Introduction to Statistical Analysis

Cancer Research UK – 24^{th} of April 2017

D.-L. Couturier / M. Dunning / M. Eldridge [Bioinformatics core]

Timeline

10:30 – Introduction

- $\sim$ 45mn Lecture
- $\sim$ 15mn Quiz

11:30 – Parametric tests

- $\sim$ 30mn Lecture
- $\sim$ 30mn Exercises

12:30 – One-hour lunch break

13:30 – Non-parametric tests

- $\sim$ 30mn Lecture
- $\sim$ 30mn Exercises

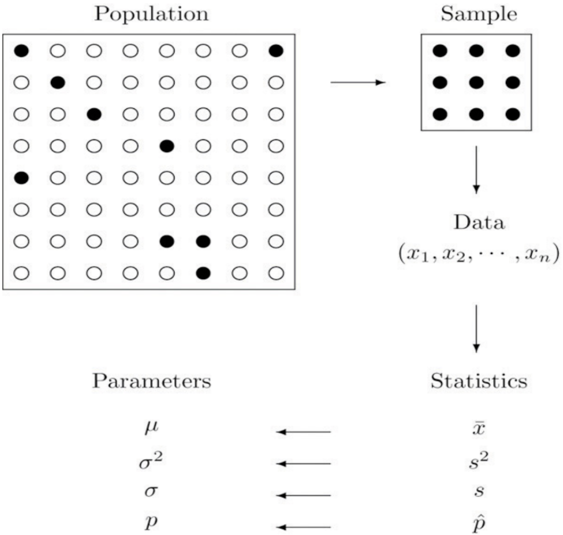
14:30 – Tests for categorical variables

- $\sim$ 15mn Lecture
- $\sim$ 45mn Exercises

15:30 – Group based exercises

- $\sim$ 60mn

Grand Picture of Statistics



Data Types

	x_1	x_2	x_3	$\cdots$	x_n
Cancer status	C	$\not\mathrm{C}$	$\not\mathrm{C}$	$\cdots$	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 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

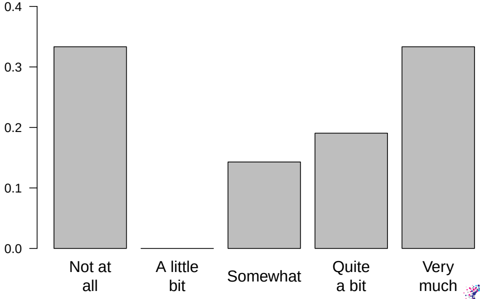
5-level answers of 21 patients to the question

"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".

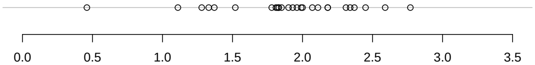


Summary statistics and plots for quantative data

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

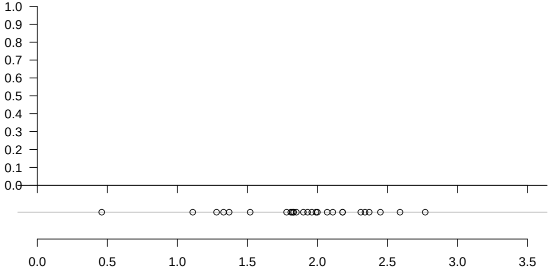
Summary statistics and plots for quantative data



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

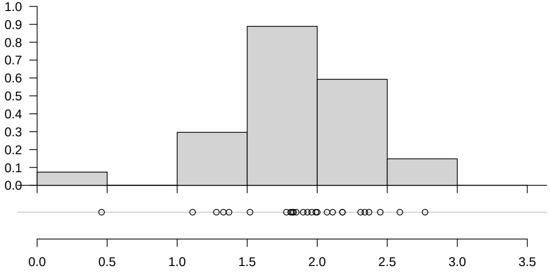
Summary statistics and plots for quantative data



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

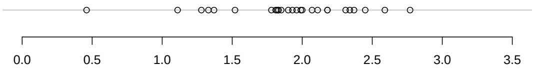
Summary statistics and plots for quantative data



