Lecture 4 – Review of Probability and Statistics

STAT/BIOF/GSAT 540: Statistical Methods for High Dimensional Biology

Sarah Merrill and Keegan Korthauer

2021/01/20

Slides by: Sarah Merrill, Sara Mostafavi, and Keegan Korthauer

Preview of next 6 lectures

- **2021/01/20 - Lecture 4: Stats Philosophy, Math/stat background & review**
- 2021/01/25 - Lecture 5: Statistical Inference - two group comparisons
- 2021/01/27 - Lecture 6: Statistical Inference - linear regression and ANOVA
- 2021/02/01 - Lecture 7: Statistical Inference - linear models (more than two groups, and interaction testing)
- 2021/02/03 - Lecture 8: Statistical Inference - continuous model + limma
- 2021/02/08 - Lecture 9: Statistical Inference - multiple testing

Outline for today

- Intro: Philosophy, goals, and central concepts
- Review: Random Variables, Probability Distributions, Sampling Distribution, Estimation, Inference, CLT, Hypothesis Testing

Your goals:

1. be familiar with the terminology
2. have a clear understanding of the concepts

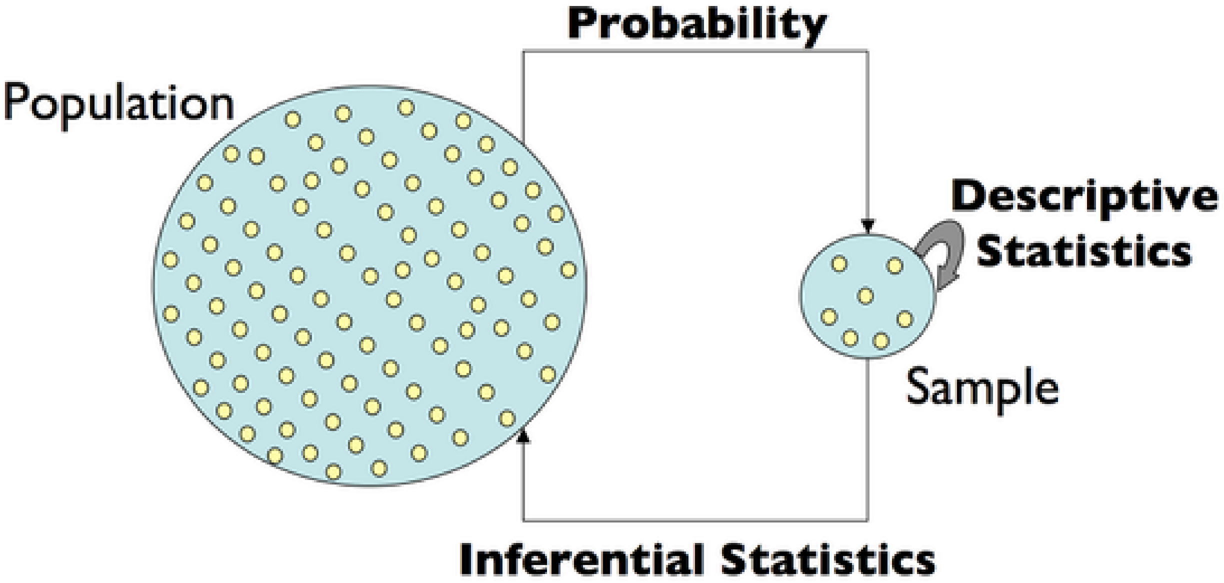
What is Statistics?

Statistics

- The field of statistics concerns the science of **collecting, analyzing/modeling, interpreting** data and **communicating uncertainty** about the results
  - Data science and machine learning have enabled application to 'big data'
- Statistical and computational methods should not be used as generic "recipes" to follow $\rightarrow$ non-robust science
- We aim for:
  - rigorous understanding to perform routine statistical analysis
  - solid foundation to follow up on specific topics

Statistical Inference

A framework for generating conclusions about a population from noisy data from a sample



- Language of **probability** enables us to make *predictions* and discuss *uncertainty*
- **Statistical inference** enables us to *understand* the data and make *conclusions*
- We need both to learn from data

Review: terminology & basic concepts

- Random variables and their distributions
- Models, parameters, and their estimators
- Central Limit Theorem (CLT)
- Hypothesis Testing

Variables

Variable *(noun)*: an element, feature, or factor that is liable to vary or change

- In statistical terminology, a **variable** is an unknown quantity that we'd like to study

- Most research questions can be formulated as

> What's the *relationship* between two or more variables?

Random variables

Random Variable (RV): A variable whose value results from the measurement of a quantity that is subject to variation (e.g. the *outcome* an experiment)

- Examples: a coin flip, a dice throw, the expression level of gene X
- An RV has a *probability distribution*

Distributions of Random Variables (RVs)

Probability: A number assigned to an outcome/event that describes the extent to which it is likely to occur

- Must satisfy certain rules (e.g. be between 0 and 1)
- Represents the (long-term) *frequency* of an event

Probability distribution: A mathematical function that maps outcomes/events to probabilities

Example experiment: Two coin tosses

- **Experiment**: Toss two coins
- **Sample space**: set of all possible outcomes $S = \{TT, HT, TH, HH\}$
- **Random Variable of interest**: number of heads

	Outcome	Number of Heads
TT		0
HT		1
TH		1
HH		2

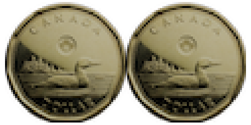
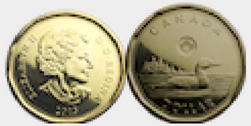
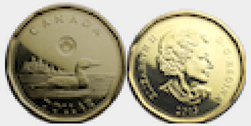
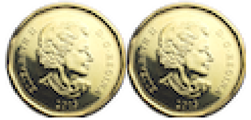
Assigning probability to outcomes

- Let:
    - $\omega =$ an outcome
    - $X(\omega) =$ number of heads in ω
- Each possible outcome is associated with a probability
- **Event:** A set of outcomes that satisfy some condition
- Each realization of the RV corresponds to an **event** (e.g. $X(\omega) = 1$ corresponds to the outcomes TH and HT)

	ω	$X(\omega)$	**Probability**
TT		0	0.25
HT		1	0.25
TH		1	0.25
HH		2	0.25

Assigning probability to events

The probability distribution of the Random Variable X tells us how likely each event (number of heads) is to occur in the experiment

Event	x	$P(X = x)$
	0	0.25
 , 	1	0.50
	2	0.25

Note on notation: $P(X = x)$ can also be written as $P_X(x)$

Two types of random variables

- A **discrete** RV has a countable number of possible values
    - e.g. throwing dice, genotype measured on a SNP chip
- A **continuous** RV takes on values in an interval of numbers
    - e.g. expression level of a gene, blood glucose level, height of individuals

Discrete or Continuous?

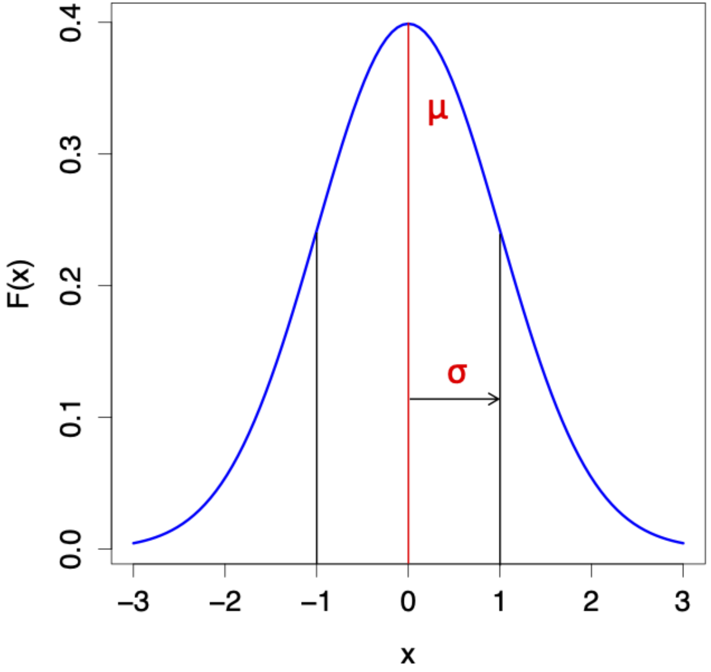
- Select a **clap** reaction if you think the example is **discrete** 👏
- Select a **thumbs up** reaction if you think the example is **continuous** 👍

