University of BRISTOL

WEEK 3

Point estimation of parameters

Felipe Campelo

bristol.ac.uk

Partly adapted from Montgomery and Runger's *Applied Statistics and Probability for Engineers* and from Dr. Rihuan Ke's *SCEM lecture notes*.

In this lecture…

In this lecture, we'll cover topics related to ***point estimation of parameters*** – basically, how to estimate the parameters of a probability distribution based on data. In summary, we'll cover:

- General concepts of estimating the parameters of a population / probability distribution / data-generating process.
- Sampling distributions and some of their characteristics.
- Important properties of point estimators, including bias, variance, and mean square error
- The central limit theorem and the important role of the normal distribution as a sampling distribution.

University of BRISTOL

Part I

Point estimators and their properties

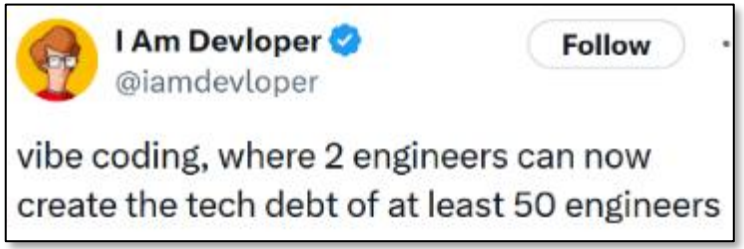
Motivation

Suppose that a data scientist is analysing the occurrence of security vulnerabilities in code generated with an AI coding assistant (such as Claude or ChatGPT).

Variability in coding security naturally occurs – depending, e.g., on the complexity of the task, availability of similar examples in the tool's training/RAG data, testing by the human developer, etc..

Imagine that this data scientist wants to estimate the expected ***proportion*** of vulnerable functions generated for a given application. In practice, they will use data from a ***sample*** of cases to compute a number that is in some sense an **estimate** of the true proportion.

This number is called a **point estimate**.



Motivation

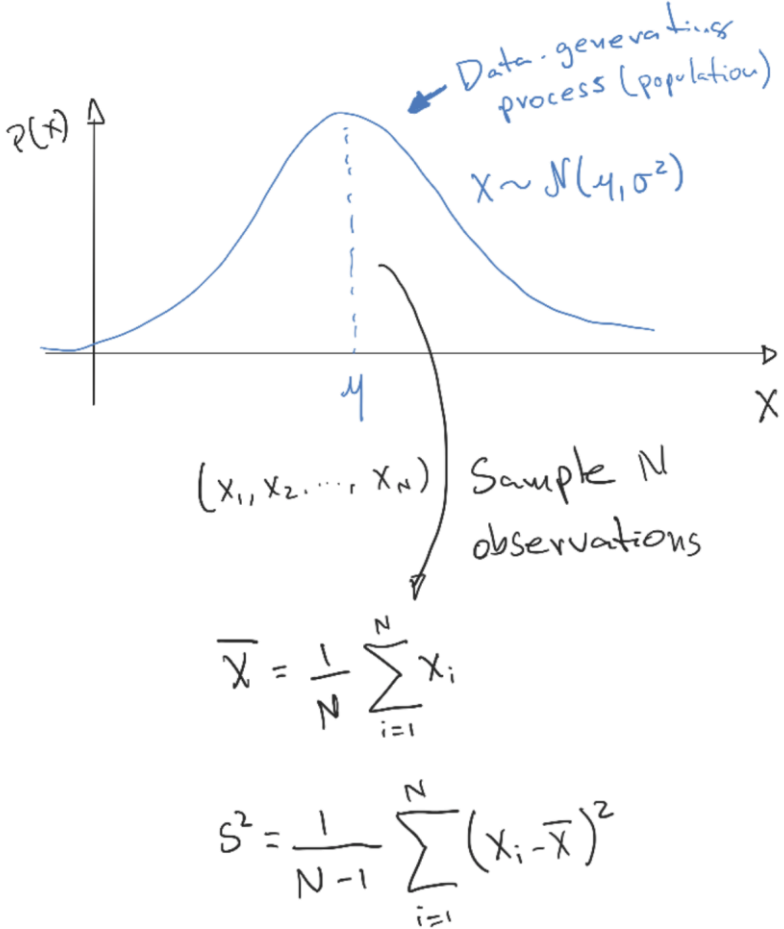
The field of *statistical inference* is concerned with **making decisions** or **drawing conclusions** about data-generating processes, commonly modelled as **(statistical) populations**.

