



CAMBRIDGE  
INSTITUTE



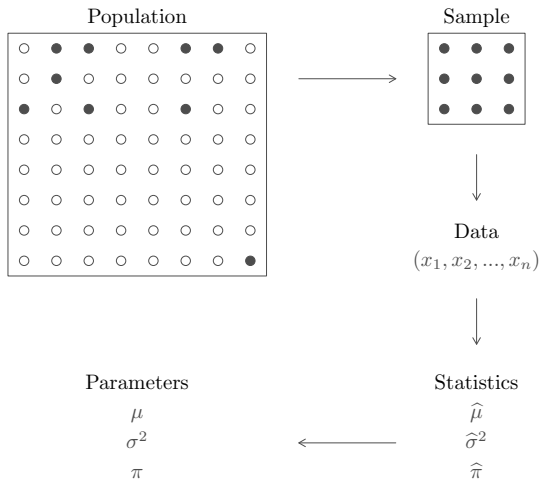
UNIVERSITY OF  
CAMBRIDGE

## Introduction to Statistical Analysis

Cancer Research UK – 12<sup>th</sup> of February 2021

D.-L. Couturier / M. Fernandes [Bioinformatics core]

# Grand Picture of Statistics



# Data Types

	$x_1$	$x_2$	$x_3$	$\dots$	$x_n$
Cancer status	C	<del>C</del>	<del>C</del>	$\dots$	C
Nucleic acid sequence	C	T	T	$\dots$	A
5-level pain score	3	1	5	$\dots$	4
# of daily admissions at A&E	16	23	12	$\dots$	17
Gene expression intensity	882.1	379.5	528.3	$\dots$	120.9

# Summary statistics and plots for qualitative data

5-level answers of 21 patients to the question

"How much did pain due to your ureteric stones interfere with your day to day activities ?":

3, 1, 5, 3, 1, 1, 1, 5, 1, 3, 4, 1, 1, 4, 5, 5, 5, 5, 5, 4, 4,

where

- ▶ 1 = "Not at all",
- ▶ 2 = "A little bit",
- ▶ 3 = "Somewhat",
- ▶ 4 = "Quite a bit",
- ▶ 5 = "Very much".

# Summary statistics and plots for qualitative data

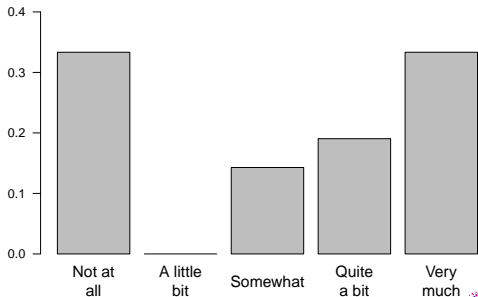
5-level answers of 21 patients to the question

"How much did pain due to your ureteric stones interfere with your day to day activities ?":

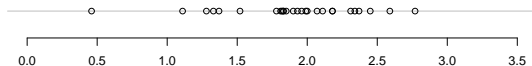
3, 1, 5, 3, 1, 1, 1, 5, 1, 3, 4, 1, 1, 4, 5, 5, 5, 5, 5, 4, 4,

where

- ▶ 1 = "Not at all",
- ▶ 2 = "A little bit",
- ▶ 3 = "Somewhat",
- ▶ 4 = "Quite a bit",
- ▶ 5 = "Very much".



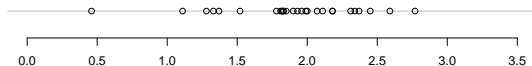
# Summary statistics and plots for quantitative data



Gene expression values of gene “CCND3 Cyclin D3” from 27 patients diagnosed with acute lymphoblastic leukaemia:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$
0.46	1.11	1.28	1.33	1.37	1.52	1.78	1.81	1.82
$x_{(10)}$	$x_{(11)}$	$x_{(12)}$	$x_{(13)}$	$x_{(14)}$	$x_{(15)}$	$x_{(16)}$	$x_{(17)}$	$x_{(18)}$
1.83	1.83	1.85	1.9	1.93	1.96	1.99	2.00	2.07
$x_{(19)}$	$x_{(20)}$	$x_{(21)}$	$x_{(22)}$	$x_{(23)}$	$x_{(24)}$	$x_{(25)}$	$x_{(26)}$	$x_{(27)}$
2.11	2.18	2.18	2.31	2.34	2.37	2.45	2.59	2.77

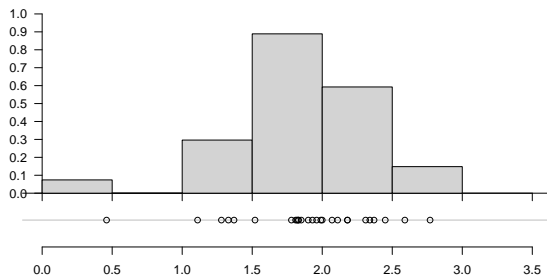
# Summary statistics and plots for quantitative data



Gene expression values of gene “CCND3 Cyclin D3” from 27 patients diagnosed with acute lymphoblastic leukaemia:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$
0.46	1.11	1.28	1.33	1.37	1.52	1.78	1.81	1.82
$x_{(10)}$	$x_{(11)}$	$x_{(12)}$	$x_{(13)}$	$x_{(14)}$	$x_{(15)}$	$x_{(16)}$	$x_{(17)}$	$x_{(18)}$
1.83	1.83	1.85	1.9	1.93	1.96	1.99	2.00	2.07
$x_{(19)}$	$x_{(20)}$	$x_{(21)}$	$x_{(22)}$	$x_{(23)}$	$x_{(24)}$	$x_{(25)}$	$x_{(26)}$	$x_{(27)}$
2.11	2.18	2.18	2.31	2.34	2.37	2.45	2.59	2.77

