Statistical Methods for High Dimensional Biology

Lecture 4 – Review of Probability and Statistics

January 16/2019

Sara Mostafavi

Lectures

Class meetings and schedule

Time : Mon Wed 9:30 - 11am

Location : ESB 2012

date	notes	instructor
Jan-07 Mon	lecture-1: Course intro	PP
Jan-09 Wed	lecture-2: Molecular Biology methods intro: RNA, DNA, methylation, ChiP-Seq	PP
Jan-14 Mon	lecture-3: Exploratory data analysis, QC	PP
Jan-16 Wed	lecture-4: Stats Philosophy & Math/stat background	SM
Jan-21 Mon	lecture-5: Statistical Inference - two groups	GFC
Jan-23 Wed	lecture-6: Statistical Inference - linear regression and ANOVA	GFC
Jan-28 Mon	lecture-7: Statistical inference - linear models (more than two groups, and interaction testing)	GFC
Jan-30 Wed	lecture-8: Statistical inference - multiple testing & non-parametric	GFC
Feb-04 Mon	lecture-9: Batch effects and resampling,	SM
Feb-06 Wed	lecture-10: Application of statistical inference to RNA-seq I	PP
Feb-11 Mon	lecture-11: Application of statistical inference to RNA-seq II	PP
Feb-13 Wed	lecture-12: Machine Learning Intro: Unsupervised learning PCA	SM
Feb-18 Mon	Midterm Break	NA
Feb-20 Wed	Midterm Break	NA
Feb-25 Mon	lecture-13: Unsupervised learning Clustering	GFC
Feb-27 Wed	lecture-14: Supervised learning I	GFC
Mar-04 Mon	lecture-15: Supervised learning II	GFC
Mar-06 Wed	lecture-16: Guest lecture	
Mar-11 Mon	lecture-17: GWAS	SM
Mar-13 Wed	lecture-18: xQTL analysis	SM
Mar-18 Mon	lecture-19: Cellular heterogeneity	SM
Mar-20 Wed	lecture-20: Gene set analysis	PP
Mar-25 Mon	lecture-21: Gene networks and function prediction	PP
Mar-27 Wed	lecture-22: Guest lecture: Andrew Roth	
Apr-01 Mon	lecture-23: Oral presentations	Oral
Apr-03 Wed	lecture-24: Oral presentations	Oral

Announcements

- GitHub account + email from TAs
- Keep on submitting seminar deliverables!
- Audit students: please go to second half of seminar.
- Project groups.

Outline

- Intro: Philosophy, goals, and central concepts
- Review: RVs, Distributions, Sampling Distribution, CLT, Hypothesis Testing

Your goals:

- Make sure a) you know the terminology (can't advance past this one if you don't know the language), b) you are not confused by any of the concepts

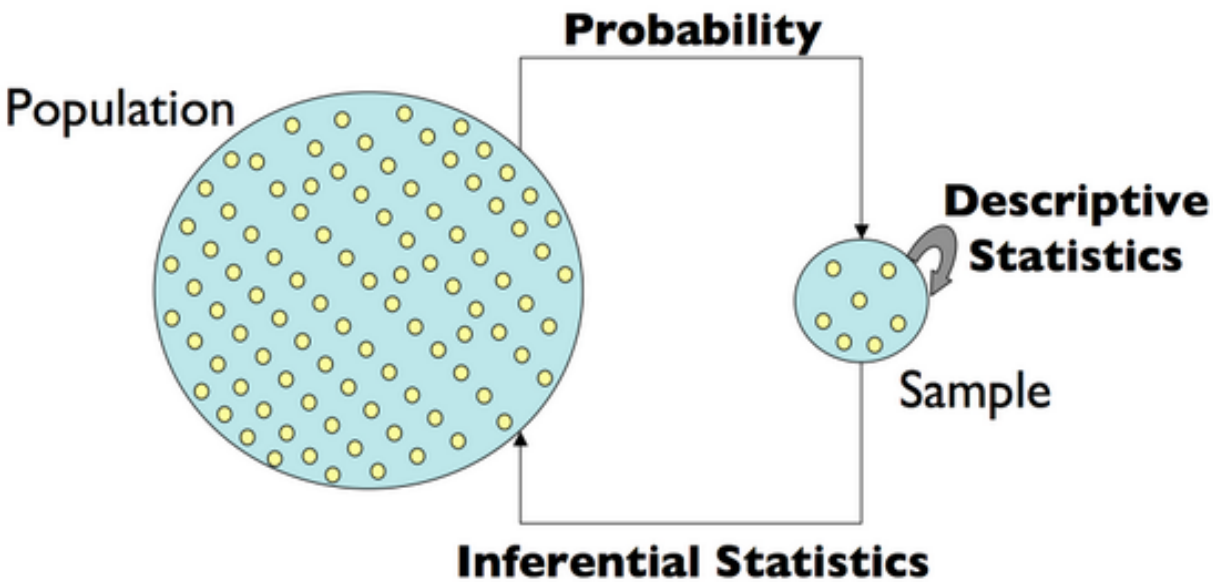
What is Statistics?

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 "on mass" application.

- Statistical and computational methods should not be used as "recipes" to follow → non robust science.
  - We emphasize: rigorous understanding to perform routine statistical analysis but also foundation to follow up on specific topics.

Statistical inference



"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.
- We need both to learn from data.

Review of terminology and basic concepts

- Random variable and its distribution
- Models, parameters and their estimators
- CLT
- Hypothesis testing

- **Variable**: An unknown quantity that we'd like to study. "Any characteristic or condition that can change or take different values".

- Most research questions can be formulated as: "What's the **relationship** between two or more variables?"

RV and its distribution

- Random Variable (RV): A variable whose value results from the measurement of a quantity that is subject to variation (*outcome* of an experiment)
  - An RV has a *probability distribution*
  - E.g., expression level of gene X.

- Probability : A number assigned to an outcome, satisfying certain rules (for now okay to think of as *frequency* of an outcome)

- Probability distribution : A mathematical function that maps outcomes to probabilities

Example:

- Experiment: Two coin tosses
- Outcome of interest: number of heads
- Possible outcomes: Sample space
- Mapping between each outcome and a probability

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

- Can you think of other RVs?

