JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Proteinbiochemie und Bioinformatik

Introduction to biostatistics

David Fournier
dfournie@uni-mainz.de

20.03.2017

Outline of the course

FIRST WEEK

Day 1: **Biostatistics** (lecture)

Day 2 (morning): **Clustering and machine learning** (lecture)

Day 2 (afternoon) – Day 5: **Initiation to programming with Python** (lecture + practical)

SECOND WEEK

Day 6 (morning): **Application of Python to statistics** (practical)

Day 6 (afternoon): **Sequence analysis and homology** (lecture)

Day 7: **Multiple sequence alignments and phylogenies** (lecture)

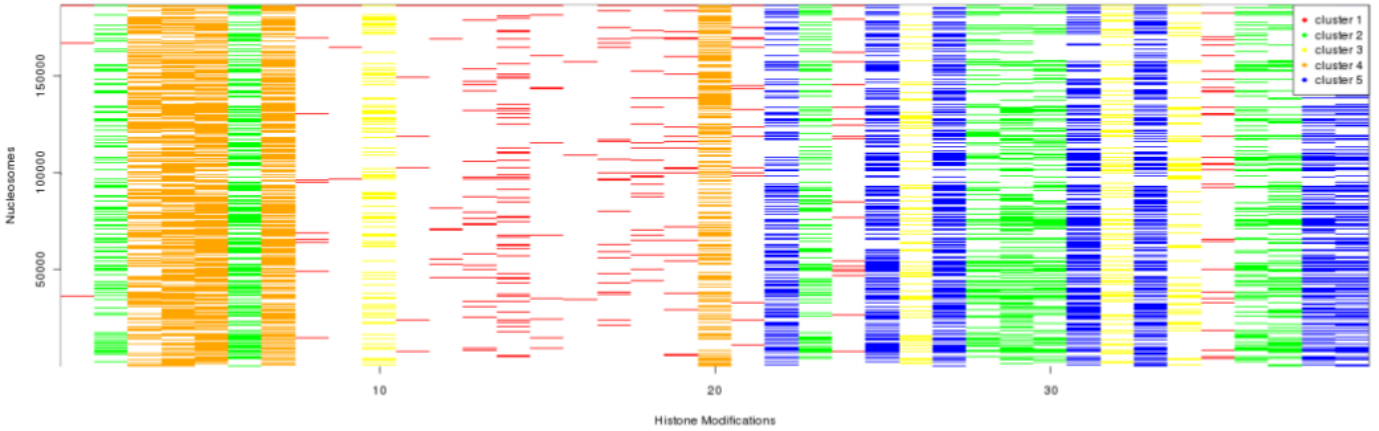
Day 8: **Protein structure** (lecture)

Day 9: **Protein interactions and dynamics** (lecture)

Day 10: **Presentation of a scientific article** (seminars 15-27)

Describing biological phenomenon

Exhaustive approaches

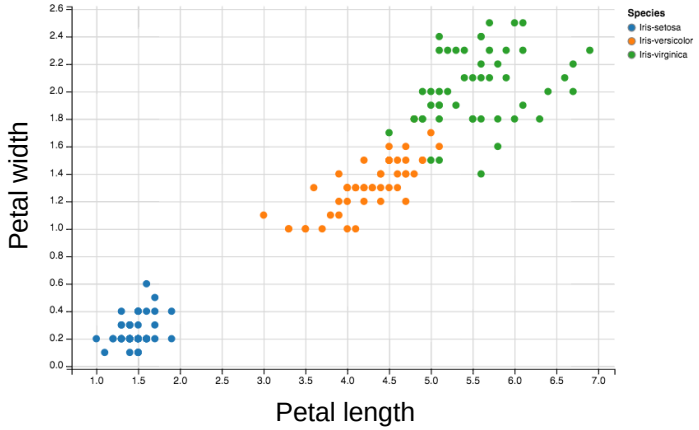


Status of various epigenetic states at all known positions on the genome
Describe the data using:

⇒ Total probability (no model error)

⇒ Clustering methods

Summary variables



Want to grasp variability in few measures
Describe the data using:

⇒ Metrics such as mean or variance

⇒ Usually using distance-based methods (ex. correlation)

⇒ Associate these values with a degree of certainty

Definition of biostatistics

Biology: find (experiment, analysis) and describe meaningful signals (i.e. that carry information). For instance, a virus in a blood sample, the change of expression of a gene.

Biostatistics: application of statistics - that is the study of random variable metrics, summarization and interpretation - to biological questions. Biostatistics will help to describe or detect signals.

Statistics can help to :

- Design studies
- Summarize data
- Representation of data
- Analyze data
- Interpret data
- Infer information from data

Usages of biostatistics

Two kinds of approach:

- Descriptive biostatistics: only the metrics of the sample are reported. This approach is usually in all cases, to present the
- Inference biostatistics: From a sample of population, derive rules about the entire population. Important approach in biomedical research.

Mathematical notations and concepts

$$\sum_{i=1}^{n} x = 1 + 2 + \cdots + \mathrm{n}$$

$$\sum_{i=1}^{n} (x_i - m)^2 = (x_1 - m)^2 + (x_2 - m)^2 + \cdots + (x_i - m)^2$$

$$\prod_{i=1}^{n} x = 1 \times 2 \times \cdots \times \mathrm{n}$$

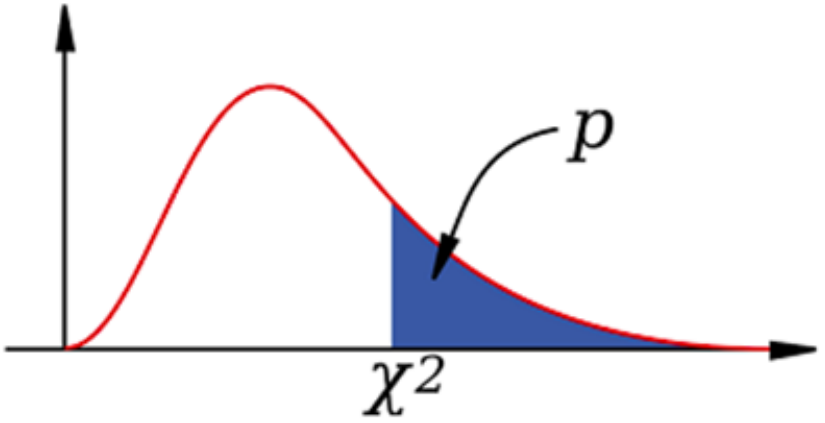
$$x\,! = 1 \times 2 \times \cdots \times \mathrm{n}$$

Mathematical notations and concepts

Density distribution:
Plot that shows the distribution of values of a given dataset. For instance, the distribution of height in a human population.

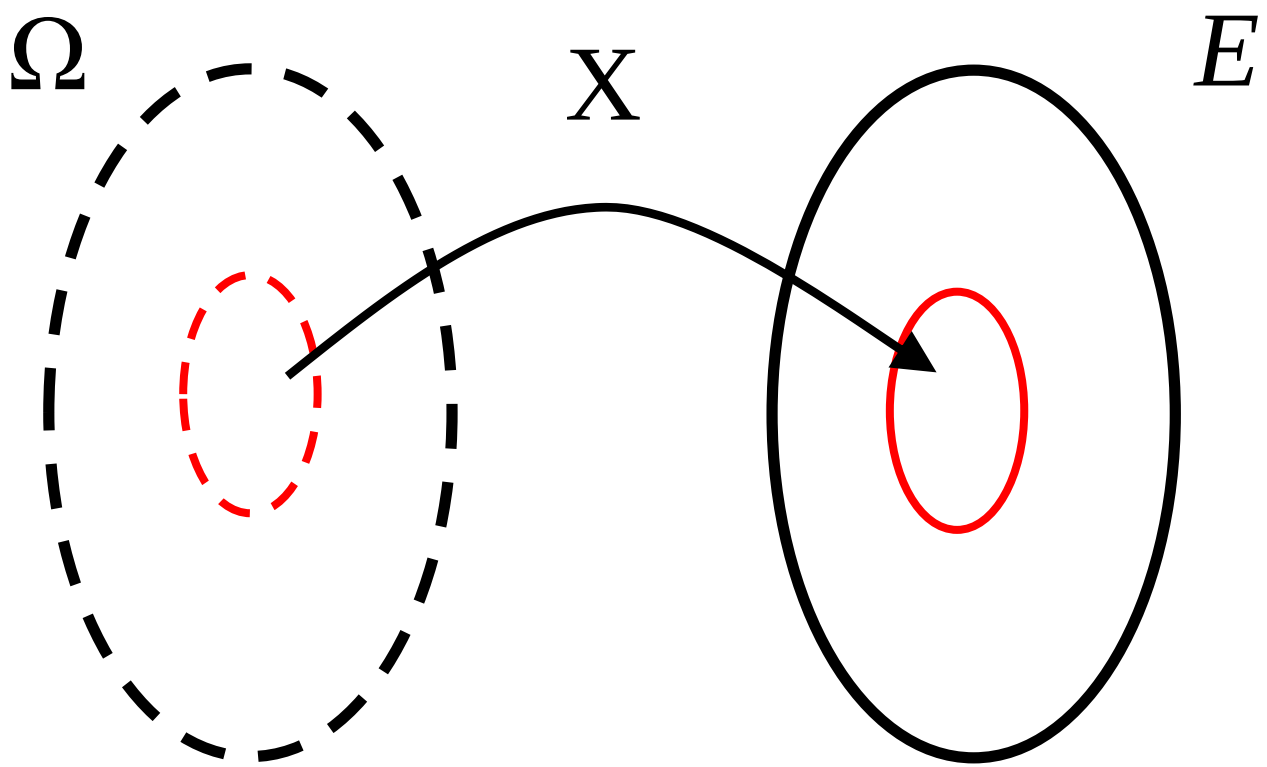
Probability distribution:
Similar distribution, with all values normalized to the sum of data values. As a result, total area under the curve is 1. The area under the curve between two values a and b is the probability to pick randomly a value between a and b.
In practical terms, these two distributions have the same shape, only the scale on y axis changes.

Example:
In the example below, probability to have a value equal or greater than X^2 is p, the area under the curve.

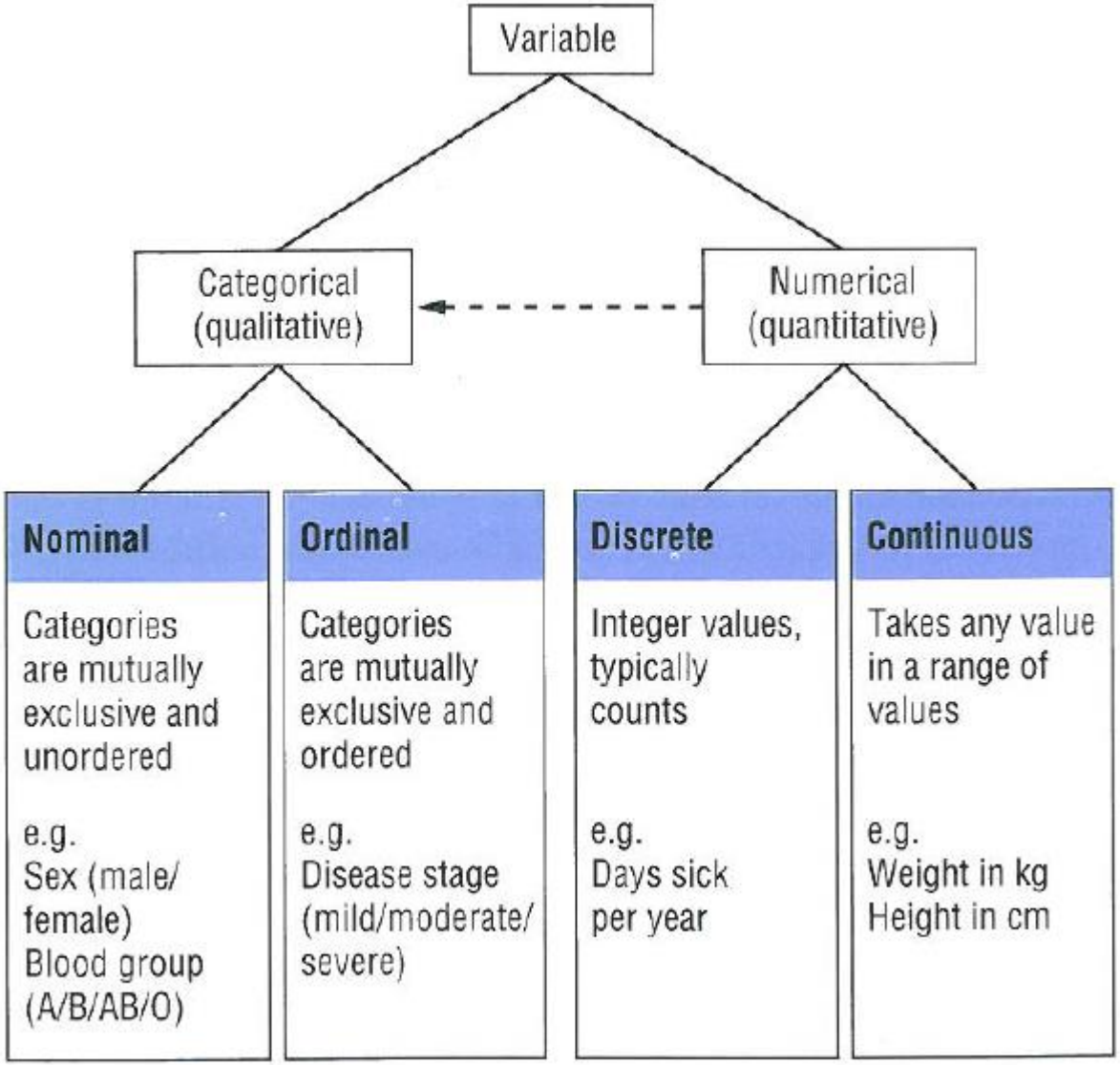


Random variables

Definition:
Value that can be any of a universe of possible outcomes.



