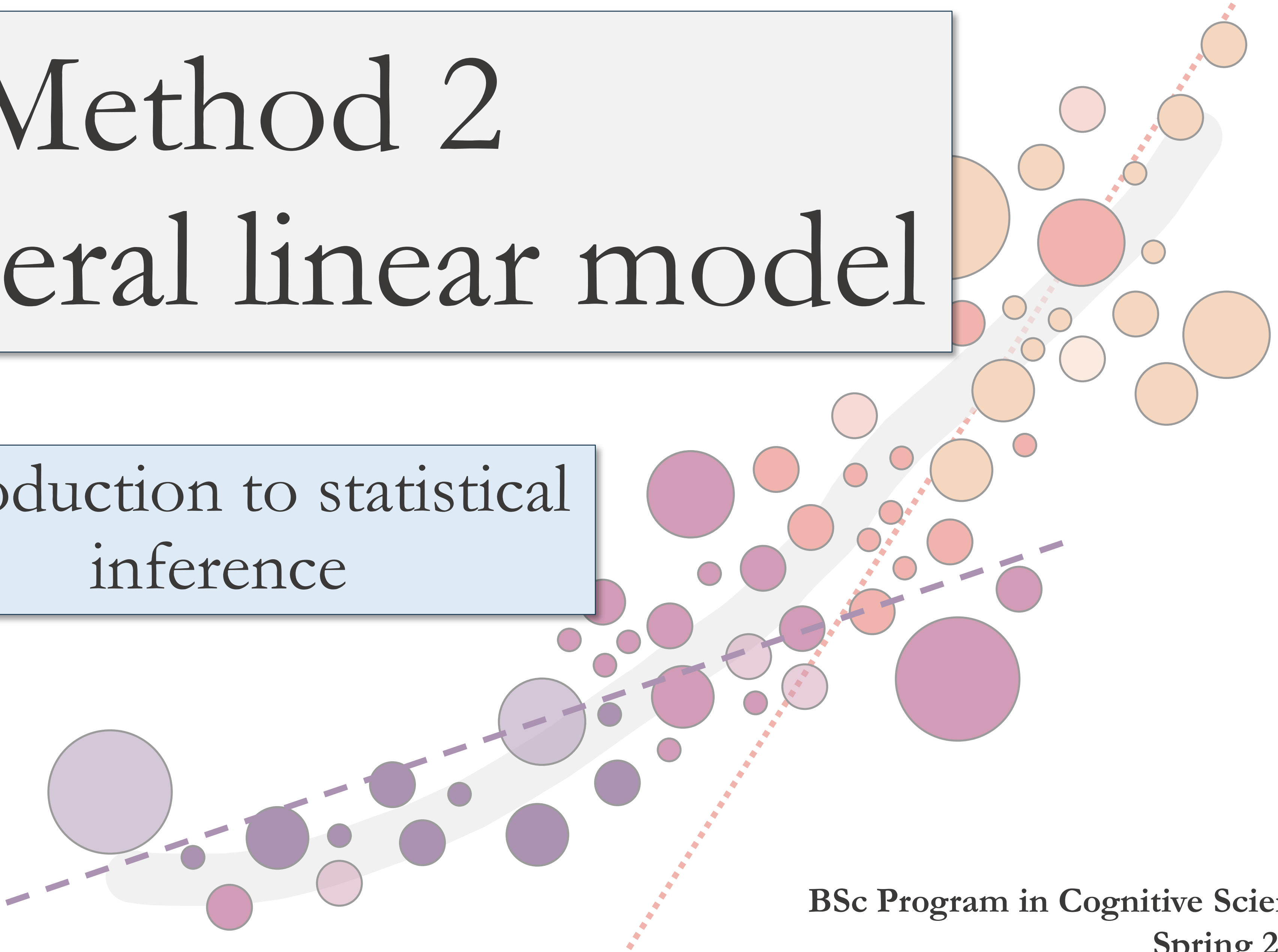
Method 2
The general linear model

Introduction to statistical inference



BSc Program in Cognitive Science
Spring 2024

Nicolas Legrand – Researcher– ILAB (Interacting Minds Centre)

Who am I?

- **PhD in Psychology**

  2014 - 2018 - Université de Caen (France)
  - *The consolidation and suppression of individual and collective memory.*
  EEG, machine learning, ECG, Python, memory suppression, fMRI, collective memory, dreams, memory schemas

- **Master in Neurobiology and behaviors**

  2013 - 2014 - Université de Caen (France)
  - *Sleep-dependent prospective memory consolidation and dreaming*

- **Master in Cognitive Science**

  2011 - 2013 - Université Lumière Lyon 2 (France)
  - *What is phenomenal consciousness?*
  - *The neuropsychoanalysis of dreaming.*

- **Bachelor in Mathematics and computer sciences**

  2009 - 2011 - Université de Grenoble 2 (France)

- **Bachelor in Philosophy**

  2008 - 2011 - Université de Grenoble 2 (France)

Dec. 2022 – Curr.

Researcher at IMC (Ilab, sup. Chris Mathys) in computational psychiatry. Using Bayesian nonparametric methods to create models of delusions. Developing a new neural network library for predictive coding (pyhgf).

Jul. 2019 – Dec. 2022

Postdoctoral fellow in computational psychiatry at CFIN (Embodied Computation Group, sup. Micah Allen). Creating new methods to measure cardiac interoception. Developing Python toolboxes for signal processing (Systole), psychophysic tasks (Cardioception), and metacognition modeling (metadpy).

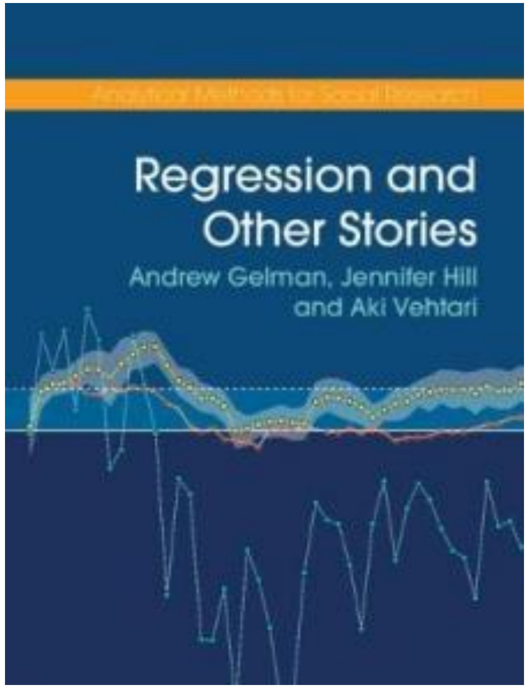
Python is my main programming language. I also have several years of experience using and teaching R and Matlab.

Overview

Topics

- The General Linear Model
- Regression modelling
- Mathematical foundations
- Linear algebra (vectors, matrices, determinants, eigen-analysis,...)
- Calculus (infinite series, derivatives, integrals,...)
- Generalized Linear Models (e.g., logistic regression)

Resources


Analytical Methods for Social Research
Regression and Other Stories
Andrew Gelman, Jennifer Hill and Aki Vehtari

ANALYTICAL METHODS FOR SOCIAL RESEARCH
Essential Mathematics for Political and Social Research
JEFF GILL
CAMBRIDGE

Textbook:
Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and Other Stories* (Analytical Methods for Social Research). Cambridge: Cambridge University Press. doi:10.1017/9781139161879
Please get a copy!
Free PDF: https://avehtari.github.io/ROS-Examples/

Textbook:
Gill, J. (2006). Essential Mathematics for Political and Social Research (Analytical Methods for Social Research). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511606656
No need to buy it! You have access to PDFs of the chapters via the Royal Library.

Resources

Code:

- This course's repository: https://github.com/methods-2-f24/

- All code and data in the book: https://github.com/avehtari/ROS-Examples

- **Please get a free GitHub account**

Videos:

- This course is on YouTube! https://www.youtube.com/playlist?list=PLvJwKACYy5_MTdnrzxx_1sN389dS9OB3S

- Order slightly different: we'll do GHV chapters 1-5 in the first 3 weeks, but after that, you can watch the videos in the order of the playlist

Schedule

Introduction to statistical inference

- **Portfolio 1**

Mathematical foundations

- **Portfolio 2**

Generalized Linear Models (GLM)

- **Portfolio 3**

Course week	Week of the year	Topics and readings
1	7	Regression and the GLM: overview, data and measurement, (GHV[1]1,2)
2	8	Basic methods, statistical inference (GHV 3,4)
3	9	Statistical inference (continued), simulation (GHV 4,5)
4	10	Math basics: functions, equations, polynomials, logarithms (Gill[2]1)
5	11	Linear algebra basics: vectors, matrices, norms, transposition (Gill 3)
6	12	More linear algebra: geometry, determinants, rank, inversion, eigenvectors (Gill 4)
7	15	Scalar calculus: derivatives, integrals, fundamental theorem (Gill 5)
8	16	More calculus: root finding, extrema, Lagrange multipliers, vector calculus (Gill 6)
9	17	Conceptual foundations and history of the GLM, model fitting (GHV 6,7,8)
10	18	Fitting GLMs: prediction, Bayesian inference (GHV 9)
11	19	Multiple predictors, interactions (GHV 10)
12	20	Model comparison, assumptions and diagnostics (GHV 11)
13	21	Transformations, predictive simulations (GHV 12) [no class, just lecture]

1 - Gelman, A., Hill, J., & Vehtari, A. (2020). Regression and Other Stories (Analytical Methods for Social Research). Cambridge: Cambridge University Press. doi:10.1017/9781139161879

2 - Gill, J. (2006). Essential Mathematics for Political and Social Research (Analytical Methods for Social Research). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511606656

Exam

- Portfolio consisting of 3 assignments

- Each assignment will require you to create an R Markdown notebook consisting of a mix of text and code.

- Due:
    1. End of week 9 (Sunday 3 March, 23:59)
    2. End of week 16 (Sunday 21 April, 23:59)
    3. End of week 17 (Sunday 26 May, 23:59)

You will receive a (short) feedback message from us on your portfolio assignments that you can use for improvements before finalizing your hand-ins.

