Introduction to Statistical analysis

Mark Dunning. Materials by Deepak Parashar, Sarah Dawson and Sarah Vowler

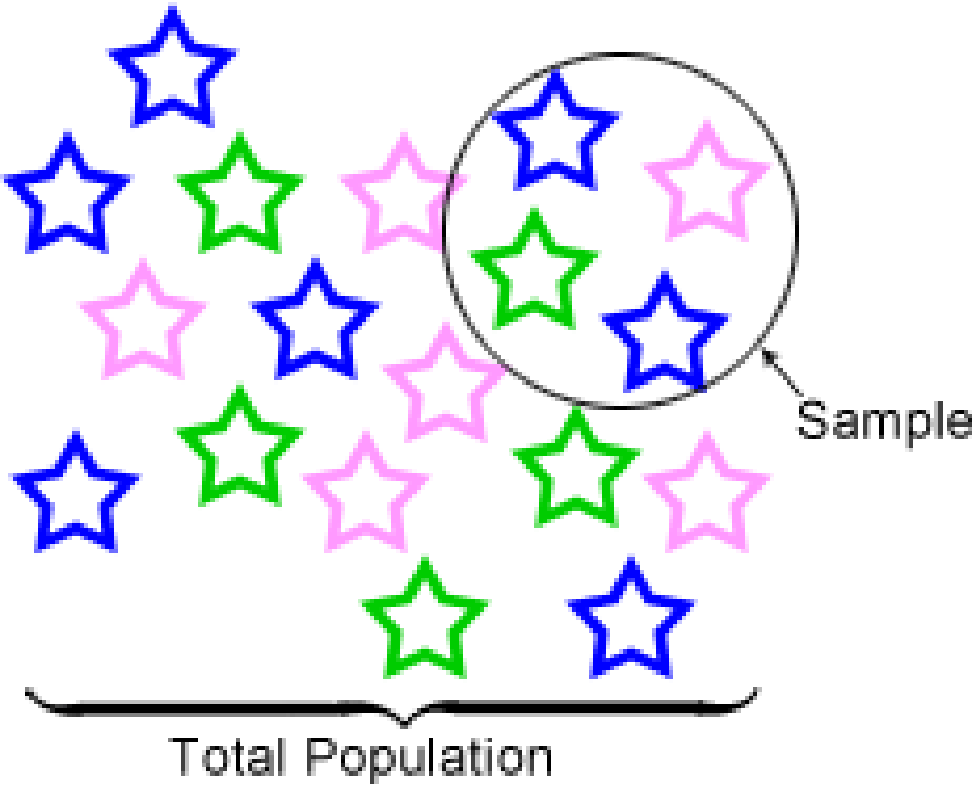
Cancer Research Uk
Cambridge Research Institute
Robinson Way
Cambridge

2nd September 2014

Outline

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

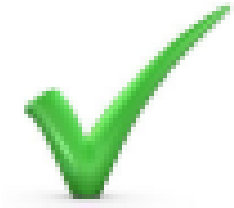
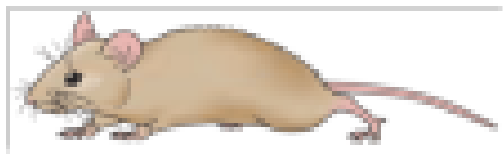
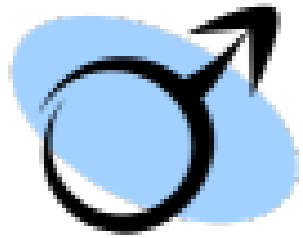


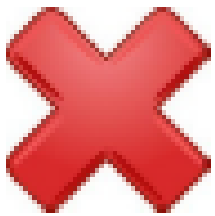
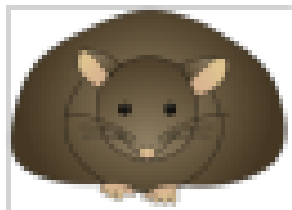
Data

- Type?
  - Categorical (nominal) , e.g. Gender
  - Categorical with ordering (ordinal), e.g. Tumour grade
  - Discrete, e.g. Shoe size
  - Continuous, e.g. Body weight in kg
- Independent or dependent measurements
- Representative of which population?
- Distribution
  - Normally distributed? Skewed? Bimodal?

Data type - example

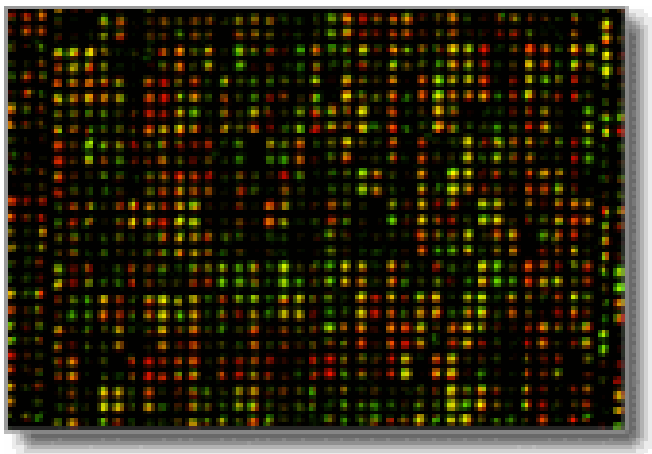
- Sucess / failure of achieving a taks for a mouse which may be wild-type of knock-out, male or female, 2, 4 or 6 months old

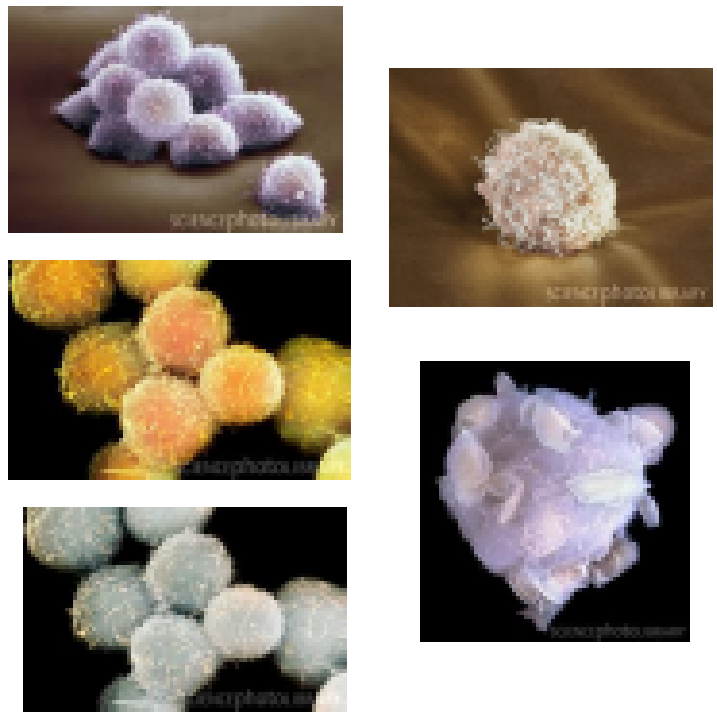




Data type - example

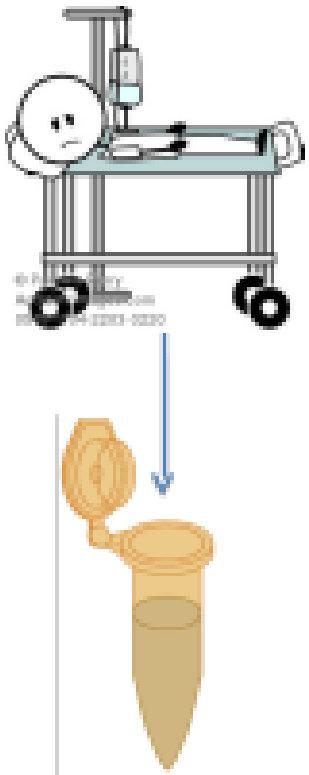
- Gene expression in each cell sample which may be one of five cell-types (A,B,C,D,E)

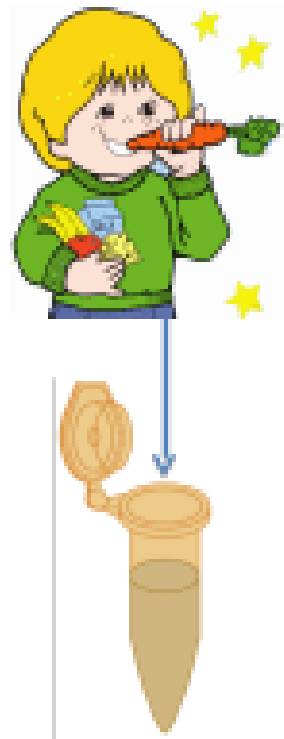




Data type - example

- The number of bacteria for each subject which may be a cancer patient or a normal





Mesurements: Dependent / Independent?

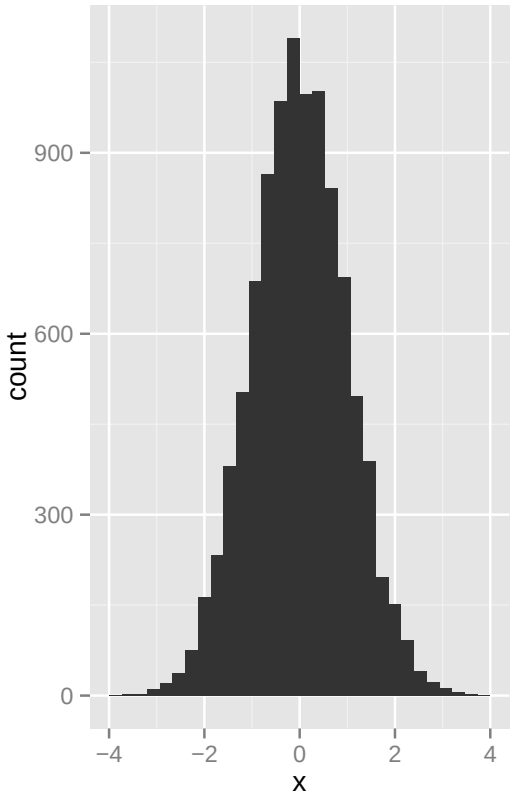
- Measurements of gene expression taken from each of 20 individuals
- Are any measurements more closely related than others?
  - Siblings / littermates?
  - Same individual measured twice?
  - Batch effects?
- If no reason - independent observations

