Basic biostatistics with SPSS

Looking beyond numbers

Jamalludin Ab Rahman MD MPH
Associate Professor
Department of Community Medicine
Kulliyyah of Medicine

Content

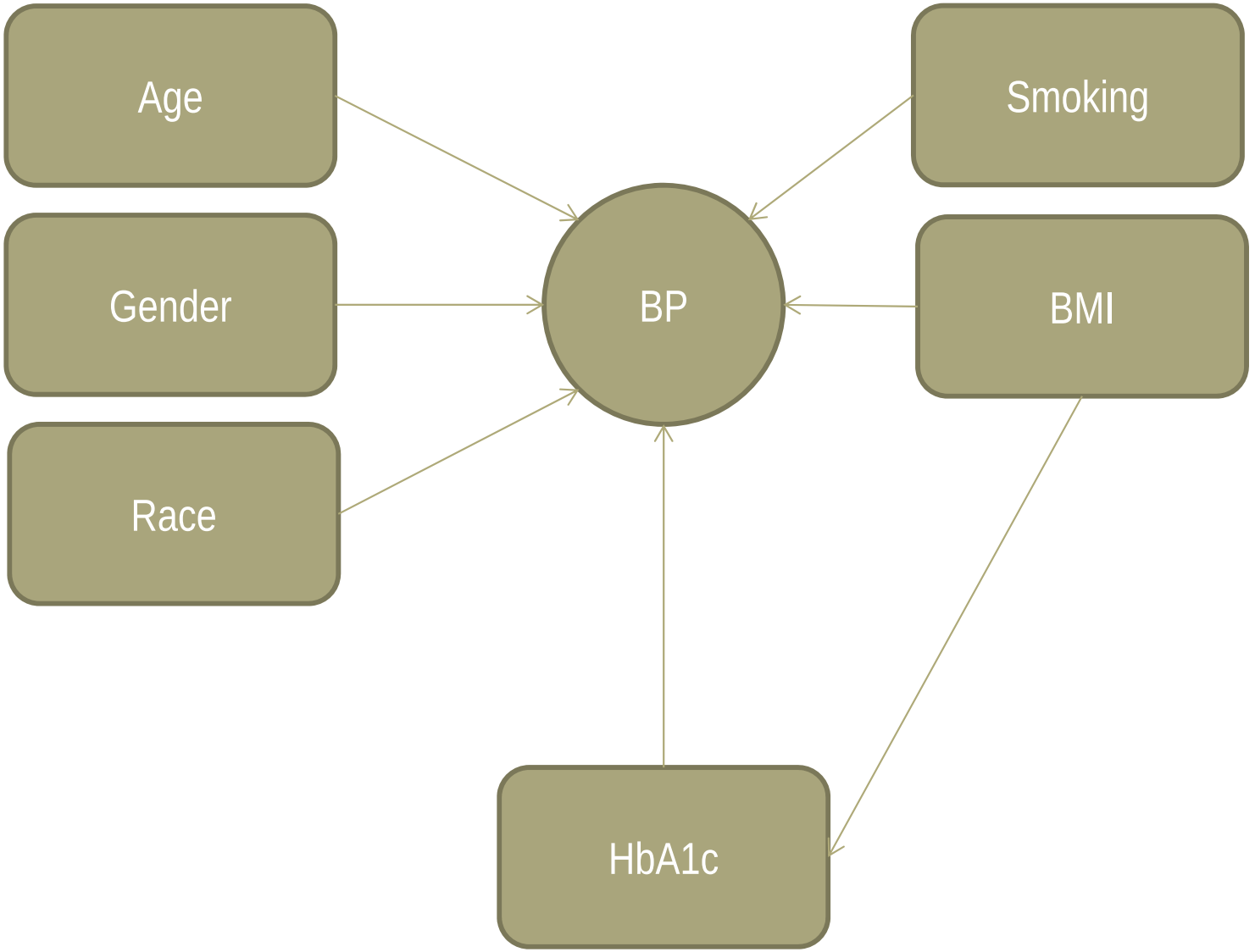
1. Describe data
2. Check distribution
3. Compare two means
4. Compare more than two means
5. Compare proportions
6. Non-parametrics
7. Correlation

Preliminary

- Use bp2.sav
- Hypothetical data consists of

Variable	*Position*	*Label*	*Measurement Level*
id	1	<none>	Scale
age	2	Age (years)	Scale
gender	3	Gender	Nominal
race	4	Race	Nominal
wt	5	Weight (kg)	Scale
ht	6	Height (cm)	Scale
smoking	7	Smoking status	Nominal
sbp	8	SBP	Scale
dbp	9	DBP	Scale
hba1c	10	HbA1c (%)	Scale

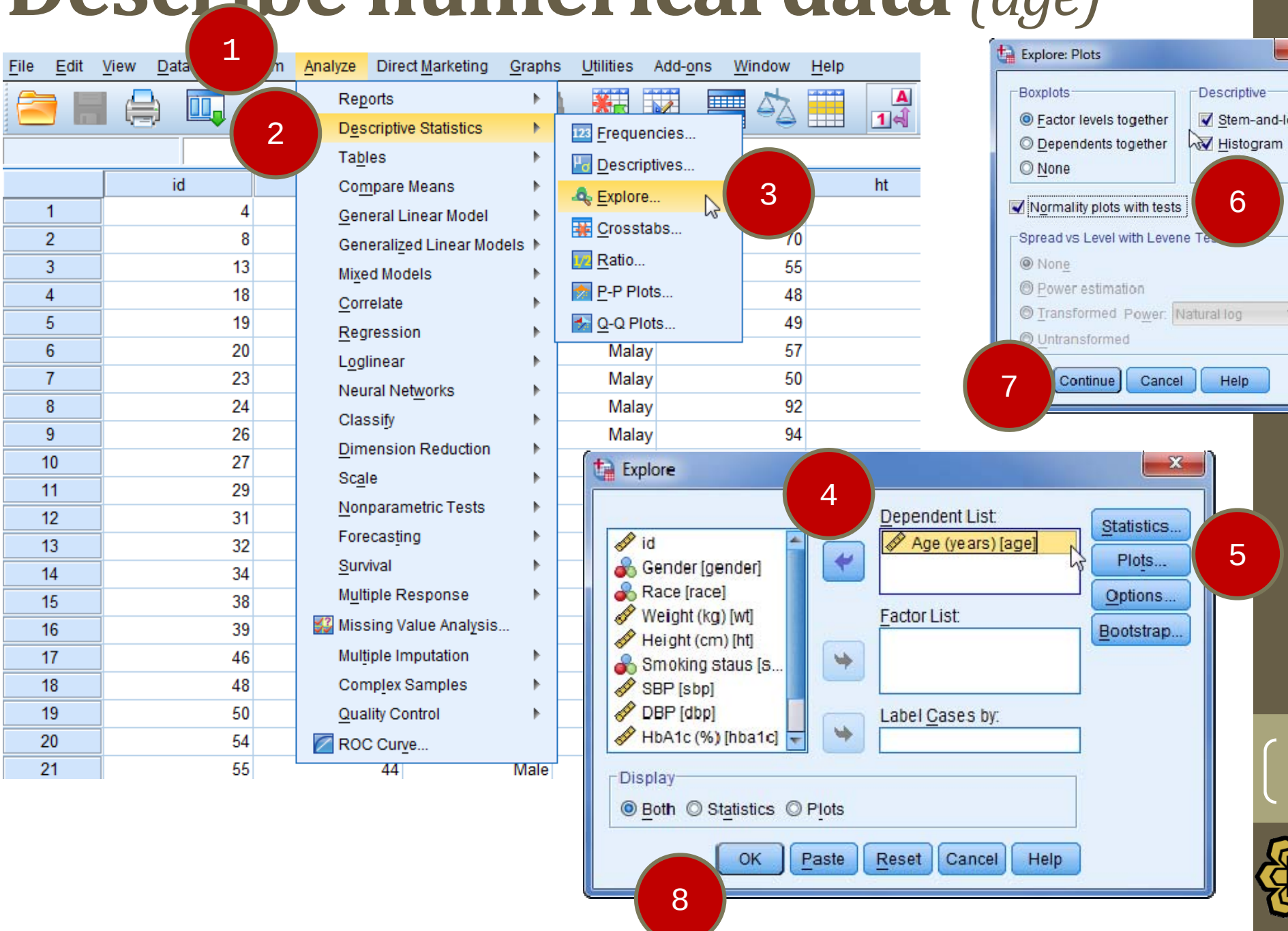
Conceptual framework



Basic biostatistics with SPSS

DESCRIBE DATA & CHECK DISTRIBUTION

Describe numerical data *(age)*



	id	ht
1	4	
2	8	70
3	13	55
4	18	48
5	19	49
6	20	57
7	23	50
8	24	92
9	26	94
10	27	
11	29	
12	31	
13	32	
14	34	
15	38	
16	39	
17	46	
18	48	
19	50	
20	54	
21	55	44 Male

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Age (years)	450	100.0%	0	0.0%	450	100.0%

Check for missing values

Descriptives

			Statistic	Std. Error
Age (years)	Mean		36.57	.534
	95% Confidence Interval for Mean	Lower Bound	35.52	
		Upper Bound	37.62	
	5% Trimmed Mean		36.33	
	Median		36.00	
	Variance		128.352	
	Std. Deviation		11.329	
	Minimum		18	
	Maximum		75	
	Range		57	
	Interquartile Range		19	
	Skewness		.281	.115
	Kurtosis		-.636	.230

Check for Normality (mean=median, skewness ± 2, kurtosis ± 2).

Decide to describe using mean (SD) or median (IQR).