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

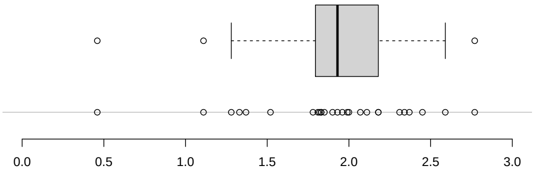
Summary statistics and plots for quantative data



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 for independent/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

Summary statistics for independent/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

Some parametric distributions: Bernoulli distribution

	x_1	x_2	x_3	$\cdots$	x_n
Cancer status	C	$\not{C}$	$\not{C}$	$\cdots$	C
	1	0	0	$\cdots$	1

If

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

then, each dichotomous experiment, X_i, follows a Bernoulli distribution with parameter π:

$$X_i \sim Bernoulli(\pi)$$

$$P(X_i = 1) = \pi$$

$$P(X_i = 0) = 1 - \pi$$

Some parametric distributions: Binomial distribution

If

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

then,

- 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:

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

- the probability of observing exactly y successes out of n experiments, is given by

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

Some parametric distributions: Binomial distribution

If

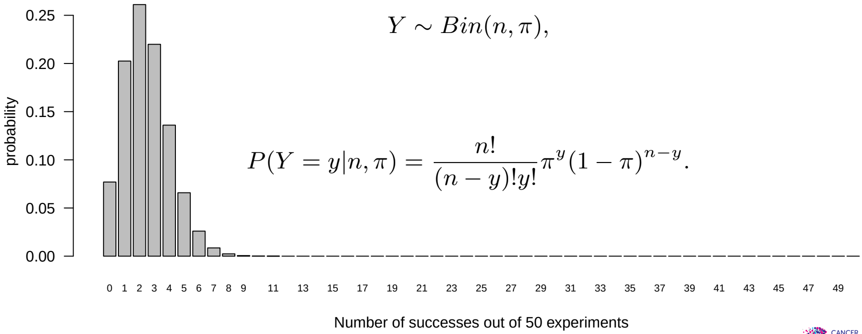
- n independent experiments,
- outcome of each experiment is dichotomous (success/failure),
- the probability of success π is the same for all experiments,

then,

- 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:

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

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



Some parametric distributions: Poisson distribution

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),

then,

- 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:

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

- the probability of observing x during a fixed time interval or in a given area is given by

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

Some parametric distributions: Poisson distribution

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

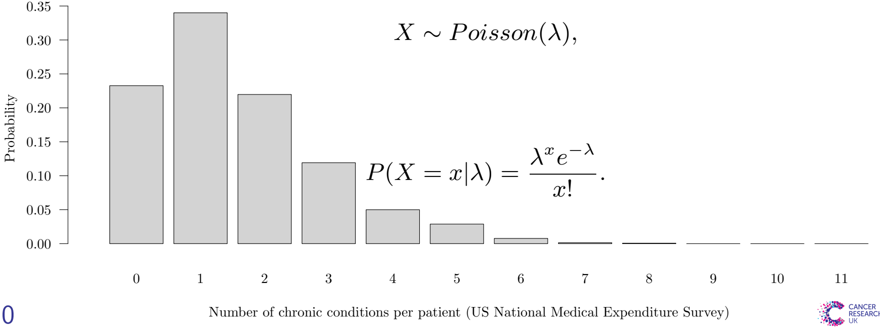
- events occur independently,
- at the same rate,
- and the probability of an event to occur in a small interval (area) is proportional to the length of the interval (size of the area),

then,

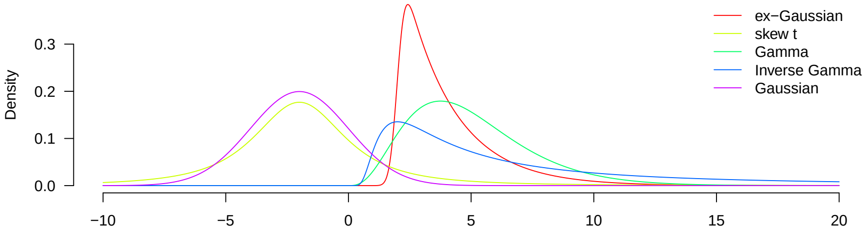
- the number of events occurring in a fixed time interval or in a given area, X, may be modelled by means of a Poisson distribution with parameter λ:

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

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



Some parametric distributions: Continuous distrib.



