

Session 11: Two-population Tests

Statistics for Data Science
Master in Business Analytics and Big Data

Professor: Rafael Ballester-Ripoll

Note on the One-population KS Test

- It doesn't matter what particular distribution we are checking
- One-sample test: **one table fits all**
- Want to get the exact ***p*-value** from **deviation D** ?
 - Install the [Real Statistics Resource Pack](#)
 - For the one-sample test: =KSPROB(D , n)
 - For the two-sample test: =KSDIST(D , n_1 , n_2)

Two-population Inference

- Parameter estimation: use one **sample** to learn about one **population**
- Now: **relate two populations** using **two samples**
 - Notation: \mathbf{x}_1 and \mathbf{x}_2
- Best to use automated statistical tools
 - [Excel Analysis ToolPak](#)
 - [KS-test calculator for two samples](#)
 - (optional) [Real Statistics Resource Pack](#)
- We will do:
 - Compare **two means**
 - Compare **two distributions**

Two Means Example 1

An economist decided to test the hypothesis that different retail prices are being charged for Japanese automobiles in Japan than in the United States. She obtained independent random samples of 50 retail sales in the United States and 50 retail sales in Japan over the same time period and for the same model of automobile and converted the Japanese sales prices from yen to dollars using current conversion rates. Get the data (autos.xlsx) from the Campus Online. *What hypothesis would you favor?*

- $H_0: \mu_{\text{US}} - \mu_{\text{Japan}} = 0$ (the retail prices are the same)
- $H_a: \mu_{\text{US}} - \mu_{\text{Japan}} \neq 0$ (the retail prices are different)

Source: *Statistics for Business and Economics* (McClave et al.), 12th Edition

Solution

- **Excel ToolPak:** t -test assuming unequal variances

The screenshot shows an Excel spreadsheet with two columns of data: 'Retail prices in US' (Column A) and 'Retail prices in Japan' (Column B). The data is as follows:

	A	B
1	Retail prices in US	Retail prices in Japan
2	28.2	28.5
3	28.5	26.8
4	28.2	26.4
5	27.3	30.4
6	26.9	28.5
7	26.2	24
8	25.5	29.8
9	26.3	28
10	28.8	27.9
11	23.3	26.9
12	27.2	28.2
13	26.2	27.3
14	26.8	27.5
15	24.9	25.5
16	26.3	27.6

The 'Data Analysis' dialog box is open, showing the 'Analysis Tools' list. The option 't-Test: Two-Sample Assuming Unequal Variances' is selected.

Data Analysis

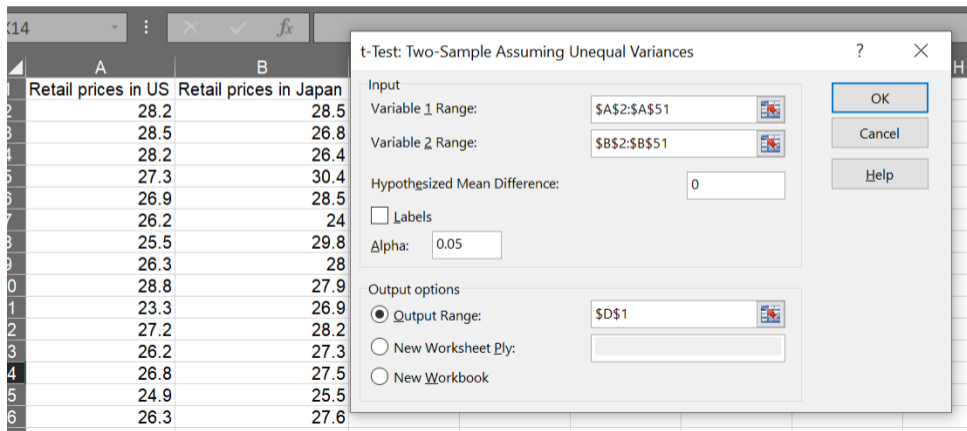
Analysis Tools

- Fourier Analysis
- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression
- Sampling
- t-Test: Paired Two Sample for Means
- t-Test: Two-Sample Assuming Equal Variances
- t-Test: Two-Sample Assuming Unequal Variances**

Buttons: OK, Cancel, Help

Solution

- Specify inputs, hypothesis, tolerance (α), output cell



The image shows an Excel spreadsheet with two columns of retail prices. Column A is labeled 'Retail prices in US' and Column B is labeled 'Retail prices in Japan'. The data ranges from row 2 to row 11. A 't-Test: Two-Sample Assuming Unequal Variances' dialog box is open, showing the following settings:

Input
Variable 1 Range: \$A\$2:\$A\$11
Variable 2 Range: \$B\$2:\$B\$11
Hypothesized Mean Difference: 0
<input type="checkbox"/> Labels
Alpha: 0.05

Output options
<input checked="" type="radio"/> Output Range: \$D\$1
<input type="radio"/> New Worksheet Ply:
<input type="radio"/> New Workbook

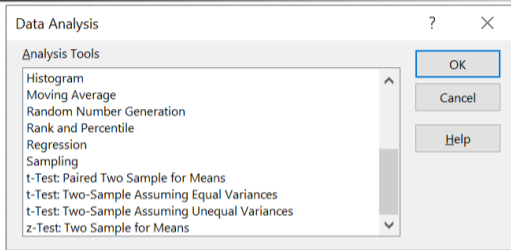
The dialog box has 'OK', 'Cancel', and 'Help' buttons on the right.