Tests of Normality

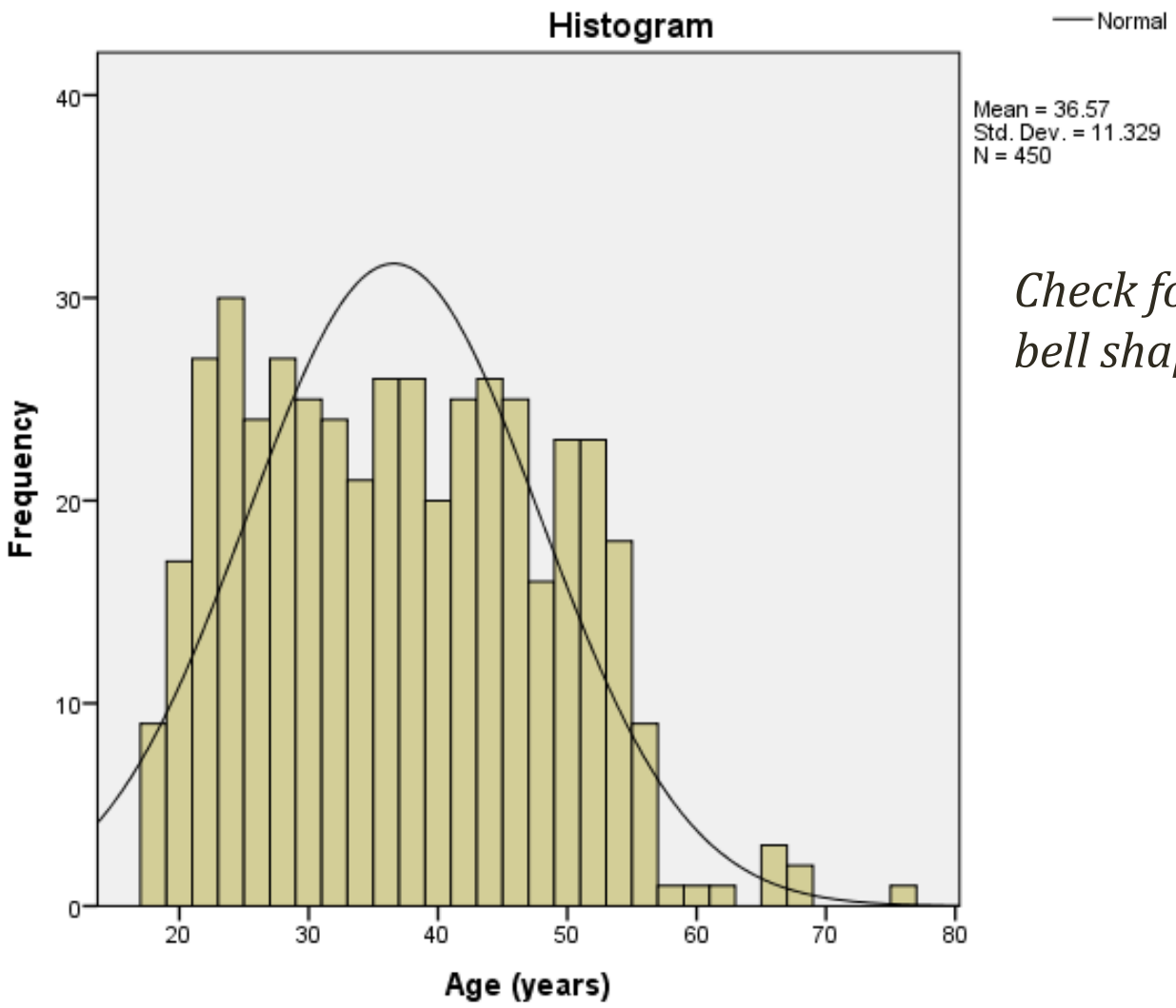
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Age (years)	.078	450	.000	.968	450	.000

a. Lilliefors Significance Correction

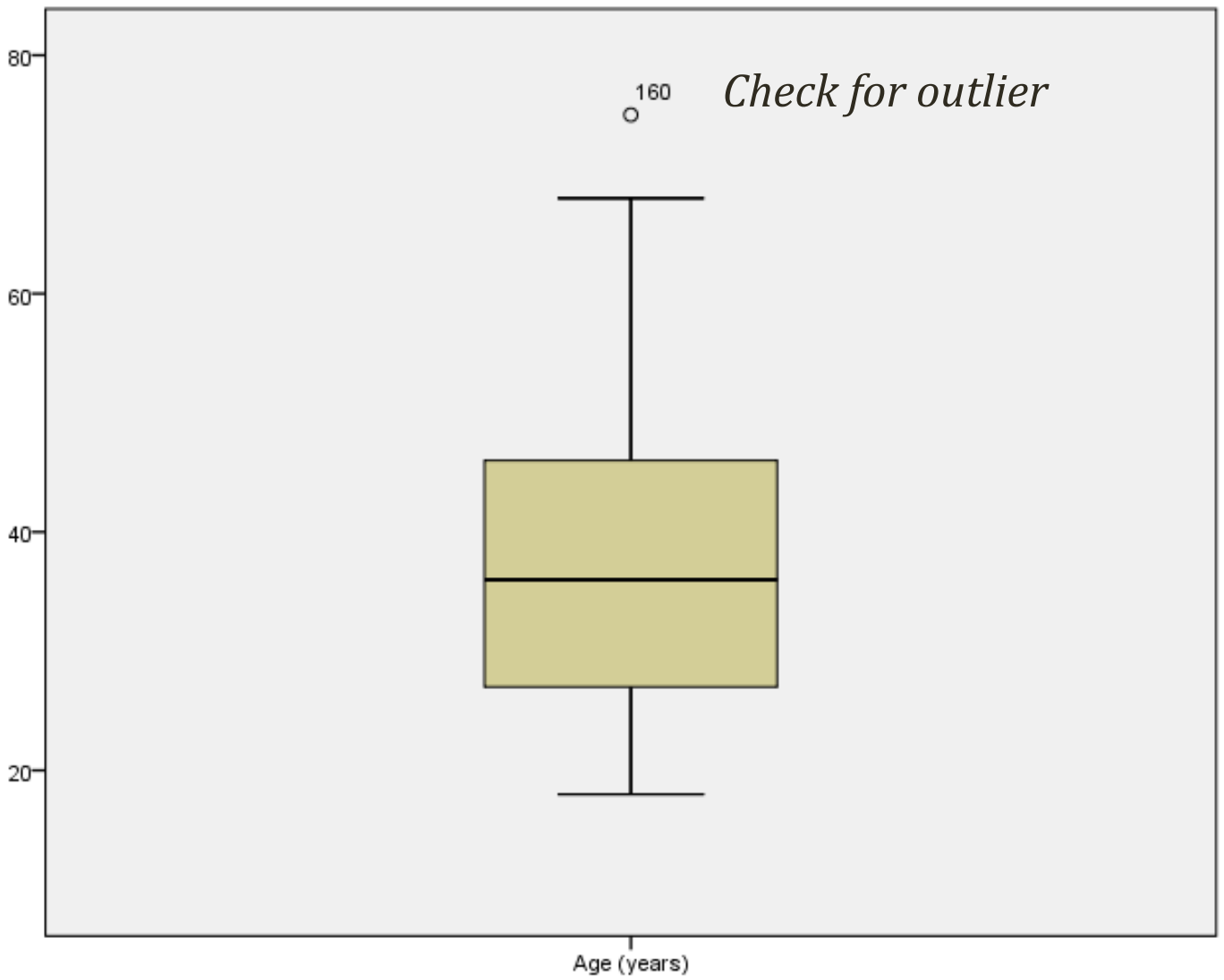
Test Normality. If P < 0.05, the distribution is not Normal.*

* *Careful when interpreting Normality test*

Age (years)

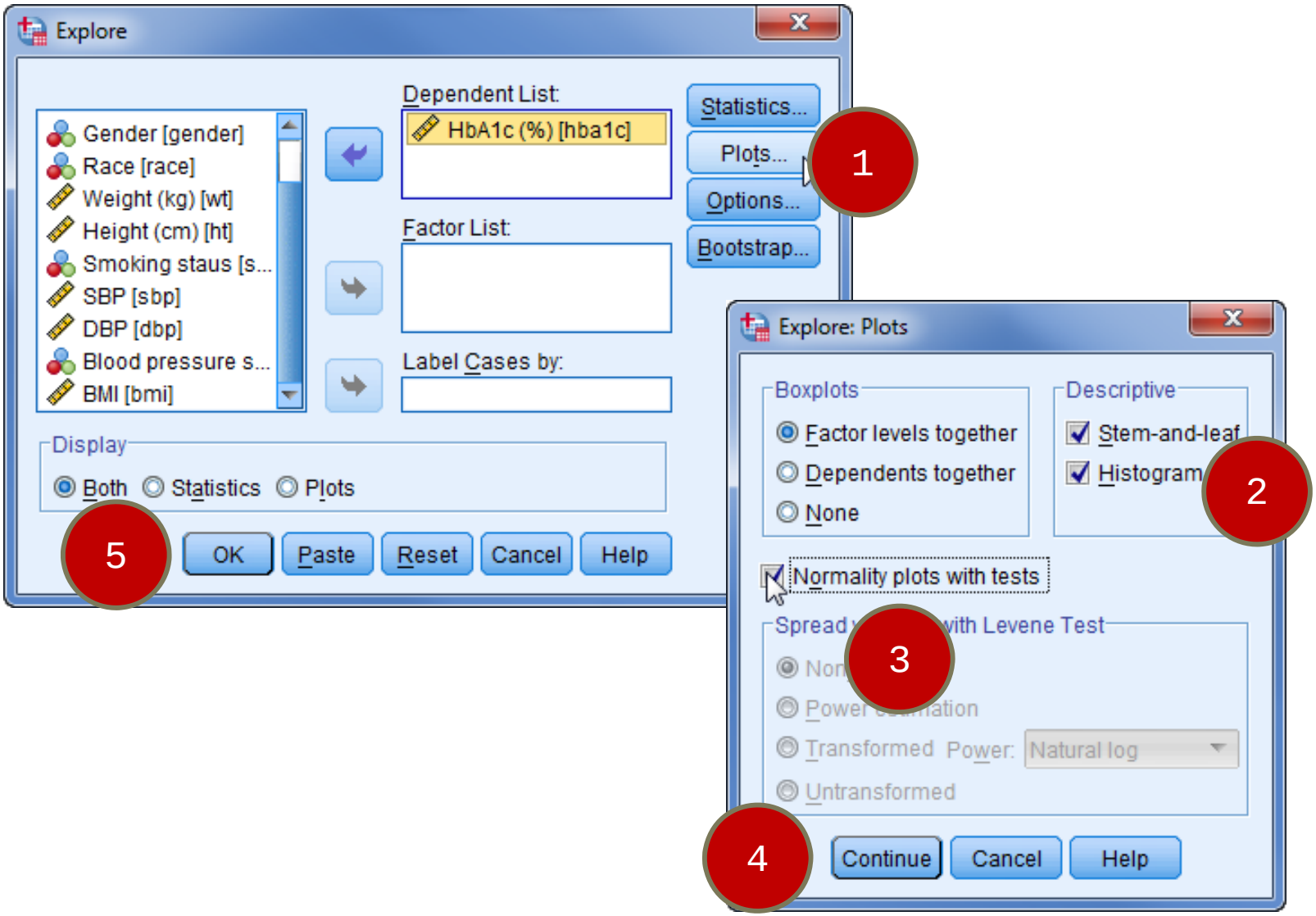


Check for symmetrical bell shape curve



Check for outlier

Describe numerical data *(HbA1c)*



Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
HbA1c (%)	450	100.0%	0	0.0%	450	100.0%

Descriptives

			Statistic	Std. Error
HbA1c (%)	Mean		6.4745	.08505
	95% Confidence Interval for Mean	Lower Bound	6.3073	
		Upper Bound	6.6416	
	5% Trimmed Mean		6.4395	
	Median		6.3511	
	Variance		3.255	
	Std. Deviation		1.80425	
	Minimum		1.91	
	Maximum		11.58	
	Range		9.66	
	Interquartile Range		2.26	
	Skewness		.280	.115
	Kurtosis		.009	.230

