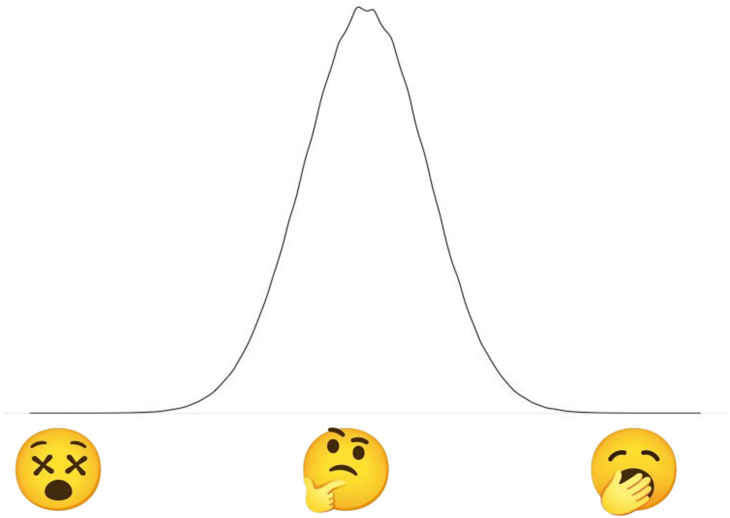
Philosophy of Science & Statistical Reasoning

3. Linear Regression & *T*-test

But first, …

📢

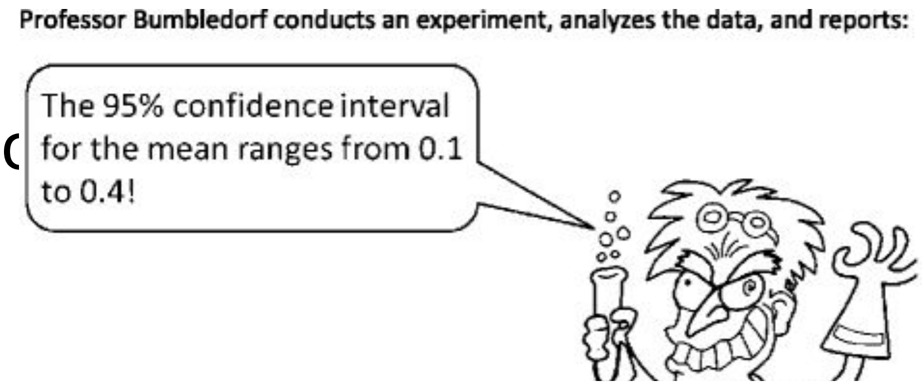
- Lecture recording
- Open book
- R programming: An Introduction to R, Base R cheat sheet, (Tidyverse cheat sheets, RStudio Education), ~~Quick-R (programming, statistics, but 💸)~~



Student Puzzlement Scale

Previously, on statistical reasoning

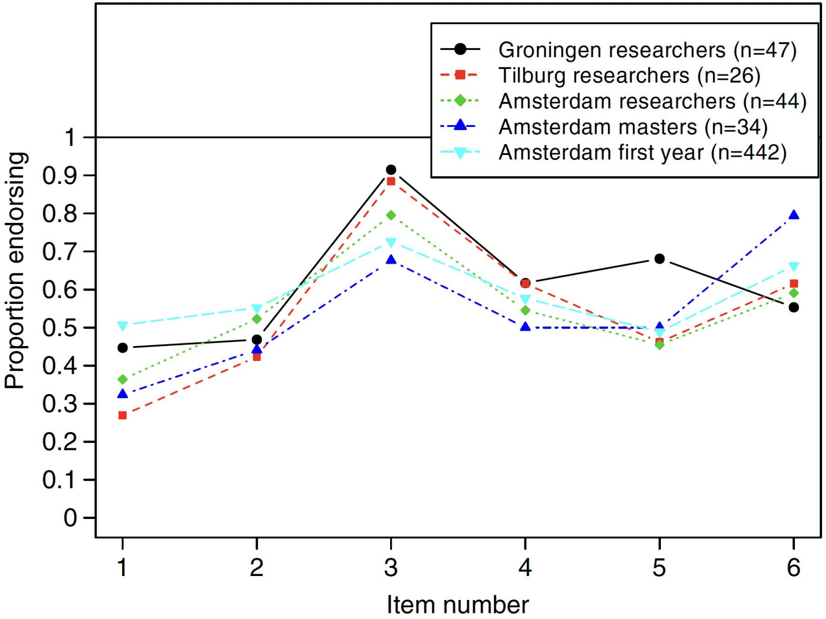
Previously, on statistic



1. The probability that the true mean is greater than 0 is at least 95%.
2. The probability that the true mean equals 0 is smaller than 5%.
3. The “null hypothesis” that the true mean equals 0 is likely to be incorrect.
4. There is a 95% probability that the true mean lies between 0.1 and 0.4.
5. We can be 95% confident that the true mean lies between 0.1 and 0.4.
6. If we were to repeat the experiment over and over, then 95% of the time the true mean falls between 0.1 and 0.4.

[Hoekstra et al., 2014]

Previously, on statistical reasoning



[Hoekstra et al., 2014]

Pub quiz

What will we learn today?

Topics

Statistical reasoning
Empirical cycle
Probability distributions
Frequentist inference
Sample / sampling distribution
Central limit theorem
Normal distribution
P-value
Type I/II errors
Effect size
Confidence interval
Power
Test statistics
Linear regression
t-Test
Moderation
ANOVA
Nonparametric inference
Bayesian inference

Questions

Estimating relationships between variables

Illustration by [Diego Mariano]

Q. Are the dimensions of the petals and sepals of the iris flower related?

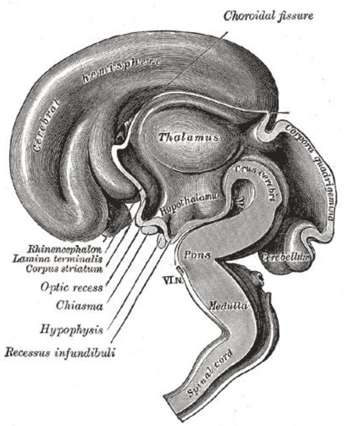
H. The length of a petal is related to the length and the width of a sepal.

