

Data Analysis and Visualisation

Lecture 5
The t Test

Introduction

- Topics covered
 - The t Test for Independent Samples
 - The Critical Value of t : t_{crit}
 - Requirements for Using the t Test for Independent Samples
 - The t Test for Dependent (Matched) Samples



- This test is often used where a sample is drawn from a population and then randomly assigned to one of two groups

Student's T-test

How can we determine, to a reasonable degree of scientific certainty, if one variety of barley yields more than another?





VOLUME VI

MARCH, 1908

No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

By STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form

Assumptions - t Test for Independent Samples

- Measurements from the two groups aren't paired in any way
- both populations considered are essentially normally distributed.

t Test for Independent Samples

$$t_{obt} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left[\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2} \right] \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

- The above equation assumes **unequal group** sample sizes
- If the sample sizes are **equal** then this simplifies to

$$t_{obt} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{SD_1^2 + SD_2^2}{n}}}$$

- Where

t_{obt} = the value of t obtained from your data

n_1, n_2 = the number of items in samples 1 and 2

SD_1^2, SD_2^2 = the estimates of the variances of the two populations

$\bar{X}_1, \bar{X}_2$ = the means of the two samples

Worked example



Does the pH of soil affects seed germination of a specific plant species?

- Group 1: eight pots with soil at pH 5.5
- Group 2: eight pots with soil at pH 7.0
- 50 seeds planted in each pot and the number that germinated in each pot was recorded.

What is the null hypothesis (H_0)?

H_0 = there is **no** statistically significant difference between the germination success of seeds in two soils of different pH

H_A = there is a significant difference between the germination of seeds in two soils of different pH

If the value for t **exceeds** the critical value ($P = 0.05$), then you can **reject** the null hypothesis.

Pot	Group 1 (pH5.5)	Group 2 (pH7.0)
1	38	39
2	41	45
3	43	41
4	39	46
5	37	48
6	38	39
7	41	46
8	36	44
Mean	39.1	43.5
Σ		

Calculate standard deviation for both groups

Group 1:

$$SD = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}} = \sqrt{\frac{38.88}{8 - 1}} = 2.36$$

Group 2:

$$SD = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}} = \sqrt{\frac{82.0}{8 - 1}} = 3.42$$

Using your means and SDs, calculate value for t

$$t_{obt} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$$

$$t_{obt} = \frac{39.1 - 43.5}{\sqrt{(2.36^2/8) + (3.42^2/8)}} = \frac{-4.4}{\sqrt{0.696 + 1.462}}$$

$$t_{obt} = 2.99$$

What does t_{obt} mean?

The Critical Value of t: t_{crit}

- What does t_{obt} mean?

Is t_{obt} statistically significant

- probably reflects random differences between the two samples
- you will have to compare it with what is known as the critical value of t (t_{crit}).

degrees of freedom (df)

In our example,

$$df = n_1 + n_2 - 2 = 8 + 8 - 2 = 14.$$

The value of t_{crit} is determined by df and by selected level of significance.



Adobe Acrobat
Document

Critical Value t_{crit}

Degrees of freedom	Significance level					
	20% (0.20)	10% (0.10)	5% (0.05)	2% (0.02)	1% (0.01)	0.1% (0.001)
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850

- Cross-referencing this with $p=0.05$ (the level of significance) gives us 2.15
- Our value for t **exceeds** the critical value, so we can reject the null hypothesis.

Conclusion: Significant difference between two groups, pH **does** affect the germination rate for this plant

Exercise

- Suppose we have two Field, A and B, and we need to measure the yield over each Field.

Field A	Field B
25	14
14	11
23	13
21	9
24	15
17	12
19	9
20	11
15	8
22	13
16	12
21	14
$\bar{X}_1 =$	$\bar{X}_2 =$
$n_1 =$	$n_2 =$
$SD_1^2 =$	$SD_2^2 =$

t Test for Independent Samples

$$t_{obt} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{SD_1^2 + SD_2^2}{n_1 + n_2}}}$$

