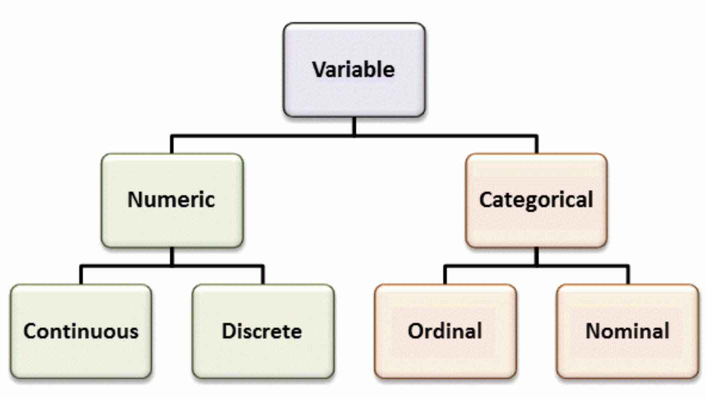
Workshop – Introduction into R

Statistical Analysis

Andreas Limacher

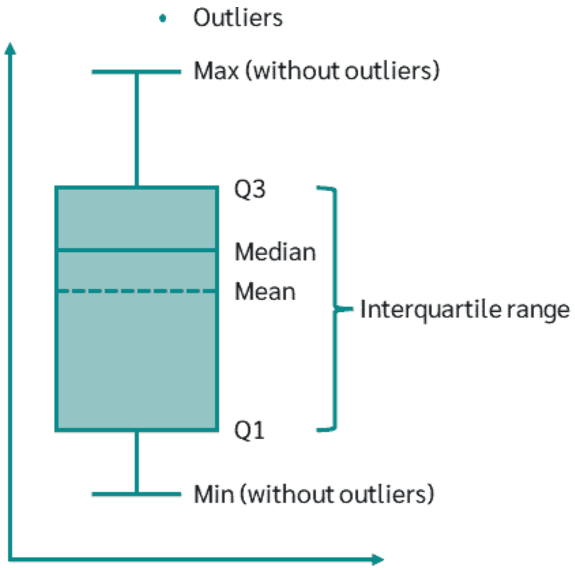
Descriptive statistics

- What measures do you use to summarize data?
  - Categorical variables
  - Numeric variables



Descriptive statistics

- Continuous variables
  - Mean, standard deviation, confidence interval
  - Median, 1^{st} quartile, 3^{rd} quartile, interquartile range
- Categorical variables
  - Numbers, proportions
- Correlation
  - Pearson and Spearman correlation



Descriptive statistics – Categorical

- Tabulation and cross-tabulation

```
> table(NHANES$Education, useNA = "always")
> table(NHANES$Education, NHANES$Gender, useNA = "always")
```

- Proportions

```
> prop.table(table(NHANES$Education))
> prop.table(table(NHANES$Education, NHANES$Gender), 2)
```

