Numerical data analysis

Dr. Wan Nor Arifin

Unit of Biostatistics and Research Methodology,
Universiti Sains Malaysia.
wnarifin@usm.my / wnarifin.github.io

Outlines

- Hypothesis Testing
- Parametric Test
- Two Independent Samples
- Two Related Samples
- More Than Two Independent Samples

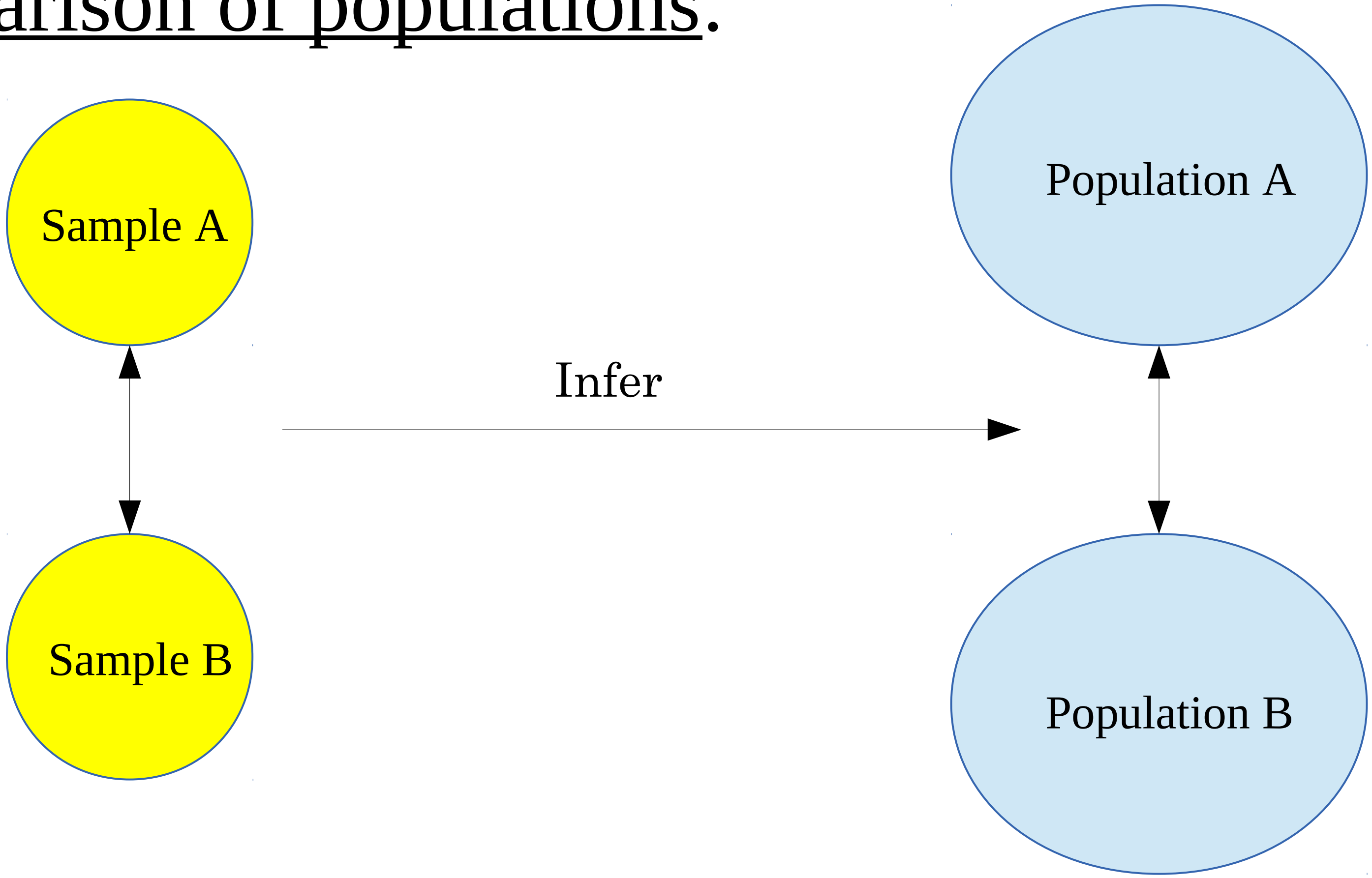
Learning outcomes

- Understand basic concept of hypothesis testing.
- Understand concept of P-value and significance level.
- Able to perform selected parametric tests for comparison of means between samples.

Hypothesis Testing

Hypothesis Testing

- In the context of comparison of samples → comparison of populations.



Hypothesis Testing

- Stated in form of **Statistical Hypothesis** → Can be tested with statistical test.

<u>Alternative Hypothesis</u>:
Population A is different from Population B

<u>Null Hypothesis</u>:
Population A is similar to Population B

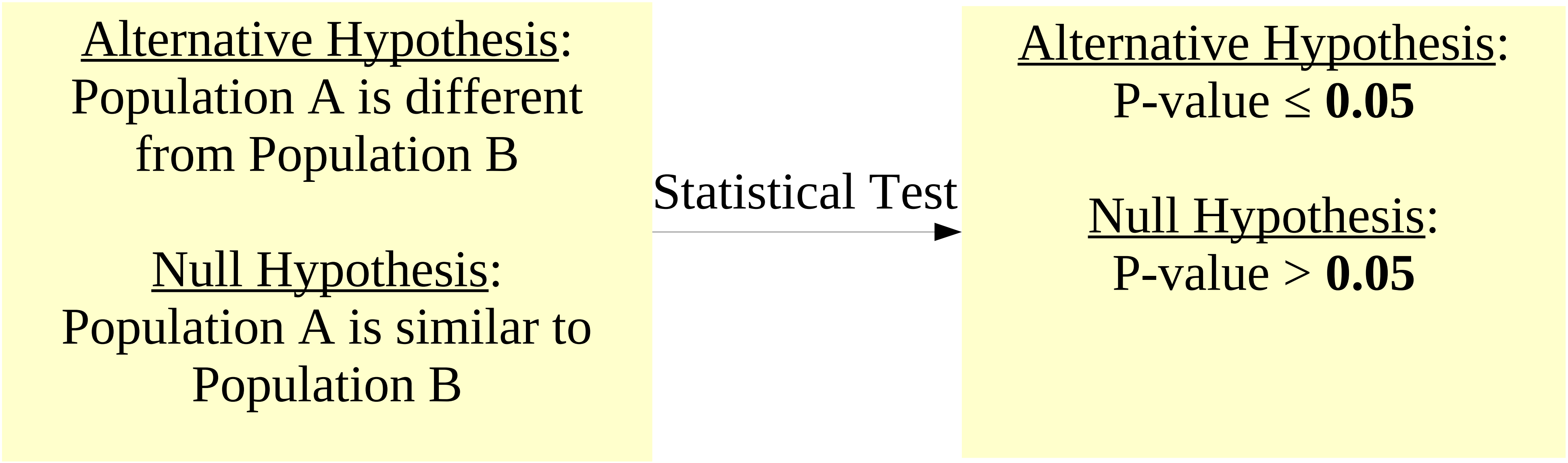
Hypothesis Testing

- **P-value** – Probability that the difference is merely by chance → Calculated from statistical test.
- Set acceptable level so called "chance" → **Significance level**, **α** (**<u>0.05</u>**, 0.01, 0.001)

<u>Alternative Hypothesis</u>:
P-value ≤ **0.05**

<u>Null Hypothesis</u>:
P-value > **0.05**

Hypothesis Testing



Alternative Hypothesis:
Population A is different from Population B

Null Hypothesis:
Population A is similar to Population B

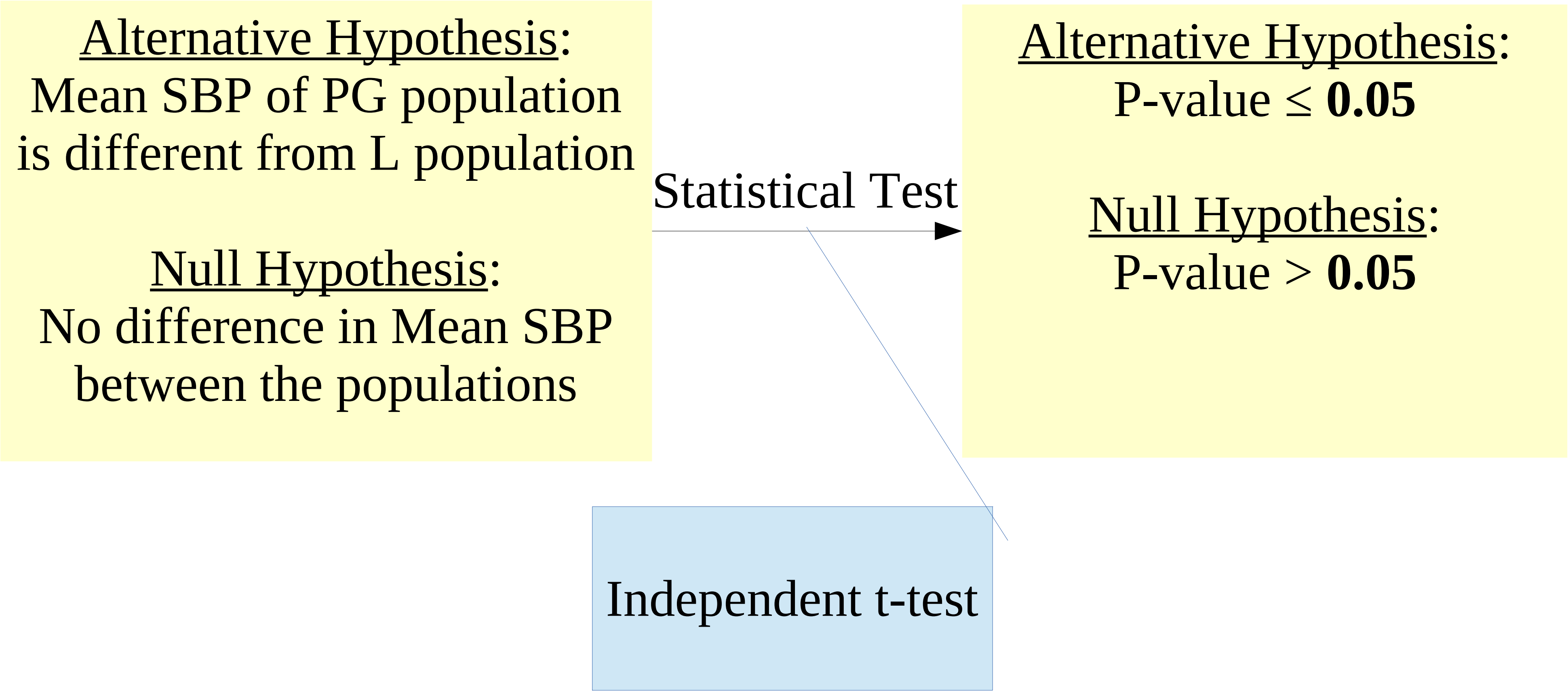
Statistical Test

Alternative Hypothesis:
P-value ≤ **0.05**

Null Hypothesis:
P-value > **0.05**

Hypothesis Testing

Comparing **mean SBP** of **postgraduate students' population** vs **lecturers' population**