E. …

```
> str(iris)
'data.frame':	150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

A data set made famous by [Ronald Fisher] and with its very own [Wikipedia page].

Statistical model



Outcome = Model + Error

- Perseverance = Student Population + Error
- Petal Length = Sepal Length + Sepal Width + Error

[Model formulae](#) in R:

y ~ model

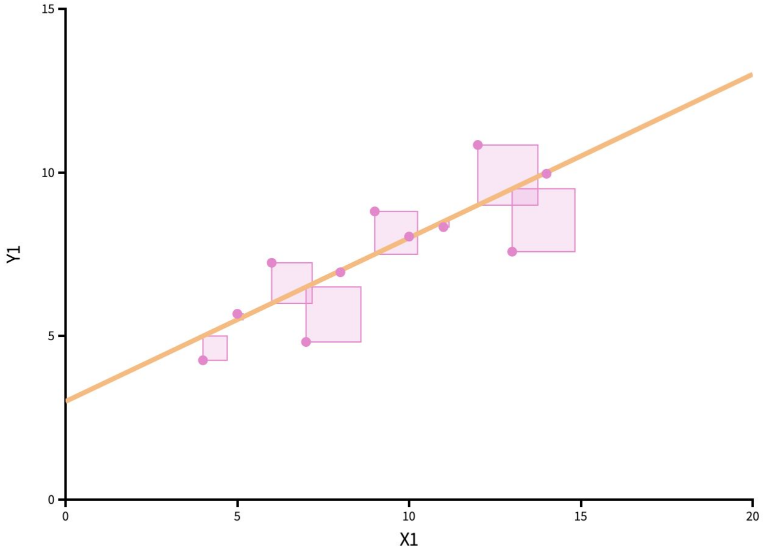
- y : dependent variable
- ~ : “is modeled by”
- model : independent variable(s)

🧑‍💻

Perseverance ~ Student_Population
Petal.Length ~ Sepal.Length + Sepal.Width
Grade ~ Attendance * Assignments

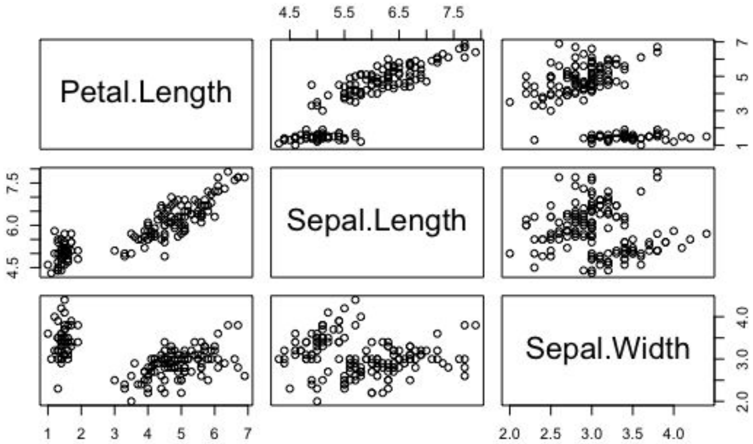
Regression analysis

> “ *Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome' or 'response' variable, or a 'label' in machine learning parlance) and one or more independent variables (often called 'predictors', 'covariates', 'explanatory variables' or 'features').*
> — [Wikipedia]



💡 Ordinary least squares (OLS) demonstration by [Seeing Theory] (used in core R function)

Multiple linear regression



Outcome	= Model	+ Error
Y_i	= …	+ e_i

$\ldots = \beta_o + \beta_1 X_i$ (simple lin. reg.)

$\ldots = \beta_o + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}$ (multiple lin. reg.)

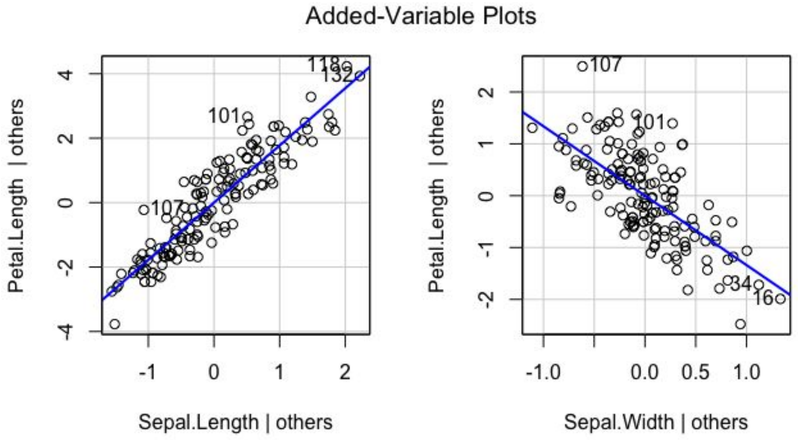
$Petal\ Length_i = \beta_o + \beta_1\ Sepal\ Length_i + \beta_2\ Sepal\ Width_i + e_i$

🧑‍💻