Standard Gaussian (Normal) distribution

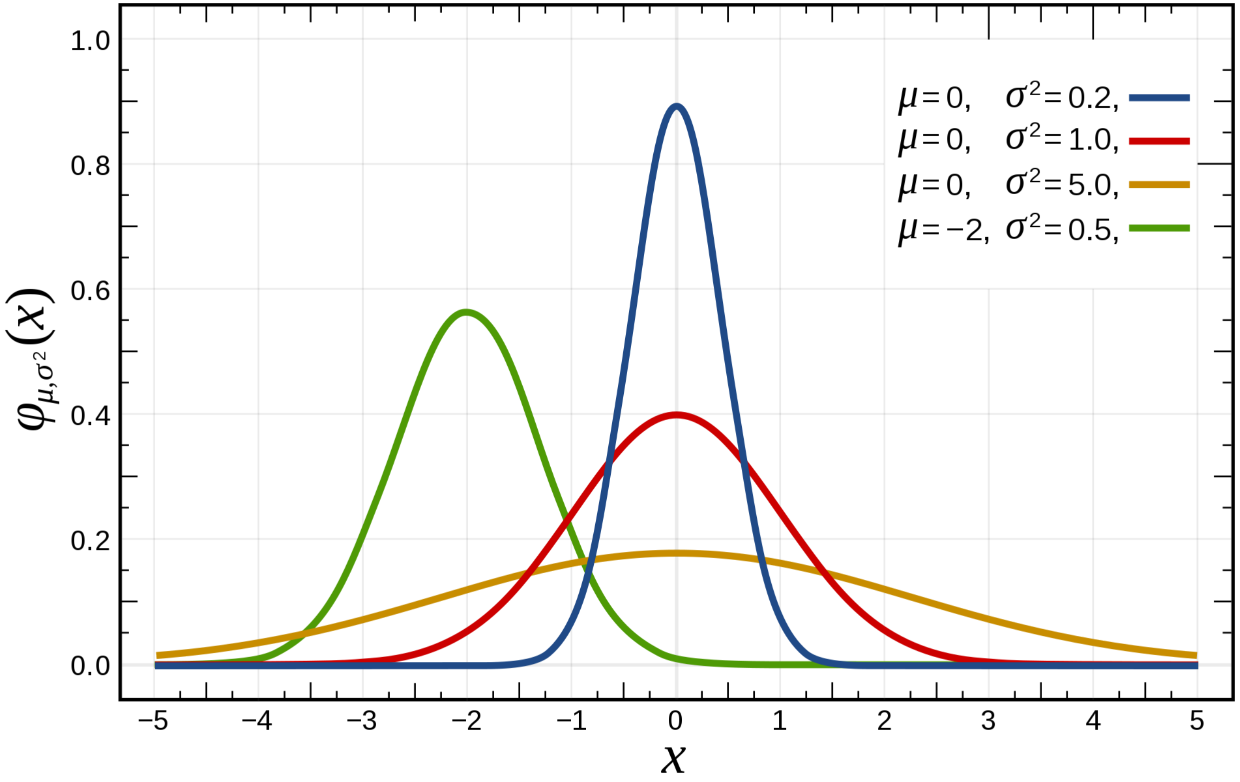
- probability density function:

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Mean $= \mu$
- Standard Deviation $= \sigma$
- For convenience, we write $N(\mu, \sigma^2)$
- When $\mu = 0$ and $\sigma = 1$, this is the *Standard* Normal distribution $N(0, 1)$

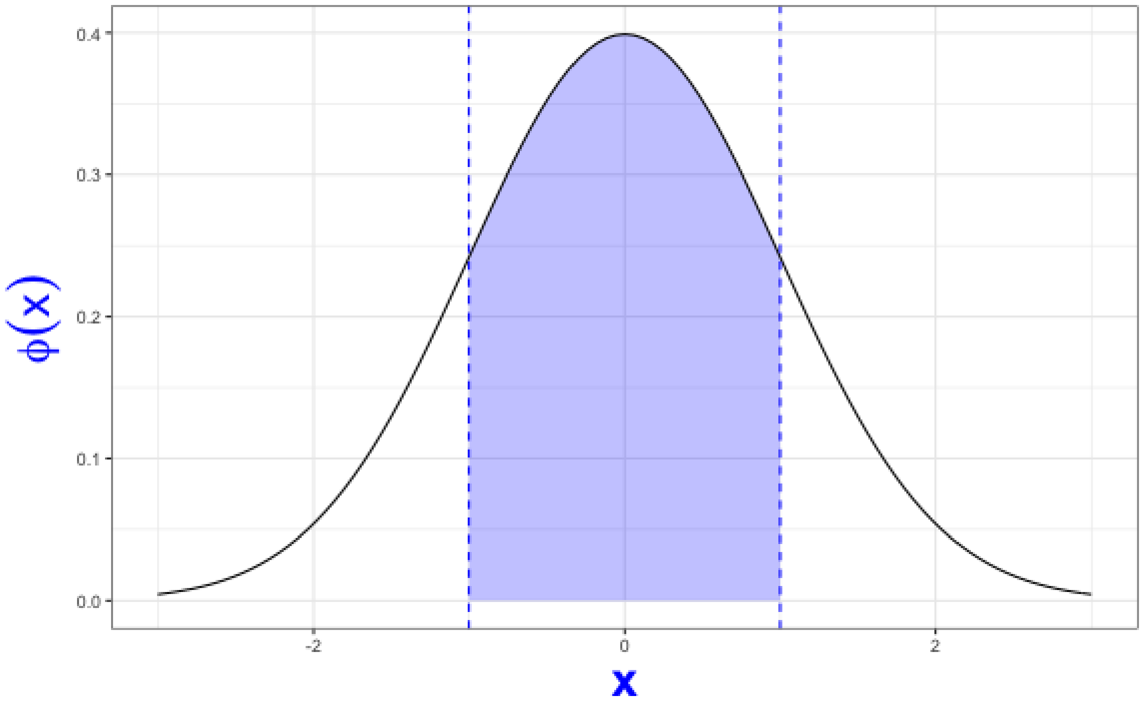


Gaussian (Normal) distribution



$$\text{Probability density function: } f(x|\mu,\sigma^2) = \phi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Density → probability requires integration



Empirical Rule for Normal Distributions

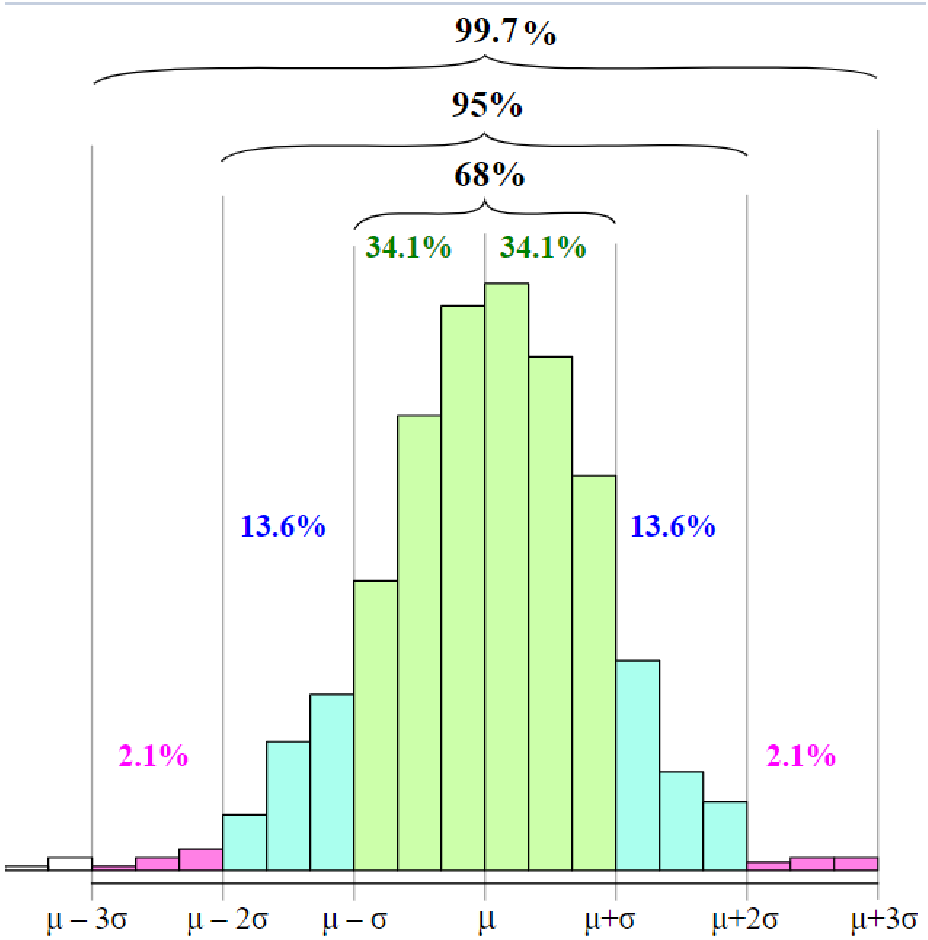
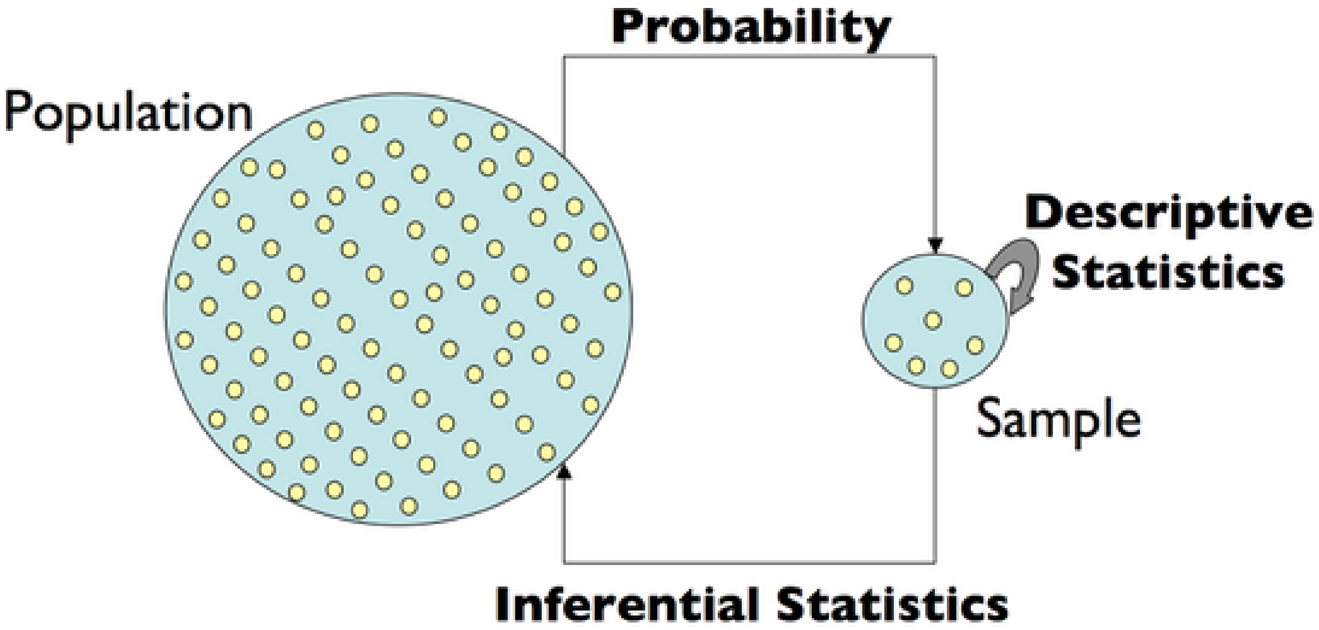


