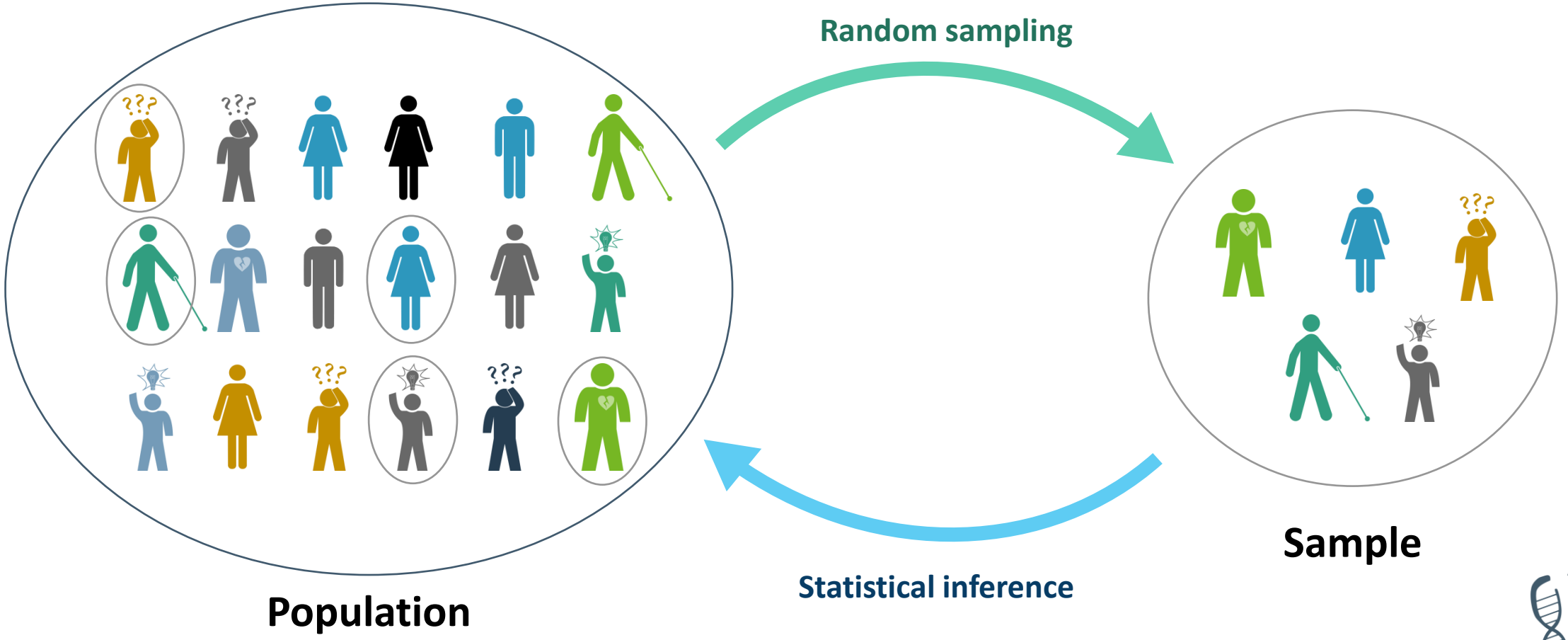
Introduction to biostatistics (with Python)

Urszula Smyczyńska

NGSeminars

Source: https://www.augusta.edu/mcg/dphs/bds/research/

Why statistics?



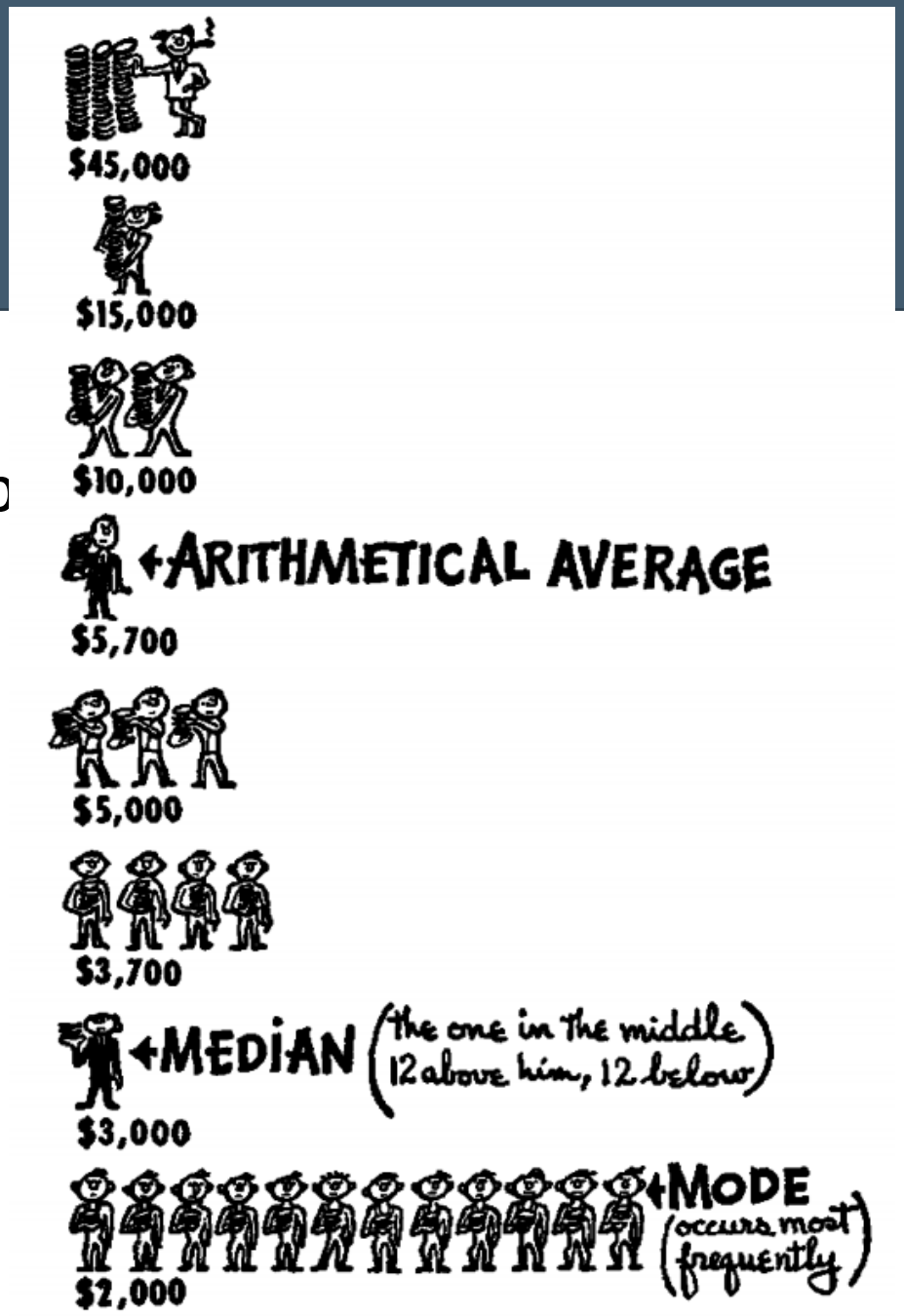
Variables

- Nominal (categorical/discrete/qualitative variables) e.g.:
  - sex
  - presence/absence of mutation in gene
  - occurrence of disease
- Ranked (ordinal variables) e.g.:
  - tumor stage
  - Apgar score
- Continuous (numeric/quantitative variables) e.g.:
  - age
  - height
  - concentrations
  - gene expression

NGSeminars

Central tendency

- Mean (arithmetic, geometric …).
- Median – value “in the middle” of samp
- Mode – most frequent value.



Source of image: Huff D., *How to lie with statistics?*

Means

Arithmetic mean (AM)

$$\bar{x}_{AM} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Geometric mean (GM)

$$\bar{x}_{GM} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Harmonic mean (HM)

$$\bar{x}_{HM} = \frac{n}{\left(\sum_{i=1}^{n} \frac{1}{x_i}\right)}$$

$$AM > GM > HM$$

More on geometric mean

How to calculate geometric mean without all that multiplication?

$$\bar{x}_{GM} = \left((-1)^m\right)^{\frac{1}{n}} \cdot \exp\left(\frac{1}{n}\sum_{i=1}^{n} \ln|x_i|\right)$$

m – numer of negative values x_i

When to use it? Whenever you have normalized measurements!
Let's see why.

$$\mathrm{GM}\left(\frac{Y_i}{X_i}\right) = \frac{GM(Y_i)}{GM(X_i)}$$

If both were normalized by Z_i before averaging:

$$\frac{GM(Y_i/Z_i)}{GM(X_i/Z_i)} = \frac{\frac{GM(Y_i)}{GM(Z_i)}}{\frac{GM(X_i)}{GM(Z_i)}} = \frac{GM(Y_i)}{GM(X_i)}$$

NGSeminars

Variability measures

- Variance
- Standard deviation
- Standard error of the mean
- Range
- Interquartile range

Variance and standard deviation

Variance

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i)^2 - \bar{x}^2$$

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

Standard deviation

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

POPULATION

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

SAMPLE

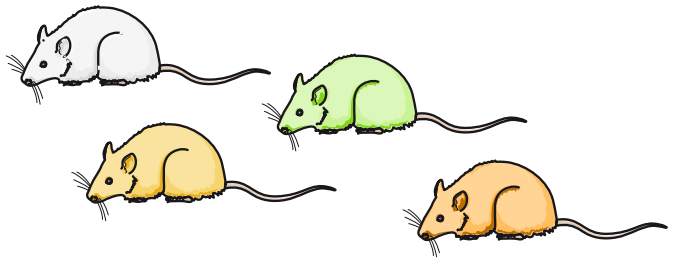
NGSeminars

Standard deviation or standard error?