Alternative Hypothesis:
Mean SBP of PG population is different from L population

Null Hypothesis:
No difference in Mean SBP between the populations

Statistical Test

Alternative Hypothesis:
P-value ≤ **0.05**

Null Hypothesis:
P-value > **0.05**

Independent t-test

Parametric Test

Parametric Test

- Statistical test that requires:
  - Sample data come from population data that can be modeled by specific statistical distribution.
  - e.g. SBP of sample ← Normally distributed SBP of population.
  - Fixed set of parameters for chosen distribution.
  - e.g. normal distribution ← mean, SD.

Parametric Test

- Statistical test that requires (cont.):
  - Specific parameters to be tested.
  - e.g. MEAN is different or not.
  - Several assumptions to be tested before performing analysis.
  - Less flexible, BUT powerful and commonly used.

Parametric Test

- Parametric tests for comparison of means:
  - Two independent samples: Independent t-test.
  - Two related samples: Paired t-test.
  - More than two independent samples: ANOVA.

Two independent samples: Independent t-test

Two independent samples: Independent t-test

- Purpose: Compare MEANS of TWO independent samples/groups.
- Assumptions:
  1. Numerical outcome.
  2. Normal data distribution for each group.
  3. Equal variance between groups.

Two independent samples: Independent t-test

Research objective:

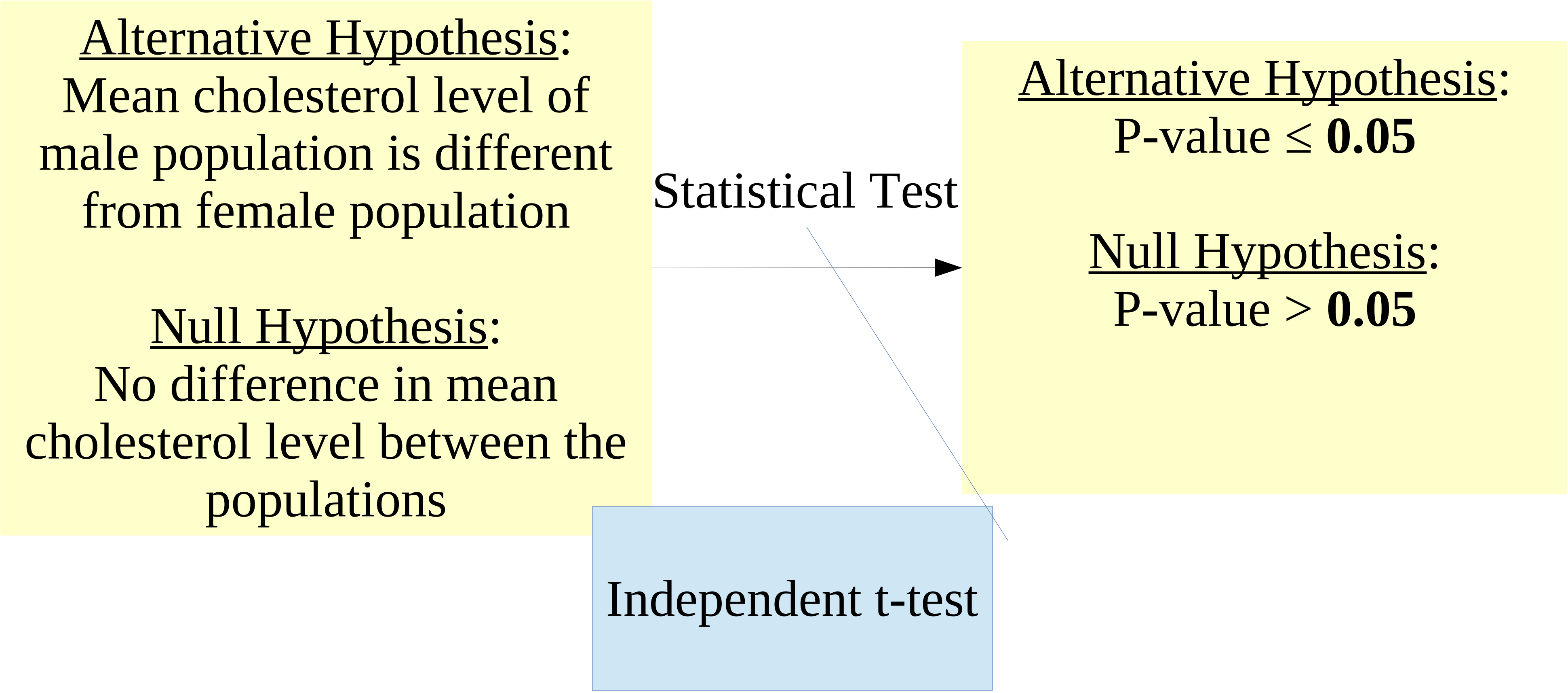
To compare mean cholesterol level between male and female.

Research question:

Is there any difference in mean cholesterol level between male and female populations?

Two independent samples: Independent t-test

RQ: Is there any difference in mean cholesterol level between male and female populations?



Alternative Hypothesis:
Mean cholesterol level of male population is different from female population

Null Hypothesis:
No difference in mean cholesterol level between the populations

Statistical Test

Alternative Hypothesis:
P-value ≤ **0.05**

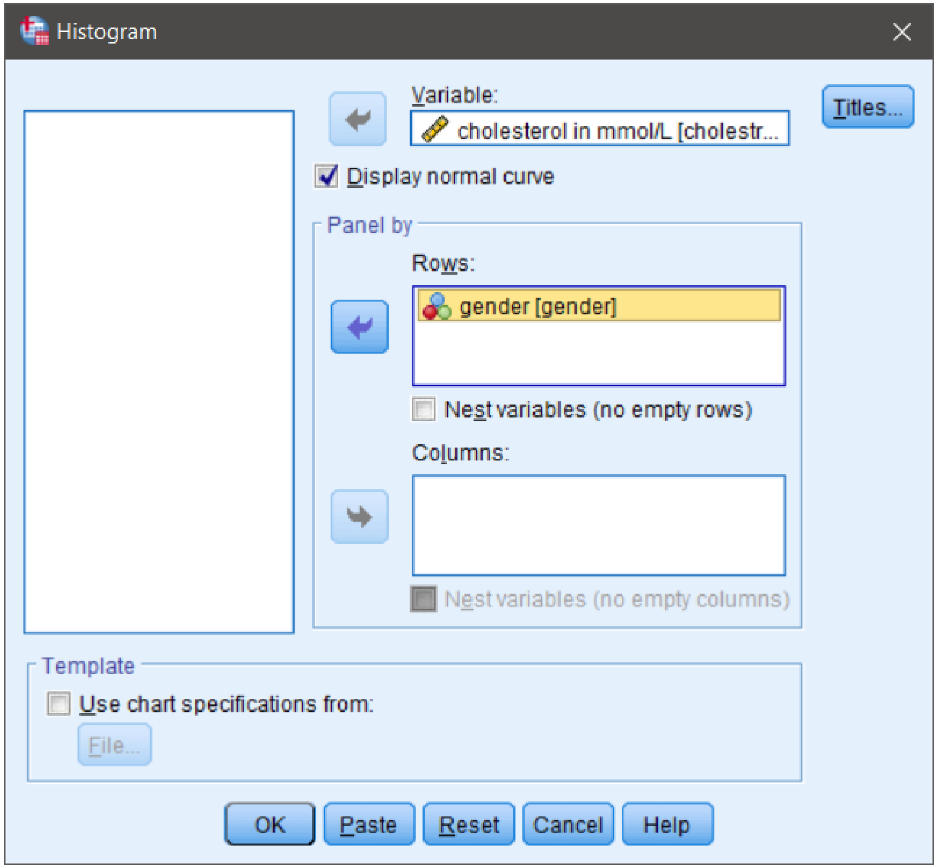
Null Hypothesis:
P-value > **0.05**

Independent t-test

Independent t-test: Practical

- Dataset: cholestrol2.sav
- Sample size: 40/group
- Group: 2 (male and female)
- Outcome: cholesterol level in mmol/L

Normality: Histogram



Variable: cholesterol in mmol/L [cholestr...