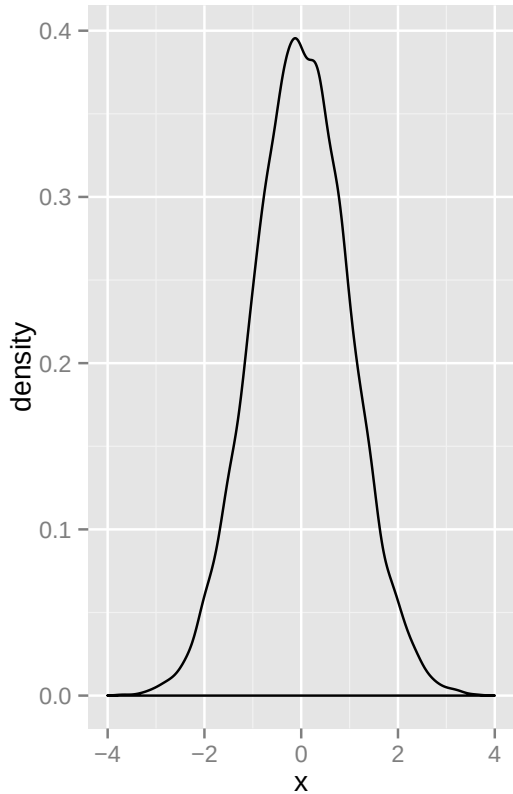
Independence is a common assumption for statistical tests

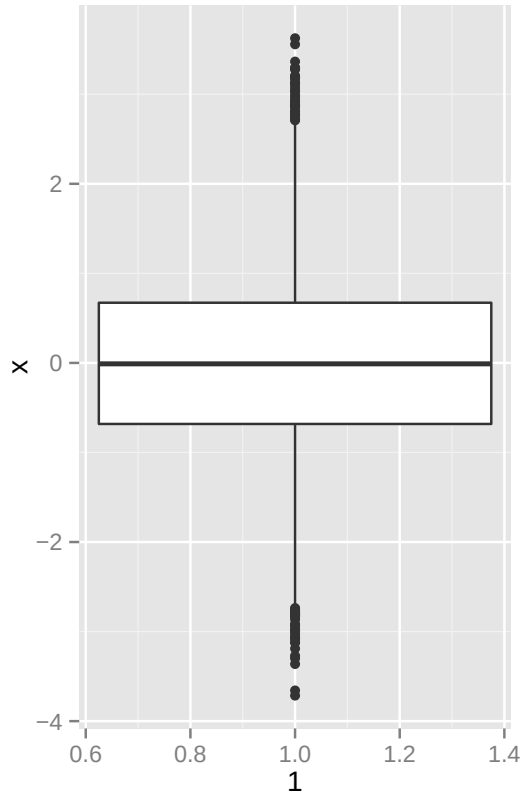
Mesurements: Dependent / Independent?

- Measuring blood pressure before and after treatment for 30 patients

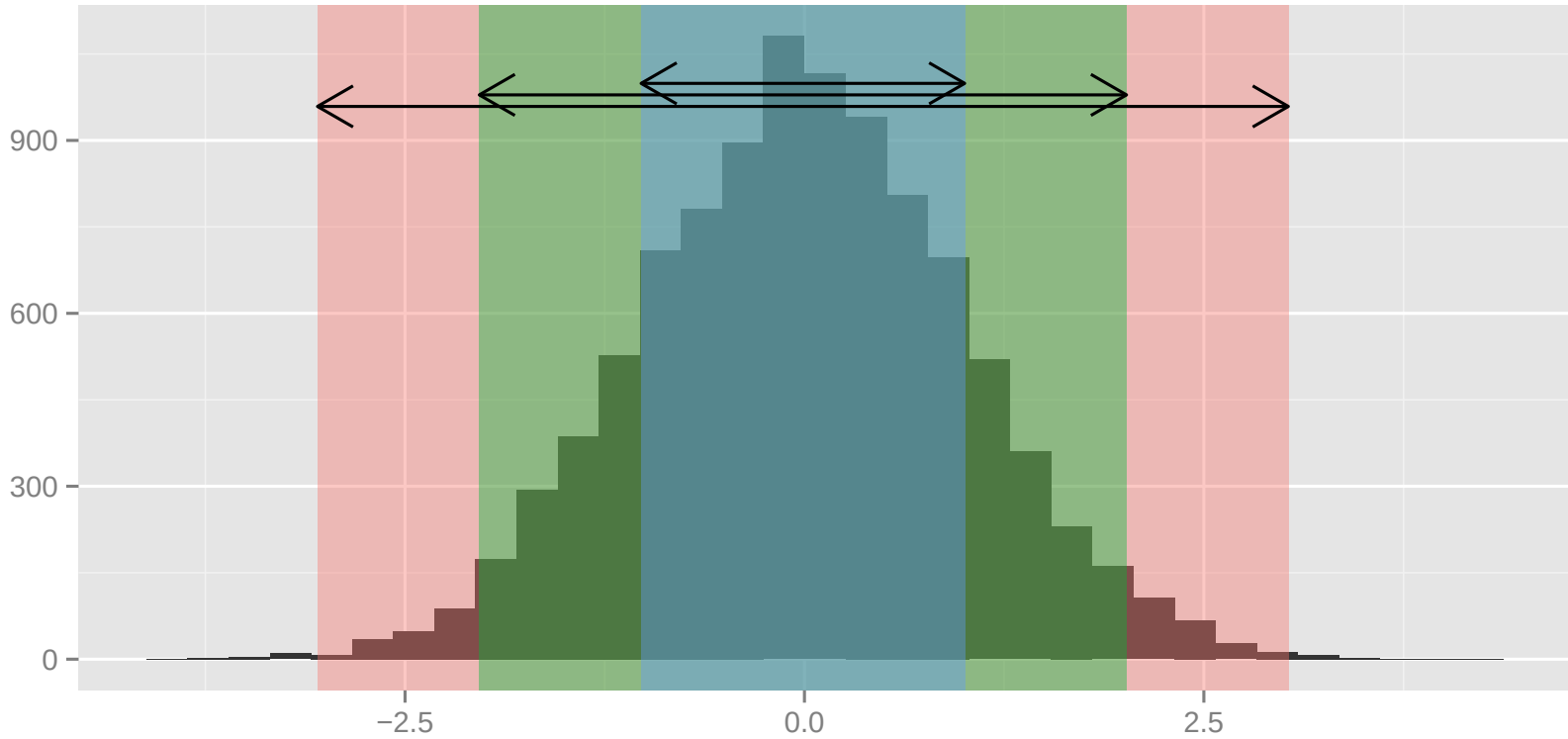
Continuous Data - Distribution





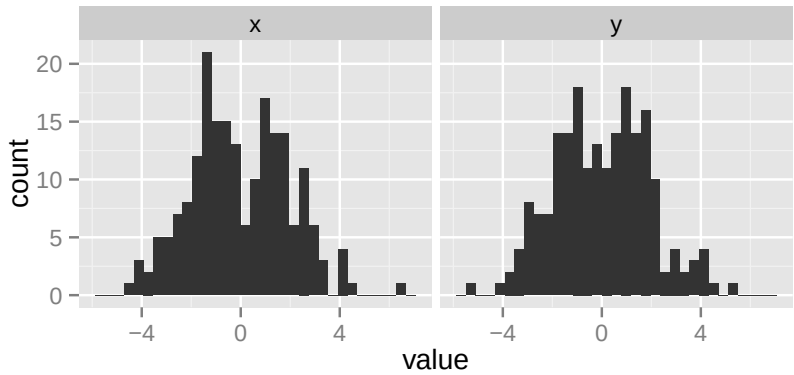


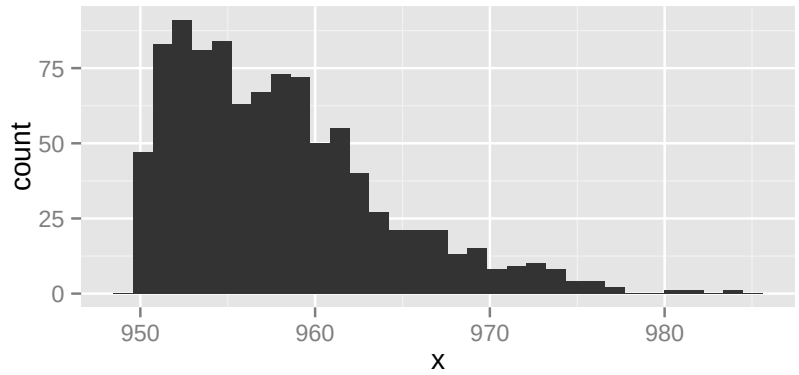
Continuous Data - Distribution

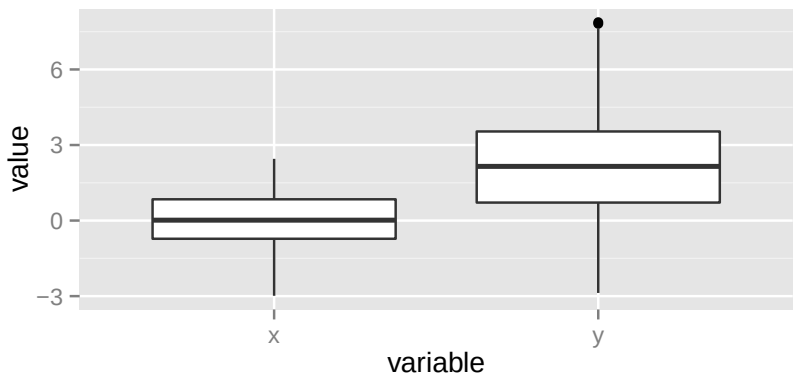


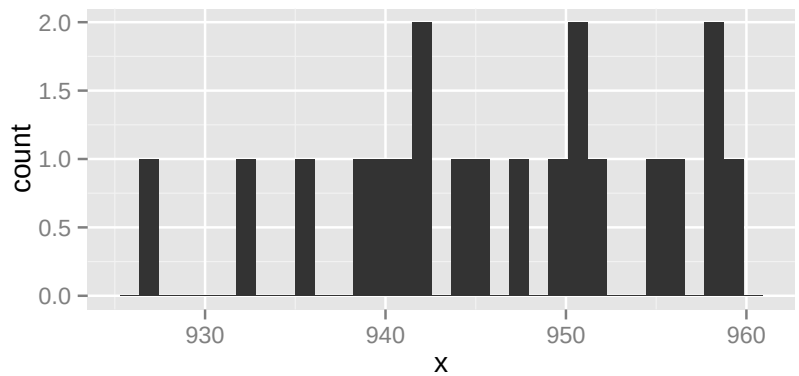
66% are within one standard deviation, 95% are within two standard deviations, 99% are within three standard deviations

Normal, or not??



