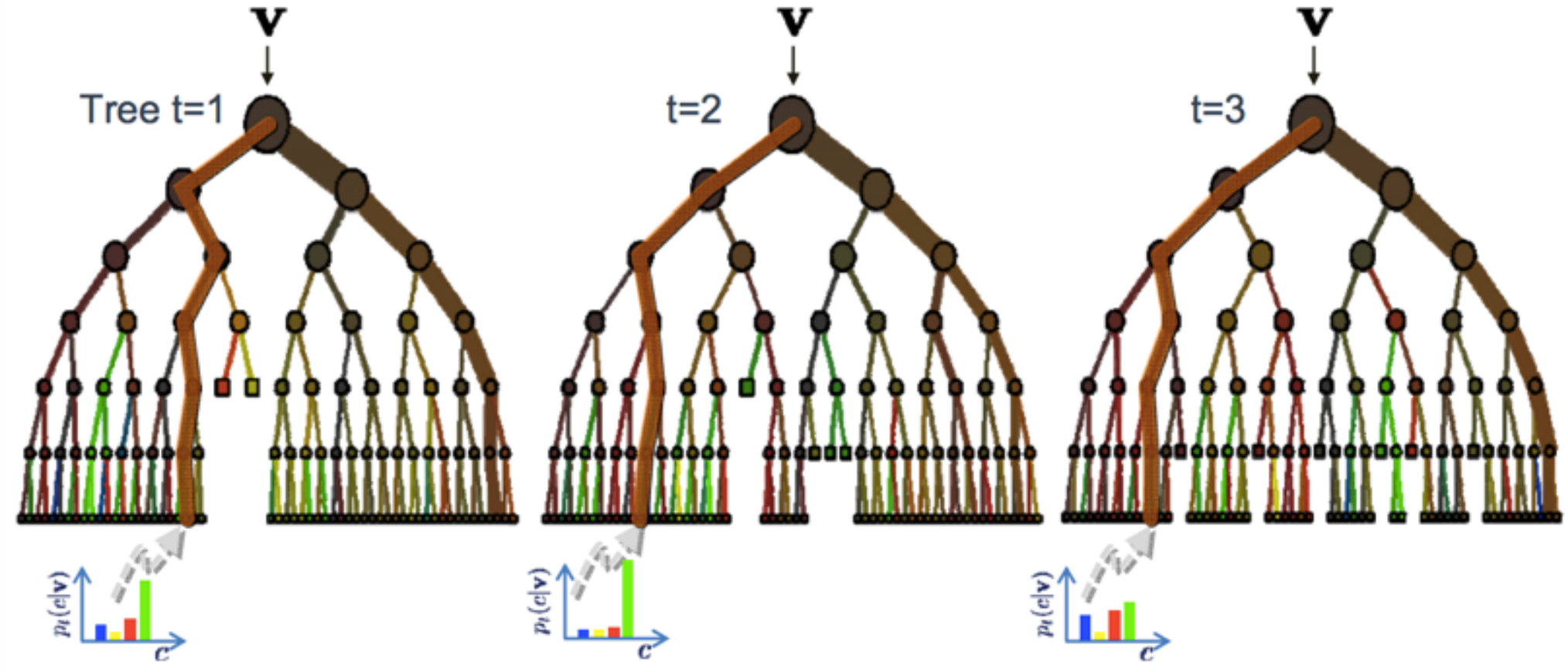
- **Predicting with trees** (often used for categorical outcomes)
    - Pros
        - Easy to interpret
        - Better performance in nonlinear settings
- Cons
    - Without cross-validation overfitting is likely
    - Harder to estimate uncertainty

Predictive data analysis

- **Predicting with trees**
  - Basic algorithm
    - Find variables/splits that best separate outcomes
    - Divide the data into two groups (leaves) based on that split (node)
    - Within each node, find the best variable/splits that best separates the outcomes
    - Repeat until groups are too small, or sufficiently "pure"

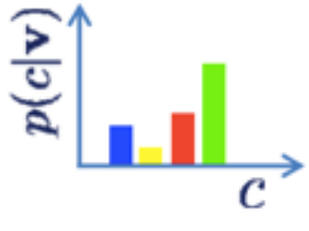
Predictive data analysis

- **Predicting with many trees** (i.e. random forests)



