Knowledge and solutions for a changing world

Be boundless

Advancing data-intensive discovery in all fields

DESCRIPTIVE STATISTICS

UW DIRECT

(Data Intensive Research Enabling Cutting-edge Tech)

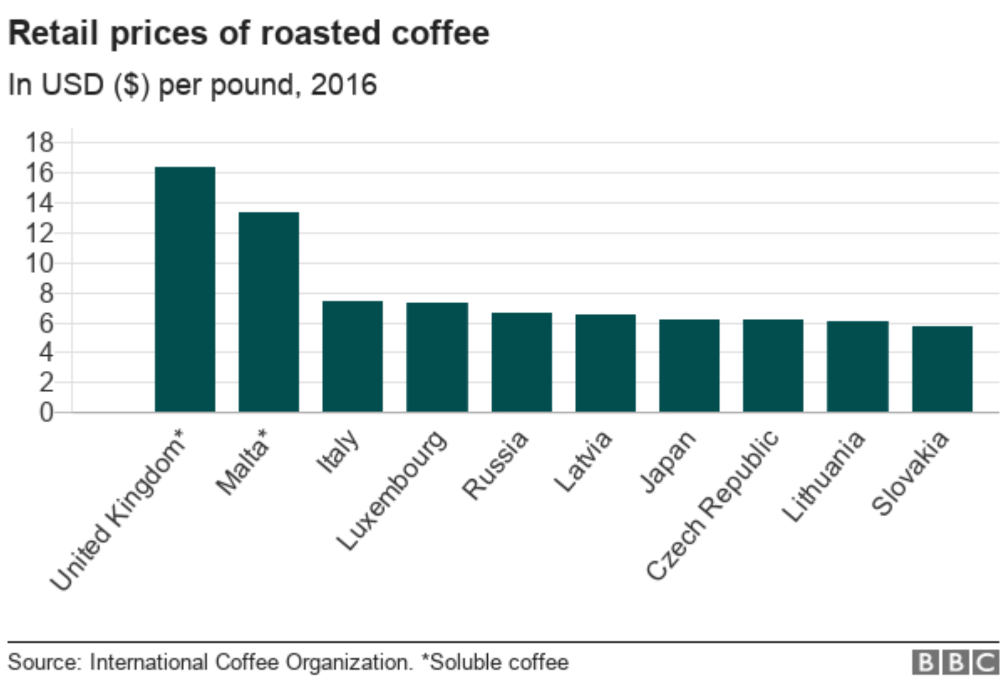
https://uwdirect.github.io

Stéphanie Valleau

Chemical Engineering

What is statistics?

The practice or science of **collecting** and **analyzing** numerical data in **large quantities**, especially for the purpose of **inferring proportions** in a whole from those in a **representative sample**.

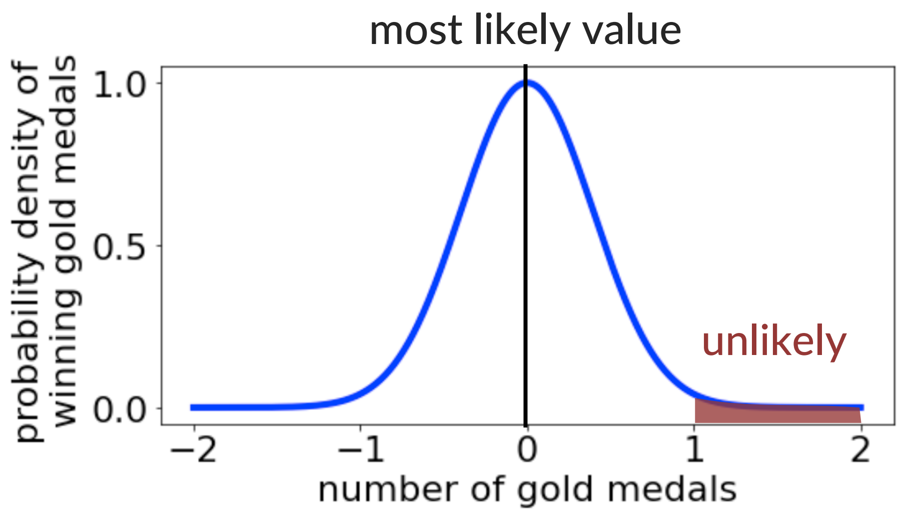


https://multimedia.journalism.berkeley.edu/tutorials/databases-of-statistical-information/

What can we learn from Stats?