Image source: By Melikamp - Own work, CC BY-SA 4.0

Statistical Inference

- The **parameter space** is the set of all possible values for the parameter
- One major goal: to "figure out" (i.e. estimate) the **parameter values**
  - i.e. *"fit the model to the data"*
- The model is a representation that (we hope) approximates the data and (more importantly) the population that the data were sampled from
- We can then use this model for:
  - hypothesis testing
  - prediction
  - simulation

Statistical Inference



IID

- A requirement (assumption) in many settings is that the data are IID: **I**ndependent and **I**dentically **D**istributed
- **Identically Distributed**: a set of observations (events) are from the same population (i.e. they have the same underlying probability distribution)
  - e.g. a t-test assumes that under the null, all observations come from the same normal distribution
- **Independent**: all samples satisfy the condition

$$P(A, B) = P(A)P(B)$$

  where A and B are events

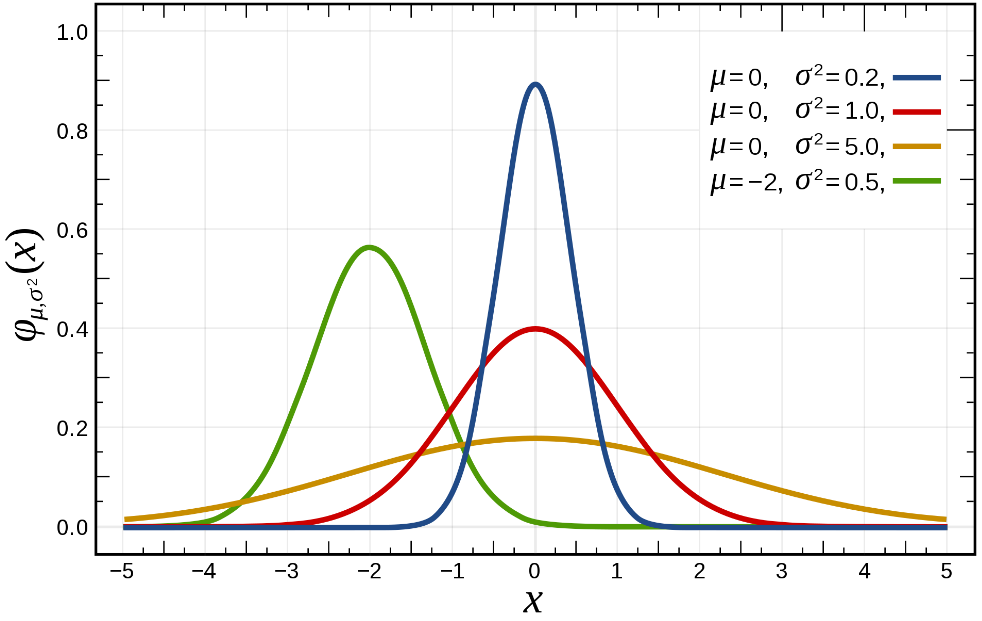
  - i.e. the joint probability is the product of the individual event probabilities
  - The above statement is for two events, but the same definition applies for any number of events (without loss of generality for any number of events)

Violations of independence

- Experimental design is in part about trying to avoid unwanted dependence
- Example of design with violation of independence assumption:

> Height measurements of individuals sampled from *related* females in a tall family

Parameters of the normal distribution



$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Mean $= \mu$
- Standard Deviation $= \sigma$
- For convenience, we write $N(\mu, \sigma^2)$
- Population parameters are unknown

Parameter estimation

- **Estimator**: A function (or rule) used to estimate a parameter of interest

- **Estimate**: A particular realization (value) of an estimator

Estimators for normally distributed data

- If we are given a sample of n observations from a normally distributed population, how do we estimate the parameter values μ and σ?

- Recall μ is the mean and σ the standard deviation of the distribution

$$\hat{\mu} = \bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma} = s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Estimators vs Parameters

	Estimators	**Parameters**
Summarize	Sample	Population (ground truth)
Value	Computed from data	Unknown
Notation	$\hat{\theta}$	θ

Normal **Mean**: Estimator vs Parameter

	Estimator	**Parameter**
Summarizes	Sample/data	Population (ground truth)
Value	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	Unknown
Notation	$\hat{\mu}$	μ

Normal **Standard Deviation**: Estimator vs Parameter

	Estimator	**Parameter**
Summarizes	Sample/data	Population (ground truth)
Value	$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$	Unknown
Notation	$\hat{\sigma}$	σ

Estimator for normally distributed data

- Let's say we collected a **sample** from a population we assume to be normal
- We estimate the mean $\hat{\mu} = \bar{x}$
- How good is the estimate?
- The answer depends on:
  - sample size
  - variability of the population

Sampling distribution

- **Statistic**: any quantity computed from values in a sample
- Any function (or statistic) of a sample (data) is a random variable
- Thus, any statistic (because it is random) has a probability distribution function $\rightarrow$ specifically, we call this the *sampling distribution*
- Example: the sampling distribution of the mean

Sampling distribution of the mean

The sample mean $\bar{x}$ is an RV, so it has associated probability or sampling distribution