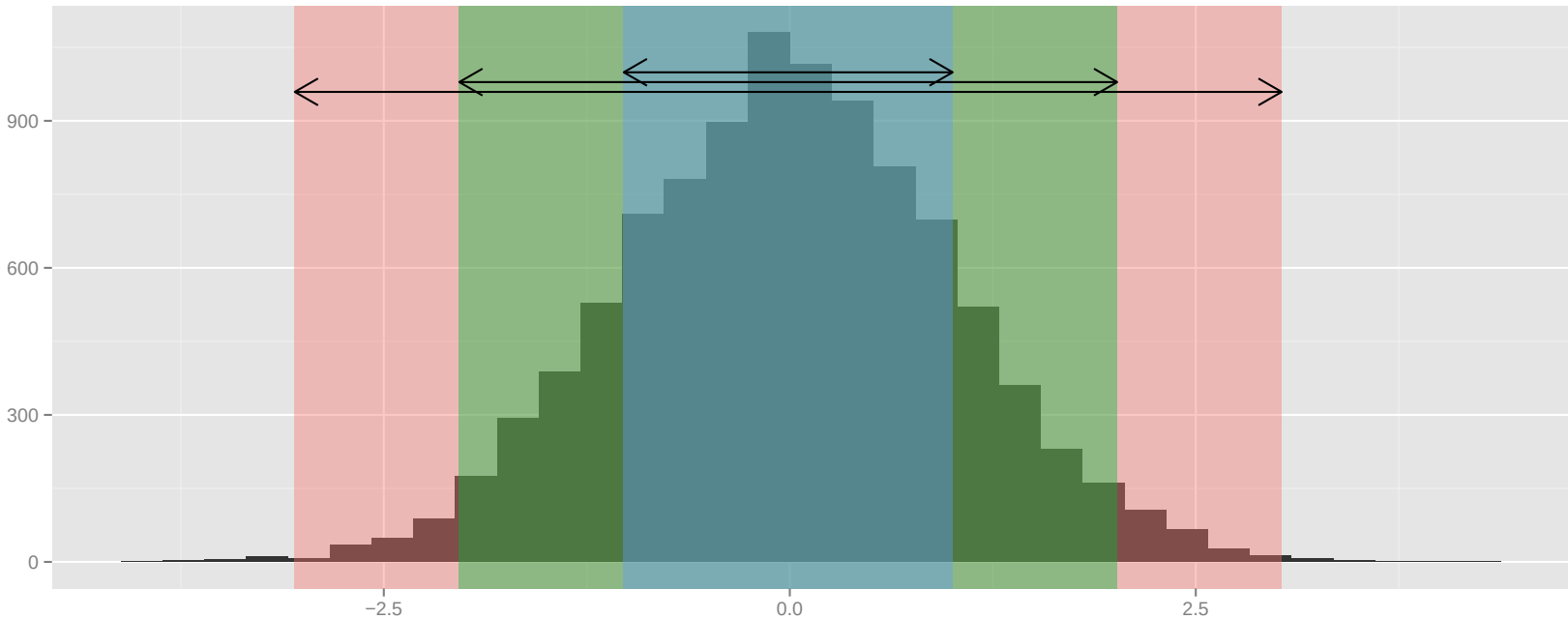






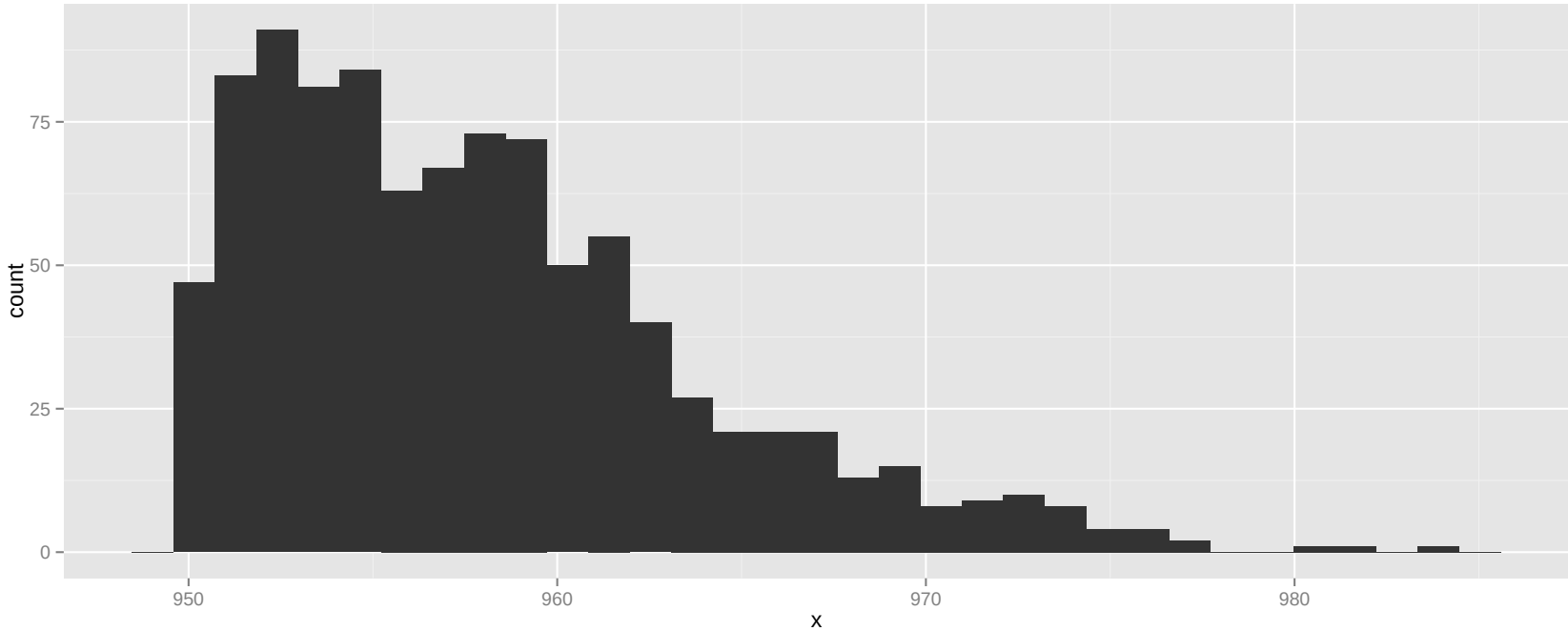
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Continuous Data - Descriptive Statistics



- Measures of location and spread
- Mean and standard deviation
  - $\bar{X} = \frac{X_1 + X_2 + \ldots X_n}{n}$
  - $s.d = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2}$

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

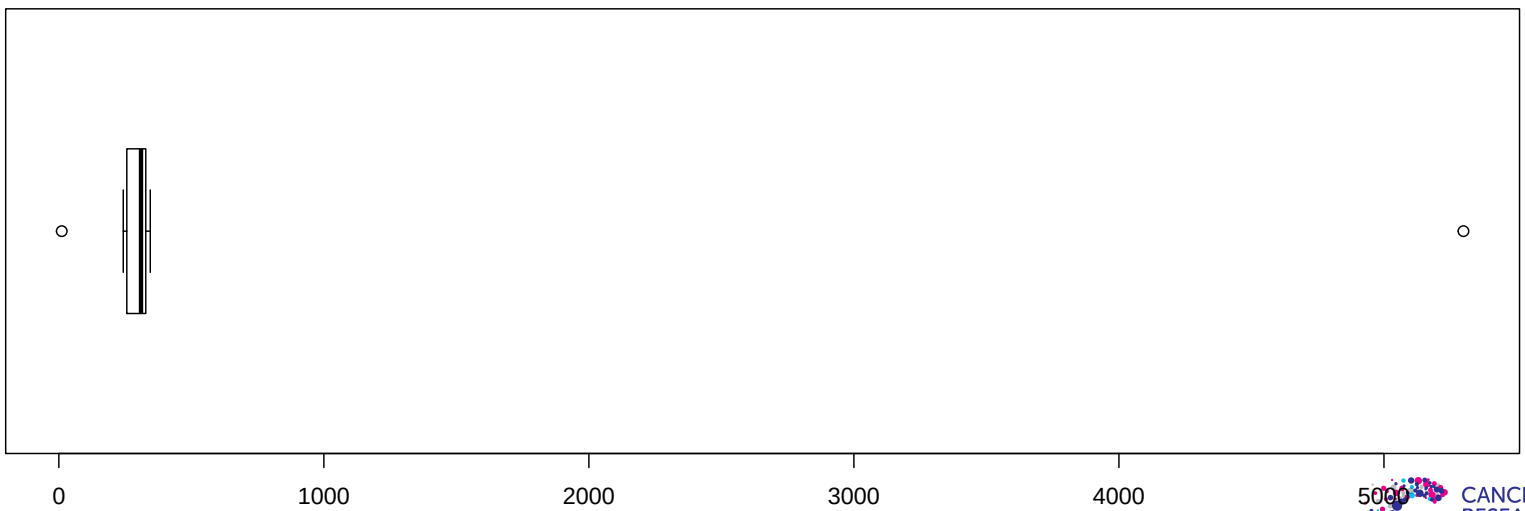
- Measures of location and spread
- Mean and standard deviation
  - $\bar{X} = \frac{X_1 + X_2 + \ldots X_n}{n} = 970$
  - $s.d = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2} = 1912.57$
- Median and interquartile range
  - 11, 243, 270, 310, 311, 345,5300

Continuous Data - Descriptive Statistics Example

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

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

- Measures of location and spread
- Mean and standard deviation
  - $\bar{X} = \frac{X_1 + X_2 + \ldots X_n}{n} = 970$
  - $s.d = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2} = 1912.57$
- 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

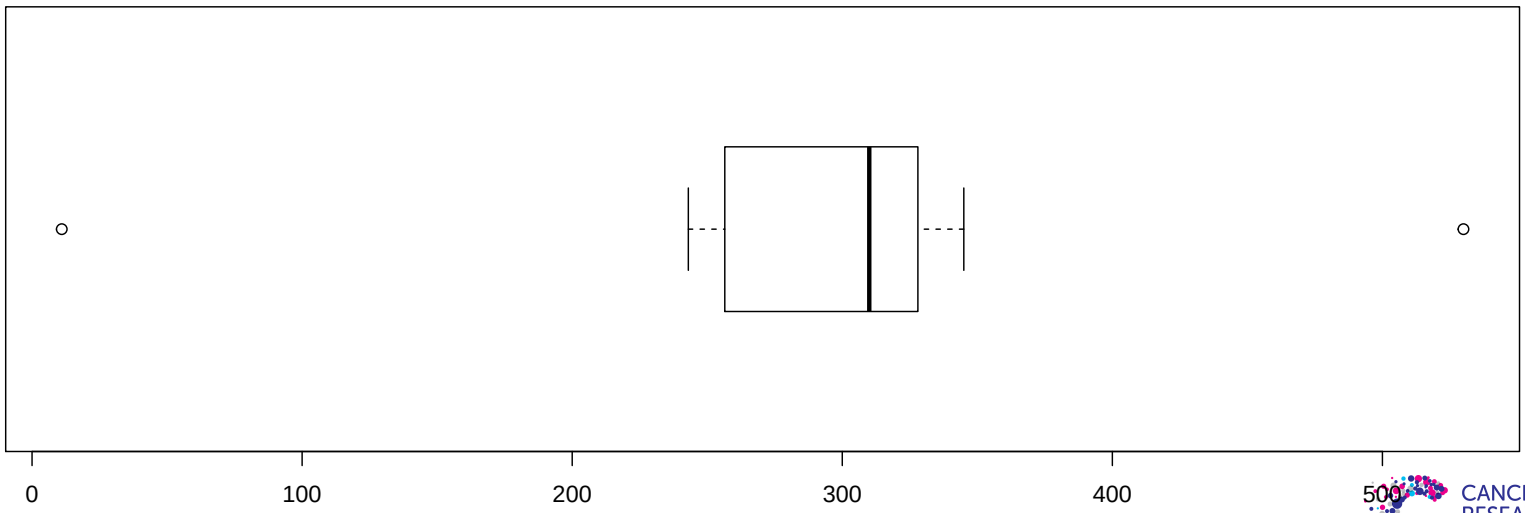
- Measures of location and spread
- Mean and standard deviation
  - $\bar{X} = \frac{X_1 + X_2 + \ldots X_n}{n} = 289$
  - $s.d = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2} = 153.79$
- Median and interquartile range
  - 11, 243, 270, 310, 311, 345,530