Statistics does not tell us **whether we are right** in coming to a conclusion (not absolute).

It tells us the **likelihood** of an outcome / the chances of being **wrong**



Two key concepts

Population

All possible values of an experimental variable (e.g. all stars in the milky way, all types of enzymes etc.)

Sample

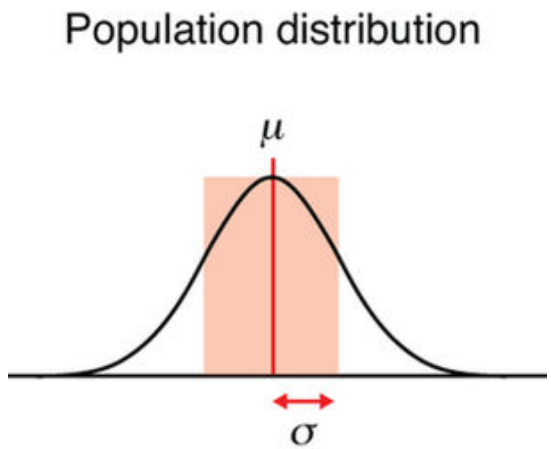
A set of data drawn from a population

Often want to know the **mean (μ)** and **standard deviation (σ)** of a population

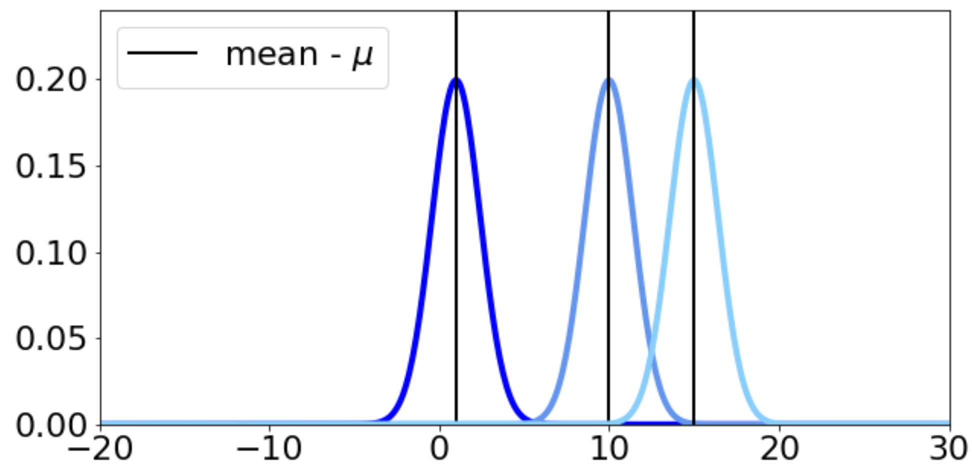
$$\mu = \sum_{i=1}^{N} \frac{X_i}{N}$$

$$\sigma = \sqrt{\frac{\sum_i^N (X_i - \mu)^2}{N}}$$

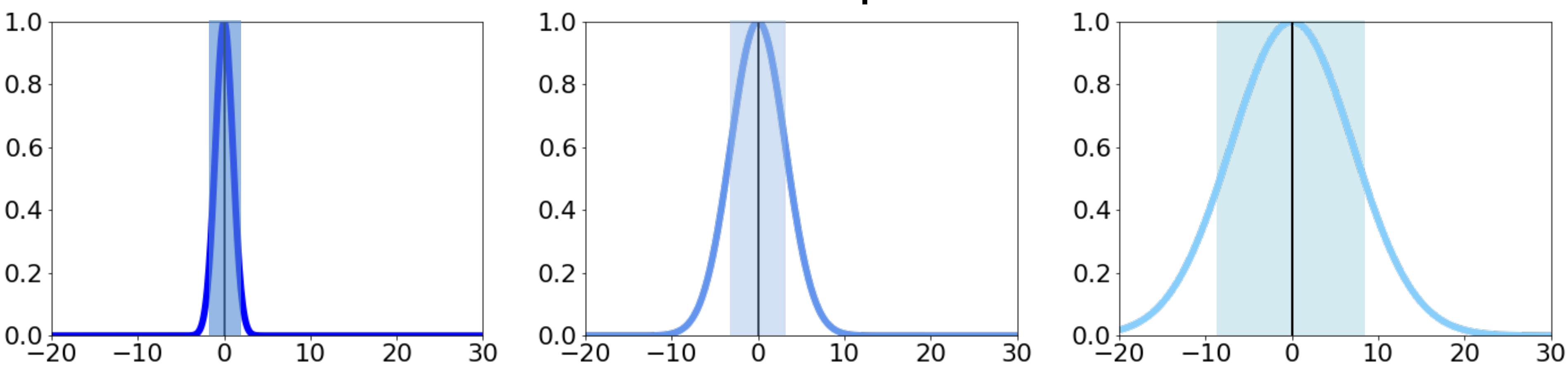
Mean & standard deviation



Mean – location



Standard deviation – spread



Standard dev. & Variance

Variance is square of std. dev.

$$\sigma = \sqrt{\frac{\sum_i^N (X_i - \mu)^2}{N}}$$

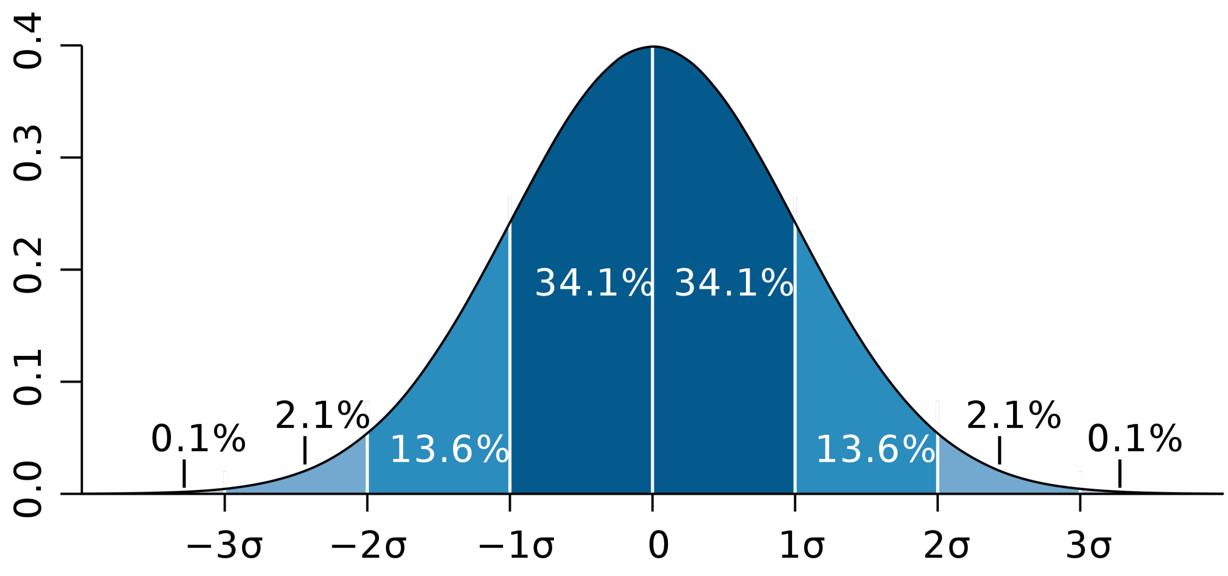
$$\sigma^2 = \frac{\sum_i^N (X_i - \mu)^2}{N}$$

