Introduction to Statistical Analysis (using Shiny Apps)

CRUK:- Tuesday 21st June 2016

Robert Nicholls & Mark Dunning

www.tiny.cc/crukStats

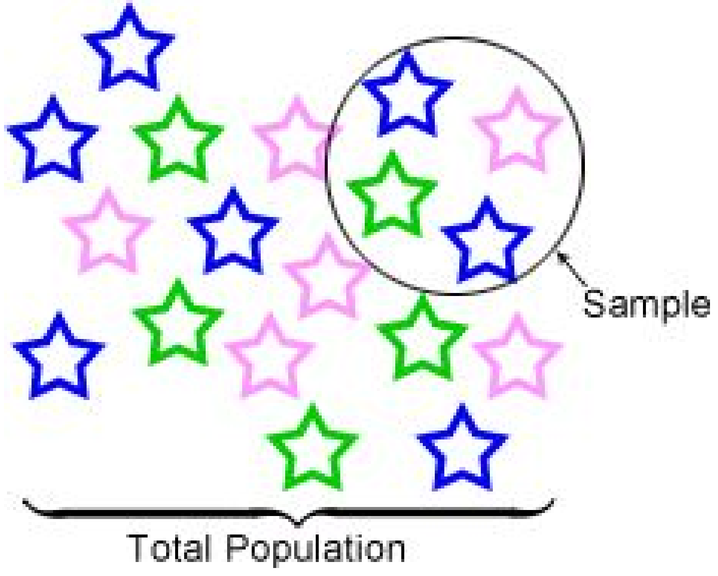
Acknowledgements:- Sarah Vowler, Sarah Dawson, Liz Merrell, Deepak Parashar

Approximate Timetable

10.30 - 11.15 – Lecture: Introduction to Statistical analysis
11.15 - 11.30 – Quiz: Variables/Dependencies/Tests/Generalisability
11.30 - 12.00 – Lecture: Parametric Tests for Continuous Variables; t-tests
12.00 - 12.30 – Examples/Practicals (computer based)
12.30 - 13.30 – Lunch (not provided)
13.30 - 14.00 – Lecture: Non-parametric tests for continuous variable
14.00 - 14.30 – Examples/Practicals (computer based)
(14:30 COFFEE)
14.30 - 14.45 – Lecture: Tests for Categorical Variables
14.45 - 15.30 – Examples/Practicals/Solutions (computer based)
15.30 - 16.25 – Group based exercise: Choosing appropriate tests
16.25 - 16.30 – Summary

The point of statistics

- Rarely feasible to study the whole population that we are interested in, so we take a sample instead
- Assume that data collected represents a larger population
- Use sample data to make conclusions about the overall population



Beginning a study

- Which samples to include?
  - Randomly selected?
  - Affects generalisability
- Always think about the statistical analysis
  - Randomised comparisons?
  - Data type?
  - Any dependency in measurements?
  - Distribution of data?
    - Normally distributed? Skewed? Bimodal?

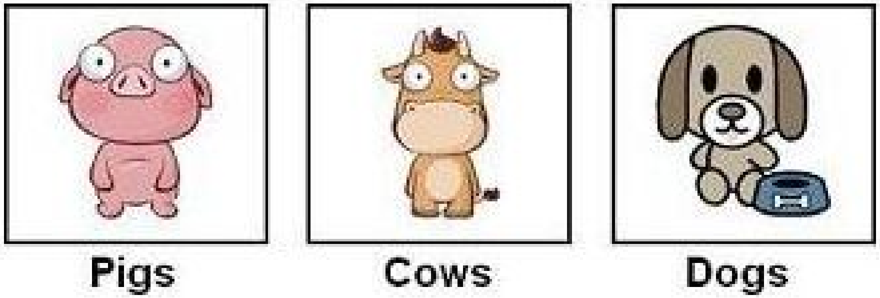
Generalisability

- How samples are selected affects interpretation
  - Generalisability is about population results apply to
- Statistical methods assume random samples
- Do not extrapolate beyond range of the data
  - i.e. don’t assume results apply to anything not represented in the data
- Examples:
  - Males only, no idea about females
  - Adults only, no idea about children

Data - types

- Several different categorisations
- Simplest:
  - Categorical (nominal)
  - Categorical with ordering (ordinal)
  - Discrete
  - Continuous

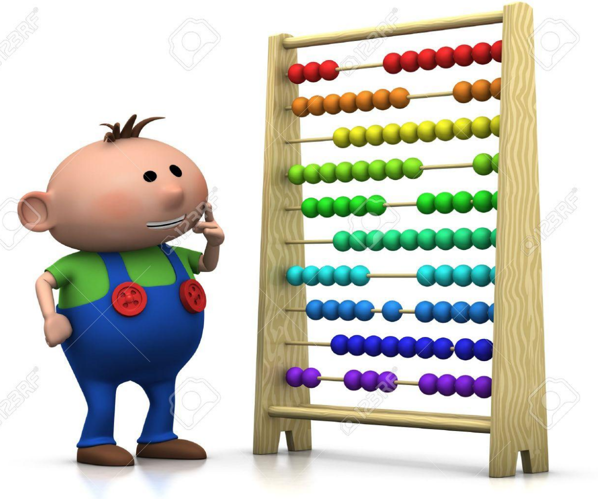
Nominal



- Most basic type of data
- Three requirements:
  - Same value assigned to all the members of level
  - Same number not assigned to different levels
  - Each observation only assigned to one level
- Example: gender, 1 = female, 2 = male
- Boils down to yes/no answer
- Others: Surgery type, cancer type, eye colour, dead/alive, ethnicity.

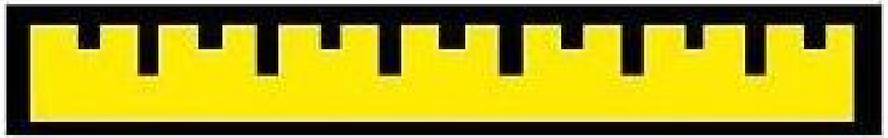
- Next type of data
- Mutually exclusive fixed categories
- Implicit order
- Can say one category higher than another
  - But not how much higher
- Example: stress level 1 = low … 7 = high
- Others: Grade, stage, treatment response, education level, pain level.

Discrete



- Third level of measurement
- Fixed categories
- Like ordinal but over bigger range
  - Can be treated as continuous if range is large
- Anything counted is discrete – *how many*?
- Example: number of tumours
- Others: Shoe size, hospital admissions, parity, number of side effects, medication dose, CD4 count, viral load, reads.

Continuous



- Final type of data
- Anything that is measured, *how much?*
- Meaningful zero: ratio, otherwise interval
  - Care required with interpretation
- Given any two observations fit one between
- Example: Height,
- Others: Weight, blood pressure, temperature, operation time, blood loss, age.

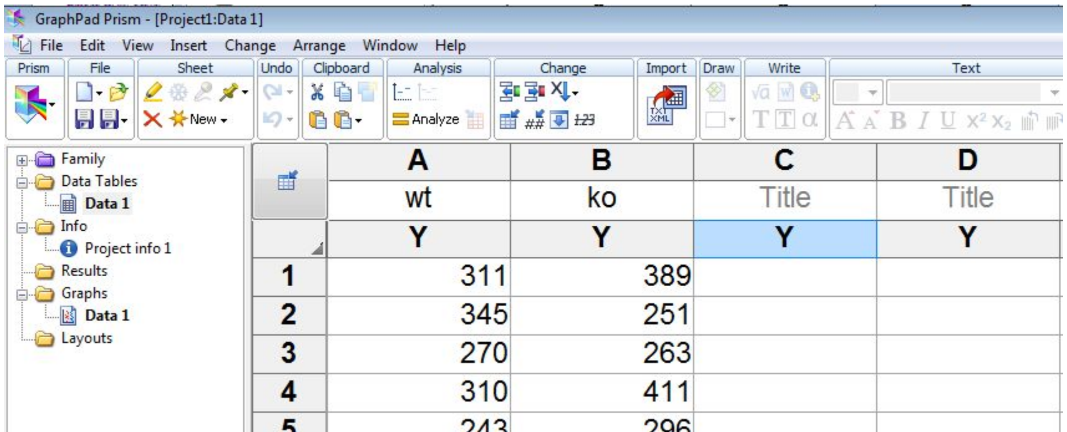
Data - types

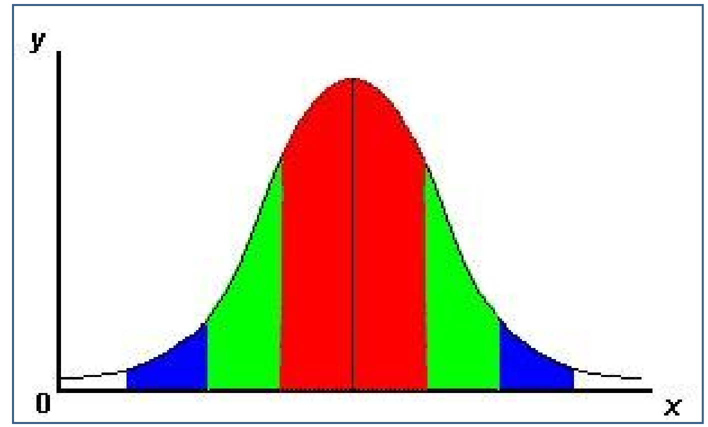
- Several different categorisations
- Simplest:
  - Categorical (nominal) – yes/no
  - Categorical with ordering (ordinal) – implicit order
  - Discrete – how many?
  - Continuous – how much?
- Write down examples

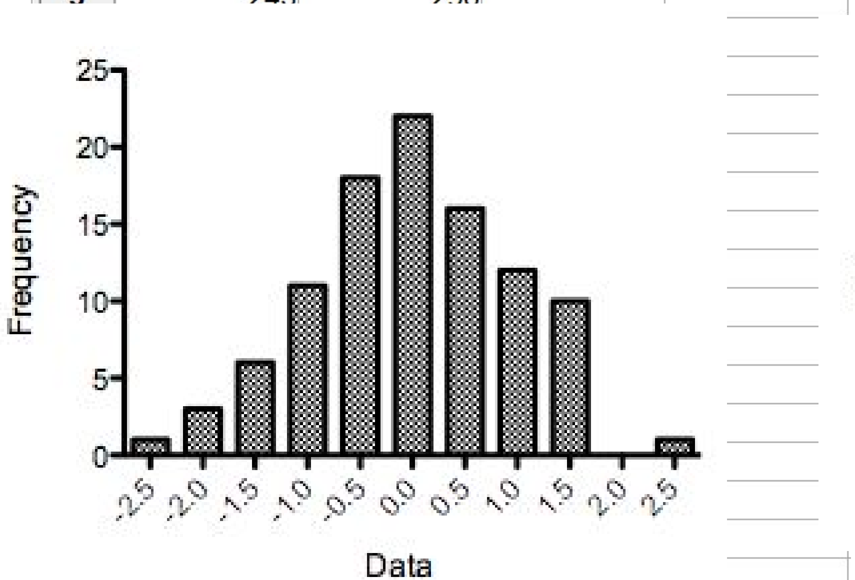
Measurements: Dependent / Independent?

