Basic methods, statistical inference

Nicolas Legrand – Researcher– ILAB (Interacting Minds Centre)

AARHUS UNIVERSITY

A case study

What does it mean?

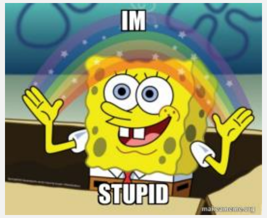
Validity

Reliability

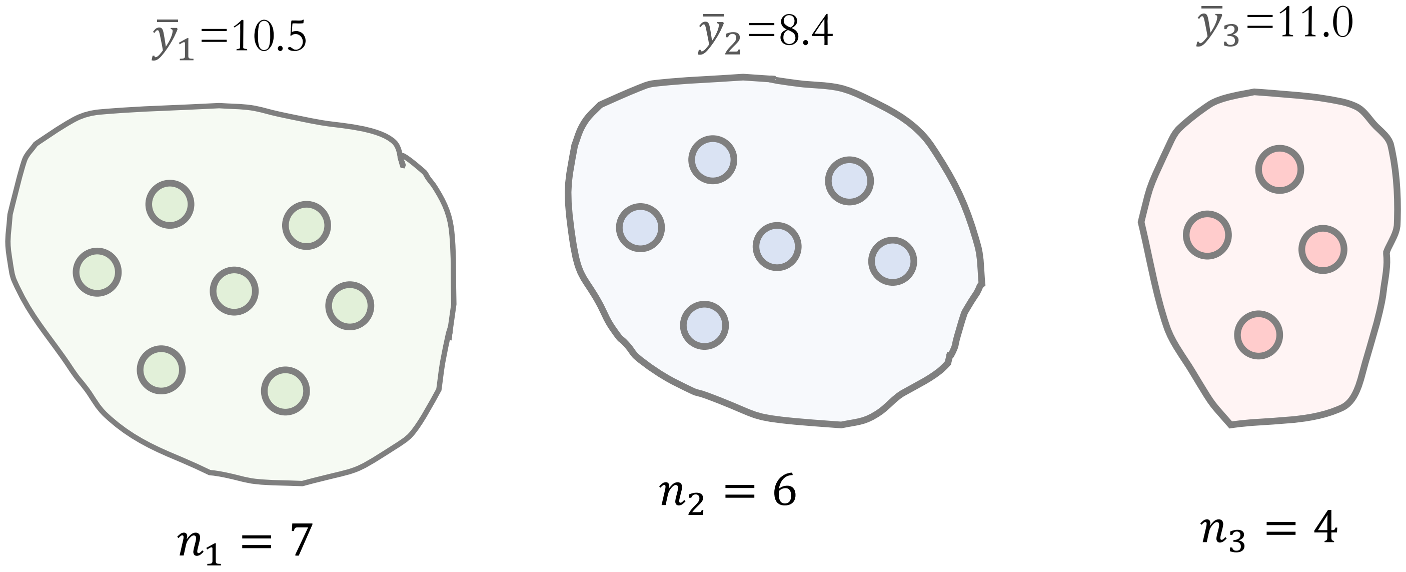
Measuring self-confidence in the population

Not confident at all

Very confident



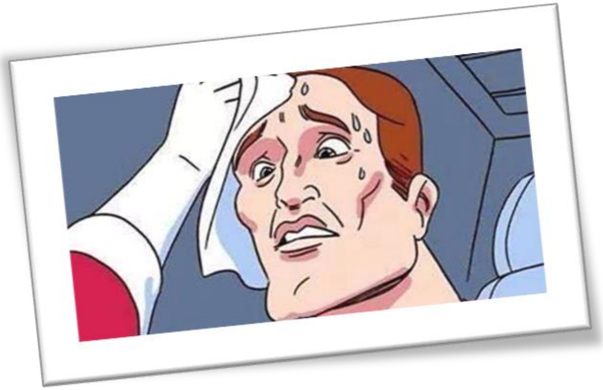
Weighted averages



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

weights

In probability, the expectation of a random variable is a generalization of the weighted average:
$E[X] = x_1 p_1 + x_2 p_2, \ldots, x_n p_n$ 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.