± 0.5 σ contains 39% of possible values

± 1 σ contains 68%

± 2 σ contains 95%

± 3 σ contains 99.7%

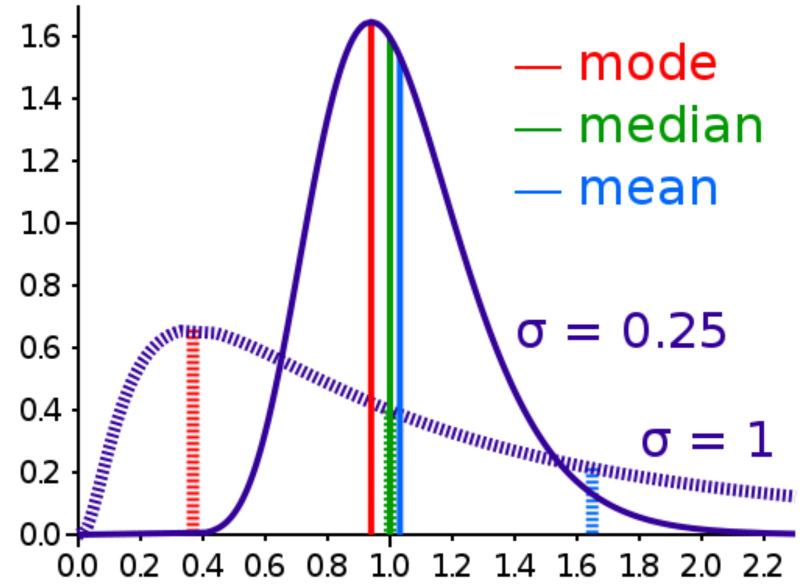


https://en.wikipedia.org/wiki/Normal_distribution

Median vs mean

The **median** is a value separating the higher half from the lower half of a data sample, a population, or a probability distribution

Median is often **more robust** than the mean in the face of **skewed distributions** and outliers



Measuring μ, σ ?

Is it possible to obtain the population for an experimental variable?

Is it possible to directly measure the mean (μ) and standard deviation (σ) of a population

If we had a population ....

Open the L3_Descriptive_Statistics.ipynb notebook

Back to reality

We could estimate the correct mean and standard deviation from samples …

Population samples

Samples are **sets of data drawn from the population**

- Described by their size *n* (number of data points)
- Notation: X indexed by sample subscript, e.g. X_1

How to choose *n*?

Sampling Bias

Do all values in a population have the **same chance of being selected**?

If not, we have **bias**.

What is an example of bias in sampling?

Example: Assess average level of knowledge of US population based on survey responses from high school students

Population samples

Population: mean (μ) and s.d. (σ)

Sample: mean - $\overline{X}$ and standard deviation - s

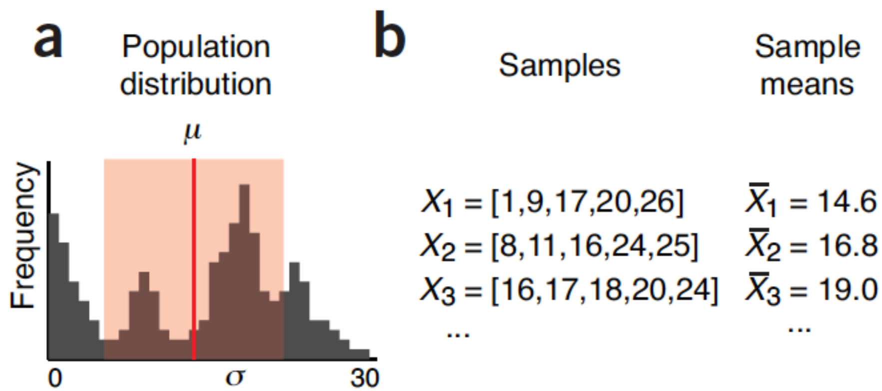


Figure 2 | Population parameters are estimated by sampling. (**a**) Frequency histogram of the values in a population. (**b**) Three representative samples taken from the population in **a**, with their sample means. (**c**) Frequency histogram of means of all possible samples of size $n = 5$ taken from the population in **a**.