The ensemble model

Forest output probability $p(c|\mathbf{v}) = \frac{1}{T}\sum_t^T p_t(c|\mathbf{v})$



Predictive data analysis

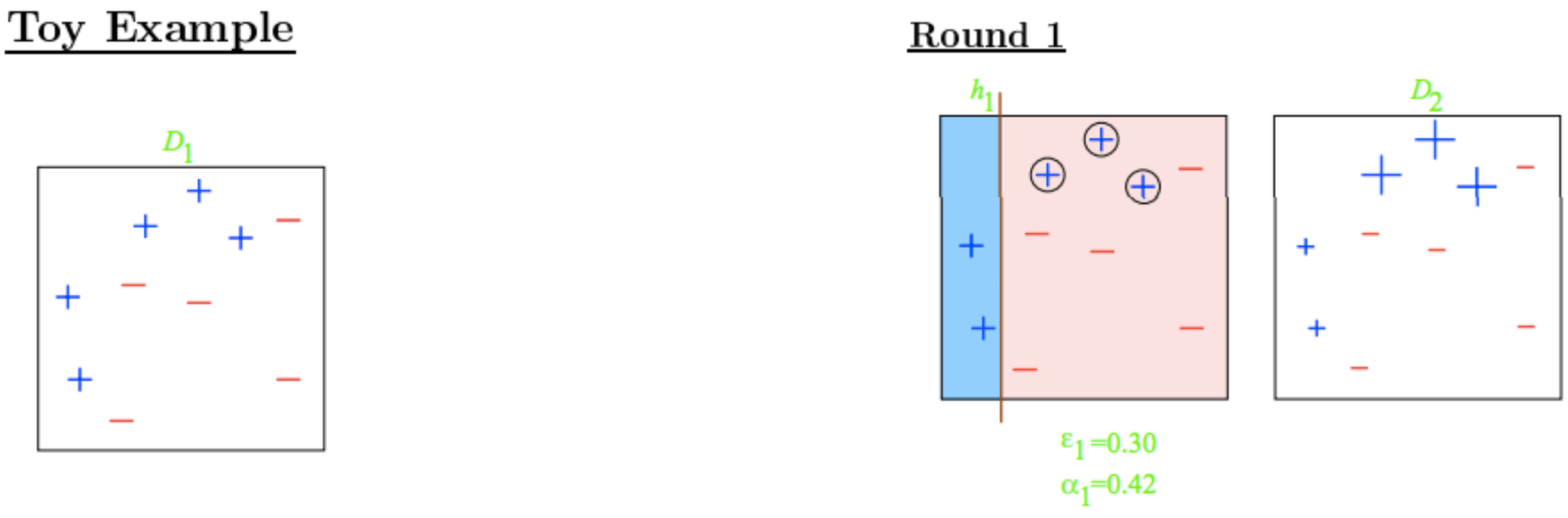
- **Predicting with many trees** (i.e. random forests)
  - Pros
    - Accuracy
  - Cons
    - Speed
    - Interpretability
    - Overfitting (use cross-validation)

Predictive data analysis

- **Boosting**
  - Take a set of classifiers (all regression models, all trees, all random forests)
  - Create a monster classifier the combines the above
    - Minimise error on training set
    - Iterative: Select one classifier per step
    - Calculate weights for each observation based on errors
    - Missed classifications get more importance for the next step