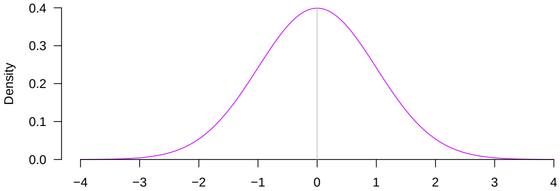
Some parametric distributions: Normal distribution

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

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

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

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



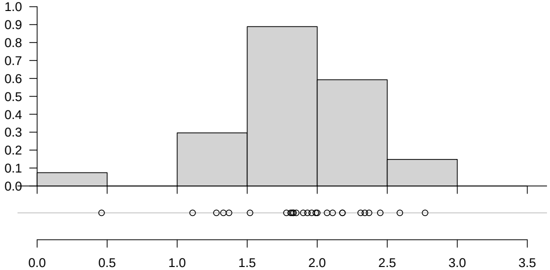
Some parametric distributions: Normal distribution

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

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

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

(i) Suitable modelling for a lot of variables



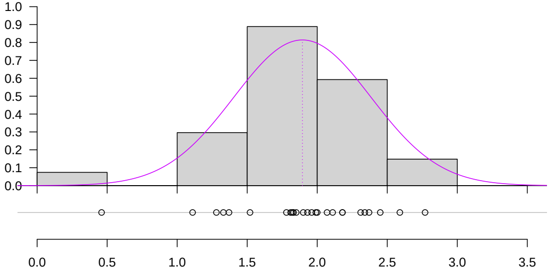
Some parametric distributions: Normal distribution

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

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

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

(i) Suitable modelling for a lot of variables



Some parametric distributions: Normal distribution

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

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

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

(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

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

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

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

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

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

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

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

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

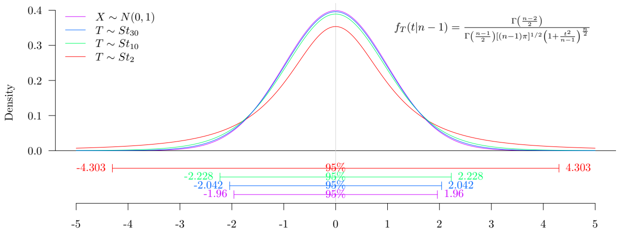
- if $X \sim N(\mu, \sigma^2)$, then $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$,
- if $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$,
- if σ unknown, then $T = \frac{X-\mu}{s} \sim St_{n-1}$.

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

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

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

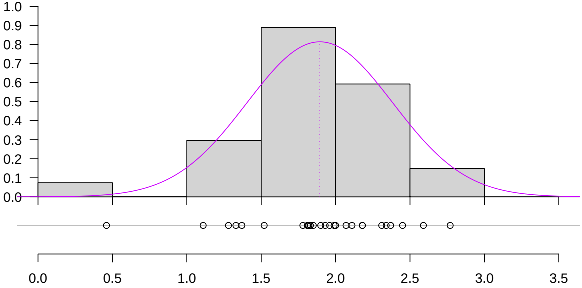
$$\mathrm{P}\left(\qquad < \qquad < \qquad \right) = 0.95$$



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

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

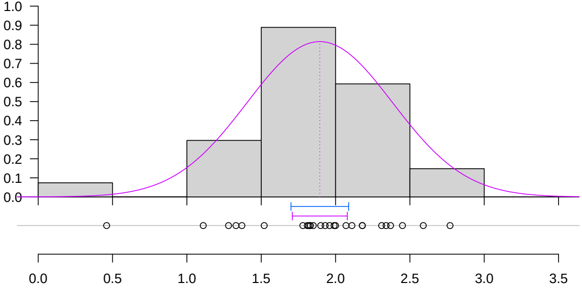
$$\text{P}\left(\qquad < \qquad < \qquad \right) = 0.95$$



