King Saud University College of Business Administration Department of Quantitative Analysis



PSPP applications on Business Statistics (QUA 107 +QUA207)

Dr. Manahil Kamal M. Eltayeb Assistant professor King Saud University Department of Quantitative Analysis (<u>maltib@ksu.edu.sa</u>)

Introduction (Variable & Data View / The data coding / The data Entering)

PSPP is a program for statistical analysis of sampled data. It is a free as in replacement for the proprietary program SPSS, and appears very similar to it with a few exceptions. (https://www.techopedia.com/definition/21531/pspp)

Opening PSPP

Start \rightarrow All Programs \rightarrow PSPP

Preparation of Data Files

Before analysis can commence, the data must be loaded into PSPP and arranged such that both PSPP and humans can understand what the data represents. There are two aspects of data:

The variables: these are the parameters of a quantity, which has been measured or estimated in some way. For example: height, weight and geographic location are all variables.

The observations (also called 'cases') of the variables — each observation represents an instance when the variables were measured or observed.



The following is a brief explanation of the main menus in the program.

- **File** includes most of the options you typically use in other programs such as open, save, create new files ...etc.
- Edit includes the cut, copy, and paste, go to variable or case and other Options...etc.
- **View** allows you to select which toolbars you want to show, select font size, add or remove the gridlines that separate each piece of data, and to select whether or not to display your raw data or the data labels.
- **Data** allows you to select several options ranging from displaying data that is sorted by a specific variable to selecting certain cases or weight cases for subsequent analyses.
- **Transform** includes several options to change current variables. For example: change the coding, compute new variables... etc.

- Analyze includes most of the commands to carry out statistical analyses. Much of this summary will focus on using commands located in this menu.
- **Graphs** includes the commands to create various types of graphs including histograms, scatterplot, and bar chart.
- **Utilities** allows you to list file information which is a list of all variables, there labels, values, locations in the data file, and type.
- Window can be used to select which window you want to view (i.e., Data Editor, Output Viewer, or Syntax
- **Help** has many useful options including a link to the SPSS homepage, a statistics coach, and a syntax guide. This is an excellent tool and can be used to troubleshoot most problems.

The Icons directly under the Menus provide shortcuts to many common commands that are available in specific menus.

When you open PSPP, you should be faced with the following screen:

1. Data view 2. Variable view

Data View	Variable View	

Variable View window (Defining Variables)

This window contains information about the variables set that is used; each row will provide information for each variable.

• Name:

PSPP has a number of rules for naming variables:

- An identifier, up to 64 bytes long. However, you should keep the variable name as short and succinct as possible.
- ✤ The name must begin with a letter. The remaining characters can be any letter, any digit, a full stop or the symbols @, #, _ or \$.
- ✤ Variable names cannot contain spaces or end with a full stop.
- Each variable name must be unique: duplication is not allowed.
- Reserved keywords cannot be used as variable names. Reserved keywords are : ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO, WITH.

• Type:

The most common choice is "**numeric**," which means the variable has a numeric value. The other common choice is "**string**," which means the variable is in text format. Below is a table showing the data types:

Туре	Width	Decimal	Label	Value Labels
	8	2		None 🗾

6	PSPPIRE.ex	œ		×
Numeric Comma	Width: 8 Decimal Places: 2	в — 2 —	+++	ОК
 Dot Scientific notation Date Dollar)	Cancel
 Custom currency String 				Help

This column enables you to specify the type of variable.

Туре	Example
Numeric	23456789
Comma	23,456,789
Dot	23.456.789
Scientific	34567 >>>3E+004
	12000>>> 1E+004
Date	01-Feb-2000
Dollar	<mark>\$</mark> 12,345,678 ,
Custom currency	<mark>SR</mark> 12,345
String	A, B, C

• Width

Width allows you to determine the number of characters PSPP will allow to be entered for the variable

• Decimals

PSPP defaults to two decimal places. Since our data does not require decimal places we can simply click in the Decimals cell and click the up or down arrows to adjust decimal places needed for that particular variable.

• Label

The Label column allows you to provide a longer description of your variable, which will be shown in the output produced by PSPP.

• Values

Values are code (number or letter) assigned to categories for nominal/ordinal variables, for example (male = 1 and female = 2).

Decimal	Label	Value Labels	Missing Values
3			None

6	PSPPIRE.exe	×
Value Labels		
Value:		
Value Label:		Cancel
Add]	Calicer
Apply		
Remove		Help

• Missing value

Sometimes it is useful to assign specific values to indicate different reasons for missing data. However, PSPP recognizes any blank cell as missing data and excludes it from any calculations, so if you intend to leave the cell blank there is no need to enter values for missing data.

el	Value Labels	Missing Values	Column	Align	М
	None 🛄		8	Right 🚍	Sc

6	PSPPIRE.exe	×						
 No missing value 	lues							
 Discrete missing values 								
Range plus o	ne optional discrete missing value							
Low:								
High:								
Discrete value:								
ок	Cancel Help							

• Columns

You can change the column width to reduce the space it takes on the screen. However, you need to allow enough space for variable names, so the default of eight is usually OK. Align :

_		
	Left	÷.
Colum	Center	÷
8	Right	=

• Measure :

		Nominal	3
Column	Align	Ordinal	
8	Right	 Scale	E

- Scale: For numeric values on an interval or ratio scale: age, sessions, satisfaction.
- Nominal: For values that represent categories with no intrinsic order: patient, sex, counsellor.
- Ordinal: For values with some intrinsic order (e.g., low, medium, high; first, second , third)

Entering Data into PSPP

Example (1):

- Switch from variable view to data view - Establish that all labels are evident across the top row of the data view window. - Once this has been established it is possible to begin inputting data.

Definition of variables & Data entry

For creating a database for a company, a questionnaire was distributed to a sample of 20 workers. The following questions are part of the questionnaire.

