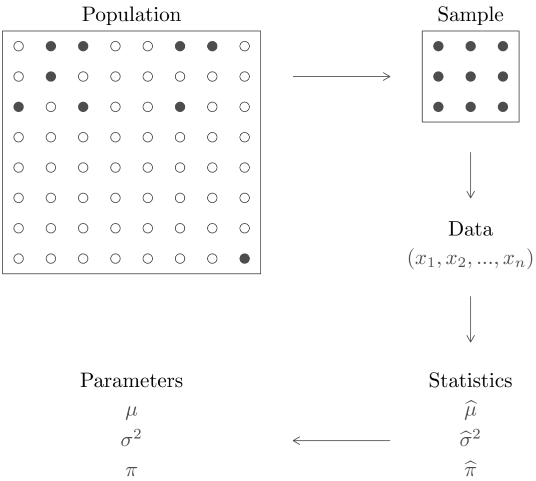
Introduction to Statistical Analysis

Cancer Research UK – 31st of January 2022

D.-L. Couturier & M. Eldridge (Bioinformatics core)

Grand Picture of Statistics



Data Types

	x_1	x_2	x_3	$\cdots$	x_n
Cancer status	C	$\not{C}$	$\not{C}$	$\cdots$	C
Nucleic acid sequence	C	T	T	$\cdots$	A
5-level pain score	3	1	5	$\cdots$	4
# of daily admissions at A&E	16	23	12	$\cdots$	17
Gene expression intensity	882.1	379.5	528.3	$\cdots$	120.9

Data Types

Two main types of data:

- Qualitative: characteristics identified by names/categories
  - binary/dichotomous data: 2 categories,
    Cancer status (Y/N)
  - categorical data: >2 categories ,
    Nucleic acid, country of birth, ethnic group
  - ordinal data: >=2 ordered categories
    stages of breast cancer (I, II, III, or IV),
    5-level pain score (minimal, moderate, severe or unbearable)

- Quantitative: expressed numerically
  - discrete (natural number)
    number of metastases, # of daily admissions at A&E,
  - continuous (real number)
    Gene expression intensity, body mass index, blood pressure

Data Types

Main properties:

- mutual exclusivity,
- rank,
- equidistance.

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

Summary statistics per data type

Two main types of summary statistics:

- Typical value:
  - qualitative:
    - mode,
  - quantitative:
    - mean,
    - median.
- Typical variability around the typical value:
  - qualitative:
    - none
  - quantitative:
    - standard deviation, variance
    - median absolute deviation,
    - inter-quartile range.

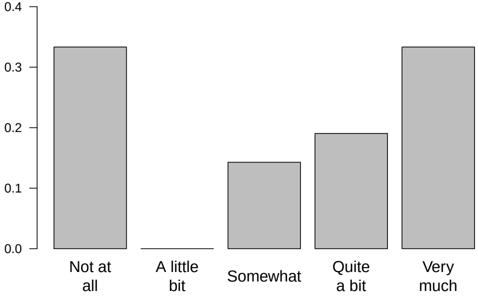
Summary statistics and plots for qualitative data

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 ?":

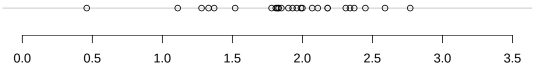
3, 1, 5, 3, 1, 1, 1, 5, 1, 3, 4, 1, 1, 4, 5, 5, 5, 5, 5, 4, 4,

where

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



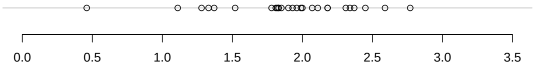
Summary statistics and plots for quantative data



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

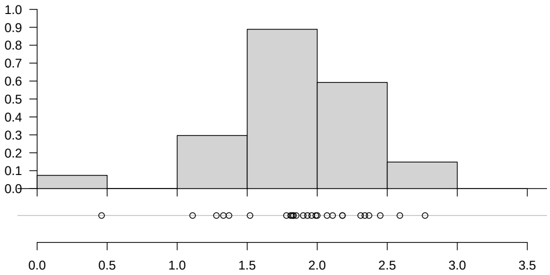
Summary statistics and plots for quantative data