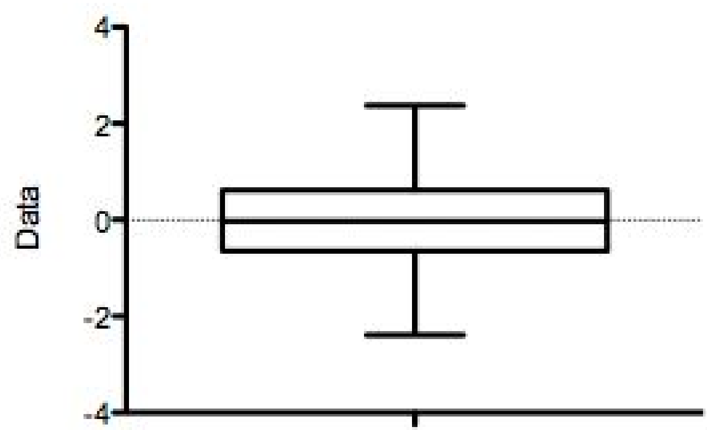
- Measurements of gene expression taken from each of 20 individuals
- Are any measurements more closely related than others?
  - Siblings/littermates?
  - Same individual measured twice?
  - Batch effects?
- If no reason – **independent observations**

Continuous Data – Distribution

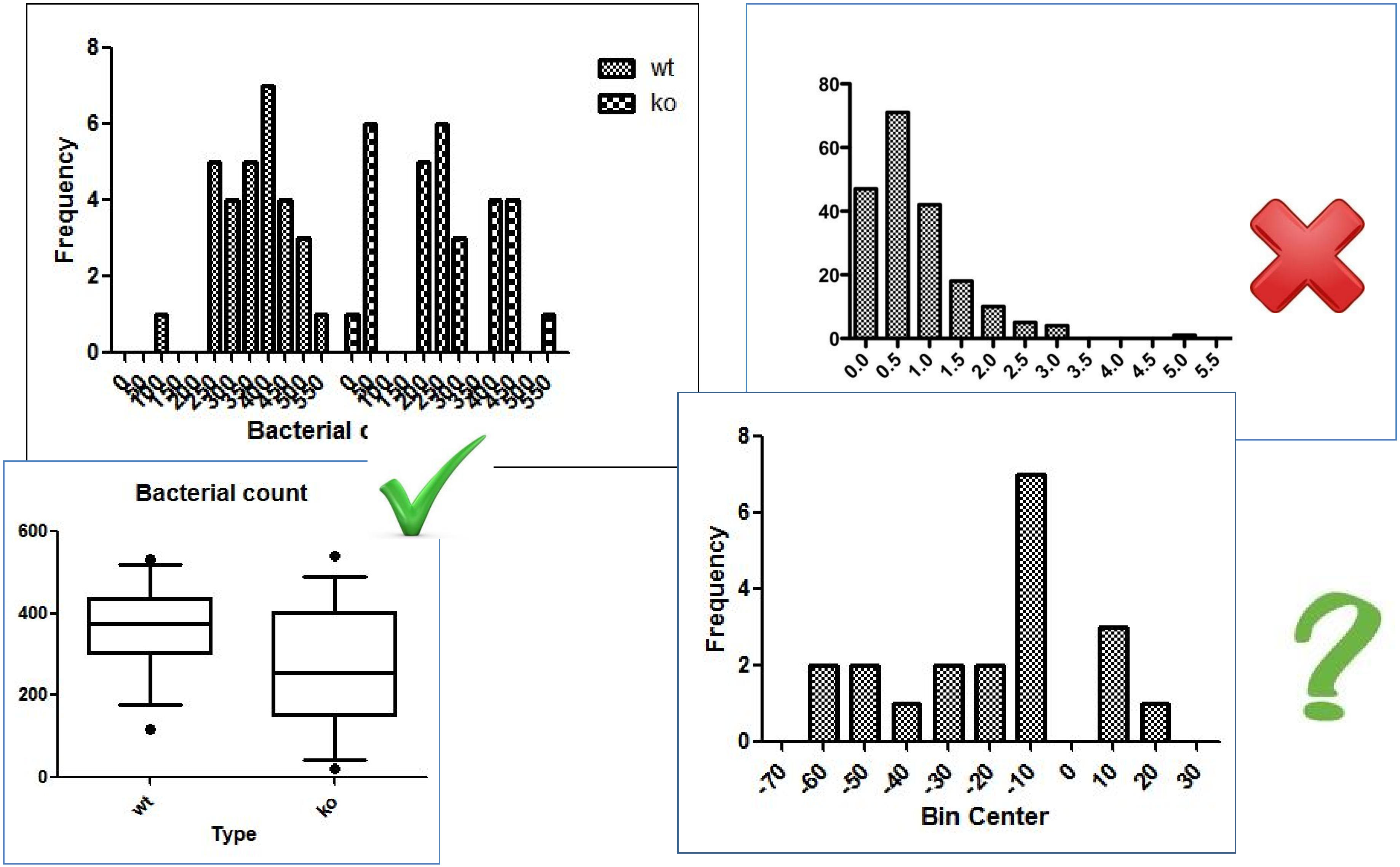






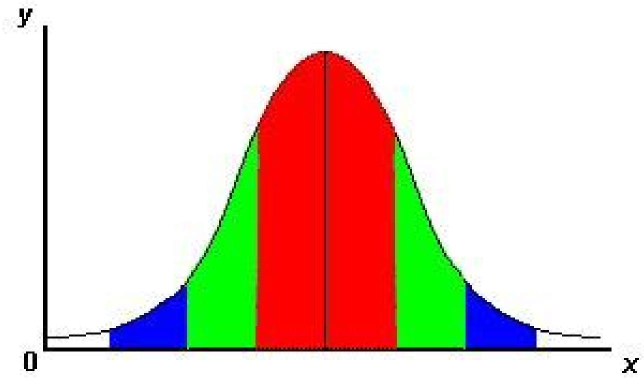


Continuous Data – Distribution?



Continuous Data – Descriptive Statistics

- Measures of location and spread

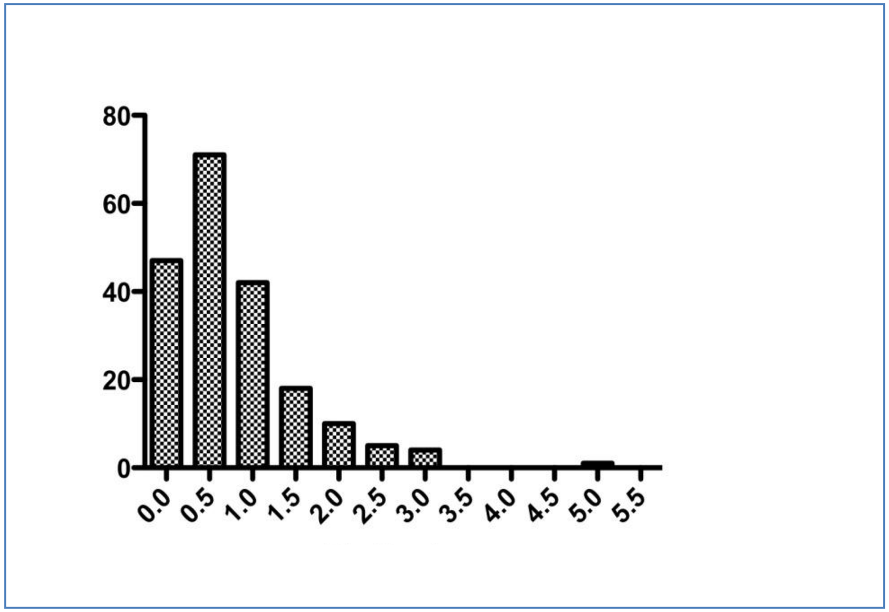


  - Mean and standard deviation

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$$s.d. = \sqrt{\frac{\left(X_1 - \overline{X}\right)^2 + \left(X_2 - \overline{X}\right)^2 + \cdots + \left(X_n - \overline{X}\right)^2}{n}}$$

Continuous Data – Descriptive Statistics



- Median: middle value
- Lower quartile: median bottom half of data
- Upper quartile: median top half of data

Continuous Data – Descriptive Statistics (Example)

E.g. No. of Facebook friends for 7 colleagues

311, 345, 270, 310, 243, 5300, 11

- Measures of location and spread
  - Mean and standard deviation

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = 970;$$

$$s.d. = \sqrt{\frac{\left(X_1 - \overline{X}\right)^2 + \left(X_2 - \overline{X}\right)^2 + \cdots + \left(X_n - \overline{X}\right)^2}{n}} = 1912.57$$

  - Median and interquartile range

11, **243**, 270, **310**, 311, **345**, 5300

Continuous Data – Descriptive Statistics (Example)

E.g. No. of facebook friends for 7 colleagues

311, 345, 270, 310, 243, **530**, 11

- Measures of location and spread
  - Mean and standard deviation

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = 289;$$

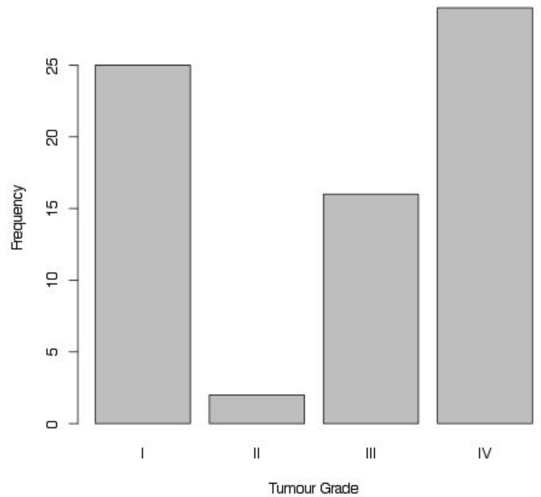
$$s.d. = \sqrt{\frac{\left(X_1 - \overline{X}\right)^2 + \left(X_2 - \overline{X}\right)^2 + \cdots + \left(X_n - \overline{X}\right)^2}{n}} = 153.79$$

  - Median and interquartile range

11, **243**, 270, **310**, 311, **345**, 530

Categorical Data

- Summarised by counts and percentages
- Examples
  - 19/82 (23%) subjects had Grade IV tumour
  - 48/82 (58%) subjects had Diarrhoea as an Adverse Event.