# Summary statistics and plots for quantitative data

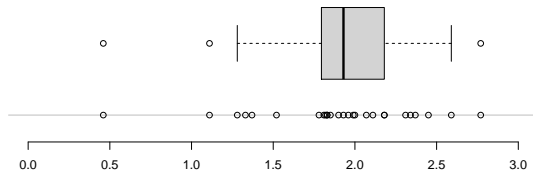


Gene expression values of gene “CCND3 Cyclin D3” from 27 patients diagnosed with acute lymphoblastic leukaemia:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$
0.46	1.11	1.28	1.33	1.37	1.52	1.78	1.81	1.82
$x_{(10)}$	$x_{(11)}$	$x_{(12)}$	$x_{(13)}$	$x_{(14)}$	$x_{(15)}$	$x_{(16)}$	$x_{(17)}$	$x_{(18)}$
1.83	1.83	1.85	1.9	1.93	1.96	1.99	2.00	2.07
$x_{(19)}$	$x_{(20)}$	$x_{(21)}$	$x_{(22)}$	$x_{(23)}$	$x_{(24)}$	$x_{(25)}$	$x_{(26)}$	$x_{(27)}$
2.11	2.18	2.18	2.31	2.34	2.37	2.45	2.59	2.77



# Summary statistics and plots for quantative data



Gene expression values of gene “CCND3 Cyclin D3” from 27 patients diagnosed with acute lymphoblastic leukaemia:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$
0.46	1.11	1.28	1.33	1.37	1.52	1.78	1.81	1.82
$x_{(10)}$	$x_{(11)}$	$x_{(12)}$	$x_{(13)}$	$x_{(14)}$	$x_{(15)}$	$x_{(16)}$	$x_{(17)}$	$x_{(18)}$
1.83	1.83	1.85	1.9	1.93	1.96	1.99	2.00	2.07
$x_{(19)}$	$x_{(20)}$	$x_{(21)}$	$x_{(22)}$	$x_{(23)}$	$x_{(24)}$	$x_{(25)}$	$x_{(26)}$	$x_{(27)}$
2.11	2.18	2.18	2.31	2.34	2.37	2.45	2.59	2.77

# Two-sample case: independent versus paired samples

Permeability constants of a placental membrane at term (X) and between 12 to 26 weeks gestational age (Y).

	1	2	3	4	5	6	7	8	9	10
X	0.80	0.83	1.89	1.04	1.45	1.38	1.91	1.64	0.73	1.46
Y	1.15	0.88	0.90	0.74	1.21					

Hamilton depression scale factor measurements in 9 patients with mixed anxiety and depression, taken at the first (X) and second (Y) visit after initiation of a therapy (administration of a tranquilizer).

	1	2	3	4	5	6	7	8	9
X	1.83	0.50	1.62	2.48	1.68	1.88	1.55	3.06	1.30
Y	0.88	0.65	0.60	2.05	1.06	1.29	1.06	3.14	1.29
Y-X	-0.95	0.15	-1.02	-0.43	-0.62	-0.59	-0.49	0.08	-0.01

# Quiz Time

## Sections 1 to 4

On <http://bioinformatics-core-shared-training.github.io/IntroductionToStats/>, select **Online quiz** under **Course Materials**

# Statistical distributions

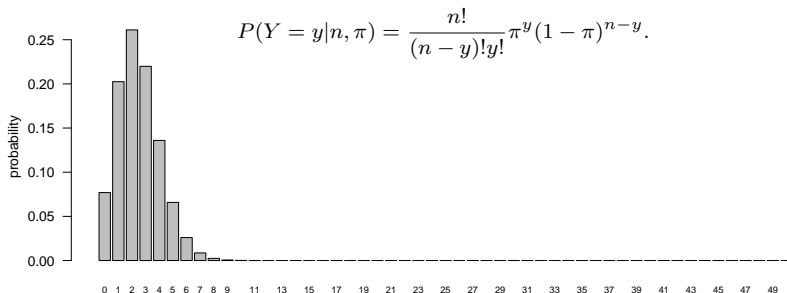
“In probability theory and statistics, a statistical distribution is a **mathematical function** that provides the **probabilities of occurrence of different possible outcomes** in an experiment” [Wikipedia].

# Statistical distributions

“In probability theory and statistics, a statistical distribution is a **mathematical function** that provides the **probabilities of occurrence of different possible outcomes** in an experiment” [Wikipedia].

For a given cancer, mutation of the nucleic acid located at position 790 of Exon 20 is assumed to occur with a probability of  $\sim 5\%$ .

Probability of observing  $y$  patients out of  $n = 50$  cancer patients with this mutation?



# Some parametric distributions: Binomial distribution

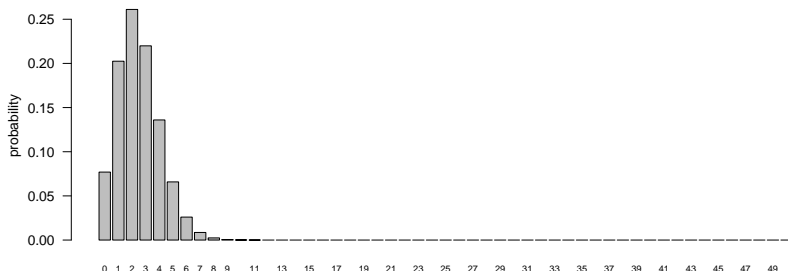
- ▶ the **number of successes out of  $n$  trials** (experiments),  $Y = \sum_{i=1}^n X_i$ , follows a binomial distribution with parameters  $n$  and  $\pi$ :

$$Y \sim \text{Bin}(n, \pi),$$

$$P(Y = y | n, \pi) = \frac{n!}{(n-y)!y!} \pi^y (1-\pi)^{n-y}.$$

IF

- ▶  $n$  independent experiments,
- ▶ outcome of each experiment is dichotomous (success/failure),
- ▶ the probability of success  $\pi$  is the same for all experiments,