Methods of statistical inference use information contained in a **sample** to draw conclusions about the population of interest.

Statistical inference may be divided into two major areas: **parameter estimation** and **hypothesis testing.**

This week, we'll cover some concepts about the task of estimating populational parameters from data.

Point estimation



A ***point estimator*** is a function that returns an estimated/probable value of a given population parameter.

If observations $X_1, \ldots, X_N$ are *independent and identically-distributed* (iid) *random variables*, they form a ***random sample***.

Functions of random variables are themselves random variables. Any function of a random sample (such as $\bar{X}$ or S^2) is a ***statistic.***

Statistics have their own distributions, which are called ***sampling distributions***.

Point estimators

A ***point estimator*** is a statistic which provides a plausible value (hopefully the most plausible) for a given (unknown) population parameter θ, given a sample $\boldsymbol{X}$

Consider a random variable X with a given distribution function $P(X|\theta)$, and a random sample of this variable, $\boldsymbol{X} = \{X_1, X_2, \ldots, X_N\}$.

A function $\widehat{\Theta} = f(\boldsymbol{X})$ that takes in a sample and returns a single numerical value $\hat{\theta}$ is called a ***point estimator*** of the parameter; and the value returned by this function for a given sample is called a ***point estimate*** of the parameter.

Point estimators

A ***point estimator*** is a statistic which provides a plausible value (hopefully the most plausible) for a given (unknown) population parameter θ, given a sample $\boldsymbol{X}$

Consider a random variable X with a given distribution function $P(X|\theta)$, and a random sample of this variable, $\boldsymbol{X} = \{X_1, X_2, \ldots, X_N\}$.

A function $\widehat{\Theta} = f(\boldsymbol{X})$ that takes in a sample and returns a single numerical value $\hat{\theta}$ is called a ***point estimator*** of the parameter; and the value returned by this function for a given sample is called a ***point estimate*** of the parameter.

Example

What is the *average accuracy of a machine learning model to classify email as spam or legit*?

The true population mean (μ) is the actual expected value across all possible spam/legit emails possible (this is impossible to measure directly).

Instead, we take a sample of 500 email messages of known class (spam / legit). The classifier labels them and gets 420 correct.

The **point estimator** of the mean accuracy is the **sample accuracy = 420/500 = 0.84 (84%)**, which serves as a best single-number guess for the classifier's true accuracy.

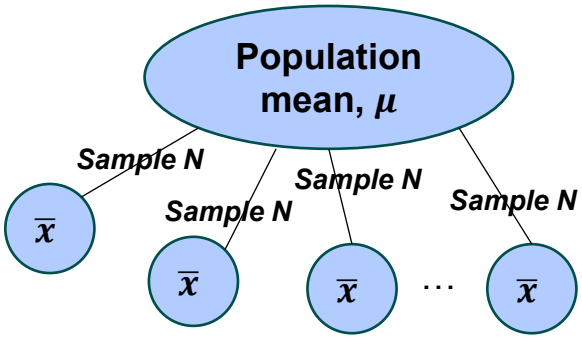
Some common parameters and their estimators