Assigning probability to outcomes

ω = an outcome of the experiment
$X(\omega)$ = number of heads

probability	$X(\omega)$
0.25	0
0.25	1
0.25	1
0.25	2
1	

Probability distribution

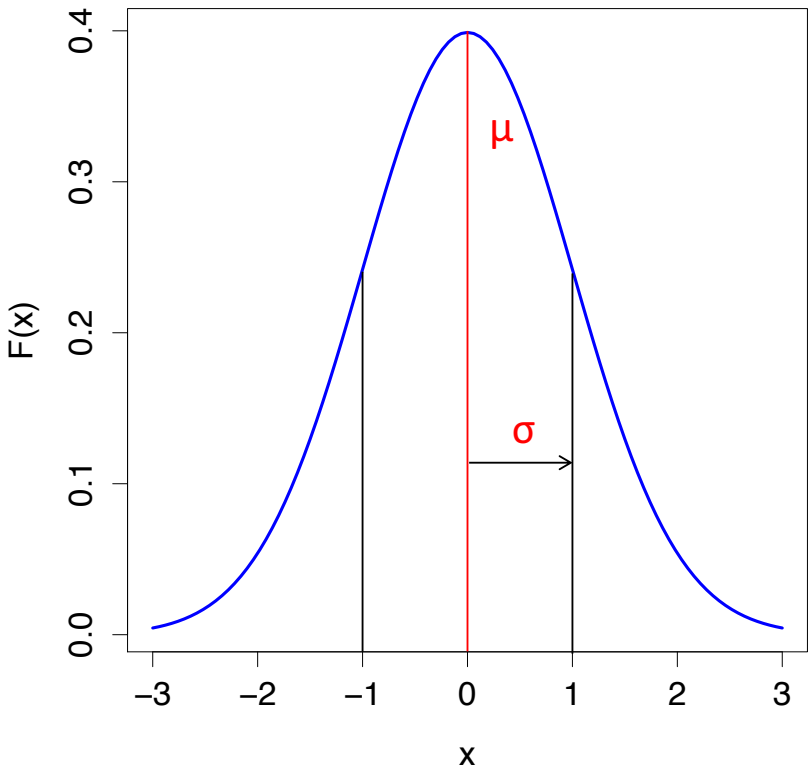
$P(X = x)$ $P_X(x)$	x
0.25	0
0.5	1
0.25	2
1	

Each realization of the random variable corresponds to an event in the sample space, and we can assign a probability to each realization ➔ RV has an associated probability distribution

Two types of random variables

- A discrete rv has a countable number of possible values
  - e.g. dice throwing outcome, 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

