PROFESSIONAL ELECTIVE COURSES: VERTICALS
VERTICAL 1: DATA SCIENCE
CCS346 EXPLORATORY DATA ANALYSIS

UNIT IV - BIVARIATE ANALYSIS

Relationships between Two Variables – Percentage Tables – Analyzing Contingency Tables – Handling Several Batches – Scatterplots and Resistant Lines.

Introduction

Data in statistics is sometimes classified according to the number of variables. For example, "height" might be one variable and "weight" might be another variable. Depending on the number of variables, the data is classified as univariate, Bivariate, Multivariate.

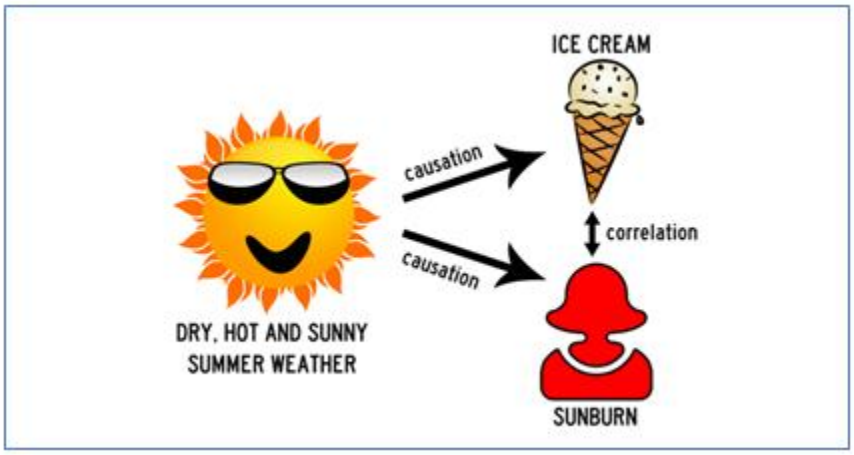
- Univariate analysis -- Analysis of one ("uni") variable.
- Bivariate analysis -- Analysis of ***exactly*** two variables.
- Multivariate analysis -- Analysis of more than two variables.

Relationships between Two Variables

Bivariate Analysis – Definition:

- Bivariate Analysis is a type of statistical analysis where two variables are observed. In this analysis, one variable is dependent and the other is independent. These two variables are mostly denoted by X and Y. Hence it is denoted as pair (X, Y).
- It is mainly used to examine the relationship between two variables. It helps in understanding how the variables are related and if there is any association or correlation between them.

  Example: Hours spent watching TV Vs time spent in doing physical exercise.
- Techniques used in bivariate analysis are:
  - Correlation analysis
  - Scatter plots
  - Regression analysis



Relationship between two variables

In statistical analysis, the terms "explanatory variable" and "response variable" are commonly used to describe the relationship between two variables in a study.

Explanatory Variable (Independent Variable)

The explanatory variable, also known as the independent variable or predictor variable, is the variable that is manipulated or controlled by the researcher. It is believed to have an influence on the response variable.

For example, in a study investigating the effect of studying time on exam scores, the amount of time spent studying would be the explanatory variable. Students study for different durations (e.g., 1 hour, 2 hours, or 3 hours).

Response Variable (Dependent Variable)

The response variable, also known as the dependent variable or outcome variable, is the variable that is observed or measured in response to changes in the explanatory variable. Its value depends on the values of the explanatory variable.

In the same example of studying time and exam scores, the exam scores would be the response variable.

The relationship between the explanatory variable and the response variable is typically analyzed, how changes in the explanatory variable affect the response variable. Statistical techniques like regression analysis, ANOVA, or correlation analysis are commonly used to quantify and analyze this relationship.

Techniques used in Bivariate Analysis

Scatter plots

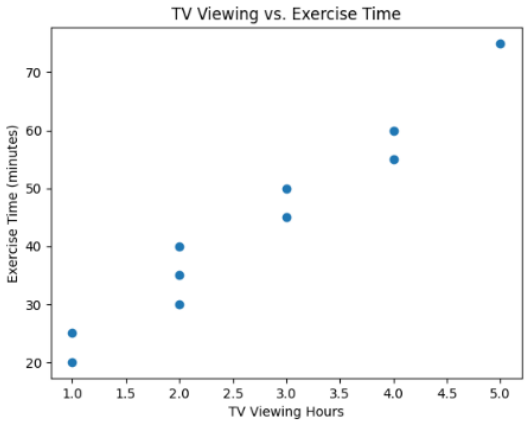
A scatter plot is a graphical representation of data points on a Cartesian plane. It shows the values of two variables as points on the graph, with one variable represented on the x-axis and the other on the y-axis. Scatter plots helps to visualize the relationship between the variables. The pattern formed by the data points can provide insights into the type and strength of the relationship.

Suppose you collect data from a group of individuals, recording the number of hours they spend watching TV (x-axis) and the amount of time they spend on physical exercise (y-axis) in a week. By plotting these data points on a scatter plot, you can visually analyze the relationship between the two variables. If the data points show a cluster towards the lower end of physical exercise time, it suggests a potential negative relationship. This implies that individuals who spend more time watching TV tend to engage in less physical exercise.

Simple Python program for scatter plot

```
import matplotlib.pyplot as plt

# Data collection
# Example TV viewing hours
tv_hours = [2, 3, 1, 4, 5, 2, 1, 3, 4, 2]

# Example exercise time
exercise_time = [30, 45, 20, 60, 75, 40, 25, 50, 55, 35]

# Creating a scatter plot
plt.scatter(tv_hours, exercise_time)
plt.xlabel("TV Viewing Hours")
plt.ylabel("Exercise Time (minutes)")
plt.title("TV Viewing vs. Exercise Time")
plt.show()
```

Output



Correlation analysis

Correlation analysis measures the strength and direction of the linear relationship between two variables. It provides a numerical value, called a correlation coefficient, which quantifies the relationship. The correlation coefficient ranges from -1 to +1. A positive correlation coefficient indicates a positive relationship, a negative correlation coefficient indicates a negative relationship, and a correlation coefficient of zero indicates no linear relationship.

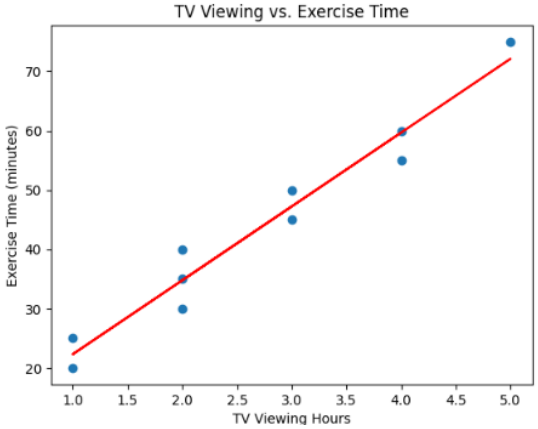
Let's say the correlation coefficient between TV viewing hours and physical exercise time is -0.6. This negative value indicates a moderate negative correlation. It suggests that as TV viewing hours increase, there tends to be a decrease in the amount of time spent on physical exercise. However, it's important to remember that correlation does not imply causation. The negative correlation does not necessarily mean that watching TV causes a decrease in physical exercise, but rather that there is an association between the two variables.

Simple Python program to explain correlation analysis

```
import numpy as np
from scipy.stats import pearsonr, linregress
import matplotlib.pyplot as plt

# Data collection
tv_hours = [2, 3, 1, 4, 5, 2, 1, 3, 4, 2]  # Example TV viewing hours
exercise_time = [30, 45, 20, 60, 75, 40, 25, 50, 55, 35]  # Example exercise
time

# Bivariate analysis
correlation_coefficient, _ = pearsonr(tv_hours, exercise_time)
slope, intercept, r_value, p_value, std_err = linregress(tv_hours,
exercise_time)

# Creating a scatter plot
```

```
plt.scatter(tv_hours, exercise_time)
plt.xlabel("TV Viewing Hours")
plt.ylabel("Exercise Time (minutes)")
plt.title("TV Viewing vs. Exercise Time")

# Regression line
x = np.array(tv_hours)
y = slope * x + intercept
plt.plot(x, y, color='red')

# Display the plot with regression line
plt.show()

# Results
print("Correlation coefficient:", correlation_coefficient)
print("Slope:", slope)
print("Intercept:", intercept)
print("R-value:", r_value)
print("P-value:", p_value)
print("Standard error:", std_err)
```

Output




```
Correlation coefficient: 0.9795036751931497
Slope: 12.45341614906832
Intercept: 9.875776397515537
R-value: 0.9795036751931494
P-value: 7.532861648764562e-07
Standard error: 0.9054273128641811
```

Regression analysis

Regression analysis is used to model the relationship between two variables. It helps predict the value of one variable based on the known value of another variable. In bivariate analysis, simple linear regression is commonly used. It assumes a linear relationship between the variables and estimates the best-fit line that minimizes the distance between the observed data points and the predicted values.

Suppose the regression analysis suggests a simple linear equation like "Physical Exercise Time = 150 - 10 * TV Viewing Hours." This equation implies that for each

additional hour spent watching TV, the expected decrease in physical exercise time is 10 minutes. Using this equation, you can estimate the physical exercise time for an individual based on the number of hours they spend watching TV.

Simple Python program to explain regression analysis

```
import numpy as np
from scipy.stats import linregress

# Data collection
tv_hours = [2, 3, 1, 4, 5, 2, 1, 3, 4, 2]  # Example TV viewing hours
exercise_time = [30, 45, 20, 60, 75, 40, 25, 50, 55, 35]  # Example exercise time

# Regression analysis
slope, intercept, r_value, p_value, std_err = linregress(tv_hours, exercise_time)

# Predicting exercise time for new TV viewing hours
new_tv_hours = [2.5, 3.5, 4.5]  # Example new TV viewing hours
predicted_exercise_time = [slope * x + intercept for x in new_tv_hours]

# Results
print("Slope:", slope)
print("Intercept:", intercept)
print("R-value:", r_value)
print("P-value:", p_value)
print("Standard error:", std_err)

# Predicted exercise time for new TV viewing hours
print("Predicted Exercise Time for New TV Viewing Hours:")
for i in range(len(new_tv_hours)):
    print(f"TV Viewing Hours: {new_tv_hours[i]}, Predicted Exercise Time: {predicted_exercise_time[i]}")
```

Output

```
Slope: 12.45341614906832
Intercept: 9.875776397515537
R-value: 0.9795036751931494
P-value: 7.532861648764562e-07
Standard error: 0.9054273128641811
Predicted Exercise Time for New TV Viewing Hours:
TV Viewing Hours: 2.5, Predicted Exercise Time: 41.00931677018633
TV Viewing Hours: 3.5, Predicted Exercise Time: 53.462732919254655
TV Viewing Hours: 4.5, Predicted Exercise Time: 65.91614906832297
```

Working with SPSS

Syntax

```
GET DATA
 /TYPE=XLS
 /FILE='C:\Users\lenovo\Desktop\cancer.xls'
 /SHEET=name 'Cancer'
 /CELLRANGE=full
```

```
  /READNAMES=on
  /ASSUMEDSTRWIDTH=32767.
EXECUTE.
DATASET NAME DataSet1 WINDOW=FRONT.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT AGE
  /METHOD=ENTER WEIGHIN
  /RESIDUALS NORMPROB(ZRESID).
```

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	WEIGHIN[b]	.	Enter

a. Dependent Variable: AGE

b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.288[a]	.083	.043	12.652

a. Predictors: (Constant), WEIGHIN

b. Dependent Variable: AGE

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	331.963	1	331.963	2.074	.163[b]
	Residual	3681.797	23	160.078		
	Total	4013.760	24			

a. Dependent Variable: AGE

b. Predictors: (Constant), WEIGHIN

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	80.375	14.620		5.498	.000
	WEIGHIN	-.116	.081	-.288	-1.440	.163

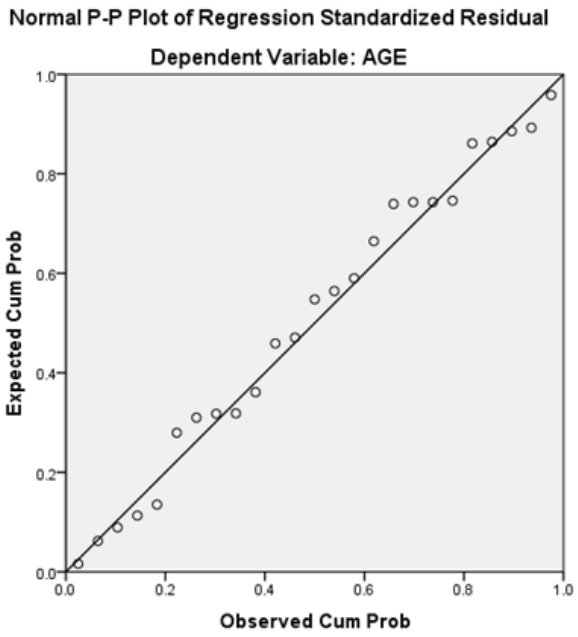
a. Dependent Variable: AGE

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	49.97	65.95	59.64	3.719	25
Residual	-27.160	21.908	.000	12.386	25
Std. Predicted Value	-2.599	1.697	.000	1.000	25
Std. Residual	-2.147	1.732	.000	.979	25

a. Dependent Variable: AGE

Charts


Normal P-P Plot of Regression Standardized Residual
Dependent Variable: AGE
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0

Proportions, percentages and probabilities

The relationship between two variables can be examined in terms of proportions, percentages, and probabilities. These measures provide insights into the distribution and likelihood of events occurring. Let's explore how these concepts relate to the relationship between two variables.

Proportions

Proportions represent the relative size or share of one category within a total. When examining the relationship between two variables, you can calculate proportions based on the occurrence or absence of specific events within each variable.

For example, consider a dataset of students and their preferred subjects: Math and Science. Calculate the proportion of students who prefer Math and the proportion of students who prefer Science. These proportions reflect the distribution of preferences within the dataset.

Percentages

Percentages represent proportions expressed as a fraction of 100. They provide a way to compare proportions on a standardized scale.

In the previous example, if 60 out of 100 students prefer Math, the proportion of students preferring Math is 60/100 or 0.6. The percentage of students preferring Math is 0.6 multiplied by 100, which equals 60%. Similarly, you can calculate the percentage of students preferring Science.

Probabilities

Probabilities quantify the likelihood of an event occurring. In the context of two variables, probabilities can be used to understand the likelihood of events happening simultaneously or independently based on the relationship between the variables. Probabilities can help determine the chances of specific outcomes or events occurring, considering the relationship between the two variables.