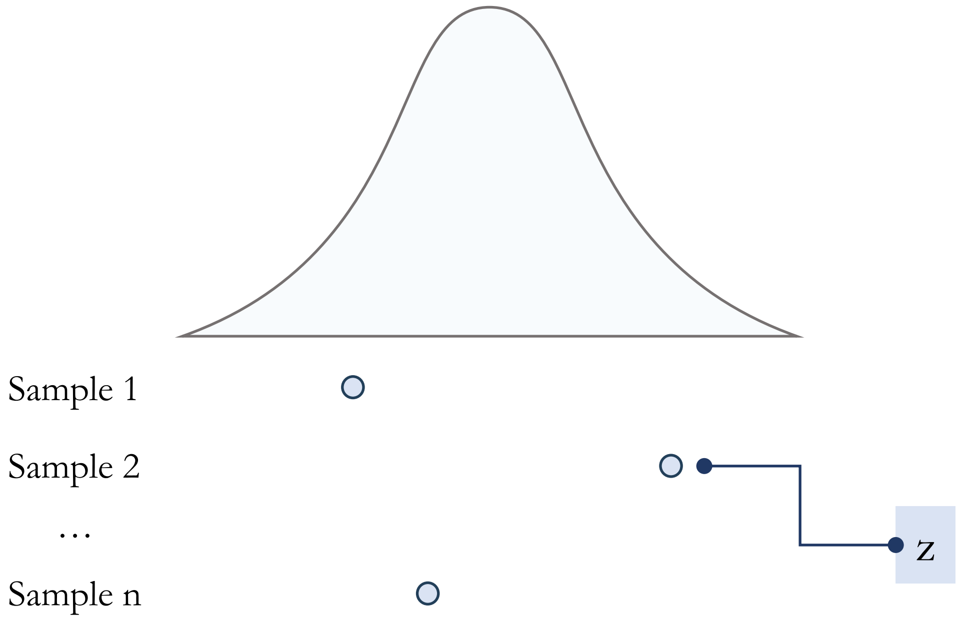
$$\bar{y} = \frac{5}{36} \cdot 7.52 + \frac{14}{36} \cdot 6.4 + \frac{5}{36} \cdot 7.2 + \frac{12}{36} \cdot 8.1 = 7.23$$

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.

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



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

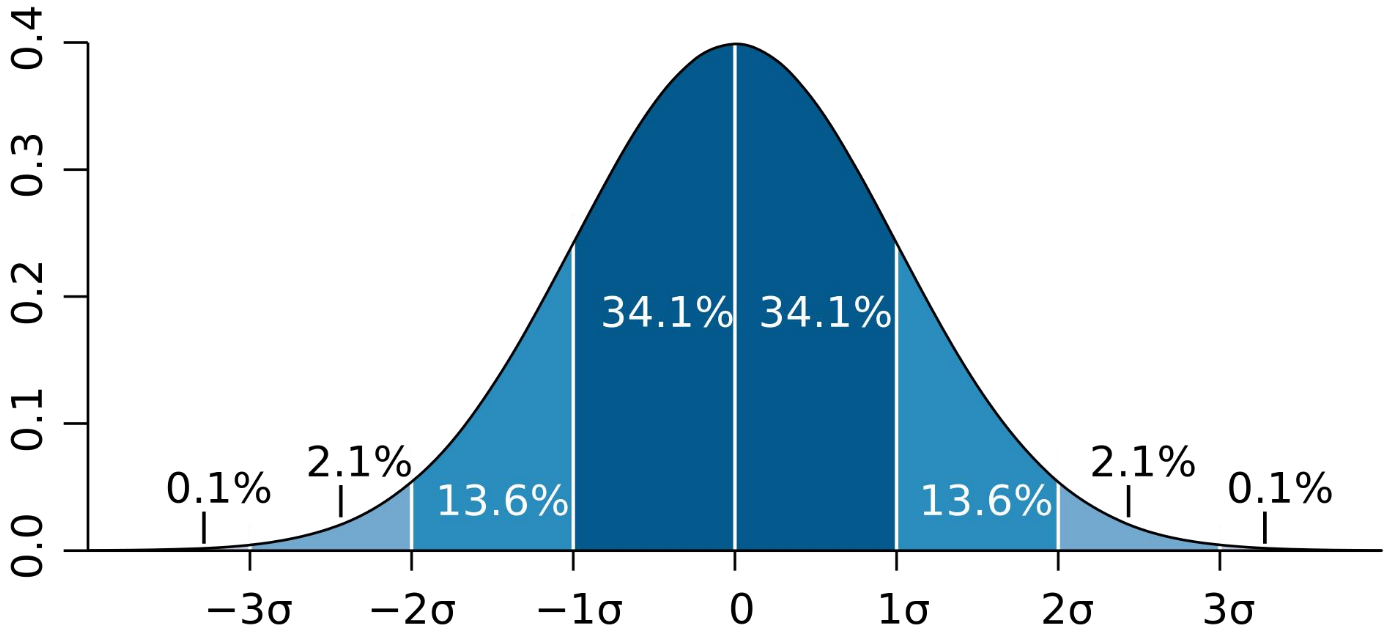
Using **R** (or any other programming language), sample 20 observations from a normal distribution using the parametrization of your choice.