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
HbA1c (%)	.049	450	.011	.992	450	.021

a. Lilliefors Significance Correction

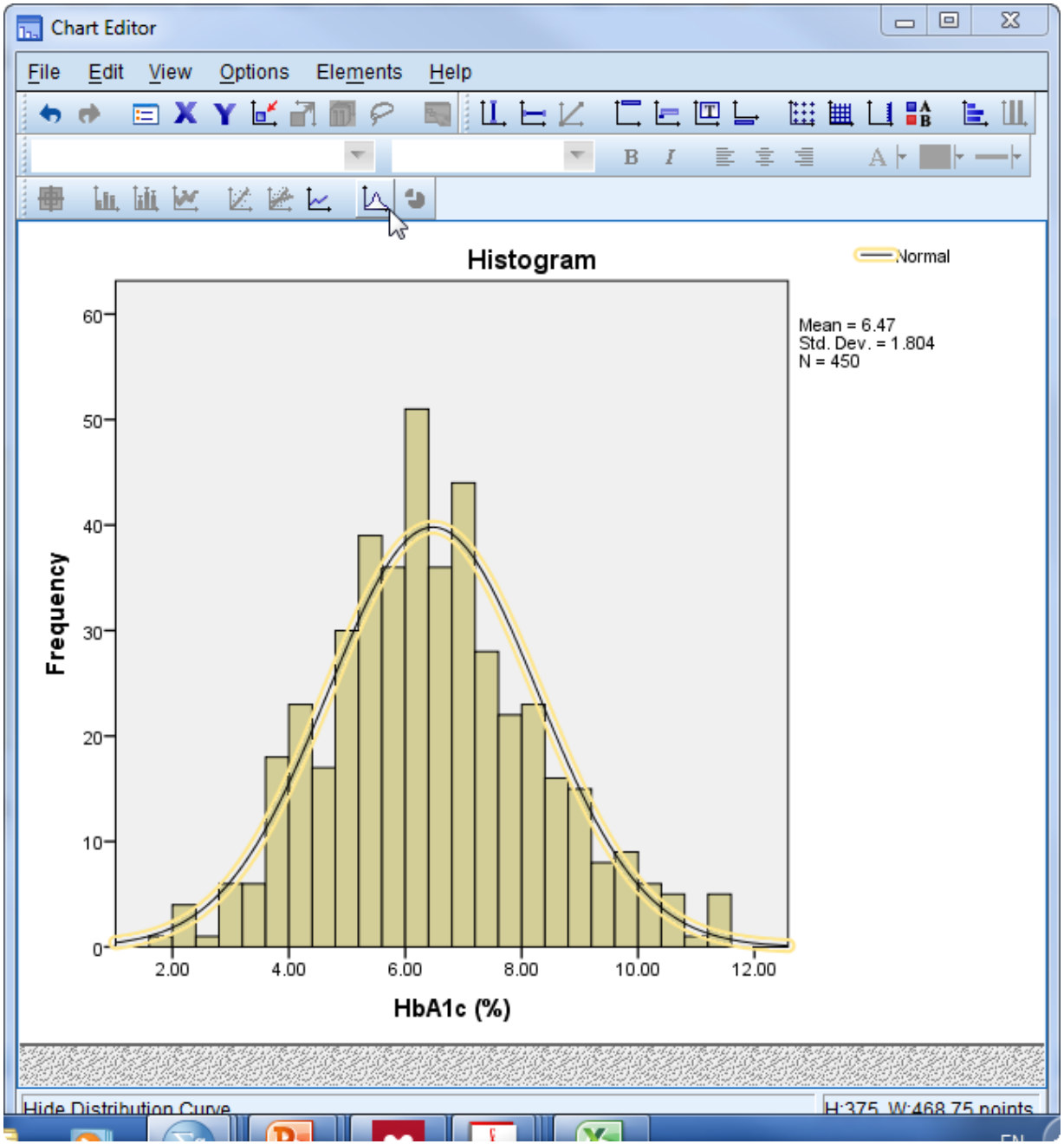
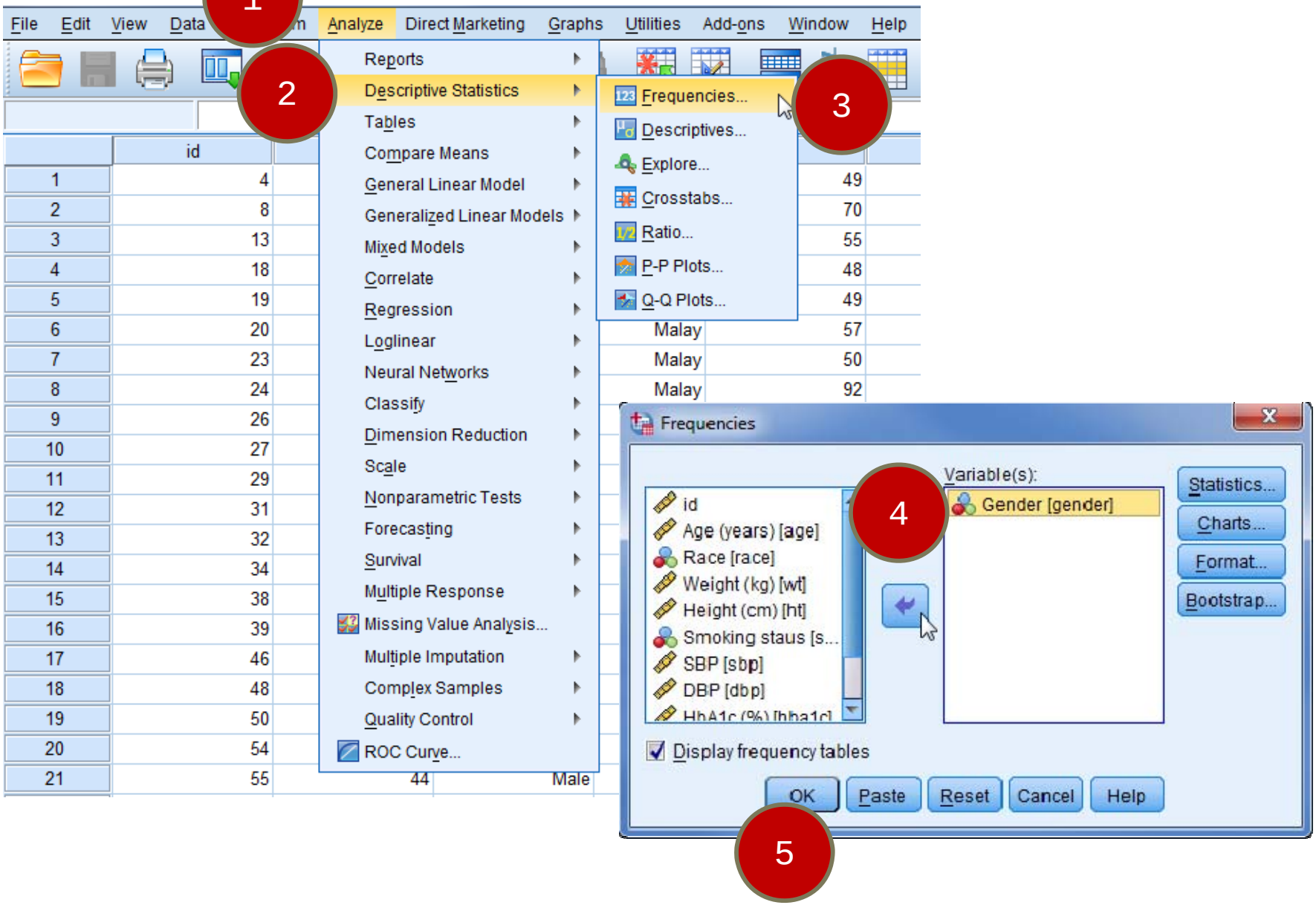


Chart Editor
File Edit View Options Elements Help
Histogram
Normal
Mean = 6.47
Std. Dev. = 1.804
N = 450
Frequency
60
50
40
30
20
10
0
2.00
4.00
6.00
8.00
10.00
12.00
HbA1c (%)
Hide Distribution Curve
H:375, W:468.75 points

Describe categorical data *(gender)*



Frequency Table

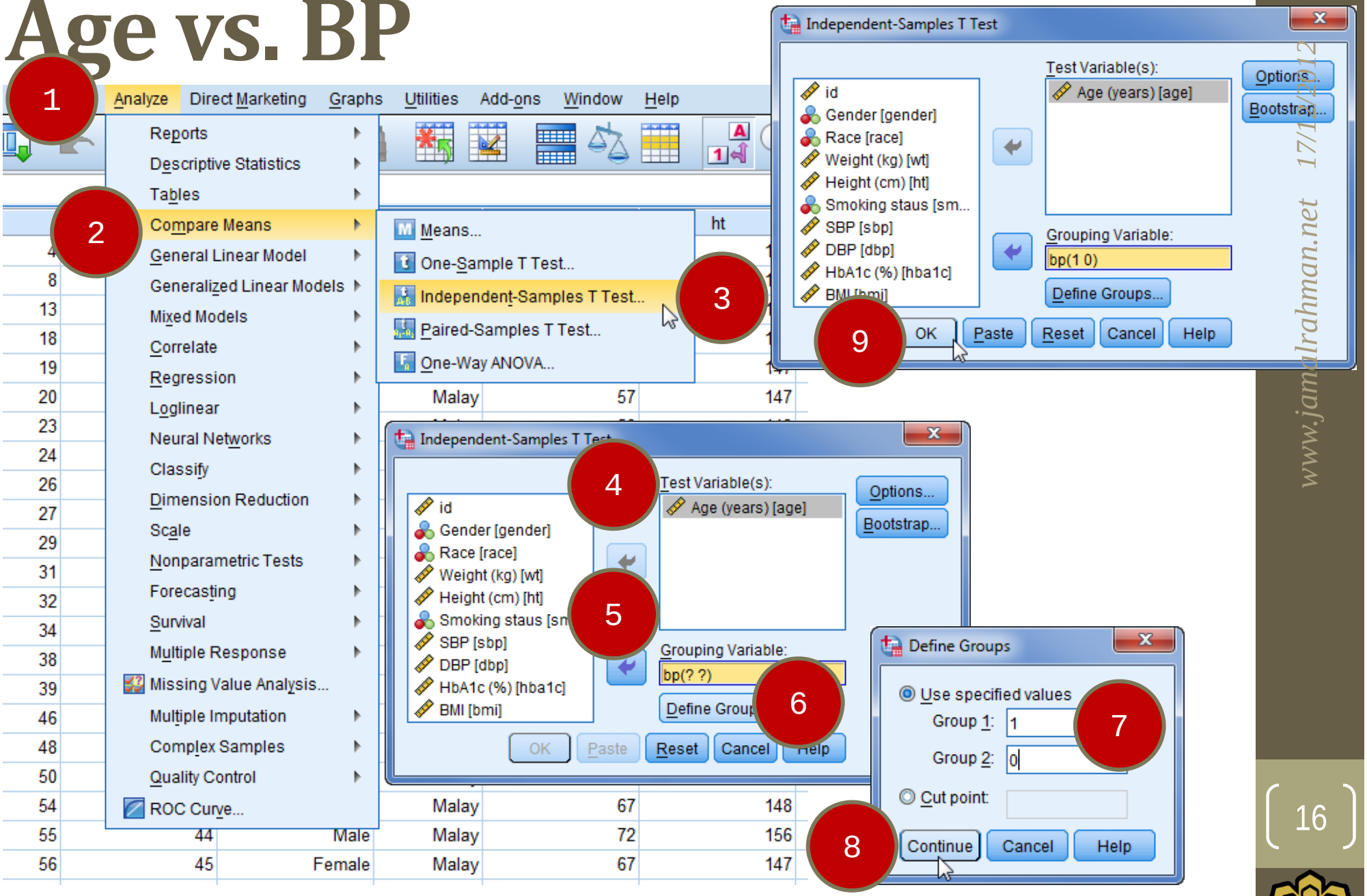
Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	240	53.3	53.3	53.3
	Male	210	46.7	46.7	100.0
	Total	450	100.0	100.0	

Basic biostatistics with SPSS

COMPARE TWO MEANS

Age vs. BP



Group Statistics

	Blood pressure status	N	Mean	Std. Deviation	Std. Error Mean
Age (years)	High	179	36.76	11.491	.859
	Normal	271	36.45	11.241	.683

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Age (years)	Equal variances assumed	.004	.950	.283	448	.777	.310	1.092	-1.837	2.456
	Equal variances not assumed			.282	375.276	.778	.310	1.097	-1.848	2.467

Is there age difference between hypertensive and normal subjects?

