

Chapter 1

Basic Biostatistics

Jamalludin Ab Rahman MD MPH
Department of Community Medicine
Kulliyah of Medicine



Content

- Basic premises – variables, level of measurements, probability distribution
- Descriptive statistics
- Inferential statistics, hypothesis testing



3



*We observe, we believe.
What we observe might not be the truth*

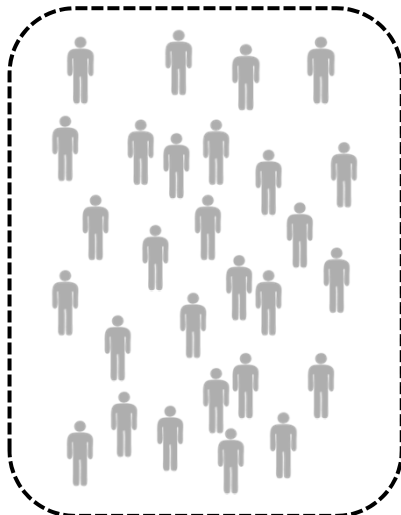
www.jamalrahman.net

29 April, 2014



4

...and we can't observe all. We sample.



Population ⇨ *Sample*

Parameter ⇨ *Statistic*

www.jamalrahman.net

29 April, 2014



5

Variable

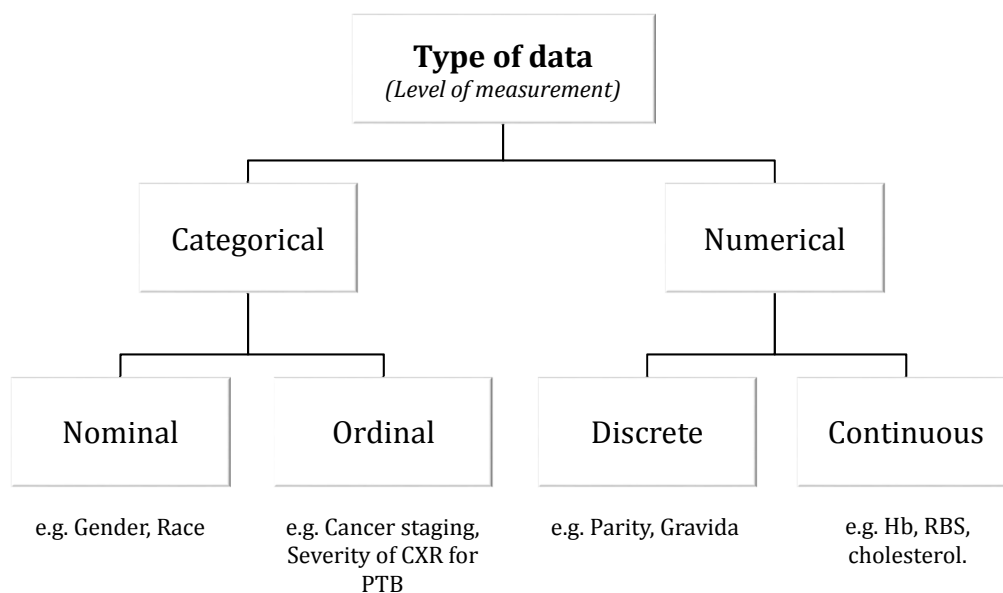
- Characteristic of a population
- Can take different values
- Data = measurements collected/observed

www.jamalrahman.net

29 April, 2014



6



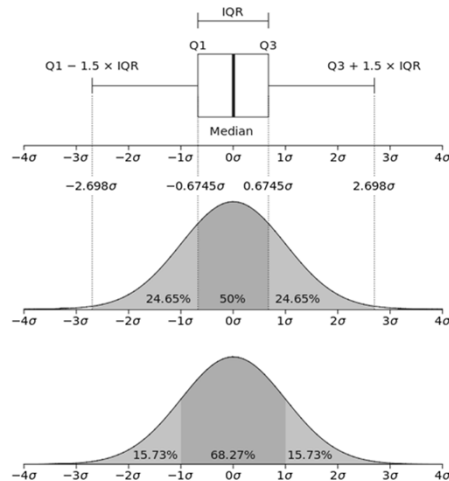
www.jamalrahman.net

29 April, 2014



7

Normal distribution



$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

www.jamalrahman.net

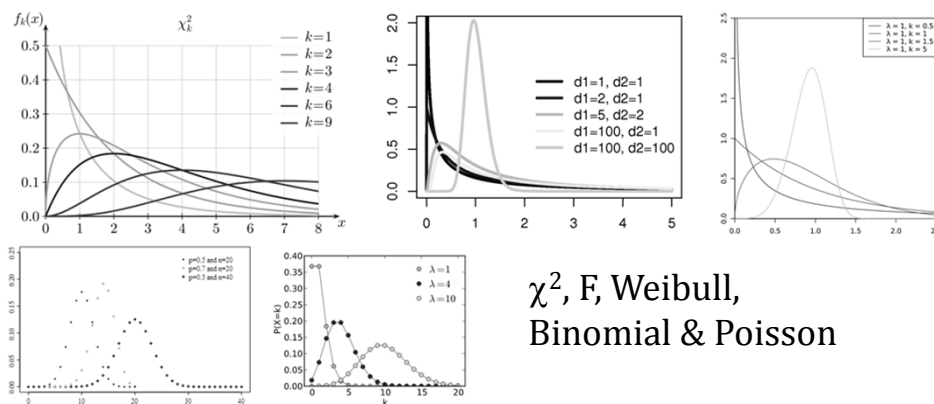
29 April, 2014



8

Other distributions

● Discrete vs. continuous probability distribution



χ^2 , F, Weibull,
Binomial & Poisson

www.jamalrahman.net

29/4/2014



9

Characteristics of Normal distribution

- Smooth, symmetrical (around the mean), uni-modal, bell shaped curve
- Mean = Median = Mode
- Skewness = 0
- Kurtosis = 0

www.jamalrahman.net

29 April, 2014



10

Test of Normality

- Anderson–Darling Test
- Corrected Kolmogorov–Smirnov Test (Lilliefors Test)
- Cramér–von-mises Criterion
- D'agostino's K-squared Test
- Jarque–Bera Test
- Pearson's Chi-square Test
- Shapiro–Francia
- Shapiro–Wilk Test

www.jamalrahman.net

29 April, 2014



11

Use Normality test with caution

- ***Small samples almost always pass a normality test.*** Normality tests have little power to tell whether or not a small sample of data comes from a Gaussian distribution.
- With large samples, ***minor deviations from normality may be flagged as statistically significant***, even though small deviations from a normal distribution won't affect the results of a t test or ANOVA.

www.jamalrahman.net

29 April, 2014



12

Statistical objectives

1. Determine presence of difference (or similarity)
2. Determine degree of difference
3. Determine the direction of changes (outcome)
4. Predict changes (outcomes)

www.jamalrahman.net

29/4/2014



13

Is there any difference between A & B?

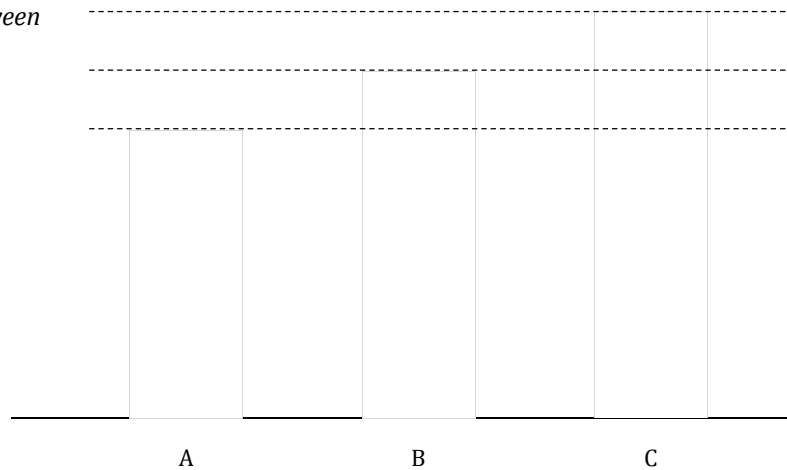
Which one is taller? A or B?

How big is the difference between A & B?

Is C different from A & B?

Is there any pattern now?

If there will be D, can you predict how tall is D?



www.jamalrahman.net

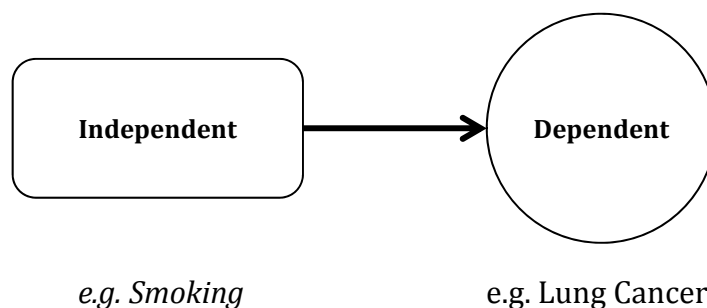
29/4/2014



14

Association

- a value and whose associated value may be changed



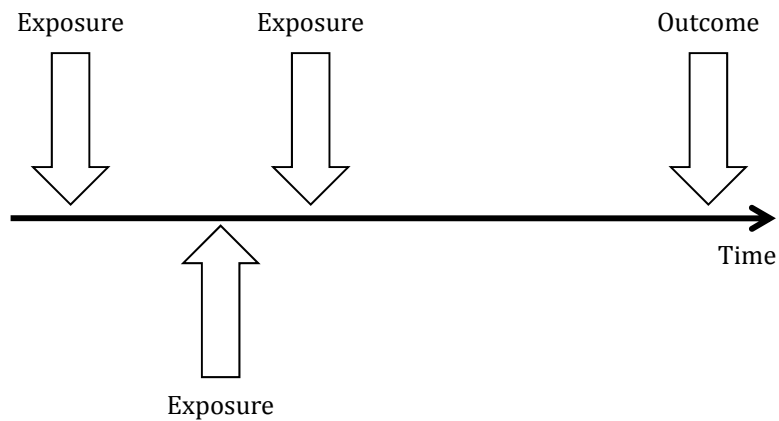
www.jamalrahman.net

29 April, 2014



15

Disease model

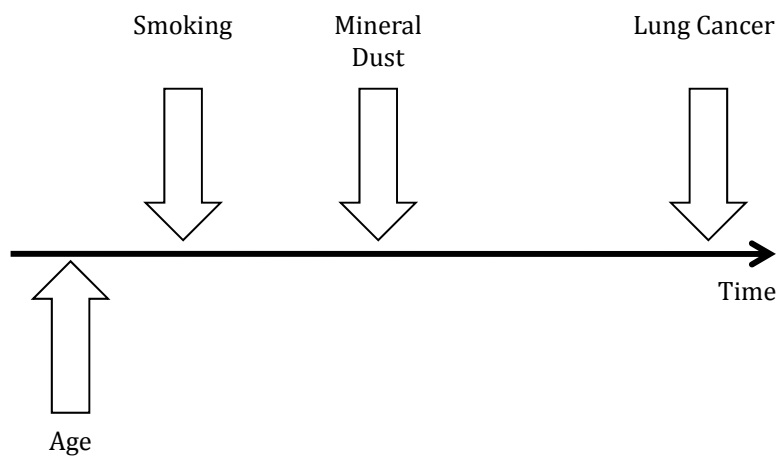
www.jamalrahman.net

29 April, 2014



16

Disease model (example)