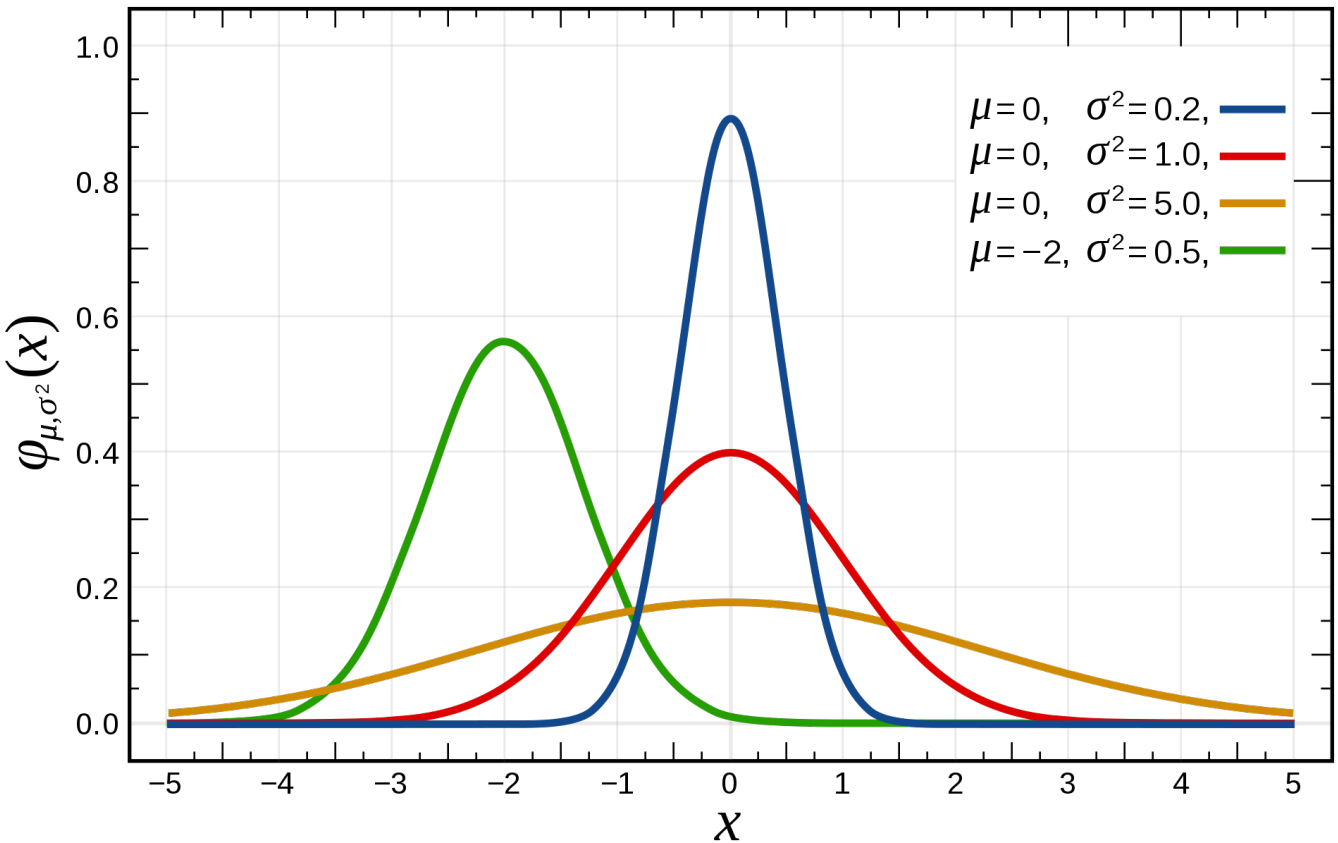
Gaussian (normal) distribution



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

mean = μ
standard deviation = σ

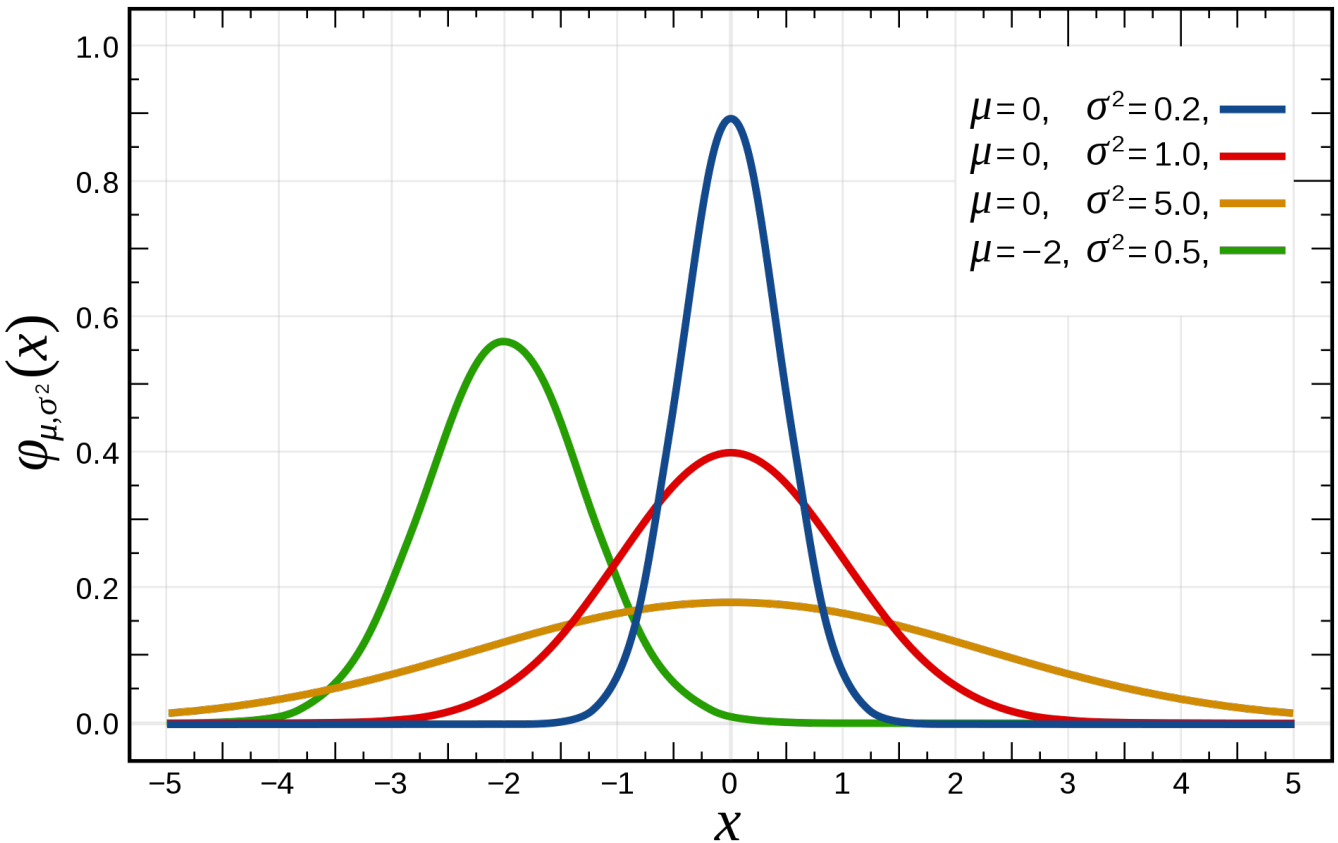
Gaussian (normal) probability density function



$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

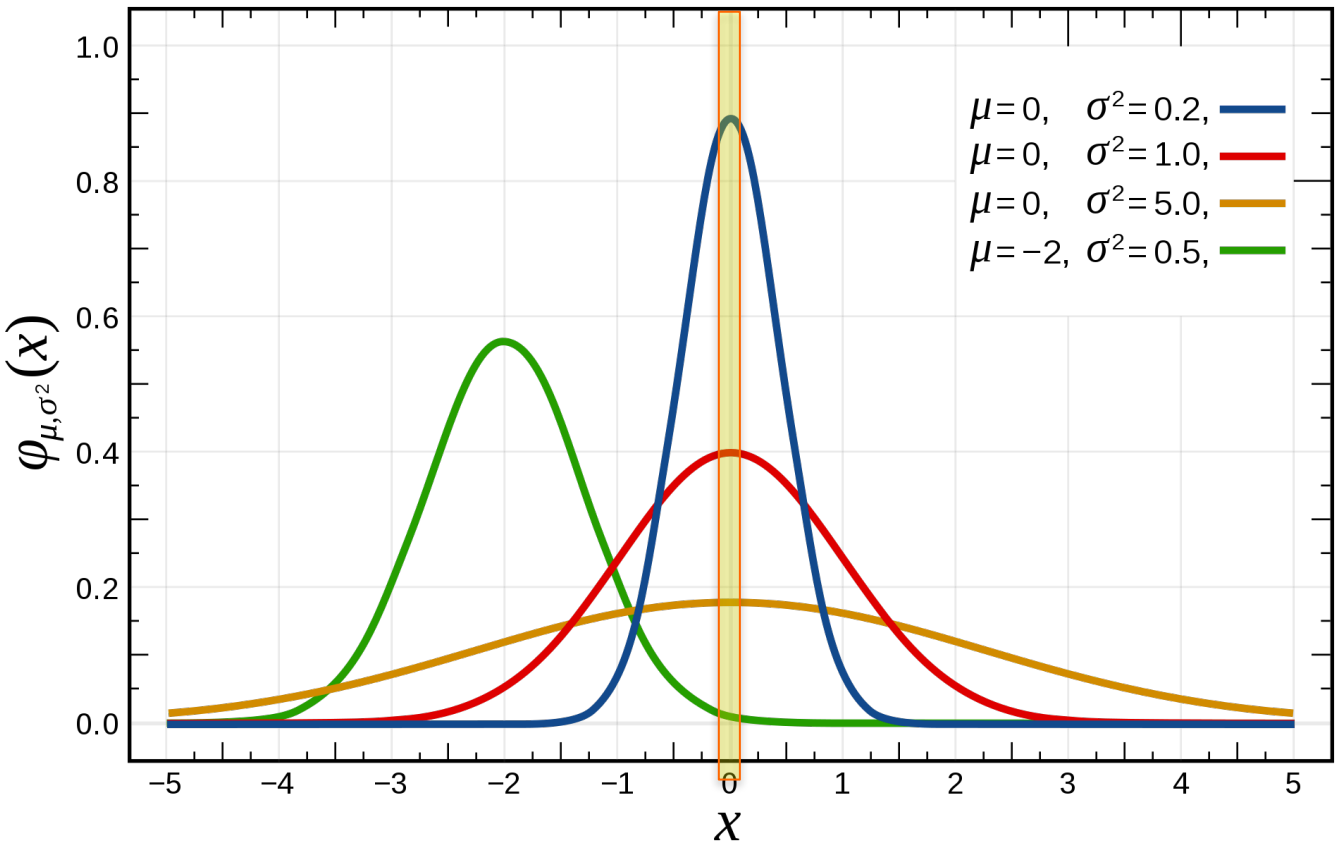
http://math.hope.edu/newsletter/2007-08/pdf-2.gif