# Some parametric distributions: Poisson distribution

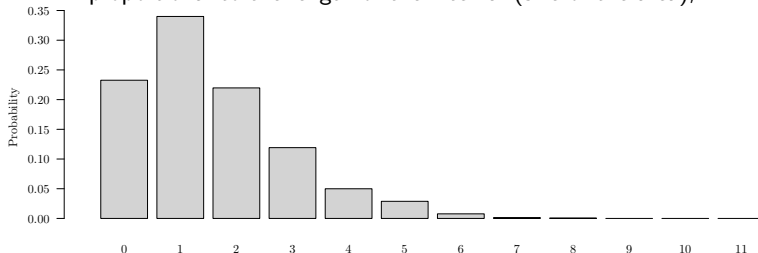
- ▶ the number of events occurring in a fixed time interval or in a given area,  $X$ , may be modelled by means of a Poisson distribution with parameter  $\lambda$ :

$$X \sim \text{Poisson}(\lambda),$$

$$P(X = x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

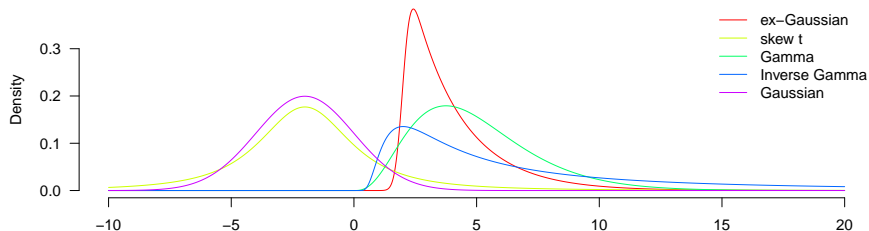
IF, during a time interval or in a given area,

- ▶ events occur independently,
- ▶ at the same rate,
- ▶ and the probability of an event to occur in a small interval (area) is proportional to the length of the interval (size of the area),



Number of chronic conditions per patient (US National Medical Expenditure Survey)

# Some parametric distributions: Continuous distrib.





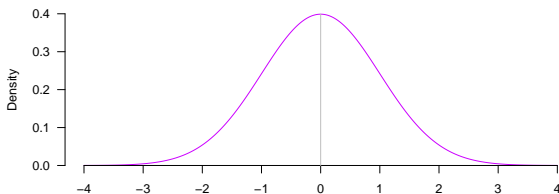
# Some parametric distributions: Normal distribution

$$X \sim N(\mu, \sigma^2), \quad f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mathbb{E}[X] = \mu, \quad \text{Var}[X] = \sigma^2,$$

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1), \quad f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

Probability density function,  $f_Z(z)$ , of a standard normal:



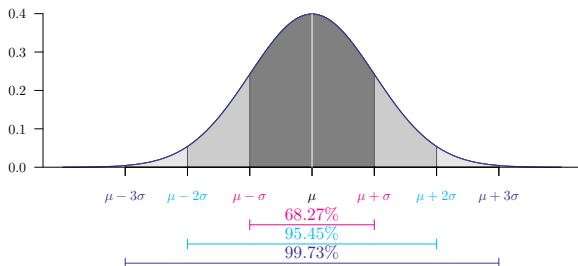
# Some parametric distributions: Normal distribution

$$X \sim N(\mu, \sigma^2), \quad f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mathbb{E}[X] = \mu, \quad \text{Var}[X] = \sigma^2,$$

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1), \quad f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

Probability density function,  $f_Z(z)$ , of a standard normal:



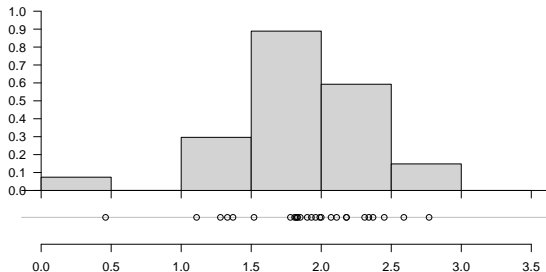
# Some parametric distributions: Normal distribution

$$X \sim N(\mu, \sigma^2), \quad f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mathbb{E}[X] = \mu, \quad \text{Var}[X] = \sigma^2,$$

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1), \quad f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

(i) Suitable modelling for a lot of variables:



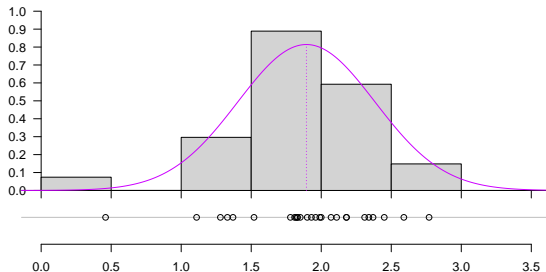
# Some parametric distributions: Normal distribution

$$X \sim N(\mu, \sigma^2), \quad f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mathbb{E}[X] = \mu, \quad \text{Var}[X] = \sigma^2,$$

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1), \quad f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