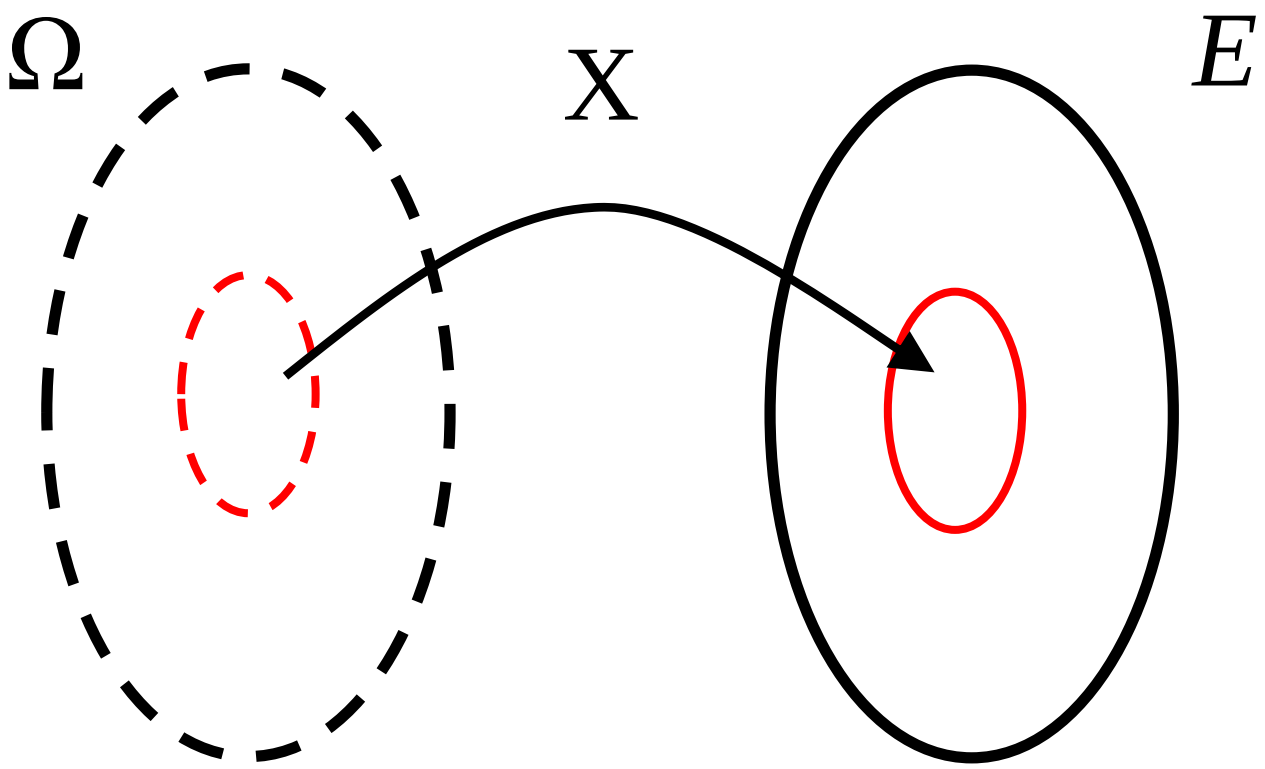
Level of measurement



Example: dice roll

Ω ={1,2,3,4,5,6}

X={1,6,6,2,6,5,6,1,4,5,5,6,1,6}



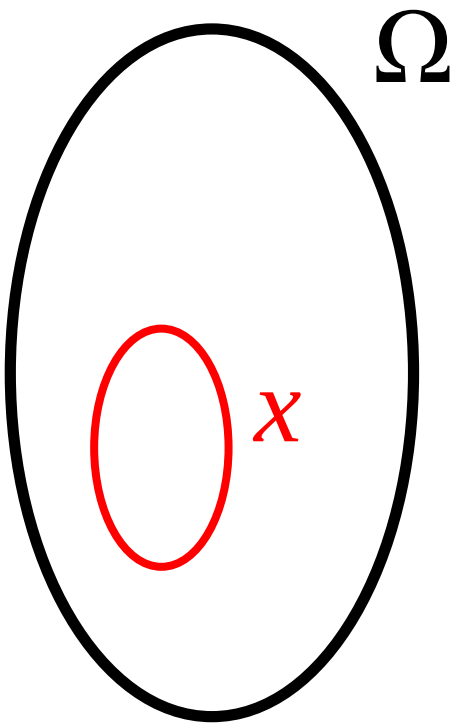
Fundamentals of probabilities

Let two ensembles Ω and *x*, Ω being the universe of all possible outcomes from a dice roll:
Ω={1,2,3,4,5,6}
and *x* being an outcome of Ω (ex: getting a 1 with a dice).

$P(x)$ is the probability of the outcome *x* to happen :

$P(x) \in [0,1]$

$$\sum_{i=1}^{n} P(x_i) = 1$$



Example of a dice roll:

$$P(1) = \frac{1}{6}$$

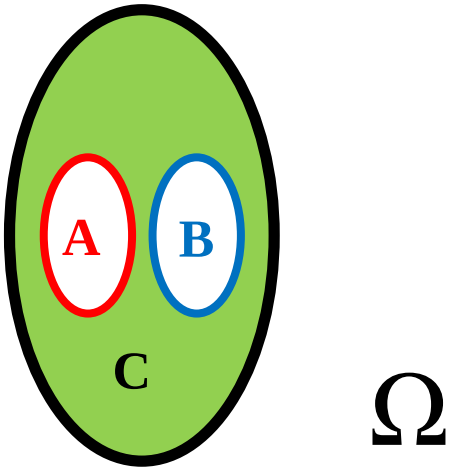
Which means that one has 1/6, that is roughly 17% chance to get a 1.

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$$

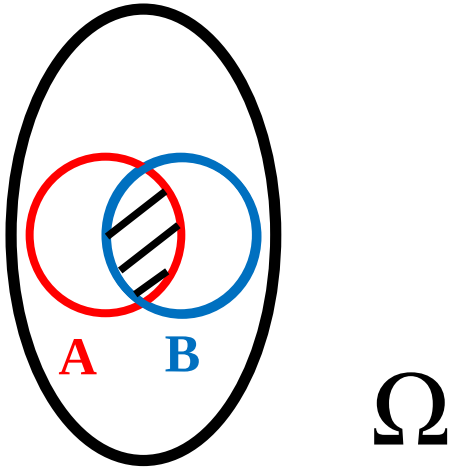
$$\sum_{i=1}^{n} P(x_i) = P(1) + P(2) + P(3) + P(4) + P(5) + P(6) = 1$$

Fundamentals of probabilities

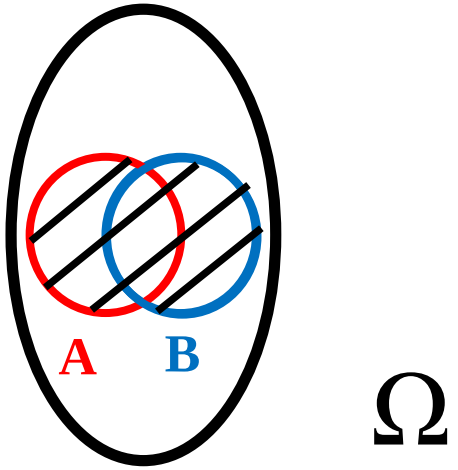
Union of two independent (i.e. not overlapping) events:
$P(A \cup B) = P(A \text{ OR } B) = P(A) + P(B)$
Union of all possible independent events is equal to one:
$P(A \cup B \cup C) = 1$



Intersection of two events signs **a dependence**:
$P(A \cap B)$
(known previously or two be estimated)
Ex: get a 6 from tice roll and getting 12 from two dice rolls overlap.



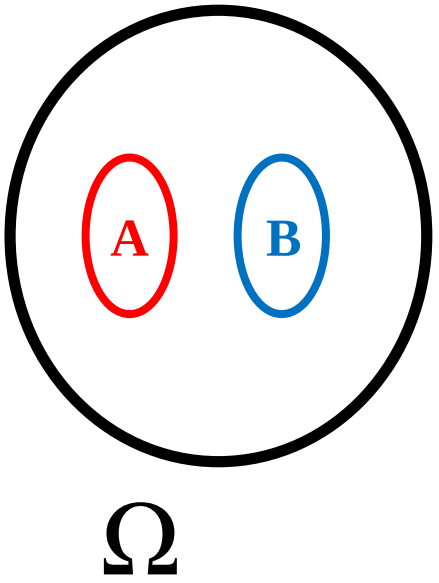
Exercise.
Draw a diagram of that specific example in the fashion of the diagrams shown on the right.

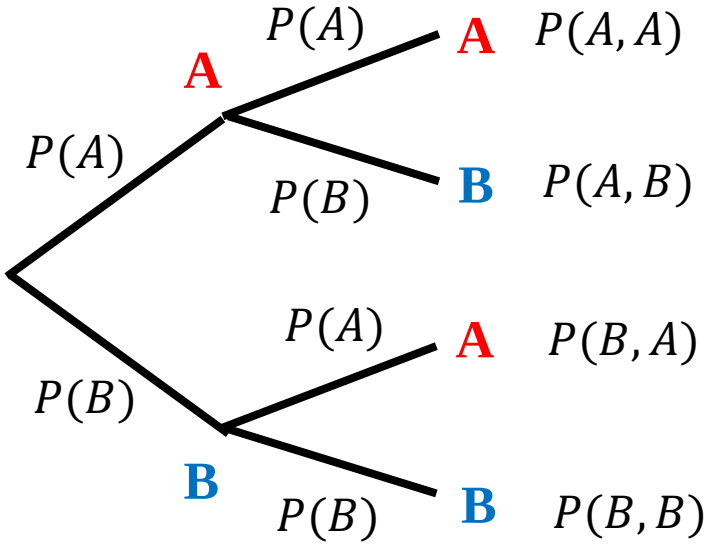


Union of two dependent (i.e. overlapping) events:
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Because $P(A) + P(B)$ includes two times $P(A \cap B)$

Fundamentals of probabilities

Product of two independent (i.e. not overlapping) events:
$P(A, B) = P(A \text{ AND } B) = P(A) \times P(B)$
Can be either from the same ensemble, or from two different ensembles (ex: two dice rolls in a row or one dice roll and one card draw. Can be same value:
$P(A, A) = P(A) \times P(A)$
Exercise: what is the probability to get twelve times a 1 in dice roll? To get 6 times a 1 in dice roll and draw four times a diamond from a card deck?

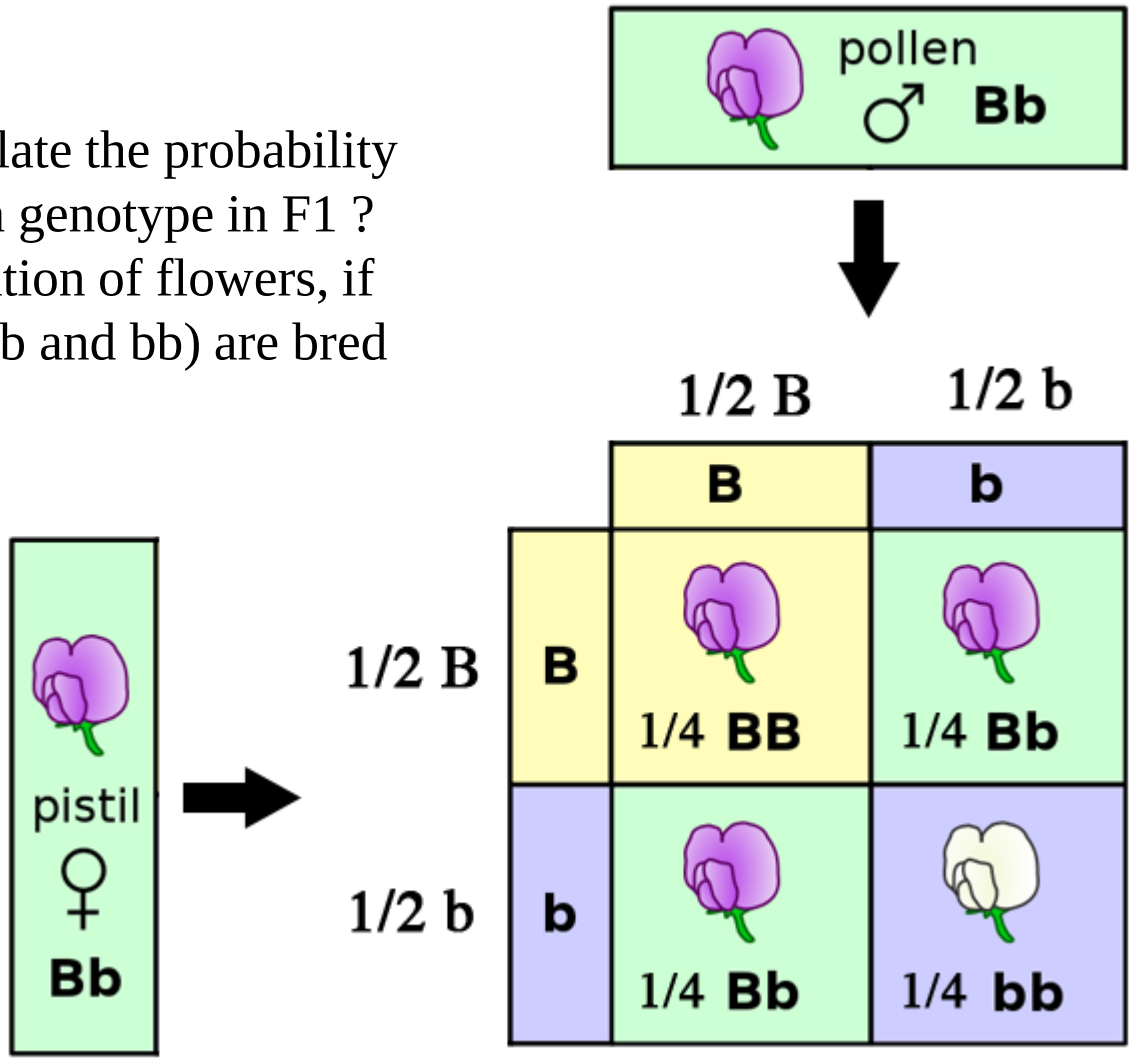




Fundamentals of probabilities

Exercise.

How do you calculate the probability Associated to each genotype in F1 ?
F1: Second generation of flowers, if F0 flowers (BB, Bb and bb) are bred together.



Fundamentals of probabilities

Exercise.

How do you calculate the probability Associated to each genotype in F1 ?
F1: Second generation of flowers, if F0 flowers (BB, Bb and bb) are bred together.

A	**B**
C	**D**

4 genotypes in F1

10 ways to combine the genotypes: A^2 or AB or AC or AD or BC or BD or B^2 or CD or C^2 or D^2

BBxBB > 1 BB
2 (BBxBB) > 2 (½ BB + ½ B/b)
BBxbb > 1 Bb
3 (BbxBb) > 3 (¼ BB + ¼ Bb + ¼ bB + ¼ bb)
2 (Bbxbb) > 2 (½ Bb + ½ bb)
bbxbb > 1 bb

1 BB + 1BB + ¾ BB = 11/4 => (11/4)/10 = 27.5% BB in F1
1 Bb + 1Bb + 3/2 Bb + 1Bb = 9/2 => (9/2)/10 = 45% Bb in F1
¾ bb + 1bb + 1bb = 11/4 => 27.5% BB in F1

Exercise

Here is a table that shows the frequency of balls of different colors present in a box.

Item	Red	Blue	green
Probability	0.2	0.2	0.6

Calculate the probability that pick one red, one blue and one green ball if one draws 3 balls randomly from the box.
Calculate the probability that get only red and blue balls if one draws 3 balls randomly.
Calculate the probability to get a red and a blue or at least a green if you draw 2 balls.

Exercise

Here is a table that shows the frequency of balls of different colors present in a box.

Item	Red	Blue	green
Probability	0.2	0.2	0.6

Calculate the probability that pick one red, one blue and one green ball if one draws 3 balls randomly from the box.

$\boldsymbol{P(r) \times P(b) \times P(g) = 0.2 \times 0.2 \times 0.6 = 0.024}$ **(2.4% chance)**

Calculate the probability that get only red and blue balls if one draws 3 balls randomly.

$\boldsymbol{P(r \cup b)^3 = (0.2 + 0.2)^3 = 0.064}$ **(6.4% chance)**

Calculate the probability to get a red and a blue or at least one green if you draw 2 balls.