Standard Deviation and Standard Error

- Commonly confused
- Standard deviation:
  - Measure of spread of the data
  - Used for describing population
- Standard error:
  - Variability of the mean from repeated sampling
  - Precision of mean
  - Used to calculate confidence interval

- SD: How widely scattered measurements are
- SE: Uncertainty in estimate of sample mean

Confidence intervals for the mean

- Confidence interval (CI) is a random interval
- In repeated experiments
  - 95% of time cover the mean
- Looser interpretation 95% of time mean in CI

$$95\%\ CI:\left(\overline{X} - 1.96 \times \text{standard error},\ \overline{X} + 1.96 \times \text{standard error}\right)$$

$$\text{Standard error} = \frac{\text{Standard deviation}}{\sqrt{n}} = \frac{154}{\sqrt{7}} = 58$$

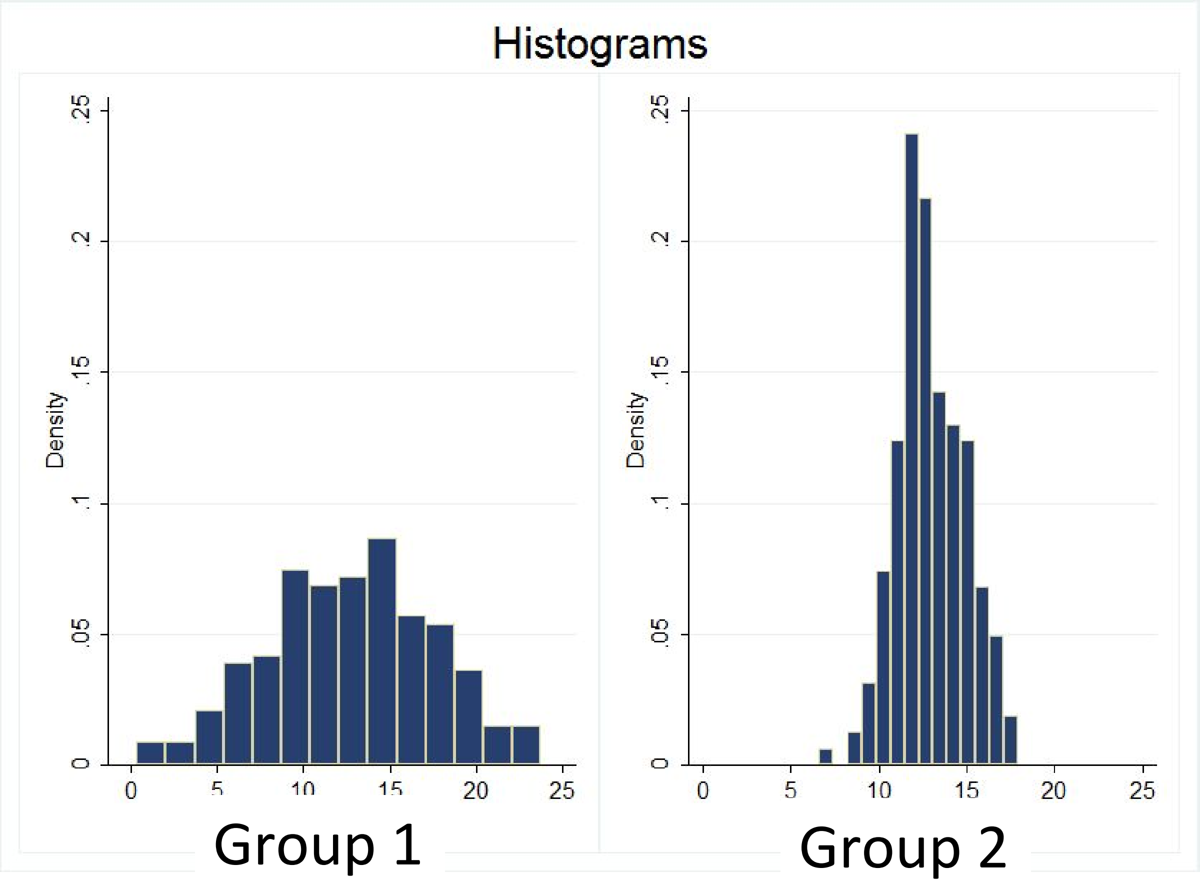
Mean 289, 95% CI (175, 402)

Confidence intervals

No. of samples/ observations

Standard deviation

Standard error of mean



Hypothesis tests – basic set-up

- Formulate a null hypothesis, H_0

The difference in gene expression before and after treatment = 0

- Calculate a test statistic from the data under the null hypothesis

$$t_{n-1} = t_{29} = \frac{\overline{X_{After-Before}}}{s.e.\left(\overline{X_{After-Before}}\right)}$$

- Determine whether the test statistic is more extreme than expected under the null hypothesis (p-value)
- Reject or do not reject the null hypothesis

Absence of evidence is not evidence of absence (Bland and Altman, 1995)

- Correction for multiple testing

Hypothesis tests – Example

Lady Tasting Tea

Randomised Experiment by Fisher

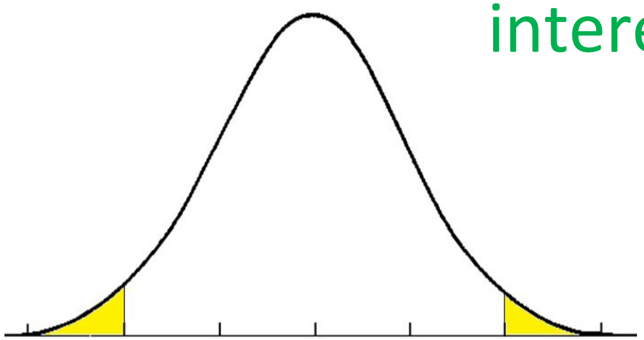
- Randomly ordered 8 cups of tea
  - 4 were prepared by first adding milk
  - 4 were prepared by first adding tea
- Task: Lady had to select the 4 cups of one particular method
- H_0: Lady had no such ability
- Test Statistic: number of successes in selecting the 4 cups.
- Result: Lady got all 4 cups right!

Reject the null hypothesis

Hypothesis tests – Errors

	Null hypothesis does not hold	Null hypothesis holds
Reject null hypothesis	Correct True positive	Wrong False positive
Do not reject null hypothesis	Wrong False negative	Correct True negative

significance level, sample size, difference of interest, variability of the observations.



Be aware of issues of multiple testing!

When to use which test

		RESPONSE		
NO OF SAMPLES		NOMINAL	ORDINAL OR NON-NORMAL	NORMALLY DISTRIBUTED
ONE SAMPLE		χ^2-test, Z-test	Kolmogorov-Smirnov Sign test	t-test
TWO SAMPLE	INDEPENDENT	χ^2-test (r x c), Fisher's exact test	Mann-Whitney U Median test	Unpaired t-test
	PAIRED	McNemar's test Stuart-Maxwell test	Wilcoxon signed rank Sign test	Paired t-test
MULTIPLE SAMPLES (K>2)	INDEPENDENT	χ^2-test (r x k) Fisher-Freeman-Halton	Kruskal-Wallis test Median Test Jonckheere-Terpstra test	Analysis of variance (ANOVA)
	PAIRED	Cochran Q test	Friedman test Page test Quade test	Repeated measures ANOVA
ASSOCIATION BETWEEN TWO VARIABLES		Contingency coefficient Phi, r_ϕ Cramér, C	Spearman's rank Kendall's tau	Pearson product moment correlation
AGREEMENT BETWEEN TWO VARIABLES		Simple kappa	Weighted kappa	Limits of agreement

Tests for continuous variables
T-tests

Statistical tests – continuous variables

- T-test:
  - **One-sample t-test**

    (e.g. H_0: mean = 5)

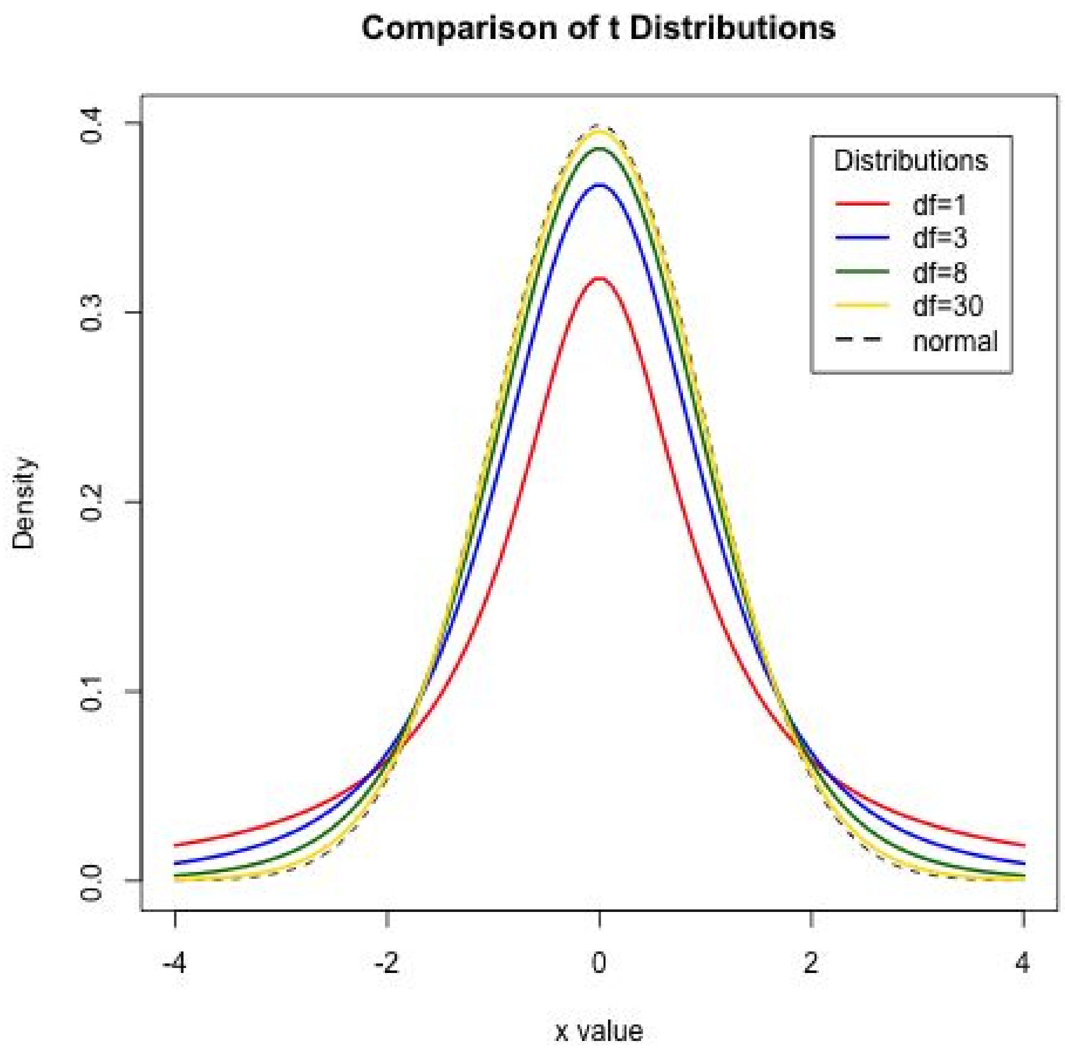
  - **Independent two-sample t-test**

    (e.g. H_0: mean of sample 1 = mean of sample 2)

  - **Paired two-sample t-test**

    (e.g. H_0: mean difference between pairs = 0)

T-distributions



One-sample t-test: does mean = X?

E.g. Research question: Published data suggests that the microarray failure rate for a particular supplier is 2.1%.

Genomics Core want to know if this holds true in their own lab?

One-sample t-test: does mean = X?

- **Null hypothesis, H_0** :
  Mean monthly failure rate = 2.1%.

- **Alternative hypothesis, H_1** :
  Mean monthly failure rate $\neq$ 2.1%.

- **Tails**: two-tailed.

- Either reject or do not reject the **null hypothesis** – <u>never accept the alternative hypothesis</u>

One-sample t-test – the data

Month	Monthly failure rate
January	2.90
February	2.99
March	2.48
April	1.48
May	2.71
June	4.17
July	3.74
August	3.04
September	1.23
October	2.72
November	3.23
December	3.40

The **mean** is the sum of all observations divided by the number of observations.

