

STATISTICAL ANALYSIS: WHICH TO CHOOSE?

1. Process of data management (follow the steps below)

- Research question(s)
- Research design
- Data collection
- Data entry
- Data exploration & cleaning
- Data analysis
- Interpretation
- Writing up

2. Role of statistics in a study

- Statistical knowledge and judgment is required at every step of a study
- What statistical analysis is appropriate to answer the research question? Points to consider to select the right statistical test:
 - Research question/ hypothesis
 - Are you clear what you want to find out and what design you have used in your study?
 - Number of variables
 - Type of data
 - Number of groups
 - Sample distribution
 - Sample type

3. Research question

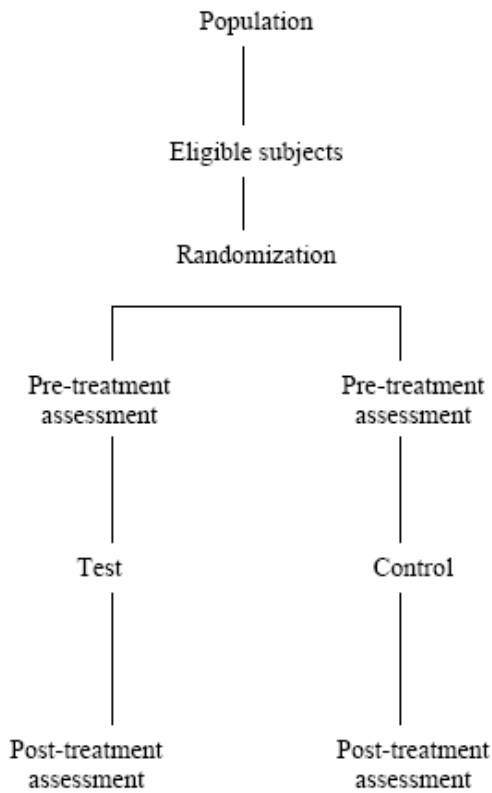
- The essential question, the study is designed to answer the question

- Most studies are concerned with answering one of four types of following questions
 - What is the magnitude of a health problem or health factor?
 - What is the efficacy of an intervention?
 - What is the casual relation between one factor (or factors) and the disease or outcome of interest?
 - What is the natural history of a disease?
- What is/are the research question (s)?
 - Common in medical research:
 - Difference between/ among means
 - Difference between/ among proportions
 - Associations between/ among factors
 - Difference between/ among treatment effects
- Hypothesis
 - This is a testable statement that describes the nature of the proposed relationship between two/ more variables interest
 - E.g. there is an association between smoking and coronary heart disease

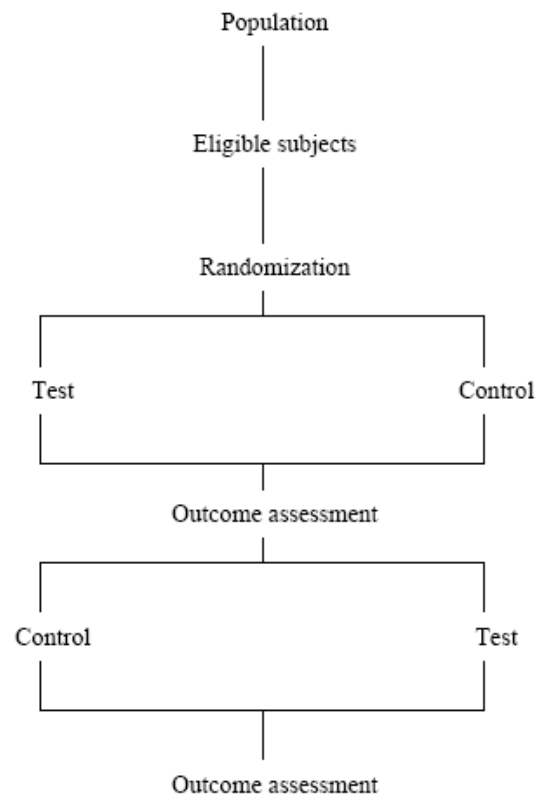
4. What is the research design applied and expected result?