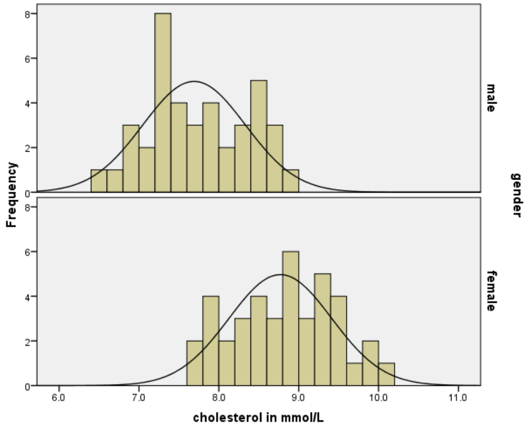
☑ Display normal curve

Rows: gender [gender]

☐ Nest variables (no empty rows)

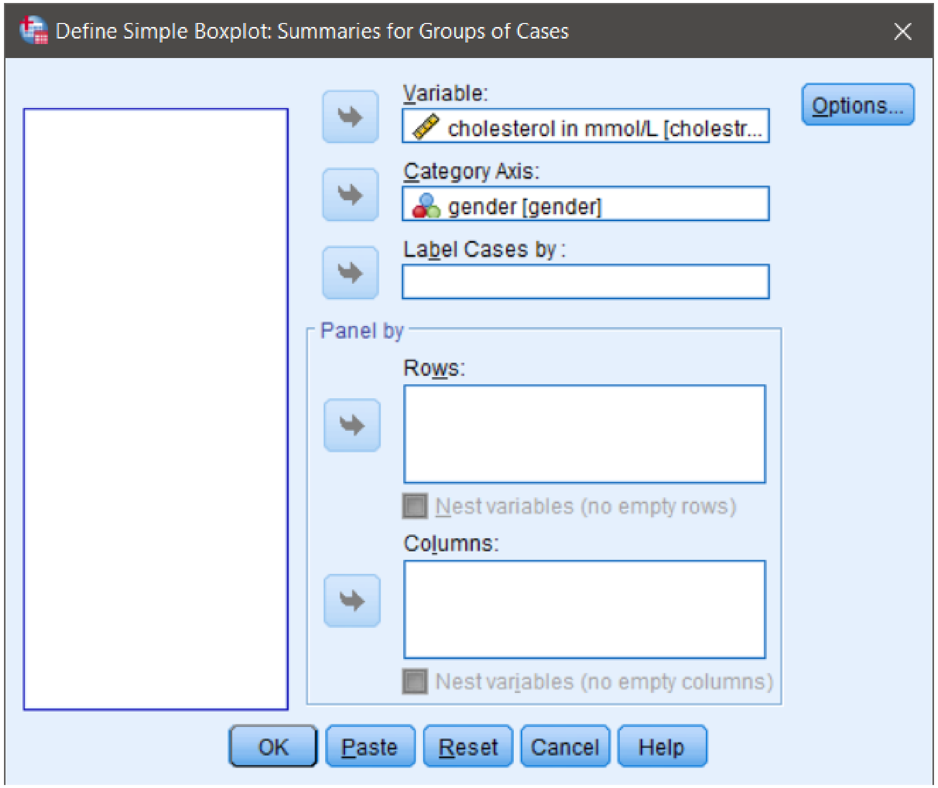
☐ Nest variables (no empty columns)

☐ Use chart specifications from:



1. **Graphs > Legacy Dialogs > Histogram**
2. **Variable: *cholestrol*, Display normal curve: [x], Rows: *gender***
3. **OK**

Normality: Boxplot


Define Simple Boxplot: Summaries for Groups of Cases

Variable: cholesterol in mmol/L [cholestr...

Category Axis: gender [gender]

Label Cases by:

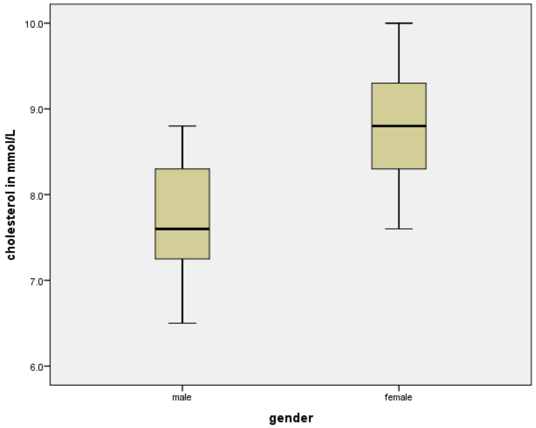
Panel by

Rows:

□ Nest variables (no empty rows)

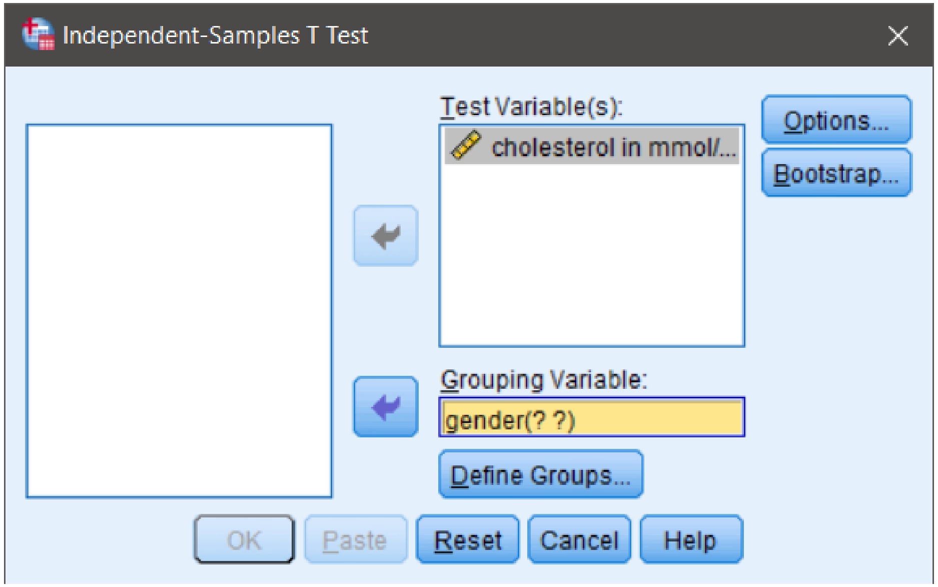
Columns:

□ Nest variables (no empty columns)



1. Graphs > Legacy Dialogs > Boxplot > Simple > Define
2. Variable: *cholestrol*, Category Axis: *gender*
3. OK

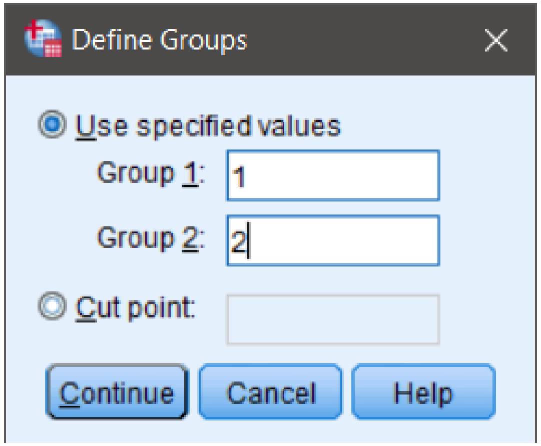
Independent t-test: Steps



Independent-Samples T Test

Test Variable(s): cholesterol in mmol/...

Grouping Variable: gender(? ?)



Define Groups

● Use specified values

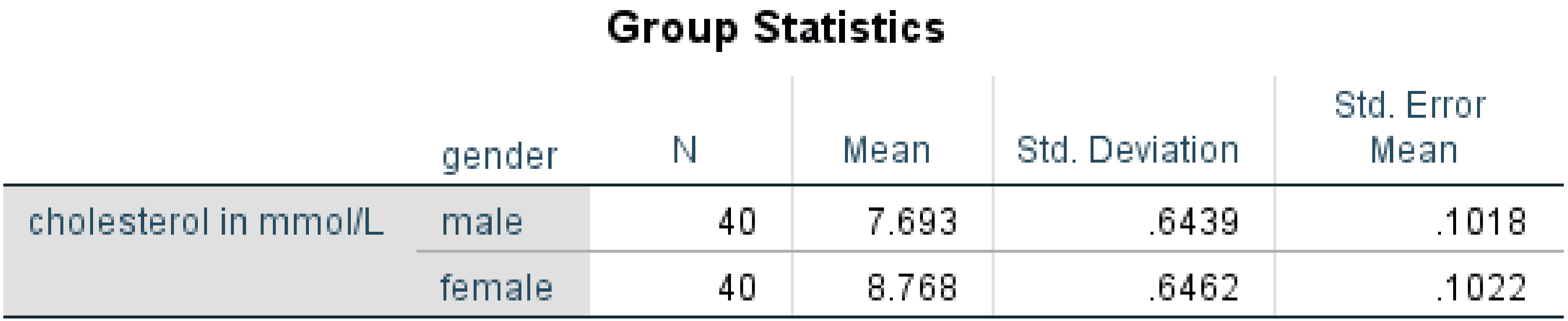
Group 1: 1

Group 2: 2

○ Cut point:

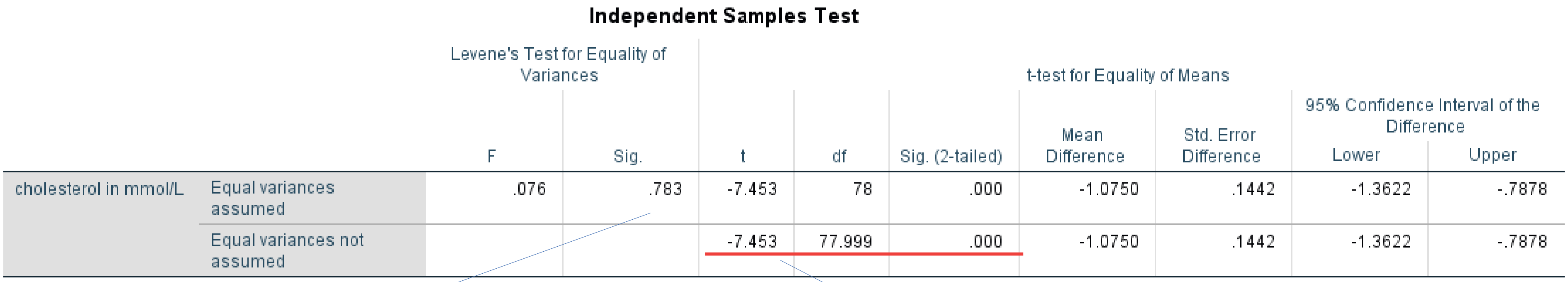
1. **Analyze > Compare Means > Independent-Samples T Test...**
2. **Test Variable(s): *cholestrol*, Grouping Variable: *gender***
3. **[Define Groups] > Group 1: 1, Group 2: 2 > Continue**
4. **OK**

Independent t-test: Results



Group Statistics

	gender	N	Mean	Std. Deviation	Std. Error Mean
cholesterol in mmol/L	male	40	7.693	.6439	.1018
	female	40	8.768	.6462	.1022



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
cholesterol in mmol/L	Equal variances assumed	.076	.783	-7.453	78	.000	-1.0750	.1442	-1.3622	-.7878
	Equal variances not assumed			-7.453	77.999	.000	-1.0750	.1442	-1.3622	-.7878

Equal: $p \geq 0.05$
Unequal: $p < 0.05$

Use Welch t-test when variance not equal

Two related samples:
Paired t-test

Two related samples: Paired t-test

- Purpose: Compare MEAN DIFFERENCE between TWO related samples, i.e. equal to ZERO if there is no difference.
- Assumptions:
  1. Numerical outcome.
  2. Normal distribution of the DIFFERENCES between TWO paired observations (e.g. SBP after treatment – SBP before treatment).