Predictive data analysis

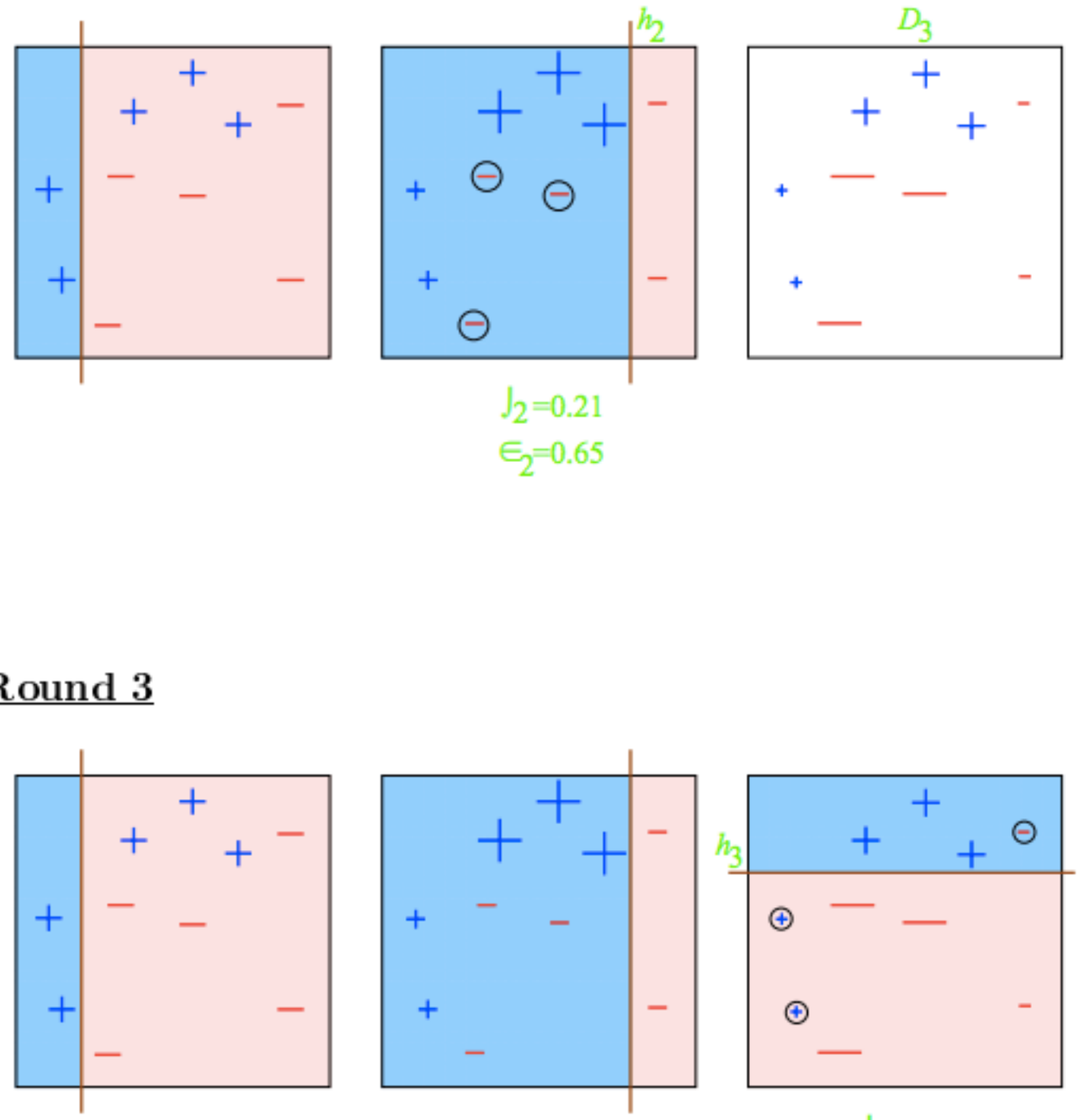
- **Boosting**



Source: http://webee.technion.ac.il/people/rmeir/BoostingTutorial.pdf

Predictive data analysis

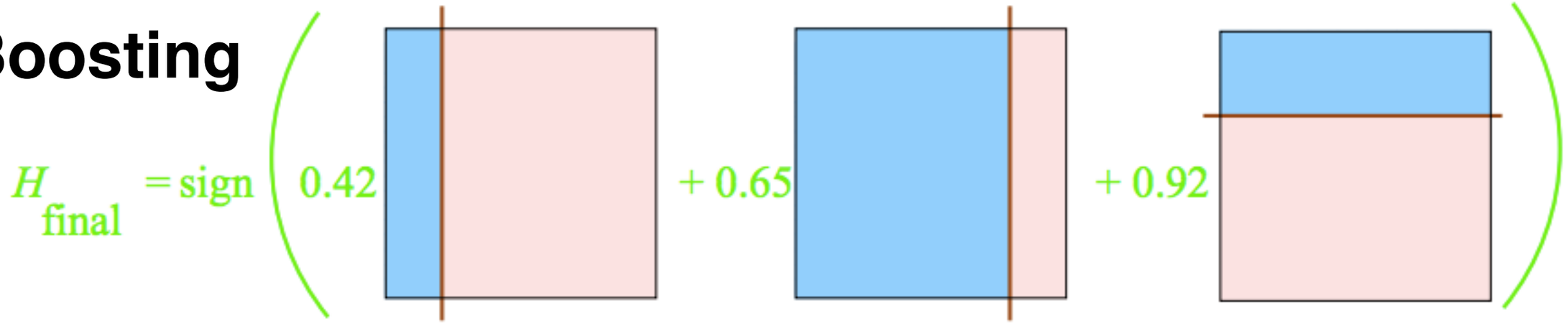
- **Boosting**

