Data analysis and hypothesis testing

Suwanna Sayruamyat

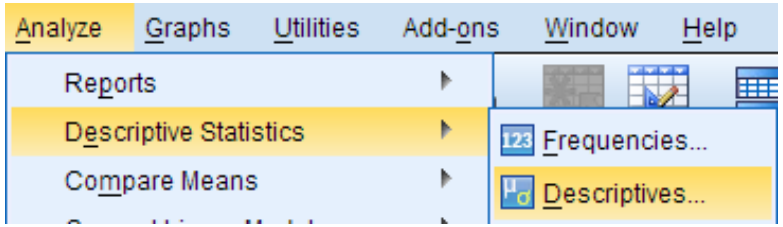
BA603

Descriptive statistics

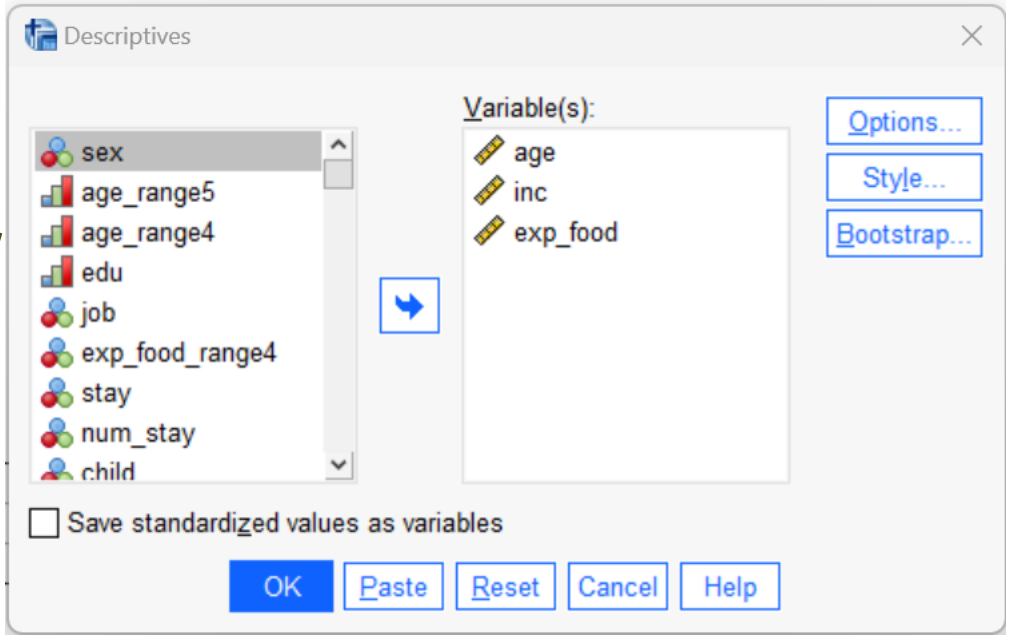
- When summarizing quantitative (continuous/interval/ratio) variables, we are typically interested in questions like:
  - What is the "center" of the data? (Mean, median)
  - How spread out is the data? (Standard deviation/variance)
  - What are the extremes of the data? (Minimum, maximum; Outliers)
  - What is the "shape" of the distribution? Is it symmetric or asymmetric? Are the values mostly clustered about the mean, or are there many values in the "tails" of the distribution? (Skewness, kurtosis)

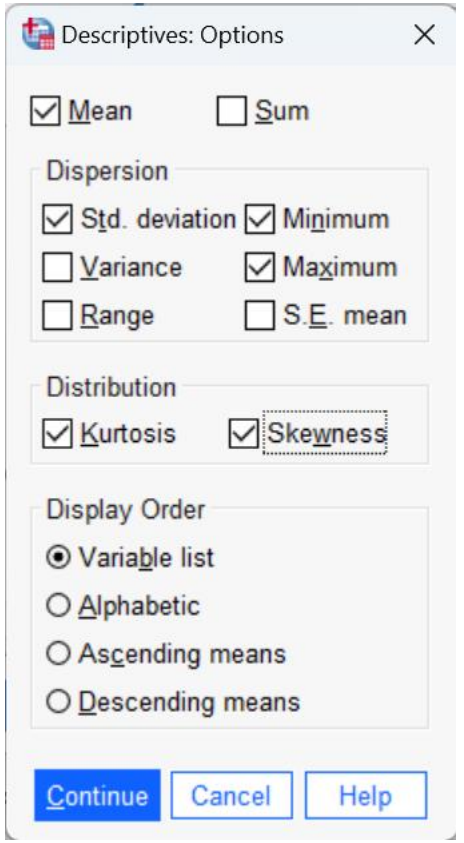
Descriptives

- To run the Descriptives procedure, select Analyze > Descriptive Statistics > Descriptives.



Select variables for analysis





Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
age	526	20.00	74.00	32.6369	10.16227	1.315	.106	1.759	.213
inc	526	600.00	300000.00	24609.5627	24439.06623	6.014	.106	53.336	.213
exp_food	526	60.00	60000.00	6878.0798	5542.18815	3.343	.106	19.879	.213
Valid N (listwise)	526								

ค่าความเบ้ (Skewness)

ค่าความเบ้ (Skewness) คือ ค่าที่ใช้วัดความไม่สมมาตรของการกระจายตัวของข้อมูล ซึ่งจะบอกเราว่าข้อมูลมีการกระจายตัวเอียงไปทางด้านใดด้านหนึ่งหรือไม่

ค่าความเบ้เป็นบวก (Positive Skewness)

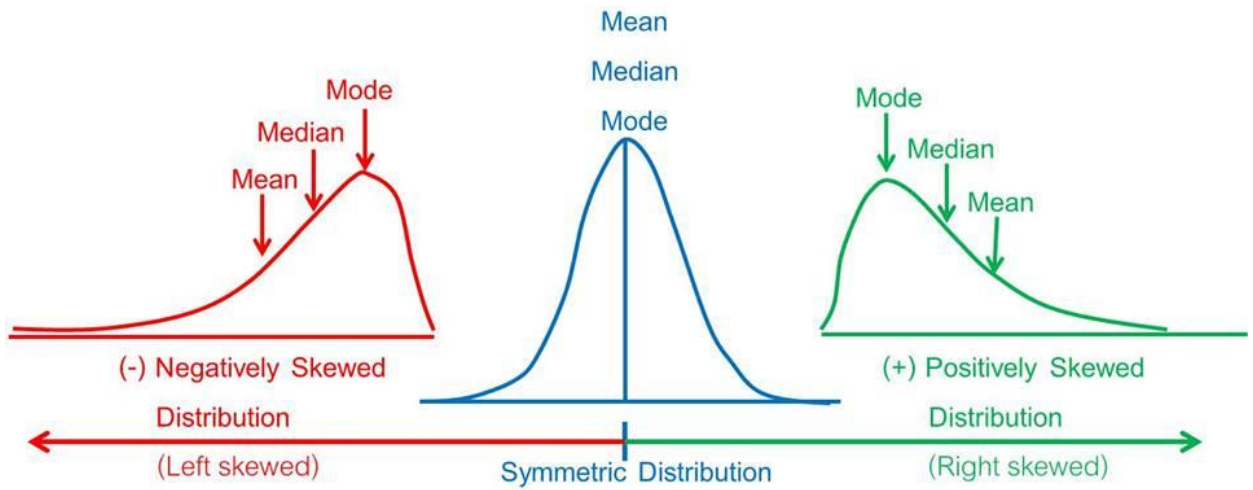
- ข้อมูลมีการกระจายตัวเอียงไปทางด้านขวา
- หางของกราฟจะยาวไปทางด้านขวา
- ค่าเฉลี่ย (Mean) จะมีค่ามากกว่าค่ามัธยฐาน (Median) และค่าฐานนิยม (Mode)
- ข้อมูลส่วนใหญ่จะกระจุกตัวอยู่ทางด้านซ้าย

ค่าความเบ้เป็นลบ (Negative Skewness

- ข้อมูลมีการกระจายตัวเอียงไปทางด้านซ้าย
- หางของกราฟจะยาวไปทางด้านซ้าย
- ค่าเฉลี่ย (Mean) จะมีค่าน้อยกว่าค่ามัธยฐาน (Median) และค่าฐานนิยม (Mode)
- ข้อมูลส่วนใหญ่จะกระจุกตัวอยู่ทางด้านขวา

ค่าความเบ้เป็นลบ (Negative Skewness

- ข้อมูลมีการกระจายตัวแบบสมมาตร (Symmetrical Distribution)
- กราฟจะมีรูปร่างเป็นระฆังคว่ำ (Bell-shaped)
- ค่าเฉลี่ย (Mean), ค่ามัธยฐาน (Median) และค่าฐานนิยม (Mode) จะมีค่าเท่ากัน



ค่าความโด่ง (Kurtosis)

Mesokurtic (ความโด่งปกติ)

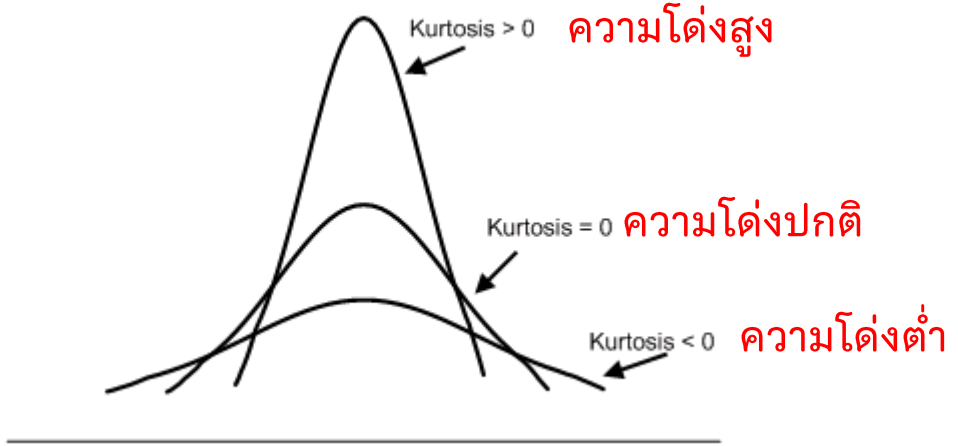
- เป็นการแจกแจงที่มีความโด่งปกติ คล้ายกับการแจกแจงแบบปกติ (Normal distribution)
- ค่าความโด่งมีค่าประมาณ 3 หรือ 0 (ขึ้นอยู่กับการคำนวณ)

Platykurtic (ความโด่งต่ำ)

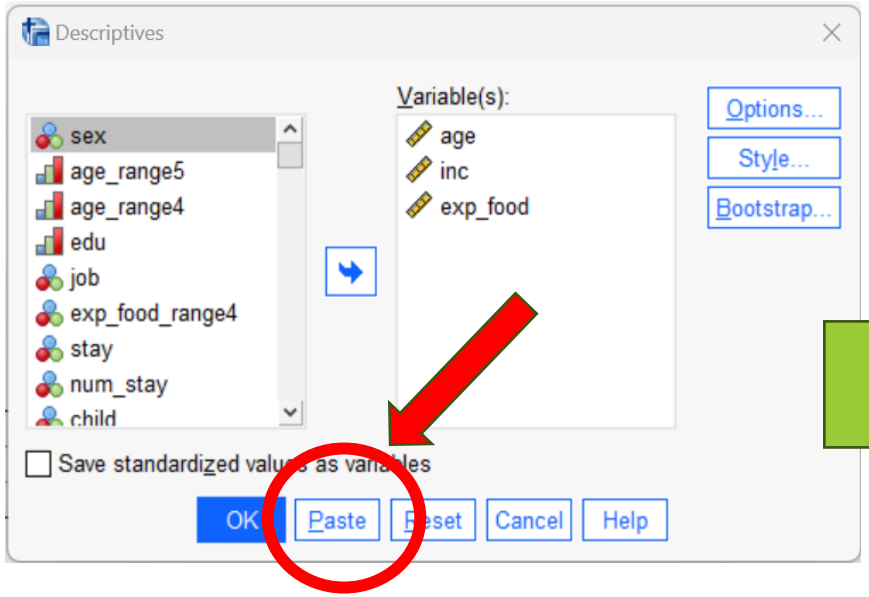
- เป็นการแจกแจงที่มีความโด่งต่ำ ยอดแบน และหางบาง
- ข้อมูลมีการกระจายตัวค่อนข้างสม่ำเสมอ และมีข้อมูลที่ห่างจากค่าเฉลี่ยน้อยกว่าการแจกแจงแบบปกติ
- ค่าความโด่งมีค่าน้อยกว่า 3 หรือน้อยกว่า 0