- Randomized control trial (RCT)
- Observational studies
 - Cross-sectional
 - Case-control
 - Prospective cohort
 - Retrospective cohort
- Case report/ series
- Diagnostic test
- E.g. 1
 - Research question: effectiveness of new anti-hypertensive drug

- o Research design: randomized controlled trial

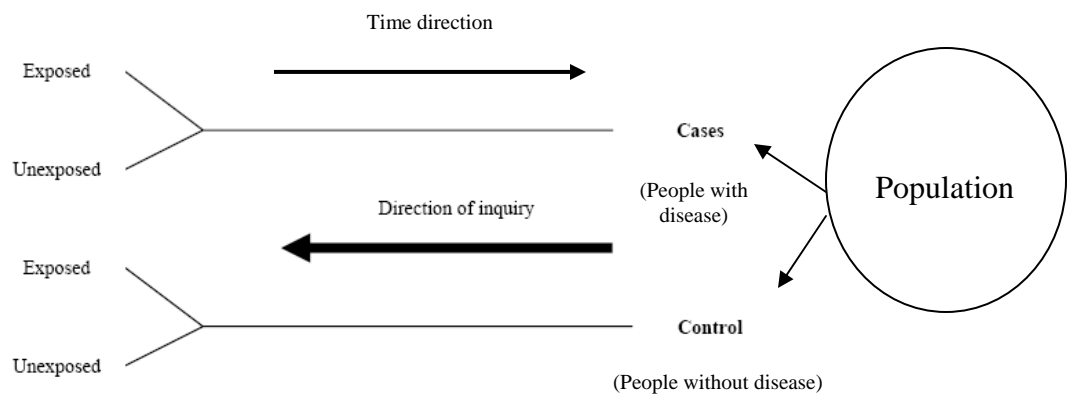


Parallel RCT

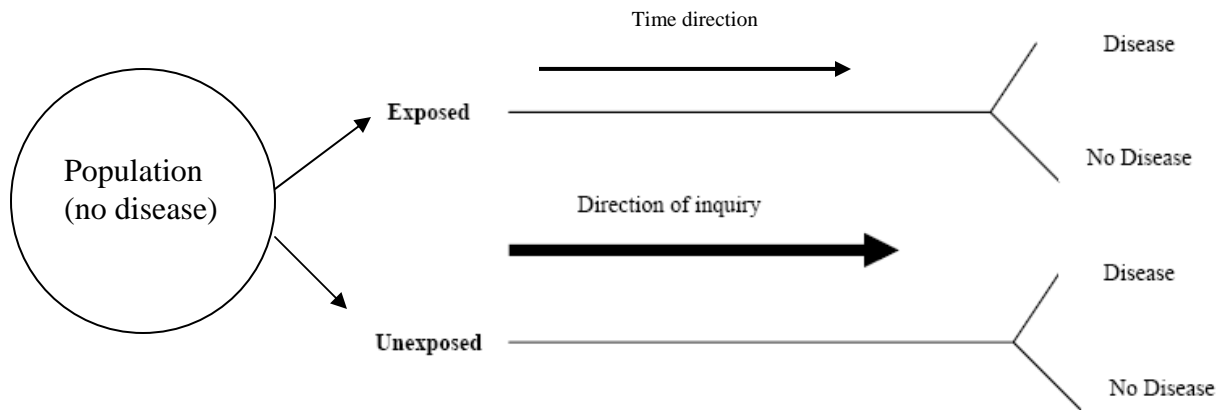


Cross-over RCT

- E.g. 2
 - o Research question: Risk factor for enteric fever
 - o Research design: Case control



- E.g. 3
 - Research question: maternal & fetal outcome in mother with PIH
 - Research design: prospective cohort



5. Study factor (s)

- Variable (s) of interest that is hypothesized to be related to health problem, disease or outcome of interest.
- Also known as the independent variables/ exposure variables/ determinants

6. Outcome factor (s)

- The event or occurrence that is supposed to have as a result of the study factor
- E.g. the outcome factor is blood pressure, as it influenced by study factors, salt intake.
- Also known as dependent variable.