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

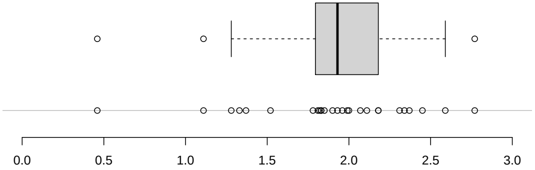
Summary statistics and plots for quantative data



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

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

https://docs.google.com/forms/d/1C3RHisRHoWXcnFqX9JhRAk3gy_aJ6FrhouJ6ljsJ-Fc

Statistical distributions

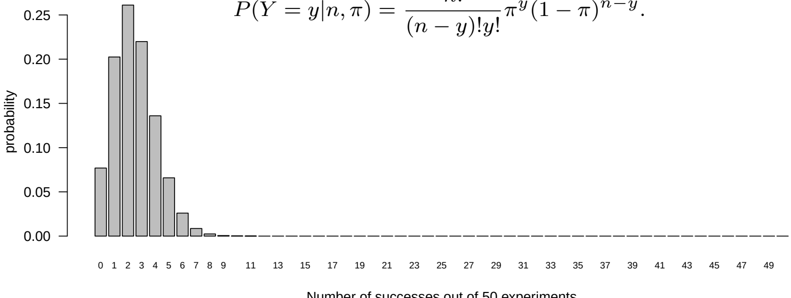
"In probability theory and statistics, a statistical distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment" [Wikipedia].

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?

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



Some parametric distributions: Binomial distribution

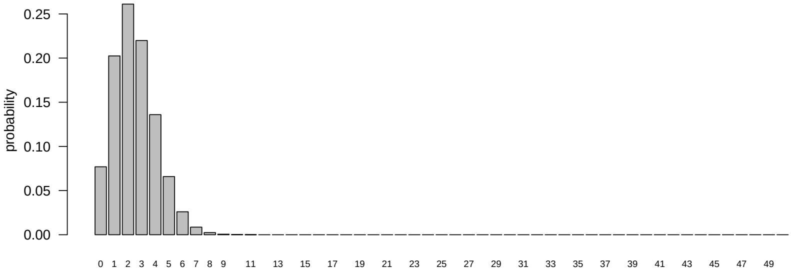
- the number of successes out of n trials (experiments), $Y = \sum_{i=1}^{n} X_i$, follows a binomial distribution with parameters n and π:

$$Y \sim Bin(n, \pi),$$

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

IF

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



Some parametric distributions: Poisson distribution

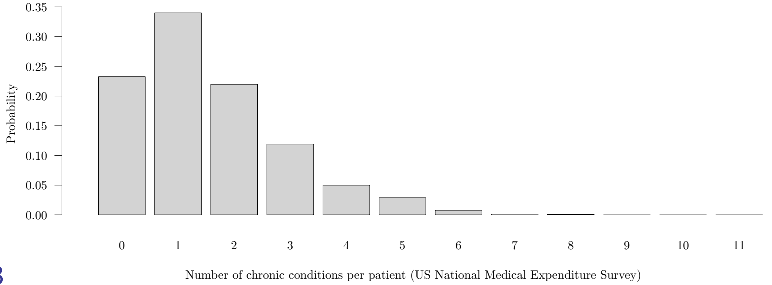
- the number of events occurring in a fixed time interval or in a given area, X, may be modelled by means of a Poisson distribution with parameter λ:

$$X \sim Poisson(\lambda),$$

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

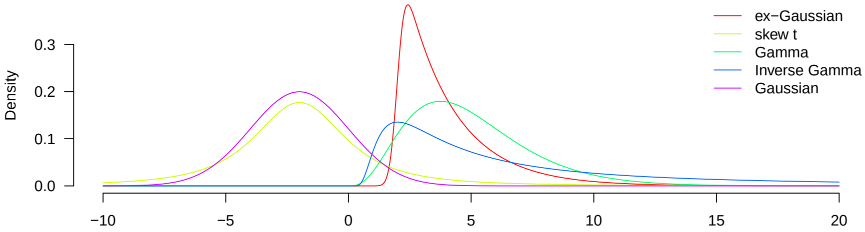
IF, during a time interval or in a given area,

- 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),



CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Some parametric distributions: Continuous distrib.