Solution

- Two-tailed p -value = 0.1088 \rightarrow **do not reject H_0**

	A	B	C	D	E	F	G	H
1	Retail prices in US	Retail prices in Japan		t-Test: Two-Sample Assuming Unequal Variances				
2	28.2	28.5						
3	28.5	26.8			Variable 1	Variable 2		
4	28.2	26.4		Mean	26.596	27.236		
5	27.3	30.4		Variance	3.92610612	3.8970449		
6	26.9	28.5		Observations	50	50		
7	26.2	24		Hypothesized	0			
8	25.5	29.8		df	98			
9	26.3	28		t Stat	-1.61798361			
10	28.8	27.9		P(T<=t) one-t	0.05444019			
11	23.3	26.9		t Critical one-t	1.66055122			
12	27.2	28.2		P(T<=t) two-t	0.10888037			
13	26.2	27.3		t Critical two-t	1.98446745			
14	26.8	27.5						
15	24.9	25.5						
16	26.3	27.6						

Tool Summary



- If the variances are unknown:
 - If known to be equal: t -test, two-sample assuming equal variances
 - Otherwise: t -test, two-sample assuming unequal variances
- If the variances are known:
 - z -test, two-sample for means
- If the samples are *the same individuals*:
 - t -test, paired two sample for means

When to Use What? Practice, Practice, Practice

Which two-sample test do we need? 1 or 2 tails?

t-Test: Paired Two Sample for Means

t-Test: Two-Sample Assuming Equal Variances

t-Test: Two-Sample Assuming Unequal Variances

z-Test: Two Sample for Means

Do Uber and Lyft rides cost the same on average?

- **Two-tailed**, because we care about any difference in costs
- **t-with unequal variance**, because we cannot assume they have the same variance (why would they?)

When to Use What? Practice, Practice, Practice

Which two-sample test do we need? 1 or 2 tails?

t-Test: Paired Two Sample for Means

t-Test: Two-Sample Assuming Equal Variances

t-Test: Two-Sample Assuming Unequal Variances

z-Test: Two Sample for Means

*We monitor a group of patients' viral load before and after treatment.
Did the load decrease?*

- **One-tailed**, because we look at decreases only
- **paired t-test**, the individuals are the same before and after

When to Use What? Practice, Practice, Practice

Which two-sample test do we need? 1 or 2 tails?

t-Test: Paired Two Sample for Means

t-Test: Two-Sample Assuming Equal Variances

t-Test: Two-Sample Assuming Unequal Variances

z-Test: Two Sample for Means

Did COVID-19 decrease rental prices in Madrid?

- **One-tailed**, because we look at decreases only
- **t-with unequal variance**, because we cannot assume they have the same variance (why would they?)

When to Use What? Practice, Practice, Practice

Which two-sample test do we need? 1 or 2 tails?

t-Test: Paired Two Sample for Means

t-Test: Two-Sample Assuming Equal Variances

t-Test: Two-Sample Assuming Unequal Variances

z-Test: Two Sample for Means

A portfolio is dynamically updated so that its risk (volatility) stays at 0.005. In terms of average log-returns, has its performance changed between 2022 and 2023?

- **Two-tailed**, because we care about any difference in average return
- **z-test: two sample for means**, because the standard deviation is known to be 0.005 in both cases

When to Use What? Practice, Practice, Practice

Which two-sample test do we need? 1 or 2 tails?

t-Test: Paired Two Sample for Means

t-Test: Two-Sample Assuming Equal Variances

t-Test: Two-Sample Assuming Unequal Variances

z-Test: Two Sample for Means

Portfolio A is dynamically updated so that it replicates the risk (volatility) of portfolio B. In terms of log-returns, do they perform equally well?

- **Two-tailed**, because we care about any difference in average return
- **t-test with equal variance**, because the standard deviation is the same

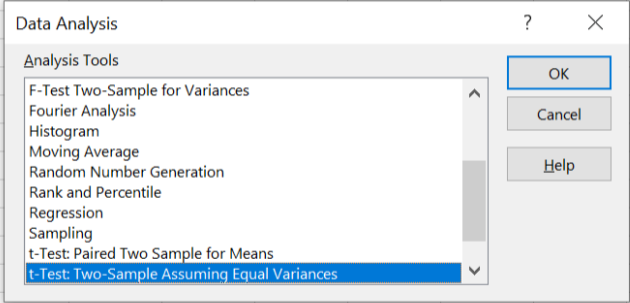
Two Means Example 2

We have selected two groups (20 people each) to study a drug's effectiveness: group A gets the real treatment, while group B is the control group and only gets placebo. For the study to be well done, the groups should be as similar as possible to each other: similar age, gender proportion, lifestyle, etc. The file `ages.xlsx` contains the age of every person in the study.

