Forum 10
Revision
Data Skills for Scientists

Welcome

Revision Weeks 1-6

Types of variables

Variables can be quantitative or categorical:

- Quantitative - two types - discrete or continuous.
- Categorical - two types - ordinal or nominal.

Example

What type of variable are these:

- Types of fruit.
- Olympic medals.
- Number of cases of SARS in a year.
- Weights of fish.

Histograms

Used to represent data by counting the number of observations in each bin.

e.g. Consider the numbers [1,1,2,2,3,5,6,6,7,8,9.] We can split this into two bins, all the observations between 0 and 5 and all those between 6 and 10:

0-5: 6 observations.

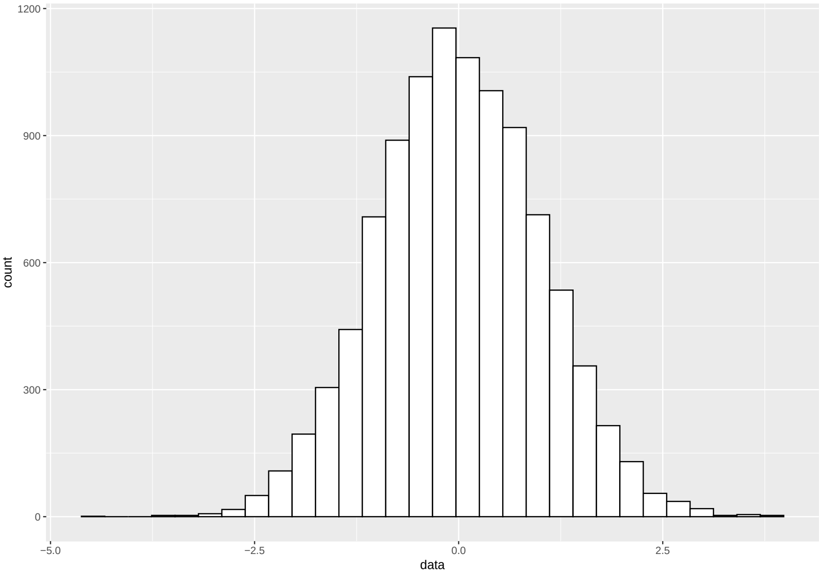
6-10: 5 observations.

Describing histograms and boxplots

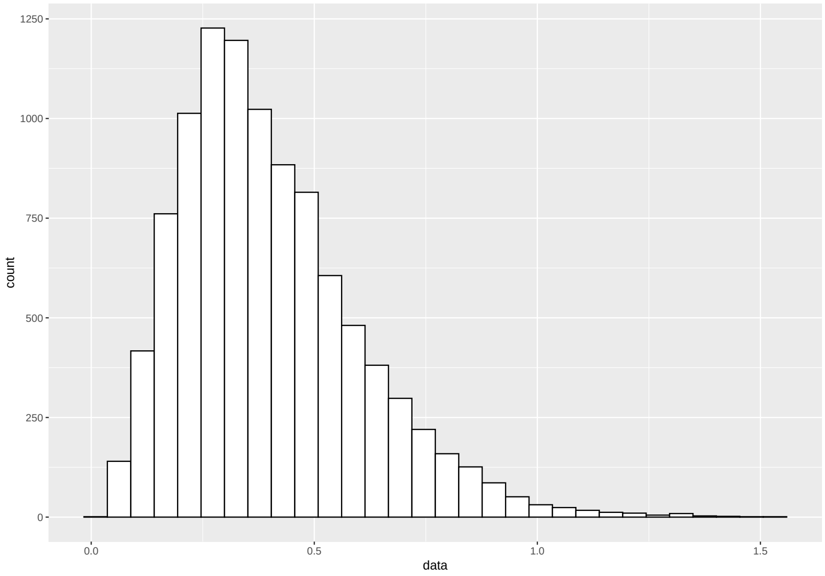
When describing a histogram or boxplot - consider the following

- Shape:
  - Number of modes (peaks),
  - symmetric,
  - left skewed,
  - right skewed.
- Location.
  - Mean.
  - Median.
- Spread.
  - IQR.
  - Standard deviation.
- Outliers.
  - How many, where?

