Community medicine

Biostatistics activity

Activity	
Statistics: Basics – Epidemiology & Biostatistics - Video	2-6
Learn how to code your data in SPSS - Video	7-9
Descriptive Statistics Examples and SPSS analysis - Slides	7-9
Running and Interpreting Chi-Square on SPSS - Video	10-13
Running and Interpreting an Independent-Sample t-test on SPSS - Video	14-16
Running and Interpreting Dependent (Paired) Sample t-test on SPSS - Video	17-18
Biostatistics Activity MCQ	19

This file includes all the lectures and slides. You can rely on it completely , The links of lectures have been placed at the end of each lecture if you want to watch it , Don't worry, this file is very simple , studying it might take less than two hours.

Done by :
Leen Mamoon

Epidemiology and Statistics

Epidemiology and statistics are closely related, but they are not the same. Statistics serves as a tool used by epidemiologists to analyze data and derive conclusions.

Data vs. Information

• Data : are raw numbers or observations that, by themselves, may lack meaning. For example:
53, 61, 62: Without context, these numbers don't provide meaningful information.

• Information: When data are associated with context, they transform into information. For instance, if these numbers represent the ages of three individuals in 2015, they now have meaning: they are the ages of people.

What is a Variable?

• A variable is a mathematical value that can change within a specific situation or scope of a problem.

In mathematics, a variable represents a value that changes within a problem.

In research, a variable can represent characteristics like age, gender, or location.

In computers, a variable might store a number or string of text.

In epidemiology, we often categorize variables as exposures (independent variables) and outcomes (dependent variables). For example:

If we're examining the relationship between smoking rates (exposure) and cancer rates (outcome), we analyze how changes in smoking rates influence cancer rates.

Math	A value that may change within the scope of a problem or situation (vs. a "constant").
Research	A logical set of attributes (gender, age, etc.).
Computers	A symbolic name given to an unknown quantity.

Types of Variables

Continuous Variables

-Accept fractions

-Examples include age, height, and temperature.

Categorical Variables (Discrete)

- Doesn't Accept fractions

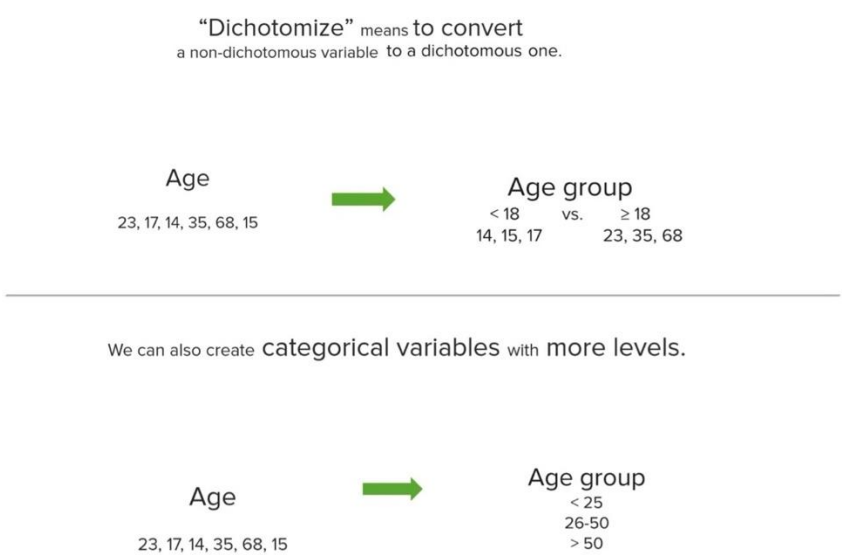
-Examples include gender, employment status, and age group.

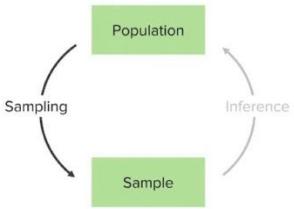
Dichotomization: This refers to converting a continuous variable into a dichotomous one (i.e., having only two categories). For example:

- Age: 23, 17, 14, 35, 68, 15
- Dichotomized Age Group: <18 vs. ≥18

Dichotomous Variables: A subtype of categorical variables with only two possible values.

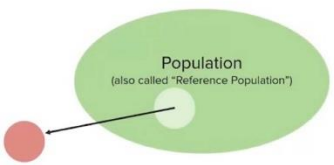
Example: Male/Female, Employed/ Unemployed.





The statistics and epidemiologic approaches we use are affected by the assumptions of the sampling strategies used.

-The topic was mentioned and explained in the main lectures (not the activities).

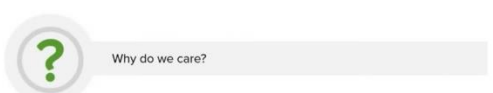


Introduction to the Null Hypothesis:

Definition: The null hypothesis is a statement that there is no relationship between the variables being tested.

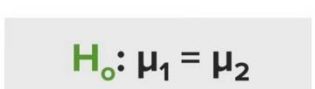
Purpose: it's crucial for performing statistical tests. It helps us determine whether to reject or fail to reject the hypothesis.

Important Note: We never "accept" the null hypothesis. Instead, we either reject it or fail to reject it.



Example: Suppose you're comparing the mean responses of two groups, such as a treatment group and a placebo group in a randomized control trial.

- The null hypothesis states that the mean measurement in the treatment group is equal to the mean measurement in the placebo group ($\mu_1 = \mu_2$).
- This indicates no difference or effect between the groups.

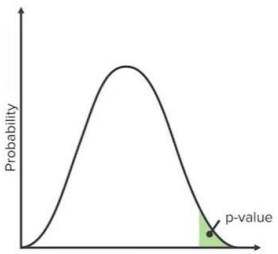


H_o: the average number of subjects getting better in the test group is no different from the average number of subjects in the placebo group.