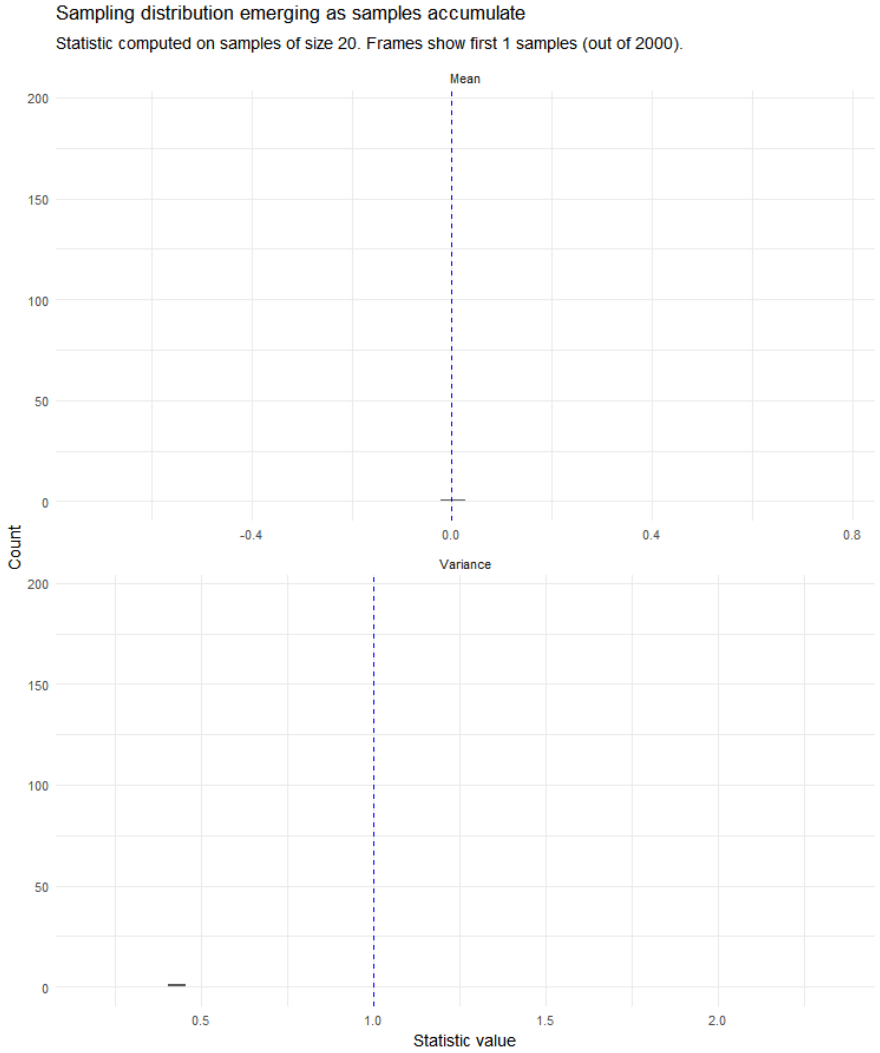
Populational parameter	Estimator
Mean of a single population, μ	Sample mean, $\hat{\mu} = \bar{x}$
Variance of a single population, σ^2	Sample variance, $\hat{\sigma}^2 = s^2$
A population proportion, p (*e.g., the proportion of correct classifications in the spam detector example*)	Sample proportion, $\hat{p} = x/n$ (x is the number of items in a random sample of size n that belong to the class of interest)
Difference of two means, $\mu_1 - \mu_2$	Difference of sample means, $\widehat{\Delta\mu} = \hat{\mu}_1 - \hat{\mu}_2 = \bar{x}_1 - \bar{x}_2$
Difference of two proportions, $p_1 - p_2$	Difference of sample proportions, $\widehat{\Delta p} = \hat{p}_1 - \hat{p}_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2}$

Note that there can be different choices of estimator for any given populational parameter. For instance, we could estimate the population mean as the ***sample mean*** (as above), or as the ***sample median*** (mid-point), as a ***winsorised mean*** (truncating extreme low/high values at a certain quantile), etc.

Sampling distributions

Given a population / data-generating process, the value of any statistic is conditional on the random sample obtained – if you take many different samples, each with N independent observations, the sample mean and of the sample variance may be different each time.





Unbiased estimators

A good estimator should consistently generate estimates that lie close to the real value of the parameter it estimates, Θ.

A given estimator $\widehat{\Theta}$ is said to be an ***unbiased estimator*** of a parameter θ if

$$\mathrm{E}\left[\widehat{\Theta}\right] = \theta \quad \text{(or, equivalently, } \mathrm{E}\left[\widehat{\Theta}\right] - \theta = 0 \text{)}$$

The quantity $\mathrm{E}\left[\widehat{\Theta}\right] - \theta$ is known as the ***bias*** of the estimator.

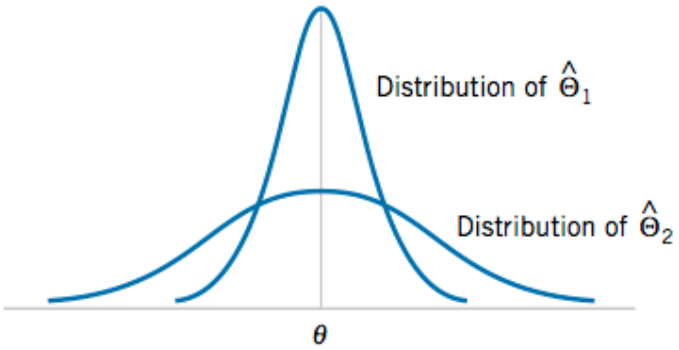
The sample mean and sample variance estimators are unbiased, i.e:

$$\mathrm{E}[\bar{X}] = \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \mu \qquad \mathrm{E}[S^2] = \mathrm{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2\right] = \sigma^2$$

Check "Why does the sample variance have N-1 in the denominator?"
https://web.ma.utexas.edu/users/mks/M358KInstr/SampleSDPf.pdf

Unbiased estimators

There are usually more than one unbiased estimator for any given parameter.