Compare the age to find out: were the groups really drawn from the same population?

Solution

- **Excel ToolPak:** t -test assuming equal variances



The image shows an Excel spreadsheet with two columns, Group A and Group B, containing numerical data. A 'Data Analysis' dialog box is overlaid on the spreadsheet, with the 't-Test: Two-Sample Assuming Equal Variances' option selected in the list of analysis tools. The dialog box includes 'OK', 'Cancel', and 'Help' buttons.

Group A	Group B
39	27
34	40
36	40
41	34
39	46
28	31
36	37
31	36
32	43
34	43
33	38
38	39
35	33
32	29
34	36
33	38

Solution

- Specify inputs, hypothesis, tolerance (α), output cell

Group A	Group B
39	27
34	40
36	40
41	34
39	46
28	31
36	37
31	36
32	43
34	43
33	38
38	39
35	33
32	29
34	36

t-Test: Two-Sample Assuming Equal Variances

Input

Variable 1 Range:

Variable 2 Range:

Hypothesized Mean Difference:

☒ Labels

Alpha:

Output options

☒ Output Range:

☐ New Worksheet Ply:

☐ New Workbook

OK Cancel Help

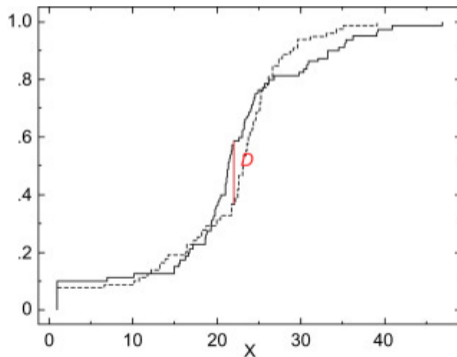
Solution

- Two-tailed p -value = 0.0343 \rightarrow **we reject H_0**

Group A	Group B	t-Test: Two-Sample Assuming Equal Variances							
39	27								
34	40		Group A	Group B					
36	40	Mean	34.3	37.25					
41	34	Variance	12.01053	24.09211					
39	46	Observations	20	20					
28	31	Pooled Variance	18.05132						
36	37	Hypothesized Mean Difference	0						
31	36	df	38						
32	43	t Stat	-2.19567						
34	43	P(T<=t) one-tail	0.017146						
33	38	t Critical one-tail	1.685954						
38	39	P(T<=t) two-tail	0.034292						
35	33	t Critical two-tail	2.024394						
32	29								
34	36								

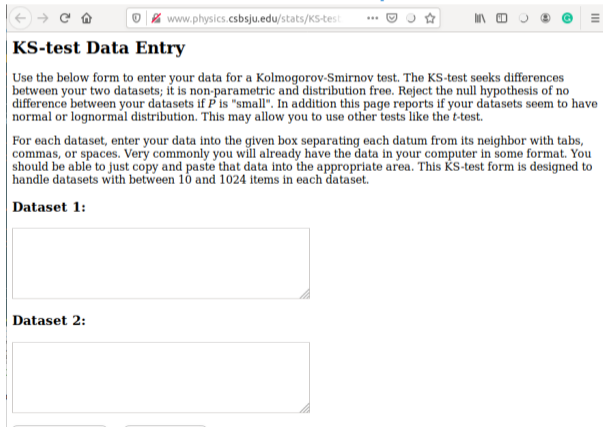
KS Test for Two Samples

- Similar to the one-sample case (last session)
 - Just compare two ECDFs!



KS Test for Two Samples

- There exist **two-population KS tables**
- Best is to use an **online resource**
 - Example: **KS-test calculator for two samples**



The screenshot shows a web browser window with the address bar displaying www.physics.csbsju.edu/stats/KS-test. The page title is "KS-test Data Entry". The main content area contains the following text:

Use the below form to enter your data for a Kolmogorov-Smirnov test. The KS-test seeks differences between your two datasets; it is non-parametric and distribution free. Reject the null hypothesis of no difference between your datasets if P is "small". In addition this page reports if your datasets seem to have normal or lognormal distribution. This may allow you to use other tests like the t -test.

For each dataset, enter your data into the given box separating each datum from its neighbor with tabs, commas, or spaces. Very commonly you will already have the data in your computer in some format. You should be able to just copy and paste that data into the appropriate area. This KS-test form is designed to handle datasets with between 10 and 1024 items in each dataset.

Dataset 1:

[Empty text input box for Dataset 1]

Dataset 2:

[Empty text input box for Dataset 2]

KS Test for Two Samples

Test the exercise from ages.xlsx with a two-sample KS