P- Valu

The p-value is the value under this normal curve that tells you the likelihood of the null hypothesis being true. It essentially measures how likely it is that you made an error in rejecting the null hypothesis.

A "p-value" is computed from a statistical test. It tells us whether we should reject the null hypothesis.



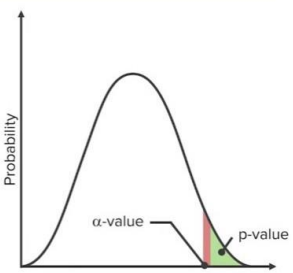
Definition of Alpha (α):

It represents the maximum probability of making a Type I Error, which means rejecting the null hypothesis when it is actually true.

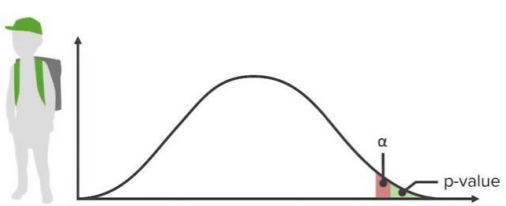
Commonly set to 0.05, meaning a 5% chance of rejecting the null hypothesis incorrectly.

Whether or not we reject the null is determined by whether the p-value is below a certain cut-off, which we call the alpha value.

Traditionally, we tend to set alpha at either 0.05 or 0.01.



For example, if we are testing whether the average heights of two groups of children are different, and perform a t-test to produce a p-value of 0.02, setting α = 0.05, we can conclude that null hypothesis is rejected and that the two groups do indeed have different average heights.



If the p-value from the test is 0.02, which is less than 0.05, this indicates that you can reject the null hypothesis.

- **Error Likelihood: The p-value of 0.02 suggests that there is only a 2% chance that you mistakenly rejected the null hypothesis. Since this is a small probability, it indicates a reliable result.**

"If the p is low, the null (hypothesis) must go."

confidence Intervals

The Confidence Level is complementary to α:

- Confidence Level = 1 - \alpha .
- For \alpha = 0.05 , the confidence level is 1 - 0.05 = 0.95 , or 95%.

- Preferred over p-values.
- Provide a range where the true value is likely to fall.
- Example: If the average age is 21, a 95% confidence interval might be (20.5, 21.5).
- This means there's a 95% chance the actual average is within that range.

statistical tests

Test Name	Purpose	Example
T-Test	Compare the means of two groups.	Comparing the average height of children in two different schools.
Chi-Square Test	Determine if two categorical variables are associated.	Testing if smoking status (smoker/non-smoker) is associated with gender (male/female).
ANOVA	Compare the means of three or more groups.	Comparing the average weight among three diet groups (low-fat, keto, Mediterranean).
Correlation	Assess the relationship between two continuous variables.	Checking if age and maximum heart rate are correlated.
Regression	Determines the influence of one or more variables on an outcome.	Modeling how smoking, age, and exercise influence heart disease risk.

Link to the lecture

Descriptive Statistics example

Exercise 1

Computer Sales (n = 12 salespeople)
Original Data: 3, 10, 2, 5, 9, 8, 7, 12, 10, 0, 4, 6
Compute the mean, median, mode, quartiles.

First order the data:
0, 2, 3, 4, 5, 6, 7, 8, 9, 10, 10, 12
$\Sigma X_i = 76$
$\overline{X} = 76 / 12 = 6.33$ computers sold
Median = 6.5 computers
Mode = 10 computers
$Q_1 = 3.5$ computers, $Q_3 = 9.5$ computers

Exercise 2

Data (n=16):

1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 5, 5, 6, 7, 8, 10

Compute the mean, median, mode, quartiles.

Answer.

1 1 2 2 ¦ 2 2 3 3 ¦ 4 4 5 5 ¦ 6 7 8 10

Mean = 65/16 = 4.06
Median = 3.5
Mode = 2
$Q_1 = 2$
Q_2 = Median = 3.5
$Q_3 = 5.5$

Exercise 3

Data – number of absences (n=13) :

0, 5, 3, 2, 1, 2, 4, 3, 1, 0, 0, 6, 12

Compute the mean, median, mode, quartiles.

Answer. First order the data:

0, 0, 0, ¦ 1, 1, 2, 2, 3, 3, 4, ¦ 5, 6, 12

Mean = 39/13 = 3.0 absences
Median = 2 absences
Mode = 0 absences
Q_1 = .5 absences
Q_3 = 4.5 absences

Inter-Quartile Range (IQR)

- $IQR = Q_3 - Q_1$
- Example (n = 15):
- **0, 0, 2, 3, 4, 7, 9, 12, 17, 18, 20, 22, 45, 56, 98**
- $Q_1 = 3$, $Q_3 = 22$

IQR = 22 – 3 = 19 (Range = 98)

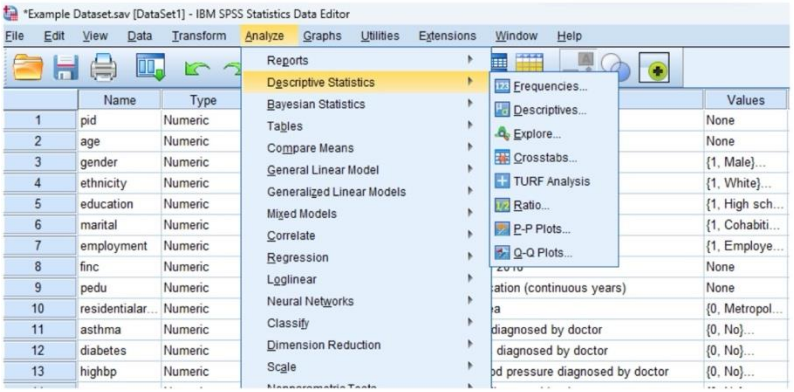
Running Descriptive Statistics in SPSS Statistics

1. STEP: Load Data into SPSS

• Commence by launching SPSS and loading your dataset, which should encompass the variables of interest – a categorical independent variable (if any) and the continuous dependent variable. If your data is not already in SPSS format, you can import it by navigating to File > Open > Data and selecting your data file.

