Linear Models II

Statistics with R

Basel R Bootcamp

April 2019

Linear Model Applications

Linear Model Equation

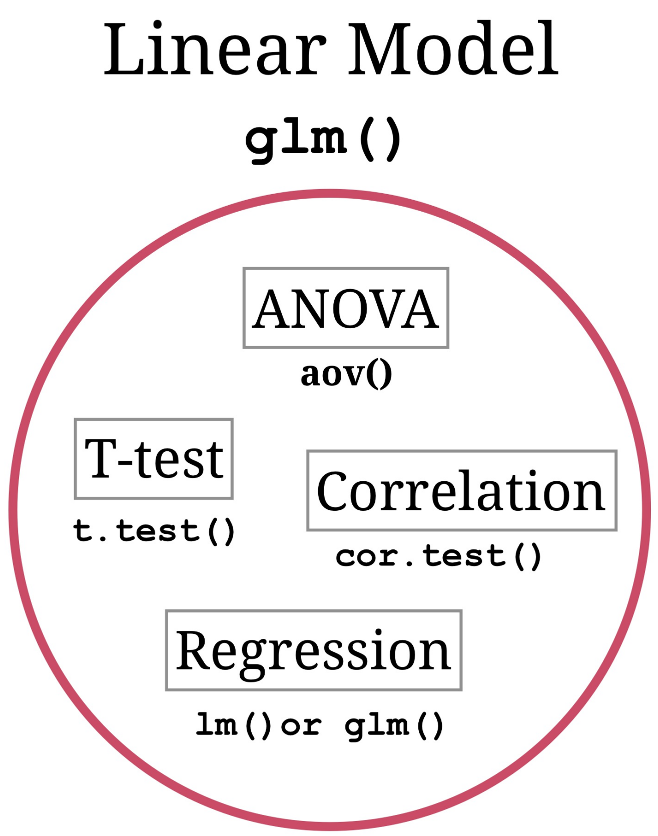
```
$$\Large y = \beta_{0} + \beta_{1}x_{1} +
\beta_{2} x_{2} + ... + \beta_{n}x_{n} +
\epsilon$$
```

Hypothesis tests are linear models!

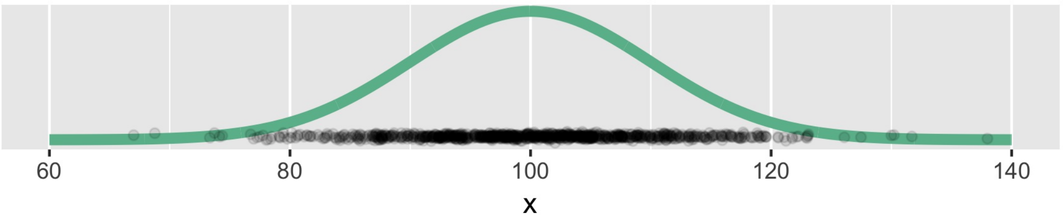
In fact, many of your favorite hypothesis tests, including **t-tests**, **correlation tests**, and **ANOVAs** can all be expressed as linear models!

This means that you can use the \`lm()\` or \`glm()\` function to do all of these tests!

However, R also has special **hypothesis test functions** with more user-friendly outputs.



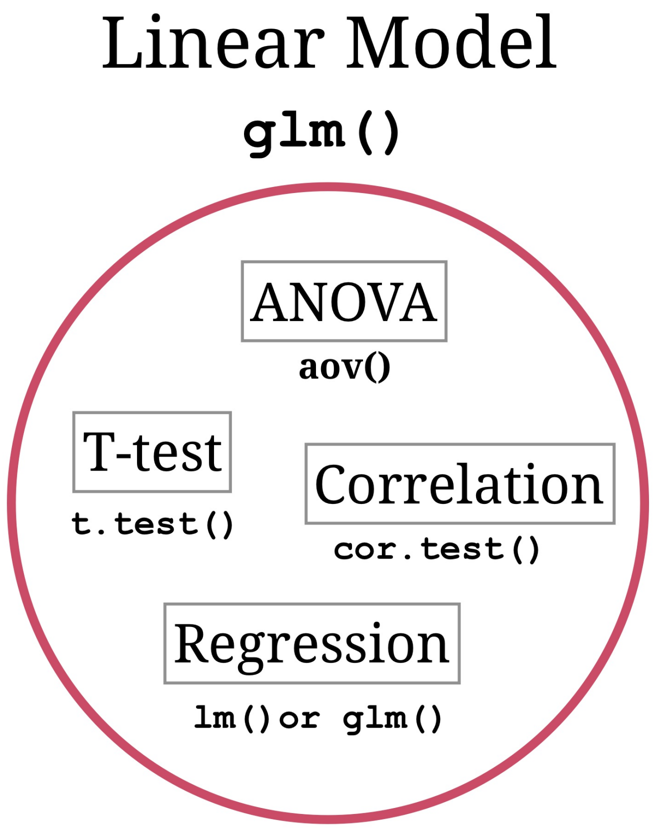
Linear Model Applications



Many of these tests assume your dependent variable is **normally distributed**. What differentiates these tests is typically the **scale of your independent variable**.

Types of predictor variables

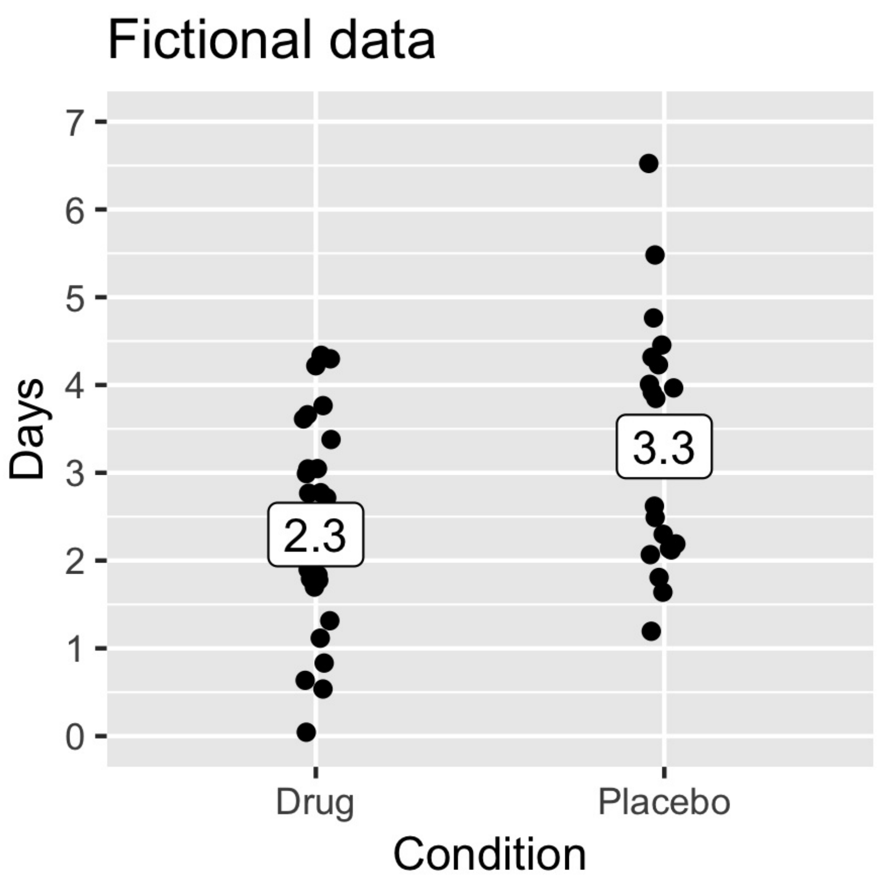
Scale	Description	Examples
Nominal	A discrete category without order	Sex, College, Favorite Color
Ratio	A continuous number	Income, Height, Weight



