STATISTICAL TESTS

By

Adebowale, S.A

UI-MEPI-J: Research Design and Methodology Workshop,

May 10 - 12, 2016

- **Population and Sample**

What is Statistical Inference?

- This involves the procedures of drawing an inference about a population on the basis of the results obtained from a sample drawn from the population.

Uses of Samples in Making Inference

- An epidemiologist wants to know the state of health of a community
- Clinical trials
- Prevalence of a disease
- To identify risk factors of a disease
- Compare blood pressure of alcohol drinkers
- e.t.c

Purpose of Health Research

- For estimation of certain parameters of the population e.g. Physiological variables (blood pressure, pulse rate, heart rate), biological variables, burden of certain conditions, haematological indices (PCV, WBC, haemoglobin.)
- Comparison of certain parameters or phenomena between different populations or the same population at different time points

- Any of the two stated purposes could be achieved through real experiments
- Therefore, there is always need to test some hypotheses

What is a Statistical Hypothesis?

- It is a statement of fact yet to be tested.
- It is a conjecture about the characteristics or phenomena of a population that has not been verified.

- For instance, an endocrinologist may say NIDDM patients are generally younger than IDDM patients.
- Is this true? The answer is either Yes or No
- **The truth will only come out after investigation**

- ***The procedures for investigating the truth in a "stated hypothesis" is called statistical Inference*.**

- ***Types of Hypothesis***

1. Null Hypothesis

It is a statement that forms the basis of investigation in a significant test. It is a statement that does not prejudge the phenomenon of interest.

- ✓ It is a statement of **no difference**
- ✓ It is a statement of **no association**
- ✓ It is a statement of **no effect**
- ✓ It is a statement of **equality**

2. Alternative Hypothesis

- ▪ This is the hypothesis formulated as an alternative to the Null hypothesis which the investigators will accept if the null hypothesis is rejected.
- ▪ It specifies what forms of departure from the null hypothesis are of potential concern.
- ▪ It is a statement **of inequality**

- **A full specification of Hypothesis must be one of the following:**

$$H_0 : \mu = a$$
$$H_1 : \mu < a$$

One sided ⇒ lower tail test

$$H_0 : \mu = a$$
$$H_1 : \mu > a$$

One sided ⇒ Upper tail test

$$H_0 : \mu = a$$
$$H_1 : \mu \neq a$$

Two sided ⇒ Two tail test

Errors in Hypothesis Tests

- **Type I error**

  This is the error committed when the null hypothesis is rejected when in actual fact it is true
- **Type II error**

  This is the error committed when we fail to reject the null hypothesis when in actual fact it is false

Power of a Test

- This is the ability to reject a null hypothesis when it is false.

  It is symbolized by **1-β** where **β** is the type II error

Significance level of a Test

- This is the maximum probability of committing type one error.
- It is symbolized as **α**

Test Statistic

- This is a particular form of data summary to be used in a test.
- Test of mean difference is **t-test**
- Test of difference in proportion, it is **z test**.

The test statistic for t-test is,

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \text{, with } n - 1 \text{ degree of freedom}$$

The test statistic for z-test is,

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

Rejection Region of a Test

- ***This is a set on real number line specified in such a way that whenever the numerical value of the test statistic falls into this region the Ho is rejected and accepted if otherwise***

Steps in Hypothesis Tests

- Write down the two hypotheses
- Determine the appropriate test statistic to be used
- Determine the significance level of the test
- Decide on the distribution of the test statistic and the sidedness of the test i.e whether single or double
- Decide the boundary or boundaries of the rejection region
- Give the decision rule

Sampling Distributions

- What is sampling distribution?
- The sampling distribution can be computed for the mean, median, standard deviation, proportion and any other statistic can be computed
- Parameter and Estimate

Central limit theorem

- *This states that the mean of the sample means is the unbiased estimate of the measure of location or the true population parameter*

Note the following Properties:

- ✓ sampling distributions are approximately normally distributed regardless of the underlying population distribution of the variable
- ✓ The mean of the sampling distribution is equal to the true population mean
- ✓ The standard deviation of the sampling distribution is directly proportional to the population standard deviation and inversely proportional to the square root of the sample size (n)