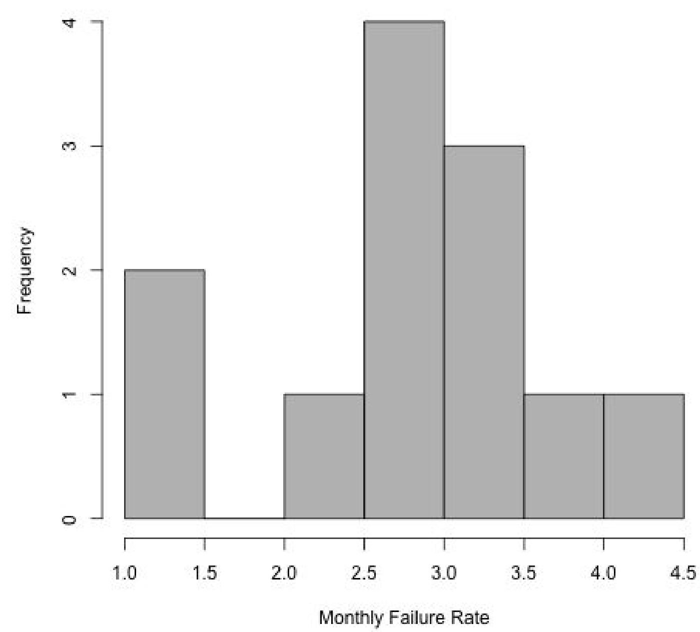
Mean = (2.90 +...+ 3.40)/12
= 2.84

Standard deviation = 0.84

Test value: 2.1

One-sample t-test – key assumptions

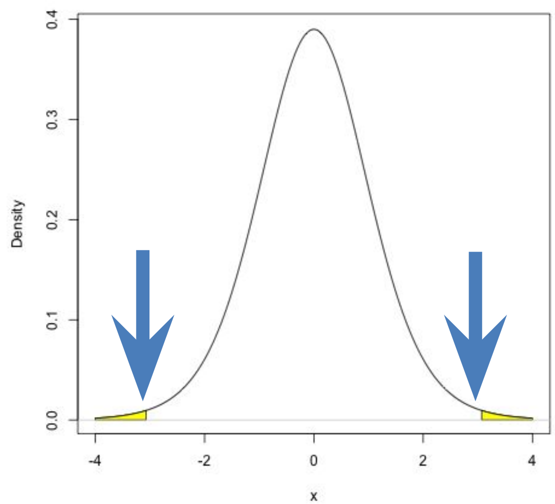
- Observations are independent
- Observations are normally distributed



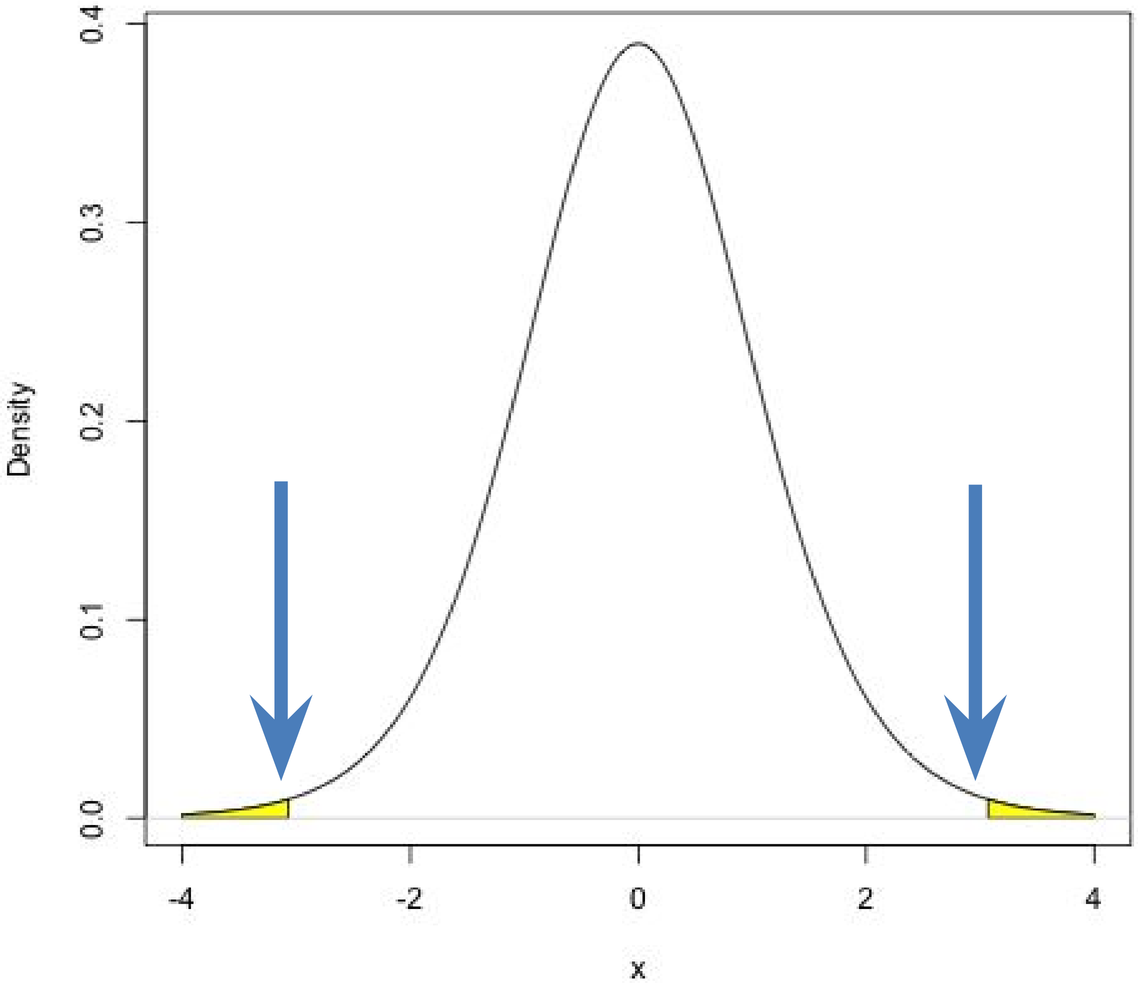
One-sample t-test - results

Test statistic:

$$t_{n-1} = t_{11} = \frac{\bar{x} - \mu_0}{s.d./\sqrt{n}} = \frac{2.84 - 2.10}{s.e.(\bar{x})} = 3.07$$



One-sample t-test - results



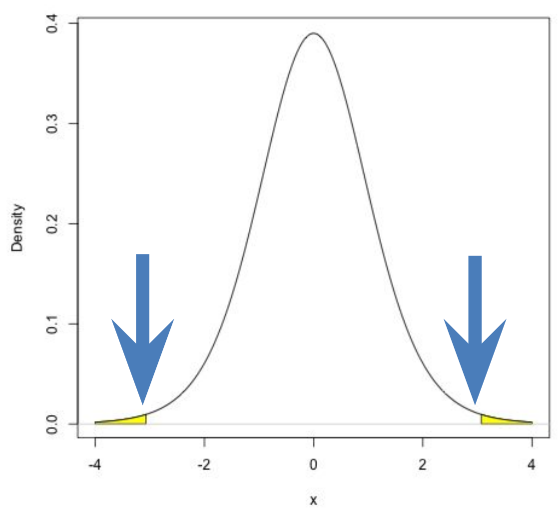
One-sample t-test - results

Test statistic:

$$t_{n-1} = t_{11} = \frac{\bar{x} - \mu_0}{s.d./\sqrt{n}} = \frac{2.84 - 2.10}{s.e.(\bar{x})} = 3.07$$

df = 11

P = 0.01



Reject H_0

(Evidence that mean monthly failure rate ≠ 2.1%.)

One-sample t-test results

- The mean monthly failure rate of microarrays in the Genomics core is 2.84 (95% CI: 2.30, 3.37).
- It is not equal to the hypothesized mean proposed by the company of 2.1.
- $t=3.07$, $df=11$, $p=0.01$

Two-sample t-test

- Two types of two-sample t-test:

  – Independent:

  e.g. the weight of two different breeds of mice.

  – Paired:

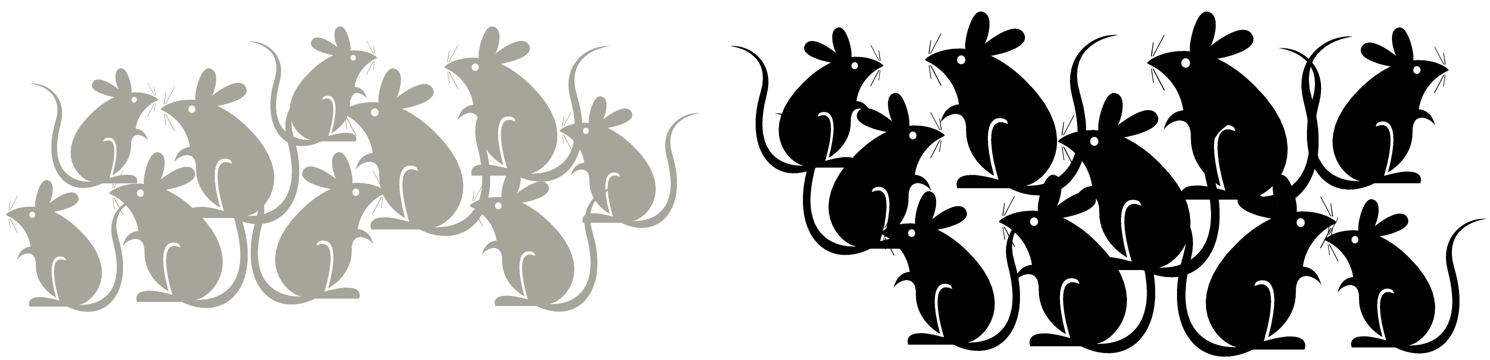
  e.g. a measurement of disease at two different parts of the body in the same patient/animal.

Independent two-sample t-test

Does mean of group A = mean of group B?

E.g. Research question: 40 male mice (20 of breed A and 20 of breed B) were weighed at 4 weeks old.

Does the weight of 4 week old male mice depend on breed?



Independent two-sample t-test

Does mean of group A = mean of group B?

- **Null hypothesis, H_0 :**
  Mean weight of breed A = Mean weight of breed B.

- **Alternative hypothesis, H_1 :**
  Mean weight of breed A $\neq$ Mean weight of breed B.

- **Tails**: two-tailed.

