



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE



UNIVERSITY OF
CAMBRIDGE

tinyurl.com/
cruk-stats

Introduction to Statistical Analysis

Cancer Research UK – 23rd of October 2017

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

Timeline

9:30 – Morning

- ▶ ~ 45mn Lecture: **data type, summary statistics and graphical displays**
- ▶ ~ 15mn Quiz

10:30 – 15mn Coffee & Tea break

- ▶ ~ 60mn Lecture: **some statistical distributions + CLT**
- ▶ ~ 15mn Exercises & discussion

12:00 – Lunch break

13:00 – Afternoon

- ▶ ~ 45mn Lecture: **Parametric tests for the mean**
- ▶ ~ 30mn Exercises with shiny apps & discussion

- ▶ ~ 30mn Lecture: **Non-parametric tests for the mean**
- ▶ ~ 15mn Exercises & discussion

15:00 – 15mn Coffee & Tea break

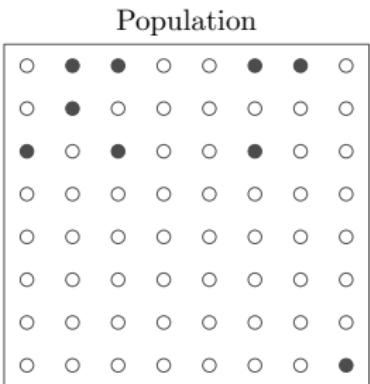
- ▶ ~ 15mn Lecture: **Tests for categorical variables**
- ▶ ~ 15mn Exercises & discussion

16:00 – Group based exercises

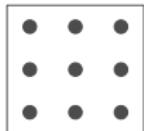
- ▶ ~ 60mn

Grand Picture of Statistics

UK SMOKERS



Sample



MEASUREMENT ERROR
MISSING VALUES

Data

(x_1, x_2, \dots, x_n)

$\text{C } \neq \neq$

POINT ESTIMATION

↓ OUTLIERS

ZERO INFLATED

Statistics

INFERENCE

$\hat{\mu}$

$\hat{\sigma}^2$

NO SUITABLE TEST

$$\hat{\pi} = \frac{1}{3} = 0.\overline{3}$$

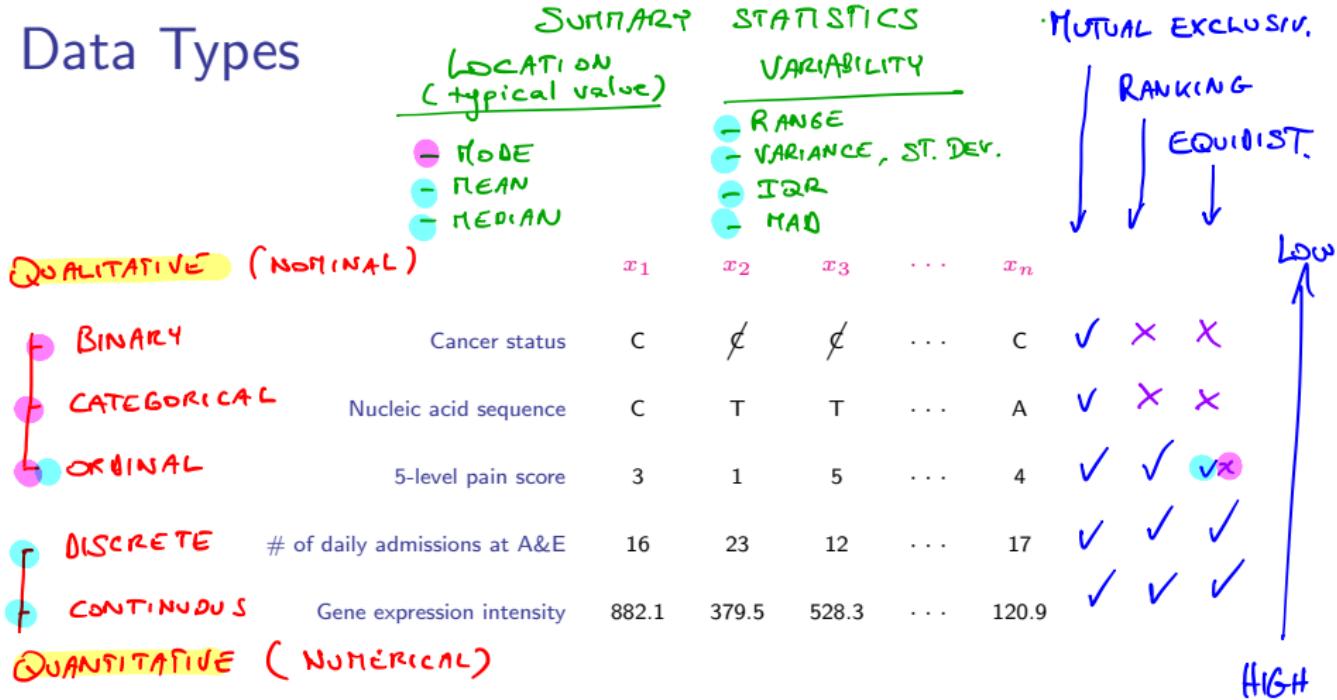
PROBABILITY OF
LUNG CANCER
AMONG SMOKERS

μ

σ^2

π

Data Types



GRAPHICAL DISPLAYS

- BARPLOT
 - HISTOGRAM
 - BOXPLOT

Summary statistics and plots for qualitative data

$$\hat{\pi}_c = \frac{\sum_{i=1}^n I(x_i = c)}{n} \quad \text{where } I(x_i = c) = 1 \text{ if } x_i = c \text{ and} \\ = 0 \text{ otherwise}$$