Q1-Gender Male 1 Female 0											
Q2 - Age Year											
Q3 - Marital status: Married 1 Not Married 0											
Q4- Years of Experience: < 5 1 $5 - 10$ 2 > 10 3											
Q5 - Educational level:											
Graduate G University U Secondary S Primary P											
Q6 - Monthly income:SR											
Q7 - Monthly expenditure:SR											

Definition of variables:

File Edit View Data Transform Analyze Graphs Utilities Windows Help

10

Ē	÷	č X	Ŝ			₽											
Variabl	Name	Туре		Width	Decimal	Label	Value Labels		Missing Values		Column	Align		Measure		Role	
1	Gender	Numeric		1	0	Gender	{0, Female}		None		6	Right	Ę	Nominal	••	Input	đ
2	Age	Numeric		2	0	Age	None		None		8	Right	Ę	Scale		Input	đ
3	Status	Numeric	:	6	0	Marital status	{0, Not Married}		None		12	Right	ų.	Nominal	••	Input	¢€
4	Experience	Numeric		8	0	Years of Experience	{1, < 5}		None		8	Right	ų.	Ordinal	ııl	Input	¢(
5	Education	String		1		Educational level	{G, Graduate}		None		10	Right	÷	Ordinal	IÅ	Input	¢
6	Income	Numeric		5	0	Monthly income	None		None		8	Right	÷	Scale		Input	¢
7	Expenditure	Numeric		5	0	Monthly expenditure	None		None		8	Right	Ę	Scale		Input	¢.
Dat	a entra	7	_				,	_		_							

Data entry

Case	Gender	Age	Status	Experienc	Education	Income	Expenditu
1	1	24	0	1	U	3500	3500
2	0	34	1	2	P	5000	4000
3	0	25	0	3	U	5500	4500
4	1	26	0	1	U	6000	5530
5	1	36	1	3	G	12300	12000
6	0	44	1	3	S	15000	15300
7	1	56	1	3	S	17500	20500
8	0	45	1	3	Р	12000	11500
9	1	29	1	2	U	7000	6500
10	0	38	0	3	G	14000	14000
11	0	43	1	3	Р	14750	13000
12	1	55	1	3	S	17000	15000
13	1	50	1	3	U	15450	14400
14	0	43	1	3	S	15000	15000
15	1	28	1	2	U	7000	<mark>6500</mark>
16	0	53	1	3	S	19000	17000
17	1	28	0	1	U	6200	7000
18	0	31	0	2	G	8500	8000
19	1	39	0	3	U	12500	12000

PSPP applications on Principles of Statistics QUA 107

Chapter (3) & (4)

Organizing Variables / Numerical Descriptive Measures Example (2): Refer to example (1):

Construct a frequency table and calculate the numerical measures for the following variables: Gender, Marital status, Age

Solation:

1) Descriptive statistics ... Frequencies

This option calculates the measures of dispersion and central tendency of quantitative variables and related drawings.

Analyze Graphs Utilities Wind	dows Help	
Descriptive Statistics	Frequencies	
Gender Marital status I Years of Experience Monthly income Monthly expenditure	Variable(s):	ок
	F	°aste
	Statistics: Mean Standard deviation Minimum Maximum Standard error of the mean Variance Skewness	ancel
	Standard error of the skewness Range Mode Kurtosis Standard error of the kurtosis Median Sum Include missing values	eset
	Charts Frequency Tables	lelp
F	requencies: Charts	
Exclude values below	0 - + Continue	
Exclude values above		
Histograms	es Percentages	
 Draw histograms Superimpose norma 	Cancel	
Bar Charts Draw bar charts Pie Charts		
Draw pie charts	ssing values Help	

The output:

e							
alue	Value	Frequency	/	Percent	t	Valid	Cum
bel						Percent	Percent
	24	1		5.00		5.00	5.00
	25	1		5.00		5.00	10.00
	26	1		5.00		5.00	15.00
	28	2	2	10.00		10.00	25.00
	29	1	.	5.00		5.00	30.00
	31	1		5.00		5.00	35.00
	34	1		5.00		5.00	40.00
	36	1		5.00		5.00	45.00
	38	1		5.00		5.00	50.00
	39	1	.	5.00		5.00	55.00
	42	1	.	5.00		5.00	60.00
	43	2	2	10.00		10.00	70.00
	44	1	.	5.00		5.00	75.00
	45	1	.	5.00		5.00	80.00
	50	1	.	5.00		5.00	85.00
	53	1	.	5.00		5.00	90.00
	55	1	.	5.00		5.00	95.00
	56	1		5.00		5.00	100.00
	Total	20)	100.0		100.0	
e							
	V	alid		20			
	М	lissina		0			
ean		5		38.42			
E. Me	ean			2.30			
ode							
d De	v			10.27			
ariand	ce			105.45			
ewn	ess			.22			
ange				32.50			
inimu	m			23.50			
aximu	um			56.00			
rcen	<i>tiles</i> 50) (Median)		39			
	e ean E. Mo d Dev ariano rewno ange inimu rcen	Ive Value bel 24 25 26 28 29 31 34 36 38 39 42 43 44 45 50 53 55 56 Total e W ean K E. Mean M ode M odd Dev M priance Songe inimum Songe inimum Songe inimum Songe	Value Frequency bel 24 1 25 1 26 1 28 2 29 1 31 1 34 1 37 1 38 1 39 1 42 1 43 2 44 1 50 1 53 1 55 1 56 1 70tal 20 e Valid Missing 1 56 1 56 1 56 1 70tal 20 e Valid Missing 1 56 1 70tal 20 e Name ean 50 E. Mean 1 57 1 58 1	Ive Value Frequency bel 24 1 25 1 26 1 28 2 29 1 31 1 34 1 36 1 38 1 39 1 42 1 43 2 44 1 50 1 53 1 55 1 56 1 70tal 20 e Valid Missing 1 56 1 56 1 70tal 20 e Valid Missing 1 56 1 70tal 20 ean 1 E. Mean 1 ode 1 d Dev 1 oriance 1 aximum	Interview Value Frequency Percent bel 24 1 5.00 25 1 5.00 26 1 5.00 28 2 10.00 29 1 5.00 31 1 5.00 34 1 5.00 36 1 5.00 38 1 5.00 39 1 5.00 42 1 5.00 43 2 10.00 44 1 5.00 50 1 5.00 55 1 5.00 55 1 5.00 55 1 5.00 70tal 20 100.0 ean . . Valid 20 0 0 38.42 2.30 0 . . . 10.27 105.45 .22 ange	Interview Value Frequency Percent 24 1 5.00 25 1 5.00 25 1 5.00 26 1 5.00 28 2 10.00 29 1 5.00 31 1 5.00 34 1 5.00 34 1 5.00 38 1 5.00 38 1 5.00 38 1 5.00 39 1 5.00 42 1 5.00 43 2 10.00 44 1 5.00 50 1 5.00 55 1 5.00 55 1 5.00 55 1 5.00 56 1 5.00 56 1 5.00 56 1 5.00 56 1 5.00 6an 2.30 0 0 0 0 ean . . 2.30	Interview Value Frequency Percent Valid Percent 24 1 5.00 5.00 25 1 5.00 5.00 26 1 5.00 5.00 28 2 10.00 10.00 29 1 5.00 5.00 31 1 5.00 5.00 34 1 5.00 5.00 36 1 5.00 5.00 38 1 5.00 5.00 39 1 5.00 5.00 42 1 5.00 5.00 43 2 10.00 10.00 44 1 5.00 5.00 50 1 5.00 5.00 55 1 5.00 5.00 55 1 5.00 5.00 55 1 5.00 5.00 56 1 5.00 5.00 56 1