```
mod <- Petal.Length ~ Sepal.Length + Sepal.Width
fit <- lm(formula = mod, data = iris, method = "qr")
summary(fit); resid(fit); confint(fit)
```

Results

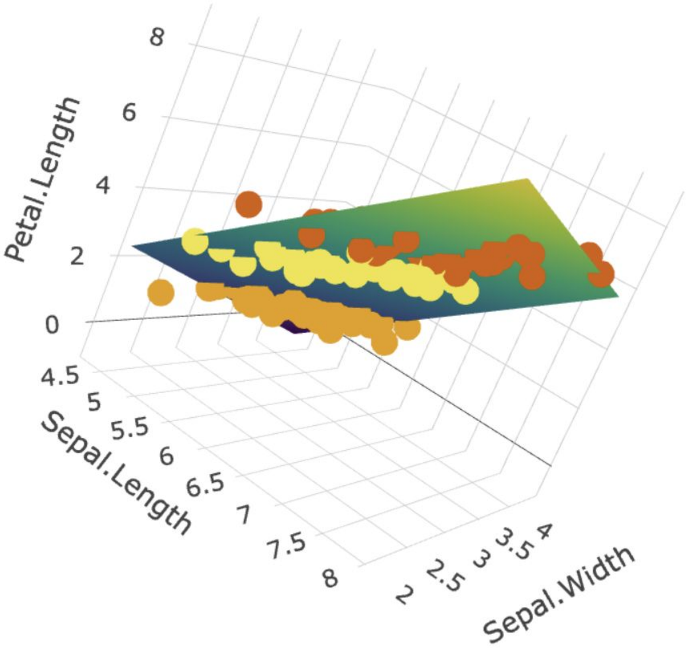
```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.25582 -0.46922 -0.05741  0.45530  1.75599 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -2.52476    0.56344  -4.481 1.48e-05 ***
Sepal.Length  1.77559    0.06441  27.569  < 2e-16 ***
Sepal.Width  -1.33862    0.12236 -10.940  < 2e-16 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6465 on 147 degrees of freedom
Multiple R-squared:  0.8677,    Adjusted R-squared:  0.8659 
F-statistic:   482 on 2 and 147 DF,  p-value: < 2.2e-16
```



Petal Length$_i$ = −2.52 + 1.78 × *Sepal Length*$_i$ + −1.34 × *Sepal Width*$_i$ + e_i

“| others” = *holding the other variables constant*

Results



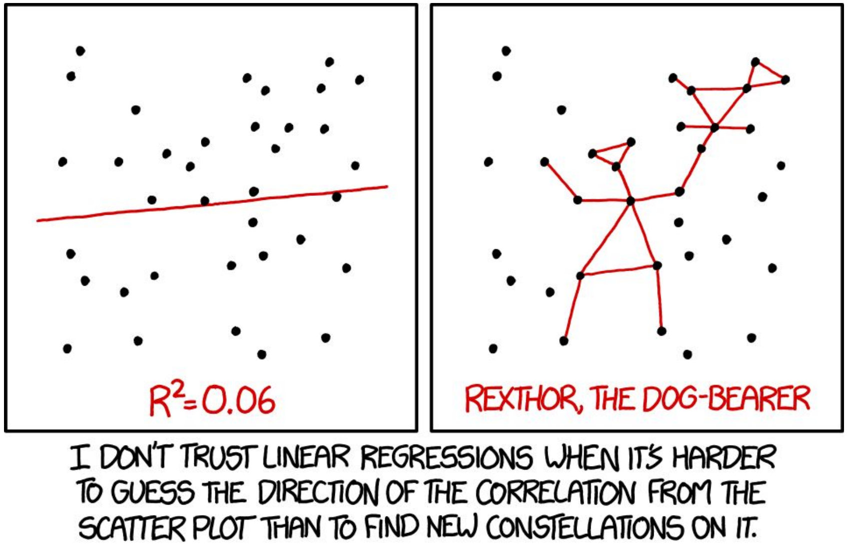
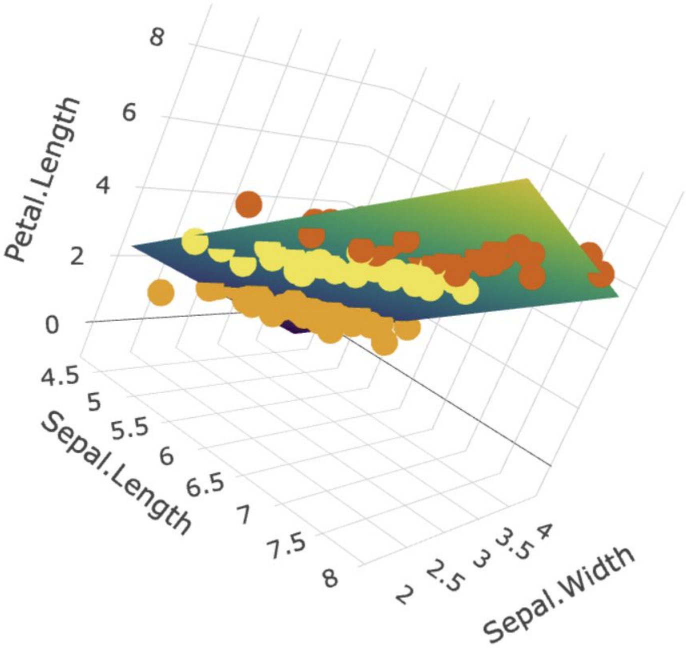


Illustration by Randall Munroe (wtf)

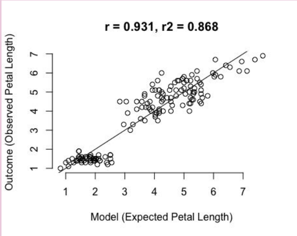
Results



🧑‍💻
```
# statistical significance of predictors
summary(fit)
```

\# multiple R^2 (explained variance)