$$t_{obt} = \frac{19.75 - 11.75}{\sqrt{\frac{12.93 + 4.93}{12}}}$$

$$t_{obt} = \frac{8}{\sqrt{\frac{17.86}{12}}}$$

$$t_{obt} = \frac{8}{\sqrt{1.49}}$$

$$t_{obt} = \frac{8}{1.22}$$

$$t_{obt} = 6.56 \quad t_{crit} = 2.07$$

Field A	Field B
25	14
14	11
23	13
21	9
24	15
17	12
19	9
20	11
15	8
22	13
16	12
21	14

$$\bar{X}_1 = 19.75$$

$$\bar{X}_2 = 11.75$$

$$n_1 = 12$$

$$n_2 = 12$$

$$SD_1^2 = 12.93$$

$$SD_2^2 = 4.93$$

Critical Value t_{crit}

- The level of significance chosen is known as α (alpha)- probability of getting a **Type I error**
- We can therefore reject $H_0: \mu_1 = \mu_2$ at the $p < 0.05$ level and accept the alternative hypothesis – yield at two fields are different.

The 5 Stages of Analysis

1. State your hypotheses
 - H_0 (the null hypothesis)
 - H_A or H_1 (the alternative hypothesis)
2. Select α , your significance level
 - ***0.05 is usually used*** but 0.01 can be used in some cases
3. Calculate t_{obt}
4. Find t_{crit}
5. Decide whether or not to reject the null hypothesis

Requirements for Independent Samples

- This form of the t Test requires two independent samples
- There are no relationships between data in group 1 and group 2;
- t Tests assume both populations are normally distributed
- For small or highly skewed samples we use non-parametric tests like the chi-squared test
- Alternatively- for paired data, t Test for non independent groups is used.

t Test for Dependent (Matched) Samples

- Suppose you gave a group of ten subjects a test both before and after a game
- You have two sets of scores, one pre-test and one post-test
- Aim: Find out if attitudes measured by the tests were more or less favorable after playing the game than they were before.

Subject	Pretest	Posttest	Posttest-Pretest (D)	
1	84	89	+5	
2	87	92	+5	$\bar{X}_{\text{pretest}} = 90.00$
3	87	98	+11	$\bar{X}_{\text{posttest}} = 95.00$
4	90	95	+5	
5	90	95	+5	
6	90	95	+5	$SD_{\text{pretest}} = 3.46$
7	90	95	+5	$SD_{\text{posttest}} = 3.46$
8	93	92	-1	$R_{12} = +0.67$
9	93	98	+5	
10	96	101	+5	

Conceptual formula -dependent sample t Test

$$t^*_{\text{obt}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(S_{\bar{x}_1})^2 + (S_{\bar{x}_2})^2 - 2 r_{12} S_{\bar{x}_1} S_{\bar{x}_2}}}$$

Where $\bar{x}_1, \bar{x}_2$ = the means of the two measurement

$S_{\bar{x}_1}, S_{\bar{x}_2}$ = the standard error of the means ($\frac{sD}{\sqrt{N}}$)

r_{12} = the correlation between the two measurements

Example – The 5 Stages of Analysis

1. State your hypotheses $H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$
2. Select α , your significance level $\alpha = 0.05$
3. Calculate t_{obt}

$$\begin{aligned}
 t_{obt}^* &= \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_{\bar{X}_1}^2 + S_{\bar{X}_2}^2 - 2r_{12}S_{\bar{X}_1}S_{\bar{X}_2}}} = \frac{90.0 - 95.0}{\sqrt{\left(\frac{3.46}{\sqrt{10}}\right)^2 + \left(\frac{3.46}{\sqrt{10}}\right)^2 - 2(.67)\left(\frac{3.46}{\sqrt{10}}\right)\left(\frac{3.46}{\sqrt{10}}\right)}} \\
 &= \frac{90.0 - 95.0}{\sqrt{\left(\frac{3.46}{\sqrt{10}}\right)^2 + \left(\frac{3.46}{\sqrt{10}}\right)^2 - 2(.67)\left(\frac{3.46}{\sqrt{10}}\right)\left(\frac{3.46}{\sqrt{10}}\right)}} \\
 &= \frac{90.0 - 95.0}{\sqrt{(1.09)^2 + (1.09)^2 - 2(.67)(1.09)(1.09)}} \\
 &= \frac{90.0 - 95.0}{\sqrt{(1.09)^2 + (1.09)^2 - 2(.67)(1.09)(1.09)}} \\
 &= \frac{90.0 - 95.0}{\sqrt{(1.19) + (1.19) - 2(.67)(1.09)(1.09)}} \\
 &= \frac{-5.0}{\sqrt{2.38 - 1.59}} = -5.62
 \end{aligned}$$

Example – The 5 Stages of Analysis

$$\begin{aligned} 4. \quad df &= N - 1 \\ &= 10 - 1 = 9 \end{aligned}$$

Therefore

$t_{crit} = 2.262$ at the .05 level of significance.