Standard Error

- The standard deviation of the sampling distribution is called the standard error.
- It is the measure of precision of the estimate i.e the price we pay for taking a sample
- The larger the sample size, the less the error in its estimate

- The sampling variability of the individual observation is measured by **standard deviation** (σ)

Whereas;

- The sampling variability of the sampling means is measured by the **standard error of the mean** $\left(\frac{\sigma}{\sqrt{n}}\right)$

Standard Error of the Mean

For single sample;

$$\mathrm{S.E}(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

For two independent samples

$$\mathrm{S.E}(\bar{x}_1 - \bar{x}_2) = \mathrm{S.E}(\bar{x}_1 + \bar{x}_2) = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Standard Error of the Proportion

- For single sample;

$$\mathrm{S.E}(\mathrm{p}) = \sqrt{\frac{\mathrm{p}(1-\mathrm{p})}{\mathrm{n}}}$$

- For two independent samples

$$\mathrm{S.E}(\mathrm{p}_1 + \mathrm{p}_2) = \mathrm{S.E}(\mathrm{p}_1 - \mathrm{p}_2) = \sqrt{\frac{\mathrm{p}_1(1-\mathrm{p}_1)}{\mathrm{n}_1} + \frac{\mathrm{p}_2(1-\mathrm{p}_2)}{\mathrm{n}_2}}$$

The estimation of population parameters

- *In an attempt to seek for the reliability of point estimates obtained for population parameters, researchers are more comfortable with interval estimates taking good advantage of the central limit theorem on sample estimates having normal distributions.*

- *It enables the researchers to build a “**confidence interval”** at appropriate probability levels for any single estimate of a population parameter*

- ***Confidence interval!***

  Confidence interval!!

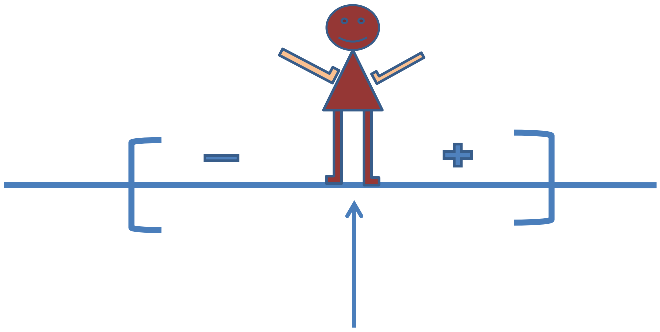
  Confidence interval!!!

  Confidence interval!!!!

  Confidence interval!!!!!

  Confidence interval!!!!!!!

Look at this picture;



From;

Study conception

to

final Data analysis

Reliability Analysis

- Reliability analysis allows you to study the properties of measurement scales and the items that compose the scales.
- It calculates a number of commonly used measures of scale reliability and also provides information about the relationships between individual items in the scale.
- Intraclass correlation coefficients can be used to compute inter-rater reliability estimates.

Example:

- Does my questionnaire measure customer satisfaction in a useful way?
- Using reliability analysis, you can determine the extent to which the items in your questionnaire are related to each other,
- you can get an overall index of the repeatability or internal consistency of the scale as a whole,
- you can identify problem items that should be excluded from the scale.

Statistics

- Descriptives for each variable and for the scale, summary statistics across items, inter-item correlations and covariances, reliability estimates, ANOVA table, intraclass correlation coefficients, Hotelling's T2, and Tukey's test of additivity.
- **Data** can be dichotomous, ordinal, or interval, but the data should be coded numerically.

- If you want to explore the dimensionality of your scale items to see whether more than one construct is needed to account for the pattern of item scores,
- **use factor analysis or multidimensional scaling.**

- To identify homogeneous groups of variables,
- **use hierarchical cluster analysis to cluster variables.**

Frequencies

- The Frequencies procedure provides statistics and graphical displays that are useful for describing many types of variables.
- It is a good place to start looking at your data.
- For a frequency report and bar chart, you can arrange the distinct values in ascending or descending order, or you can order the categories by their frequencies.
- The frequencies report can be suppressed when a variable has many distinct values.
- You can label charts with frequencies (the default) or percentages.