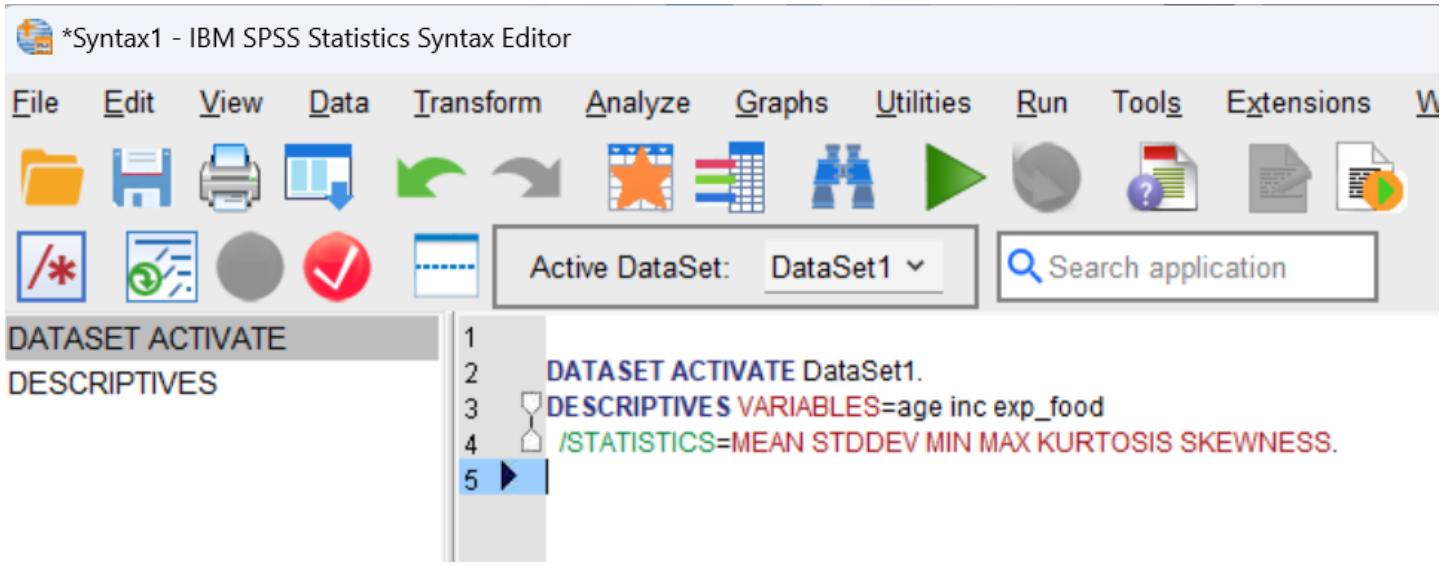
Leptokurtic (ความโด่งสูง)

- เป็นการแจกแจงที่มีความโด่งสูง ยอดแหลม และหางหนามีข้อมูลกระจุกตัวอยู่บริเวณค่าเฉลี่ยมาก และมีข้อมูลที่ห่างจากค่าเฉลี่ยมาก (Outliers) มากกว่าการแจกแจงแบบปกติค่าความโด่งมีค่ามากกว่า 3 หรือมากกว่า 0



Using Syntax

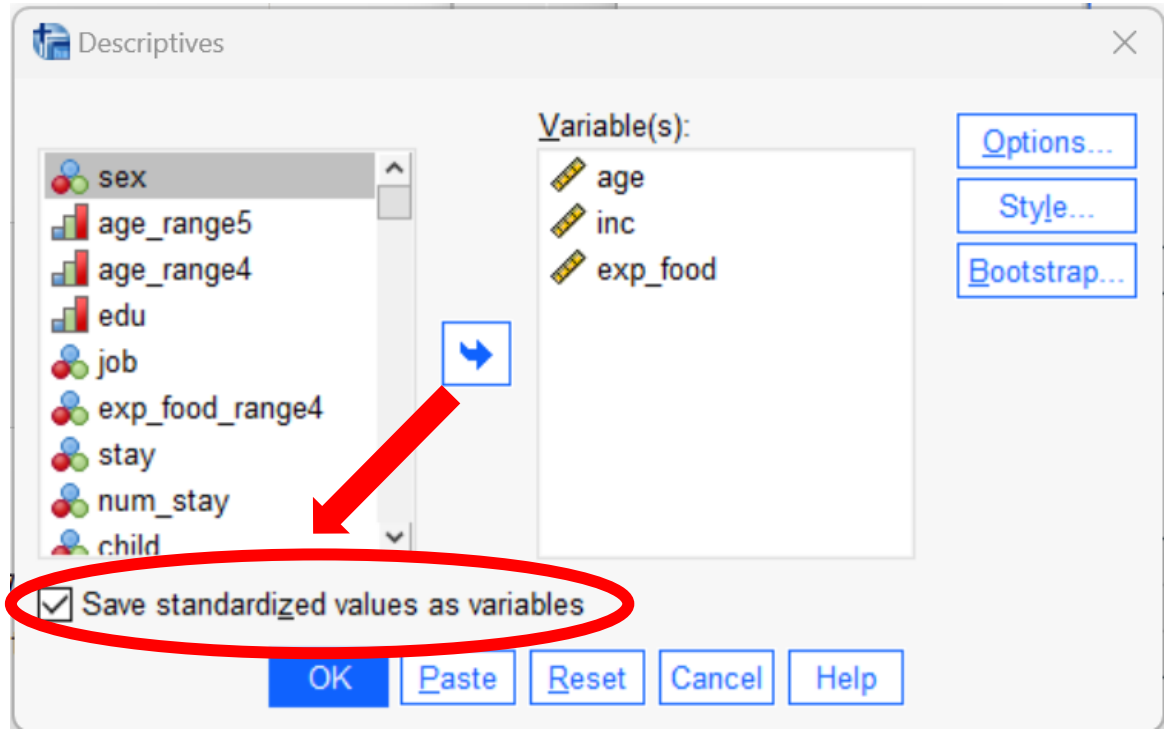




Compute standardized scores (Z-score) for numeric variables

Using the Descriptives Dialog Window

1. Click **Analyze > Descriptive Statistics > Descriptives**.
2. Add the variables *age, inc and exp_food* to the Variables box.
3. Check the box **Save standardized values as variables**.
4. Click OK when finished.

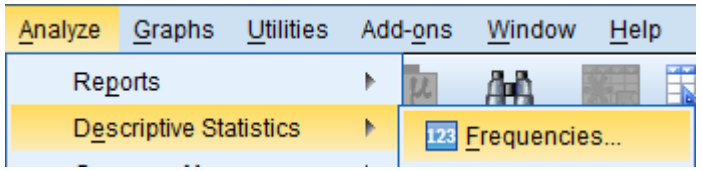


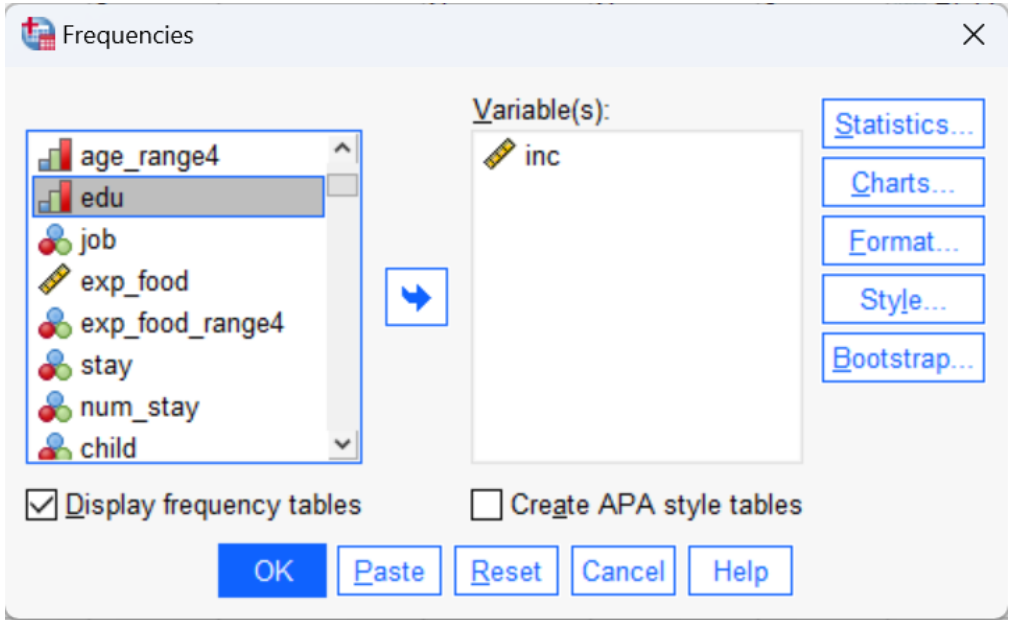
83	avg_PBC	Numeric	8	2		None	None	10	Right	Scale	Input
84	avg_INT	Numeric	8	2		None	None	10	Right	Scale	Input
85	avg_AVAI	Numeric	8	2		None	None	10	Right	Scale	Input
86	avg_HC	Numeric	8	2		None	None	10	Right	Scale	Input
87	Zage	Numeric	11	5	Zscore(age)	None	None	13	Right	Scale	Input
88	Zinc	Numeric	11	5	Zscore(inc)	None	None	13	Right	Scale	Input
89	Zexp_food	Numeric	11	5	Zscore(exp_fo...	None	None	13	Right	Scale	Input

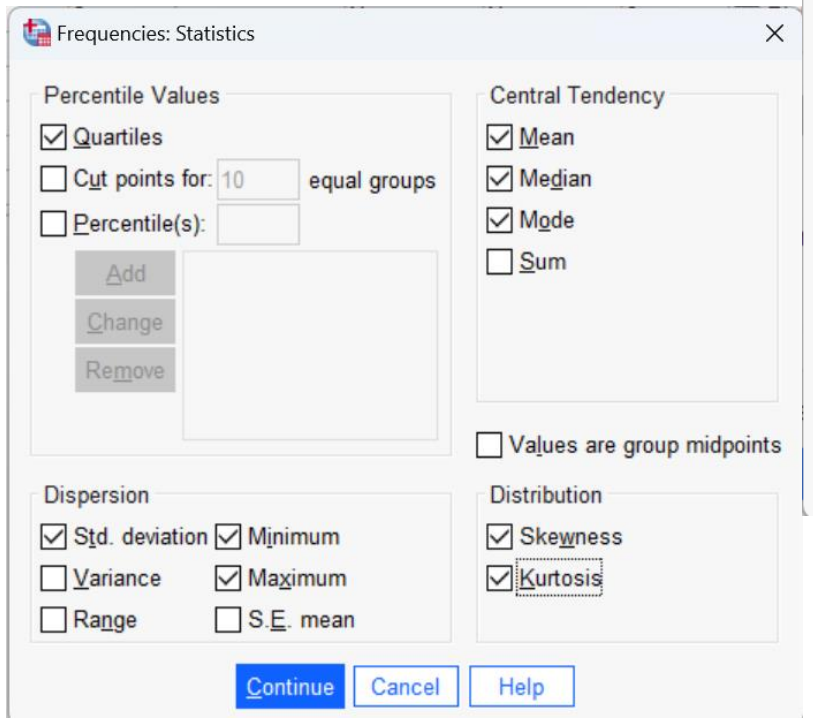
ค่า Z-score ของตัวแปรที่ต้องการจะเพิ่มใน variable lists

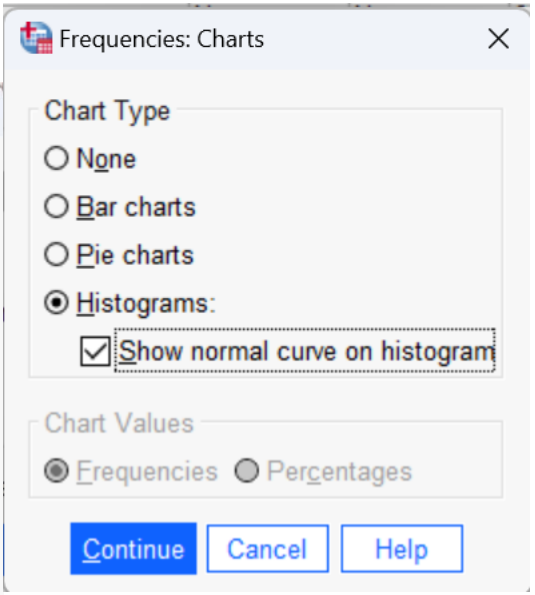
Frequencies with Scale Variables

click Analyze > Descriptive Statistics > Frequencies.