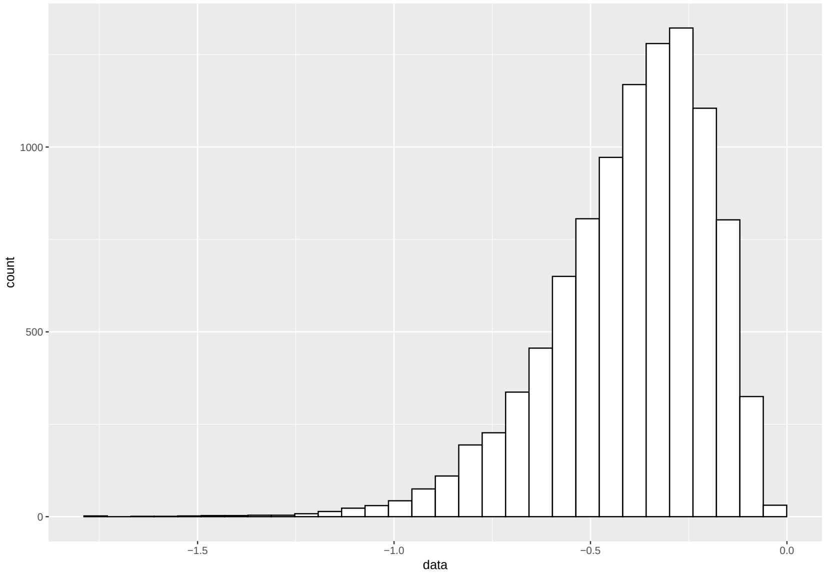
Shape



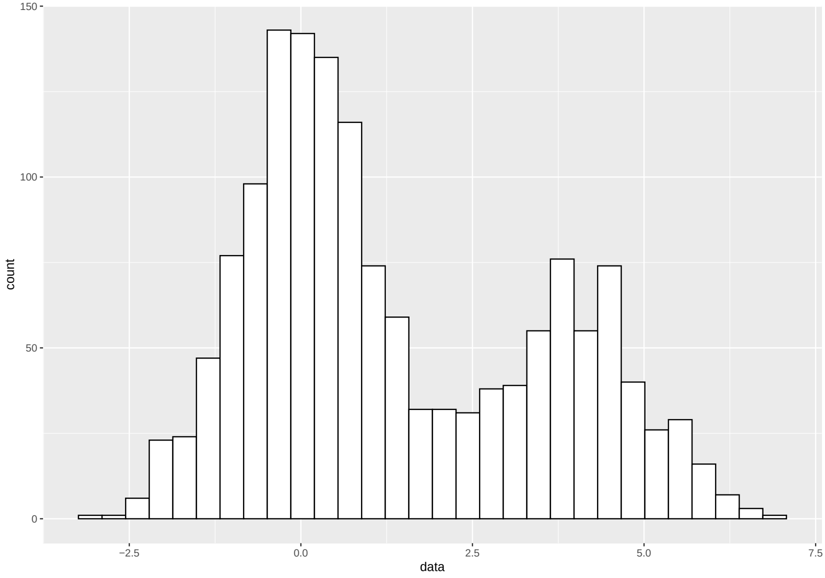
Shape



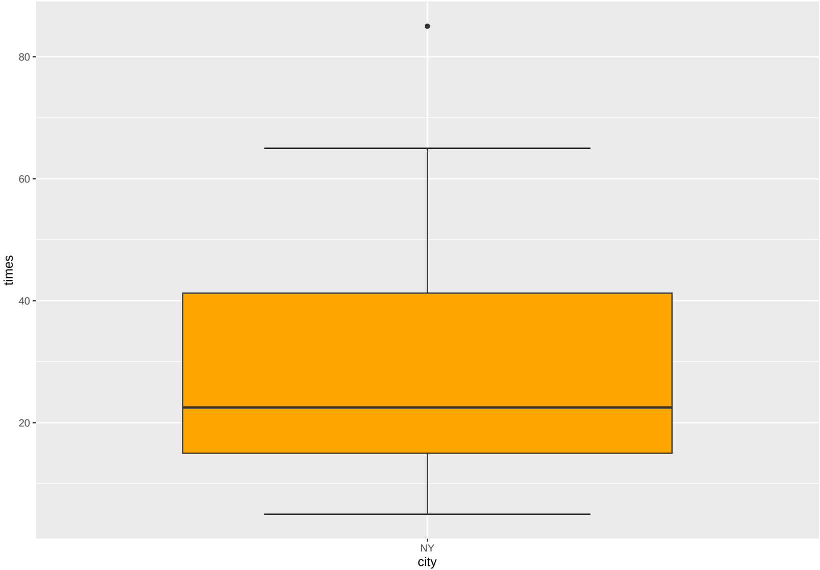