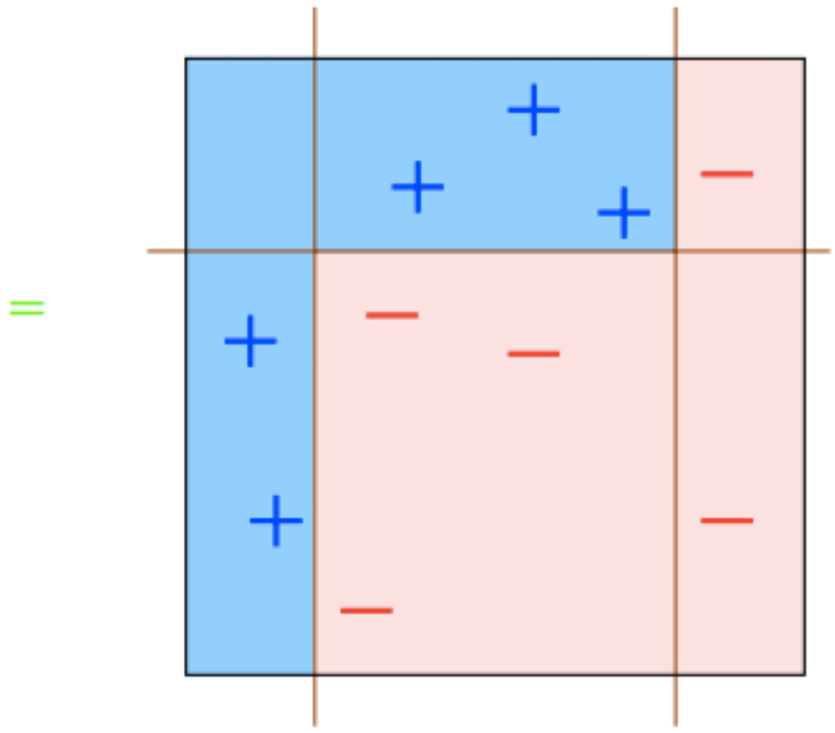


Source: http://webee.technion.ac.il/people/rmeir/BoostingTutorial.pdf

Predictive data analysis

- **Boosting**

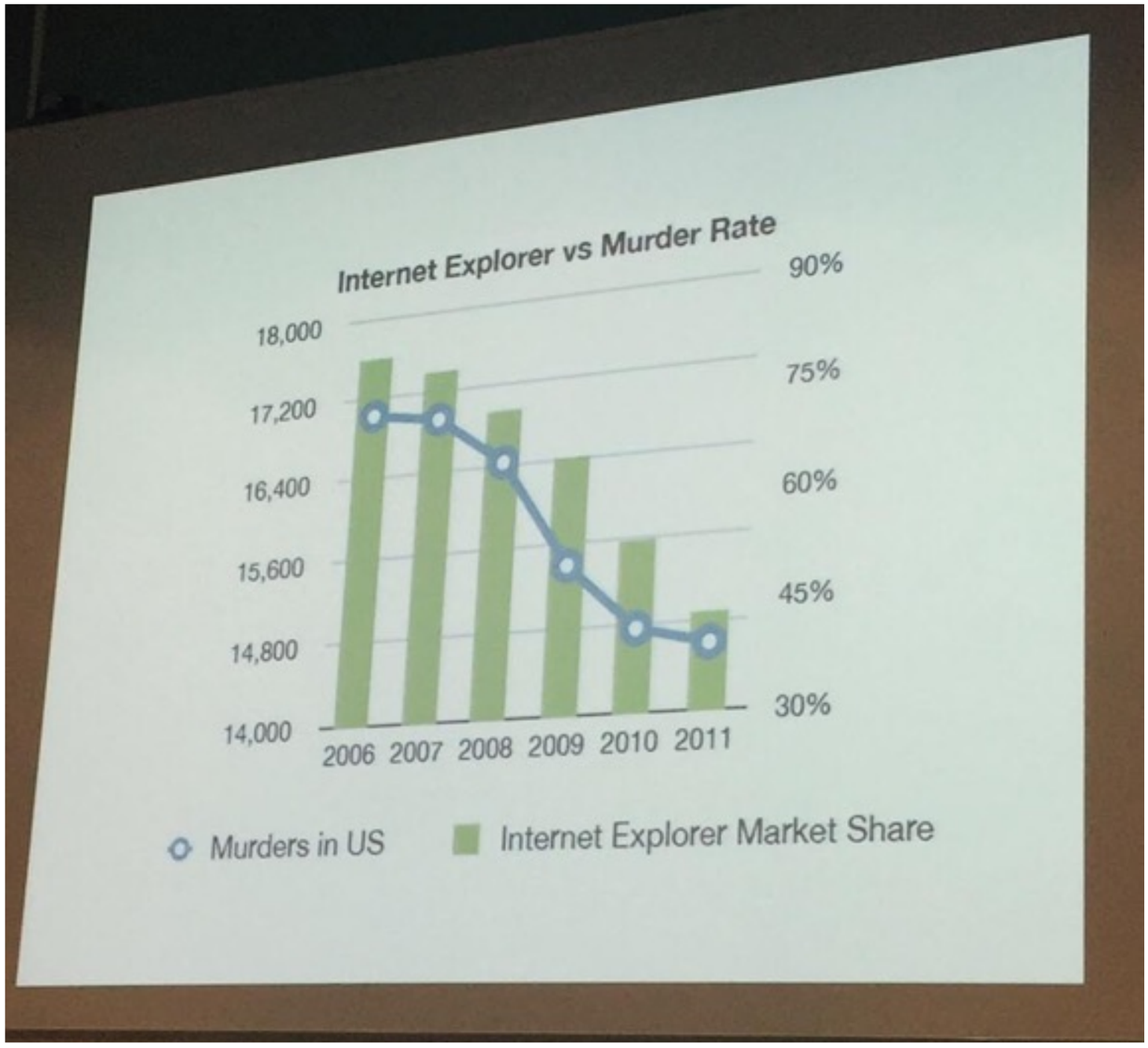




Source: http://webee.technion.ac.il/people/rmeir/BoostingTutorial.pdf

