Statistical inference (continued) Hypothesis testing.

Statement for Audio and Video Learning Resources

Video and audio content at the University uses closed captions generated by automatic speech recognition (ASR). The ASR process is based on machine learning algorithms which automatically transcribe voice to text. According to our technology providers, this process is approximately 70-90% accurate depending on the quality of the audio, and consequently video and audio closed captions may include some transcription errors. It is therefore important to recognise that the original recording is the most accurate reflection of the content, and not the captions.

If you require accurate captions as part of your reasonable adjustments, please contact the Inclusion Centre to discuss your requirements.

Content

- [Saw confidence intervals in previous lecture]
- Hypothesis testing
    - Definitions
    - T-tests
    - Anova
- Sample sizes

Hypothesis testing

- **Hypothesis test:** uses data to test a hypothesis (claim) which is believed to be true.
- **NULL hypothesis:** the default hypothesis in the test.
    - Use data to test whether the NULL hypothesis is true.
- **Alternative hypothesis:** an alternative claim to the null hypothesis

Types of errors

- NULL hypothesis: the hypothesis which stands unless there is lots of evidence for a different hypothesis.
- Types of errors
  - **Type 1:** reject the null hypothesis when it is true
  - **Type 2:** fail to reject null hypothesis when it is false (i.e. alternative hypothesis is true)
  - If we try to reduce type 1 errors, the number of type 2 errors will normally increase
  - If we try to reduce type 2 errors, the number of type 1 errors will normally increase
- **T-score:** how many standard errors the observed mean is from the null value.

  - $T\text{-}score = \frac{(\bar{x} - nullValue)}{SE}$

    where *nullValue* is the value from the null hypothesis

Example 1

- For example dataset with following 21 observations regarding coin tossing 10 times
  - 5 5 5 2 5 5 3 4 4 4 5 4 8 6 6 3 5 4 3 6 6
  - Null hypothesis – coin is NOT biased so null value is 5 heads
  - $\bar{x} = \frac{5+5+5+2+5+5+3+4+4+4+5+4+8+6+6+3+5+4+3+6+6}{21} = 4.666667$
  - $s = \sqrt{\frac{\sum(\mathrm{x} - \bar{x})^{2}}{n-1}} = 1.354006$
  - $SE = \frac{s}{\sqrt{n}} = 0.2954684$
- $T\text{-}score = \frac{(\bar{x} - nullValue)}{SE} = \frac{(4.666667 - 5)}{0.2954684} = -1.1282$
  - *Observed mean is -1.1282 standard errors from null value*

Example 2

- For example assume dataset with 101 observations regarding coin tossing 10 times with
  - $\bar{x} = 4.356436$
  - $s = 1.269521$
- Null hypothesis is that the coin is not biased so mean is 5 heads.
  - $SE = \frac{1.269521}{\sqrt{101}} = 0.126322$
  - $T\text{-score} = \frac{(\bar{x} - nullValue)}{SE} = \frac{(4.356436 - 5)}{0.126322} = -5.0946$
  - *The observed value is -5.0946 standard errors from the null value (5)*

Hypotheses

- H_0: NULL (default) hypothesis uses equality to a value x = nullValue
- H_1: Alternative hypothesis - the one “difficult to believe and accept”
  - **One sided** x > nullValue **or** x < nullValue
  - **Two sided** x ≠ nullValue
- **Null hypothesis** – sceptical position, i.e. only rejected if evidence for alternative hypothesis is very strong.
  - It is not a hypothesis that is accepted
    - If there is enough evidence in favour of the alternative hypothesis, it is rejected.
    - Otherwise, it is not rejected.

… hypothesis test

- Assume NULL hypothesis cannot be rejected.
- Check answers data and compare to what would be expected if NULL hypothesis is true.
  - If the answer is likely to happen, the NULL hypothesis cannot be rejected
  - Only if there is very strong evidence for the alternative hypothesis the NULL hypothesis is rejected

Example

- A student survey with 30 data about respondents suggest students travel 3 miles to work.
  - Assume you suspect this is an over-estimate
- Null hypothesis
  - Average distance = 3 miles
- Alternative hypothesis
  - Average distance <3 miles (one sided)

p-value

- **P-Value:** value produced by hypothesis testing
  - Probability of seeing a result at least as extreme as the one in the sample data when the null hypothesis is true
- **High p-value:** null hypothesis cannot be rejected.
- **Low p-value:** null hypothesis likely to be false
  - Alternative hypothesis likely to be true.
  - So reject null hypothesis in favour of alternative hypothesis.
- How low does the p-value need to be to reject the null hypothesis?
  - Depends of significance level (1- confidence level) used, α.
  - If p-value < significance level (α), reject the null hypothesis

… p-value

- Use significance level (α) to compare with p-value
- The smaller the p-value, the weaker the support for the null hypothesis
- At 5% significance p-value < 0.05 used
  - p-value < 0.05
    - The probability of seeing the result given that the null hypothesis is true is less than 5%
    - There is a ***statistically significant*** difference
    - Reject null hypothesis in favour of the alternative hypothesis.
  - But other α values can be used.
  - Common α values are 0.1, **0.05** and 0.01

t-table

DF	One tail	0.1	0.05	0.025	0.01	0.005	0.001	0.00025
	Two tail	0.2	0.10	0.05	0.02	0.01	0.002	0.001
1		3.078	6.314	12.706	31.821	63.656	318.289	636.578
2		1.886	2.920	4.303	6.965	9.925	22.328	31.600
3		1.638	2.353	3.182	4.541	5.841	10.214	12.924
4		1.533	2.132	2.776	3.747	4.604	7.173	8.610
5		1.476	2.015	2.571	3.365	4.032	5.894	6.869
6		1.440	1.943	2.447	3.143	3.707	5.208	5.959
7		1.415	1.895	2.365	2.998	3.499	4.785	5.408
8		1.397	1.860	2.306	2.896	3.355	4.501	5.041
9		1.383	1.833	2.262	2.821	3.250	4.297	4.781
10		1.372	1.812	2.228	2.764	3.169	4.144	4.587
11		1.363	1.796	2.201	2.718	3.106	4.025	4.437
12		1.356	1.782	2.179	2.681	3.055	3.930	4.318
13		1.350	1.771	2.160	2.650	3.012	3.852	4.221
14		1.345	1.761	2.145	2.624	2.977	3.787	4.140
15		1.341	1.753	2.131	2.602	2.947	3.733	4.073
16		1.337	1.746	2.120	2.583	2.921	3.686	4.015
17		1.333	1.740	2.110	2.567	2.898	3.646	3.965
18		1.330	1.734	2.101	2.552	2.878	3.610	3.922
19		1.328	1.729	2.093	2.539	2.861	3.579	3.883
20		1.325	1.725	2.086	2.528	2.845	3.552	3.850
…								

Hypothesis testing vs confidence intervals

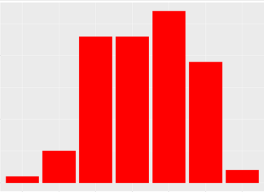
- If NULL hypothesis rejected
  - Value compared against is not in confidence interval.
  - E.g. if Null hypothesis is difference in means is zero, zero is in not interval.
- If NULL hypothesis cannot be rejected
  - Value compared against lies in confidence interval.
- A significance cut off of $x\%$ matches a confidence interval with $(100 - x)\%$ chance of containing true mean.

Example 1 – what is the null hypothesis?

- Peter has been accused of stealing from a shop till. There is data covering a period of 44 days showing the difference between the amount of cash in the till and the amount expected. A court is trying to establish if Peter is innocent or guilty.
    - Formulate the null hypothesis and the alternative hypothesis
- Null hypothesis – Peter is innocent (difference in cash is 0).
- Alternative – Peter is guilty. Only when evidence is very strong this hypothesis is concluded.
- Reverse would be wrong as then guilty would be accepted unless very strong evidence of innocence is found.

Example 2 – hypotheses and info

- RGU health is conducting a survey of their alumni levels of activity. A questionnaire was sent to 10000 alumni. 400 alumni responded. On average alumni took 5750 steps a day.
  - Formulate null and alternative hypothesis for the claim that alumni took 5750 steps a day
  - Can we apply the t-test (or model for a normal distribution) to the sample?
  - Sample distribution is



  - Shapiro-Wilk test shows
    - P-value = 0.11
  - Is there any other information needed?

… example 2

- Hypothesis
  - Null hypothesis: alumni take, on average 5750 steps/day
  - Alternative hypothesis: the average number of steps alumni take in a day is not 5750
- T-model
  - Independence – 400 alumni, <10% of 10000, presumed independent
  - Distribution not entirely normal but sample size is large.
  - Normality test shows a p-value= 0.11 which greater than the normal significance level of 0.05 so it is reasonable to assume that the distribution is normal.
  - Q-Q plot may have been useful for extra evidence regarding distribution.