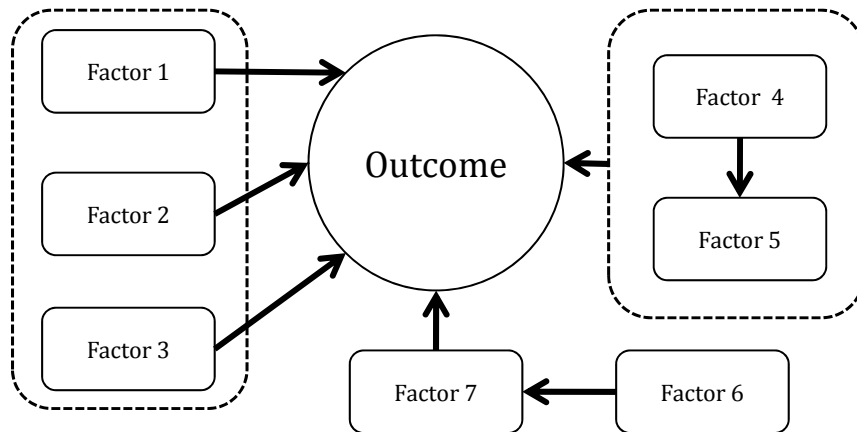
www.jamalrahman.net

29 April, 2014



17

Causal relationship



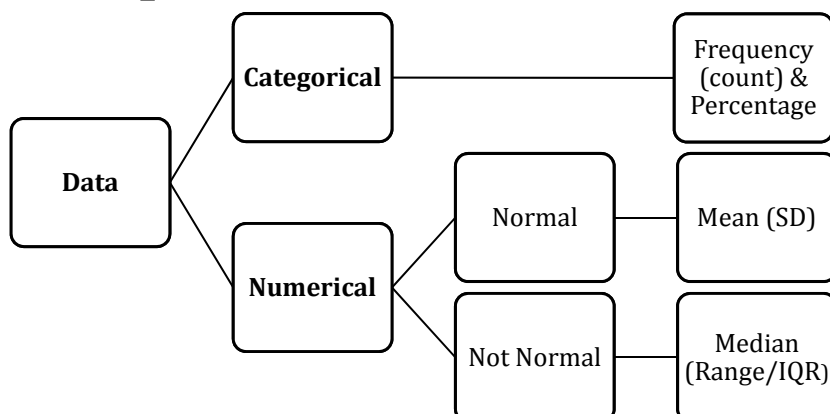
www.jamalrahman.net

29 April, 2014



18

Descriptive statistics



www.jamalrahman.net

29 April, 2014



Hypothesis Testing

- Involve more than one variables
 - exposure & outcome,
 - predictor & criterion,
 - risk & disease
- Try to prove that
Exposure causes the Disease
e.g. Smoking causing Lung Cancer
- Example ~ H_0 : No difference of risk to get Lung Cancer between smoker and non-smoker



	Lung Cancer	No Lung Cancer
Smoking	20 (18.2%)	90 (81.8%)
Not Smoking	5 (4.5%)	105 (95.5%)

$$\chi^2 (df=1) = 10.150, p = 0.001, OR = 4.7 (CI 95\% 1.7 - 13.0)$$

Because $p < 0.05$, we reject H_0 . Therefore there is a difference between smoker & non smoker



21

Statistical Test

- Univariate ~ One dependent & one independent
- Multivariate ~ Multiple dependent & multiple independent variable

www.jamalrahman.net

29 April, 2014



22

What test to use?

Variable 1	Variable 2	Test
Categorical	Categorical	Chi-square
Categorical (2 pop)	Numerical (Normal)	Independent sample t-test
Categorical (2 pop)	Numerical (Not Normal)	Mann-Whitney U test
Categorical (> 2 pop)	Numerical (Normal)	One-way ANOVA
Categorical (> 2 pop)	Numerical (Not Normal)	Kruskal-Wallis test
Numerical (Normal)	Numerical (Normal)	Pearson Correlation Coefficient Test
Numerical (Normal/ Not Normal)	Numerical (Not Normal)	Spearman Correlation Coefficient Test
Numerical (Normal)	Numerical (Normal) – Paired	Paired t-test
Numerical (Not Normal)	Numerical (Not Normal) – Paired	Friedman test

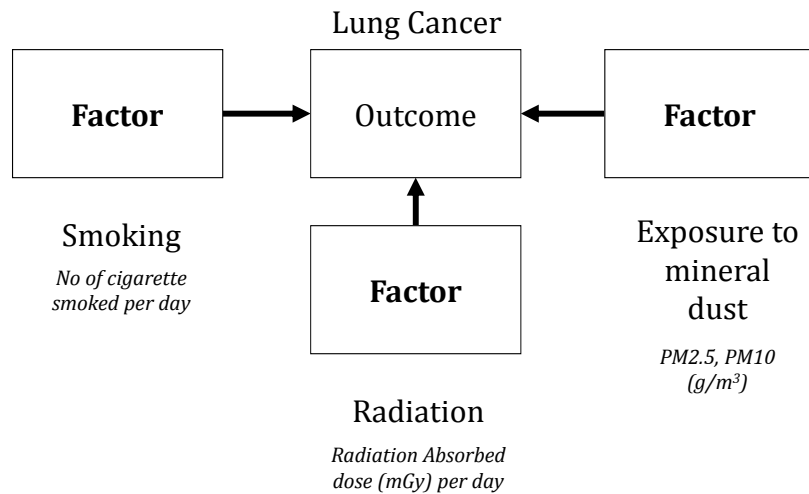
www.jamalrahman.net

29 April, 2014



23

But life is not simple!



www.jamalrahman.net

29 April, 2014



24

The multivariate model

Lung CA = Smoking + Radiation + Mineral dust + Others

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

↑
↑
 Regression Residual

A good model is when Regression > Residual

www.jamalrahman.net

29/4/2014



25

Multivariate Analysis

- Hypothesis testing & control for confounders
 - e.g. General Linear Model, Logistic Regression
- Modeling
 - e.g. Linear Regression
- Data reduction
 - e.g. Factor Analysis, Cluster Analysis

www.jamalrahman.net

29 April, 2014



26

Writing plan for statistical analysis #1

Data were analyzed using the complex sample function of SPSS (version 13.0). Sampling errors were estimated using the primary sampling units and strata provided in the data set. Sampling weights were used to adjust for nonresponse bias and the oversampling of blacks, Mexican Americans, and the elderly in NHANES. The prevalence of hypertension, as well as the awareness, treatment, and control rates, were age adjusted by direct standardization to the US 2000 standard population.¹⁰ To analyze differences over time, the 2003–2004 data were compared with the 1999–2000 data. Estimates with a coefficient of variation >0.3 were considered unreliable. A 2-tailed P value <0.05 was considered statistically significant.

(Ong et al. 2009)

www.jamalrahman.net

29 April, 2014



Writing plan for statistical analysis #2

To assess the effect of the selection process on the characteristics of the cases, we compared cases included in the final analysis to the rest of the cases. Since controls included in the present analysis were different from the rest of the diabetes free participants by design, no similar comparisons were performed for that group. To compare baseline characteristics of cases and controls appropriate univariate statistics were used. Similar binary logistic and multiple linear regression models were built with incident diabetes or HbA1c as respective outcomes and additive block entry of adiponectin and potential confounders. For linear regression CRP and triglycerides were log transformed. Since HbA1c could be modified by drug treatment, we ran a sensitivity analysis excluding all participants on antidiabetic medication. A p-value of <0.05 was considered significant. Analyses were performed with SPSS 14.0 for Windows.