Standard deviation

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

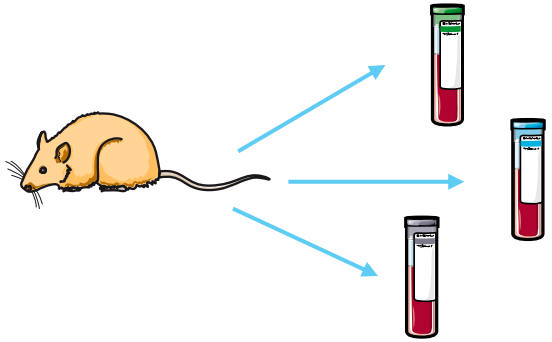
BIOLOGICAL REPLICATES



Standard error

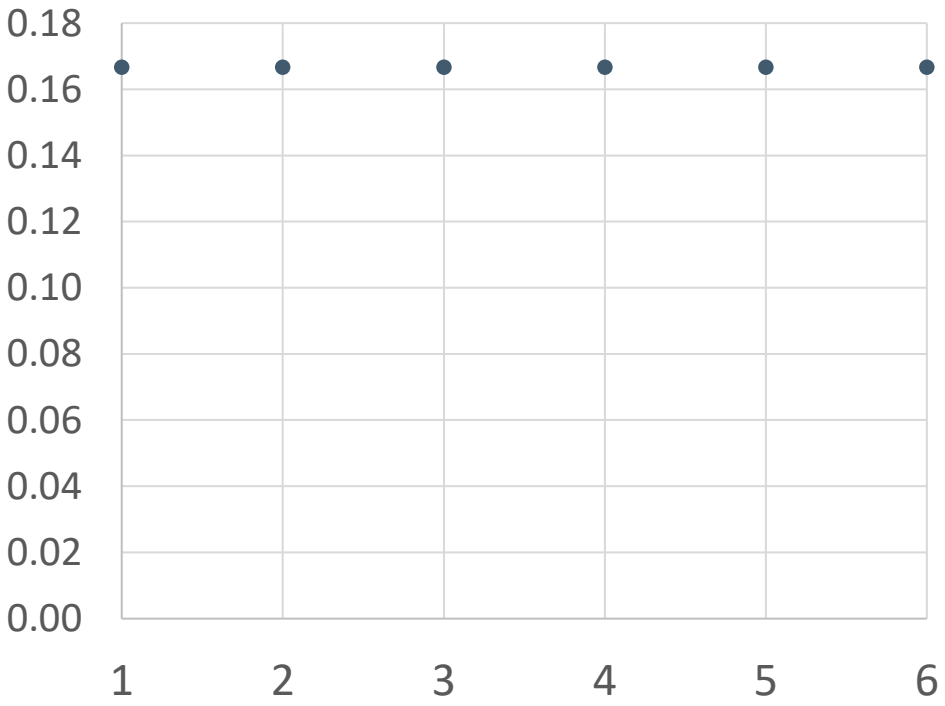
$$SE = \frac{\sigma}{\sqrt{n}}$$

TECHNICAL REPLICATES



Images from: smart.servier.com

Distributions



Images from: wikipedia.org & unsplash.com

Distribution

Any function that fulfils following condition can be a probability distribution:

Continuous

$$\int_{-\propto}^{\propto} P(x)dx = 1$$

Discrete

$$\sum_{x \in X} P(x) = 1$$

X – set of all possible events ,
e.g. for dice roll X = {1, 2, 3, 4, 5, 6}

We also define cumulative distribution function as:

Continuous

$$F(z) = \int_{-\propto}^{z} P(x)dx$$

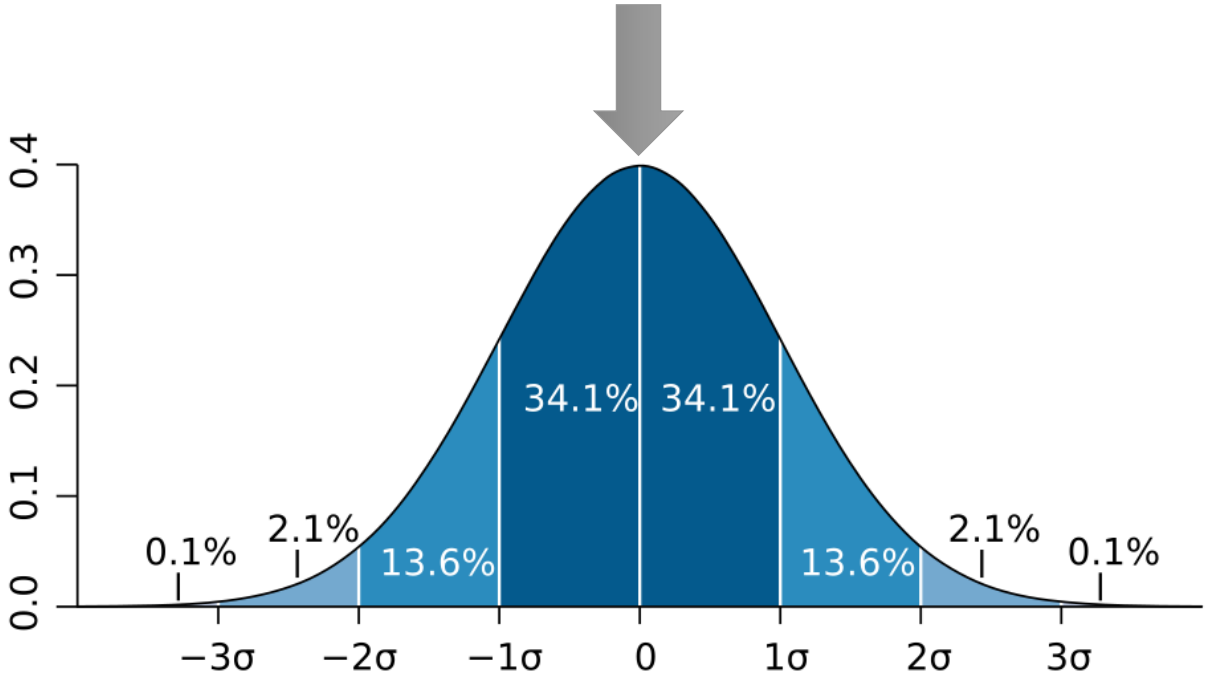
Discrete

$$F(z) = \sum_{x<z} P(x)$$

NGSeminars

Normal distribution

mean = median = mode



Probability density function:

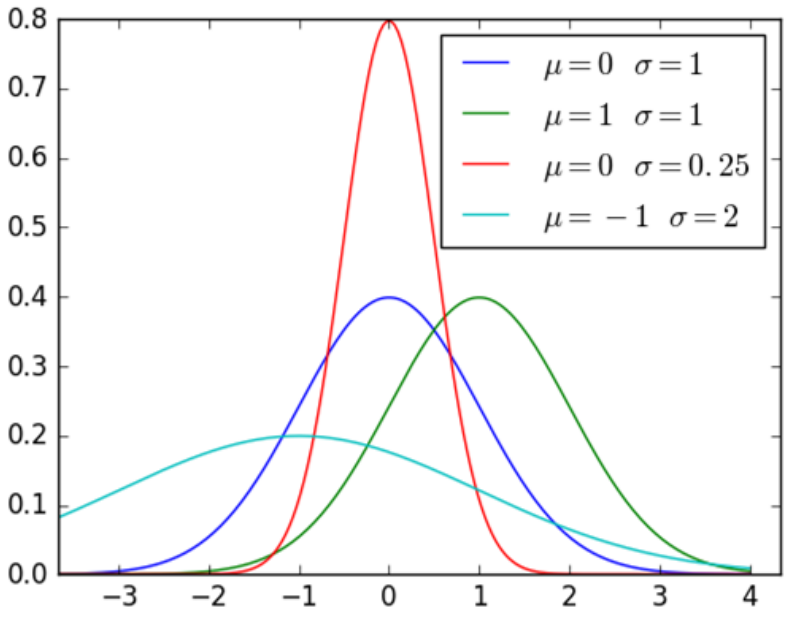
$$P(x) = \frac{1}{\sigma\sqrt{(2\pi)}} \cdot \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

Looks complicated but is fully described by only 2 parameters: mean (μ) and standard deviation (σ).