Example 3

- Assume you have a sample set with 15 observations with zero mean. Estimate proportion of the observations below -2.145.
  - This assumes a mean of zero
- Go to table
- df = 15 – 1 = 14
- Find column with 2.145
- Choose the one-tail corresponding column value, i.e. 0.025
  - So 2.5% of the values
- Why one-tail? We are only interested in whether the value is below.
  - Not interested in value above -2.145

T-table

DF	One tail	0.1	0.05	0.025	0.01	0.005	0.001	0.00025
	Two tail	0.2	0.10	0.05	0.02	0.01	0.002	0.001
1		3.078	6.314	12.706	31.821	63.656	318.289	636.578
2		1.886	2.920	4.303	6.965	9.925	22.328	31.600
3		1.638	2.353	3.182	4.541	5.841	10.214	12.924
4		1.533	2.132	2.776	3.747	4.604	7.173	8.610
5		1.476	2.015	2.571	3.365	4.032	5.894	6.869
6		1.440	1.943	2.447	3.143	3.707	5.208	5.959
7		1.415	1.895	2.365	2.998	3.499	4.785	5.408
8		1.397	1.860	2.306	2.896	3.355	4.501	5.041
9		1.383	1.833	2.262	2.821	3.250	4.297	4.781
10		1.372	1.812	2.228	2.764	3.169	4.144	4.587
11		1.363	1.796	2.201	2.718	3.106	4.025	4.437
12		1.356	1.782	2.179	2.681	3.055	3.930	4.318
13		1.350	1.771	2.160	2.650	3.012	3.852	4.221
14		1.345	1.761	2.145	2.624	2.977	3.787	4.140
15		1.341	1.753	2.131	2.602	2.947	3.733	4.073
16		1.337	1.746	2.120	2.583	2.921	3.686	4.015
17		1.333	1.740	2.110	2.567	2.898	3.646	3.965
18		1.330	1.734	2.101	2.552	2.878	3.610	3.922
19		1.328	1.729	2.093	2.539	2.861	3.579	3.883
20		1.325	1.725	2.086	2.528	2.845	3.552	3.850
…								

Example 4

- Assume you have a sample set with 20 observations. Estimate proportion of the observations is falling above 2.145
- Go to table
- df = 19
- Find columns with values below and above as exact value not in table, i.e. 2.093 and 2.539
  - Corresponding one-tail values , i.e. 0.025 and 0.01
    - So between 1% and 2.5% of the values

T-table

DF	One tail	0.1	0.05	0.025	0.01	0.005	0.001	0.00025
	Two tail	0.2	0.10	0.05	0.02	0.01	0.002	0.001
1		3.078	6.314	12.706	31.821	63.656	318.289	636.578
2		1.886	2.920	4.303	6.965	9.925	22.328	31.600
3		1.638	2.353	3.182	4.541	5.841	10.214	12.924
4		1.533	2.132	2.776	3.747	4.604	7.173	8.610
5		1.476	2.015	2.571	3.365	4.032	5.894	6.869
6		1.440	1.943	2.447	3.143	3.707	5.208	5.959
7		1.415	1.895	2.365	2.998	3.499	4.785	5.408
8		1.397	1.860	2.306	2.896	3.355	4.501	5.041
9		1.383	1.833	2.262	2.821	3.250	4.297	4.781
10		1.372	1.812	2.228	2.764	3.169	4.144	4.587
11		1.363	1.796	2.201	2.718	3.106	4.025	4.437
12		1.356	1.782	2.179	2.681	3.055	3.930	4.318
13		1.350	1.771	2.160	2.650	3.012	3.852	4.221
14		1.345	1.761	2.145	2.624	2.977	3.787	4.140
15		1.341	1.753	2.131	2.602	2.947	3.733	4.073
16		1.337	1.746	2.120	2.583	2.921	3.686	4.015
17		1.333	1.740	2.110	2.567	2.898	3.646	3.965
18		1.330	1.734	2.101	2.552	2.878	3.610	3.922
19		1.328	1.729	2.093	2.539	2.861	3.579	3.883
20		1.325	1.725	2.086	2.528	2.845	3.552	3.850
…								

Example 5

- Assume you have a sample set with 20 observations.
  - Estimate proportion of the observations falling below 2.145
- In examples 4 we calculated between 1% and 2.5% of the values fall above
  - Therefore between 97.5% and 99% fall below

Example 6

- Assume a t-distribution with 10 degrees of freedom.
  - Estimate the proportion of the distribution falling more than 3 units from the mean.
- df = 10
- Look for 3 in table with df= 10. Find 2.764 and 3.169
- This corresponds to 2 tail 0.02 and 0.01, i.e.
- Between 1 and 2%
- Note: 2-tail as interested in both below and above tails.

T-table

DF	One tail	0.1	0.05	0.025	0.01	0.005	0.001	0.00025
	Two tail	0.2	0.10	0.05	0.02	0.01	0.002	0.001
1		3.078	6.314	12.706	31.821	63.656	318.289	636.578
2		1.886	2.920	4.303	6.965	9.925	22.328	31.600
3		1.638	2.353	3.182	4.541	5.841	10.214	12.924
4		1.533	2.132	2.776	3.747	4.604	7.173	8.610
5		1.476	2.015	2.571	3.365	4.032	5.894	6.869
6		1.440	1.943	2.447	3.143	3.707	5.208	5.959
7		1.415	1.895	2.365	2.998	3.499	4.785	5.408
8		1.397	1.860	2.306	2.896	3.355	4.501	5.041
9		1.383	1.833	2.262	2.821	3.250	4.297	4.781
10		1.372	1.812	2.228	2.764	3.169	4.144	4.587
11		1.363	1.796	2.201	2.718	3.106	4.025	4.437
12		1.356	1.782	2.179	2.681	3.055	3.930	4.318
13		1.350	1.771	2.160	2.650	3.012	3.852	4.221
14		1.345	1.761	2.145	2.624	2.977	3.787	4.140
15		1.341	1.753	2.131	2.602	2.947	3.733	4.073
16		1.337	1.746	2.120	2.583	2.921	3.686	4.015
17		1.333	1.740	2.110	2.567	2.898	3.646	3.965
18		1.330	1.734	2.101	2.552	2.878	3.610	3.922
19		1.328	1.729	2.093	2.539	2.861	3.579	3.883
20		1.325	1.725	2.086	2.528	2.845	3.552	3.850
...								

Example 7

- Assume a significance level of 0.05
- If the null hypothesis is true, how often should the p-value < 0.05?
- Around 5% of the time
  - The data has 5% chance of being data which favours the alternative hypothesis

Example 8

- Assume p-value = 0.07 at a significance level of 0.05.
- Can we reject the null hypothesis?
- What if the significance level is 1%?
- And if the significance level is 10%?
- 0.07 > 0.05 so we cannot reject null hypothesis at 5%
- 0.07 > 0.01 so we cannot reject null hypothesis at 1%
- 0.07 < 0.1 so we can reject null hypothesis at 10%

Example 9

- 19 samples of student performance in each of the last 2 years reveals
  - In 2016, the average number of A grades was 3.2
  - In 2017, the average number of A grades was 3.3
- What are the hypotheses if you want to test if students are getting better?
- Null hypothesis
  - Average = 3.2
- Alternative
  - Average > 3.2

Example 10

- Assume that a 19 sample of average distances travelled by students from the School of Computing shows an average number of miles travelled to attend lectures is 4.4 with a sample standard deviation of 2.3, a minimum number of miles travelled of 1.7 and a maximum of 9.2.
  - Calculate the 95% confidence interval
- 19 samples are random and < 10% of population
- $\bar{x} = 4.4, s = 2.3$
- $df = 18$
- $\alpha = 1 - 0.95 = 0.05$
- $t_{df} = 2.1$ *(from table, find value for two-tailed with* $\alpha = 0.05$)
- $SE = \frac{s}{\sqrt{n}} = 0.528$
- $\bar{x} \pm (t_{df} * SE) = 4.4 \pm (2.1 * 0.528)$ → ***(3.29, 5.51)***
- *We are 95% confident that students travel between 3.**29 and 5.51** miles to attend lectures*

Example 11

- Assume that you have a null hypothesis stating that dogs are smarter than cats.
  - Which of the following values makes you most confident that the null hypothesis is not true?
  - p=0.22
  - p=0.46
  - p=0.03
  - p=0.002
- The p-value is the probability of getting a sample with a difference as large as the one in the null hypothesis.
  - The lower the p-value, the less likely the null hypothesis is true.
  - Lower p-values make us more confident of null hypothesis falsehood.
  - The lowest value is p=0.002.