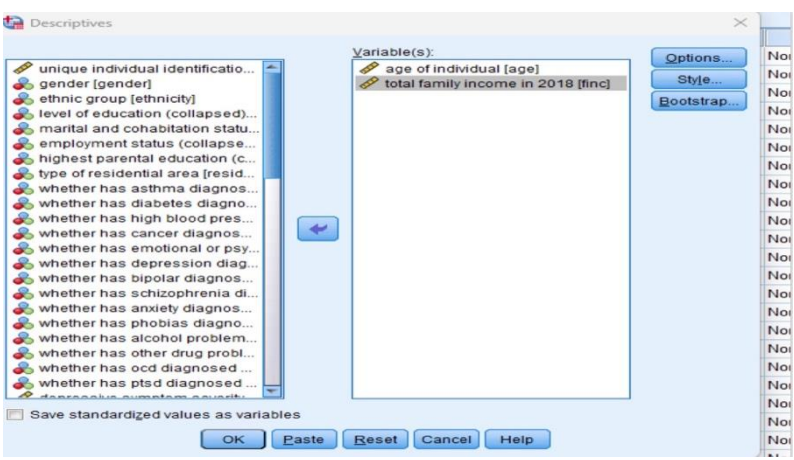
2. STEP: Access the Analyze Menu

• In the top menu, locate and click on “Analyze. ” Within the “Analyze” menu, navigate to “Descriptive Statistics” and choose “Descriptives.
” Analyze > Descriptive Statistics > Descriptives



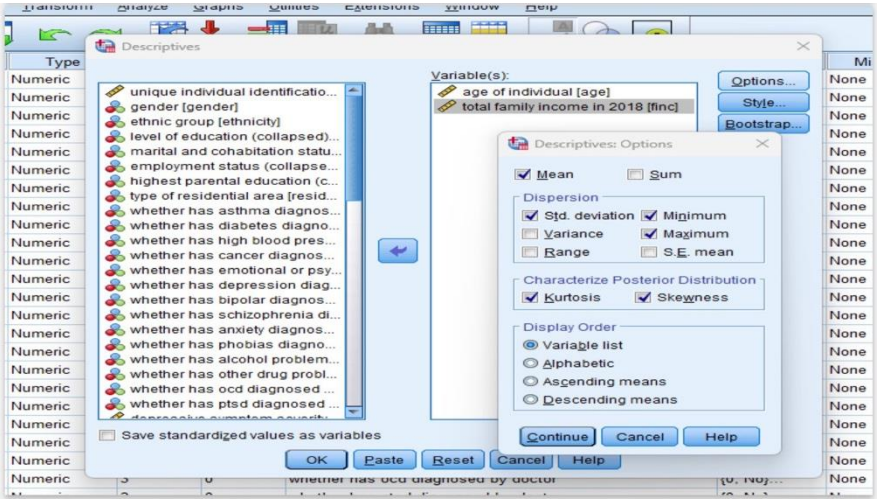
3. STEP: Specify Variables

• Upon selecting “Descriptives, ” a dialog box will appear. Transfer the continuous variable you wish to analyze into the “Variable(s)” box.



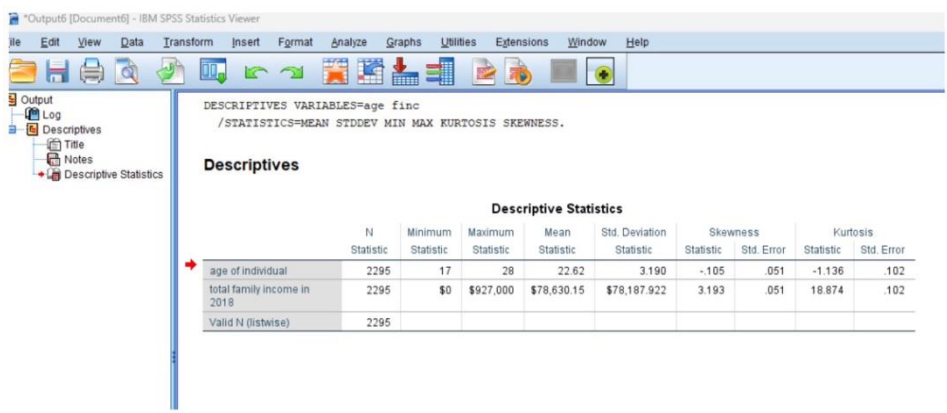
4. STEP: Define Options

• Click on the “Options” button within the “Descriptives” dialog box to access additional settings. Here, you can request various descriptive statistics such as mean, median, mode, standard deviation, and more. Adjust the settings according to your analytical requirements



5. STEP: Generate Descriptive Statistics:

Once you have specified your variables and chosen options, click the "OK" button to execute the analysis. SPSS will generate a comprehensive output, including the requested descriptive statistics for your dataset.



Link to the lecture

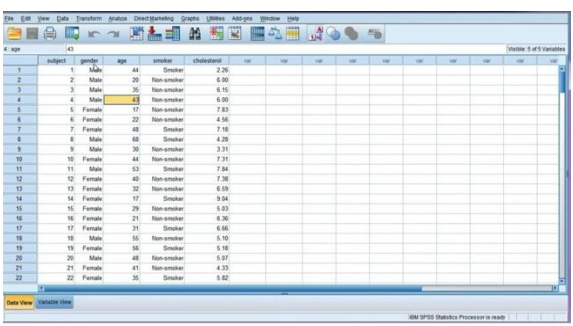
This lecture includes these slides . However, it only teaches you how to input information into the computer, which is something that is impossible to be tested on or come up in the exam.