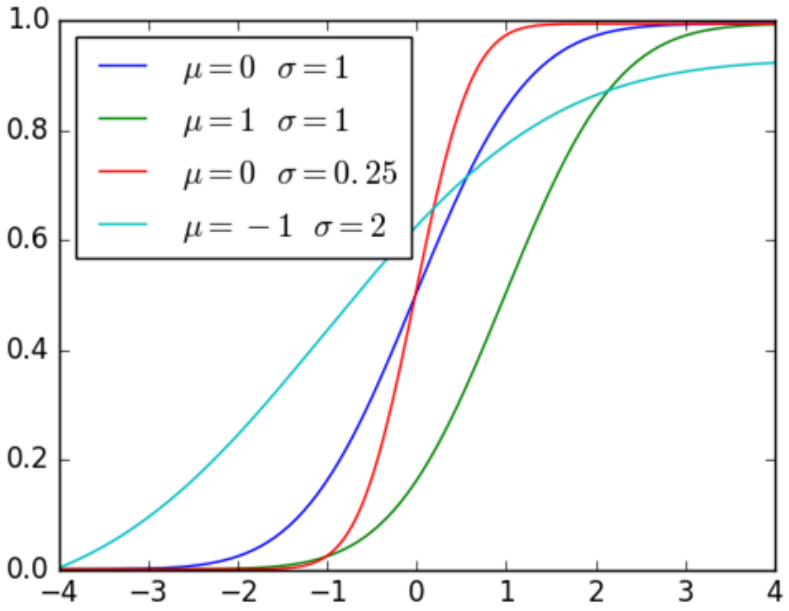
Image from: wikipedia.org

Normal distribution

Probability distribution

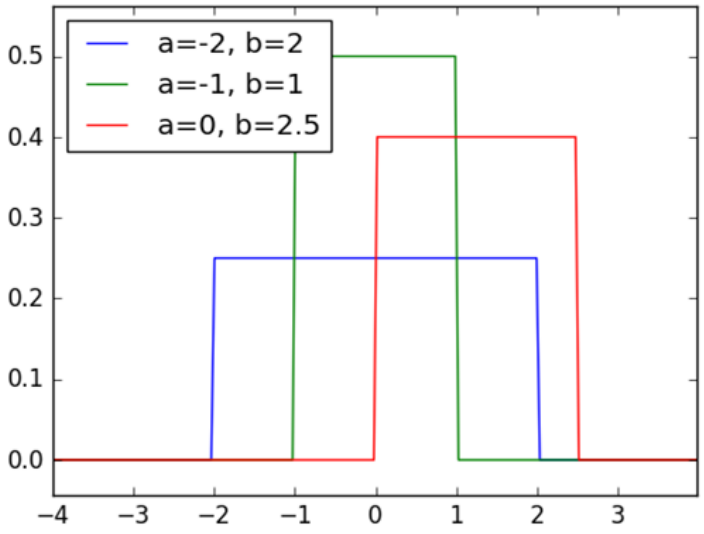


Cumulative distribution



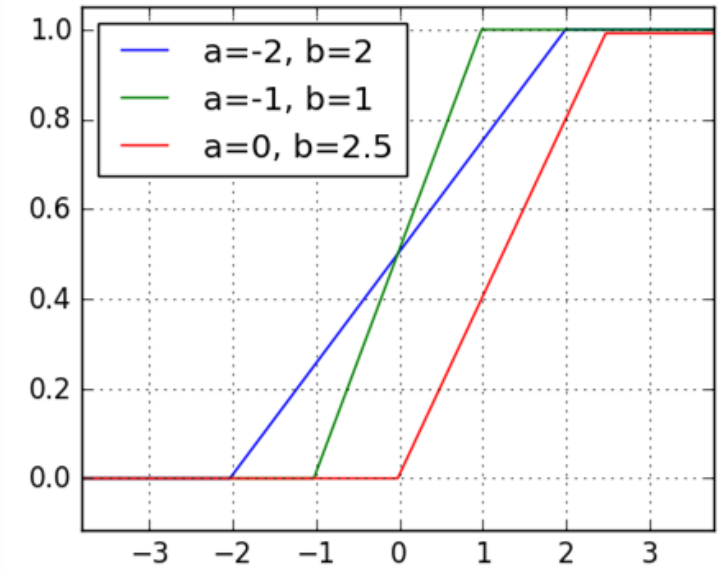
Uniform distribution

Probability distribution



$$P(x) = \begin{cases} \dfrac{1}{b-a} & for\ x \in [a,b] \\ 0 & for\ x \notin [a,b] \end{cases}$$

Cumulative distribution

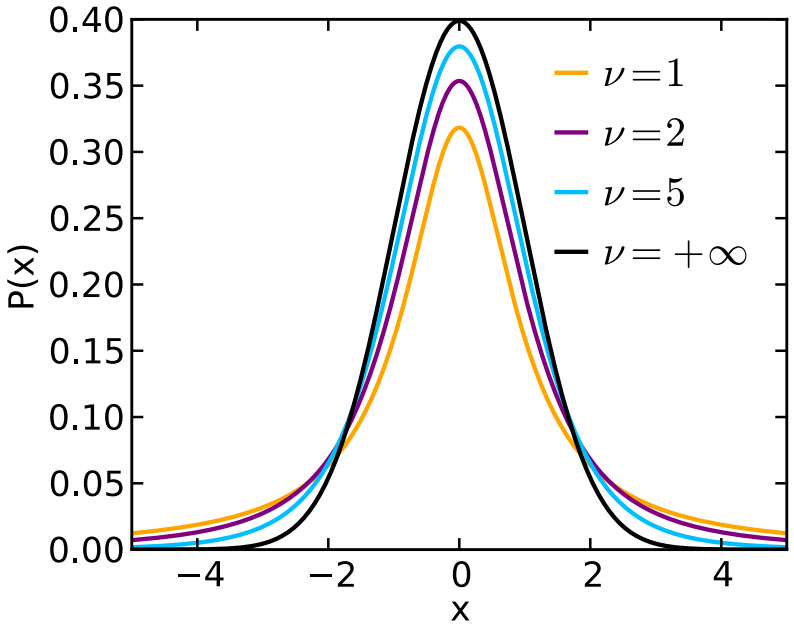


$$F(x) = \begin{cases} 0 & for\ x \in (-\infty, a) \\ \dfrac{x-a}{b-a} & for\ x \in [a,b) \\ 1 & for\ x \in [b,\infty) \end{cases}$$