Shape



Shape

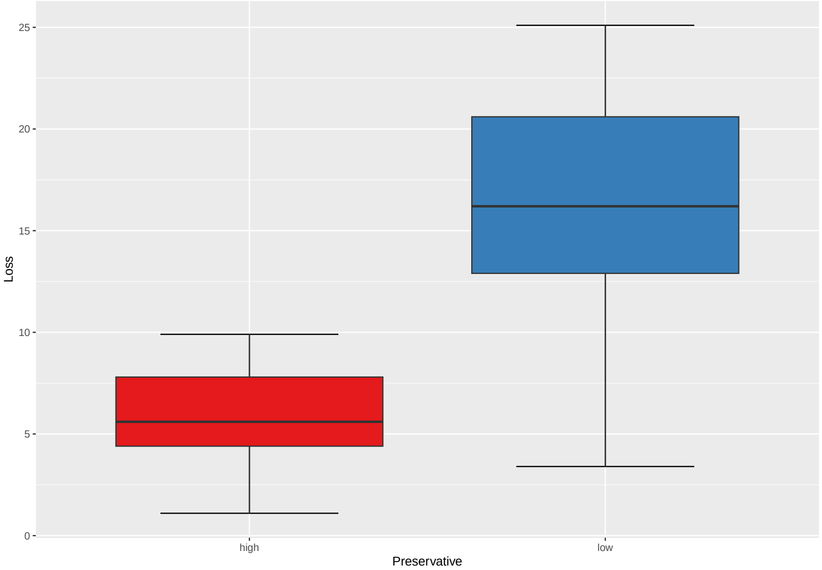


Boxplots



- Mark on the median, Q1, Q3, IQR, and outliers.