[Nature, 2013. Importance of being uncertain]

Population samples

Sample parameters like $\bar{X}$ have their **own distributions** – e.g. panel **C**

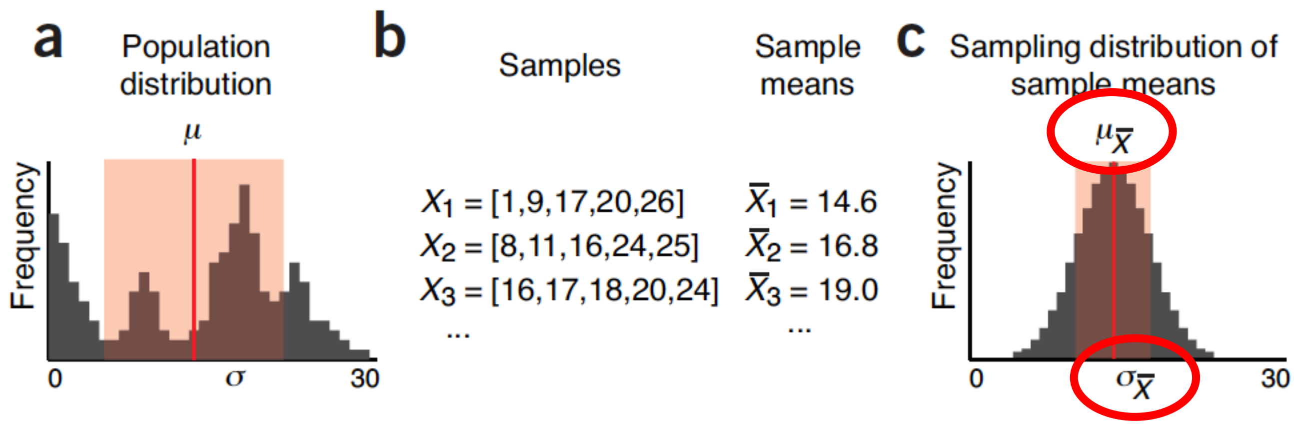


Figure 2 | Population parameters are estimated by sampling. (**a**) Frequency histogram of the values in a population. (**b**) Three representative samples taken from the population in **a**, with their sample means. (**c**) Frequency histogram of means of all possible samples of size $n = 5$ taken from the population in **a**.

Nature, 2013. Importance of being uncertain

Sampling **with** replacement

N=7 population values = [12, 13, 14, 15, 16, 17, 18]

Sample **with** replacement (n=2)

- First sample, each item has 1/7 probability
- Second sample, each item has 1/7 probability
- How many total possibilities (assuming order is important)?

$$N^n = 49$$

Each time we sample our choice is **independent** from the prior choice!

Sampling **without** replacement

Population values = [12, 13, 14, 15, 16, 17, 18]

Sample **without** replacement (n=2)

- First sample, each item has 1/7 probability
- Second sample? 1/6
- How many total possibilities (assuming order is important)?

$$\frac{N!}{(N-n)!} = \frac{7!}{(5-2)!} = \frac{7 \cdot 6 \cdot 5 \cdot \ldots \cdot 1}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 7 \cdot 6 = 42$$

Note: Each time we sample our choice **depends** on the previous choice!

Will sampling work?