Two related samples: Paired t-test

Research objective:

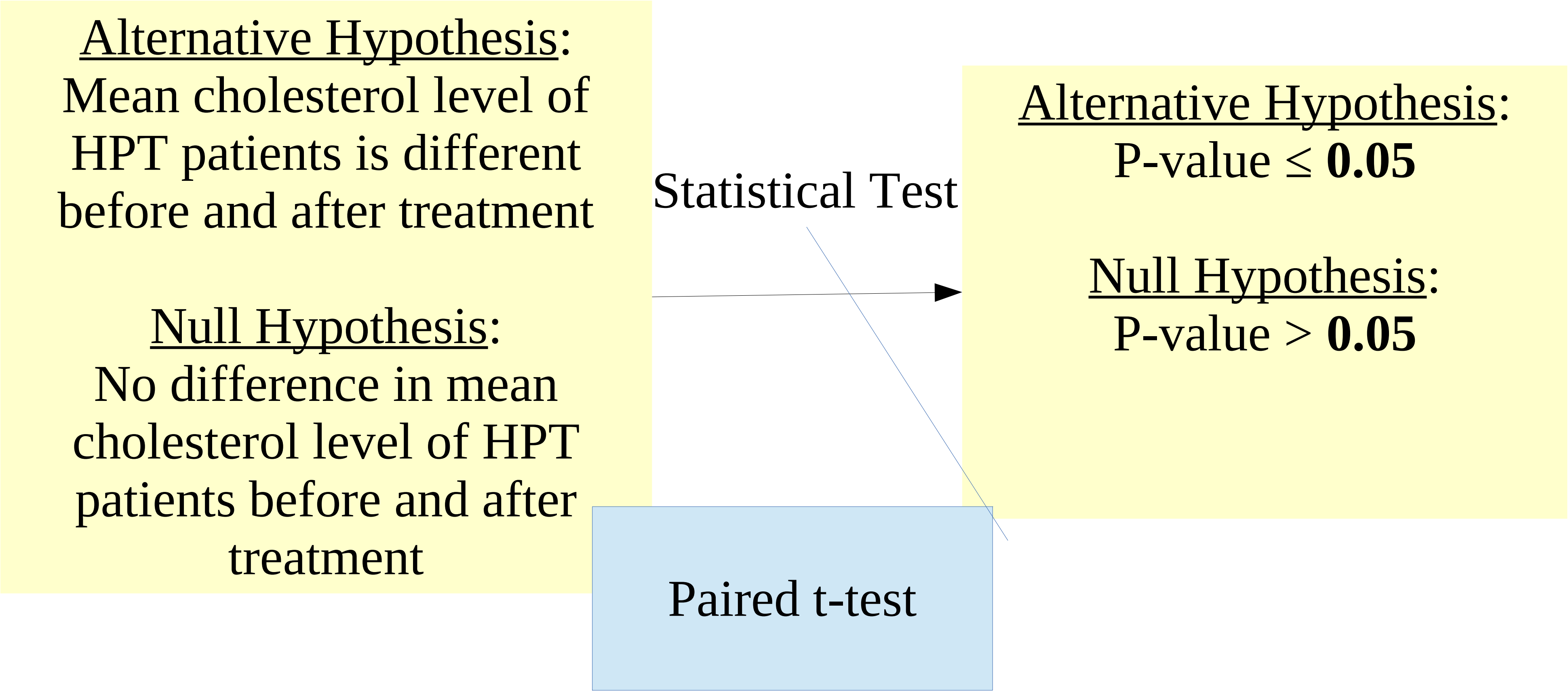
To compare mean cholesterol level of hypertensive patients before and after treatment.

Research question:

Is there any difference in mean cholesterol level of hypertensive patients before and after treatment?

Two related samples: Paired t-test

RQ: Is there any difference in mean mean cholesterol level of hypertensive patients before and after treatment?



Alternative Hypothesis:
Mean cholesterol level of HPT patients is different before and after treatment

Null Hypothesis:
No difference in mean cholesterol level of HPT patients before and after treatment

Statistical Test

Paired t-test

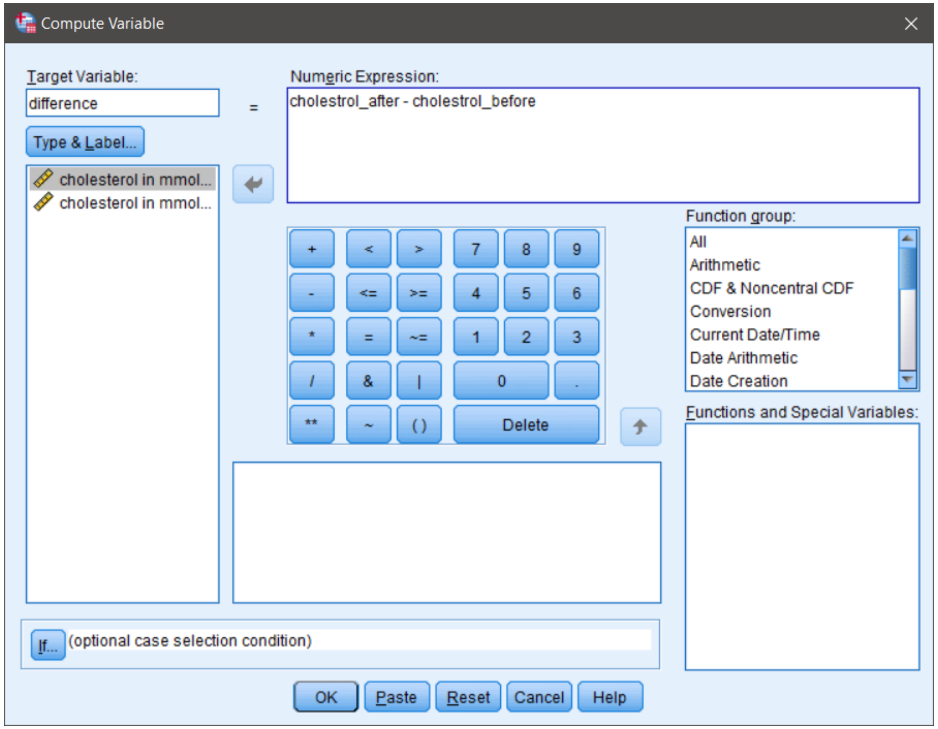
Alternative Hypothesis:
P-value ≤ **0.05**

Null Hypothesis:
P-value > **0.05**

Paired t-test: Practical

- Dataset: cholestrol_prepost.sav
- Sample size: 30 paired observations
- Repetition: 2 (before and after treatment)
- Outcome: cholesterol level in mmol/L

Compute difference

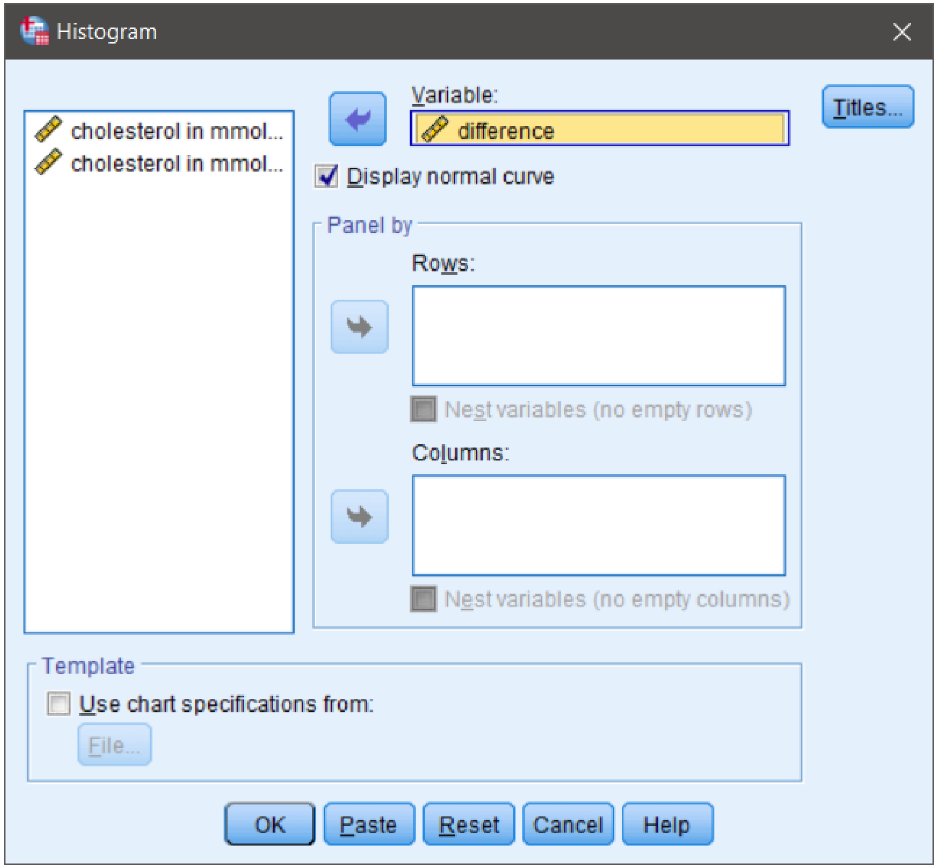


Target Variable: difference

Numeric Expression: cholestrol_after - cholestrol_before

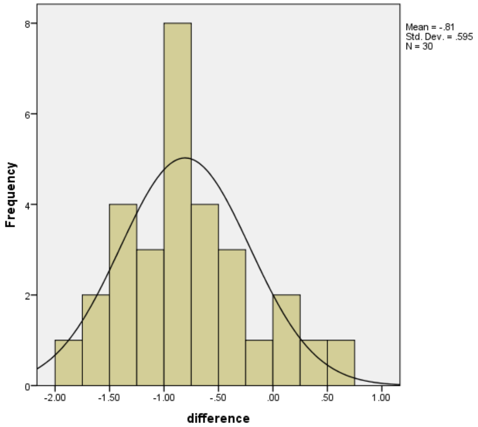
1. **Transform > Compute Variable...**
2. **Target Variable: *difference*, Numeric Expression: *cholestrol_after - cholestrol_before***
3. **OK**