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

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

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



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

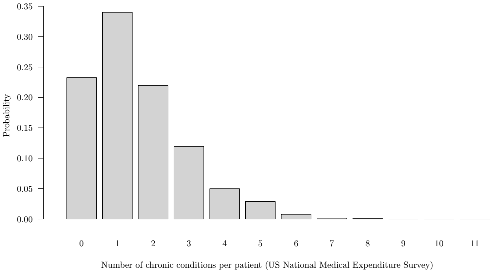
- CLT: $\overline{X} \quad \xrightarrow{d} \quad N\left(\mu, \frac{\sigma^2}{n}\right)$,
- if $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$,
- if σ unknown, then $T = \frac{X-\mu}{s} \sim St_{n-1}$.

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

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

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

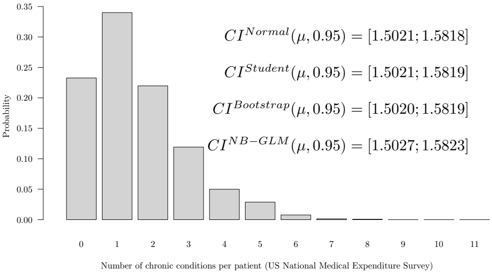
$$\text{P}\left(\qquad < \qquad < \qquad \right) = 0.95$$



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

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

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



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 $\sigma_X^2 = \sigma_Y^2$ [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 $s_p = \frac{(n_X-1)s_X^2 + (n_Y-1)s_Y^2}{n_X+n_Y-2}$,

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 $\sigma_X^2 = \sigma_Y^2$ [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 $s_p = \frac{(n_X-1)s_X^2 + (n_Y-1)s_Y^2}{n_X+n_Y-2}$,
- if $\sigma_X^2 \neq \sigma_Y^2$ [Welch-Satterthwaite equation],
  - $CI\,(\mu_Y - \mu_X, 0.95) = (\overline{Y} - \overline{X}) \pm t_{1-\frac{\alpha}{2}, \mathrm{df}} \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$, where

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

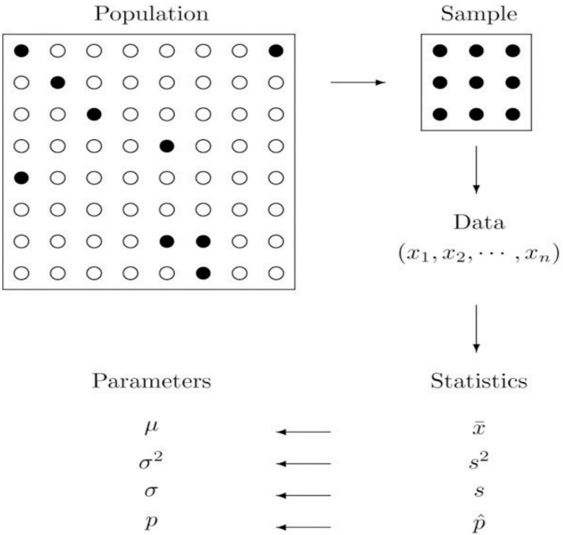
Quiz Time

PART II: Parametric tests

Cancer Research UK – 24^{th} of April 2017

D.-L. Couturier / M. Dunning / M. Eldridge [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),
- the alternative hypothesis (H1).

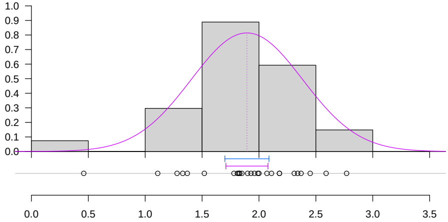
The aim of the test is to reject the null hypothesis in favour of the alternative hypothesis, and conclude, with a probability α of being wrong, that the idealised model/description of H1 is true.

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),
- Define n, the sample size, allowing you to reject H0 when H1 is true with a probability $1 - \beta$ (Power),
- Determine the test statistic to be used,
- Collect the data,
- Perform the statistical test and reject (or not) the null hypothesis.