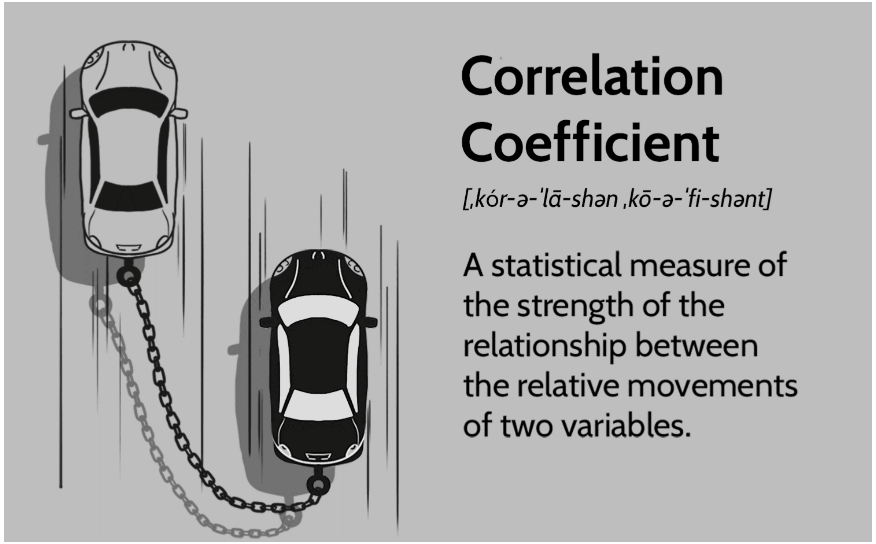
Descriptive statistics – Numeric

- Mean, sd, CI, median, and quartiles using tidyverse

```
> age_summary <- NHANES %>%
    summarise(
      mean_age = mean(Age, na.rm = TRUE),
      sd_age = sd(Age, na.rm = TRUE),
      n = sum(!is.na(Age)),
      ci_lower = mean_age - qnorm(0.975) * sd_age / sqrt(n),
      ci_upper = mean_age + qnorm(0.975) * sd_age / sqrt(n),
      median_age = median(Age, na.rm = TRUE),
      q1 = quantile(Age, 0.25, na.rm = TRUE),
      q3 = quantile(Age, 0.75, na.rm = TRUE) )
```

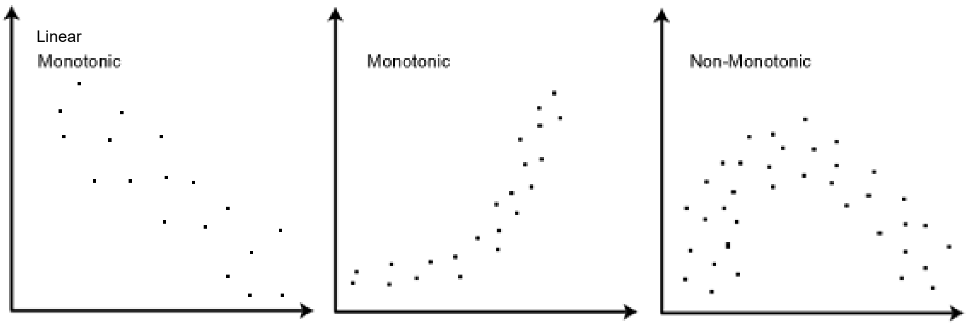
Correlation – Assumptions

- Which correlation coefficients can be used for numeric data?
- What are the assumptions?



Correlation – Assumptions

- Pearson correlation
  - Normal distribution (approximately)
  - Linear relationship between two variables
  - Observations are independent of each other
- Spearman correlation
  - Monotonic relationship between two variables
  - Observations are independent of each other



Correlation – R code

- Pearson correlation

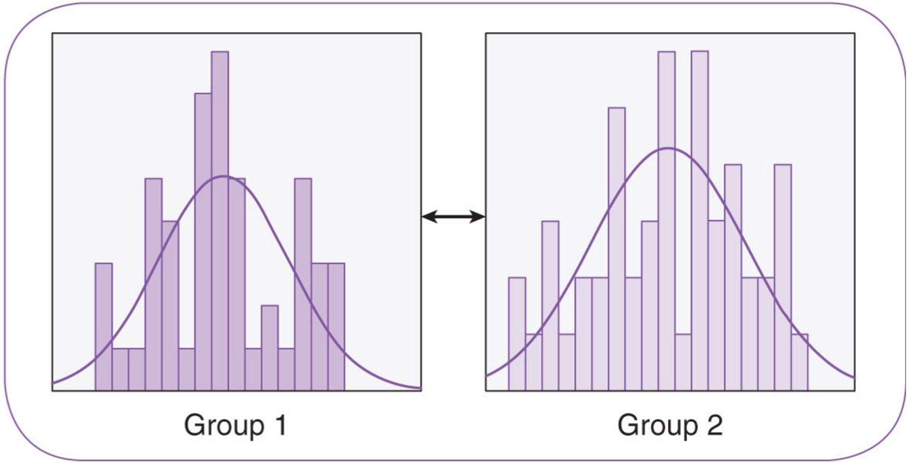
```
> NHANES %>%
    select(Age, BMI) %>%
    drop_na() %>%
    summarise(correlation = cor(Age, BMI, method = "pearson"))
```