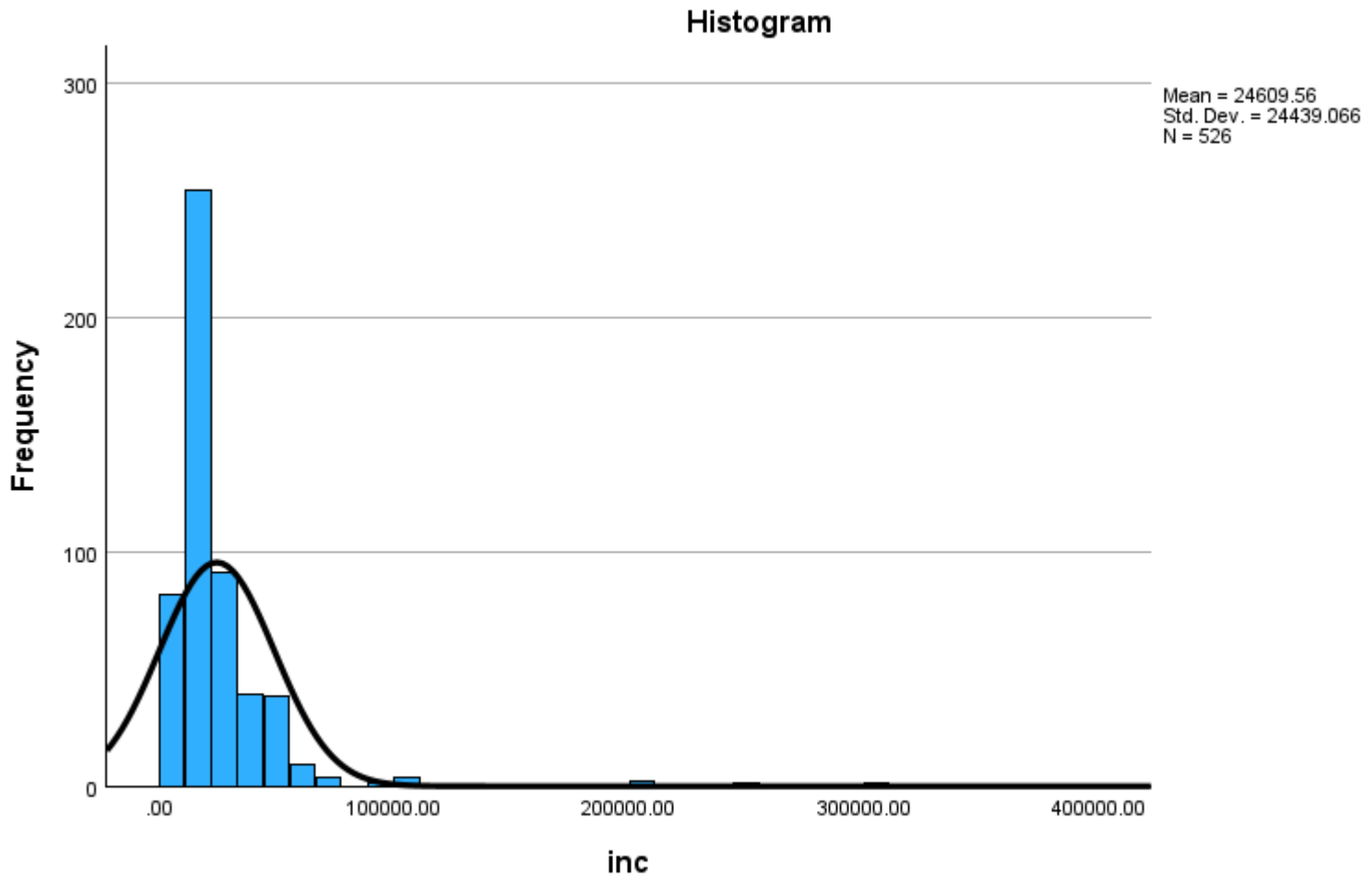


Frequencies with Scale Variables

Statistics

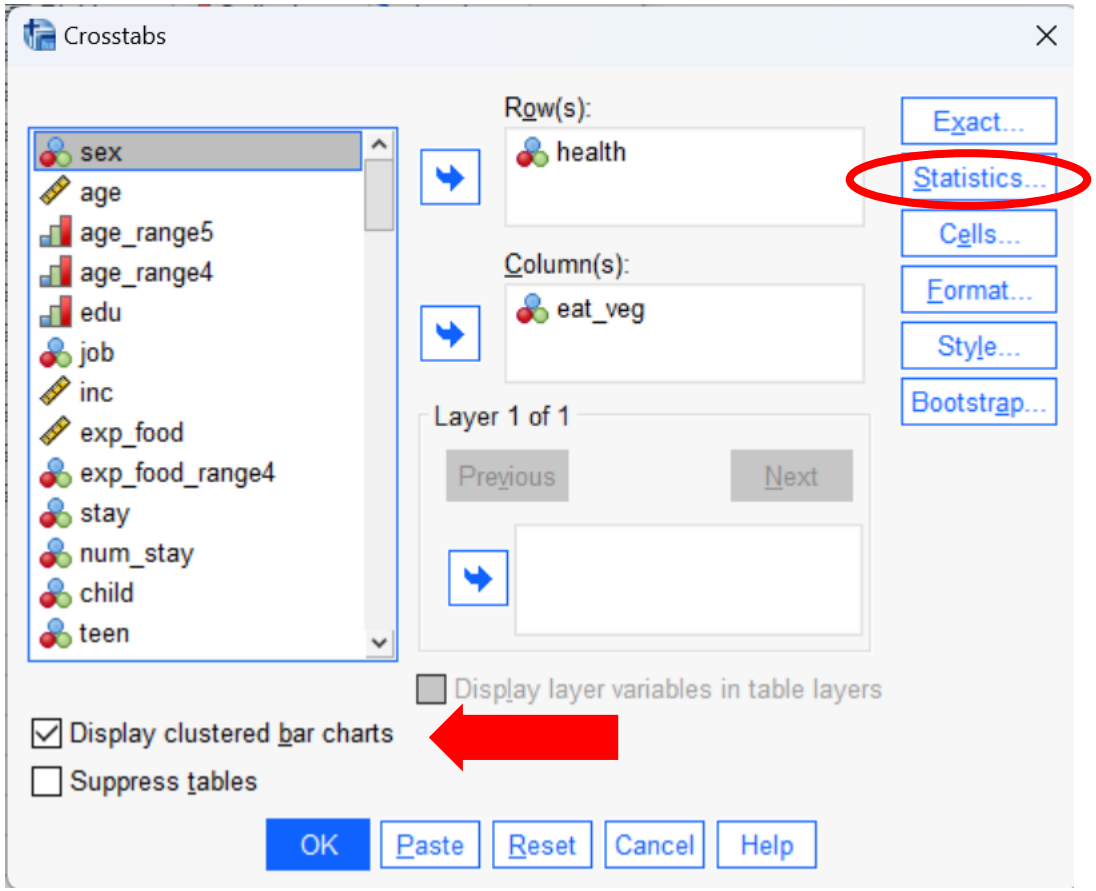
inc

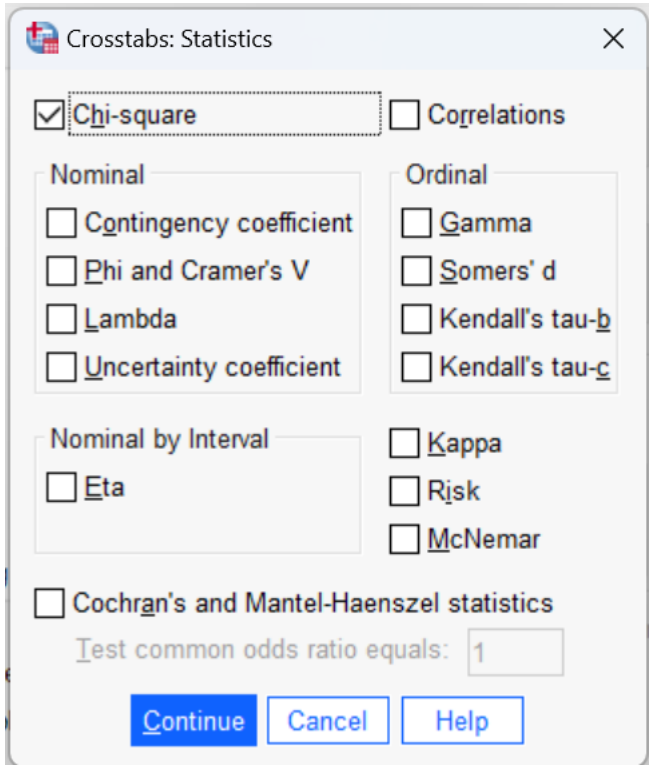
N	Valid	526
	Missing	0
Mean		24609.5627
Median		19000.0000
Mode		15000.00
Std. Deviation		24439.06623
Skewness		6.014
Std. Error of Skewness		.106
Kurtosis		53.336
Std. Error of Kurtosis		.213
Minimum		600.00
Maximum		300000.00
Percentiles	25	15000.0000
	50	19000.0000
	75	30000.0000



Crosstabs: Two categorical variables.

- To create a crosstab, click Analyze > Descriptive Statistics > Crosstabs.





Crosstabs: output

health * eat_veg Crosstabulation

Count

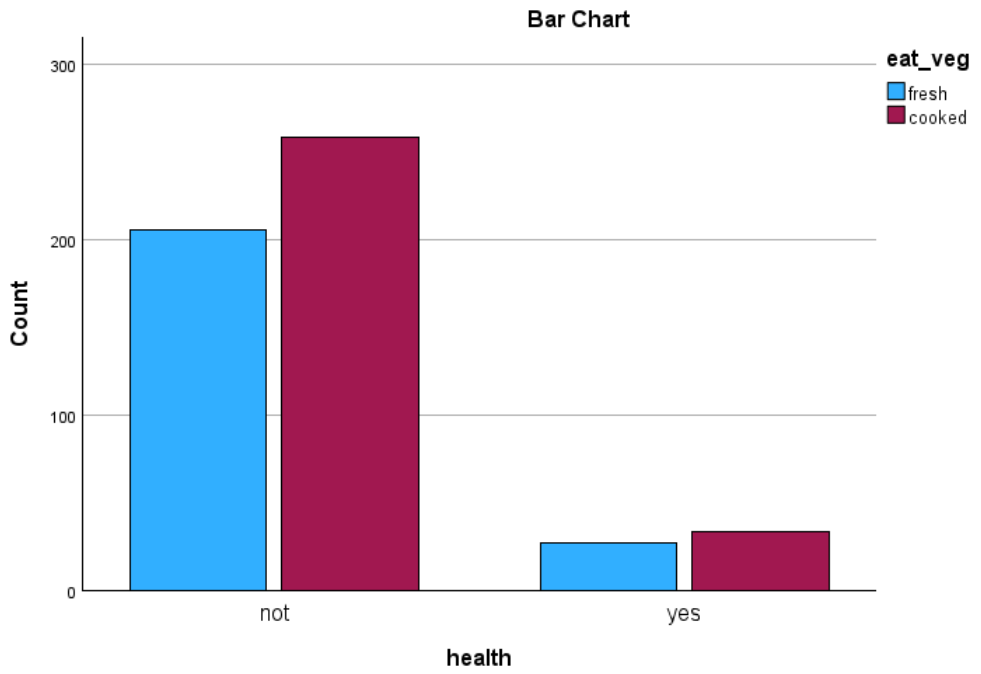
		eat_veg		
		fresh	cooked	Total
health	not	206	259	465
	yes	27	34	61
Total		233	293	526