Null hypothesis testing

Null hypothesis testing is a statistical framework where one hypothesis (H_0) is tested to defend the other, alternative hypothesis (H_1).

Hypothesis	**Description**	**Example**
Null (H_0)	A proposed effect **does not exist** and variation **is not systematic**.	Drug and placebo have the same effect.
Alternative (H_1)	A proposed effect **does exist** and variation **is systematic**	Drug and placebo do *not* have the same effect



Correlation test

Does Y tend to change when X changes?

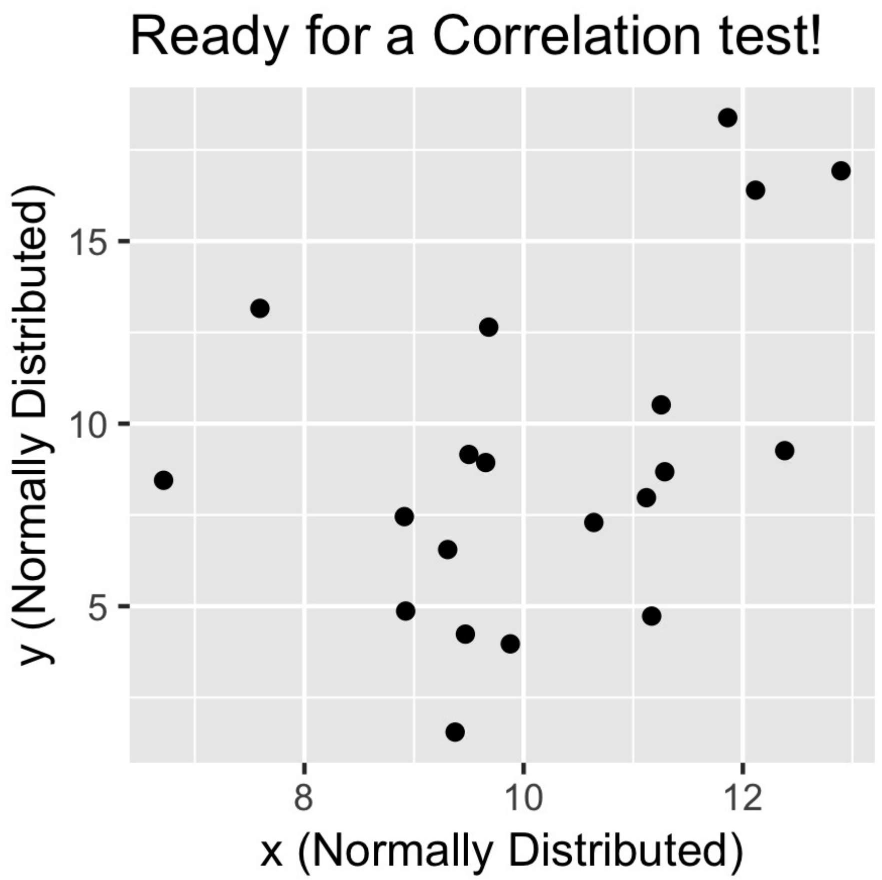
Conduct a **correlation test** when you have 2 continuous, Normally distributed independent variables X and Y.

Formula

$$\LARGE Y=\beta_{0}+\beta_{1}x$$

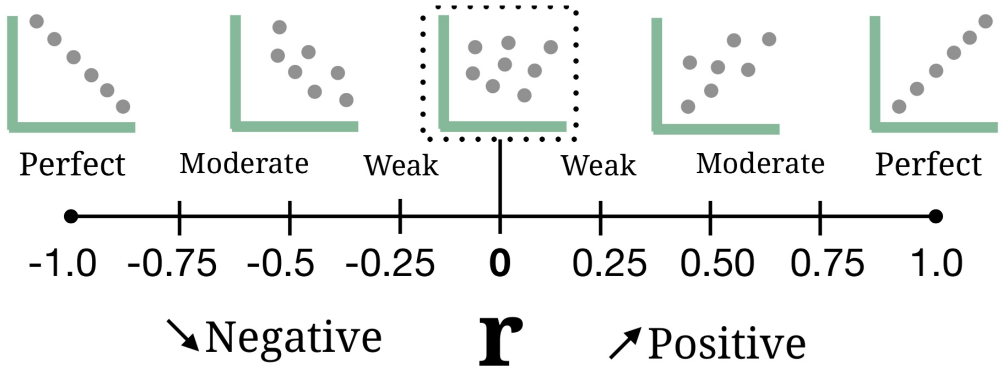
$$\LARGE \beta_{1}=\rho\frac{\sigma_{y}}{\sigma_{x}}$$

ρ	The **population correlation** between x and Y
r	The **sample correlation** between x and Y