Example 12

- You have a null hypothesis which states that dogs and cats are just as smart.
  - Their means for “smartness” are equal
  - Your significance level is 0.05
- You calculate a p-value of 0.15.
- Can you reject the null hypothesis?
  - Not if we assume a confidence level of 0.05
- Can you accept the alternative hypothesis?
  - No, the p-value is not below the significance level
- Can you fail to reject the null hypothesis?
  - Yes

Types of tests (t-tests)

Types of tests

- Parametric
  - Compare mean values
  - Conditions must be true for the test to be applicable
- Non parametric
  - Compare median values
- Always use parametric tests if the conditions for the test are satisfied.
  - They are more powerful
- In this lecture **we only cover parametric tests**.

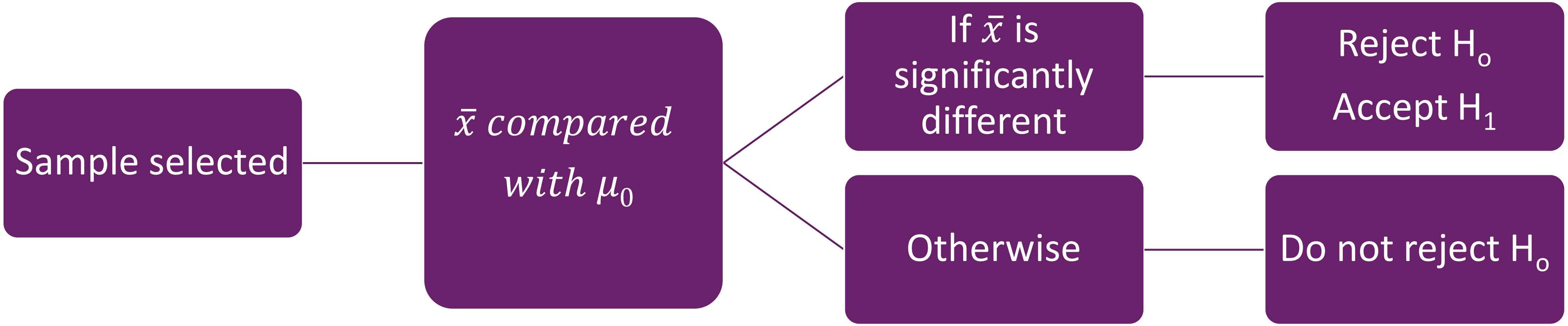
Conditions

- Quantitative data
- Either
  - Sample size > 30
- OR
  - Data normally distributed. Note that ordinal data is not normal.
    - Check data is normal
    - With paired data check for normal distribution of the differences in values
    - With ANOVA check normality of residuals (after test).

Types of test

- **One sample t-test**
  - Test difference between a sample mean and a known or hypothesized value of the mean in the population
- **Paired t-test**
  - Test whether the mean difference between two sets of paired observations is zero.
  - Two related variables tested.
- **Two sample t-test**
  - Test means between 2 population samples
  - Is there a difference between the means?
- ANOVA
  - Tests whether there are differences among group means (more than 2).

Process



Sample selected

$\bar{x}\ compared\ with\ \mu_0$

If $\bar{x}$ is significantly different

Reject H_o

Accept H_1

Otherwise

Do not reject H_o

Tests used for comparison - parametric

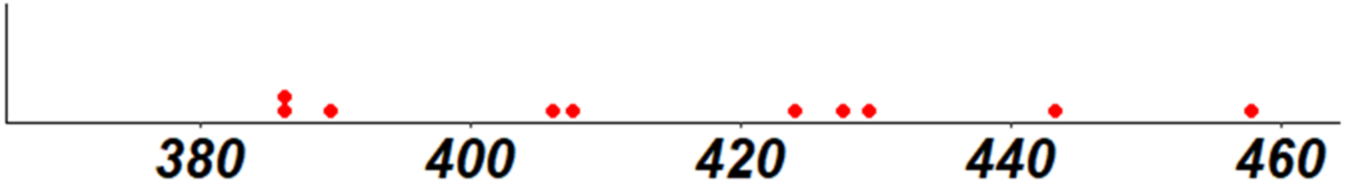
- Parametric – assumes normal data or quantitative with n ≥ 30 in each sample.
    - t-test one mean against specified value
        - $H_0 : \mu = \mu_0$
    - t-test for matched pairs (d is difference in matched pair values)
        - $H_0 : \mu_d = 0$
    - t-test of two means
        - $H_0: \mu_1 = \mu_2$
    - Analysis of variance of k means
        - $H_0: \mu_1 = \mu_2 = \ldots = \mu_k$

Example of one-sample t-test

- A software program is updated with new security features. Ten program runs give the following times:

386.1 386.1 389.6 406.1 407.5 424.0 427.5 429.5 443.3 457.8

- Before the upgrade, the software run in 400 seconds (μ_0) on average. With a 5% significance level, is there evidence of a change in the mean runtime?
- Hypothesis
  - H_0: there is no difference in means (before and after upgrade)
  - H_1: there is a difference in means
- Dot plot of time



Example of one-sample t-test (cont)

- Checking distribution for normality

```
times <- c(386.1, 386.1, 389.6, 406.1 , 407.5 , 424.0, 427.5, 429.5,
443.3, 457.8 )
```

```
shapiro.test(times)
```

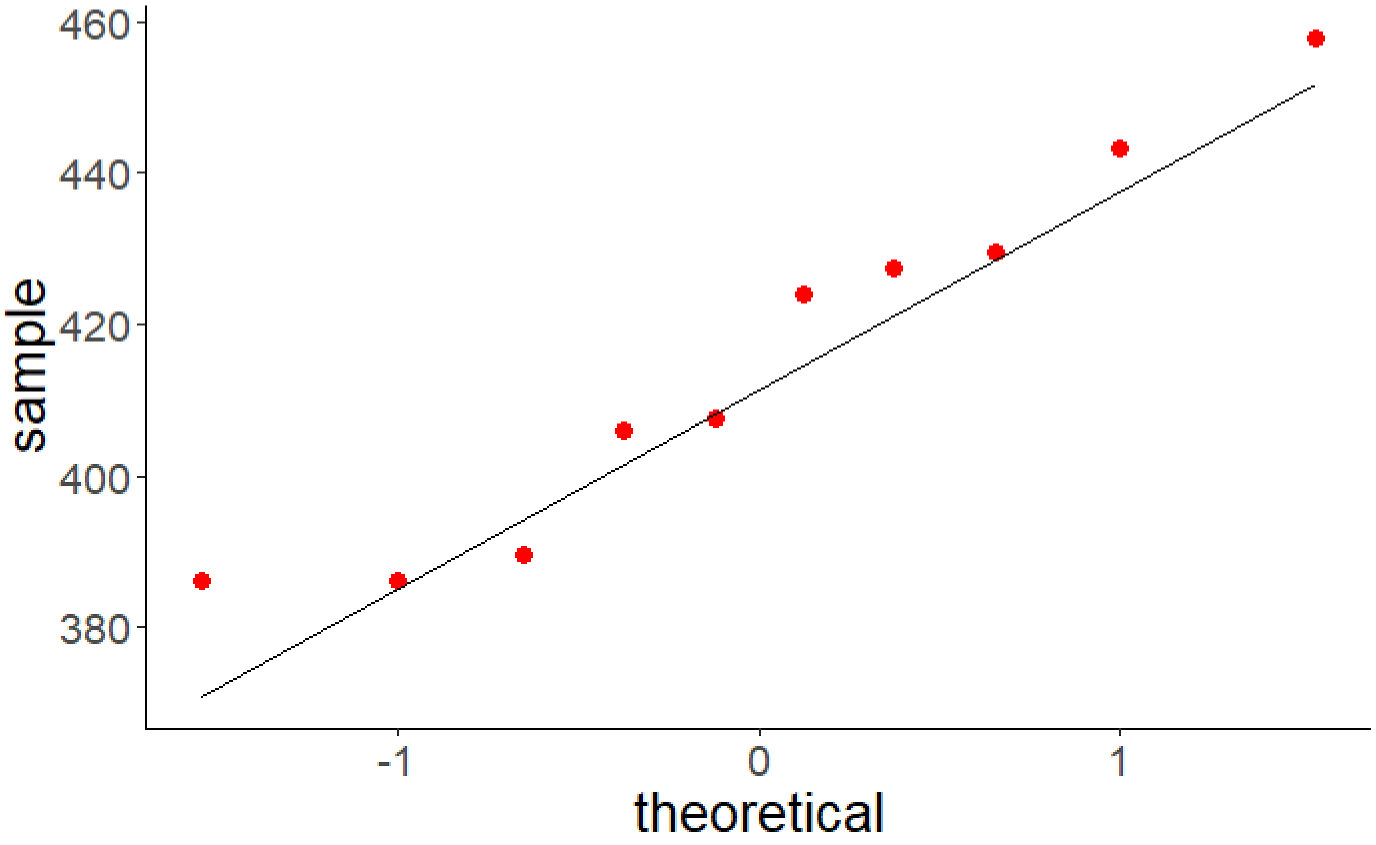
```
     Shapiro-Wilk normality test


data:  c(386.1, 386.1, 389.6, 406.1, 407.5, 424, 427.5,
429.5, 443.3, 457.8)
W = 0.93281, p-value = 0.4761
```

p-value > 0.05 so it is reasonable to assume that the data is normally distributed.