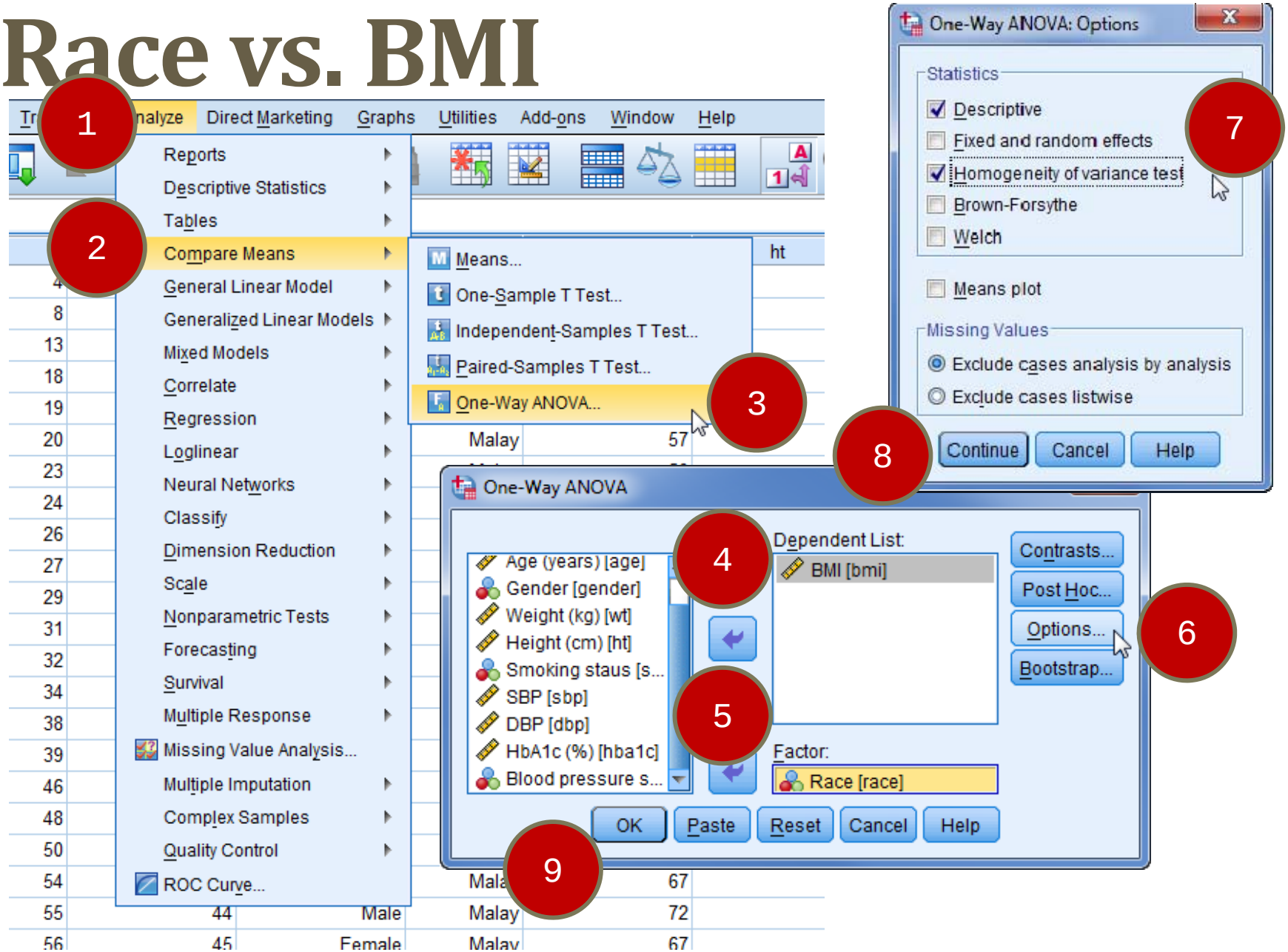
36.8 (11.5) years vs. 36.4 (11.2) years (t=0.283, df=448, P=0.777).

Choose P=0.777 instead of P=0.778 if P for Levene's test > 0.05 (meaning equal variances assumed)

Basic biostatistics with SPSS

COMPARE MORE THAN TWO MEANS

Race vs. BMI


1
2
3
4
5
6
7
8
9
Analyze
Direct Marketing
Graphs
Utilities
Add-ons
Window
Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Quality Control
ROC Curve...
Means...
One-Sample T Test...
Independent-Samples T Test...
Paired-Samples T Test...
One-Way ANOVA...
One-Way ANOVA
Dependent List:
BMI [bmi]
Factor:
Race [race]
Age (years) [age]
Gender [gender]
Weight (kg) [wt]
Height (cm) [ht]
Smoking staus [s...
SBP [sbp]
DBP [dbp]
HbA1c (%) [hba1c]
Blood pressure s...
Contrasts...
Post Hoc...
Options...
Bootstrap...
OK
Paste
Reset
Cancel
Help
One-Way ANOVA: Options
Statistics
Descriptive
Fixed and random effects
Homogeneity of variance test
Brown-Forsythe
Welch
Means plot
Missing Values
Exclude cases analysis by analysis
Exclude cases listwise
Continue
Cancel
Help

Descriptives

BMI

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Malay	234	26.2103	5.91398	.38661	25.4486	26.9720	18.90	66.10
Chinese	182	27.3701	6.69782	.49648	26.3904	28.3497	18.49	62.53
Indian	32	27.1894	7.19586	1.27206	24.5950	29.7838	19.56	48.90
Total	448	26.7514	6.34901	.29996	26.1619	27.3409	18.49	66.10

Is there differences of mean BMI between Malay, Chinese or Indian? 26.2 vs. 27.3 vs. 27.2?

Test of Homogeneity of Variances

BMI

Levene Statistic	df1	df2	Sig.
2.333	2	445	.098

ANOVA

BMI

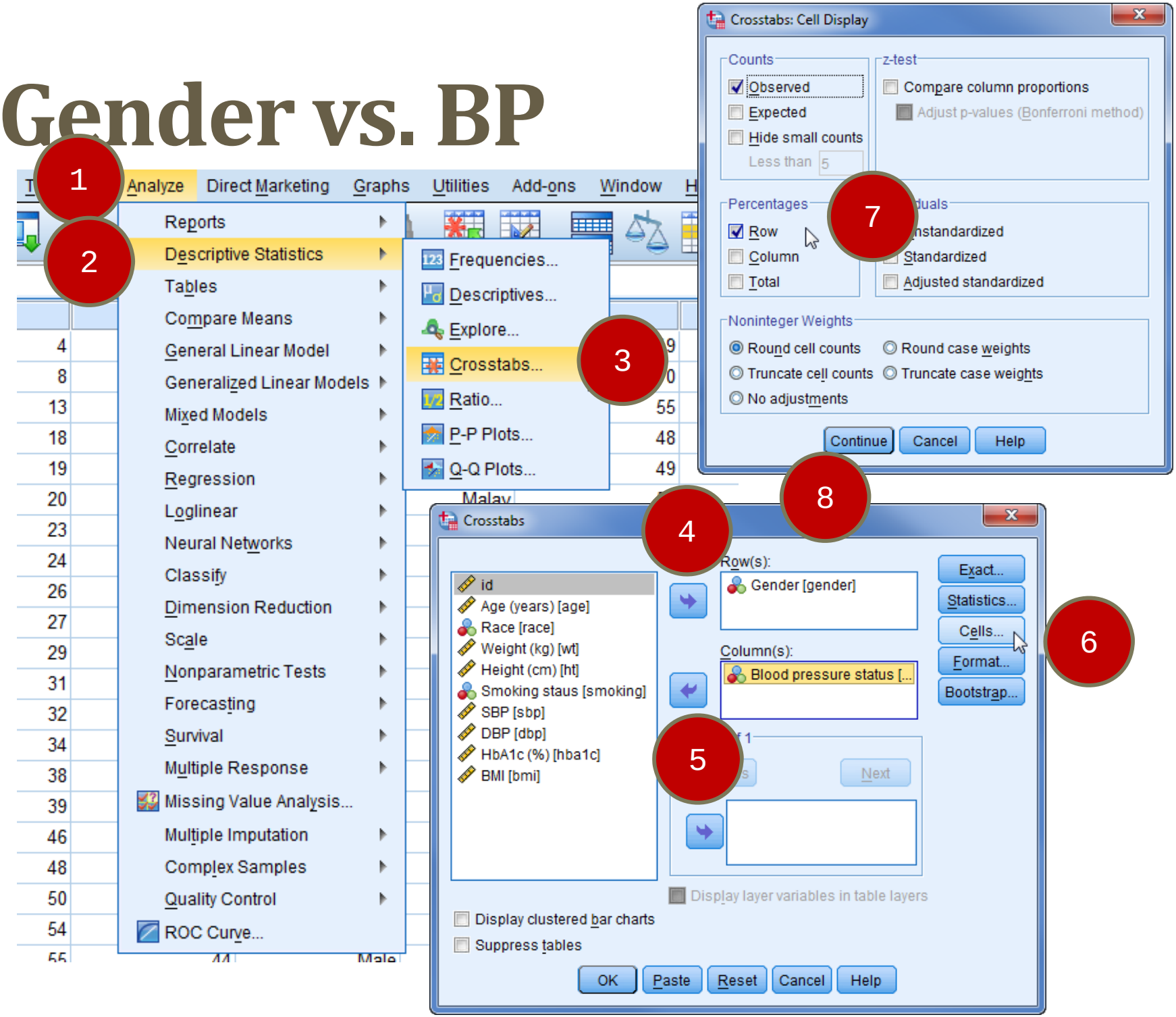
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	144.315	2	72.158	1.796	.167
Within Groups	17874.221	445	40.167		
Total	18018.536	447			