Correlation test

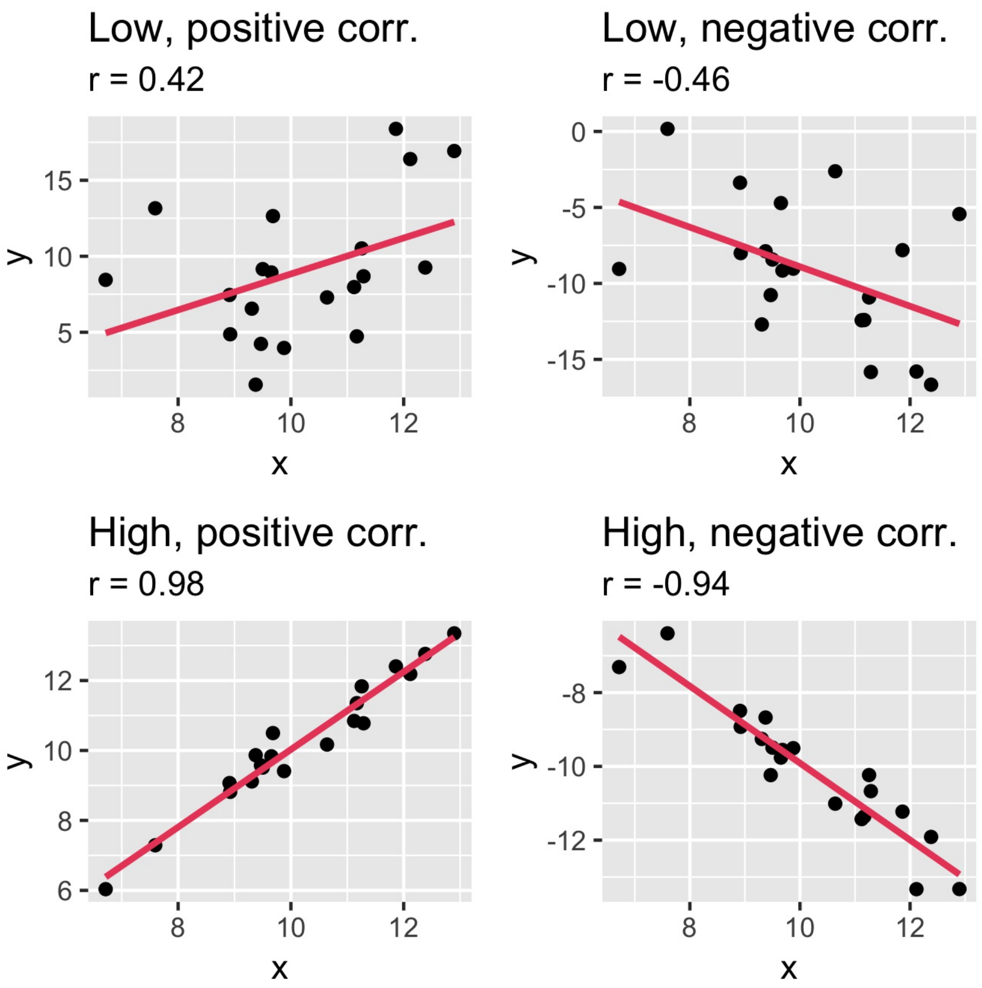
Correlation Coefficient



Hypotheses

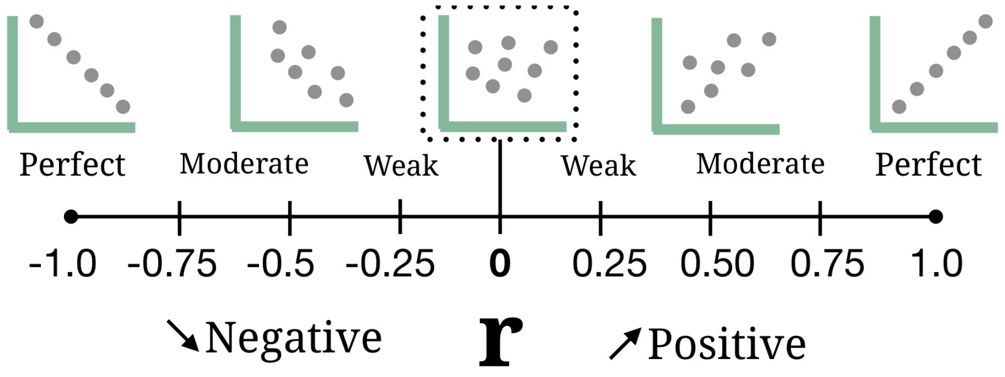
Null: $H_{0}:\rho = 0$, "There **is no** correlation in the population"

Alternative: $H_{A}:\rho \neq 0$, "There **is a** (non-zero) correlation in the population!"