5-level answers of 21 patients to the question

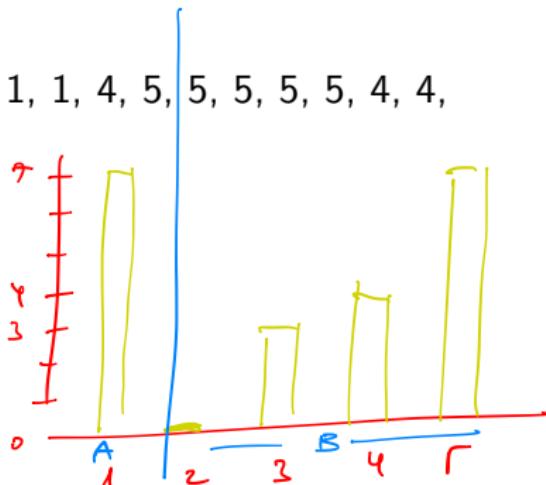
"How much did pain due to your ureteric stones interfere with your day to day activities ?":

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", 7/21
- 2 = "A little bit", 0/21
- 3 = "Somewhat", 3/21
- 4 = "Quite a bit", 4/21
- 5 = "Very much". 7/21

$$\sum = 1$$



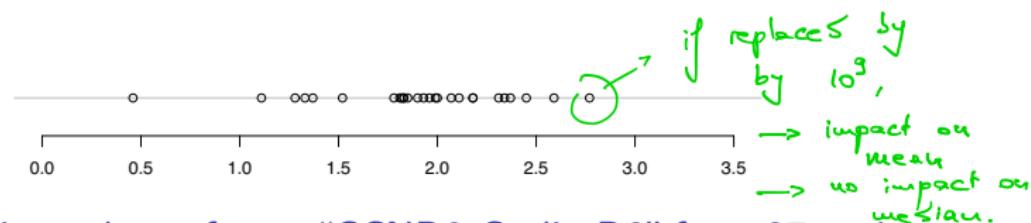
A: TWO - STAGE PROCESS :
INTERFERENCE: YES/NO MODE = YES
B: IF INTERFERENCE : INTERF. LEVEL MODE = 5

Summary statistics and plots for quantitative data

LOCATION:

$$\text{MEAN: } \hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 1.83$$

$$\text{MEDIAN: } \hat{med} = \hat{q}_{0.5} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{(x_{\frac{n}{2}} + x_{\frac{n}{2}+1})}{2} & \text{if } n \text{ is even} \end{cases} = 1.93$$



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

VARIABILITY

$$\text{RANGE} : x_{(u)} - x_{(1)} = 2.77 - 0.46 = 2.31$$

$$\text{VARIANCE} : \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{(0.46 - 1.83)^2 + (1.11 - 1.83)^2 + \dots}{n} = 0.5$$

$$\text{ST. DEV} : \hat{\sigma} = \sqrt{\hat{\sigma}^2} = (0.25)$$

$$\text{MEDIAN ABS. DEV} : \hat{MAD} = \text{med}(|x_i - \text{med}(x)|) = 0.25$$

$$\text{INTERQUARTILE RANGE} : IQR = \hat{Q}_{0.75} - \hat{Q}_{0.25}$$

ROBUST TO OUTLIERS

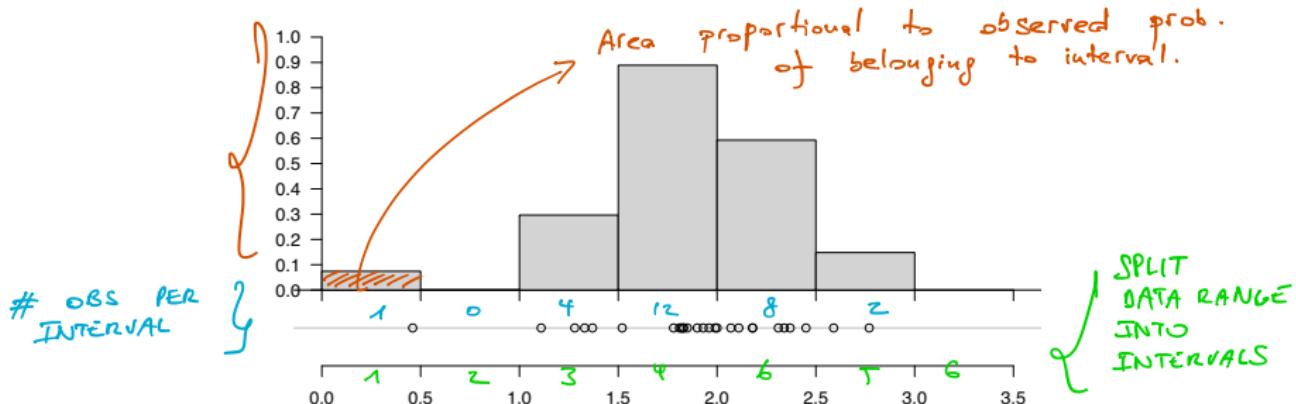


Gene expression values of gene "CCND3 Cyclin D3" from 27 patients diagnosed with acute lymphoblastic leukaemia:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$
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

HISTOGRAM

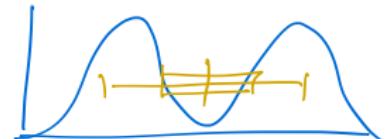


Gene expression values of gene "CCND3 Cyclin D3" from 27 patients diagnosed with acute lymphoblastic leukaemia:

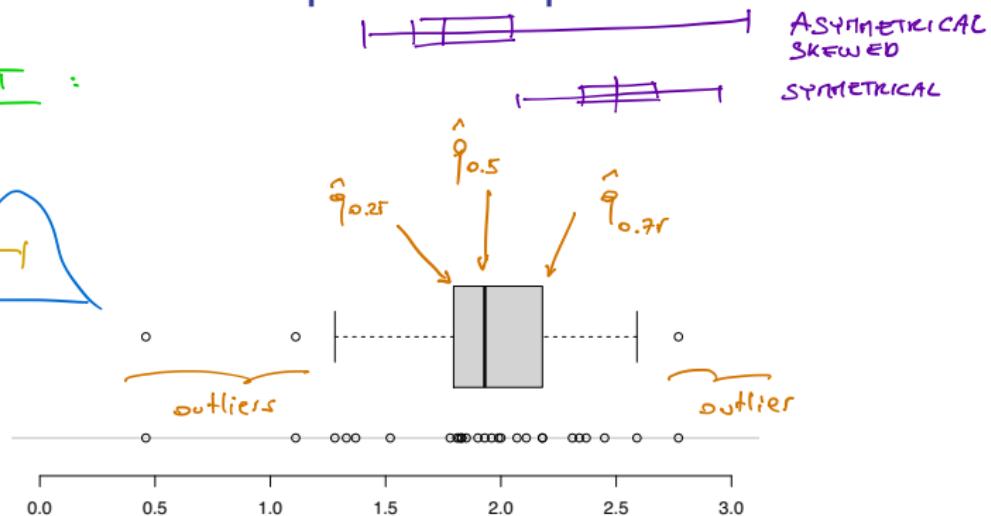
$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$
0.46	1.11	1.28	1.33	1.37	1.52	1.78	1.81	1.82
$x_{(10)}$	$x_{(11)}$	$x_{(12)}$	$x_{(13)}$	$x_{(14)}$	$x_{(15)}$	$x_{(16)}$	$x_{(17)}$	$x_{(18)}$
1.83	1.83	1.85	1.9	1.93	1.96	1.99	2.00	2.07
$x_{(19)}$	$x_{(20)}$	$x_{(21)}$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

Box PLOT :



not suitable
for BIMODAL
VARIABLES



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

DIFFERENT AND UNRELATED PARTICIPANTS

SAME PARTICIPANTS MEASURED TWICE

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					

$$\overline{Y} - \overline{X}$$

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

mean of the differences. { $\frac{\sum (y_i - x_i)}{n}$ }

Quiz Time

Sections 1 to 4

https://docs.google.com/forms/d/e/1FAIpQLScblQ_-ISfSCGp_EIVPPI_mnrJHttaKxln8vVoyjJFvS8BL1w/viewform

Some parametric distributions: Bernoulli distribution

	x_1	x_2	x_3	\dots	x_n
Cancer status	1	0	0	\dots	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 :

$$X_i \sim \text{Bernoulli}(\pi)$$

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

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

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 :

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

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

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

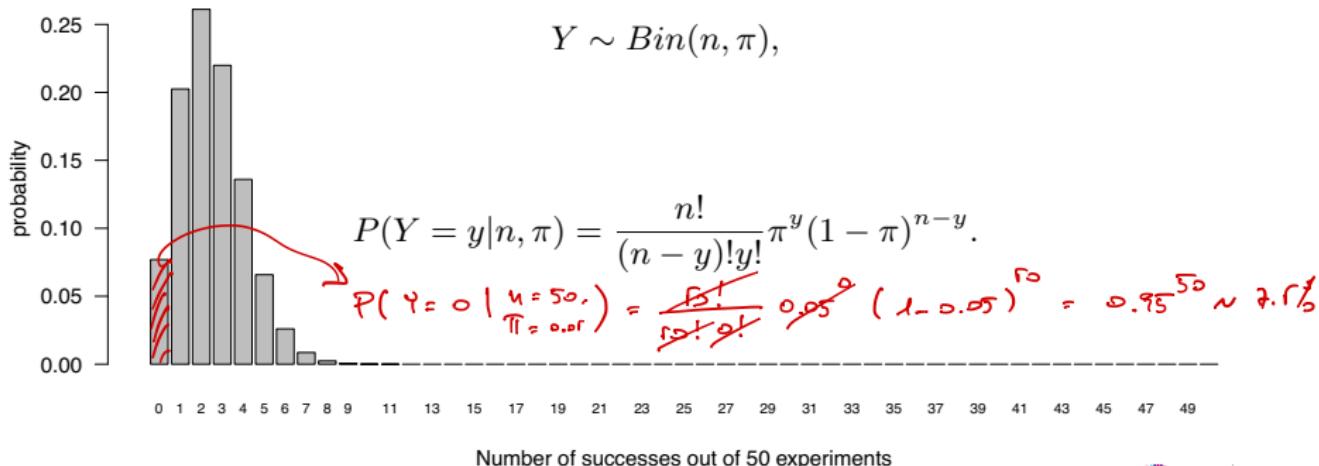
Some parametric distributions: Binomial distribution

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 :



Some parametric distributions: Poisson distribution

EVENTS PER
UNIT OF TIME AND/OR
PER AREA

HOURLY ADMISSION AT ANE
OF DAILY ORGAN DONORS IN THE UK

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 :