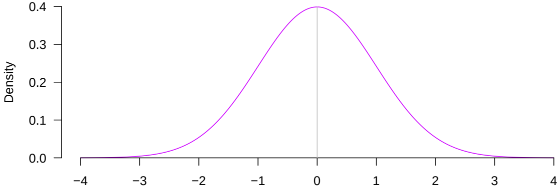
Some parametric distributions: Normal distribution

$$X \sim N(\mu, \sigma^2)$$

$$\mathsf{E}[X] = \mu, \qquad \mathsf{Var}[X] = \sigma^2,$$

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

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



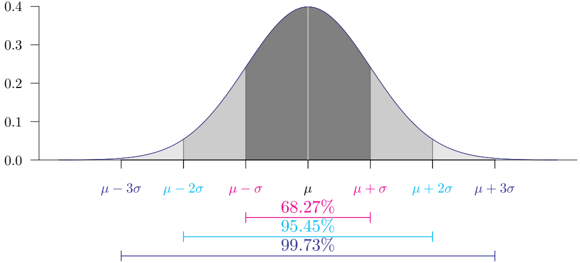
Some parametric distributions: Normal distribution

$$X \sim N(\mu, \sigma^2)$$

$$\mathsf{E}[X] = \mu, \qquad \mathsf{Var}[X] = \sigma^2,$$

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

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



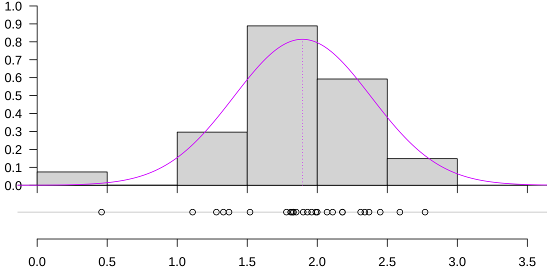
Some parametric distributions: Normal distribution

$$X \sim N(\mu, \sigma^2)$$

$$\mathsf{E}[X] = \mu, \qquad \mathsf{Var}[X] = \sigma^2,$$

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

(i) Suitable modelling for a lot of variables:



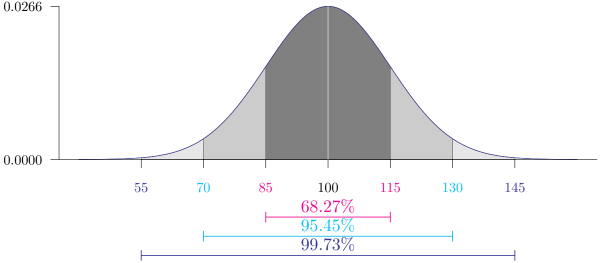
Some parametric distributions: Normal distribution

$$X \sim N(\mu, \sigma^2)$$

$$\mathsf{E}[X] = \mu, \qquad \mathsf{Var}[X] = \sigma^2,$$

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

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



Some parametric distributions: Normal distribution

$$X \sim N(\mu, \sigma^2)$$

$$\mathsf{E}[X] = \mu, \qquad \mathsf{Var}[X] = \sigma^2,$$

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

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

- Let $(X_1, ..., X_n)$ be n independent and identically distributed (iid) random variables drawn from distributions of expected values given by μ and finite variances given by σ^2,
- then

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

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 μ, the population mean, when $X_i \sim N(\mu, \sigma^2)$ and σ is known or n is large

we know that $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ so that $Z = \frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$.

Therefore,

$$\mathrm{P}\left(\qquad < \qquad < \qquad \right) = 0.95$$

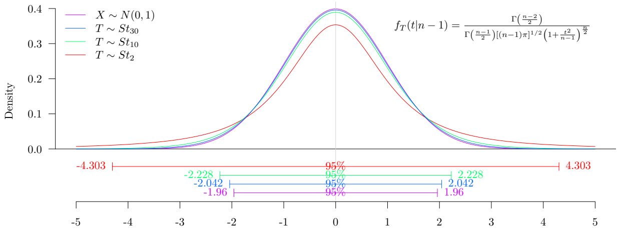
$$\mathrm{P}\left(\qquad < \qquad < \qquad \right) = 0.95$$