Correlation test

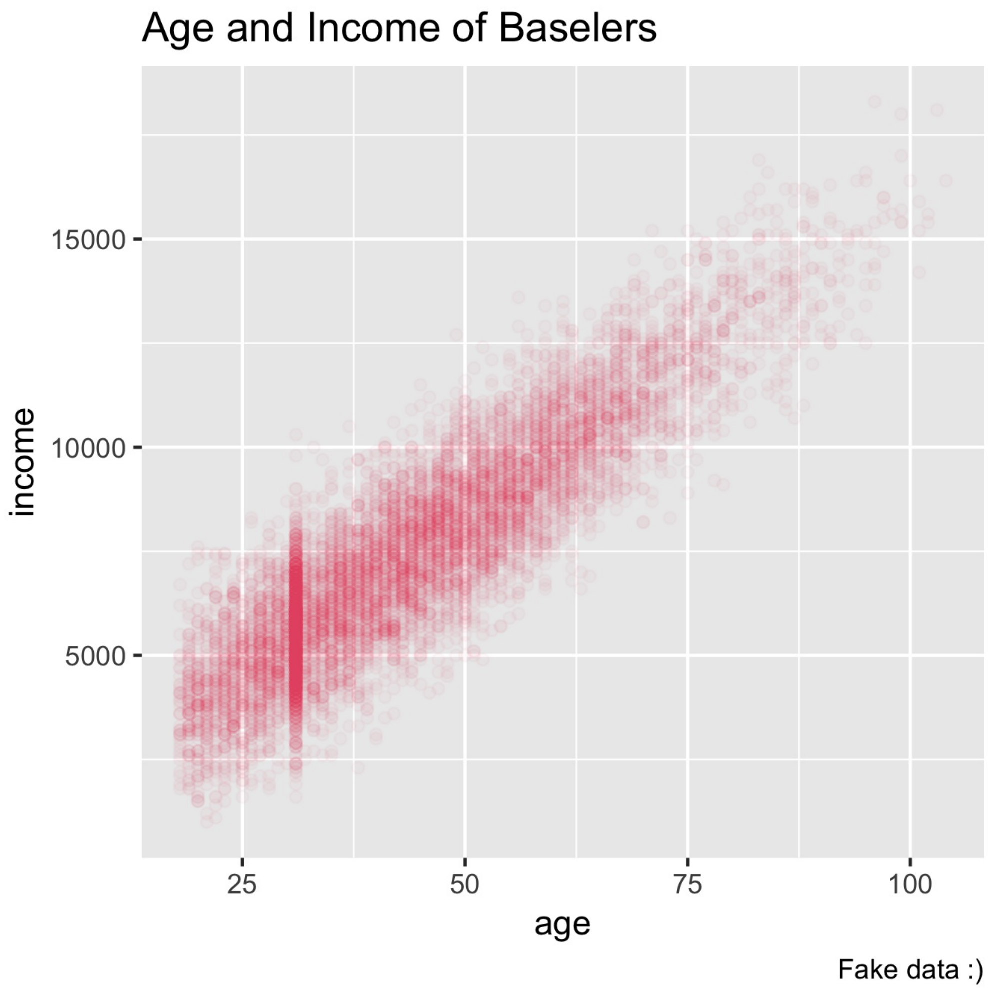
Correlation Coefficient



Hypotheses

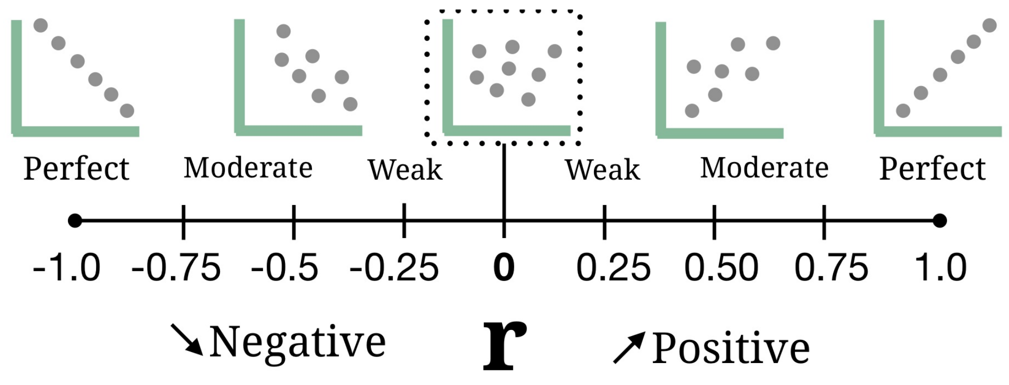
Null: $H_{0}:\rho = 0$, "There **is no** correlation in the population"

Alternative: $H_{A}:\rho \neq 0$, "There **is a** (non-zero) correlation in the population!"



Correlation test

Correlation Coefficient



Hypotheses

Null: $H_{0}:\rho = 0$, "There **is no** correlation in the population"

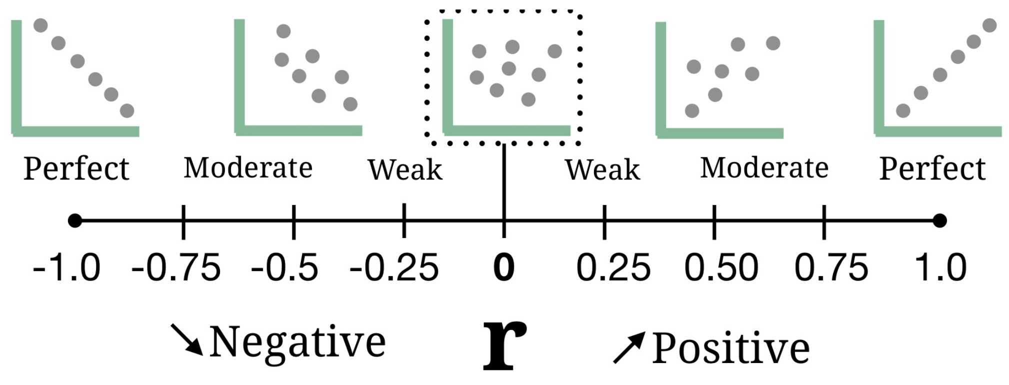
Alternative: $H_{A}:\rho \neq 0$, "There **is a** (non-zero) correlation in the population!"

```
# Relationship between age and income?

inc_ht <- cor.test(formula = ~ age + income,
                   data = baselers)

# Print result
inc_ht
```

```
##
##      Pearson's product-moment correlation
##
## data:  age and income
## t = 180, df = 8500, p-value <2e-16
## alternative hypothesis: true correlation is not equ
## 95 percent confidence interval:
##  0.8882 0.8968
## sample estimates:
##    cor
## 0.8926
```

Correlation test

Correlation Coefficient



Hypotheses

Null: \(H_{0}:\rho = 0\), "There **is no** correlation in the population"

Alternative: \(H_{A}:\rho \neq 0\), "There **is a** (non-zero) correlation in the population!"

```
# Show all named elements
names(inc_ht)
```

```
## [1] "statistic"   "parameter"
## [3] "p.value"     "estimate"
## [5] "null.value"  "alternative"
## [7] "method"      "data.name"
## [9] "conf.int"
```

```
# Show estimated correlation coefficient
inc_ht$estimate
```

```
##    cor
## 0.8926
```

```
# Show p-value
inc_ht$p.value
```

```
## [1] 0
```

t-test

Does the mean of group A differ from group B?

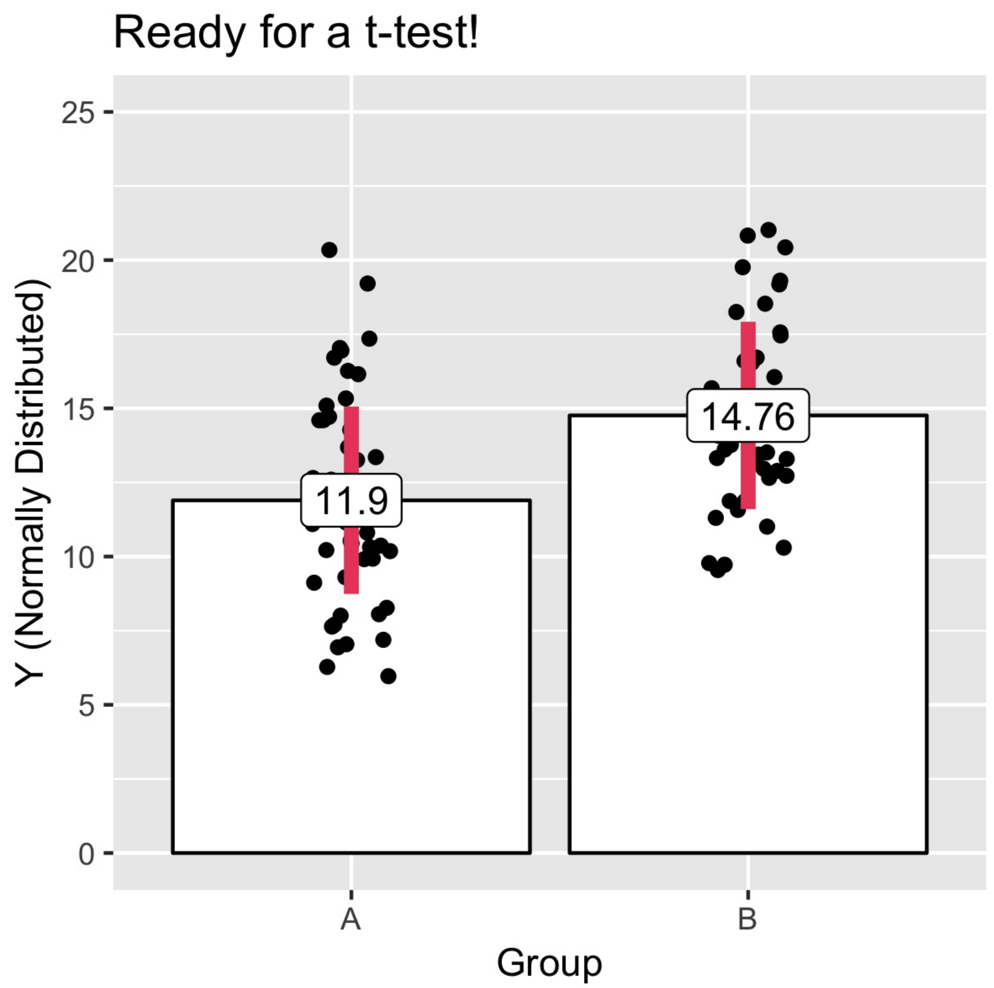
Conduct a t-test when you have one nominal independent variable with 2 levels A and B