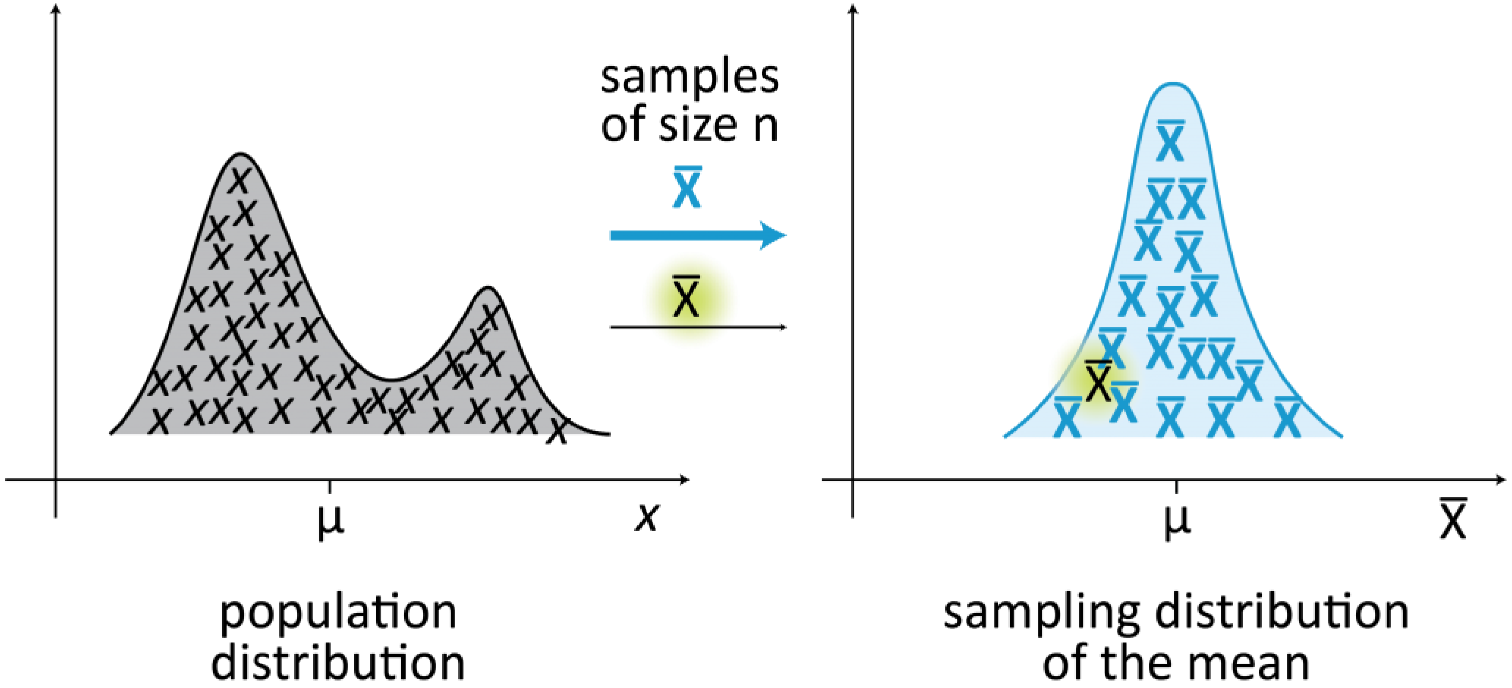


Image source: incertitudes.fr/book.pdf

Central Limit Theorem (CLT)

By the *Central Limit Theorem (CLT)*, we know that the sampling distribution of the mean is Normal:

- with mean $\mu_{\bar{X}} = \mu$ and standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

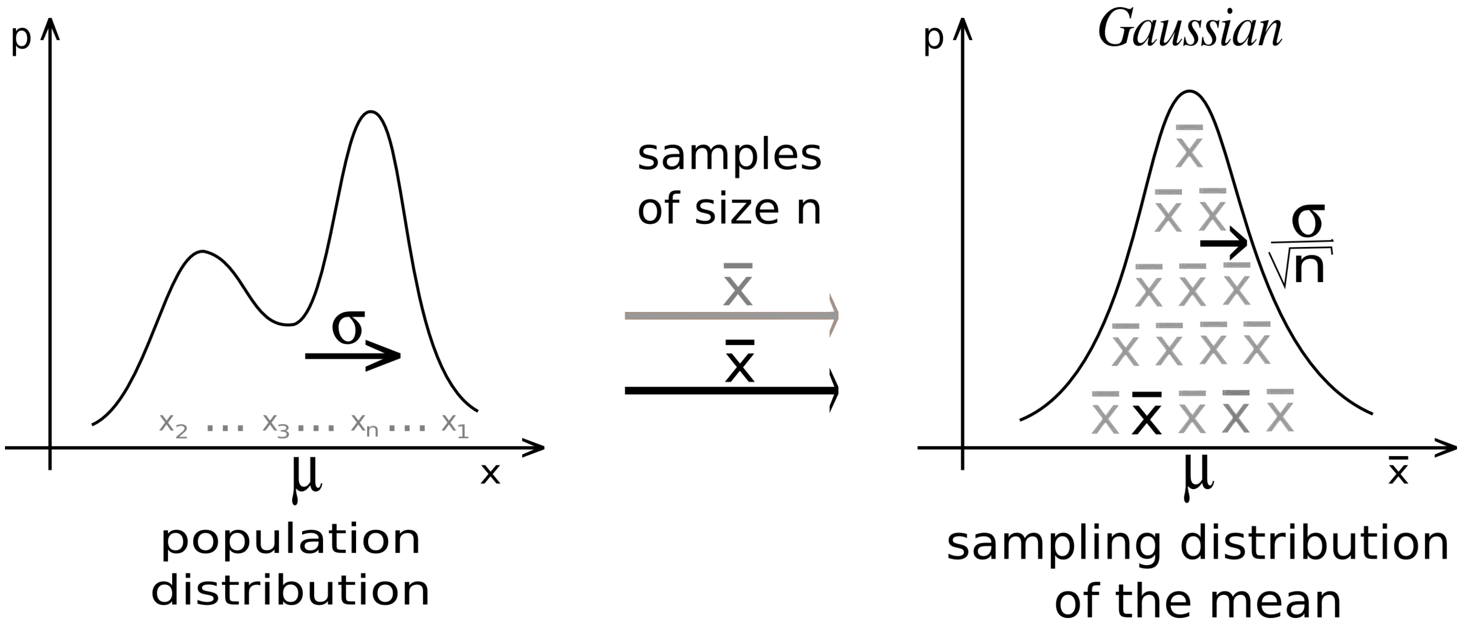


Image source: incertitudes.fr/book.pdf

⚠️ Standard deviation vs Standard error ⚠️

- The sampling distribution of the mean (by CLT):

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

- The *standard error* of the mean is $\frac{\sigma}{\sqrt{n}}$

- The *standard deviation* of X is σ

Estimation of parameters of the sampling distribution of the mean

Just as we estimated μ and σ before, we can estimate $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$

- $\hat{\mu}_{\bar{X}} = \hat{\mu} = \bar{x}$
- $\hat{\sigma}_{\bar{X}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n}}$

Standard error of the mean

$$\hat{\sigma}_{\bar{X}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n}}$$

- The standard error (SE) of the mean reflects uncertainty about the value of the population mean μ
- The CLT assumes a 'large enough' sample:
  - when the sample size is ~30 or more, the normal distribution is a good approximation for the sampling distribution of the mean
  - for smaller samples, the SE $\frac{s}{\sqrt{n}}$ is an underestimate

CLT applies to any population (regardless of distribution)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with a non-normal distribution. If the sample size n is sufficiently large, then the sampling distribution of the mean will be approximately normal: $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