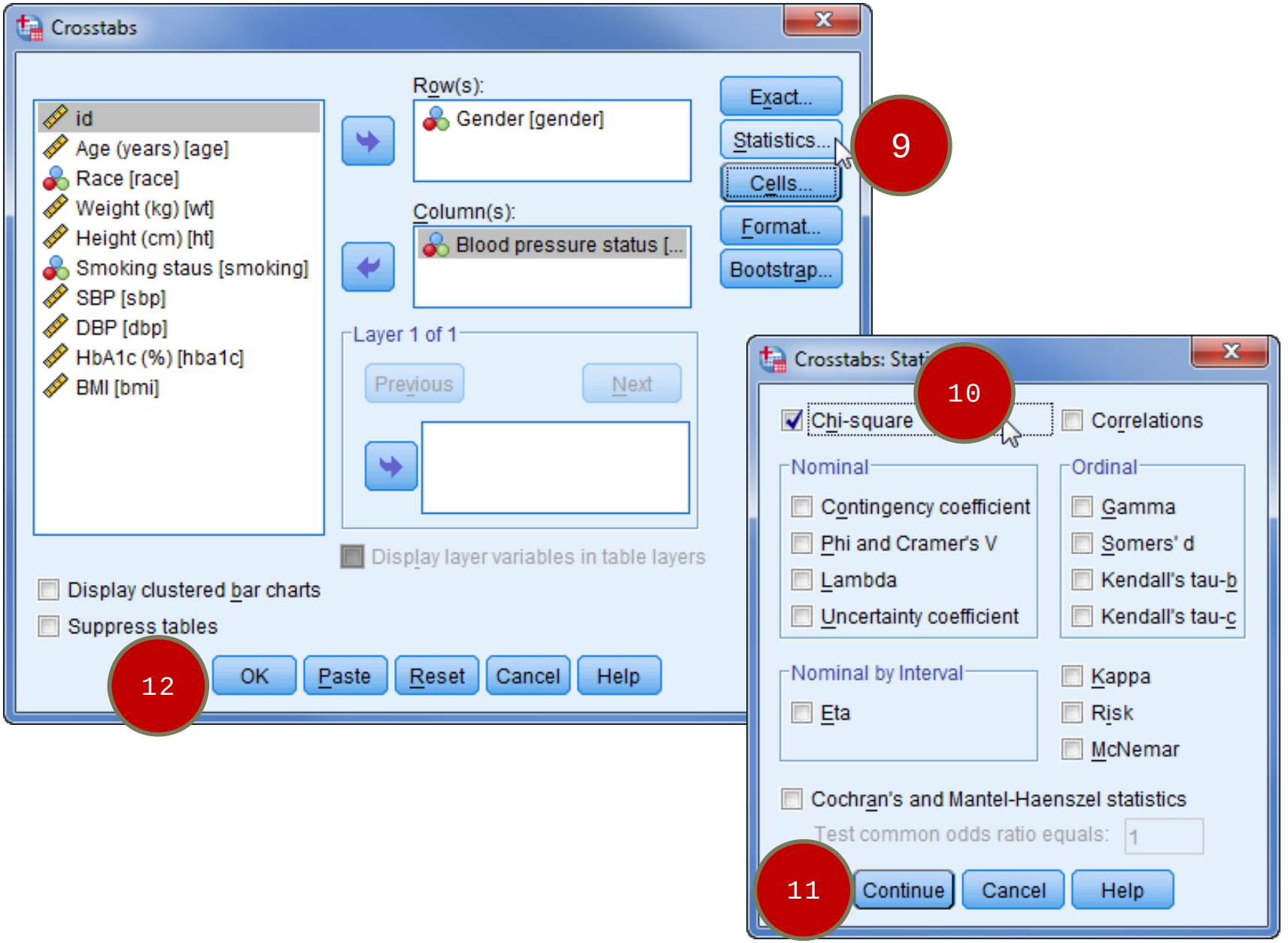
F (2; 445) = 1.796, P = 0.167 i.e. P > 0.05.

Basic biostatistics with SPSS

COMPARE PROPORTIONS

Gender vs. BP




Crosstabs
id
Age (years) [age]
Race [race]
Weight (kg) [wt]
Height (cm) [ht]
Smoking staus [smoking]
SBP [sbp]
DBP [dbp]
HbA1c (%) [hba1c]
BMI [bmi]
Row(s):
Gender [gender]
Column(s):
Blood pressure status [...
Layer 1 of 1
Previous
Next
Display layer variables in table layers
Display clustered bar charts
Suppress tables
Exact...
Statistics...
Cells...
Format...
Bootstrap...
9
12
OK
Paste
Reset
Cancel
Help
Crosstabs: Stati
10
Chi-square
Correlations
Nominal
Contingency coefficient
Phi and Cramer's V
Lambda
Uncertainty coefficient
Ordinal
Gamma
Somers' d
Kendall's tau-b
Kendall's tau-c
Nominal by Interval
Eta
Kappa
Risk
McNemar
Cochran's and Mantel-Haenszel statistics
Test common odds ratio equals:
1
11
Continue
Cancel
Help

Gender * Blood pressure status

Crosstab

			Blood pressure status		Total
			Normal	High	
Gender	Female	Count	145	95	240
		% within Gender	60.4%	39.6%	100.0%
	Male	Count	126	84	210
		% within Gender	60.0%	40.0%	100.0%
Total		Count	271	179	450
		% within Gender	60.2%	39.8%	100.0%

Who has higher prevalence of high blood pressure? Male or female? 40% vs. 39.6%.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.008[a]	1	.928		
Continuity Correction[b]	.000	1	1.000		
Likelihood Ratio	.008	1	.928		
Fisher's Exact Test				1.000	.502
Linear-by-Linear Association	.008	1	.928		
N of Valid Cases	450				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 83.53.

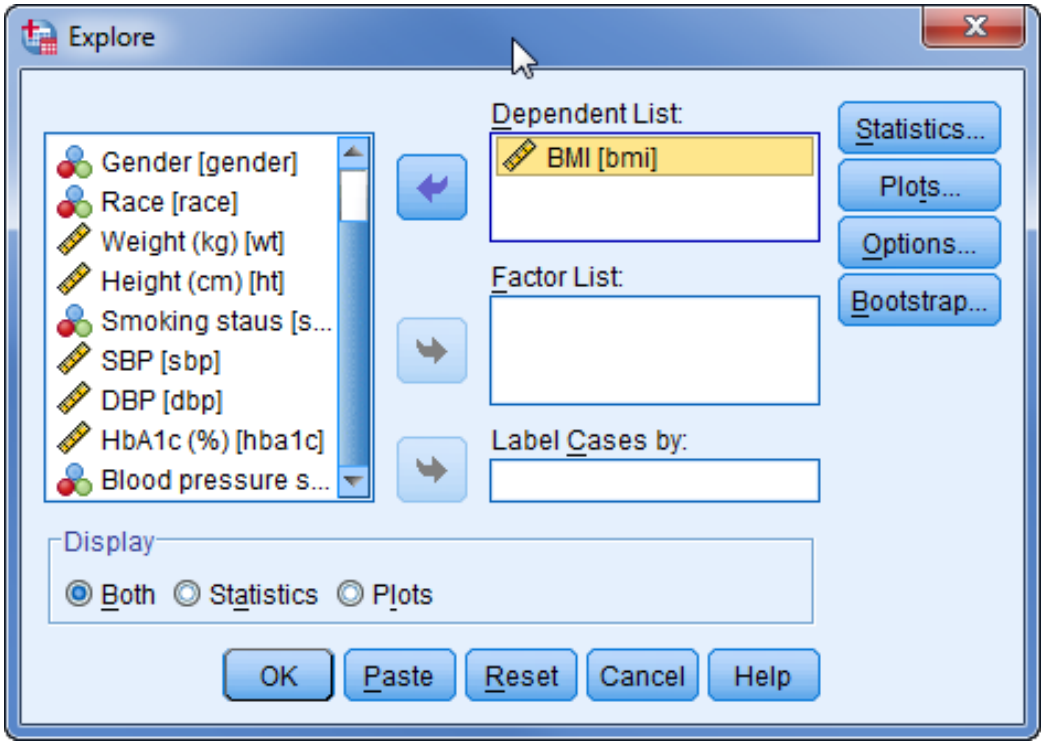
b. Computed only for a 2x2 table

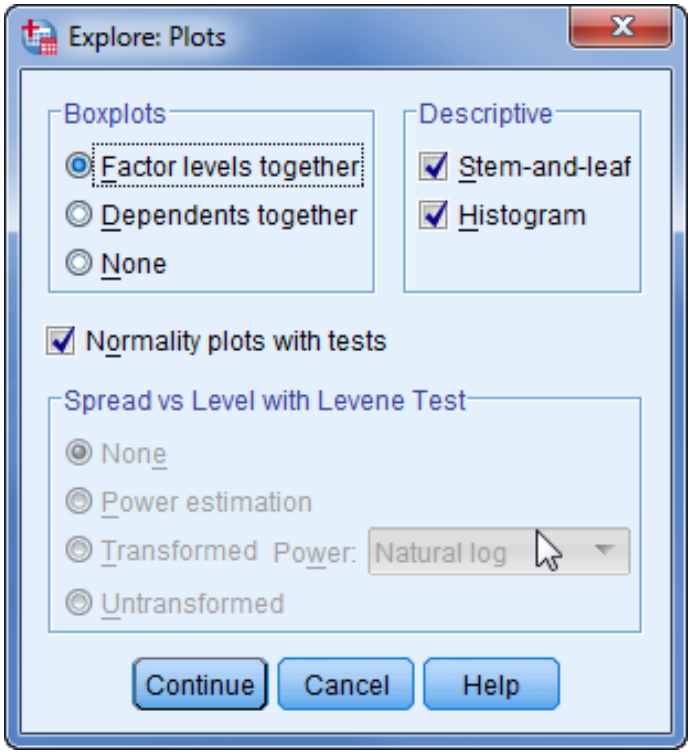
$\chi 2=0.928$, $df=1$, $P=0.928$ i.e. $P > 0.05$.