```
observed <- iris$Petal_Length
expected <- fitted(fit)
cor(observed, expected)^2
```



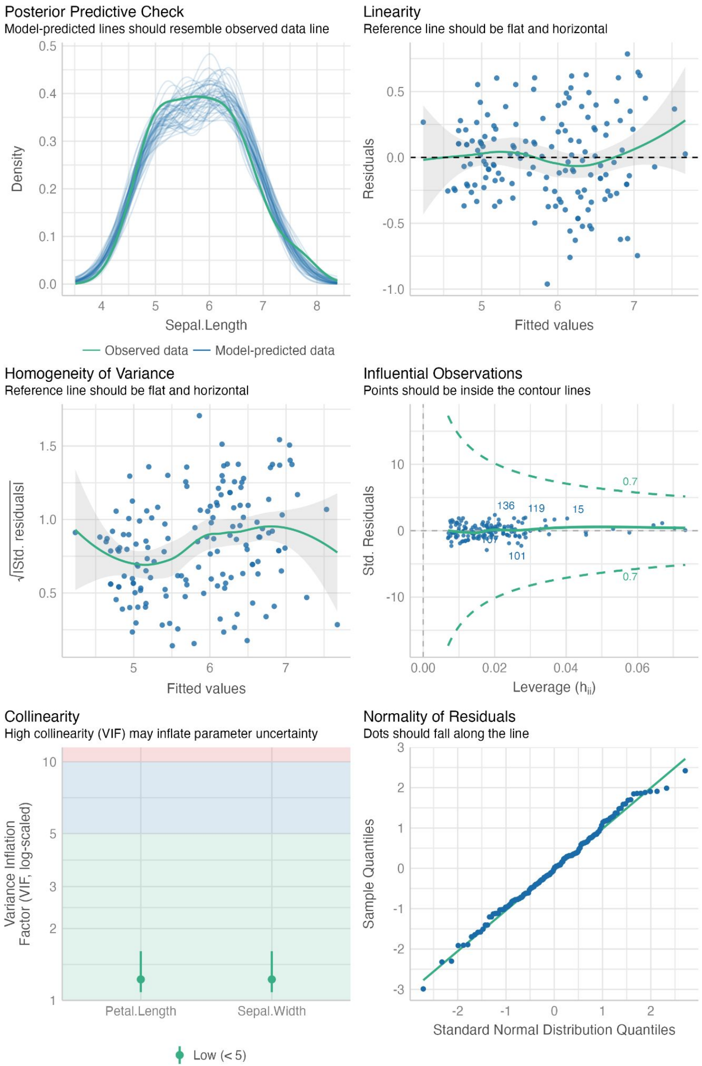