Frequencies contd.

- Statistics and plots. Frequency counts, percentages, cumulative percentages, mean, median, mode, sum, standard deviation, variance, range, minimum and maximum values, standard error of the mean, skewness and kurtosis (both with standard errors), quartiles, user-specified percentiles, bar charts, pie charts, and histograms.
- **Data:** Use numeric codes or short strings to code categorical variables (nominal or ordinal level measurements).

Frequencies contd.

- The tabulations and percentages provide a useful description for data from any distribution, especially for variables with ordered or unordered categories.
- Most of the optional summary statistics, such as the mean and standard deviation, are based on normal theory and are appropriate for quantitative variables with symmetric distributions.
- **Robust statistics,** such as the median, quartiles, and percentiles, are appropriate for quantitative variables that may or may not meet the assumption of normality.

Descriptive

- The Descriptives procedure displays univariate summary statistics for several variables in a single table and calculates standardized values (z scores).
- Variables can be ordered by the size of their means (in ascending or descending order), alphabetically, or by the order in which you select the variables (the default).

Descriptive (Contd.)

- **Statistics:** Sample size, mean, minimum, maximum, standard deviation, variance, range, sum, standard error of the mean, and kurtosis and skewness with their standard errors.

Crosstabs

- The Crosstabs procedure forms two-way and multi-way tables and provides a variety of tests and measures of association for two-way tables. The structure of the table and whether categories are ordered determine what test or measure to use.
- For example, if gender is a layer factor for a table of married (yes, no) against life (is life exciting, routine, or dull).

Crosstabs (Contd.)

- **Statistics and measures of association:** Pearson chi-square, likelihood-ratio Chi-square, linear-by-linear association test, Fisher's exact test, Yates' corrected chi-square, Pearson's r, Spearman's rho, contingency coefficient, phi, Cramér's V, symmetric and asymmetric lambdas, Goodman and Kruskal's tau, uncertainty coefficient, gamma, Somers' d, Kendall's tau-b, Kendall's tau-c, eta coefficient, Cohen's kappa, **relative risk estimate**, **odds ratio**, **McNemar test**, and Cochran's and **Mantel-Haenszel statistics**.

Crosstabs (Contd.)

- To define the categories of each table variable, use values of a numeric or short string (eight or fewer characters) variable.
- For example, for gender, you could code the data as 1 and 2 or as male and female.

Crosstabs (Contd.)

- **Note:** Ordinal variables can be either numeric codes that represent categories (for example, 1 = low, 2 = medium, 3 = high) or string values.
- The alphabetic order of string values must reflect the true order of the categories.
- For example, for a string variable with the values of low, medium, high, the order of the categories is interpreted as high, low, medium--which is not the correct order.
- In general, it is more reliable to use numeric codes to represent ordinal data.

Ratio Statistics

- The Ratio Statistics procedure provides a comprehensive list of summary statistics for describing the ratio between two scale variables.
- You can sort the output by values of a grouping variable in ascending or descending order.

- ***Example:*** *Is there good uniformity in the ratio between the appraisal price and sale price of homes in each of five counties? From the output, you might learn that the distribution of ratios varies considerably from county to county.*

Ratio Statistics (Contd.)

- **Statistics:** Median, mean, weighted mean, confidence intervals, coefficient of dispersion (COD), median-centered coefficient of variation, mean-centered coefficient of variation, price-related differential (PRD), standard deviation, average absolute deviation (AAD), range, minimum and maximum values, and the concentration index computed for a user-specified range or percentage within the median ratio.
- Use numeric codes or short strings to code grouping variables (nominal or ordinal level measurements).

Choosing a Statistical Test

- In terms of selecting a statistical test, the most important question is "what is the main study hypothesis?" In some cases there is no hypothesis; the investigator just wants to "see what is there".

- For example, in a prevalence study there is no hypothesis to test, and the size of the study is determined by how accurately the investigator wants to determine the prevalence.