- Spearman correlation

```
> NHANES %>%
    select(Age, BMI) %>%
    drop_na() %>%
    summarise(correlation = cor(Age, BMI, method = "spearman"))
```

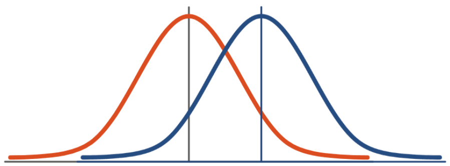
Common statistical tests

- Which tests do you know?
- When are they used?

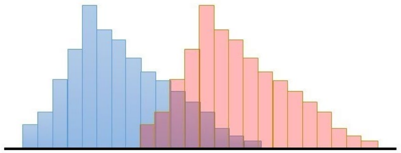


Common statistical tests - Numeric

- Normal assumption
  - T-test (one-group, 2 independent groups)
  - Paired t-test (2 paired groups)
  - ANOVA (multiple independent groups)
  - Repeated measure ANOVA (multiple paired groups)



- Non-parametric (no normal assumption)
  - Wilcoxon rank sum test / Mann-Whitney-U test (2 independent groups)
  - Wilcoxon signed rank test (2 paired groups)
  - Kruskal-Wallis test (multiple independent groups)
  - Friedman test (multiple paired groups)



Common statistical tests – Numeric normal

- One-sample t-test (based on average US population age)

```
> t.test(NHANES$Age, mu = 38.5)
```

- Two-sample t-test

```
> t.test(Age ~ Gender, data = NHANES)
```

- Paired t-test

```
> t.test(NHANES$BPSys1, NHANES$BPSys2, paired = TRUE)
```

- ANOVA

```
> summary(aov(Age ~ Education, data = NHANES))
```

Common statistical tests – Numeric non-parametric

- Wilcoxon ranksum test

```
> wilcox.test(Age ~ Gender, data = NHANES)
```

- Wilcoxon signed rank test

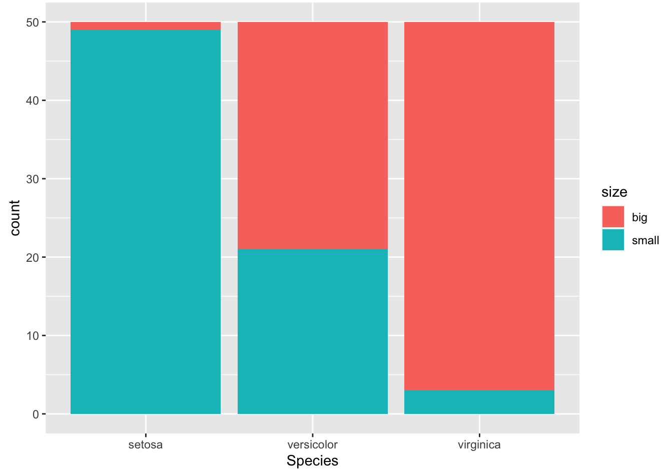
```
> wilcox.test(NHANES$BPSys1, NHANES$BPSys2, paired = TRUE)
```

- Kruskal-Wallis test

```
> kruskal.test(Age ~ Education, data = NHANES)
```

Common statistical tests - Categorical

- Categorical data
  - Chi-square test (independent groups)
  - Fisher's exact test (independent groups, sparse data)
  - McNemar's test (paired data)