Reporting analysis (example)

TABLE 1. Characteristics of the cohort

	No known diabetes	Known diabetes	Total	P value
Admissions	62.72 (710)	37.28 (422)	1132	
Patients	64.78 (629)	35.22 (342)	971	
HbA1c (%)	6.05 \pm 0.87	8.49 \pm 2.56	6.96 \pm 2.08	<0.001
HbA1c \geq 7.0 (yes)	9.44 (67)	69.43 (293)	31.80 (360)	<0.001
Admission glucose (mg/dl)	118.39 \pm 52.65	220.68 \pm 175.32	156.52 \pm 125.01	<0.001
Maximum glucose (mg/dl)	158.48 \pm 87.85	318.98 \pm 177.09	218.32 \pm 150.13	<0.001
Glucose \geq 200 mg/dl (yes)	17.61 (125)	72.04 (304)	37.90 (429)	<0.001
Age (yr)	56.62 \pm 18.16	61.17 \pm 14.70	58.32 \pm 17.09	0.001
Sex (male)	50.70 (360)	47.63 (201)	49.56 (561)	0.40
Race/ethnicity				0.04
Black	27.32 (194)	30.57 (129)	28.53 (323)	
White	15.35 (109)	9.95 (42)	13.34 (151)	
Hispanic	41.97 (298)	41.23 (174)	41.70 (474)	
Other	15.35 (109)	18.25 (77)	16.43 (186)	
Prior medication/hospital/clinic (yes)	63.66 (452)	76.25 (321)	68.35 (773)	<0.001
HTN (yes)	53.10 (377)	87.68 (370)	65.99 (747)	<0.001
Systolic BP (mm/Hg)	136.51 \pm 23.75	144.50 \pm 25.93	139.50 \pm 24.88	<0.001
Diastolic BP (mm/Hg)	76.43 \pm 15.12	76.90 \pm 14.68	76.61 \pm 14.95	0.64
BMI (kg/m ²)	27.55 \pm 7.62	28.88 \pm 7.46	28.04 \pm 7.58	0.03
LDL (mg/dl)	103.81 \pm 42.81	98.04 \pm 43.52	101.48 \pm 43.16	0.13
HDL (mg/dl)	50.70 \pm 20.88	45.45 \pm 16.94	48.54 \pm 19.52	<0.01
Triglycerides (mg/dl)	118.60 \pm 89.33	161.19 \pm 187.01	136.09 \pm 139.56	<0.01

Data are presented as mean \pm sd for continuous variables and percentage (number) for categorical variables based on total number of admissions. Totals may not equal 100% due to rounding. P values were calculated by generalized estimating equations. HTN, Hypertension; BP, blood pressure; LDL, low-density lipoprotein; HDL, high-density lipoprotein.



Reporting analysis (example)

29 April, 2014

29

Table 1. Sociodemographic Characteristics of the Participants

Sociodemographic Characteristics	Women (n=1800)			Men (n=1281)		
	n	Unweighted, %*	Weighted, %*	n	Unweighted, %*	Weighted, %*
Place of residence						
Urban	893	49.6	30.9	652	50.9	33.0
Rural	907	50.4	69.1	629	49.1	67.0
Age, y						
25 to 34	725	40.3	42.6	470	36.7	35.8
35 to 44	500	27.8	27.9	350	27.3	27.3
45 to 54	367	20.4	18.8	276	21.6	21.1
55 to 64	208	11.6	10.7	185	14.4	15.8
Education, y†						
None	685	38.1	44.7	196	15.3	21.2
1 to 4	642	35.7	37.3	423	33.1	36.1
5	191	10.6	7.5	248	19.4	19.4
6 to 7	164	9.1	6.1	199	15.6	12.2
≥8	115	6.4	4.4	213	16.6	11.2

*Within each variable, the sum of the proportions may not be 100% because of rounding.

†The sum of the number of participants in each category is <1800 for women and 1281 for men because of missing data.

www.jamalrahman.net



Reporting analysis (example)

30

Table 2. Prevalence of Hypertension Among Women and Men From Urban and Rural Areas According to Age, Education, Body Mass Index, Waist Circumference, and Current Alcohol Drinking

Participant Characteristics	Hypertension							
	Women				Men			
	Urban		Rural		Urban		Rural	
	%	95% CI	%	95% CI	%	95% CI	%	95% CI
All participants	41.0	36.4 to 45.7	26.8	19.5 to 34.1	40.0	34.0 to 46.1	33.5	27.8 to 39.2
Age, y								
25 to 34	17.6	12.7 to 22.5	11.1	5.9 to 16.2	31.8	25.0 to 38.6	32.8	26.4 to 39.2
35 to 44	43.2	31.8 to 54.5	27.1	18.8 to 35.5	35.3	25.7 to 44.9	27.7	18.8 to 36.6
45 to 54	69.5	57.6 to 81.3	45.5	35.6 to 55.4	49.8	39.6 to 60.1	32.0	20.3 to 43.7
55 to 64	73.0	64.0 to 81.9	57.9	44.9 to 70.9	59.4	38.3 to 80.5	46.0	46.0 to 60.8
Education, y								
0 to 4	45.8	38.0 to 53.6	27.0	19.7 to 34.3	42.7	26.3 to 59.1	33.7	26.1 to 41.3
5 to 7	37.6	32.4 to 42.9	18.5	7.6 to 29.4	39.0	28.3 to 49.7	32.8	21.8 to 43.8
≥8	27.2	20.8 to 33.6	46.8	4.6 to 89.0	38.3	29.5 to 47.2	35.0	4.0 to 66.0
Body mass index, kg/m ²								
<25.0	33.0	29.4 to 36.7	25.6	18.7 to 32.4	34.4	29.3 to 39.6	30.5	23.8 to 37.2
25.0 to 29.9	54.1	45.2 to 63.1	42.2	26.1 to 58.2	53.9	39.6 to 68.2	62.5	48.3 to 76.7
≥30	54.9	44.5 to 65.3	31.8	8.4 to 55.1	78.6	67.0 to 90.2	69.7	66.6 to 100.0
Waist circumference, cm								
Women <8 and men <102	38.0	33.2 to 42.8	27.4	19.9 to 34.9	38.3	32.5 to 44.1	32.5	27.0 to 38.0
Women ≥8 and men ≥102	60.0	50.4 to 69.6	50.9	26.1 to 75.7	79.3	65.3 to 93.2	100	*
Current drinking								
No	40.4	33.8 to 47.0	24.4	16.2 to 32.6	37.8	28.9 to 46.8	28.5	23.1 to 33.9
<1 d/wk	40.3	33.6 to 47.0	33.0	23.3 to 42.8	43.4	33.2 to 53.6	35.4	20.7 to 50.2
≥1 d/wk	48.2	33.1 to 63.3	34.6	20.1 to 49.1	38.6	27.7 to 49.5	40.5	31.4 to 49.6

*Only 1 subject was in this category.

www.jamalrahman.net

29 April, 2014



Summary

1. Identify & define variables
2. Type – independent vs. dependent
3. Level of measurements – nominal, ordinal or continuous
4. Check distribution – Normal vs. Not Normal
5. Decide what to do - descriptive vs. analytical



Chapter 2 Introduction to SPSS

IBM SPSS Statistics v21 for Windows

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



IBM SPSS Statistics



Figure 1: IBM SPSS Statistics integrates with a broad range of capabilities for the entire analytical process.

www.jamalrahman.net

IBM Corporation
 Software Group
 Route 100
 Somers, NY 10589
 Produced in the United States of America
 May 2012



29/4/2014

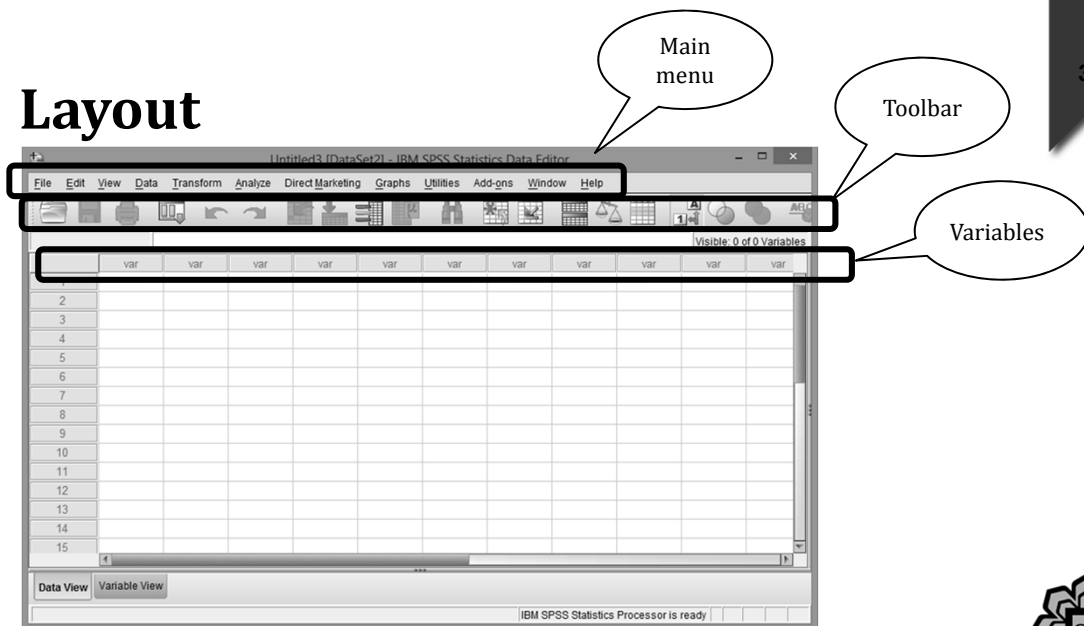
SPSS LAYOUT

www.jamalrahman.net



29/4/2014

Layout



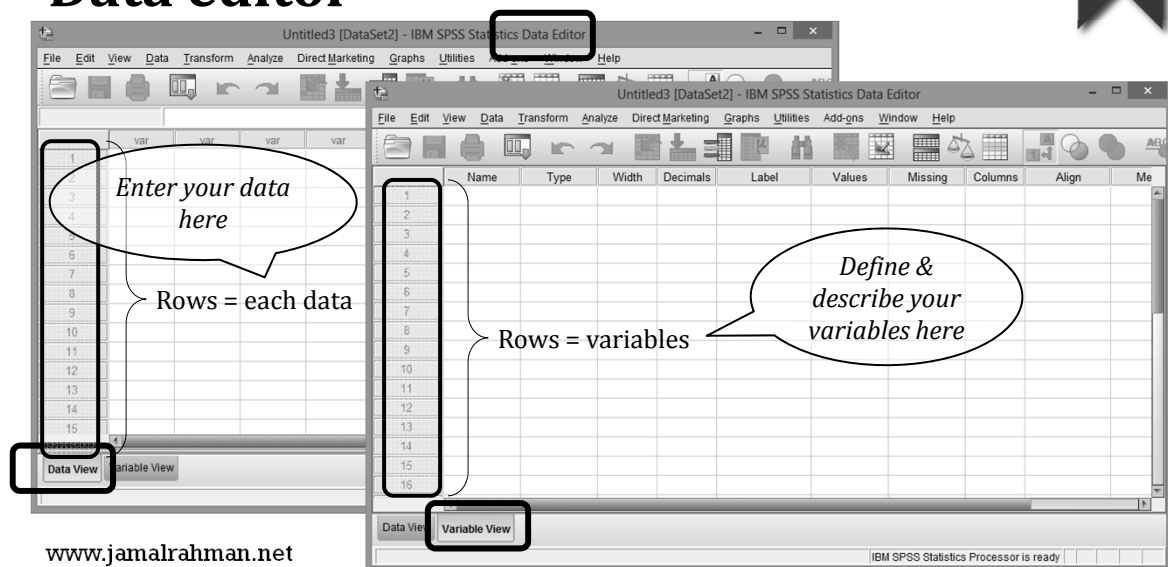
www.jamalrahman.net

29/4/2014



35

Data editor



www.jamalrahman.net

37

Viewer

Statistics

	Gender	Race	Exercise (paq)	Smoking	BMI Status
N	Valid 301	301	301	301	301
	Missing 0	0	0	0	0

