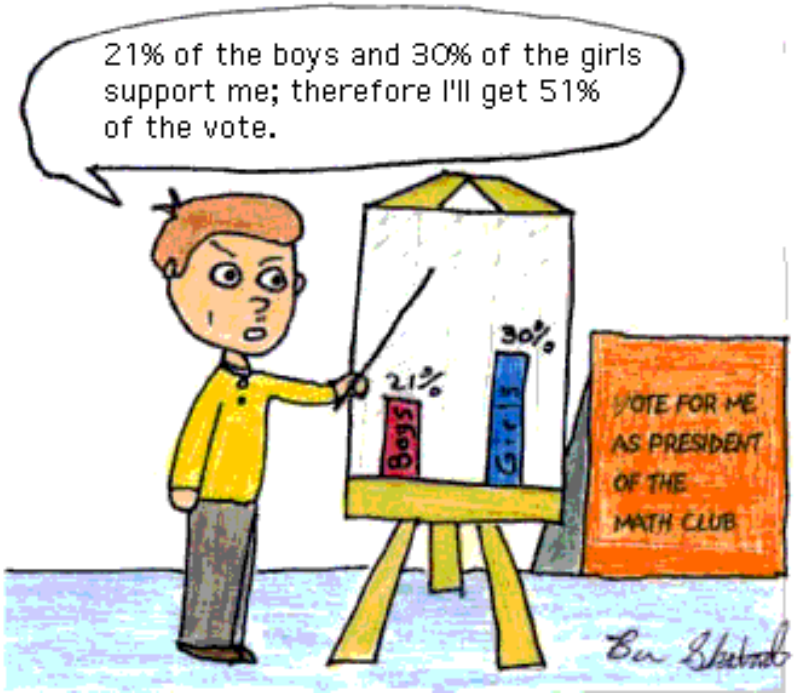
Introduction to Statistics and Quantitative Research Methods

fraserhealth

Better health.
Best in health care.

Purpose of Presentation

- To aid in the understanding of basic statistics, including terminology, common terms, and common statistical methods.
- To help those interested in research feel more comfortable with statistics.
- To encourage potential researchers to undertake research projects to facilitate the production of knowledge.



Statistics Defined

- Statistics is the science and practice of developing human knowledge through the use of empirical data expressed in quantitative form. It is based on statistical theory which is a branch of applied mathematics. Within statistical theory, randomness and uncertainty are modelled by probability theory (Wikipedia Encyclopedia).

What is statistics?

- The collecting, summarizing, and analyzing of data.
- The term also refers to raw numbers, or “stats”, and to the summarization of data.
- Example: Frequencies

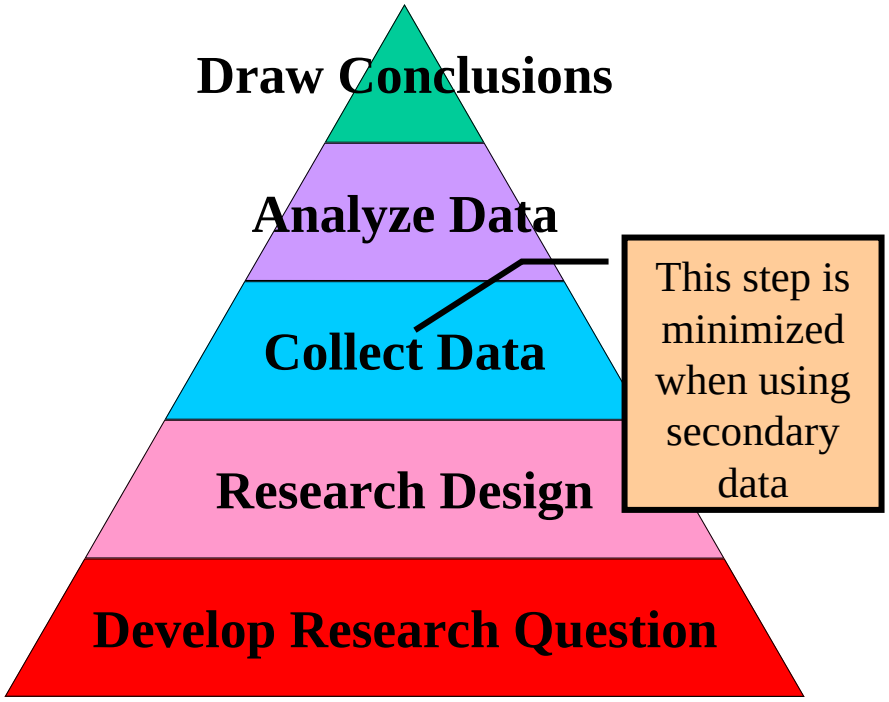
Statistics humour

- Why is a physician held in much higher esteem than a statistician?

- A physician makes an analysis of a complex illness whereas a statistician makes you ill with a complex analysis!

Research Methods

- Research is structural. There are basic steps depending on the subject matter and researcher.
- It is also possible to conduct research using pre-collected data, this is called secondary data analysis. There are many advantages to using secondary data, and Fraser Health has a large number of data sets available for analysis.



Basic Steps

- The following are the basic steps of most research.

- 1) Develop a research question
- 2) Conduct thorough literature review
- 3) Re-define research question → hypothesis
- 4) Design research methodology/study
- 5) Create research proposal
- 6) Apply for funding
- 7) Apply for ethics approval
- 8) Collect and analyze data
- 9) Draw conclusions and relate findings

Research begins when there is a question.

- Different kinds of questions:

Descriptive:

How many men work at Fraser Health?

How many hours a week do employees spend at their desks?