2) Descriptive statistics ... Descriptive This option calculates the measures of dispersion and central tendency of quantitative variables (except some) and Z values.

Analyze	Graphs	Utilities	Windows	He	elp
Descri	ptive Stati	stics			Frequencies
Comp	are Means	•	Descriptives		

Gender	Variab	les:	ОК
🐣 Marital status	Age		
II Years of Experience			
Monthly income			
Monthly expenditure			
	•		
			Paste
	Statisti	ics: Mean	
		Standard dev	Cancel
		Minimum	
		Maximum	
		Range	
		Sum	
		Standard err	
		Variance	Reset
		Kurtosis	
2		Skewness	
Deptions: Exclude entire case if any s	elected variable	e is missing	
Include user-missing data in	n analysis		
Save Z-scores of selected v	ariables as nev	v variables	Help

The output:

Mapping of variables to corresponding Z-scores.

Source	Target
Age	ZAge

Valid cases = 20; cases with missing value(s) = 0.

Variabl	e N	1	Mean	S.E. Mean	Std Dev	Varian	се	Skewnes	55	S.E. Skew	Ra	nge	Minim	um	Maximu	m	Sum
Age	20	3	8.42	2.30	10.27	105.4	15	.2	2	.51	32	.50	23.	50	56.0	00	768.50
Case	Gend	er	Age		Status		E	xperienc	E	ducatio	n	Inc	ome	Ex	penditu	Z	Age
1		1		24		0		1			U		3500		3500		-1.45
2		0		34		1		2			Ρ		5000		4000		43
3		0		25		0		3			U		5500		4500		-1.31
4		1		26		0		1			U		6000		5530		-1.21
5		1		36		1		3			G		12300		12000		24
6		0		44		1		3			s		15000		15300		.54
7		1		56		1		3			s		17500		20500		1.71
8		0		45		1		3			Ρ		12000		11500		.64
9		1		29		1		2			U		7000		6500		92

3)

Descriptive statistics ... Explore



Outputs by Quantitative Variable (Age)

	Cuber	1000	soning o	unn	initian y						
						Cases				1	
			Valid		/	Missing		То	tal		
		N	Perce	ent	N	Percent	N	P	ercent		
	Age	20	100	%	0	0%	20		100%		
	Extron	no 1/2	luos							4	
	LAUEI	ne va	lues		Card	Number	1/2/	100	1		
ł	4.00	Lliah	oct 1	╬	Case	7	100	56	1		
	Aye	nigi	1051 1	,		12		55			
			3	3		16		53			
			4	ŧ		13		50			
			5	5		8	4	45			
		Low	est 1	L		1	:	24			
			2	2		3		25			
			3	8		4		26			
			4			15	:	28			
U						17		20			
	Percer	ntiles	<u> </u>	_							
					_			Pe	ercentile	?S	
				L	5	10	25	_	50	75	90
	Age	HAV	erage	2	3.57	25.10	28.2	5	38.50	44.75	54.80
		Hind	ey s 165				28.5		38.50	44.50	
1		, mig	100								
	Desc	riptive	s								
											Statistic
	Age	Me	an								38.42
		95	% Conf	Ide	nce I	nterval for	mean	1	Lower E	sound	33.62
		5%	Trimn	nod	l Moa	n			opper e	ouna	43.23
		Me	dian	lee	i meu						38.50
		Va	riance								105.45
		Sto	l. Devia	tio	n						10.27
		Mir	nimum								23.50
		Ma	ximum								56.00
		Ra	nge		_						32.50
		Int	erquart	ile	Rang	e					16.50
1		SK	ewness								.22

Case Processing Summary

Kurtosis

<u>95</u>

55.95

Std. Error

2.30

.51

.99

-1.08

Outputs by qualitative variable (Age & Gender)

Case I	Case Processing Summary											
		Cases										
			Valid		Missing	Total						
	Gender	N	Percent	N	Percent	N	Percent					
Age	Female	9	100%	0	0%	9	100%					
	male	11	100%	0	0%	11	100%					