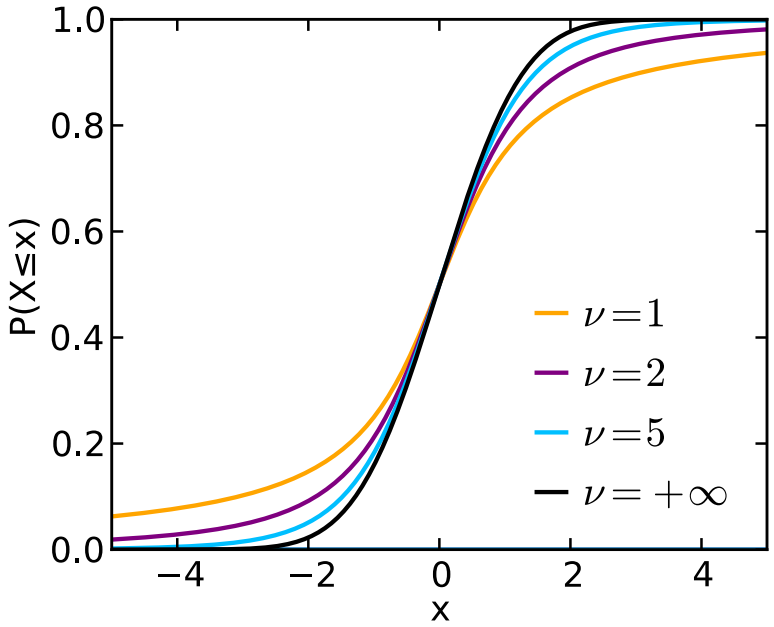
NGSeminars

t-Student distribution

Probability distribution



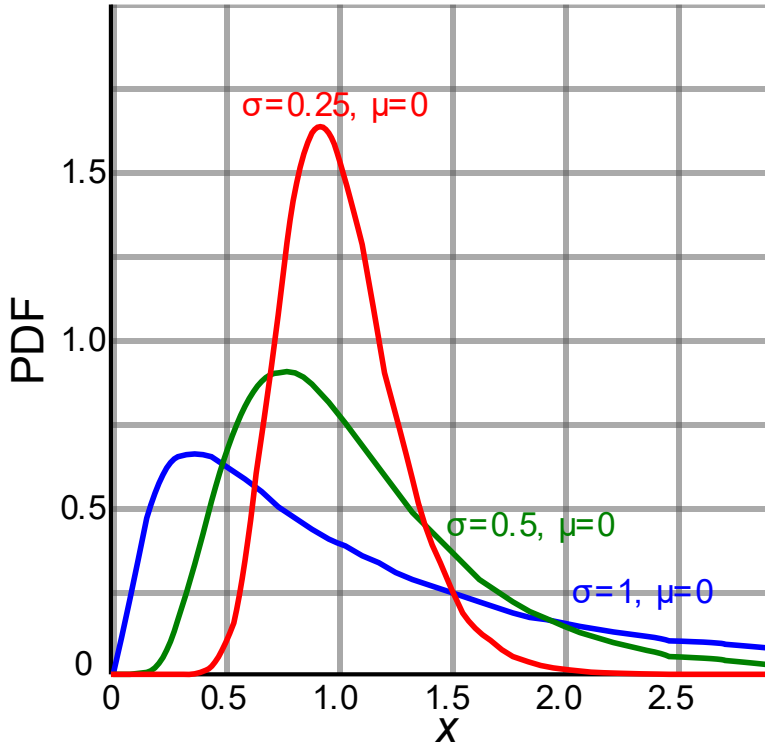
Cumulative distribution



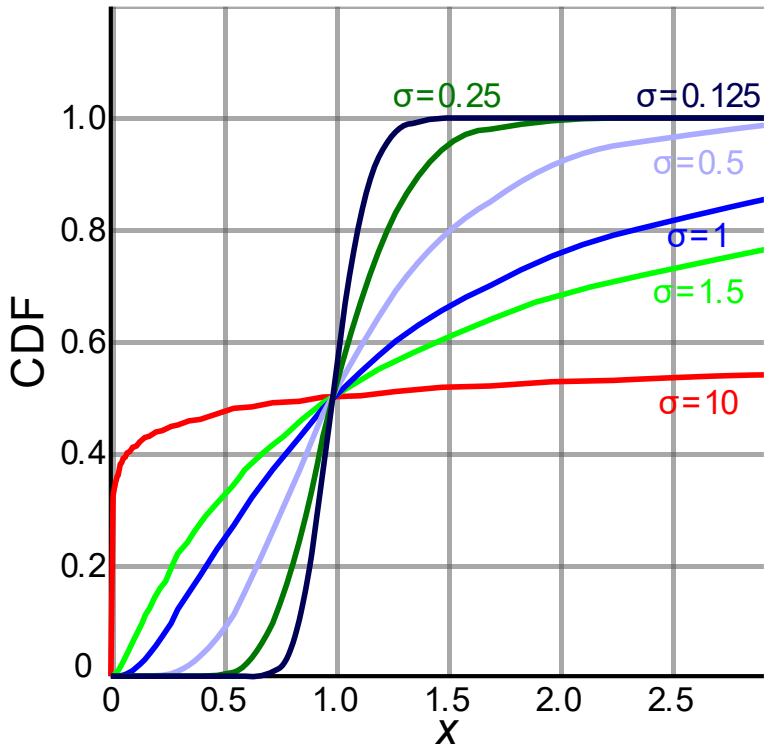
Images from: wikipedia.org

log-normal distribution

Probability distribution



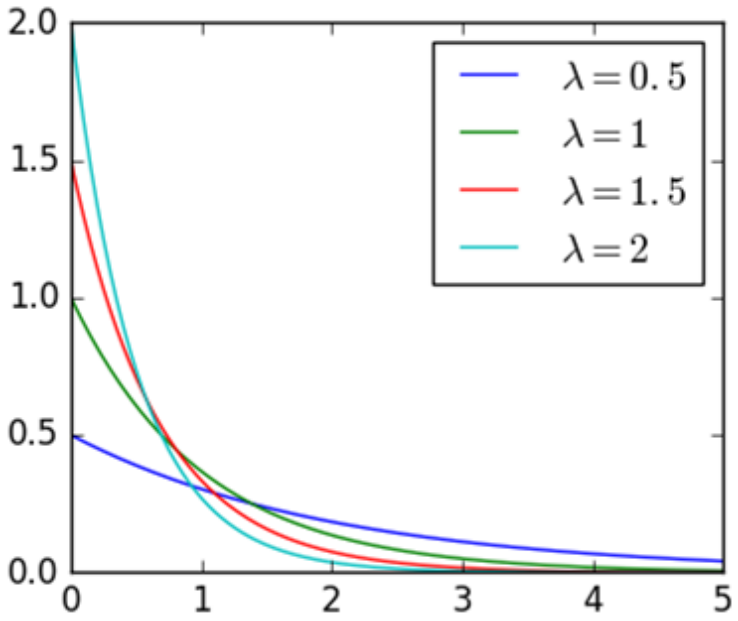
Cumulative distribution



Images from: wikipedia.org

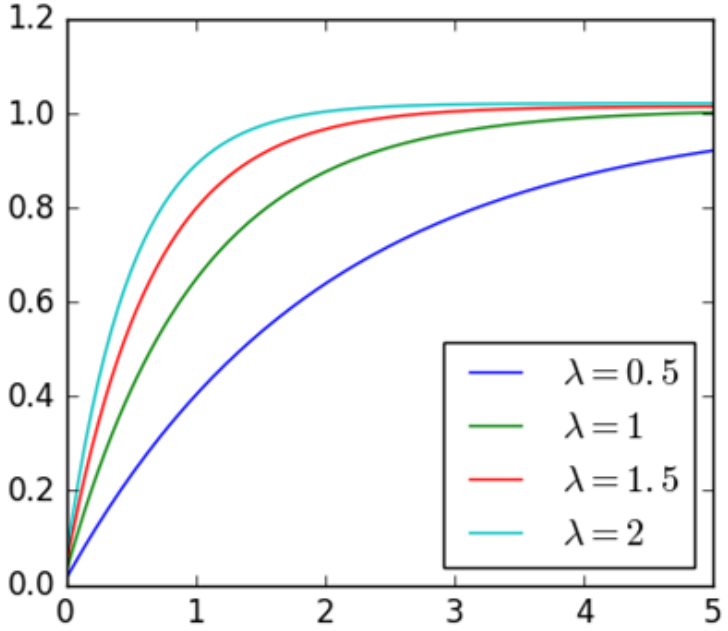
Exponential distribution

Probability distribution



$$P(x) = \begin{cases} \lambda \cdot e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Cumulative distribution

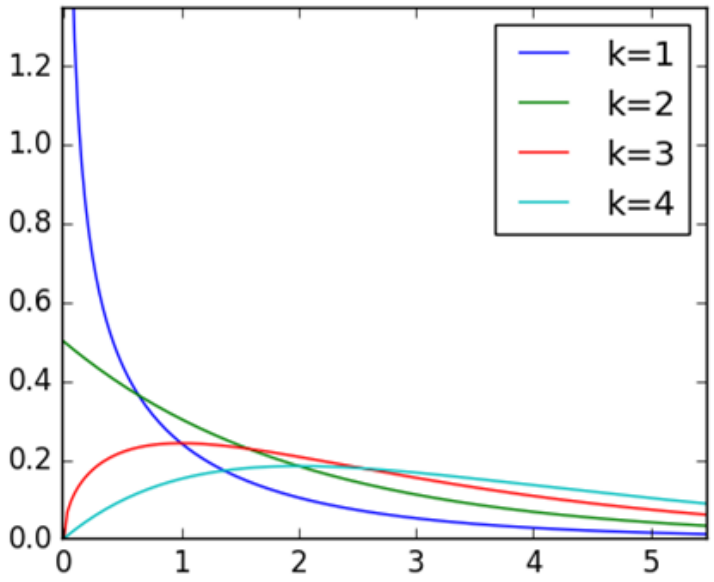


$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

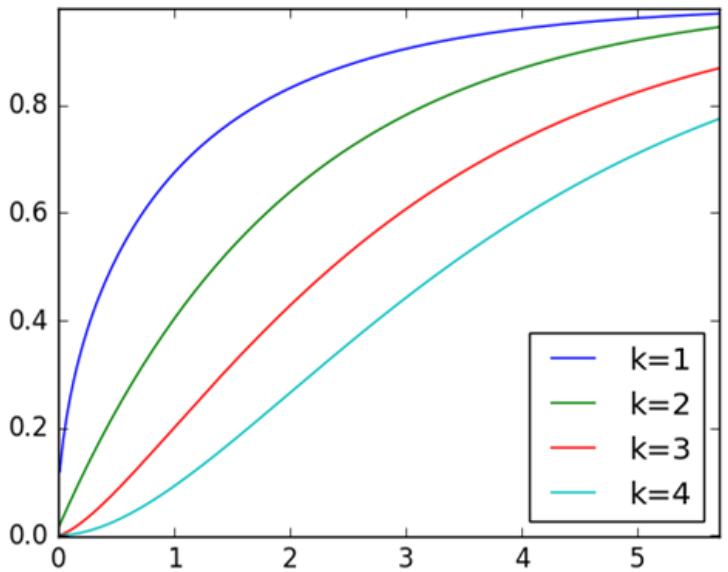
NGSeminars

Chi-square (X^2) distribution

Probability distribution



Cumulative distribution



Let’s try it!

NGSeminars

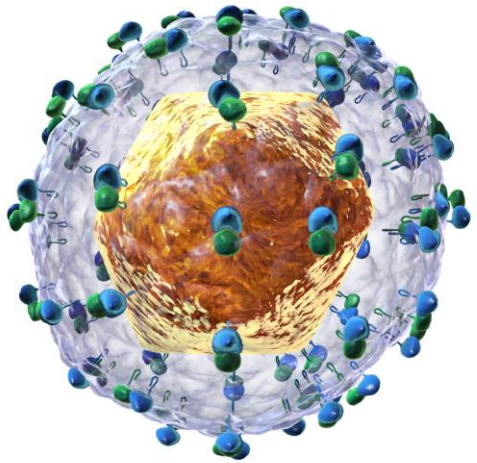
Dataset 1: HCV Data from Germany

ID	Category	Healthy	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
1	Blood Donor	YES	32	m	38.5000	52.5000	7.7000	22.1000	7.5000	6.9300	3.2300	106.0000	12.1000	69.0000
2	Blood Donor	YES	32	m	38.5000	70.3000	18.0000	24.7000	3.9000	11.1700	4.8000	74.0000	15.6000	76.5000
3	Blood Donor	YES	32	m	46.9000	74.7000	36.2000	52.6000	6.1000	8.8400	5.2000	86.0000	33.2000	79.3000
4	Blood Donor	YES	32	m	43.2000	52.0000	30.6000	22.6000	18.9000	7.3300	4.7400	80.0000	33.8000	75.7000
5	Blood Donor	YES	32	m	39.2000	74.1000	32.6000	24.8000	9.6000	9.1500	4.3200	76.0000	29.9000	68.7000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
610	Cirrhosis	NO	59	f	39.0000	51.3000	19.6000	285.8000	40.0000	5.7700	4.5100	136.1000	101.1000	70.5000
612	Cirrhosis	NO	64	f	24.0000	102.8000	2.9000	44.4000	20.0000	1.5400	3.0200	63.0000	35.9000	71.3000
613	Cirrhosis	NO	64	f	29.0000	87.3000	3.5000	99.0000	48.0000	1.6600	3.6300	66.7000	64.2000	82.0000
614	Cirrhosis	NO	46	f	33.0000	nan	39.0000	62.0000	20.0000	3.5600	4.2000	52.0000	50.0000	71.0000
615	Cirrhosis	NO	59	f	36.0000	nan	100.0000	80.0000	12.0000	9.0700	5.3000	67.0000	34.0000	68.0000

Source of data: http://archive.ics.uci.edu/ml/datasets/HCV+data

What is HCV?