Q-Q Plot – sample against theoretical normal distribution

- Points reasonably close to theoretical line so it is reasonable to assume the distribution is normal.



Defining the test

- Test: one-sample test so only have value for x.
  - x = timedata
  - y = NULL
    - This is true by default so no need to include it.
- Value compared with is 400
  - mu=400
- Test is two sided: the alternative hypothesis refers to the mean not being 400 – so it can be below 400 or above 400 (i.e. just not 400).
  - So alternative = "two.sided "
- Confidence level is 95% (significance level is 0.05)
  - conf.level=0.95

t.test of the time data. $H_0 : \mu = 400$

- timeData: 386.1, 386.1, 389.6, 406.1, 407.5, 424.0, 427.5, 429.5, 443.3, 457.8

$n = 10$ $\bar{X} = 415.75$ $s = 24.78$

```
t.test(x= timeData, alternative = "two.sided", mu= 400, conf.level=0.95)
```

```
  t = 2.0101, df = 9, p-value = 0.07532
  alternative hypothesis: true mean is not equal to 0
  95 percent confidence interval:
        ( 398.0247,  433.4753)
  sample estimates:
        mean of x
        415.75
```

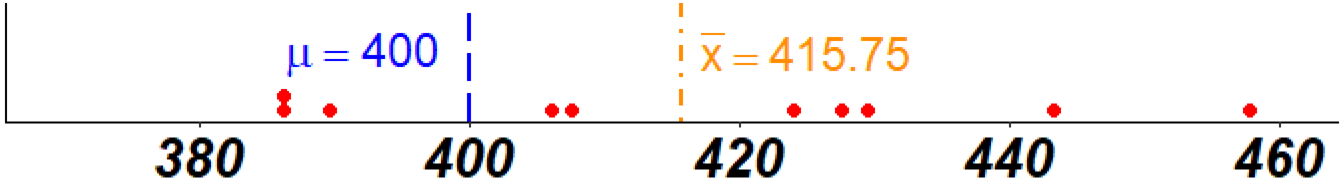
Mean tested against.
$H_0 : \mu = 400$
$H_1 : \mu \neq 400$

P-value > 0.05, null hypothesis is NOT rejected. Not enough evidence of change in time.

Note that t is NOT the t value but the t-statistic = $\frac{\bar{X}-\mu}{s/\sqrt{n}} = = \frac{415.75-400}{24.78/\sqrt{10}} = 2.01$

… t-test

- There is not enough evidence to reject the null hypothesis that $\mu = 400$ with α **= 0.05 (confidence level 95%).**



- Confidence interval (398.0247, 433.4753)

$$\bar{X} = 415.75 \in (398.0247, 433.4753)$$

Test for matched pairs

- Pair of observations for each experimental unit.
  - E.g.
    - Measure BEFORE event
    - Measure AFTER event
    - Has event affected value of measure?

- **A single sample test carried out on the set of differences.**
  - **t-test for matched pairs**
  - **If conditions for t-test are valid.**

Example – paired t-test

- 2 search engines are used to identify relevant documents for 12 different searches.
- For each search engine, the number of relevant documents found within the top 50 documents returned is noted.
- Call this dataset engines.

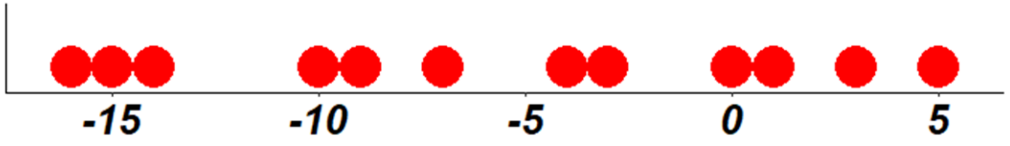
Task	Search engine A	Search engine B	Difference
1	13	20	-7
2	30	25	5
3	36	40	-4
4	29	45	-16
5	28	28	0
6	17	16	1
7	23	26	-3
8	21	18	3
9	27	41	-14
10	20	35	-15
11	40	49	-9
12	28	38	-10

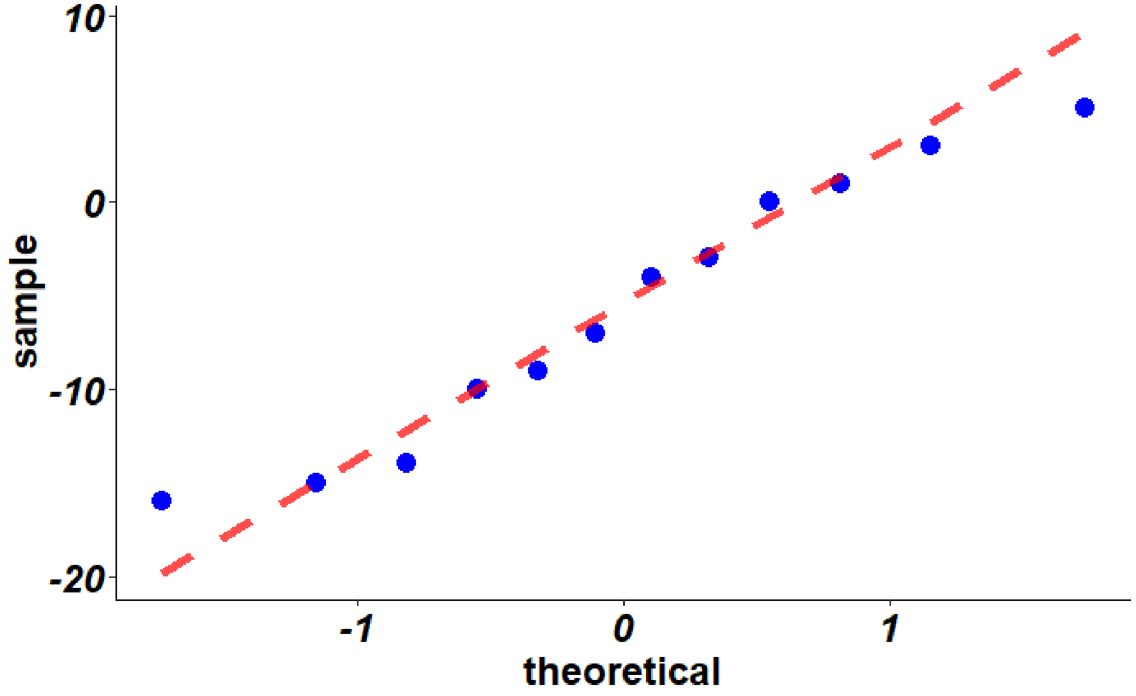
Testing for normality

- `shapiro.test(engines$Difference)`

```
     Shapiro-Wilk normality test

data:  engines$Difference
W = 0.9467, p-value = 0.5894
```





- Checking differences.
  - P-value > 0.05
  - Points close to line in Q-Q plot
  - Paired t-test applicable

Defining the hypotheses and test

- $H_{0:}$ means are same.
- H_1: means are different.
- Test: matched pairs as there are pairs of observations (one value per search engine).
  - paired = T
  - Value tested against: mean difference is zero.
    - mu = 0
- Test is two sided: the alternative hypothesis refers to the mean difference being below zero or above zero (i.e. just not zero)
  - So alternative = "two.sided“
- Confidence level is 95% (significance level is 0.05)
  - conf.level=0.95

Paired t-test. $H_{0:}$ means are same

```
t.test(x=engines$Search.engine.A, y=engines$Search.engine.B,
alternative = "two.sided", paired=T, mu= 0, conf.level=0.95)
          Paired t-test
data:  engines$Search.engine.A and engines$Search.engine.B
t = -2.7664, df = 11, p-value = 0.01834
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 -10.324702  -1.175298
sample estimates:
mean of the differences
                  -5.75
```

Test statistic

P-value <0.05, **reject NULL hypothesis H_o**

At 95% confidence level the mean number of relevant documents retrieved by Engine B is between 1.18 and 10.32 higher than the Engine A mean. Interval does not contain Zero.