Statistical hypothesis testing
Example: One-sample two-sided t-test

We test:
H0: $\mu = \mu_0$,
H1: $\mu \neq \mu_0$.



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

From the CLT, we know

- $\overline{X} \quad \xrightarrow{d} \quad N\left(\mu, \frac{\sigma^2}{n}\right),$
- $Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1),$
- $T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \sim St_{n-1}.$

Thus, if H0 is true, we have:

- $T = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}} \sim St_{n-1}.$

Statistical hypothesis testing
Example: One-sample two-sided t-test

We test:
H0: $\mu = \mu_0$,
H1: $\mu \neq \mu_0$.

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

From the CLT, we know

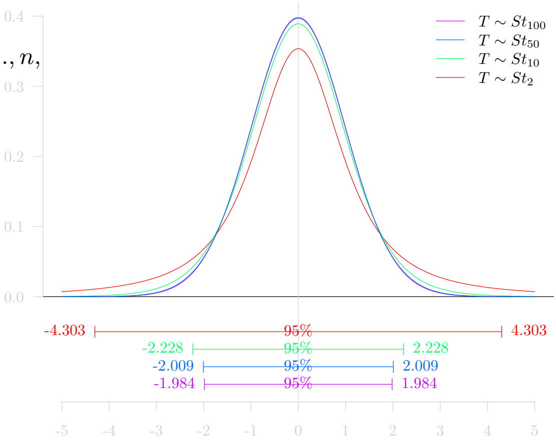
- $\overline{X} \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right)$,
- $Z = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$,
- $T = \frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}} \sim St_{n-1}$.

Thus, if H0 is true, we have:

- $T = \frac{\overline{X}-\mu_0}{\frac{s}{\sqrt{n}}} \sim St_{n-1}$.

Define the p-value:

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



Statistical hypothesis testing
4 possible outcomes

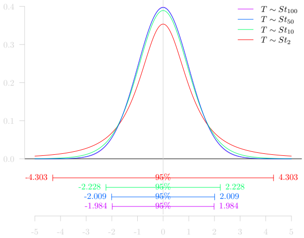
Conclude:

- 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	$1-\alpha$	α
	H1 true	β	$1-\beta$

where

- α is the type I error,
- β is the type II error.



Statistical hypothesis testing
Example: One-sided binomial exact test

We test:
H0: $\pi = 5\%$,
H1: $\pi > 5\%$.

We have $X_i \sim Bernoulli(\pi), i = 1, ..., n,$

We know

- $Y = \sum_{i=1}^{n} X_i \quad \sim \quad Binomial\left(\pi, n\right),$

Thus, if H0 is true, we have:

- $Y = \sum_{i=1}^{n} X_i \quad \sim \quad Binomial\left(5\%, n\right),$

Statistical hypothesis testing
Example: One-sided binomial exact test

We test:
H0: $\pi = 5\%$,
H1: $\pi > 5\%$.

We have $X_i \sim Bernoulli(\pi), i = 1, ..., n,$

We know

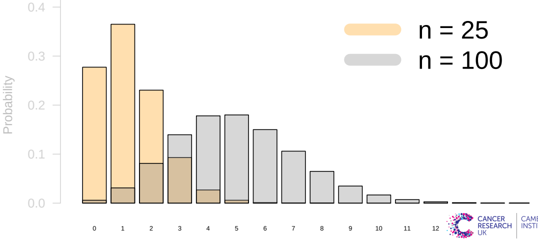
- $Y = \sum_{i=1}^{n} X_i \quad \sim \quad Binomial\left(\pi, n\right),$

Thus, if H0 is true, we have:

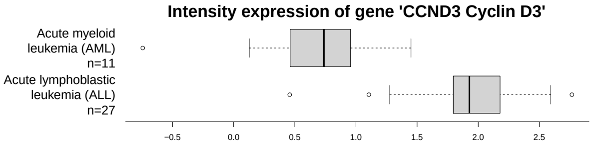
- $Y = \sum_{i=1}^{n} X_i \quad \sim \quad Binomial\left(5\%, n\right),$

Define the p-value:

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



Two-sample two-sided t-test & Welch test



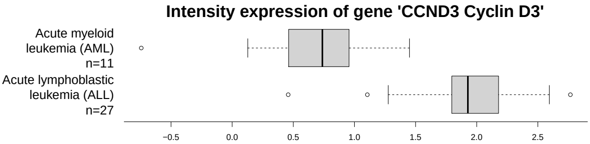
We test **H0**: $\mu_Y - \mu_X = 0$ against **H1**: $\mu_Y - \mu_X \neq 0$.