7. Number of variables

- One independent variable only – univariate analysis
- More than one independent factor variables – multivariate analysis
- Less likely to conduct and conclude a study with only focusing on univariate analysis in health sciences

- If there is a multi-factorial effect on the outcome, univariate gives misleading results
- Example risk factors for coronary heart disease
- Multivariate analysis can eliminate confounding effect

8. Type of data

- Numerical
 - Continuous (e.g. weight)
 - Discrete (e.g. number of patients admitted)
- Categorical
 - Nominal (e.g. occupation, gender)
 - Ordinal (e.g. disease severity, socioeconomic status)
- Statistical tests applied are different based on the type of variables (must consider both independent and dependent variables)

9. Number of group

- Two group (two levels) (e.g. diabetic and non-diabetic group)
- More than two group (more than two levels) (e.g. race – Malay, Chinese, Indian, Others)

10. Sample distribution

- Normal distribution → parametric test
- Non-normal distribution → non-parametric test
- Suggested procedure for assessing normality
 - Compare the mean & median (for normal distribution mean = median)
 - Construct a histogram overlaid with normal curve
 - Construct a box and whisker plot
 - Statistical test
 - Kolmogorov-Sminov test

- Shapiro-wilk test
- Non-parametric test are appropriate when:
 - Data is ordinal
 - Data is non-normal distribution and cannot be easily transformed
 - Data may contain outlier
- Non-parametric methods have two general limitations
 - Not as powerful as parametric counterparts
 - Test for complex design are not readily available in standard computer packages

11. Sample type

- Independent sample (e.g. disease and non-disease groups, male and female)
- Dependent/ paired/ matched sample (e.g. difference of blood pressure measurements before and after treatment, age and sex matched samples)

12. What to be asked before choosing a statistical test?

- What is the research question/ hypothesis?
- What is the outcome factor and what are the study factors?
- How many variables?
- How many groups?
- What is the distribution like?
- Are the samples independent?
- Is the data numerical/ categorical?

13. Data exploration and cleaning

- Compulsory to do
- Do not rush to analyze data

- Clean and explore first
- Get acquaintance with the data
- Check duplications
- Out-of-range values and location of error
- Distribution of variables
- Missing data checking consistency errors
- Exploring the relationship between variables
- Transformations
- To get acquaintance with data set before the major analysis is carried out
 - Read the protocol again
 - Recall the objectives
 - Identify major outcome, exposure and potential confounders/ effect modifier
 - To check records with duplicating ID number (to prevent repeated data entry)
- Error checking
 - Respondent's mis-marking answers
 - Coder's miscoding response
 - Marking errors by data personnel
- Out-of-range values and location errors
 - Measurement error
 - Recording error
 - Genuine observation
- What to do?
 - Check again original measurements where possible
 - If original measurements suspicious → repeat the measurement
 - If not possible to check → common sense
 - If the value is impossible/implausible → justifiable to set as "missing"

14. Distribution of the variables

- Examine each variable
 - Continuous
 - Normal distribution
 - If not
 - ? transformation
 - ? categorization
 - Categorical
 - Frequency distribution

15. Missing data

- Occur when respondent would/could not answer
- Too much missing data
 - Threat the study
 - Indicate a problem with a question
- Should not be entered as a blank as some statistical packages interpret blanks as zeroes
- Common practice – coded as 9, 99 or 999

16. Consistency errors

- Situations where respondents answered a question for which they were ineligible or when codes were entered incorrectly
- Countercheck with questionnaire/data collection form
- Can be prevented by proper programming in some statistical software