$$\mathrm{P}\left(\qquad < \qquad < \qquad \right) = 0.95$$

95% Confidence interval for μ, the population mean, when $X_i \sim N(\mu, \sigma^2)$, σ is unknown and n is small/moderate

in this case, we can show that $Z = \frac{\overline{X} - \mu}{\sqrt{\frac{s^2}{n}}} \sim St_{n-1}$. Therefore,

$$\mathsf{P}\left(\qquad\qquad < \qquad\qquad < \qquad\qquad \right) = 0.95$$



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

If we have

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

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

If we have

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

then

- if $\sigma_X^2 = \sigma_Y^2$ [Student's t-test equation],
  - $CI\,(\mu_Y - \mu_X, 0.95) = (\overline{Y} - \overline{X}) \pm t_{1-\frac{\alpha}{2}, n_X+n_Y-2} s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}$

    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

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

then

- if $\sigma_X^2 = \sigma_Y^2$ [Student's t-test equation],
  - $CI\left(\mu_Y - \mu_X, 0.95\right) = (\overline{Y} - \overline{X}) \pm t_{1-\frac{\alpha}{2}, n_X + n_Y - 2} s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}$
    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],
  - $CI\left(\mu_Y - \mu_X, 0.95\right) = (\overline{Y} - \overline{X}) \pm t_{1-\frac{\alpha}{2}, \text{df}} \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$, where
    $$\text{df} = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{\left(\frac{s_X^2}{n_X}\right)^2}{n_X - 1} + \frac{\left(\frac{s_Y^2}{n_Y}\right)^2}{n_Y - 1}}.$$

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

PART II:
Parametric and non-parametric one-sample location tests

Cancer Research UK – 31^{st} of January 2022

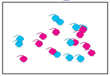
D.-L. Couturier & M. Eldridge (Bioinformatics core)

Grand Picture of Statistics

Statistical Hypotheses

H0: $\mu_{Tamoxifen} = \mu_{Control}$
H1: $\mu_{Tamoxifen} < \mu_{Control}$

Sample



Idea:
Tamoxifen represses the progression of ER+ Breast cancer

Data: Tumour size at day 42

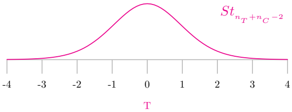
$$(x_{T,1}; x_{T,2}; ...; x_{T,n_T})$$
$$(x_{C,1}; x_{C,2}; ...; x_{C,n_T})$$

Point estimation

$$\widehat{\mu}_{Tamoxifen} - \widehat{\mu}_{Control}$$

Inference: Under H0

$$T_{obs} = \frac{\widehat{\mu}_{Tamoxifen} - \widehat{\mu}_{Control}}{s_p \sqrt{\frac{1}{n_T} + \frac{1}{n_C}}} \sim St_{n_T + n_C - 2}$$



$$p - value = P(T < T_{obs})$$

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 α of being wrong, that the idealised model/description of H1 is true.

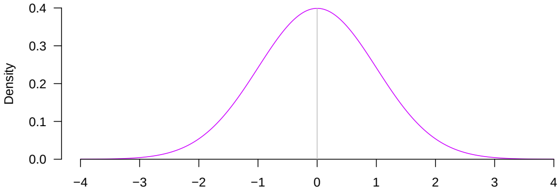
Theory 1: Dieters lose more fat than the exercisers

Theory 2: There is no majority for Brexit now

Theory 3: Serum vitamin C is reduced in patients

Test statistic distribution under H0 and p-value

Compare the observed test statistics, Z_{obs}, to its distribution under H0 to assess how likely it is to observe such a value if there is no effect:



P-value for a two-sided test:
p-value $= 2 \ \text{min} \left[P(Z \leq Z_{obs}|\text{H0}), P(Z \geq Z_{obs}|\text{H0})\right]$
i.e. the probability of observing a test statistic which is as extreme or more extreme than the observed one if H0 is true

Conclusion at the α level

Conclude:

- if p-value $> \alpha$ $\rightarrow$ do not reject H0.
- if p-value $< \alpha$ $\rightarrow$ reject H0 in favour of H1.