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?

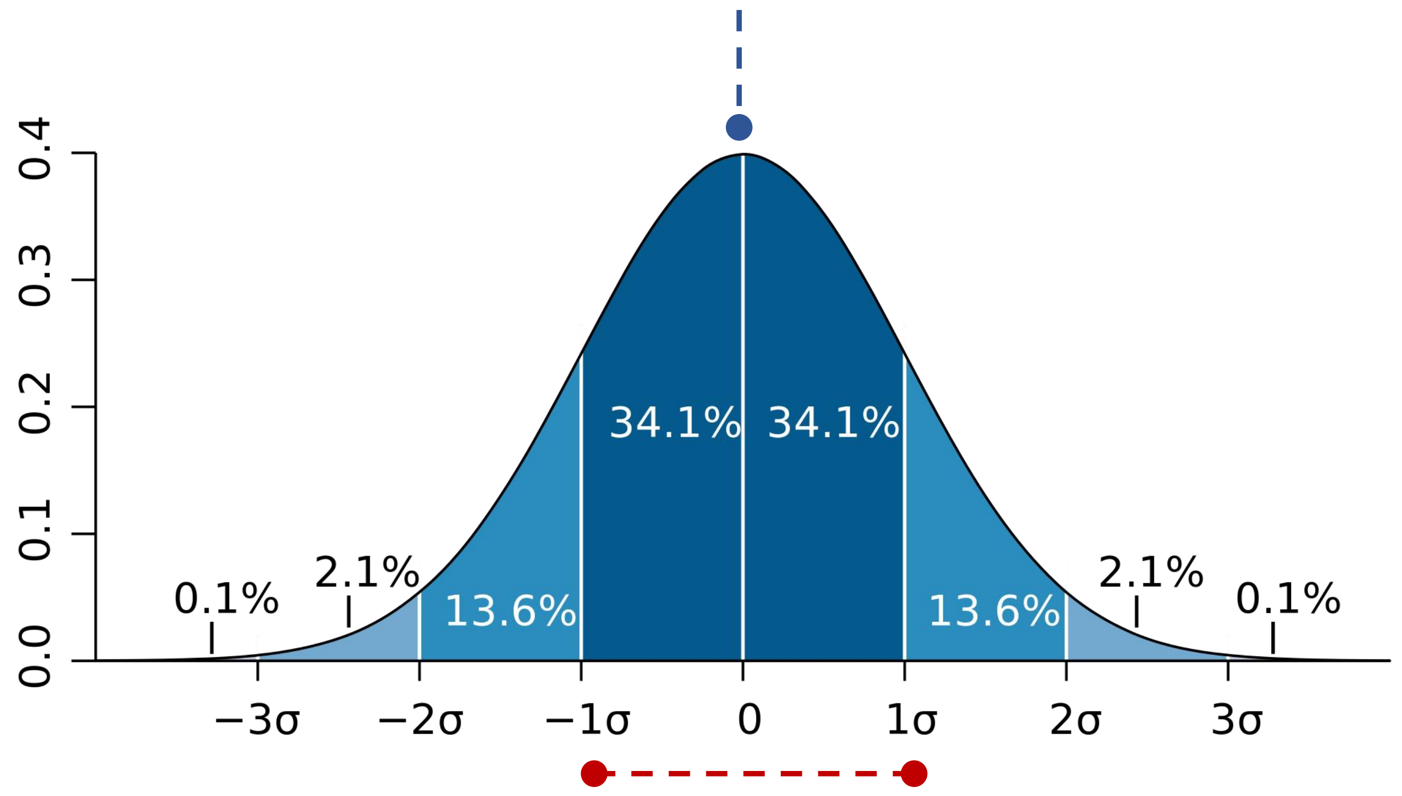
The normal distribution

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 $f(x) > 1$?