17. Exploring the relationship between variables

- Cross tabulation useful for categorical variables (sometimes better to categorize)
- Should consider confounding & interaction

- Graphs – mostly for continuous variables
- Relationship between the outcome variable and other variables
 - E.g. scatter plot

18. Transformation

- Severely skewed data – two approaches
 - Use nonparametric methods
 - Apply transformation
- Many distributions in medicine – skewed to the right
- Involve performing a mathematical operation on every value of the variable
- Improves the symmetry of the distribution

Transformation	Name	Effect
X^3	Cube	Reduce extreme skewness to left
X^2	Square	Reduce skewness to left
$X^{1/2}$	Square root	Reduce mild skewness to right
$\log_{10}(X)$	Log	Reduce skewness to right
$-1/\sqrt{X}$	-ve reciprocal root	Reduce extreme skewness to right
$-1/X$	-ve reciprocal	Reduce events more extreme skewness to right.

- Check the symmetry of the distribution after transformation
- If sufficiently improved → use the transformed data
- If resistant to transformation → use nonparametric methods

19. Interpretation

- Most confusing part of researchers
- May be the most difficult part for those who are not familiar with statistical applications

- Should interpret only when considered to be results of final analysis stage
 - E.g. in multivariate analysis, final model should be interpreted for writing regardless of the prior more-favorable results towards the hypothesis
- Recall statistical theory and concepts whenever applicable
- May need help from a medical statistician

20. Univariate analysis

- Test hypothesis between one independent and one dependent variable

21. Multivariate analysis

- Why we need multivariate analysis?
- Purpose of using multivariate analysis
- Common multivariate analysis methods in health sciences research.

