Overview & Concept of

1

Multivariable Analyses

Professor Dr Norsa'adah Bachok
Unit Of Biostatistics & Research Methodology,
School Of Medical Sciences,
Universiti Sains Malaysia

Things to consider when selecting a statistical test

2

- Research questions
- Study design
- Number dependent/independent variables
- Type of variables: categorical/numerical
- Number of groups/categories
- Normality of distribution
- Sample size
- Related samples

Research questions / Study hypotheses / Objectives

3

- Difference of means between groups
- Difference of proportions between groups
- Association between variables
- Relationship between variables
- Correlation between variables
- Effectiveness of an intervention

Study Design
related sample / match / paired / pre post

4

Design	Variables	Test
Independent	Numerical vs categorical	Independent t test
Match case control Pre post (same cases measured twice)	2 Numerical	Paired t test
Match case control Pre post (same cases measured twice)	2 Categorical	Mc Nemar test
Repeatedly measured	Several Numerical	Repeated measure ANOVA

Type of variable for each independent & dependent variables

5

Independent Variable	Dependent Variable	Test
Age (continuous)	Lung cancer (categorical yes & no)	Independent t test
Smoking (categorical yes & no)	Blood cholesterol (continuous)	Independent t test
Smoking (categorical yes & no)	Lung cancer (categorical yes & no)	Chi square test
Age (continuous)	Body mass index (continuous)	Correlation / Linear Regression

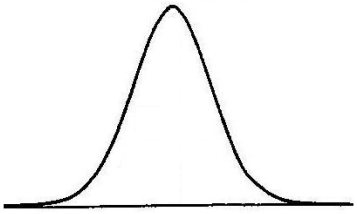
Number of groups

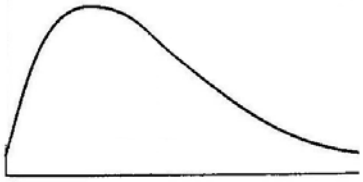
6

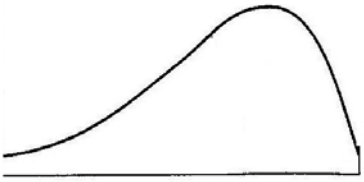
Independent Variable	Dependent Variable	Test
Smoking (categorical) Yes No	Blood cholesterol (continuous)	Independent t test
Smoking (categorical) Currently smoker Ex-smoker Never smoke	Blood cholesterol (continuous)	One way ANOVA

Normal distribution

7







Skewed to right

Skewed to left

Normal distribution

8

Type Of Variable	Parametric Test	Non-parametric Test
2 independent samples Continuous vs categorical 2 levels	Independent t-test	The Mann-Whitney Test / Wilcoxon Rank Sum Test
2 paired samples Continuous Categorical	Paired t-test	The Sign Test The Wilcoxon Signed-Rank Test Mc Nemar Test
>2 independent samples Continuous vs categorical >2 levels	One-way ANOVA	The Kruskal-Wallis Test
Correlation Continuous	Pearson Correlation	Spearman Correlation

Univariate versus multivariable & multivariate

9

Dependent variable	Independent variable	Name
One	One	Univariate
One	Many	multivariable
>one	Many	multivariate

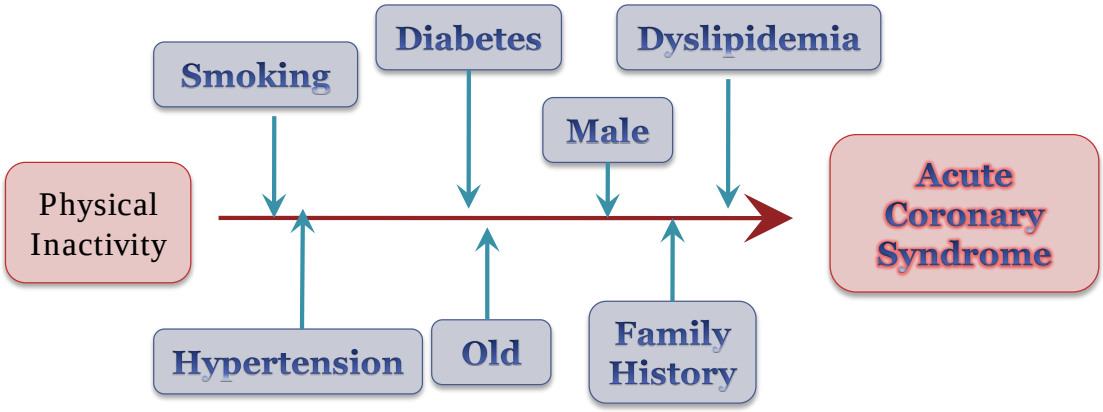
- Univariate analysis:
  - Cannot make conclusion
  - Do not control confounders (bias)

Confounders

- Is a distortion of a exposure-disease relationship brought about by the association of other factors with both exposure and disease.
- Confounding effects can be eliminated when multivariable analysis is carried out.
- Proved by comparing parameter estimates & confidence interval in univariable and multivariable analysis.

Multiple causation of disease

11



Why should aim multivariate analysis?

12

- Reality: not possible one factor causes one outcome
- Many interference of the relationship
- Take account on confounders, covariates, effect modifier & interactions
- Simultaneously assess the impact of multiple independent variables on outcome
- Quality publication

Uses of multivariable models

13

- Identify associated/prognostic factors while adjusting confounders
- Predict the outcome
- Adjust for differences in baseline characteristics
  - Especially when randomization is not possible
- Provide estimation of risks
  - Eg chances of survival on 5 years time
- Determine the best combination of diagnostic information
  - The likelihood of a patient presenting to A&E with chest pain has acute ischaemia

Which multivariable test?