- getting a red and a blue: $\boldsymbol{P(r \times b) + P(b \times r) = x = 0.2^2 + 0.2^2 = 0.08}$

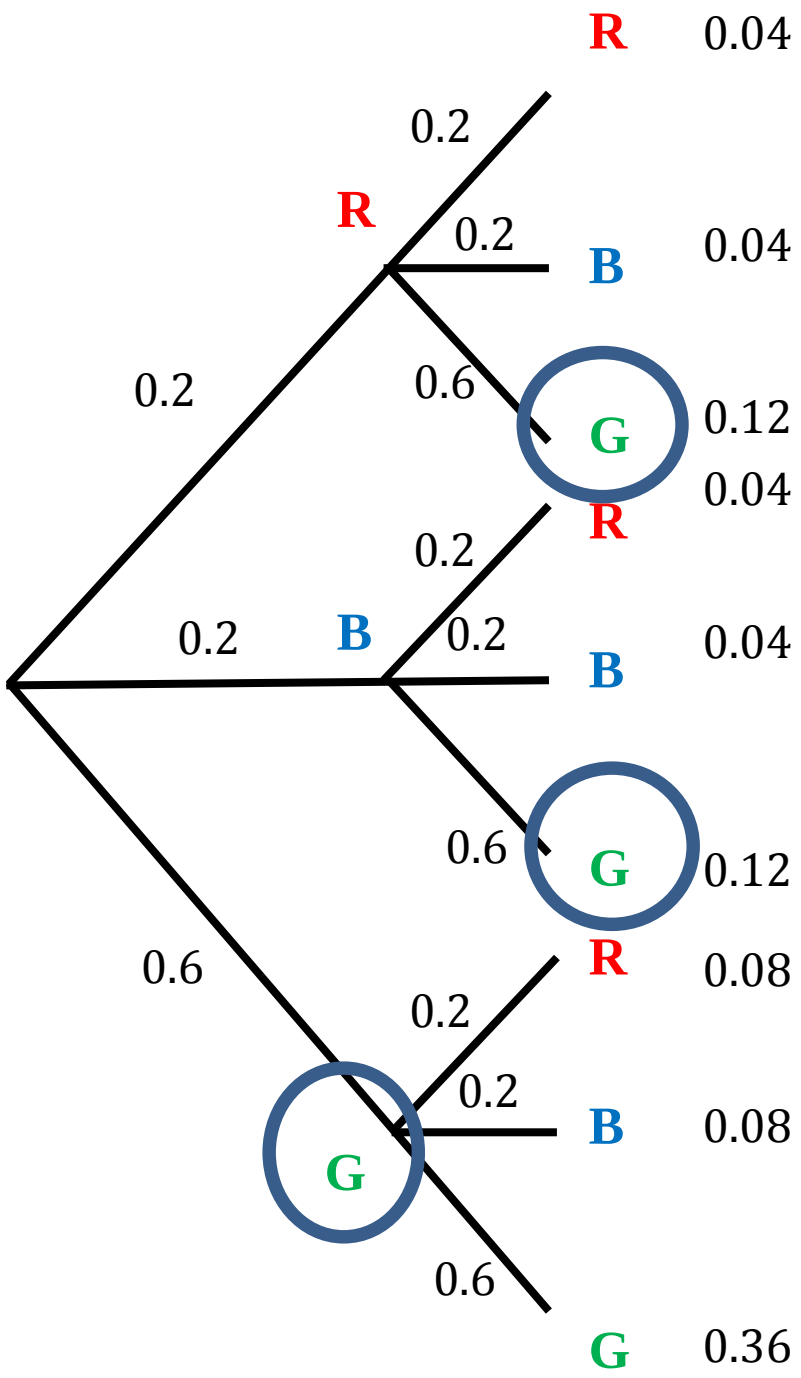
[either you get a red first and blue second or a blue first and red second].

- getting at least a green: $\boldsymbol{P(g) = y = 0.72}$.

[see explanation on next slide for $\boldsymbol{P(g)}$**]**

- Probability to get a read and a blue or at least one green:

$\boldsymbol{P = x + y = 0.08 + 0.72 = 0.8}$



$P(g) = P(r, g) + P(b, g) + P(g, rgb) =$ $0.12 + 0.12 + 0.6 = 0.72$, 72% chance of getting at least a green ball.

Conditional probabilities

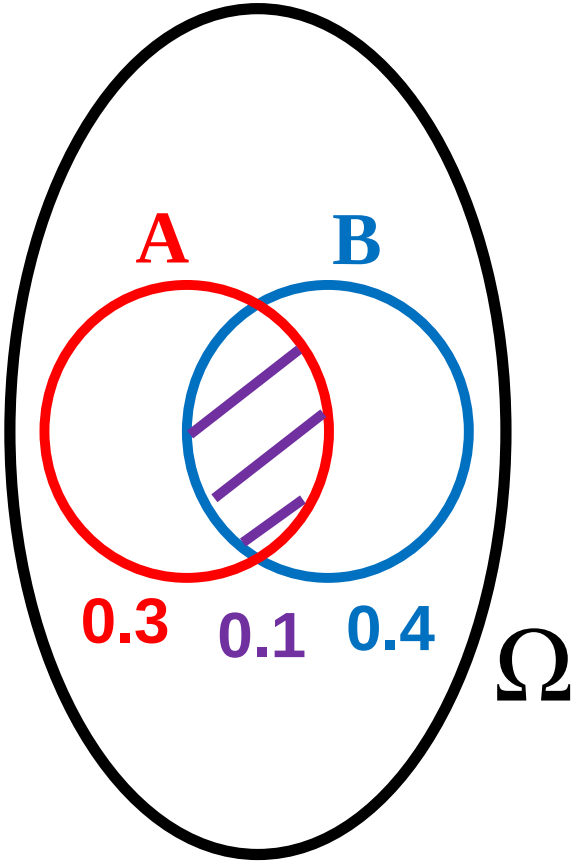
Case where two variables are **dependent**.
Notation: $P(A|B)$, with:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (1)$$

i.e. Probability of A happening if B has already happened.
The equation is intuitive if you consider that $P(A|B)$ is the percentage of $P(B)$ that $P(A \cap B)$ represents.
In other words: what is the chance that we are in the intersection if B has happened ?

Note that (1) cannot be calculated if B has zero probability to appear ; conditional probability is then useless as B never happens.

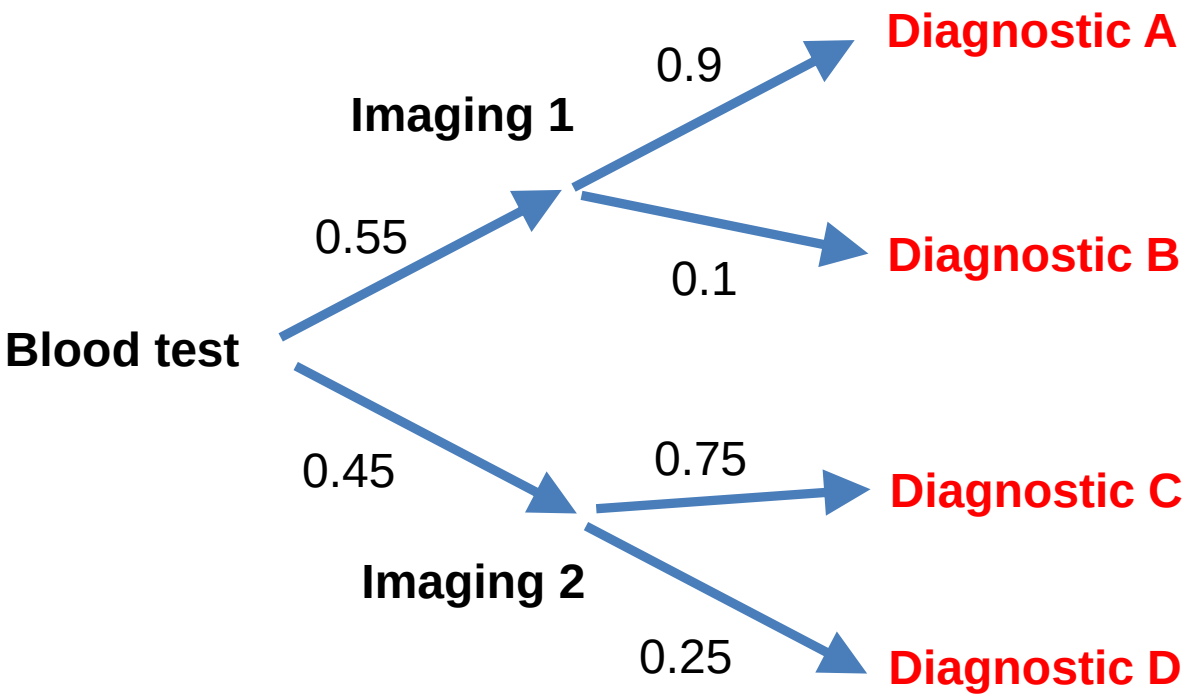
Case where two variables are **independent**.
$P(A|B) = P(A)$



Exercise

Here is a tree for medical diagnostics where each decision leads to more possible decisions. Knowing the probability of each transitions, calculate the probability to reach each diagnostics.

Is the sum of probabilities of diagnostics equal to 1? Why?



Application: outcomes of a screening test

Here is a table presenting the number of occurrences of people either baring a disease or not, and either getting positive or negative at a screening test.

Question: What is the probability to be sick if a person is tested positive ?

		Disease		
		Yes	No	
Test	Positive	**0.11**	0.06	0.17
	Negative	0.04	**0.79**	0.83
		0.15	0.85	1

Application: outcomes of a screening test

Here is a table presenting the number of occurrences of people either baring A disease or not, and either getting positive or negative at a screening test.

Question: What is the probability to be sick if a person is tested positive ?

		Disease		
		Yes	No	
Test	Positive	**0.11**	0.06	0.17
	Negative	0.04	**0.79**	0.83
		0.15	0.85	1

Answer: $\boldsymbol{P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.11}{0.17} = 0.647}$**, which represents 64,7% to have the disease.**

Posterior probabilities

Posterior probability: Conditional probability of having parameters of a certain value given the data.

$P(\theta|x)$, with θ the parameter(s) of your model (the dice) and *x* the data or observations (the outcomes of dice rolls), i.e. "probability of the model given the data".

As a conditional probability, $P(\theta|x)$ is equal to:

$$P(\theta|x) = \frac{P(x \cap \theta)}{P(x)}$$

Example:
Say you have several dices, one being loaded (i.e. probabilities of each number not equal). You choose a dice randomly and roll it ten times.
$P(\theta|x)$ will be the probability to have chosen the fair dice in hand given the results of the dice rolls. The dice is the model you consider (θ) and *x* are the data, i.e. the results of dice rolls.

Theorem of Bayes

- Usually posterior probability cannot be directly calculated, and one needs to introduce Another probability called **likelihood**, $P(x|\theta)$.
- Likelihood is related to posterior probability via the theorem of Bayes:

$$P(\theta|x) = \frac{P(x|\theta).P(\theta)}{P(x)}$$

With $P(\theta)$ being the prior probability, i.e. the probability of the model and $P(x)$ the probability of the data.

Likelihood: probability of having the data given a certain model: $P(x|\theta)$. This is usually easier to calculate than $P(\theta|x)$.

Probability of the model or Prior probability (i.e. before having looked at data)
Can be for instance picking up one of two dices (each of the dices is a model).

Notion of likelihood

Example of likelihood:
The probability of drawing a 6 using a fair dice is $P(6|\theta)$, that is 1/6. The likelihood of Getting a 6 with model θ where $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$ is 1/6.

Exercise.
- Let θ be a loaded dice for which number 6 has two times more chance to appear than other values, which are equally probable. Explicit the model: what are the probabilities of each number to appear? What is the likelihood of getting a 6 three times in a row ?

Odd ratio

Definition

Generally speaking, odd ratios are helpful to know how much more likely one event compared to the other. Example: odds of a football team A to win versus another football team B wins is 2:1, which means chances to win are 66%.

Ratio of probabilities can be useful to get rid off values such as P(x), probability of the data, which are usually not known a priori (it is the case only when you know everything about your system, which is rarely the case).
in practical terms, we can compute the ratio between two models θ_1 and θ_2, which will eliminate the P(x) value:

Given $P(\theta_1|x) = \frac{P(x|\theta_1).P(\theta_1)}{P(x)}$ and $P(\theta_2|x) = \frac{P(x|\theta_2).P(\theta_2)}{P(x)}$

Odd ratio OR of the two likelihoods is:

$$OR = \frac{P(\theta_1|x)}{P(\theta_2|x)} = \frac{\frac{P(x|\theta_1).P(\theta_1)}{P(x)}}{\frac{P(x|\theta_2).P(\theta_2)}{P(x)}} = \frac{P(x|\theta_1).P(\theta_1)}{P(x|\theta_2).P(\theta_2)}$$

With all four quantities $P(\theta_1)$, $P(x|\theta_1)$, $P(\theta_2)$ and $P(x|\theta_2)$ usually computable.
One can then derived the probability of both posterior probability can be calculated
As their sum is 1. Ex: odd ratio of 2gives $P(\theta_1|x)$ /$P(\theta_2|x) = P(\theta_1|x)$ / (1- $P(\theta_1|x)$) = 2, so $P(\theta_1|x)$ = 2/3, 66,7% chance to have model 1 (and 33.3 to have model 2).

Posterior probabilities

Exercise:
You have two dices, one is fair (i.e. all numbers have an equal probability), one is loaded and only gives 5 and 6 with associated probabilities 1/2 and 1/2. One picks randomly one of the two dices.

An example of posterior probability is the probability to be using the loaded dice if the outcome of four dice rolls is {6,6,6,6}.

Calculate the prior probability in the example above.
Calculate the odd ratio of the dice to be loaded versus not loaded.
Calculate the probability that the dice to be loaded.

Posterior probabilities

Answer:

- Calculate the prior probability in the example above.

Prior probability is the probability of the model, which is ½ , as there is 50% chance to choose one of the two dices.

- Calculate the odd ratio of the dice to be loaded versus not loaded.

θ_1 being the model "loaded dice" and θ_1 the fair dice.

$$OR = \frac{P(\theta_1|x)}{P(\theta_2|x)} = \frac{\frac{P(x|\theta_1).P(\theta_1)}{P(x)}}{\frac{P(x|\theta_2).P(\theta_2)}{P(x)}} = \frac{P(x|\theta_1).P(\theta_1)}{P(x|\theta_2).P(\theta_2)}$$

$P(x|\theta_1) = (\frac{1}{2})^4$ **(50% chances to have a 6 with loaded dice)**

$P(x|\theta_2) = (\frac{1}{6})^4$;

As $P(\theta_1) = P(\theta_2)$, $OR = \frac{(\frac{1}{2})^4}{(\frac{1}{6})^4} = 74.83$

- Calculate the probability for the dice to be loaded.

$P(\theta_1|x)$ / $P(\theta_2|x) = P(\theta_1|x)$ / (1- $P(\theta_1|x)) = 74.83$

$\rightarrow P(\theta_1|x) = 0.9868$, i.e. dice has 98.68% chance to be loaded.

Application of likelihood: statistical tests

Definition of a statistical test
A statistical test is a procedure that tries to find the likelihood of observing the data x if hypothesis H0 (called "null hypothesis") is true. As a likelihood, this probability is easy to calculate providing that we can describe the parameter of the model (H0). Examples of null hypothesis: Mean of the population is 1,75 m, A gene expression does not change during stress.