Variables	Variables
Independent Predictor Explanatory	Dependent Outcome Response

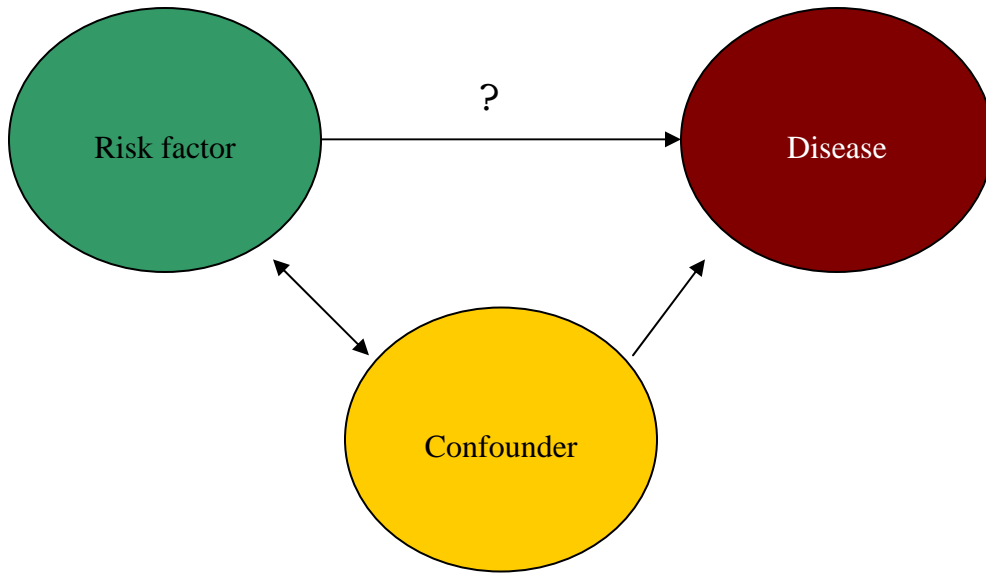
{

 Covariates
 Confounders
 Controls
 Effect modifiers

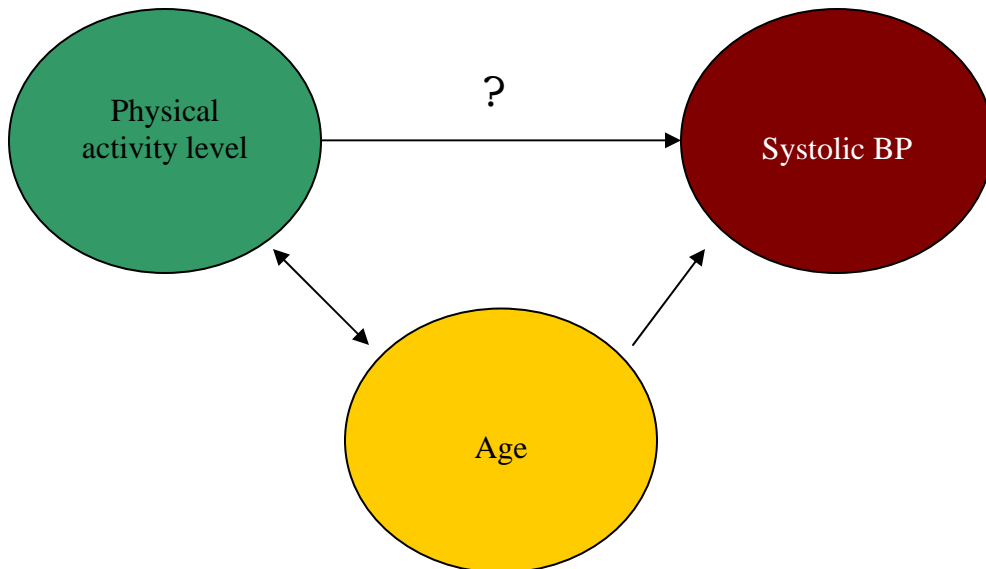
}

Not the primary interest
 Must be recognized

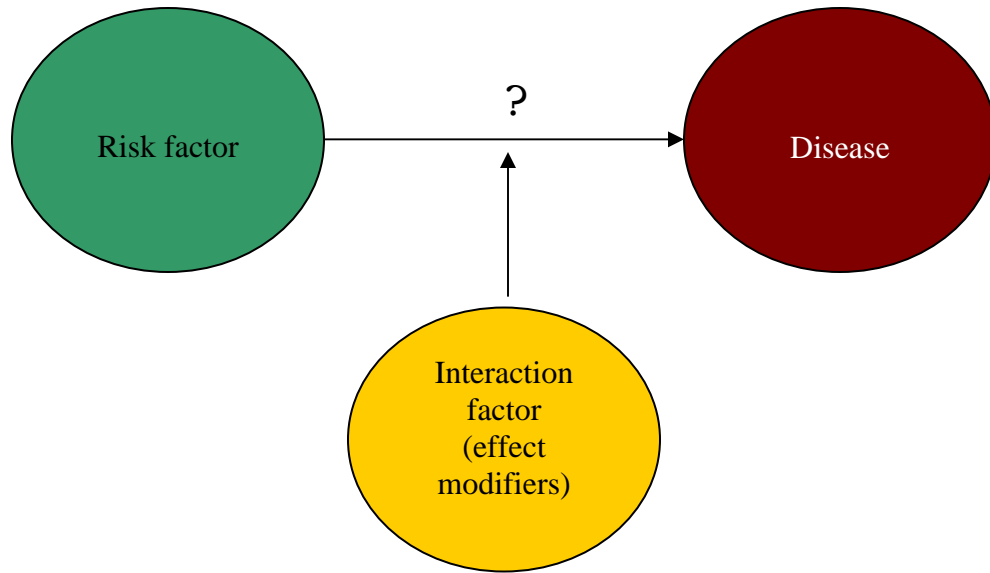
- Confounding



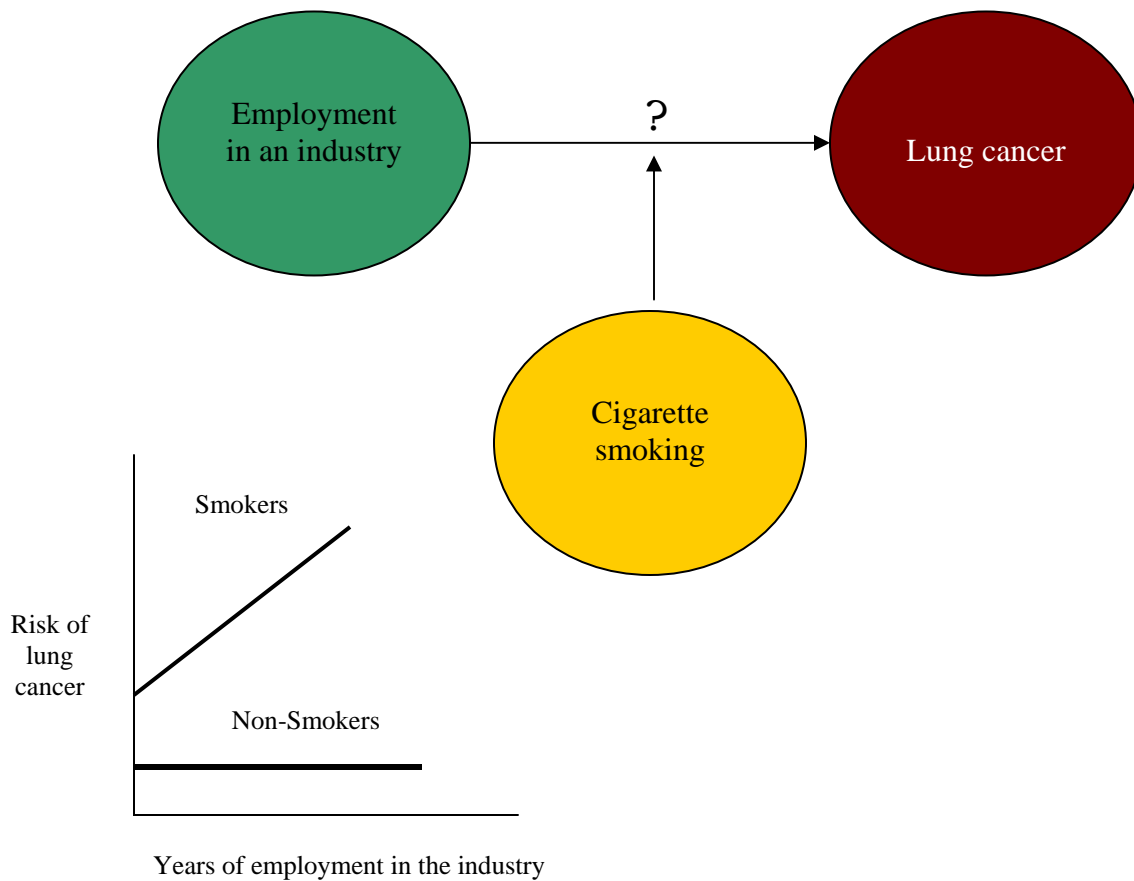