Example 1

To demonstrate how proportions and percentages can be calculated from a students' database, let's consider a simple example. Assume we have a database containing information about students' favorite subjects: Math, Science, and English. We want to calculate the proportion and percentage of students who prefer each subject. Here's an example database:

Student ID	Favorite Subject
1	Math
2	Science
3	Math

4	English
5	Math
6	Science
7	Science
8	English
9	English
10	Math

Solution

To calculate the proportions and percentages, we need to determine the number of students who prefer each subject and divide it by the total number of students.

Step 1: Count the number of students who prefer each subject:

Math: 4 students

Science: 3 students

English: 3 students

Step 2: Calculate the proportions:

Proportion of students who prefer Math: 4/10 = 0.4

Proportion of students who prefer Science: 3/10 = 0.3

Proportion of students who prefer English: 3/10 = 0.3

Step 3: Calculate the percentages:

Percentage of students who prefer Math: 0.4 * 100 = 40%

Percentage of students who prefer Science: 0.3 * 100 = 30%

Percentage of students who prefer English: 0.3 * 100 = 30%

So, in this example, 40% of the students prefer Math, 30% prefer Science, and 30% prefer English.

By calculating proportions and percentages, we can gain insights into the distribution of favorite subjects among students in the database. This information can help in understanding the preferences and patterns within the student population.

Example 2

Let's work through a solved example of analyzing a frequency table related to gender (male/female) and job satisfaction (satisfied/dissatisfied) data. We collected data on job satisfaction from a sample of 200 employees.

Calculating the Frequency table

Step 1:

Here is the frequency table:

Frequency table			
		Job Satisfaction	
		Satisfied	Dissatisfied
Gender	Male	80	30
	Female	70	20

Step 2: Calculate the totals

Analyzing the table, we find 80 males are satisfied with their job, 30 males are dissatisfied, 70 females are satisfied, and 20 females are dissatisfied.

- *Calculate the row totals*. The row totals represent the total number of employees for each gender.
  Row total for Male = 80 + 30 = 110
  Row total for Female = 70 + 20 = 90
- *Calculate column totals:* The column totals represent the total number of employees for each job satisfaction level.
  Column total for Male = 80 + 70 = 150
  Column total for Female = 30 + 20 = 50

Step 3: Calculate percentages or proportions

Convert the frequencies to percentages or proportions to analyze the relative distribution of job satisfaction levels among male and female employees.

Proportion of males who are satisfied is 80/110 = 0.73 or 73%,

Proportion of females who are satisfied is 70/90 = 0.78 or 78%.

Step 4: Interpret the results

Analyze the frequencies, row totals, column totals, and percentages to gain insights into the distribution of job satisfaction levels among male and female employees.

- In this example, it appears that a higher proportion of females (78%) are satisfied compared to males (73%).

Calculate the probability

Calculate the probability of job satisfaction for males

$$Probability\ of\ Male\ being\ Satisfied = \frac{Frequency\ of\ Male\ Satisfied}{Total\ number\ of\ observations}$$

= 80 / (80 + 30 + 70 + 20)
= 80 / 200
= 0.4 (or 40%)

Calculate the probability of job dissatisfaction for males

$$Probability\ of\ Male\ being\ dissatisfied = \frac{Frequency\ of\ Male\ dissatisfied}{Total\ number\ of\ observations}$$
= 30 / (80 + 30 + 70 + 20)
= 30 / 200
= 0.15 (or 15%)

Calculate the probability of job satisfaction for females

$$Probability\ of\ Female\ being\ Satisfied = \frac{Frequency\ of\ Female\ Satisfied}{Total\ number\ of\ observations}$$
= 70 / (80 + 30 + 70 + 20)
= 70 / 200
= 0.35 (or 35%)

Calculate the probability of job dissatisfaction for females

$$Probability\ of\ Female\ being\ dissatisfied = \frac{Frequency\ of\ Female\ dissatisfied}{Total\ number\ of\ observations}$$
= 20 / (80 + 30 + 70 + 20)
= 20 / 200
= 0.1 (or 10%)

Analysis

By calculating the probabilities, the probability of job satisfaction is higher for females (35%) compared to males (40%), while the probability of job dissatisfaction is higher for males (15%) compared to females (10%).

Working with SPSS

Frequency, percentage & proportions

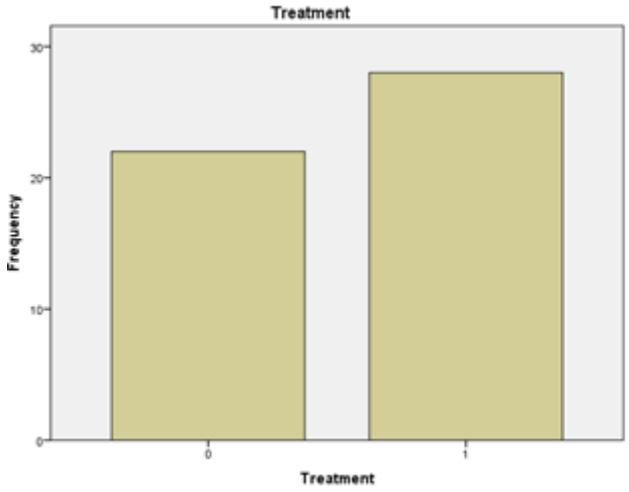
Syntax

```
GET DATA /TYPE=XLSX
  /FILE='C:\Users\lenovo\Desktop\newdrug(excel2007).xlsx'
  /SHEET=name 'Sheet1'
  /CELLRANGE=full
  /READNAMES=on
  /ASSUMEDSTRWIDTH=32767.
EXECUTE.
DATASET NAME DataSet5 WINDOW=FRONT.
FREQUENCIES VARIABLES=Treatment
  /BARCHART FREQ
  /ORDER=ANALYSIS.
```

Output

Treatment

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	22	44.0	44.0	44.0
	1	28	56.0	56.0	100.0
	Total	50	100.0	100.0	



Syntax

```
FREQUENCIES VARIABLES=Age Treatment
  /STATISTICS=STDDEV VARIANCE RANGE MINIMUM MAXIMUM SEMEAN
  /HISTOGRAM
  /ORDER=ANALYSIS.
```

Frequencies

Statistics

		Age	Treatment
N	Valid	50	50
	Missing	0	0
Std. Error of Mean		.920	.071
Std. Deviation		6.507	.501
Variance		42.336	.251
Range		30	1
Minimum		45	0
Maximum		75	1

Frequency Table

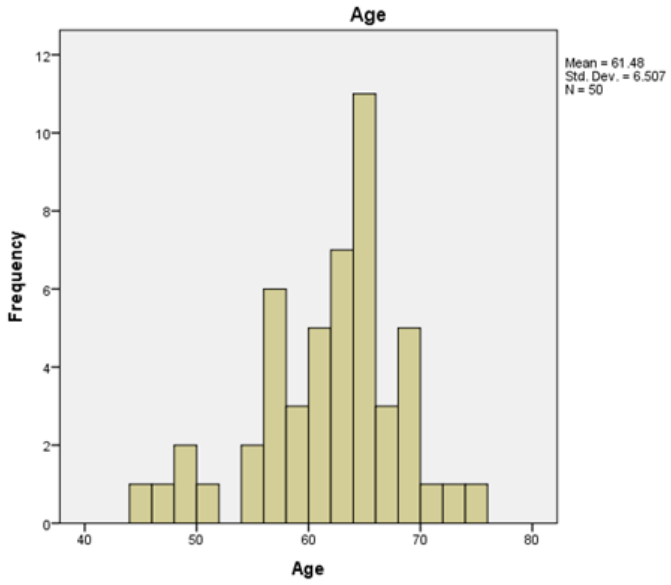
Age

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	45	1	2.0	2.0	2.0
	46	1	2.0	2.0	4.0
	48	1	2.0	2.0	6.0
	49	1	2.0	2.0	8.0
	51	1	2.0	2.0	10.0
	54	2	4.0	4.0	14.0
	56	3	6.0	6.0	20.0
	57	3	6.0	6.0	26.0
	58	1	2.0	2.0	28.0
	59	2	4.0	4.0	32.0
	60	1	2.0	2.0	34.0
	61	4	8.0	8.0	42.0
	62	3	6.0	6.0	48.0
	63	4	8.0	8.0	56.0
	64	4	8.0	8.0	64.0
	65	7	14.0	14.0	78.0
	66	3	6.0	6.0	84.0
	68	2	4.0	4.0	88.0
	69	3	6.0	6.0	94.0
	70	1	2.0	2.0	96.0
	73	1	2.0	2.0	98.0
	75	1	2.0	2.0	100.0
	Total	50	100.0	100.0	

Treatment

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	22	44.0	44.0	44.0
	1	28	56.0	56.0	100.0
	Total	50	100.0	100.0	

Histogram



Analyzing contingency tables

Contingency tables, also known as cross-tabulation tables or two-way frequency tables, are used to analyze the relationship between two categorical variables. They helps us to summarize and visualize the data, and helps in finding any associations or dependencies between the variables.

Step-by-step guide to analyze contingency tables

Step 1: Understand the Variables: Determine the two categorical variables involved in the contingency table. Identify the categories or levels of each variable.

- Suppose you are studying the relationship between gender (male/female) and job satisfaction (satisfied/dissatisfied) among employees in a company.
    - Here, the two categorical variables are gender (with two levels: male and female) and job satisfaction (with two levels: satisfied and dissatisfied).

Step 2: Create the Contingency Table: Create a table with rows representing one variable and columns representing the other variable. Each cell in the table will contain the frequency or count of observations falling into that particular combination of categories.

- You collect data from a sample of 200 employees and create a contingency table to analyze the relationship.
- Construct the table with gender as rows and job satisfaction as columns.

Contingency table			
		Job Satisfaction	
		Satisfied	Dissatisfied
Gender	Male	60	40
	Female	70	30

Step 3: Calculate Row and Column Totals: Add row and column totals to the contingency table

Contingency table				
		Job Satisfaction		**TOTAL**
		Satisfied	Dissatisfied	
Gender	Male	60	40	100
	Female	70	30	100
TOTAL		130	70	**200**

Step 4: Calculate Expected Frequencies (Optional): If you want to assess whether the observed frequencies deviate significantly from what would be expected, you can calculate the expected frequencies.

$$Expected\ Frequency = \frac{Row\ total * Column\ Total}{Total\ number\ of\ Observations}$$

- Expected frequency for the "Male-Satisfied" cell = (100 × 130) / 200 = 65.
- Expected Frequency for "Male-Dissatisfied" cell = (100 × 70) / 200 = 35.
- Expected frequency for the "Female-Satisfied" cell = (100 × 130) / 200 = 65.
- Expected Frequency for "Female-Dissatisfied" cell = (100 × 70) / 200 = 35.

Step 5: Interpret the Observed Frequencies: Examine the observed frequencies in the contingency table. Look for any patterns or differences between the levels of the variables, as they may indicate a significant relationship.

- In our example, we can observe that there are 60 males who are satisfied with their job and 70 females who are satisfied. Similarly, there are 40 males and 30 females who are dissatisfied.

Step 6: Calculate Percentages or Proportions: Convert the frequencies in the contingency table to percentages or proportions. This helps in understanding the relative distribution of observations across the categories and facilitates comparison between the levels of the variables.

- Here, the proportion of males satisfied with their job is 60/100 = 0.6 or 60%, while the proportion of females satisfied is 70/100 = 0.7 or 70%.

Step 7: Conduct Chi-Square Test (Optional):
Chi-square test

- The chi-square test is commonly used to analyze contingency tables and determine if there is a significant association between two categorical variables. The test compares the observed frequencies in the contingency table with the expected frequencies.
- The chi-square test assesses whether the observed frequencies differ significantly from the expected frequencies. It provides a statistical measure of the association between the variables.

$$X^2 = \sum \frac{\text{(Observed value - Expected value)}^2}{\text{Expected value}}$$

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

- Here's how you can conduct a chi-square test using a contingency table:

Intervention	Outcome	Observed (O)	Expected (E)	(O-E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
Male	Satisfied	60	65	-5	25	0.38
	Unsatisfied	40	35	5	25	0.71
Female	Satisfied	70	65	5	25	0.38
	Unsatisfied	30	35	-5	25	0.71
					TOTAL – X^2	**2.20**

- Calculate the degrees of freedom:
  Since there are two intervention groups (male, female) and two outcome groups (satisfied, unsatisfied) there are (2 – 1) * (2 – 1) = 1 degrees of freedom.
  DF = (Number of rows - 1) × (Number of columns - 1) = (2 - 1) × (2 - 1) = 1

Step 8: Interpret the Results

For a test of significance at α = .05 and df = 1, the X^2 critical value is 3.841

Critical values of chi-square (right tail)

Degrees of freedom (*df*)	Significance level (α)							
	.99	.975	.95	.9	.1	.05	.025	.01
1	--------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277

- Compare the chi-square statistic with the critical value or p-value:
  X^2 = 2.20
  Critical value = 3.841
  The X^2 value is lesser than the critical value.

- If the chi-square statistic is greater than the critical value or the p-value is less than the significance level ($p < \alpha$), reject the null hypothesis. This indicates that there is evidence of an association between gender and job satisfaction.
- If the chi-square statistic is smaller than the critical value or the p-value is greater than the significance level ($p > \alpha$), fail to reject the null hypothesis.

Step 9: Visualize the Contingency Table: To further explore the relationship between the variables, you can create visual representations of the contingency table, such as stacked bar charts or heatmaps.

Contingency Table using SPSS

Syntax

```
CROSSTABS
  /TABLES=Treatment BY Gender
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT
  /COUNT ROUND CELL.
```

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Treatment * Gender	50	100.0%	0	0.0%	50	100.0%

Treatment * Gender Crosstabulation

Count

		Gender		Total
		F	M	
Treatment	0	11	11	22
	1	15	13	28
Total		26	24	50

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.063[a]	1	.802		
Continuity Correction[b]	.000	1	1.000		
Likelihood Ratio	.063	1	.802		
Fisher's Exact Test				1.000	.513
N of Valid Cases	50				

- 0 cells (0.0%) have expected count less than 5. The minimum expected count is 10.56.
- Computed only for a 2x2 table

Recoding, reordering, and collapsing categorical variables

Recoding, reordering, and collapsing categorical variables are common data manipulation techniques used to transform and organize categorical variables in a dataset.