Formula

$$\LARGE Y=\beta_{0}+\beta_{1}x$$

$$\beta_{0} = Mean\;of\;group\;A, \beta_{1} = Difference\;between \;groups$$

Group	x
Group = A	x = 0
Group = B	x = 1



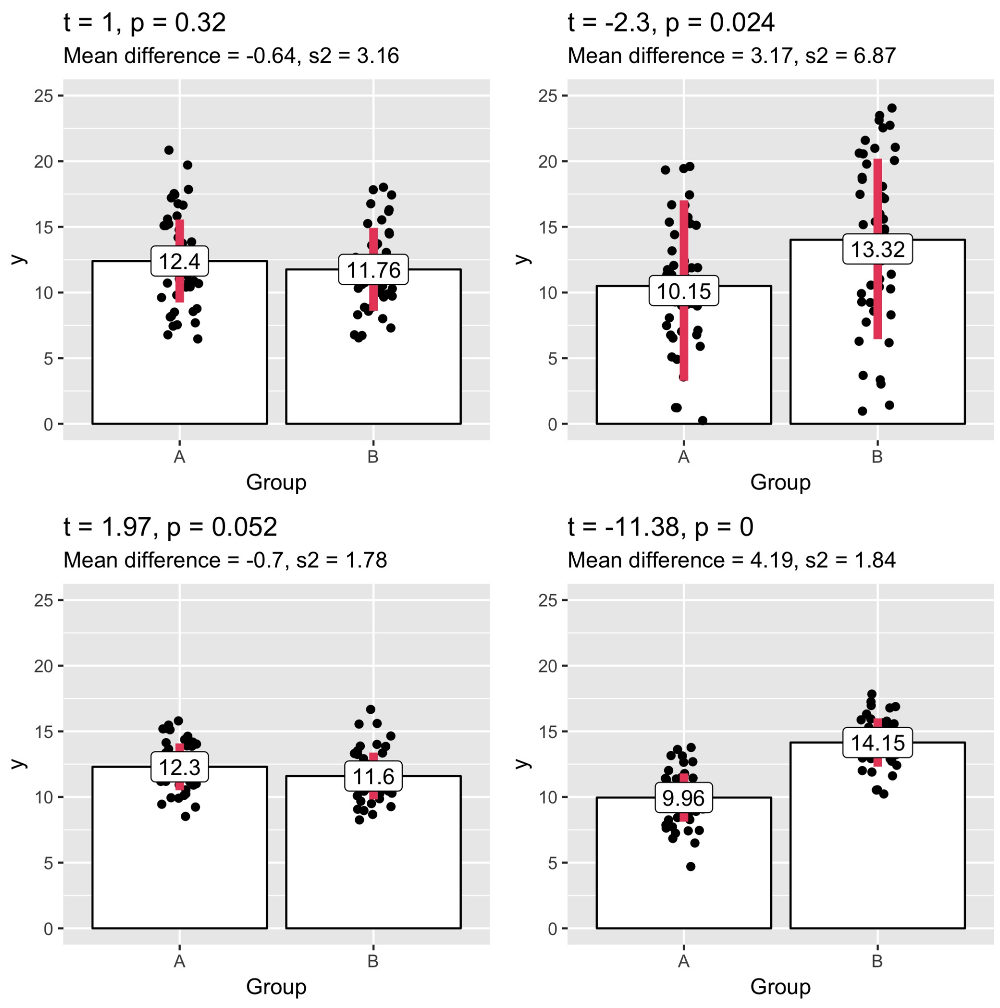
t-test

Does the mean of group A differ from group B?

Conduct a **t-test** when you have one nominal independent variable with **2 levels** A and B

Formula

$$\Large t = \frac{\bar{x}_{A}-\bar{x}_{B}}{\sqrt{s^{2}(\frac{1}{n_{A}}+\frac{1}{n_{B}})}}$$ $$\Large s^{2} = Pooled\;variance$$