```
# F statistic

# model comparison
mod_0 <- Petal.Length ~ Sepal.Length
fit_0 <- lm(formula = mod_0, data = iris)
anova(fit, fit_0)

# predictive validity
```

predict(fit, *new_data*)

Assumptions

🧑‍💻

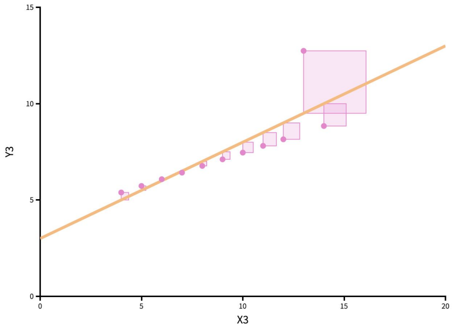
```
library("easystats")
performance::check_model(fit)
```



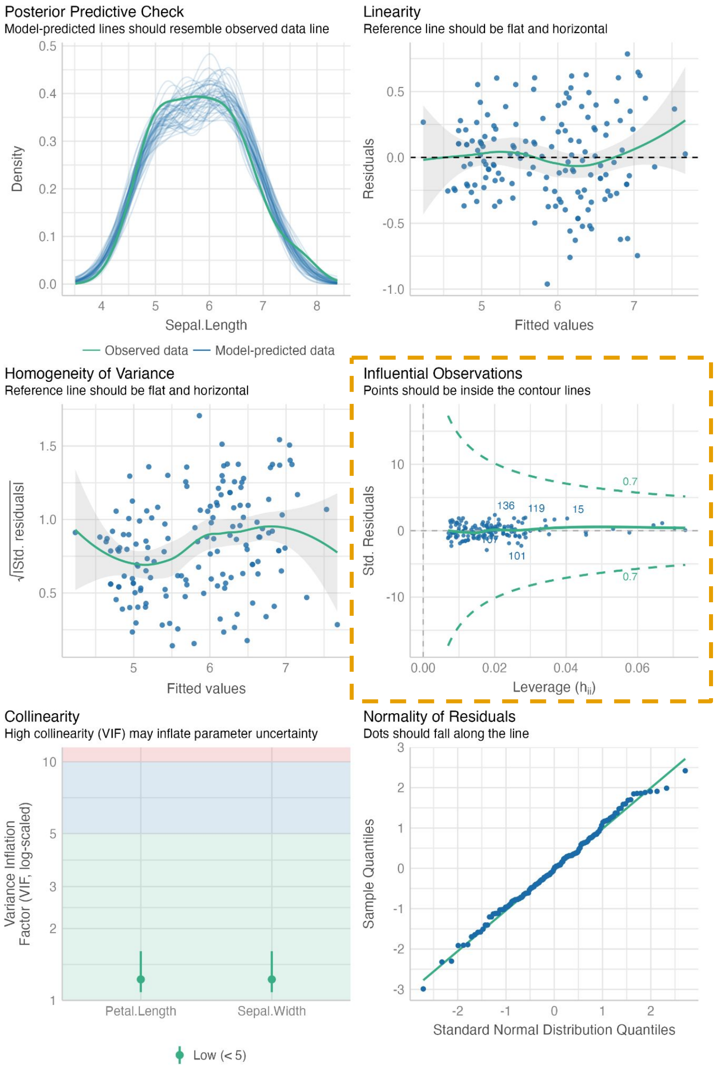
Influential observations

[Leys et al., 2019](#):

- Error outliers
- Interesting outliers
- Random outliers



Cook's distance

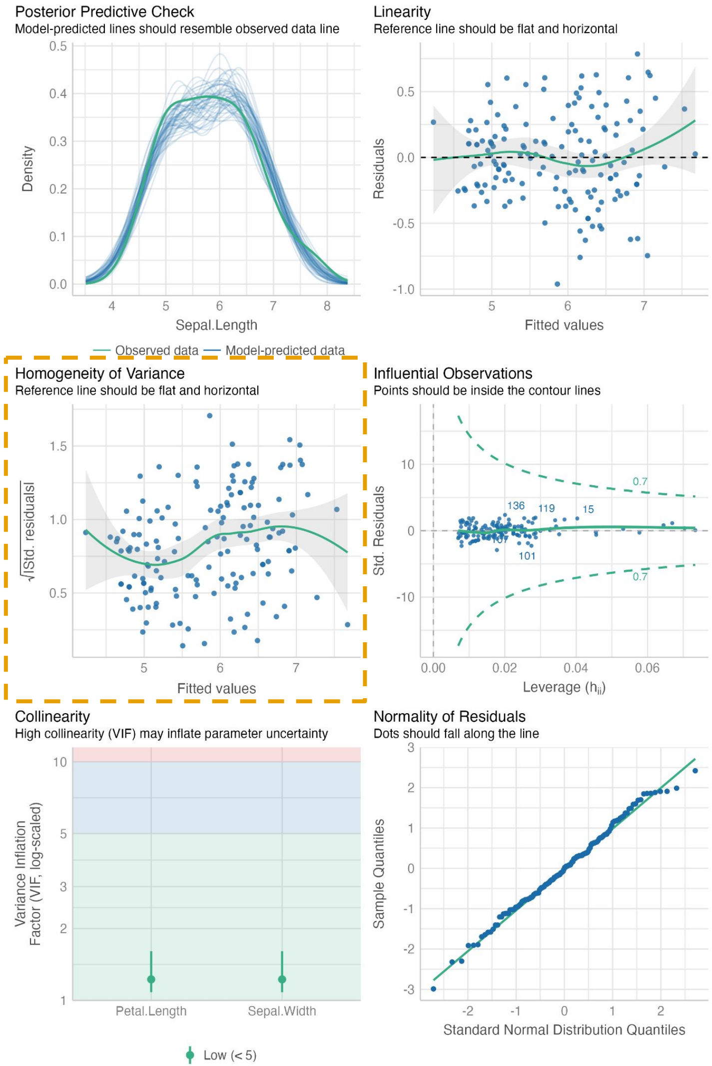


💭 [Interpretation and solutions](#)

Homogeneity of variance

> [T]he variance of the residuals across different values of predictors is similar and does not notably increase or decrease.
> — Performance package

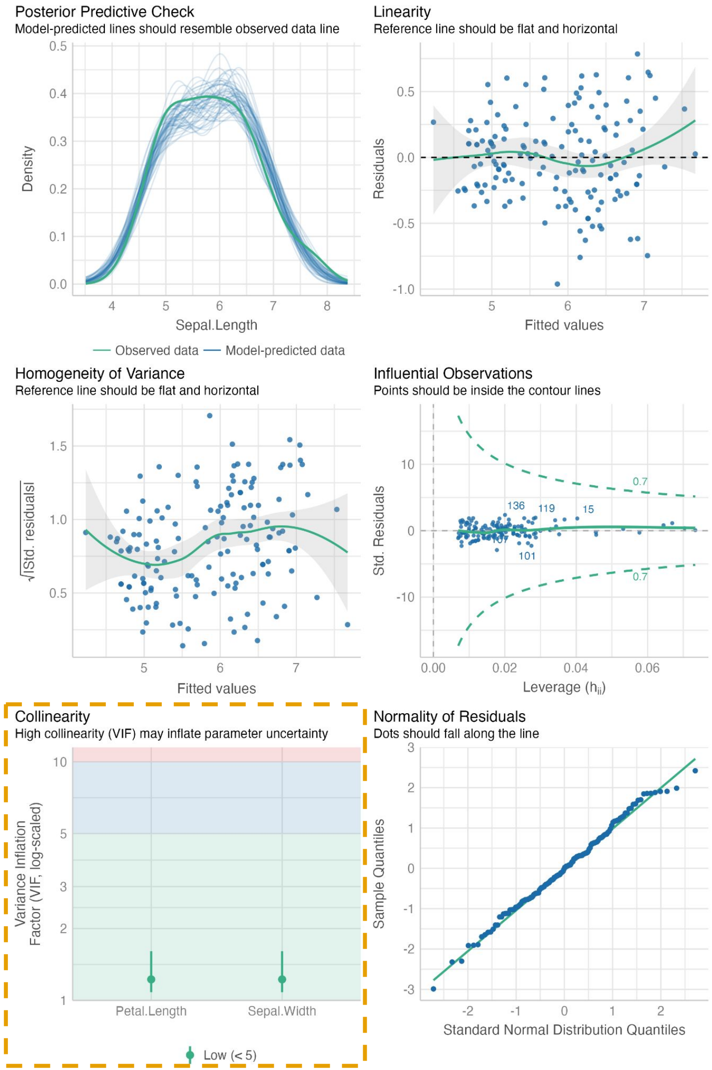
Also, *homoscedasticity*.



💭 Interpretation and solutions

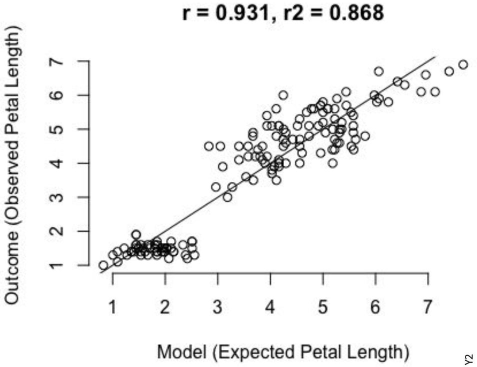
Multicollinearity

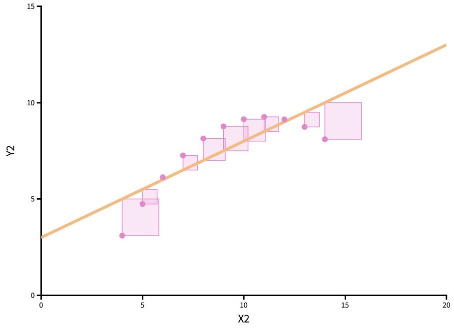
Explanation vs. prediction

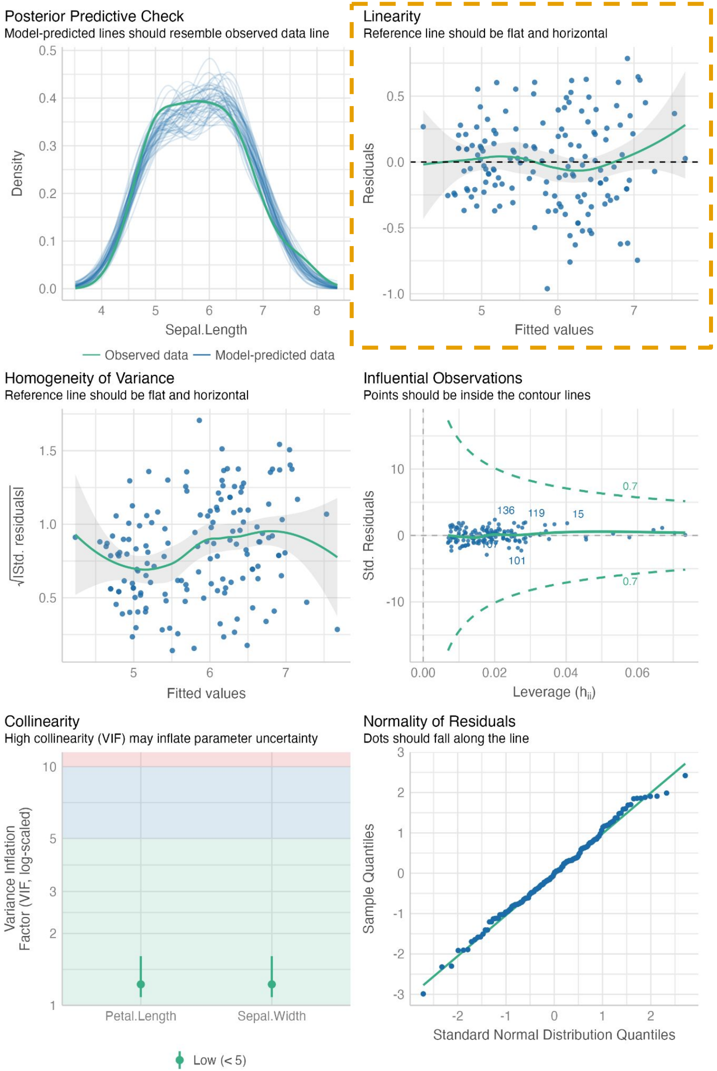


💭 [Interpretation and solutions]

Linearity







[Interpretation and solutions]

Statistical model II

🧑‍💻

y ~ x # with intercept
y ~ 1 + x # with intercept
y ~ 0 + x # without intercept

y ~ x + z # add a term
y ~ x - z # remove a term
y ~ I(x + z) # sum two terms
y ~ x:z # create an interaction term
y ~ x * z # create crossed terms (x + z + x:z)
y ~ x %in% z) # create nested terms (x + x:z)

and there's more…

Traditional name	Model formula	R code
Bivariate regression	Y ~ X1 (continuous)	`lm(Y ~ X)`
One-way ANOVA	Y ~ X1 (categorical)	`lm(Y ~ X)`
Two-way ANOVA	Y ~ X1 (cat) + X2(cat)	`lm(Y ~ X1 + X2)`
ANCOVA	Y ~ X1 (cat) + X2(cont)	`lm(Y ~ X1 + X2)`
Multiple regression	Y ~ X1 (cont) + X2(cont)	`lm(Y ~ X1 + X2)`
Factorial ANOVA	Y ~ X1 (cat) * X2(cat)	`lm(Y ~ X1 * X2)` or `lm(Y ~ X1 + X2 + X1:X2)`

Table from An Introduction to R

🚁 Nearly anything can be described with a (generalized linear) regression model. A cheat sheet for model formulae. Understand the *t*-test and ANOVA as a linear model (cheat sheet).

☕

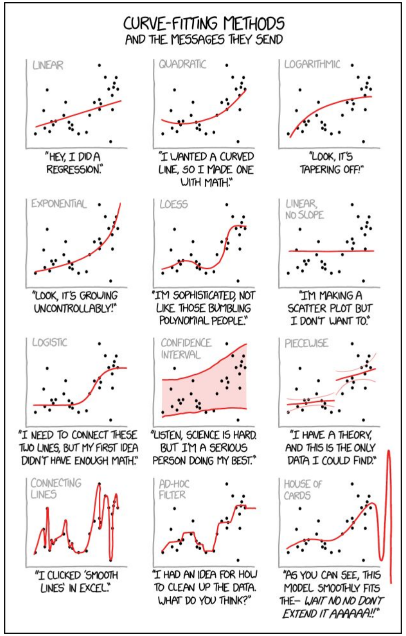
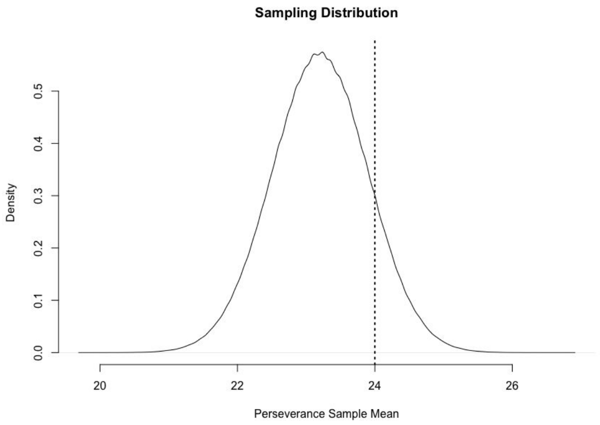


Illustration by Randall Munroe (wtf)

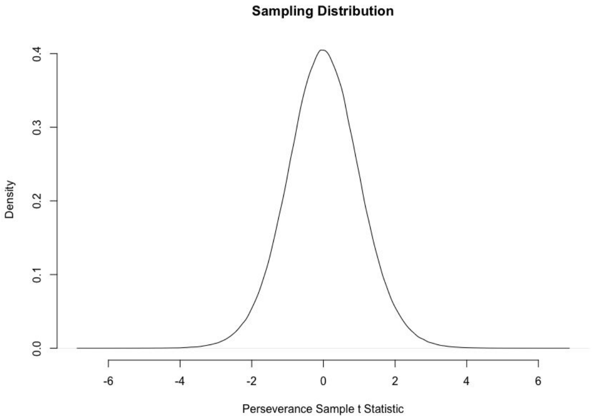
15:00

Comparing two means with Student's *t*-test

Déjà vu?



Normal distribution
mu = population mean
sd = standard error of sample mean



Student's *t*-distribution
df = n − 1

Student's *t*-statistic 🍻

Standardization (mean 0; sd 1)

Sample: (observation − sample mean) / sd

t = (sample mean − μ) / se

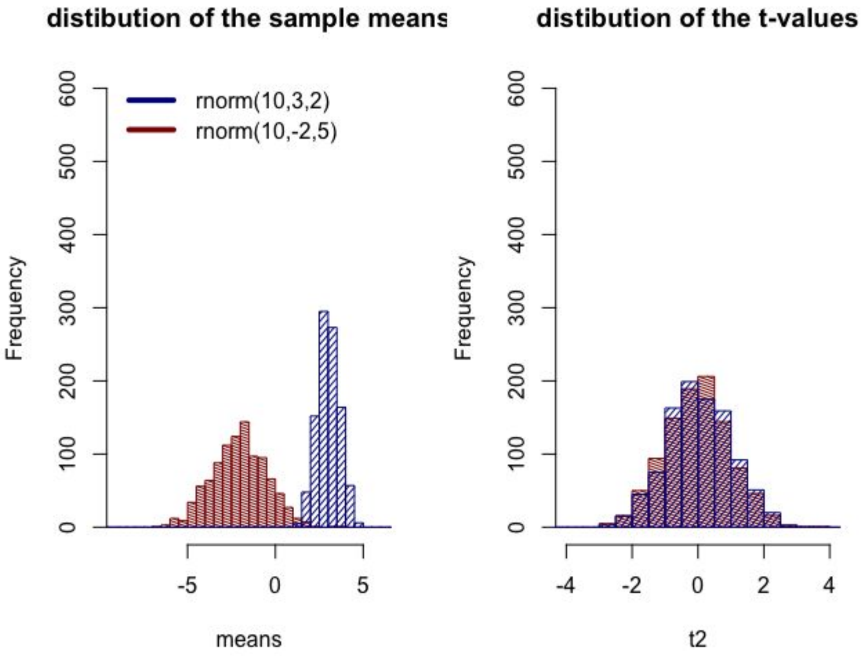
sample mean = 24
μ = 23.2
sample sd = 2.75
n = 20

t = (24 − 23.2) / (2.75 / √(20)) = 1.2996

🧑‍💻