Exam

https://kursuskatalog.au.dk/en/course/115680/Methods-2-The-General-Linear-Model

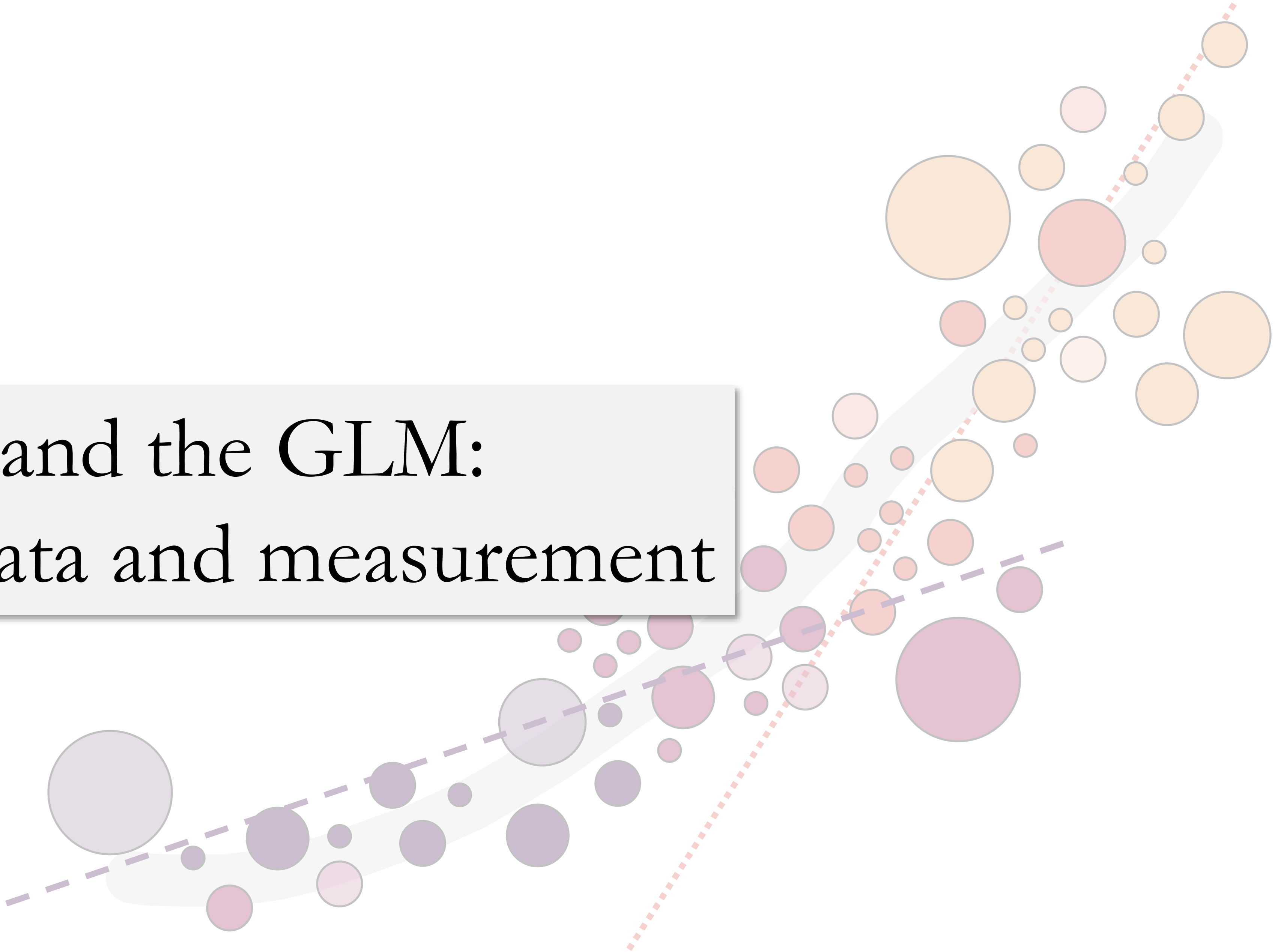
Ordinary examination and re-examination:
The exam consists of a portfolio containing some assignments. The total length of the portfolio is 3-7 assignments.

Their form and length will be announced on Blackboard by the teacher at the start of the semester. The portfolio may include products. Depending on their length, and subject to the teacher’s approval, these products can replace some of the standard pages in the portfolio.

It must be possible to carry out an individual assessment. So if some parts of the portfolio have been produced by a group, it must be stated clearly which parts each student is responsible for, and which parts the group as a whole is responsible for.

The complete portfolio must be submitted for assessment in the Digital Exam system.
Each student submits a portfolio.”

Regression and the GLM: overview, data and measurement



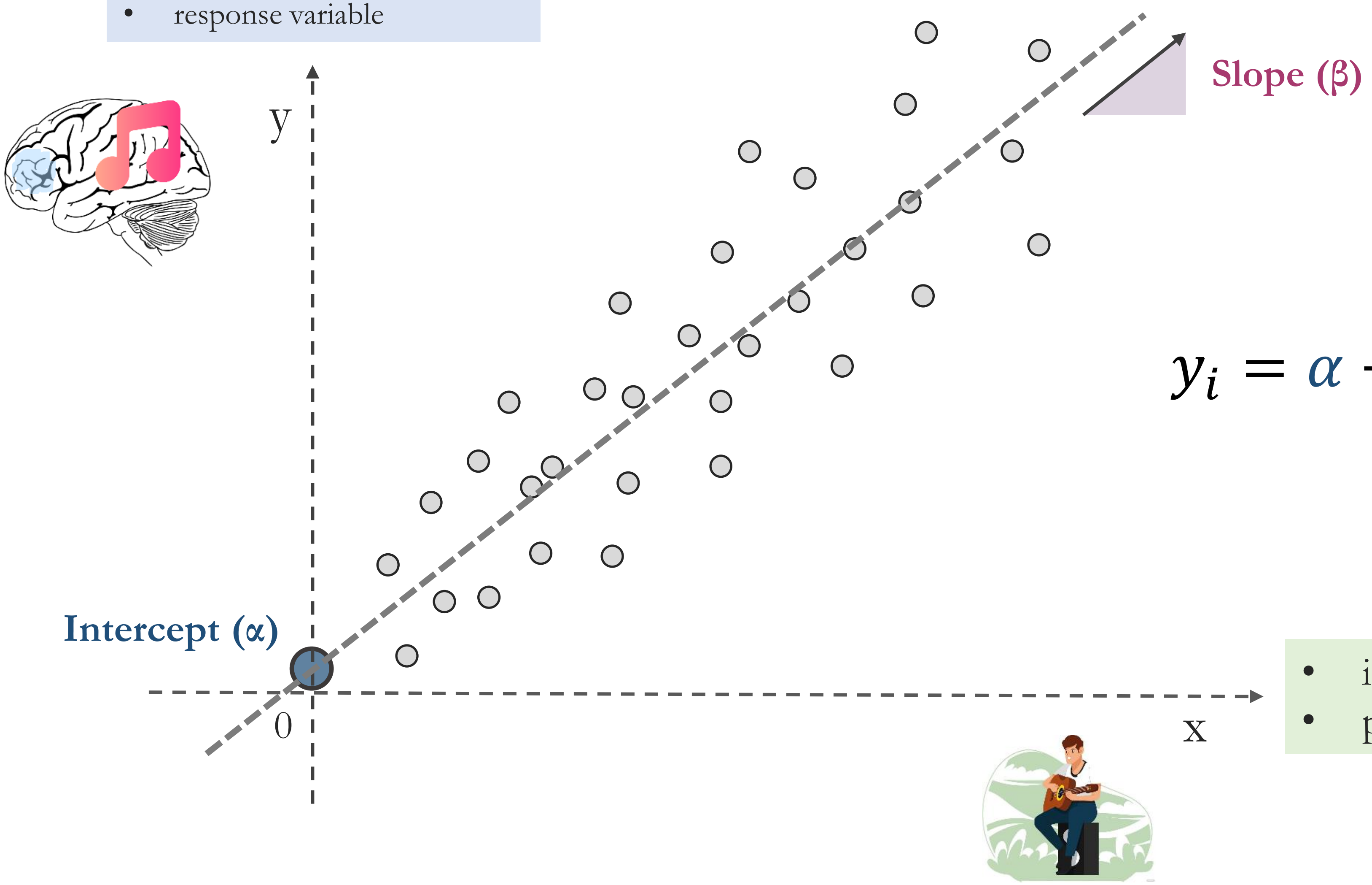
Nicolas Legrand – Researcher– ILAB (Interacting Minds Centre)

Linear regressions

When is something linear in mathematics?

What is a regression?

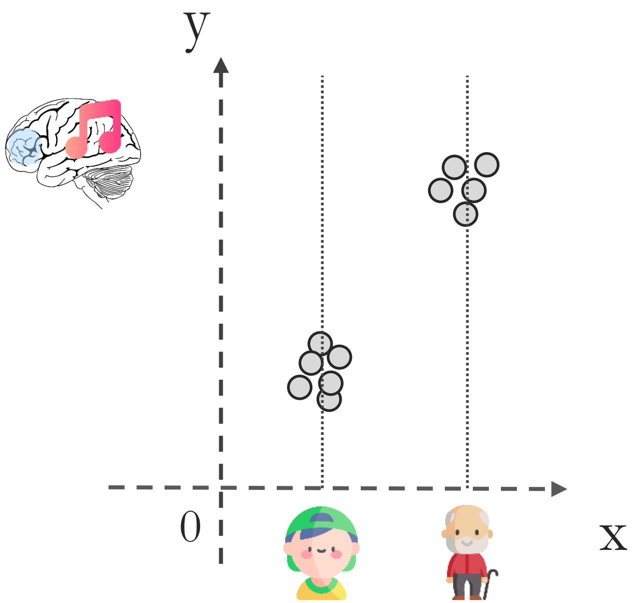
- dependent variable
- outcome variable
- response variable