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

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

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

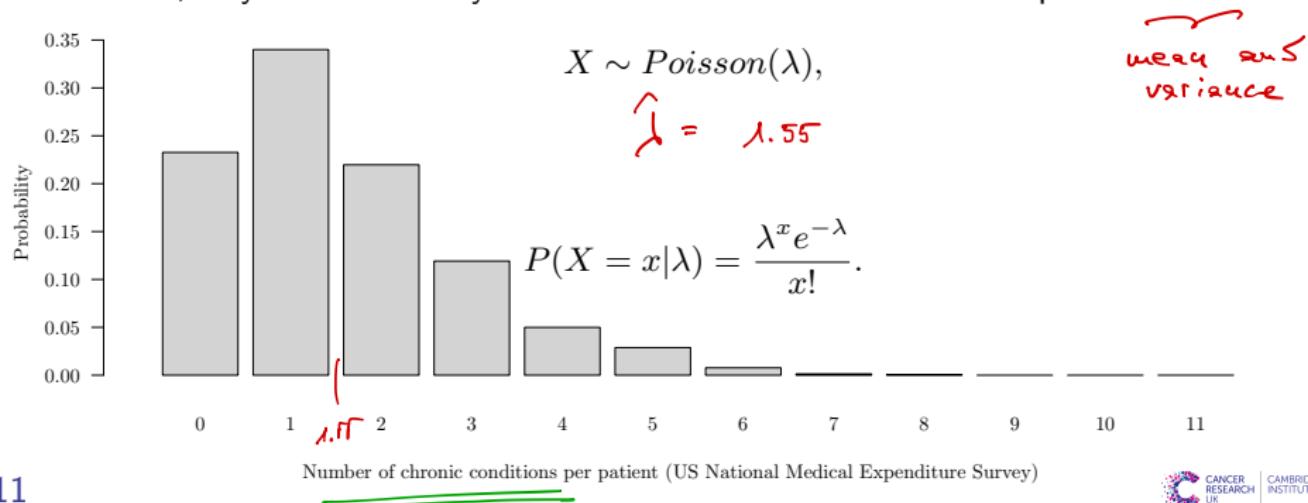
Some parametric distributions: Poisson distribution

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

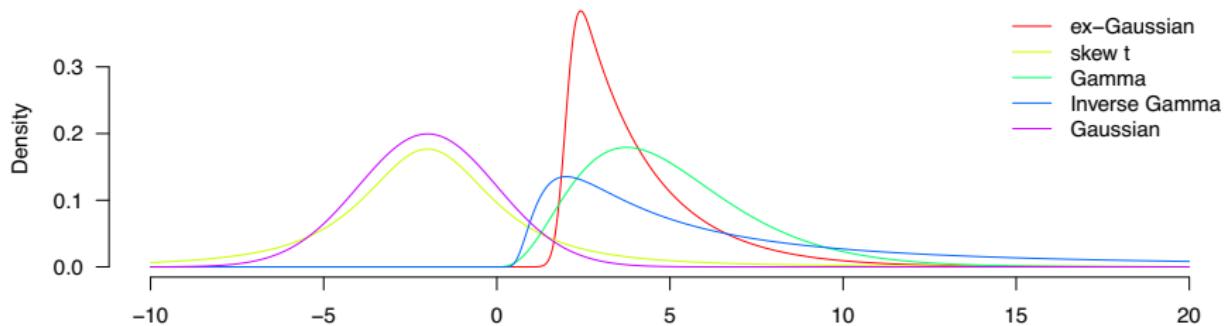
- ▶ events occur independently,
- ▶ at the same rate,
- ▶ and the probability of an event to occur in a small interval (area) is proportional to the length of the interval (size of the area),

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 :



Some parametric distributions: Continuous distrib.