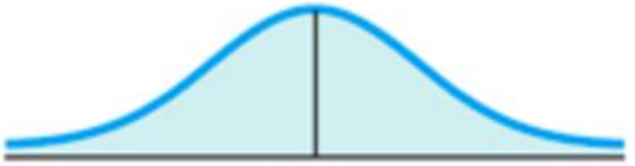
Test statistic
A score derived from the data. Each of its possible values is associated with a likelihood that H0 is true.

p-value
Likelihood of the data, given that model "H0" is true or likelihood $P(x|\theta)$ of having the data *x* knowing model θ (H0).

Level of significance α.
Threshold where you decide that H0 cannot be true anymore. Usually, if there is a likelihood of 5% to observe the data, H0 is rejected to favor alternative hypothesis H1, which can be in an example above that the expression of the gene is changing. Note: usually, we state as H0 what we don't want, and under statistical significance, we accept the hypothesis we think is the best, H1.

Distribution of random variables

Unimodal and symmetrical
Normally distributed variable: “bell-shaped” distribution, one summit (i.e. one mode).

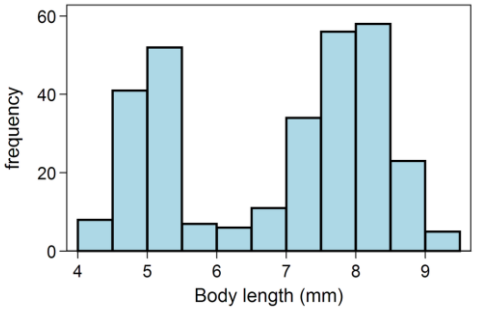


Normally distributed variable

Unimodal and asymmetrical
Chi-square distribution. Represents the distribution of a sum of square of variables.

Chi-square distribution

Bimodal distribution
Usually describes sub-categories of the variable
Example: distribution of ants sizes in a ant nest (nurse vs. fighters).



Bimodal distribution

Normal distribution

Definition

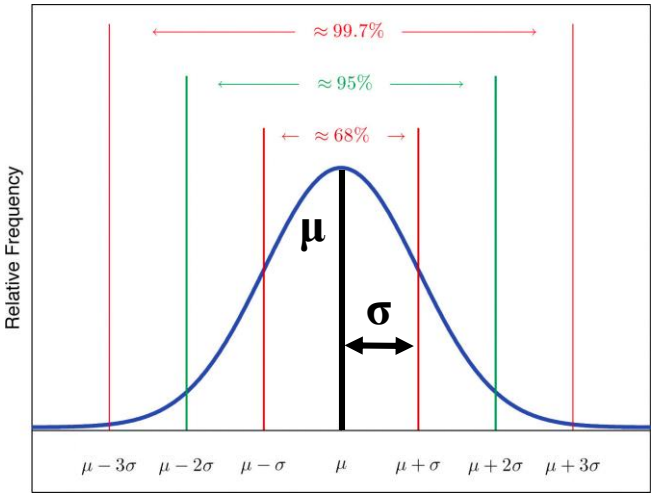
Symmetrical continuous probability density of mean μ and standard deviation σ whose formula is:

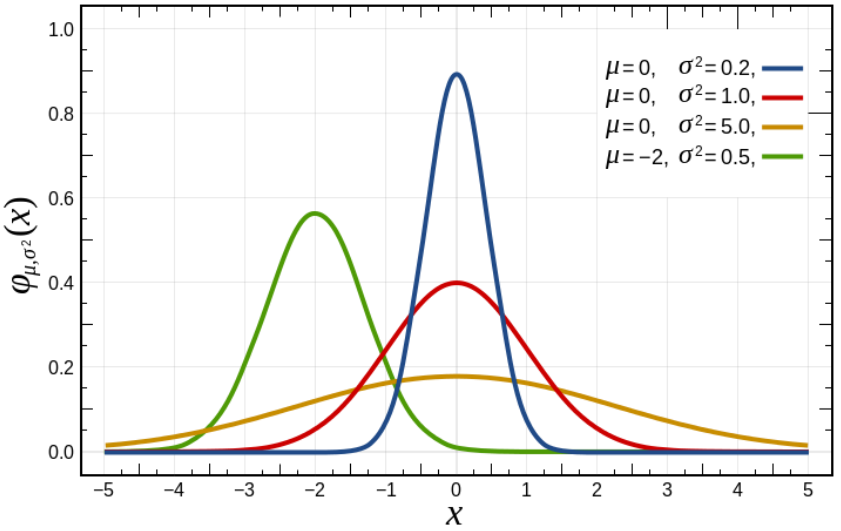
$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Symmetric around the point x = μ, with same quantity of data on each side.
The area under the curve and over the x-axis is 1, as probability distribution.
Density has two inflection points, one per side, located one standard deviation away from the mean, at x = μ − σ and x = μ + σ.

Empirical rule

68% of data falls within the first standard deviation from the mean.
95% fall within two standard deviations.
99.7% fall within three standard deviations.



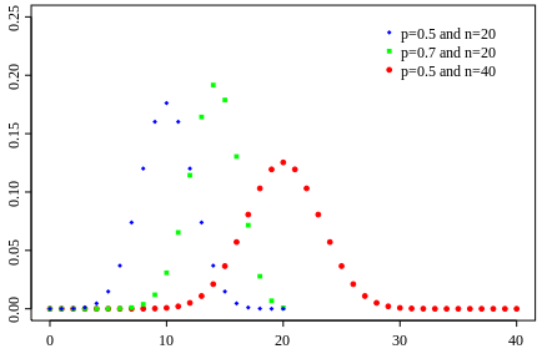


Famous distributions

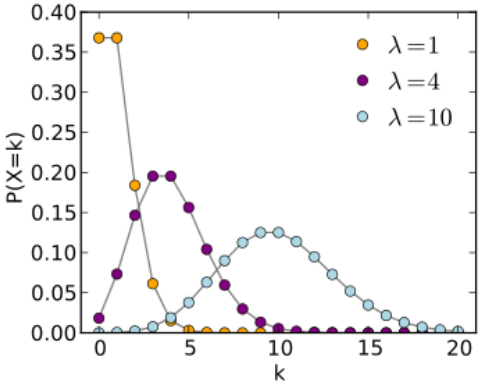
Binomial distribution
Distribution of a binary variable (of type yes/no 1/0).

Poisson
A famous distribution to describe the number of events occurring in space or time in a given interval. One remarkable property is the fact that in this case mean is equal to σ^2.

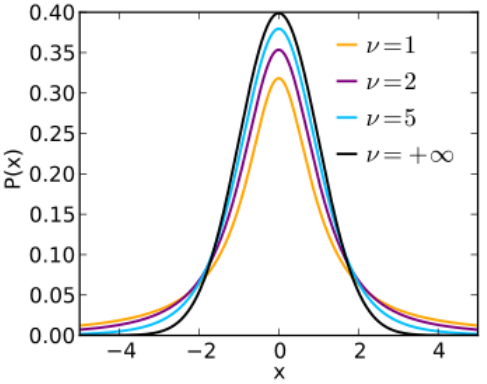
Student
Used to describe a population whose standard deviation is not known. In this case the normal law is not directly applicable.



Binomial



Poisson



Student

Other laws

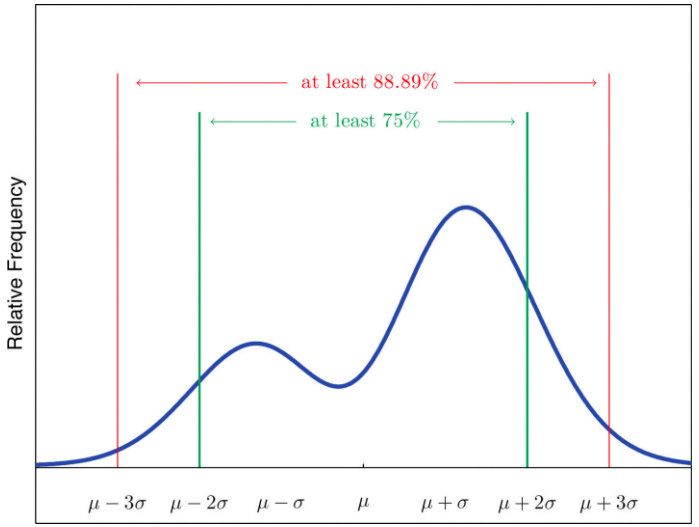
Chebyshev's theorem

The theorem is similar to the empirical rule, but applies to all kinds of distribution, and gives the following rules**:**

75% of the data fall within two standard deviations of the mean
89% of the data (8/9) fall within three standard deviations.
$100(1-1/k^2)$ % of the data fall within k standard deviations.

With general case formulated as $P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$

- Applicable only when μ and σ are available, which is not always the case.
- In case of distributions with values at the extreme, such as the bimodal distributions, where this rule may not apply, other methods can be used, for instance studying the possibility of a Gaussian mixture (the bimodality probably comes from the fact that this observer considers two merged populations of very different biological functions, not one).



Source: https://stats.libretexts.org

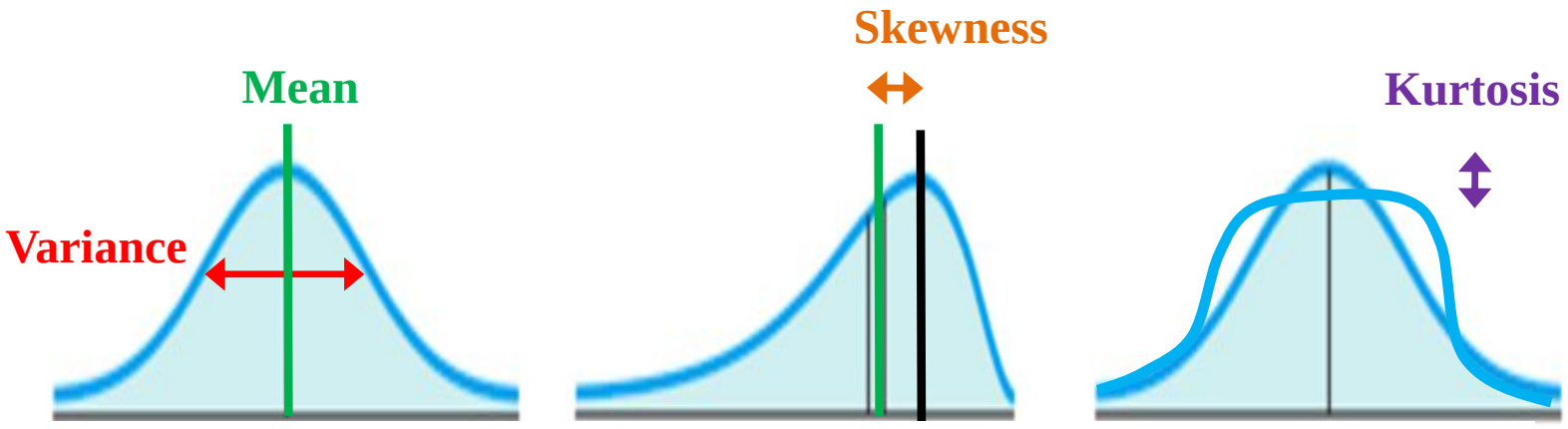
Moments of a distribution

Definition
Moments are quantities which help describe the shape of a distribution.

In the context of probability distribution, moments are:

1st moment: centrality of the distribution (mean)
2nd moment: dispersion of the distribution (variance)
3rd moment: asymmetry of the distribution (skewness)
4th moment: kurtosis or "central flatness"

Normal distribution moments are μ, σ, 0 (symmetry) and 0 (no flatness around mean).



Main summary values for normally distributed data of a population

Measures of central tendency

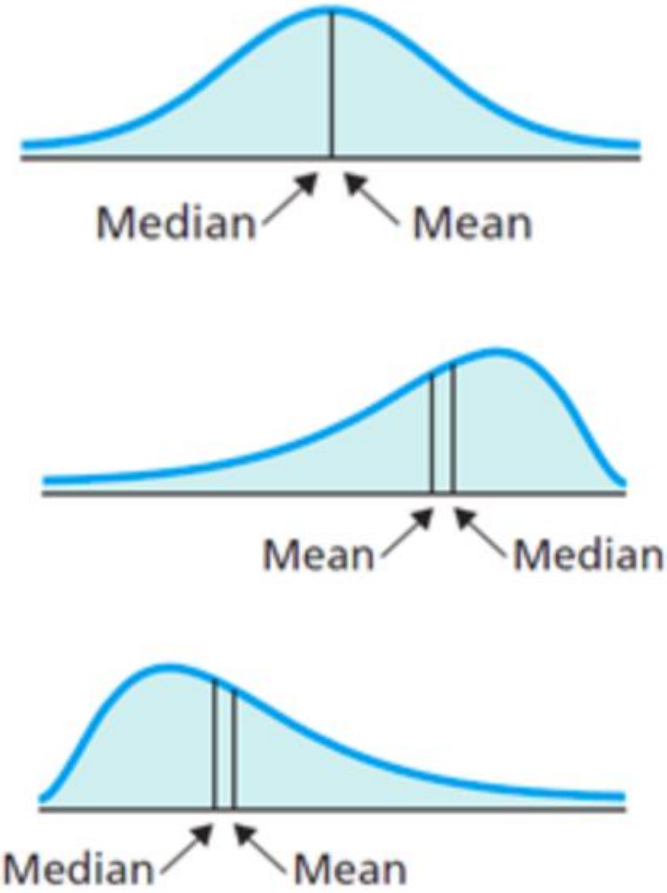
Mean (1st moment):

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Median:

Let a dataset be $(x_1, x_2, \ldots, x_n)$.
- If n is an odd number, median is item $x_{(n+1)/2}$
- If n is an even number, median is $\frac{x_{n/2} + x_{n/2+1}}{2}$

Exercise.
What is the mean and the variance of the values of a regular dice?

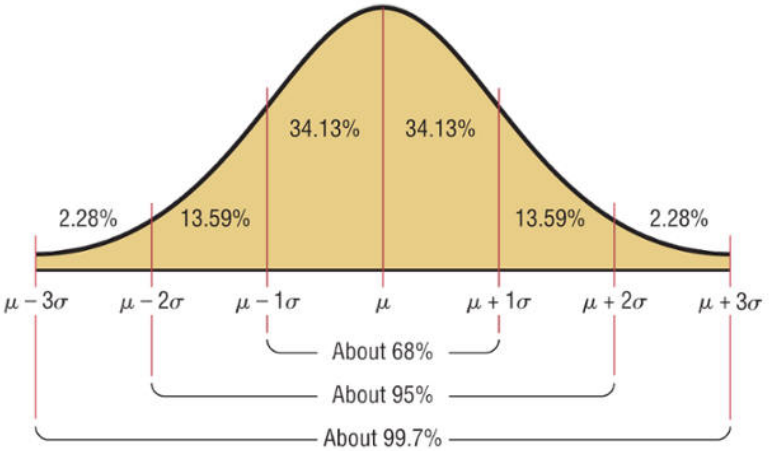


Main summary values for normally distributed data of a population

Measures of Dispersion

Standard deviation (2nd moment):

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$



Quartiles:

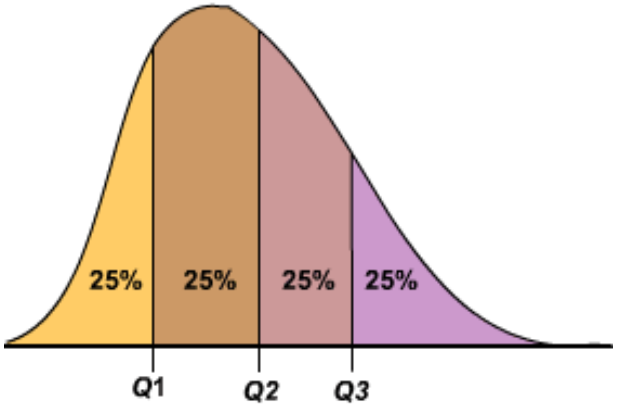
First quartile (Q_1): value that divide the data into the 25% lowest values and 75% highest.