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

where

- α is the Type I error, the probability of rejecting H0 when H0 is correct,
- β is the Type II error, the probability of not rejecting H0 when H1 is correct.

Warnings

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

α level, the Type I error

Definition:

- the (pre-defined) probability of rejecting H0 when H0 is correct,
- probability of finding an effect when there is none.

The **Type I error** occurs when the random sampling lead to a difference/association/correlation large enough to be a statistically significant. It is a false positive [FP].

Choice of α level:

- $\alpha = 0.05 = 1/20$ used as convention in many scientific fields

  'It is convenient to draw the line at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials". If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point)'.

  (Fisher, R., 1935)

- $\alpha = 0.005 = 1/200$ often suggested

  'We propose to change the default P value threshold for statistical significance for claims of new discoveries from 0.05 to 0.005'.

  (Benjamin, D.J. et al, 2017, Redefine statistical significance)

- $\alpha = 0.0000003 =$1 in 3.5 million used to claim the discovery of Higgs boson

β level, the Type II error

Definition:

- the probability of not rejecting H0 when H1 is correct,
- probability of not detecting an effect when there is one.

The **Type II error** occurs when the random sampling doesn't lead to a difference/association/correlation large enough to be a statistically significant. It is a false negative [FN].

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Statistical hypothesis testing steps

Several-step process:

- Define H0 and H1 according to a theory
- Set α, the probability of rejecting H0 when it is true (Type I error),
- Determine the test statistic to be used,
- Define n, the sample size, allowing you to reject H0 when H1 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.

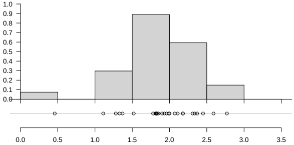
Parametric location test

Student's test

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

$$X_i = \mu + e_i,$$

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



Interest for **H0**: $\mu = \mu_0$ against **H1**: $\mu < \mu_0$ or $\mu \neq \mu_0$ or $\mu > \mu_0$.

Test statistics : $T = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$.

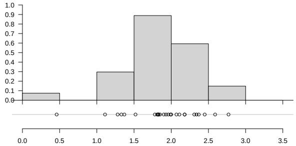
Parametric location test

Student's test

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

$$X_i = \mu + e_i,$$

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



Interest for **H0**: $\mu = \mu_0$ against **H1**: $\mu < \mu_0$ or $\mu \neq \mu_0$ or $\mu > \mu_0$.

Test statistics : $T = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$.

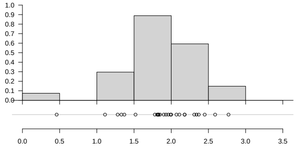
Distribution of W under H0: $T \sim Student(df = n-1)$.

```
One Sample t-test

data:  golub[1042, gol.fac == "ALL"]
t = 4.172, df = 26, p-value = 0.0002982
alternative hypothesis: true mean is not equal to 1.5
95 percent confidence interval:
 1.699817 2.087948
sample estimates:
mean of x
 1.893883
```

Non-parametric location test

Wilcoxon sign-rank test



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

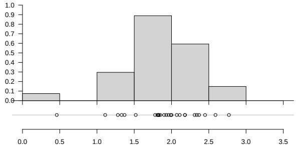
$$X_i = \theta + e_i,$$

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

Non-parametric location test
Wilcoxon sign-rank test



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

$$X_i = \theta + e_i,$$

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

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

```
Wilcoxon signed rank exact test

data:  golub[1042, gol.fac == "ALL"]
V = 333, p-value = 0.0002363
alternative hypothesis: true location is not equal to 1.5
95 percent confidence interval:
 1.73868 2.09106
sample estimates:
(pseudo)median
      1.926475
```


CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

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

Introduction to Shiny Apps and Exercises

PART III: Parametric and non-parametric two-sample location tests

Cancer Research UK – 31^{st} of January 2022

D.-L. Couturier & M. Eldridge (Bioinformatics core)

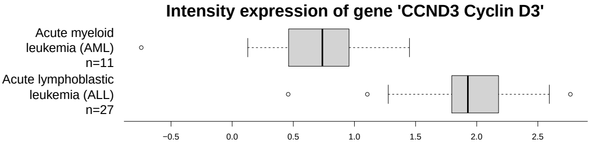
Two-sample case

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



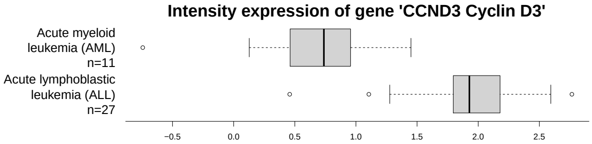
We test **H0**: $\mu_Y - \mu_X = 0$ against **H1**: $\mu_Y - \mu_X \neq 0$.