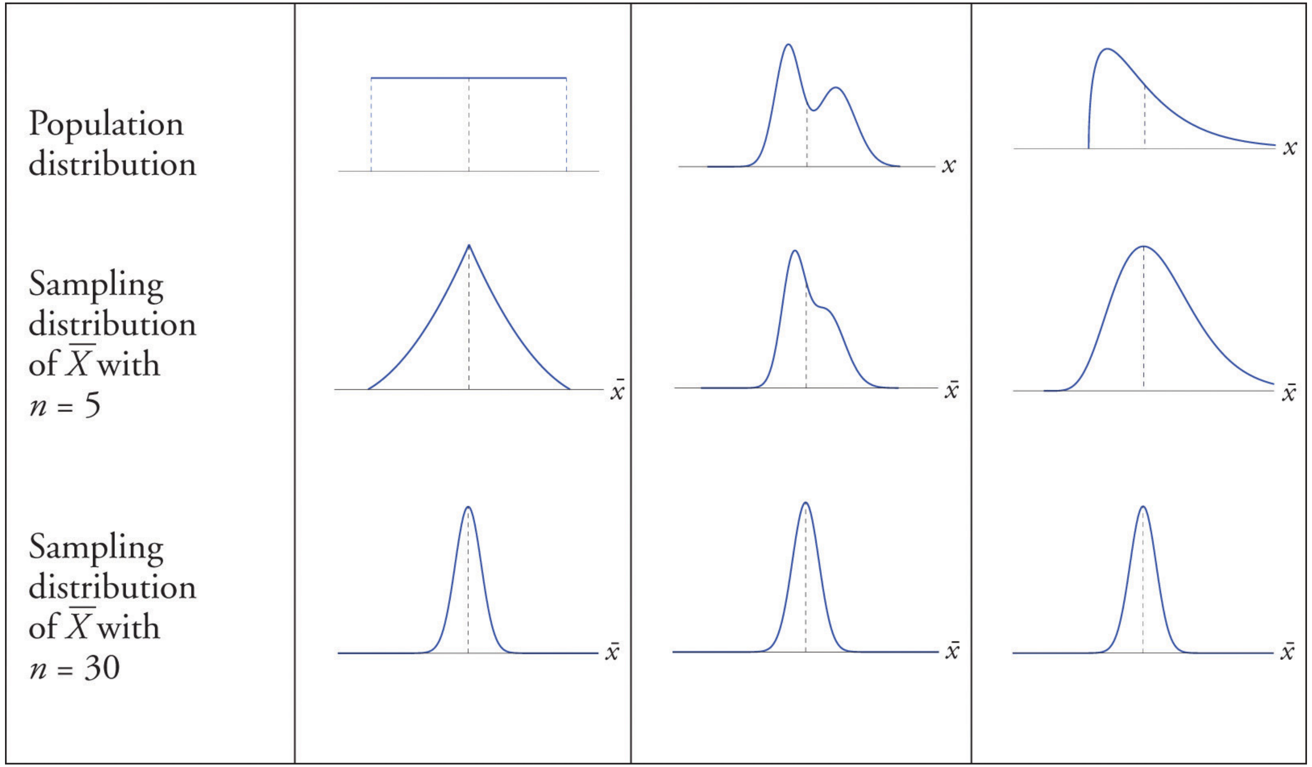
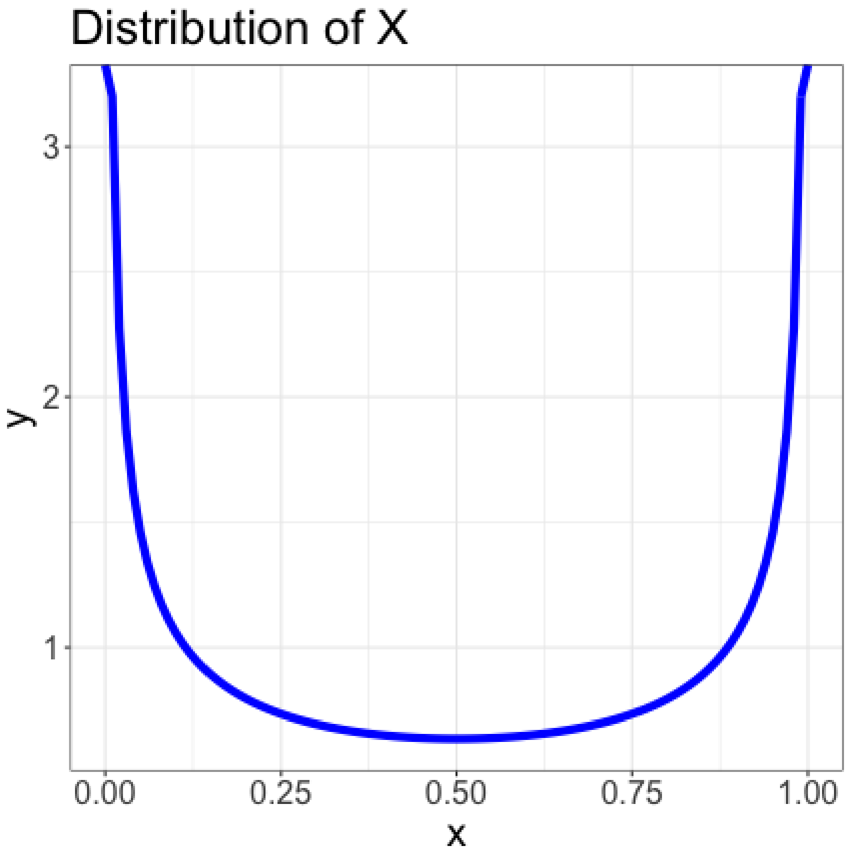
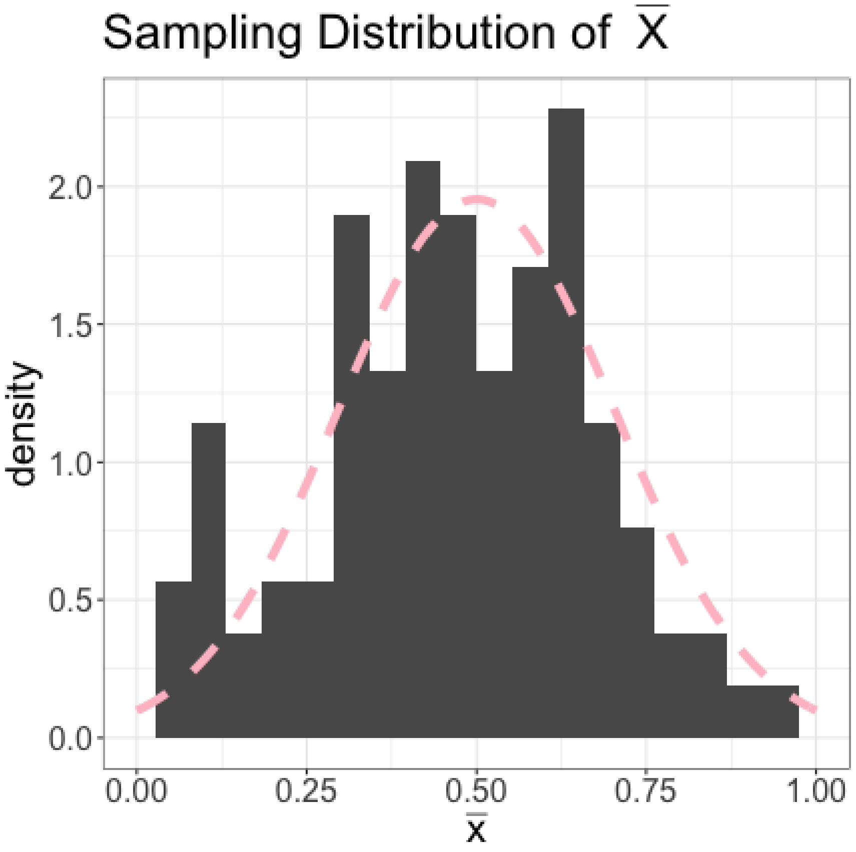


Image source: saylordotorg.github.io/text_introductory-statistics

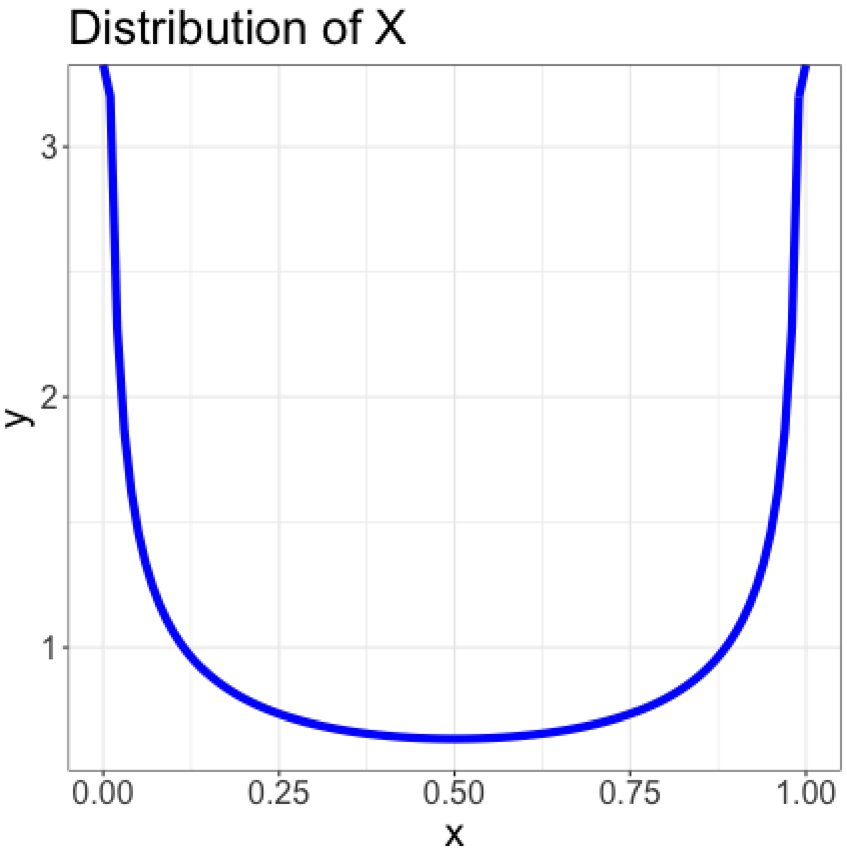
Illustration (n = 3)