- If there is no hypothesis, then there is no statistical test.

Comparing Means

- The Means procedure calculates subgroup means and related univariate statistics for dependent variables within categories of one or more independent variables.
- Optionally, you can obtain a one-way analysis of variance, eta, and tests for linearity.

Example:

- Measure the average amount of fat absorbed by three different types of cooking oil, and perform a one-way analysis of variance to see whether the means differ.

Comparing Means (Contd.)

Data consideration:

- The dependent variables are quantitative, and the independent variables are categorical. The values of categorical variables can be numeric or short string.
- Some of the optional subgroup statistics, such as the mean and standard deviation, are based on normal theory and are appropriate for quantitative variables with symmetric distributions.
- **Robust statistics,** such as the median, are appropriate for quantitative variables that may or may not meet the assumption of normality. Analysis of variance is robust to departures from normality, but the data in each cell should be symmetric.

Independent-Samples T Test

- The Independent-Samples T Test procedure compares means for two groups of cases.
- Ideally, for this test, the subjects should be randomly assigned to two groups, so that any difference in response is due to the treatment (or lack of treatment) and not to other factors.
- This is not the case if you compare average income for males and females. A person is not randomly assigned to be a male or female.
- In such situations, you should ensure that differences in other factors are not masking or enhancing a significant difference in means.

Independent-Samples T Test (Contd.)

Example:

Patients with high blood pressure are randomly assigned to a placebo group and a treatment group. The placebo subjects receive an inactive pill, and the treatment subjects receive a new drug that is expected to lower blood pressure. After the subjects have been treated for two months, the two-sample t test is used to compare the average blood pressures for the placebo group and the treatment group. Each patient is measured once and belongs to one group.

Independent-Samples T Test (Contd.)

Statistics

- For each variable: sample size, mean, standard deviation, and standard error of the mean. For the difference in means: mean, standard error, and confidence interval (you can specify the confidence level).

- Tests: Levene's test for equality of variances, and both pooled-variances and separate-variances t tests for equality of means.

Independent-Samples T Test (Contd.)

Data Consideration:

- The values of the quantitative variable of interest are in a single column in the data file.
- The grouping variable can be numeric (values such as 1 and 2 or 6.25 and 12.5) or short string (such as yes and no).
- As an alternative, you can use a quantitative variable, such as age, to split the cases into two groups by specifying a cut point (cut point 21 splits age into an under-21 group and a 21-and-over group).

Paired-Samples T Test:

- The Paired-Samples T Test procedure compares the means of two variables for a single group.
- The procedure computes the differences between values of the two variables for each case and tests whether the average differs from 0.

Paired-Samples T Test (Contd.)

- **Example:** In a study on high blood pressure, all patients are measured at the beginning of the study, given a treatment, and measured again. Thus, each subject has two measures, often called before and after measures.

- An alternative design for which this test is used is a **matched-pairs or case-control study**, in which each record in the data file contains the response for the patient and also for his or her matched control subject.

Paired-Samples T Test (Contd.)

Data:

- For each paired test, specify two quantitative variables (interval level of measurement or ratio level of measurement).
- For a matched-pairs or case-control study, the response for each test subject and its matched control subject must be in the same case in the data file.

Assumptions:

- Observations for each pair should be made under the same conditions. The mean differences should be normally distributed. Variances of each variable can be equal or unequal.

One-Way ANOVA

- The One-Way ANOVA procedure produces a one-way analysis of variance for a quantitative dependent variable by a single factor (independent) variable. ANOVA is used to test the hypothesis that several means are equal. This technique is an extension of the two-sample t test.
- In addition to determining that differences exist among the means, you may want to know which means differ. There are two types of tests for comparing means: **a priori contrasts and post hoc tests.**
- Contrasts are tests set up before running the experiment, and post hoc tests are run after the experiment has been conducted. You can also test for trends across categories.

One-Way ANOVA (Contd.)