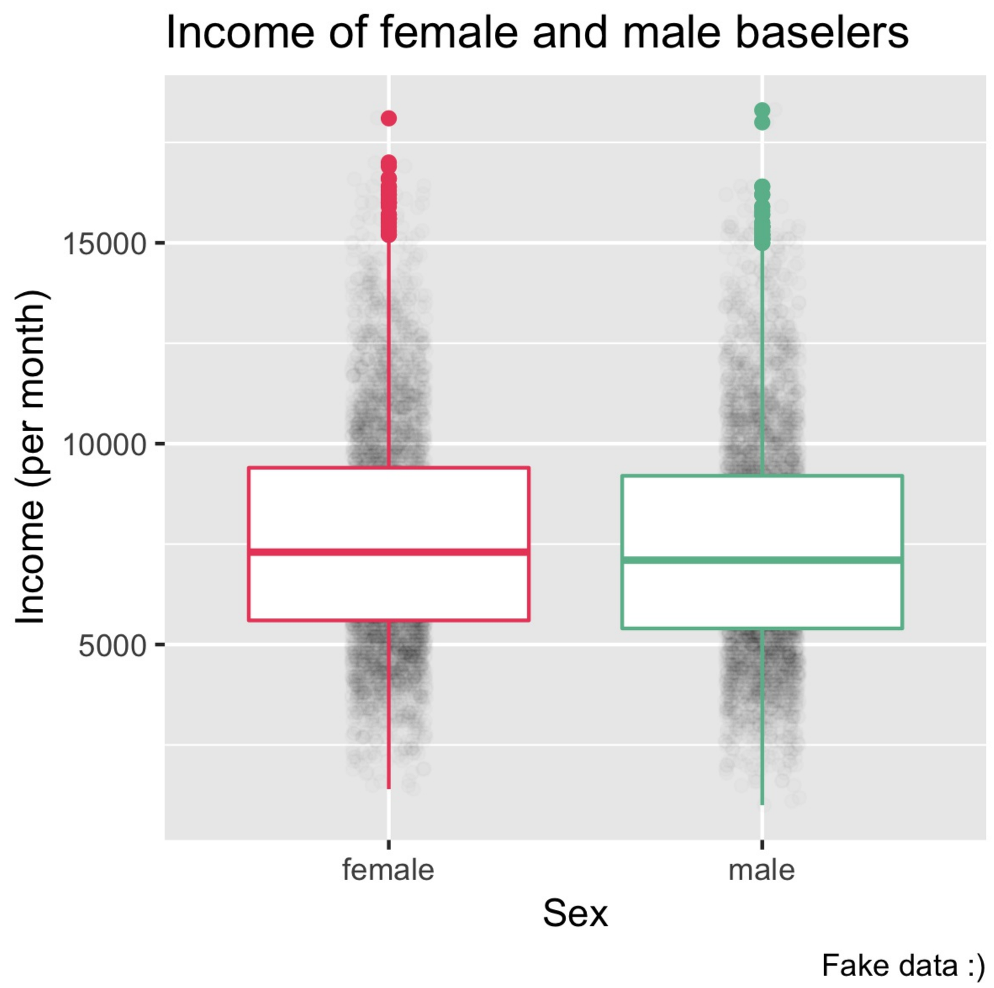
t-test

Does the mean of group A differ from group B?

Conduct a **t-test** when you have one nominal independent variable with **2 levels** A and B

Formula

$$\Large t = \frac{\bar{x}_{A}-\bar{x}_{B}}{\sqrt{s^{2}(\frac{1}{n_{A}}+\frac{1}{n_{B}})}}$$ $$\Large s^{2} = Pooled\;variance$$



t-test

Does the mean of group A differ from group B?

Conduct a **t-test** when you have one nominal independent variable with **2 levels** A and B

Formula

$$\Large t = \frac{\bar{x}_{A}-\bar{x}_{B}}{\sqrt{s^{2}(\frac{1}{n_{A}}+\frac{1}{n_{B}})}}$$ $$\Large s^{2} = Pooled\;variance$$

```
# 2-sample t-test
inc_ht <- t.test(formula = income ~ sex,
                 data = baselers)

# Print
inc_ht
```

```
##
##     Welch Two Sample t-test
##
## data:  income by sex
## t = 4, df = 8500, p-value = 6e-05
## alternative hypothesis: true difference in means is
## 95 percent confidence interval:
##  120.6 352.2
## sample estimates:
## mean in group female   mean in group male
##                 7650                 7414
```

t-test

Does the mean of group A differ from group B?

Conduct a **t-test** when you have one nominal independent variable with **2 levels** A and B

Formula

$$\Large t = \frac{\bar{x}_{A}-\bar{x}_{B}}{\sqrt{s^{2}(\frac{1}{n_{A}}+\frac{1}{n_{B}})}}$$ $$\Large s^{2} = Pooled\;variance$$

```
# Show all named elements
names(inc_ht)
```

```
## [1] "statistic"   "parameter"
## [3] "p.value"     "conf.int"
## [5] "estimate"    "null.value"
## [7] "alternative" "method"
## [9] "data.name"
```

```
# Print the test statistic
inc_ht$statistic
```

```
##     t
## 4.001
```

```
# Print the p.value
inc_ht$p.value
```

```
## [1] 6.366e-05
```

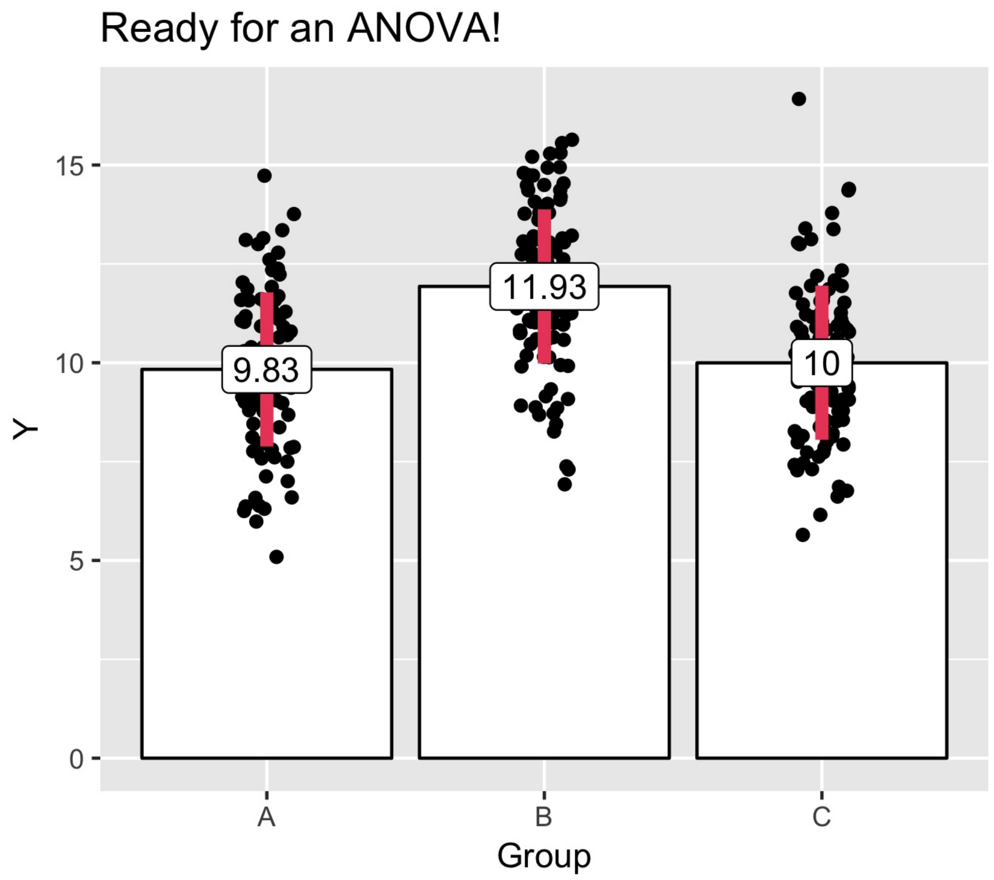
ANOVA

Do the means of my (many) groups differ?

Conduct an **ANOVA** when you have one nominal independent variable with **more than 2 levels** A, B, C, ...