Normality: Histogram



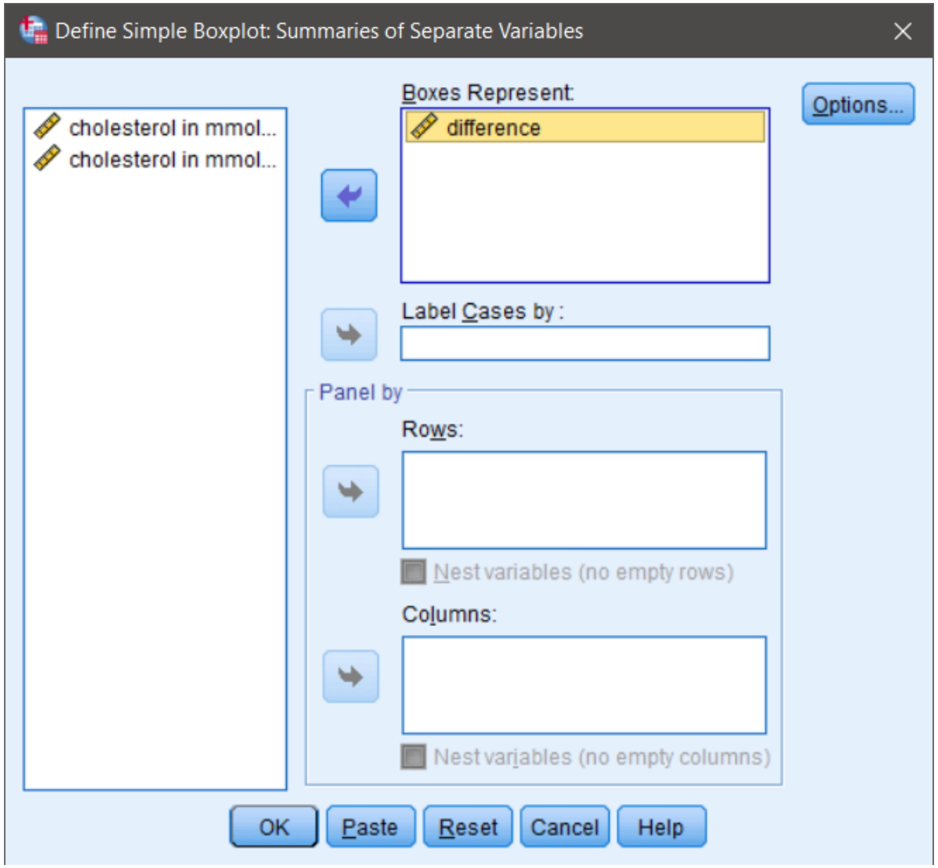
Variable: difference

☑ Display normal curve



1. **Graphs > Legacy Dialogs > Histogram**
2. **Variable: *difference*, Display normal curve: [x]**
3. **OK**

Normality: Boxplot



Define Simple Boxplot: Summaries of Separate Variables

cholesterol in mmol...
cholesterol in mmol...

Boxes Represent: difference

Label Cases by :

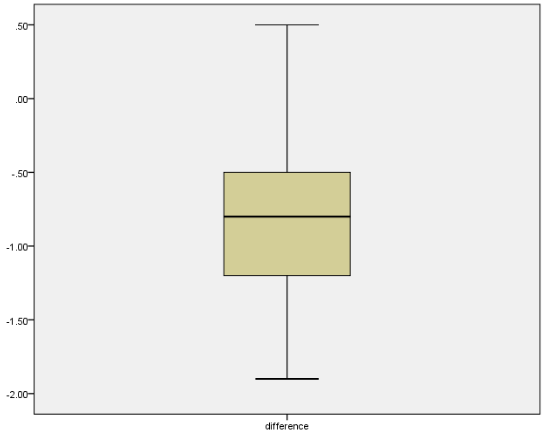
Panel by

Rows:

☐ Nest variables (no empty rows)

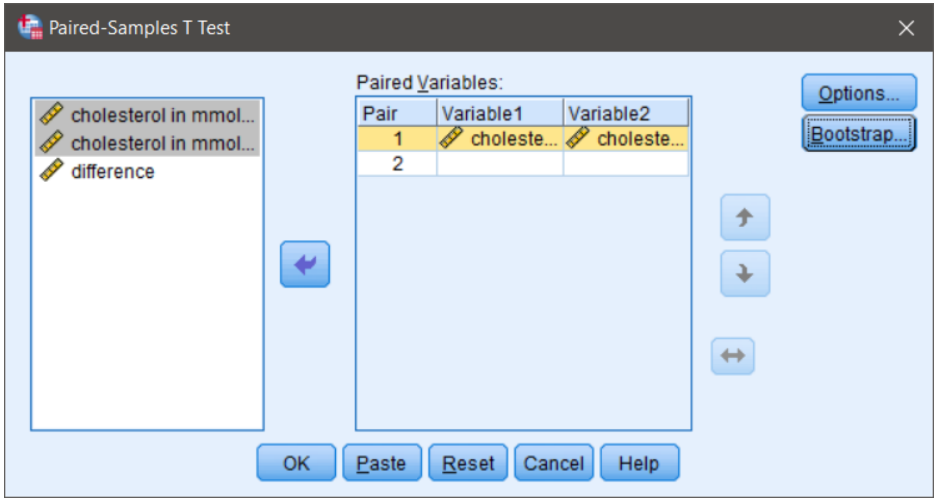
Columns:

☐ Nest variables (no empty columns)



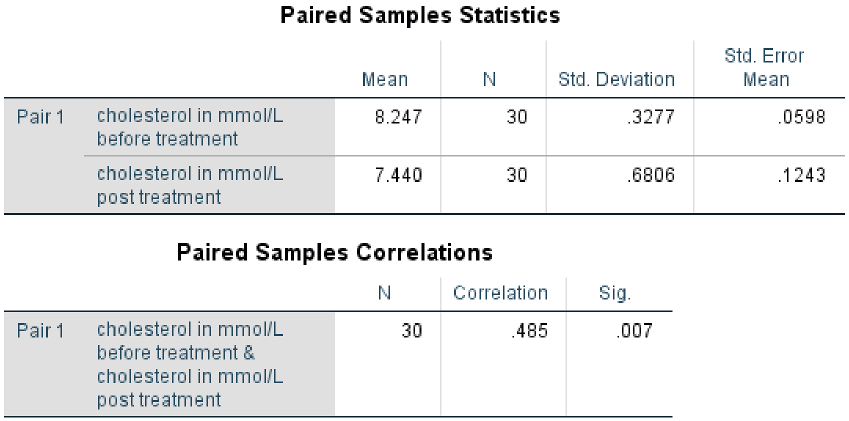
1. **Graphs > Legacy Dialogs > Boxplot > Simple, Data in Chart Are: Summaries of separate variables [x] > Define**
2. **Boxes Represent: *difference***
3. **OK**

Paired t-test: Steps



1. **Analyze > Compare Means > Paired-Samples T Test...**
2. **Select both *cholestrol_before, cholestrol_after* → Paired Variables**
3. **OK**

Paired t-test: Results

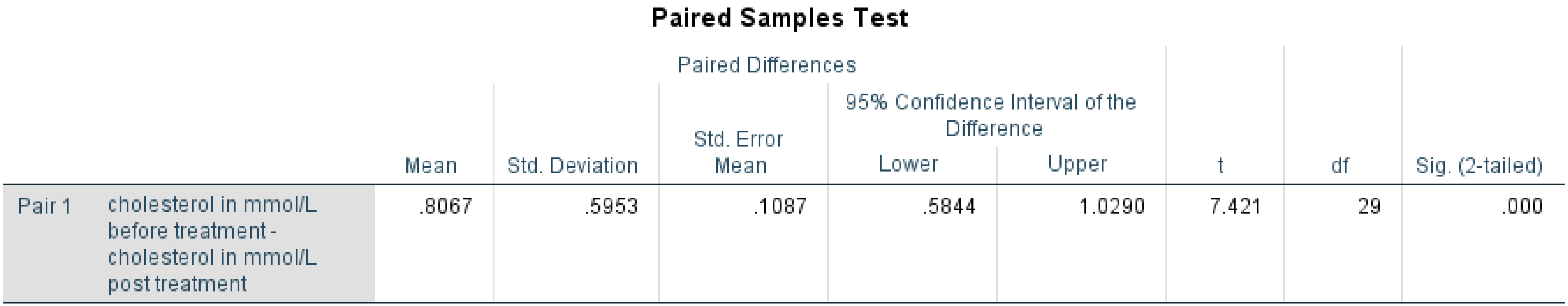


Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	cholesterol in mmol/L before treatment	8.247	30	.3277	.0598
	cholesterol in mmol/L post treatment	7.440	30	.6806	.1243

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	cholesterol in mmol/L before treatment & cholesterol in mmol/L post treatment	30	.485	.007



Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	cholesterol in mmol/L before treatment - cholesterol in mmol/L post treatment	.8067	.5953	.1087	.5844	1.0290	7.421	29	.000

More than two independent samples: ANOVA

More than two independent samples: ANOVA

- ANalysis Of VAriance.
- Purpose: Compare MEANS of THREE/MORE independent samples/groups.
- Assumptions:
  1. Numerical outcome.
  2. Normal data distribution for each group.
  3. Equal variance between groups.