Side note: density to probability requires integration



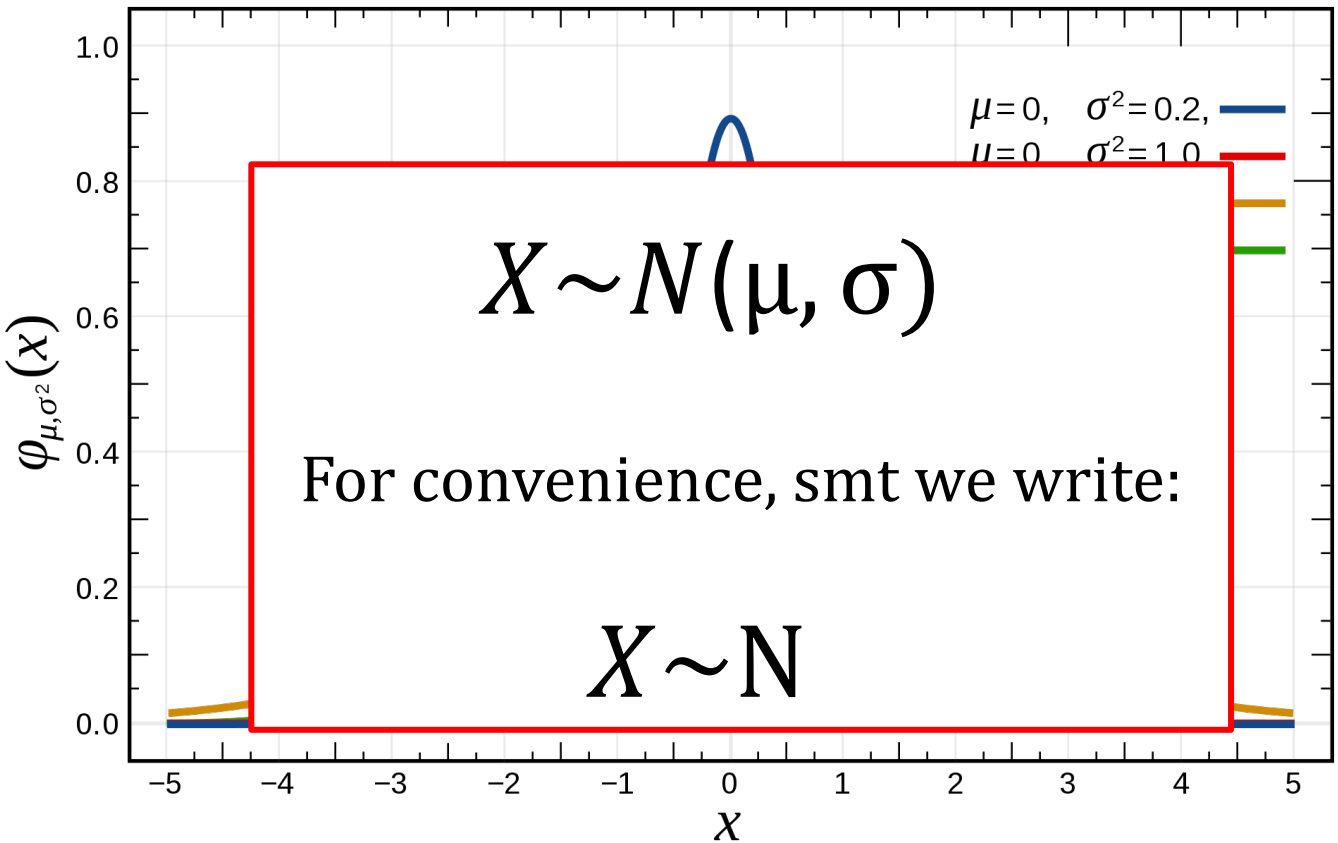
$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Side note: density to probability requires integration



$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Gaussian (normal) probability density function



$$X \sim N(\mu, \sigma)$$

For convenience, smt we write:

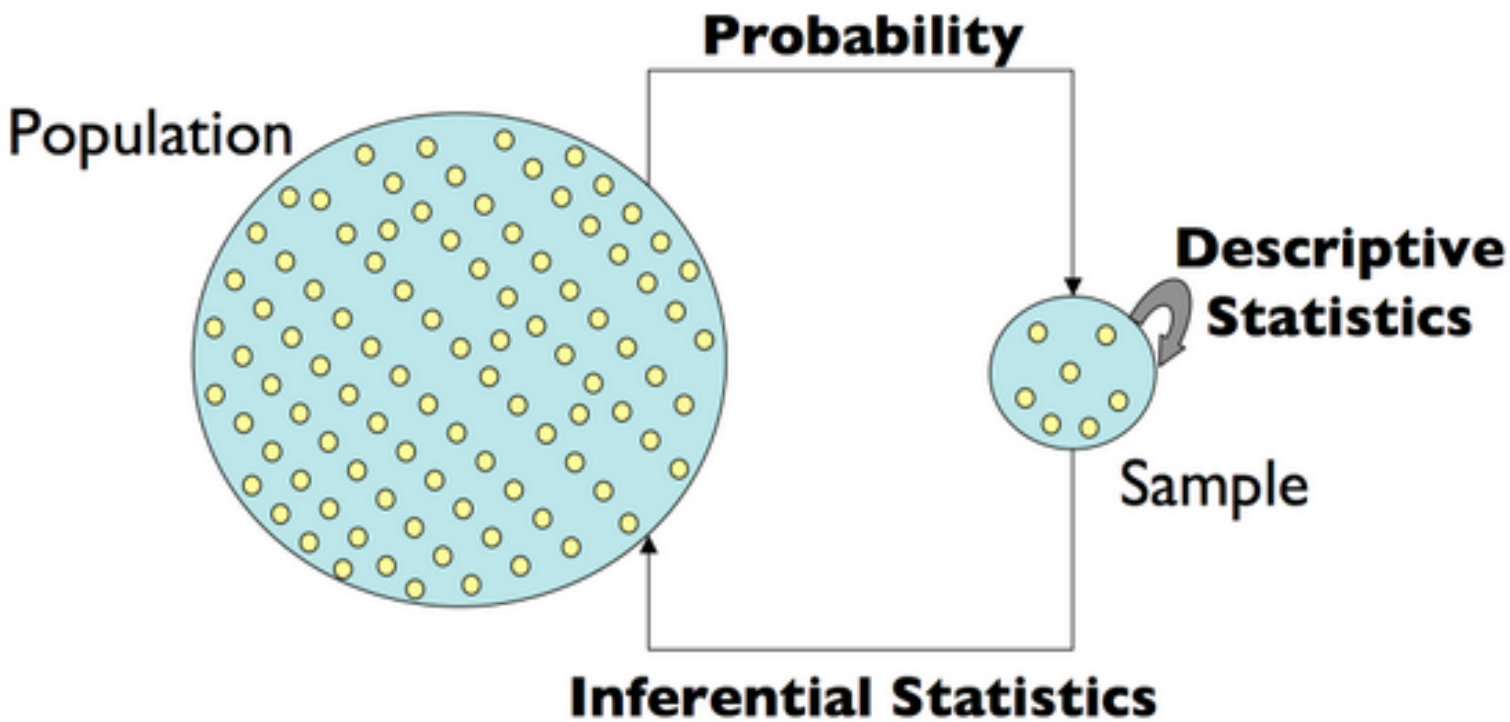
$$X \sim \mathrm{N}$$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Statistical inference

- The ***parameter space*** is the set of all possible values for the parameter

- One major goal: *"figure out"* (i.e., estimate) the **parameter values**; "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
  - For prediction
  - For simulation


Probability
Population
Descriptive
Statistics
Sample
Inferential Statistics

IID

- A {requirement, assumption} in numerous settings is that the data are IID: **I**ndependent and **I**dentically **D**istributed.