Basic biostatistics with SPSS

NON-PARAMETRICS

Describe BMI





Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
BMI	448	99.6%	2	0.4%	450	100.0%

Descriptives

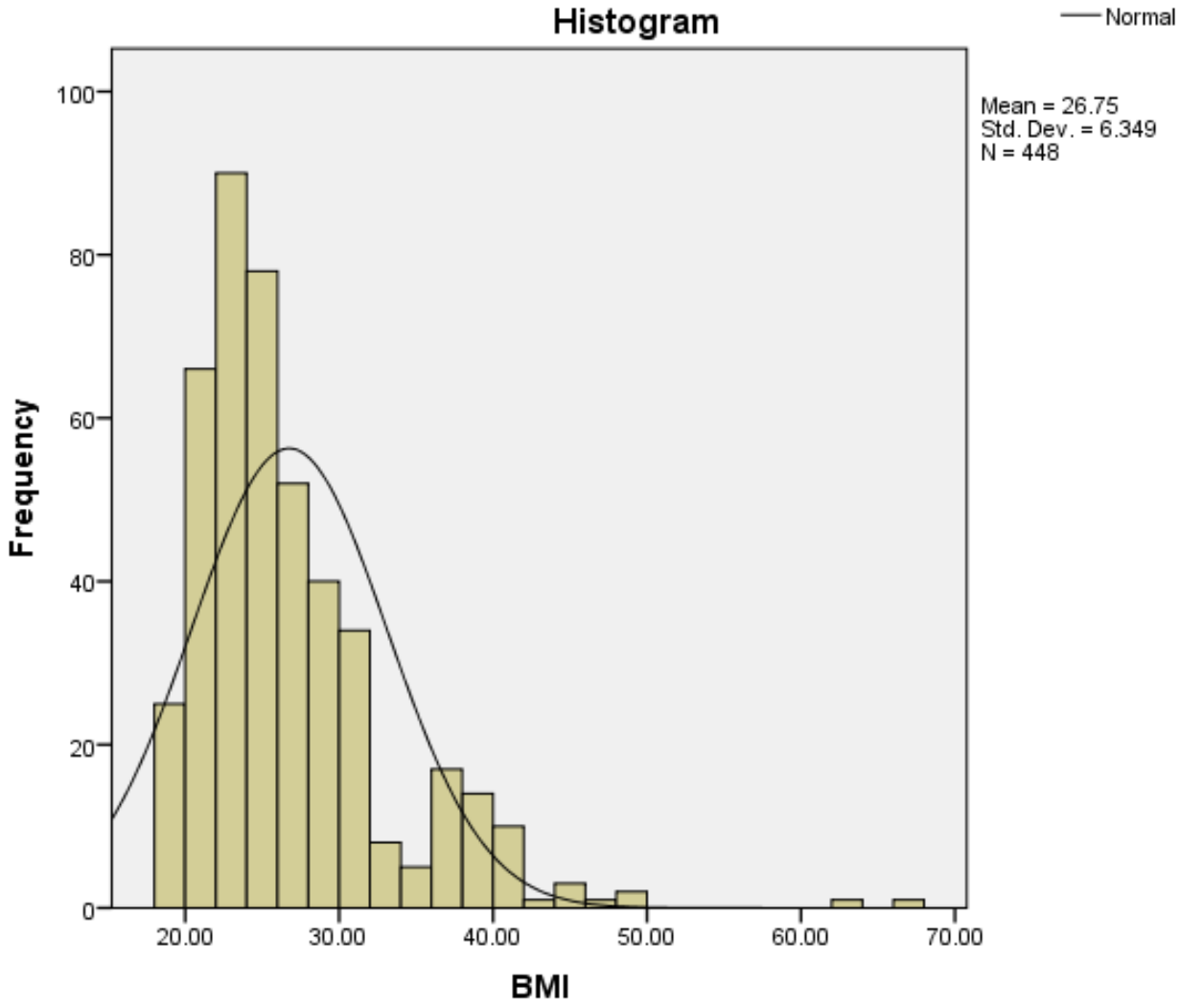
			Statistic	Std. Error
BMI	Mean		26.7514	.29996
	95% Confidence Interval for Mean	Lower Bound	26.1619	
		Upper Bound	27.3409	
	5% Trimmed Mean		26.1638	
	Median		25.0194	
	Variance		40.310	
	Std. Deviation		6.34901	
	Minimum		18.49	
	Maximum		66.10	
	Range		47.60	
	Interquartile Range		6.20	
	Skewness		1.896	.115
	Kurtosis		5.746	.230

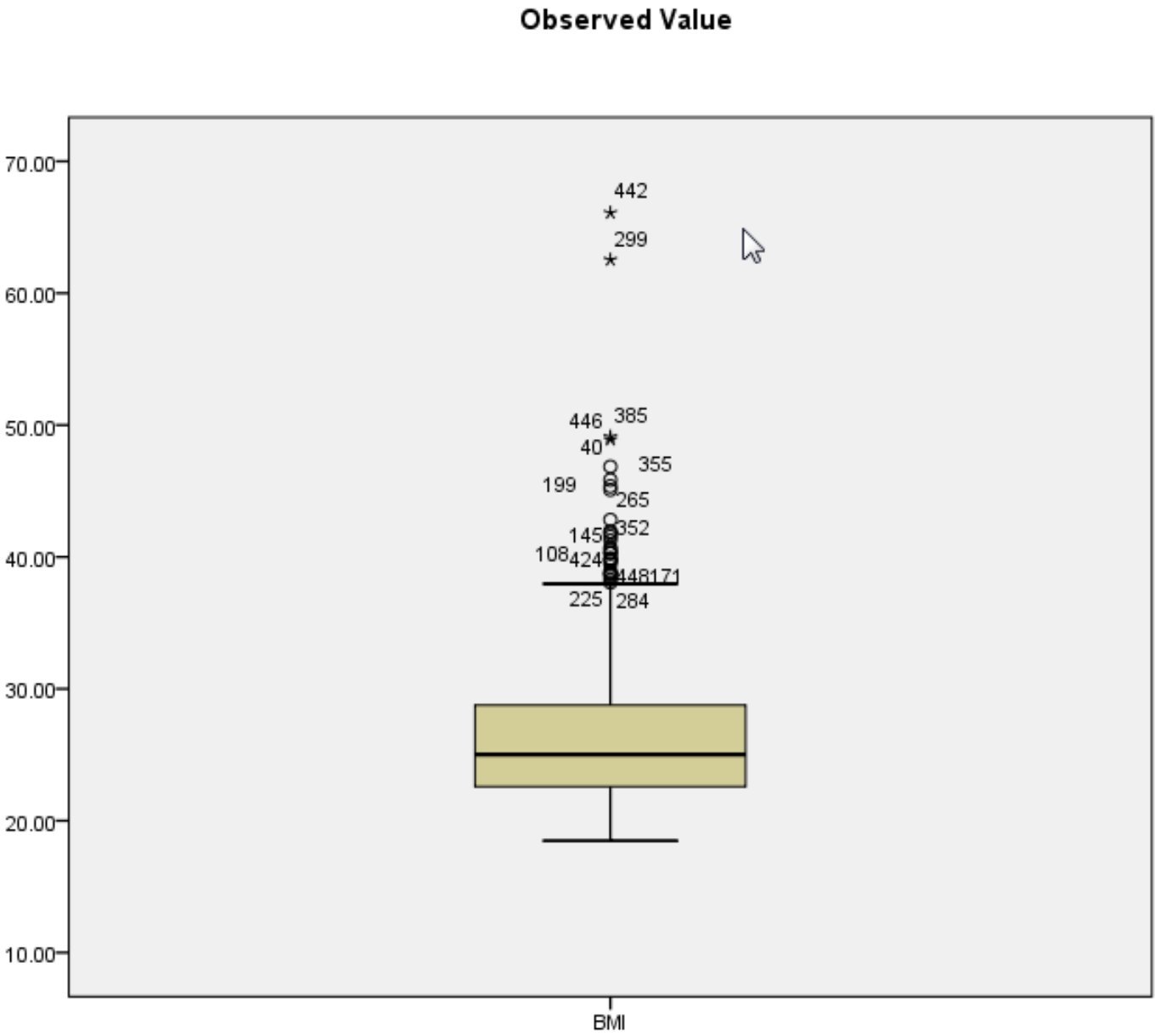
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
BMI	.137	448	.000	.844	448	.000

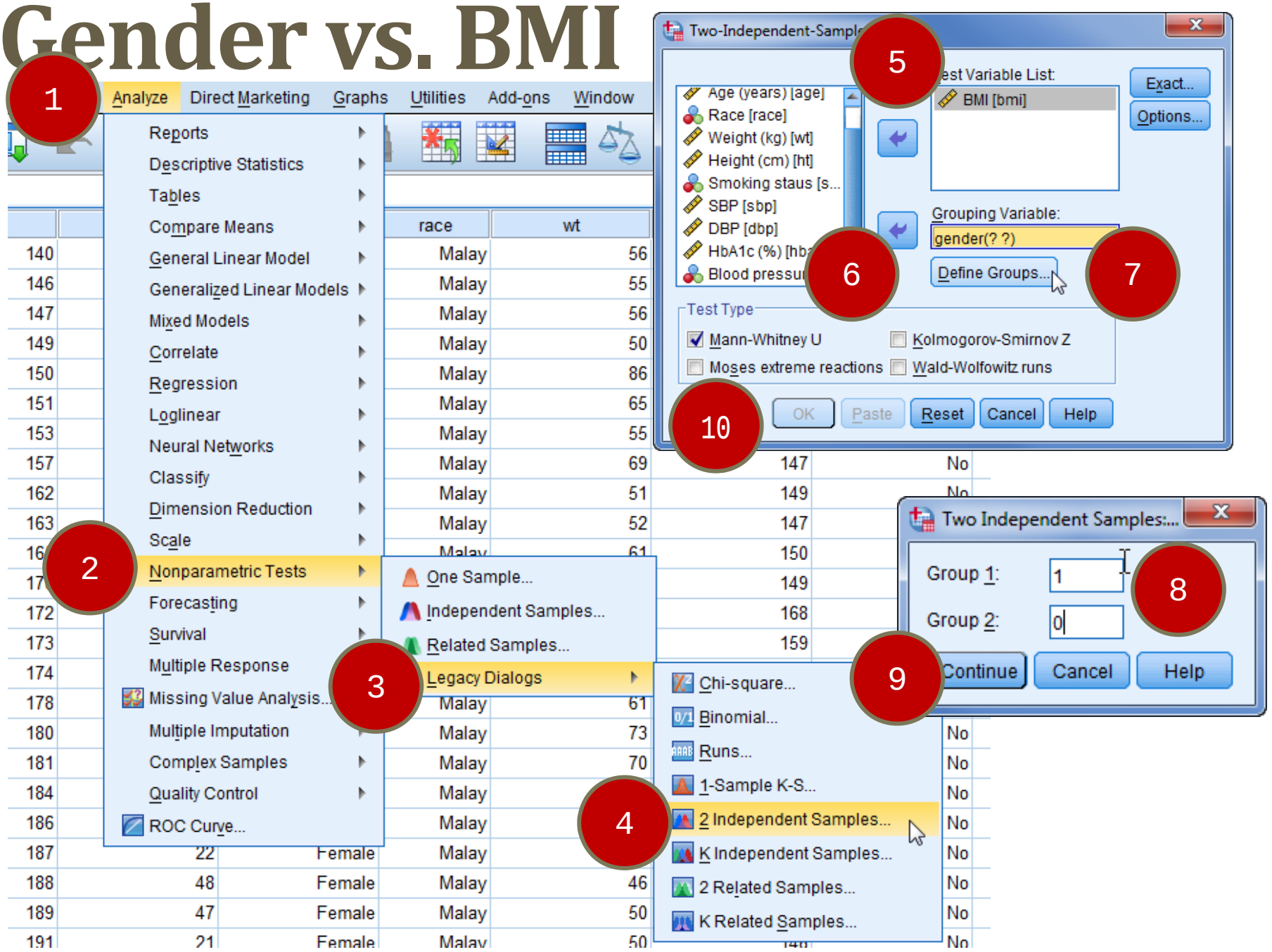
a. Lilliefors Significance Correction

BMI is not Normally distributed.