Extreme Values

	Gender			Case Number	Value
Age	Female	Highest	1	16	53
			2	8	45
			3	6	44
			4	14	43
			5	11	43
		Lowest	1	3	25
			2	18	31
			3	2	34
			4	10	38
			5	11	43
	male	Highest	1	7	56
			2	12	55
			3	13	50
			4	20	42
			5	19	39
		Lowest	1	1	24
			2	4	26
			3	15	28
			4	17	28
			5	9	29

Percentiles

			Percentiles									
	Gender		5	10	25	50	75	90	<i>95</i>			
Age	Female	HAverage	12.50	25.00	32.50	43.00	44.50	53.00	53.00			
		Tukey's Hinges			34.00	43.00	44.00					
	male	HAverage	14.10	24.00	28.00	36.00	50.00	55.80	56.00			
		Tukey's Hinges			28.00	36.00	46.00					

	Gender			Statistic	Std. Error
Age	Female	Mean		39.56	2.82
		95% Confidence Interval for Mean	Lower Bound	33.05	
			Upper Bound	46.06	
		5% Trimmed Mean		39.62	
		Median		43.00	
		Variance		71.53	
		Std. Deviation		8.46	
		Minimum		25.00	
		Maximum		53.00	
		Range		28.00	
		Interquartile Range		12.00	
		Skewness		30	.72
		Kurtosis		10	1.40
	male	Mean		37.50	3.58
		95% Confidence Interval for Mean	Lower Bound	29.52	
			Upper Bound	45.48	
		5% Trimmed Mean		37.25	
		Median		36.00	
		Variance		141.05	
		Std. Deviation		11.88	
		Minimum		23.50	
		Maximum		56.00	
		Range		32.50	
		Interquartile Range		22.00	
		Skewness		.52	.66
		Kurtosis		-1.28	1.28

4) **Descriptive statistics ... Crosstabs**

This option is for designing intersecting tables (contingency tables) of qualitative variables and illustrates percentages by columns, rows, and totals and chi square test.





1.10

Summary.

				Cases		
		Valid		Missing	Total	
	N	Percent	N	Percent	N	Percent
Gender * Educational level	20	100.0%	0	0.0%	20	100.0%

Gender * Educational level [count, row %, column %, total %].

Gender	Graduate	Primary	Secondary	University	Total
Female	2.00	3.00	3.00	1.00	9.00
	22.22%	33.33%	33.33%	11.11%	100.00%
	66.67%	100.00%	60.00%	11.11%	45.00%
	10.00%	15.00%	15.00%	5.00%	45.00%
male	1.00	.00	2.00	8.00	11.00
	9.09%	.00%	18.18%	72.73%	100.00%
	33.33%	.00%	40.00%	88.89%	55.00%
	5.00%	.00%	10.00%	40.00%	55.00%
Total	3.00	3.00	5.00	9.00	20.00
	15.00%	15.00%	25.00%	45.00%	100.00%
	100.00%	100.00%	100.00%	100.00%	100.00%
	15.00%	15.00%	25.00%	45.00%	100.00%

Chi-square tests.

Statistic	Value	df	Asymp. Sig. (2-tailed)
Pearson Chi-Square	8.87	3	.031
Likelihood Ratio	10.70	3	.013
N of Valid Cases	20		

Example (3): Refer to example (1):

Draw a bar chart to the following:

- 1. The number of males and females in the sample
- 2. Average monthly income by level of education.
- 3. Average monthly expenditure by level of education and Marital status

Solation:

The number of males and females in the sample

-	Table(1) Exam	ipie (1,2,3,3,0).sav
File Edit View Data Tra	ansform Analyze Graphs Utilities Windows Help	
🕒 🛓 🔞	Scatterplot Histogram	•
	Barchart	
<i>•</i>	Barchart	×
Gender	Bars Represent N of cases O % of cases	ОК
Marital status	Cum. n of cases Cum. % of cases	Paste
Monthly income	Variable:	Cancel
Monthly expenditure Z-score of Age	Category Axis: Gender	Reset
	Category Cluster:	Help



Average monthly income by level of education.

6 Barchart Bars Represent 🖧 Gender ОК N of cases % of cases
 🛛 Age Cum. n of cases Cum. % of cases 🖧 Marital status Paste Years of Experience Other summary function Mean . Ŧ Educational level Variable: Monthly income Cancel Income Monthly expenditure Category Axis: Z-score of Age ٠ Reset Education Category Cluster: ۲ Help



3. Average monthly expenditure by level of education and Marital status

é	Barchart	×
Gender	Bars Represent O % of cases	ОК
Marital status	Cum. n of cases Cum. % of cases	Paste
Monthly income Monthly expenditure	Expenditure	Cancel
Z-score of Age	Category Axis: Experience	Reset
	Status	Help



Years of Experience

Example (4): Refer to example (1): Draw a pie chart to the education level variable

Analyze Graphs Guides Windows help	Utilities Windows Help	Help	Windows	Utilities	Graphs	Analyze		
Descriptive Statistics Frequencies	istics Frequencies.	•	Descriptive Statistics					

6	Frequencies	×
 Gender Age Marital status Years of Experience Monthly income 	Variable(s): Education Statistics:	OK Paste Cancel
 Monthly income Monthly expenditure Z-score of Age 	Standard deviation Include missing values Charts Frequency Tables	Reset

🗉 F	Frequencies: Charts								
Chart Formatting	Chart Formatting								
Exclude values below	0	—	+	Continue					
Exclude values above	Exclude values above 100 - +								
Scale: 💿 Frequencie	s 🕘 Percei	ntages	5						
Histograms									
Draw histograms				Cancel					
Superimpose norma	l curve								
Bar Charts									
Draw bar charts									
Pie Charts	Pie Charts								
🗹 Draw pie charts	🗹 Draw pie charts								
Include slices for mis	ssing values			Help					