Chi- square in SPSS

The Chi-square test is applied when we are working with categorical variables "Nominal Variables" (e.g., gender, disease status, education level) rather than continuous data (e.g., height, weight, or age).

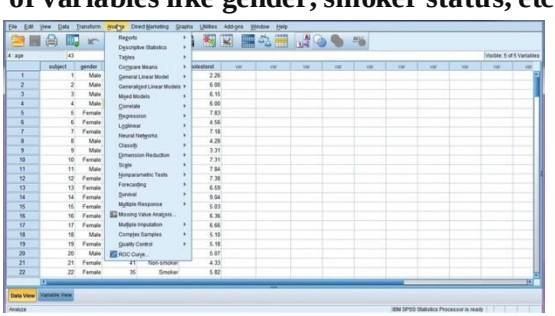
1. **Open SPSS:**
2. **Enter your data into SPSS with the following columns:**

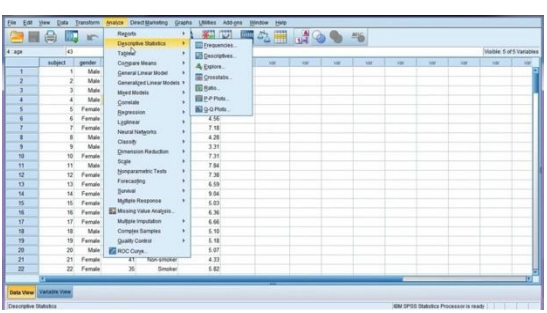
Subject: An identifier for each participant , Gender: Categorical variable (Male , Female) , Age ,Smoker: Categorical variable (Smoker ,Non-smoker) , Cholesterol.

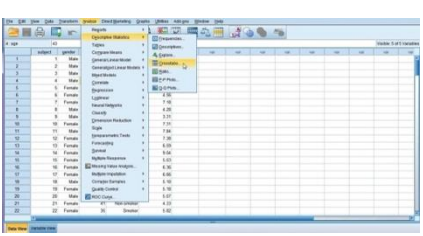


3. Descriptive Statistics:

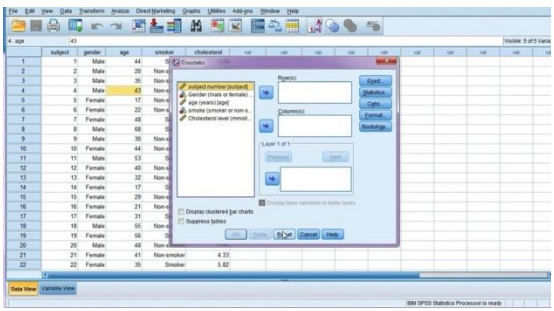
Go to Analyze → Descriptive Statistics → Crosstabs to check the distribution of variables like gender, smoker status, etc.

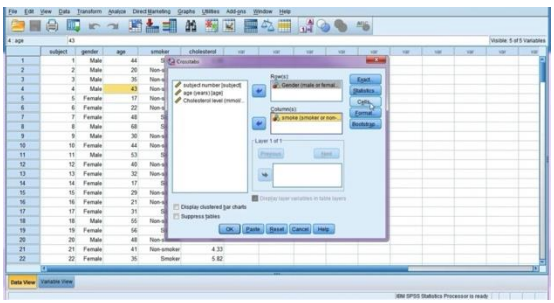


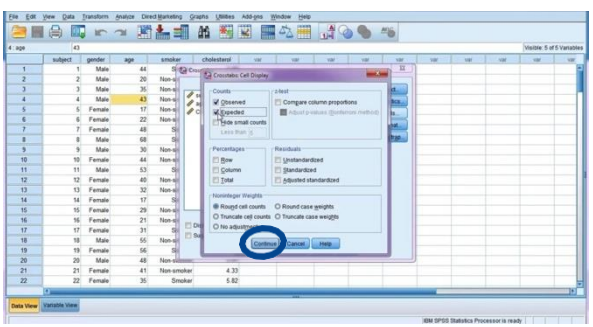




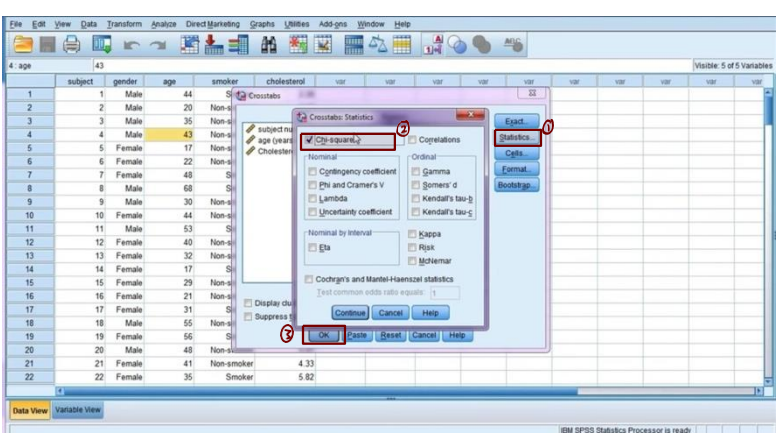
4. Reset → Select Gender for the Row(s) & Select Smoker for the Column(s) → Click on Cells and check the box for Expected to see the expected frequencies in the table (this is important for the Chi-square test) → continue