Histogram
Normal
Mean = 26.75
Std. Dev. = 6.349
N = 448
Frequency
100
80
60
40
20
0
20.00
30.00
40.00
50.00
60.00
70.00
BMI


Observed Value
70.00
60.00
50.00
40.00
30.00
20.00
10.00
442
299
446
385
40
355
199
265
352
145
108
424
448
171
225
284
BMI

Gender vs. BMI



Mann-Whitney Test

Ranks

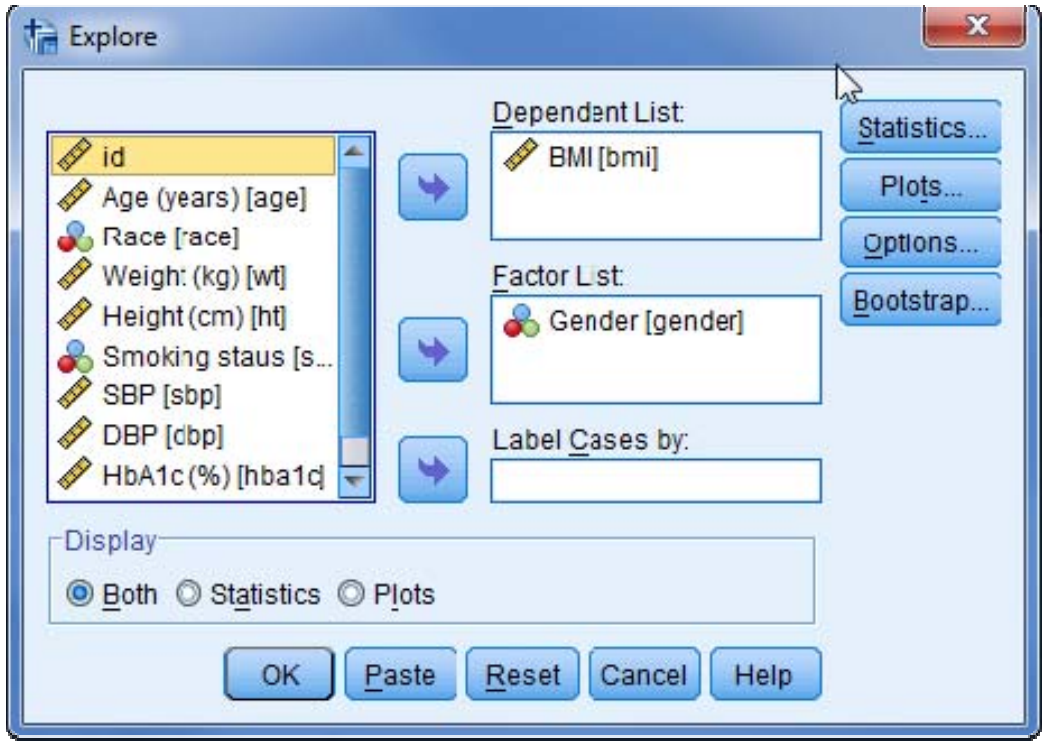
	Gender	N	Mean Rank	Sum of Ranks
BMI	Female	238	168.59	40124.00
	Male	210	287.87	60452.00
	Total	448		

Mann-Whitnet tests between 'rank'. Don't have to cite the mean 'rank'. They have no meaning in the actual context. Instead describe their medians.

Test Statistics[a]

	BMI
Mann-Whitney U	11683.000
Wilcoxon W	40124.000
Z	-9.731
Asymp. Sig. (2-tailed)	.000

a. Grouping Variable: Gender

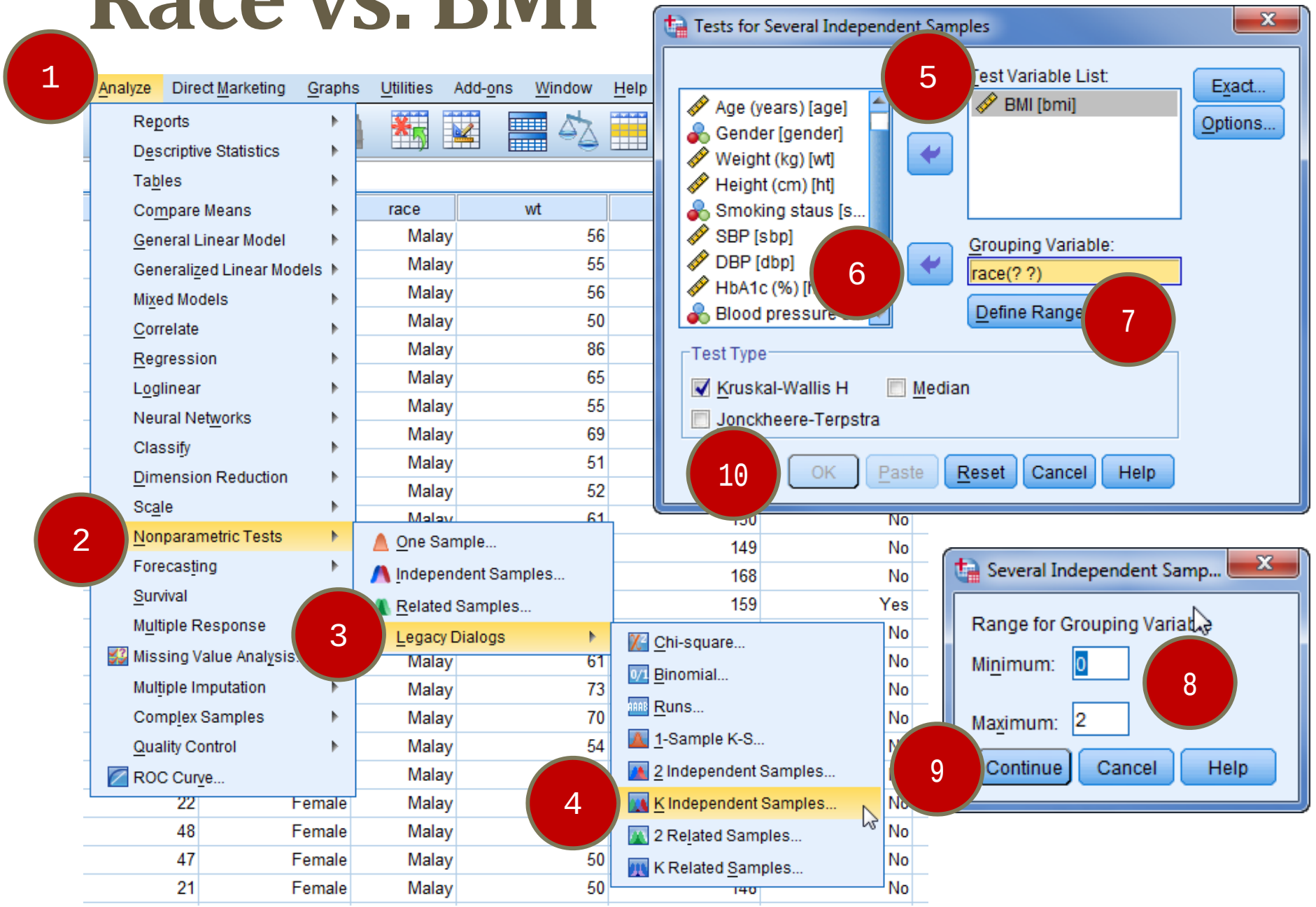


Descriptives

	Gender			Statistic	Std. Error
BMI	Female	Mean		24.0304	.22031
		95% Confidence Interval for Mean	Lower Bound	23.5963	
			Upper Bound	24.4644	
		5% Trimmed Mean		23.8684	
		Median		23.3391	
		Variance		11.552	
		Std. Deviation		3.39877	
		Minimum		18.49	
		Maximum		32.82	
		Range		14.33	
		Interquartile Range		4.75	
		Skewness		.684	.158
		Kurtosis		-.314	.314
	Male	Mean		29.8352	.51236
		95% Confidence Interval for Mean	Lower Bound	28.8252	
			Upper Bound	30.8453	
		5% Trimmed Mean		29.2583	
		Median		27.5397	
		Variance		55.127	
		Std. Deviation		7.42476	
		Minimum		19.92	
		Maximum		66.10	
		Range		46.18	
		Interquartile Range		11.02	
		Skewness		1.463	.168
		Kurtosis		3.319	.334

23.3 (IQR=4.8) vs. 27.5 (IQR=11.0) kg/m^2, Z (Mann-Whitney)=9.731, P<0.001

Race vs. BMI



		BMI	
		Count	Median
Race	Malay	234	24.53
	Chinese	184	25.91
	Indian	32	23.87

Kruskal-Wallis Test

Ranks

	Race	N	Mean Rank
BMI	Malay	234	214.29
	Chinese	182	238.32
	Indian	32	220.58
	Total	448	

Test Statistics[a,b]

