

# 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)

# Solution

---

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

The screenshot shows a Microsoft Excel spreadsheet with two columns of data: 'Retail prices in US' and 'Retail prices in Japan'. The Data Analysis dialog box is open, listing various statistical tools. The 't-Test: Two-Sample Assuming Unequal Variances' option is highlighted with a blue selection bar.

	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
0	28.8	27.9
1	23.3	26.9
2	27.2	28.2
3	26.2	27.3
4	26.8	27.5
5	24.9	25.5
6	26.3	27.6

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**

OK Cancel Help

# Solution

- Specify inputs, hypothesis, tolerance ( $\alpha$ ), output cell

The screenshot shows a Microsoft Excel spreadsheet and the 't-Test: Two-Sample Assuming Unequal Variances' dialog box.

**Data Table:**

	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
0	28.8	27.9
1	23.3	26.9
2	27.2	28.2
3	26.2	27.3
4	26.8	27.5
5	24.9	25.5
6	26.3	27.6
	---	---

**t-Test Dialog Box:**

**Input**

- Variable 1 Range: \$A\$2:\$A\$51
- Variable 2 Range: \$B\$2:\$B\$51
- Hypothesized Mean Difference: 0
- Labels
- Alpha: 0.05

**Output options**

- Output Range: \$D\$1
- New Worksheet Ply: (empty)
- New Workbook

Buttons: OK, Cancel, Help

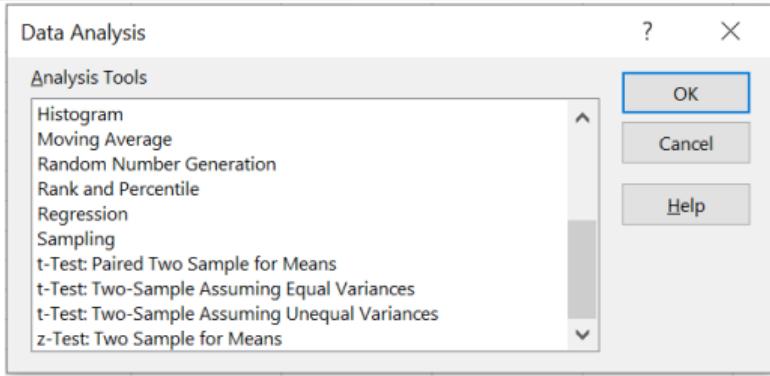
# Solution

---

- Two-tailed  $p$ -value = 0.1088 → **do not reject  $H_0$**

		t-Test: Two-Sample Assuming Unequal Variances			
				Variable 1	Variable 2
	Retail prices in US	Retail prices in Japan	Mean	26.596	27.236
1	28.2	28.5	Variance	3.92610612	3.8970449
2	28.5	26.8	Observations	50	50
3	28.2	26.4	Hypothesized	0	
4	27.3	30.4	df	98	
5	26.9	28.5	t Stat	-1.61798361	
6	26.2	24	P(T<=t) one-t	0.05444019	
7	25.5	29.8	t Critical one-t	1.66055122	
8	26.3	28	P(T<=t) two-t	0.10888037	
9	28.8	27.9	t Critical two-t	1.98446745	
0	23.3	26.9			
1	27.2	28.2			
2	26.2	27.3			
3	26.8	27.5			
4	24.9	25.5			
5	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 screenshot shows a Microsoft Excel spreadsheet with two columns of data labeled "Group A" and "Group B". The "Group A" column contains values: 39, 34, 36, 41, 39, 28, 36, 31, 32, 34, 33, 38, 35, 32, 34, 33. The "Group B" column contains values: 27, 40, 40, 34, 46, 31, 37, 36, 43, 43, 38, 39, 33, 29, 36, 38. To the right of the spreadsheet, the "Data Analysis" dialog box is open. The "Analysis Tools" list includes: F-Test Two-Sample for Variances, Fourier Analysis, Histogram, Moving Average, Random Number Generation, Rank and Percentile, Regression, Sampling, t-Test: Paired Two Sample for Means, and t-Test: Two-Sample Assuming Equal Variances. The "t-Test: Two-Sample Assuming Equal Variances" option is highlighted with a blue selection bar. The dialog box has standard buttons for OK, Cancel, and Help.