```
pt(1.2996, df = 20-1, lower.tail = FALSE)  # probability of t or higher
t.test(dat, mu = 23.2, alternative = "greater")  # one sample t-test
```

Student’s *t*-statistic 🍻

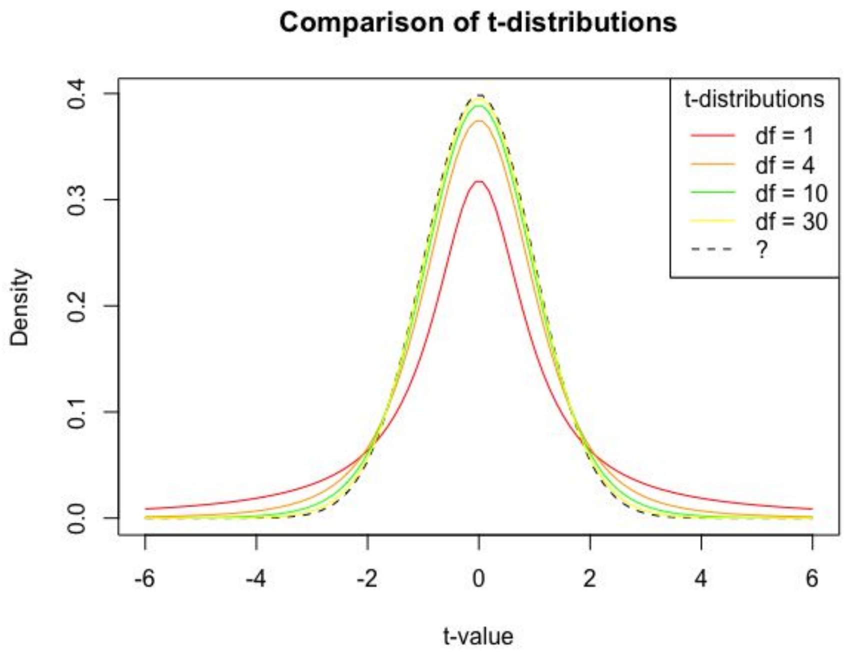


> “ The key property of the *t*-statistic is that it is a pivotal quantity – while defined in terms of the sample mean, its sampling distribution does not depend on the population parameters, and thus it can be used regardless of what these may be.
> — Wikipedia

Student's *t*-distribution

💭 What are the degrees of freedom at the dashed line?

💭 If there is a difference in population means, is it easier to find a significant effect with a larger sample size?



🛠️ Web simulation by Kristoffer Magnusson.

Effect size

R^2

$R^2 = t^2 / (t^2 + df)$

Cohen's *d*

R Psychologist

Linear regression

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.25582 -0.46922 -0.05741  0.45530  1.75599

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.52476    0.56344  -4.481 1.48e-05 ***
Sepal.Length  1.77559    0.06441  27.569  < 2e-16 ***
Sepal.Width  -1.33862    0.12236 -10.940  < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6465 on 147 degrees of freedom
Multiple R-squared:  0.8677,    Adjusted R-squared:  0.8659
F-statistic:   482 on 2 and 147 DF,  p-value: < 2.2e-16
```