Example

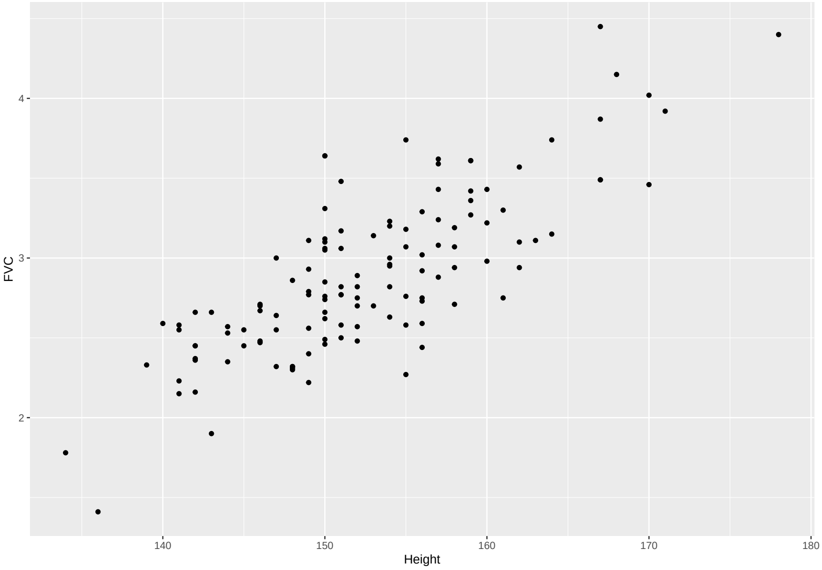


Scatter plots

Each subject is represented by a point, the predictor (explanatory) variable is on the x-axis, while the response variable is on the y-axis.

- Description of the relationship:
  - Negative or positive.
    - Strength (strong, moderate, weak).
    - Linear or curved.
    - Outliers.

Example



- Describe the relationship.

Summary statistics

- Measures of location:
  - Median.
  - Mean - more sensitive to skewness.
- Measures of spread:
  - Standard deviation.
  - IQR.

Example

Consider the numbers: [3, 2, 2, 7, 2, 6, 2.] What is the mean, median, Q_1, Q_3, and IQR?

If I change the 7 to 700, what changes?

Correlation

Lies between -1 and 1. Gives a measure of the linear relationship between the predictor and the response variables.

The correlation squared, r^2**:** The proportion of the variation in the response variable that is explained by the linear relationship with the predictor (explanatory) variable.

"how much of the story x is telling about y"

Experimental design

- Good principles:
  - Control.
    - Randomisation.
    - Replication.
- Other concepts:
  - Placebo.
  - Double-blind.

Probability distribution

These are often tables that tell you the probability of each value.

Probability distributions must follow certain rules to be valid:

- Each probability must lie between 0 and 1.
- The total of the probabilities must be 1.

Example

Value of X	1	2	3	4
Probability	0.1	0.1	0.3	??

What is the missing value?

What is the probability of 3 or more?

Mean from probability mass function

The mean, denoted μ_x is calculated by [$x =$ {all, x_i}x_ip_i,] where p_i is the probability that X is x_i.

Example

Value of X	1	2	3	4
Probability	0.1	0.1	0.3	0.5

What is the value of μ_x?

Transformation of RVs

Sometimes we are given a formula for obtaining a random variable from another. If these are of the form [Y = a + bX,] then we can get μ_Y and σ_Y from μ_X and σ_X:

[_Y = a + b _X, and] [_Y = |b| _X.]

Example

Given [Y $=$ 2 $+$ 3X] and [_X $=$ 2 and _X$=$0.5.] What is μ_Y and σ_Y?

Example

If $Y \sim N(3, 1)$, which of the following R commands calculates the probability that Y is less than or equal to 2, $P(Y \leq 2)$?

- `pnorm(2,3,1)`
- `qnorm(2,3,1)`
- `1 - pnorm(2,3,1)`
- `pbinom(2,3,1)`

Which gives $P(Y \geq 2)$?

Binomial

- (B)inary.
- (I)ndependent.
- (N)umber.
- (S)uccess.

Example

If I toss a fair coin 10 times, which of the following R commands calculates the probability of exactly three heads?

- `pbinom(3,10,0.5)`
- `1 - pbinom(3,10,0.5)`
- `dbinom(3,10,0.5)`
- `qbinom(3,10,0.5)`
- `dbinom(3,10,0.65)`

Revision weeks 7-12

Hypothesis testing framework

- State the null and alternative hypotheses.
- Calculate the value of the test statistic.
- Identify the reference distribution.
- Find the P-value for the observed test statistic.
- Check the assumptions.
- State a conclusion.

