DATA SCIENCE WITH R

DATA SCIENCE WITH R

HYPOTHESIS TESTING

Introduction to Hypothesis Testing

Basic Framework of a Hypothesis Test

Distance Measures

Central Limit Theorem

Types of Hypothesis Tests

Two Sample Tests

Two Sample T-Tests

Two Sample T-Tests

Example:

Company A has been facing a lot of customer complaints about resolution time when clients contact the customer care center.

Two Sample T-Tests

Example:

Company A has been facing a lot of customer complaints about resolution time when clients contact the customer care center.

The VP, Call Center has been asked to find a solution on reducing issue resolution time.

Two Sample T-Tests

Example:

Company A has been facing a lot of customer complaints about resolution time when clients contact the customer care center.

The VP, Call Center has been asked to find a solution on reducing issue resolution time.

Approach

Two Sample T-Tests

Example:

Company A has been facing a lot of customer complaints about resolution time when clients contact the customer care center.

The VP, Call Center has been asked to find a solution on reducing issue resolution time.

Approach

1. Take a sample of calls, note average current resolution time

Two Sample T-Tests

Example:

Company A has been facing a lot of customer complaints about resolution time when clients contact the customer care center.

The VP, Call Center has been asked to find a solution on reducing issue resolution time.

Approach

1. Take a sample of calls, note average current resolution time

2. Implement project, and note average call resolution time for another sample

Two Sample T-Tests

Example:

Company A has been facing a lot of customer complaints about resolution time when clients contact the customer care center.

The VP, Call Center has been asked to find a solution on reducing issue resolution time.

Approach

1. Take a sample of calls, note average current resolution time
2. Implement project, and note average call resolution time for another sample
3. Check if sample means are significantly different

Two Sample T-Tests

The two-sample t-test is used when means across two groups (or samples) are compared

	Avg Call Time Before Implementation	Avg Call Time Post Implementation
	8.5	6.5
	9.5	5.5
	10	6.5
	7	3
	12.5	5
	9	11
	4	8.5
	9	9.5
	9	8.5
	9	9
Average	8.75	7.3

Two Sample T-Tests

The two-sample t-test is used when means across two groups (or samples) are compared

In most business situations we check for differences across samples rather than a sample outcome against a population outcome

	Avg Call Time Before Implementation	Avg Call Time Post Implementation
	8.5	6.5
	9.5	5.5
	10	6.5
	7	3
	12.5	5
	9	11
	4	8.5
	9	9.5
	9	8.5
	9	9
Average	8.75	7.3

Two Sample T-Tests

The two-sample t-test is used when means across two groups (or samples) are compared

In most business situations we check for differences across samples rather than a sample outcome against a population outcome

In this example, can we use the data to show a significant improvement (i.e reduction) in average call time post project implementation?

	Avg Call Time Before Implementation	Avg Call Time Post Implementation
	8.5	6.5
	9.5	5.5
	10	6.5
	7	3
	12.5	5
	9	11
	4	8.5
	9	9.5
	9	8.5
	9	9
Average	8.75	7.3

Two Sample T-Tests

The test statistic is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Two Sample T-Tests

The test statistic is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Assuming equal variance between the two samples.

Two Sample T-Tests

The test statistic is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Assuming equal variance between the two samples.

Variances are unequal:

Two Sample T-Tests

The test statistic is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Assuming equal variance between the two samples.

Variances are unequal:

$$s^2 = \frac{\sum_{j=1}^{n_1}(x_j - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_i - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Two Sample T-Tests

If we calculate the Test Statistic assuming equal variance:

1. Mean of group 1 = 8.75
2. Mean of group 2 = 7.3
3. Std Deviation Group 1 = 2.16
4. Std Deviation Group 2 = 2.41
5. DF = 10 + 10 -2 = 18
6. T-stat = 1.41

Two Sample T-Tests

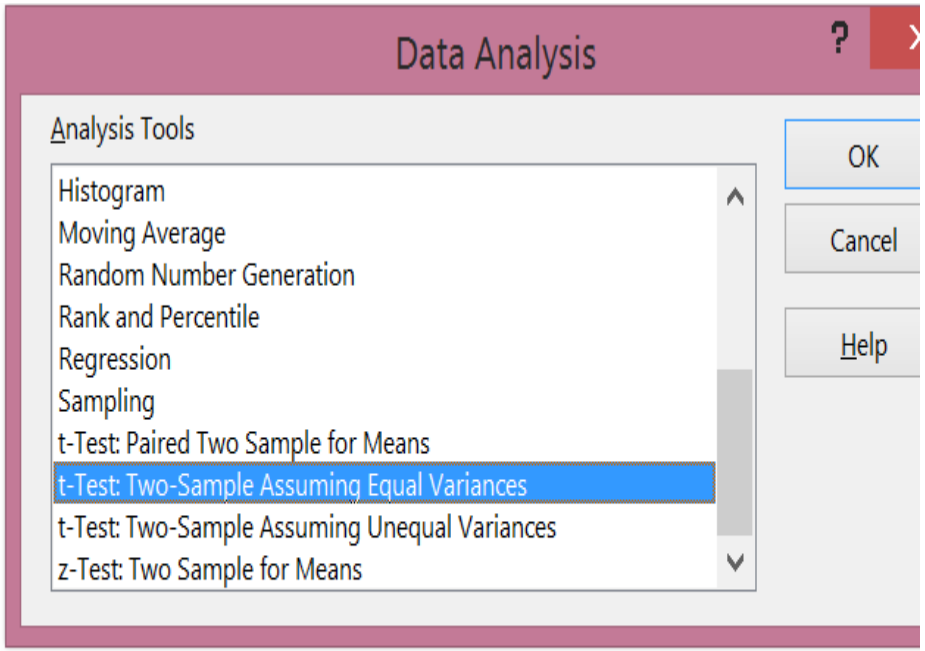
If we calculate the Test Statistic assuming equal variance:

1. Mean of group 1 = 8.75
2. Mean of group 2 = 7.3
3. Std Deviation Group 1 = 2.16
4. Std Deviation Group 2 = 2.41
5. DF = 10 + 10 -2 = 18
6. T-stat = 1.41

We can also use Excel built in options: **Data Analysis: Two Sample Test**

Two Sample T-Tests

	Avg Call Time Before Implementation	Avg Call Time Post Implementation
	8.5	6.5
	9.5	5.5
	10	6.5
	7	3
	12.5	5
	9	11
	4	8.5
	9	9.5
	9	8.5
	9	9
Average	8.75	7.3



Two Sample T-Tests

t-Test: Two-Sample Assuming Equal Variances		
	Avg Call Time Before Implementation	*Avg Call Time Post Implementation*
Mean	8.75	7.3
Variance	4.680555556	5.844444444
Observations	10	10
Pooled Variance	5.2625	
Hypothesized Mean Difference	0	
df	18	
t Stat	1.413373519	
P(T<=t) one-tail	0.087303288	
t Critical one-tail	1.734063607	
P(T<=t) two-tail	0.174606576	
t Critical two-tail	2.10092204	

Two Sample T-Tests

We compared two samples that had equal observations

Two Sample T-Tests

We compared two samples that had equal observations

1. If there are unequal observations, we can still use the t-test, but the degrees of freedom used should be 1 less the small sample size

Two Sample T-Tests

We compared two samples that had equal observations

1. If there are unequal observations, we can still use the t-test, but the degrees of freedom used should be 1 less the small sample size

2. We can also assume similar variance across the two samples, in which case the t-stat will be simplified, but if variance is not similar, use the test for unequal variance in Excel

Paired Difference T-Tests

We looked at comparisons of average call time for a *random 10 calls* before the project was implemented, and a *random 10 calls* after a solution was designed

Sometimes we may want to test **observations that are paired** to see if there is a true difference in their means before and after the experiment

Respondent #	Weight - Pre	Weight - Post
1	162	168
2	170	158
3	184	186
4	164	155
5	172	143
6	176	161
7	159	160
8	170	135

Paired Difference T-Tests

We looked at comparisons of average call time for a *random 10 calls* before the project was implemented, and a *random 10 calls* after a solution was designed

Sometimes we may want to test **observations that are paired** to see if there is a true difference in their means before and after the experiment

For example:

Testing the efficacy of a particular drug that claims it will help patients lose weight -

Respondent #	Weight - Pre	Weight - Post
1	162	168
2	170	158
3	184	186
4	164	155
5	172	143
6	176	161
7	159	160
8	170	135

Paired Difference T-Tests

We looked at comparisons of average call time for a *random 10 calls* before the project was implemented, and a *random 10 calls* after a solution was designed

Sometimes we may want to test **observations that are paired** to see if there is a true difference in their means before and after the experiment

For example:

Testing the efficacy of a particular drug that claims it will help patients lose weight -

We record average 'before & after' weight for 8 respondents for 20 weeks

Respondent #	Weight - Pre	Weight - Post
1	162	168
2	170	158
3	184	186
4	164	155
5	172	143
6	176	161
7	159	160
8	170	135