More than two independent samples: ANOVA

Research objective:

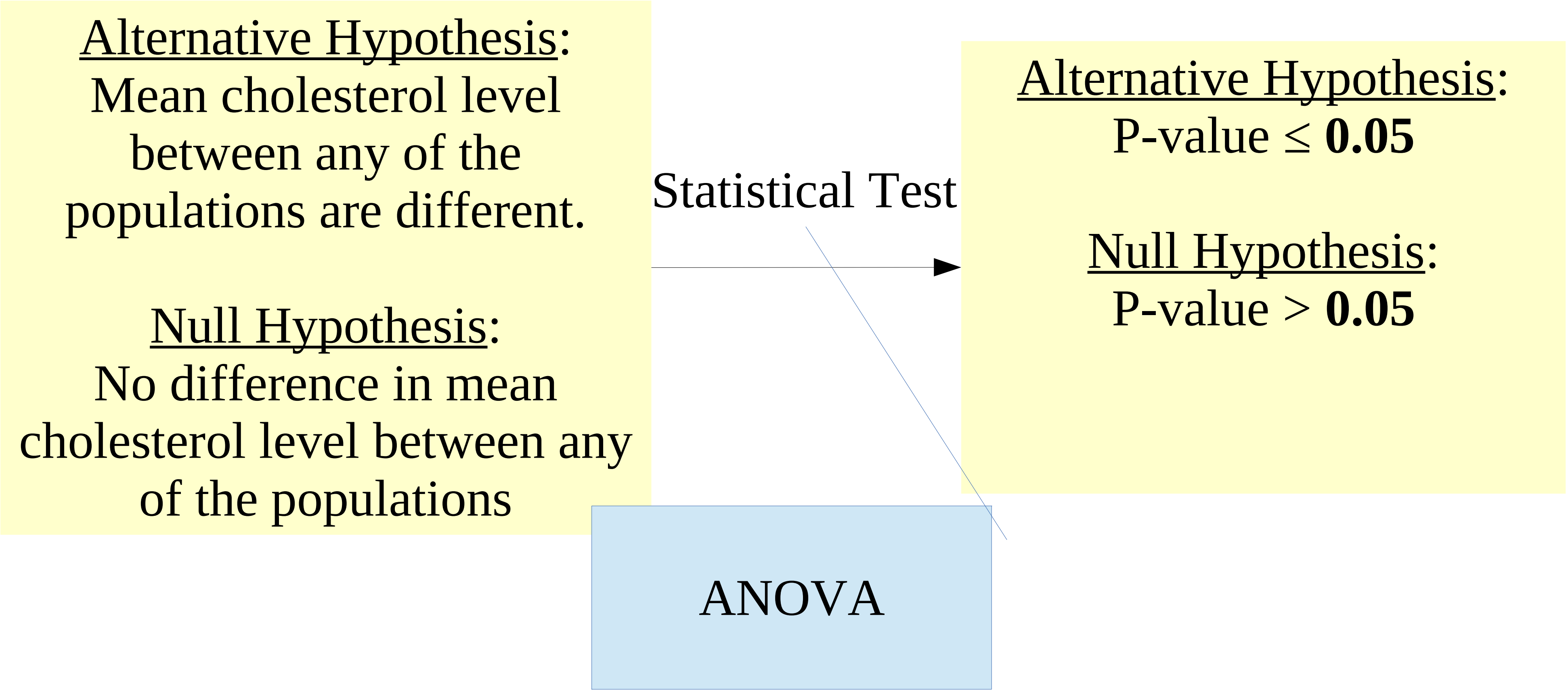
To compare mean cholesterol level between Group A, B and C treatment groups.

Research question:

Is there any difference in mean cholesterol level between Group A, B and C treatment groups?

More than two independent samples: ANOVA

RQ: Is there any difference in mean cholesterol level between Group A, B and C treatment groups?



Alternative Hypothesis:
Mean cholesterol level between any of the populations are different.

Null Hypothesis:
No difference in mean cholesterol level between any of the populations

Statistical Test

ANOVA

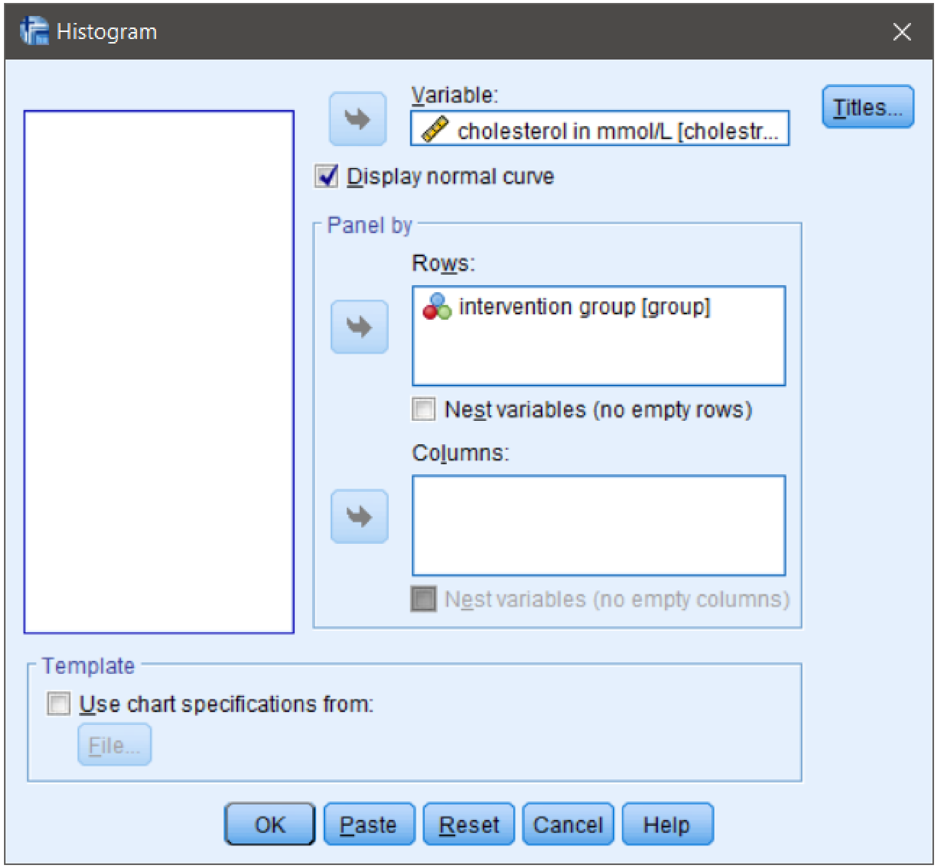
Alternative Hypothesis:
P-value ≤ **0.05**

Null Hypothesis:
P-value > **0.05**

ANOVA: Practical

- Dataset: cholestrol3.sav
- Sample size: 25/group
- Group: 3 (Grp A, B and C)
- Outcome: cholesterol level in mmol/L

Normality: Histogram



Histogram

Variable:
cholesterol in mmol/L [cholestr...

☑ Display normal curve

Panel by

Rows:
intervention group [group]

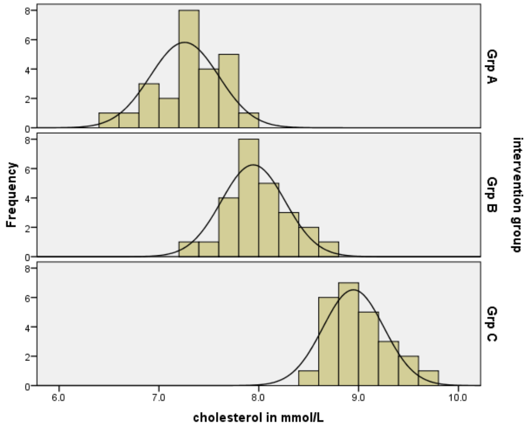
☐ Nest variables (no empty rows)

Columns:

☐ Nest variables (no empty columns)

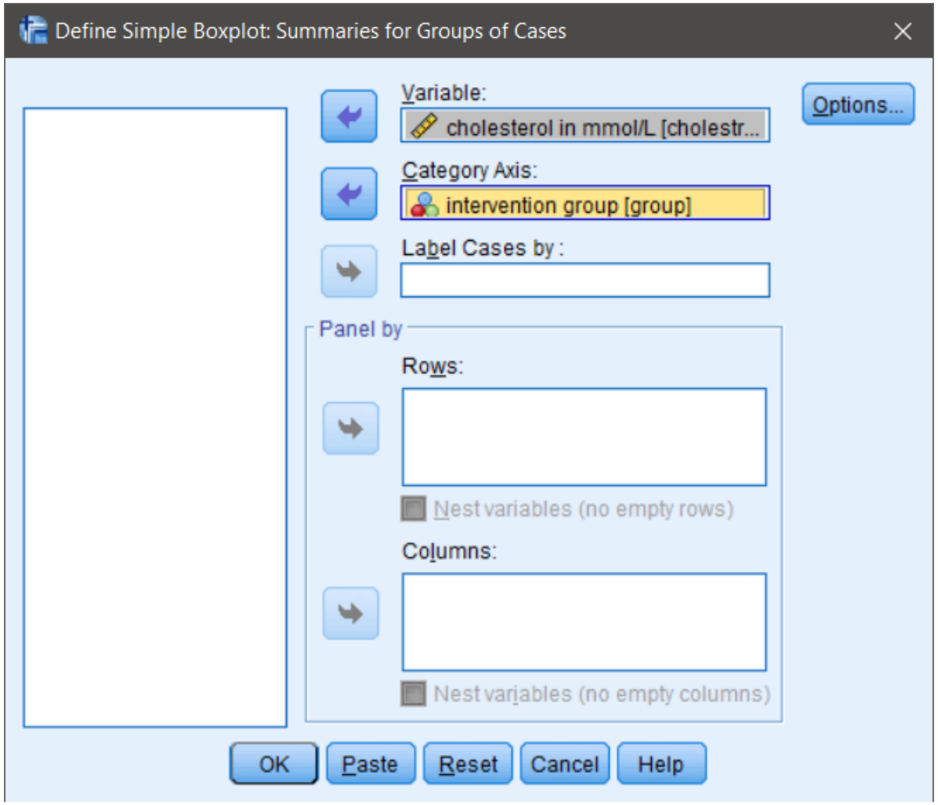
Template

☐ Use chart specifications from:



1. **Graphs > Legacy Dialogs > Histogram**
2. **Variable: *cholestrol*, Display normal curve: [x], Rows: *group***
3. **OK**

Normality: Boxplot

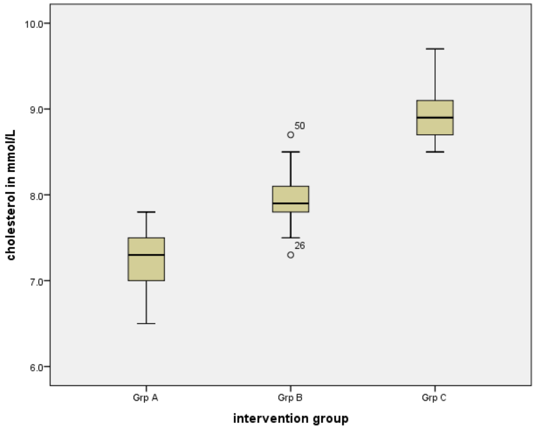

Define Simple Boxplot: Summaries for Groups of Cases

Variable: cholesterol in mmol/L [cholestr...

Category Axis: intervention group [group]

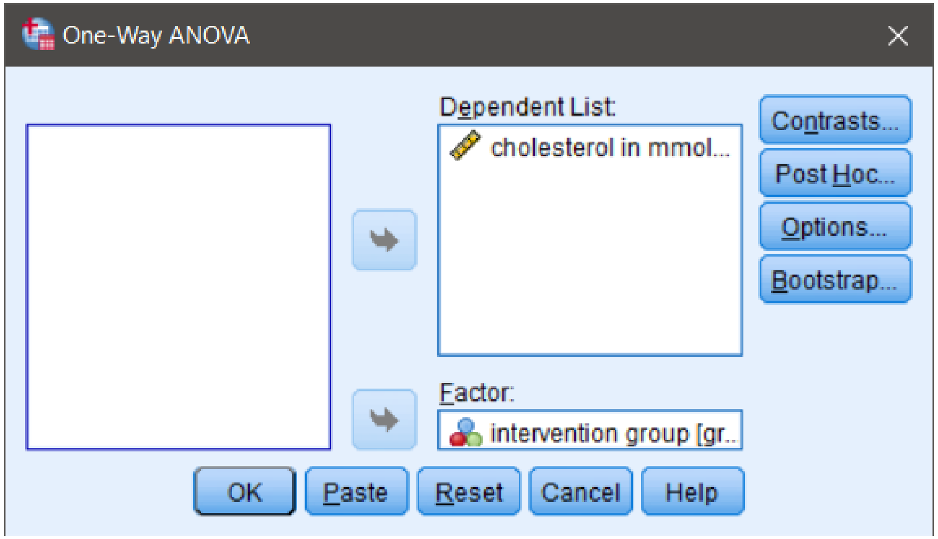
Label Cases by:

Options...
Panel by
Rows:
□ Nest variables (no empty rows)
Columns:
□ Nest variables (no empty columns)
OK
Paste
Reset
Cancel
Help



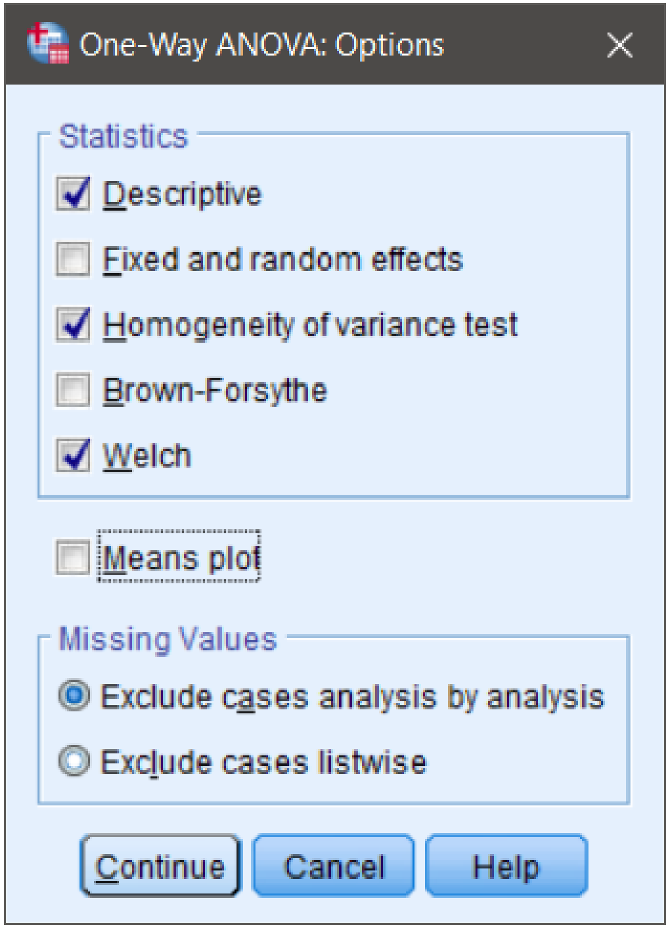
1. **Graphs > Legacy Dialogs > Boxplot > Simple > Define**
2. **Variable: *cholestrol*, Category Axis: *group***
3. **OK**

ANOVA: Steps



1. **Analyze > Compare Means > One-Way ANOVA...**
2. **Dependent List: *cholestrol*, Factor: *group***

ANOVA: Steps



One-Way ANOVA: Options

Statistics

☑ Descriptive

□ Fixed and random effects

☑ Homogeneity of variance test

□ Brown-Forsythe

☑ Welch

□ Means plot

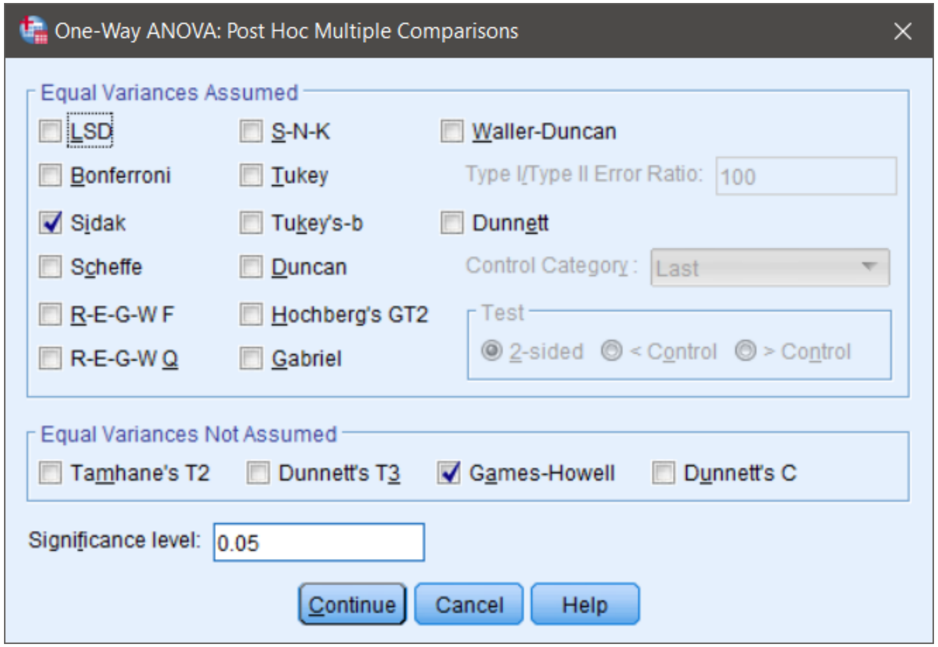
Missing Values

● Exclude cases analysis by analysis

○ Exclude cases listwise

3. **[Options...] > Statistics: Descriptive [x] Homogeneity of variance test [x] Welch [x] > Continue**

ANOVA: Steps



One-Way ANOVA: Post Hoc Multiple Comparisons

Equal Variances Assumed

□ LSD □ S-N-K □ Waller-Duncan

□ Bonferroni □ Tukey Type I/Type II Error Ratio: 100

☑ Sidak □ Tukey's-b □ Dunnett

□ Scheffe □ Duncan Control Category: Last

□ R-E-G-W F □ Hochberg's GT2 Test

□ R-E-G-W Q □ Gabriel ● 2-sided ○ < Control ○ > Control

Equal Variances Not Assumed

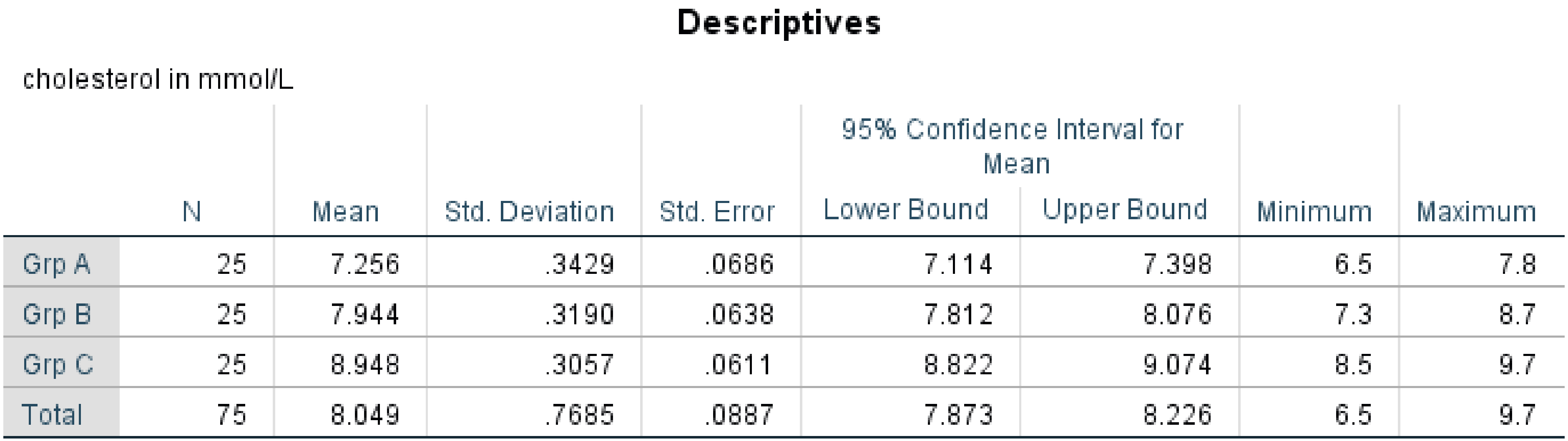
□ Tamhane's T2 □ Dunnett's T3 ☑ Games-Howell □ Dunnett's C