5. *Decide Whether to Reject H_0 .*

t_{obt} is much larger than t_{crit} so we can reject the null hypothesis.

Conclusion:

- Game appeared to have a positive effect on attitudes toward public schools.

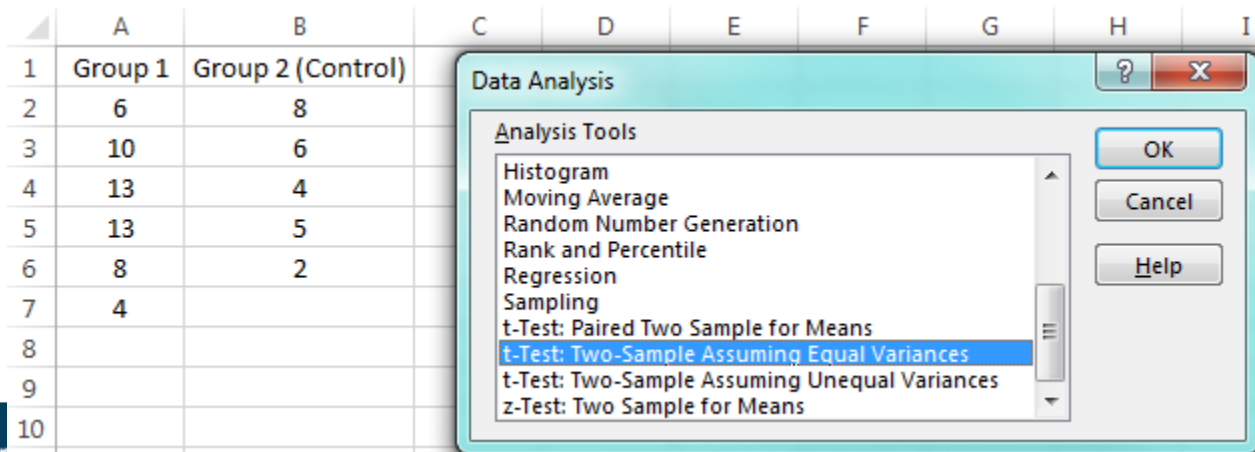
Exercise

- Using hospital and agency records, a researcher locates six pairs of twins, one of whom was adopted at birth and one of whom was in foster care for at least three years. All of the twins are now five years old. The researcher wants to show that early adoption leads to better intellectual ability so researcher test all the twins with the Wechsler Intelligence Scale for Children (WISC). The results are as shown.
- Define null and alternative hypotheses for this study
- Run the appropriate test
- Can you reject the null hypothesis at $p=0.05$?

Twin pair #	Adopted twin	Fostered twin
1	105	103
2	99	97
3	112	105
4	101	99
5	124	104
6	100	110