Frequency Table

Gender

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid Female	180	59.8	59.8	59.8
Male	121	40.2	40.2	100.0

The output of analyses will be displayed here. Output is separated from data

www.jamalrahman.net

29/4/2014



38

Syntax

FREQUENCIES VARIABLES=gender race exercise smoking bmistat
/ORDER=ANALYSIS.

ONEWAY hcy hba1c BY exercise
/STATISTICS DESCRIPTIVES
/MISSING ANALYSIS.

T-TEST GROUPS=smoking(1 0)
/MISSING=ANALYSIS
/VARIABLES=hba1c hcy
/CRITERIA=CI(.95).

We can compile all the steps of the analyses here. Extend the programming function in SPSS. Ability to perform complex steps e.g. "looping"

www.jamalrahman.net

29/4/2014



39

CREATING DATASET

www.jamalrahman.net

29/4/2014



40

Before even you start SPSS!

- You must identify & define relevant variables
- Define means
 1. Name – preferably short single name, begins with alphabet, no special character, no space
 2. Type of data – e.g. Numeric, Date, String
 3. Width & Decimal Places (if numeric)
 4. Label – description for the Name (*will be displayed in Viewer*)
 5. Values – labels for value e.g. 1=Male, 2=Female
 6. Missing – define missing value e.g. 999 for N/A

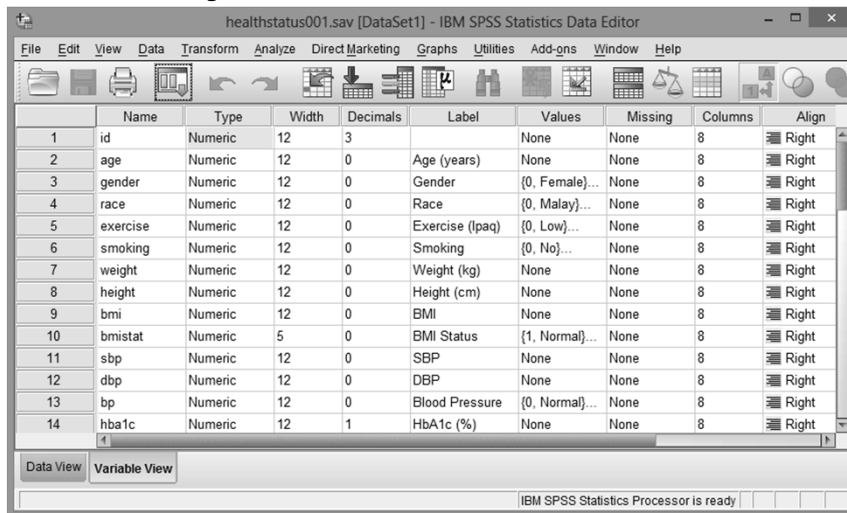
www.jamalrahman.net

29/4/2014



41

Define your variables



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align
1	id	Numeric	12	3		None	None	8	Right
2	age	Numeric	12	0	Age (years)	None	None	8	Right
3	gender	Numeric	12	0	Gender	{0, Female}...	None	8	Right
4	race	Numeric	12	0	Race	{0, Malay}...	None	8	Right
5	exercise	Numeric	12	0	Exercise (Ipaq)	{0, Low}...	None	8	Right
6	smoking	Numeric	12	0	Smoking	{0, No}...	None	8	Right
7	weight	Numeric	12	0	Weight (kg)	None	None	8	Right
8	height	Numeric	12	0	Height (cm)	None	None	8	Right
9	bmi	Numeric	12	0	BMI	None	None	8	Right
10	bmiestat	Numeric	5	0	BMI Status	{1, Normal}...	None	8	Right
11	sbp	Numeric	12	0	SBP	None	None	8	Right
12	dbp	Numeric	12	0	DBP	None	None	8	Right
13	bp	Numeric	12	0	Blood Pressure	{0, Normal}...	None	8	Right
14	hba1c	Numeric	12	1	HbA1c (%)	None	None	8	Right

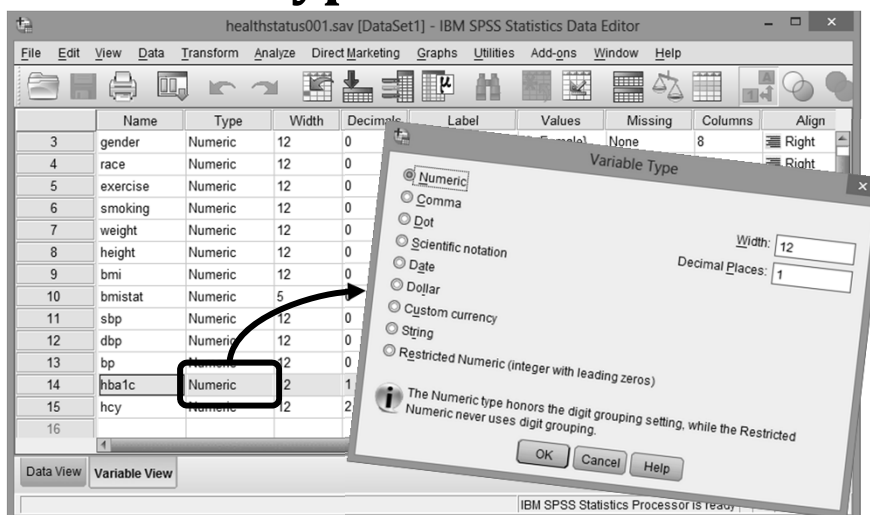
www.jamalrahman.net

29/4/2014



42

Variable Types



The screenshot shows the 'Variable Type' dialog box in IBM SPSS Statistics. The 'Numeric' radio button is selected. The 'Width' is set to 12 and 'Decimal Places' is set to 1. The dialog box also includes a note: 'The Numeric type honors the digit grouping setting, while the Restricted Numeric never uses digit grouping.' The background shows the Variable View of the same dataset as in slide 41, with the 'hba1c' variable highlighted.

www.jamalrahman.net

29/4/2014



Variable Type

Variable Type

For numeric, determine Width & Decimal. Decimal < Width

New option for Numerics with leading zeros

For String, no option for Decimal Place

Decide the suitable variable type

www.jamalrahman.net

29/4/2014

Value Labels

healthstatus001.sav [DataSet1] - IBM SPSS Statistics Data Editor

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align
1	id	Numeric	12	3		None	None	8	Right
2	age	Numeric	12	0	Age (years)	None	None	8	Right
3	gender	Numeric	12	0	Gender	{0, Female}...	None	8	Right
4	race	Numeric	12	0	Race	{0, Malay}...	None	8	Right
5	exercise	Numeric	12	0	Exercise (lapa)	{0, Low}...	None	8	Right
6	smoking	Numeric	12	0	Smoking	{0, No}...	None	8	Right
7	weight	Numeric	12	0	Weight (kg)	None	None	8	Right
8	height	Numeric	12	0	Height (cm)	None	None	8	Right
9	bmi	Numeric	12	0	BMI	None	None	8	Right
10	bmiestat	Numeric	5	0	BMI Status	{1, Normal}...	None	8	Right
11	sbp	Numeric	12	0	SBP	None	None	8	Right
12	dbp	Numeric	12	0	DBP	None	None	8	Right
13	bp	Numeric	12	0	Blood Pressure	{0, Normal}...	None	8	Right
14	hba1c	Numeric	12	1	HbA1c (%)	None	None	8	Right

Value Labels

Value:

Label:

0 = "Low"
1 = "Moderate"
2 = "High"

www.jamalrahman.net

29/4/2014

Missing Value

This question is Not Applicable to male

e.g. Assign 999 to represent N/A value & this won't be included in any analysis

IBM SPSS Statistics Processor is ready

29/4/2014

Chapter 3 Descriptive Statistics

IBM SPSS Statistics v21 for Windows

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

47

Exercise data

- Hypothetical
- Study to describe factors related to HbA1c & Homocystein (HCY)
- N=301
- 13 variables