$$y_i = \alpha + \beta x_i + \epsilon_i$$

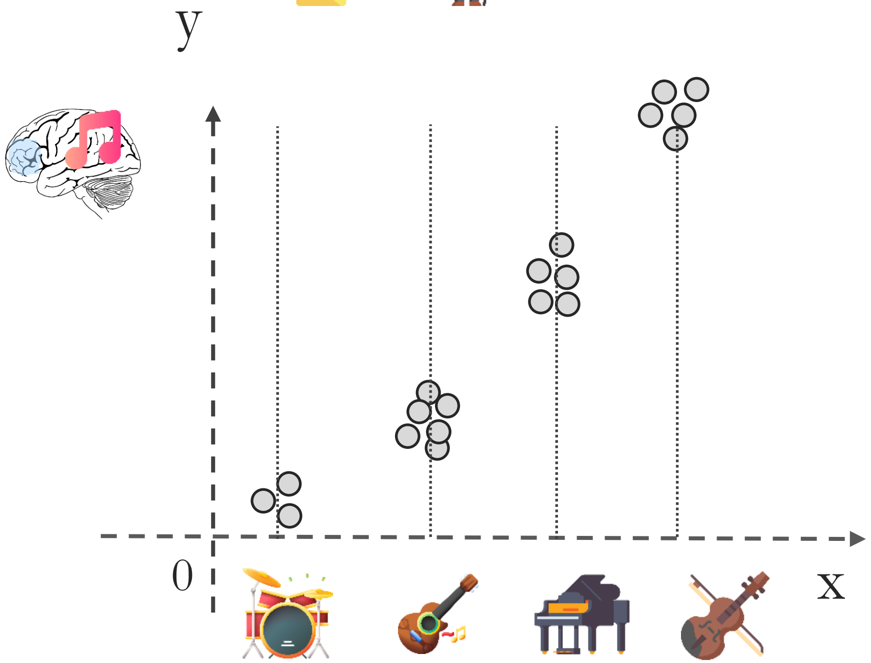
- independent variable
- predictor

Generalizations of linear models



Two sample t-test

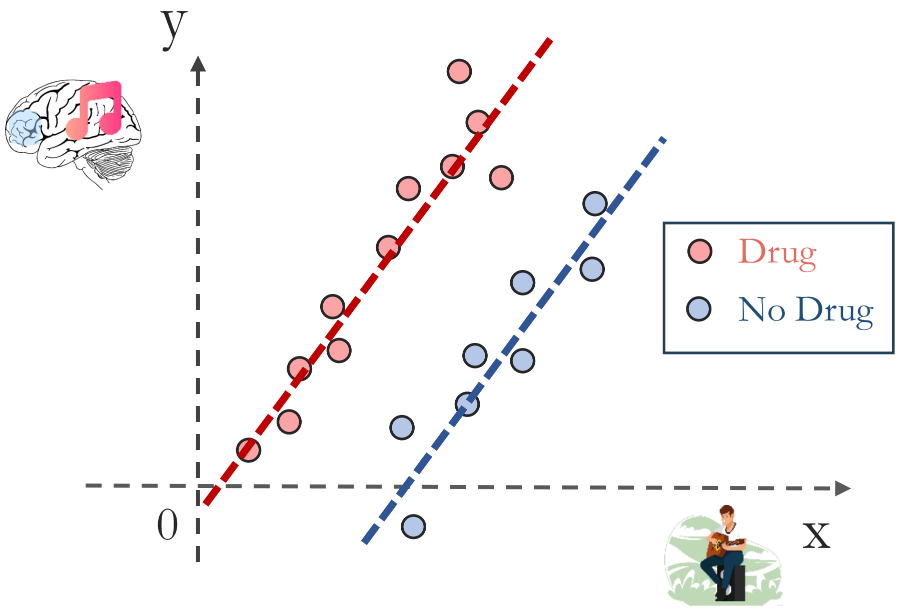
$$y_i = \alpha + \beta x_i + X_i^T \cdot \epsilon$$



Factorial ANOVA

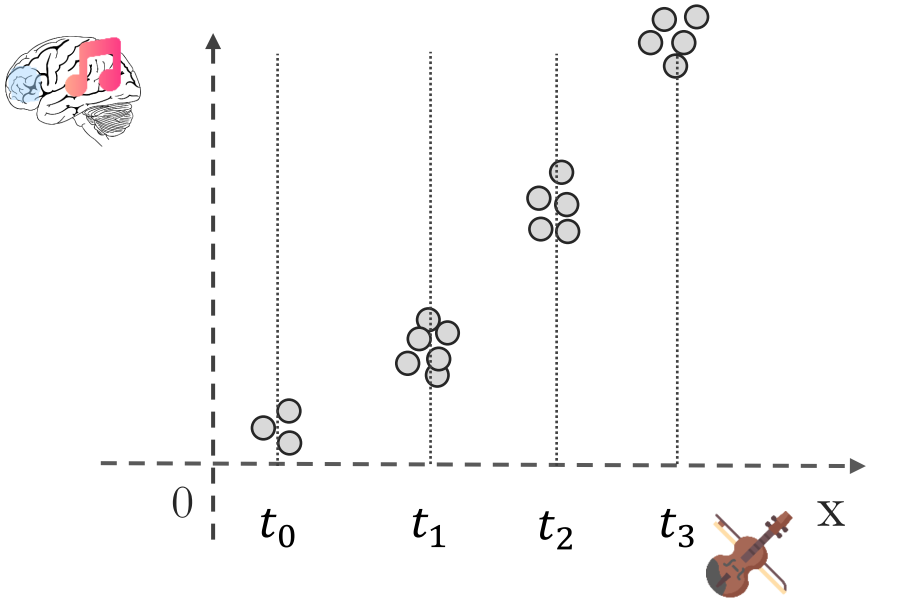
$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + X_i^T \epsilon$$

Generalizations of linear models



ANCOVA

$$y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x_j}) + \epsilon_{ij}$$



Repeated-measure ANOVA

$$y_{ij} = \alpha + \beta_j + \epsilon_i$$

The generalized linear model (GLM)

$$\mathbf{Y} = \mathbf{XB} + \mathrm{E}$$

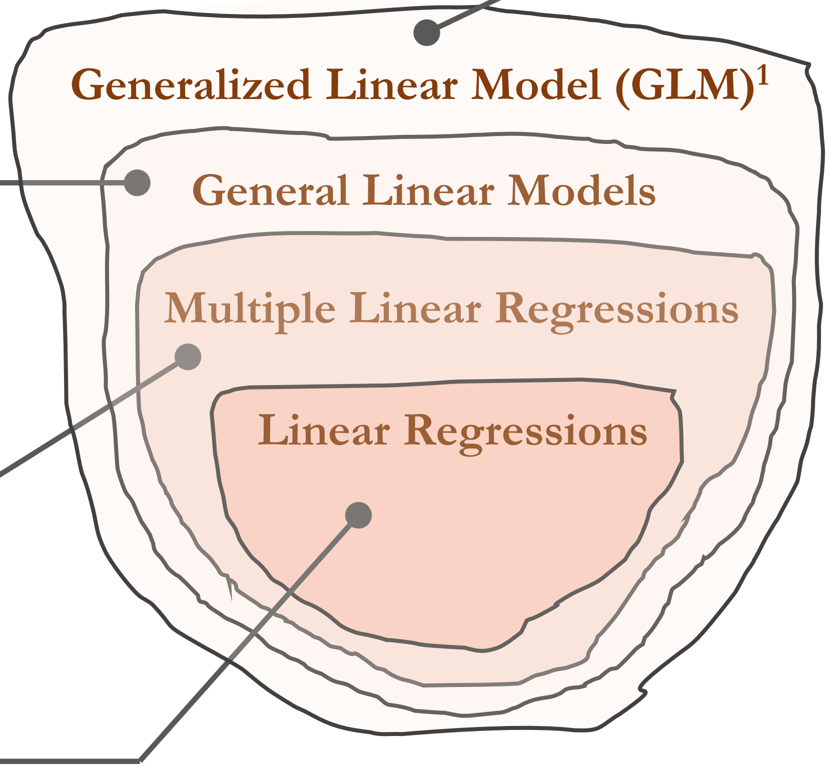
Multiple predictors and outcome variables. Include ANOVA, t-test, ANCOVA…

$$y = \mathbf{X}\beta + \epsilon$$

One outcome, multiple predictors variables

$$y_i = \alpha + \beta x_i + \epsilon_i$$

One predictor, one outcome variable


Generalized Linear Model (GLM)[1]
General Linear Models
Multiple Linear Regressions
Linear Regressions

The residuals can come from any distribution from the exponential family

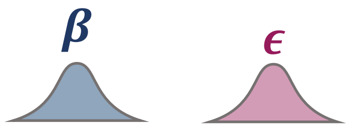
The link between the predictor variables and the outcomes can be any function

$$\mu_i = \mathbf{X}_i^T \cdot \beta \rightarrow g(\mu_i) = \mathbf{X}_i^T \beta$$

Statistical inference

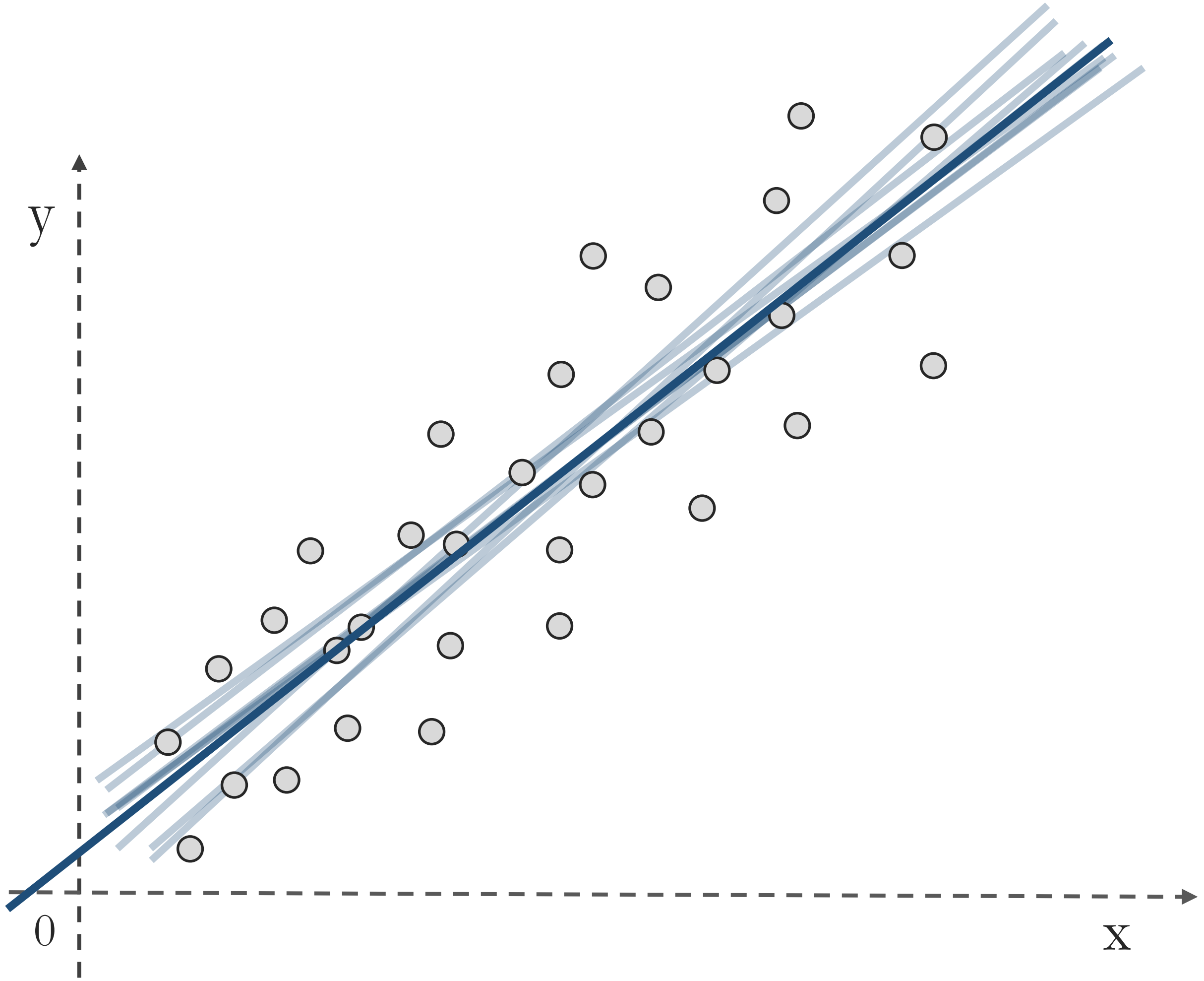
Three challenges of statistics:

1. Generalizing from sample to population
2. Generalizing from treatment to control group
3. Generalizing from observed measurements to underlying constructs of interest



[1] Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. In Journal of the Royal Statistical Society. Series A (General) (Vol. 135, Issue 3, p. 370). JSTOR. https://doi.org/10.2307/2344614

Statistical inference



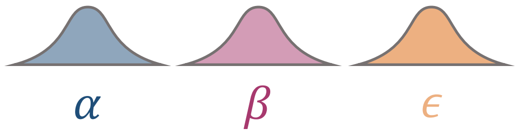
Linear Algebra

$$y_i = \alpha + \beta x_i + \epsilon_i$$

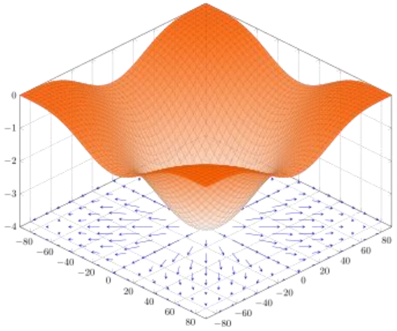
Probability Theory

Bayesian inference

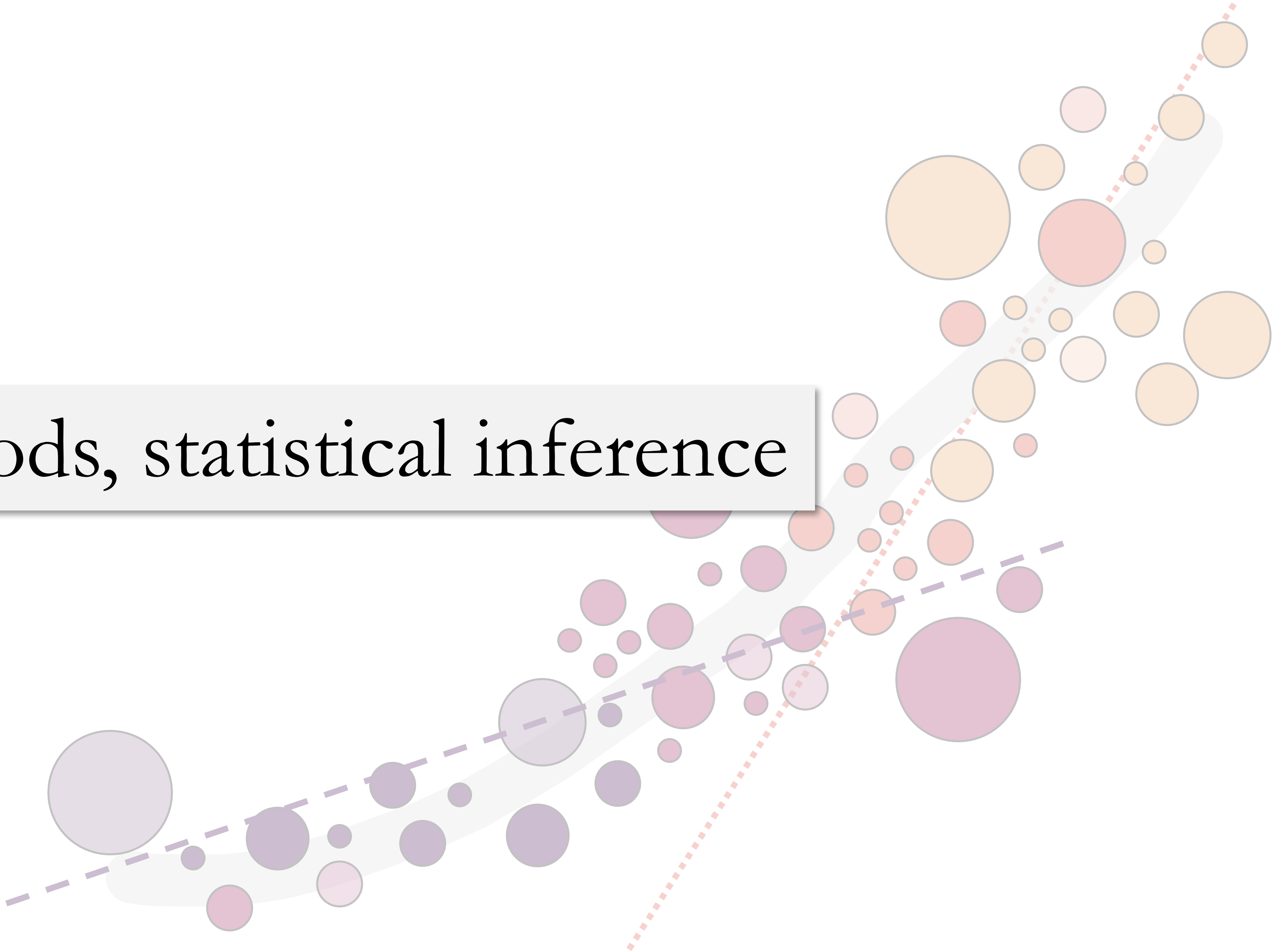
$$P(\alpha|y) = \frac{P(y|\alpha)P(\alpha)}{P(y)}$$



Calculus



Basic methods, statistical inference



A case study

What does it mean? →

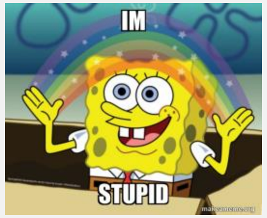
✓ **Validity**

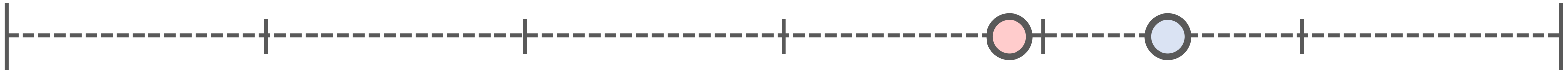
✓ **Reliability**

Measuring self-confidence in the population

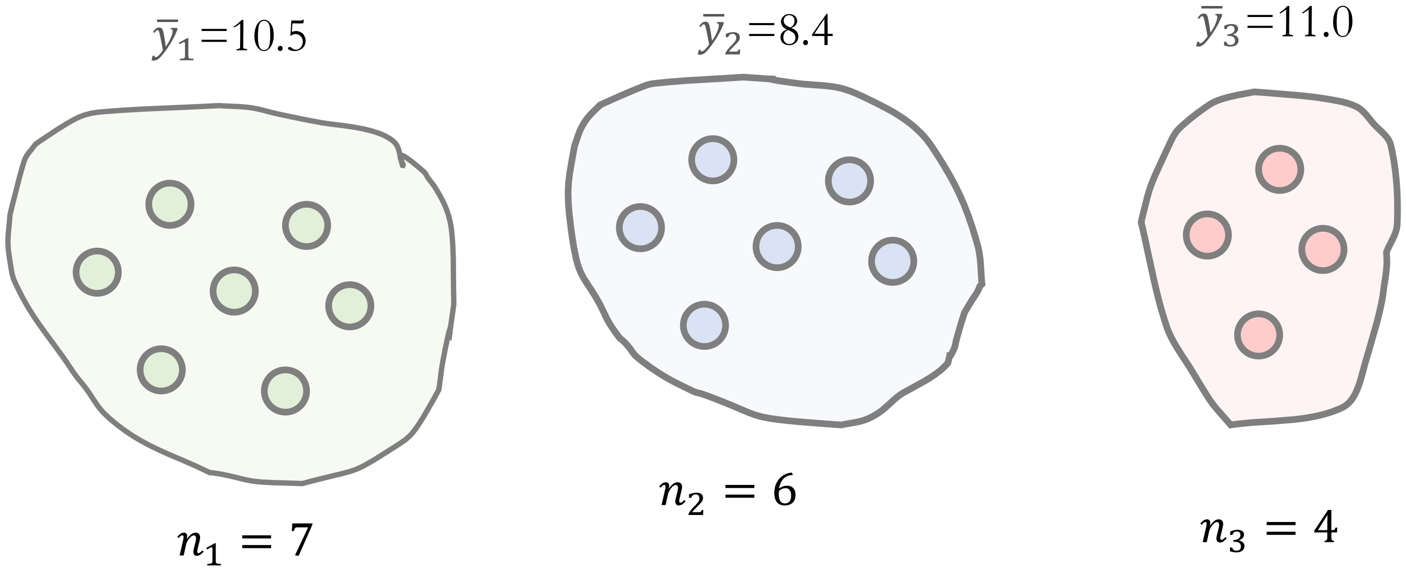
Not confident at all

Very confident





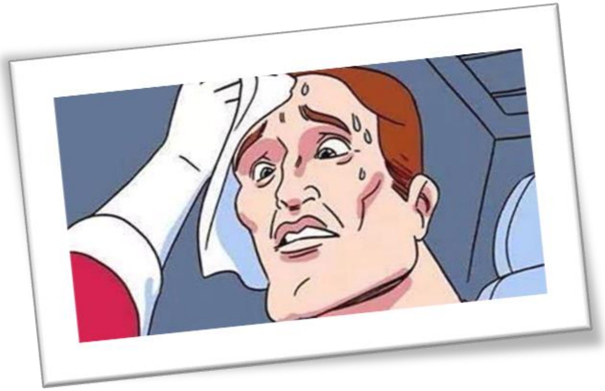
Weighted averages



$$weighted\ average = \frac{\sum_j N_j \bar{y}_j}{\sum_j N_j} = \sum_j \frac{N_j \bar{y}_j}{\sum_j N_j} = \frac{7}{17} \cdot 10.5 + \frac{6}{17} \cdot 8.4 + \frac{4}{17} \cdot 11.0$$

weights

In probability, the expectation of a random variable is a generalization of the weighted average:
$E[X] = x_1 p_1 + x_2 p_2, \dots, x_n p_n$ for discrete varialbes,
$E[X] = \int_{-\infty}^{+\infty} x f(x) dx$ for continuous variables



The data collection for Study 2 was a bit chaotic and you only managed to collect the average scores for the following groups of participants:

$$\bar{y}=\{6.4, 7.2, 8.1\}, n=\{14, 5, 12\}$$

As well as the following scores for individual participants:

$$\{5.0, 6.7, 8.8, 8.1, 9.0\}$$

What is the average score for the whole group of participants?