Recoding Categorical Variables:

Recoding involves changing the values of a categorical variable to new values based on certain criteria or rules. This technique is useful when you want to group similar categories together or create new categories. For example, suppose you have a categorical variable "Education" with categories "High School," "Bachelor's Degree," and "Master's Degree." You may want to recode these categories as "High School or Less," "Bachelor's Degree," and "Advanced Degree" to create broader categories. Recoding can be performed manually or using software like SPSS, R, or Python.

In SPSS, there are three basic options for recoding variables:

- Recode into Different Variables
- Recode into Same Variables
- DO IF syntax

Recode into Different Variables and DO IF syntax

- Create a new variable without modifying the original variable

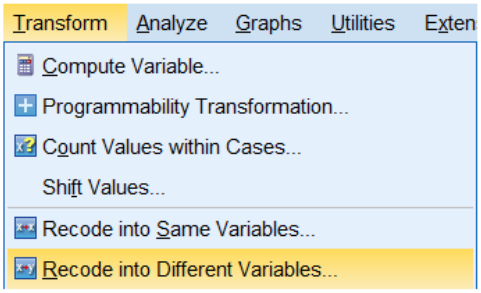
Recode into Same Variables

- Permanently overwrite the original variable.

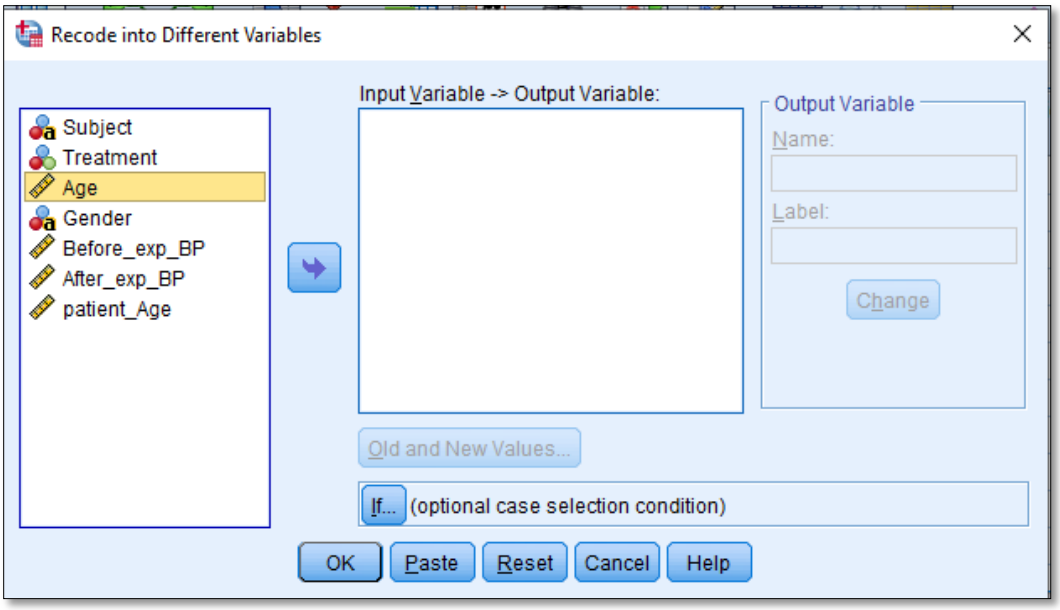
Recode into Different Variables

Recoding into a different variable transforms an original variable into a new variable. That is, the changes do not overwrite the original variable; instead changes are applied to a copy of the original variable under a new name.

To recode into different variables, click Transform > Recode into Different Variables.

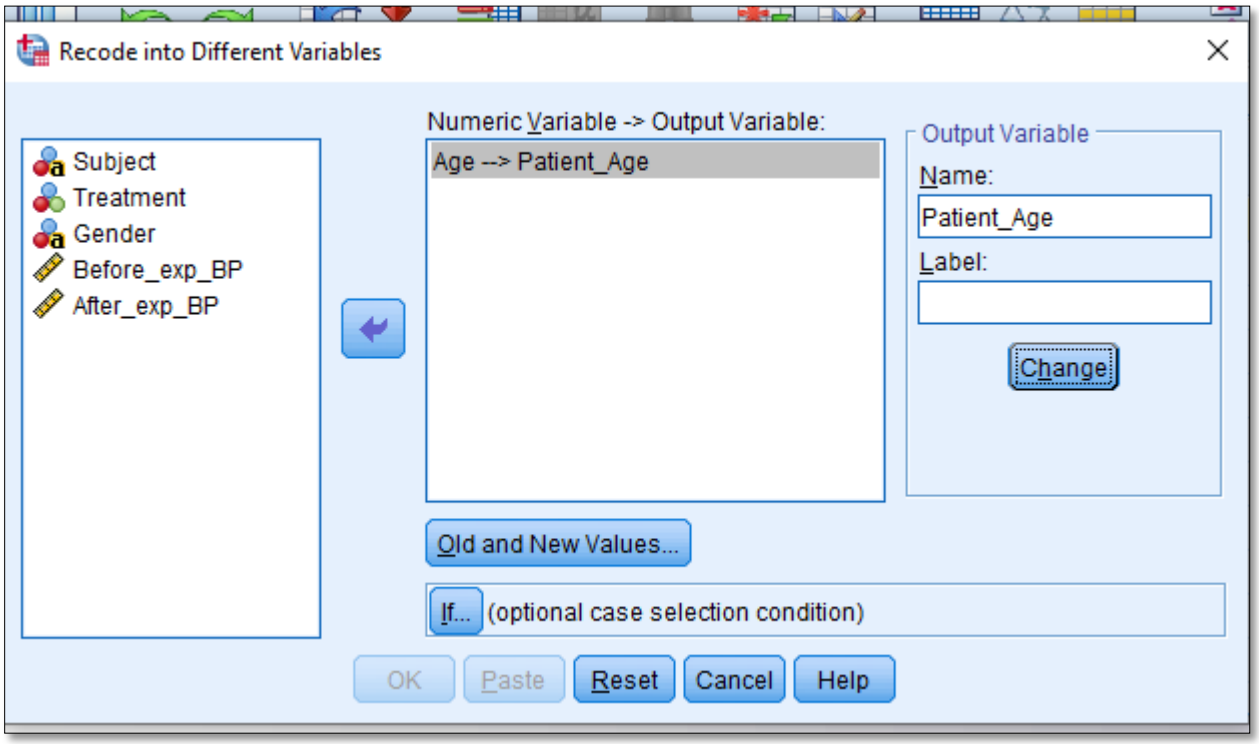


The Recode into Different Variables window will appear.



Input Variable -> Output Variable: The center text box lists the variable(s) you have selected to recode, as well as the name your new variable(s) will have after the recode

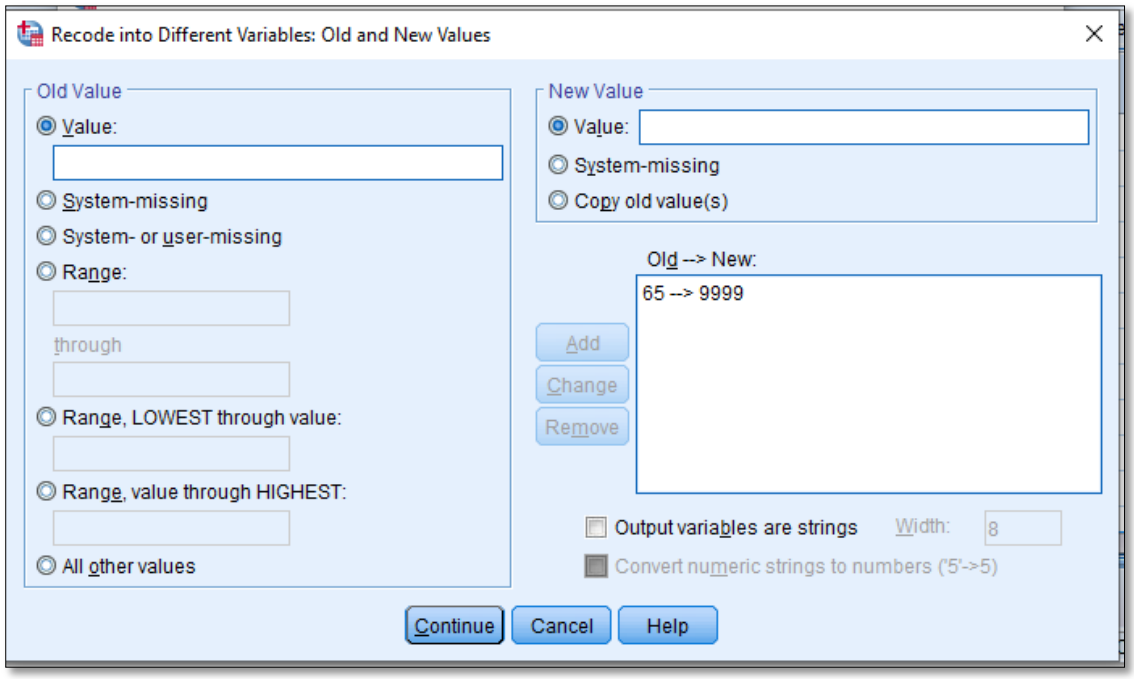
Output Variable: Define the name and label for your recoded variable(s) by typing them in the text fields. Once you are finished, click Change. Here, we have changed as "Age --> Patient_Age").



Old And New Values

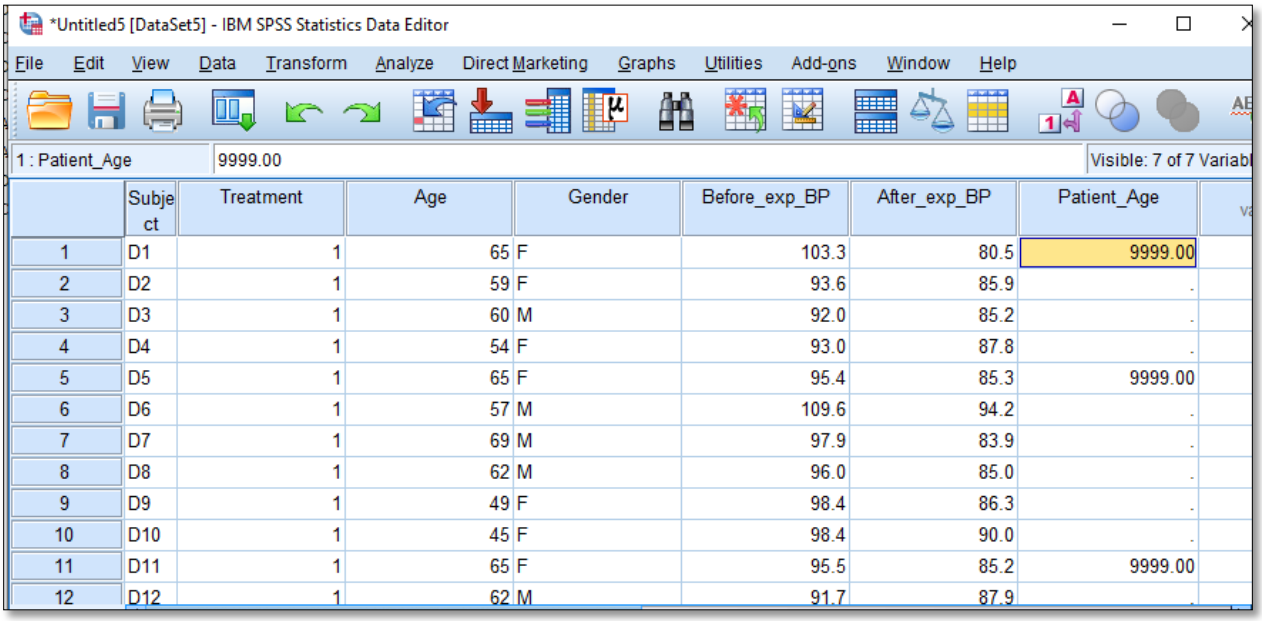
Once you click Old and New Values, a new window where you will specify how to transform the values will appear.

- Old Value: Specify the type of value you wish to recode
- New Value: Specify the new value for your variable



Syntax

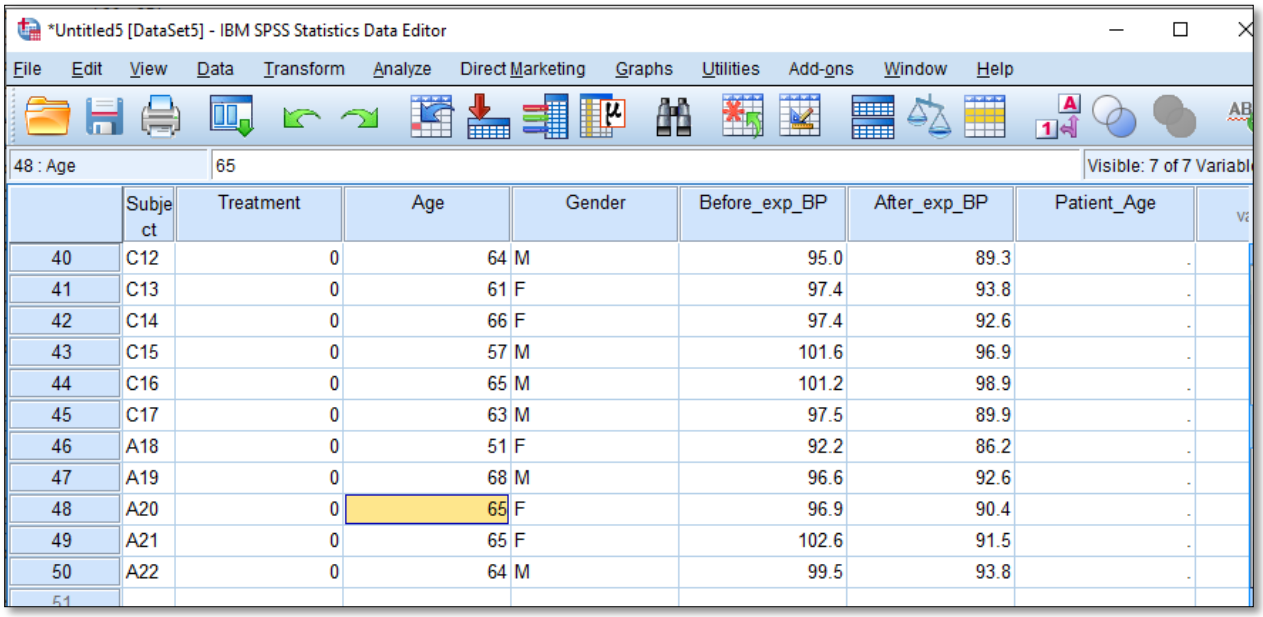
```
RECODE Age (65=9999) INTO Patient_Age.
EXECUTE.
```



	Subject	Treatment	Age	Gender	Before_exp_BP	After_exp_BP	Patient_Age
1	D1	1	65	F	103.3	80.5	9999.00
2	D2	1	59	F	93.6	85.9	.
3	D3	1	60	M	92.0	85.2	.
4	D4	1	54	F	93.0	87.8	.
5	D5	1	65	F	95.4	85.3	9999.00
6	D6	1	57	M	109.6	94.2	.
7	D7	1	69	M	97.9	83.9	.
8	D8	1	62	M	96.0	85.0	.
9	D9	1	49	F	98.4	86.3	.
10	D10	1	45	F	98.4	90.0	.
11	D11	1	65	F	95.5	85.2	9999.00
12	D12	1	62	M	91.7	87.9	.