One-sample Z-test

- **Scenario** One mean, know σ
- **Null and alternative hypotheses**

$$H_0 : \mu = \mu_0$$
$$H_a : \mu \neq \mu_0$$

- **Test statistic**

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- **Reference distribution**

$$N(0, 1)$$

- P**-value**

$$2 * pnorm(-|z|, 0, 1)$$

- **CI**

$$\bar{x} \pm z^{*}\sigma/\sqrt{n}$$

One-sample t-test

- **Scenario** One mean, do not know σ
- **Null and alternative hypotheses**

$$H_0 : \mu = \mu_0$$
$$H_a : \mu \neq \mu_0$$

- **Test statistic**

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- **Reference distribution**

$$t_{n-1}$$

- P**-value**

$$2 * pt(-|t|, n-1)$$

- **CI**

$$\bar{x} \pm t^* s/\sqrt{n}$$

R one-sample T-test

```
    One Sample t-test

data:  DO
t = -0.94256, df = 14, p-value = 0.3619
alternative hypothesis: true mean is not equal to 5
95 percent confidence interval:
 4.251002 5.291664
sample estimates:
mean of x
 4.771333
```

One-sample proportion

- **Scenario** One proportion
- **Null and alternative hypotheses**

$$H_0 : p = p_0$$
$$H_a : p \neq p_0$$

- **Test statistic**

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- **Reference distribution**

$$N(0, 1)$$

- P**-value**

$$2 * pnorm(-|z|, 0, 1)$$

- **CI**

$$\hat{p} \pm z^* \sqrt{\hat{p}(1 - \hat{p})/n}$$

Two-sample t-test

- **Scenario** Two means
- **Null and alternative hypotheses**

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

- **Test statistic**

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- **Reference distribution**

$$t_{min(n_1-1, n_2-1)}$$

- **P-value**

$$2 * pt(-|t|, min(n_1 - 1, n_2 - 1))$$

- **CI**

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

R two-sample T-test

```
    Welch Two Sample t-test

data:  weight by group
t = 2.4403, df = 32.877, p-value = 0.02023
alternative hypothesis: true difference in means between group Control
95 percent confidence interval:
  1.885639 20.810013
sample estimates:
mean in group Control   mean in group Ozone
             22.34783              11.00000
```

Matched pairs t-test

- **Scenario** Two measurements on each subject
- **Null and alternative hypotheses**

$$H_0 : \mu_D = 0$$
$$H_a : \mu_D \neq 0$$

- **Test statistic**

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}}$$

where $D = X - Y$

- **Reference distribution**

$$t_{n-1}$$

- **P-value**

$$2 * pt(-|t|, n-1)$$

- **CI**

$$\bar{d} \pm t^* s_d/\sqrt{n}$$

R matched-pairs T-test

```
        One Sample t-test

data:  moon$D
t = 6.4518, df = 14, p-value = 1.518e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 1.623968 3.241365
sample estimates:
mean of x
 2.432667
```

Two-sample proportion

- **Scenario** Two proportions
- **Null and alternative hypotheses**

$$H_0 : p_1 = p_2$$
$$H_a : p_1 \neq p_2$$

- **Test statistic**

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$$

- **Reference distribution**

$$N(0,1)$$

- P**-value**

$$2 * pnorm(-|z|, 0, 1)$$

- **CI**

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$$

Chi-square test

- **Scenario** Two categorical random variables measured on each subject.
- **Null and alternative hypotheses**

$$H_0 : \text{no association between the two random variables}$$
$$H_a : \text{an association between the two random variables}$$

- **Test statistic**

$$X^2 = \sum_{\text{all cells}} \frac{(observed - expected)^2}{expected}$$

- **Reference distribution**

$$\chi^2_{(r-1)\times(c-1)}$$

- P**-value**

$$pchisq(X^2, df = (r-1) \times (c-1), lower.tail = FALSE)$$

R Chi-squared test

```
    Pearson's Chi-squared test

data:  mytable
X-squared = 19.763, df = 6, p-value = 0.003051
```

Linear regression (testing for a significant linear relationship)