Although their bias is the same (zero), the variance of the unbiased estimators can be different. Of all unbiased estimators, the one with the lowest **variance** is known as the ***minimum variance unbiased estimator*** (MVUE)

MVUE (when available) are usually relatively good estimators, since they tend not to systematically over- or under-estimate the target parameter. (*But sometimes a slightly larger bias can be compensated by a much lower variance*).

***Image**: Montgomery & Runger's Design and Analysis of Experiments, 3rd ed., ch. 7.*

Mean squared error and bias-variance decomposition

The mean squared error of an estimator $\widehat{\Theta}$ of a given parameter θ is defined as

$$\mathrm{MSE}(\widehat{\Theta}) = \mathrm{E}\left[(\widehat{\Theta} - \theta)^2\right]$$

Which can be decomposed into bias and variance terms:

$$\mathrm{MSE}(\widehat{\Theta}) = \mathrm{E}\left[(\widehat{\Theta} - \theta)^2\right] = \mathrm{E}\left[\left((\widehat{\Theta} - \mathrm{E}[\widehat{\Theta}]) + (\mathrm{E}[\widehat{\Theta}] - \theta)\right)^2\right]$$

$$= \mathrm{E}\left[(\widehat{\Theta} - \mathrm{E}[\widehat{\Theta}])^2\right] + \mathrm{E}\left[(\mathrm{E}[\widehat{\Theta}] - \theta)^2\right] + 2\mathrm{E}\left[(\widehat{\Theta} - \mathrm{E}[\widehat{\Theta}])(\mathrm{E}[\widehat{\Theta}] - \theta)\right]$$

$$= \underbrace{\mathrm{E}\left[(\widehat{\Theta} - \mathrm{E}[\widehat{\Theta}])^2\right]}_{Variance} + \underbrace{(\mathrm{E}[\widehat{\Theta}] - \theta)^2}_{bias^2}$$

Mean squared error and bias-variance decomposition

The bias-variance decomposition of an estimator is an important aspect that will appear often in the context of statistical modelling and machine learning.

For now, it suffices to know that the MSE is often a useful quantity to compare estimators: for two estimators $\widehat{\Theta}_1$ and $\widehat{\Theta}_2$, the quantity $MSE(\widehat{\Theta}_1)/MSE(\widehat{\Theta}_2)$ is known as the **relative efficiency** of $\widehat{\Theta}_2$ in relation to $\widehat{\Theta}_1$.

If this quantity is less than 1, then $\widehat{\Theta}_1$ is generally considered a better estimator than $\widehat{\Theta}_2$.

(*Notice that the estimator with the smallest MSE may not be the MVUE*).

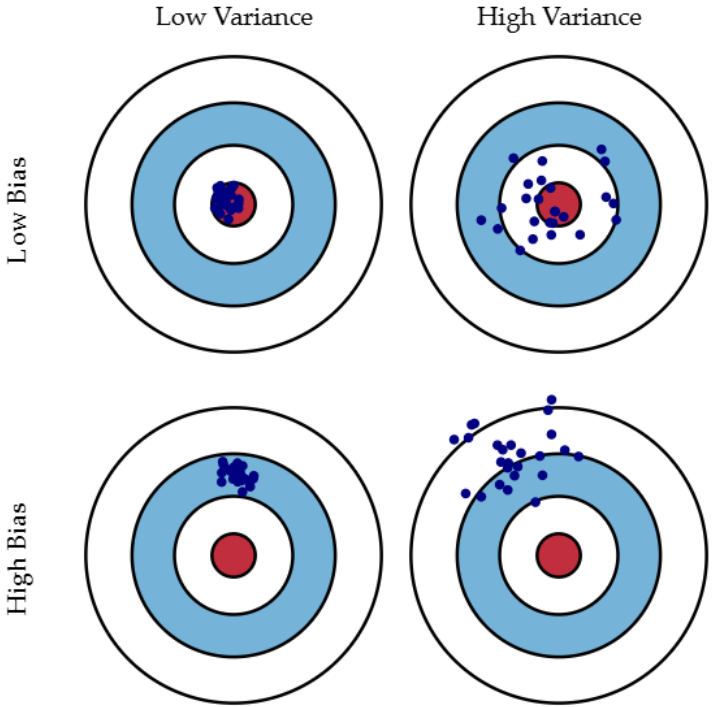


Figure: https://scott.fortmann-roe.com/docs/BiasVariance.html

Standard errors

The ***standard error*** of an estimator $\widehat{\Theta}$ is its standard deviation, given by