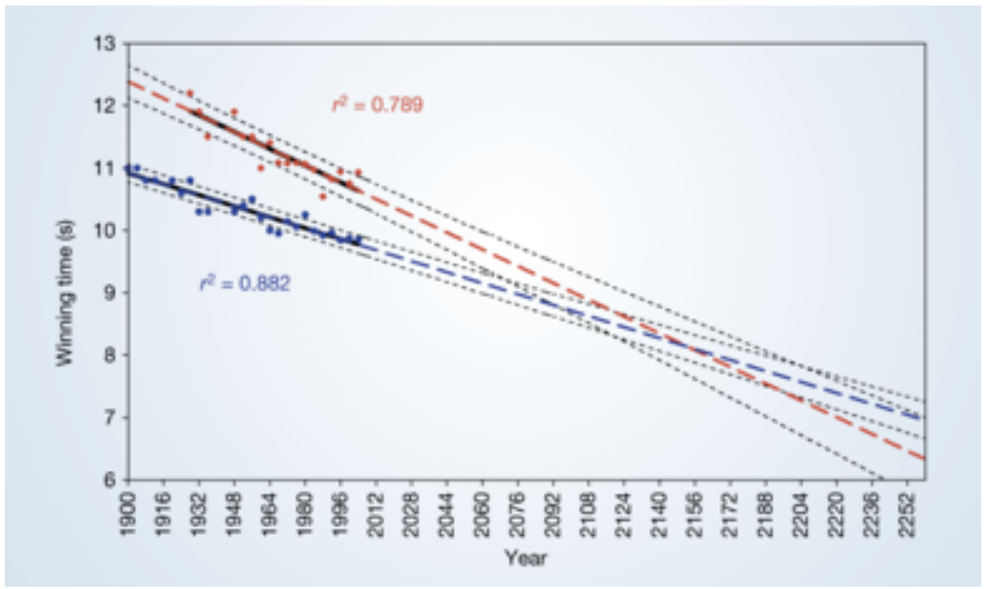
Predictive data analysis

- **Forecasting with time series**
  - dependent data
  - with patterns
    - trend
    - season
    - cycle



Predictive data analysis