Significance level: 0.05

4. **[Post Hoc...] > Equal Variances Assumed: Sidak [x], Equal Variances Not Assumed: Games-Howell [x] > Continue**

5. **OK**

ANOVA: Results

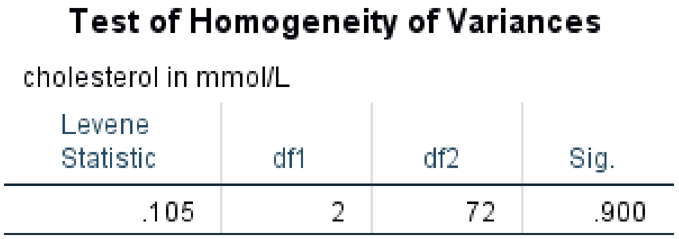


Descriptives

cholesterol in mmol/L

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Grp A	25	7.256	.3429	.0686	7.114	7.398	6.5	7.8
Grp B	25	7.944	.3190	.0638	7.812	8.076	7.3	8.7
Grp C	25	8.948	.3057	.0611	8.822	9.074	8.5	9.7
Total	75	8.049	.7685	.0887	7.873	8.226	6.5	9.7

ANOVA: Results

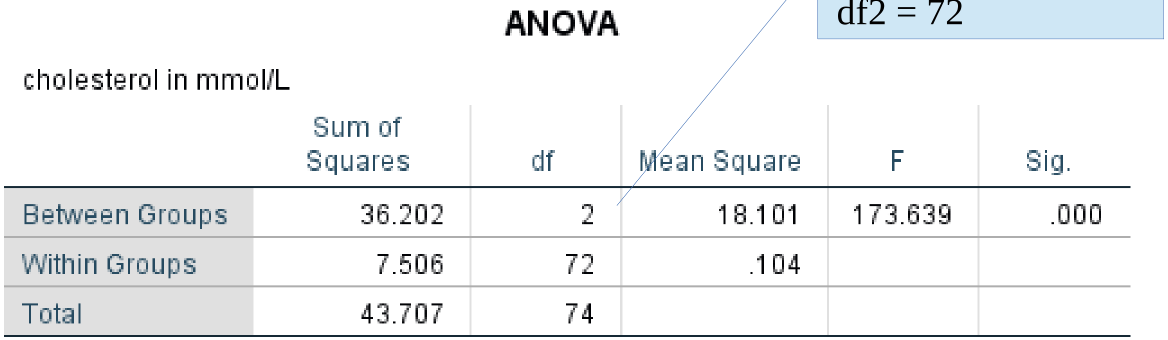


Test of Homogeneity of Variances

cholesterol in mmol/L

Levene Statistic	df1	df2	Sig.
.105	2	72	.900

Equal: p ≥ 0.05
Unequal: p <0.05

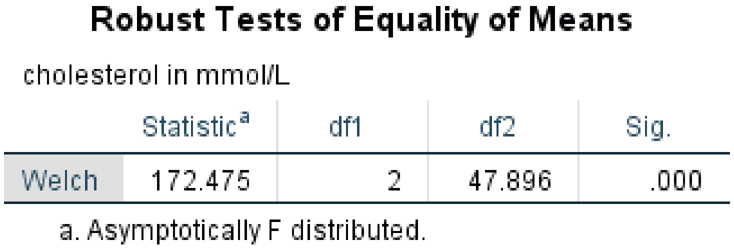


ANOVA

cholesterol in mmol/L

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	36.202	2	18.101	173.639	.000
Within Groups	7.506	72	.104		
Total	43.707	74			

df1 = 2
df2 = 72



Robust Tests of Equality of Means

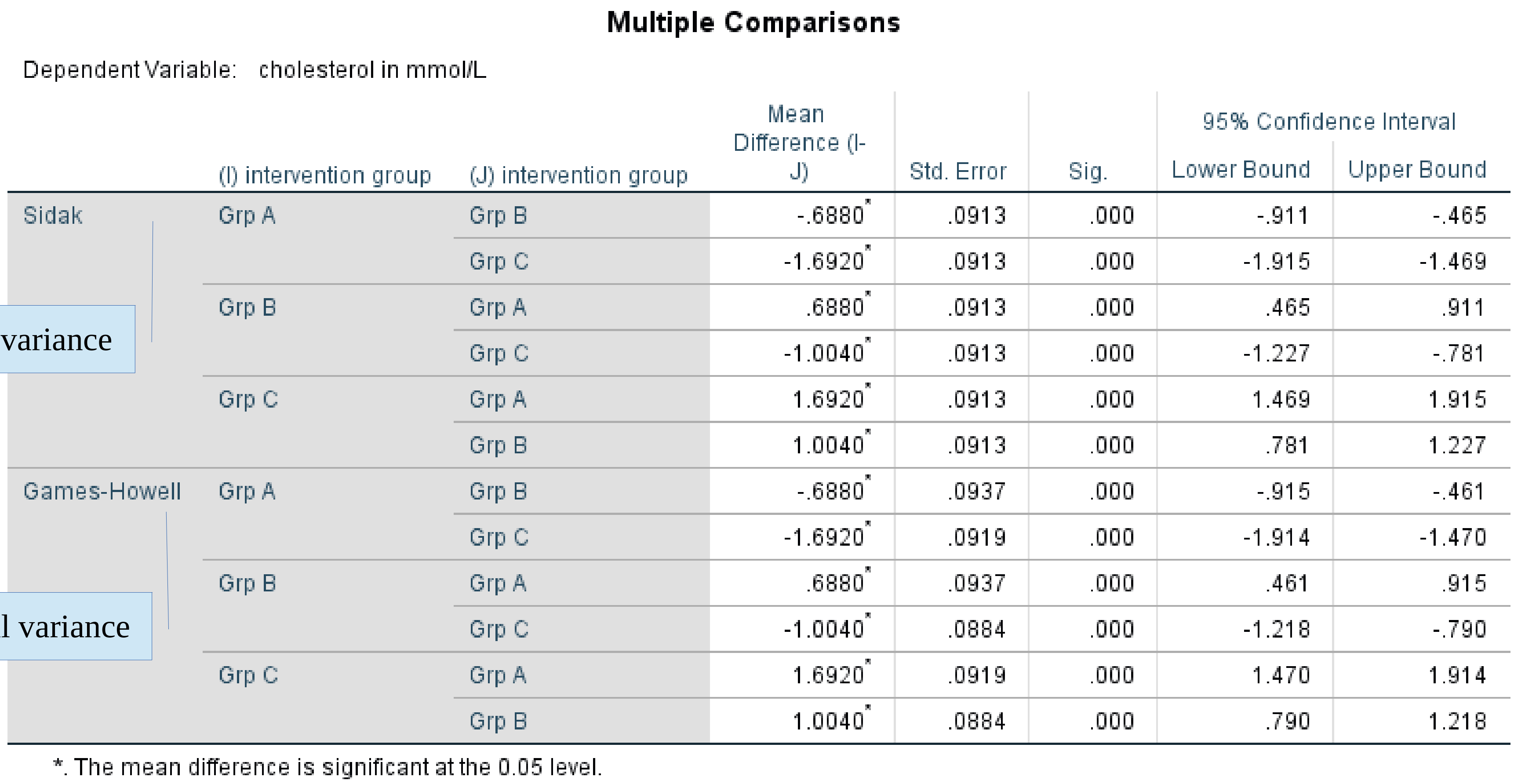
cholesterol in mmol/L

	Statistic[a]	df1	df2	Sig.
Welch	172.475	2	47.896	.000

a. Asymptotically F distributed.

Use Welch ANOVA when variance not equal

ANOVA: Results



Multiple Comparisons

Dependent Variable: cholesterol in mmol/L

	(I) intervention group	(J) intervention group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Sidak	Grp A	Grp B	-.6880*	.0913	.000	-.911	-.465
		Grp C	-1.6920*	.0913	.000	-1.915	-1.469
	Grp B	Grp A	.6880*	.0913	.000	.465	.911
		Grp C	-1.0040*	.0913	.000	-1.227	-.781
	Grp C	Grp A	1.6920*	.0913	.000	1.469	1.915
		Grp B	1.0040*	.0913	.000	.781	1.227
Games-Howell	Grp A	Grp B	-.6880*	.0937	.000	-.915	-.461
		Grp C	-1.6920*	.0919	.000	-1.914	-1.470
	Grp B	Grp A	.6880*	.0937	.000	.461	.915
		Grp C	-1.0040*	.0884	.000	-1.218	-.790
	Grp C	Grp A	1.6920*	.0919	.000	1.470	1.914
		Grp B	1.0040*	.0884	.000	.790	1.218

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

Equal variance

Unequal variance

Q&A