Compute t-statistic for β_1 (same procedure as for the mean): t = (1.776 − 0) / 0.064 = 27.569

Common name	Built-in function in R	Equivalent linear model in R
y is independent of x P: One-sample t-test N: Wilcoxon signed-rank	t.test(y) wilcox.test(y)	lm(y ~ 1) lm(signed_rank(y) ~ 1)
P: Paired-sample t-test N: Wilcoxon matched pairs	t.test(y_1, y_2, paired=TRUE) wilcox.test(y_1, y_2, paired=TRUE)	lm(y_2 - y_1 ~ 1) lm(signed_rank(y_2 - y_1) ~ 1)
y ~ continuous x P: Pearson correlation N: Spearman correlation	cor.test(x, y, method='Pearson') cor.test(x, y, method='Spearman')	lm(y ~ 1 + x) lm(rank(y) ~ 1 + rank(x))
y ~ discrete x P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	t.test(y_1, y_2, var.equal=TRUE) t.test(y_1, y_2, var.equal=FALSE) wilcox.test(y_1, y_2)	lm(y ~ 1 + G_2)[A] gls(y ~ 1 + G_2, weights=...[B])[A] lm(signed_rank(y) ~ 1 + G_2)[A]

Table by Jonas Kristoffer Lindeløv

🚁 Always use the Welch's *t*-test (for unequal variances).