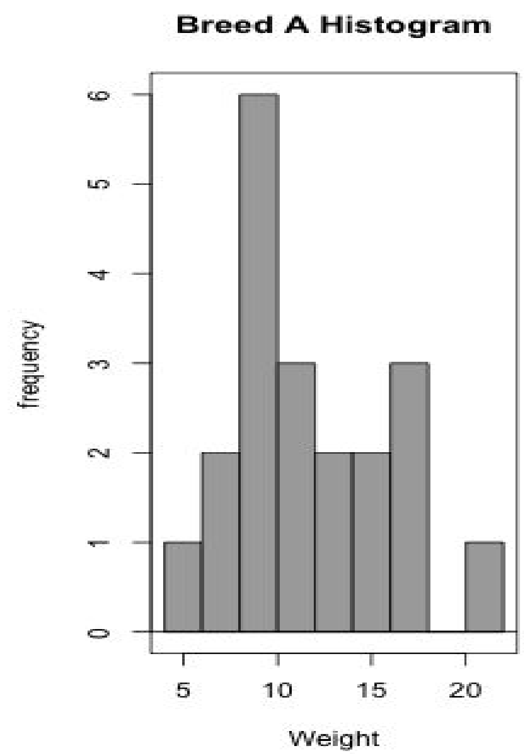
- Either reject or do not reject the **null hypothesis** – <u>never accept the alternative hypothesis</u>

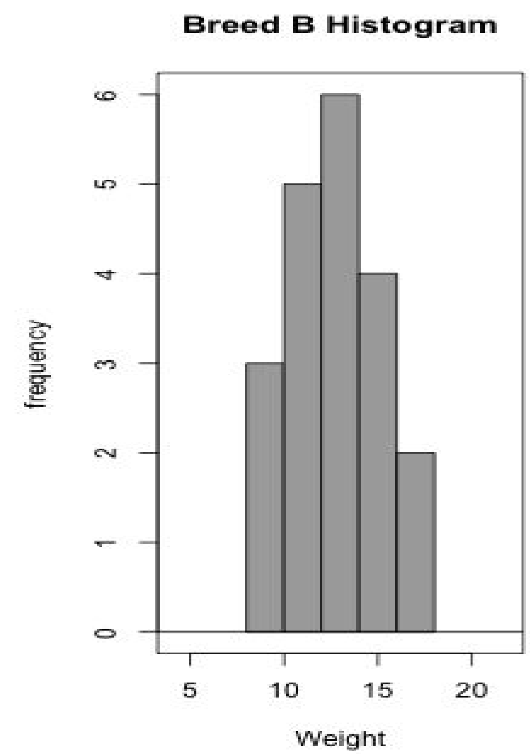
Independent two-sample t-test – the data

Breed A		Breed B	
Subject	**Weight at 4 weeks (g)**	**Subject**	**Weight at 4 weeks (g)**
1	20.77	21	15.51
2	9.08	22	12.93
3	9.80	23	11.50
4	8.13	24	16.07
5	16.54	25	15.51
6	11.36	26	17.66
7	11.47	27	11.25
8	12.10	28	13.65
9	14.04	29	14.28
10	16.82	30	13.21
11	6.32	31	10.28
12	17.51	32	12.41
13	9.87	33	9.63
14	12.41	34	14.75
15	7.39	35	9.81
16	9.23	36	13.02
17	4.06	37	12.33
18	8.26	38	11.90
19	10.24	39	8.98
20	14.64	40	11.29
Mean	11.50	**Mean**	12.80
Standard deviation	4.18	**Standard deviation**	2.33

Independent two-sample t-test – key assumptions

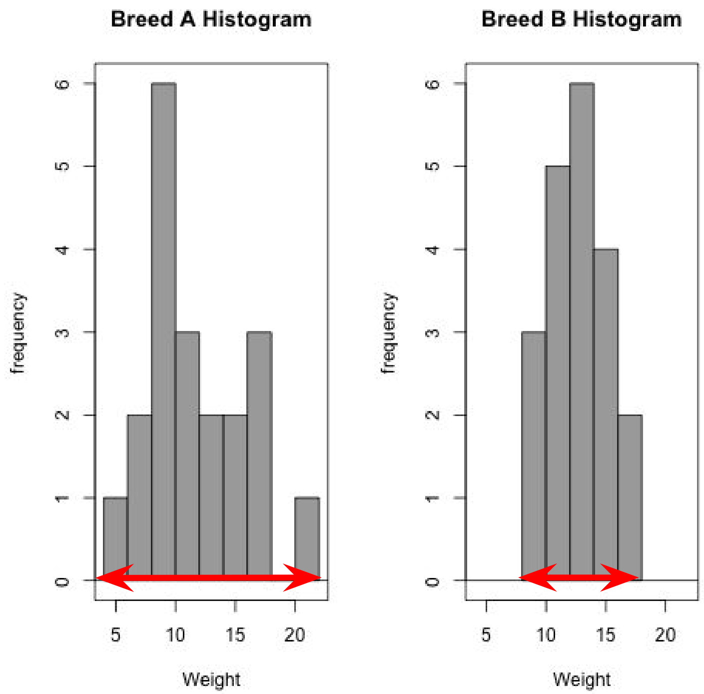
- Observations are independent
- Observations are normally distributed





Independent two-sample t-test -More key assumptions...

- Equal variance in the two comparison groups
  - Use Welch's correction if variances are different
    - » Alters the t-value and degrees of freedom

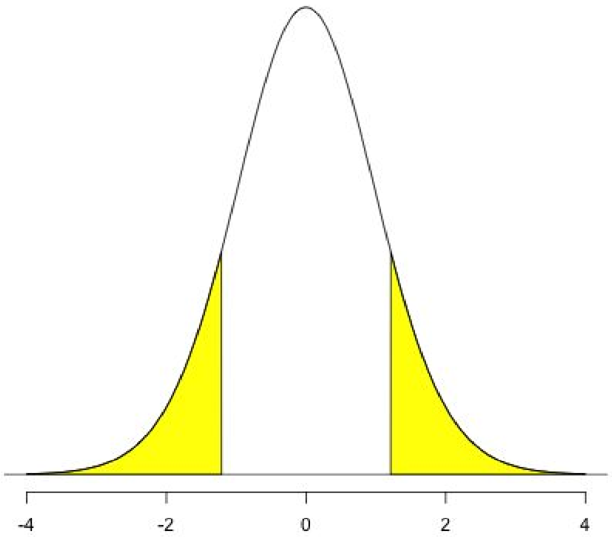


Standard deviation **4.18** **2.33**

Independent two-sample t-test - results

Test statistic: $$t_{df} = \frac{\overline{X_A} - \overline{X_B}}{s.e.\left(\overline{X_A} - \overline{X_B}\right)} = 1.21$$

df = 29.78
(Welch's correction)



P-value: **0.24**

Do not reject H_0
(No evidence that mean weight of breed A ≠ mean weight of breed B)

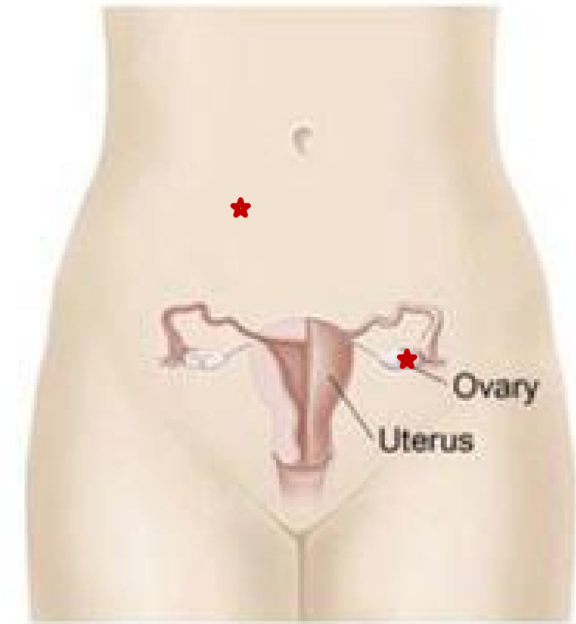
Independent two-sample t-test - results

- The difference in mean weight between the two breeds is -1.30 (95% CI: -3.48, 0.89)
  - [NB this is negative breed B weights tend to be bigger than breed A weights].
- There is no evidence of a difference in weights between breed A and breed B.
- $t=1.21$, $df= 29.78$ (Welch's correction), $p=0.24$.

Paired two-sample t-test:
Does the mean difference = 0?

E.g. Research question: 20 patients with ovarian cancer were studied using MRI imaging. Cellularity was measured for each patient at two sites of disease.

Does the cellularity differ between two different sites of disease?



Paired two-sample t-test:
Does the mean difference = 0?

- **Null hypothesis, H_0** :
  Cellularity at site A = Cellularity at site B

- **Alternative hypothesis, H_1** :
  Cellularity at site A ≠ Cellularity at site B

- **Tails**: two-tailed.

- Either reject or do not reject the **null hypothesis** – never accept the alternative hypothesis

Paired two-sample t-test – Null hypothesis

H_0 : Cellularity at site A = Cellularity at site B

OR

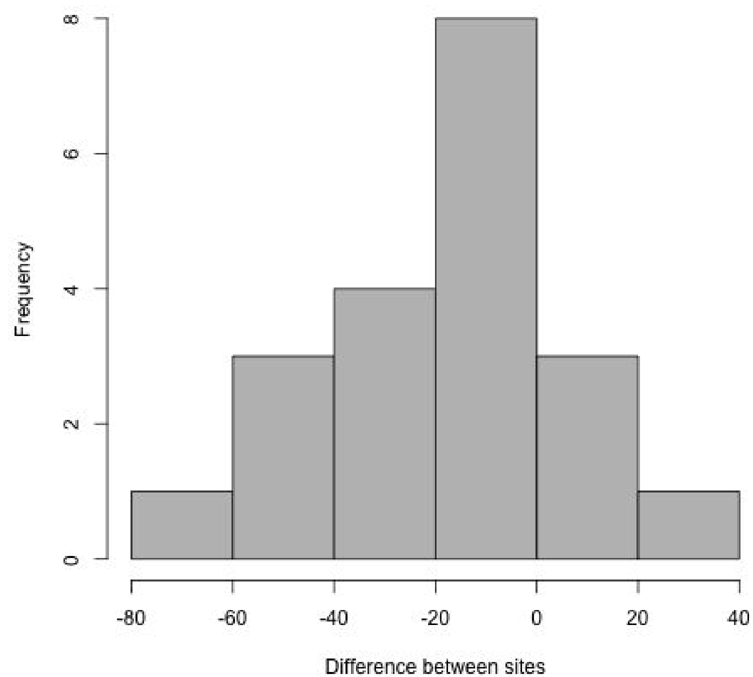
H_0 : Cellularity at site A - Cellularity at site B = 0

H_0 : Cellularity at site A - Cellularity at site B = 0

Subject	Cellularity		
	Site A: Primary ovarian mass	Site B: Peritoneal deposits	Difference
1	1201.33	1155.98	-45.35
2	1029.64	1020.82	-8.82
3	895.57	881.21	-14.37
4	842.14	830.78	-11.36
5	903.07	897.06	-6.01
6	1311.57	1262.73	-48.84
7	833.52	823.06	-10.46
8	1007.66	951.01	-56.65
9	1465.51	1450.98	-14.53
10	967.82	978.15	10.33
11	812.72	778.26	-34.46
12	884.08	823.57	-60.51
13	1358.56	1335.78	-22.78
14	1280.10	1293.91	13.80
15	942.38	925.75	-16.63
16	884.33	891.34	7.01
17	930.09	892.02	-38.07
18	1146.75	1132.80	-13.95
19	881.50	847.78	-33.72
20	1315.22	1337.80	22.58
		Mean difference	19.14
		Standard deviation	23.37

Paired two-sample t-test – key assumptions

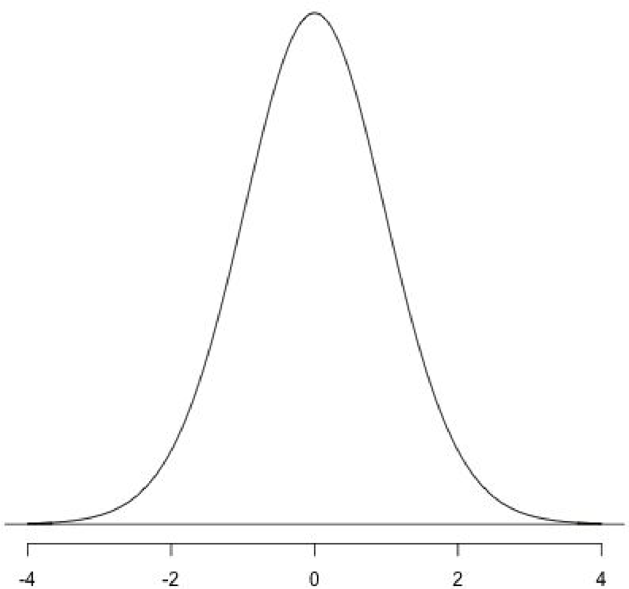
- Observations are independent
- The **paired differences** are normally distributed



Paired two-sample t-test - results

Test statistic $$t_{n-1} = t_{19} = \frac{\overline{X_{A-B}}}{s.e.\left(\overline{X_{A-B}}\right)} = 3.66$$

df = 19



P-value: **0.002**

Reject H_0

(Evidence that cellularity at site A ≠ Cellularity at site B)

Paired two-sample t-test - results

- The difference in cellularity between the two sites is 19.14 (95% CI: 8.20, 30.08).
- There is evidence of a difference in cellularity between the two sites.
- $t=3.66$, $df=19$, $p=0.0017$.