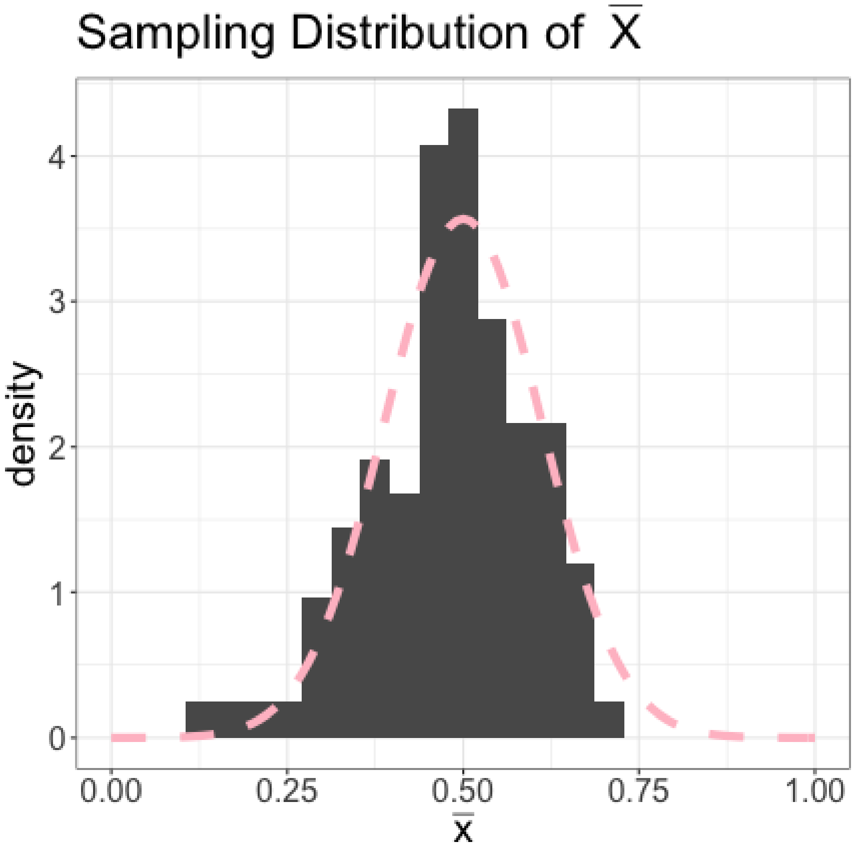




On right: dashed pink line is $N(\mu, \sigma^2/n)$

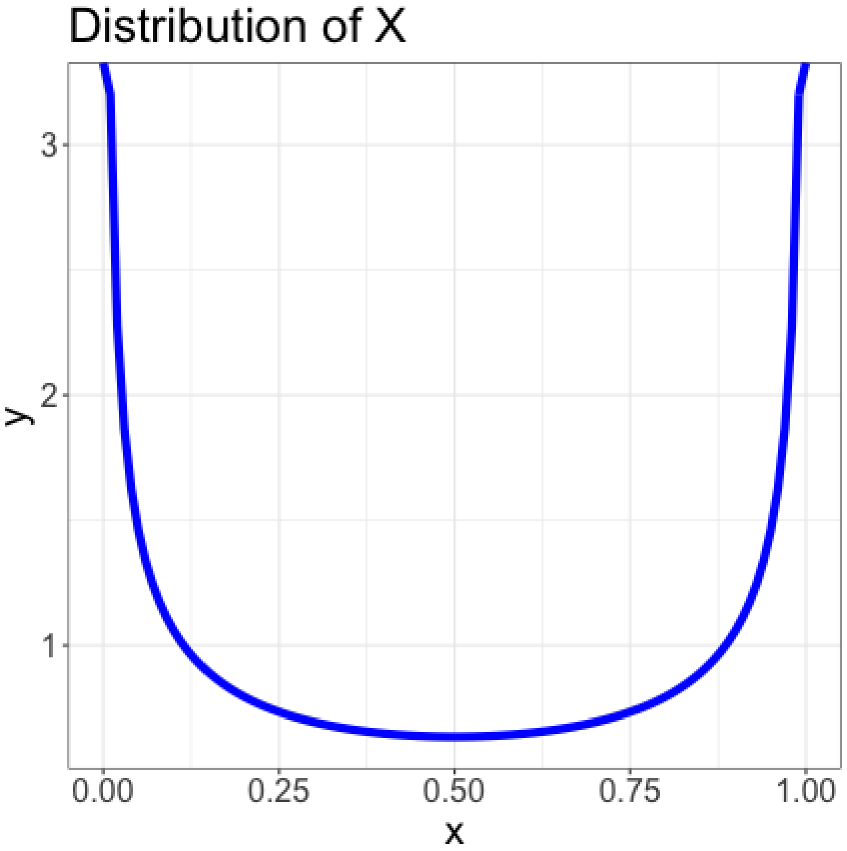
Illustration (n = 10)

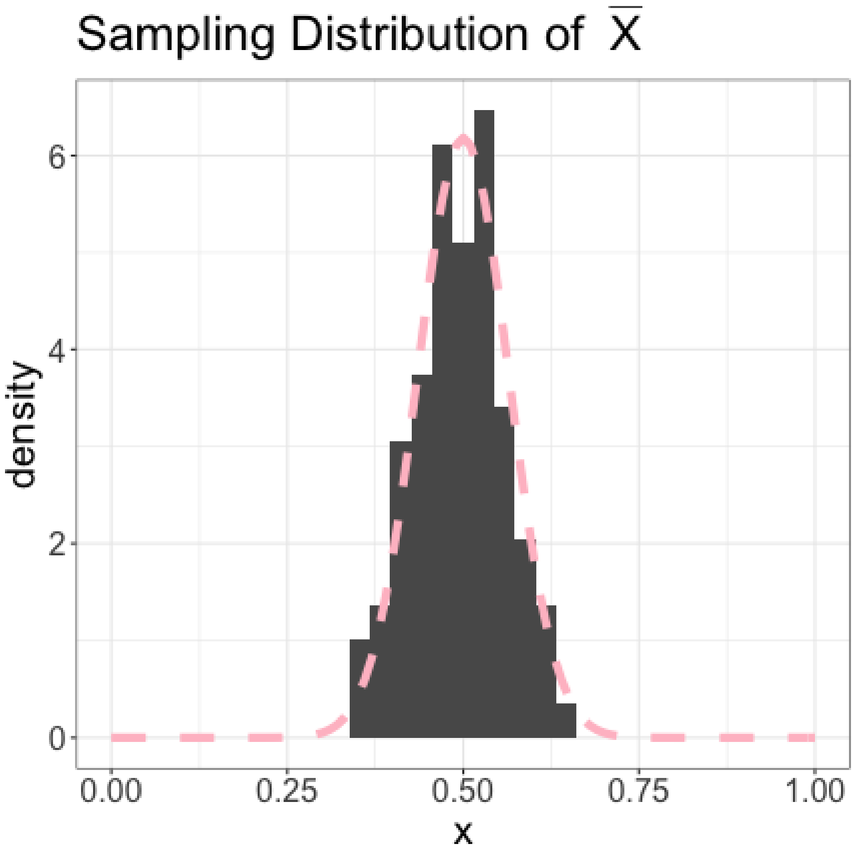




On right: dashed pink line is $N(\mu, \sigma^2/n)$

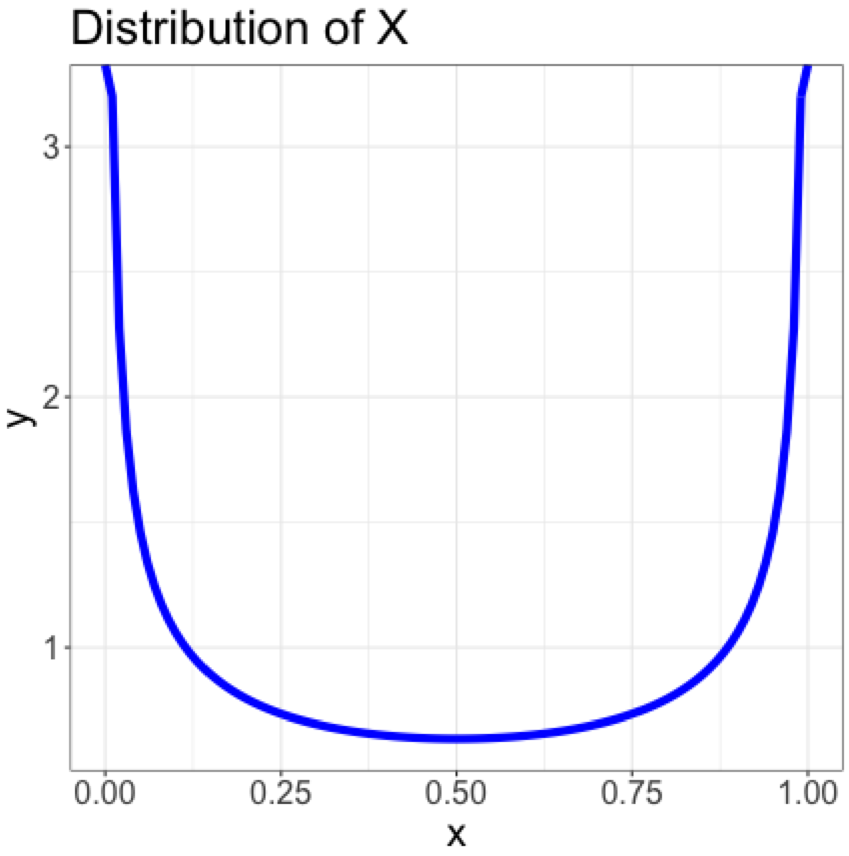
Illustration (n = 30)