We know:

- 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-test [assume $\sigma_X^2 \neq \sigma_Y^2$]: $\frac{(\overline{Y}-\overline{X})-(\mu_Y-\mu_X)}{\sqrt{\frac{s_X^2}{n_X}+\frac{s_Y^2}{n_Y}}} \sim t_{1-\frac{\alpha}{2},df}$

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Two-sample two-sided t-test & Welch test



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

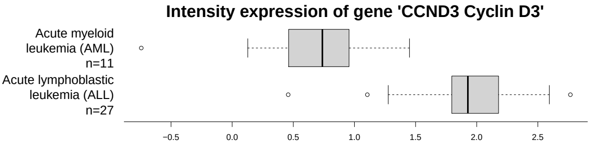
We know:

- 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-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
```

Two-sample two-sided t-test & Welch test



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

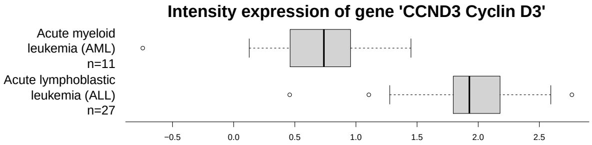
We know:

- 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-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
```

F-test of equality of variances

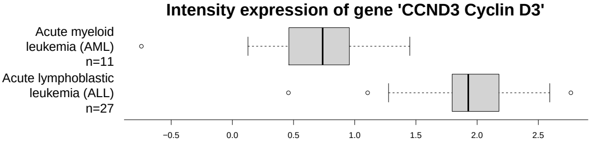


We test **H0**: $\sigma_Y^2 = \sigma_X^2$ against **H1**: $\sigma_Y^2 \neq \sigma_X^2$.

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
```

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: $\alpha = 1 - (1 - \alpha_k)^{1/k}$.

Introduction to Shiny Apps and Exercises

PART III: Non-parametric tests

Cancer Research UK – 24^{th} of April 2017

D.-L. Couturier / M. Dunning / M. Eldridge [Bioinformatics core]

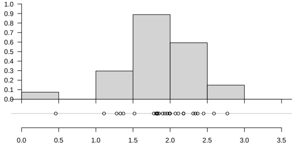
Parametric or non-parametric ?

T-test		**Outcome(s) normally distributed**		
		Yes	Mildly	No
Sample size	Small			
	Mild			
	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.

Sign test



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

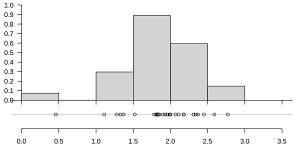
$$X_i = \theta + e_i,$$

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

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

Test statistics: $S = \sum_{i=1}^{n} \iota(X_i - \theta_0 > 0)$.

Sign test



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

$$X_i = \theta + e_i,$$

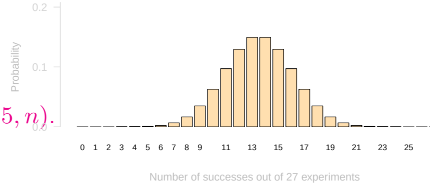
where $e_i \sim iid(\mu_e = 0, \sigma_e^2)$.

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

Test statistics: $S = \sum_{i=1}^{n} \iota(X_i - \theta_0 > 0)$.

Distribution of S under H0:

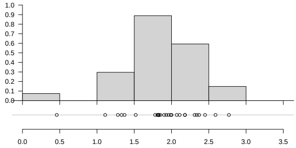
$$S \sim Binomial(0.5, n).$$