Formula

$$\Large F = \frac{Variance\;Between\;Groups}{Variance\;Within\;Groups}$$



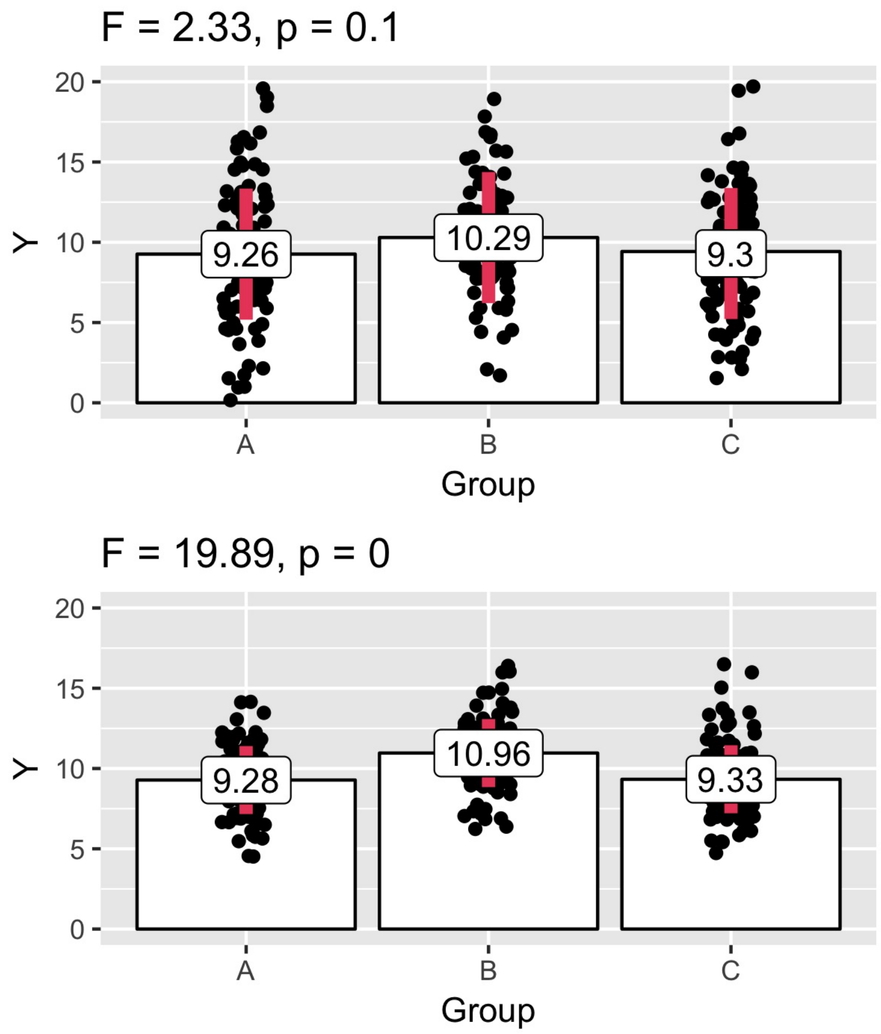
ANOVA

Do the means of my (many) groups differ?

Conduct an **ANOVA** when you have one nominal independent variable with **more than 2 levels** A, B, C, …

Formula

$$\Large F = \frac{Variance\;Between\;Groups}{Variance\;Within\;Groups}$$



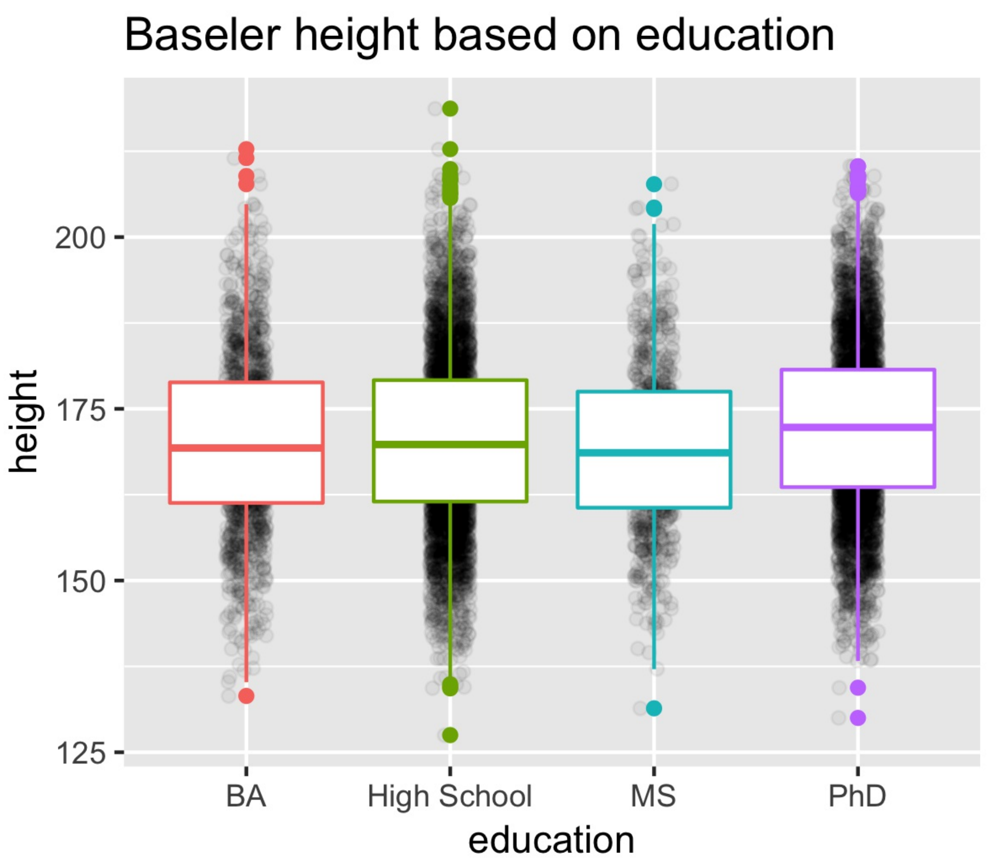
ANOVA with `aov()`

Do the means of my (many) groups differ?

Conduct an **ANOVA** when you have one nominal independent variable with **more than 2 levels** A, B, C, ...

Formula

```
$$\Large F =
\frac{Variance\;Between\;Groups}
{Variance\;Within\;Groups}$$
```



Fake data :)

ANOVA with `aov()`

Do the means of my (many) groups differ?

Conduct an **ANOVA** when you have one nominal independent variable with **more than 2 levels** A, B, C, …

Formula

```
$$\Large F =
\frac{Variance\;Between\;Groups}
{Variance\;Within\;Groups}$$
```

```r
# Relationship height and education?
height_ht <-
  aov(formula = height ~ education,
                data = baselers)

# Print result
height_ht
```

```
## Call:
##    aov(formula = height ~ education, data = baseler
##
## Terms:
##                    education Residuals
## Sum of Squares        12440   1582357
## Deg. of Freedom           3      9996
##
## Residual standard error: 12.58
## Estimated effects may be unbalanced
```

ANOVA with aov()

Do the means of my (many) groups differ?