Solution

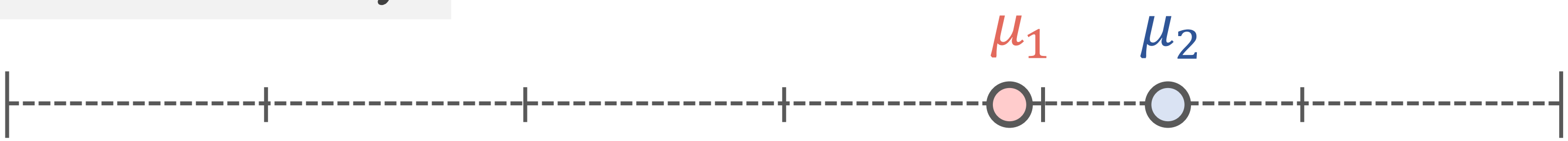
Let's first compute the mean from individual scores:

$$\overline{y_4} = \frac{5.0 + 6.7 + 8.8 + 8.1 + 9.0}{5} = \frac{37.6}{5} = 7.52$$

Using this value to compute the weighted sum (don't forget to include the number of individuals). We have 14 + 5 + 12 + 5 = 36 participant in total.

$$\bar{y} = \frac{5}{36} \cdot 7.52 + \frac{14}{36} \cdot 6.4 + \frac{5}{36} \cdot 7.2 + \frac{12}{36} \cdot 8.1 = 7.23$$

Quantifying uncertainty

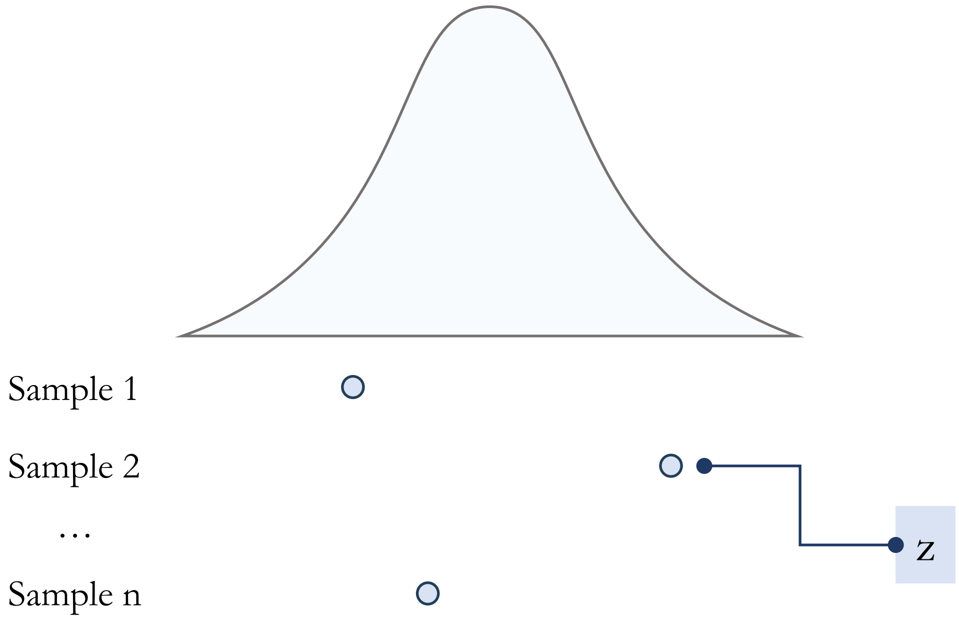


But what is a probability distribution?

A probability distribution corresponds to an urn with a potentially infinite number of balls inside. When a ball is drawn at random, the "random variable" is what is written on this ball.

Probabilistic distributions are used in regression modeling to help us characterize the variation that remains *after* predicting the average.

$z \sim N(\mu_z, \sigma_z^2)$



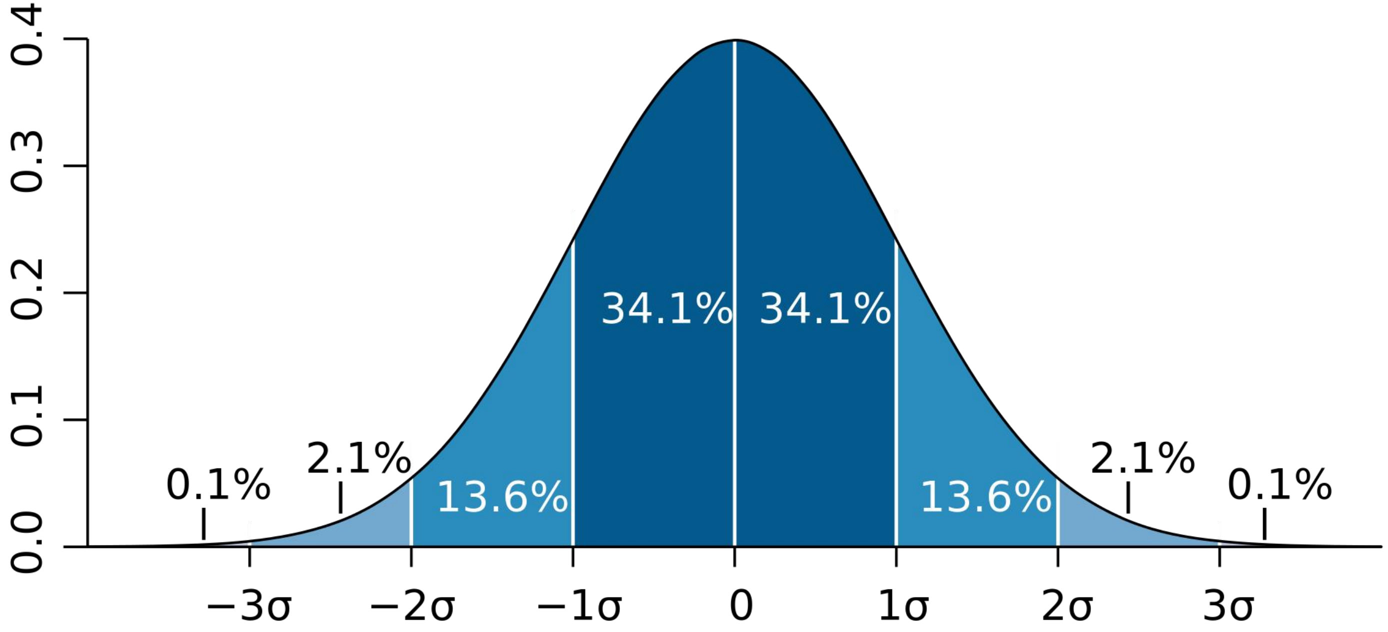
Using **R** (or any other programming language), sample 20 observations from a normal distribution using the parametrization of your choice.

The normal distribution

Probability density functions

$$f_1(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$f_2(x) = \sqrt{\frac{\tau}{2\pi}} e^{\frac{-\tau(x-\mu)^2}{2}}$$



Using **R** (or any other programming language), plot the functions f_1 and f_2 in the range -5.0 to 5.0 as described above using the following parameters: $\mu = 0.0, \sigma = 1.0, \tau = 1.0$. Can you spot any difference?

Do the same, but this time using the following parameters: $\mu = 0.0, \sigma = 3.0, \tau = 3.0$. How would you describe the influence of τ and σ on the width of the distribution?