(i) Suitable modelling for a lot of variables



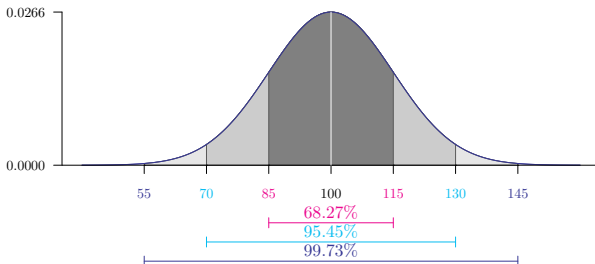
# Some parametric distributions: Normal distribution

$$X \sim N(\mu, \sigma^2), \quad f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mathbb{E}[X] = \mu, \quad \text{Var}[X] = \sigma^2,$$

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1), \quad f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

(i) Suitable modelling for a lot of variables: IQ



# Some parametric distributions: Normal distribution

$$X \sim N(\mu, \sigma^2), \quad f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$E[X] = \mu, \quad \text{Var}[X] = \sigma^2,$$

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1), \quad f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

## (ii) Central limit theorem (Lindeberg-Lévy CLT)

- ▶ Let  $(X_1, \dots, X_n)$  be  $n$  independent and identically distributed (iid) random variables drawn from distributions of expected values given by  $\mu$  and finite variances given by  $\sigma^2$ ,
- ▶ then

$$\hat{\mu} = \overline{X} = \frac{\sum_{i=1}^n X_i}{n} \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right).$$

If  $X_i \sim N(\mu, \sigma^2)$ , this result is true for all sample sizes.

# Central limit theorem shiny app:

## Distribution of the mean

<https://bioinformatics.cruk.cam.ac.uk/apps/stats/central-limit-theorem/>

<https://pauljudge.shinyapps.io/central-limit-theorem-master/>

95% Confidence interval for  $\mu$ , the population mean,  
when  $X_i \sim N(\mu, \sigma^2)$

- ▶ if  $X \sim N(\mu, \sigma^2)$ , then  $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ ,
- ▶ if  $X \sim N(\mu, \sigma^2)$ , then  $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$ ,

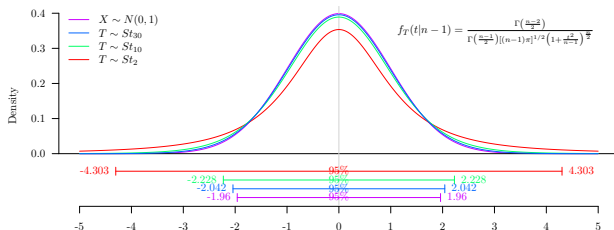
$$P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 0.95$$



# 95% Confidence interval for $\mu$ , the population mean, when $X_i \sim N(\mu, \sigma^2)$

- ▶ if  $X \sim N(\mu, \sigma^2)$ , then  $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ ,
- ▶ if  $X \sim N(\mu, \sigma^2)$ , then  $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ ,
- ▶ if  $\sigma$  unknown, then  $T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim St_{n-1}$ .

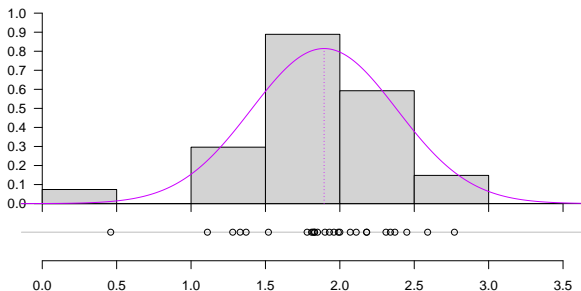
$$P\left( \bar{X} - \frac{s}{\sqrt{n}} \cdot t_{n-1, \alpha/2} < \mu < \bar{X} + \frac{s}{\sqrt{n}} \cdot t_{n-1, \alpha/2} \right) = 0.95$$



95% Confidence interval for  $\mu$ , the population mean,  
when  $X_i \sim N(\mu, \sigma^2)$

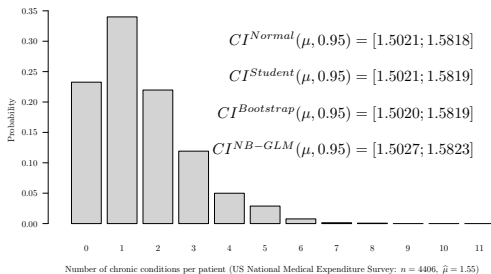
- ▶ if  $X \sim N(\mu, \sigma^2)$ , then  $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ ,
- ▶ if  $X \sim N(\mu, \sigma^2)$ , then  $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$ ,
- ▶ if  $\sigma$  unknown, then  $T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim St_{n-1}$ .

$$P\left( \quad < \quad < \quad \right) = 0.95$$



# 95% Confidence interval for $\mu$ , the population mean, when $X_i \sim iid(\mu, \sigma^2)$

- ▶ CLT:  $\bar{X} \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right)$ ,
- ▶ if  $X \sim N(\mu, \sigma^2)$ , then  $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$ ,
- ▶ if  $\sigma$  unknown, then  $T = \frac{X-\mu}{s} \sim St_{n-1}$ .