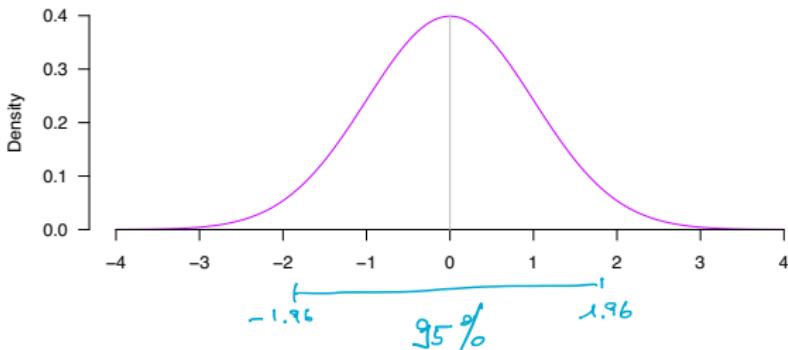


Some parametric distributions: Normal distribution

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

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

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

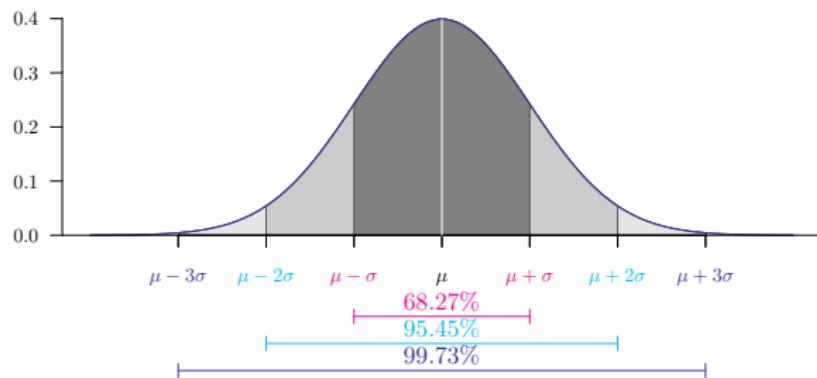


Some parametric distributions: Normal distribution

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

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

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

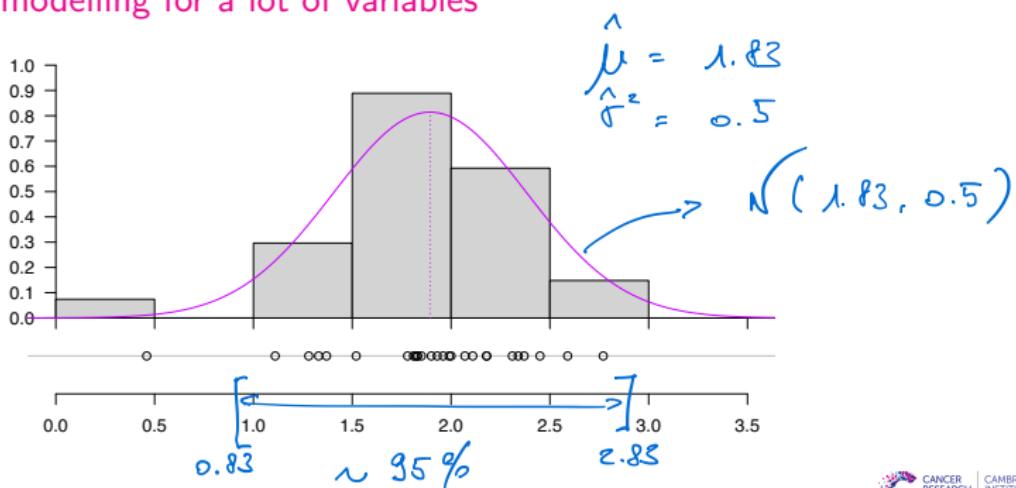


Some parametric distributions: Normal distribution

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

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

(i) Suitable modelling for a lot of variables

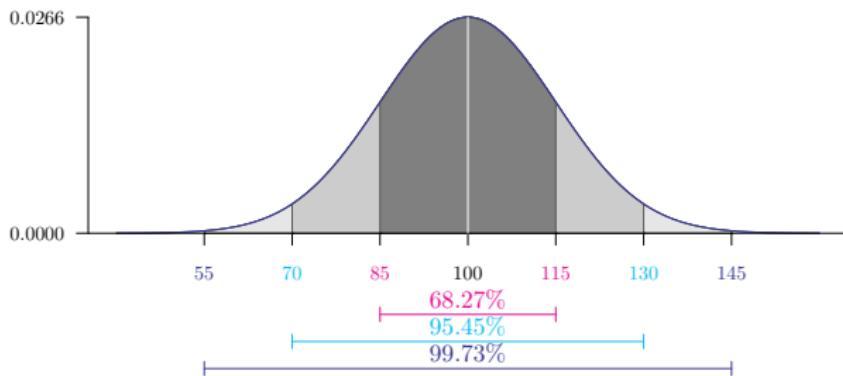


Some parametric distributions: Normal distribution

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

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

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



Some parametric distributions: Normal distribution

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

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

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

(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 μ and finite variances given by σ^2 ,
- ▷ then

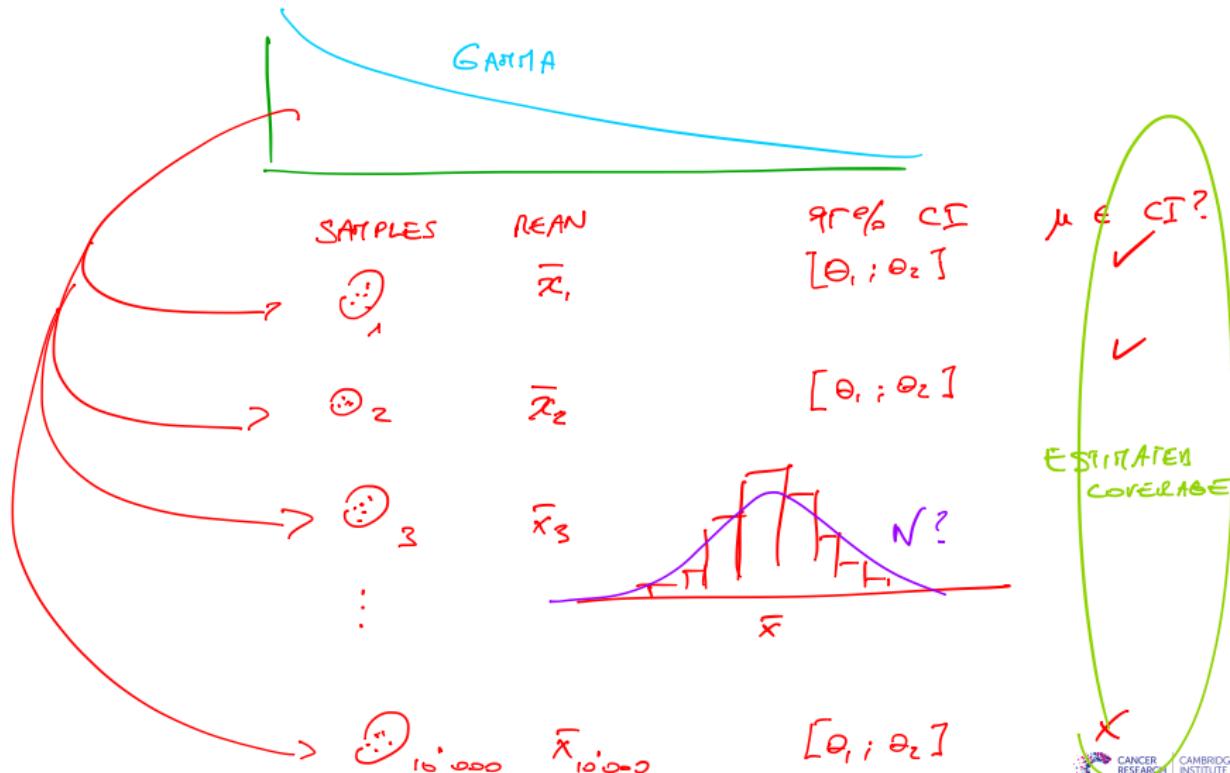
$$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^n X_i}{n} \quad \xrightarrow{d} \quad N\left(\mu, \frac{\sigma^2}{n}\right).$$