	BMI
Chi-Square	3.557
df	2
Asymp. Sig.	.169

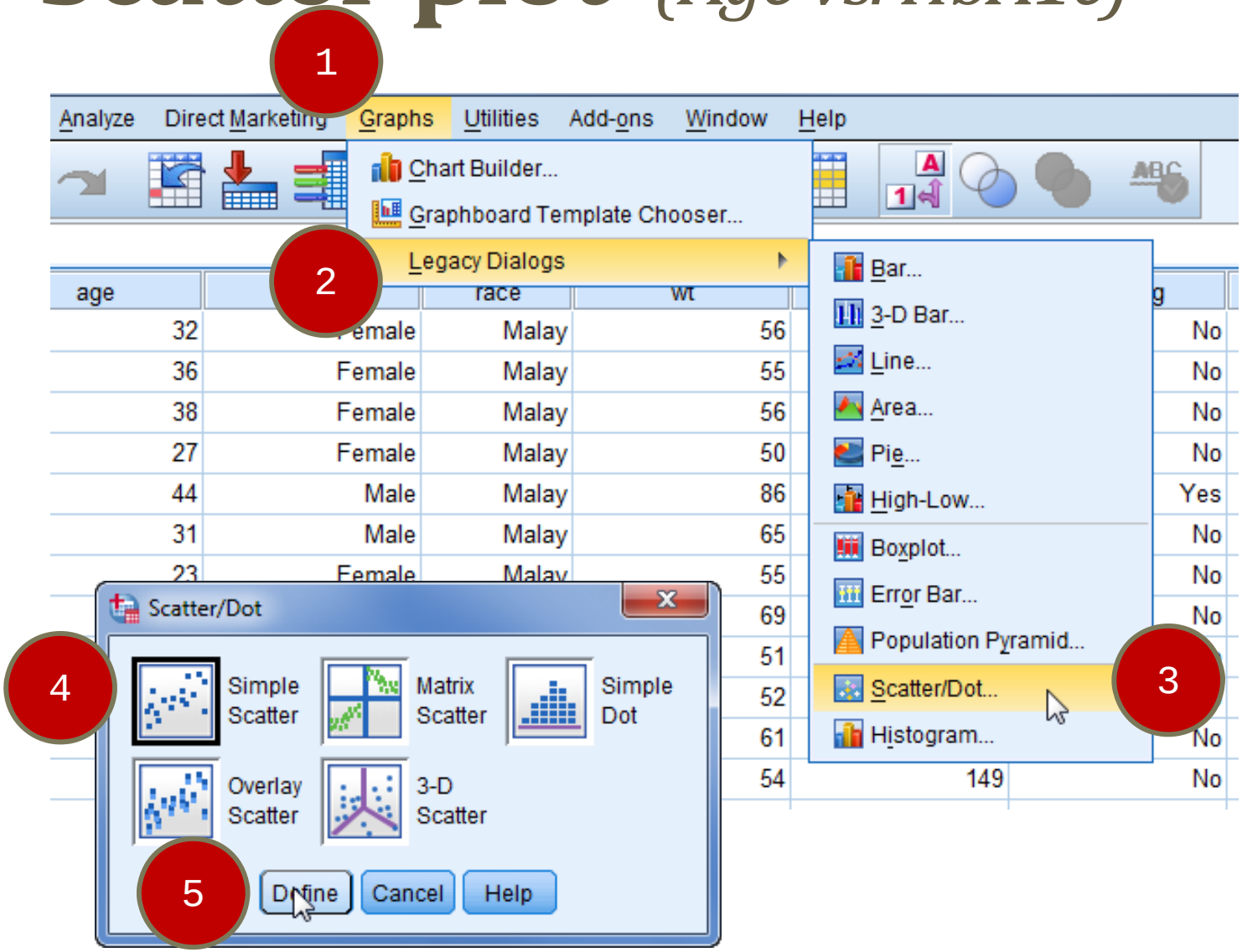
a. Kruskal Wallis Test

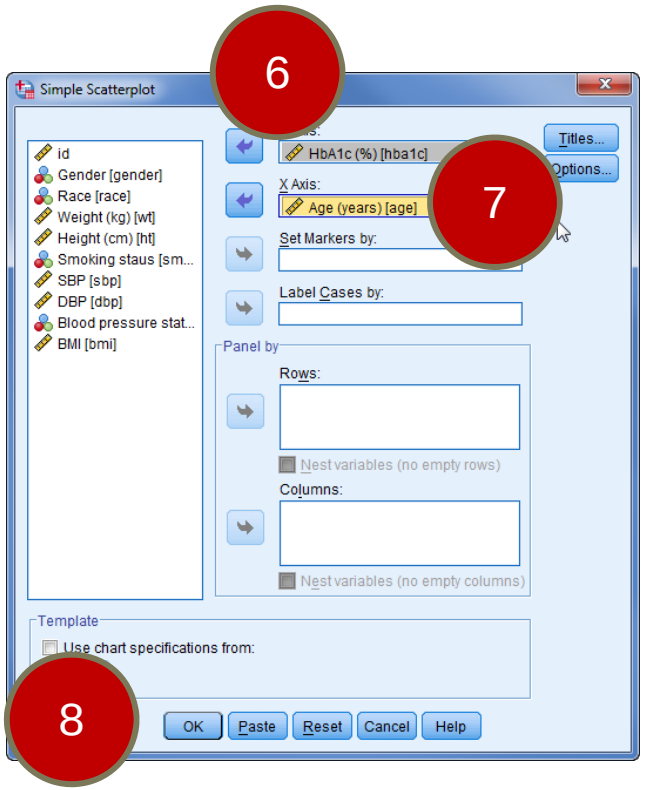
b. Grouping Variable: Race

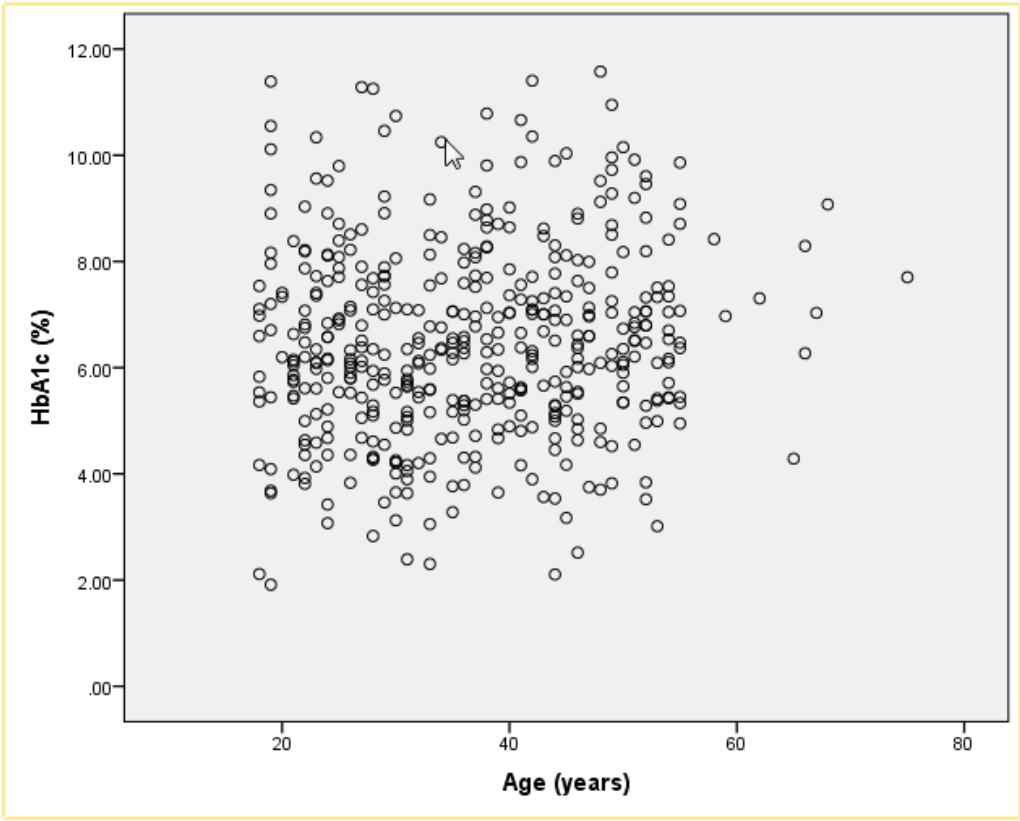
Basic biostatistics with SPSS

CORRELATION

Scatter plot *(Age vs. HbA1c)*

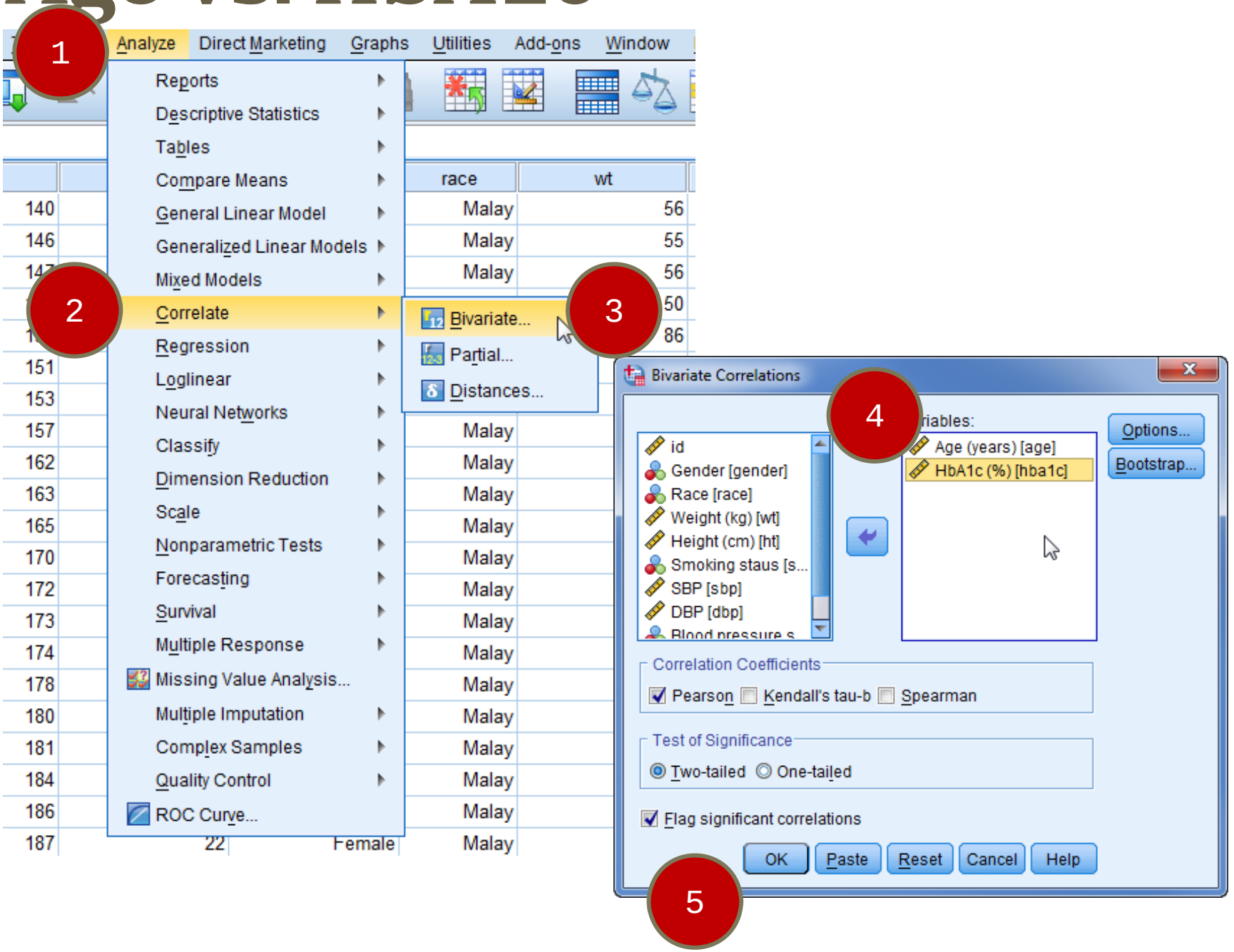






Describe the relationship of two numerical variables.
Dependent variable in Y-axis, explanatory (exposure) in x-axis.

Age vs. HbA1c


1
Analyze
Direct Marketing
Graphs
Utilities
Add-ons
Window
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
2
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Quality Control
ROC Curve...
3
Bivariate...
Partial...
Distances...
race
wt
Malay
Female
4
Bivariate Correlations
id
Gender [gender]
Race [race]
Weight (kg) [wt]
Height (cm) [ht]
Smoking staus [s...
SBP [sbp]
DBP [dbp]
Age (years) [age]
HbA1c (%) [hba1c]
Options...
Bootstrap...
Correlation Coefficients
Pearson
Kendall's tau-b
Spearman
Test of Significance
Two-tailed
One-tailed
Flag significant correlations
OK
Paste
Reset
Cancel
Help
5

Correlations

		Age (years)	HbA1c (%)	BMI
Age (years)	Pearson Correlation	1	.096*	-.016
	Sig. (2-tailed)		.043	.734
	N	450	450	448
HbA1c (%)	Pearson Correlation	.096*	1	.314**
	Sig. (2-tailed)	.043		.000
	N	450	450	448
BMI	Pearson Correlation	-.016	.314**	1
	Sig. (2-tailed)	.734	.000	
	N	448	448	448

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

There is a weak correlation between age & HbA1c (r=0.96, P=0.43).

< 0.3 = Weak
0.3-0.7 = Moderate
> 0.7 = Strong