Common statistical tests - Categorical

- Chi-square test

```
> chisq.test(table(NHANES$Gender, NHANES$Education))
```

- Fisher's exact test

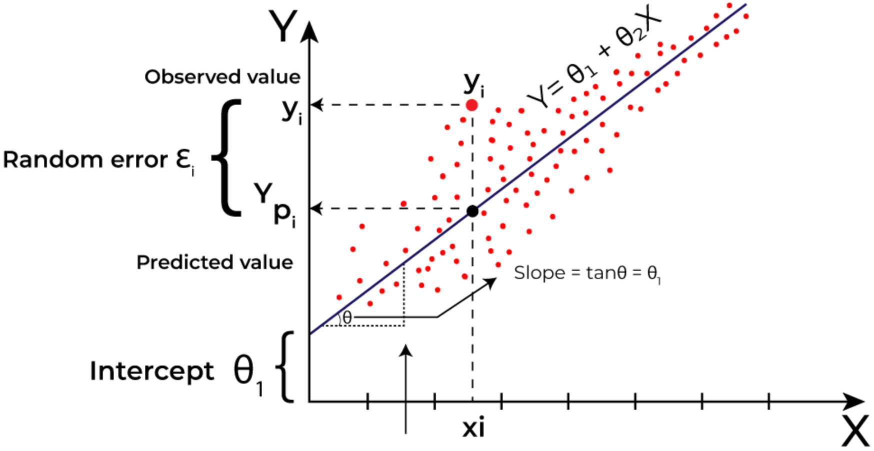
```
> fisher.test(table(NHANES$Gender[1:50], NHANES$Education[1:50]))
```

- McNemar test for high blood pressure

```
> bp_high1 <- NHANES$BPSys1 > 140
> bp_high2 <- NHANES$BPSys2 > 140
> mcnemar.test(bp_high1, bp_high2)
```

Linear regression

- Is age associated with BMI? Is there an influence of age on BMI?



Linear regression

```
> model <- lm(BMI ~ Age, data = NHANES)
> summary(model)
```

How do you interpret the model output?

```
Residuals:
    Min      1Q  Median      3Q     Max 
-16.794  -4.803  -1.236   3.466  55.697 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 21.451688   0.137193  156.36   <2e-16 ***
Age          0.138033   0.003148   43.84   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.735 on 9632 degrees of freedom
Multiple R-squared:  0.1664,	Adjusted R-squared:  0.1663 
F-statistic:  1922 on 1 and 9632 DF,  p-value: < 2.2e-16
```

```
> confint(model, level = 0.95)
```

Linear regression

```
> model <- lm(BMI ~ Age, data = NHANES)
> summary(model)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-16.794  -4.803  -1.236   3.466  55.697

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 21.451688   0.137193  156.36   <2e-16 ***
Age          0.138033   0.003148   43.84   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.735 on 9632 degrees of freedom
Multiple R-squared:  0.1664,	Adjusted R-squared:  0.1663

F-statistic:  1922 on 1 and 9632 DF,  p-value: < 2.2e-16
```