# 95% Confidence interval for $\mu_Y - \mu_X$ , the difference between population means

If we have

- ▶  $X_i \sim iid(\mu_X, \sigma_X^2), i = 1, \dots, n_X,$
- ▶  $Y_i \sim iid(\mu_Y, \sigma_Y^2), i = 1, \dots, n_Y,$

# 95% Confidence interval for $\mu_Y - \mu_X$ , the difference between population means

If we have

- ▶  $X_i \sim iid(\mu_X, \sigma_X^2)$ ,  $i = 1, \dots, n_X$ ,
- ▶  $Y_i \sim iid(\mu_Y, \sigma_Y^2)$ ,  $i = 1, \dots, n_Y$ ,

then

- ▶ if  $\sigma_X^2 = \sigma_Y^2$  [Student's t-test equation],

$$\triangleright CI(\mu_Y - \mu_X, 0.95) = (\bar{Y} - \bar{X}) \pm t_{1-\frac{\alpha}{2}, n_X+n_Y-2} s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}$$

$$\text{where } s_p = \frac{(n_X-1)s_X^2 + (n_Y-1)s_Y^2}{n_X+n_Y-2},$$

# 95% Confidence interval for $\mu_Y - \mu_X$ , the difference between population means

If we have

- ▶  $X_i \sim iid(\mu_X, \sigma_X^2)$ ,  $i = 1, \dots, n_X$ ,
- ▶  $Y_i \sim iid(\mu_Y, \sigma_Y^2)$ ,  $i = 1, \dots, n_Y$ ,

then

- ▶ if  $\sigma_X^2 = \sigma_Y^2$  [Student's t-test equation],

$$\triangleright CI(\mu_Y - \mu_X, 0.95) = (\bar{Y} - \bar{X}) \pm t_{1-\frac{\alpha}{2}, n_X+n_Y-2} s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}$$

$$\text{where } s_p = \frac{(n_X-1)s_X^2 + (n_Y-1)s_Y^2}{n_X+n_Y-2},$$

- ▶ if  $\sigma_X^2 \neq \sigma_Y^2$  [Welch-Satterthwaite's t-test equation],

$$\triangleright CI(\mu_Y - \mu_X, 0.95) = (\bar{Y} - \bar{X}) \pm t_{1-\frac{\alpha}{2}, df} \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}, \text{ where}$$

$$df = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{\left(\frac{s_X^2}{n_X}\right)^2}{n_X-1} + \frac{\left(\frac{s_Y^2}{n_Y}\right)^2}{n_Y-1}}.$$

# Central limit theorem shiny app:

## Coverage of Student's asymptotic confidence intervals

<https://bioinformatics.cruk.cam.ac.uk/apps/stats/central-limit-theorem/>

<https://pauljudge.shinyapps.io/central-limit-theorem-master/>

# Quiz Time

## Practical 1

[https://bioinformatics-core-shared-training.github.io/  
IntroductionToStats/practical.html](https://bioinformatics-core-shared-training.github.io/IntroductionToStats/practical.html)



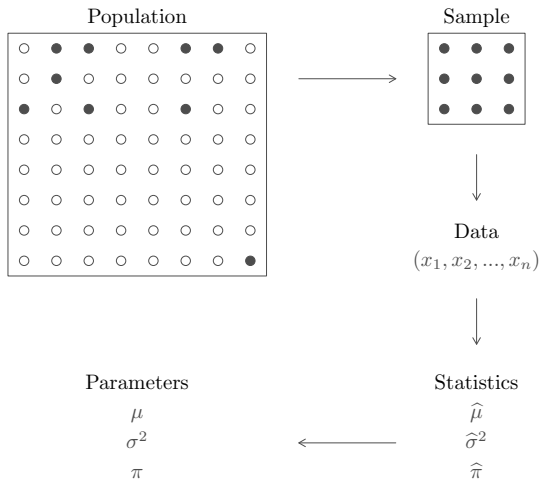
## PART II:

### Parametric and non-parametric one-sample location tests

Cancer Research UK – 12<sup>th</sup> of February 2021

D.-L. Couturier / M. Fernandes [Bioinformatics core]

# Grand Picture of Statistics



# Statistical hypothesis testing

A hypothesis test describes a phenomenon by means of two non-overlapping idealised models/descriptions:

- ▶ the null hypothesis **H0**, “generally assumed to be true until evidence indicates otherwise”
- ▶ the alternative hypothesis **H1**.

The aim of the test is to reject the null hypothesis in favour of the alternative hypothesis, and conclude, with a probability  $\alpha$  of being wrong, that the idealised model/description of H1 is true.

Theory 1: Dieters lose more fat than the exercisers

Theory 2: There is no majority for Brexit now

Theory 3: Serum vitamin C is reduced in patients

# Statistical hypothesis testing

Several-step process:

- ▶ Define  $H_0$  and  $H_1$  according to a theory
- ▶ Set  $\alpha$ , the probability of rejecting  $H_0$  when it is true (type I error),
- ▶ Determine the test statistic to be used,
- ▶ Define  $n$ , the sample size, allowing you to reject  $H_0$  when  $H_1$  is true with a probability  $1 - \beta$  (Power),
- ▶ Collect the data,
- ▶ Perform the statistical test, define the  $p$ -value, and reject (or not) the null hypothesis.

# Statistical hypothesis testing

Many options:

- ▶ One-sided versus two-sided tests,
- ▶ Exact versus asymptotic tests,
- ▶ Parametric versus non-parametric tests.

# Parametric or non-parametric ?

		T-test		
		Outcome(s) normally distributed		
		Yes	Mildly	No
Sample size	Small			
	Medium			
	Large			

Situations which may suggest the use of non-parametric statistics:

- ▶ When there is a small sample size or **very unequal groups**,
- ▶ When the data has **notable outliers**,
- ▶ When one outcome has a **distribution other than normal**,
- ▶ When the data are **ordered** with many ties or are rank ordered.