	A	B
1	Group A	Group B
2	39	27
3	34	40
4	36	40
5	41	34
6	39	46
7	28	31
8	36	37
9	31	36
10	32	43
11	34	43
12	33	38
13	38	39
14	35	33
15	32	29
16	34	36
17	33	38

# Solution

- Specify inputs, hypothesis, tolerance ( $\alpha$ ), output cell

The screenshot shows a Microsoft Excel spreadsheet with two columns of data labeled "Group A" and "Group B". Below the table is the "t-Test: Two-Sample Assuming Equal Variances" dialog box.

**Data Table:**

	Group A	Group B
1	39	27
2	34	40
3	36	40
4	41	34
5	39	46
6	28	31
7	36	37
8	31	36
9	32	43
10	34	43
11	33	38
12	38	39
13	35	33
14	32	29
15	34	36

**t-Test Dialog Box:**

**Input**

- Variable 1 Range: \$A\$1:\$A\$21
- Variable 2 Range: \$B\$1:\$B\$21

Hypothesized Mean Difference: 0

Labels checkbox is checked.

Alpha: 0.05

**Output options**

- Output Range: \$D\$1 (radio button selected)
- New Worksheet Ply: (radio button)
- New Workbook: (radio button)

OK, Cancel, and Help buttons are visible in the dialog box.

## Solution

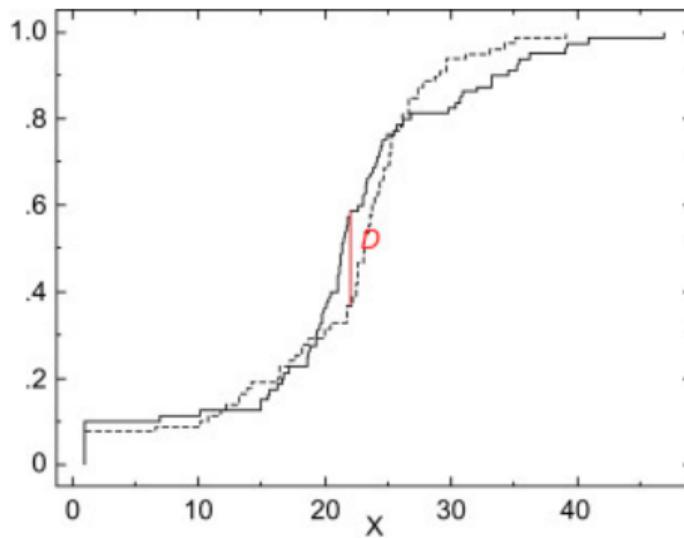
- Two-tailed  $p$ -value = 0.0343 → we reject  $H_0$

Group A	Group B	t-Test: Two-Sample Assuming Equal Variances		
		Group A	Group B	
39	27			
34	40			
36	40	Mean	34.3	37.25
41	34	Variance	12.01053	24.09211
39	46	Observation	20	20
28	31	Pooled Va	18.05132	
36	37	Hypothesi	0	
31	36	df	38	
32	43	t Stat	-2.19567	
34	43	P(T<=t) on	0.017146	
33	38	t Critical o	1.685954	
38	39	P(T<=t) tw	0.034292	
35	33	t Critical tw	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 URL [www.physics.csbsju.edu/stats/KS-test](http://www.physics.csbsju.edu/stats/KS-test). The page title is "KS-test Data Entry". The content explains the purpose of the test and provides instructions for entering data. It includes two large text input fields labeled "Dataset 1:" and "Dataset 2:".

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:**

**Dataset 2:**

# KS Test for Two Samples

---

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