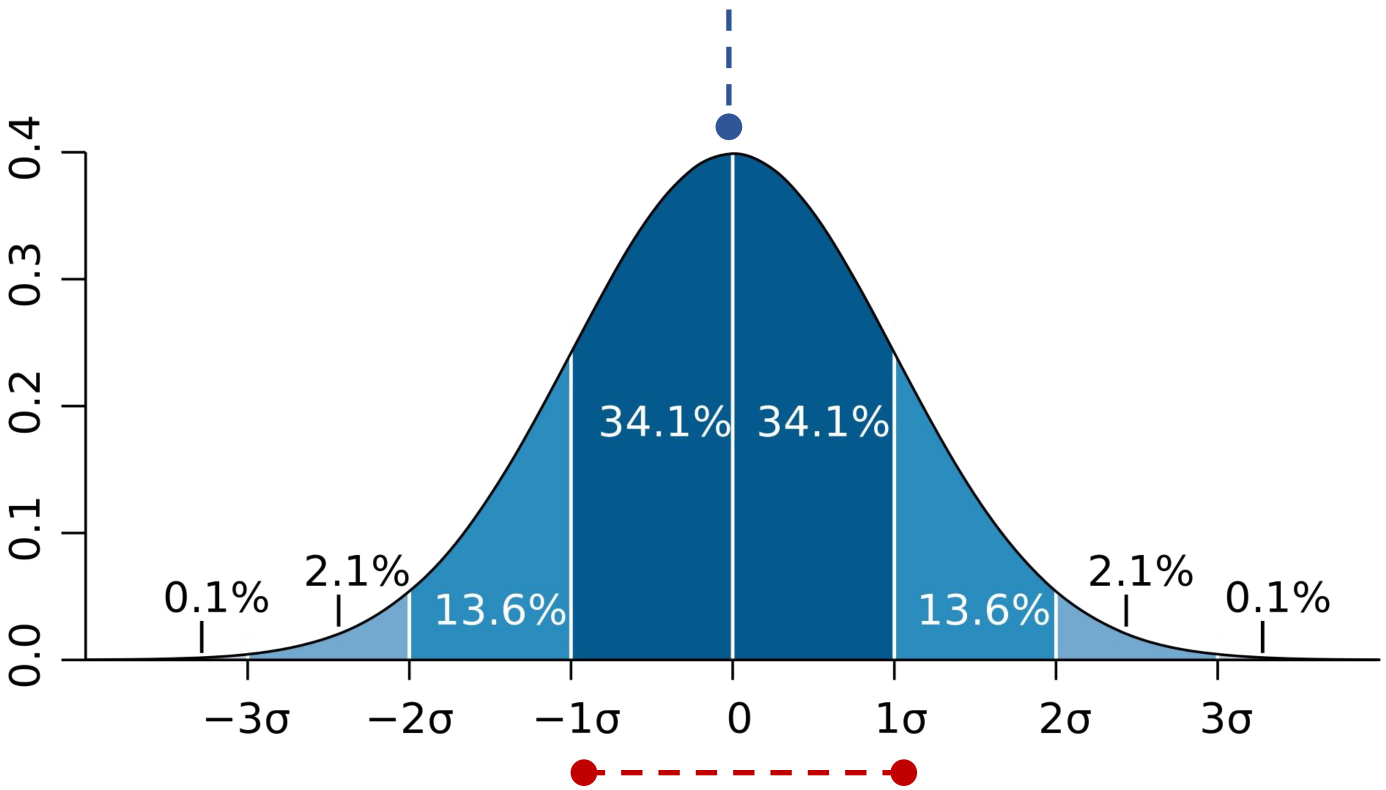
The normal distribution

Probability density functions

$$f_1(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$f_2(x) = \sqrt{\frac{\tau}{2\pi}} e^{\frac{-\tau(x-\mu)^2}{2}}$$

$\mu_z = E[z] = mean$

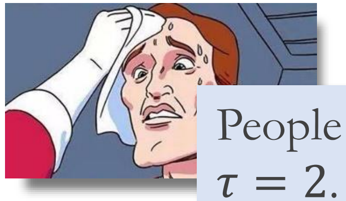


$\sigma_z^2 = E[(z-\mu_z)^2] = variance$

$\sigma_z = \sqrt{E[(z-\mu_z)^2]} = standard\ deviation$

$\tau_z = \frac{1}{variance} = \frac{1}{E[(z-\mu_z)^2]} = precision$

Let $f_1(x)$ be the probability density function of the normal distribution as defined above. Can we find x, μ, σ such as $f(x) > 1$?



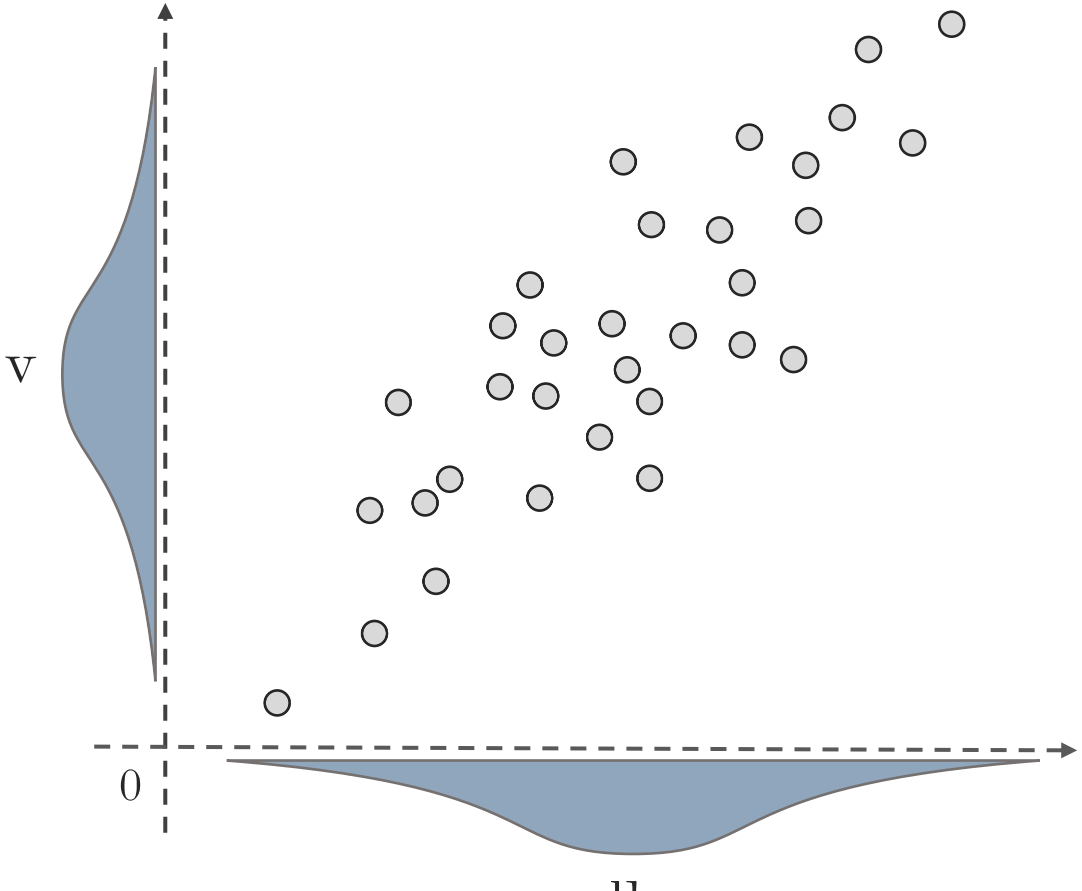
People from Switzerland have scores distributed normally with $\mu = 7.0$ and $\tau = 2$. Assuming that this is the real distribution, if I talk to 100 people in Switzerland, how many would have a score higher than 8.4?

Can you simulate this using R?

Manipulating random variables

Correlation between two random variables

$$\rho_{uv} = \frac{E[(u-\mu_u)(v-\mu_v)]}{\sigma_u\sigma_v}$$



Let $f_1(x)$ be the probability density function of the normal distribution as defined above. Can we find x, μ, σ such as $f(x) > 1$?

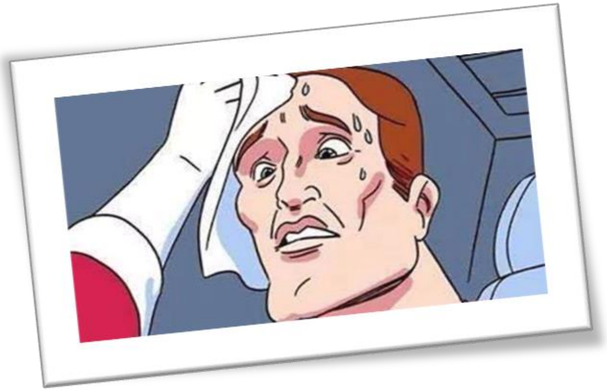
What is the difference between a correlation and a linear regression?

Summing two random variables

$$w = au + bv$$

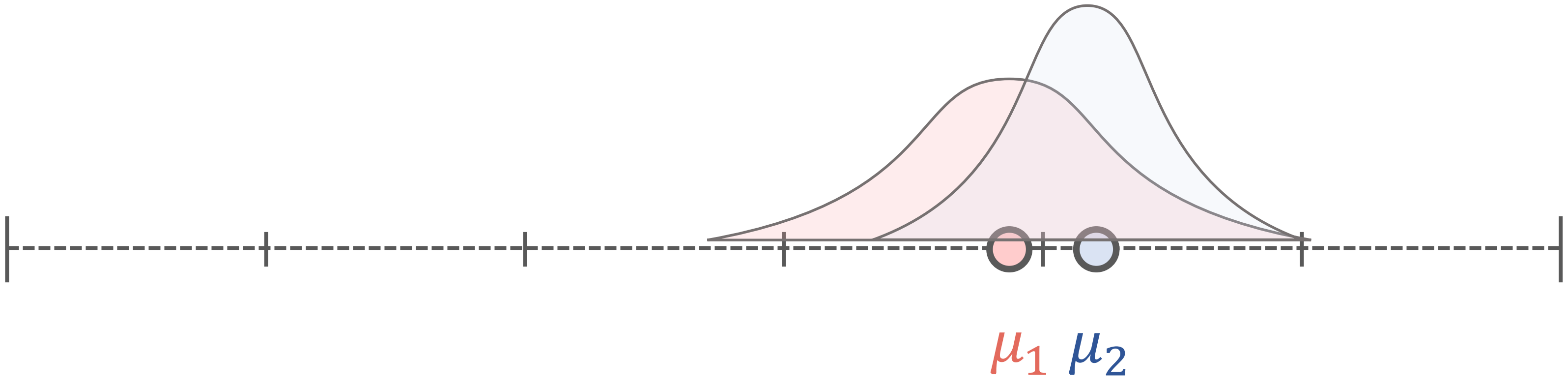
$$\mu_w = a\mu_u + b\mu_v$$

$$\sigma_w = \sqrt{a^2\sigma_u^2 + b^2\sigma_v^2 + 2ab\rho\sigma_u\sigma_v}$$



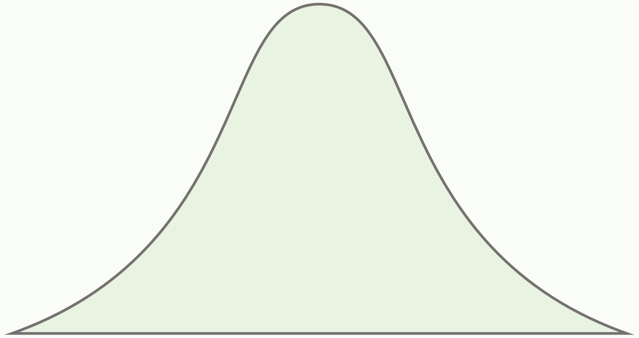
Data collection for Study 3 is better, but again, things were registered chaotically. We only know that French are as self-confident as 0.75 time Italians plus 0.5 time Spanish, whose distributions are given by: $\mu = \{5.8,\ \ 6.6\}$ and $\sigma = \{2.5,\ \ 1.6\}$. Danes have a distribution with $\mu = 7.0$ and $\tau = 0.081$.

Plot the distributions corresponding to the two populations.