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

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



95% Confidence interval for μ , the population mean,
when $X_i \sim N(\mu, \sigma^2)$ or n large and $X_i \sim i.i.d. N(\mu, \sigma^2)$

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

$$P\left(-1.96 < Z < 1.96\right) = 0.95$$

$$P\left(-1.96 < \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} < 1.96\right) = 0.95$$

$$P\left(-1.96 \sqrt{\frac{\sigma^2}{n}} - \bar{x} < \bar{x} - \mu < 1.96 \sqrt{\frac{\sigma^2}{n}} - \bar{x}\right) = 0.95$$

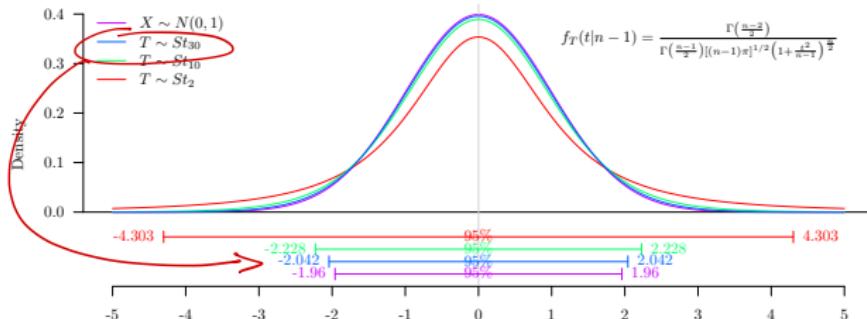
$$P\left(\bar{x} - 1.96 \sqrt{\frac{\sigma^2}{n}} < \mu < \bar{x} + 1.96 \sqrt{\frac{\sigma^2}{n}}\right) = 0.95$$

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

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

$$P\left(\bar{X} - T^{-1}_{n(0.975)} \frac{s}{\sqrt{n}} < \mu < \bar{X} + T^{-1}_{n(0.975)} \frac{s}{\sqrt{n}}\right) = 0.95$$

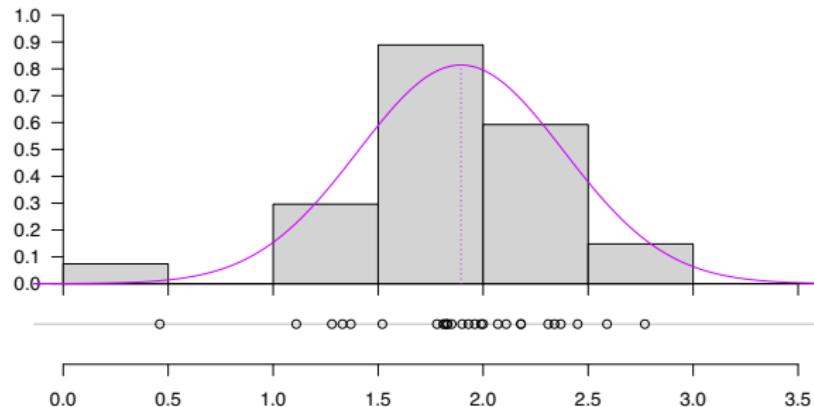
$$T^{-1}_{n(0.975)} = 2.042 \text{ for } n = 31$$



95% Confidence interval for μ , 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 $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$,
- if σ unknown, then $T = \frac{X-\mu}{s} \sim St_{n-1}$.

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

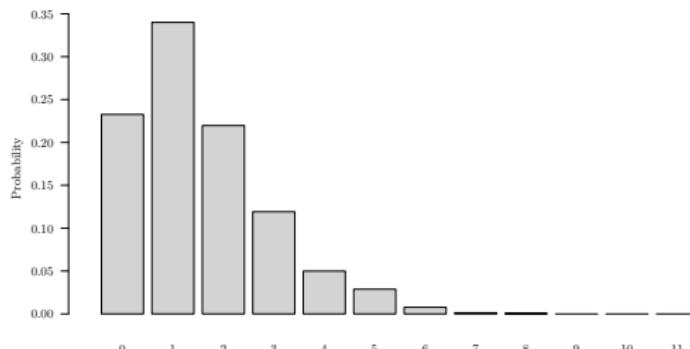


95% Confidence interval for μ , the population mean, when $X_i \sim iid(\mu, \sigma^2)$

- 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 σ unknown, then $T = \frac{X-\mu}{s} \sim St_{n-1}$.