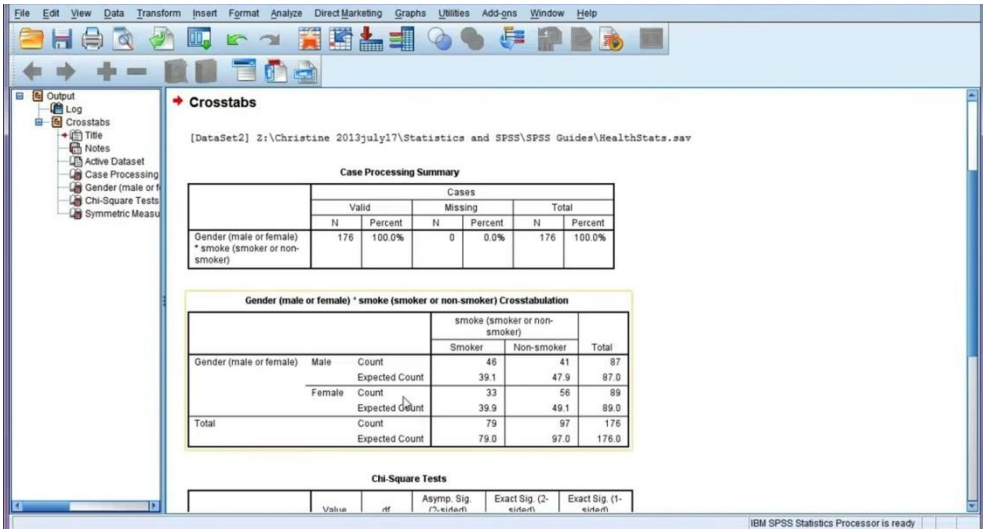




5. In the Crosstabs dialog box, click on the Statistics button → Select Chi-square to run the test →Click Continue → ok



The Chi-square test for independence or for association is used to determine if there is a statistically significant relationship or association between two categorical variables. However, it does not provide any information about the strength or magnitude of that relationship



Gender (male or female) * smoke (smoker or non-smoker) Crosstabulation

			smoke (smoker or non-smoker)		
			Smoker	Non-smoker	Total
Gender (male or female)	Male	Count	46	41	87
		Expected Count	39.1	47.9	87.0
	Female	Count	33	56	89
		Expected Count	39.9	49.1	89.0
Total		Count	79	97	176
		Expected Count	79.0	97.0	176.0

Look at the table showing the Observed and Expected frequencies for each combination of Gender and Smoker categories.

- **Observed Counts:** **These are the actual counts in your data for each combination of gender and smoking status.**

- **Expected Counts:** **These are the counts you would expect if there was *no* association between gender and smoking status (i.e., if the two variables were independent).**

- **large difference between observed and expected counts suggests that the variables (gender and smoking status) are not independent.**

- **If the difference is not large, it suggests that gender and smoking status are independent**

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4.437[a]	1	.035		
Continuity Correction[b]	3.821	1	.051		
Likelihood Ratio	4.455	1	.035		
Fisher's Exact Test				.049	.025
Linear-by-Linear Association	4.411	1	.036		
N of Valid Cases	176				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 39.05.

b. Computed only for a 2x2 table

Double-click to activate

At the bottom of the Chi-square test table in SPSS, you'll see a line that tells you the percentage of cells that have an expected count less than 5. You want this percentage to be 0% (or very close to 0%). This means no cell has an expected count less than 5, and the Chi-square test is valid.

If this percentage is greater than 20%, it indicates that the assumption has been violated, and the Chi-square test results may not be reliable

0 cells have an expected count less than 5, which means the assumption is not violated.

The Pearson Chi-square test gives a statistically significant result (p = 0.035), which is less that alpha (0.05) so we reject the null hypothesis meaning gender and smoking status are dependent on each other.

All of the following about degree of freedom is correct except:

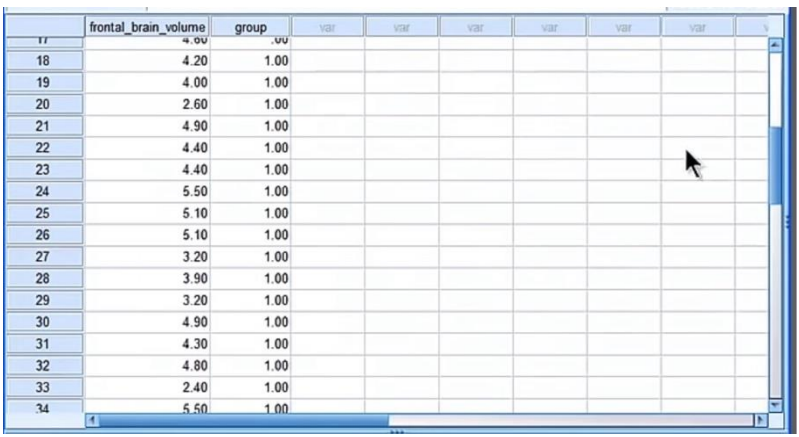
- Advantages of large degree of freedom often depends on the type of the analysis
- Degrees of freedom typically (but not always) relate the size of the sample.
- A higher degree of freedom means more power to reject a false null hypothesis and find a significant result.
- ✔ Higher degrees of freedom generally mean smaller sample sizes

Link to the lecture

T-test on SPSS

You have two variables:

- Dependent variable: Frontal brain volume (measured on an interval scale).
- Grouping variable: Smoker vs. Non-smoker (categorical variable coded as 0 for non-smokers and 1 for smokers).