$$\sigma_{\widehat{\Theta}} = \sqrt{\mathrm{Var}[\widehat{\Theta}]}$$

If the standard error involves unknown parameters that can be estimated from the data, substitution of those values into $\sigma_{\widehat{\Theta}}$ produces an **estimated standard error**, denoted as $se_{\widehat{\Theta}}$ or $\hat{\sigma}_{\widehat{\Theta}}$.

In almost all applied scenarios, “standard error” commonly refers to the sample estimate.

Standard errors

The ***standard error*** of an estimator $\widehat{\Theta}$ is its standard deviation, given by

$$\sigma_{\widehat{\Theta}} = \sqrt{\mathrm{Var}[\widehat{\Theta}]}$$

If the standard error involves unknown parameters at can be estimated from the data, substitution of those values into $\sigma_{\widehat{\Theta}}$ produces an **estimated standard error**, denoted as $se_{\widehat{\Theta}}$ or $\hat{\sigma}_{\widehat{\Theta}}$.

In almost all applied scenarios, "standard error" commonly refers to the sample estimate.

Example

Suppose that we are sampling from a normal distribution with known mean μ and known variance σ^2.
In this scenario, the distribution of $\bar{X}$ is normal with mean μ and variance σ^2/N, so the standard error of $\bar{X}$ is

$$\sigma_{\bar{X}} = \sigma/\sqrt{N}$$

If the variance were unknown and we had to estimate it from the data, we could substitute s for σ in the previous equation and get the estimated standard error,

$$se_{\bar{X}} = s/\sqrt{N}$$

Standard errors

Standard errors provide an estimate of the uncertainty of a given estimate.

Smaller standard errors indicate estimates that are likely to be closer to the *expected value of the* estimator.
(If the estimator is unbiased, this means the true population parameter. If not, then the estimated value is close to the true parameter value plus bias.)

Larger standard errors indicate greater uncertainty and a wider range of plausible values for the true parameter.

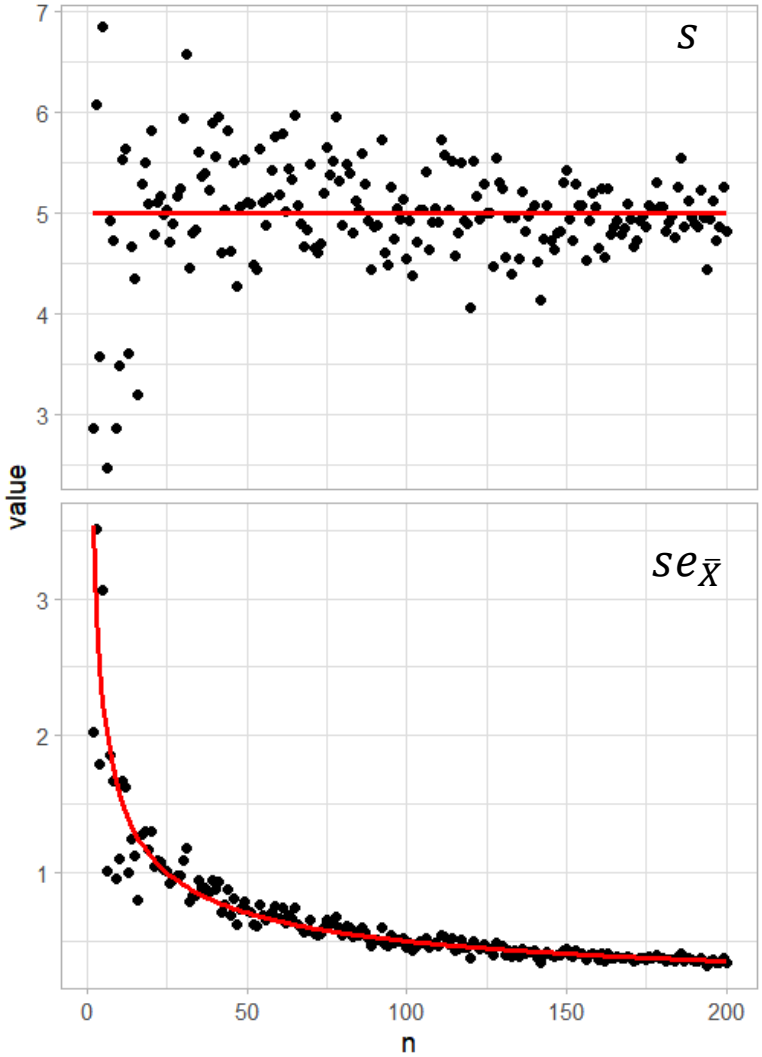
When reporting estimates, it is good practice to report them as $\hat{\theta} \pm se_{\hat{\boldsymbol{\theta}}}$ to clearly communicate the uncertainty associated with the estimated value.