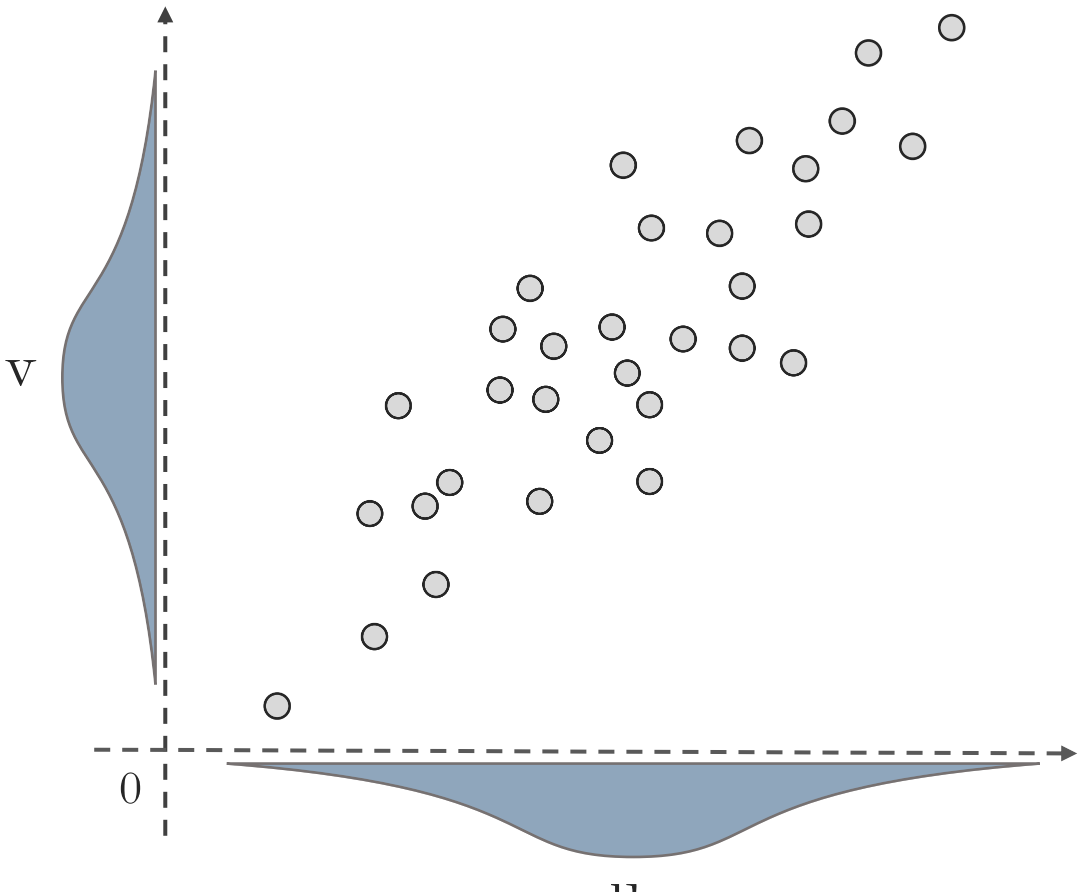
People from Switzerland have scores distributed normally with $\mu = 7.0$ and $\tau = 2$. Assuming that this is the real distribution, if I talk to 100 people in Switzerland, how many would have a score higher than 8.4?

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 $f(x) > 1$?

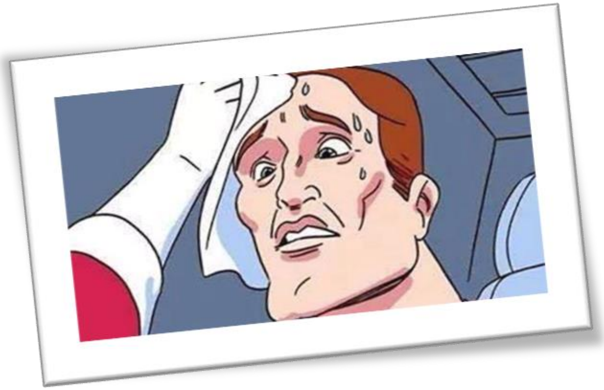
What is the difference between a correlation and a linear regression?