What if normality is not reasonable?

- Transform your data, e.g. Ln transformation

- Non-parametric tests:

Parametric test	Non-parametric test
One-sample t-test	One-sample Wilcoxon signed rank test One-sample sign test
Independent two-sample t-test	Mann-Whitney U test/ Wilcoxon rank sum test
Paired two-sample t-test	Matched-pairs Wilcoxon signed rank test Two-sample sign test

Summary – continuous variables

- **One-sample t-test**

  Use when we have one group.

- **Independent two-sample t-test**

  Use when we have two independent groups. A Welch correction may be needed if the two groups have different spread.

- **Paired two-sample t-test**

  Use when we have two non-independent groups.

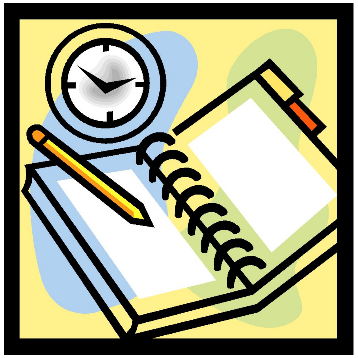
- **Non-parametric tests or transformations**

  Use when we cannot assume normality.

Summary – t-test

- Turn scientific question to null and alternative hypothesis
- Think about test assumptions
- Calculate summary statistics
- Carry out t-test if appropriate

T-tests practical



- Work through examples in the slides
- Complete the t-test practical

    First three datasets only

- We will start the next lecture after lunch (13:30pm)
- Feel free to take a short break if you want to

Tests for continuous variables non-parametric methods

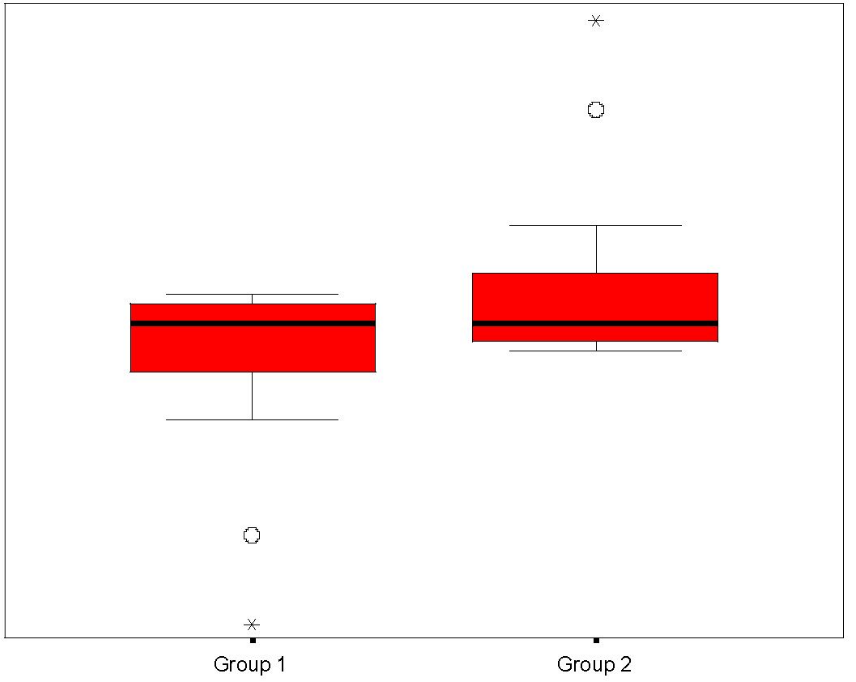
When to use which test

		RESPONSE		
NO OF SAMPLES		**NOMINAL**	**ORDINAL OR NON-NORMAL**	**NORMALLY DISTRIBUTED**
ONE SAMPLE		χ^2-test, Z-test	Kolmogorov-Smirnov Sign test	t-test
TWO SAMPLE	**INDEPENDENT**	χ^2-test (r x [illegible]), Fisher's exact test	Mann-Whitney U Median test	Unpaired t-test
	PAIRED	McNemar's test Stuart-Maxwell test	Wilcoxon signed rank Sign test	Paired t-test
MULTIPLE SAMPLES (K>2)	**INDEPENDENT**	χ^2-test (r x k) Fisher-Freeman-Halton	Kruskal-Wallis test Median Test Jonckheere-Terpstra test	Analysis of variance (ANOVA)
	PAIRED	Cochran Q test	Friedman test Page test Quade test	Repeated measures ANOVA
ASSOCIATION BETWEEN TWO VARIABLES		Contingency coefficient Phi, r_ϕ Cramér, C	Spearman's rank Kendall's tau	Pearson product moment correlation
AGREEMENT BETWEEN TWO VARIABLES		Simple kappa	Weighted kappa	Limits of agreement

Mann-Whitney U test

- Wilcoxon, Mann-Whitney
- Assumptions:
  - Two independent groups
  - At least ordinal dependent variable
  - Randomly selected observations
  - Population distributions same shape
- Hypotheses:
  - H_0: populations have the same median
  - H_0: populations have the same spread and shape

Misunderstood test



Statistics	Group 1	Group 2
Minimum	9.03	0.40
Median	9.94	9.94
Maximum	19.48	10.85
Mann-Whitney U	U=303, p=0.03	

Method

- Construct hypotheses and decide on α
- Rank whole sample from smallest to largest
- Assign average rank to ties
- Calculate sum of ranks for each group
- Calculate:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \qquad U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

$$U = \min(U_1, U_2)$$

- Compare U to critical value in the tables

Example

- Fisher’s book, coronary artery surgery study
- Exercise times in seconds, control and 3 vessel’s group
- Is there a difference in exercise times between the two groups, two-sided test

Group	Treadmill times in seconds									
Control	1014	684	810	990	840	978	1002	1110		
3 vessels	864	636	638	708	786	600	1320	750	594	750

Example

- Fisher’s book, coronary artery surgery study
- Exercise times in seconds, control and 3 vessel’s group
- Is there a difference in exercise times between the two groups, two-sided test

Group	Treadmill times in seconds									
Control	1014	684	810	990	840	978	1002	1110		
Rank	17	5	10	14	11	13	15	16		
3 vessels	864	636	638	708	786	600	1320	750	594	750
Rank	12	3	4	6	9	2	18	7.5	1	7.5

- Sum of ranks: control group =101, 3 vessels =70

$$U_1 = 8 \times 10 + \frac{8(8+1)}{2} - 101 = 15$$

$$U_2 = 8 \times 10 + \frac{10(10+1)}{2} - 70 = 65$$

$$U = \min(U_1, U_2) = \min(15, 65) = 15$$

- Look up n_1= 8, n_2= 10, p=2.5 (as 2–sided)
- U = 15 < 17, from tables
- Presentation of the results:
  – The Mann-Whitney U test showed that the individuals in the control group exercised for significantly longer than the individual in the 3 vessels disease group (U=15, p=0.025)

Advantages and limitations

- Almost as powerful as t-test
- Therefore almost as likely as t-test to reject H_0 if false
- Sensitive to central tendencies of scores
- Often misinterpreted
- Difference in medians if same shape distributions
- Otherwise tests for difference in spread and shape

When to use which test

		RESPONSE		
NO OF SAMPLES		**NOMINAL**	**ORDINAL OR NON-NORMAL**	**NORMALLY DISTRIBUTED**
ONE SAMPLE		χ^2-test, Z-test	Kolmogorov-Smirnov Sign test	t-test
TWO SAMPLE	**INDEPENDENT**	χ^2-test (r x c), Fisher's exact test	Mann-Whitney U Median test	Unpaired t-test
	PAIRED	McNemar's test Stuart-Maxwell test	Wilcoxon signed rank Sign test	Paired t-test
MULTIPLE SAMPLES (K>2)	**INDEPENDENT**	χ^2-test (r x k) Fisher-Freeman-Halton	Kruskal-Wallis test Median Test Jonckheere-Terpstra test	Analysis of variance (ANOVA)
	PAIRED	Cochran Q test	Friedman test Page test Quade test	Repeated measures ANOVA
ASSOCIATION BETWEEN TWO VARIABLES		Contingency coefficient Phi, r_ϕ Cramér, C	Spearman's rank Kendall's tau	Pearson product moment correlation
AGREEMENT BETWEEN TWO VARIABLES		Simple kappa	Weighted kappa	Limits of agreement

Sign Test

- A very simple test
  - Based on binomial distribution
- Uses directions of differences
- One-sample case: compares to fixed value
- Two-sample case: compares medians
- Can be used when it’s possible to say one quantity is greater than another

Sign Test

- Assumptions:
  - Order in coding system
  - Randomly selected observations
  - Paired data in two-sample case
- Hypotheses:
  - H_0: medians equal in two groups
  - H_A: medians in two groups differ

Method

- One-sample: compare values to m
  - + if bigger, – if smaller, = equal
- Two-sample: compare values to each other
  - + if 1st largest, – if 2nd largest, = equal
- Count +, –, =
  - x = number of smaller values
  - r = number of non-ties
  - p = 0.5 (probability, not p-value)
- Compare to binomial tables

One-Sample Example

- General health section of SF-36 collected in a breast cancer study
- Expected value in general population 72

GH value	60	55	75	100	55	60	50	60	72	40	90	75	70	75	55
Sign	-	-	+	+	-	-	-	-	=	-	+	+	-	+	-

- Number of non–ties = 14
- 9 – < 5 + □smaller value = 5
- Look up n=14, p=0.5, x=5 in binomial tables

One-Sample Example

- General health section of SF-36 collected in a breast cancer study
- Expected value in general population 72

GH value	60	55	75	100	55	60	50	60	72	40	90	75	70	75	55
Sign	-	-	+	+	-	-	-	-	=	-	+	+	-	+	-

- P = 0.42
- Therefore insufficient evidence to reject H_0
- Conclude median value not different to 72

Two-Sample Example

- General health values collected in same study at a 2nd time point
- Is there a difference between the time points?

Time 1	60	55	75	100	55	60	50	60	72	40	90	75	70	75	55
Time 2	70	65	100	50	70	95	95	65	85	55	95	45	75	65	60
Sign	-	-	-	+	-	-	-	-	-	-	-	+	-	+	-

- Number of non-ties = 15
- 12 – > 3 + □ smaller value is 3
- Look up n=15, p=0.5, x=3 in binomial tables

Two-Sample Example

- General health values collected in same study at a 2nd time point
- Is there a difference between the time points?