Common misconceptions

"The standard error is the standard deviation of the sample", or "***A small standard error means my data points are tightly clustered."***

This confuses the standard error $se_{\hat{\Theta}}$ with the sample standard deviation s.

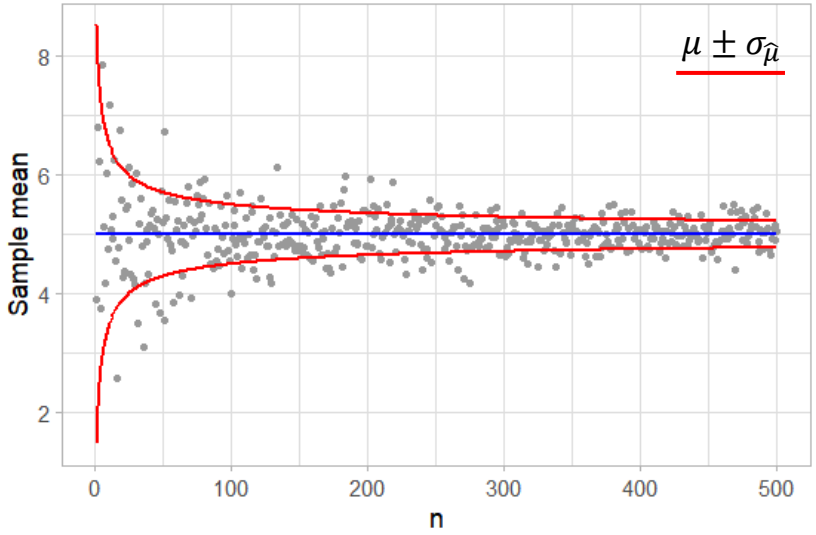
The *standard error* measures the precision of the estimate, and its value decreases with increases in the sample size. The *standard deviation* is a property of the data-generating process, and its value is independent of the sample size.



Common misconceptions

"The standard error tells us the range where most of the data lies"

$\bar{X} \pm se_{\bar{X}}$ gives a broad indication of the *uncertainty of the estimate of the mean*, not a prediction on individual data points. Ranges of the individual data points are given by characteristics of the data-generating distribution, and sometimes estimated using prediction or tolerance intervals.



"There is one universal standard error"

Standard error can refer to the standard deviation of the sampling distribution of any statistic. There are standard errors of the mean, of the median, proportions, differences between means, regression coefficients, etc.

University of BRISTOL

Part II

The Central Limit Theorem

The central limit theorem (CLT)

*Note: this formulation of the CLT is also known as the **Lindeberg-Lévy CLT**.*

Suppose a sequence of *i.i.d.* random variables, $\{\mathrm{X_i}\} = X_1, X_2, \ldots, X_n$, with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$. As the sample size n approaches infinity, the random variables defined by $\sqrt{n}(\bar{X} - \mu)$ converge in distribution to a normal variable with mean zero and variance σ^2,

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

with $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ being the sample mean.

Notice that this also means that $\bar{X} \xrightarrow{d} \mathcal{N}(\mu, \sigma^2/n)$ and, for variables with $\sigma^2 > 0$,

$$z = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

For a proof of the CLT, see https://faculty.washington.edu/fscholz/DATAFILES394_2019/CLT.pdf

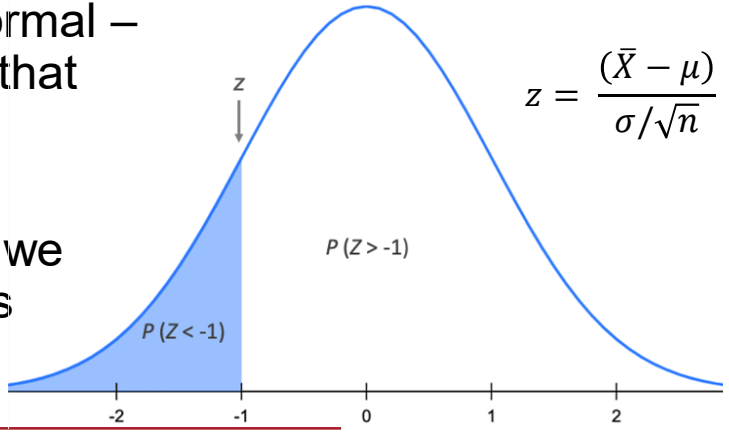
The central limit theorem (CLT)

Although it may seem like a mere statistical curiosity, the CLT has important practical implications for statistical estimation and inference.