Chi-Square Tests

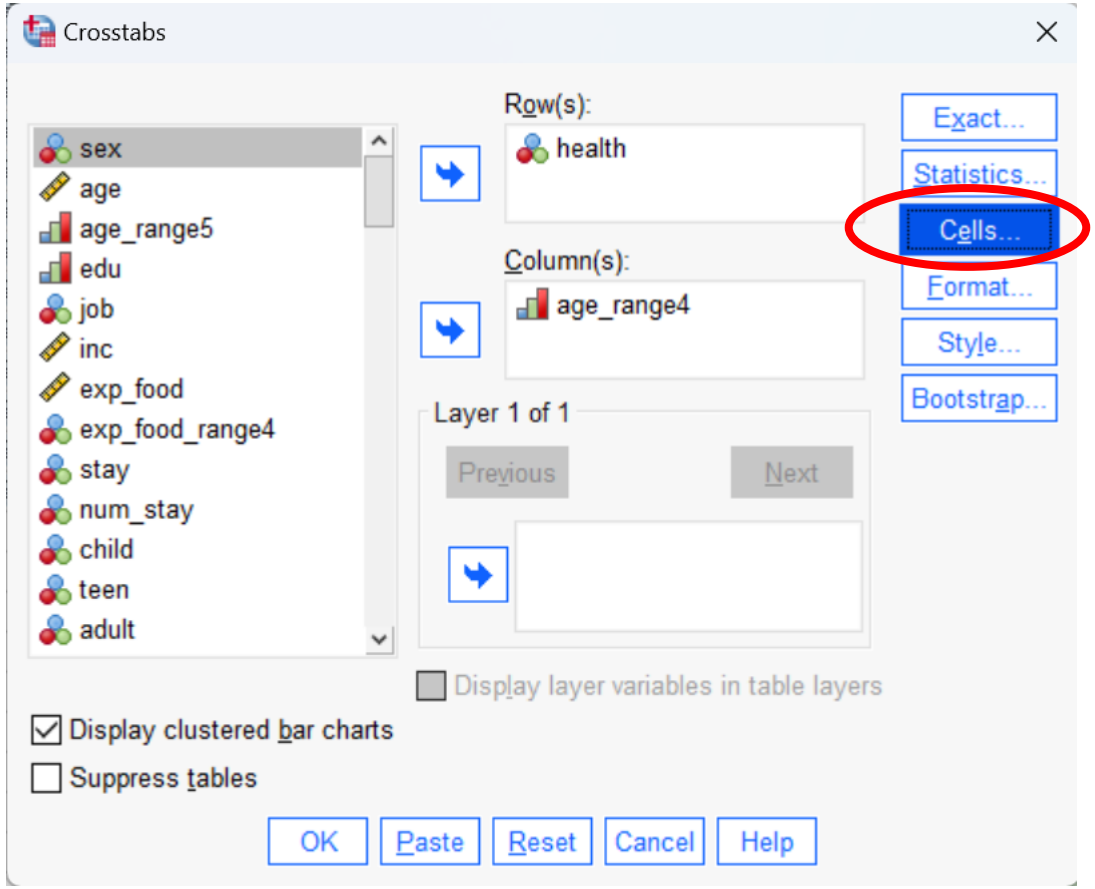
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.000[a]	1	.995		
Continuity Correction[b]	.000	1	1.000		
Likelihood Ratio	.000	1	.995		
Fisher's Exact Test				1.000	.554
Linear-by-Linear Association	.000	1	.995		
N of Valid Cases	526				

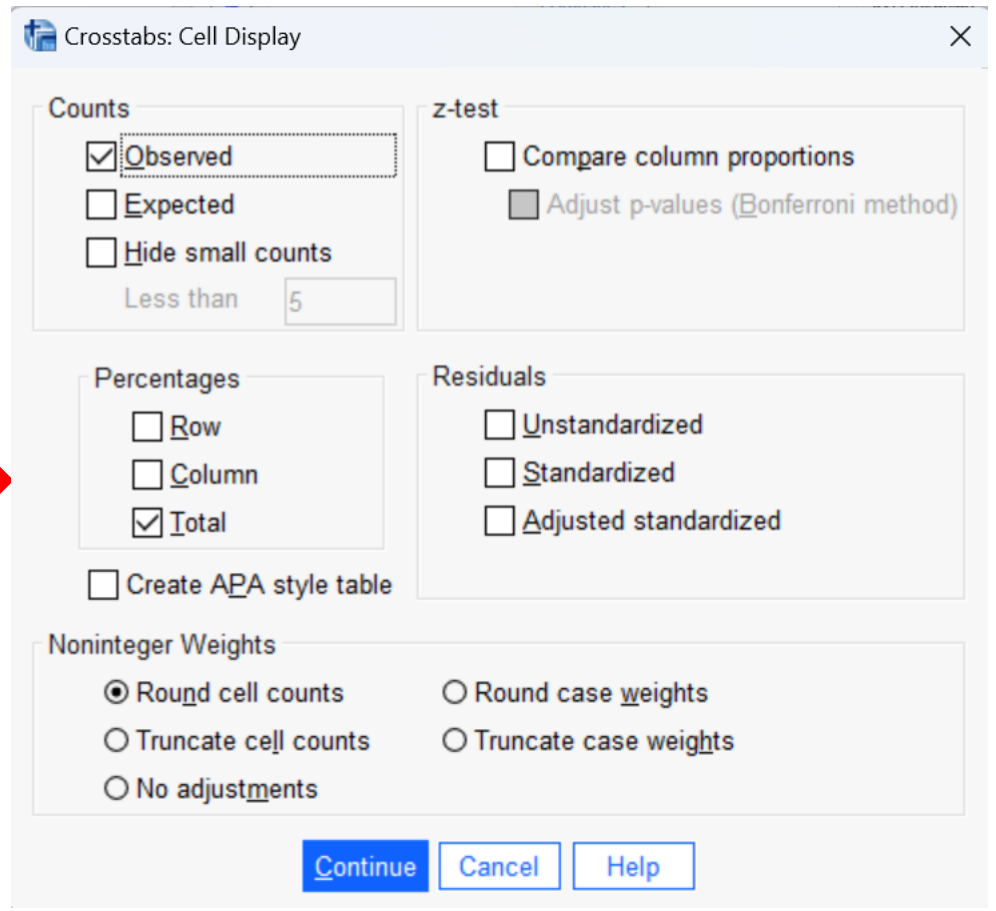
a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 27.02.

b. Computed only for a 2x2 table



Crosstabs: Two or more categories (groups) for each variable.





Crosstabs: Two or more categories (groups) for each variable.

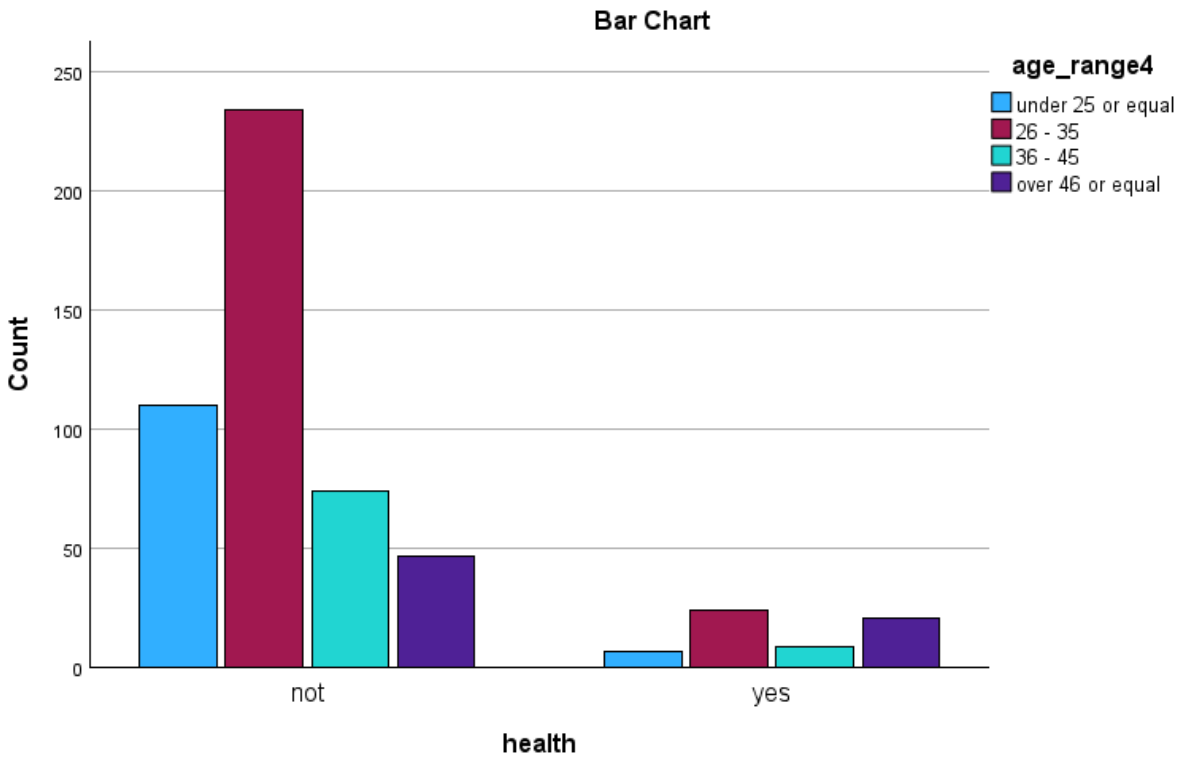
health * age_range4 Crosstabulation

			age_range4				
			under 25 or equal	26 - 35	36 - 45	over 46 or equal	Total
health	not	Count	110	234	74	47	465
		% of Total	20.9%	44.5%	14.1%	8.9%	88.4%
	yes	Count	7	24	9	21	61
		% of Total	1.3%	4.6%	1.7%	4.0%	11.6%
Total		Count	117	258	83	68	526
		% of Total	22.2%	49.0%	15.8%	12.9%	100.0%