- HCV is a hepatitis C virus that causes infectious disease hepatitis C.
- HCV can cause acute or chronic infection.
- It can cause liver failure, liver cancer, blood vessel complications in digestive tract.
- There is no vaccine for HCV.
- Treatment is often expensive, in extreme cases patients may require liver transplant.



Hepatis C Virus (HCV)

NGSeminars

Dataset 1: HCV Data from Germany

Categories:

- Blood Donor
- Cirrhosis – very advanced fibrosis
- Fibrosis
- Hepatitis – inflammation of liver tissue

Simplified categories (Healthy):

- Healthy (YES) = blood donors
- Ill (NO) = all other groups.

Variables:

- Age
- Sex
- ALB – albumin
- ALP – alkaline phosphatase
- ALT – alanine amino-transferase
- AST – aspartate amino-transferase
- BIL – bilirubin
- CHE – choline esterase
- CHOL – cholesterol
- CREA – creatinine
- GGT – gamma-glutamyl transferase
- PROT – total protein

NGSeminars

Dataset 2: HCV Data from Egypt

ID	Age	Gender	BMI	Fever	Nausea/Vomiting	Headache	Diarrhea	Fatigue & generalized bone ache	Jaundice	Epigastric pain	...	ALT 36	ALT 48	ALT after 24 w	RNA Base	RNA 4
1	56	M	35	YES	NO	NO	NO	YES	YES	YES	...	5	5	5	655330	634536
2	46	M	29	NO	YES	YES	NO	YES	YES	NO	...	57	123	44	40620	538635
3	57	M	33	YES	YES	YES	YES	NO	NO	NO	...	5	5	5	571148	661346
4	49	F	33	NO	YES	NO	YES	NO	YES	NO	...	48	77	33	1041941	449939
5	59	M	32	NO	NO	YES	NO	YES	YES	YES	...	94	90	30	660410	738756
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1381	44	M	29	NO	YES	YES	YES	NO	NO	NO	...	63	44	45	387795	55938
1382	55	M	34	NO	YES	YES	NO	NO	NO	NO	...	97	64	41	481378	152961
1383	42	M	26	YES	YES	NO	NO	NO	YES	NO	...	87	39	24	612664	572756
1384	52	M	29	YES	NO	NO	YES	YES	YES	NO	...	48	81	43	139872	76161
1385	55	F	26	NO	YES	YES	YES	NO	YES	NO	...	64	71	34	1190577	628730

Source of data: http://archive.ics.uci.edu/ml/datasets/Hepatitis+C+Virus+HCV+for+Egyptian+patients

Dataset 2: HCV Data from Egypt

Nominal variables:

- Gender
- Fever
- Nausea/Vomiting
- Headache
- Diarrhea
- Fatigue & generalized bone ache
- Jaundice
- Epigastric pain
- Baseline histological Grading
- Baseline histological staging

Continuous variables:

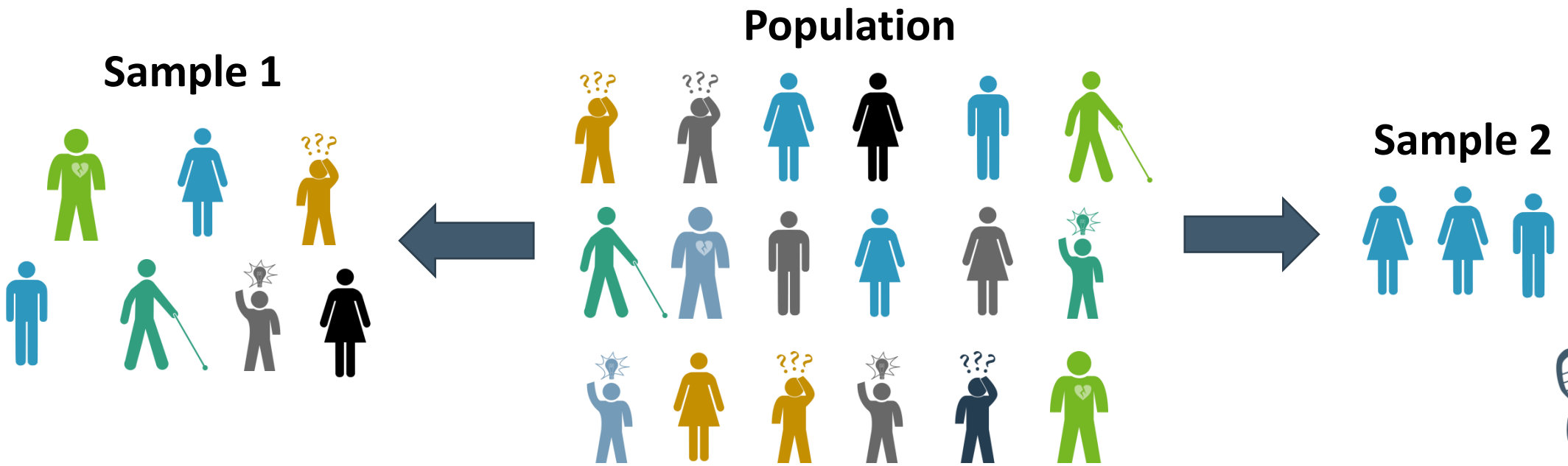
- Age
- BMI
- WBC – white blood cells
- RBC – red blood cells
- HGB – hemoglobin
- Plat – platelets
- AST - aspartate amino-transferase
- ALT - alanine amino-transferase at 7 timepoints
- RNA – viral RNA at 4 timepoints
- RNA EOF – RNA elongation factor

NGSeminars

Statistical tests

Statistical testing

- A way to answer the question “Can what I see in my sample be generalized to whole population?”.
- Done by showing that it is not probable that observed effect is caused by chance.



Statistical testing

Each statistical test has:

- Null hypothesis (H_0): it usually states that there is no difference or dependence in the underlying population.
- Alternative hypothesis (H_A): it usually states that there is difference or dependence in the population. It is what a researcher believes in or wants to prove by rejecting H_0.

The two hypotheses must be complementary and cover all possible options, like these two:

$$H_0: x_1 = x_2$$
$$H_A: x_1 \neq x_2$$

NGSeminars

Statistical testing - possible outcomes

		Reality	
Decision		H_0 false	H_0 true
	H_0 rejected	☺ (true positive)	**Type I error (false positive, α)**
	H_0 not rejected	**Type II error (false negative, β)**	☺ (true negative)

Statistical testing - procedure

1. Calculate test statistics (***s***) basing in your sample.
2. Calculate the probability of obtaining such or more extreme ***s*** assuming that $\boldsymbol{H_0}$ is true. This calculation is done with the use of theoretical distribution of ***s*** in this setting (and you don't need to bother much with equations, any statistical software can do it for you). What you will get is a ***p*-value** (often just ***p***).
3. Decide if you reject null hypothesis and accept alternative one.

Statistical testing – how to decide?

To decide, you need predefined significance level **α** that is an acceptable probability of type I error (false positive).

You will decide this way:

- $p \geq \alpha$ – there is not enough evidence to reject H_0.
  It may be true, but not necessarily.
- $p < \alpha$ – reject H_0, accept H_A and treat it as true.

In biomedical sciences we usually use $\alpha = 0.05$, but you can use other values if there are reasons to do so.

NGSeminars

What is p-value (or what it is not)?

- ***p*** is probability of your sample (or more extreme one) coming from population in which H_0 is true.
- But it is not a probability of you committing any type of error while testing.
- It is a value derived based on your sample that depends on effect size and sample size.
- But it is not a measure of effect size! Lower ***p*** does not automatically mean stronger dependence or greater difference.

NGSeminars

How to chose statistical test?

- What do you want to test: difference of means, equality of variances, normality of distribution?
- What type of variables you have (numeric, categorical)?
- How many variables do you want to analyze?
- Are your measurements/groups related/dependent?
- Is the distribution of values normal?

NGSeminars

Why is normal distribution important?

- Some tests, called **parametric**, are specially designed to work well on normally distributed data.
- They require fulfillment of assumptions.
- We use them for normally distributed data, because they are more powerful than nonparametric test.
- **Non-parametric** test work well with any distribution. They operate on ranks rather that real values of data.