```
Exact binomial test

data:  21 and 27
number of successes = 21, number of trials = 27, p-value = 0.005925
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5774169 0.9137831
sample estimates:
probability of success
             0.7777778
```

Wilcoxon sign-rank test



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

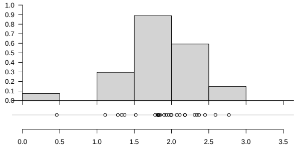
$$X_i = \theta + e_i,$$

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

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$(|X_i - \theta_0|)$.

Wilcoxon sign-rank test



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

$$X_i = \theta + e_i,$$

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

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.

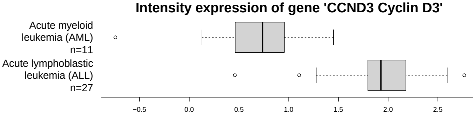
```
Wilcoxon signed rank test

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
```

Mann-Whitney-Wilcoxon test: Shift in location

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 $\delta > \delta_0$.

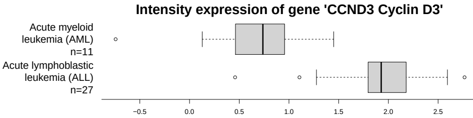
Standardised test statistic: $z = \frac{\sum_{i=1}^{n_Y} R(Y_i) - [n_Y(n_X+n_Y+1)/2]}{\sqrt{n_X n_Y (n_X+n_Y+1)/12}}$,

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

Mann-Whitney-Wilcoxon test: Shift in location

Let

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

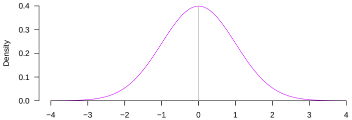


Interest for **H0**: $\delta = \delta_0$ against **H1**: $\delta < \delta_0$ or $\delta \neq \delta_0$ or $\delta > \delta_0$.

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
```

Non-parametric is not assumption free

Shift in location tests when H0 is true

Simulate 2500 samples with

- $X_i \sim Uniform(1.5, 2.5),\ i = 1, ..., n_X,$
- $Y_i \sim Uniform(-2, 6),\ i = 1, ..., n_Y,$

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),
- Fligner-Policello test (FP),
- T-test,
- Welch-test.

$\widehat{\alpha}$		**Tests**			
		MWW	F-P	t-test	Welch
Sample size	$n_X = 200, n_Y = 70$	0.145	0.056	0.202	0.055
	$n_X = 20, n_Y = 7$	0.148	0.120	0.240	0.062

Exercises

PART IV: Tests for categorical variables

Cancer Research UK – 24^{th} of April 2017

D.-L. Couturier / M. Dunning / M. Eldridge [Bioinformatics core]

χ^2 goodness-of-fit test

A trial to assess the effectiveness of a new treatment versus a placebo in reducing tumour size in patients with ovarian cancer:

Observed frequencies		Binary outcome		
		Tumour did not shrink	Tumour did shrink	
Group	Treatment	44	40	(84)
	Placebo	24	16	(40)
		(68)	(56)	(124)

- **H0** : No association between treatment group and tumour shrinkage,
- **H1** : Some association.

χ^2 goodness-of-fit test

A trial to assess the effectiveness of a new treatment versus a placebo in reducing tumour size in patients with ovarian cancer:

Observed frequencies		Binary outcome		
		Tumour did not shrink	Tumour did shrink	
Group	Treatment	44	40	(84)
	Placebo	24	16	(40)
		(68)	(56)	(124)

- **H0** : No association between treatment group and tumour shrinkage,
- **H1** : Some association.

Expected frequencies under H0		Binary outcome		
		Tumour did not shrink	Tumour did shrink	
Group	Treatment			(84)
	Placebo			(40)
		(68)	(56)	(124)

We have 2 categorical variables with a total of $J = 4$ cells (categories).

- **H0** : $\pi_j = \pi_{j_0}, j = 1, ..., J$,
- **H1** : $\pi_j \neq \pi_{j_0}, j = 1, ..., J$.

χ^2-test: $\sum_{j=1}^{J} \frac{(O_j - E_j)^2}{E_j} \sim \chi^2(J-1)$.