We know:

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

Parametric two-sample location test

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



We test **H0**: $\mu_Y - \mu_X = 0$ against **H1**: $\mu_Y - \mu_X \neq 0$.

We know:

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

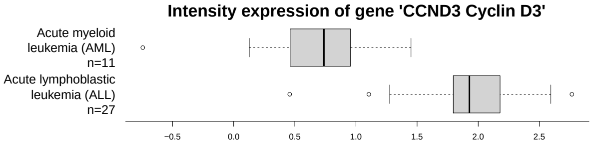
```
Two Sample t-test

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
```


CANCER RESEARCH UK CAMBRIDGE INSTITUTE

Parametric two-sample location test

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



We test **H0**: $\mu_Y - \mu_X = 0$ against **H1**: $\mu_Y - \mu_X \neq 0$.

We know:

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

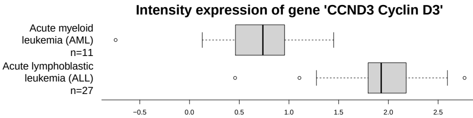
```
Welch Two Sample t-test

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

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



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

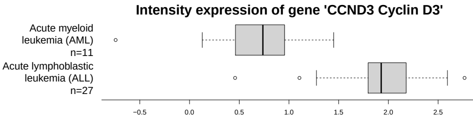
Standardised test statistic: $z = \frac{\sum_{i=1}^{n_Y} R(Y_i) - [n_Y(n_X+n_Y+1)/2]}{\sqrt{n_X n_Y (n_X+n_Y+1)/12}}$,

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

Non-parametric two-sample location test
Mann-Whitney-Wilcoxon test

Let

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

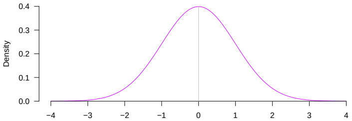


Interest for **H0**: $\delta = \delta_0$ against **H1**: $\delta < \delta_0$ or $\delta \neq \delta_0$ or $\delta > \delta_0$.

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 $(X_1, ..., X_{n_X}, Y_1, ..., Y_{n_Y})$.

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

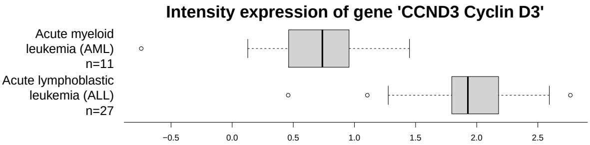



```
Implementation 1:
statistic =  -4.361334 , p-value =  1.292716e-05

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
```


CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

F-test of equality of variances

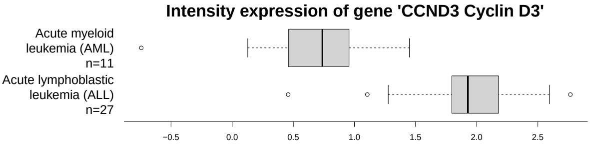


We test **H0**: $\sigma_Y^2 = \sigma_X^2$ against **H1**: $\sigma_Y^2 \neq \sigma_X^2$.

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 **H0**: $\sigma_Y^2 = \sigma_X^2$ against **H1**: $\sigma_Y^2 \neq \sigma_X^2$.

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 H0 (and accept H1) when H0 is true equals α.

For k tests, the probability of rejecting H0 (and accept H1) at least 1 time when H0 is true, α_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

- $X_i \sim Uniform(1.5, 2.5),\ i = 1, ..., n_X,$
- $Y_i \sim Uniform(0, 4),\ i = 1, ..., n_Y,$

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

Assume

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

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.

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

Exercises