Syntax

```
DO IF  (Treatment= 1).
RECODE Age (65=8888) INTO Patient_Age.
END IF.
EXECUTE.
```



	Subject	Treatment	Age	Gender	Before_exp_BP	After_exp_BP	Patient_Age
40	C12	0	64	M	95.0	89.3	.
41	C13	0	61	F	97.4	93.8	.
42	C14	0	66	F	97.4	92.6	.
43	C15	0	57	M	101.6	96.9	.
44	C16	0	65	M	101.2	98.9	.
45	C17	0	63	M	97.5	89.9	.
46	A18	0	51	F	92.2	86.2	.
47	A19	0	68	M	96.6	92.6	.
48	A20	0	65	F	96.9	90.4	.
49	A21	0	65	F	102.6	91.5	.
50	A22	0	64	M	99.5	93.8	.

Reordering Categorical Variables:

Reordering categorical variables involves changing the order of categories to align with a specific criterion or logical sequence. This technique is useful when you want to present the categories in a specific order for analysis or visualization purposes. For instance, if you have a categorical variable "Income Level" with categories "Low," "Medium," and "High," you may want to reorder them as "Low," "Medium," and "High" for consistency or to reflect a natural order. Reordering can be done manually or using data analysis software.

Collapsing Categorical Variables:

Collapsing involves combining multiple categories of a categorical variable into a single category. This technique is used when you want to simplify the analysis or reduce the number of categories. For example, if you have a categorical variable "Age Group" with categories "18-24," "25-34," "35-44," and "45-54," you may want to collapse them into broader categories like "18-34" and "35-54" to have fewer groups for analysis. Collapsing can be done manually or programmatically in statistical software.

When applying these techniques, it is important to consider the context and objectives of the analysis. Be mindful of preserving the integrity of the data and ensure that the recoded, reordered, or collapsed categories accurately represent the underlying information.

Handling Several Batches

Boxplots are widely used in data exploration to visualize the distribution and statistical properties of a dataset. They provide a concise summary of the data's central tendency, dispersion, and skewness. Here's how boxplots are constructed and interpreted:

Construction of a boxplot

- The median (50th percentile) of the dataset is represented by a horizontal line inside a box.
- The box extends from the lower quartile (25th percentile) to the upper quartile (75th percentile), indicating the interquartile range (IQR).
- Whiskers, represented by vertical lines, extend from the box's edges to the furthest data points within a certain range. The exact range depends on the specific rules used for whisker calculation.
- Data points outside the whiskers are considered outliers and are typically plotted individually.

Interpretation of a boxplot

- *Center:* The median (line inside the box) represents the dataset's central tendency. If the median is closer to the bottom of the box, the data is skewed towards lower values, while if it's closer to the top, the data is skewed towards higher values.
- *Spread:* The length of the box (IQR) gives an indication of the data's dispersion. A longer box implies greater variability, while a shorter box indicates less variability.
- *Skewness:* The symmetry of the boxplot helps identify skewness in the data. If the median line is not in the center of the box, the data may be skewed.
- *Outliers:* Data points plotted individually outside the whiskers are considered outliers and may indicate unusual or extreme observations.
- Boxplots are particularly useful when comparing distributions between different groups or variables. By placing multiple boxplots side by side, you can visually compare their central tendencies, spreads, and skewness.

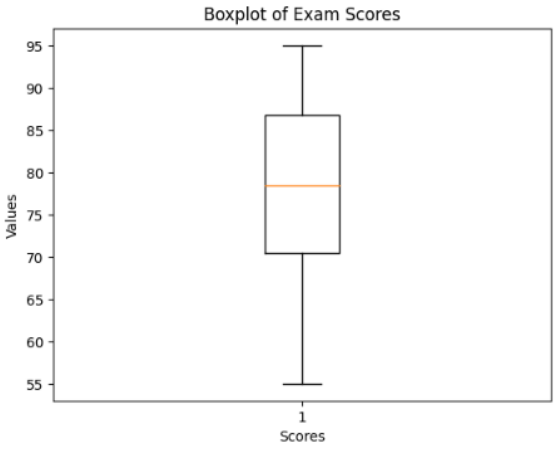
In summary, boxplots offer a visual summary of key statistical characteristics of a dataset, making them a valuable tool for data exploration and initial analysis.

Simple Python Program to display Boxplot

```
import matplotlib.pyplot as plt

# Example dataset
scores = [55, 58, 60, 62, 65, 66, 68,
          70, 72, 73, 74, 75, 76, 77,
          78, 79, 80, 81, 82, 83, 85,
```

```
          86, 87, 88, 89, 90, 91, 92, 93, 95]

# Creating the boxplot
plt.boxplot(scores)

# Adding labels and title
plt.xlabel('Scores')
plt.ylabel('Values')
plt.title('Boxplot of Exam Scores')

# Displaying the plot
plt.show()
```

Output



How to Create and Interpret Box Plots in SPSS

A box plot is used to visualize the five number summary of a dataset, which includes:

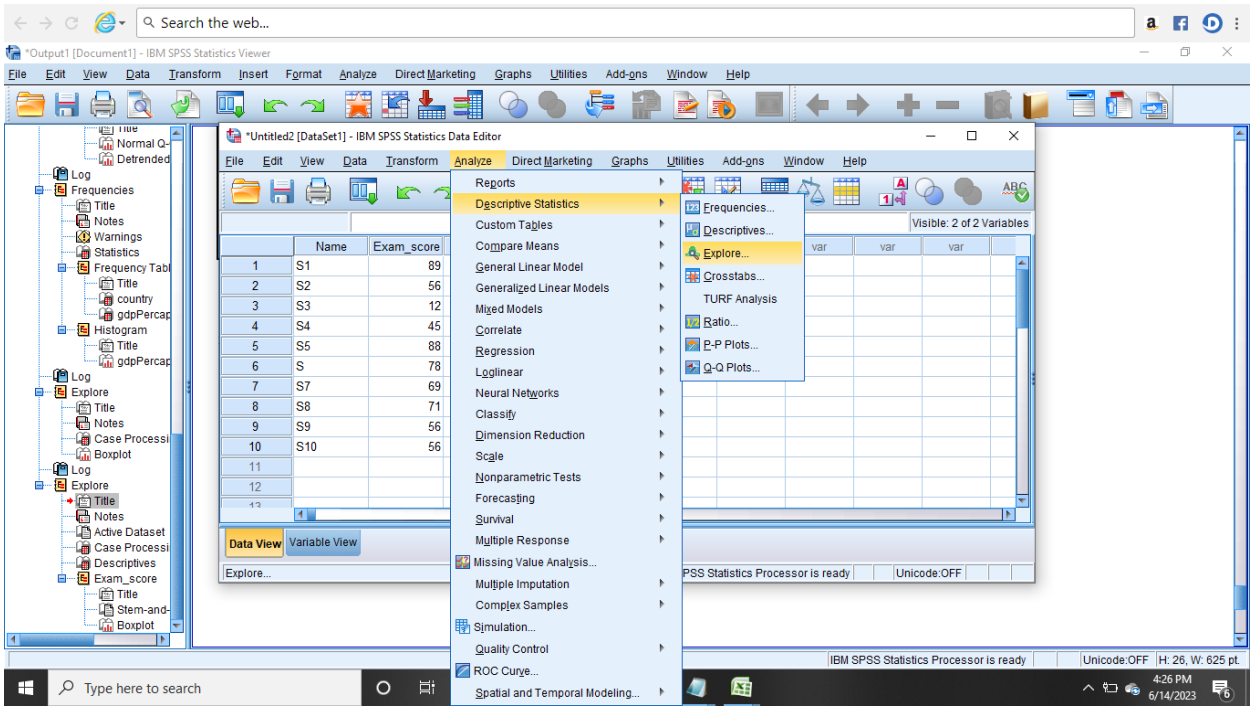
- The minimum
- The first quartile
- The median
- The third quartile
- The maximum

Creating a Single Box Plot in SPSS

Suppose we have the following dataset that shows the exam marks scored by 10 students:

	Name	Exam_score
1	S1	89
2	S2	56
3	S3	12
4	S4	45
5	S5	88
6	S	78
7	S7	69
8	S8	71
9	S9	56
10	S10	56

To create a box plot to visualize the distribution of these data values, we can click the Analyze tab, then Descriptive Statistics, then Explore:



To create a box plot, drag the variable points into the box labelled Dependent List. Then make sure Plots is selected under the option that says Display near the bottom of the box.

Case Processing Summary

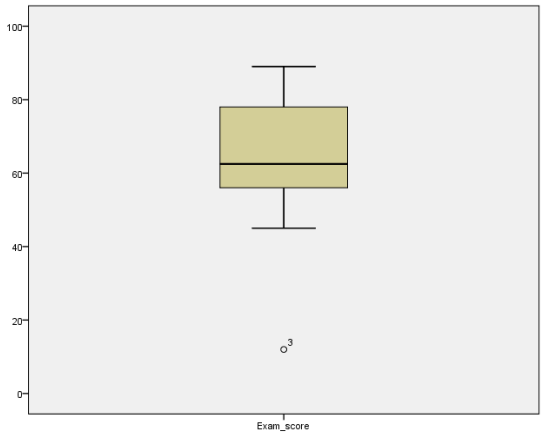
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Exam_score	10	100.0%	0	0.0%	10	100.0%

Descriptives

			Statistic	Std. Error
Exam_score	Mean		62.00	7.217
	95% Confidence Interval for Mean	Lower Bound	45.67	
		Upper Bound	78.33	
	5% Trimmed Mean		63.28	
	Median		62.50	
	Variance		520.889	
	Std. Deviation		22.823	
	Minimum		12	
	Maximum		89	
	Range		77	
	Interquartile Range		27	
	Skewness		-1.030	.687
	Kurtosis		1.614	1.334

Exam Score

```
Exam_score Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     1.00 Extremes    (=<12)
     1.00         4 .  5
     3.00         5 .  666
     1.00         6 .  9
     2.00         7 .  18
     2.00         8 .  89

 Stem width:        10
 Each leaf:        1 case(s)
```



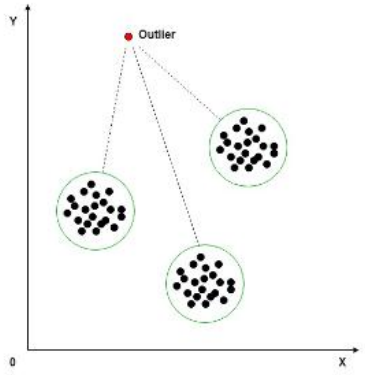
Outliers

In statistics, outliers are data points that deviate significantly from the rest of the dataset. These observations are considered to be unusual, extreme, or inconsistent with the overall pattern or distribution of the data.

3, 35, 37, 38, 40, 42, 44, 76

Outliers can occur due to various reasons, including measurement errors, data entry mistakes, natural variability, or genuinely unusual observations. They can have a significant impact on statistical analysis, as they can skew results and affect the validity of assumptions made about the data.

Identifying outliers is important because they can distort statistical measures such as the mean and standard deviation, as well as affect data modeling and analysis. Outliers can be detected through various methods, such as graphical exploration, statistical tests, or using domain knowledge.



Some commonly used techniques to identify outliers include:

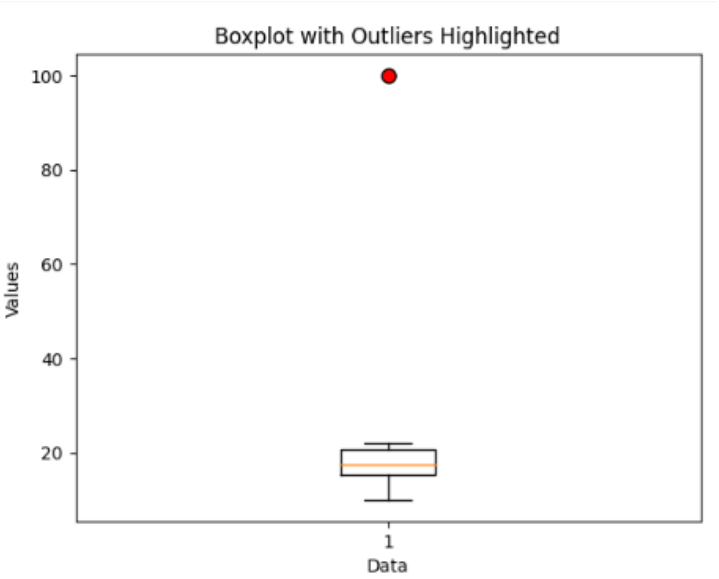
- *Visual inspection:* Plotting the data using graphs like scatter plots, boxplots, or histograms can reveal any data points that appear significantly different from the majority of the data.
- *Statistical tests:* Methods like the z-score or modified z-score can be used to identify observations that fall beyond a certain threshold of standard deviations from the mean. Values that exceed a specific cutoff (e.g., z-score greater than 3 or 3.5) can be considered outliers.
- *Interquartile Range (IQR) method:* This method uses the quartiles (Q1 and Q3) and the IQR to determine outliers. Data points below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR are considered outliers.
- *Domain knowledge:* Sometimes, outliers are known to exist based on the context or subject matter expertise. In such cases, domain knowledge can be used to identify and handle outliers appropriately.

Outliers can be treated by removing them, transforming them, or treating them as missing values, depending on the circumstances and impact on the analysis. In summary, outliers are data points that are significantly different from the rest of the dataset and can have an impact on statistical analysis. Detecting and managing outliers is crucial to ensure accurate and reliable results.