**Non-parametric does not mean assumption free**

# Parametric location test

## (One-sided) t-test

We test:

$$H_0: \mu_{IQ} = 100,$$

$$H_1: \mu_{IQ} > 100.$$

We have  $X_i \sim N(\mu, \sigma^2), i = 1, \dots, n,$

We know

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

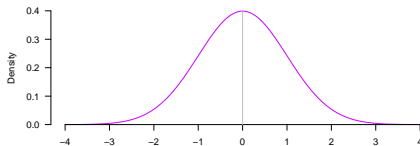
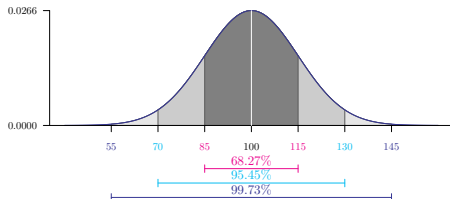
$$\blacktriangleright Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1),$$

Thus, if  $H_0$  is true, we have:

$$\blacktriangleright Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1).$$

Define the p-value:

$$\blacktriangleright p\text{-value} = P(T > T_{obs})$$



# Statistical tests

## 4 possible outcomes

Conclude:

- ▶ if  $p\text{-value} > \alpha \rightarrow$  do not reject  $H_0$ .
- ▶ if  $p\text{-value} < \alpha \rightarrow$  reject  $H_0$  in favour of  $H_1$ .

		Test Outcome	
		H0 not rejected	H1 accepted
Unknown Truth	H0 true	$1 - \alpha$ [TN]	$\alpha$ [FP]
	H1 true	$\beta$ [FN]	$1 - \beta$ [TP]

where

- ▶  $\alpha$  is the type I error, the probability of rejecting  $H_0$  when  $H_0$  is correct,
- ▶  $\beta$  is the type II error, the probability of not rejecting  $H_0$  when  $H_1$  is correct.

Warnings

- ▶ 'absence of evidence is not evidence of absence',
- ▶ design may help minimising FP and FN (ie, maximising TN and TP).



# Parametric location test

## (One-sided) binomial exact test

We test:

H0:  $\pi = 5\%$ ,

H1:  $\pi > 5\%$ .

We have  $X_i \sim \text{Bernoulli}(\pi), i = 1, \dots, n$ ,

We know

$$\blacktriangleright Y = \sum_{i=1}^n X_i \sim \text{Binomial}(\pi, n),$$

Thus, if H0 is true, we have:

$$\blacktriangleright Y = \sum_{i=1}^n X_i \sim \text{Binomial}(5\%, n),$$

# Parametric location test

## (One-sided) binomial exact test

We test:

$H_0: \pi = 5\%$ ,

$H_1: \pi > 5\%$ .

We have  $X_i \sim \text{Bernoulli}(\pi), i = 1, \dots, n$ ,

We know

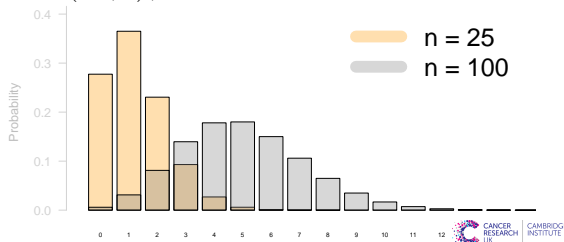
$$\blacktriangleright Y = \sum_{i=1}^n X_i \sim \text{Binomial}(\pi, n),$$

Thus, if  $H_0$  is true, we have:

$$\blacktriangleright Y = \sum_{i=1}^n X_i \sim \text{Binomial}(5\%, n),$$

Define the p-value:

$$\blacktriangleright p\text{-value} = P(Y > Y_{\text{obs}})$$



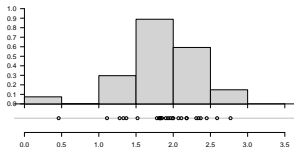
# Non-parametric location test

## Wilcoxon sign-rank test

A location model is assumed for  $X_i$ ,  $i = 1, \dots, n$ :

$$X_i = \theta + e_i,$$

where  $e_i \sim iid(\mu_e = 0, \sigma_e^2)$ , a symmetrical distribution.



Interest for **H0**:  $\theta = \theta_0$  against **H1**:  $\theta < \theta_0$  or  $\theta \neq \theta_0$  or  $\theta > \theta_0$ .

Test statistics :  $W^+ = \sum_{i=1}^n \iota(X_i - \theta_0 > 0) \text{ Rank}(|X_i - \theta_0|)$ .

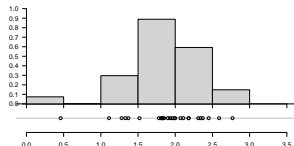
# Non-parametric location test

## Wilcoxon sign-rank test

A location model is assumed for  $X_i$ ,  $i = 1, \dots, n$ :

$$X_i = \theta + e_i,$$

where  $e_i \sim iid(\mu_e = 0, \sigma_e^2)$ , a symmetrical distribution.



Interest for **H0**:  $\theta = \theta_0$  against **H1**:  $\theta < \theta_0$  or  $\theta \neq \theta_0$  or  $\theta > \theta_0$ .

Test statistics :  $W^+ = \sum_{i=1}^n \iota(X_i - \theta_0 > 0) \text{Rank}(|X_i - \theta_0|)$ .

Distribution of  $W$  under H0:  $W^+$  has no closed-form distribution.