Remember that the z statistic converges to a *standard normal distribution*, $\mathcal{N}(0,1)$, regardless* of the original probability distribution of the population.

Think about it: although there are infinitely many possible distributional shapes, there is only one standard normal – ***and its parameters are fully known***. This means that we know the exact probability of a value of z being, e.g., smaller or larger than a certain threshold.

This will be a central point (no pun intended) when we discuss confidence intervals and statistical intervals of the mean.



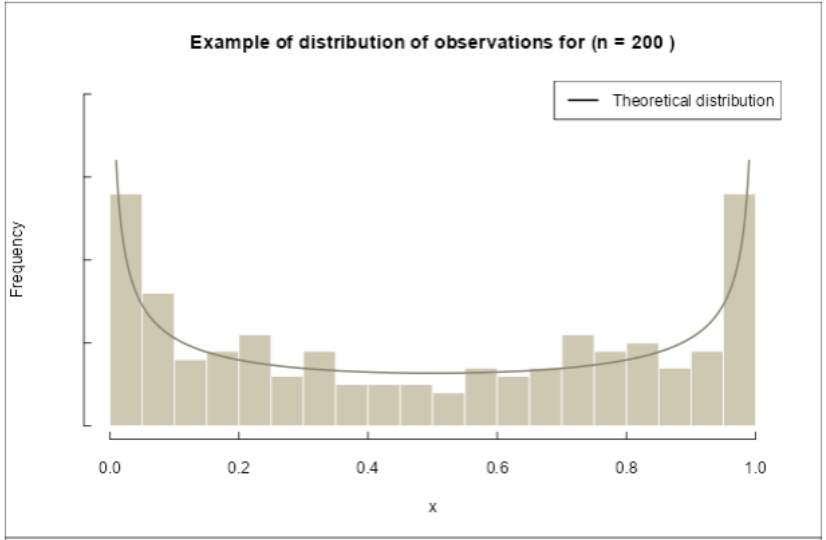
* As long as the mean and variance are defined and finite and the variance is nonzero.

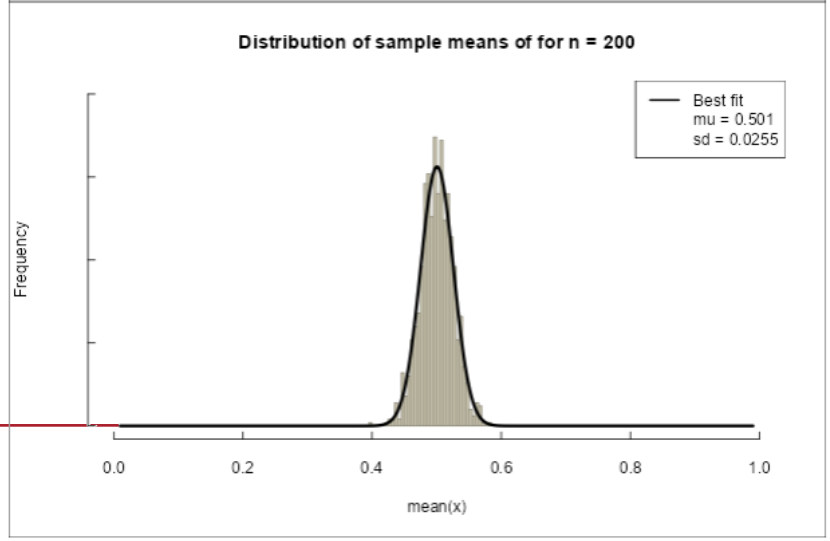
Common misconceptions

The CLT applies to the original data itself: *"If I have a large enough sample size, my data will look normally distributed*"

The CLT describes the behaviour of the **sample mean** (or sum), not of the individual data points. The underlying distribution (e.g., exponential, binomial) does not change with sampling.