- **Scenario** Two quantitative random variables measured on each subject.
- **Null and alternative hypotheses**

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

- **Test statistic**

$$t = \frac{b_1}{SE_{b_1}}$$

- **Reference distribution**

$$t_{n-2}$$

- **CI**

$$b_1 \pm t^* SE_{b_1}$$

R linear regression

```
Call:
lm(formula = FVC ~ Height, data = FVC)

Residuals:
     Min       1Q   Median       3Q      Max
-0.75507 -0.23898 -0.00411  0.21238  0.87589

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.064961   0.552593  -9.166 1.24e-15 ***
Height       0.052194   0.003618  14.426  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3137 on 125 degrees of freedom
Multiple R-squared:  0.6248,    Adjusted R-squared:  0.6218
F-statistic: 208.1 on 1 and 125 DF,  p-value: < 2.2e-16
```

One-way ANOVA

- **Scenario** 3 or more means to compare
- **Null and alternative hypotheses**

$$H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_K$$
$$H_a : \text{not all of the } \mu_i\text{s are equal}$$

- **Test statistic**

$$F = \frac{MSM}{MSE}$$

- **Reference distribution**

$$F(K-1, N-K)$$

- P**-value**

$$P(F > f)$$

R ANOVA

```
             Df Sum Sq Mean Sq F value Pr(>F)    
class         6   2295   382.5    45.1 <2e-16 ***
Residuals   227   1925     8.5                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