NGSeminars

Parametric vs non-parametric tests

To compare heights of women and men…

parametric t-test will use actual values

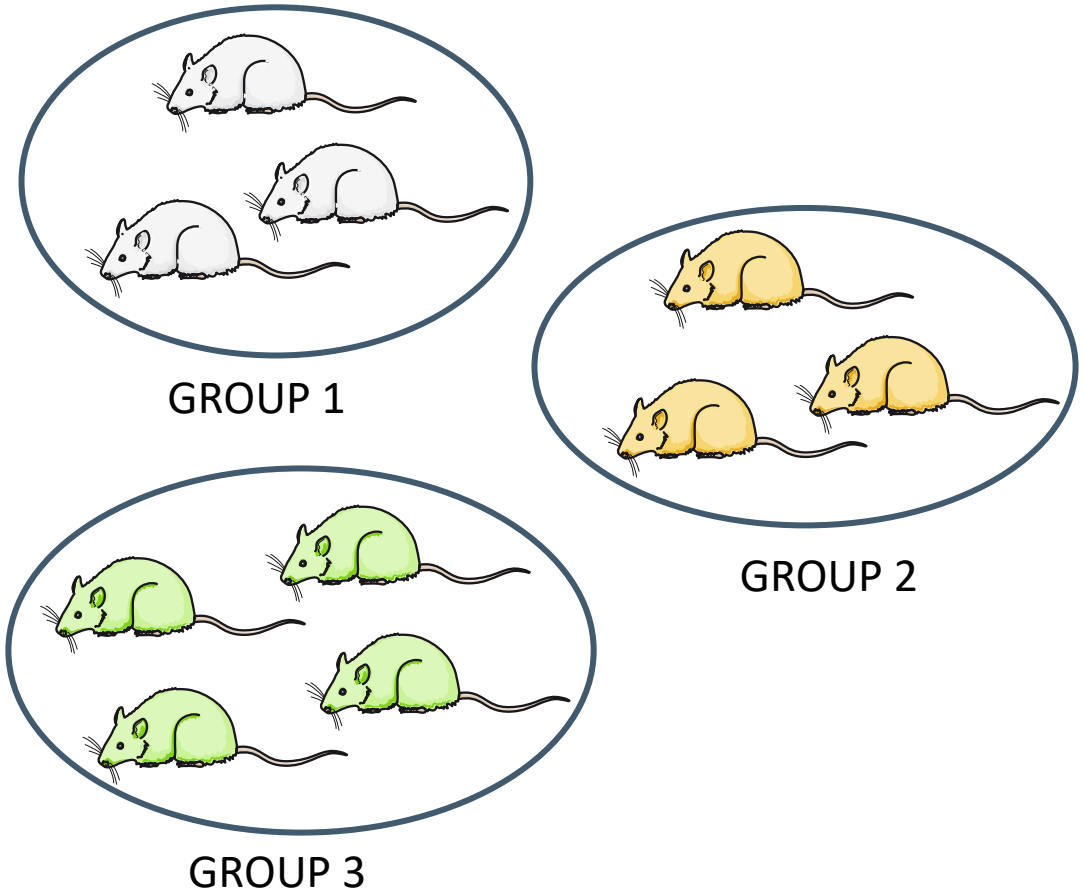
non-parametric Mann-Whitney test will use ranks

Height [cm]	Sex	Rank
156.0	F	1
162.7	F	2
165.4	F	3
169.3	M	4
174.7	M	5
177.2	F	6
179.5	M	7
182.1	F	8
186.9	M	9
193.4	M	10

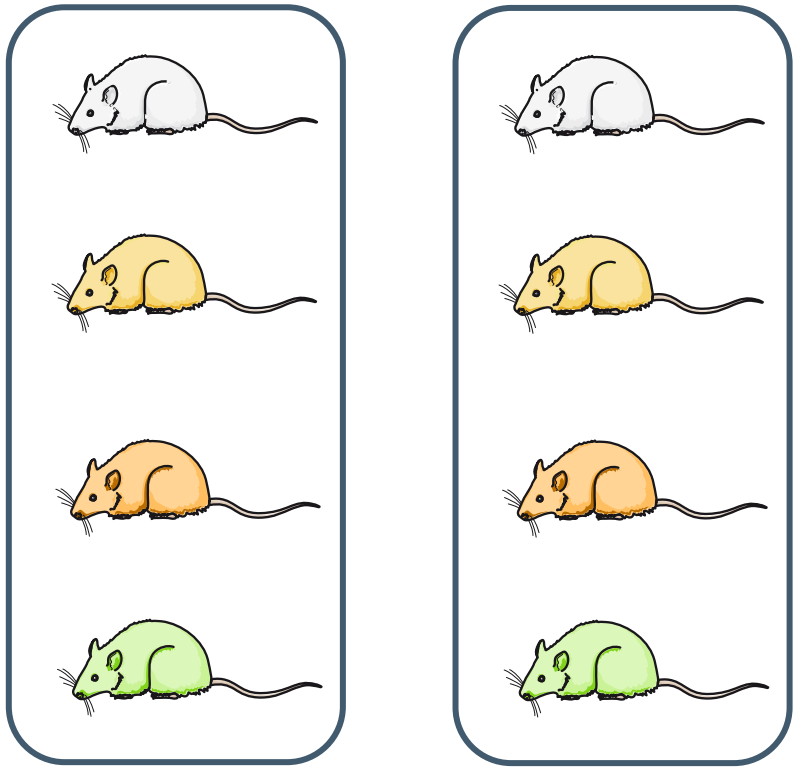
NGSeminars

What are related (paired) measurements?

Unpaired measurements

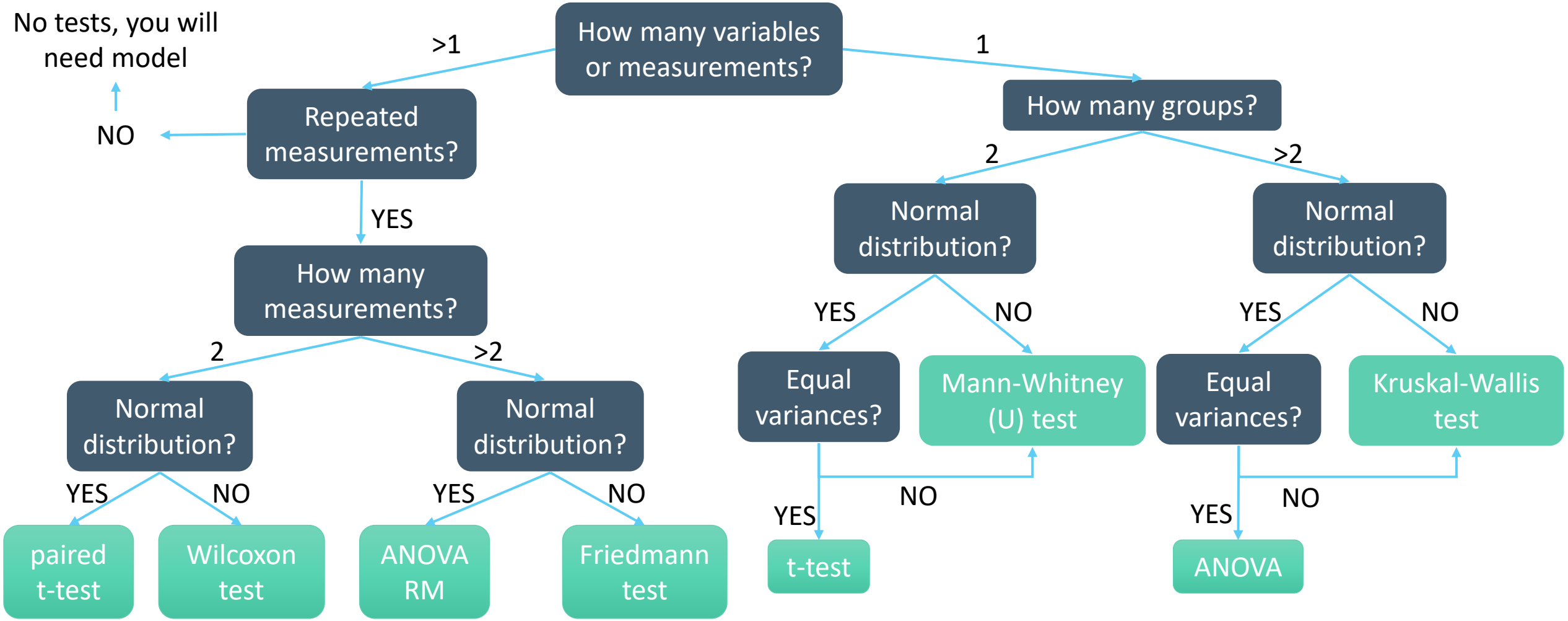


Paired measurements



NGSeminars

Tests for continuous variables



Let’s try it!

NGSeminars

Tests for nominal variables

If you need to test dependence between 2 nominal variables you can chose one of following tests:

- Fisher exact test for small samples (<5 cases in some group),
- Chi^2 test with Yates correction for middle size samples (<15 cases in some group),
- Chi^2 test for larger samples.

	Smokers	Non-smokers
Lung cancer	10	2
Healthy	546	623

Correlations

Correlation are used to describe dependence between 2 numeric variables (continuous or ordinal). Typically, we use 2 correlation measures:

- Person (linear) correlation for continuous, normally distributed variables.
- Spearman correlation for different distributions and ordinal variables.