Assuming $X_i \sim Poisson$
 $\hat{\mu} = \hat{\sigma}^2 \approx 1.55$

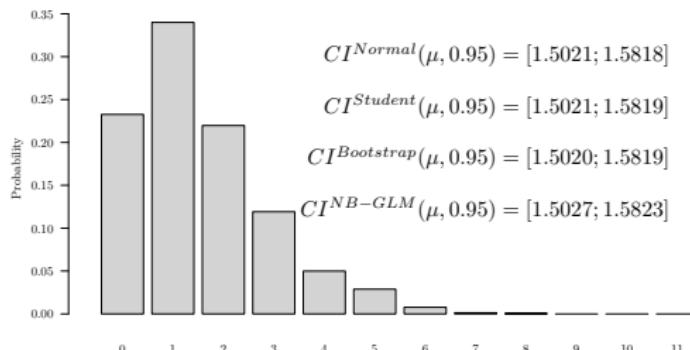
$$P\left(1.96 - 1.96 \sqrt{\frac{1.55}{4406}} < \mu < 1.96 + 1.96 \sqrt{\frac{1.55}{4406}}\right) = 0.95$$



Number of chronic conditions per patient (US National Medical Expenditure Survey: $n = 4406$, $\hat{\mu} = 1.55$)

95% Confidence interval for μ , 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 σ 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

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

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

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

Quiz Time

Practical 1

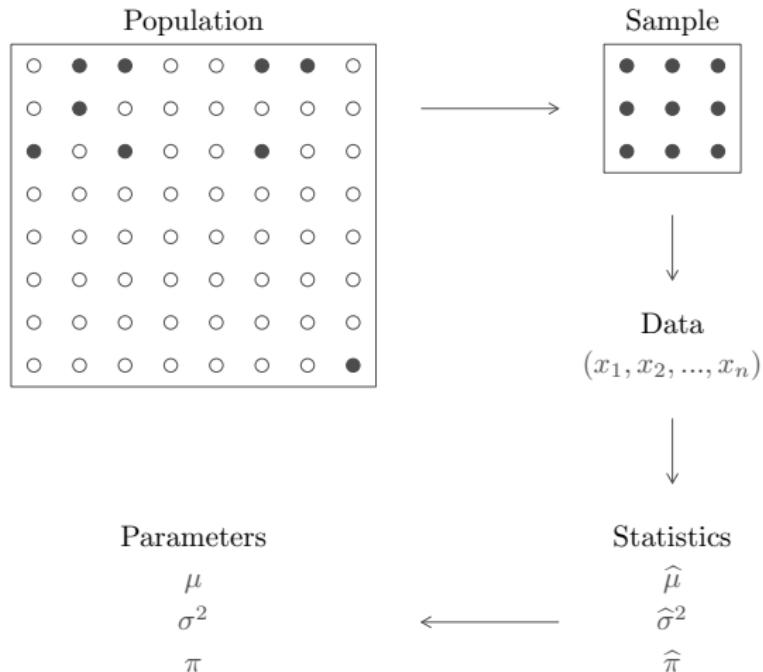
[http://bioinformatics-core-shared-training.github.io/
IntroductionToStats/practical.html](http://bioinformatics-core-shared-training.github.io/IntroductionToStats/practical.html)

PART II: Parametric tests

Cancer Research UK – 24th 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**, “generally assumed to be true until evidence indicates otherwise”
- ▶ the alternative hypothesis **H1**.

$$\begin{aligned} H_0 : \pi_B &= 0.5 \\ H_1 : \pi_B &> 0.5 \end{aligned}$$

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

$$\begin{aligned} H_0 : \mu_D &= \mu_E \\ H_1 : \mu_D &> \mu_E \end{aligned}$$

Theory 2: There is no majority for Brexit now

Theory 3: Serum vitamin C is reduced in patients

$$\begin{aligned} H_0 : \mu_P &= \mu_F \\ H_1 : \mu_P &< \mu_F \end{aligned}$$

Statistical hypothesis testing

Several-step process:

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

Statistical hypothesis testing

Example: One-sample two-sided t-test

We test:

$$H_0: \mu_{IQ} = 100,$$
$$H_1: \mu_{IQ} > 100.$$

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

$$\sigma^2 = 1\sigma^2$$

We know

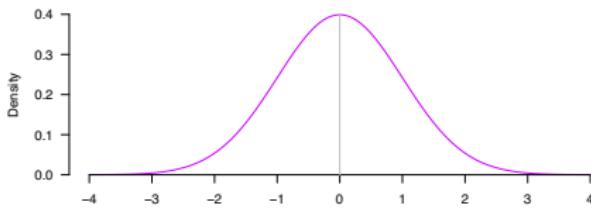
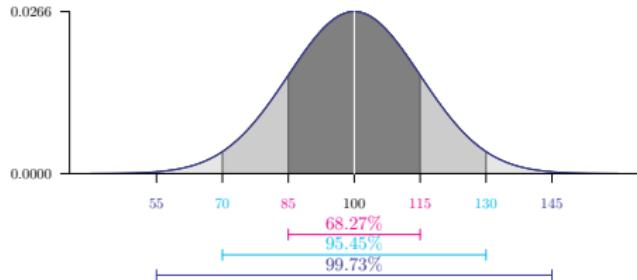
- ▶ $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right),$
- ▶ $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1),$

Thus, if H_0 is true, we have:

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

Define the p-value:

$$\text{▶ } p\text{-value} = P(|Z| > z_{obs})$$
$$= P(Z > z_{obs})$$



$$z_{obs} = \frac{120 - 100}{\sqrt{\frac{1}{30}}}$$

Statistical hypothesis testing

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 .