Educational le	evel				
Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
Graduate	G	3	15.00	15.00	15.00
Primary	Р	3	15.00	15.00	30.00
Secondary	S	5	25.00	25.00	55.00
University	U	9	45.00	45.00	100.00
	Total	20	100.0	100.0	



Example (5): Refer to example (1): Draw histogram to the income





PSPP applications on statistical tests hypotheses & Confidence Interval 207 QUA

Chapter 8 & 9: Confidence Interval and One-Sample Tests

Hypothesis Tests for one mean One sample T- test

Example (1):

The McFarland Insurance Company Claims Department reports the mean cost to process a claim is \$60. An industry comparison showed this amount to be larger than most other insurance companies, so the company instituted cost-cutting measures. To evaluate the effect of the cost-cutting measures, the Supervisor of the Claims Department selected a random sample of 26 claims processed last month. The sample information is reported below.

No	1	2	3	4	5	6	7	8	9	10	11	12	13
Mean Cost (Claims) \$	45	49	62	40	43	61	48	53	67	63	78	64	48
No	14	15	16	17	18	19	20	21	22	23	24	25	26
Mean Cost (Claims) \$	54	51	56	63	69	58	51	58	59	56	57	38	76

- Determine the 95 percent confidence interval for the population mean.
- Calculate the mean and standard error of mean.
- At the .05 significance level, is it reasonable a claim is now less than \$60? **Solution:**

	<u></u>	*one t-test 207.sav [DataSet1] — PSPPIRE Data Editor
n	Analyze Graphs Utilities Windows	Help
1	Descriptive Statistics	
9	Compare Means	Means
	Univariate Analysis	One Sample T Test
_	Bivariate Correlation	Independent Samples T Test
	K-Means Cluster	Paired Samples T Test
	Factor Analysis	One Way ANOVA
	Reliability	
	Regression	•
	Non-Parametric Statistics	•
	ROC Curve	
	Factor Analysis Reliability Regression Non-Parametric Statistics ROC Curve	one way ANOVA



$$H_0: \ \mu \ge \$ \ 60 \qquad H_1: \ \mu < \$ \ 60$$

P-value (0.081/2) > 0.01 Accept the null hypotheses

We have not demonstrated that the cost-cutting measures reduced the mean cost per claim to less than \$60. The difference of \$3.58 (\$56.42 - \$60) between the sample mean and the population mean could be due to sampling error.

Accept the null hypothesis and conclude there is not sufficient evidence that the cost-cutting measures reduced the mean cost per claim to less than \$60.

Hypothesis Tests for one Proportion

First method (Binomial test)

Example (2)

A person who claims to possess extrasensory perception (ESP) says she can guess more often than not the outcome of a flip of a fair coin. Out of 25 flips, she guesses correctly 15 times. Would you conclude that she truly has ESP?





			*one p	prop	ortion by Binomal.sav [DataSet1]
Analyze	Graphs	Utilities	Windows	Не	p
Descr	iptive Stati	stics		•	
Comp	are Means	5		•	H
Univa	riate Analy	sis			
Bivaria	ate Correla	ation			
K-Mea	ans Cluster	r			
Factor	r Analysis.				
Reliab	oility				
Regre	ssion		_	•	
Non-P	arametric	Statistics		•	Chi Square
ROC (Curve				Binomial
				٦.	Runs
					1 Sample K-S
					2 Related Samples
					K Related Samples
					K Independent Samples

e de la companya de la compa	Bi	nomial Test		×
The correct number of guesses	Þ	Test Variable List:		OK Paste
Oet from data		Test Proportion: 0.5	5	Reset
				Help

Binomial Test						
		Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
Total number of flips	Group1	successes	15	.60	.50	.424 P-Value
	Group2	failures.	10	.40		
	Total		25	1.00		

The p-value of .424, which is the value for a two-tailed test. The appropriate p-value for the test we are conducting is half of .424, or p = .212

Note that SPSS uses a method based on the binomial distribution, which may not exactly match the values from the other calculation methods. However, these results will still lead to the appropriate decision to fail to reject the null hypothesis

Second method (Chi-square/Z-test) Example (3)

A marketing company claims that it receives 8% responses from its mailing. To test this claim, a random sample of 500 were surveyed with 25 responses. Test at the $\alpha = 0.05$ significance level.



H₀: $\pi = 0.08$ H₁: $\pi \neq 0.08$

é	Weight Cases	×
response	Do not weight cases	OK Paste
	Frequency Variable	Cancel
	Current Status: Weight cases by Frequance	Reset Help

				*one pro	port	tion by	y Z (chi squ	iare).sav [Dat	as
I	Analyze	Graphs	Utilities	Windows	Hel	p			
I	Descri	ptive Stati	istics		۲	55			
	Compa	are Means	5		•	-			
	Univar	iate Analy	sis						
I	Bivaria	ate Correla	ation						
I	K-Mea	ns Cluster	r						
	Factor	Analysis.							
	Reliab	ility							
	Regre	ssion			•				_
	Non-P	arametric	Statistics			Chi	Square		
	ROC C	Curve				Bin	omial		I
						Rur	15		I
						1 S	ample K-S		I
						2 R	elated Samp	oles	L
						KR	elated Samp	oles	
						K Ir	ndependent	Samples	

response			
	Observed N	Expected N	Residual
response	25	40.00	-15.00
non-response	475	460.00	15.00
Total	500		
Test Statistics			
	response		
Chi-Square	6.11		
Cill Square			
df	1	P-value	
df	1	P-value	

Z-test

Rejection Region: Reject the null hypothesis if p-value ≤ 0.05 . Test Statistic: $Z = \sqrt{\chi^2} = \sqrt{6.11} = 2.47$ P-value = Asymp. Sig. (2-tailed) = 0.013 P-value (0.013) < 0.05 (Reject H₀) There is sufficient evidence to reject the company's claim of 8% response rate.