Simple Python Program to show outlier in Boxplot

```
import matplotlib.pyplot as plt

# Example dataset
data = [10, 12, 15, 16, 17, 18, 20, 21, 22, 100]

# Creating the boxplot
plt.boxplot(data, flierprops={'marker': 'o', 'markerfacecolor':
'red', 'markersize': 8})

# Adding labels and title
plt.xlabel('Data')
plt.ylabel('Values')
plt.title('Boxplot with Outliers Highlighted')

# Displaying the plot
plt.show()
```

Output



Note:

The flierprops parameter is used to customize the appearance of the outliers. In this example, we set the marker style to a red circle ('o'), the marker face color to red, and the marker size to 8.

Dependents

"Dependents together" means that all dependent variables are shown together in each boxplot. If you enter a factor -say, sex- you'll get a separate boxplot for each factor level -

female and male respondents. “Factor levels together” creates a separate boxplot for each dependent variable, showing all factor levels together in each boxplot.

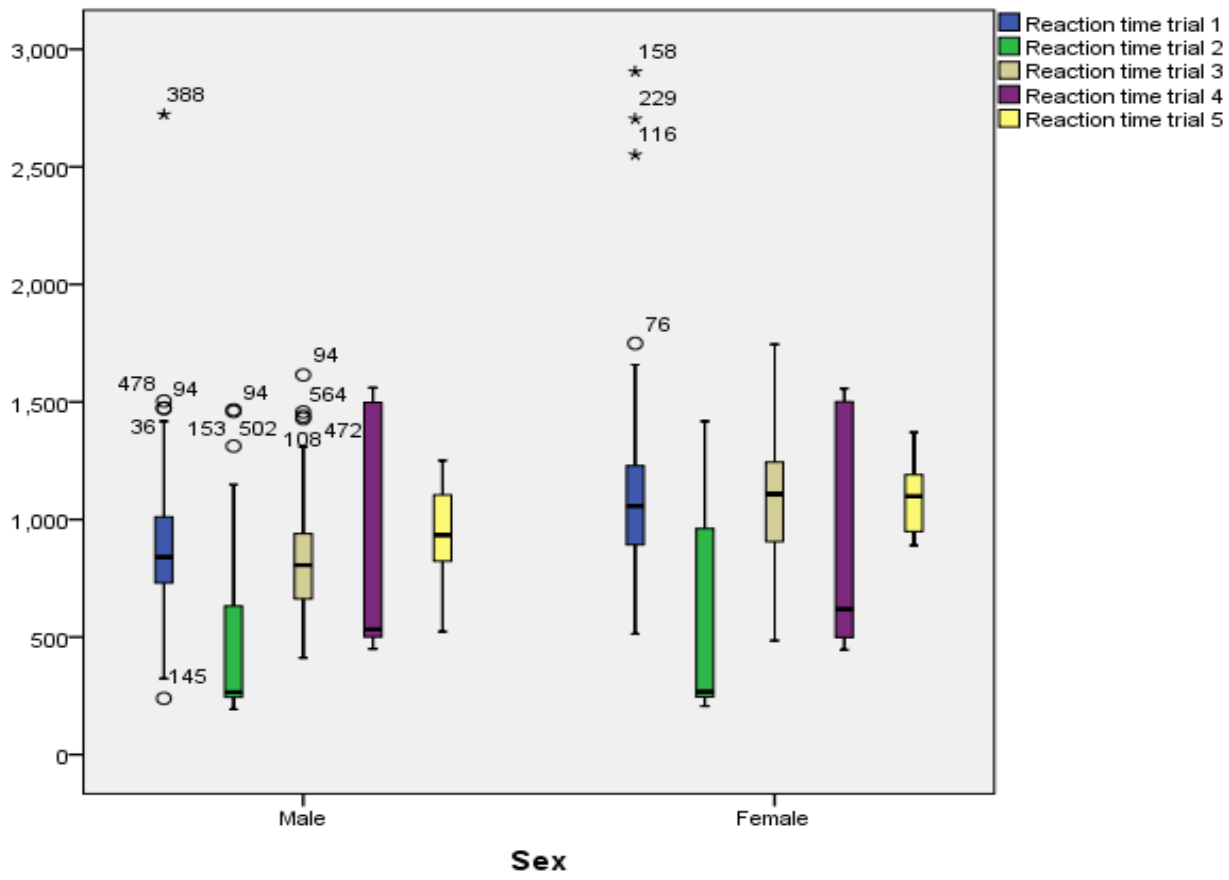
“Exclude cases pairwise” means that the results for each variable are based on all cases that don't have a missing value for that variable. “Exclude cases listwise” uses only cases without any missing values on all variables.

Syntax

```
EXAMINE VARIABLES=r01 r02 r03 r04 r05 BY sex
  /COMPARE VARIABLE
  /PLOT=BOXPLOT
  /STATISTICS=NONE
  /NOTOTAL
  /ID=id
  /MISSING=PAIRWISE.
```

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	Sex	N	Percent	N	Percent	N	Percent
Reaction time trial 1	Male	115	96.6%	4	3.4%	119	100.0%
	Female	118	99.2%	1	0.8%	119	100.0%
Reaction time trial 2	Male	117	98.3%	2	1.7%	119	100.0%
	Female	118	99.2%	1	0.8%	119	100.0%
Reaction time trial 3	Male	115	96.6%	4	3.4%	119	100.0%
	Female	118	99.2%	1	0.8%	119	100.0%
Reaction time trial 4	Male	119	100.0%	0	0.0%	119	100.0%
	Female	119	100.0%	0	0.0%	119	100.0%
Reaction time trial 5	Male	20	16.8%	99	83.2%	119	100.0%
	Female	13	10.9%	106	89.1%	119	100.0%



Case Processing Summary

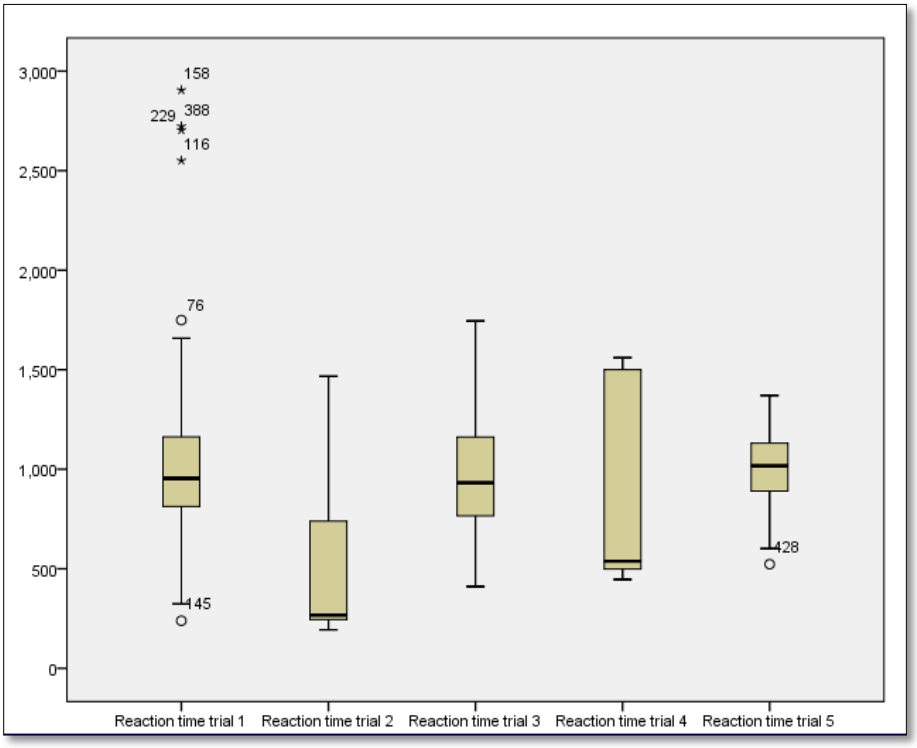
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Reaction time trial 1	233	97.9%	5	2.1%	238	100.0%
Reaction time trial 2	235	98.7%	3	1.3%	238	100.0%
Reaction time trial 3	233	97.9%	5	2.1%	238	100.0%
Reaction time trial 4	238	100.0%	0	0.0%	238	100.0%
Reaction time trial 5	33	13.9%	205	86.1%	238	100.0%

Descriptives

			Statistic	Std. Error
Reaction time trial 1	Mean		1001.57	22.372
	95% Confidence Interval for Mean	Lower Bound	957.49	
		Upper Bound	1045.64	
	5% Trimmed Mean		977.53	
	Median		954.00	
	Variance		116617.479	
	Std. Deviation		341.493	
	Minimum		239	
	Maximum		2905	
	Range		2666	
	Interquartile Range		353	
	Skewness		2.211	.159
	Kurtosis		9.654	.318
Reaction time trial 2	Mean		476.77	23.839
	95% Confidence Interval for Mean	Lower Bound	429.80	
		Upper Bound	523.73	
	5% Trimmed Mean		442.37	
	Median		267.00	
	Variance		133544.556	
	Std. Deviation		365.437	
	Minimum		193	
	Maximum		1467	
	Range		1274	
	Interquartile Range		498	
	Skewness		1.244	.159
	Kurtosis		.032	.316

Reaction time trial 1 Stem-and-Leaf Plot

Frequency	Stem &	Leaf
1.00	Extremes	(=<**)
2.00	3 .	27
1.00	4 .	2
7.00	5 .	0114566
15.00	6 .	111444456677999
29.00	7 .	00000111222333334444555677889
47.00	8 .	00111111222222233344444455555667777777788899999
27.00	9 .	000000112334555566777788899
33.00	10 .	000011111222333444555556788888999
25.00	11 .	011122224456666778888899
14.00	12 .	00012223356789
14.00	13 .	01222233445888
9.00	14 .	011357778
3.00	15 .	058
1.00	16 .	5
5.00	Extremes	(>=**)



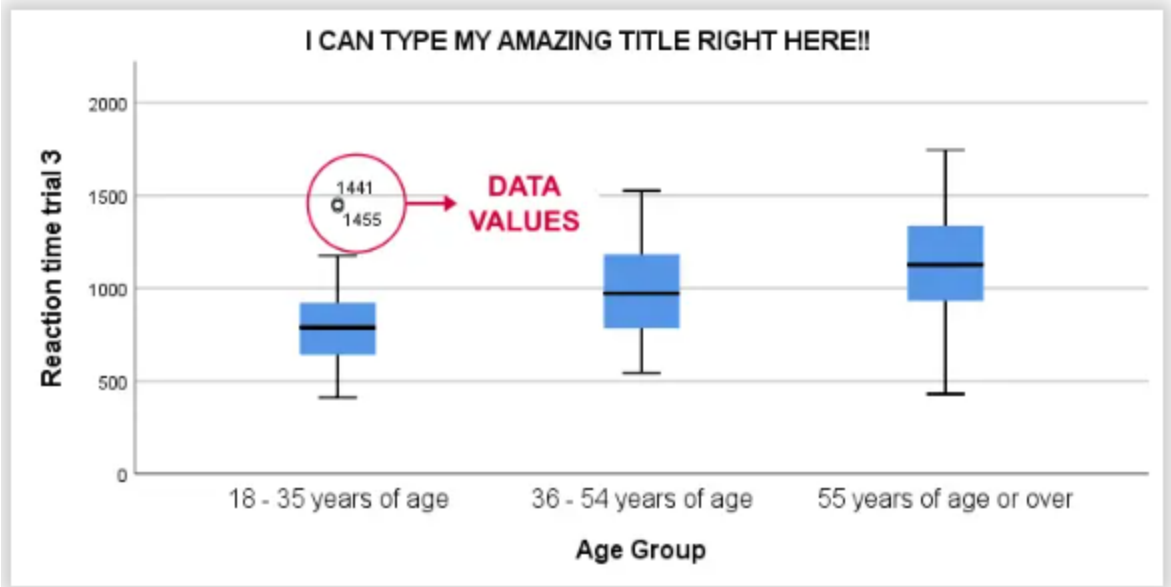
Case Processing Summary

	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Reaction time trial 1	233	97.9%	5	2.1%	238	100.0%
Reaction time trial 2	235	98.7%	3	1.3%	238	100.0%
Reaction time trial 3	233	97.9%	5	2.1%	238	100.0%
Reaction time trial 4	238	100.0%	0	0.0%	238	100.0%
Reaction time trial 5	33 !	13.9%	205 !	86.1%	238	100.0%

- The first columns tells how many cases were used for each variable.
- Note that trial 5 has N = 205 or 86.1% missing values.

<u>Boxplot for 1 Variable - Multiple Groups of Cases</u>

```
GGRAPH
/GRAPHDATASET NAME="graphdataset" VARIABLES=agegroup r03 MISSING=LISTWISE
REPORTMISSING=NO
/GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
SOURCE: s=userSource(id("graphdataset"))
DATA: agegroup=col(source(s), name("agegroup"), unit.category())
DATA: r03=col(source(s), name("r03"))
GUIDE: axis(dim(1), label("Age Group"))
GUIDE: axis(dim(2), label("Reaction time trial 3"))
GUIDE: text.title(label("I CAN TYPE MY AMAZING TITLE RIGHT HERE!"))
SCALE: cat(dim(1), include("1", "2", "3"))
SCALE: linear(dim(2), include(0))
ELEMENT: schema(position(bin.quantile.letter(agegroup*r03)), label(r03))
END GPL.
```



T - Test

A T-test is a statistical method of comparing the means or proportions of two samples gathered from either the same group or different categories. It is aimed at hypothesis testing, which is used to test a hypothesis pertaining to a given population. It is the difference between population means and a hypothesized value.

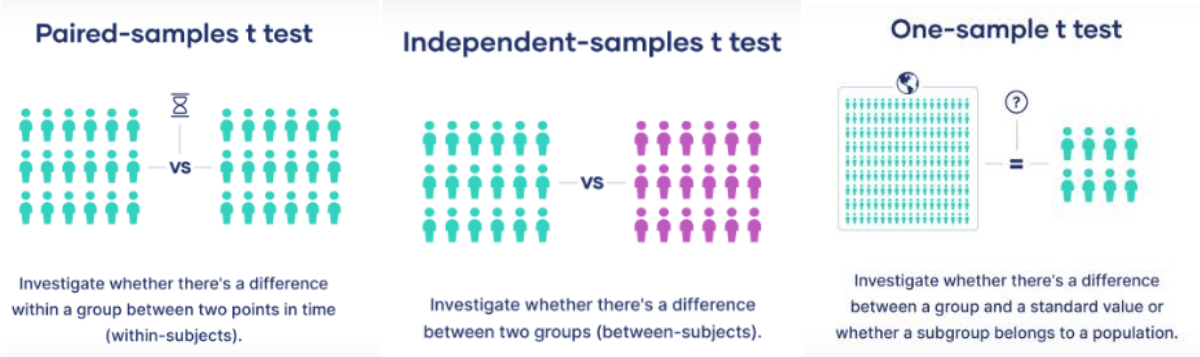
There are several types of t-tests, but the most commonly used ones are:

- Independent samples t-test: This test compares the means of two independent groups to determine if there is a significant difference between them. It assumes that the two groups are independent of each other.