Example:

- Doughnuts absorb fat in various amounts when they are cooked. An experiment is set up involving three types of fat: peanut oil, corn oil, and lard. Peanut oil and corn oil are unsaturated fats, and lard is a saturated fat. Along with determining whether the amount of fat absorbed depends on the type of fat used, you could set up an a priori contrast to determine whether the amount of fat absorption differs for saturated and unsaturated fats.

One-Way ANOVA (Contd.)

Statistics:

- For each group: number of cases, mean, standard deviation, standard error of the mean, minimum, maximum, and 95%-confidence interval for the mean.
- Levene's test for homogeneity of variance, analysis-of-variance table and robust tests of the equality of means for each dependent variable, user-specified a priori contrasts, and post hoc range tests and multiple comparisons:

  Bonferroni, Sidak, Tukey's honestly significant difference, Hochberg's GT2, Gabriel, Dunnett, Ryan-Einot-Gabriel-Welsch F test (R-E-G-W F), Ryan-Einot-Gabriel-Welsch range test (R-E-G-W Q), Tamhane's T2, Dunnett's T3, Games-Howell, Dunnett's C, Duncan's multiple range test, Student-Newman-Keuls (S-N-K), Tukey's b, Waller-Duncan, Scheffé, and least-significant difference.

One-Way ANOVA (Contd.)

Data:

- Factor variable values should be integers, and the dependent variable should be quantitative (interval level of measurement).
- Assumptions.
- Each group is an independent random sample from a normal population.
- ANOVA is robust to departures from normality, although the data should be symmetric.
- The groups should come from populations with equal variances. To test this assumption, use Levene's homogeneity-of-variance test.

Linear Regression

- Linear Regression estimates the coefficients of the linear equation, involving one or more independent variables, that best predict the value of the dependent variable.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon_i$$

Linear Regression (Contd.)

Statistics:

- For each variable: number of valid cases, mean, and standard deviation.
- For each model: regression coefficients, correlation matrix, part and partial correlations, multiple R, R2, adjusted R2, change in R2, standard error of the estimate, analysis-of-variance table, predicted values, and residuals.
- Also, 95%-confidence intervals for each regression coefficient, variance-covariance matrix, variance inflation factor, tolerance, Durbin-Watson test, distance measures (Mahalanobis, Cook, and leverage values), DfBeta, DfFit, prediction intervals, and casewise diagnostics.
- Plots: scatterplots, partial plots, histograms, and normal probability plots.

Linear Regression (Contd.)

Data:

- The dependent and independent variables should be quantitative. Categorical variables, such as religion, major field of study, or region of residence, need to be recoded to binary (dummy) variables or other types of contrast variables.

Assumptions:

- For each value of the independent variable, the distribution of the dependent variable must be normal.
- The variance of the distribution of the dependent variable should be constant for all values of the independent variable.
- The relationship between the dependent variable and each independent variable should be linear, and all observations should be independent.

Logistic Regression

- It is useful for situations in which you want to be able to predict the presence or absence of a characteristic or outcome based on values of a set of predictor variables.
- It is similar to a linear regression model but is suited to models where the dependent variable is dichotomous.
- Its coefficients can be used to estimate odds ratios for each of the independent variables in the model.
- It is applicable to a broader range of research situations than discriminant analysis.

Logistic Regression (Contd.)

Example:

- What lifestyle characteristics are risk factors for coronary heart disease (CHD)? Given a sample of patients measured on smoking status, diet, exercise, alcohol use, and CHD status, you could build a model using the four lifestyle variables to predict the presence or absence of CHD in a sample of patients.
- The model is then be used to derive estimates of the odds ratios for each factor to tell you, for example, how much more likely smokers are to develop CHD than non-smokers.

Logistic Regression (Contd.)

Statistics:

- For each categorical variable: parameter coding. For each step: variable(s) entered or removed, iteration history, –2 log-likelihood, goodness of fit, Hosmer-Lemeshow goodness-of-fit statistic, model chi-square
- For each variable in the equation: coefficient (B), standard error of B, Wald statistic, estimated odds ratio (exp(B)), confidence interval for exp(B), log-likelihood if term removed from the model.

Logistic Regression (Contd.)