Third quartile (Q_3): value that divide the data into the 75% lowest values and 25% highest.

Second quartile is the median. (divides set into two parts).

Exercise.

What is the standard deviation and the first and third quartiles of the values of a regular dice?



Test for normality

- Null hypothesis H0: “The distribution of the data is following a normal distribution”.
- Important for certain procedure in statistics, for instance comparing two populations and ask whether they are similar or not.

Shapiro-Wilk test

Test statistics:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

With :

$\bar{x}$ the mean of the dataset ;

x_i the observed values ;

a_i the associated expected values (if the dataset was following a normal distribution).

Associated p-value to W is the likelihood that the distribution is normal; if it is below 0.05, H0 is rejected and distribution is considered non-normal.

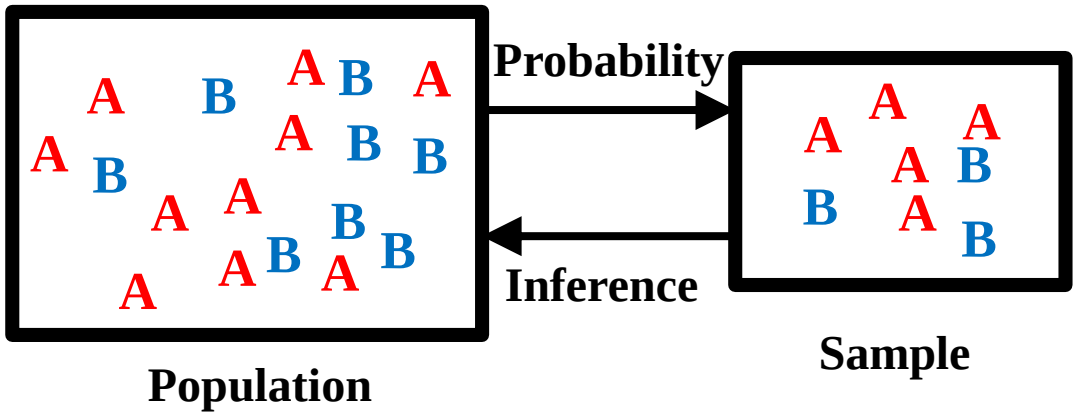
Sampling

Definition: A sample is a collection of data selected from a population.

Examples:

A “direct” sample for biological measure. Hormone measurement, abundance of a species, etc.
Replicates for an experiment.
Sample of human population in case of a medical treatment

- **Descriptive biostatistics**: only the metrics of the sample are reported. This approach is usually in all cases, to present the general characteristics of a sample (number of individuals, mean, etc.)
- **Inference biostatistics**: From a sample of population, derive rules about the entire population. Important approach in biomedical research.



Parameters estimation

Biological model

Any entity that is defined by parameters: a dice (parameters are the probabilities to get a 1, 2, 3, etc.), a genotype (parameters are the sequences at different positions), a medical condition (a parameter could be the blood pressure level).

Parameters can be the mean or a variance of a distribution, for instance the distribution of blood pressure in the general population.

Distinction between the real parameter of the model and the estimation we get from the sampling.

An **estimator** is the value which is predicted to be the closest to the real parameter θ is the value that minimizes the error:

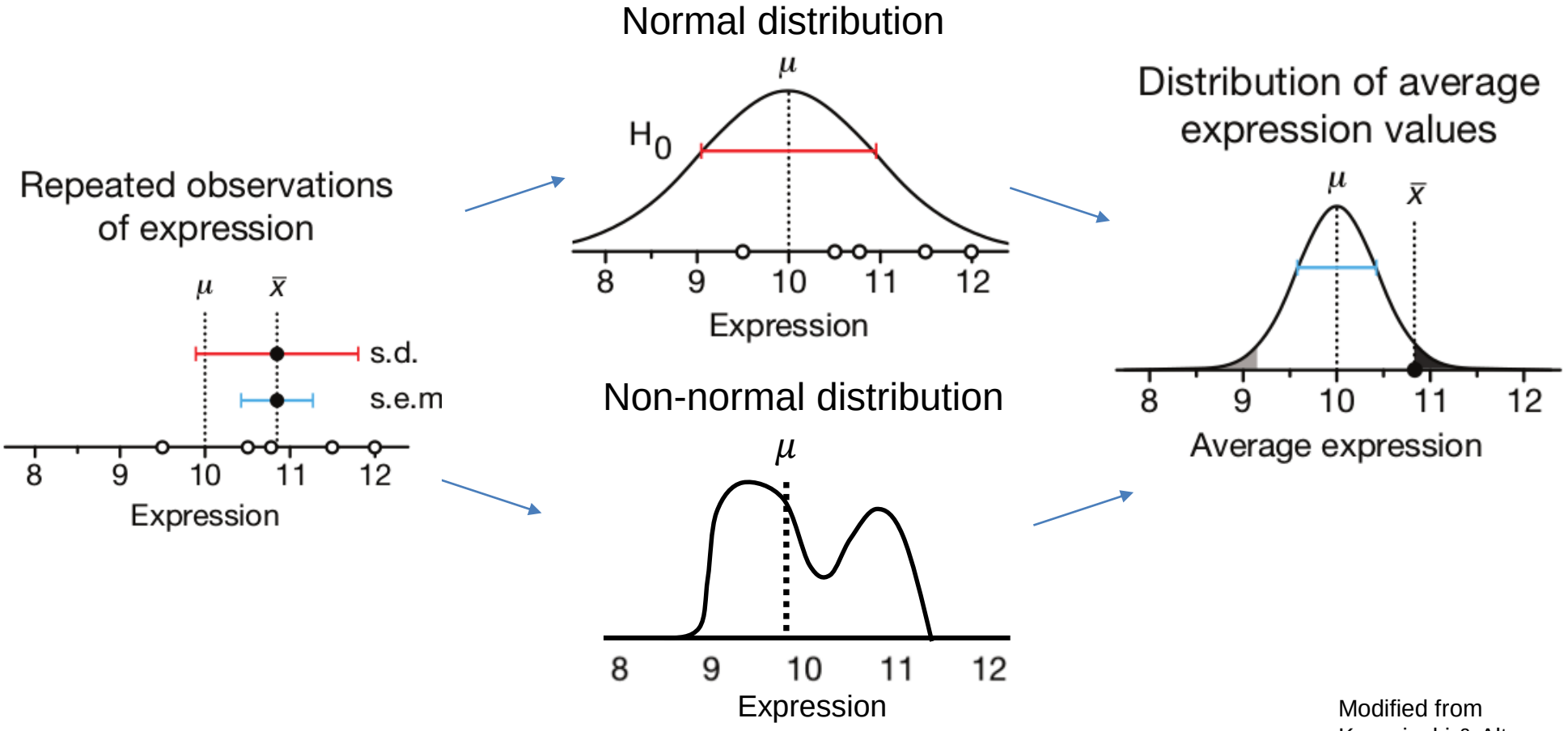
$$e = \hat{\theta} - \theta$$

θ being the real value of the parameter. e is the error or bias and needs to be as small as possible, i.e. real mean and estimator should be as similar as possible.

Limit central theorem

Definition:

Predicts that the average value of a sampling strategy is always following a normal law, whatever the shape of the initial distribution. As a result, the sampling is compared to the distribution of means of possible sampling results, not to the initial distribution.



Modified from Krzywinski & Altman

Estimators of the mean

Definition of an estimator:

Estimators or parameter values of a population are based on the likelihood to see the data in the sample under model “there is no difference between parameter of the sample and parameter of the population” (for instance the mean).

Mean of the population:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1+x_2+\cdots+x_n}{n}$$

Estimator x of the mean of the population:

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

Estimators of the standard deviation

Standard deviation σ of a population (theoretical of when you know all the values of the population):

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$$

Unbiased estimator s of the standard deviation σ of the population:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-m)^2}{n-1}}$$

Division in this case is by n-1: the number of degree of freedom (i.e. minimal number of variables to explain all the variable dependences) is n-1 as the sum of residuals, i.e. $(x_i - m)^2$ is equal to 0. In other words, if you know n-1 variables, you know the last one as the sum is 0, so you have only n-1 degrees of freedom.

Confidence interval and precision

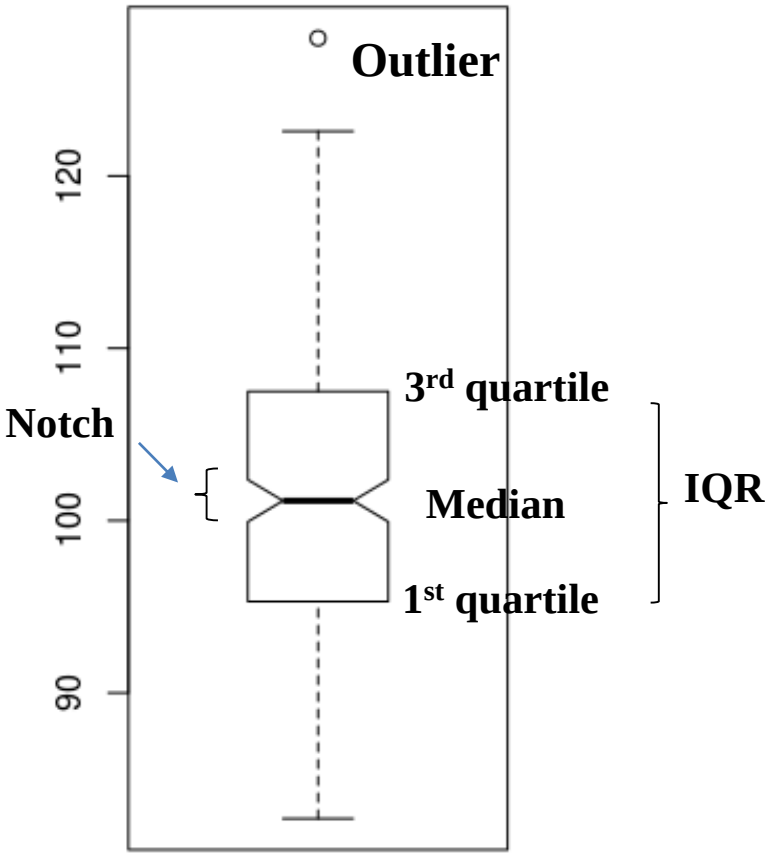
Data representation

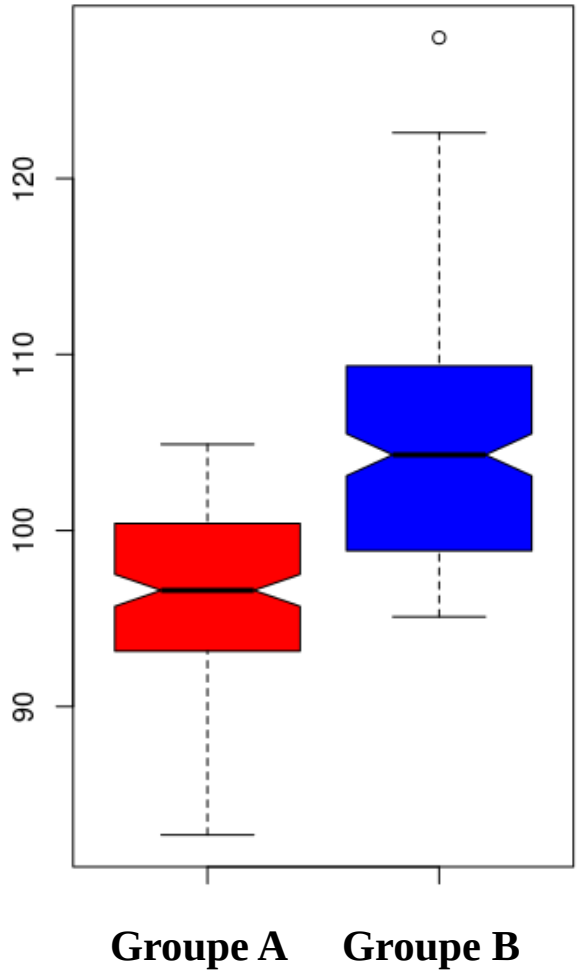
Boxplots

Interquartile (IQR): region between the 1st and 3rd quartiles, contains 50% of data.

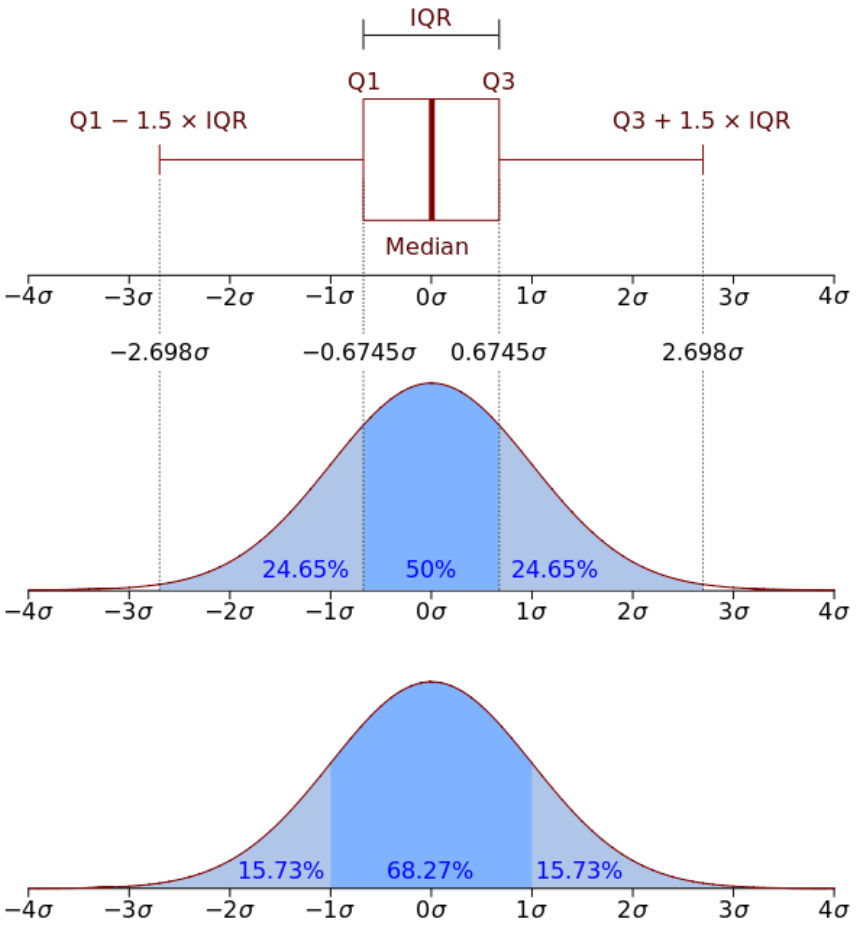
Notch: 95% confidence interval of the median (Median +/- 1.57 x IQR/$n^{0.5}$). If the notches of two distributions do not overlap, they are significantly different.

Example below: the red and blue distribution overlap but means are significantly as the notches do not overlap.