HIV test

		Test Outcome	
		H_0 (−) not rejected	H_1 (+) accepted
Unknown Truth	H_0 true (−)	$1 - \alpha$	α FALSE POS.
	H_1 true (+)	β FALSE NEG.	$1 - \beta$

where

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

Statistical hypothesis testing

Example: One-sided binomial exact test

We test:

$$H_0: \pi = 5\%,$$

$$H_1: \pi > 5\%.$$

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

We know

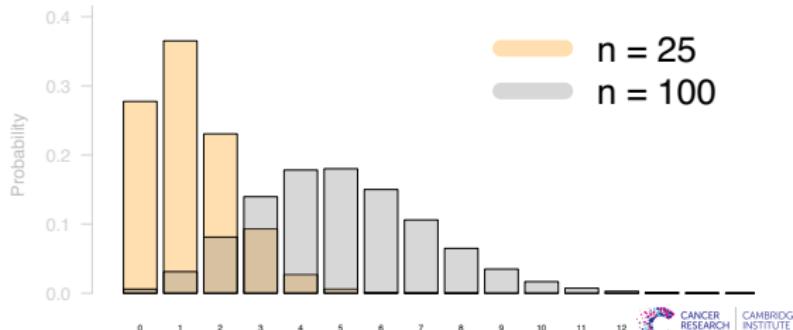
$$\blacktriangleright Y = \sum_{i=1}^n X_i \sim Binomial(\pi, n),$$

Thus, if H_0 is true, we have:

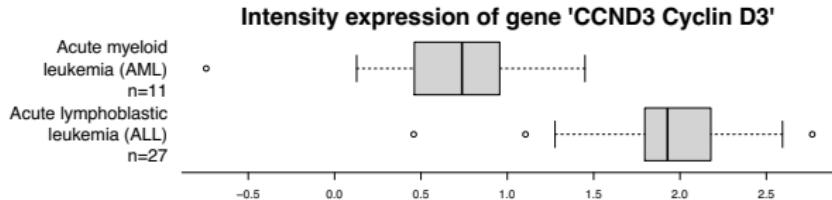
$$\blacktriangleright Y = \sum_{i=1}^n X_i \sim Binomial(5\%, n),$$

Define the p-value:

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



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



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

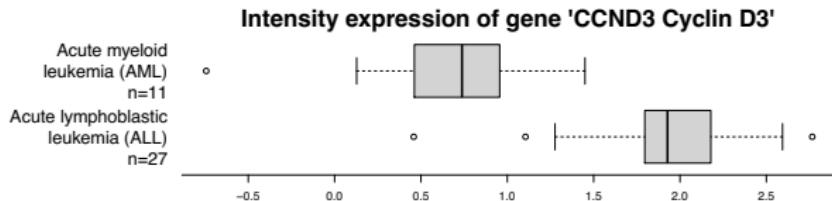
We know:

- ▶ 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
```

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
```

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 – 24th 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	✓	/ / / /	✗
	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.

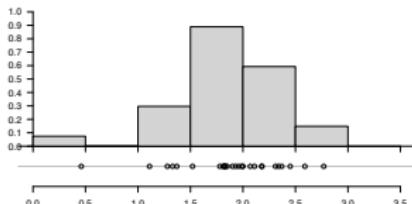
Sign test

RATHER INEFFICIENT

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

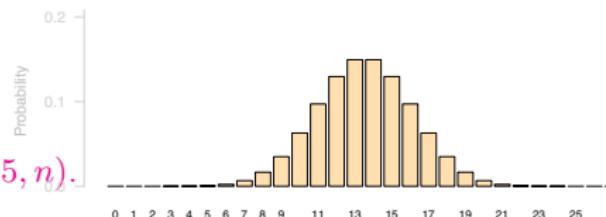
$$X_i = \theta + e_i,$$

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



Interest for $H_0: \theta = \theta_0$ against $H_1: \theta < \theta_0$ 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 H_0 :

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

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
```

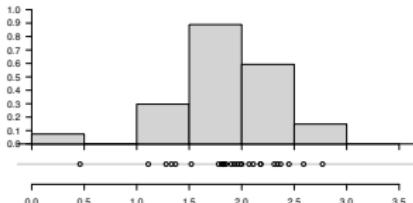
Number of successes out of 27 experiments

Wilcoxon sign-rank test

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

$$X_i = \theta + e_i,$$

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



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

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

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

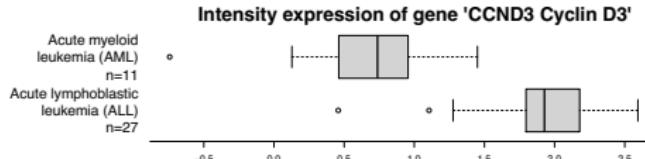
Wilcoxon signed rank test

```
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
```

Mann-Whitney-Wilcoxon test: Shift in location

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

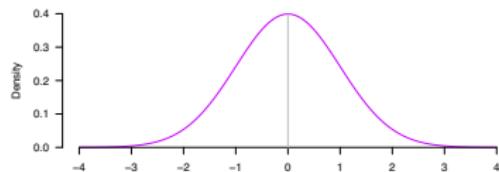
where $R(Y_i)$ denotes the rank of Y_i amongst the combined samples, i.e., amongst $(X_1, \dots, X_{n_X}, Y_1, \dots, Y_{n_Y})$.

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



Non-parametric is not assumption free

Shift in location tests when H0 is true

Simulate 2500 samples with

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

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

Assume

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

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		
		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

PART IV: Tests for categorical variables

Cancer Research UK – 24th 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	
		Treatment	Placebo	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	
		Treatment	Placebo	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** : $\pi_j = \pi_{j0}, j = 1, \dots, J$,
- **H1** : $\pi_j \neq \pi_{j0}, j = 1, \dots, J$.

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

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

Fisher showed that, under H_0 (independence),

$P(\text{observed table} \mid H_0) = P(X = a)$ and $X \sim \text{Hypergeometric}(n, a + c, a + b)$.

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
```