Difference between observed and predicted values

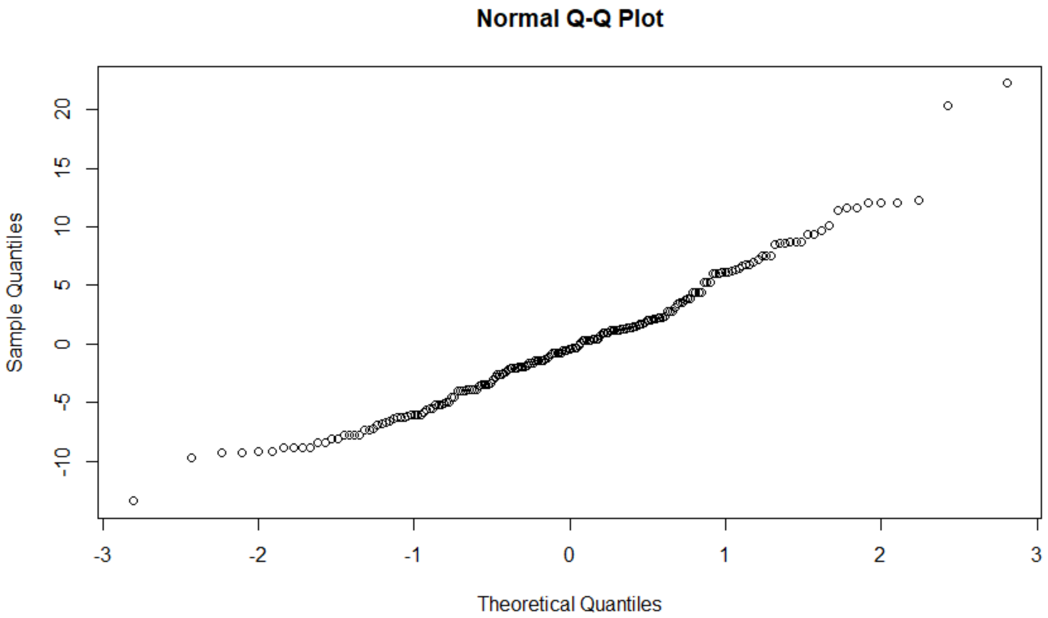
Change in dependent variable

BMI increases by 0.14 if age increases by 1 year

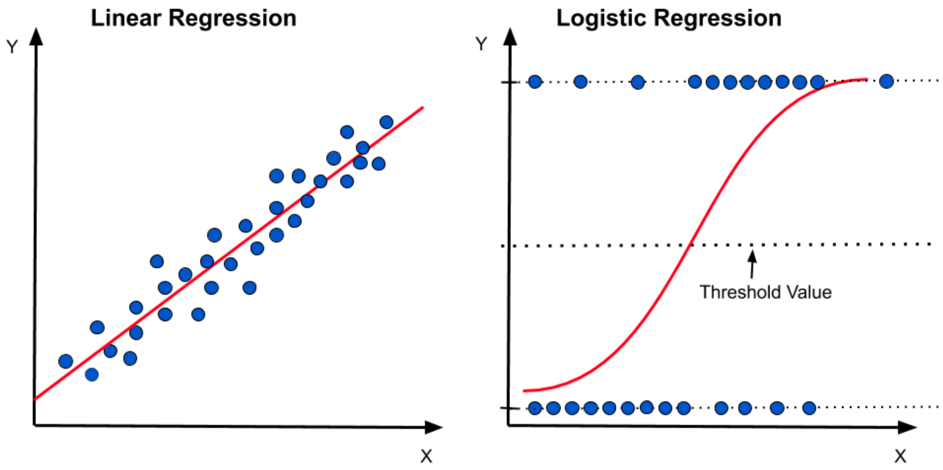
Variance explained by the model (%)
(adjusted for number of predictors)

Linear regression – Normal Q-Q Plot

```
> qqnorm(residuals(model))
```



Is age associated with obesity?



Logistic model

```
> model <- glm(HighBMI ~ Age_centered, data = NHANES, family = binomial)
> summary(model)
```

How do you interpret the model output?

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.988679   0.023903  -41.36   <2e-16 ***
Age_centered  0.023872   0.001084   22.01   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null deviance: 11552  on 9633  degrees of freedom
Residual deviance: 11040  on 9632  degrees of freedom
AIC: 11044
```

```
> exp(cbind(OR = coef(model), confint(model)))

OR       2.5 %     97.5 %
(Intercept)  0.372068 0.3549663 0.3898363
Age_centered 1.024159 1.0219920 1.0263459
```

Logistic model

```
> model <- glm(HighBMI ~ Age_centered, data = NHANES, family = binomial)
> summary(model)
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.988679   0.023903  -41.36   <2e-16 ***
Age_centered  0.023872   0.001084   22.01   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null deviance: 11552  on 9633  degrees of freedom
Residual deviance: 11040  on 9632  degrees of freedom
AIC: 11044
```

Change on logit-scale

Unexplained variation (lower values indicate a better model fit)

Model fit adjusted for number of predictors

```
> exp(cbind(OR = coef(model), confint(model)))
```

```
OR             2.5 %     97.5 %
(Intercept)  0.372068 0.3549663 0.3898363
Age_centered 1.024159 1.0219920 1.0263459
```

Relative change expressed as odds ratio

Exercise