Chi-Square Tests

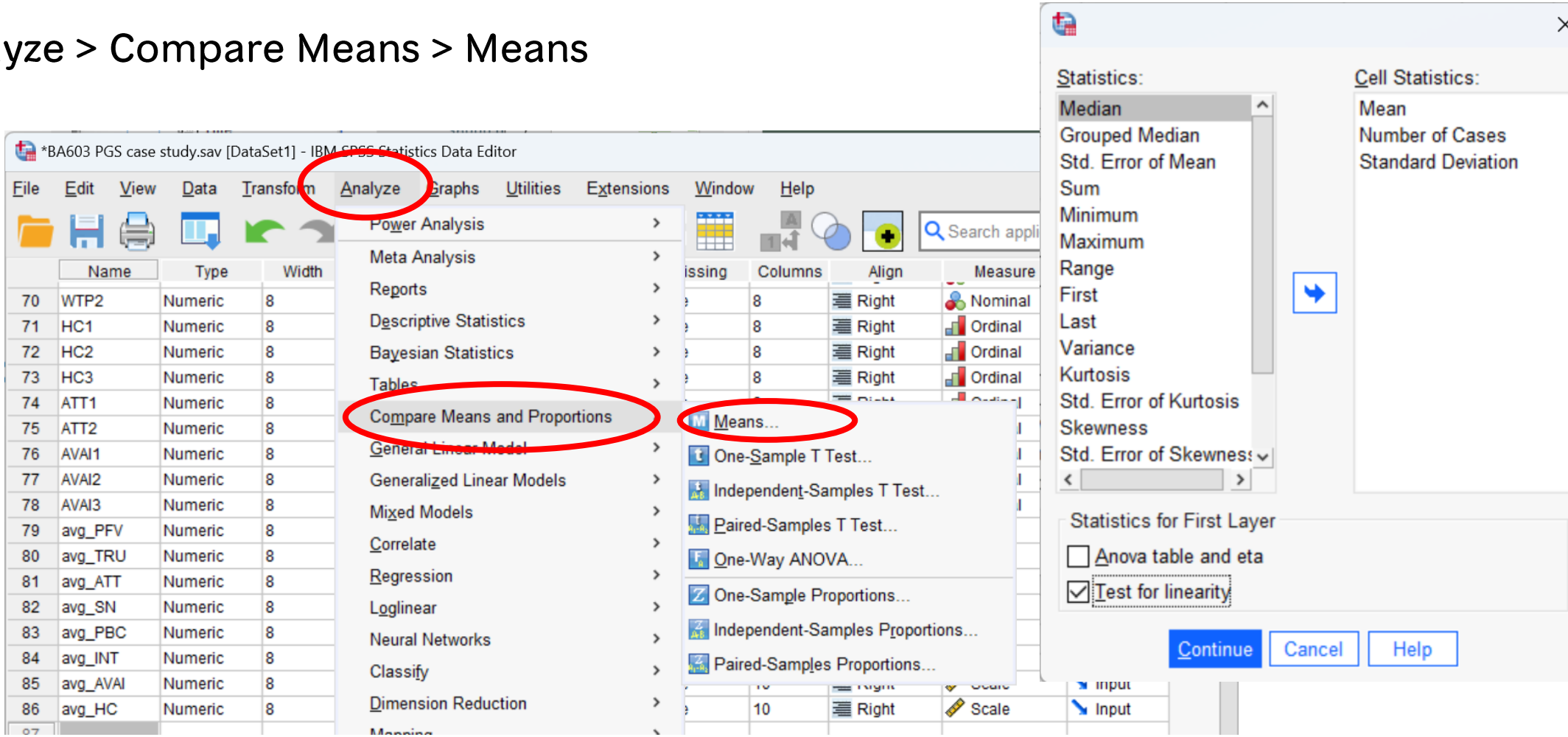
	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	29.637[a]	3	<.001
Likelihood Ratio	23.740	3	<.001
Linear-by-Linear Association	22.281	1	<.001
N of Valid Cases	526		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 7.89.



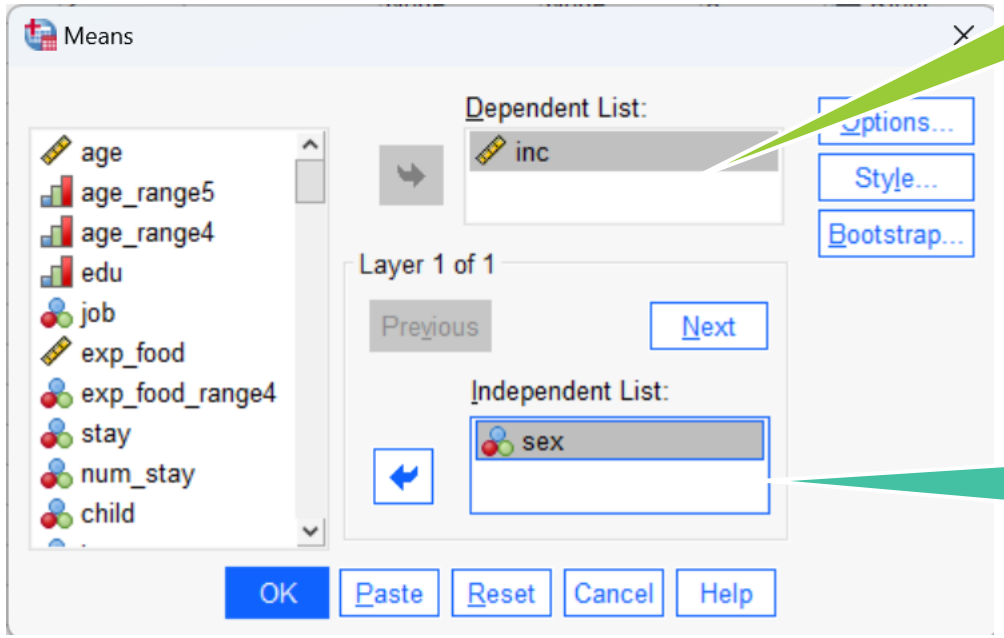
Compare Mean: Comparing averages across related demographic variables

Click: Analyze > Compare Means > Means

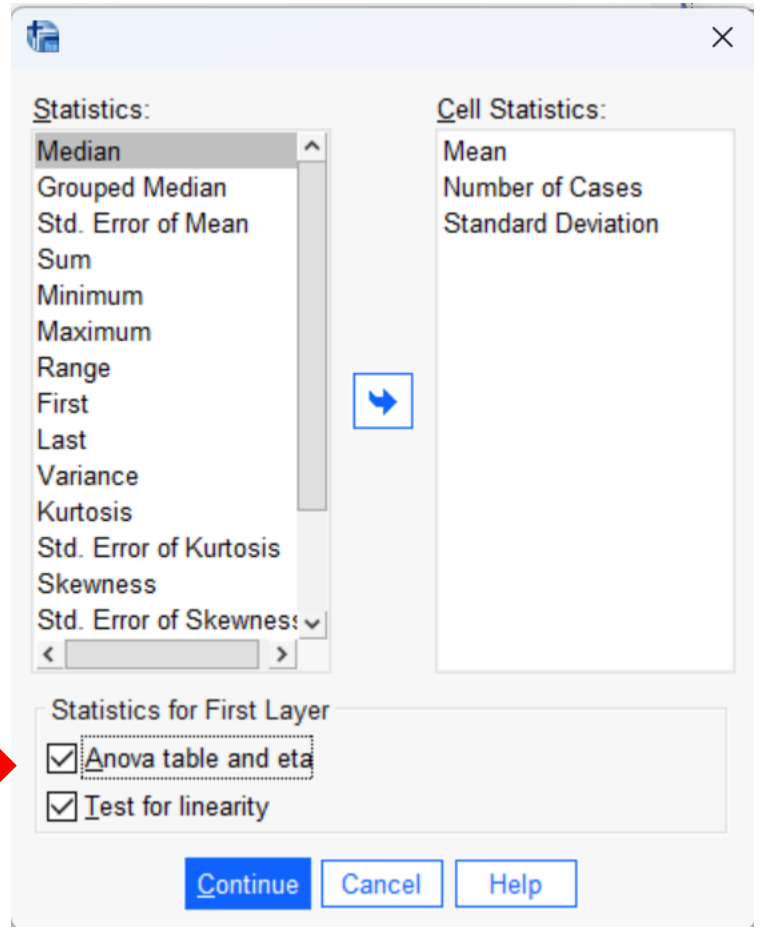


Compare Mean:

ตัวแปรตามที่ต้องการเปรียบเทียบ



ตัวแปรกลุ่มที่ต้องการเปรียบเทียบ



Compare Mean: Output

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
inc * sex	526	100.0%	0	0.0%	526	100.0%

Report

inc

sex	Mean	N	Std. Deviation
male	28844.5665	173	33170.82010
female	22534.0510	353	18439.86499
Total	24609.5627	526	24439.06623

ANOVA Table[a]

			Sum of Squares	df	Mean Square	F	Sig.
inc * sex	Between Groups	(Combined)	4623434799.9	1	4623434799.9	7.842	.005
	Within Groups		3.089E+11	524	589584433.59		
	Total		3.136E+11	525			

a. With fewer than three groups, linearity measures for inc * sex cannot be computed.

Measures of Association

	Eta	Eta Squared
inc * sex	.121	.015

- "inc * sex" หมายถึงความสัมพันธ์ระหว่างรายได้ (inc) กับเพศ (sex)
- ค่า Eta = .121 แสดงถึง ความสัมพันธ์ในระดับที่ต่ำถึงปานกลางระหว่างรายได้และเพศ
- Eta Squared = .015 หมายถึง เพศสามารถอธิบายความผันแปรของรายได้ได้ 1.5%

Compare Means: Test for linearity

- ค่าสัมประสิทธิ์อีต้า (Eta: η) เป็นค่าสถิติที่ใช้วัดความสัมพันธ์ระหว่างตัวแปรสองตัว เมื่อตัวแปรหนึ่งเป็นตัวแปรเชิงกลุ่ม (categorical variable) และอีกตัวแปรหนึ่งเป็นตัวแปรเชิงปริมาณ (continuous variable)
- ค่า Eta จะอยู่ในช่วง 0 ถึง 1
  - 0 หมายถึงไม่มีความสัมพันธ์
  - 1 หมายถึงมีความสัมพันธ์สมบูรณ์แบบ
  - Eta จะคล้ายกับค่าสัมประสิทธิ์สหสัมพันธ์ (correlation coefficient) แต่ใช้ในกรณีที่ตัวแปรหนึ่งเป็นเชิงกลุ่ม

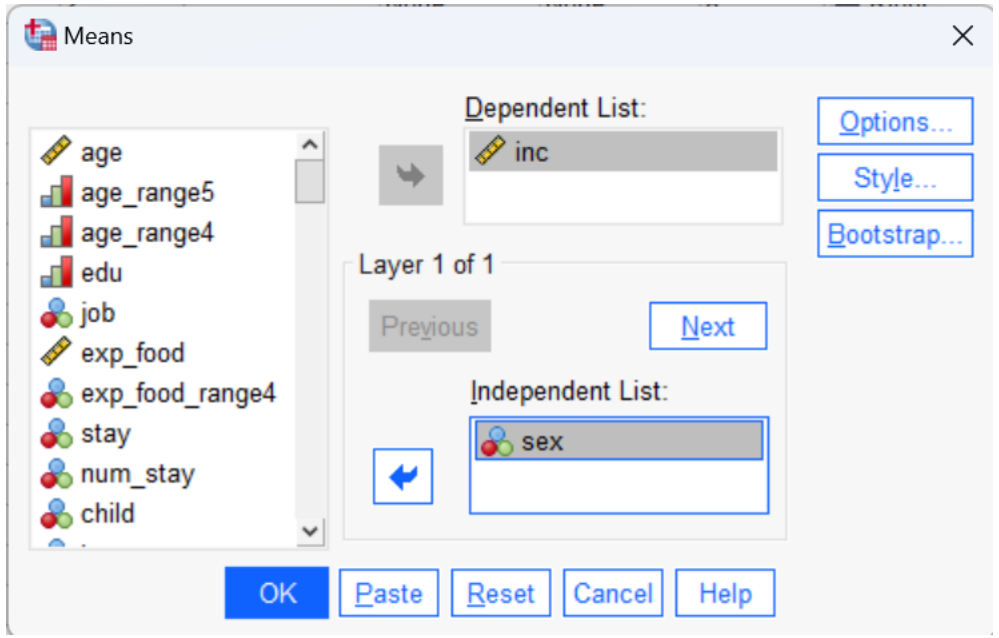
Measures of Association

	Eta	Eta Squared
inc * sex	.121	.015

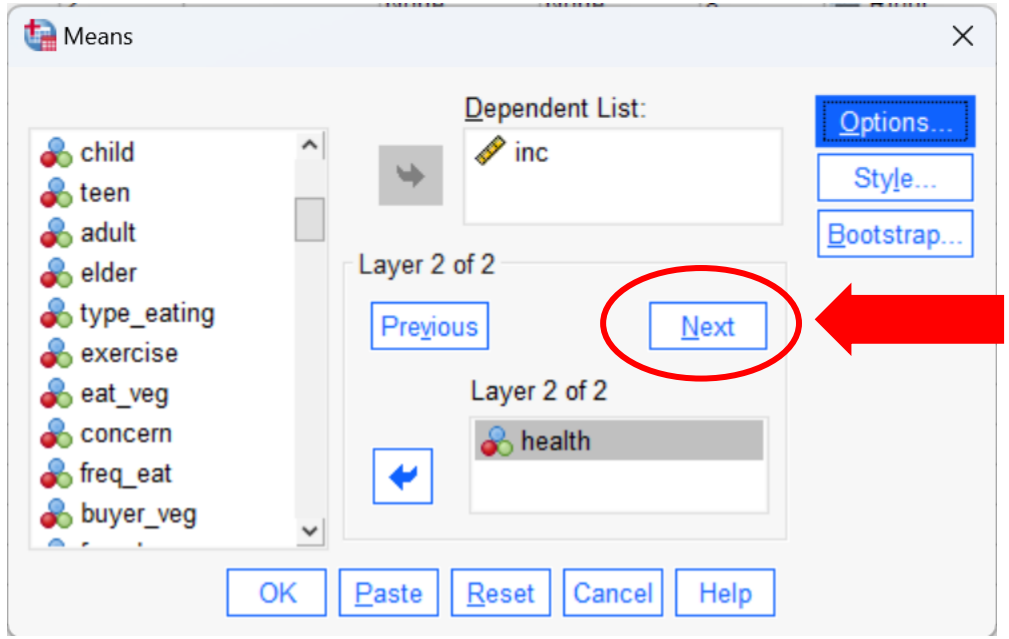
- "inc * sex" หมายถึงความสัมพันธ์ระหว่างรายได้ (inc) กับเพศ (sex)
- ค่า Eta = .121 แสดงถึง ความสัมพันธ์ในระดับที่ต่ำถึงปานกลางระหว่างรายได้และเพศ
- Eta Squared = .015 หมายถึง เพศสามารถอธิบายความผันแปรของรายได้ได้ 1.5%