Test for 2 independent samples

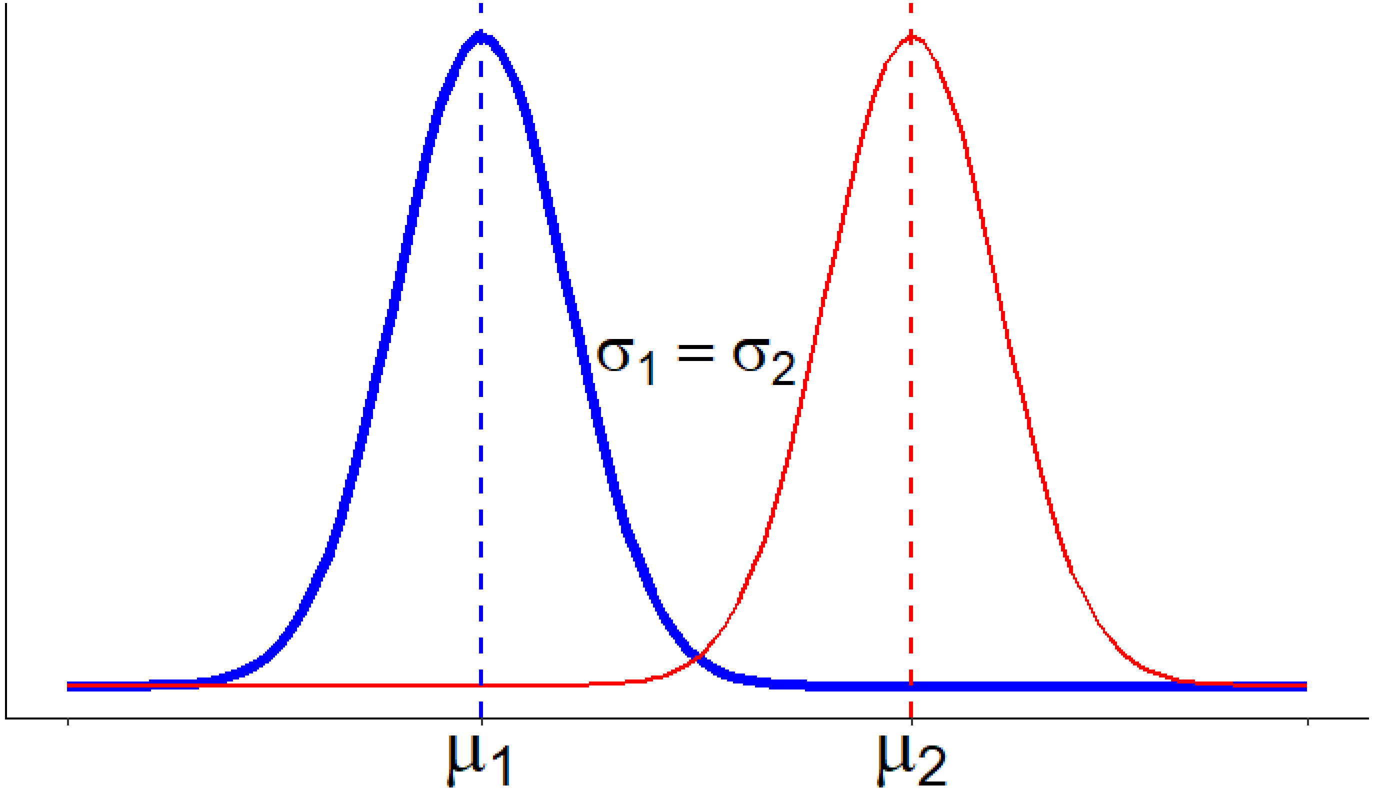
- Two populations
- One sample per population, with sample means calculated.
- Can difference between population means be inferred?

2-sample t-test: comparison of 2 population means

- Two versions:
  - **Pooled 2-sample t-test** assumes equal variances (so equal standard deviations)
  - **Fisher-Behrens 2-sample t-test** (less powerful) does not require equal variances.
- Test of equal variances needed to select correct version of the 2-sample t-test.
  - A test of equal variances can provide useful information
    - Is there more variation in the ratings of program A than the ratings of Program B?.
  - But **normally equal variances are assumed** (so pooled 2-sample t-test conducted).

Pooled 2-sample t-test

- 2 populations
- Same standard deviation assumed
  - Is this a reasonable assumption?
  - If it is not, do not use the pooled 2-sample test.

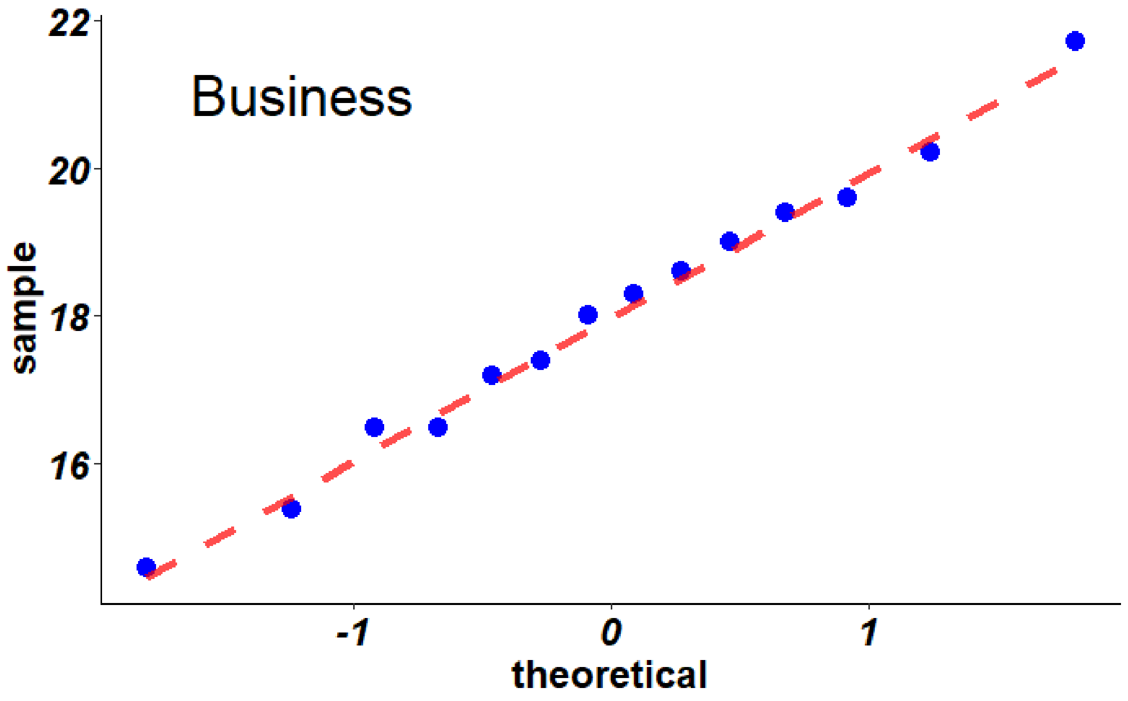


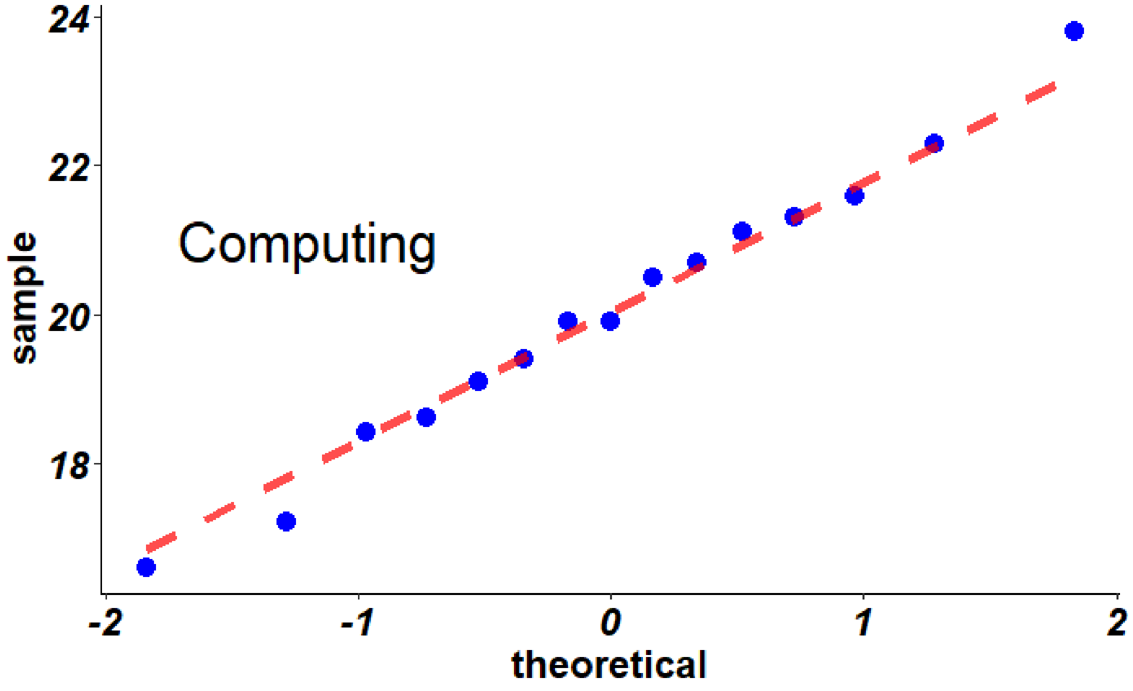
Comparing starting salaries

- Job 1: business graduate
- Job 2: computing graduate
- Do computing graduates and business graduates have the same starting salaries?
- Data measured in 1000£/year
- Use a significance level of 1%, i.e.
  - $\alpha = 0.01$