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
  - $\bar{X} = \frac{X_1 + X_2 + \ldots X_n}{n} = 289$
  - $s.d = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2} = 153.79$
- Median and interquartile range
  - 11, 243, 270, 310, 311, 345,530



Standard Deviation and Standard Error

- Commonly confused
- Standard deviation
  - Measure of spread of data
  - Used for describing **population**
- Standard error
  - Variability of the mean from repeated sampling
  - Precision of the 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 95% CI:
  $(\bar{X} - 1.96 \times se, \bar{X} + 1.96 \times se)$

For facebook friends data:

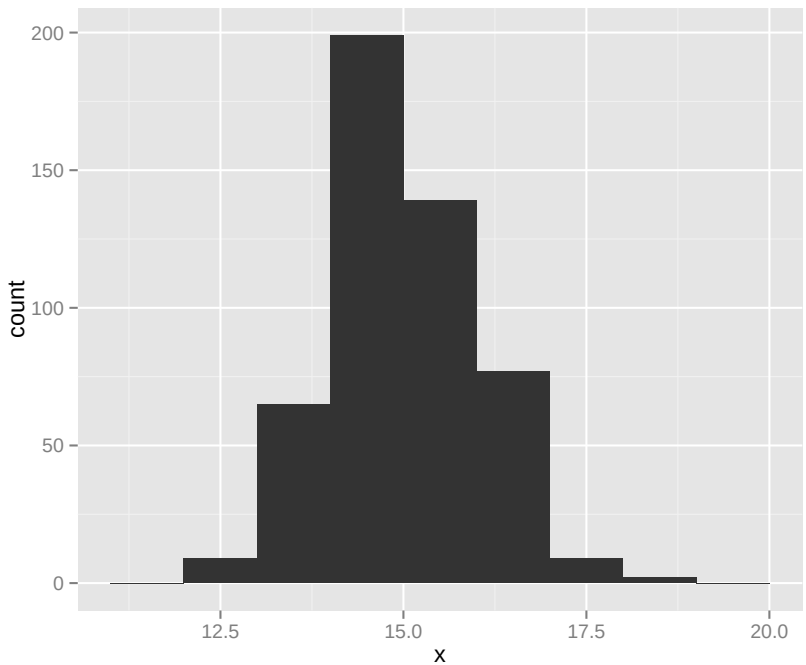
$se = \frac{sd}{\sqrt{(n)}} = \frac{154}{\sqrt{7}} = 58$

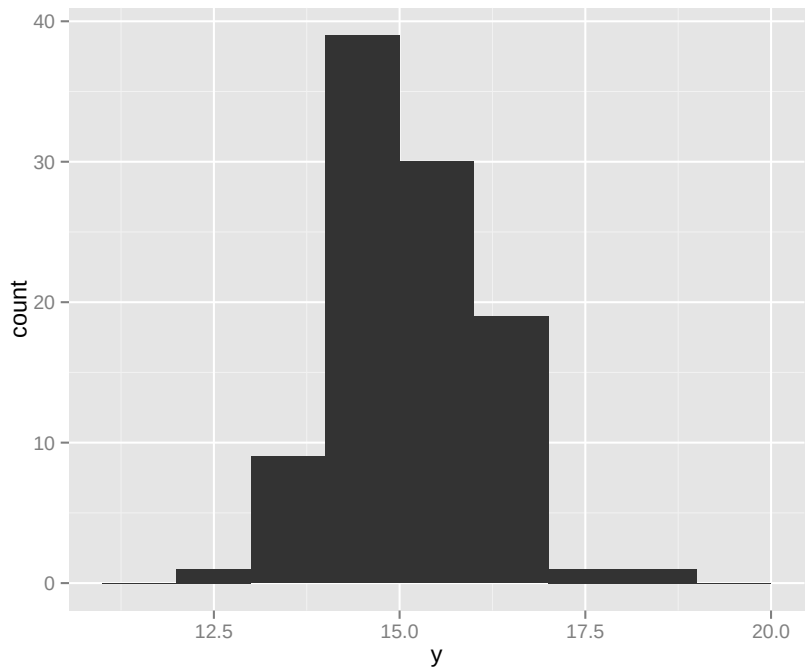
Mean 289, 95% CI ($289 - (1.96 \times 58), 289 + (1.96 \times 58)$)

Mean 289, 95% CI (175, 402)

Confidence intervals

- As number of observations goes up....
- Standard deviation stays **the same**...
- But standard error **goes down....**





Basic set-up

- Formulate a *null hypothesis* - H_0
- e.g. *"The difference in treatment before and after treatment = 0"*
- Calculate a test statistic from the data under the null hypothesis
- Determine whether the test statistic is more extreme than expected under the null hypothesis
- Reject or do no reject the null hypothesis
- Absence of evidence is not evidence of absense (Bland and Altman, 1995)

Example

Lady Tasting Tea (Randomised experiment by Fisher)

- Randomly-ordered 8 cups of tea
  - 4 were prepared by adding milk first
  - 4 were prepared by adding tea first
- Task: Lady had to select the 4 cups of one particular method
- H_0 Lady no such ability
- Test Statistic: number of sucessses in selecting 4 cups
- Result: Lady got all 4 cups correct
- **Reject the null hypothesis**

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

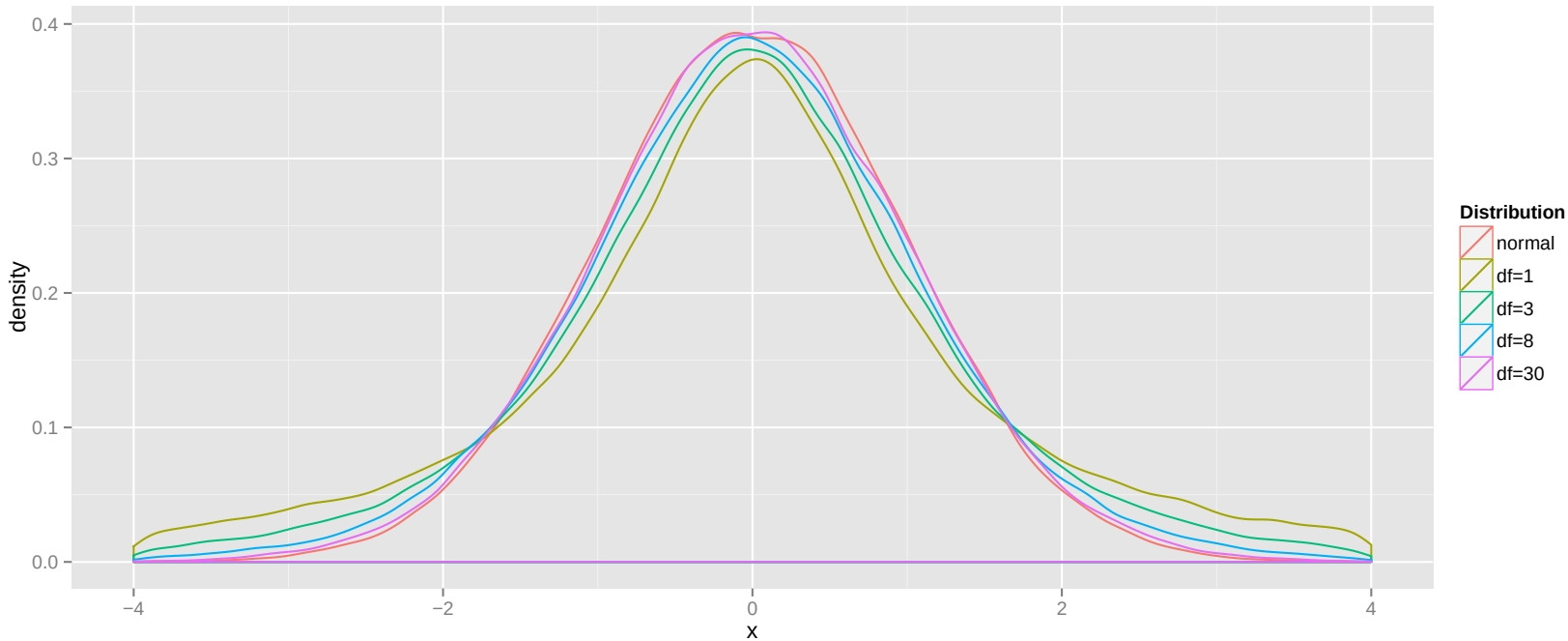
Significance level, sample size, different of interest, variability of the observations

Be aware of issues of multiple testing

Various flavours

- 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



Does mean = X?