Summing two random variables

$$w = au + bv$$

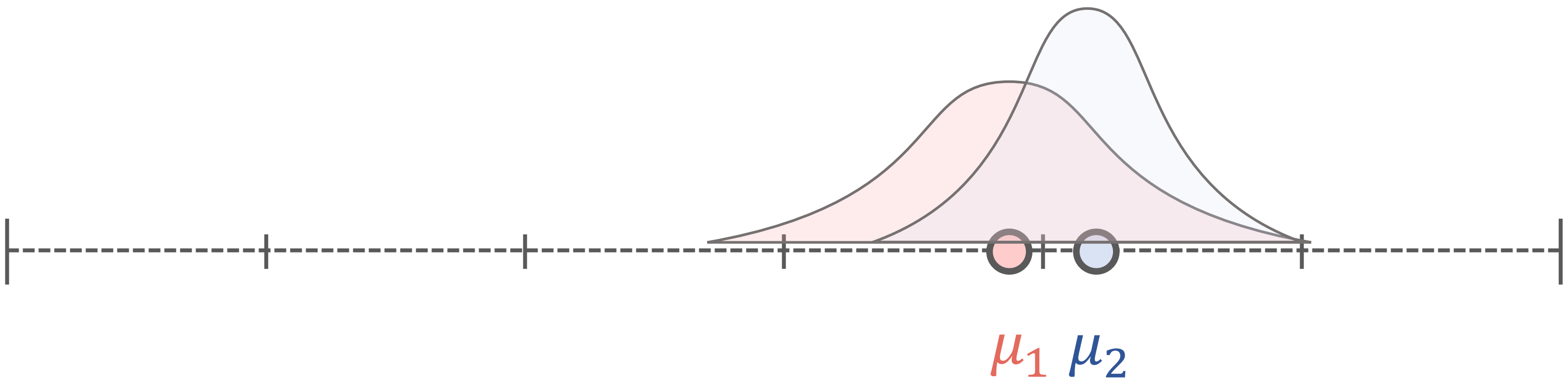
$$\mu_w = a\mu_u + b\mu_v$$

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



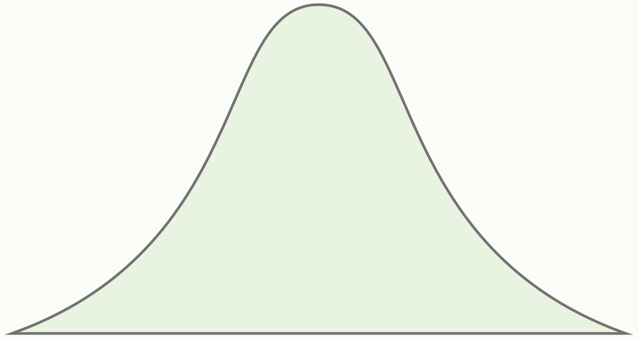
Data collection for Study 3 is better, but again, things were registered chaotically. We only know that French are as self-confident as 0.75 time Italians plus 0.5 time Spanish, whose distributions are given by: $\mu = \{5.8,\ 6.6\}$ and $\sigma = \{2.5,\ \ 1.6\}$. Danes have a distribution with $\mu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

Plot the distributions corresponding to the two populations.



The standard error (of the mean)

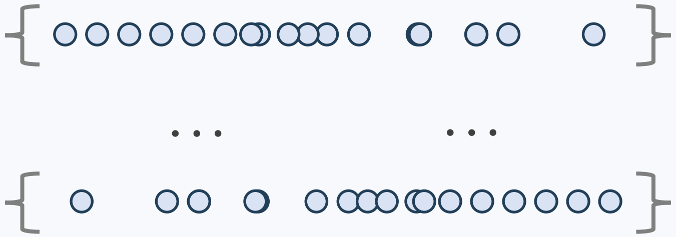
Population



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

Sampling distributions

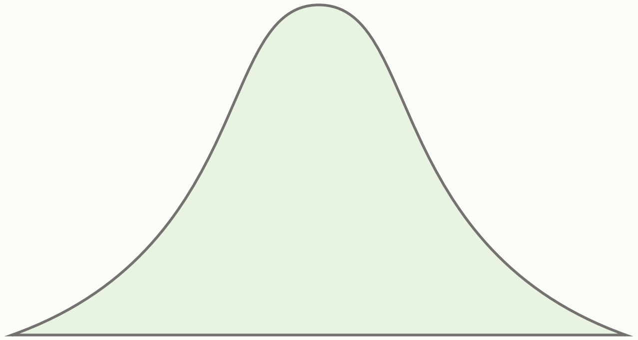
(n = 18)



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

The standard error (of the mean)

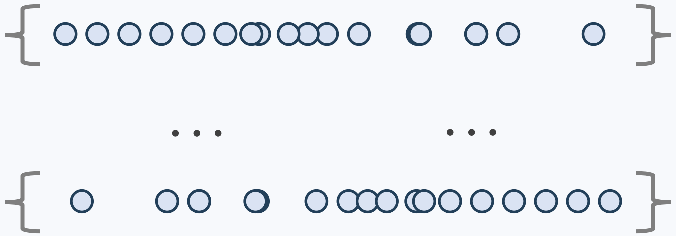
Population



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

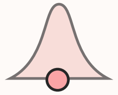
Sampling distributions

(n = 18)



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

Estimate of the mean



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

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