Job 1	Job 2
17.2	17.2
18.3	19.9
17.4	21.6
19.6	21.1
19	19.1
14.6	18.4
18.6	22.3
16.5	19.9
15.4	21.3
20.2	18.6
18	19.4
16.5	16.6
21.7	20.7
19.4	23.8
	20.5

Checking for normality





Points are very close to straight line

Normality tests

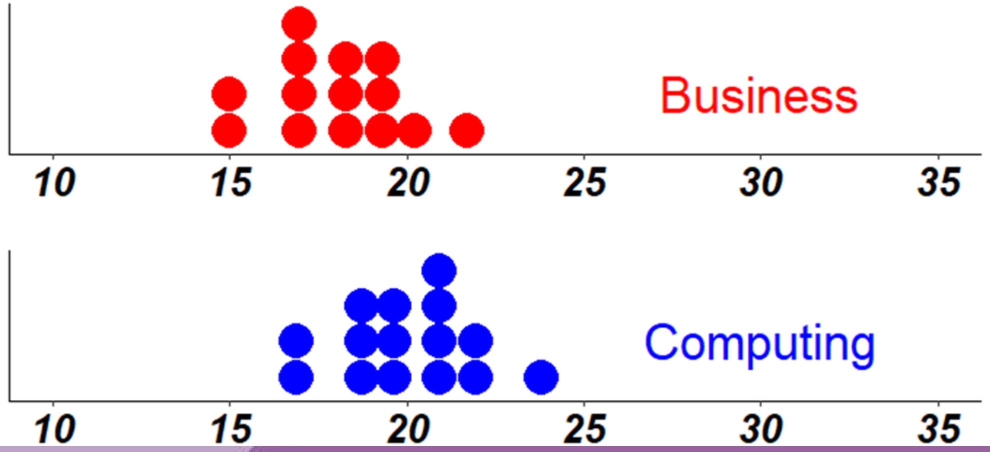
- Computing graduates

```
    Shapiro-Wilk normality test
data:  jobs$Job.1
W = 0.99231, p-value = 0.9999
```

- Business graduates

```
    Shapiro-Wilk normality test
data:  jobs$Job.2
W = 0.99095, p-value = 0.9997
```

- All normality checks are passed.



Hypotheses and t-test definition.

- H_0: means are same
- H_1: means are different
- Not matched pairs as the observations are independent
  - paired = F
- Value tested against: mean difference is zero.
  - mu = 0
- Pooled 2-sample t-test assumes equal variance so
  - var.equal= T
- Test is two sided: the alternative hypothesis refers to the mean difference being below zero or above zero (i.e. just not zero)
  - So alternative = "two.sided"

Pooled t-test. H_0: means are same

-
  ```
  t.test(x=jobs$Job.1, y=jobs$Job.2 ,
   alternative = "two.sided", paired=F,
   var.equal= T, mu= 0, conf.level=0.99)
  ```

Two Sample t-test

```
data:  jobs$Job.1 and jobs$Job.2
t = -2.8032, df = 27, p-value = 0.009254
alternative hypothesis: true difference in means is not
equal to 0
99 percent confidence interval:
 -3.97302566 -0.02316481
sample estimates:
mean of x mean of y
 18.02857  20.02667
```

Test statistic t = $\dfrac{\overline{x_1} - \overline{x_2}}{s(x_1 - x_2)}$

P-value < 0.01 so reject NULL hypothesis H_0

Evidence that Computing graduates have a higher salary

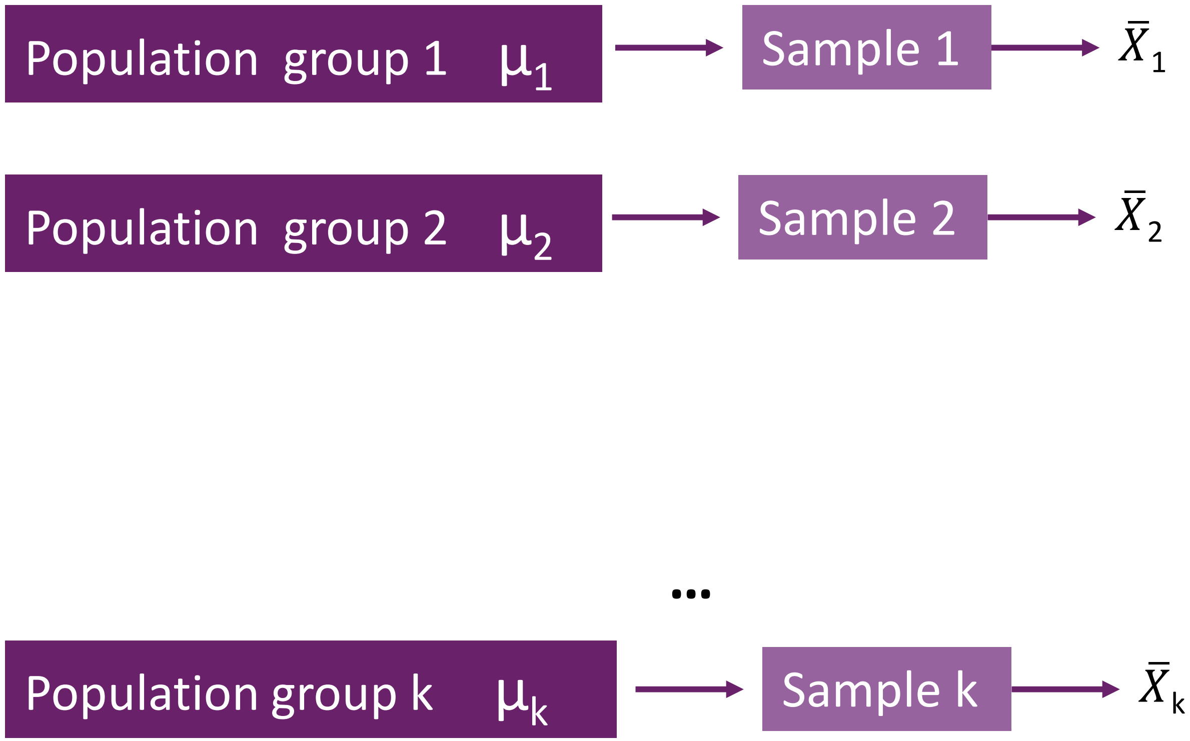
Alternative hypothesis

- In the previous examples, the alternative hypothesis has been two sided.
  - alternative = "two.sided"
- If the alternative hypothesis is one sided, there are two options
  - The mean of x is greater than the mean of y (or the stated value)
    - alternative = "greater"
  - The mean of x is less than the mean of y
    - alternative = "less"

Types of tests - ANOVA

ANOVA – Tests for k independent samples

- ANOVA – one-way ANalysis Of VAriance



Population group 1 μ_1 → Sample 1 → $\bar{X}_1$

Population group 2 μ_2 → Sample 2 → $\bar{X}_2$

…

Population group k μ_k → Sample k → $\bar{X}_k$

k populations

Given the sample means, is there evidence of **differences in population means between groups**?

ANOVA assumptions

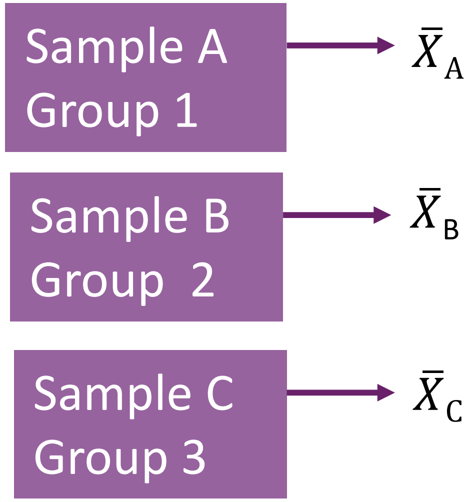
- The observations are obtained independently and randomly from the population defined by the groups.
- The data of each group are normally distributed.
- The samples have the same variance.

ANOVA

- Parametric test for comparing independent samples
- Extension of t-test for more than 2 independent groups
- Aim: distinguish between the
  - **Random variation:** due to experimental error
    - It cannot be controlled.
  - **Systematic variation:** due to identifiable, contributory factors
    - E.g. different users, algorithms, hardware.
    - The sources can be compared using F-tests: check whether the variation is significantly higher than the random variation (see below).

ANOVA (cont)

- Assume 3 group values with one random sample of each.



Sample A Group 1 → $\bar{X}_A$

Sample B Group 2 → $\bar{X}_B$

Sample C Group 3 → $\bar{X}_C$

- E.g.
  - Groups are: Computing students, Business students and Art students.
  - Measure is time spent working on assignments
  - Is there a difference in mean time according to student group?