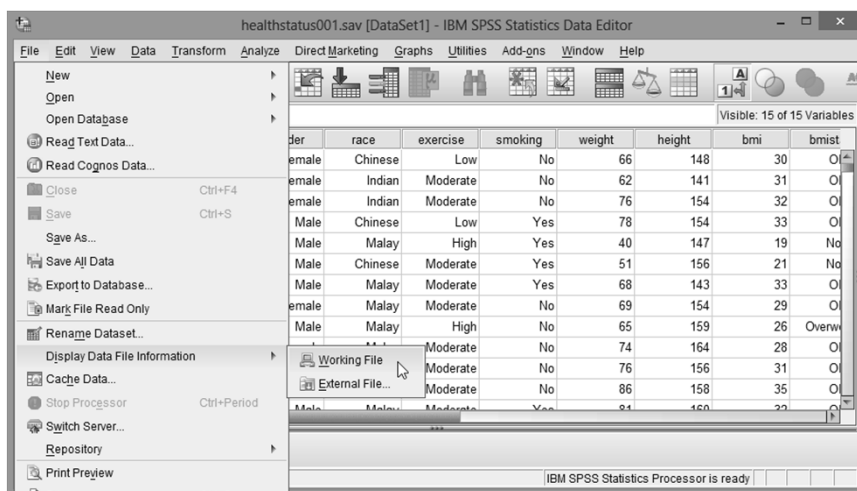
www.jamalrahman.net

29/4/2014



48

Retrieve file information



www.jamalrahman.net

29/4/2014



healthstatus001

Variable Information

Variable	Position	Label	Measurement Level	Role	Column Width	Alignment	Print Format	Write Format
id	1	<none>	Scale	Input	8	Right	F12.3	F12.3
age	2	Age (years)	Scale	Input	8	Right	F12	F12
gender	3	Gender	Nominal	Input	8	Right	F12	F12
race	4	Race	Nominal	Input	8	Right	F12	F12
exercise	5	Exercise (Ipaq)	Nominal	Input	8	Right	F12	F12
smoking	6	Smoking	Nominal	Input	8	Right	F12	F12
weight	7	Weight (kg)	Scale	Input	8	Right	F12	F12
height	8	Height (cm)	Scale	Input	8	Right	F12	F12
sbp	9	SBP	Scale	Input	8	Right	F12	F12
dbp	10	DBP	Scale	Input	8	Right	F12	F12
hba1c	11	HbA1c (%)	Scale	Input	8	Right	F12.1	F12.1
hcy	12	Homocysteine (umol/L)	Scale	Input	8	Right	F12.2	F12.2

Variables in the working file

Variable Values

Value	Label
gender 0	Female
1	Male
race 0	Malay
1	Chinese
2	Indian
exercise 0	Low
1	Moderate
2	High
smoking 0	No
1	Yes

www.jamalrahman.net

29/4/2014



Task 1

1. Describe socio-demographic characteristics of the respondent (age, gender & race)
2. Describe the explanatory variables
 1. Exercise
 2. smoking status
 3. BMI status &
 4. BP status
3. Describe HbA1c (taking cut-off for Poor HbA1c $\geq 6.5\%$) & HCY

www.jamalrahman.net

29/4/2014



50

51

DESCRIBE NUMERICAL DATA

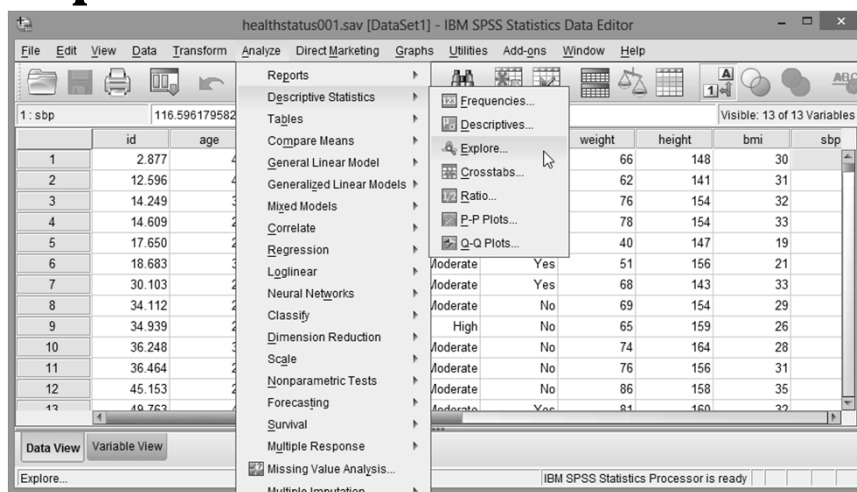
www.jamalrahman.net

29/4/2014



52

Explore

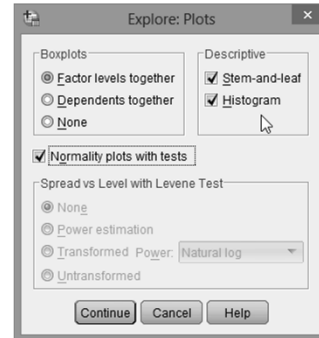
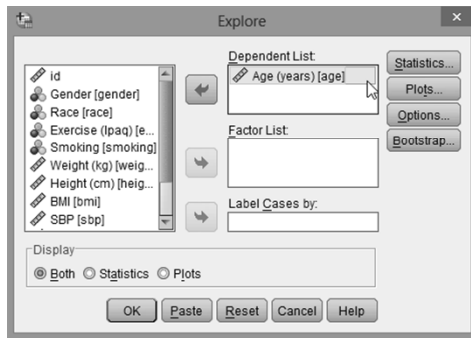


www.jamalrahman.net

29/4/2014



53



www.jamalrahman.net

29/4/2014



Results

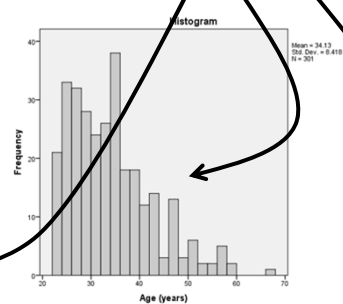
Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Age (years)	301	100.0%	0	0.0%	301	100.0%

Descriptives

		Statistic	Std. Error
Age (years)	Mean	34.13	.485
	95% Confidence Interval for Mean	Lower Bound	33.15
		Upper Bound	35.09
	5% Trimmed Mean	32.60	
	Median	32.60	
	Variance	8.418	
	Std. Deviation	2.901	
	Minimum	11	
	Maximum	67	
	Range	56	
Interquartile Range		11	
Skewness		1.002	.140
Kurtosis		.837	.280

Check for Normality.
Is Age data distributed Normally?



Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Age (years)	.093	301	.000	.925	301	.000

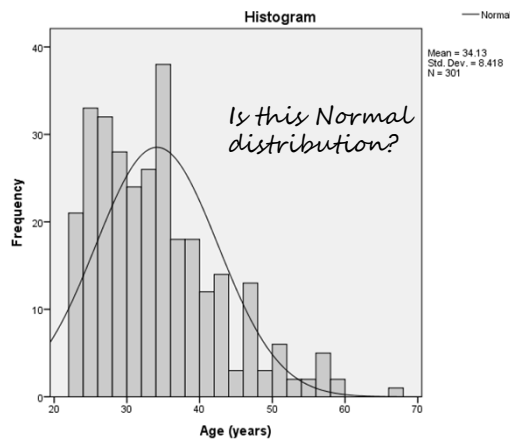
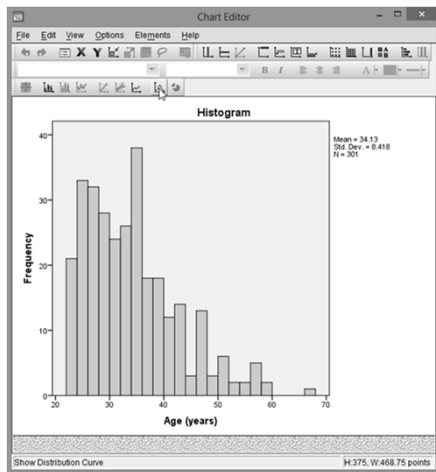
a. Lilliefors Significance Correction

www.jamalrahman.net

29/4/2014



55



www.jamalrahman.net

29/4/2014



56

Describe age

Normal

- The subjects distributed between 23-67 years old with the average of 34 (SD=8) years.

If not Normal

- The subjects distributed between 23-67 years old with the median of 33 (IQR=11) years

www.jamalrahman.net

29/4/2014



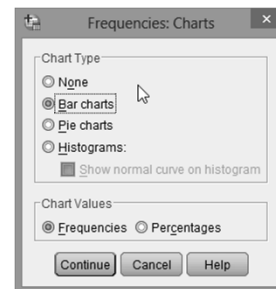
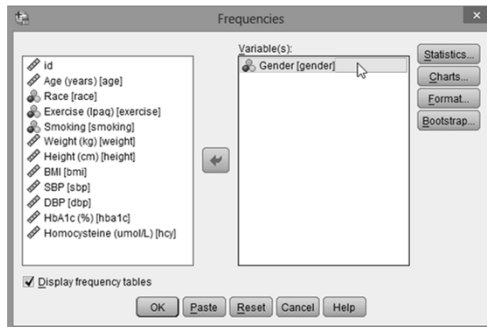
29/4/2014

58



29/4/2014

59



www.jamalrahman.net

29/4/2014



60

Results

Statistics

Gender

N	Valid	301
	Missing	0

Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	180	59.8	59.8	59.8
	Male	121	40.2	40.2	100.0
	Total	301	100.0	100.0	