Let’s go back to the notebook

Sampling distributions

We observed that $\sigma_{\bar{X}} < \sigma$

and also that $\mu \cong \mu_{\bar{X}}$

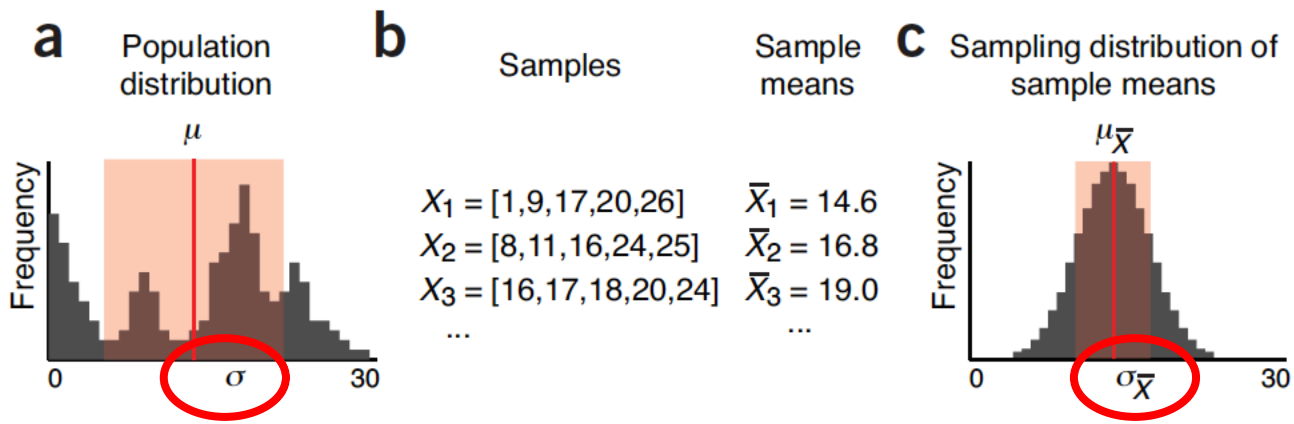
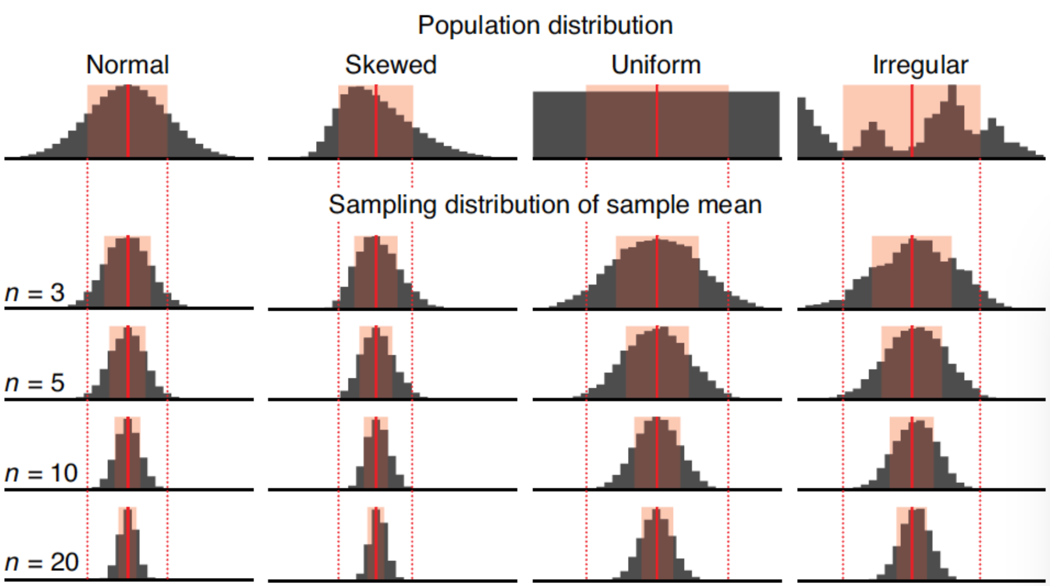


Figure 2 | Population parameters are estimated by sampling. (**a**) Frequency histogram of the values in a population. (**b**) Three representative samples taken from the population in **a**, with their sample means. (**c**) Frequency histogram of means of all possible samples of size $n = 5$ taken from the population in **a**.