… ANOVA

- Between group variation – systematic variation due to group
  - E.g. variation between Computing, Business and Art groups
- Within group variation – random variation
  - E.g. variation between Computing students.
- If the means for the groups are significantly different
  - Between group variation significantly greater than within group variation
- If the means for the groups are not significantly different
  - Between group variation not significantly greater than within group variation

Procedure

- Hypotheses definition.
  - H_0: null hypothesis
    - All means are the same $\mu_1 = \mu_2 = \mu_3$
  - H_1: alternative hypothesis
    - At least 2 means are different.

- Apply test.
- Check residual distribution.
  - The procedure is only valid if the residual errors are normally distributed.

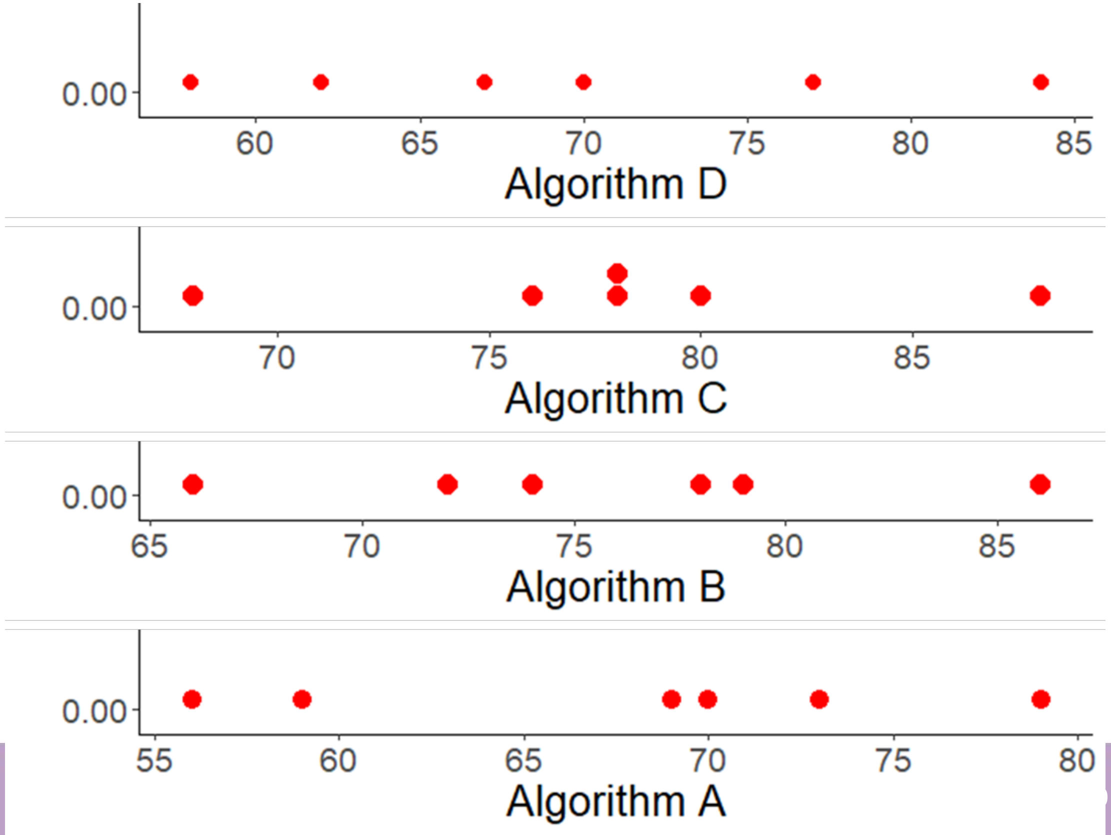
Example - ANOVA

- Sample of 24 databases selected from a large family of databases.
  - 6 databases are allocated, at random, to each of 4 search algorithms (A, B, C and D).
  - Search times are measured.
  - Is there evidence of any differences between the mean search times of the 4 algorithms? Use α = 0.05

Algorithm A	Algorithm B	Algorithm C	Algorithm D
56	86	80	62
69	66	78	58
59	79	68	67
73	74	76	84
70	78	88	77
79	72	78	70

Defining the hypotheses

- H_0: $\mu_A = \mu_B = \mu_C = \mu_D$
- H_1: at least two means are different



algorithm	time
Algorithm A	56
Algorithm A	69
Algorithm A	59
Algorithm A	73
Algorithm A	70
Algorithm A	79
Algorithm B	86
Algorithm B	66
Algorithm B	79
Algorithm B	74
Algorithm B	78
Algorithm B	72
Algorithm C	80
Algorithm C	78
…	…
Algorithm D	70

One-way ANOVA – function aov

- Is there a difference in mean time value according to group (algorithm)?

```
anova <- aov( time ~ algorithm, data = dataAlgo)
summary(anova)
            Df Sum Sq Mean Sq F value Pr(>F)
algorithm    3  434.5  144.82   2.267  0.112
Residuals   20 1277.5   63.87
```

p-value > 0.05 so the NULL hypothesis cannot be rejected.

- There is no evidence of differences between the population mean search times of the algorithms.

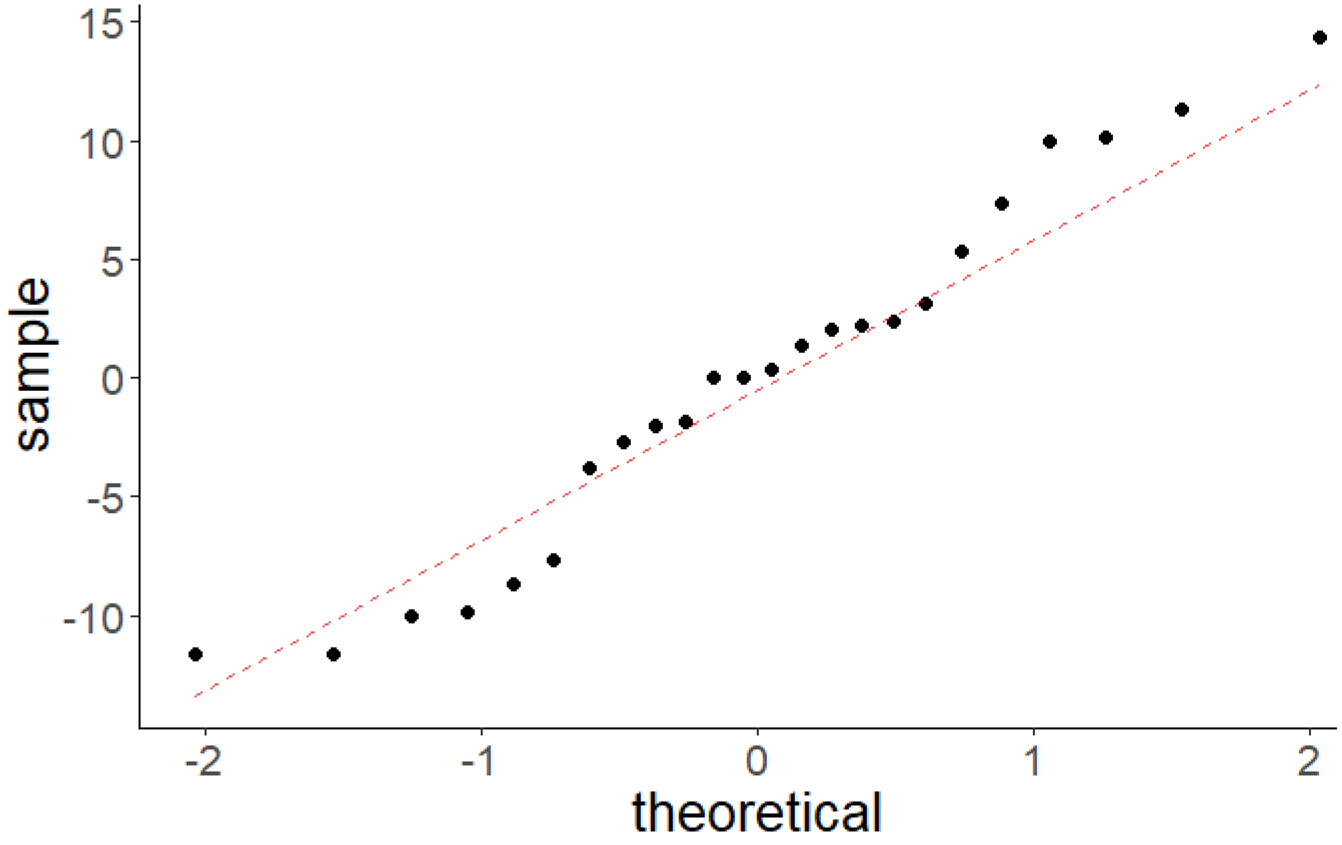
Checking validity of ANOVA

- Are residuals normally distributed?
- Q-Q plot of residuals
  - All points close to line
- Shapiro-Wilk test on residuals

```
shapiro.test(anova$residuals)
      Shapiro-Wilk normality test


data:  anova$residuals
W = 0.96022, p-value = 0.4428
```

P-value > 0.