Wilcoxon signed rank test

```
data: golub[1042, gol.fac == "ALL"]
V = 268, p-value = 0.05847
alternative hypothesis: true location is not equal to 1.75
95 percent confidence interval:
 1.73868 2.09106
sample estimates:
(pseudo)median
 1.926475
```

# Introduction to Shiny Apps and Exercises

## PART III:

### Parametric and non-parametric two-sample location tests

Cancer Research UK – 12<sup>th</sup> of February 2021

D.-L. Couturier / M. Fernandes [Bioinformatics core]

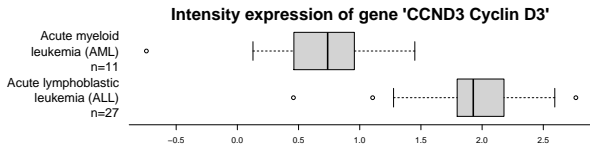
# Two-sample case

Many options:

- ▶ One-sided versus two-sided tests,
- ▶ Exact versus asymptotic tests,
- ▶ Parametric versus non-parametric tests,
- ▶ Tests for paired versus independent data.

# Parametric two-sample location test

## Two-sample two-sided Student-s & Welch's t-tests



We test  $H_0: \mu_Y - \mu_X = 0$  against  $H_1: \mu_Y - \mu_X \neq 0$ .

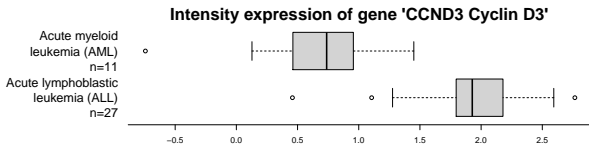
We know:

- ▶ Student's t-test [assume  $\sigma_X^2 = \sigma_Y^2$ ]: 
$$\frac{(\bar{Y} - \bar{X}) - (\mu_Y - \mu_X)}{s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim t_{1-\frac{\alpha}{2}, n_X + n_Y - 2}$$
- ▶ Welch's t-test [assume  $\sigma_X^2 \neq \sigma_Y^2$ ]: 
$$\frac{(\bar{Y} - \bar{X}) - (\mu_Y - \mu_X)}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} \sim t_{1-\frac{\alpha}{2}, df}$$



# Parametric two-sample location test

## Two-sample two-sided Student-s & Welch's t-tests



We test  $H_0: \mu_Y - \mu_X = 0$  against  $H_1: \mu_Y - \mu_X \neq 0$ .

We know:

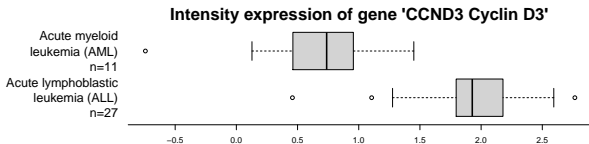
- ▶ Student's t-test [assume  $\sigma_X^2 = \sigma_Y^2$ ]:  $\frac{(\bar{Y} - \bar{X}) - (\mu_Y - \mu_X)}{s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim t_{1-\frac{\alpha}{2}, n_X + n_Y - 2}$
- ▶ Welch's t-test [assume  $\sigma_X^2 \neq \sigma_Y^2$ ]:  $\frac{(\bar{Y} - \bar{X}) - (\mu_Y - \mu_X)}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} \sim t_{1-\frac{\alpha}{2}, df}$

Two Sample t-test

```
data: golub[1042, gol.fac == "ALL"] and golub[1042, gol.fac == "AML"]
t = 6.7983, df = 36, p-value = 6.046e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.8829143 1.6336690
sample estimates:
mean of x mean of y
1.8938826 0.6355909
```

# Parametric two-sample location test

## Two-sample two-sided Student-s & Welch's t-tests



We test  $H_0: \mu_Y - \mu_X = 0$  against  $H_1: \mu_Y - \mu_X \neq 0$ .

We know:

- ▶ Student's t-test [assume  $\sigma_X^2 = \sigma_Y^2$ ]:  $\frac{(\bar{Y} - \bar{X}) - (\mu_Y - \mu_X)}{s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim t_{1-\frac{\alpha}{2}, n_X + n_Y - 2}$
- ▶ Welch's t-test [assume  $\sigma_X^2 \neq \sigma_Y^2$ ]:  $\frac{(\bar{Y} - \bar{X}) - (\mu_Y - \mu_X)}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} \sim t_{1-\frac{\alpha}{2}, df}$

Welch Two Sample t-test

```
data: golub[1042, gol.fac == "ALL"] and golub[1042, gol.fac == "AML"]
t = 6.3186, df = 16.118, p-value = 9.871e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.8363826 1.6802008
sample estimates:
mean of x mean of y
1.8938826 0.6355909
```

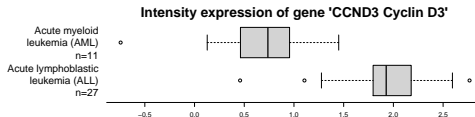
# Non-parametric two-sample location test

## Mann-Whitney-Wilcoxon test

Let

►  $X_i \sim iid(\mu_X, \sigma^2), i = 1, \dots, n_X,$

►  $Y_i \sim iid(\mu_X + \delta, \sigma^2), i = 1, \dots, n_Y.$



Interest for **H0**:  $\delta = \delta_0$  against **H1**:  $\delta < \delta_0$  or  $\delta \neq \delta_0$  or  $\delta > \delta_0$ .

Standardised test statistic:  $z = \frac{\sum_{i=1}^{n_Y} R(Y_i) - [n_Y(n_X + n_Y + 1)/2]}{\sqrt{n_X n_Y (n_X + n_Y + 1)/12}},$

where  $R(Y_i)$  denotes the rank of  $Y_i$  amongst the combined samples, i.e., amongst  $(X_1, \dots, X_{n_X}, Y_1, \dots, Y_{n_Y})$ .