Nature, 2013. Importance of being uncertain

The Central limit theorem

As ***n*** increases, the distribution of sample means tends to become a ***normal distribution***, **regardless of population distribution shape**



[Nature, 2013. Importance of being uncertain]

Central limit theorem

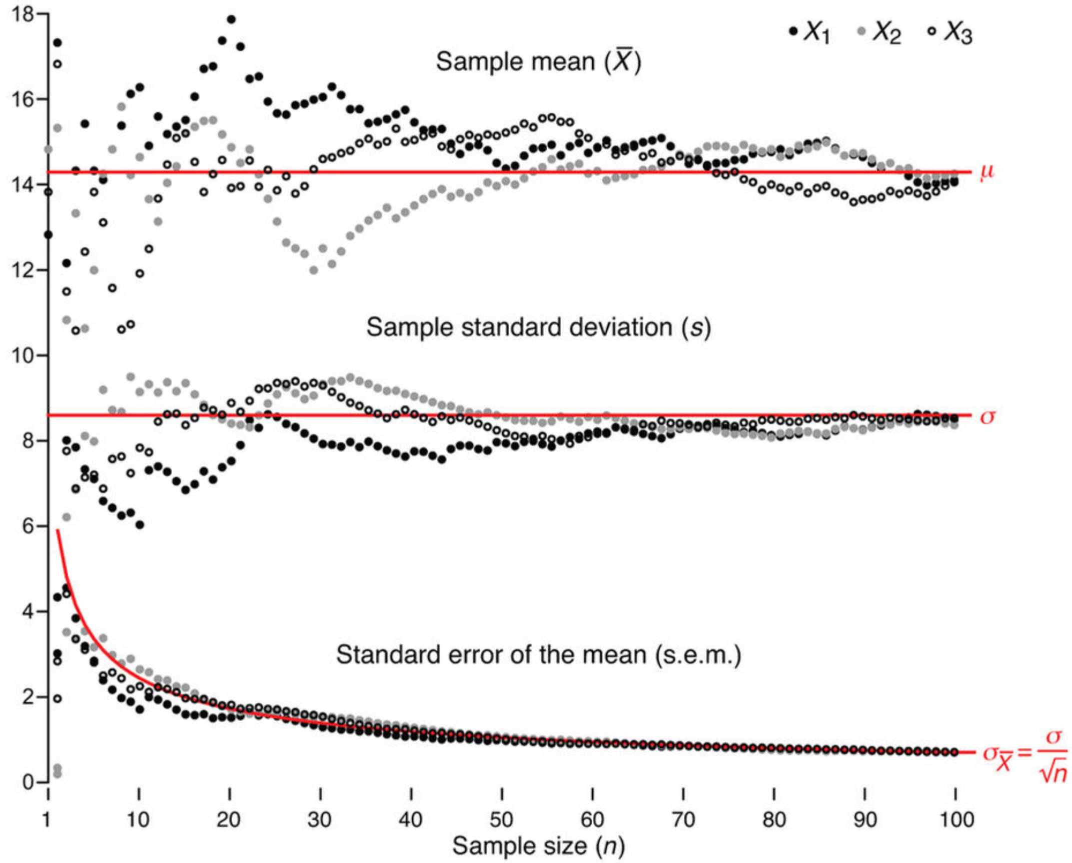
- As ***n*** increases, $\mu_{\bar{X}}$ decreases, i.e. we get **better and better estimates** of the population mean μ
- Thus big ***n*** makes $\mu \cong \mu_{\bar{X}}$
- As ***n*** increases, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

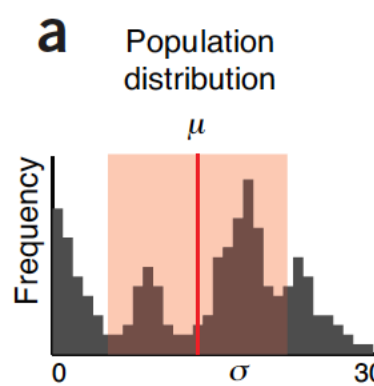
	n = 50	n = 2000	exact
mean	415.30	415.22	415.23

[Nature, 2013. Importance of being uncertain]()

Central limit theorem

Increasing sample size improves estimates





[Nature, 2013. Importance of being uncertain]

That's all for today!