- Both normality checks suggest it is reasonable to assume residuals are normally distributed.

ANOVA test is valid

Sample size

Sample size – minimum size required

- There is a minimum sample size required.
- Depends on
  - The difference in true response to be detected.
  - The variation between experimental units.
    - The standard deviation.
  - The level of risk to be allowed.
    - Significance level.
    - Power of test.
- Can be obtained in advance if there is an estimate of the error standard deviation, e.g. obtained from a preliminary study.

Significance level and power of a test.

- **Significance level:** the probability of wrongly rejecting a true NULL hypothesis.
  - Typical levels are 0.1, **0.05**, 0.01 and 0.001
- **Power of a test:** probability of correctly rejecting a false NULL hypothesis
  - Typical values are 0.8, 0.85, 0.9, 0.95, 0.99

Standardised difference

- **Standardised difference:** minimum size of true difference which is deemed to be of practical significance.
  - The detection of such difference is important.
  - Standardised difference = (true difference) / (estimated standard deviation)

Example – minimum sample size

- Find the minimum sample sizes for a two-sided two-sample t-test to detect an actual difference of 8 units at a significance level of 0.05 with power 0.90. A preliminary study has shown a standard deviation of approximately 10 units.
- Standardised difference = 8/10 = 0.8

```
power.t.test(power=0.9,delta=8,sd=10,sig.level=0.05,type="two.sample")
```

```
 Two-sample t test power calculation
              n = 33.82555
          delta = 8
             sd = 10
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE:n is number in *each* group
```

Always round up

- So the recommended sample size is 34 for each algorithm.

How does power level affect sample size?

- Assume desired difference is 8 units with α= 0.05 and estimated σ =10

```
a <- power.t.test(power=0.8,delta=8,sd=10,sig.level=0.05,type="two.sample")
B <- power.t.test(power=0.85,delta=8,sd=10,sig.level=0.05,type="two.sample")
c <- power.t.test(power=0.9,delta=8,sd=10,sig.level=0.05,type="two.sample")
d <- power.t.test(power=0.95,delta=8,sd=10,sig.level=0.05,type="two.sample")
e <- power.t.test(power=0.99,delta=8,sd=10,sig.level=0.05,type="two.sample")
f <- power.t.test(power=0.999,delta=8,sd=10,sig.level=0.05,type="two.sample")
cat(
  paste("",
       "n = ", ceiling(a$n), "units = ", a$delta, "power = ",  a$power, "\n",
       "n = ", ceiling(b$n), "units = ", b$delta, "power = ",  b$power, "\n",
       "n = ", ceiling(c$n), "units = ", c$delta, "power = ",  c$power, "\n",
       "n = ", ceiling(d$n), "units = ", d$delta, "power = ",  d$power, "\n",
       "n = ", ceiling(e$n), "units = ", e$delta, "power = ",  e$power, "\n",
       "n = ", ceiling(f$n), "units = ", f$delta, "power = ", f$power,
       sep = "\t"))
```

new line

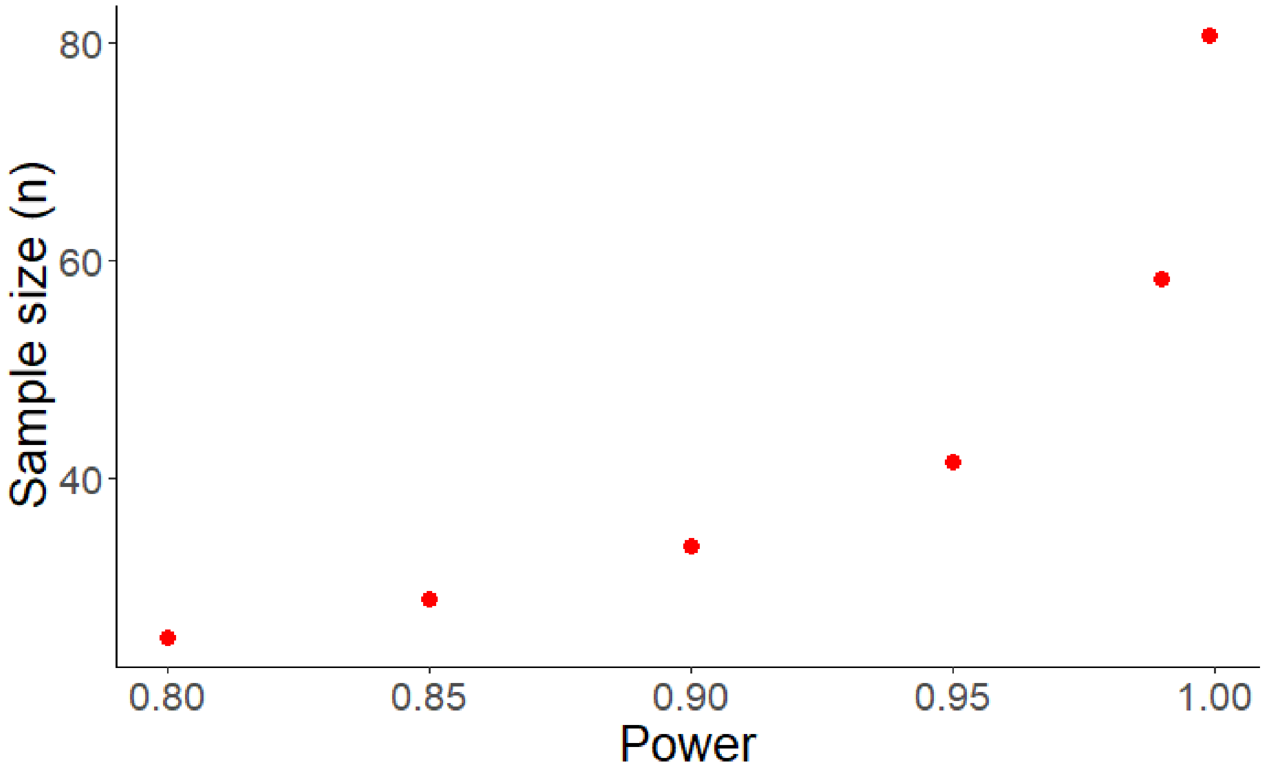
tab separator

Sample size vs. power

Assume

- sig.level = 0.05
- estimated σ =10

```
n  =   26          power  =    0.8
n  =   30          power  =    0.85
n  =   34          power  =    0.9
n  =   42          power  =    0.95
n  =   59          power  =    0.99
n  =   81          power  =    0.999
```



Statistical approach to design and analysis

- State problem. E.g. compare means of 4 algorithms.
  - Objectives of the experiment clearly stated.
- Choose **factors** and **levels**
  - Independent variables (factors) to be investigated.
    - Qualitative or
    - Quantitative: consider how factors can be controlled at required values and measures taken.
  - Levels (values) of factors to be used in the experiment. May be chosen or selected at random from set of possible levels.
  - E.g.
    - Qualitative: type of algorithm with 2 levels, A and B.
    - Quantitative: time-out thresholds. Time allowed for successful retrieval. Levels: 60, 90 and 120 seconds.

… approach to design and analysis

- Select **response** (dependent) variable
  - It must be possible to measure the response variable. Need to state
    - how it will be measured.
    - Level of accuracy.
  - It must provide information about the problem investigated.
  - E.g. runtime for retrieval in seconds or; number of successful retrievals.
- Choose experimental design: given difference in true response to be detected and risks to be tolerated find
  - Sample size (i.e. number of replicates).
  - Order of data collection.
  - Method of randomisation.
  - Balance between accuracy and cost.
  - Mathematical model for the experiment: analysis of data.

… approach to design and analysis (cont)

- Perform experiment
  - Data collection process.
  - Ensure experiment goes according to plan including: randomisation, measurement of accuracy, uniform environment throughout the experiment.
- Data analysis
  - Use of appropriate statistical methods.
  - Also check for normality and outliers as appropriate.
- Report on results, i.e. conclusions and recommendations:
  - Interpret statistical inferences, evaluating their practical significance.
  - May recommend more experiments
  - Ensure report includes relevant information.

Reporting on results

- Report:
  - Comparison under study
  - Statistic used in analysis: value obtained.
  - p-value

Summary

- Statistics can be used to infer conclusions from sample data
- Confidence intervals – show mean and uncertainty (margin or error)
- Hypothesis testing – parametric tests
  - T-test
  - Anova
  - p-value – used to accept or reject null hypothesis
- Sample sizes
  - Minimum sizes depend on various different aspects including power, delta, estimated standard deviation, significance level and whether the test is two-sided or one-sided.