- **Research question:** Published data suggests that the microarray failure rate for a particular supplier is 2.1%
- Genomics want to know if this holds true in their own lab

Formulating the question

- 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 - **never accept the alternative hypothesis**

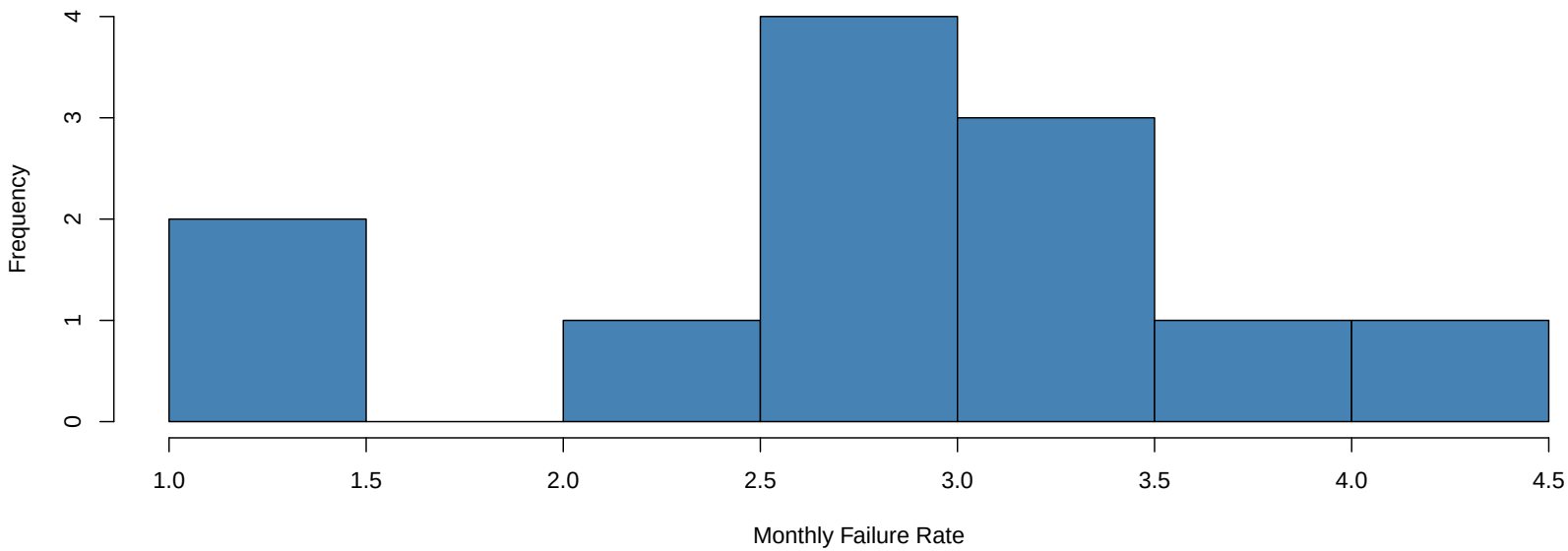
The Data

	Month	Monthly.failure.rate
1	January	2.90
2	February	2.99
3	March	2.48
4	April	1.48
5	May	2.71
6	June	4.17
7	July	3.74
8	August	3.04
9	September	1.23
10	October	2.72
11	November	3.23
12	December	3.40

Summary Statistics

mean $= (2.9 + \cdots + 3.40)/12 = 2.841$
Standard deviation $= 0.837$

- Observations must be independent
- Observations must be normally distributed

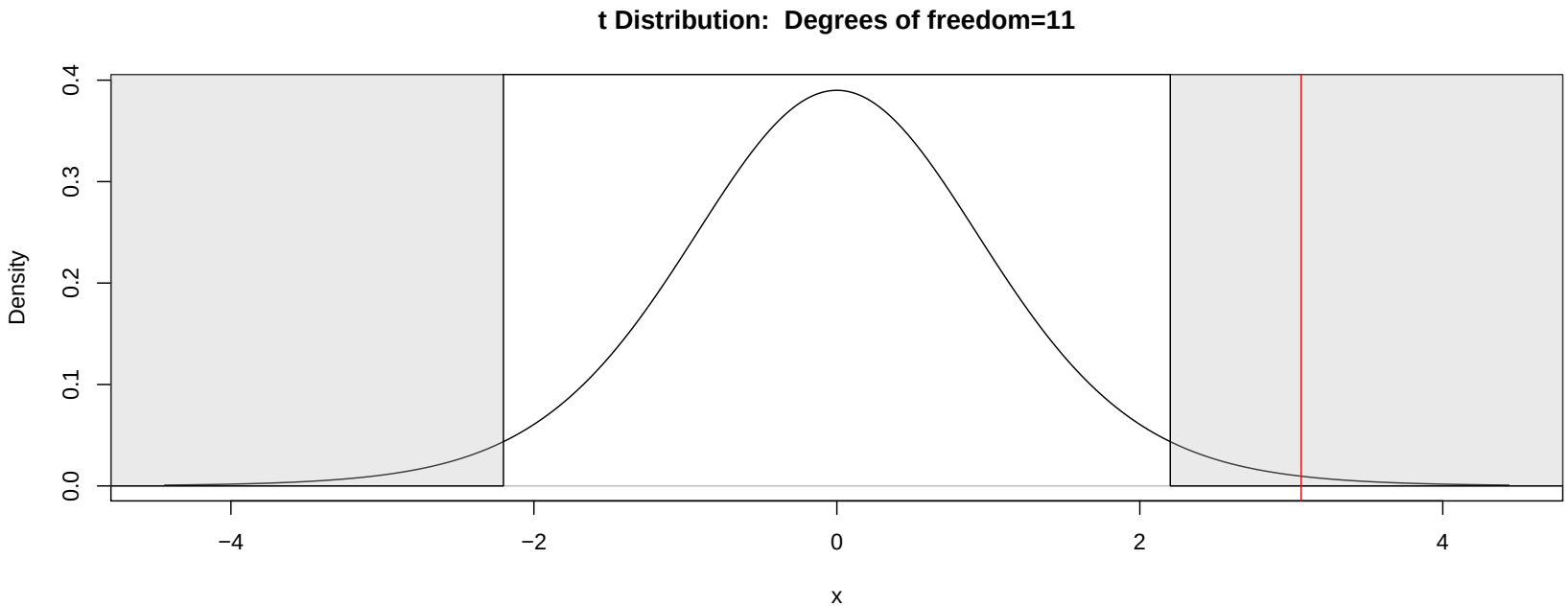


Results

Test statistic $t_{n-1} = t_{11} = \frac{\bar{x}-\mu_0}{s.d./\sqrt{n}} = \frac{2.84-2.10}{s.e.(\bar{x})} = 3.065$

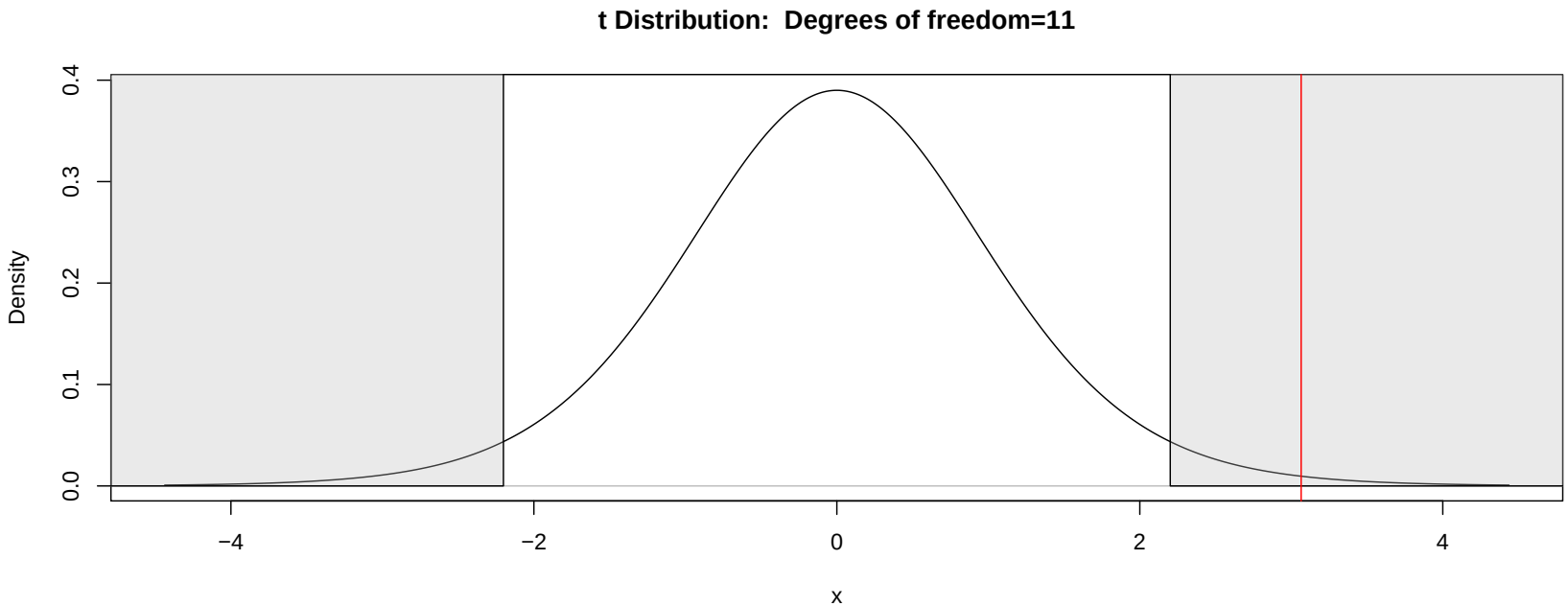
Results

Test statistic $t_{n-1} = t_{11} = \frac{\bar{x}-\mu_0}{s.d./\sqrt{n}} = \frac{2.84-2.10}{s.e.(\bar{x})} = 3.065$



Results

Test statistic $t_{n-1} = t_{11} = \frac{\bar{x}-\mu_0}{s.d./\sqrt{n}} = \frac{2.84-2.10}{s.e.(\bar{x})} = 3.065$



p:value = 0.011 Reject H_0 and conclude that mean monthly failure rate in Genomics is not 2.1%

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

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

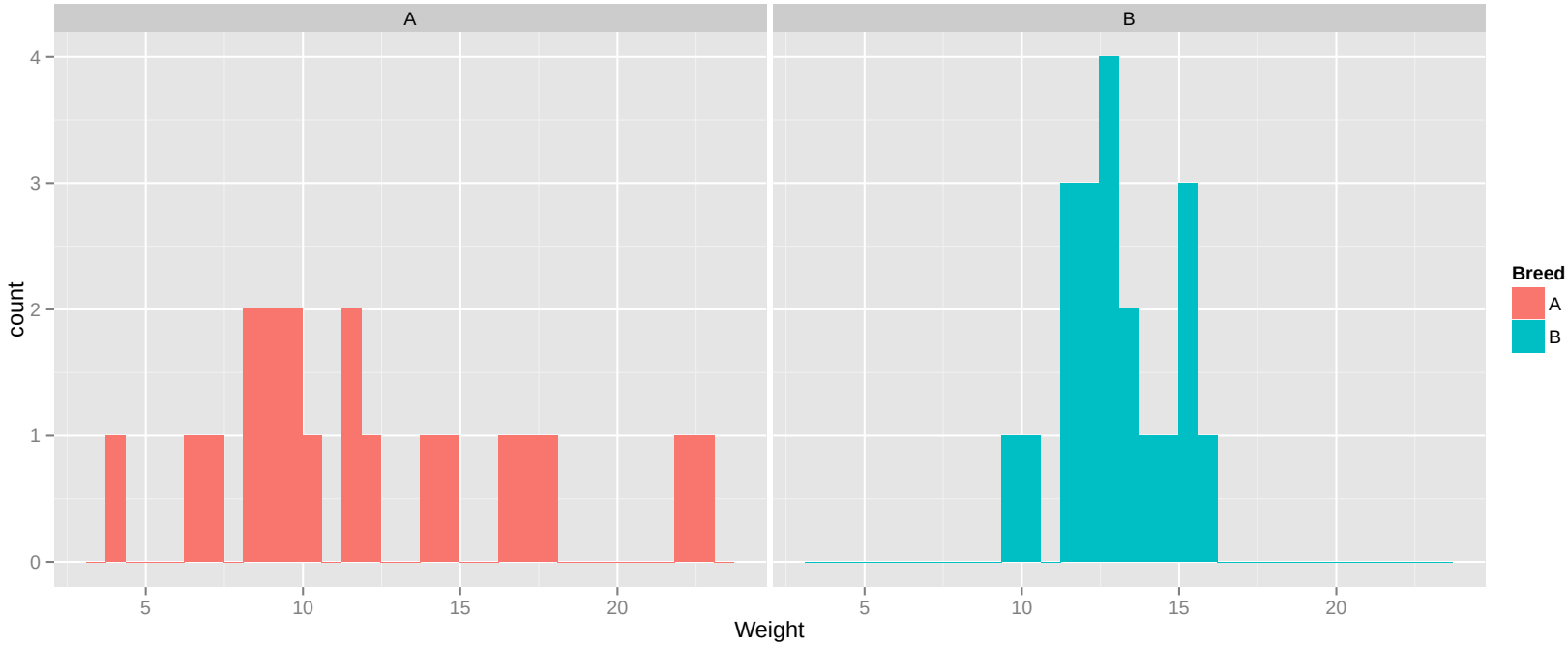
- Null hypothesis, H_0 Mean weight of breed A = Mean weight of breed B
- Alternative hypothesis, H_1 Mean weight of breed A $\neq$ Mean weight of breed B
- Tails: two-tailed
- Either *reject* or *do not reject* the null hypothesis - **never accept the alternative hypothesis**