Chapter 10 & 11: Two-Samples Tests, One-Way ANOVA and Chi -square

Difference between Two Means Independent Samples ''T-Test''

Example (1)

We have the following data, which represents the statistic test scores for the two groups .Assuming both populations are approximately normal, is there a difference in the mean scores in a statistics test between psychology students and law students $(\alpha = 0.05)$?

Score(1)	Score(2)
" psychology students"	" law students"
95	72
84	69
93	65
79	73
78	75
87	80
83	85
91	68
78	
89	
91	
93	

 $H_0: \mu_1 - \mu_2 = 0$ VS $H_1: \mu_1 - \mu_2 \neq 0$

6						*t-test	(independent samp	les).sav [DataSet1] — PSP	PIRE Data Editor							-
File E	dit View Da	ta Transf	orm	Analy	yze Gra	phs Utilities Windows	Help									
Ŀ	L.	ß	Ŝ			P										
Variabl	Name	Туре	1	Width	Decimal	Label	(Value Labels	Missing Values	Colum	Align	1	Measure		Role	
1	Score	Numeric	8		0	All scores (psychology &	law students)	None	None .	8	Right	= s	Scale		Input	¢
2	Students	Numeric	8		0	psychology & law student	:S		None	. 19	Right	= N	lominal	••	Input	ц,
3		(
							6	PSPPIRE.	exe		×					
							Value Labels Value:			ОК						
							Add	1 = `Score(1) " psychology "		Cancel						
							Remove	2 = `Score(2) " law		Help						

File	Edit	View	Data	Tra	ansform	Ar	nalyze	Gra	aphs	Utilitie	es	Wind	lows	Help)		
Ľ		L	6	×	õ		Q		ô	Ę	ŝ				Þ	٩	D
Case	S	core	Studer	nts													
	1	95	5	Core	e(1) ychology '	II											
	2	84	9	Core psy	e(1) ychology '	"											
	3	93	9	Core psy	e(1) ychology '	"											
	4	79	9	Core psy	e(1) ychology '	"											
	5	78	9	Core psy	e(1) ychology '	H											
	6	87	9	Core psy	e(1) ychology '	n											
	7	83	9	Core psy	e(1) ychology '	"											
	8	91	9	Core psy	e(1) ychology '	"											

	<u> </u>	*t-test(inc	lependent samples).sav [DataSet1] —
n	Analyze Graphs Utilities	Windows	He	elp
1	Descriptive Statistics		►	
2	Compare Means		•	Means
	Univariate Analysis			One Sample T Test
	Bivariate Correlation			Independent Samples T Test
	K-Means Cluster			Paired Samples T Test
JY	Factor Analysis			One Way ANOVA
_	Reliability			
IV	Regression		►	
,,	Non-Parametric Statistics		►	
	ROC Curve			

6	Independent-Samples T Test		×			
🌲 psychologyw students	Test Variable(s):		ОК			
	Score					
	Grouping Variable:		Cancel			
	Students		Reset			
	Define Groups Op	tions	Help			
6	Define Groups	×				
• Use specified v	alues:	Continue				
Group1 value:	Score(1) " psychology "					
Group2 value:	Score(2) " law" 🔹					
O Cut point:	•	Help				

The output

Group Statistics			X-Ba	ar (1 & 2)								
	psychology & lat	v students	N	Mean	Std. Devia	tion 5	S.E. Mean	1				
All scores (psychology & law stude	nts) Score(1) " psychology		12	86.75		5.20	1.79					
	Score(2) " law"		8	73.38		5.57	2.32					
Independent Samples Test						,				- 10 5		
		Leve	ne's T	est or E	quality of				t-test for	Equality of Means		
					,						95% Confidence Differ	Interval of the ence
		,	F		Sig.	t	df	Sig. (2- ta <u>iled</u>)	Mean Difference	Std. Error Difference	Lower	Upper
All scores (psychology & law	Equal variances	.0)5		.830	4.62	18.00	.000	13.38	2.90	7.29	19.46
students)	assumed Equal variances not assumed				Τ	4.56	14.52	.000	13.38	2.93	7.11	19.64
•				^	/			, /				

Here, you see there are two results from two different t-tests :

- Equal variance assumed (Pooled-Variance t test.)
- Equal variance not assumed (Separate-variance t test)

The choice of the result depends on the Levene's test. Since from A, the p-value of Levene's test is $0.83 \ge \alpha$ (0.05) we can assume that the variances of two groups are the equal. (If the p-value of Levene's test is < 0.05, we have to use the "Equal variance not assumed" result.

From **B**, since the p-value of t-test is $0.000 \le \alpha (0.05)$ we reject the null hypothesis and conclude that there is difference between the mean score of psychology students and law students at 5% significance level.

95% Confidence Interval for $\mu_1 - \mu_2 = (7.29 - 19.46)$

Difference between Two Means Related Populations the Paired Difference t-Test

Example (2)

For answering the question: Does a treatment reduce the level of anxiety? A sample of 7 people was taken and anxiety levels were measured before and after treatment. Is there a change in the result?