Paired Difference T-Tests

Test statistic is: $$t = \frac{\bar{d}}{\sqrt{s^2 / n}}$$

Where d is the difference in scores

Paired Difference T-Tests

Test statistic is: $$t = \frac{\bar{d}}{\sqrt{s^2 / n}}$$

Where d is the difference in scores

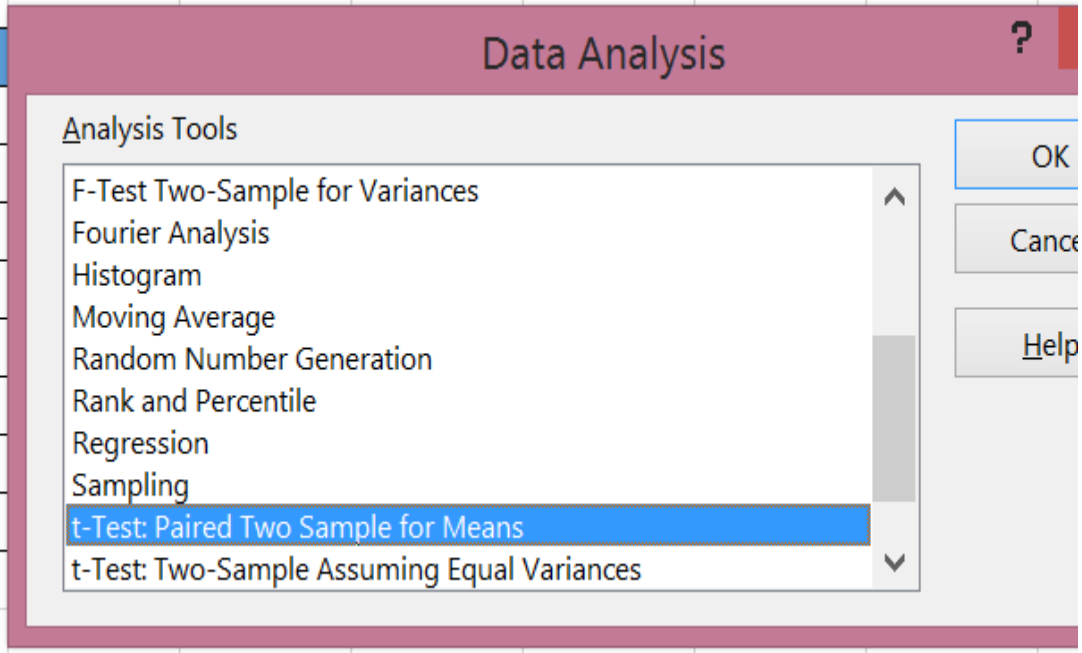
We need to calculate the mean and standard error of the pre-post differences for each pair and then use that for the test statistic

Let's say our hypothesis is that the drug does have a positive impact on weight loss, and we would like to use a 95% confidence level -

How would you test the hypothesis?

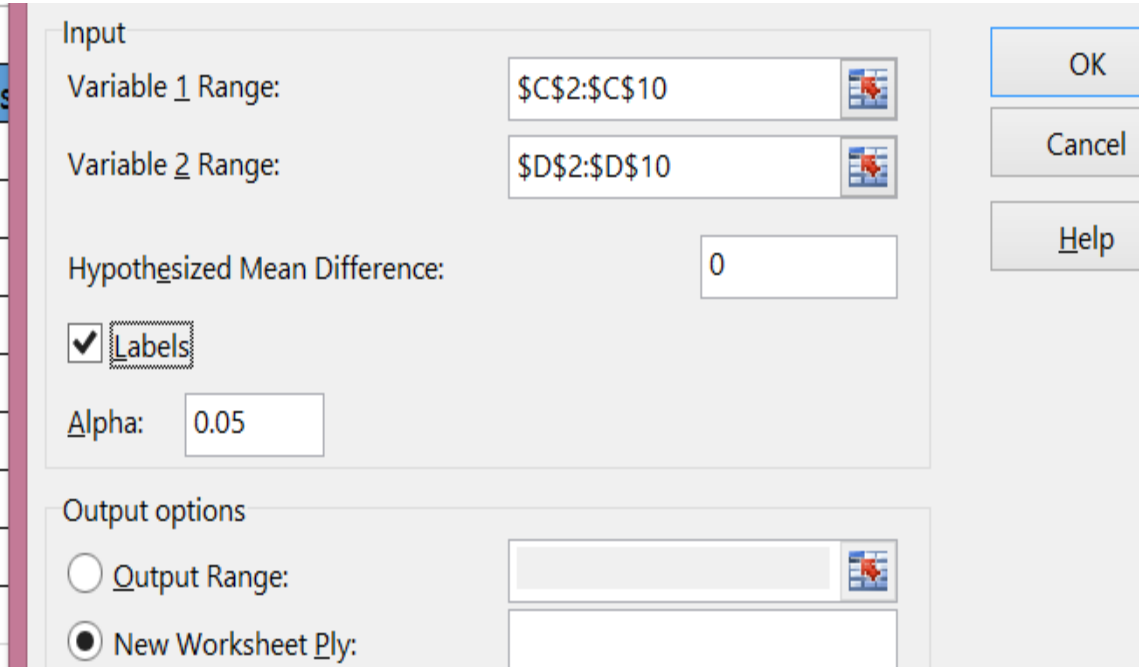
Paired Difference T-Tests

Respondent #	Weight - Pre	Weight - Post
1	162	168
2	170	158
3	184	186
4	164	155
5	172	143
6	176	161
7	159	160
8	170	135



Paired Difference T-Tests

Respondent #	Weight - Pre	Weight - Pos
1	162	168
2	170	158
3	184	186
4	164	155
5	172	143
6	176	161
7	159	160
8	170	135



Coming Up

Types of Hypothesis Tests:

Multiple Sample Tests

THANK YOU