(*Just think about it: the outcomes of a single roll of a fair 6-sided dice remain discrete and uniformly distributed, regardless of how many rolls are done.*)





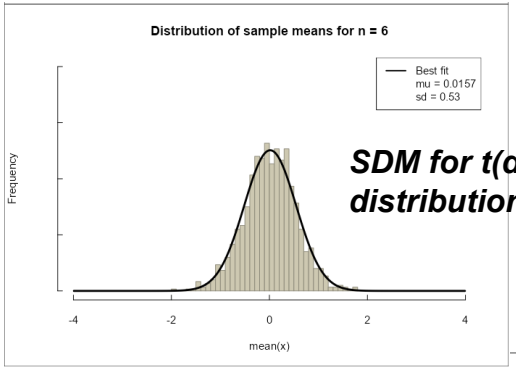
You can see the CLT by yourself running the Shiny app available in https://github.com/fcampelo/SCEM-materials/tree/main/Demo01-CLT

Common misconceptions

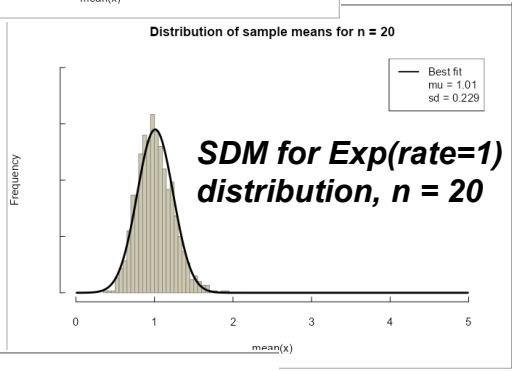
***$n = 30$ is a universal magic threshold**: "I have 30 data points, so my sample mean is certainly normal" or "I only have 6 observations, my mean is certainly not normal"*

The required sample size for the sampling distribution of the means to be approximately normal depends entirely on the **shape of the populational distribution**. For highly skewed or extremely heavy-tailed distributions, n needs to be much larger. For most symmetric distributions, a smaller n is usually sufficient.

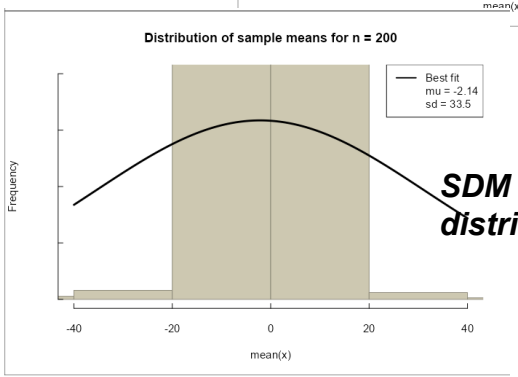
(*Just think about it: if the original distribution is already normal, the sampling distribution of the means is also normal even with $n = 1$*)



SDM for t(df=2) distribution, n = 6



SDM for Exp(rate=1) distribution, n = 20



SDM for Cauchy distribution, n = 200

Common misconceptions

The CLT explains why many datasets in the world are normally distributed:
"*Biological traits like height are normally distributed because of the CLT*."

This confuses cause and effect. The normal distribution of many natural phenomena is often the *result* of the additive combination of many small, independent factors. The CLT *describes and formalises why* such additive processes lead to a normal distribution, but it's not the cause of the phenomena themselves.

Galton box by Matemateca/USP, taken from https://en.wikipedia.org/wiki/Galton_board

Common misconceptions

The CLT applies to any statistic. *"The sampling distribution of the median, variance, or maximum will also be normal for large enough* n.“

The standard CLT specifically concerns the **sample mean** (and by extension, the sample sum). Other statistics (such as the sample maximum) have their own distinct limiting distributions.

A large sample size can fix a biased sampling method. *"As long as my sample is huge, it doesn't matter how I collected it.“*

If your sampling method is biased (e.g., voluntary response survey, faulty measurement instrument), a larger sample size will only give you a **more precise estimate of the wrong value**. Garbage in → garbage out (even in large quantities).

In this lecture we discussed…

- General concepts of estimating the parameters of a population / probability distribution / data-generating process.
- Sampling distributions and some of their characteristics.
- Important properties of point estimators, including bias, variance, and mean square error
- The central limit theorem and the important role of the normal distribution as a sampling distribution.

Further reading

DC Montgomery, GC Runger, *Applied statistics and probability for engineers, 5th ed.*
[Chapters 7.1-7.3]

- *Read the chapter and try to solve the worked examples by yourself*.