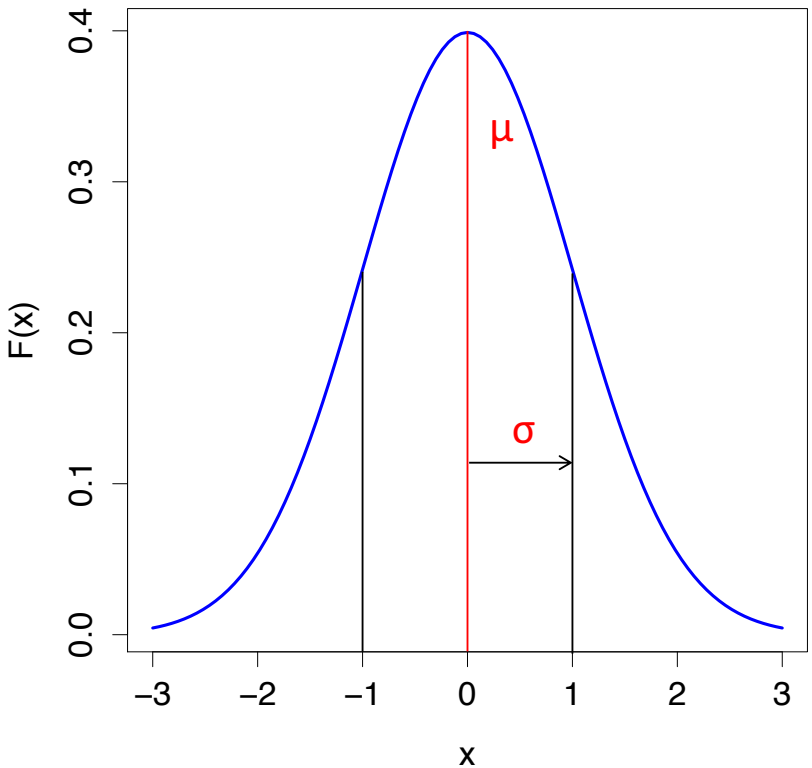
- **Identically Distributed**: a set of observations (events) are from the same population (that is, they have the same underlying probability distribution)
  - E.g. a t-test assumes that under the null, all observations come from the same normal distribution

- **Independent**: all samples satisfy the condition P(A,B) = P(A)P(B) where A and B are events (without loss of generality for any number of events) – that is, the joint probability is the product of the individual event probabilities.

Violations of Independence

- Experimental design is in part about trying to avoid unwanted dependence
- E.g., experiments sampled *related* females from a tall family

Parameters of normal distribution



What two parameters define a normal distribution?

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

mean = μ
standard deviation = σ

Parameter Estimation

- **Estimator**: A function (rule) used to estimate a parameter of interest
- **Estimate**: A particular realization of an estimator

Estimators for normally distributed data

- Given a sample from a normally distributed population, what estimator would you use for μ, σ?

Parameter estimates from our sample/data

$$\hat{\mu} = \bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

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

Estimators for normally distributed data

- Given a sample from a normally distributed population, what estimator would you use for μ, σ?

Parameter estimates from our sample/data

$$\hat{\mu} = \bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

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

True/population parameters

$$\mu$$

$$\sigma$$

Estimator for normally distributed data

- Let's say that we collected a <u>sample</u> from our normal looking population.
- We estimated the mean from our sample, but how good is the estimate?

- What would it depend on?

Estimator for normally distributed data

- Let's say that we collected a <u>sample</u> from our normal looking population.
- We estimated the mean from our sample, but how good is the estimate?

- What would it depend on:
  - Sample size
  - Variability of the population (hence variance)

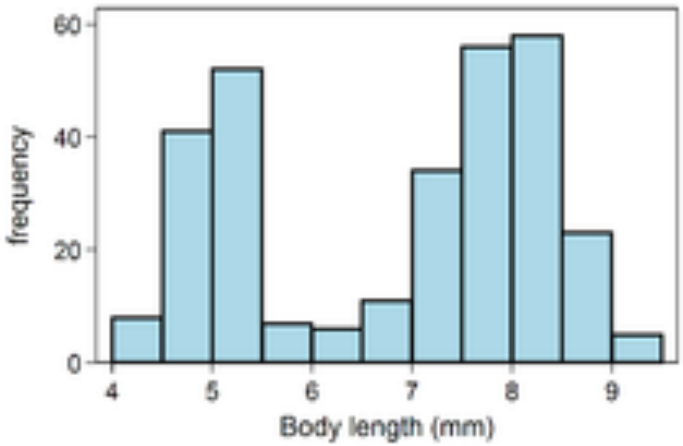
Sampling distribution

- Any function (statistic) of a sample (data) is a random variable
  - Statistic: single measure of some attribute of a sample/data.

- Thus, any statistic, because it's random, has a probability distribution function – this is called the sampling distribution

- Let's focus on the sampling distribution of the sample mean

Central Limit Theorem
Sampling distribution of the Sample Mean



We have a population distribution.

Let's take random samples from it

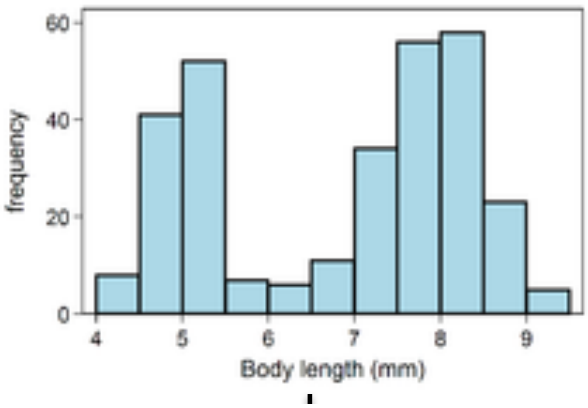
Sample 1 = [5, 7 , 8 , 8]
Sample 2 = [5, 12 , 9 , 10]

$\bar{x_1} = 7$ Sample 1 mean

$\bar{x_2} = 9$ Sample 2 mean

[slide credit: Dr Fowler, UW]

Sampling distribution of the Sample Mean



A population distribution

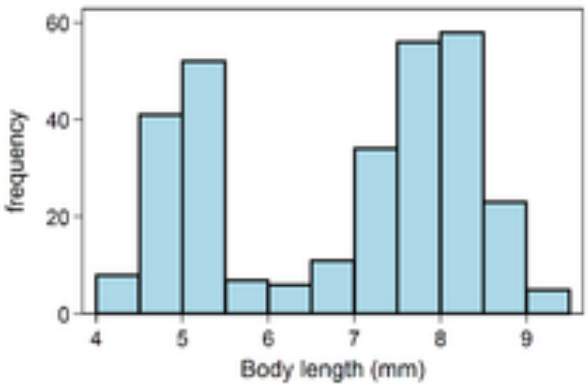
$\bar{x_1} = 7$ A sample mean

$\bar{x_2} = 9$ Another

... Many more

[slide credit: Dr Fowler, UW]