14

- Type of regression depends on type of dependent v.
  - Continuous (linear regression)
  - Binary (logistic regression)
  - Time-to-event (Cox regression)
- Analysis of variance – ANCOVA, MANOVA
  - Dependent v is continuous
  - The aim is to determine mean difference between groups

Multivariate tests

15

No Of Independent Variable	No Of Dependent Variable	Test Example
Many (continuous / Mix)	One numerical	Multiple Linear Regression
Many categorical with covariate	One numerical	Multi-factorial ANCOVA
Many categorical	One numerical	Multi-factorial ANOVA
Many (categorical / continuous / Mix)	One binary categorical	Multiple Logistic Regression
Many (continuous / categorical / Mix)	One ordinal categorical	Multinomial Logistic Regression
Many mix type	>one categorical	Polytomous Logistic Regression
Many mix type	>one numerical	MANCOVA

Examples of statistical modelling

16

Dependent variable	Example	Type
Continuous	Blood pressure, weight, temperature	Multiple Linear Regression
Dichotomous	Present vs absent of disease	Multiple Logistic Regression
Time to occurrence of dichotomous event	Time to death (alive), Time to recurrence of cancer	Proportional Hazards Analysis, Survival Analysis
Ordinal	Stage of cancer (l, ll, lll & lV)	Ordinal Logistic Regression
Nominal	Disease outcome of obesity (cancer, heart disease, osteoarthitis, hpt, diabetes)	Polytomous/ multinomial Logistic Regression

ANCOVA

17

- Evaluates whether the population means on the dependent variable, adjusted for differences on the covariate, differ across levels of one or several factor/s.
- Dependent v: continuous
- Independent v: categorical (one or several, with two or more groups)
  - If the factor has more than 2 levels, need to do post hoc test.
- Covariate: continuous (usually not the main interest of study), used to adjust dependent variable.
- Used to confirm relationship; non exploratory
- There is no model selection, should report all although no significant difference.

MANOVA

18

- Multiple numerical dependent variables: called multivariate.
- Independent variables are factors (categorical) with two or more levels.
- Dependent variables: several numerical variables.
  - Need to be meaningful biological & theoretical.
  - Moderately correlated.
- Need follow up multivariable ANOVA for individual dependent variable, discriminant function analysis, post hoc analysis.
- No variable selection.

Purposes of Regression

19

- Describe association between dependent and indep v
  - As number of cigarette smoked increases, the birth weight of newborn decreases..
  - How much decrease in birth weight of newborn for one cigarette smoked increase?
- Make predictions
  - What is the mean birth weight we would expect if the mother smoked a pack daily?
  - How precise is our estimate of newborn birth weight?
- Adjust or control for confounding variables
  - What happen to the association between maternal smoking and newborn birth weight when adjustment for other factors is done such as age, gender, prenatal care, maternal morbidity etc.

Concept of modelling

20

- Select independent variable in the model by using selection method.
- The goal is to find the best fitting, simplest model possible describing the relationship between an outcome variable and a set of independent variables.
- Independent variables in the model can be statistically non significant but clinically important.

Variable selection technique

21

- Forward
  - Start with empty model. Enter variables into the model sequentially. Starts with the strongest association with the outcome, one by one. Adjustment is done for variables already in the model.
- Backward
  - Start with full model. All variables in the model. Deletes variables from the model sequentially. Starts with the weakest association.
- Both not necessarily produce same model.

Assumptions to be checked

22

- Normality: histogram, residuals
- Equal variance: Levene test, Box's test
- Multicolinearity: SE, CI, VIF
- Linearity: for continuous variables
- Interactions between variables
- Model fitness: chi-square, classification tables, area under ROC curve

Multiple Linear Regression

- Outcome is a CONTINUOUS variable
- All independent variables are numerical
- Mix numerical and categorical → General Linear Regression.
- Assumes association between Y and X is a 'straight line'

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

Multiple Linear Regression

- The mean value of the outcome increases / decreases linearly with multiple independent variable
- What being modelled: the mean value of the outcome
- Results presented as: b coefficient, 95% CI, F test, p value, R2
- Conclusion:
  - There is a significant linear relationship between independent v and dependent v.
  - How much dependent v increases as independent variable increases/decreases (based on b or slope of the regression line).
  - How much the selected model explains the variation of dependent v (based on R^2).

Multiple Logistic Regression

25

- Dependent variable: categorical with 2 groups only eg. Disease: yes, no
- Independent variables can be mix of numerical & categorical.
- The logit of the outcome changes linearly with multiple independent variables
- What being modelled: the logarithm of the odds of the outcome
- Logit = logarithm of the odds of the outcome
- Odds of the outcome = the probability of having divided by the probability of not having the outcome

Multiple Logistic Regression

26

- Coding of dependent variable
  - 1 for high risk
  - 0 for low risk
- Odds Ratio = Odds of a factor among exposed / odds of the factor among non-exposed
- Results presented as: OR, 95% CI, Wald test, p value
- Conclusion:
  - Strength of association, eg. Those who smoked have 5 times more likely to develop lung cancer compared to non-smokers

Cox Proportional Hazards Analysis

27

- Time variable
  - Is an interval for a subject's participation in the study to the date the subject experienced an outcome / lost follow up / withdrawn / completed the study
- Concept of censor
  - Lost to follow up / withdraw
  - Unknown status
  - End of study

Multivariable analysis

28

- Compulsory for an academic thesis / dissertation.
- Depends on journal for manuscript.
  - High impact factor journal usually requires multivariable analysis.
  - Usually no need details checking of assumptions.