Correlations are described by:

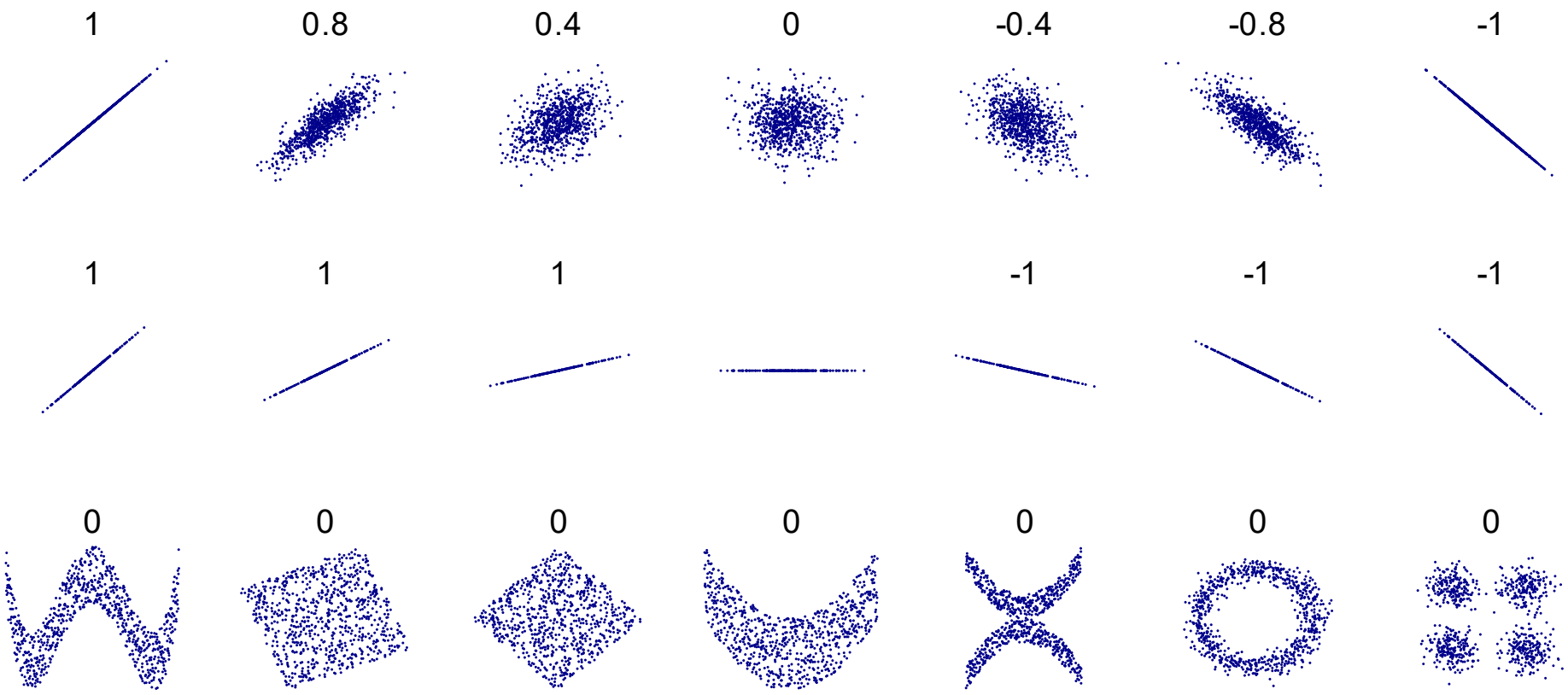
- correlation coefficient $r \epsilon [-1, 1]$
- *p*-value.

Correlations

Correlation are used to describe dependence between 2 numeric variables (continuous or ordinal). Typically, we use 2 correlation measures:

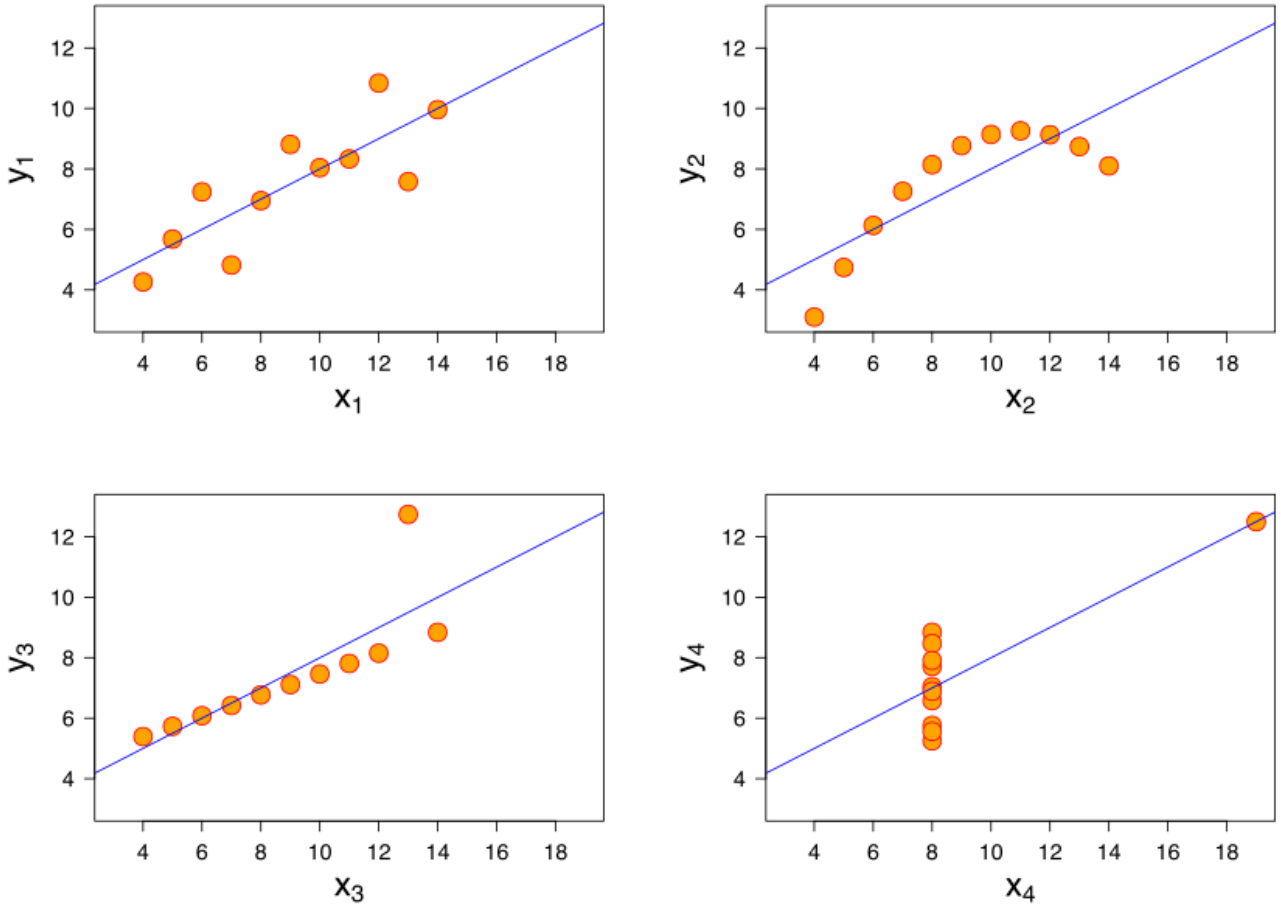
- Person (linear) correlation for continuous, normally distributed variables.
- Spearman correlation for different distributions and ordinal variables.

Pearson correlation



Be carful with correlations!

Anscombe's quartet

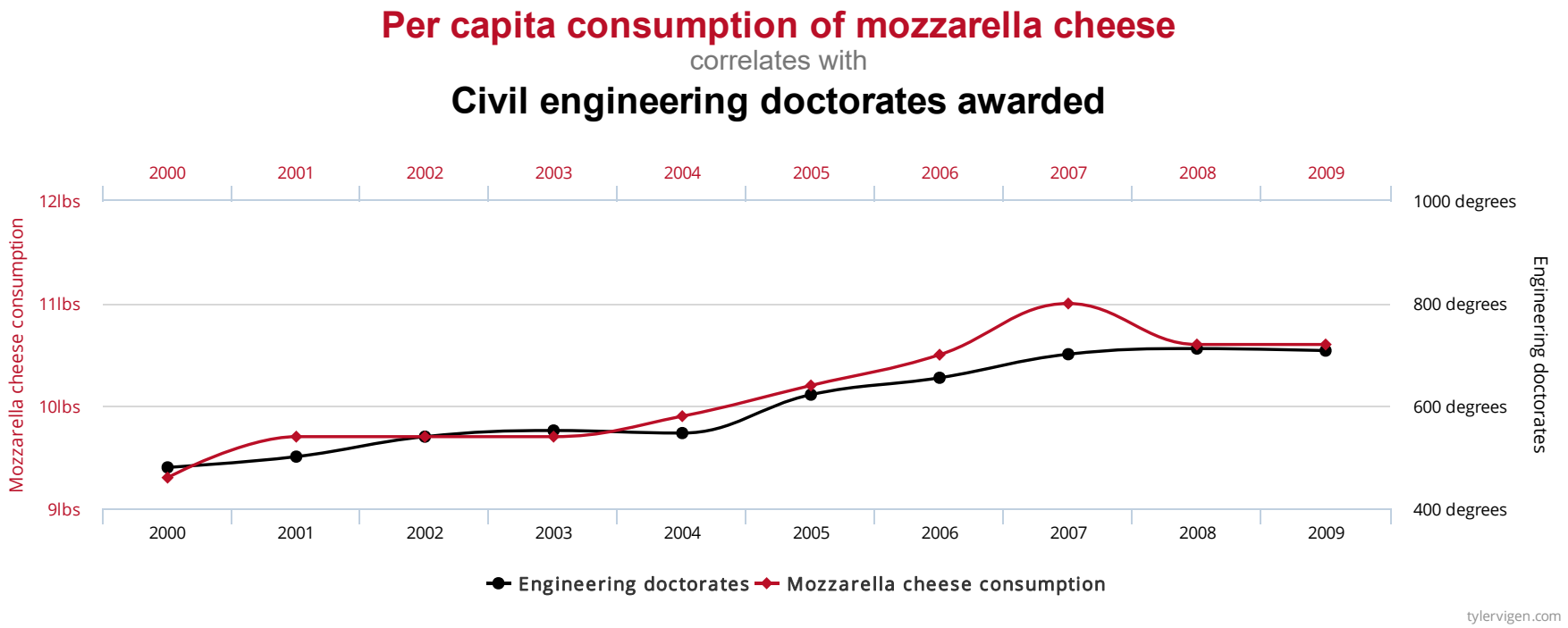


All od this datasets have the same mean, variance and identical Pearson correlation coefficient **r=0.816**.