Compare Means for 2 layers

Layer 1: sex



Layer 2: health



Compare Means for 2 layers: output

Report

inc

sex	health	Mean	N	Std. Deviation
male	not	28913.3333	156	34482.19753
	yes	28213.5294	17	17582.41271
	Total	28844.5665	173	33170.82010
female	not	22012.3301	309	18376.12966
	yes	26197.9545	44	18682.91292
	Total	22534.0510	353	18439.86499
Total	not	24327.5054	465	25139.32745
	yes	26759.6721	61	18259.82794
	Total	24609.5627	526	24439.06623

ANOVA Table[a]

			Sum of Squares	df	Mean Square	F	Sig.
inc * sex	Between Groups	(Combined)	4623434799.9	1	4623434799.9	7.842	.005
	Within Groups		3.089E+11	524	589584433.59		
	Total		3.136E+11	525			

a. With fewer than three groups, linearity measures for inc * sex cannot be computed.

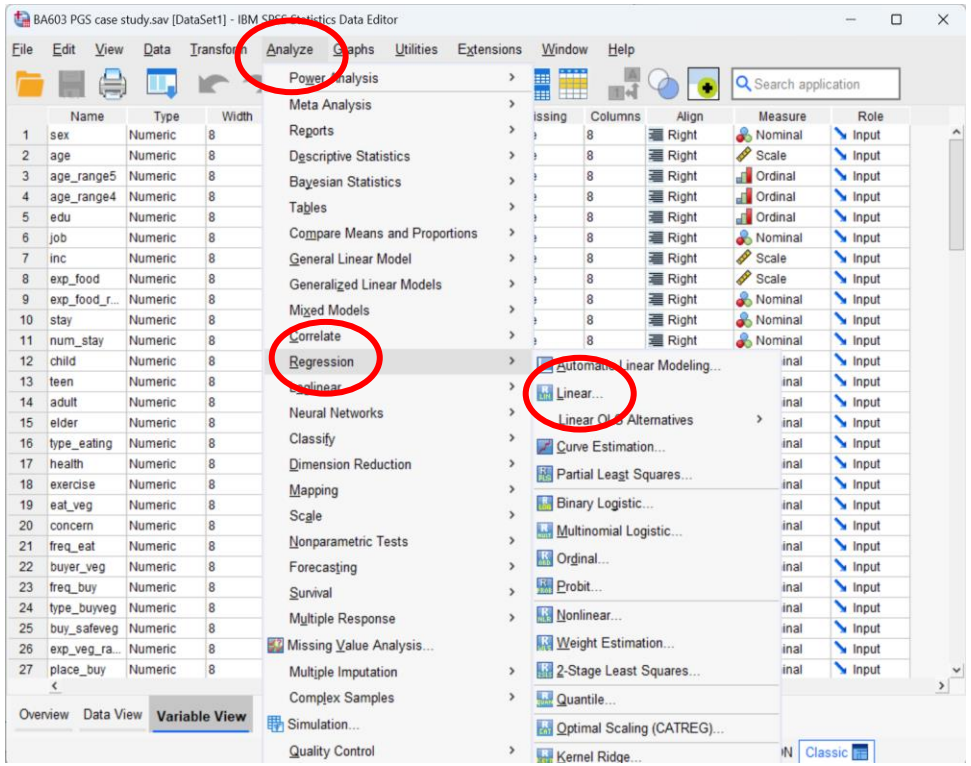
Measures of Association

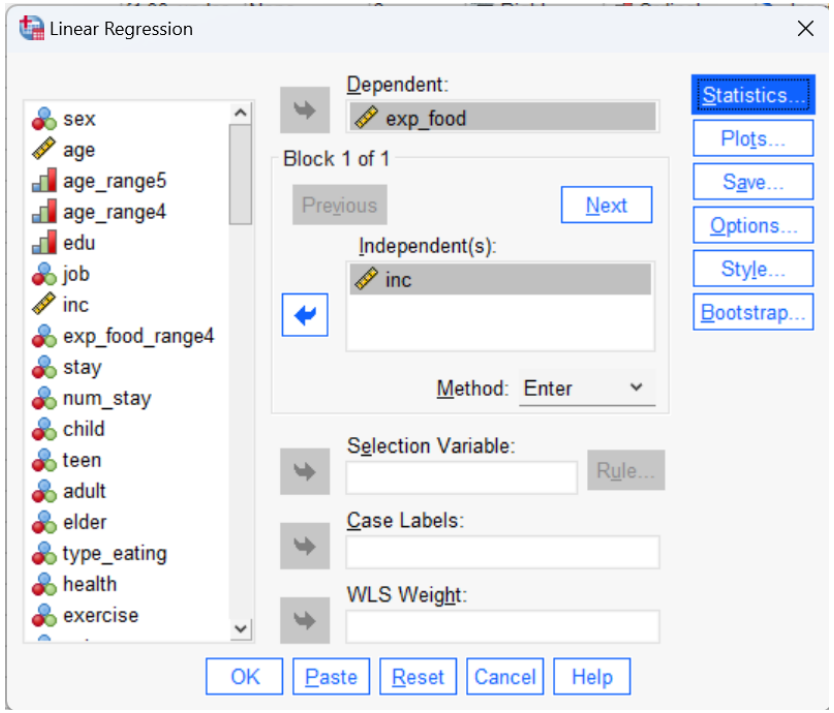
	Eta	Eta Squared
inc * sex	.121	.015

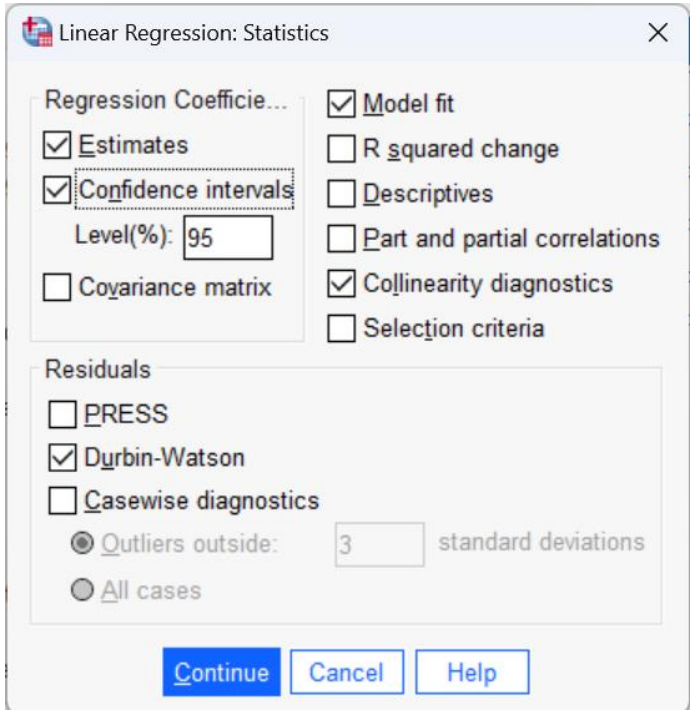
ผลส่วนนี้จะเหมือนกับ Layer 1

Simple Linear regression

- Click Analyze > Regression > Linear… on the top menu, as shown below:







Simple Linear regression

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.514[a]	.264	.263	4757.60882	1.985

a. Predictors: (Constant), inc

b. Dependent Variable: exp_food

Income สามารถอธิบายการเปลี่ยนแปลงของ exp_food ได้ 26.4%

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4265163940.2	1	4265163940.2	188.434	<.001[b]
	Residual	11860657020	524	22634841.642		
	Total	16125820961	525			

a. Dependent Variable: exp_food

b. Predictors: (Constant), inc

The F-ratio in the ANOVA table (see below) tests whether the **overall regression model is a good fit for the data.**

โดยรวมแล้ว โมเดลการถดถอยทำนายตัวแปรผลลัพธ์ได้อย่างมีนัยสำคัญทางสถิติ [$F(1,524)=188.34$, $p < .001$]

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	4007.911	294.533		13.608	<.001	3429.300	4586.521		
	inc	.117	.008	.514	13.727	<.001	.100	.133	1.000	1.000

a. Dependent Variable: exp_food

คอลัมน์ "Sig." ระบุ ค่าสัมประสิทธิ์ตัวแปรอิสระมีความแตกต่างอย่างมีนัยสำคัญทางสถิติจาก 0 (ศูนย์) ถึงแม้ว่าค่าตัดขวาง (intercept) B0 จะได้รับการทดสอบความสำคัญทางสถิติแล้วก็ตาม แต่ผลลัพธ์นี้ไม่สำคัญ

To present the regression equation as:
Exp_food = 4,007 +0.117(inc)

รายได้เพิ่มขึ้น 1 บาท ส่งผลให้ค่าใช้จ่ายอาหารเพิ่มขึ้น .117 บาท

Simple Linear regression

คอลัมน์ "Sig." ระบุ ค่าสัมประสิทธิ์ตัวแปรอิสระมีความแตกต่างอย่างมีนัยสำคัญทางสถิติจาก 0 (ศูนย์) ถึงแม้ว่าค่าตัดขวาง (intercept) B0 จะได้รับการทดสอบความสำคัญทางสถิติแล้วก็ตาม แต่ผลลัพธ์นี้ไม่สำคัญ

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	4007.911	294.533		13.608	<.001	3429.300	4586.521		
	inc	.117	.008	.514	13.727	<.001	.100	.133	1.000	1.000

a. Dependent Variable: exp_food

To present the regression equation as:
Exp_food = 4,007 +0.117(inc)