Patient	Before	After
1	40	24
2	42	30
3	36	37
4	31	21
5	55	32
6	45	40
7	46	47

$H_0: \mu_D = 0 \quad VS \quad H_1: \mu_D \neq 0$

Variabl	Name	Туре	Width	Decimal	Label	Value Labels	Missing Values	Column	Align	Measure		Role	
1	Before	Numeric	8	0	Before therapy	None	None	8	Right -	Scale	Ð	Input 🧣	÷
2	After	Numeric	8	0	After therapy	None	None	8	Right -	Scale	2	Input 🧣	÷

File Edit View Data Tr



45

46

40

47

6

7

8





Paired S	Sample Statistics	X-E	ar										
		Mean	N	Std. De	viation	S.E. Mea	an						
Pair 1	Before therapy	42.14	7		7.69	2.9	91						
	After therapy	33.00	7		9.09	3.4	14						
Paired S	Samples Correlatio	ons											
				N	Correla	tion Sig	ι.						
Pair 1	Before therapy 8	& After th	erapy	7		.45 .30	8						
Paired S	Samples Test												
								Paired D	ifferences				P-value
				_					95% Confidence Inte	rval of the Difference			- Value
				Mear	n Std.	Deviation	St	d. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	Before therapy -	After the	erapy	9.14	ł	8.86		3.35	.95	17.34	2.73	6	.034
			D-	Bar			SE(D)					

P-value (0.034) < α (0.05) (reject null hypotheses) The confidence interval for μ_D is : (0.95 – 17.34)

Difference between two proportions

Example (3)

In a test of the reliability of products produced by two machines, machine a produced 15 defective parts in a run of 280, while machine B produced 10 defective parts in a run of 200. Do these results imply a difference in the reliability of these two machines? (Use $\alpha = 0.05$.)

Solution:

Enter the group values (Machine: 1 = Machine A, 2=Machine B) into one variable, the quality values (Quality: 1=Defective, 2=Acceptable) into another variable, and the observed counts into a third variable.

 $H_0: \pi_1 - \pi_2 = 0$ (The two proportions are equal)

 $H_1: \pi_1 - \pi_2 \neq 0$ (There is a significant difference between proportions)



5							*2-porpo	rtions by	Chi square 8
File I	Edit View I	Data Transfo	orm Ana	alyze (Graphs	Utilities	Windows	Help	
Ŀ	Ţ.	ß	õ	Q	ô	ŝ		T e	•
				_					
Case	Machine	Quality	counts						
1	Machine A	Defective	15						
2	Machine A	Acceptable	265						
3	Machine B	Defective	10						
4	Machine B	Acceptable	190						

6										
File E	dit	View	D	ata	Transfo	rm	Ana	lyze	Grap	hs
Ŀ		Ŧ		Sor Tra	t Cases Inspose	•		Q		ô
				Ag	gregate					
Case	Ma	chine		Spl	it File					
1	м	achine		Sel	ect Cases	s				
2	м	achine		We	ight Case	s				
3	м	lachine	В	[Defective		10			
4	М	lachine	В	Ac	ceptable		190			
5										







6	Crosstabs: Statistics	×
Statistics Chisa		Continue
Phi		Cancel
CC		Help

=			Cro	sstabs	;			×	
Counts	;		.]	Rows	1.		ОК		
			•	ма	cnine		Paste		
			•	Colum	nns ality		Cance		
				Qui	uncy		Reset	=	
Form	nat	Statistic	cs		Ce	ls	Help	=	
	•	С	ross	tabs: C	Cells		×		
Summany	Colum Total Expect	n ed					Cancel		
Summary.			Ca	ases			7		
		Valid	Mis	ssing		Total			
		N Percent	N	Percent	N	Percent	-		
Machine * (Juality 48	0 100.0%	0	0.0%	480	100.0%			
Machine * Q	uality coun	t, total %].]					
Marchine	Q	uality							
Machine A	15.00	265.00	28						
indenne / (3.13%	55.21%	58	.33%					
Machine B	10.00	190.00	20	00.00					
Tabal	2.08%	39.58%	41	.67%					
I OLAI	5.21%	455.00 94.79%	100	.00%					
Chi-square t	ests.	chi -sta	t		Р	-value			
Statistic		Value	df	Asymp	. Sia. ((2-tailed)	Exact Sig. (2-ta	iled)	Exact Sig. (1-tailed)
Pearson Chi	-Square	.03	1	, r	~ ~ \	.862		-	
Likelihood R	atio	.03	1			.862			
Fisher's Exa	ct lest	00	1			1 000	1	.000	.518
Linear-by-Li	near Associa	ation .00	1			.862			
N of Valid C	ases	480							

You should use the output information in the following manner to answer the question:

Rejection Region: Reject the null hypothesis if p-value ≤ 0.05 .

P-value = Asymp. Sig. (2-tailed) = 0.862

Since p-value $(0.8622) > \alpha$ (0.05), we fail to reject the null hypothesis.

At the = 0.05 level of significance, there is not enough evidence to conclude that there is a difference in the reliability of the two machines.

Note: If you used the Z-test:

Test Statistic: $Z = \sqrt{\chi^2} = \sqrt{0.030} = 0.1735$

If the test were one-tailed, the p-value would be 1/2 (Asymp. Sig. (2-tailed)).

Analysis of variance One –way ANOVA

Example (4)

A manufacturer suspects that the batches of raw material furnished by her supplier differ significantly in calcium content. There is a large number of batches currently in the warehouse. Five of these are randomly selected for study. A chemist makes five determinations on each batch and obtains the following data.

Batch 1	Batch 2	Batch 3	Batch 4	Batch 5
23.46	23.59	23.51	23.28	23.29
23.48	23.46	23.64	23.40	23.46
23.56	23.42	23.46	23.37	23.37
23.39	23.49	23.52	23.46	23.32
23.40	23.50	23.49	23.39	23.38

Is there a significant variation in calcium content from batch to batch? Use $\alpha = 0.05$.