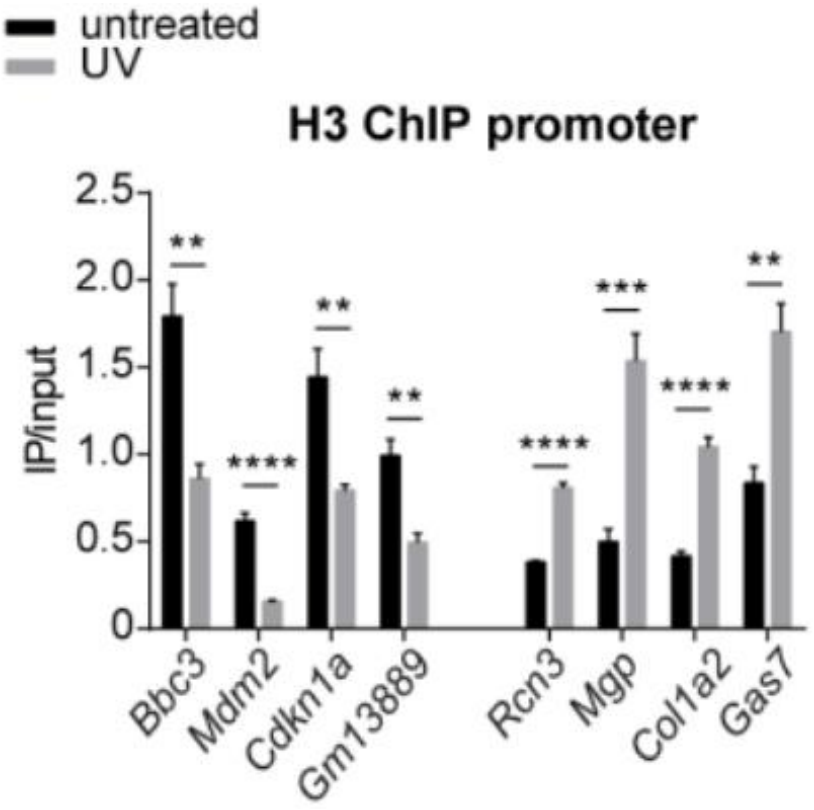


Boxplots and distribution

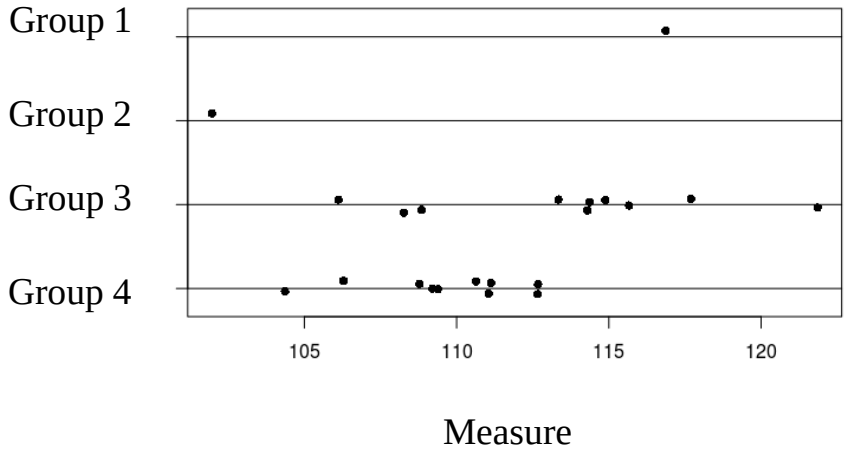


Relation between boxplot and distribution

Other representations



Barplots



Stripchart

What to choose for what?

Some rules:

If you wan to focus on the mean, do a barplot

To focus on the extend of a distribution, do a boxplot.

If you want to focus on outliers, do a stripchart or a 2D plot.

If you want to do clustering, do a 2D/3D plot, or a stripchart if data are to be represented in one dimension.

Student's t-test

Definition:
Tests the null hypothesis: "are two normal distributions identical?". Ex: comparison of blood pressure of two groups, disease and control.

Statistics:
Calculates a distance between the two means of distributions 1 and 2.

$t = \frac{\overline{X_1} - \overline{X_2}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, with $s_p = \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}}$ and

$\overline{X_1}$, $\overline{X_1}$means of the samples, s_{X_1}, $s_{X_1}$the standard deviations, n_1, n_2 the means.

Interpretation of the p-value:

P-value associated with the statistics gives the probability that H0 is true, so the probability for the two distributions to be similar.

Prior:
- The two samples have to follow Normal laws (to be tested with the Shapiro-Wilk test.
- The number of individuals in each sample has to reach a minimum for the test to be applicable.

Experimental design

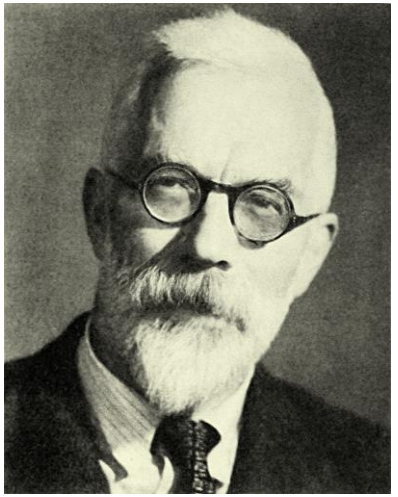
Comparison/control

Replication

Randomization

Blocking

Factorial experiments



Ronald Fisher

The Arrangement of Field Experiments (1926) The Design of Experiments (1935)

Comparison condition/control

Definition: A scientific control is an experiment or observation designed to minimize the effects of variables in order to eliminate alternative explanation.

Negative control. A group where no phenomenon is observed.
Ex: A patient does receives a placebo.

Positive control. A group associated with an expected outcome.
Ex: A group with a given blood pressure or a given hormone concentration.

Blind experiment. The operator does not know what case is the condition.

Double-blind experiment. In biomedical research, neither patient or doctor know whether the medicine or the placebo is being given.

Sham groups. A control group in the context of an experiment requiring surgery.

Comparison condition/control

Principle
Most of experiments do not have a standard of measurement so one needs a reference for comparison. A treatment or an intervention is operated on a sample, while the other gets a mock treatment.

Examples:

Comparison of systolic blood pressure between two groups. One group is given a treatment and the other one a placebo.

In a molecular biology experiment, one cell line gets a vector for a mutated gene, while the other cell line gets an empty vector.

Note:
The control group and the group with the condition can be the same.

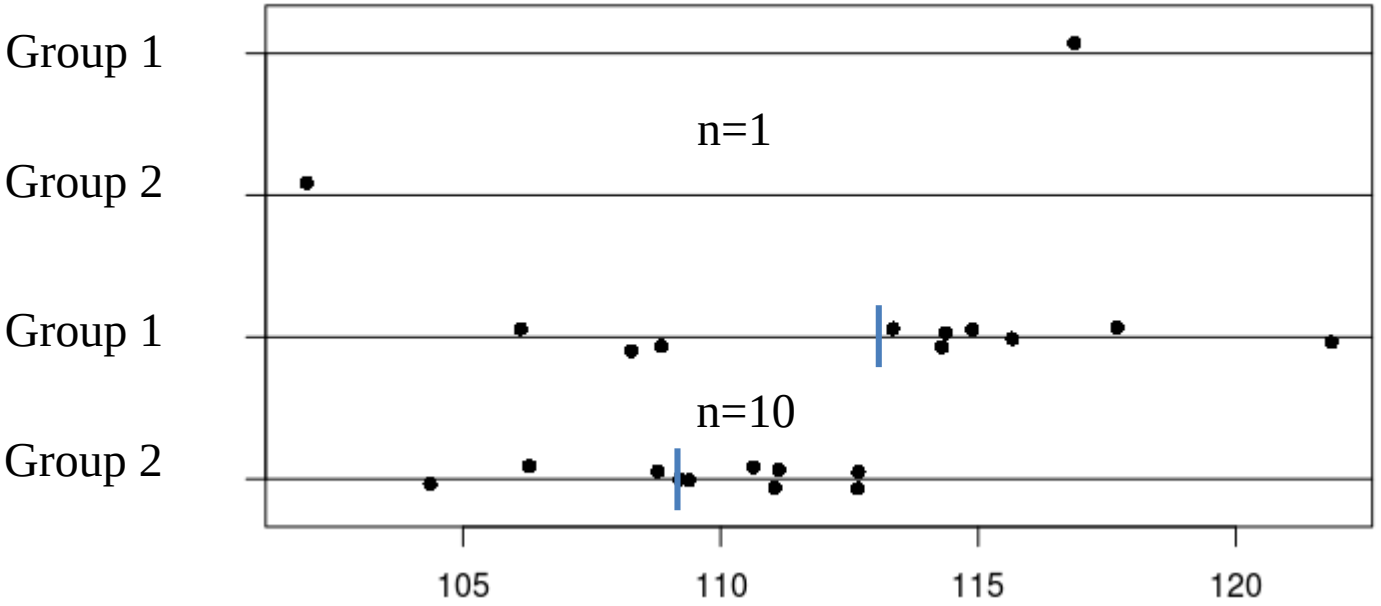
Replication

Goals:

Is an experiment reproducible?

Reduce the effect of unpredicted variation

Approach the true mean of the population



Replication

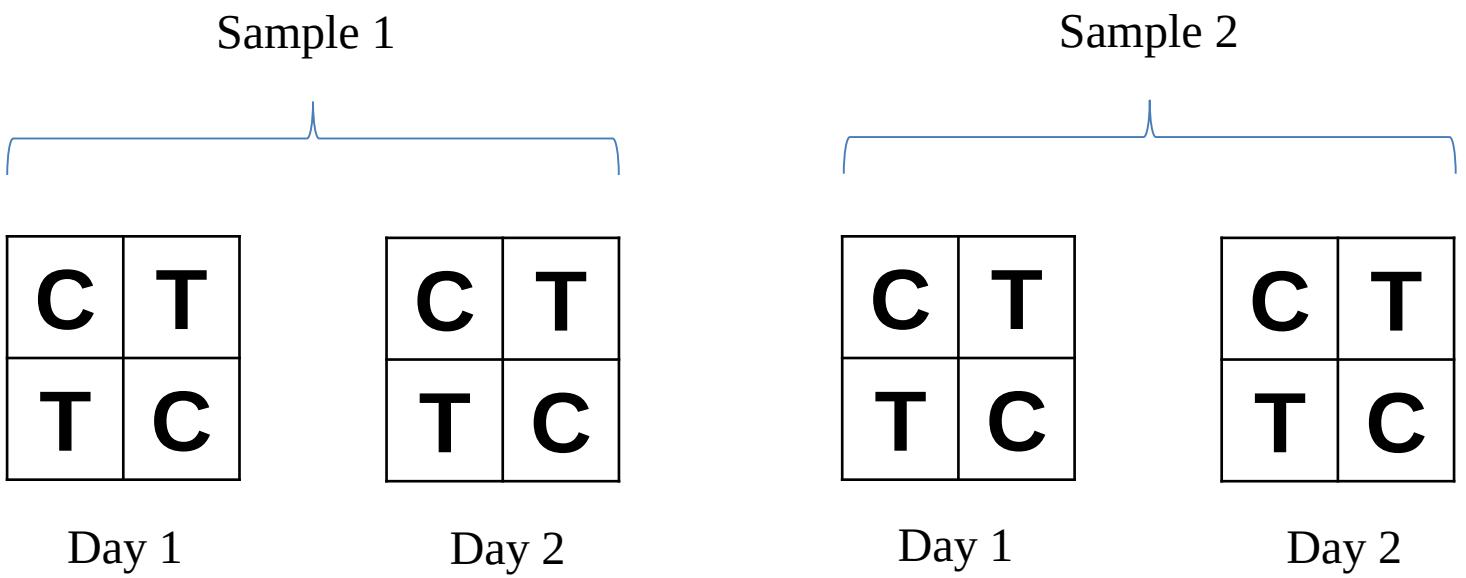
Technical replicate: Measure variability coming from performing experiments. repeating operation on same sample will give a technical replicate. Usually if no variation detected, technical replicates are pooled.

Biological replicate: Measure variability coming from the biological properties of the system you are observing. This is the only replication that gives the number n of replicates in your experiment.

Experimental replicate: Replication at different times, in “batches”. In this case groups of control/condition are done simultaneously, at different times, each time being a replicate.

Other kinds of replication:
Experimenter replicate
Method replicate
Laboratory replicate

Replication

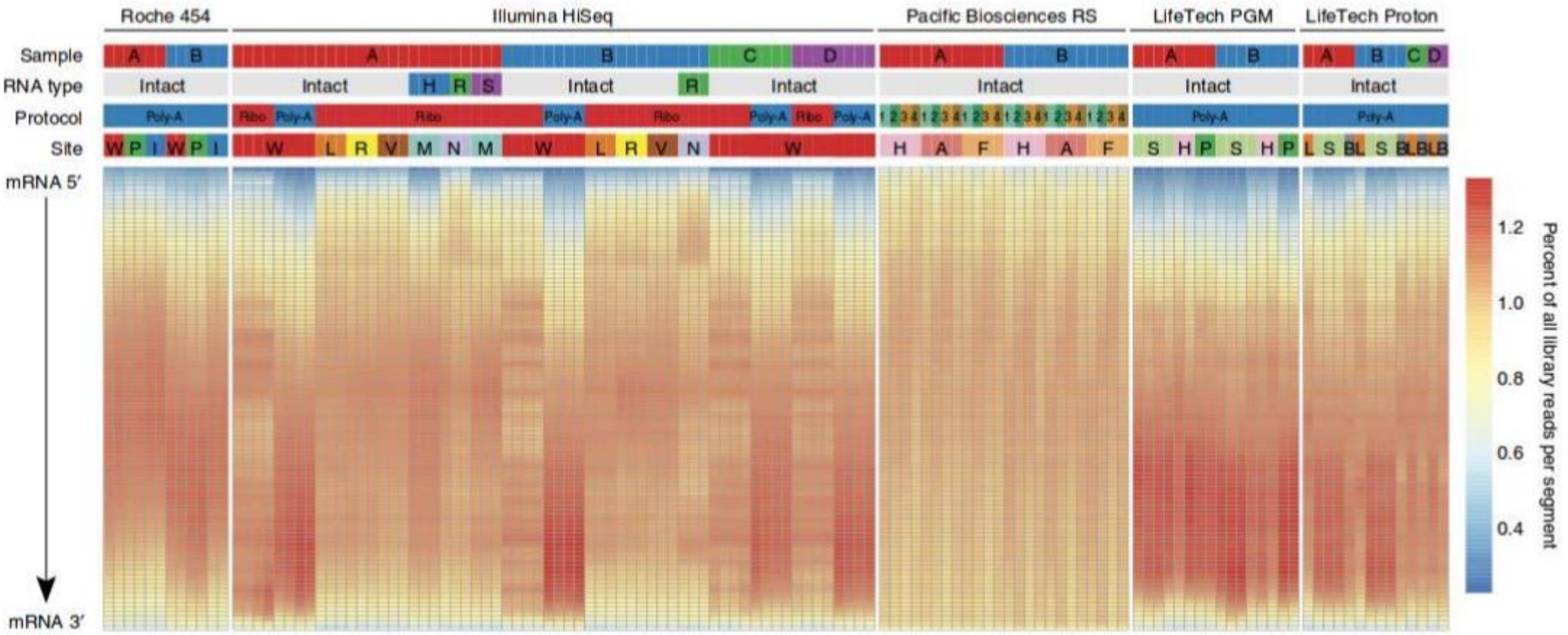


C: Control
T: Treatment

Replication

Exercise.