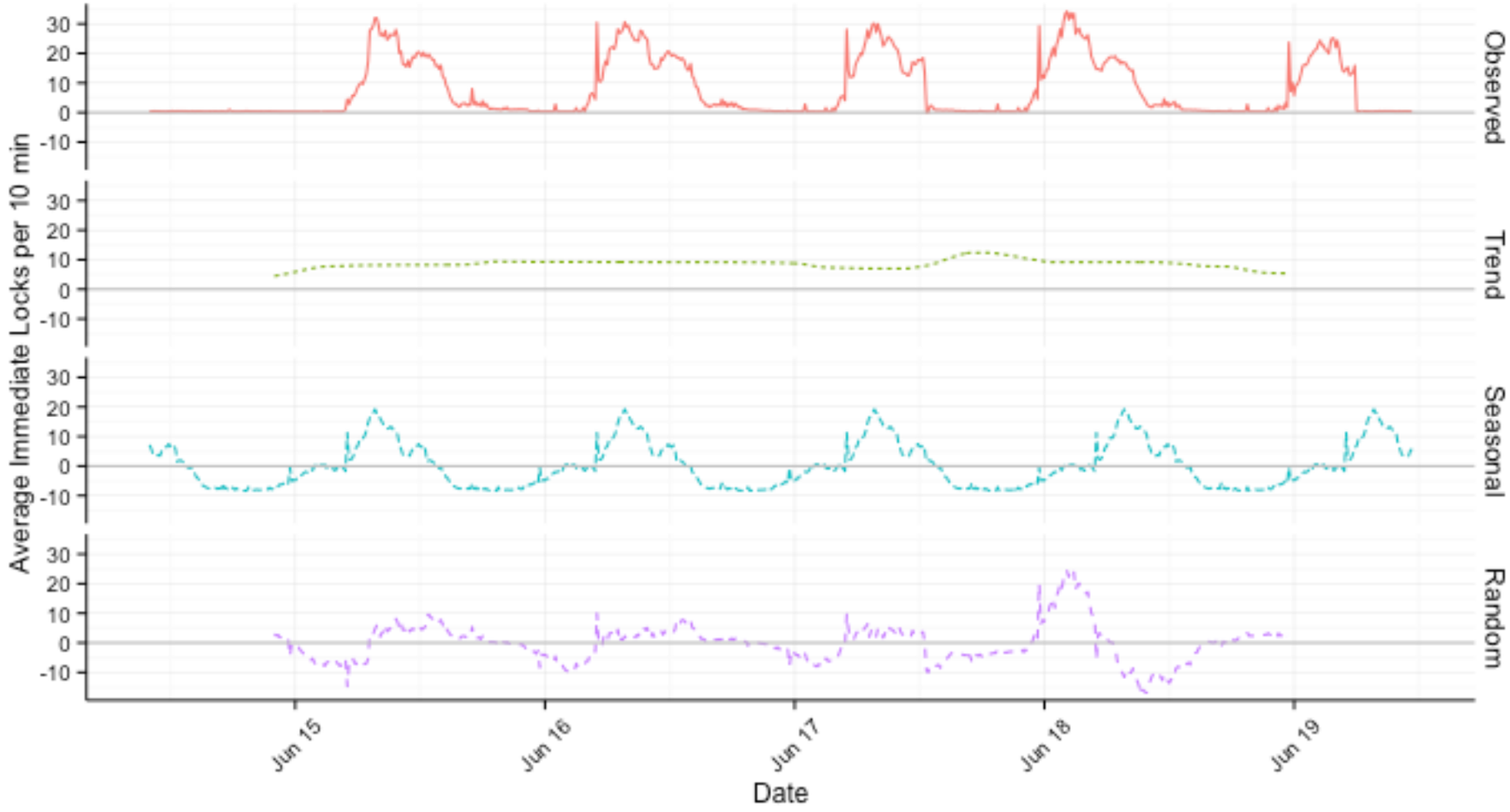
- **Forecasting with time series**
  - Subsampling is more complicated
    - Can only use earlier data to predict later data
  - Beware of spurious correlations
  - Don’t predict too far in the future