Plots tell you more than numbers!

Images from: wikipedia.org

Once again, beware of correlations!



r = 0.9586

Correlation is not causation!
(*Unless you really need a lot of mozzarella cheese to write PhD thesis in civil engineering.*)

Image from: www.tylervigen.com/spurious-correlations

Let’s try it!

NGSeminars

Multiple testing

- When we perform one test and take α=0.05, so the risk of false positive is 5%. In the other words, probability of our rejecting null hypothesis being correct is 95%.
- What happens if we perform 2 test? What is the probability of committing at least one error?
  - Let's start with probability of not committing any type I error. It will be:
    $$0.95 \cdot 0.95 = (0.95)^2 = 0.9025$$
  - Then probability of at least one type I error is:
    $$1 - 0.9025 = 0.0975 \approx 10\%$$
- With 6 tests probability of at least one type I error would be:
  $$1 - (0.95)^6 = 0.2649 \approx 26\%$$
- That's why we use post-hoc tests, not just pairwise comparisons.

NGSeminars

Multiple testing

- Now, let's imagine that we should compare expression of 100 genes between healthy and ill people. How probably will we discover at least one difference that does not exist?

$$1 - (0.95)^{100} = 0.9941 \approx 99.4\%$$

- Ok, lets improve it and take α=0.01. Now, we have probability of at least one false positive equal to:

$$1 - (0.99)^{100} = 0.6340 \approx 63.4\%$$

- But, in real experiments we can have as much as 20,000 or even more genes. And then, we can be practically sure to have false findings, even with α=0.01.

NGSeminars

Multiple testing - FWER

- What can we do? Make α even lower!
- Fine, that's what Bonferroni correction does. It corrects p-values in such way that probability of at least one false positive (called familywise error rate, FWER) is α.
- We should only reject null hypothesis when:

$$p < \frac{\alpha}{m} \text{ where } m \text{ is number of comparisons.}$$

- With 20,000 comparisons we would use:

$$p < \frac{0.05}{20\,000} = 2.5 \cdot 10^{-6}$$

Multiple comparisons - FDR

- But, wasn't that to strict? How often can we see $p < 2.5 \cdot 10^{-6}$?
- Then, there are some other corrections.
- Many of the control false discovery rate (FDR), which is an expected fraction of false positives, not just probability of making even single such discovery.
- One of the options is Benjamini-Hochberg procedure:
  1. Sort *p*-values for *m* comparisons in ascending order.
  2. Find largest *k* such that $p(k) \leq \frac{k}{m} \cdot \alpha$.
  3. Reject first *k* H_0, those with lowest p.

NGSeminars

Let’s try it!

NGSeminars

Dataset 3: rats' gene expression

ID	healthy1	healthy2	healthy3	KO1	KO2	KO3	HET1	HET2	HET3	Sarcoma1	Sarcoma2	Sarcoma3
Group	healthy	healthy	healthy	ko	ko	ko	het	het	het	sarcoma	sarcoma	sarcoma
1700016D06Rik	1.8790	1.7172	1.6952	1.7642	1.5305	1.6763	1.4656	1.9429	2.2067	2.3327	2.5214	2.3348
2310003L06Rik	1.2736	1.4688	1.3084	1.2316	1.3412	1.3780	1.0981	1.1020	1.4581	1.2956	1.2950	1.4108
3100002H09Rik	1.4809	1.4581	1.6102	1.5838	1.6039	1.4140	1.1908	1.3091	1.5157	1.1004	1.3716	1.1029
4930513O06Rik	1.0668	1.0712	1.1849	1.2200	1.0701	1.1558	1.1400	1.7675	1.1055	1.2958	1.2896	1.1254
A1bg	1.6433	1.3979	2.0608	2.0007	2.0726	1.7918	1.9064	1.6826	1.8136	1.4766	1.8878	1.6590
...	...	...	...	...	...	...	...	...	...	...	...	...
Zyg11a	1.3978	1.4769	1.3731	1.4883	1.7974	1.3870	1.2953	1.2322	1.5306	1.6531	1.2419	1.5963
Zyg11b	8.8109	8.5008	8.6023	8.7563	8.4042	8.7173	8.7391	8.8345	8.4033	4.8837	6.0576	5.3475
Zyx	3.1944	2.5035	3.0898	3.4528	3.0730	3.1539	3.2964	2.8927	3.1548	3.7506	5.2061	5.0407
Zzef1	3.5122	3.1699	3.2266	3.7369	3.3165	3.0692	3.3376	3.1566	2.9069	2.9550	3.3947	3.0491
Zzz3	3.9735	3.4899	3.3508	3.7601	3.1962	3.5507	3.5866	2.3868	2.9908	2.7987	4.1192	4.6615

Source of data: Smyczyńska et al, TP53-Deficient Angiosarcoma Expression Profiling in Rat Model, *Cancers*, 12(6):1525, 2020.

Dataset 3: rats' gene expression

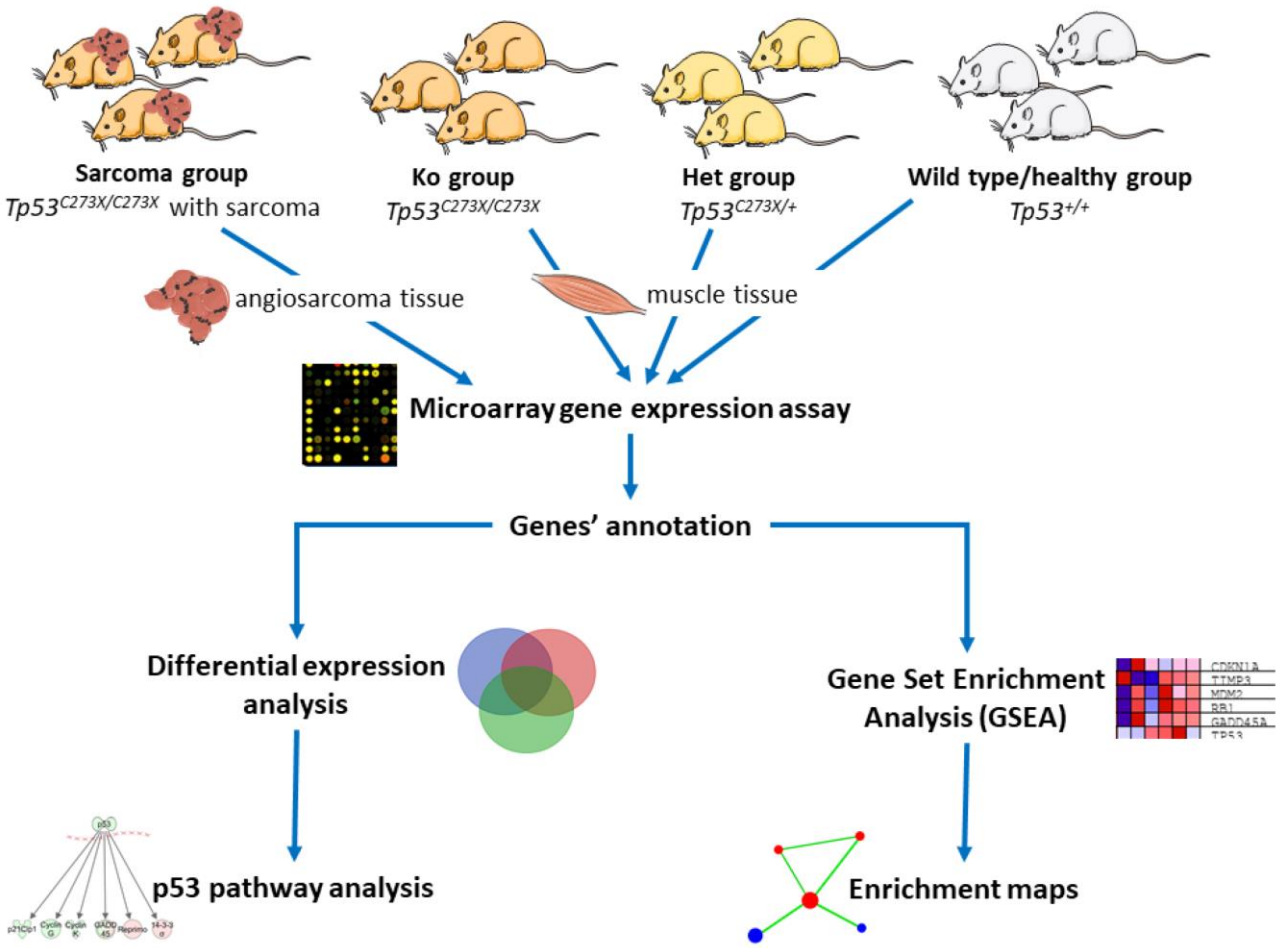


Image from: Smyczyńska et al, TP53-Deficient Angiosarcoma Expression Profiling in Rat Model, *Cancers*, 12(6):1525, 2020.