Conduct an **ANOVA** when you have one nominal independent variable with **more than 2 levels** A, B, C, …

Formula

$$\Large F = \frac{Variance\;Between\;Groups}{Variance\;Within\;Groups}$$

```
# Show summary results
summary(height_ht)
```

```
##                Df  Sum Sq Mean Sq F value
## education       3   12440    4147    26.2
## Residuals    9996 1582357     158        
##              Pr(>F)    
## education    <2e-16 ***
## Residuals              
## ---
## Signif. codes:  
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
##   0.1 ' ' 1
```

ANOVA post-hoc with TukeyHSD()

Which groups differ?

After conducting an **ANOVA**, conduct a **post-hoc test** to see which specific pairs of groups differ .

Formula

$$\Large t = \frac{\bar{x}_{A}-\bar{x}_{B}}{Total\;Variability}$$

```
# Conduct post-hoc tests
#   Which pairs of groups differ?

TukeyHSD(height_ht)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = height ~ education, data = baselers)
##
## $education
##                       diff     lwr    upr
## High School-BA      0.2708 -0.7808 1.3223
## MS-BA              -0.7759 -2.2772 0.7254
## PhD-BA              2.2936  1.2435 3.3436
## MS-High School     -1.0467 -2.3400 0.2466
## PhD-High School     2.0228  1.3008 2.7448
## PhD-MS              3.0695  1.7774 4.3616
##                     p adj
## High School-BA     0.9115
## MS-BA              0.5451
## PhD-BA             0.0000
## MS-High School     0.1599
## PhD-High School    0.0000
## PhD-MS             0.0000
```

tidy()

The `tidy()` function from the `broom` package **converts** the most important results of many statistical objects to a **data frame**.

Try `tidy()` on your favorite statistical object and see what you get!

```
# Load broom package
library(broom)    # For tidy()

# Conduct correlation test
income_htest <- cor.test(formula = ~ height + income,
                         data = baselers)

# Standard printout
income_htest
```

```
##
##      Pearson's product-moment correlation
##
## data:  height and income
## t = -2.5, df = 8500, p-value = 0.01
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.048577 -0.006115
## sample estimates:
##      cor
## -0.02736
```

tidy()

The `tidy()` function from the `broom` package **converts** the most important results of many statistical objects to a **data frame**.

Try `tidy()` on your favorite statistical object and see what you get!

```
# Load broom package
library(broom)    # For tidy()

# Conduct correlation test
income_htest <- cor.test(formula = ~ height + income,
                         data = baselers)

tidy(income_htest)
```

```
## # A tibble: 1 x 8
##   estimate statistic p.value parameter
##      <dbl>     <dbl>   <dbl>     <int>
## 1  -0.0274     -2.52  0.0116      8508
## # … with 4 more variables: conf.low <dbl>,
## #   conf.high <dbl>, method <chr>,
## #   alternative <chr>
```

tidy()

The `tidy()` function from the `broom` package **converts** the most important results of many statistical objects to a **data frame**.

Try `tidy()` on your favorite statistical object and see what you get!

```
# Load broom package
library(broom)    # For tidy()

# Conduct t.test
height_sex_ttest <- t.test(formula = height ~ sex,
                           data = baselers)
height_sex_ttest
```

```
##
##     Welch Two Sample t-test
##
## data:  height by sex
## t = -67, df = 9900, p-value <2e-16
## alternative hypothesis: true difference in means is not equal
## 95 percent confidence interval:
##  -14.41 -13.59
## sample estimates:
## mean in group female   mean in group male
##                  164                  178
```

tidy()

The `tidy()` function from the `broom` package **converts** the most important results of many statistical objects to a **data frame**.

Try `tidy()` on your favorite statistical object and see what you get!

```r
# Load broom package
library(broom)    # For tidy()

# Conduct t.test
height_sex_ttest <- t.test(formula = height ~ sex,
                           data = baselers)

# tidy results
tidy(height_sex_ttest)
```

```
## # A tibble: 1 x 10
##   estimate estimate1 estimate2 statistic
##      <dbl>     <dbl>     <dbl>     <dbl>
## 1    -14.0      164.      178.     -66.6
## # … with 6 more variables: p.value <dbl>,
## #   parameter <dbl>, conf.low <dbl>,
## #   conf.high <dbl>, method <chr>,
## #   alternative <chr>
```

Distribution functions

R has a several functions that allow you to draw **random samples** data from specified distributions:

Type	CDF	CDF^{-1}	Simulate
Normal	`pnorm()`	`qnorm()`	`rnorm()`
Uniform	`punif()`	`qunif()`	`runif()`
F	`pf()`	`qf()`	`rf()`
Binomial	`pbinom()`	`qbinom()`	`rbinom()`
Chi-square	`pchisq()`	`qchisq()`	`rchisq()`

CDF - Cumulative Density Function
CDF^{-1} - Inverse Cumulative Density Function

```r
# Pr(z ≤ 2)
pnorm(q = 2, mean = 0, sd = 1)
```

```
## [1] 0.9772
```

```r
# z for p(z ≤ x) = 95%
qnorm(p = .95, mean = 0, sd = 1)
```

```
## [1] 1.645
```

```r
# simulate z
rnorm(n = 23, mean = 0, sd = 1)
```

```
##  [1] -0.27253 -0.36898  0.35302  0.87006
##  [5] -0.09277  0.45343 -0.21054 -1.96924
##  [9] -0.61462 -0.33692  2.13845  0.32882
## [13] -1.05609  0.10568  0.26327  0.31299
## [17]  0.43195 -0.50643 -1.21301  0.36599
## [21]  0.11673 -1.35680 -2.07270
```

Distribution functions

R has a several functions that allow you to draw **random samples** data from specified distributions:

Type	CDF	CDF^{-1}	Simulate
Normal	`pnorm()`	`qnorm()`	`rnorm()`
Uniform	`punif()`	`qunif()`	`runif()`
F	`pf()`	`qf()`	`rf()`
Binomial	`pbinom()`	`qbinom()`	`rbinom()`
Chi-square	`pchisq()`	`qchisq()`	`rchisq()`

CDF - Cumulative Density Function
CDF^{-1} - Inverse Cumulative Density Function

```
# Pr(t ≤ 2)
pt(q = 2, df = 99)
```

```
## [1] 0.9759
```

```
# t for p(t ≤ x) = 95%
qt(p = .95, df = 99)
```

```
## [1] 1.66
```

```
# simulate t
rt(n = 20, df = 99)
```

```
##  [1] -0.61673  0.08941  0.49011  1.33736
##  [5] -0.61308  0.74276  0.77337 -1.15292
##  [9]  0.67891  0.98415 -1.16118  1.24822
## [13]  0.98867 -1.09769 -0.46602 -0.52857
## [17] -0.96983  1.32325  0.34523  1.03262
```

Practical