- Distortion of a risk factor-disease relationship brought about by the association of other factors with both risk factor and disease
- Example of confounding:



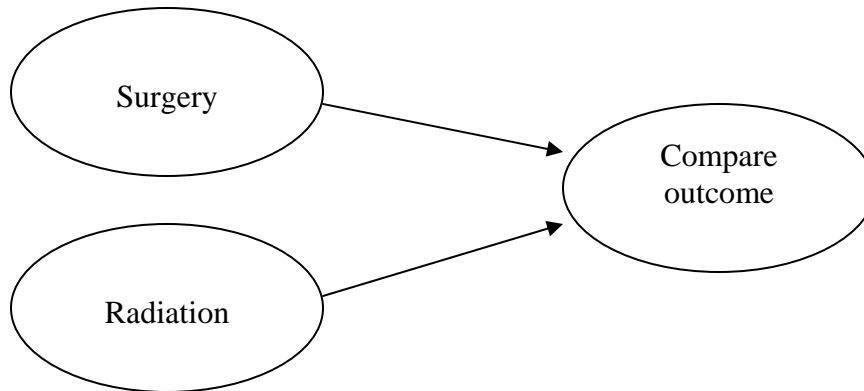
• Interaction



- Exist when the primary relationship of interest between a risk factor and a disease is different at different levels of the interaction factor
- Example of interaction