Time 1	60	55	75	100	55	60	50	60	72	40	90	75	70	75	55
Time 2	70	65	100	50	70	95	95	65	85	55	95	45	75	65	60
Sign	-	-	-	+	-	-	-	-	-	-	-	+	-	+	-

- Pr = 0.035, sufficient evidence to reject H_0
- There is a difference in General health between the two time points

Presentation of the Results

- One-sample case:
  - There is no evidence of a difference in median general health value of 60 in this population and that of 72 in the general population ($p=0.42$, sign test).
- Two-sample case
  - The median general health value at the second time point, 70 was significantly higher than the median of 60 at the first time point, ($p=0.035$, sign test).

Advantages and Limitations

- Simple
- Probability can be adjusted
- Quick assessment of direction
- Less powerful than other tests
  - Does not consider magnitude

When to Use Which Test

		RESPONSE		
NO OF SAMPLES		NOMINAL	ORDINAL OR NON-NORMAL	NORMALLY DISTRIBUTED
ONE SAMPLE		χ^2-test, Z-test	Kolmogorov-Smirnov Sign test	t-test
TWO SAMPLE	INDEPENDENT	χ^2-test (r x c), Fisher's exact test	Mann-Whitney U Median test	Unpaired t-test
	PAIRED	McNemar's test Stuart-Maxwell test	Wilcoxon signed rank Sign test	Paired t-test
MULTIPLE SAMPLES (K>2)	INDEPENDENT	χ^2-test (r x k) Fisher-Freeman-Halton	Kruskal-Wallis test Median Test Jonckheere-Terpstra test	Analysis of variance (ANOVA)
	PAIRED	Cochran Q test	Friedman test Page test Quade test	Repeated measures ANOVA
ASSOCIATION BETWEEN TWO VARIABLES		Contingency coefficient Phi, r_ϕ Cramér, C	Spearman's rank Kendall's tau	Pearson product moment correlation
AGREEMENT BETWEEN TWO VARIABLES		Simple kappa	Weighted kappa	Limits of agreement

Wilcoxon Signed Rank Test

- Alternative to sign test
- Assumptions:
  - Single sample in pairs, matched or before/after
  - Continuous or ordinal data (no normality assump)
  - Symmetry of difference scores about true median difference (test with plot)
- Hypothesis:
  - H_0: sum positive ranks equals sum negative ranks
  - H_A: sum positive ranks is not equal sum negative ranks

Method

- Construct hypotheses and decide α
- Find difference for each subject
- Rank magnitude of differences
- Put sign of difference with rank
- Find sum of positive and negative ranks
- Compare smaller sum to critical value from tables

Example

- Taken from Glanz' book, data are urine production before/after diuretic
- Is there a difference? Two-sided test

Person	Daily urine production ml/day	
	Before drug	After drug
1	1600	1490
2	1850	1300
3	1300	1400
4	1500	1410
5	1400	1350
6	1010	1000
7	1750	1750

Example

Person	Daily urine production ml/day			Rank of difference	Signed rank of difference
	Before drug	After drug	Difference		
1	1600	1490	-110	5	-5
2	1850	1300	-550	6	-6
3	1300	1400	+100	4	+4
4	1500	1410	-90	3	-3
5	1400	1350	-50	2	-2
6	1010	1000	-10	1	-1
7	1750	1750	0	-	-

- $W^{+}=4 < W^{-}=17$ look up n= 6, P=2.5 in tables

Results

- As $W^+ > 0$ not sufficient evidence to reject null hypothesis
- Conclude that there is no evidence of a change in urine production before and after drug
- Presentation of the results:
  - The Wilcoxon signed rank test showed that there was no evidence of a change in urine production before and after treatment ($W=4$, $p=0.22$).

Advantages and Limitations

- Easy to apply
- Powerful
  - Takes into account more information
- Computer output confusing
- Sometimes misinterpreted

Summary-Two Independent Samples

- **t-test**: a test for comparing means in two independent groups when the data are consistent with a normal distribution.
- **Mann-Whitney U test (Wilcoxon Rank Sum test)**: If the assumption of similarity of distributions holds it is a test for comparing medians in two independent groups. Otherwise it compares the shape and spread of the two groups.

Summary-Paired Groups

- **One-sample sign test**: is for comparing the median to a proposed value in the population.
- **Two-sample sign test**: is for comparing the medians between matched pairs.
- **Wilcoxon signed rank test**: if there is a symmetry of difference scores about the true median difference it compares means.
- **Paired t-test**: if the within pair differences are consistent with a normal distribution compares means.

Tests for categorical variables

Associations between categorical variables

- All about frequencies!
- Row x Column table (2 x 2 simplest)
- Categorical data

Treatment group	Tumour shrinkage	
	No	Yes
Treatment	44	40
Placebo	24	16

2 x 2

- Look for association (relationship) between row variable and column variable

Chi-square test

- **E.g. Research question:** A trial to assess the effectiveness of a new treatment versus a placebo in reducing tumour size in patients with ovarian cancer.

Treatment group	Tumour shrinkage	
	No	Yes
Treatment	44	40
Placebo	24	16

- Is there an association between treatment group and tumour shrinkage?
- **Null hypothesis, H_0 :** No association
- **Alternative hypothesis, H_1 :** Some association

Chi-square test

Calculating expected frequencies:

Treatment group	Tumour shrinkage		Total
	No	Yes	
Treatment	44 46.1	40 37.9	84
Placebo	24 21.9	16 18.1	40
Total	68	56	124

$$E = \frac{\text{row total} \times \text{col total}}{\text{overall total}}$$

$$\text{e.g. } \frac{84}{124} \times \frac{68}{124} \times 124 = \frac{84 \times 68}{124} = 46.1$$

Chi-square test

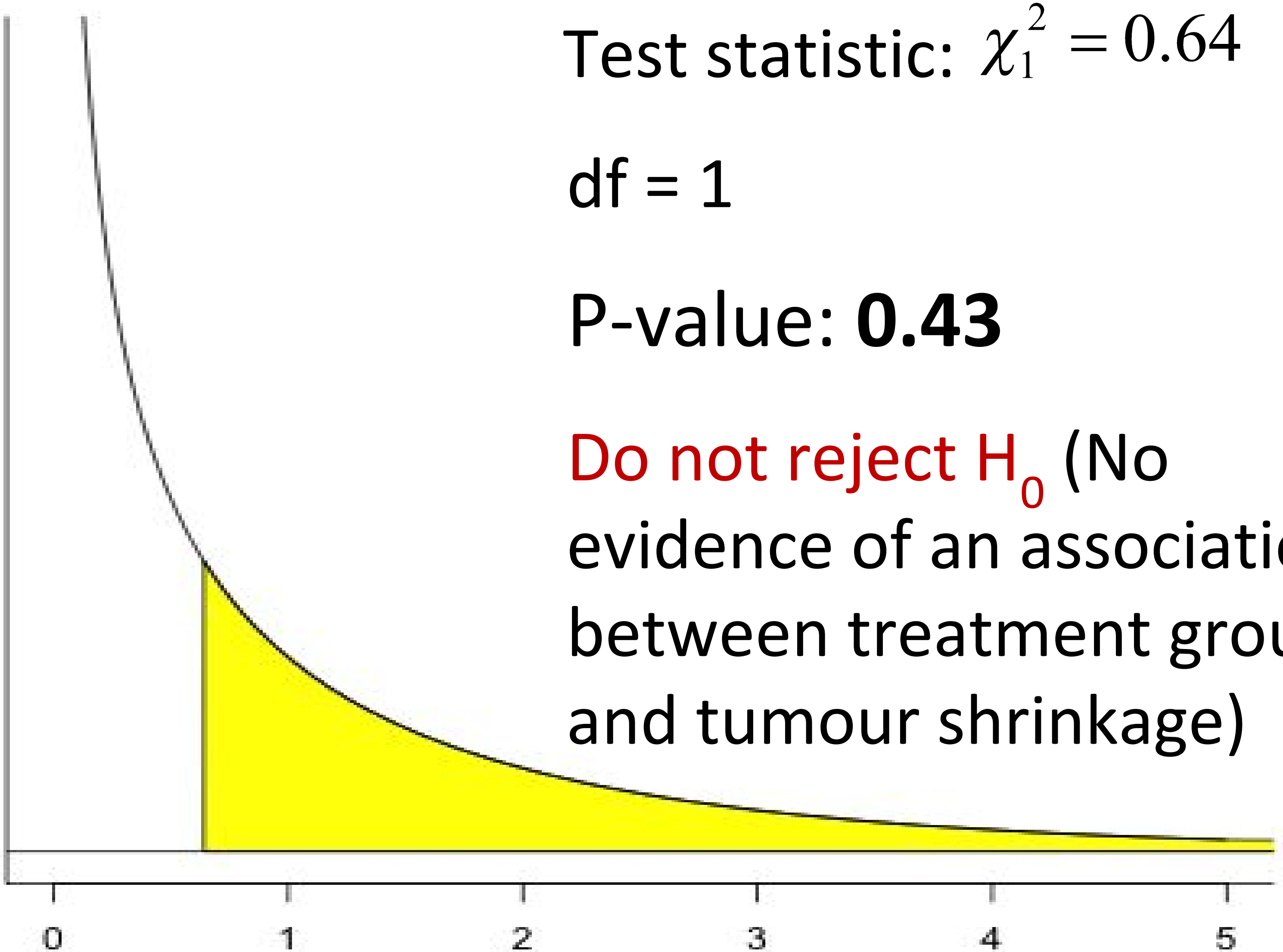
Calculating the chi-square statistic:

Treatment group	Tumour shrinkage		Total
	No	Yes	
Treatment	44 46.1	40 37.9	84
Placebo	24 21.9	16 18.1	40
Total	68	56	124

$$\chi^2_{(r-1)\times(c-1)} = \sum \frac{(O-E)^2}{E}$$

$$\chi^2_{(r-1)\times(c-1)} = \sum \frac{(O-E)^2}{E} = \frac{(44-46.1)^2}{46.1} + \frac{(40-37.9)^2}{37.9} + \frac{(24-21.9)^2}{21.9} + \frac{(16-18.1)^2}{18.1} = 0.64$$

Chi-square test



Test statistic: $\chi_1^2 = 0.64$

df = 1

P-value: **0.43**

Do not reject H_0 (No evidence of an association between treatment group and tumour shrinkage)

Limitations of the chi-square test

- In general, a Chi-square test is appropriate when:
  - at least 80% of the cells have an expected frequency of 5 or greater
  - none of the cells have an expected frequency less than 1
- If these conditions aren't met, Fisher's exact test should be used.

Same question, smaller sample size

- **E.g. Research question**: Is there an association between treatment group and tumour shrinkage?