www.jamalrahman.net

29/4/2014



61

TRANSFORM

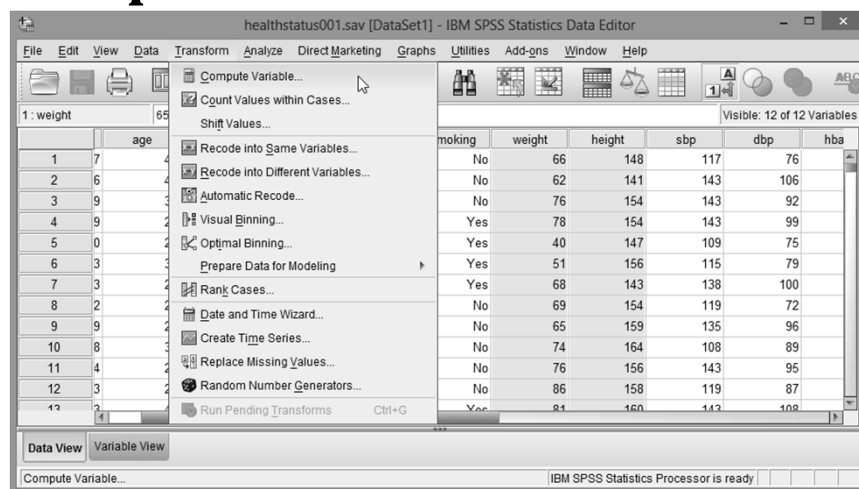
www.jamalrahman.net

29/4/2014



62

Compute

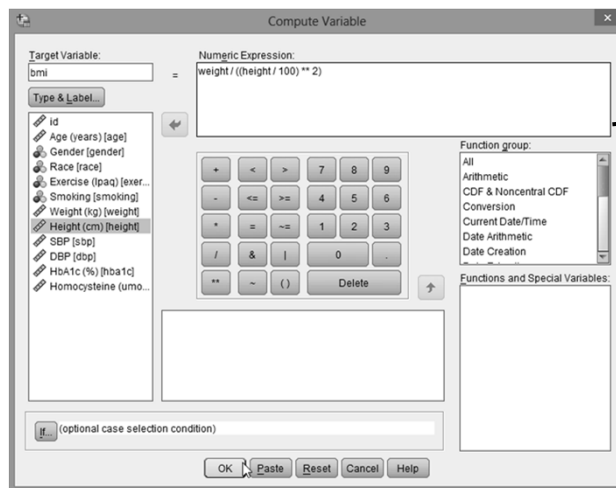


www.jamalrahman.net

29/4/2014



63



$$\text{weight} / ((\text{height} / 100) ** 2)$$

1. bmi	rcise	smoking	weight	height	sbp	dbp	hba1c	hcy	bmi
1	Low	No	66	148	117	76	5.8	6.67	29.38
2	oderate	No	62	141	143	106	9.3	12.78	27.24
3	oderate	No	76	154	143	92	10.1	15.35	32.19
4	Low	Yes	78	154	143	99	9.7	12.63	33.02
5	High	Yes	40	147	109	75	4.2	9.63	18.65
6	oderate	Yes	51	156	115	79	11.3	8.36	20.65
7	oderate	Yes	68	143	138	100	12.0	15.75	33.30
8	oderate	No	69	154	119	72	4.1	8.18	29.13
9	High	No	60	159	135	96	8.2	8.64	25.77
10	oderate	No	74	164	108	89	5.0	7.69	27.74
11	oderate	No	76	156	143	95	8.5	12.04	31.40
12	oderate	No	85	160	119	87	5.1	8.54	34.63
13	oderate	No	51	155	113	102	7.4	14.41	16.44

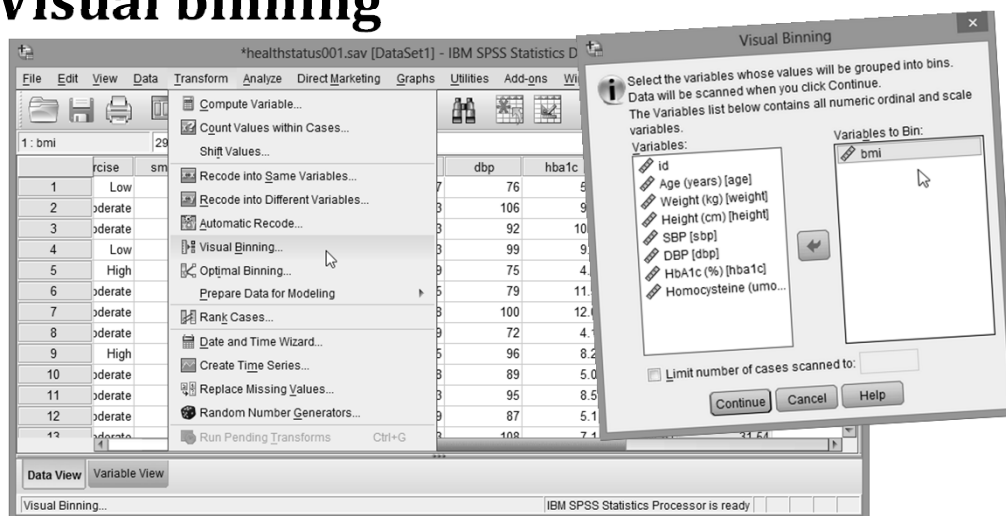
www.jamalrahman.net

29/4/2014



Visual binning

64



www.jamalrahman.net

29/4/2014



65

Visual Binning

Scanned Variable List: **bmi**

Current Variable: **bmi** Name: **bmi** Label: **bmi**

Binned Variable: **bmicat** BMI Status

Minimum: 17.92 Nonmissing Values Maximum: 40.45

Enter interval cutpoints or click Make Cutpoints for automatic intervals. A cutpoint for example, defines an interval starting above the previous interval and ending at the specified value.

Grid:

	Value	Label
1	23.000	Normal
2	27.500	Overweight
3	HIGH (Obese)	
4		

Upper Endpoints: ☒ Included (\leq) ☒ Excluded ($<$)

Buttons: OK, Paste, Reset, Cancel, Help

IBM SPSS Statistics 21

Binning specifications will create 1 variables.

Buttons: OK, Cancel

Normal < 23
Overweight 23 to < 27.5
Obese >= 27.5

www.jamalrahman.net

29/4/2014



66

*healthstatus001.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

1: bmicat 3 Visible: 14 of 14 Variables

	smoking	weight	height	sbp	dbp	hba1c	hcy	bmi	bmicat	var
1	No	66	148	117	76	5.8	6.67	29.98	Obese	
2	No	62	141	143	106	9.3	12.78	31.24	Obese	
3	No	76	154	143	92	10.1	15.35	32.19	Obese	
4	Yes	78	154	143	99	9.7	12.63	33.02	Obese	
5	Yes	40	147	109	75	4.2	9.65	18.65	Normal	
6	Yes	51	156	115	79	11.3	8.36	20.65	Normal	
7	Yes	68	143	138	100	12.0	15.75	33.30	Obese	
8	No	69	154	119	72	4.1	8.18	29.13	Obese	
9	No	65	159	135	96	8.2	8.04	25.77	Overweight	
10	No	74	164	108	89	5.0	7.69	27.74	Obese	
11	No	76	156	143	95	8.5	12.04	31.40	Obese	
12	No	86	158	119	87	5.1	8.94	34.63	Obese	
13	No	84	160	143	100	7.1	11.51	31.51	Obese	

Data View Variable View

IBM SPSS Statistics Processor is ready

www.jamalrahman.net

29/4/2014



Chapter 4

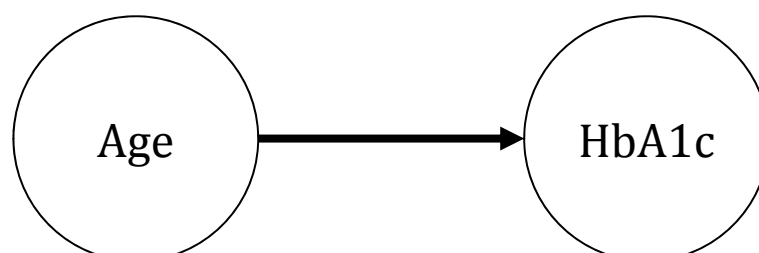
Bivariable analyses

IBM SPSS Statistics v21 for Windows

Jamalludin Ab Rahman MD MPH
Department of Community Medicine
Kulliyah of Medicine



To check association of two variables?



69

The steps

1. Determine which is dependant & which is independent
2. Determine level of measurements
3. Determine Normality of the numerical measurement
4. Determine the suitable statistical test

www.jamalrahman.net

29/4/2014



70

What are the tests?

Variable 1	Variable 2	Test
Categorical	Categorical	Chi-square
Categorical (2 pop)	Numerical (Normal)	Independent sample t-test
Categorical (2 pop)	Numerical (Not Normal)	Mann-Whitney U test
Categorical (> 2 pop)	Numerical (Normal)	One-way ANOVA
Categorical (> 2 pop)	Numerical (Not Normal)	Kruskal-Wallis test
Numerical (Normal)	Numerical (Normal)	Pearson Correlation Coefficient Test
Numerical (Normal/ Not Normal)	Numerical (Not Normal)	Spearman Correlation Coefficient Test
Numerical (Normal)	Numerical (Normal) – Paired	Paired t-test
Numerical (Not Normal)	Numerical (Not Normal) – Paired	Friedman test

www.jamalrahman.net

29 April, 2014



71

Tasks 2

1. Determine association between socio-demographic characteristics & all the risk factors with HbA1c
2. Determine association between socio-demographic characteristics & all the risk factors with HCY

Note: It would be good if you could construct **dummy table** for the answers even before the analyses started

www.jamalrahman.net

29/4/2014



72

HCY normal range

Blood reference ranges for homocysteine:

Sex	Age	Lower limit	Upper limit	Unit	Elevated	Therapeutic target
Female	12–19 years	3.3 ^[12]	7.2 ^[12]	μmol/L	> 10.4 μmol/L or > 140 μg/dl	< 6.3 μmol/L ^[13] or < 85 μg/dL ^[13]
		45 ^[14]	100 ^[14]	μg/dL		
	>60 years	4.9 ^[12]	11.6 ^[12]	μmol/L		
		66 ^[14]	160 ^[14]	μg/dL		
Male	12–19 years	4.3 ^[12]	9.9 ^[12]	μmol/L	> 11.4 μmol/L or > 150 μg/dL	< 6.3 μmol/L ^[13] or < 85 μg/dL ^[13]
		60 ^[14]	130 ^[14]	μg/dL		
	>60 years	5.9 ^[12]	15.3 ^[12]	μmol/L		
		80 ^[14]	210 ^[14]	μg/dL		