- Paired samples t-test: This test compares the means of two related groups, where each observation in one group is paired with an observation in the other group. It is used when the same individuals or objects are measured before and after an intervention or treatment.

- One-sample t-test: This test compares the mean of a single group to a known population mean or a hypothesized value. It is used to determine if the observed mean significantly differs from a specific value.

One-sample, two-sample, paired, equal, and unequal variance are the types of T-tests users can use for mean comparisons.



Independent samples t tests

Independent samples t tests have the following hypotheses:

- Null hypothesis: The means for the two populations are equal.
- Alternative hypothesis: The means for the two populations are not equal.
    - If the p-value is less than your significance level (e.g., 0.05), you can reject the null hypothesis. The difference between the two means is statistically significant. Your sample provides strong enough evidence to conclude that the two population means are not equal.

Simple Example to calculating an Independent Samples T Test

Calculate an independent samples t test for the following data sets:

Data set A: 1,2,2,3,3,4,4,5,5,6

Data set B: 1,2,4,5,5,5,6,6,7,9

Solution

Step 1: Sum the two groups:

A: 1 + 2 + 2 + 3 + 3 + 4 + 4 + 5 + 5 + 6 = 35

B: 1 + 2 + 4 + 5 + 5 + 5 + 6 + 6 + 7 + 9 = 50

Step 2: Square the sums from Step 1:

Set A total = 35^2 = 1225

Set B total = 50^2 = 2500

Step 3: Calculate the means for the two groups:

A: (1 + 2 + 2 + 3 + 3 + 4 + 4 + 5 + 5 + 6)/10 = 35/10 = 3.5

B: (1 + 2 + 4 + 5 + 5 + 5 + 6 + 6 + 7 + 9) = 50/10 = 5

Step 4: Square the individual scores and then add them up:

A: 11 + 22 + 22 + 33 + 33 + 44 + 44 + 55 + 55 + 66 = 145

B: 12 + 22 + 44 + 55 + 55 + 55 + 66 + 66 + 77 + 99 = 298

Step 5: Insert your numbers into the following formula and solve:

$$t = \frac{\mu_A - \mu_B}{\sqrt{\left[\frac{\left(\Sigma A^2 - \frac{(\Sigma A)^2}{n_A}\right) + \left(\Sigma B^2 - \frac{(\Sigma B)^2}{n_B}\right)}{n_A + n_B - 2}\right] \cdot \left[\frac{1}{n_A} + \frac{1}{n_B}\right]}}$$

$(\Sigma A)^2$: Sum of data set A, squared (Step 2).
$(\Sigma B)^2$: Sum of data set B, squared (Step 2).
μA: Mean of data set A (Step 3)
μB: Mean of data set B (Step 3)
ΣA^2: Sum of the squares of data set A (Step 4)
ΣB^2: Sum of the squares of data set B (Step 4)
nA: Number of items in data set A
nB: Number of items in data set B

Applying all the values,

$$t = \frac{3.5 - 5}{\sqrt{\left[\frac{\left(145 - \frac{1225}{10}\right) + \left(298 - \frac{2500}{10}\right)}{10 + 10 \quad -2}\right] \cdot \left[\frac{1}{10} + \frac{1}{10}\right]}}$$

t = -1.69

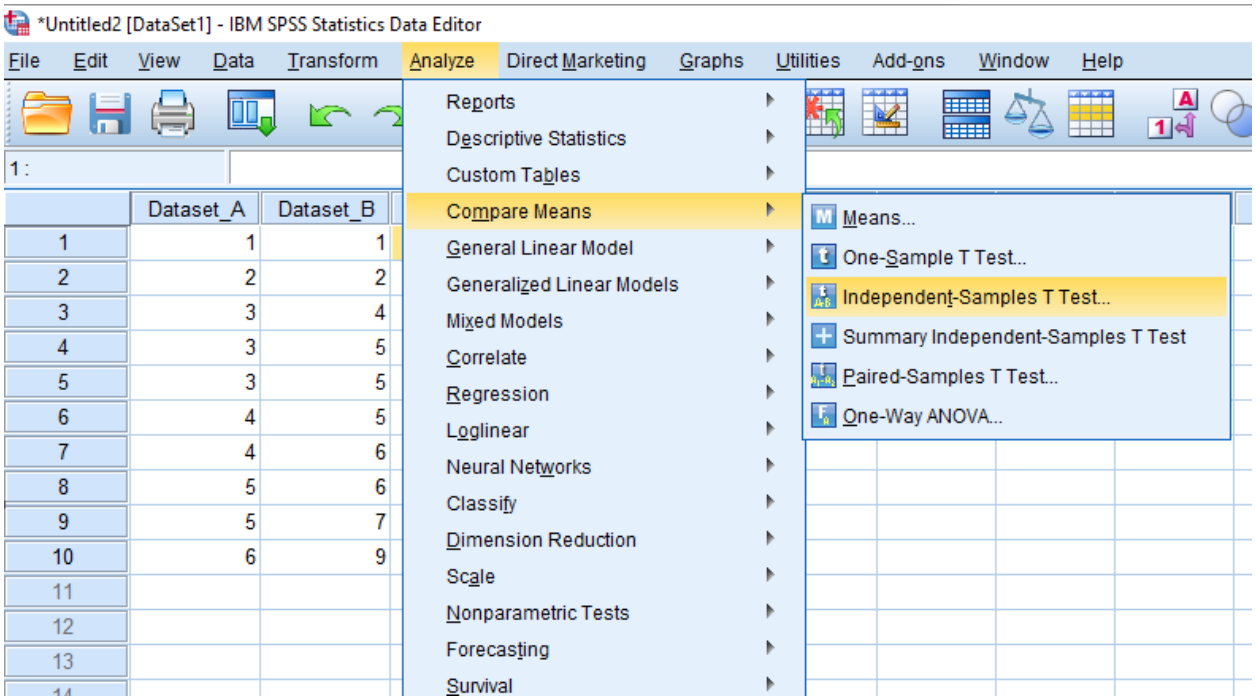
Step 6: Find the Degrees of freedom (nA-1 + nB-1) = 18

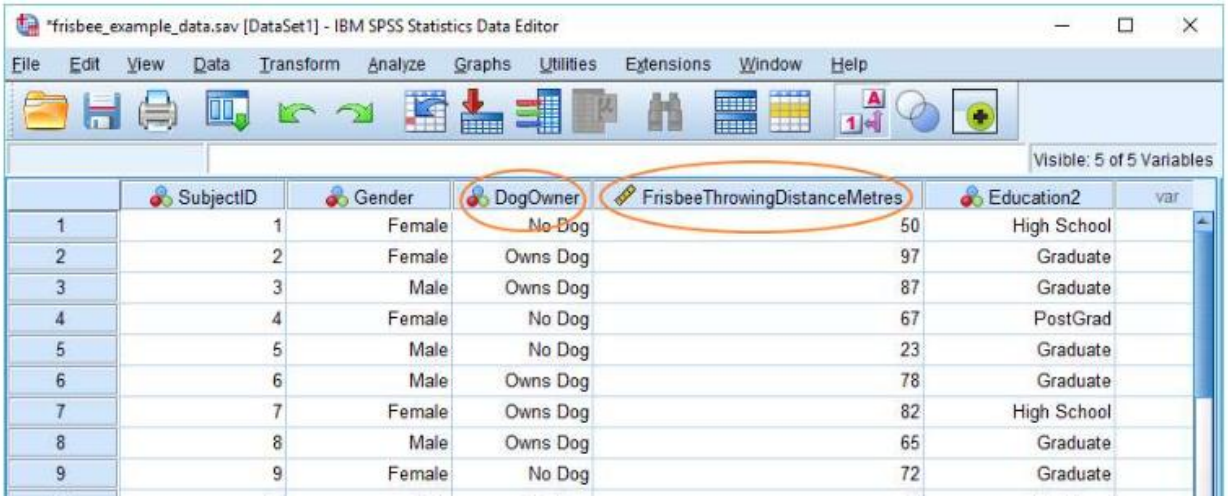
Step 7: Look up your degrees of freedom (Step 6) in the t-table. If you don't know what your alpha level is, use 5% (0.05).
18 degrees of freedom at an alpha level of 0.05 = 2.10.

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
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.043	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.158	2.617	3.373
∝	1.282	1.645	1.960	2.326	2.576	3.291

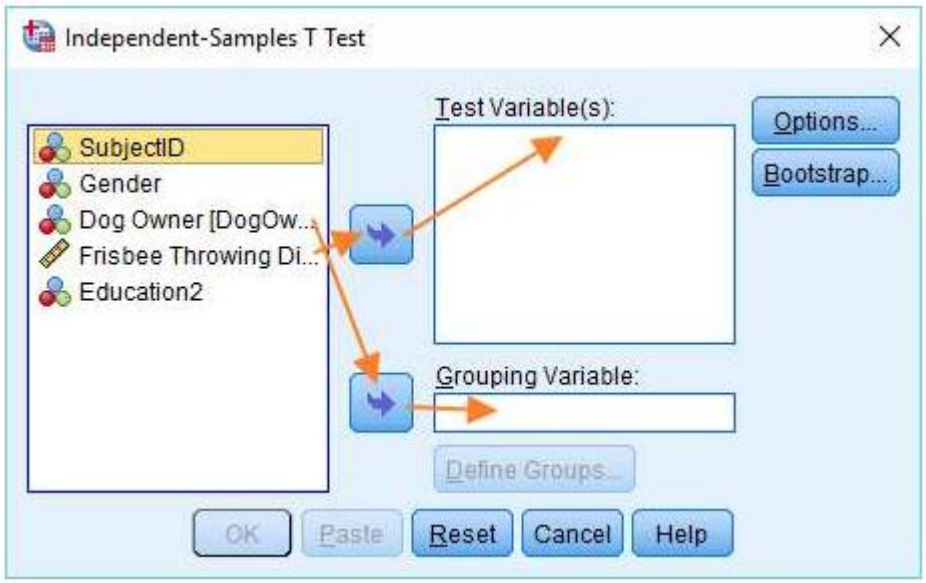
Step 8: Compare your calculated value (Step 5) to your table value (Step 7). The calculated value of -1.79 is less than the cutoff of 2.10 from the table. Therefore $p > .05$. As the p-value

is greater than the alpha level, So we can reject the null hypothesis that there is no difference between means.



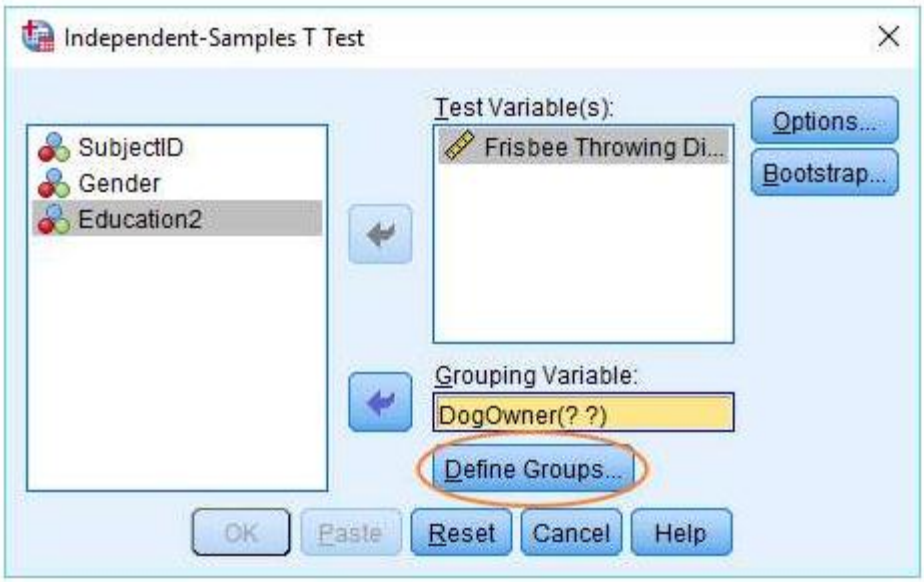


Click on Analyze -> Compare Means -> Independent-Samples T Test. This will bring up the following dialog box.

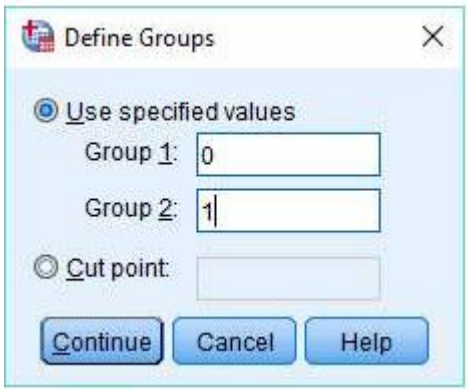


To perform the t test, we've got to get our dependent variable (Frisbee Throwing Distance) into the Test Variable(s) box, and our grouping variable (Dog Owner) into the Grouping Variable box. To move the variables over, you can either drag and drop, or use the arrows, as above.