- Logistic regression solution may be more stable if your predictors have a multivariate normal distribution.
- Additionally, as with other forms of regression, multi-collinearity among the predictors can lead to biased estimates and inflated standard errors.
- The procedure is most effective when group membership is a truly categorical variable
- If group membership is based on values of a continuous variable (for example, "high IQ" versus "low IQ"), you should consider using linear regression to take advantage of the richer information offered by the continuous variable itself.

Multinomial Logistic Regression

- Multinomial Logistic Regression is useful for situations in which you want to be able to classify subjects based on values of a set of predictor variables.
- This type of regression is similar to logistic regression, but it is more general because the dependent variable is not restricted to two categories.

Multinomial Logistic Regression (Contd.)

Methods:

- A multinomial logit model is fit for the full factorial model or a user-specified model. Parameter estimation is performed through an iterative maximum-likelihood algorithm.

Data:

- The dependent variable should be categorical. Independent variables can be factors or covariates. In general, factors should be categorical variables and covariates should be continuous variables.

Probit Analysis

- This procedure measures the relationship between the strength of a stimulus and the proportion of cases exhibiting a certain response to the stimulus.
- It is useful for situations where you have a dichotomous output that is thought to be influenced or caused by levels of some independent variable(s) and is particularly well suited to experimental data.
- This procedure will allow you to estimate the strength of a stimulus required to induce a certain proportion of responses, such as the median effective dose.

Probit Analysis (Contd):

Example:

- How effective is a new pesticide at killing ants, and what is an appropriate concentration to use?
- You might perform an experiment in which you expose samples of ants to different concentrations of the pesticide and then record the number of ants killed and the number of ants exposed.
- Applying probit analysis to these data, you can determine the strength of the relationship between concentration and killing.
- you can determine what the appropriate concentration of pesticide would be if you wanted to be sure to kill, say, 95% of exposed ants.

Probit Analysis (Contd):

Data:

- For each value of the independent variable (or each combination of values for multiple independent variables), your response variable should be a count of the number of cases with those values that show the response of interest, and the total observed variable should be a count of the total number of cases with those values for the independent variable.

- The factor variable should be categorical, coded as integers.

Probit Analysis (Contd):

- The probit analysis procedure reports estimates of effective values for various rates of response (including median effective dose), while the logistic regression procedure reports estimates of odds ratios for independent variables.

Chi-Square Test

- The Chi-Square Test procedure tabulates a variable into categories and computes a chi-square statistic.
- Association
- This goodness-of-fit test compares the observed and expected frequencies in each category to test that all categories contain the same proportion of values or test that each category contains a user-specified proportion of values.

Chi-Square Test (Contd.)

Data

- Use ordered or unordered numeric categorical variables (ordinal or nominal levels of measurement).
- To convert string variables to numeric variables, use the Automatic Recode procedure, which is available on the Transform menu.

Assumptions.

- Nonparametric tests do not require assumptions about the shape of the underlying distribution.
- The data are assumed to be a random sample.
- The expected frequencies for each category should be at least 1.
- No more than 20% of the categories should have expected frequencies of less than 5

Survival Data
or
Follow-up Studies

Life Table

- The basic idea of the life table is to subdivide the period of observation into smaller time intervals.
- For each interval, all people who have been observed at least that long are used to calculate the probability of a terminal event occurring in that interval.
- The probabilities estimated from each of the intervals are then used to estimate the overall probability of the event occurring at different time points.

Life Table (Contd.)

Example:

- Is a new nicotine patch therapy better than traditional patch therapy in helping people to quit smoking?
- You could conduct a study using two groups of smokers, one of which received the traditional therapy and the other of which received the experimental therapy.
- Constructing life tables from the data would allow you to compare overall abstinence rates between the two groups to determine if the experimental treatment is an improvement over the traditional therapy.
- You can also plot the survival or hazard functions and compare them visually for more detailed information.

Life Table (Contd.)