The data

	BreedAMouse	WeightA	BreedBMouse	WeightB
1	1	22.77	21	15.51
2	2	9.08	22	12.93
3	3	9.80	23	11.50
4	4	8.13	24	16.07
5	5	16.54	25	15.51
6	6	11.36	26	15.16
7	7	11.47	27	11.25
8	8	22.25	28	13.65
9	9	14.04	29	14.28
10	10	17.12	30	13.21
11	11	6.32	31	10.28
12	12	17.51	32	12.41
13	13	9.87	33	9.63
14	14	12.41	34	14.75
15	15	7.39	35	12.56
16	16	9.23	36	13.02
17	17	4.06	37	12.33

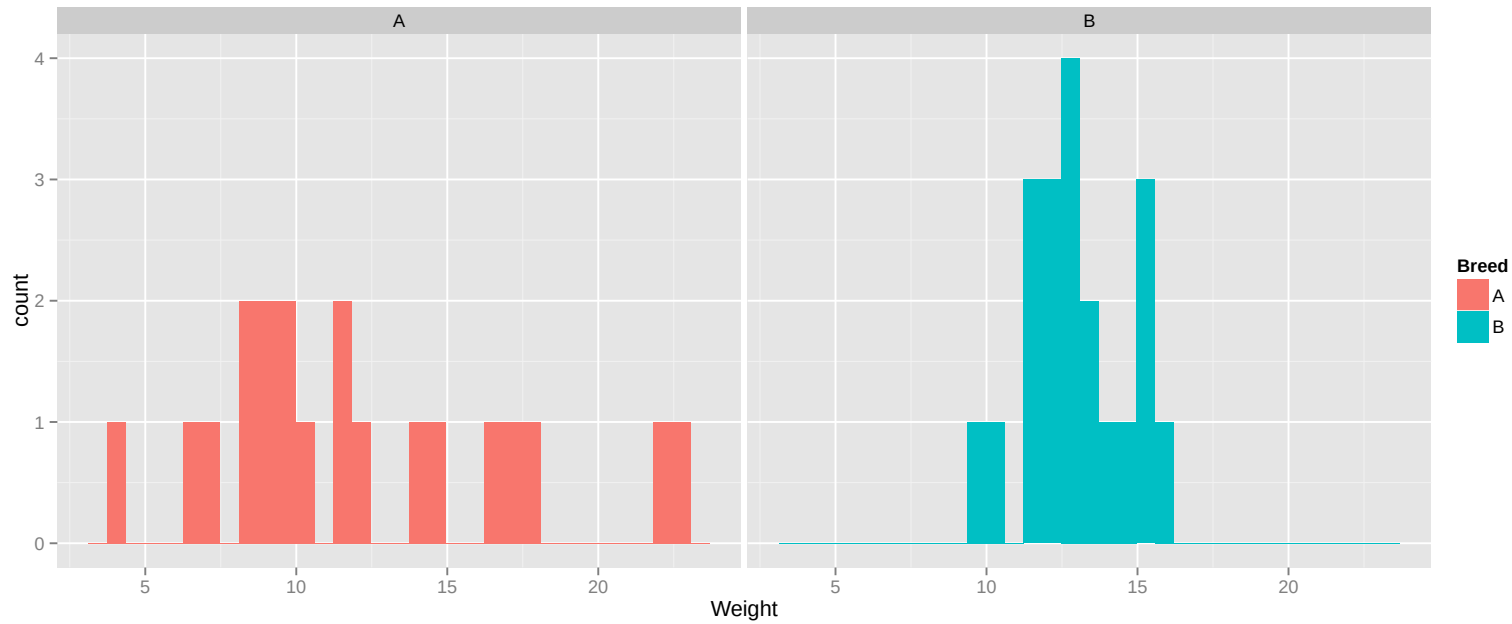
Data summary

Mean of breed A: 12.12 **Mean** of breed B: 12.99
Standard Deviation of breed A: 5.05
Standard Deviation of breed B: 1.8



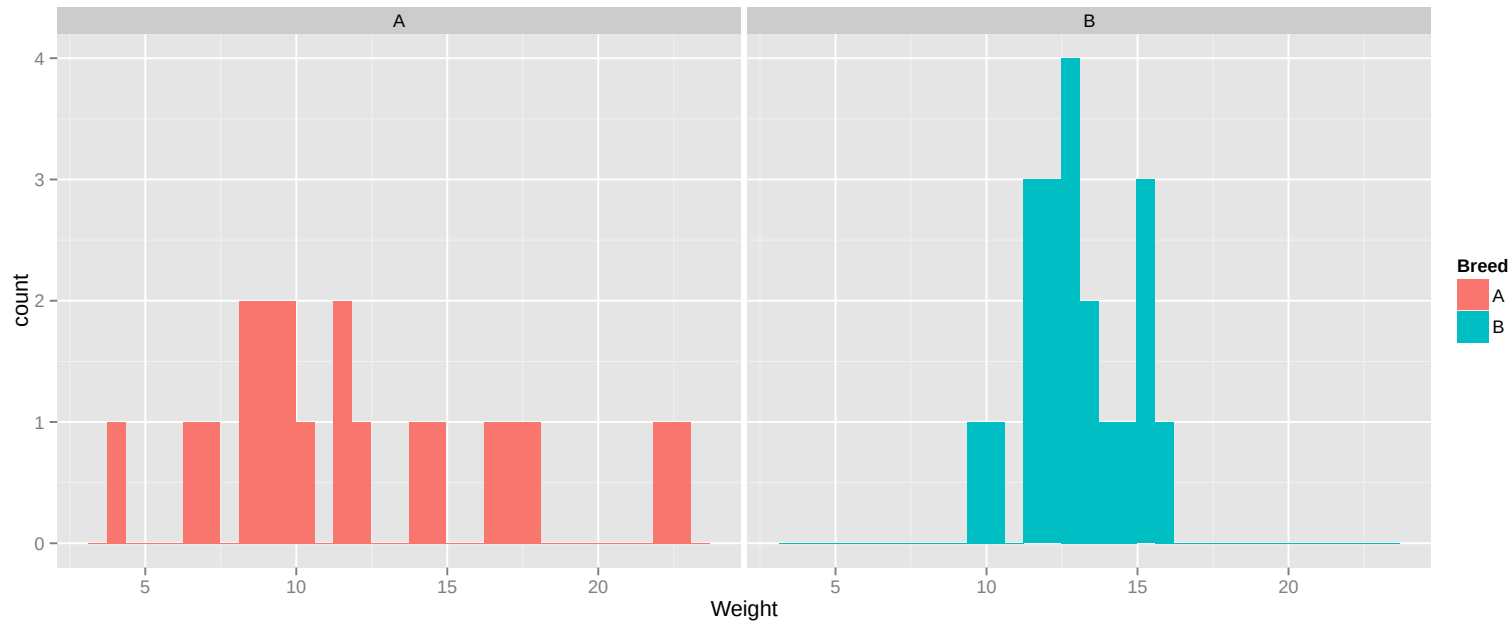
Checking assumptions

- Observations are independent
- Observations are normally distributed



Checking assumptions

- Observations are independent
- Observations are normally distributed

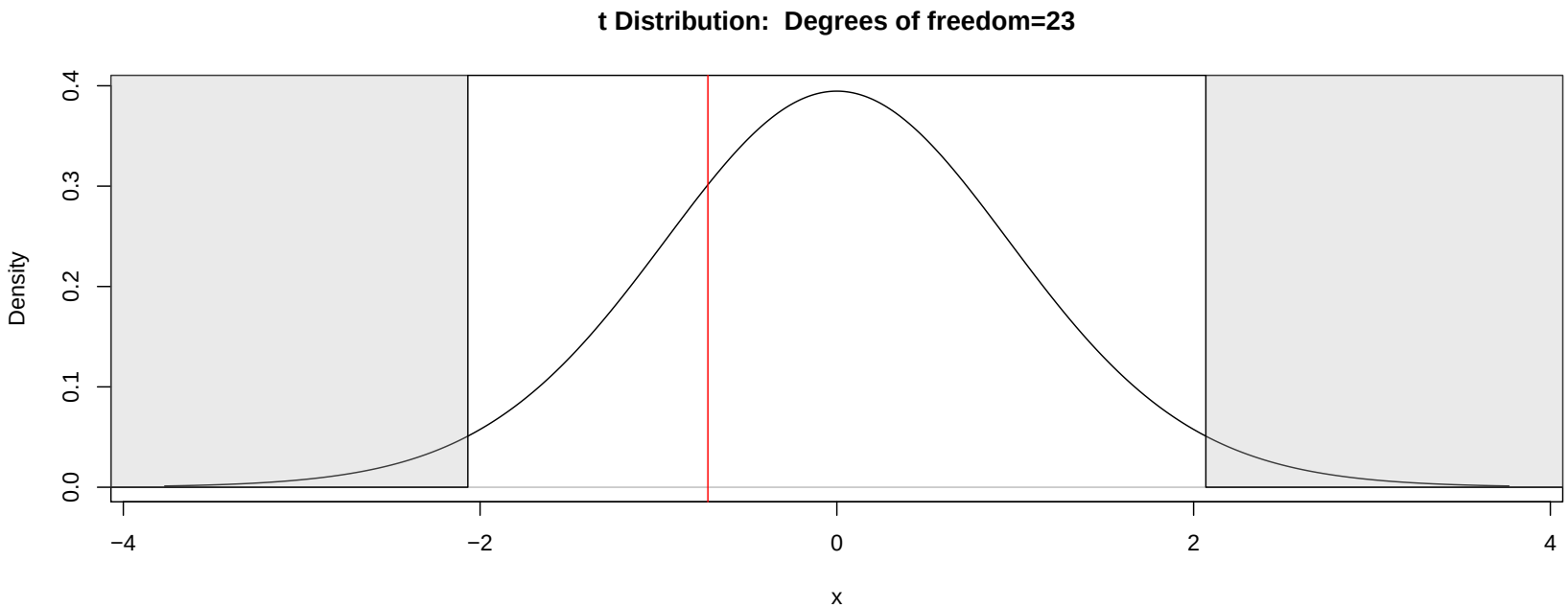


- Use Welch's correction if variances are different (alters the t-statistic and degrees of freedom)