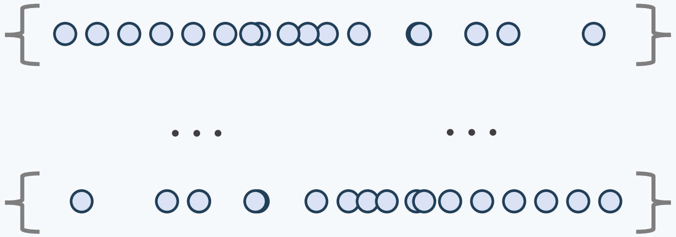
The standard error (of the mean)

Population



$N(\mu_p, \sigma_p^2)$

Sampling distributions

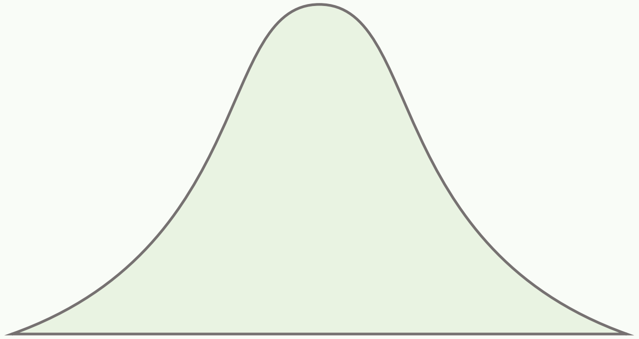


(n = 18)

$N(\mu_s, \sigma_s^2)$

The standard error (of the mean)

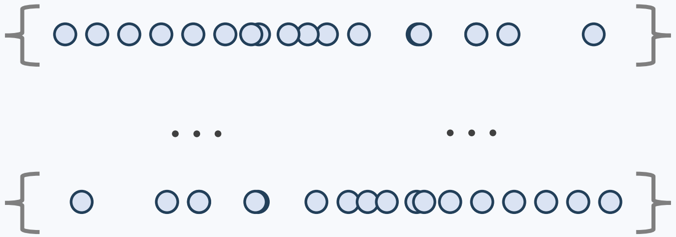
Population



$$N(\mu_p, \sigma_p^2)$$

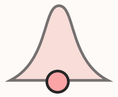
Sampling distributions

(n = 18)



$$N(\mu_s, \sigma_s^2)$$

Estimate of the mean



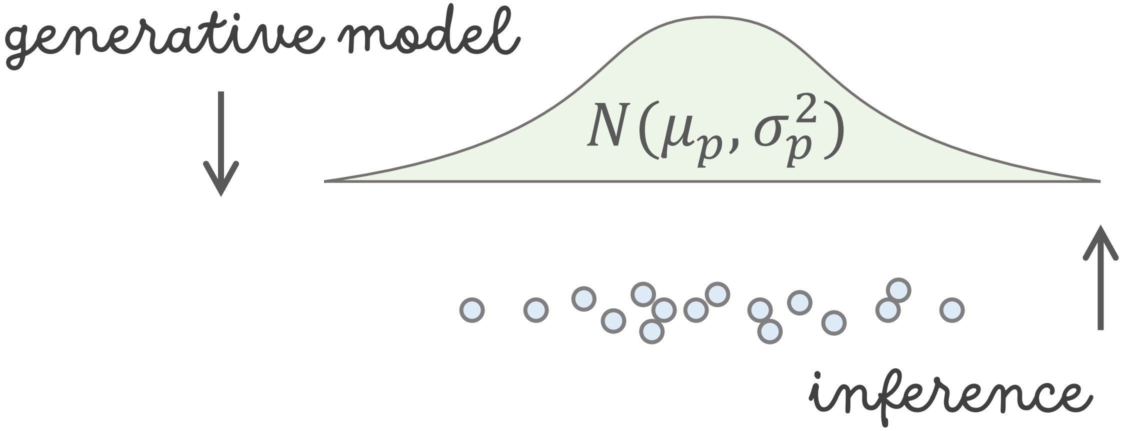
$$N(\mu_e, \sigma_e^2)$$

$$\sigma_e = \frac{\sigma_p}{\sqrt{n}} \approx \frac{\sigma_s}{\sqrt{n}}$$

Statistical inference (continued), simulation

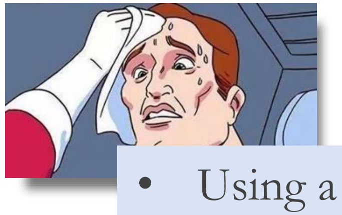
Nicolas Legrand – Researcher– ILAB (Interacting Minds Centre)

Simulations – Measurement error


generative model

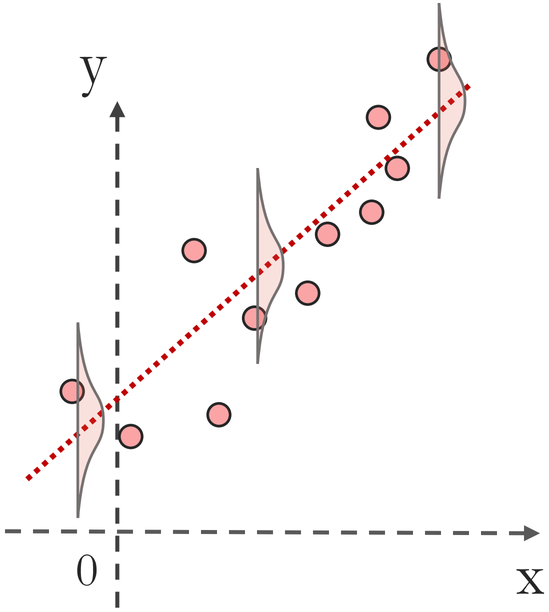
$N(\mu_p, \sigma_p^2)$

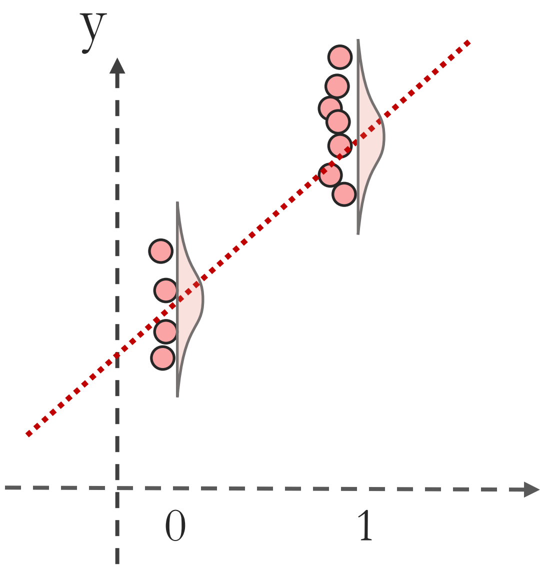
inference



- Using a normal distribution with the parameters of your choice, sample 200 times 5 observations and plot the distribution of their mean using a histogram.
- Do the same, but this time using 50 observations each time.
- Which estimate is the most reliable?

Simulations – Sampling distribution





$$\mu_z = 5.0$$
$$\sigma_z = 3.0$$
$$\alpha = 1.5$$
$$\beta = 2.0$$
$$z \sim N(\mu_z, \sigma_z^2)$$
$$y_i = \alpha + \beta x_i$$

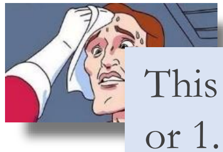
Using R, create a scatterplot of the variables *x* and *y* as defined by the model above.

Stop using linear regression
You know what? I'm gonna start using linear regression even harder

…

$$\sigma_\epsilon = 0.5$$
$$\epsilon_i \sim N(0, \sigma_\epsilon^2)$$
$$y_i = \alpha + \beta x_i + \epsilon_i$$

Using R, create a scatterplot of the variables *x* and *y* as defined by the model above that adds noise to the previous definition.



This is the same, except that *x* is either 0 or 1. We can do this by "hard-coding" the values, or by sampling from a binomial distribution (try it).

Summarizing a set of simulations

Median Absolute Deviation

$$M = med(z_1, \dots, z_n)$$

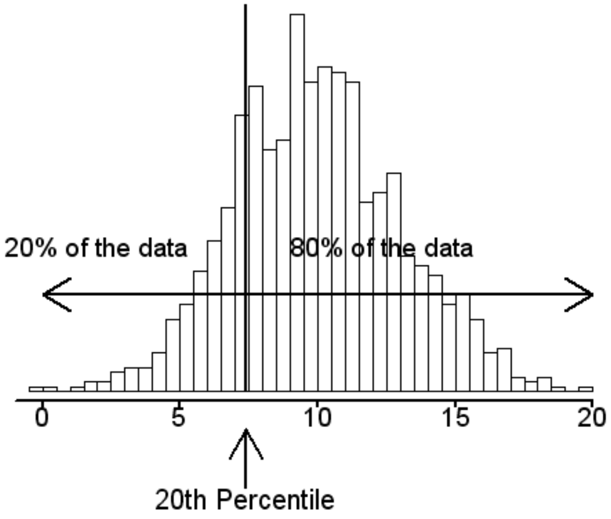
$$mad_X = med_{i=1}^{n}(|z_i - M|)$$

$$sd_X = mad_X * 1.483$$

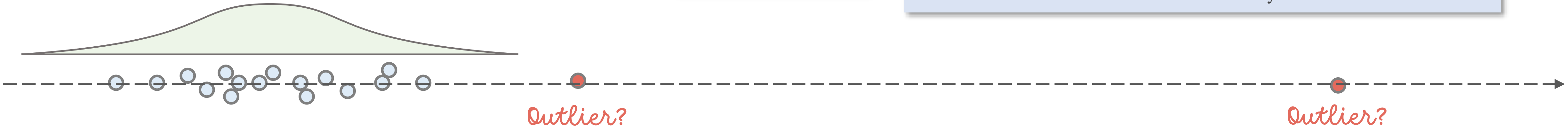
Robust standard deviation

Uncertainty intervals

Using R, **quantile(z, 0.25, 0.75)** returns a central 50% interval and **quantile(z, 0.025, 0.975)** returns a central 95% interval.



Generate 200 observations from a Gaussian distribution with mean=8.0 and standard deviation = 2.5. Compute the robust standard deviation and 95% uncertainty interval.