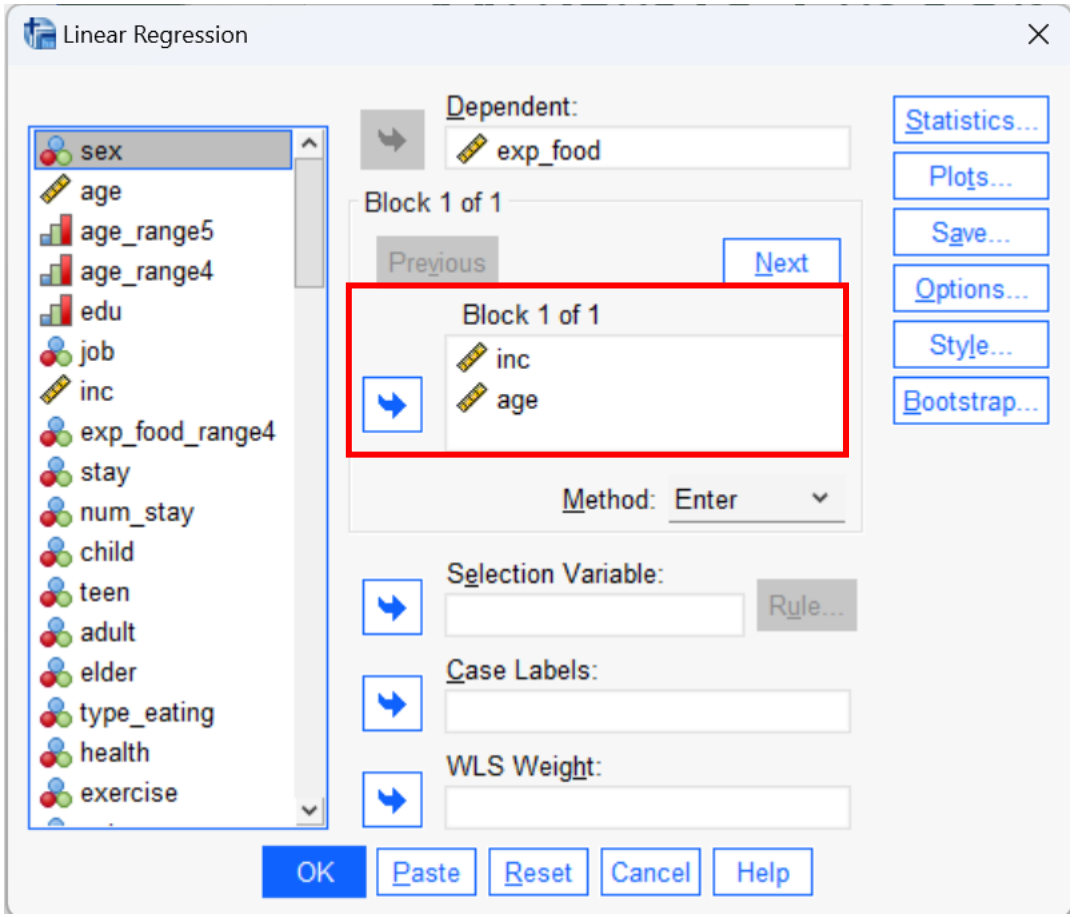
รายได้เพิ่มขึ้น 1 บาท ส่งผลให้ค่าใช้จ่ายอาหารเพิ่มขึ้น .117 บาท

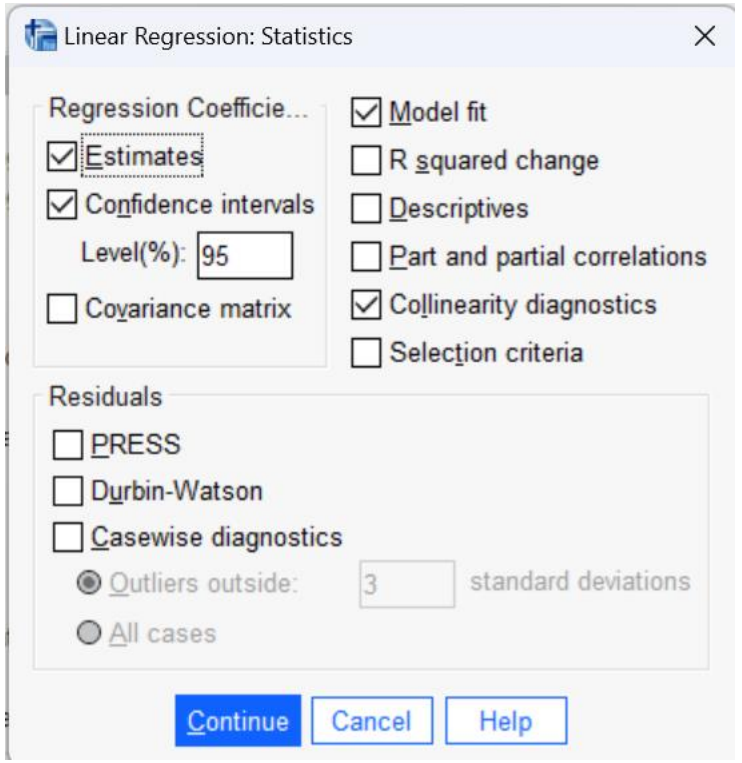
Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	Variance Proportions inc
1	1	1.710	1.000	.15	.15
	2	.290	2.428	.85	.85

a. Dependent Variable: exp_food

Multiple Linear Regression





Multiple Linear Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	age, inc[b]	.	Enter

a. Dependent Variable: exp_food

b. All requested variables entered.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4525536375.9	2	2262768187.9	102.017	<.001[b]
	Residual	11600284585	523	22180276.453		
	Total	16125820961	525			

a. Dependent Variable: exp_food

b. Predictors: (Constant), age, inc

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.530[a]	.281	.278	4709.59409

a. Predictors: (Constant), age, inc

R^2 และ Adjusted R^2 เพิ่มขึ้น เมื่อเพิ่มตัวแปรอิสระ “age” แสดงว่า ตัวแปรที่เพิ่มขึ้น ส่งผลให้แบบจำลองดีขึ้น

Multiple Linear Regression

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	1859.352	691.560		2.689	.007	500.776	3217.929		
	inc	.108	.009	.476	12.299	<.001	.091	.125	.918	1.090
	age	72.346	21.115	.133	3.426	<.001	30.864	113.827	.918	1.090

a. Dependent Variable: exp_food

ค่า “VIF” ควร < 10

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				(Constant)	inc	age
1	1	2.607	1.000	.01	.05	.01
	2	.349	2.734	.04	.93	.03
	3	.044	7.683	.94	.02	.96

a. Dependent Variable: exp_food

- **Eigenvalue**
  - ค่า Eigenvalue ใกล้ 0 ชี้ว่ามีปัญหา Multicollinearity
- **Condition index**
  - ค่ามากกว่า 15 แสดงว่า มีปัญหา Multicollinearity
  - ค่ามากกว่า 30 แสดงว่า มีปัญหา Multicollinearity รุนแรง ต้องแก้ไข โดยพิจารณาจาก Variance Proportions
- **Variance Proportions**
  - ค่าVariance Proportions ในแถวเดียวกันมีค่า > .90 อย่างน้อย 2 ตัวแปร แสดงว่าตัวแปรเหล่านั้นมีปัญหา Multicollinearity กัน ซึ่งต้องแก้ไข (Hair et al, 2014)
  - ค่า Variance Proportions > .90 เพียงค่าเดียวในแถว แสดงว่าไม่มีปัญหา Multicollinearity

ตัวอย่าง Collinearity diagnostics

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-10,753	5,211		-2,064	,046		
	x1	-,306	,790	-,165	-,387	,701	,018	54,560
	x2	,538	,803	,283	,670	,507	,019	53,972
	x3	,265	,970	,131	,273	,786	,014	69,555
	x4	-,228	,950	-,117	-,240	,811	,014	71,438
	x5	,578	,106	,326	5,445	,000	,924	1,082
	x6	2,586	,180	,851	14,374	,000	,944	1,059

a. Dependent Variable: y

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions						
				(Constant)	x1	x2	x3	x4	x5	x6
1	1	6,257	1,000	,00	,00	,00	,00	,00	,00	,00
	2	,265	4,857	,00	,00	,00	,00	,00	,24	,21
	3	,232	5,196	,00	,00	,00	,00	,00	,39	,39
	4	,187	5,791	,00	,00	,00	,00	,01	,00	,33
	5	,058	10,384	,03	,01	,00	,00	,01	,30	,04
	6	,001	80,295	,13	,96	,95	,04	,05	,04	,00
	7	,001	109,115	,85	,02	,05	,96	,94	,03	,02

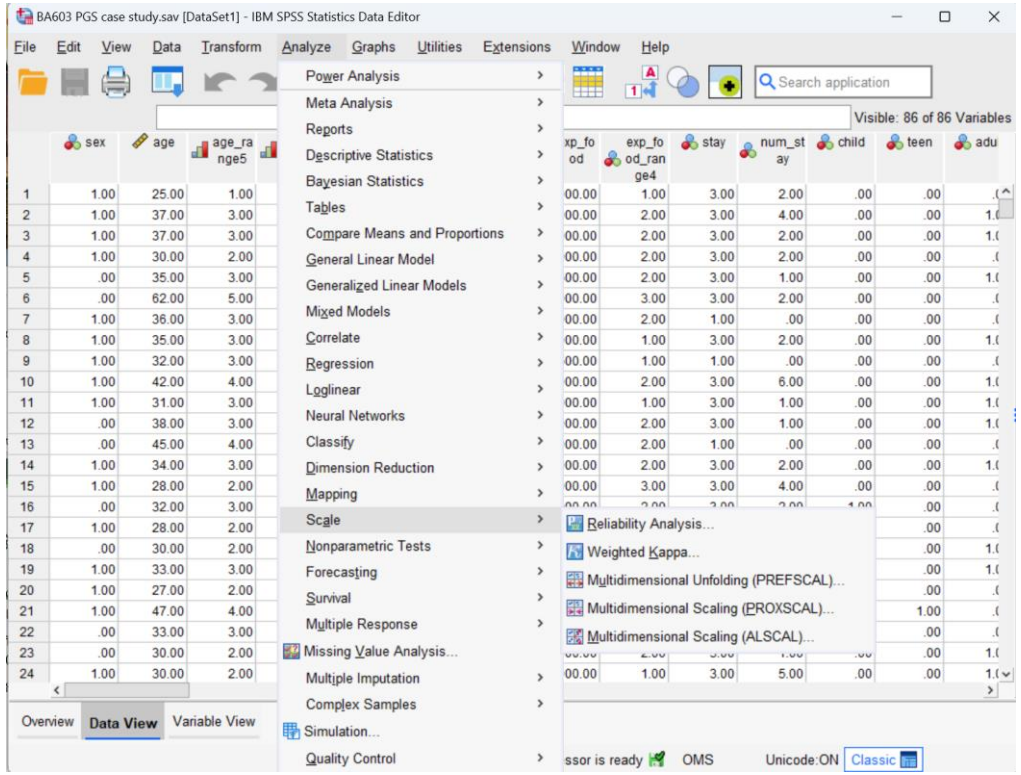
a. Dependent Variable: y

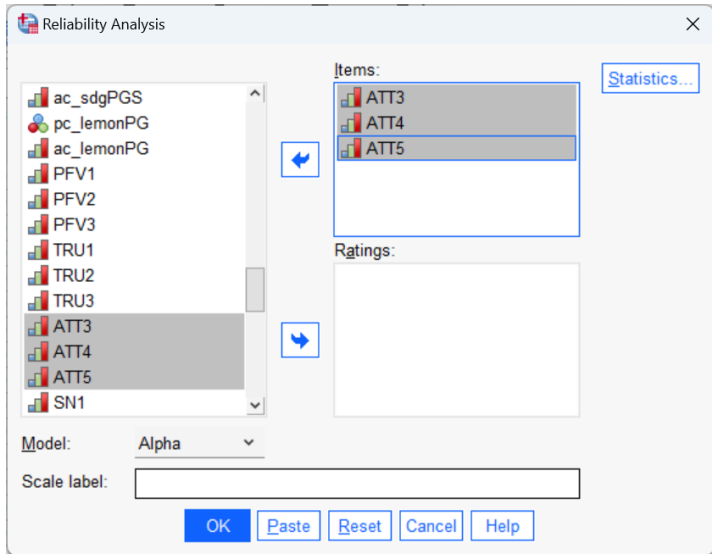
1. ตัวแปรอิสระ (predictors) ที่มีค่า VIF สูงกว่า 10 (x1, x2, x3, x4)
2. predictors มากกว่าสองตัว (ในที่นี้คือสี่ตัว) ที่สามารถนำไปใช้กับสิ่งนี้ได้
3. Dimension 6 และ 7 มีค่า index > 15
4. ให้ค้นหาค่าที่สูงกว่า 0.90 สำหรับ Dimension 6 เราจะพบค่าเหล่านี้สำหรับตัวทำนาย x1 และ x2 สำหรับ Dimension 7 เราจะพบค่าเหล่านี้สำหรับตัวทำนาย x3 และ x4 ถือว่ามีปัญหาความสัมพันธ์เชิงเส้นตรงสองแบบที่แตกต่างกันในแบบจำลอง: ระหว่าง x1 และ x2 และระหว่าง x3 และ x4
   - อย่างไรก็ตาม หากค่าทั้งหมดที่สูงกว่า 0.90 สำหรับตัวทำนายทั้งสี่นี้อยู่ในบรรทัดเดียวกัน นั่นจะบ่งชี้ว่ามีปัญหาความสัมพันธ์เชิงเส้นตรงหลายแบบของตัวแปรทั้งสี่ตัว