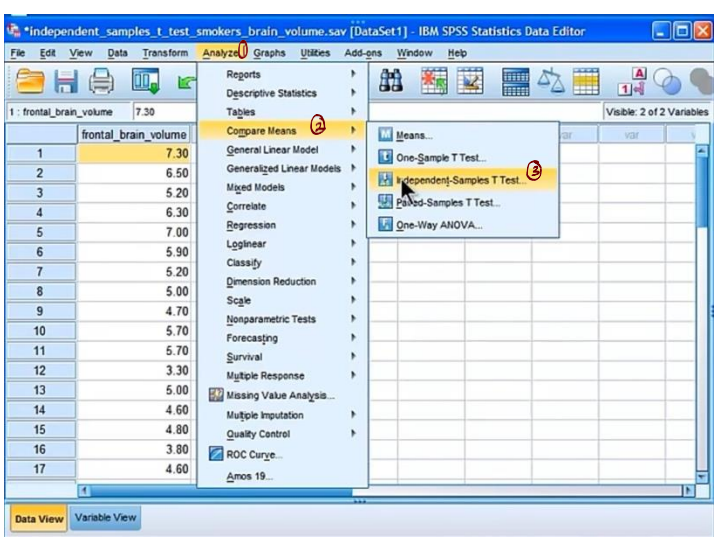
	frontal_brain_volume	group
18	4.20	1.00
19	4.00	1.00
20	2.60	1.00
21	4.90	1.00
22	4.40	1.00
23	4.40	1.00
24	5.50	1.00
25	5.10	1.00
26	5.10	1.00
27	3.20	1.00
28	3.90	1.00
29	3.20	1.00
30	4.90	1.00
31	4.30	1.00
32	4.80	1.00
33	2.40	1.00
34	5.50	1.00

الدراسة كانت تهدف إلى مقارنة حجم الدماغ الأمامي (Frontal Brain Volume) بين المدخنين وغير المدخنين لمعرفة ما إذا كان هناك اختلاف إحصائي بين المجموعتين.

البيانات المستخدمة:

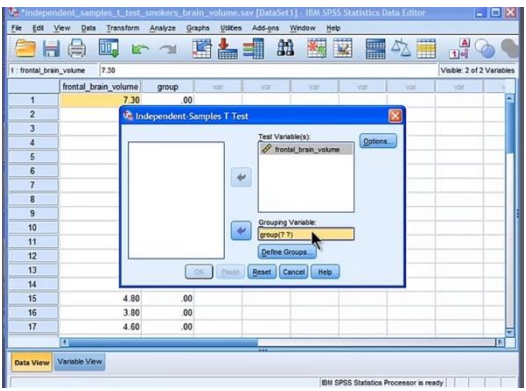
- تم جمع البيانات من 36 مشاركًا في الدراسة.
- تم تقسيم المشاركين إلى مجموعتين: 17 شخصًا غير مدخنين و 19 شخصًا مدخنًا.

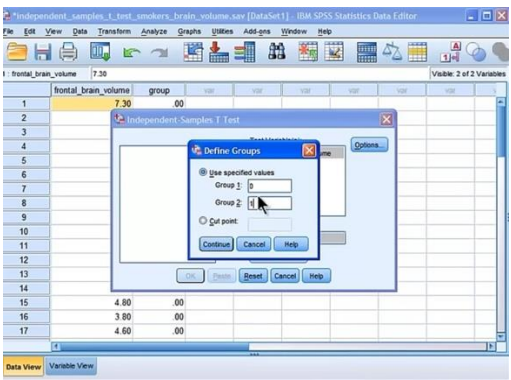
1. Go to Analyze > Compare Means > Independent Samples T-Test.

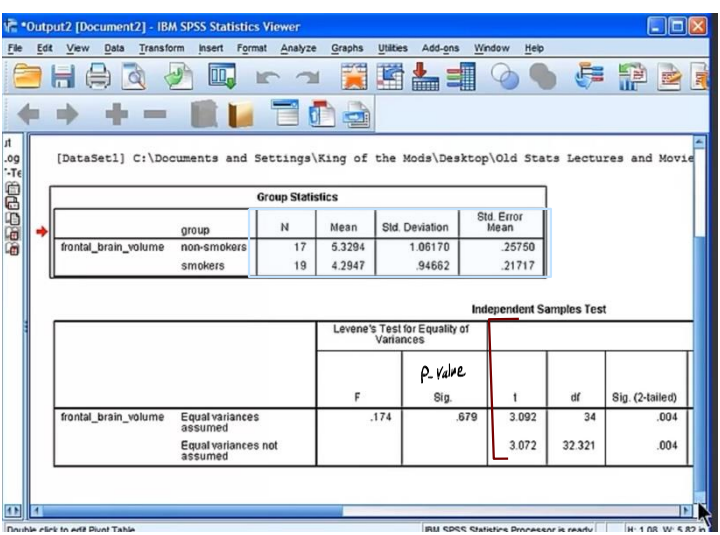


2. Place the dependent variable (frontal brain volume) into the "Test Variables" box & Place the grouping variable (smoker vs. non-smoker) into the "Grouping Variable" box.

• Define the groups by selecting "0" for non-smokers and "1" for smokers.. continue







*Output2 [Document2] - IBM SPSS Statistics Viewer

[DataSet1] C:\Documents and Settings\King of the Mods\Desktop\Old Stats Lectures and Movie

Group Statistics

	group	N	Mean	Std. Deviation	Std. Error Mean
frontal_brain_volume	non-smokers	17	5.3294	1.06170	.25750
	smokers	19	4.2947	.94662	.21717

Independent Samples Test

		Levene's Test for Equality of Variances: F	Sig. (p-value)	t	df	Sig. (2-tailed)
frontal_brain_volume	Equal variances assumed	.174	.679	3.092	34	.004
	Equal variances not assumed			3.072	32.321	.004

1. Group: Non-smokers (17 participants) and Smokers (19 participants).

2. Standard Deviation:

This tells you the spread or variability of the brain volume measurements within each group. A larger standard deviation means more variability in measurements.