Sampling distribution of the Sample Mean

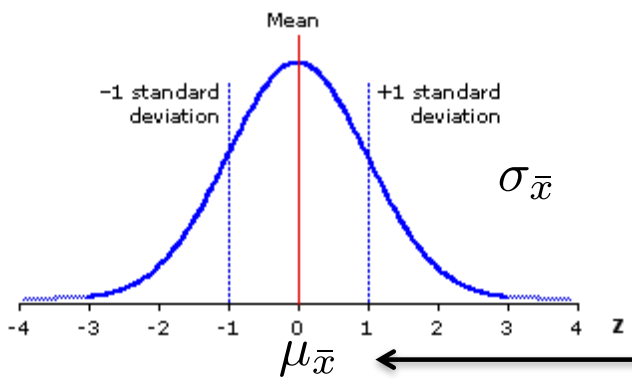


A population distribution

$\bar{x_1} = 7$ A sample mean

$\bar{x_2} = 9$ Another

... Many more



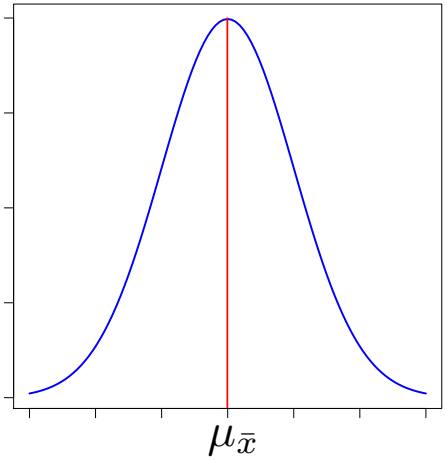
Pop'n mean of the sample means

[slide credit: Dr Fowler, UW]

Sampling distribution

- Recall that the sample mean, $\bar{x}$ is an RV, and hence has an associated distribution

- By CLT, the sampling distribution of the mean is normal:



$$\mu_{\bar{x}} = \mu = \bar{x}$$

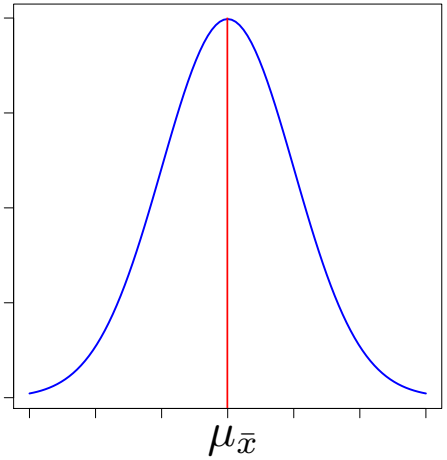
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{s}{\sqrt{n}}$$

[slide credit: Dr Fowler, UW]

Sampling distribution

- Recall that the sample mean, $\bar{x}$ is an RV, and hence has an associated distribution

- By CLT, the sampling distribution of the mean is normal:



Standard Error (SE)

$$\mu_{\bar{x}} = \mu = \bar{x}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{s}{\sqrt{n}}$$

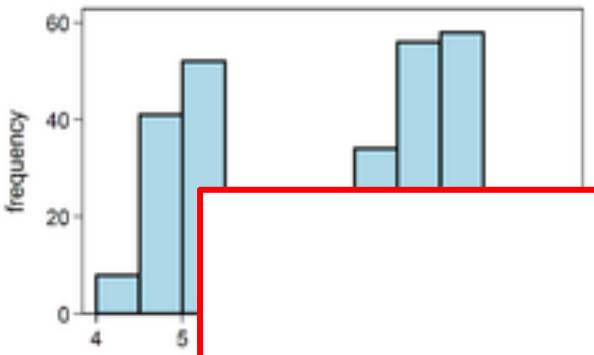
[slide credit: Dr Fowler, UW]

Standard Error of the Mean

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{s}{\sqrt{n}}$$

- SE is the standard deviation of the sampling distribution of the mean
- Often get's confused in literature as "standard deviation" – pay attention to this, given that SE is smaller than SD.
- SE reflects the uncertainty about where the population mean be located, given a sample.
- When sample size ~30, then the normal distribution is a good approximation for the sampling distribution of the sample mean. With smaller samples, the SE $\frac{s}{\sqrt{n}}$ is an underestimate.

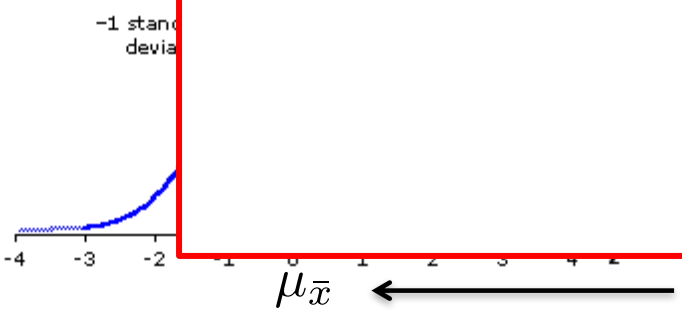
Sampling distribution of the Sample Mean



A population distribution

$\bar{x_1} = 7$ A sample mean

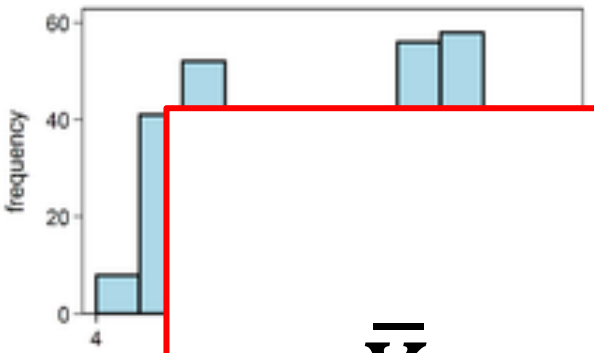
$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$