Test results

$$t_{df} = \frac{\bar{X}_A - \bar{X}_B}{s.e.(\bar{X}_A - \bar{X}_B)} = -0.7228852$$

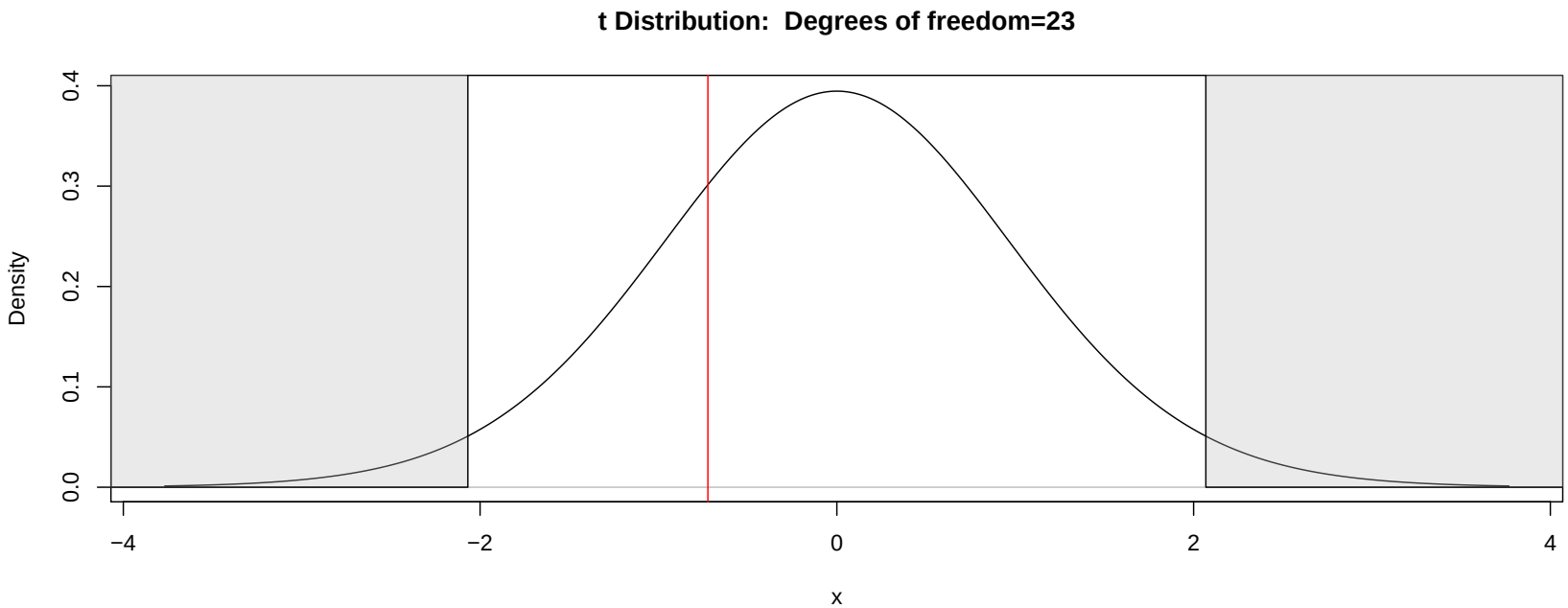
df $= 23$ (with Welch's correction)



Test results

$$t_{df} = \frac{\bar{X}_A - \bar{X}_B}{s.e.(\bar{X}_A - \bar{X}_B)} = -0.7228852$$

df = 23 (with Welch's correction)

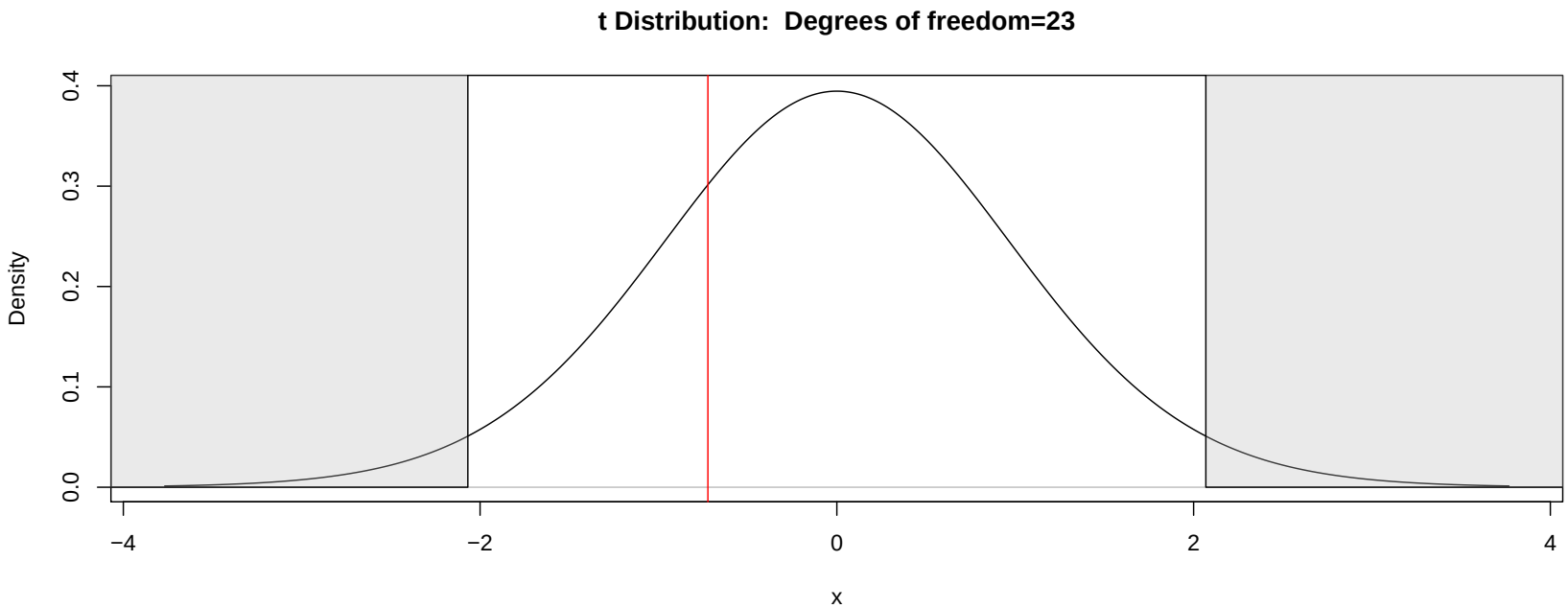


p-value: 0.4768151

Test results

$$t_{df} = \frac{\bar{X}_A - \bar{X}_B}{s.e.(\bar{X}_A - \bar{X}_B)} = -0.7228852$$

df = 23 (with Welch's correction)



p-value: 0.4768151 **Do not reject H_0. There is no evidence for a difference in weight between breeds**

Paired two-sample t-test

Does the mean difference = 0?

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 the two different sites of disease

- Null hypothesis, H_0 Cellularity at site A = Cellularity at site B
- Alternative hypothesis, H_1 Cellularity at site A $\neq$ Cellularity at site B
- Tails: two-tailed
- Either *reject* or *do not reject* the null hypothesis - **never accept the alternative hypothesis**

Null hypothesis

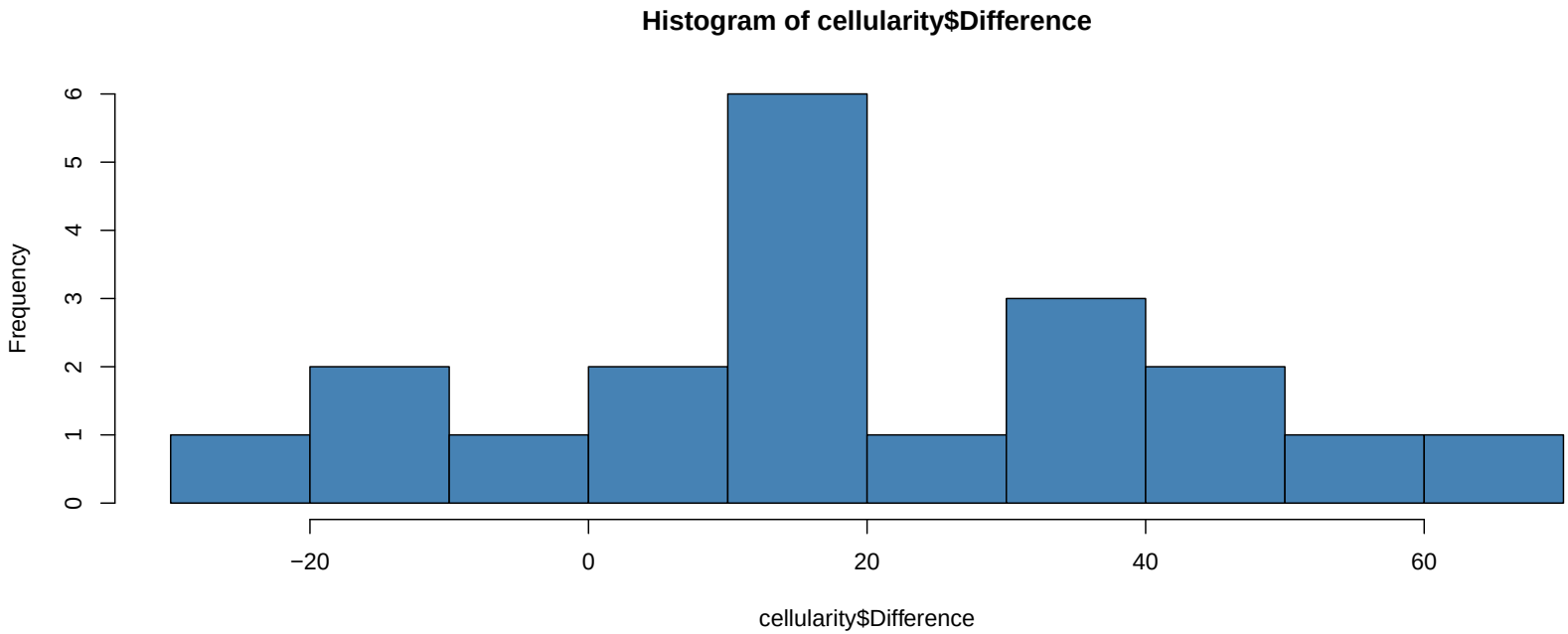
- H_0 Cellularity at site A = Cellularity at site B
  OR
- H_0 Cellularity at site A - Cellularity at site B = 0