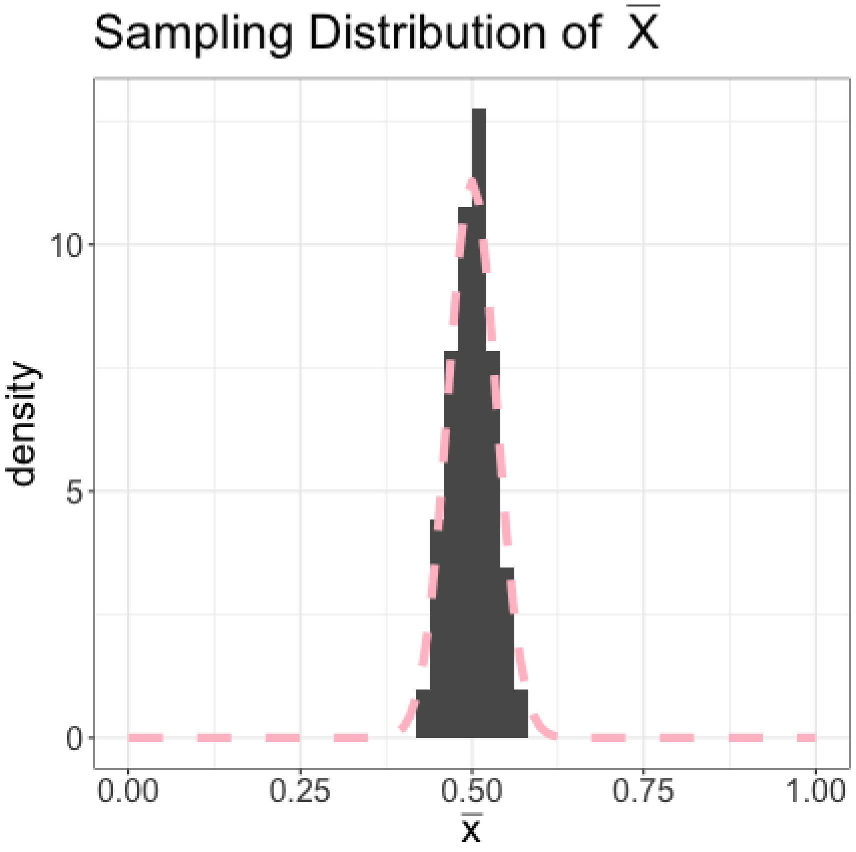




On right: dashed pink line is $N(\mu, \sigma^2/n)$

Illustration (n = 100)


Distribution of X


Sampling Distribution of $\bar{X}$

On right: dashed pink line is $N(\mu, \sigma^2/n)$

Hypothesis Testing

- **Hypothesis:** A *testable (falsifiable)* idea for explaining a phenomenon
- **Statistical hypothesis:** A hypothesis that is testable on the basis of observing a process that is modeled via a set of random variables
- **Hypothesis Testing:** A formal procedure for determining whether to *accept* or *reject* a statistical hypothesis
- Requires comparing two hypotheses:
  - H_0: null hypothesis
  - H_A or H_1: alternative hypothesis

Hypothesis Testing: Motivating Example

- The expression level of gene g is measured in n patients with disease (e.g. cancer), and m healthy (control) individuals:
  - $z_1, z_2, \ldots, z_n$ and $y_1, y_2, \ldots, y_m$
- Is gene g differentially expressed in cancer vs healthy samples?
  - $H_0 : \mu_Z = \mu_Y$
  - $H_A : \mu_Z \neq \mu_Y$
- In this setting, hypothesis testing allows us to determine whether observed differences between groups in our data are *significant*

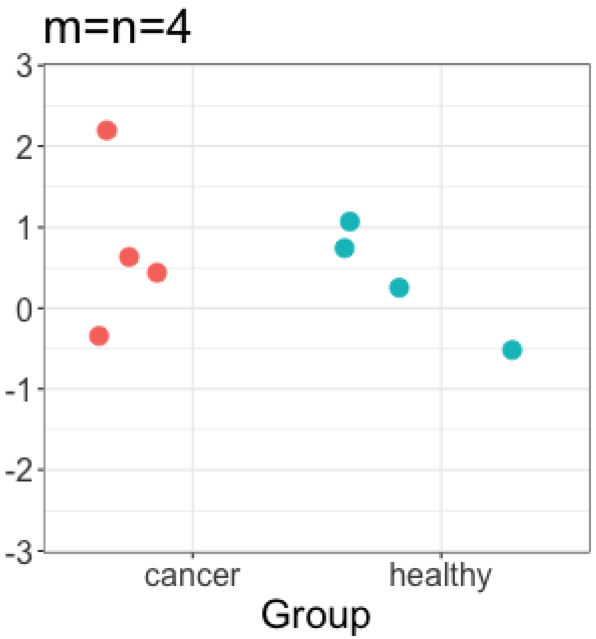
Steps in Hypothesis Testing

1. Formulate your hypothesis as a statistical hypothesis

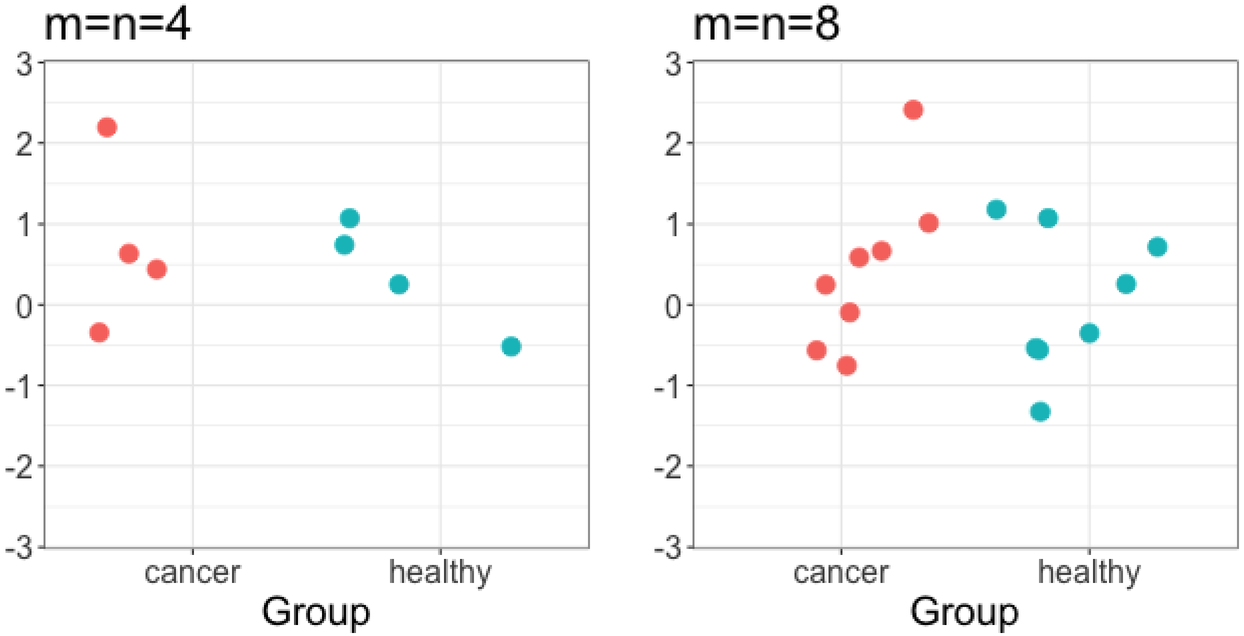
2. Define a test statistic (RV) that corresponds to the question. You typically know the expected distribution of the test statistic *under the null*