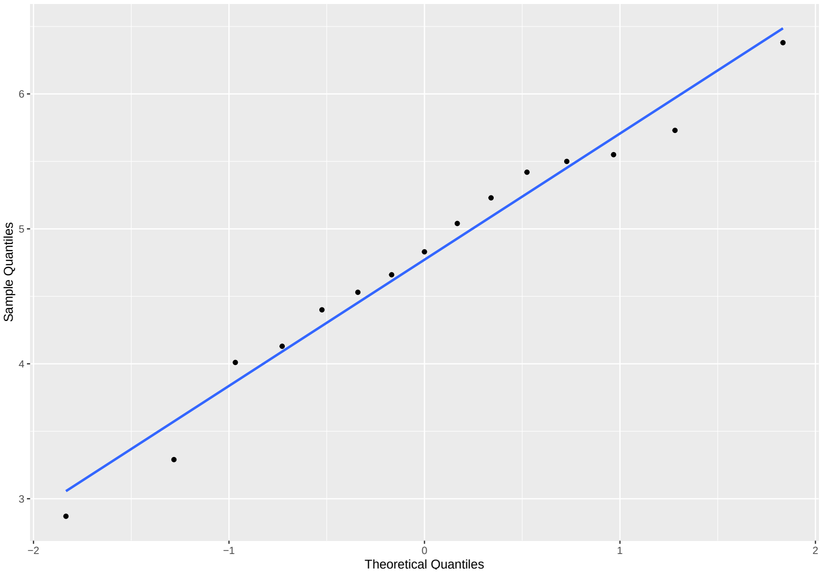
Multiple Comparison

	diff	lwr	upr	p adj
compact-2seater	4.728	0.652	8.803	0.012
midsize-2seater	3.356	-0.748	7.460	0.190
minivan-2seater	0.418	-4.255	5.091	1.000
pickup-2seater	-2.400	-6.558	1.758	0.605
subcompact-2seater	4.971	0.829	9.114	0.008
suv-2seater	-1.900	-5.928	2.128	0.800
midsize-compact	-1.372	-3.223	0.480	0.298
minivan-compact	-4.309	-7.211	-1.408	0.000
pickup-compact	-7.128	-9.095	-5.160	0.000
subcompact-compact	0.244	-1.691	2.178	1.000
suv-compact	-6.628	-8.303	-4.952	0.000
minivan-midsize	-2.938	-5.880	0.004	0.051
pickup-midsize	-5.756	-7.782	-3.730	0.000
subcompact-midsize	1.615	-0.379	3.609	0.199
suv-midsize	-5.256	-7.000	-3.512	0.000
pickup-minivan	-2.818	-5.835	0.198	0.084
subcompact-minivan	4.553	1.558	7.548	0.000
suv-minivan	-2.318	-5.153	0.516	0.190
subcompact-pickup	7.371	5.269	9.474	0.000
suv-pickup	0.500	-1.367	2.367	0.985
suv-subcompact	-6.871	-8.703	-5.040	0.000

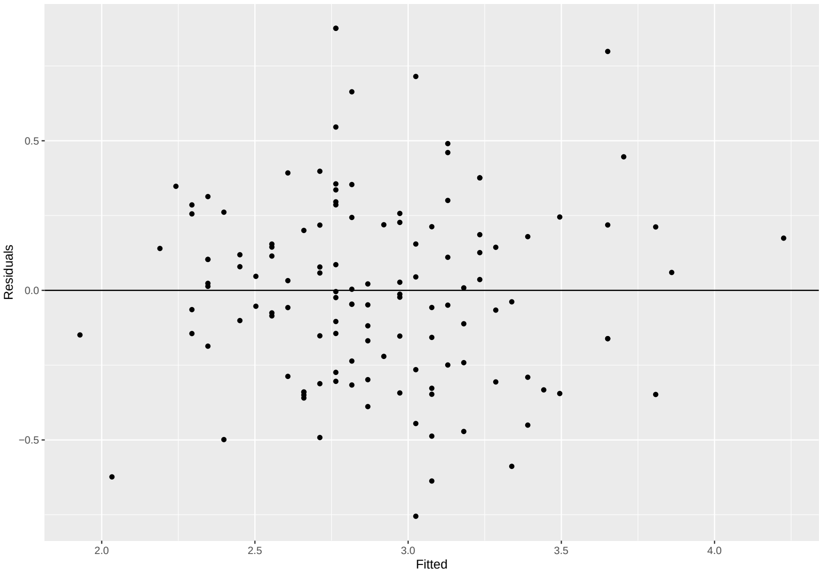
Checking the assumptions

- What is the assumption?
- Where do you look to check the assumptions?
- What do you expect to see if the assumption is valid?
- What do you see?
- What is your conclusion?

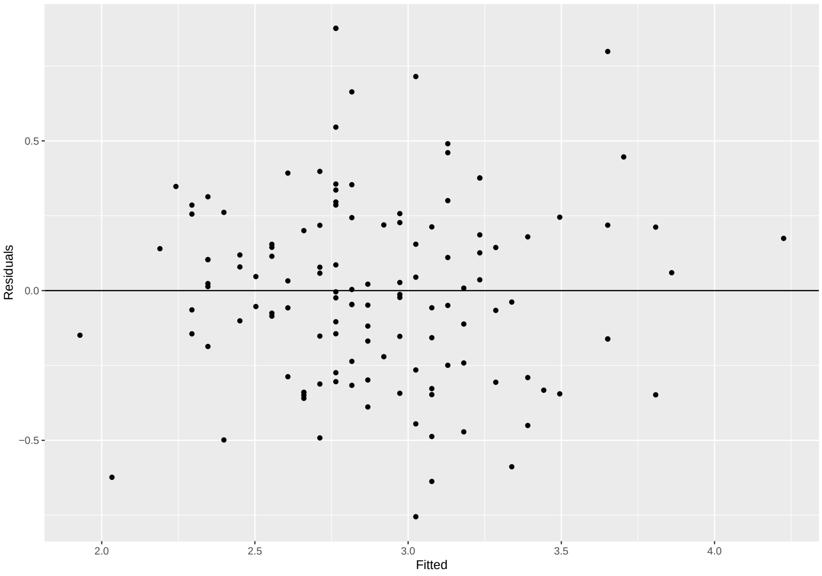
Normality



Linearity



Constant spread



Independence

Check experimental design.