Solution:

H_o : $\mu_1 = \mu_2 = \mu_3 = \mu_4$ H_1 : Not all of the population means are equal



File I	Edit Vie	w Data	Transfor	m Analyze
	4			ð Q
Case	Calciun	batches		
1	23.46		1	
2	23.48		1	
3	23.56		1	
4	23.39		1	
5	23.40		1	
6	23.59		2	
7	23.46		2	
8	23.42		2	
9	23.49		2	
10	23.50		2	
11	23.51		3	
12	23.64		3	
13	23.46		3	
14	23.52		3	
15	23.49		3	

			*ANO	VA t	est.sav [Datas	Set1] -	— PSP	PI
Analyze Gra	phs Utilities	Windows	Help					
Descriptive	Statistics		•	7	•			
Compare N	1eans		► r	Mean	s			
Univariate	Analysis		•	One S	Sample T Test.			
Bivariate C	orrelation		1	indep	endent Sample	es T Te	st	
K-Means C	luster		F	aire	d Samples T Te	est		
Factor Ana	lysis			Dne \	Vay ANOVA			
Reliability								
Regression			•					
Non-Param	etric Statistics		•					
ROC Curve								
ANOVA								
		Sum of	Squares	df	Mean Square	F	Sig.	
Calcium content	Between Group	<i>ns</i>	.10	4	.02	5.54	.004	┝
	Within Groups		.09	20	.00			
	Total		.18	24				

Since the *p*-value (Sig) $(0.004) < \alpha (0.05)$, one can reject the null hypothesis that all means are equal. Thus, there is a significant variation in calcium content from batch to batch, for $\alpha = 0.05$.

(Chapter 12) Simple Linear Regression and correlation

Example (1)

A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet) .A random sample of 10 houses is selected (Dependent variable (Y) = house price in 1000s, Independent variable (X) = square feet).

Find:

- 1) The estimate regression equation (prediction line), and Interpret the slope and intercept of this problem.
- 2) Sum of square of regression (SSR) and Error sum of square (SSE).
- 3) Coefficient of Determination (R^2) and Interpret it.
- 4) Standard error of the estimate.
- 5) Is there a linear relationship between X and Y? (Use t-test)
- 6) Construct ANOVA table for regression to test that there is no significance relationship between X and Y by using F-ratio and t test.
- 7) Find \widehat{Y}
- 8) The correlation between X and Y.

House Price in \$1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

File Ed	dit View Da	ta Trans	form	Anal	yze Gra	phs Utilities Windows	Help					
	Ł	X X	Ŝ			4						
Variabl	Name	Туре		Width	Decimal	Label	Value Labels	Missing Values	Column	Align	Measure	Role
1	Y	Numeric		8	0	House Price in \$1000s	None .	 None	8	Right 🚍	Scale 📘	Input 🔒
2	Х	Numeric		8	0	Square Feet (X)	None .	 None	8	Right 🗧	Scale 📘	Input 🔒

10 c	10 cases × 1 variable									
Case	Y	×								
1	245	1400								
2	312	1600								
3	279	1700								
4	308	1875								
5	199	1100								
6	219	1550								
7	405	2350								
8	324	2450								
9	319	1425								
10	255	1700								





6	Regressior	Statistics	×
Statistics Coeff		[Continue
Conf. Inter		Cancel	
🗹 Anova			
Bcov			Help

Model Summary					
				Std. Error of the	
		R Square	Adjusted R	Estimate	
Model	R	(Coefficient of Determination)	Square	Syx	
1	.762 ^a	.581	.528	41.330	

NOVA ^a							
				df degrees of freedom		F	Sig.
Model		Sum	of Squares		Mean Square	(F-Stat)	P-value
1	Regressio n	<mark>SSR</mark>	18934.935	<mark>k =</mark> 1	18934.935	11.085	.010 ^b
	Residual	<mark>SSE</mark>	13665.565	$\frac{(n-k-1)}{8} =$	1708.196		
	Total	<mark>SST</mark>	32600.500	<mark>n-1</mark> = 9			

Coefficients^a

		Unstandardized Coefficients		Standardized Coefficients		Sig.
Model		В	Std. Error	Beta	t - Stat	P-value
1	(Constant)	<mark>t (b0)</mark> 98.248	58.033		1.693	.129
	Square Feet (X)	<mark>Slope(b1</mark>) .110	<mark>Sы</mark> .033	.762	<mark>3.329</mark>	<mark>.010</mark>

 $\hat{Y} = 98.25 + 0.1098X$ House price =98.25 + 0.1098 (sq.ft)

7) Find \widehat{Y}

Dependent OK Y Paste Independent Cancel X Statistics Save Help	e Regressio	ion	×
Independent Cancel X Reset Statistics Save Help	Depe	oendent OK	
Independent Cancel		Paste	
Statistics Save Help	Inde	ependent Cance	el
Statistics Save Help		Reset	t
	Statistics Sa	ave Help	
😂 Regression: Save 🗙	🧉 Regression	n: Save	
Predicted values Continue	Predicted values	Continue	
Cancel		Cancel	
Residuals Help	Residuals	Help	

Case	Y	х	PRED1	
1	245	1400	251.92	
2	312	1600	273.88	
3	279	1700	284.85	`` Y^
4	308	1875	304.06	
5	199	1100	218.99	
6	219	1550	268.39	
7	405	2350	356.20	
8	324	2450	367.18	
9	319	1425	254.67	
10	255	1700	284.85	
11				

The correlation between X and Y.

			/	
י 🕻	Analyze Graphs	Utilities	Windows	Hel
Γ	Descriptive Stati	stics		۲
2	Compare Means			•
L	Univariate Analy	sis		
K	Bivariate Correla	ition		
0	K-Means Cluster			
-	Factor Analysis			
1	Reliability			
8	Regression			•
9	Non-Parametric	Statistics		•
9	ROC Curve			

 Bivariate Correlations 	×
Y	ОК
	Paste
Test of Significance	Cancel
• Two-tailed 🔘 One-tailed	Reset
Flag significant correlations	Help

Correlations

		House Price in \$1000s	Square Feet (X)
House Price in \$1000s	Pearson Correlation	1.00	.76
	Sig. (2- tailed)		.010
	N	10	10
Square Feet (X)	Pearson Correlation	.76	1.00
	Sig. (2- tailed)	.010	
	N	10	10

There is a relatively strong positive linear relationship between XY