Source: http://www.nature.com/nature/journal/v431/n7008/full/431525a.html

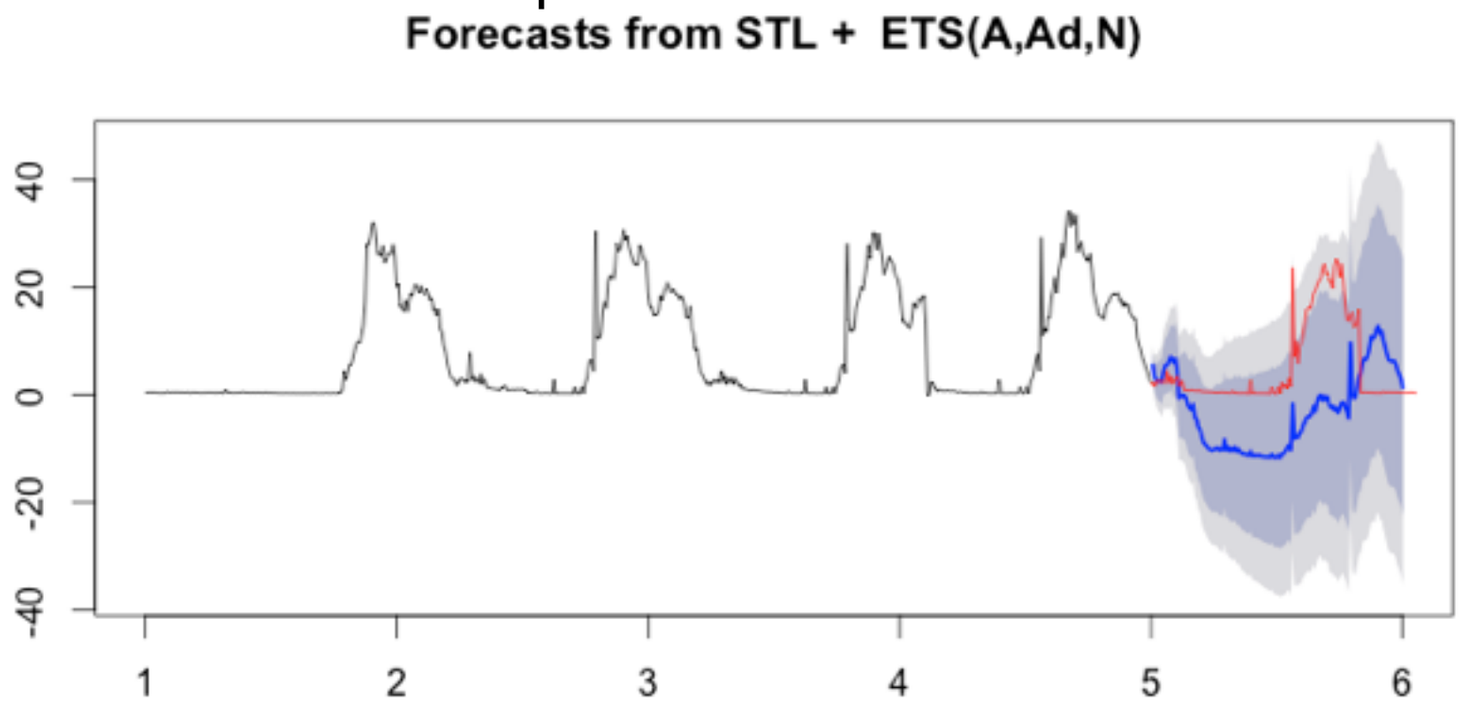
Predictive data analysis

- **Forecasting with time series**



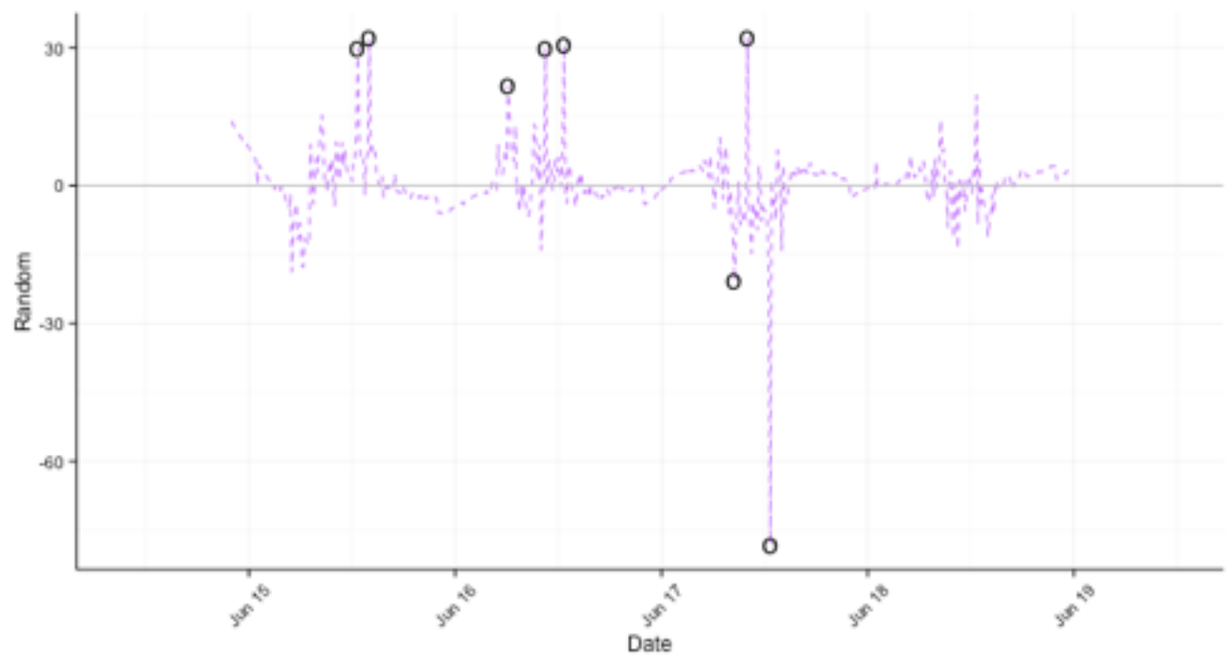
Predictive data analysis

- **Forecasting with time series**
  - Model the time series (many options!)
    - Moving average
    - Exponential moving average
  - Use the model to predict future observations



Predictive data analysis

- **Time series for anomaly detection**
  - Use some property of your time series to find out whether an observation is an anomaly or a normal value.
  - Twitter: AnomalyDetection (R package)
  - Etsy: Kale



The end.

Image sources

- Think photo: https://stocksnap.io/photo/X8I9SUI6DZ
- First slide: https://stocksnap.io/photo/S3JE5YAMND
- Cat image descriptives: https://stocksnap.io/photo/055MVXJ6N9
- Cat image explore: https://stocksnap.io/photo/Y7HV18OPPE
- 4 regression image: http://en.wikipedia.org/wiki/Anscombe%27s_quartet
- Break image: https://stocksnap.io/photo/D322D377AD
- Inferential data analysis: https://stocksnap.io/photo/29648714AF
- Predictive modeling: https://flic.kr/p/euq7N7
- The end: https://stocksnap.io/photo/AWTG760M6W