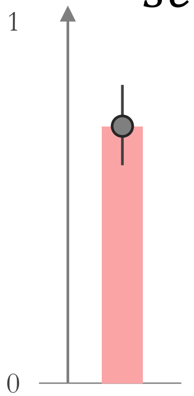


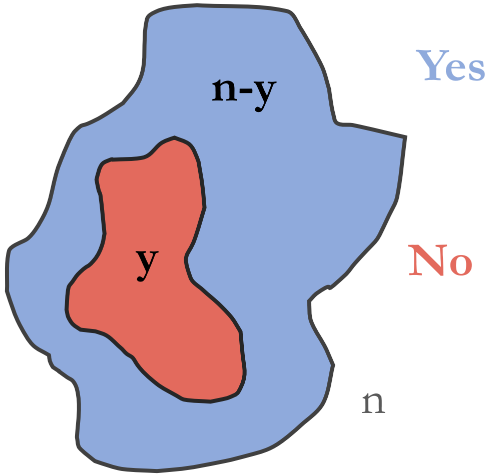
Standard errors

Standard errors for proportions

$$p = \frac{y}{n}$$

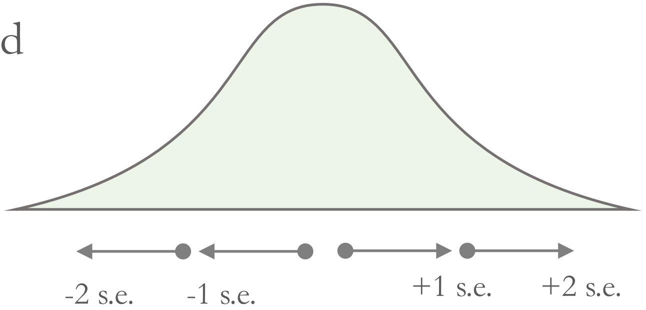
$$se = \sqrt{\frac{p(1-p)}{n}}$$



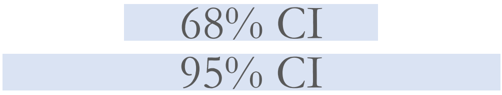


- We asked 50 people whether they would accept to participate in our study. 28 have declined. Compute *p*, the proportion of people that would accept to participate in the population, and the standard error around this estimate.
- We now asked 1000 and 569 have declined. How does this change your estimate?

The standard error is the estimated standard deviation of an estimate of interest.

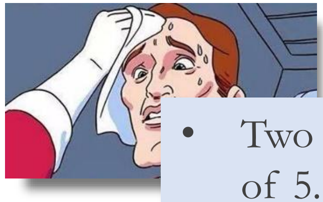


The x% confidence interval will include the true value of the parameter x% of the time.



Standard errors for differences

$$se = \sqrt{se_1^2 + se_2^2}$$



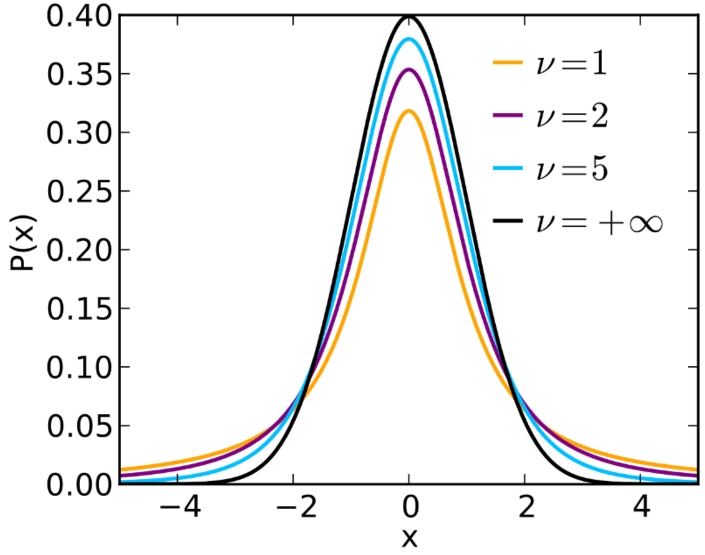
- Two groups were tested and have scores of 5.6 ±1.1 s.e. and 6.3 ±0.8 s.e. Compute their difference and the standard error around this difference.

Degrees of Freedom

t distribution

Generally, the more degrees of freedom you have, the more closely your sampling distribution resembles a normal distribution, which has narrower tails and less variability. This means that your confidence interval will be narrower and more precise.

On the other hand, the fewer degrees of freedom you have, the more skewed and fat-tailed your sampling distribution will be, which means that your confidence interval will be wider and less precise.



This might help

This as well

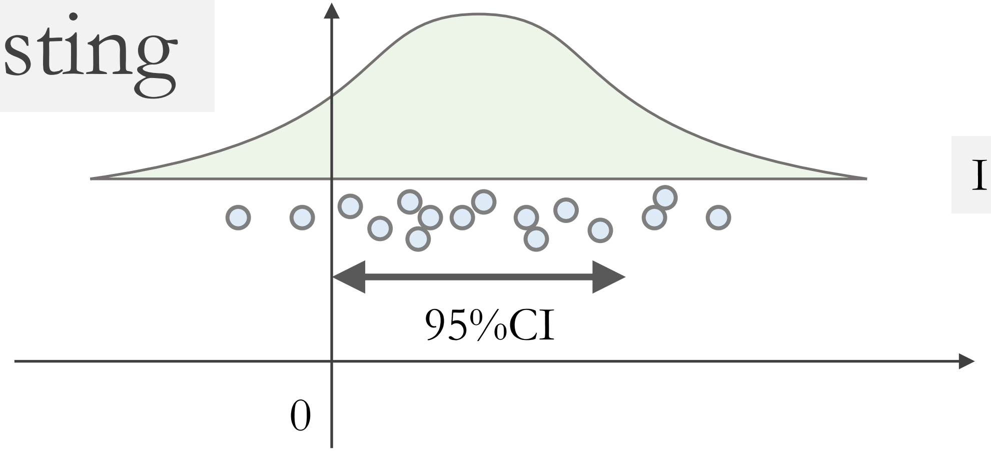
link

We have observed the following scores from Study 4:

$x = \{6.5, 5.5, 6.0, 5.7, 8.1, 7.7, 6.9, 9.2, 5.6\}$

Compute the 95% CI using base R (no *lm()*).

Hypothesis testing



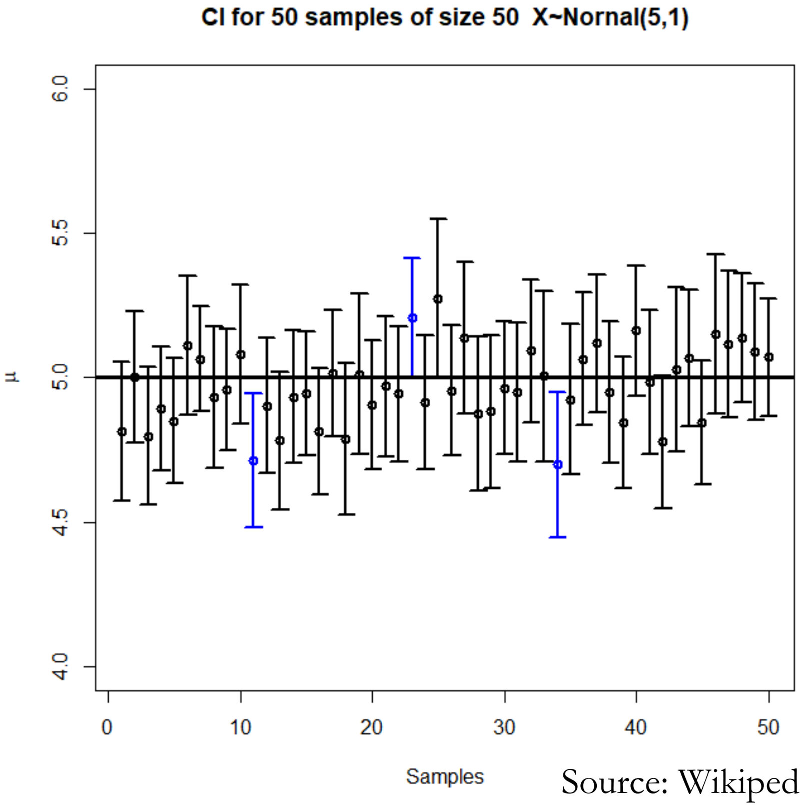
Is it different from 0?

Yes

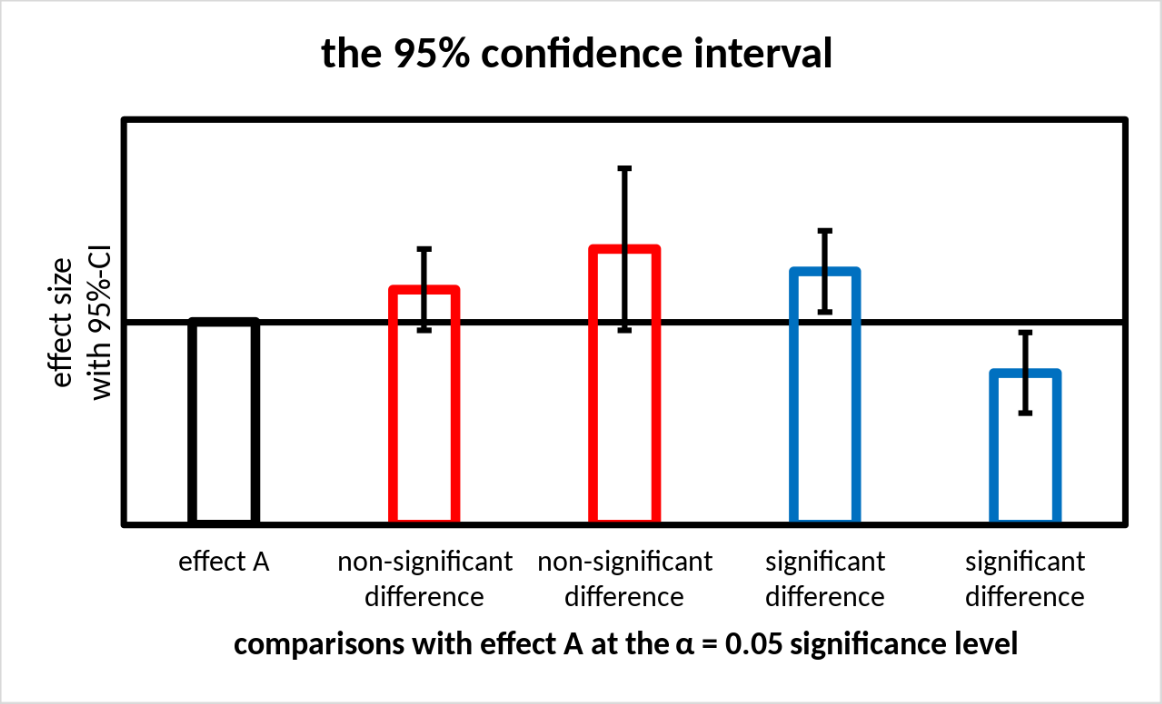
No

Let's see how it works

[link]



Source: Wikipedia



Source: Wikipedia