Identify the different kinds of replication exemplified in the below figure.



Li et al. 2014, Nat biotech.

Randomization

Goals:

- Reduce confounding effects.
- Allow to perform statistical analysis on the data.

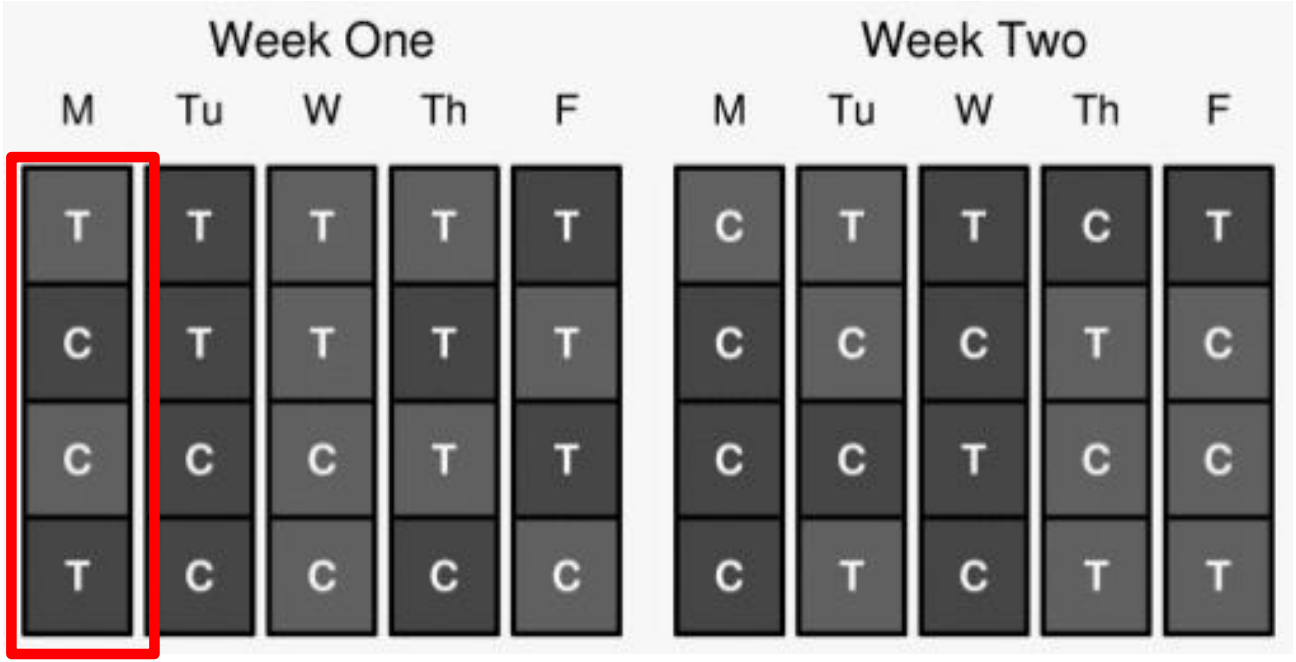
In biology: the treatment and the absence of treatment are distributed
randomly between different samples. This usually requires to associate
identifiers instead of explicit labels.

In medicine: a patient is given randomly the treatment or the placebo.

Randomization

Goal:

Do not find differences between samples because of external influences (for instance, the difference you see may not come from treatment but from the time of the day when your performed the experiments).

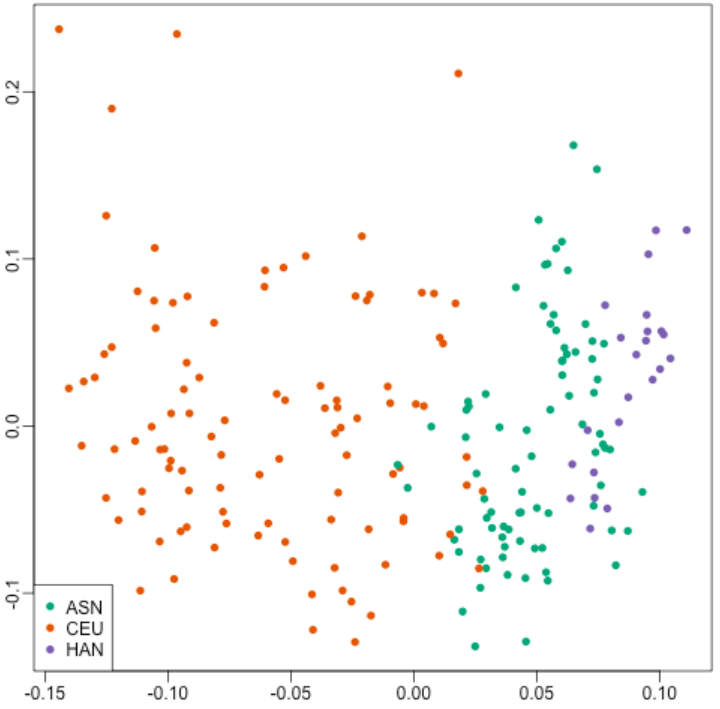


C: control **Dark grey: Male**

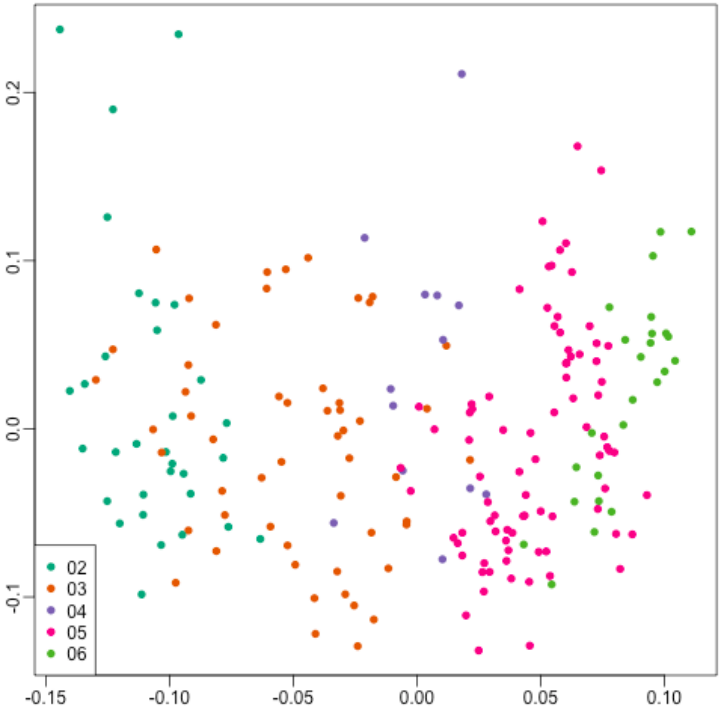
T: treatment **Light grey: Female**

Batch effect

Annotation according to condition



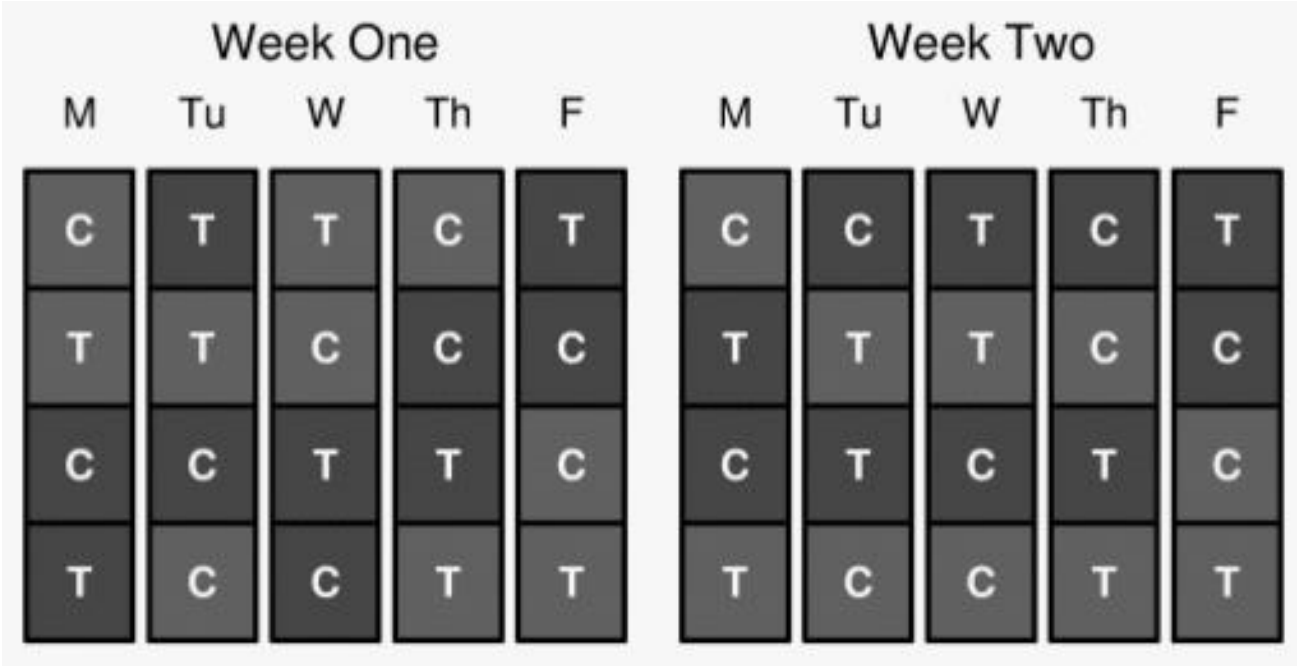
Annotation according to batch



Blocking (stratification)

Definition of a block:

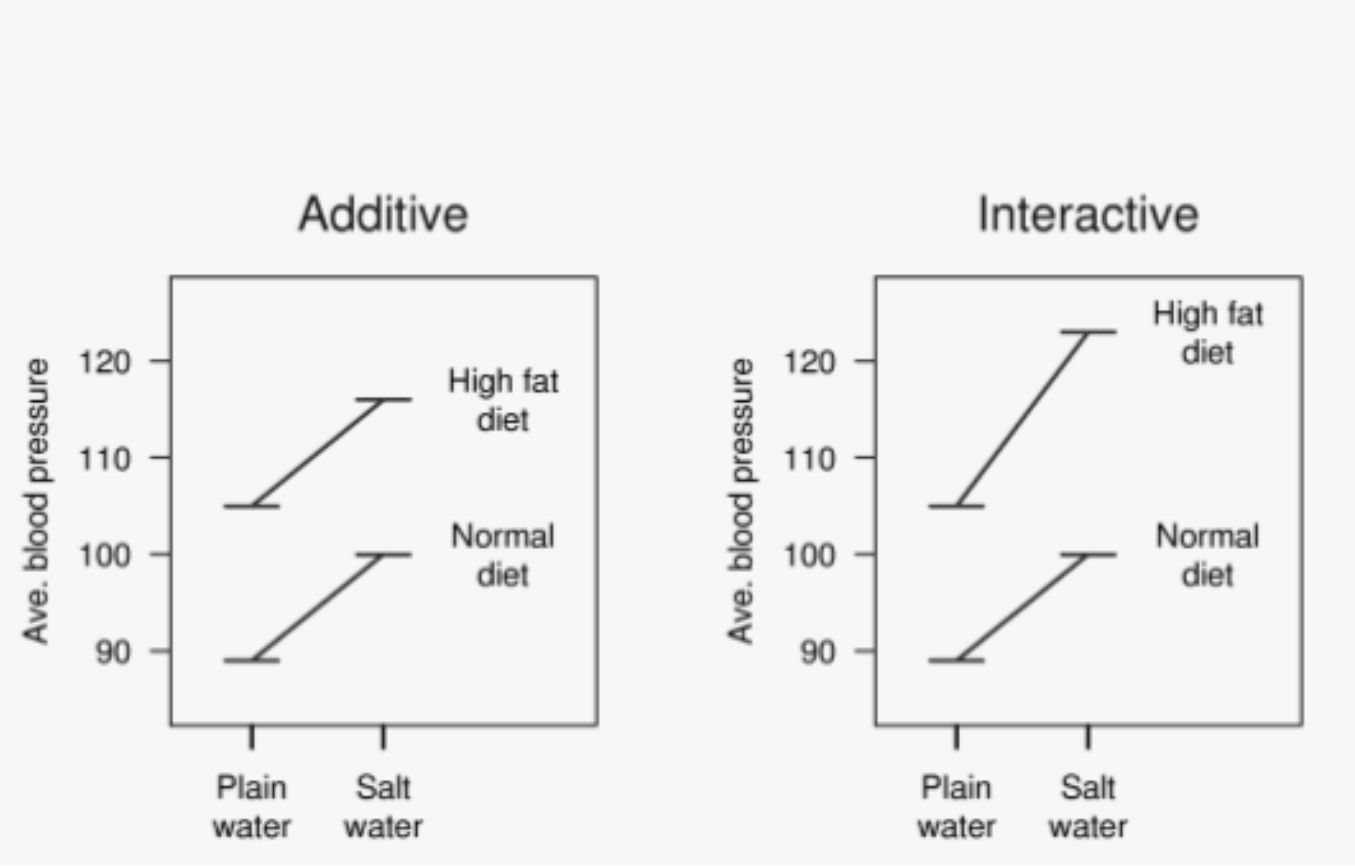
Captures the possible influence of the second factor (gender) to improve precision.