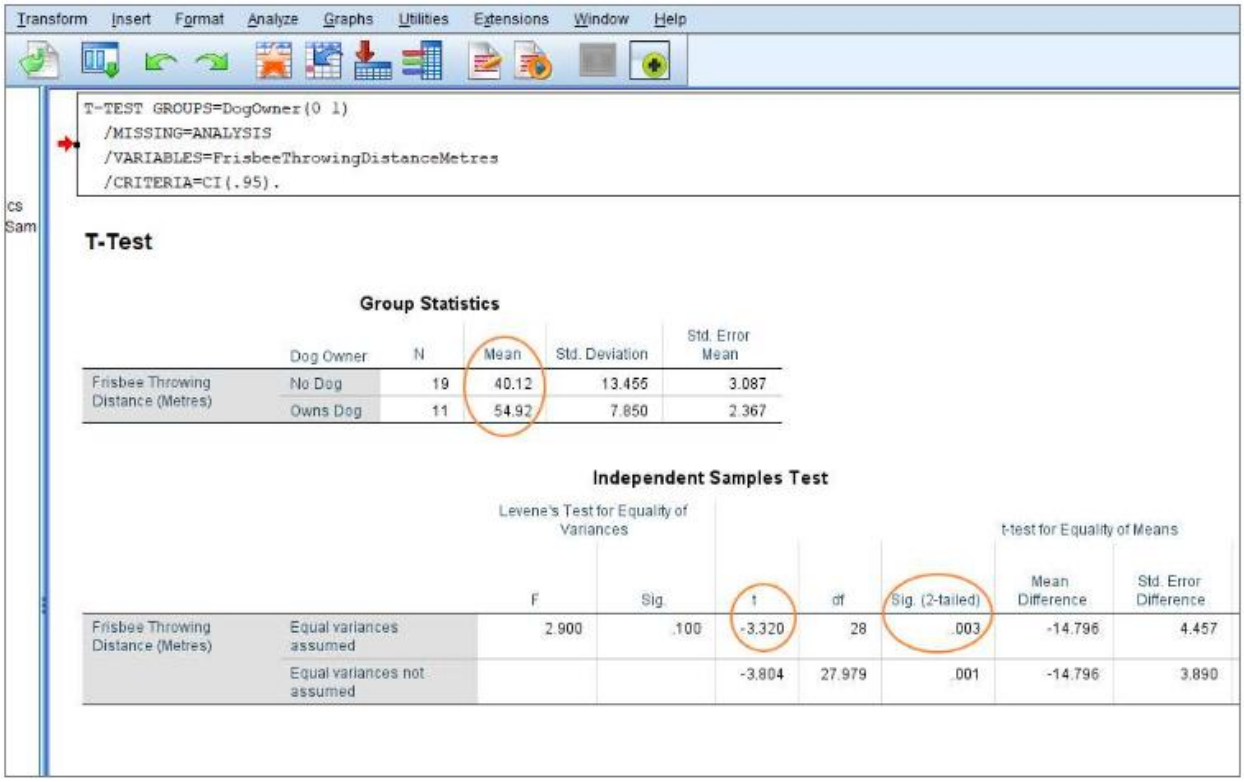
The dialog box should now look like this.



You'll notice that the Grouping Variable, DogOwner, has two question marks in brackets after it. This indicates that you need to define the groups that make up the grouping variable. Click on the Define Groups button.



We're using 0 and 1 to specify each group, 0 is No Dog; and 1 is Owns Dog.



The first thing to note is the mean values in the Group Statistics table. Here you can see that on average people who own dogs throw frisbees further than people who don't own dogs (54.92 metres as against only 40.12 metres).

SPSS is reporting a t value of -3.320 and a 2-tailed p-value of .003. This would almost always be considered a significant result (standard alpha levels are .05 and .01). Therefore, we can be confident in rejecting the null hypothesis that holds that there is no difference between the frisbee throwing abilities of dog owners and non-owners.

One Sample t Test

The One Sample t Test examines whether the mean of a population is statistically different from a known or hypothesized value. The One Sample t Test is a parametric test.
In a One Sample t Test, the test variable's mean is compared against a "test value", which is a known or hypothesized value of the mean in the population.
Example:
A particular factory's machines are supposed to fill bottles with 150 milliliters of product. A plant manager wants to test a random sample of bottles to ensure that the machines are not under- or over-filling the bottles.

Note: The One Sample t Test can only compare a single sample mean to a specified constant. It cannot compare sample means between two or more groups. If you wish to compare the means of multiple groups to each other, you will likely want to run an Independent Samples t Test (to compare the means of two groups) or a One-Way ANOVA (to compare the means of two or more groups).
Data Requirements

- Test variable that is continuous (i.e., interval or ratio level)
- Scores on the test variable
- Random sample of data from the population
- No outliers

Hypotheses
The null hypothesis (H0) and (two-tailed) alternative hypothesis (H1) of the one sample T test can be expressed as:

- H0: μ = μ0 ("the population mean is equal to the [proposed] population mean")
- H1: μ ≠ μ0 ("the population mean is not equal to the [proposed] population mean")

where μ is the "true" population mean and μ0 is the proposed value of the population mean.
Test statistic
The test statistic for a One Sample t Test is denoted t, which is calculated using the following formula:

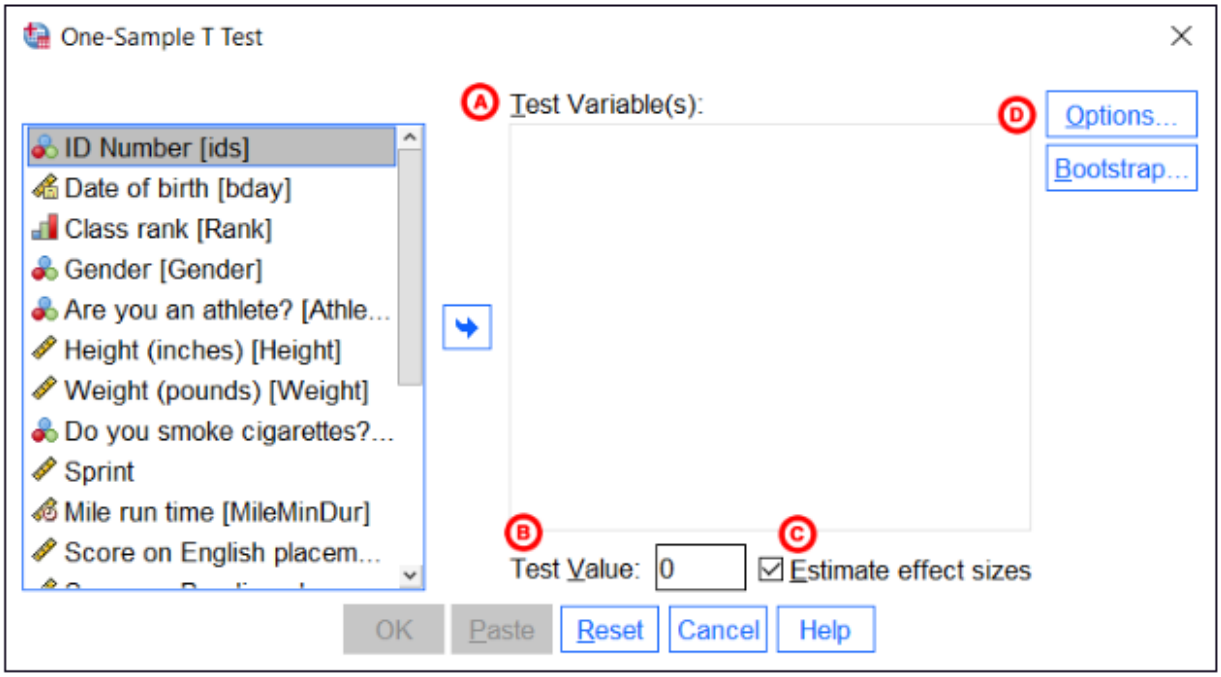
$$t = \frac{\overline{x} - \mu_0}{s_{\overline{x}}} \qquad s_{\overline{x}} = \frac{s}{\sqrt{n}}$$ v

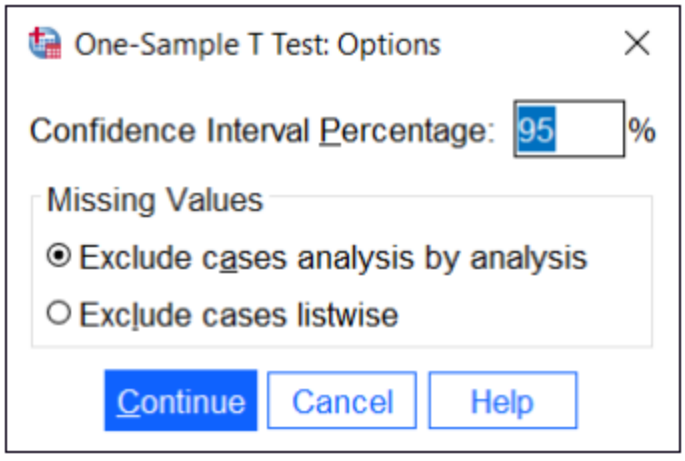
μ0 = The test value -- the proposed constant for the population mean
$\bar{x}$ = Sample mean
n = Sample size (i.e., number of observations)
s = Sample standard deviation

$s_{\bar{x}}$ = Estimated standard error of the mean (s/sqrt(n))

The calculated t value is then compared to the critical t value from the t distribution table with degrees of freedom df = n - 1 and chosen confidence level. If the calculated t value > critical t value, then we reject the null hypothesis.

Run a One Sample t Test
To run a One Sample t Test in SPSS, click Analyze > Compare Means > One-Sample T Test.



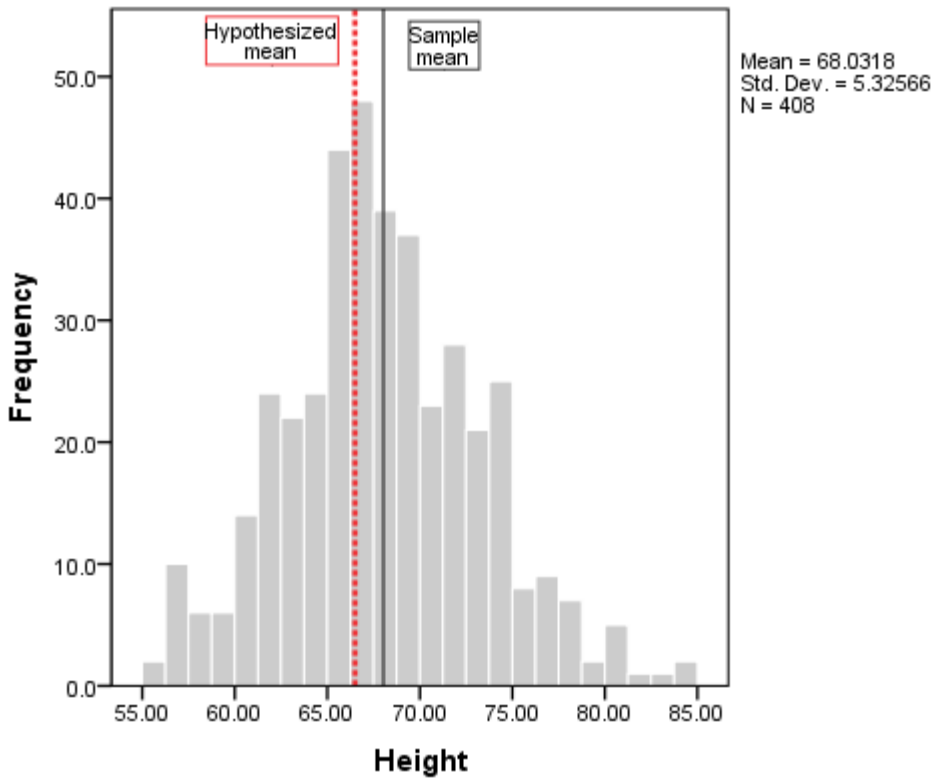


Problem Statement
The mean height of adults ages 20 and older is about 6.5 inches (69.3 inches for males, 63.8 inches for females).

In our sample data, we have a sample of 435 college students from a single college. Let's test if the mean height of students at this college is significantly different than 66.5 inches using a one-sample t test. The null and alternative hypotheses of this test will be:

H0: μHeight = 66.5 ("the mean height is equal to 66.5")
H1: μHeight ≠ 66.5 ("the mean height is not equal to 66.5")

before the test
In the sample data, we will use the variable Height, which a continuous variable representing each respondent's height in inches. The heights exhibit a range of values from 55.00 to 88.41 (Analyze > Descriptive Statistics > Descriptives).
Let's create a histogram of the data to get an idea of the distribution, and to see if our hypothesized mean is near our sample mean. Click Graphs > Legacy Dialogs > Histogram. Move variable Height to the Variable box, then click OK.



To add vertical reference lines at the mean (or another location), double-click on the plot to open the Chart Editor, then click Options > X Axis Reference Line. In the Properties window, you can enter a specific location on the x-axis for the vertical line, or you can choose to have the reference line at the mean or median of the sample data (using the sample data). Click Apply to make sure your new line is added to the chart.

Here, we have added two reference lines: one at the sample mean (the solid black line), and the other at 66.5 (the dashed red line).

running the test
To run the One Sample t Test, click Analyze > Compare Means > One-Sample T Test. Move the variable Height to the Test Variable(s) area. In the Test Value field, enter 66.5.

output
Tables

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Height (inches)	408	68.0318	5.32566	.26366

One-Sample Test

	(B) t	(C) df	(A) Test Value = 66.5 Significance (D) One-Sided p	Significance (D) Two-Sided p	(E) Mean Difference	(F) 95% Confidence Interval of the Difference Lower	Upper
Height (inches)	5.810	407	<.001	<.001	1.53176	1.0135	2.0501

A→ Test Value: which we have entered in T Test window.
B → t Statistic: In this example, t = 5.810. Note that t is calculated by dividing the mean difference (E) by the standard error mean (from the One-Sample Statistics box).
C →df: The degrees of freedom for the test. For a one-sample t test, df = n - 1; so here, df = 408 - 1 = 407.
D → Significance (One-Sided p and Two-Sided p): The p-values corresponding to one of the possible one-sided alternative hypotheses (in this case, μHeight > 66.5) and two-sided alternative hypothesis (μHeight ≠ 66.5), respectively.
E → Mean Difference: The difference between the "observed" sample mean (from the One Sample Statistics box) and the "expected" mean (the specified test value (A)). The positive t value in this example indicates that the mean height of the sample is greater than the hypothesized value (66.5).
F →Confidence Interval for the Difference: The confidence interval for the difference between the specified test value and the sample mean.

Decision and Conclusions
Since p < 0.001, we reject the null hypothesis that the mean height of students at this college is equal to the hypothesized population mean of 66.5 inches and conclude that the mean height is significantly different than 66.5 inches.

Based on the results, we can state the following:

- There is a significant difference in the mean height of the students at this college and the overall adult population. (p < .001).

Regression line in a scatter plot

The regression line in a scatter plot is used to visualize the relationship between two variables and to estimate the trend or pattern in the data. It represents the best-fit line through the scatter plot points, indicating the average or expected value of the dependent variable (y) for a given value of the independent variable (x).