3. The standard error measures how accurately the sample mean represents the population mean. A smaller standard error indicates more precision in the sample mean.

Levene's Test for Equality of Variances:

- **Purpose: Tests whether the variances of the two groups (smokers and non-smokers) are equal.**
- **If the significance (p-value) from Levene's test is greater than 0.05, we can assume that the variances are equal. In this case, the p-value is 0.679, so the assumption of equal variances is not violated.**

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means		
		F	Sig.	t	df	Sig. (2-tailed)
frontal_brain_volume	Equal variances assumed	.174	.679	3.092	34	.004
	Equal variances not assumed			3.072	32.321	.004

T-test Results:

- **T-statistic: 3.092**
- **Degrees of freedom (df): 34**
- **P-value: 0.004**

Since the p-value (0.004) is less than the typical significance level (0.05), we reject the null hypothesis that there is no difference between the brain volumes of smokers and non-smokers. The

Statistical significance: This means that the difference you observed between the groups (e.g., frontal brain volume between smokers and non-smokers) is unlikely to be due to random variability or chance. In other words, the result is probably not due to random sampling or measurement errors.

وجود علاقة إحصائية بين التدخين وحجم الدماغ الأمامي (أي أن المدخنين وغير المدخنين يختلفون في الحجم).

لكن لا يمكن الجزم أن التدخين هو السبب المباشر لهذا الاختلاف دون إجراء مزيد من الأبحاث ودراسات طويلة المدى

Deviation	Std. Error Mean
1.06170	.25750
.94662	.21717

Independent Samples Test

or Equality of ces	t-test for Equality of Means						
						95% Confidence Interval of the Difference	
Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
.679	3.092	34	.004	1.03467	.33466	.35457	1.71478
	3.072	32.321	.004	1.03467	.33685	.34880	1.72055

Cohen's d (Effect Size):

- **Cohen's d measures the size of the effect or the magnitude of the difference between the two groups (smokers vs. non-smokers).**
- **In this case, Cohen's d is about 1.03, which indicates a large effect size. A value above 0.8 is typically considered a large effect.**

SPSS also provides confidence intervals for the mean difference between groups, but this information is rarely reported.

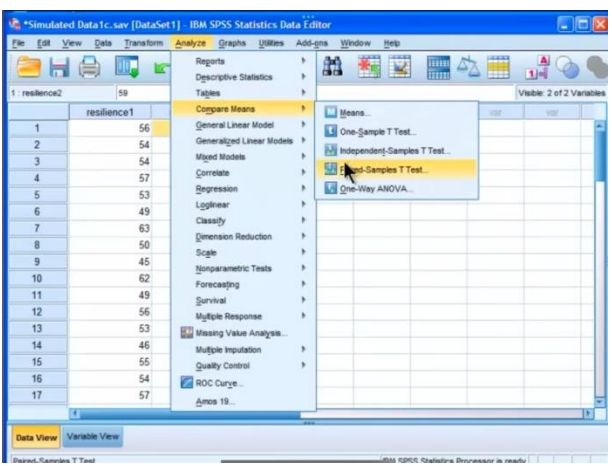
Link to the lecture

Paired samples T-test on SPSS

The study involves 50 participants, and their resilience scores were measured before and after a training program (Time 1 and Time 2).

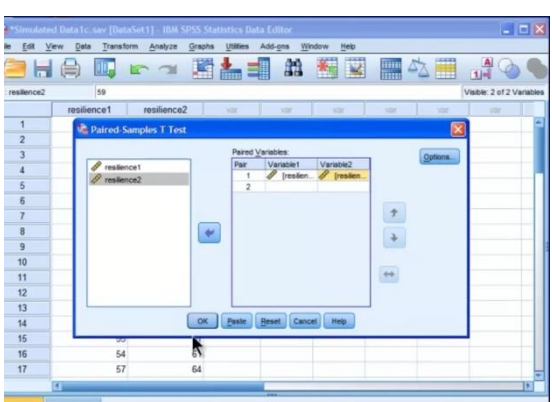
The goal is to determine if there is a statistically significant difference between the resilience scores before and after the training.

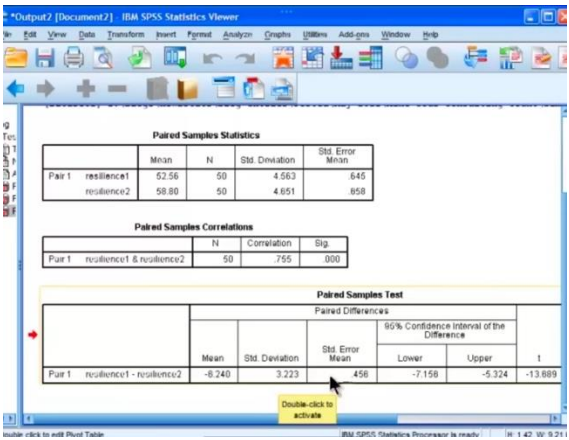
1. **In the top menu, click on Analyze →Compare Means → Select Paired-Samples T Test.**

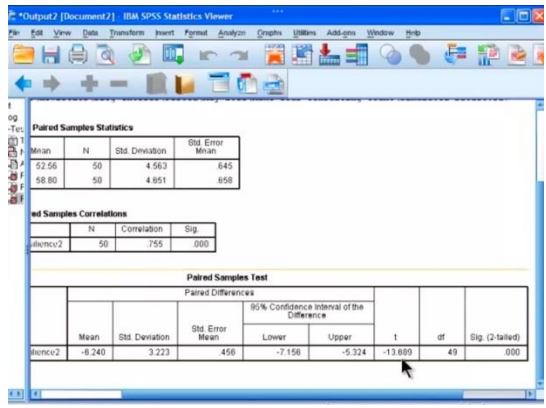


هذا الاختبار يساعدنا في معرفة ما إذا كانت الاختلافات بين القياسين (قبل وبعد التدريب) كبيرة بما يكفي لكي نعتبرها ذات دلالة إحصائية

2. Select the column for Resilience Time 1 and move it to Variable 1 & Select the column for Resilience Time 2 and move it to Variable 2 → ok







Paired Samples Statistics Table:

- This will display the means, standard deviations, and sample sizes for both Resilience Time 1 and Resilience Time 2.