C: control **Dark grey: Male**

T: treatment **Light grey: Female**

Blocking (stratification)



Interaction between variables

Summary

Characteristics of good experiments:

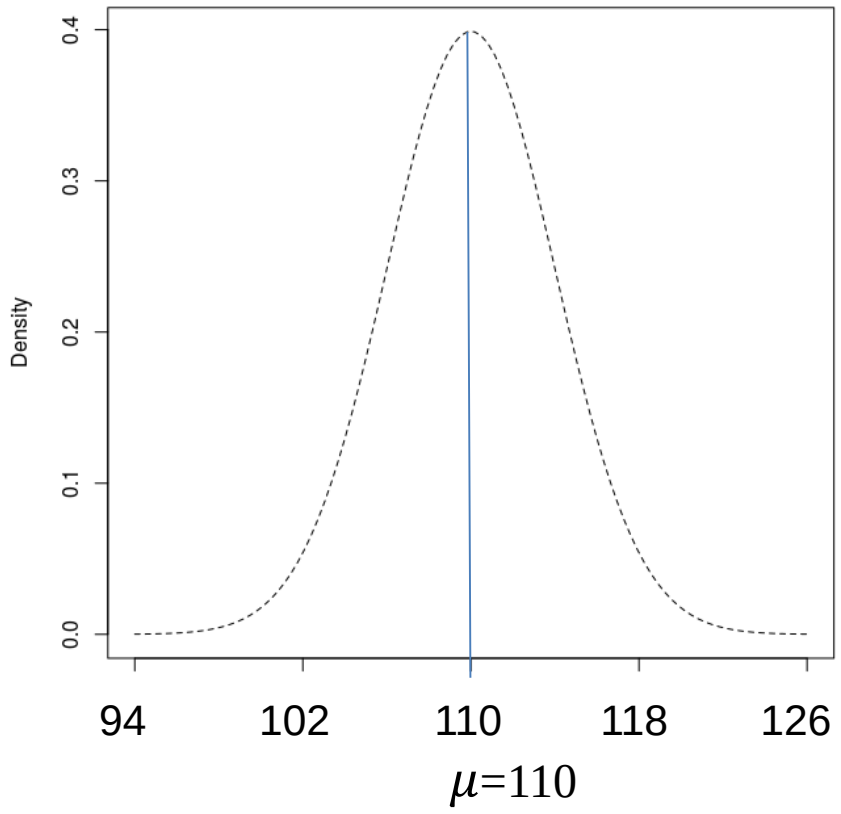
• Unbiased (i.e. distribution of sample centered on distribution of population)
– Randomization
– Blinding

• High precision (value of statistics is close to the mean of the population)
– Uniform material
– Replication
– Randomization
– Blocking

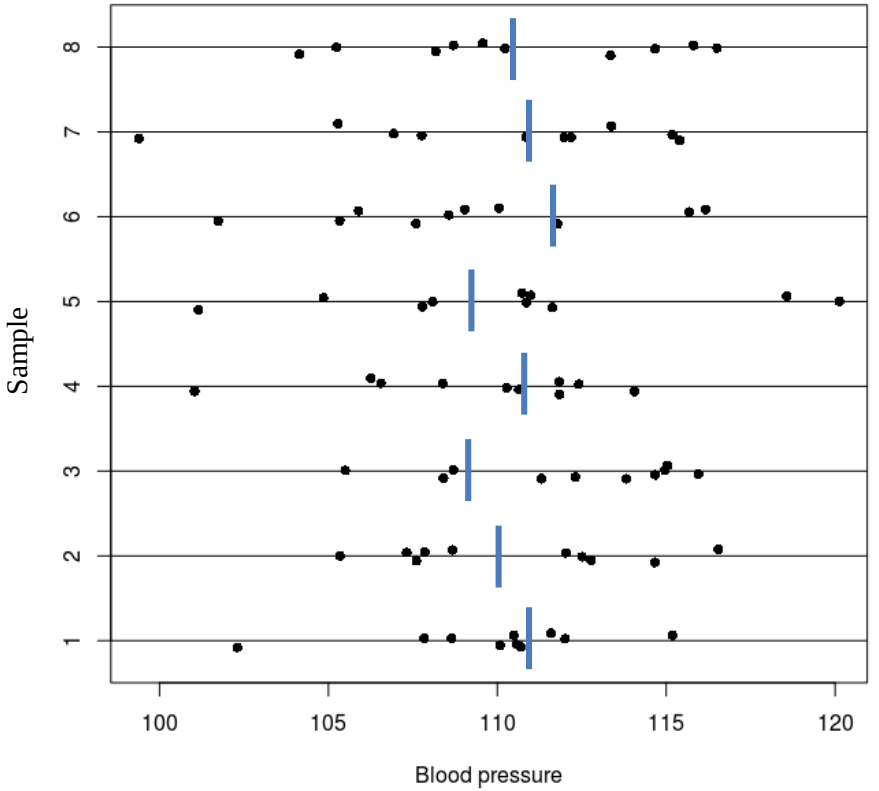
• Study variables interaction
– Replication
– Randomization + blocking
– Factorial designs

• Simple

Sampling

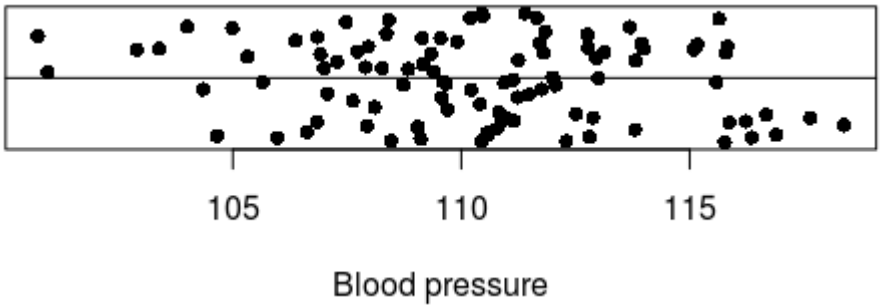


Population
(Real or hypothetical)

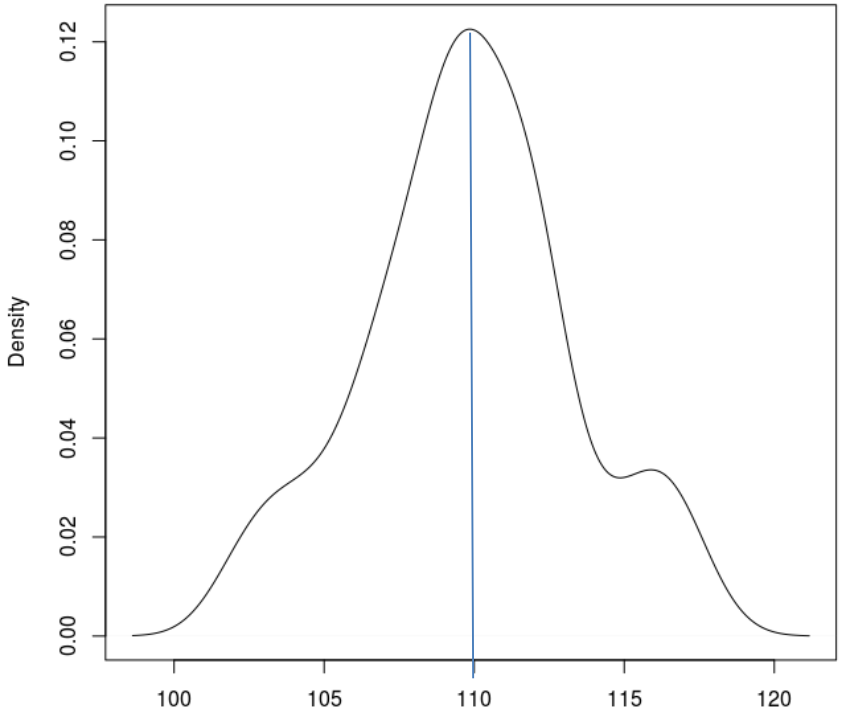


Different sampling

Estimation of the mean



Result of sampling



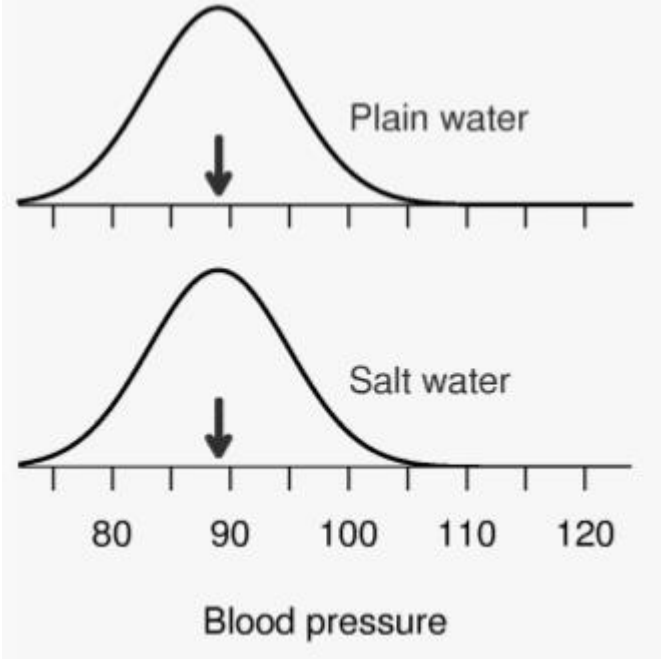
$\hat{\mu}$=109.8317

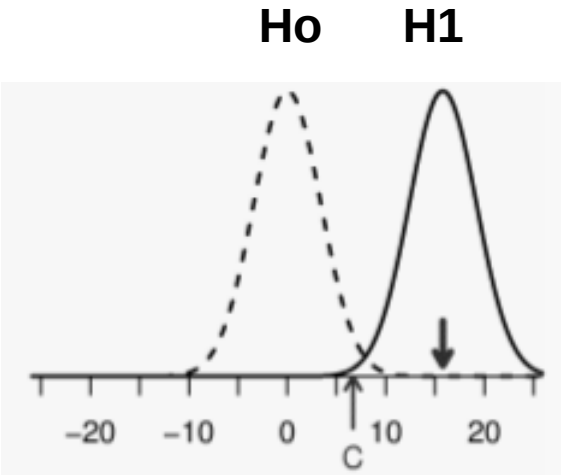
Sampling distribution
(Real or hypothetical)

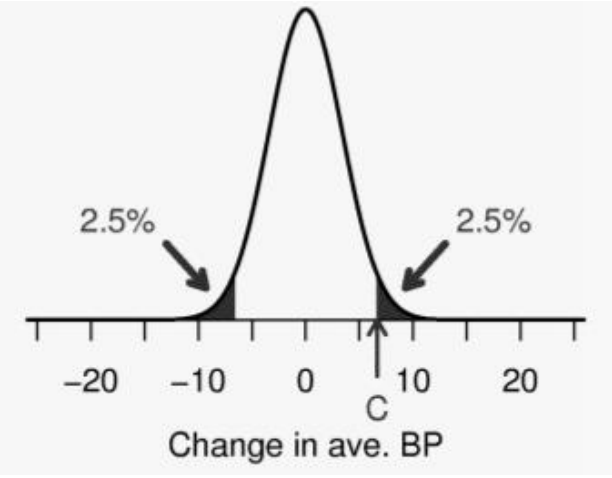
t-test applied to experimental data

Goals:

- Distribution of differences between the two distribution of values should be normal and centered on 0
- Null hypothesis H_0: no difference
- If diff > C, we reject H_0
- C chosen so the chance to reject H_0, if $H_0$is true, is 5%.







Types of errors

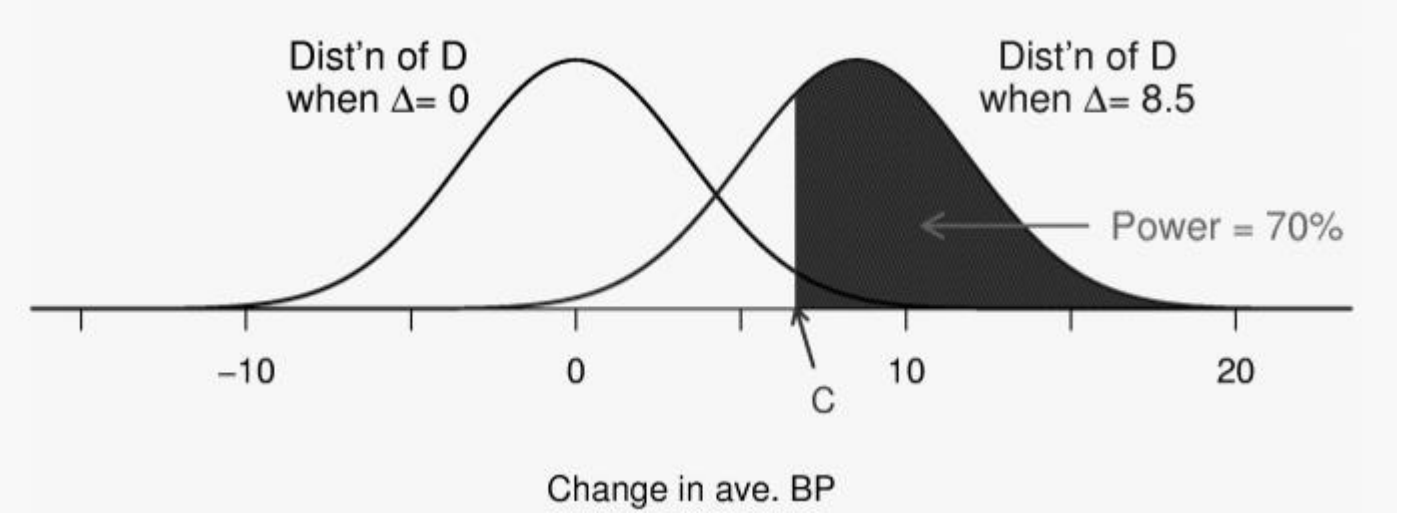
Type I error (“false positive”)
Conclude that salt water has an effect on BP when, in fact, it does not have an effect.

Type II error (“false negative”)
Fail to demonstrate the effect of salt water when salt water really does have an effect on BP.

	The truth	
Conclusion	No effect	Has an effect
Reject H_0	Type I error	✓
Fail to reject H_0	✓	Type II error

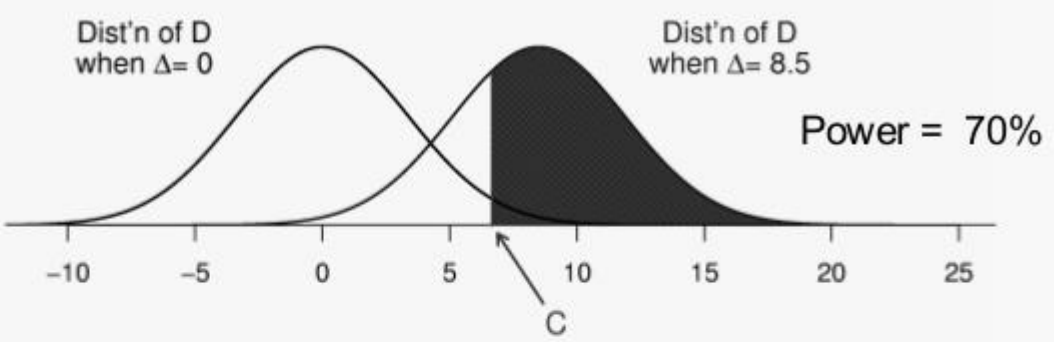
Statistical power

Definition:
Chance to reject H_0 when H_0 is false, or predict correctly that treatment has an effect.

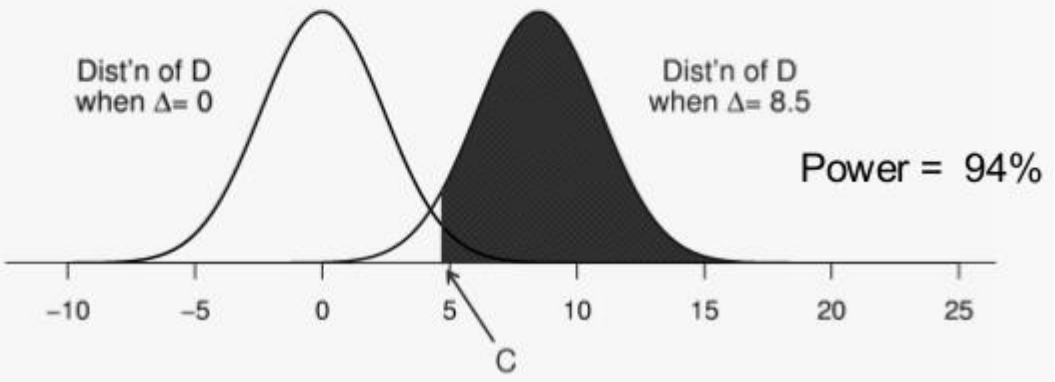


Power and sample size

6 per group:



12 per group:



Images copyrights

Slide 5: top: Kirti Prakash, 2012. ;
bottom: https://www.datacamp.com/community/tutorials/machine-learning-in-r#gs.vZWccEQ

Slide 30: example from the Scientific method in brief. Chapter 9 on Bayesian inference.

Krzywinski & Altman, Nat. methods. 2013. See also Krzywinski, M. & Altman, N. Nat. Methods 10, 809–810 (2013). On sampling