$\mu_{\bar{x}}$ ← Pop'n mean of the sample means

Sampling distribution of the Sample Mean

A population distribution

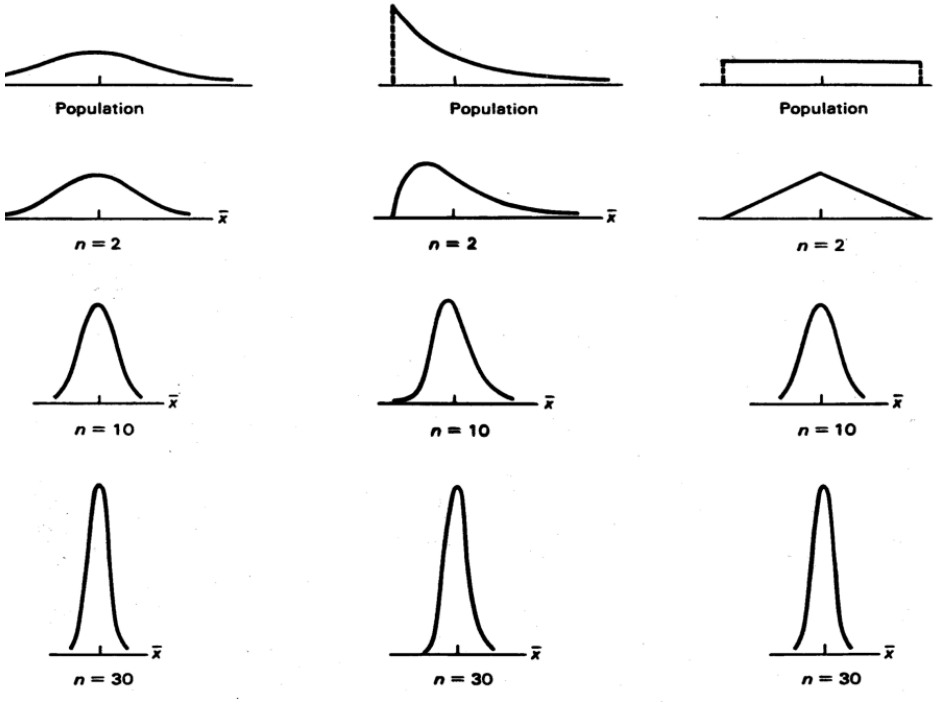


$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(?\,,?\,)$$

$\mu_{\bar{x}}$ ← Pop'n mean of the sample means

Central Limit Theorem

Let $X_1, X_2, \ldots$ be an iid random sample from some population with non-normal distribution . If the sample size is sufficiently large, then the sampling distribution of the mean will be normal.



$$\overline{X} \sim N(\mu \ \frac{\sigma}{\sqrt{n}})$$

CLT Summary

- If sample size is large, the sample mean follows a normal distribution centered at population average and with standard deviation following population standard deviation.

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

Hypothesis testing

- Hypothesis: A *testable (falsifiable)* idea for explaining a phenomenon.
- Statistical hypothesis: "is 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.

Hypothesis testing

- Motivating example: given the expression level of gene g in some disease (cancer) and some healthy (control) samples, determine if gene g is differentially expressed in cancer vs. healthy.

- Requires comparing two hypotheses: null hypothesis H_0 and alternative hypothesis H_A.

- Allows determining if differences between observed data in two groups are significant.

Steps in hypothesis testing

1. Formulate your hypothesis as a statistical hypothesis.

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

3. Compute the p-value associated with the observed test-statistics under the null distribution. $p(t_0 | H_0)$

Hypothesis testing: T-test

- Expression level of gene g measured for n cancer and m healthy samples:
  - $z_1, z_2, \ldots, z_n$ & $y_1, y_2, \ldots, y_m$

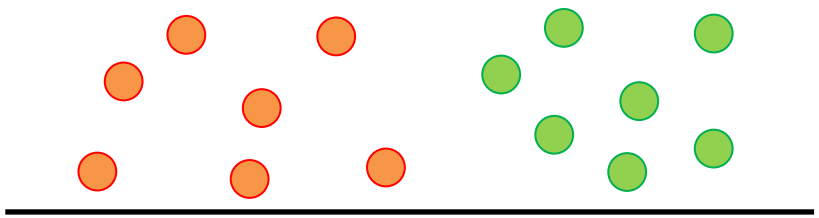
- Population mean expression of gene g is **significantly** different in cancer and healthy

- Null and alternative hypothesis

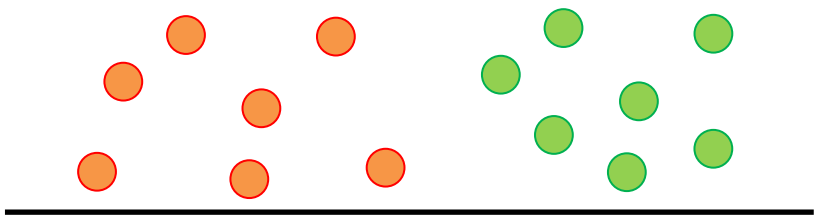
  H_0: $\boldsymbol{\mu}_z = \boldsymbol{\mu}_y$

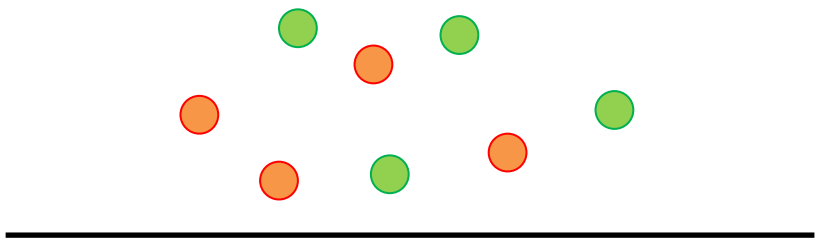
  H_A: $\boldsymbol{\mu}_z \neq \boldsymbol{\mu}_y$