-This will give you an idea of the central tendency of each group.

Paired Samples Correlations Table:

- This shows the correlation between the two sets of scores (before and after).
- You might see a correlation value of 0.755, indicating that participants' scores tend to be similar between the two time points.

Paired Samples Test Table:

- This is where you'll find the results of the t-test.
  - Mean Difference: The difference between the means of Time 1 and Time 2 (e.g., -6.24).
  - Standard Deviation of the Difference: The variability of the difference scores.
  - Standard Error Mean: The error associated with the mean difference.
  - t-value: The calculated t-statistic (e.g., -13.69).
  - Degrees of Freedom (df): The number of participants minus 1 (e.g., 49 if you have 50 participants).
  - Sig. (2-tailed): The p-value (e.g., 0.000). If this value is less than 0.05, the difference is statistically significant.

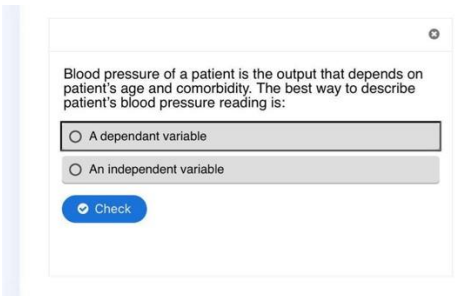
كان لدى المشاركين متوسط مرونة قبل التدريب قدره 52.56، ومتوسط مرونة بعد التدريب قدره 58.80.

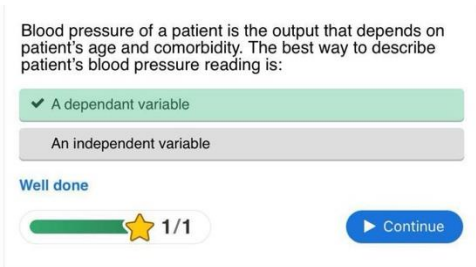
- الفرق بين هذين المتوسطين هو 6.24.
- القيمة $p = 0.000$ تشير إلى أن هذا الفرق ليس صدفة، وأن التدريب بالفعل أدى إلى زيادة في مستوى المرونة

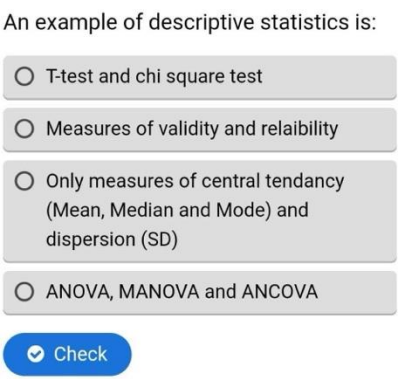
Link to the lecture

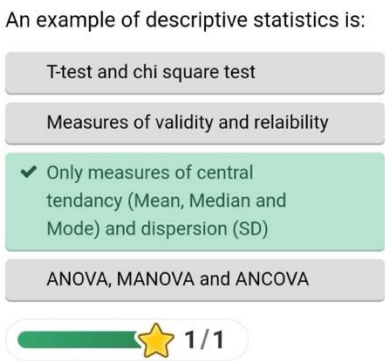
p-value (Sig.) is less than 0.05, you can reject the null hypothesis, indicating that there is a statistically significant difference between the two time points. the p-value of 0.000 indicates a very statistically significant difference

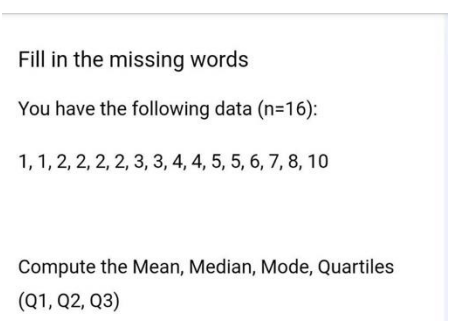
Biostatistics activity MCQ

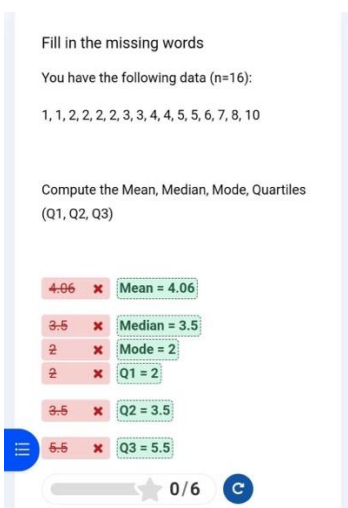












تم بحمدلله

اللهم يسر لي كل صعب، واجعل الخير في عملي ونتائجه، واجعلني من الذين يسيرون على طريق الحق والرشاد. اللهم إني أسالك علمًا نافعًا، ورزقًا طيبًا، وعملًا متقبلاً، وقلبًا سليمًا، ولسانًا ذاكرًا، ورزقًا مباركًا في الدنيا والآخرة ، اللهم افتح لي أبواب الفهم والنجاح، وبارك لي في سعيي واجعل توفيقك رفيقي في كل خطوة أخطوها.

For any feedback , click here