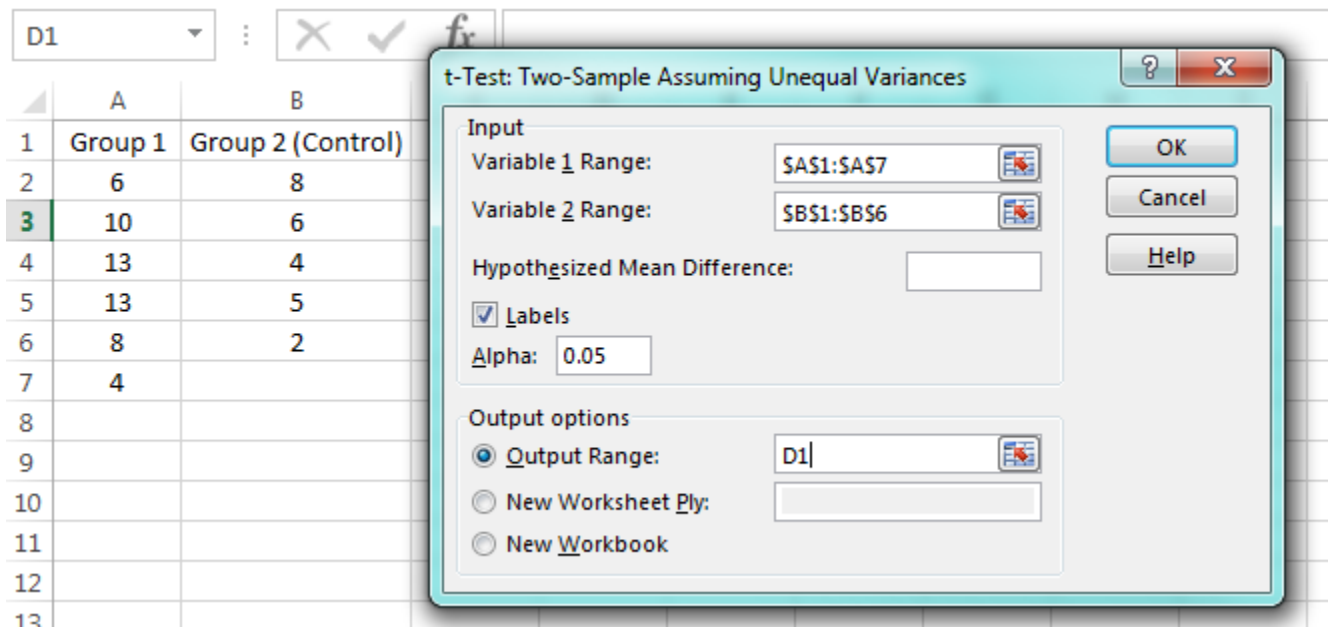
t Tests in Excel

- The t Test is part of the Data Analysis toolpak
- It provides 3 t Test options
 - t Test: Paired Two Sample for Means
 - t Test: Two-Sample Assuming Equal Variances
 - t Test: Two-Sample Assuming Unequal Variances
- Depending on your analysis you will need the first or third of these tools



t Tests in Excel

- Input the data range for analysis, input a location for the data to be output, and tick the labels box if appropriate
- Excel assumes that Alpha is 0.05 but allows you to change it if required



t Tests in Excel

	A	B	C	D	E	F
1	Group 1	Group 2 (Control)		t-Test: Two-Sample Assuming Equal Variances		
2	6	8				
3	10	6			Group 1	Group 2 (Control)
4	13	4		Mean	9	5
5	13	5		Variance	13.6	5
6	8	2		Observations	6	5
7	4			Pooled Variance	9.777777778	
8				Hypothesized Mean Difference	0	
9				df	5	
10				t Stat	2.112536371	
11				P(T<=t) one-tail	0.031903219	
12				t Critical one-tail	1.833112933	
13				P(T<=t) two-tail	0.063806438	
14				t Critical two-tail	2.262157163	
15						

- t_{opt} is shown as t Stat, t_{crit} as t Critical one-tail or t Critical two-tail
- $P(T \leq t)$ is the probability that the difference has appeared by chance
- In this case there would be a statistically significant difference if we had chosen the one-tail test but not the two-tail test

Problem

- A researcher tests the high-frequency hearing acuity of a group of teens two days before they attend a rock concert. Two days after the concert she tests them again. Her results are in file DAV Datafile 16. She expects to find that the teens hear better before the concert than afterwards (the higher the score on the test the poorer the hearing)
 - Using the appropriate Excel function, analyse the data according to the 5 stage process
1. State your hypotheses
 2. Select α , your significance level
 3. Calculate t_{obt}
 4. Find t_{crit}
 5. Decide whether or not to reject the null hypothesis

Summary

- Topics covered
 - The t Test for Independent Samples
 - The Critical Value of t : t_{crit}
 - Requirements for Using the t Test for Independent Samples
 - The t Test for Dependent (Matched) Samples
- Next week we will cover
 - Computational steps
 - Strength of Association
 - Post hoc Analyses
 - The Scheffé Method of Post hoc Analysis
- The ANOVA test allows us to compare the means of multiple samples

Coming soon

- **Next session:** Analysis of Variance (ANOVA)
- **Homework:** Read chapter 11 of *Statistics for the Terrified*.