Three scenarios

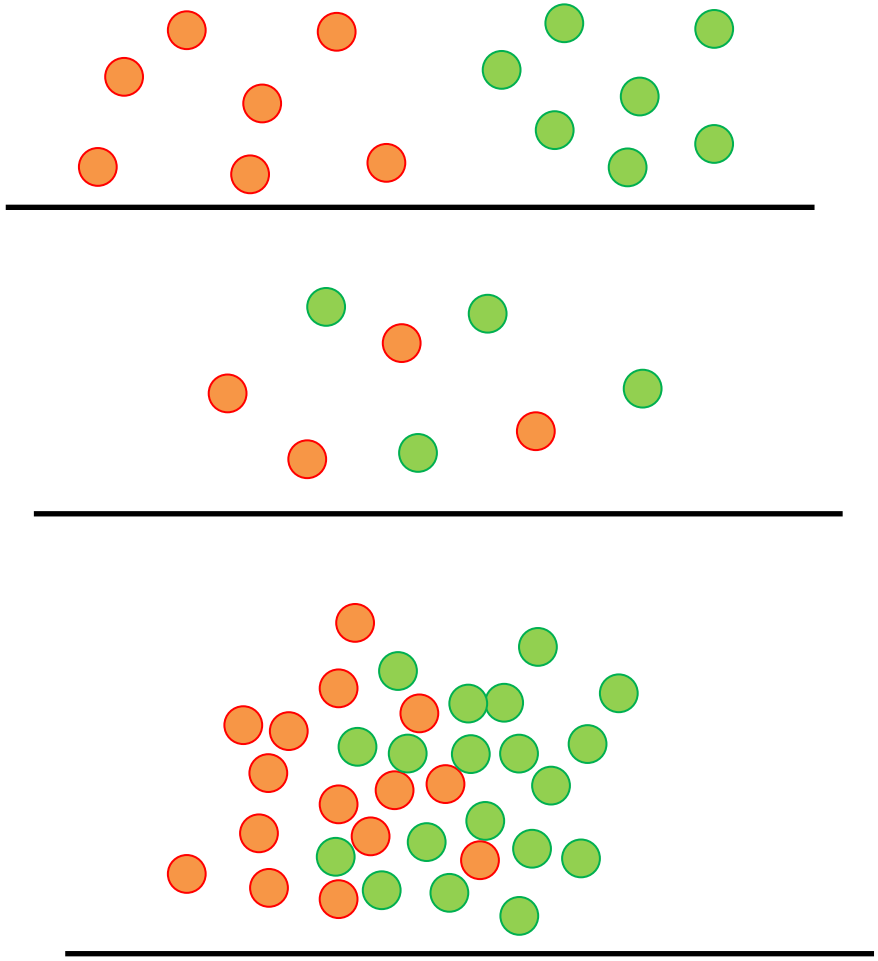


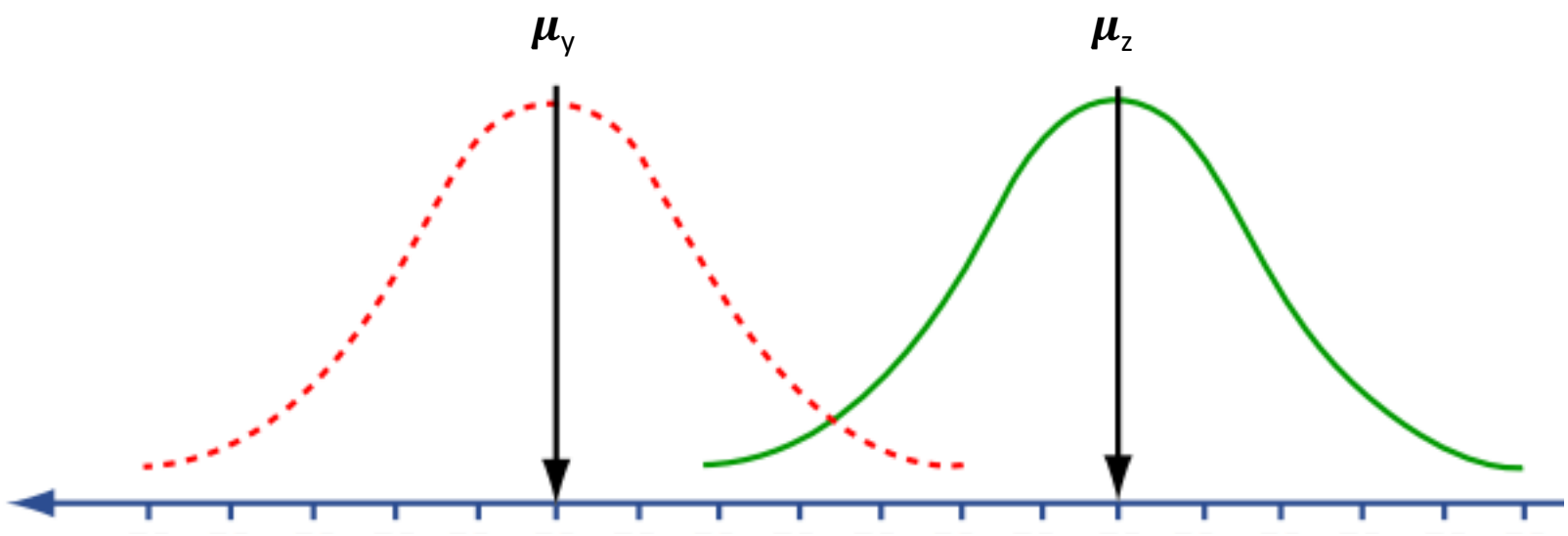
Three scenarios





Three scenarios





Is there a **significance** different between the two means?

Occurs when the mean difference is put in the context of *spread (standard deviation)* of the data. Also depends on the sample size.

T statistic: $\frac{\bar{Y}-\bar{Z}}{SE}$

Standard Error:
standard deviation/sqrt(sample size)

Hypothesis testing: T-test

- Expression level of gene g measured for n cancer and m healthy samples:
  - $z_1, z_2, \ldots, z_n$ & $y_1, y_2, \ldots, y_m$

- Null and alternative hypothesis

$$t = \frac{\bar{z} - \bar{y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

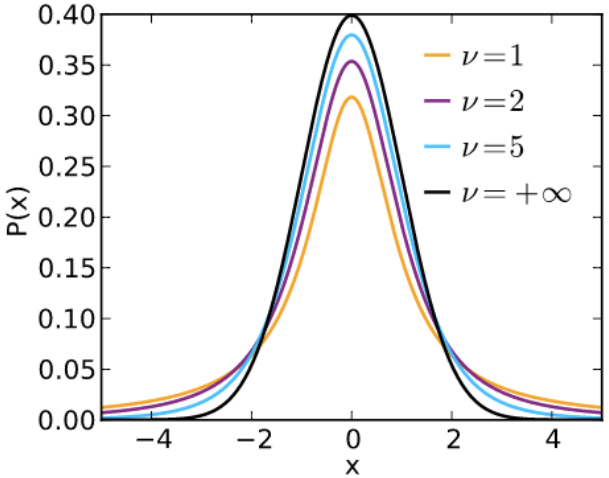
$$S_p^2 = \frac{\sum_{i=1}^{n} (z_i - \bar{z})^2 + \sum_{i=1}^{m} (y_i - \bar{y})^2}{(n-1) + (m-1)}$$

Hypothesis testing: T-test

- From theory we know the distribution of our test-statistics, if we are willing to make some assumptions:
  - Assuming normal distribution for X and Y, with equal variance then

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

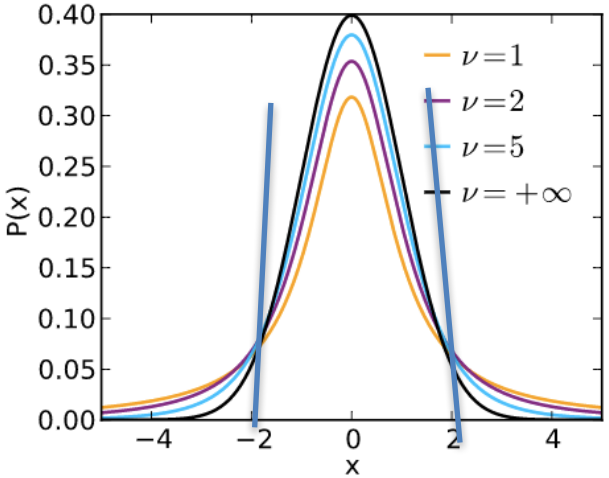
Degrees of freedom



Hypothesis testing: T-test

- Plug in the observed t-statistic to find the probability of observing a value as larger or larger than the one observed.

e.g., t-stat = 2 ; two sided



Summary

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