www.jamalrahman.net

29/4/2014



COMPARING TWO MEANS

73

INDEPENDENT SAMPLE t-TEST

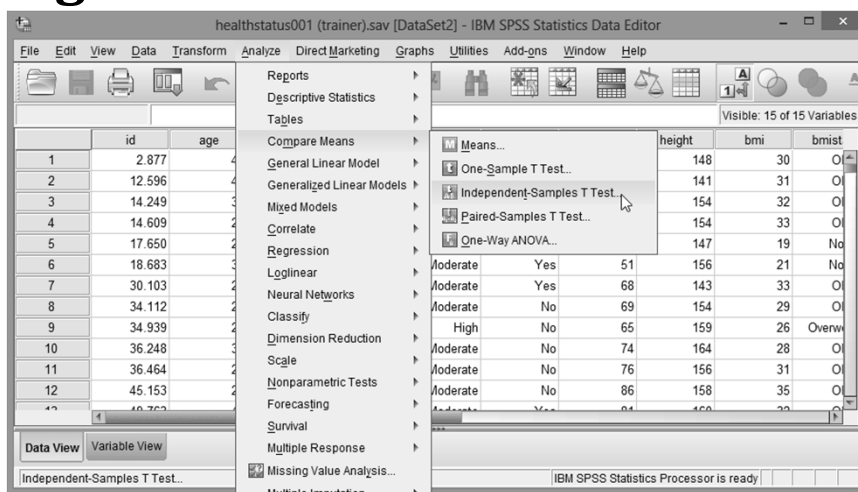
www.jamalrahman.net

29/4/2014



Age vs. BP

74

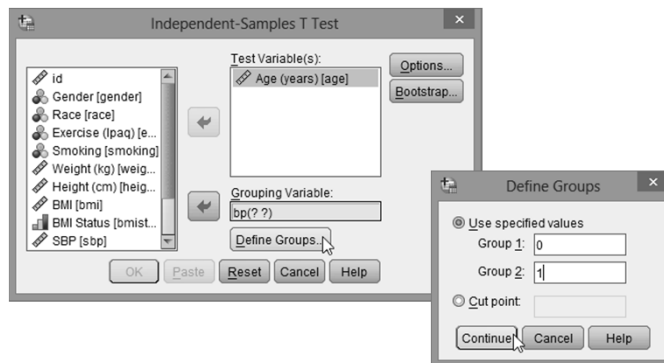


www.jamalrahman.net

29/4/2014



75



www.jamalrahman.net

29/4/2014



Results

The original *t*-test (Student's *t*-test) assumes equal variances for equal sample sizes. However if the variances are equal, it is robust for different sizes.

Group Statistics				
	Blood Pressure	N	Mean	Std. Deviation
Age (years)	Normal	156	33.93	7.928
	High	145	34.35	8.937

Independent Samples Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Age (years)	Equal variances assumed	3.522	.062	-.431	299	.667	-.419	.972	-2.332	1.495
	Equal variances not assumed			-.429	288.372	.668	-.419	.977	-2.341	1.503

Welch's correction

Levene's test check equality between variances. Ho: There is no difference of variances. So if P is significant, we reject Ho, and therefore equal variances assumed.

www.jamalrahman.net

29/4/2014



77

Table – Distribution of age by blood pressure status

	N	Mean	SD	Statistics	df	P
Normal BP	156	33.9	7.9	t=0.431	299	0.667
High BP	145	34.4	8.9			

www.jamalrahman.net

29/4/2014



DIFFERENCE OF TWO PROPORTIONS

78

CHI-SQUARED TEST

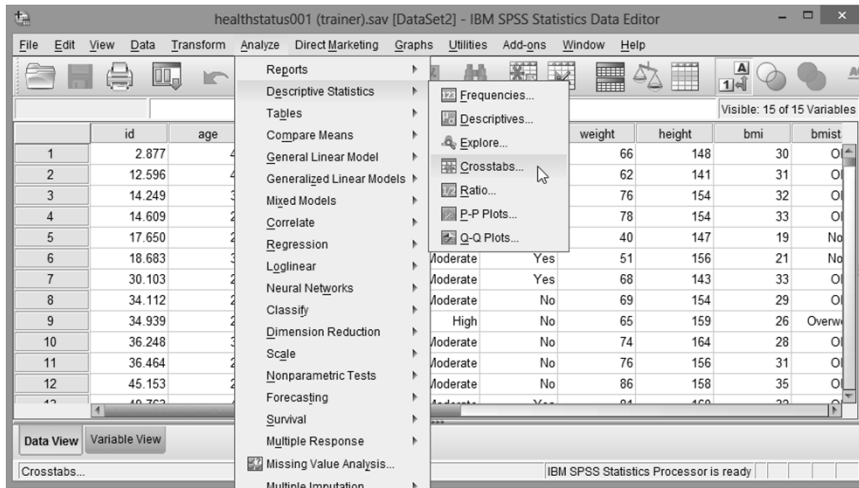
www.jamalrahman.net

29/4/2014



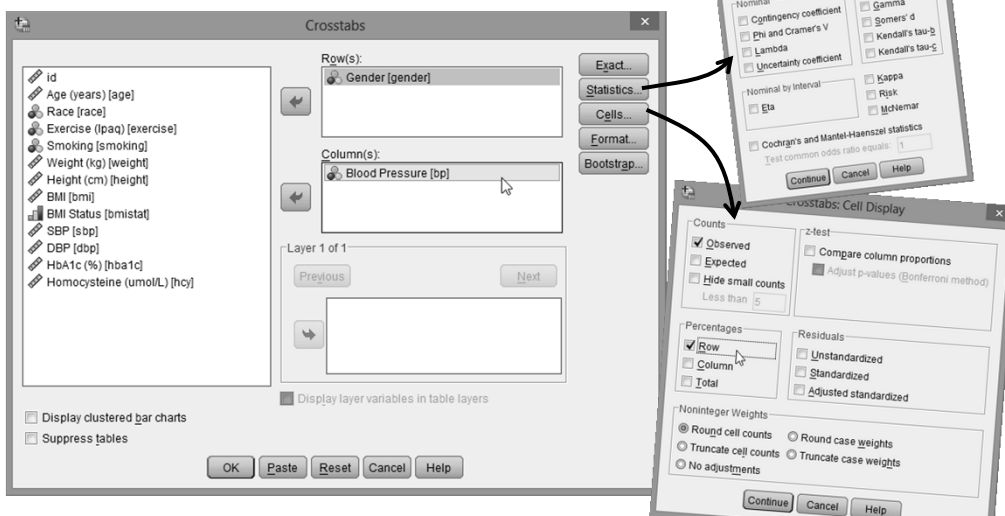
Gender vs. BP

79



www.jamalrahman.net

29/4/2014



www.jamalrahman.net

29/4/2014



Results

81

Some books may suggest the use of Continuity Correction at ALL time, but recent simulations showed that CC (or Yate's correction) is OVERCONSERVATIVE. Hence, use Pearson χ^2 when $< 20\%$ of cells have expected count < 5

Case Processing Summary						
		Cases				
		Valid		Missing		Total
		N	Percent	N	Percent	N
Gender * Blood Pressure		301	100.0%	0	0.0%	301

Gender * Blood Pressure Crosstabulation						
Gender			Blood Pressure		Total	
			Normal	High		
Female	Count		92	88	180	
	% within Gender		51.1%	48.9%	100.0%	
Male	Count		64	57	121	
	% within Gender		52.9%	47.1%	100.0%	
Total	Count		156	145	301	
	% within Gender		51.8%	48.2%	100.0%	

Describe this table first. What is your impression? 49% women vs. 47% men with high BP

www.jamalrahman.net

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.092 ^a	1	.762		
Continuity Correction ^b	.034	1	.858		
Likelihood Ratio	.092	1	.762		
Fisher's Exact Test				.814	.426
Linear-by-Linear Association	.092	1	.762		
N of Valid Cases	301				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 58.29.
b. Computed only for a 2x2 table

When $\geq 20\%$ of cells have $EC < 5$, use Fisher's Exact Test

This is given because we code the variables using numbers. Can be used to measure P-trend

29/4/2014



COMPARING MORE THAN TWO MEANS

82

ONE-WAY ANOVA

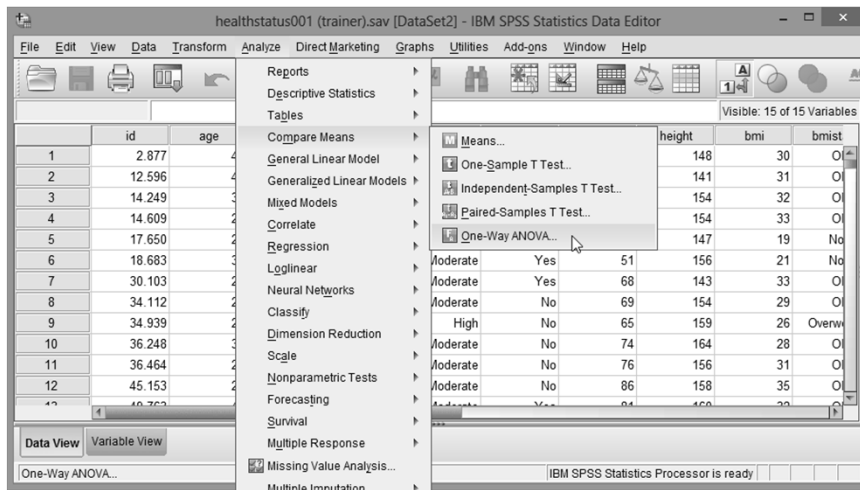
www.jamalrahman.net

29/4/2014



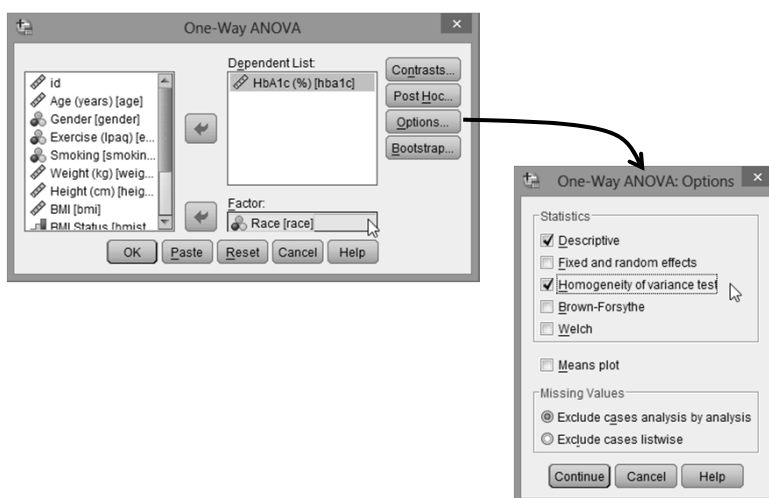
Race vs. HbA1c

83



www.jamalrahman.net

29/4/2014



www.jamalrahman.net

29/4/2014



85

Results

Describe these results first. What is your impression? HbA1c between races? 6.4 (SD 2.1) vs. 6.7 (SD 2.2) vs. 6.5 (SD 2.2)