Cooling down

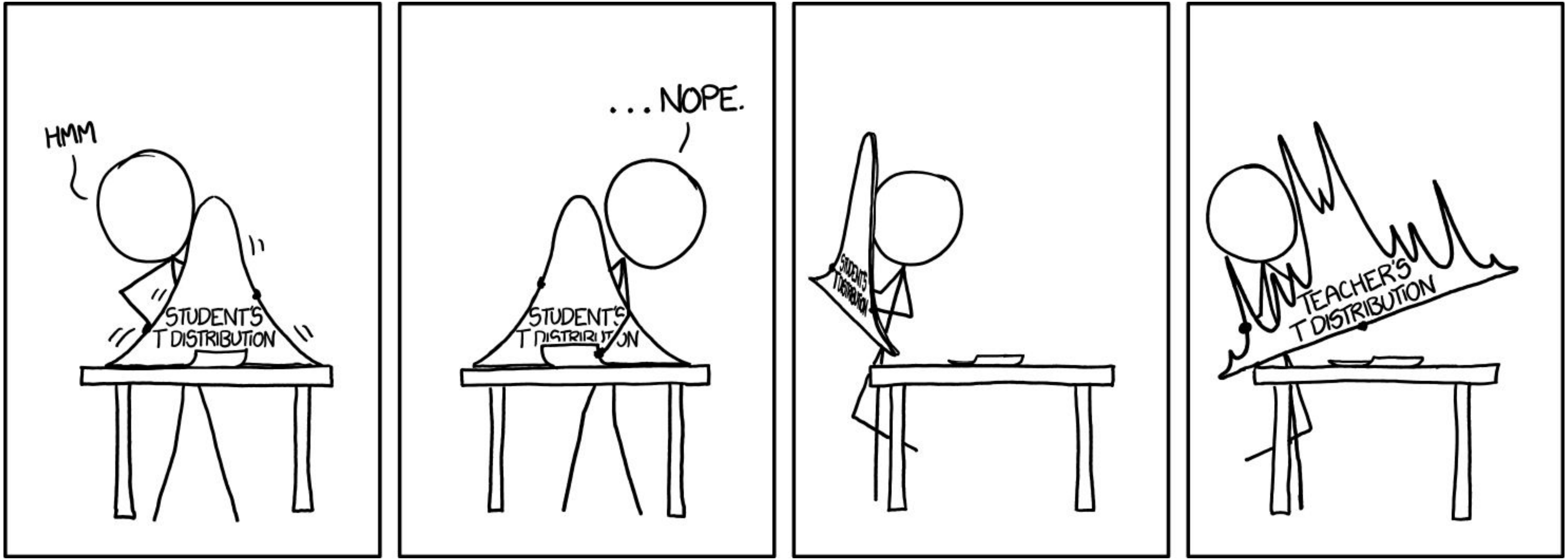


Illustration by [Randall Munroe](#) ([wtf](#))

What did we learn?

Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say 20 subjects in each sample). Further, suppose you use a simple independent means *t*-test and your result is significant (t = 2.7, d.f. = 18, p = 0.01). Please mark each of the statements below as "true" or "false." "False" means that the statement does not follow logically from the above premises. Also note that

1. You have absolutely disproved the null hypothesis (that is, there is no difference between the population means).
2. You have found the probability of the null hypothesis being true.
3. You have absolutely proved your experimental hypothesis (that there is a difference between the population means).
4. You can deduce the probability of the experimental hypothesis being true.
5. You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.
6. You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.

Gigerenzer, 2004 🔒

4. You can deduce the probability of the experimental hypothesis being true.

Take-home assignments

📆 Weekly assignment

🍻 Pub quiz

Create an *informative* four-choice question about the content of today's lecture.

An informative question has a large spread in responses across answer options.

Clarify answer options (which are (in)correct and why).

Illustration adapted from [Snippets.com]

Take-home assignments

📰 Fake news

Find a headline that incorrectly states a causal relationship instead of a correlation, and post it in the discussion forum.

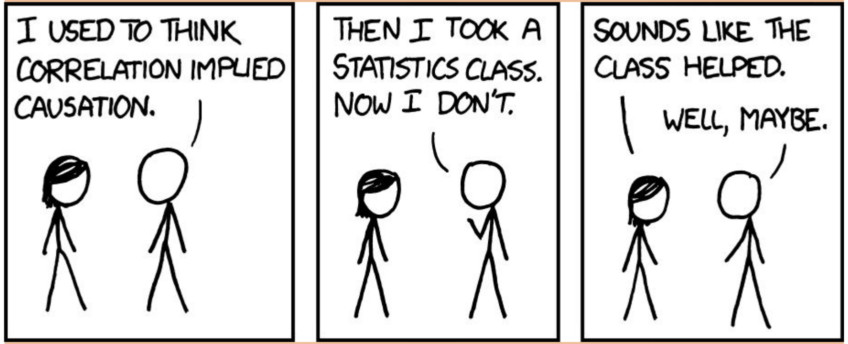


Illustration by [Randall Munroe](#) ([wtf](#))

"News consumption helps against polarization"

Video van de week

Nieuwsconsumptie helpt tegen polarisatie

Polarisatie wordt vaak in verband gebracht met desinformatie en echokamers. Onderzoeker Magdalena Wojcieszak stelt dat er een factor is waarover we ons misschien meer zorgen moeten maken: het gebrek aan onze online consumptie van kwaliteitsnieuws.

Video from [University of Amsterdam](#)