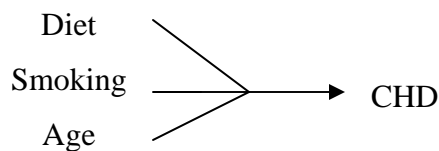
e.g. multivariate analysis



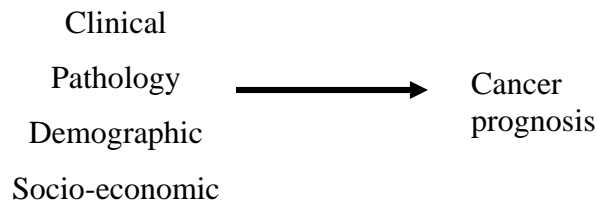
- Are these two groups comparable?
- What are the role covariates?

22. Purpose Multivariate Analysis

- To statistically adjust the effect on variable Y by change in a particular variable x when others are controlled
 - $X_1 \rightarrow Y$ ($X_2, X_3, X_4 \dots$ statistically adjusted for e.g. diet \rightarrow CHD {smoking and age adjusted for})
- To discover the variable X which has the most influence on outcome variable Y



- To predict the outcome Y



- Whole ideas of multivariate analysis are “How to separate independent effect of each X and Y”
- Common multivariate analysis methods in health related sciences research
- Multivariate models
- Modeling strategies

MTV

	Independent Variable	Dependent variable
Multiple linear regression	> 1	1
Multiple logistic regression	> 1	1
Log-linear regression	> 1	1
Survival analysis	> 1	1

<u>Independent variables</u>	<u>Dependent variables</u>	<u>methods</u>
Continuous	Continuous	Multiple linear reg.
Categorical	Categorical	Multiple logistic reg.
Continuous	Categorical	Multiple logistic reg.
Continuous/ categorical	Continuous (survival time)	Survival analysis
Continuous/ categorical	Continuous	Log-linear analysis

23. Multivariate analysis → General Linear Model (GLM)

- The GLM is a flexible statistical model incorporating analysis involving normally distributed dependent variables and combinations of categorical and continuous predictor variables.
- The GLM Univariate model procedure provides regression analysis and analysis of variance one dependent variable by one or more factors or covariates

- The GLM Multivariate model procedure provides regression analysis and analysis of variance for multiple dependent variable by one or more factor or covariates
- The GLM Repeated Measures procedure provides analysis of variance when the same measurement is made several times on each subject or case.

<u>GLM</u>	Independent Variable	Dependent variable
Univariate GLM	≥ 1	1
Multivariate GLM	≥ 1	> 1

24. Repeated measures in categorical outcome

- When the dependent variable is a numerical variable

Independent Variable	Dependent variable	Statistical test
Categorical	Numerical	Repeated measures ANOVA (parametric)
Categorical	Numerical	Friedman test (non-parametric)

- When the dependent variable is a categorical variable

Independent Variable	Dependent variable	Statistical test
Repeated measure 2 measures	2 outcomes categories	Mc Nemar's test
2 measures	3++ outcomes categories	Test of marginal Homogeneity
3++ measures	2 outcomes categories	Cochran's Q test

<p>Repeated measure with independent variables</p>	<p>Binary Ordinal Multiple</p> <p>Count</p>	<p>Cross-sectional time series (xt)</p> <p>} Logistic regression (xt logic)</p> <p>Loglinear regression (xt poisson)</p> <p>General estimating equation (GEE) model (xtgee)</p>
--	---	---