χ^2 goodness-of-fit test

A trial to assess the effectiveness of a new treatment versus a placebo in reducing tumour size in patients with ovarian cancer:

Observed frequencies		Binary outcome		
		Tumour did not shrink	Tumour did shrink	
Group	Treatment	44	40	(84)
	Placebo	24	16	(40)
		(68)	(56)	(124)

- **H0** : No association between treatment group and tumour shrinkage,
- **H1** : Some association.

Expected frequencies under H0		Binary outcome		
		Tumour did not shrink	Tumour did shrink	
Group	Treatment	$\frac{84 \times 68}{124} = 46.06$	$\frac{84 \times 58}{124} = 37.94$	(84)
	Placebo	$\frac{40 \times 68}{124} = 21.94$	$\frac{40 \times 56}{124} = 18.71$	(40)
		(68)	(56)	(124)

We have 2 categorical variables with a total of $J = 4$ cells (categories).

- **H0** : $\pi_j = \pi_{j_0}, j = 1, ..., J$,
- **H1** : $\pi_j \neq \pi_{j_0}, j = 1, ..., J$.

χ^2-test: $\sum_{j=1}^{J} \frac{(O_j - E_j)^2}{E_j} \sim \chi^2(J-1)$.

χ^2 goodness-of-fit test

A trial to assess the effectiveness of a new treatment versus a placebo in reducing tumour size in patients with ovarian cancer:

Observed frequencies		Binary outcome		
		Tumour did not shrink	Tumour did shrink	
Group	Treatment	44	40	(84)
	Placebo	24	16	(40)
		(68)	(56)	(124)

- **H0** : No association between treatment group and tumour shrinkage,
- **H1** : Some association.

Expected frequencies under H0		Binary outcome		
		Tumour did not shrink	Tumour did shrink	
Group	Treatment			(84)
	Placebo			(40)
		(68)	(56)	(124)

We have 2 categorical variables with a total of $J = 4$ cells (categories).

- **H0** : $\pi_j = \pi_{j_0}, j = 1, ..., J$,
- **H1** : $\pi_j \neq \pi_{j_0}, j = 1, ..., J$.

χ^2-test: $\sum_{j=1}^{J} \frac{(O_j - E_j)^2}{E_j} \sim \chi^2(J-1)$.

```
Pearson's Chi-squared test with Yates' continuity correction

data:  M
X-squared = 0.36474, df = 1, p-value = 0.5459
```

Fisher's exact test of independence

χ^2 goodness-of-fit test not suitable when

- n is small
- $E_j < 5$ for at least one cell.

Observed frequencies		Variable 1		
		Category 1	Category 2	
Variable 2	Category 1	a	b	(a+b)
	Category 2	c	d	(c+d)
		(a+c)	(b+d)	(a+b+c+d=n)

Fisher showed that, under H0 (independence),
$P(\text{observed table} \mid \text{H0}) = P(X = a)$ and $X \sim Hypergeometric(n, a + c, a + b)$.
To compute the Fisher's test:

- Define $P(X = a)$ for all possible tables having the observed marginal counts,
- Calculate the $p - value$ by defining the percentage of these tables that get a probability equal to or smaller than the one observed.

Fisher's exact test of independence

χ^2 goodness-of-fit test not suitable when

- n is small
- $E_j < 5$ for at least one cell.

Observed frequencies		Binary outcome		
		Tumour did not shrink	Tumour did shrink	
Group	Treatment	44	40	(84)
	Placebo	24	16	(40)
		(68)	(56)	(124)

Fisher showed that, under H0 (independence),
$P(\text{observed table} \mid \text{H0}) = P(X = a)$ and $X \sim Hypergeometric(n, a + c, a + b)$.
To compute the Fisher's test:

- Define $P(X = a)$ for all possible tables having the observed marginal counts,
- Calculate the $p - value$ by defining the percentage of these tables that get a probability equal to or smaller than the one observed.

```
Fisher's Exact Test for Count Data

data:  M
p-value = 0.4471
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.3160593 1.6790135
sample estimates:
odds ratio
 0.7351707
```