3. Compute the p-value associated with the observed test statistic under the null distribution $p(t|H_0)$

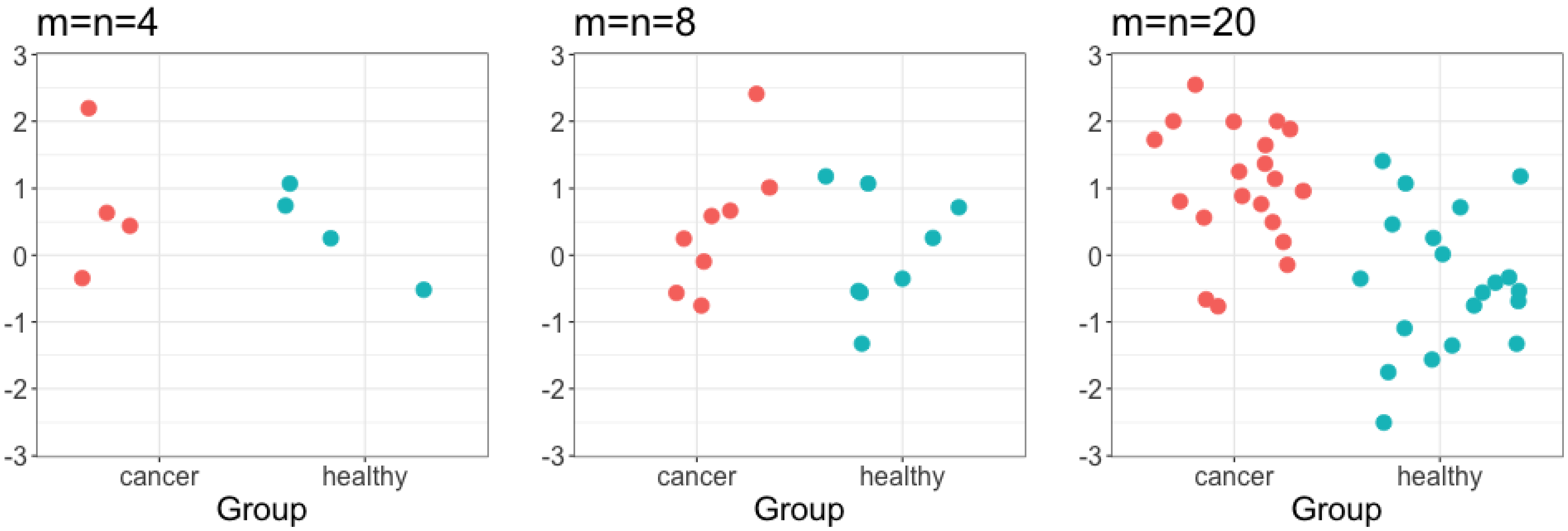
Motivating example (cancer vs healthy gene expression)



Motivating example (cancer vs healthy gene expression)

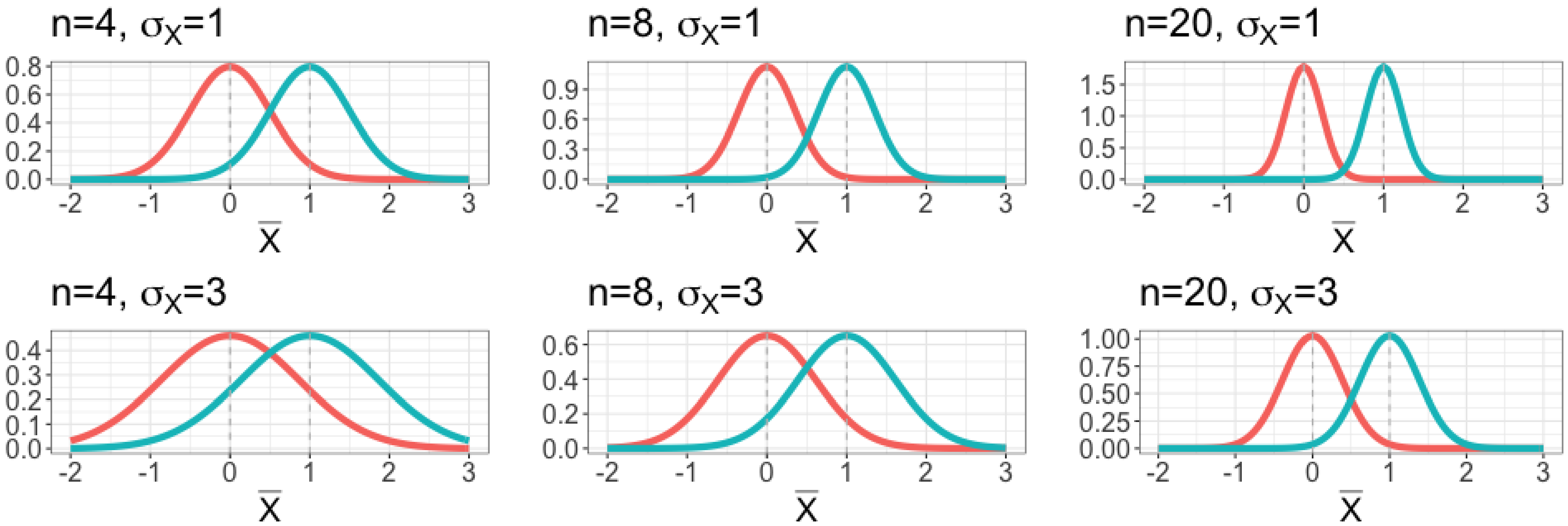


Motivating example (cancer vs healthy gene expression)



All three samples drawn from iid Normal distributions with equal variance and $\mu_Z - \mu_Y = 1$

Is there a **significant** difference between the two means?



- Mean difference needs to be put into context of the *spread (standard deviation)*

- Also depends on the sample size

t-statistic / t-test

- Measures difference in means, adjusted for spread/standard deviation:

$$t = \frac{\bar{z} - \bar{y}}{SE_{\bar{z}-\bar{y}}}$$

for $z_1, z_2, \ldots, z_n$ expression measurements in healthy samples and $y_1, y_2, \ldots, y_m$ cancer samples

- Standard error estimate for the difference in means:

$$SE_{\bar{z}-\bar{y}} = s_p \sqrt{\frac{1}{n} + \frac{1}{m}} \text{, where } s_p^2 = \frac{s_z^2 + s_y^2}{(n-1) + (m-1)}$$

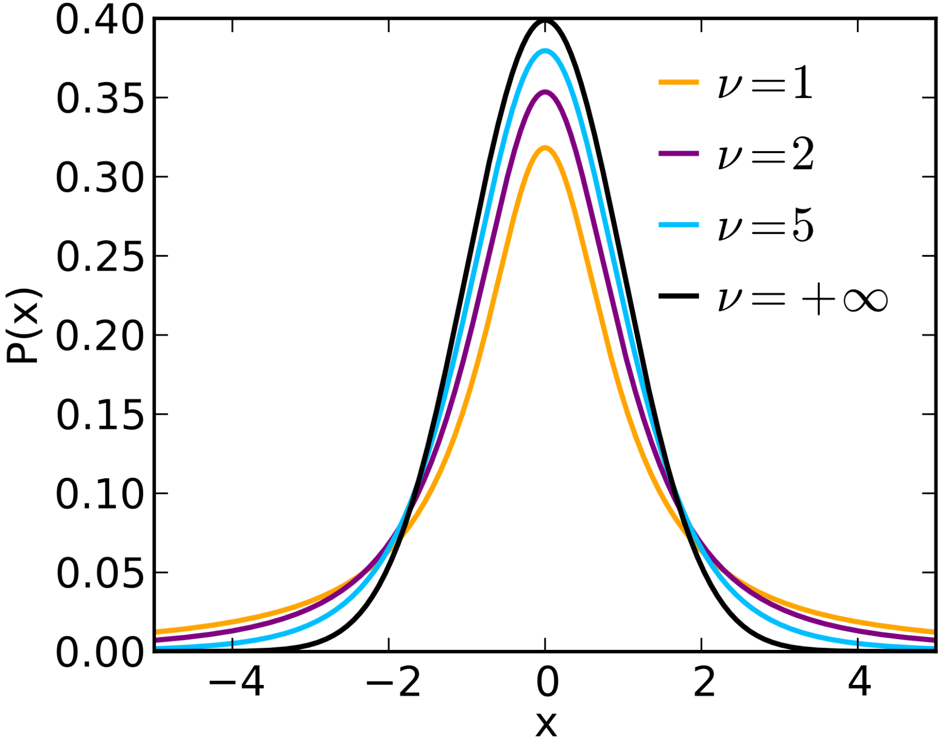
t-test

- From the theory, we know the distribution of our test statistic, if we are willing to make some assumptions
- If we assume:
    - Z and Y are normally distributed
    - Z and Y have equal variance

Then our t-statistic follows a t distribution with m+n-2 degrees of freedom

$$t \sim t_{n+m-2}$$

t distribution



- statistic value tells us how extreme our observed data is relative to the null

- obtain p-value by computing area to the left and/or right of the t statistic (one-sided vs two-sided)

Summary

- Random variables are variables that have an associated probability distribution
- Any statistic of sampled data is an RV, and hence has an associated probability distribution
- The CLT gives us the sampling distribution of the mean
- Hypothesis testing gives us a framework to assess a statistical hypothesis under the null