Descriptives								
HbA1c (%)								
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Malay	195	6.386	2.1212	.1519	6.086	6.685	1.3	12.0
Chinese	100	6.684	2.1839	.2184	6.251	7.118	1.2	12.4
Indian	6	8.295	2.9488	1.2038	5.201	11.390	2.7	11.1
Total	301	6.523	2.1708	.1251	6.277	6.769	1.2	12.4

Test of Homogeneity of Variances			
HbA1c (%)			
Levene Statistic	df1	df2	Sig.
.171	2	298	.843

ANOVA					
HbA1c (%)					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	25.137	2	12.568	2.697	.069
Within Groups	1388.561	298	4.660		
Total	1413.698	300			

The F test shows that there is no single significant difference between any two groups

www.jamalrahman.net

29/4/2014



86

Results – BMI status vs. HbA1c

Descriptives								
HbA1c (%)								
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Normal	85	6.143	2.1100	.2289	5.688	6.598	1.3	12.4
Overweight	73	5.733	2.0389	.2386	5.257	6.208	1.2	9.6
Obese	143	7.152	2.0995	.1756	6.805	7.499	2.2	12.0
Total	301	6.523	2.1708	.1251	6.277	6.769	1.2	12.4

Test of Homogeneity of Variances			
HbA1c (%)			
Levene Statistic	df1	df2	Sig.
.082	2	298	.922

ANOVA					
HbA1c (%)					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	114.483	2	57.242	13.129	.000
Within Groups	1299.215	298	4.360		
Total	1413.698	300			

The F test shows that at least there is ONE pair with significant difference. Either N vs. OW, N vs. OB or OW vs. OB

We need to run Post-hoc test to determine which of the PAIR is significant.

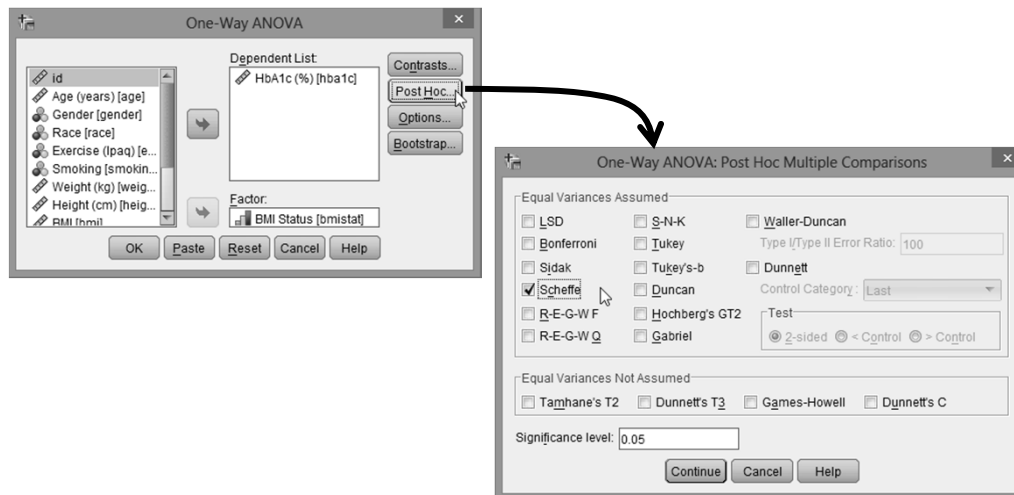
To decide which Post-hoc test to choose, we have to test for equality of variances i.e. Homogeneity of variances (Levene's test)

www.jamalrahman.net

29/4/2014



87



www.jamalrahman.net

29/4/2014



Results – Post hoc

88

Post Hoc Tests

Multiple Comparisons

Dependent Variable: HbA1c (%)
Scheffe

(I) BMI Status	(J) BMI Status	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Normal	Overweight	.4100	.3332	.470	-.410	1.230
	Obese	-1.0094*	.2860	.002	-1.579	-.440
Overweight	Normal	-.4100	.3332	.470	-1.230	.410
	Obese	-1.4194*	.3004	.000	-2.158	-.681
Obese	Normal	1.0094*	.2860	.002	.306	1.713
	Overweight	1.4194*	.3004	.000	.681	2.158

*. The mean difference is significant at the 0.05 level.

The significant difference is only for Normal vs. Obese (P=0.002)

www.jamalrahman.net

29/4/2014



Report

There is a significant association between BMI Status and HbA1c ($F(2,298)=13.129$, $P<0.001$). Post-hoc test showed that Obese subjects have significantly higher HbA1c compared to Normal and Overweight subjects ($P=0.001$ and $P < 0.001$ respectively).



NON PARAMETRIC TESTS

MANN-WHITNEY U



Gender vs. HCY

91

Two-Independent-Samples Tests

Test Variable List: Homocysteine (umo)

Grouping Variable: gender(0 1)

Test Type: ☒ Mann-Whitney U

OK Paste Reset Cancel Help

www.jamalrahman.net

29/4/2014



Results

92

Mann-Whitney Test

Ranks				
Gender	N	Mean Rank	Sum of Ranks	
Homocysteine (umol/L)				
Female	180	151.83	27329.00	
Male	121	149.77	18122.00	
Total	301			

This ranks table is not to be cited in the research paper. Instead, describe their MEDIAN

Test Statistics ^a	
	Homocysteine (umol/L)
Mann-Whitney U	10741.000
Wilcoxon W	18122.000
Z	-.203
Asymp. Sig. (2-tailed)	.840

a. Grouping Variable: Gender

www.jamalrahman.net

29/4/2014



NON-PARAMETRIC TESTS

93

KRUSKAL WALLIS

www.jamalrahman.net

29/4/2014



Race vs. HCY

94

The screenshot shows the IBM SPSS Statistics Data Editor window with a dataset named 'healthstatus001 (trainer).sav'. The dataset contains variables: exercise, smoking, bmisat, sbp, dbp, bp, hba1c, hcy, and race. The 'Analyze' menu is open, and the 'Nonparametric Tests' submenu is selected. The 'Kruskal-Wallis H' test is chosen for the variable 'hcy'. The 'Tests for Several Independent Samples' dialog box is also visible, showing the 'Test Variable List' with 'Homocysteine (umo...)' and the 'Grouping Variable' set to 'race(0 2)'. The 'Test Type' section shows 'Kruskal-Wallis H' selected.

www.jamalrahman.net

29/4/2014



Results

95

Kruskal-Wallis Test

Ranks			
	Race	N	Mean Rank
Homocysteine (umol/L)	Malay	195	145.85
	Chinese	100	156.44
	Indian	6	227.58
	Total	301	

Test Statistics ^{a,b}	
	Homocysteine (umol/L)
Chi-Square	5.718
df	2
Asymp. Sig.	.057

a. Kruskal Wallis Test

b. Grouping Variable:
Race
www.jamalrahman.net

29/4/2014



RELATIONSHIP OF TWO NUMERICAL DATA

96

CORRELATION TEST

www.jamalrahman.net

29/4/2014



Age vs. HbA1c

97


www.jamalrahman.net

29/4/2014



Results

98

Correlations			
		Age (years)	HbA1c (%)
Age (years)	Pearson Correlation	1	.064
	Sig. (2-tailed)		.271
	N	301	301
HbA1c (%)	Pearson Correlation	.064	1
	Sig. (2-tailed)	.271	
	N	301	301

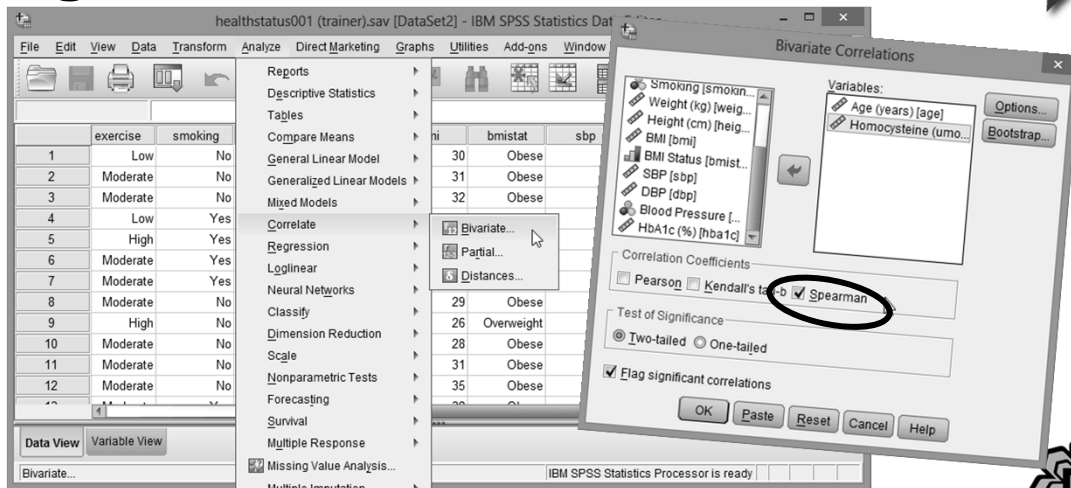
www.jamalrahman.net

29/4/2014



99

Age vs. HCY



www.jamalrahman.net

29/4/2014



100

Correlations

			Age (years)	Homocysteine (umol/L)
Spearman's rho	Age (years)	Correlation Coefficient	1.000	-.058
		Sig. (2-tailed)	.	.314
		N	301	301
	Homocysteine (umol/L)	Correlation Coefficient	-.058	1.000
		Sig. (2-tailed)	.314	.
		N	301	301

www.jamalrahman.net

29/4/2014