Treatment group	Tumour shrinkage		Total
	No	Yes	
Treatment	8	3	11
Placebo	9	4	13
Total	17	7	24

- **Null hypothesis, H_0** : No association
- **Alternative hypothesis, H_1 :** Some association

Expected frequencies

$$E = \frac{\text{row total} \times \text{col total}}{\text{overall total}}$$

Treatment group	Tumour shrinkage		Total
	No	Yes	
Treatment	8 7.8	3 3.2	11
Placebo	9 9.2	4 3.8	13
Total	17	7	24

Expected frequency less than 5

Only 50% of cells have an expected frequency greater than 5 → use Fisher's exact test

e.g. $\frac{11}{24} \times \frac{17}{24} \times 24 = \frac{11 \times 17}{24} = 7.8$

Fisher's exact test - results

Treatment group	Tumour shrinkage		Total
	No	Yes	
Treatment	8 7.8	3 3.2	11
Placebo	9 9.2	4 3.8	13
Total	17	7	24

- Test statistic: **N/A**
- P-value: **1.00**
- Interpretation: Do not reject H_0 (No evidence of an association between treatment group and tumour shrinkage).

Chi-square test for trend

- **E.g. Research question:** Is there a linear association between tumour grade and the incidence of tumour shrinkage?

Tumour grade	Tumour shrinkage		Total
	No	Yes	
2	18	5	23
3	15	14	27
4	11	21	34
Total	44	40	84

- **Null hypothesis, H_0:** No linear association
- **Alternative hypothesis, H_1 :** Some linear association

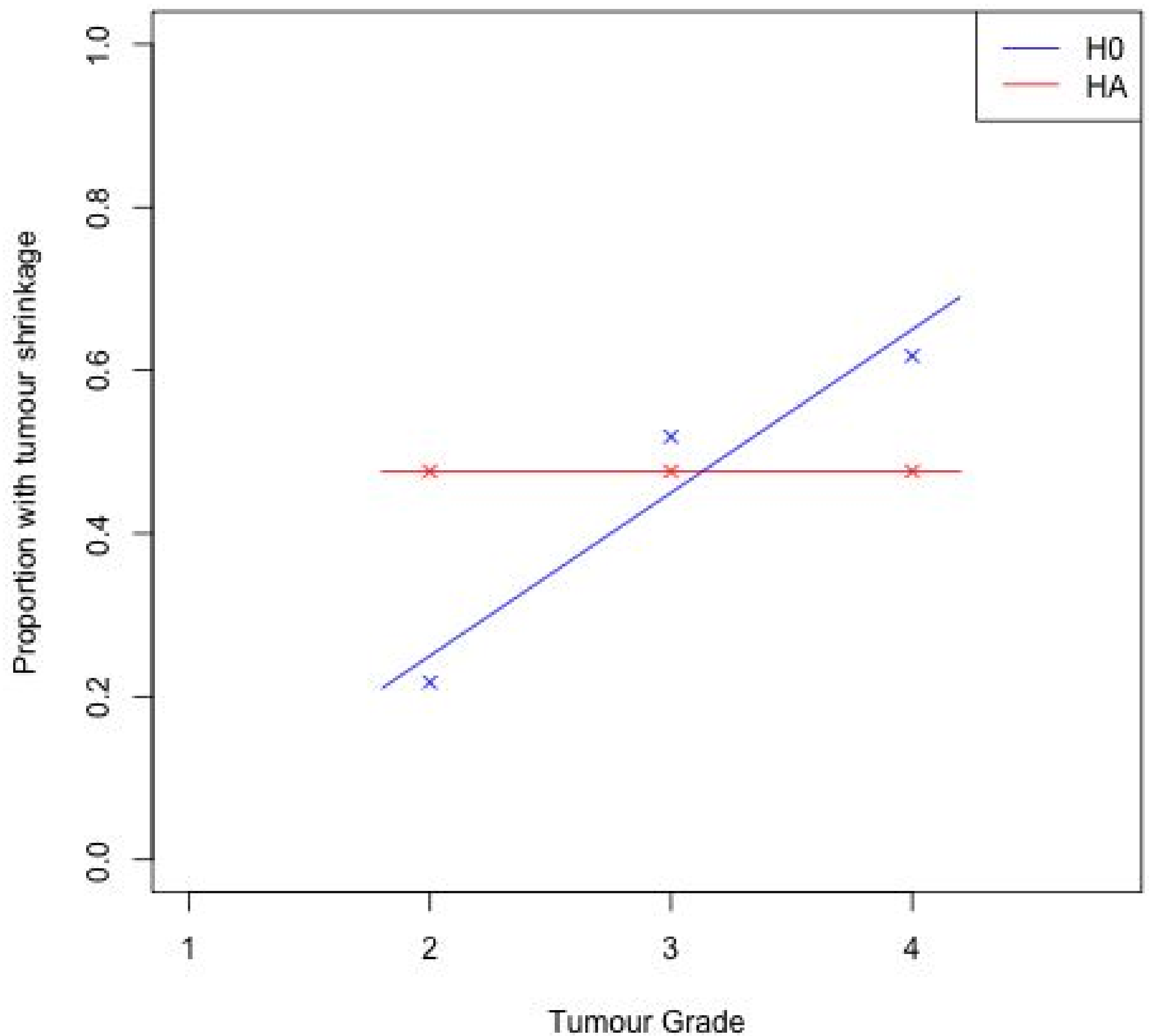
Expected frequencies

$$E = \frac{\text{row total} \times \text{col total}}{\text{overall total}}$$

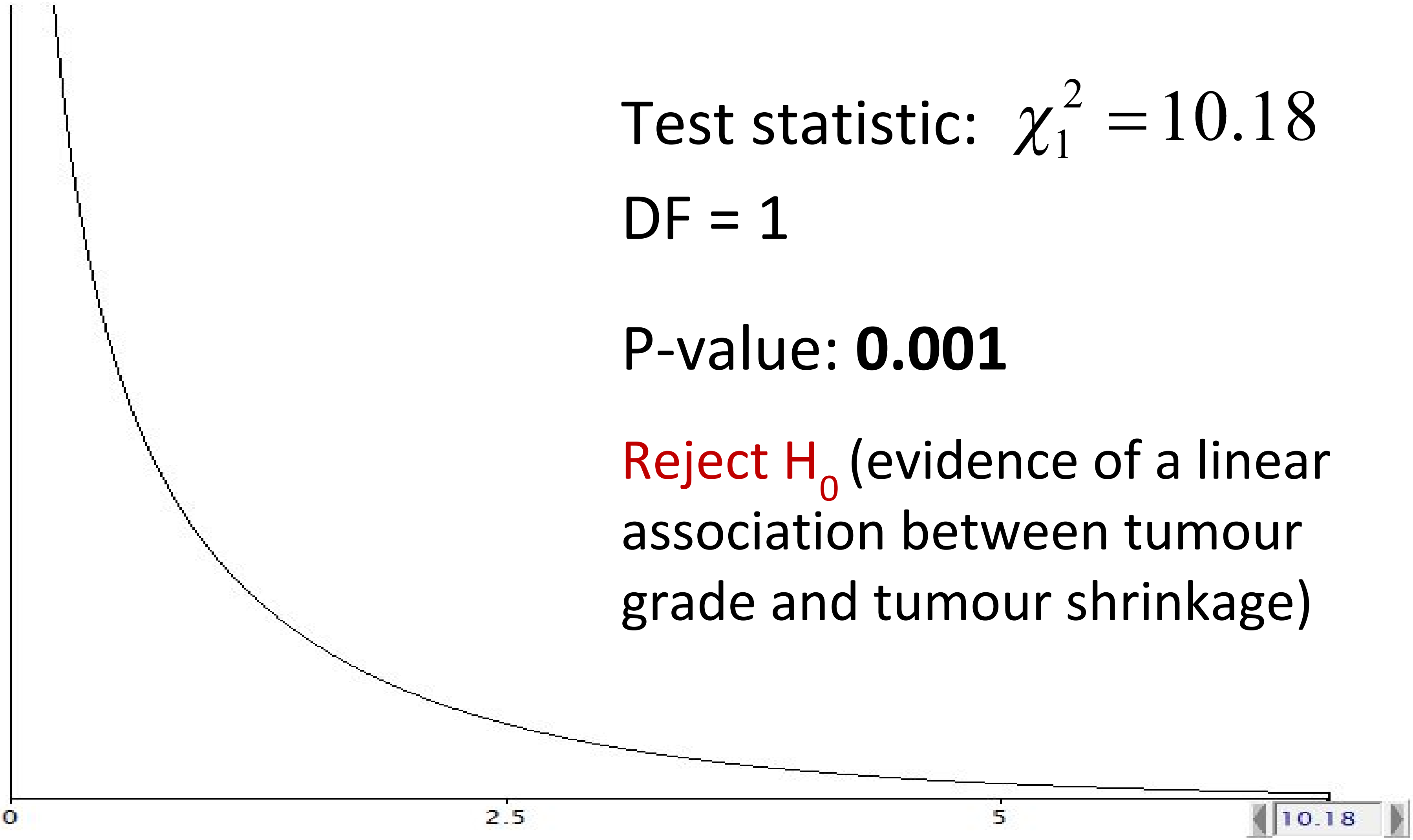
Tumour grade	Tumour shrinkage		Total
	No	Yes	
2	18 12.0	5 11.0	23
3	15 14.1	14 12.9	27
4	11 17.8	21 16.2	34
Total	44	40	84

e.g. $\frac{44}{84} \times \frac{23}{84} \times 84 = \frac{44 \times 23}{84} = 12.0$

Chi-square test for trend



Chi-square test for trend - results



Test statistic: $\chi_1^2 = 10.18$

DF = 1

P-value: **0.001**

Reject H_0 (evidence of a linear association between tumour grade and tumour shrinkage)

Summary – categorical variables

- **Chi-square test**

  Use when we have two categorical variables, each with two or more levels, and our expected frequencies **are not** too small.

- **Fishers exact test**

  Use when we have two categorical variables, each with two levels, and our expected frequencies **are** small.

- **Chi-square test for trend**

  Use when we have two categorical variables, where one or both are naturally ordered and the ordered variable has at least three levels, and our expected frequencies **are not** too small.

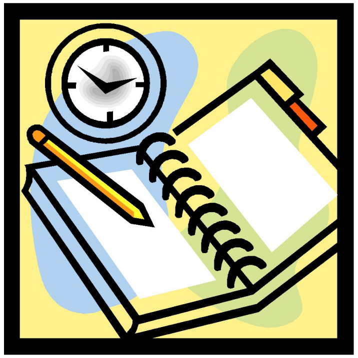
- **McNemar's test**

  Use when we have two categorical paired variables.

Summary – contingency tables

- Turn scientific question to null and alternative hypothesis
- Calculate expected frequencies
- Think about test assumptions
- Carry out chi-square or Fishers test if appropriate

Contingency table practical



- Work through examples in the slides
- Complete contingency table practical
- We will have solutions before moving to the group exercise

Summary

- For independent observations
- For normally distributed continuous outcomes - T-tests
- For categorical outcomes - Chi-squared tests
- Confidence interval tell us more of story than p-value
- Limitations
  - Confounding – can adjust for important factors by stratification or regression

References

1. *Essential Medical Statistics, Betty Kirkwood and Jonathan Sterne, Wiley-Blackwell, 2nd Edition 2003.*
2. *Practical Statistics for Medical Research, Douglas G. Altman, Chapman & Hall / CRC, 1999.*
3. *Ten Simple Rules for Effective Statistical Practice, Robert E. Kass, Brian S. Caffo, Marie Davidian, Xio-Li Meng, Bin Yu, Nancy Reid* http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004961

Finally…

- Course Materials:-
- http://tiny.cc/crukStats
- Course Feedback:-
- http://tinyurl.com/stats-june21