The data

	Ovarian	Peritoneal	Difference
1	1201.33	1155.98	45.35
2	1029.64	1020.82	8.82
3	895.57	881.21	14.36
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	-13.81
15	942.38	925.75	16.63
16	884.33	891.34	-7.01
17	930.09	892.02	38.07

Mean difference: 19.139

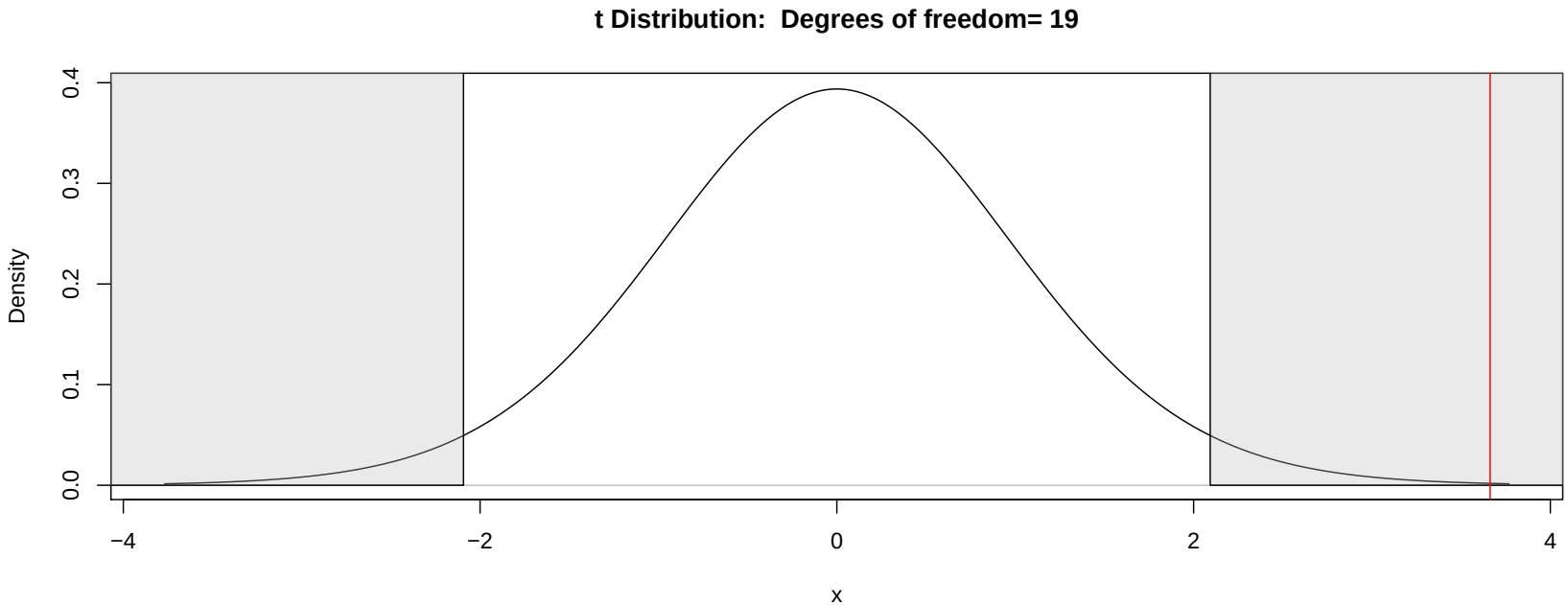
- Observations are independent
- The paired differences are normally distributed



Test results

$$t_{n-1} = t_{19} = \frac{\bar{X}_{A-B}}{s.e.(\bar{X}_{A-B})} = 3.6624$$

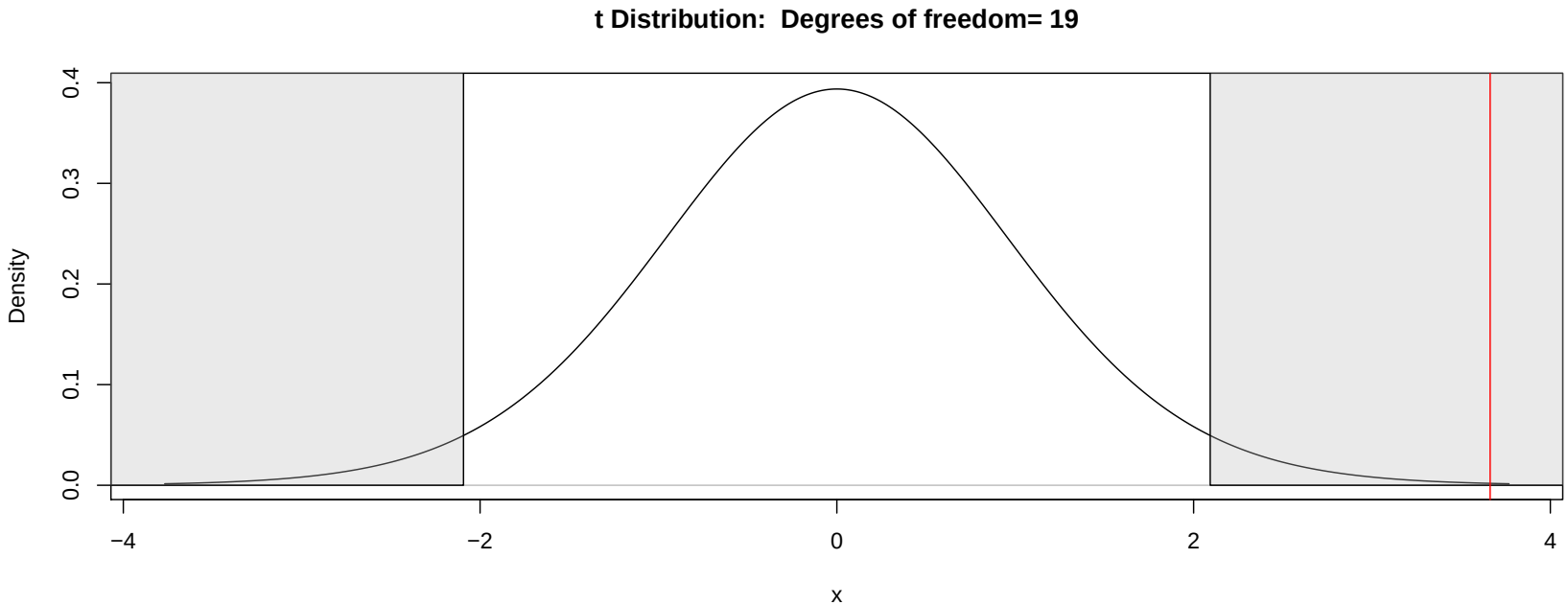
$$\text{df} = 19$$



Test results

$$t_{n-1} = t_{19} = \frac{\bar{X}_{A-B}}{s.e.(\bar{X}_{A-B})} = 3.6624$$

$$\text{df} = 19$$

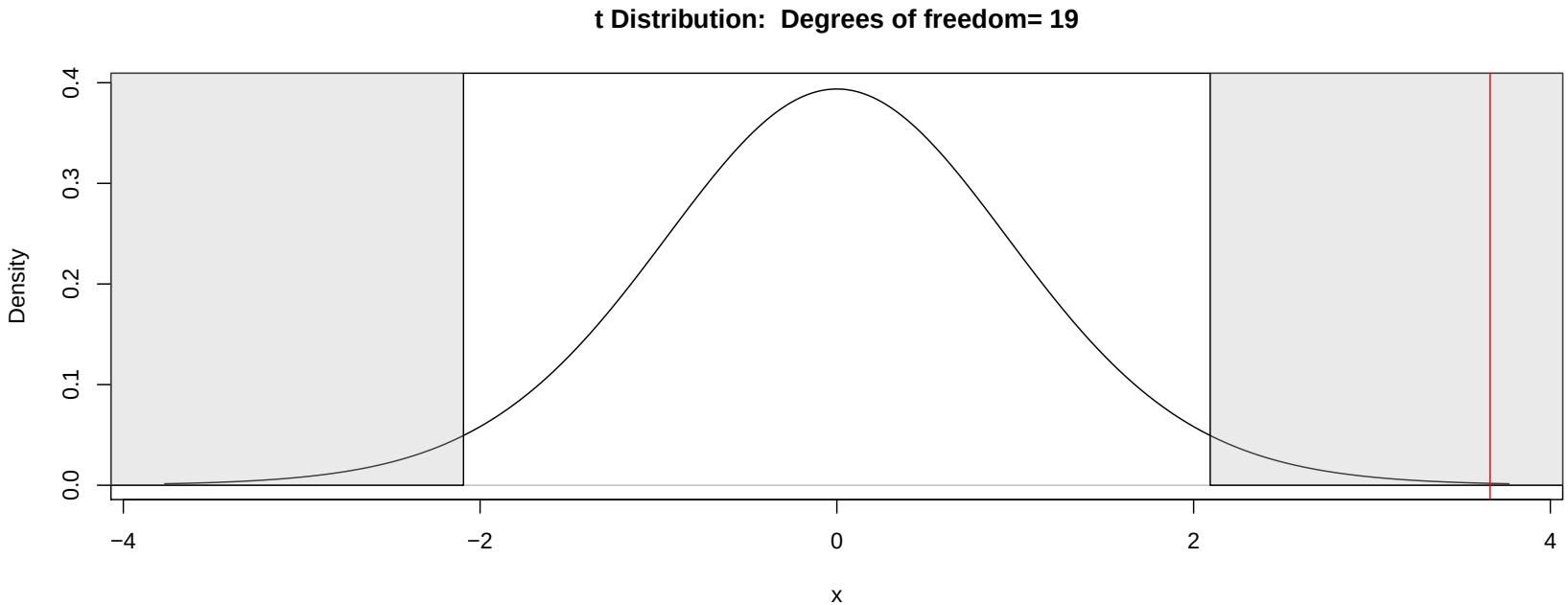


p-value: 0.0016558

Test results

$$t_{n-1} = t_{19} = \frac{\bar{X}_{A-B}}{s.e.(\bar{X}_{A-B})} = 3.6624$$

$$\text{df} = 19$$



p-value: 0.0016558 **Reject H_0.** There is evidence that cellularity at site A $\neq$ cellularity at site B

What if normality is not reasonable?

- Transform your data. e.g. log transformation
- Non parametric tests:
  - One-sample test $\rightarrow$ One-sample Wilcoxon signed rank test
  - Independent two-sample t-test $\rightarrow$ Mann-Whitney U test / Wilcoxon rank sum test
  - Paired two-sample t-test $\rightarrow$ Matched-pairs Wilcoxon signed rank 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*