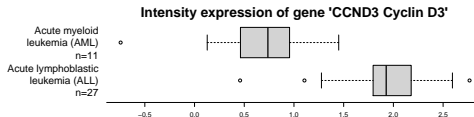
# Non-parametric two-sample location test

## Mann-Whitney-Wilcoxon test

Let

►  $X_i \sim iid(\mu_X, \sigma^2), i = 1, \dots, n_X,$

►  $Y_i \sim iid(\mu_X + \delta, \sigma^2), i = 1, \dots, n_Y.$



Interest for **H0**:  $\delta = \delta_0$  against **H1**:  $\delta < \delta_0$  or  $\delta \neq \delta_0$  or  $\delta > \delta_0$ .

Standardised test statistic:  $z = \frac{\sum_{i=1}^{n_Y} R(Y_i) - [n_Y(n_X + n_Y + 1)/2]}{\sqrt{n_X n_Y (n_X + n_Y + 1)/12}},$

where  $R(Y_i)$  denotes the rank of  $Y_i$  amongst the combined samples, i.e., amongst  $(X_1, \dots, X_{n_X}, Y_1, \dots, Y_{n_Y})$ .

Distribution of  $Z$  under H0:  $Z \sim N(0, 1)$ .

Implementation 1:

statistic = -4.361334 , p-value = 1.292716e-05

Implementation 2:

W = 284, p-value = 6.15e-07

alternative hypothesis: true location shift is not equal to 0

95 percent confidence interval:

0.89647 1.57023

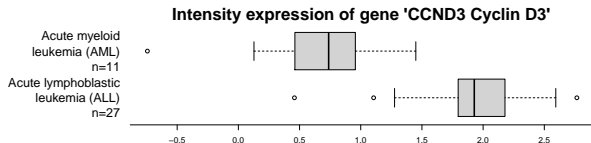
sample estimates:

difference in location

1.21951



# F-test of equality of variances

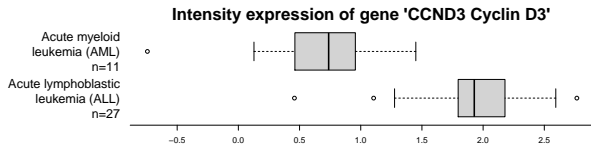


We test  $H_0: \sigma_Y^2 = \sigma_X^2$  against  $H_1: \sigma_Y^2 \neq \sigma_X^2$ .

We know:

► F-test [assume  $X_i \sim N(\mu_X, \sigma_X)$  and  $Y_i \sim N(\mu_Y, \sigma_Y)$ ]:  $\frac{s_Y^2}{s_X^2} \sim F_{n_Y-1, n_X-1}$

# F-test of equality of variances



We test  $H_0: \sigma_Y^2 = \sigma_X^2$  against  $H_1: \sigma_Y^2 \neq \sigma_X^2$ .

We know:

► F-test [assume  $X_i \sim N(\mu_X, \sigma_X)$  and  $Y_i \sim N(\mu_Y, \sigma_Y)$ ]:  $\frac{s_Y^2}{s_X^2} \sim F_{n_Y-1, n_X-1}$

F test to compare two variances

```
data: golub[1042, gol.fac == "ALL"] and golub[1042, gol.fac == "AML"]
F = 0.71164, num df = 26, denom df = 10, p-value = 0.4652
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2127735 1.8428387
sample estimates:
ratio of variances
 0.7116441
```

# Warning

## Multiplicity correction

For each test, the probability of rejecting  $H_0$  (and accept  $H_1$ ) when  $H_0$  is true equals  $\alpha$ .

For  $k$  tests, the probability of rejecting  $H_0$  (and accept  $H_1$ ) at least 1 time when  $H_0$  is true,  $\alpha_k$ , is given by

$$\alpha_k = 1 - (1 - \alpha)^k.$$

Thus, for  $\alpha = 0.05$ ,

- ▶ if  $k = 1$ ,  $\alpha_1 = 1 - (1 - \alpha)^1 = 0.05$ ,
- ▶ if  $k = 2$ ,  $\alpha_2 = 1 - (1 - \alpha)^2 = 0.0975$ ,
- ▶ if  $k = 10$ ,  $\alpha_{10} = 1 - (1 - \alpha)^{10} = 0.4013$ .

Idea: change the level of each test so that  $\alpha_k = 0.05$ :

- ▶ Bonferroni correction :  $\alpha = \frac{\alpha_k}{k}$ ,
- ▶ Dunn-Sidak correction:  $\alpha = 1 - (1 - \alpha_k)^{1/k}$ .

# Warning

## Non-parametric is not assumption free: Type I error

Simulate 2500 samples with

- ▶  $X_i \sim \text{Uniform}(1.5, 2.5)$ ,  $i = 1, \dots, n_X$ ,
- ▶  $Y_i \sim \text{Uniform}(0, 4)$ ,  $i = 1, \dots, n_Y$ ,

so that  $E[X_i] = E[Y_i] = 2$  (i.e., same mean, same median).

Assume

- ▶  $X_i \sim \text{iid}(\mu_X, \sigma^2)$ ,  $i = 1, \dots, n_X$ ,
- ▶  $Y_i \sim \text{iid}(\mu_X + \delta, \sigma^2)$ ,  $i = 1, \dots, n_Y$ .

Test **H0**:  $\delta = \delta_0$  against **H1**:  $\delta \neq \delta_0$ , at the 5% level, by means of

- ▶ Mann-Whitney-Wilcoxon test (MWW),
- ▶ T-test,
- ▶ Welch-test.

$\hat{\alpha}$		Tests		
Sample size		MWW	Student's t-test	Welch's test
	$n_X = 200, n_Y = 70$	0.145	0.202	0.055
	$n_X = 20, n_Y = 7$	0.148	0.240	0.062



# Exercises