Statistics

- Number entering, number leaving, number exposed to risk, number of terminal events, proportion terminating, proportion surviving, cumulative proportion surviving (and standard error), probability density (and standard error), and hazard rate (and standard error) for each time interval for each group; median survival time for each group; and Wilcoxon (Gehan) test for comparing survival distributions between groups.
- **Plots:** function plots for survival, log survival, density, hazard rate, and one minus survival.

Life Table (Contd.)

Assumptions:

- Probabilities for the event of interest should depend only on time after the initial event--they are assumed to be stable with respect to absolute time.
- That is, cases that enter the study at different times (for example, patients who begin treatment at different times) should behave similarly.
- There should also be no systematic differences between censored and uncensored cases. If, for example, many of the censored cases are patients with more serious conditions, your results may be biased.

Kaplan-Meier Survival Analysis

- The Kaplan-Meier procedure is a method of estimating time-to-event models in the presence of censored cases.
- The Kaplan-Meier model is based on estimating conditional probabilities at each time point when an event occurs and taking the product limit of those probabilities to estimate the survival rate at each point in time.

Kaplan-Meier Survival Analysis (Contd.)

Example:

- Does a new treatment for AIDS have any therapeutic benefit in extending life?
- You could conduct a study using two groups of AIDS patients, one receiving traditional therapy and the other receiving the experimental treatment.
- Constructing a Kaplan-Meier model from the data would allow you to compare overall survival rates between the two groups to determine whether the experimental treatment is an improvement over the traditional therapy.
- You can also plot the survival or hazard functions and compare them visually for more detailed information.

Kaplan-Meier Survival Analysis (Contd.)

Statistics:

- Survival table, including time, status, cumulative survival and standard error, cumulative events, and number remaining; and mean and median survival time, with standard error and 95% confidence interval. Plots: survival, hazard, log survival, and one minus survival.

Data:

- The time variable should be continuous, the status variable can be categorical or continuous, and the factor and strata variables should be categorical.

Kaplan-Meier Survival Analysis (Contd.)

Assumptions:

- Probabilities for the event of interest should depend only on time after the initial event--they are assumed to be stable with respect to absolute time.
- That is, cases that enter the study at different times (for example, patients who begin treatment at different times) should behave similarly.
- There should also be no systematic differences between censored and uncensored cases.
- If, for example, many of the censored cases are patients with more serious conditions, your results may be biased.

Cox Regression Analysis

- Like Life Tables and Kaplan-Meier survival analysis, Cox Regression is a method for modelling time-to-event data in the presence of censored cases.
- It allows inclusion of predictor variables (covariates) in your models.
- **For example**, you could construct a model of length of employment based on educational level and job category.
- Cox Regression will handle the censored cases correctly, and it will provide estimated coefficients for each of the covariates, allowing you to assess the impact of multiple covariates in the same model.
- You can also use Cox Regression to examine the effect of continuous covariates.

Cox Regression Analysis (Contd.)

Example:

- Do men and women have different risks of developing lung cancer based on cigarette smoking?
- By constructing a Cox Regression model, with cigarette usage (cigarettes smoked per day) and gender entered as covariates, you can test hypotheses regarding the effects of gender and cigarette usage on time-to-onset for lung cancer.

Statistics:

- For each model: –2LL, the likelihood-ratio statistic, and the overall chi-square.
- For variables in the model: parameter estimates, standard errors, and Wald statistics.
- For variables not in the model: score statistics and residual chi-square.

Cox Regression Analysis (Contd.)

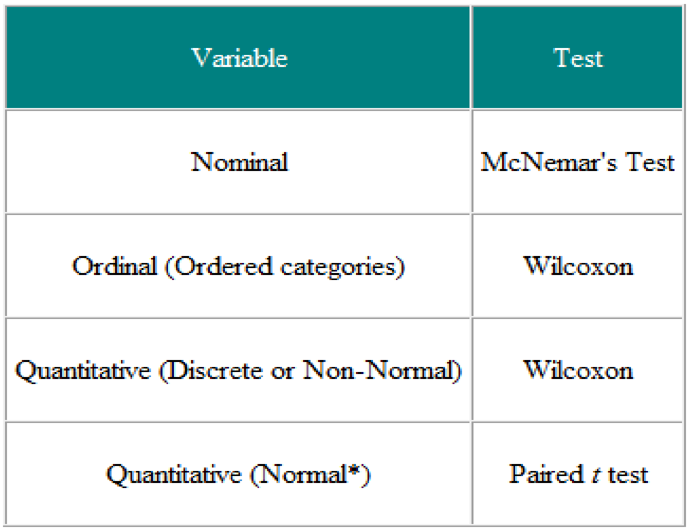
Data:

- Your time variable should be quantitative and your status variable can be categorical or continuous.
- Independent variables (covariates) can be continuous or categorical; if categorical, they should be dummy- or indicator-coded (there is an option in the procedure to recode categorical variables automatically).
- Strata variables should be categorical, coded as integers or short strings.

Assumptions:

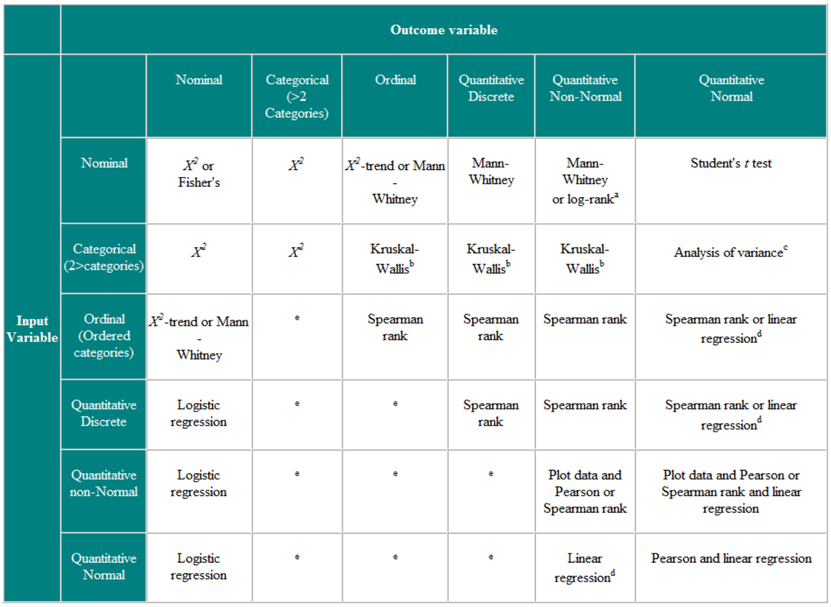
- Observations should be independent, and the hazard ratio should be constant across time; that is, the proportionality of hazards from one case to another should not vary over time. The latter assumption is known as the proportional hazards assumption.

Choice of statistical test from paired or matched observation



Variable	Test
Nominal	McNemar's Test
Ordinal (Ordered categories)	Wilcoxon
Quantitative (Discrete or Non-Normal)	Wilcoxon
Quantitative (Normal*)	Paired *t* test

Choice of statistical test for independent observations



		Outcome variable					
Input Variable		Nominal	Categorical (>2 Categories)	Ordinal	Quantitative Discrete	Quantitative Non-Normal	Quantitative Normal
	Nominal	X^2 or Fisher's	X^2	X^2-trend or Mann - Whitney	Mann-Whitney	Mann-Whitney or log-rank[a]	Student's *t* test
	Categorical (2>categories)	X^2	X^2	Kruskal-Wallis[b]	Kruskal-Wallis[b]	Kruskal-Wallis[b]	Analysis of variance[c]
	Ordinal (Ordered categories)	X^2-trend or Mann - Whitney	[e]	Spearman rank	Spearman rank	Spearman rank	Spearman rank or linear regression[d]
	Quantitative Discrete	Logistic regression	[e]	[e]	Spearman rank	Spearman rank	Spearman rank or linear regression[d]
	Quantitative non-Normal	Logistic regression	[e]	[e]	[e]	Plot data and Pearson or Spearman rank	Plot data and Pearson or Spearman rank and linear regression
	Quantitative Normal	Logistic regression	[e]	[e]	[e]	Linear regression[d]	Pearson and linear regression

- **Thank you for listening**