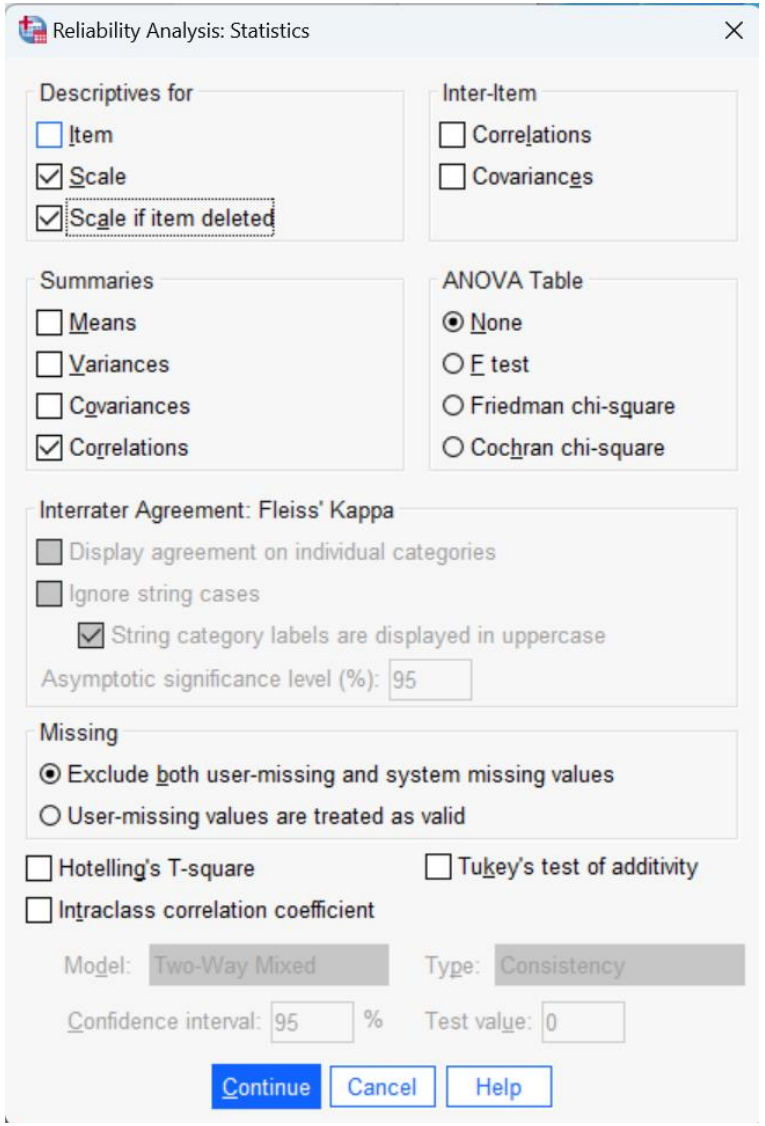
https://www.regorz-statistik.de/en/collinearity_diagnostics_table_SPSS.html

Reliability test (Cronbach's alpha coefficient)

1. Click: Analyze > Scale > Reliability Analysis...
2. **เลือกตัวแปรที่ต้องการทดสอบ**
3. Click: Statistics
4. **เลือกค่าสถิติที่ต้องการ**
5. Click: Continue







Reliability test (Cronbach's alpha coefficient)

Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	526	100.0
	Excluded[a]	0	.0
	Total	526	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	N of Items
.914	3

Cronbach's alpha coefficient

ช่วยพิจารณาการตัดข้อคำถาม

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
ATT3	8.2319	2.544	.825	.879
ATT4	8.2034	2.471	.876	.836
ATT5	8.0932	2.645	.783	.913

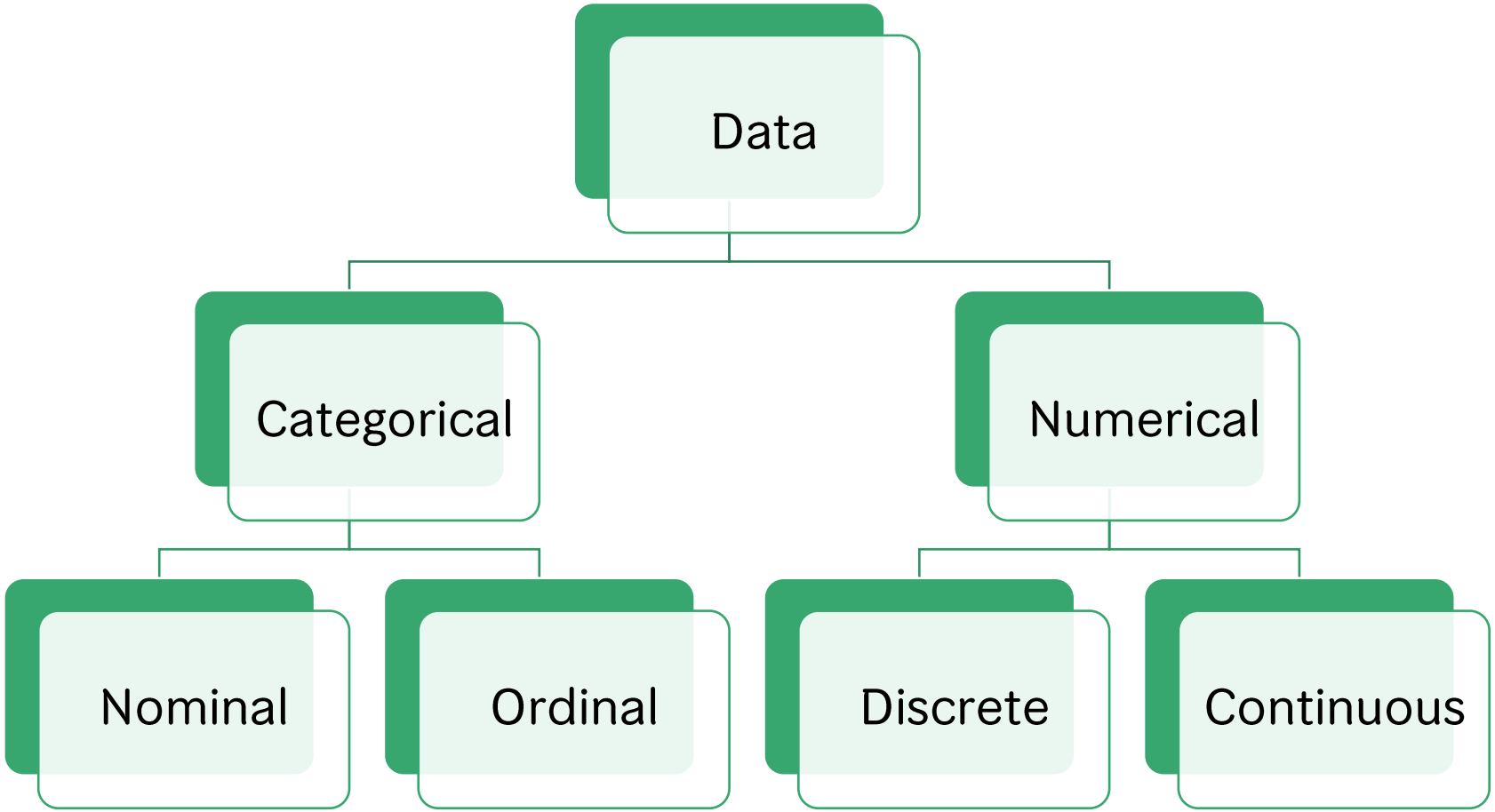
Scale Statistics

Mean	Variance	Std. Deviation	N of Items
12.2643	5.509	2.34714	3

There are no universally agreed-upon "perfect" cutoffs, but general guidelines are often used:

Cronbach's alpha	Meaning
$\alpha \geq 0.9$	Excellent internal consistency
$.0.8 \leq \alpha < 0.9$	Good internal consistency
$.0.7 \leq \alpha < 0.8$	Acceptable internal consistency
$.0.6 \leq \alpha < 0.7$	Questionable internal consistency
$.0.5 \leq \alpha < 0.6$	Poor internal consistency.
$\alpha < 0.5$	Unacceptable internal consistency

Type of data


Data
Categorical
Numerical
Nominal
Ordinal
Discrete
Continuous

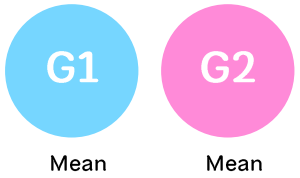
Types of Data Analysis

Descriptive

ความถี่ ร้อยละ
ค่าเฉลี่ย ส่วน
เบี่ยงเบนมาตรฐาน
ความแปรปรวน

Comparative

T-test



Independent t-test
Ho: mean G1 = mean G2
Ha: mean G1 != mean G2

One-way ANOVA

Post Hoc Comparisons
Ho: two groups mean equal
Ha: two groups mean not equal

One-way ANOVA
Ho: all group means equal
Ha: at least one group different

Associative

Correlation

Ho: no linear relationship bet Two variables
Ha: there is linear relationship bet two variables
* Correlation does not imply causation

Crosstabulation

Ho: no relationship bet Two variables
Ha: there is relationship bet two variables

Chi-square test

Predictive

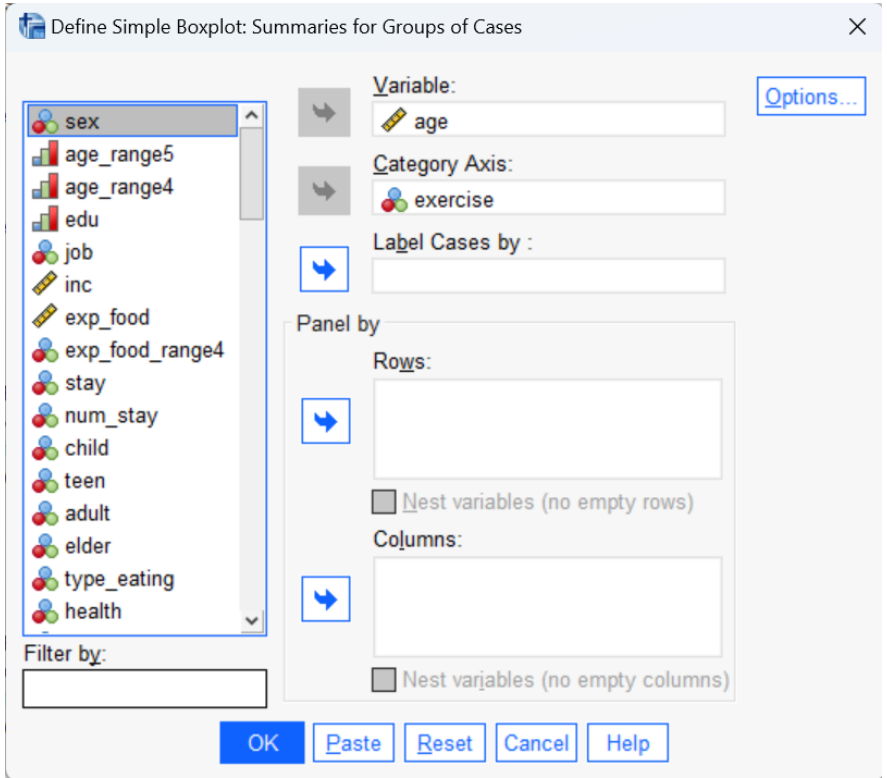
Regression

Ho: all coeff. = 0
Ha: at least one coif. != 0

Structural Equation modelling (SEM)

Boxplot

```
EXAMINE VARIABLES=age BY exercise
  /PLOT=BOXPLOT
  /STATISTICS=NONE
  /NOTOTAL.
```



เปรียบเทียบอายุระหว่างกลุ่มคนที่ออกกำลังกาย