Here are a few key uses and interpretations of the regression line in a scatter plot:

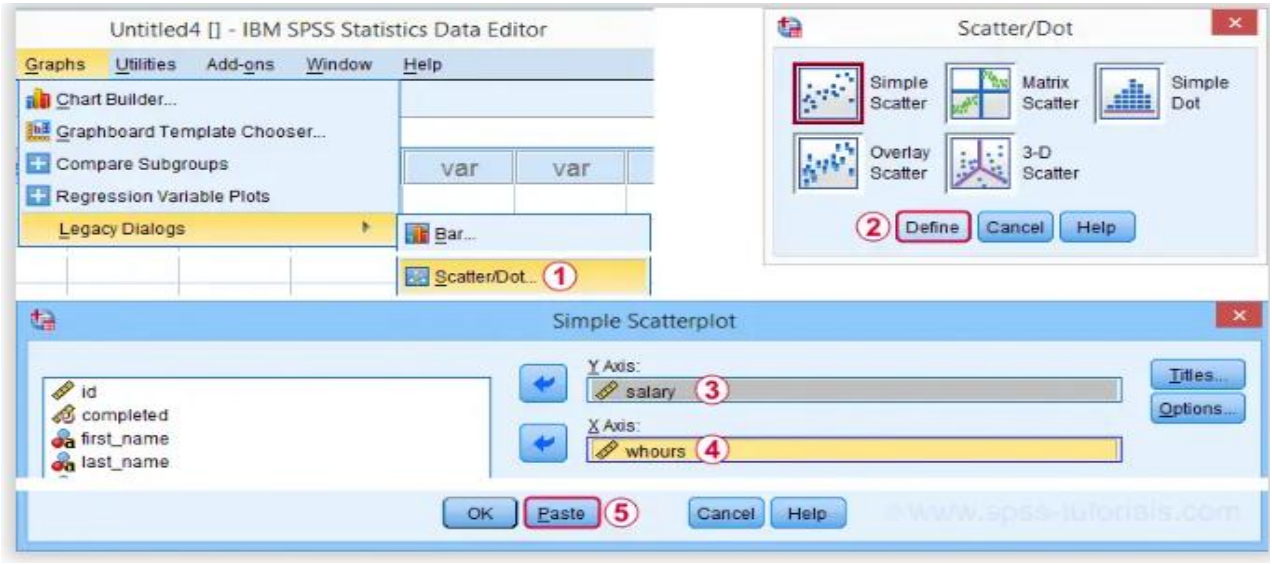
- Trend Identification: The regression line helps identify the general trend or direction of the relationship between the variables. If the line slopes upwards from left to right, it suggests a positive relationship, indicating that as the independent variable increases, the dependent variable tends to increase as well. Conversely, a downward-sloping line indicates a negative relationship.
- Prediction: The regression line can be used for predicting the dependent variable (y) based on a given value of the independent variable (x). By plugging in an x-value into the equation of the line, you can estimate the corresponding y-value. However, it's important to note that predictions become less reliable as you move further away from the observed data points.
- Strength of Relationship: The steepness or slope of the regression line provides information about the strength of the relationship between the variables. A steeper line indicates a stronger association between x and y, while a shallower line suggests a weaker relationship.
- Outlier Detection: The regression line can help identify outliers or data points that deviate significantly from the overall trend. Points that fall far away from the regression line may represent unusual or exceptional observations that warrant further investigation.
- Model Evaluation: The regression line also serves as a benchmark for evaluating the goodness-of-fit of a regression model. Various statistical measures, such as the coefficient of determination (R-squared), can be used to assess how well the regression line represents the data and the proportion of the variation in the dependent variable that is explained by the independent variable.

It's important to note that the regression line represents the average relationship between the variables and may not perfectly capture the behavior of individual data points. Additionally, other types of regression models, such as polynomial regression or multiple regression, can be used to capture more complex relationships between variables.

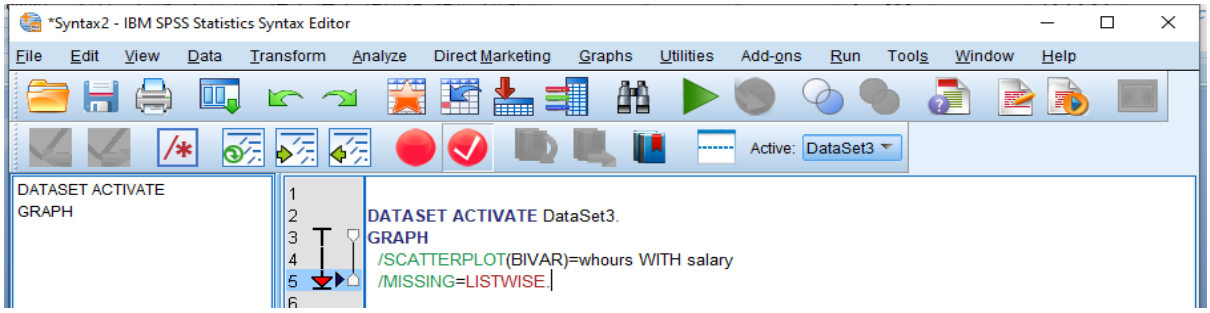
Overall, the regression line in a scatter plot provides a visual representation and estimation of the relationship between two variables, helping to understand patterns, make predictions, and evaluate the strength of the association.

How to make a scatter plot chart in SPSS

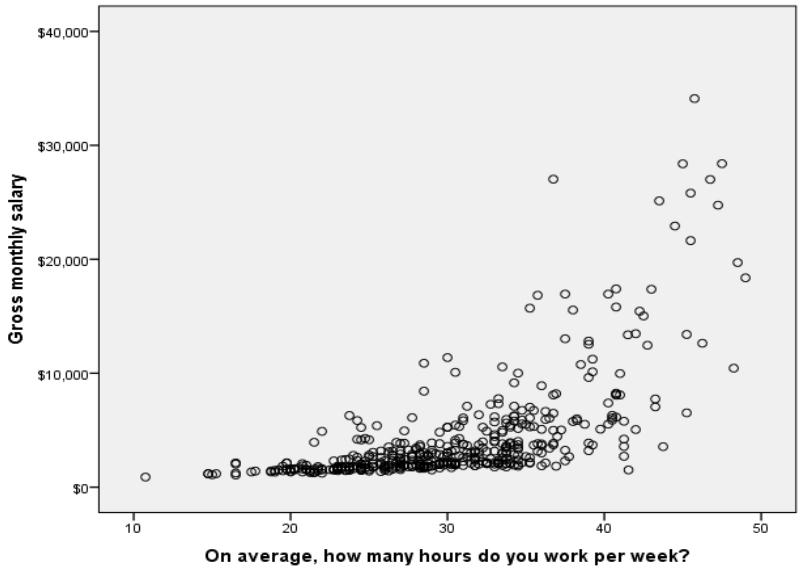
A simple option for drawing linear regression lines is found under Graphs SPSS Menu Arrow Legacy Dialogs SPSS Menu Arrow Scatter/Dot as shown below.



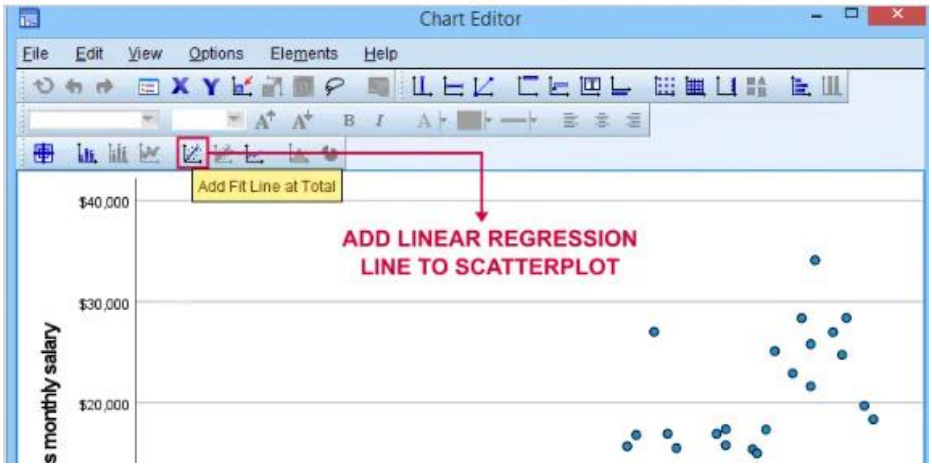
Syntax



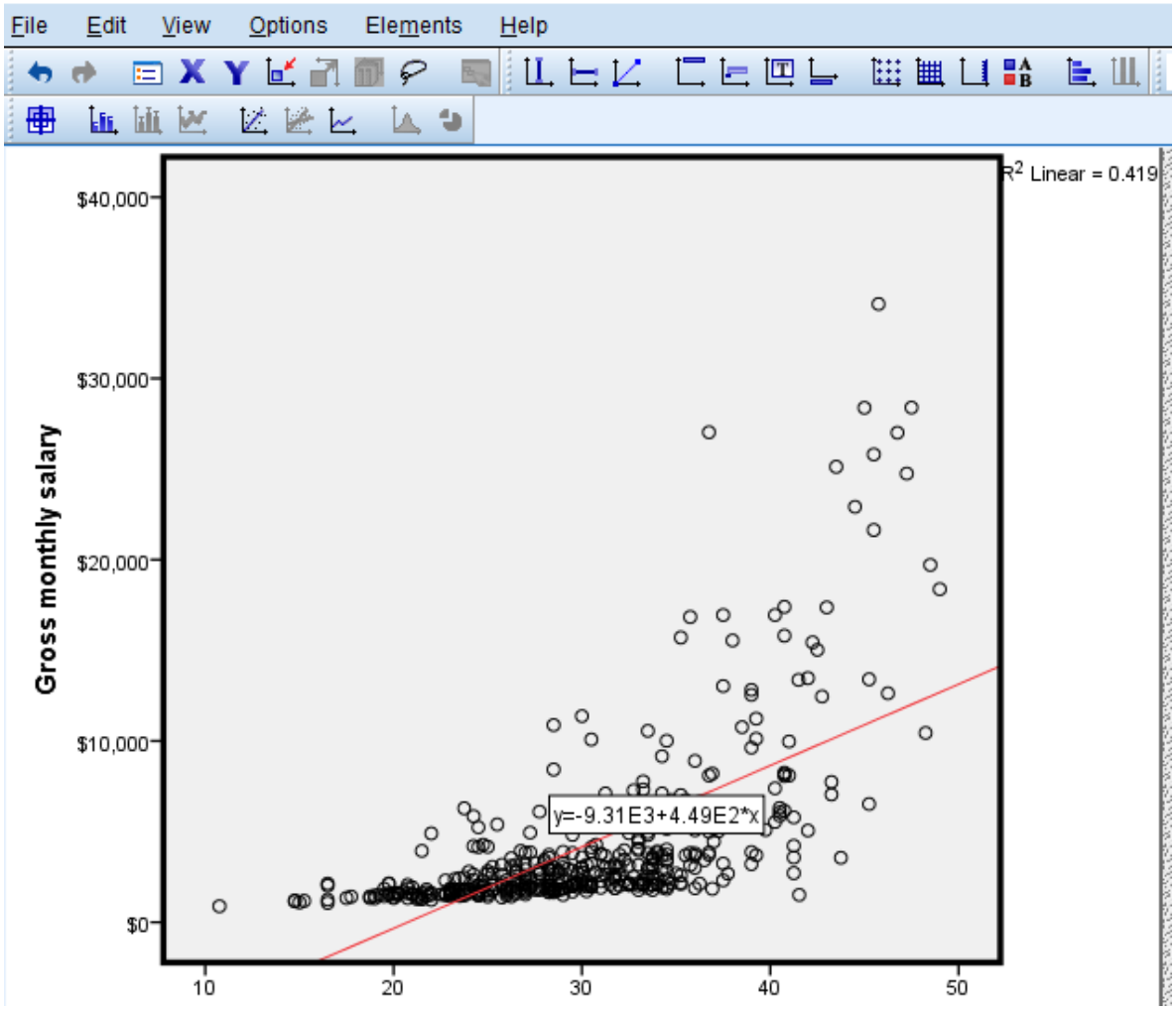
Output



For adding a regression line, first double click the chart to open it in a Chart Editor window. Next, click the "Add Fit Line at Total" icon as shown below.



Adding Regression line to Scatter Plot

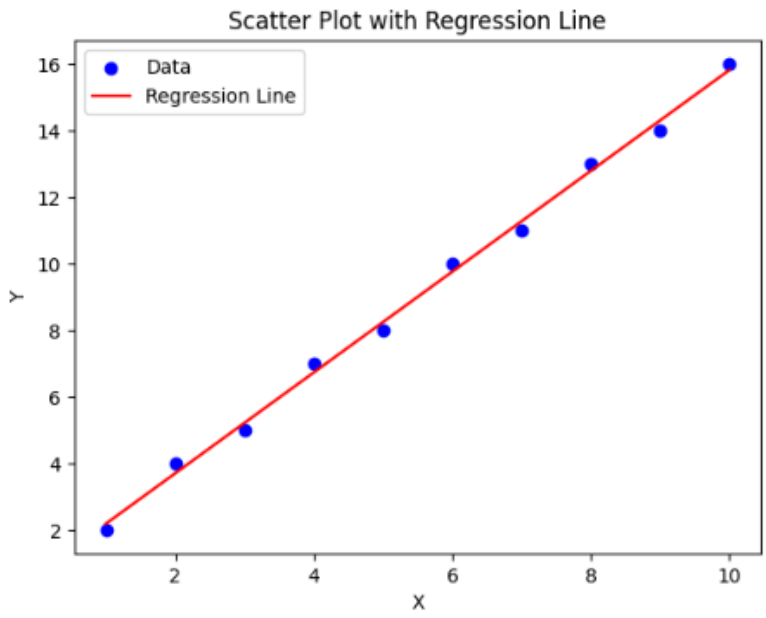


The linear regression equation is shown in the label on our line:
y = 9.31E3 + 4.49E2*x
which means that

Salary′=9310+449·Hours

Simple Python program to add regression line in Scatter Plot

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import linregress

# Example data
x = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
y = [2, 4, 5, 7, 8, 10, 11, 13, 14, 16]

# Perform linear regression
slope, intercept, r_value, p_value, std_err = linregress(x, y)

# Create scatter plot
plt.scatter(x, y, color='blue', label='Data')

# Add regression line
plt.plot(x, intercept + slope*np.array(x), color='red',
label='Regression Line')

# Add labels and title
plt.xlabel('X')
plt.ylabel('Y')
plt.title('Scatter Plot with Regression Line')

# Add legend
plt.legend()

# Display the plot
plt.show()
```



Note:

In this example, we have two arrays, x and y, representing the data points for the scatter plot. We then use the linregress() function from SciPy to perform linear regression on the x and y data.

After that, we create a scatter plot using plt.scatter(x, y, color='blue', label='Data'), specifying the data points, color, and label for the scatter plot.

Next, we add the regression line using plt.plot(x, intercept + slope*np.array(x), color='red', label='Regression Line'). This line is created by evaluating the equation y = intercept + slope*x for each x value.