Inferential:

Does having a science degree help students learn statistical concepts?

What risk factors most predict heart disease?

Types of Statistics

- Descriptive Statistics: describe the relationship between variables.
  - E.g. Frequencies, means, standard deviation

- Inferential Statistics: make inferences about the population, based on a random sample.

Variables

- In research, the characteristic or phenomenon that can be measured or classified is called a variable. There are 4 levels of variables:
  - Nominal
  - Ordinal
  - Interval
  - Ratio

Levels of Data

- Nominal= categorical
- E.g. Apples and pears, gender, eye colour, ethnicity.
- Data that is classified into categories and cannot be arranged in any particular order.
  - Nominal=Categorical=Dichotomous

- Ordinal= data ordered, but distance between intervals not always equal. E.g. Low, middle and high income, or rating a brand of soft drink on a scale of 1-5.

- Interval= equal distance between each interval. E.g. 1,2,3. Arbitrary zero point (ex. Fahrenheit scale for temperature - temperature does not cease to exist at 0 degrees.

- Ratio= similar to interval scale, but has true zero point E.g. Weight, salary ($0=$0).

Types of Variables

- Variables can be classified as independent or dependent.
- An independent variable is the variable that you believe will influence your outcome measure.
- A dependent variable is the variable that is dependent on or influenced by the independent variable(s). A dependent variable may also be the variable you are trying to predict.

Types of Variables

- An intervening variable is the variable that links the independent and dependent variable

Independent Variable → Intervening variable → Dependent variable

E.g. Educational level → Occupational type → Income level

- A confounding variable is a variable that has many other variables, or dimensions built into it.
- Not sure what it contains or measures.
- For example: Socio Economic Status (SES)
  - How can we measure SES? Income, Employment status, etc.
  - Need to be careful when using confounding variables…

Example

A researcher wants to study the effect of Vitamin C on cancer.

Vitamin C would be the independent variable because it is hypothesized that it will have an affect on cancer, and cancer would be the dependent variable because it is the variable that may be influenced by Vitamin C .

Independent Variable	→ Direction of Affect	→ Dependent Variable
Vitamin C	→ Increase or decrease of certain affect	→ Cancer

5 minute group exercise

3 Questions:

For each question:

- What is the dependent variable in this study?
- What is the independent variable?
- What is the level of data?

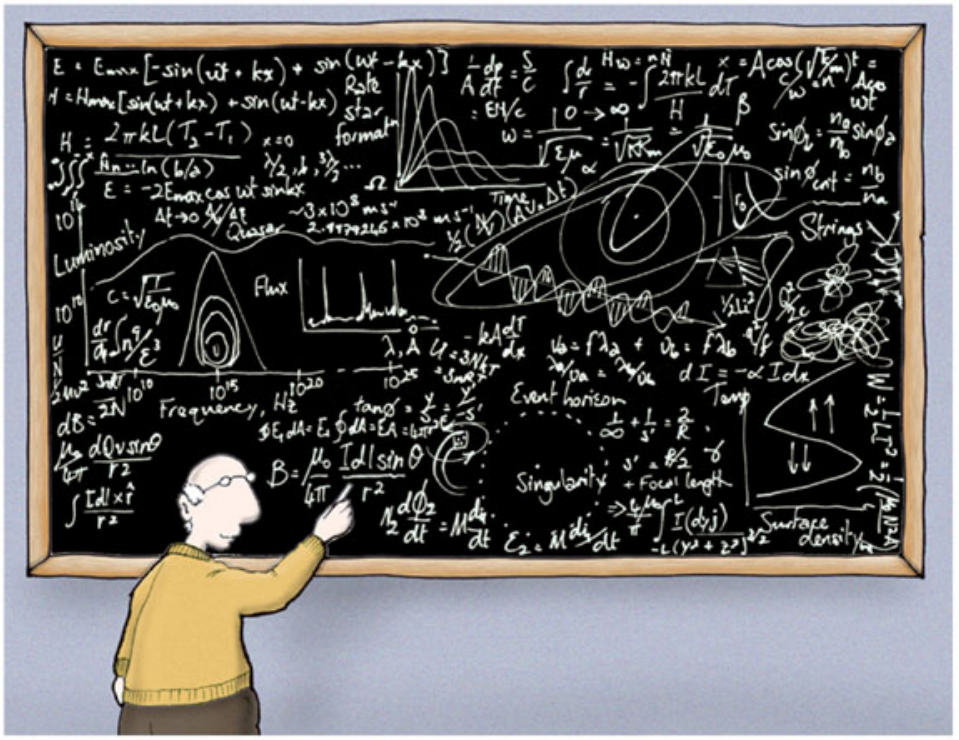
5 minute group exercise

1) Researcher Purple wants to examine if a women's consumption of calcium is related to large foot size. Calcium is measured in milligrams, and foot size is measured in centimetres. Researcher Purple hypothesizes that calcium affects foot size.

2) Researcher Orange wants to know if a man's consumption of orange juice is related to an increase in male pattern baldness. Consumption of orange juice is measured in millilitres, and male pattern baldness is measured on a scale of 1-3 (1=totally bald, 2=some balding, 3=no balding). Researcher Orange hypothesizes that orange juice affects male pattern baldness.

3) Researcher Blue wants to know if pet type has a relationship with happiness. Pet type is measured on a scale of 1-5 (1=cat, 2=dog, 3=bird, 4=fish, 5=other). Happiness is measured on a scale of 1-3 (1=not happy, 2=somewhat happy, 3=very happy). Researcher Blue hypothesizes that pet type will affect level of happiness.

Back to stats.....

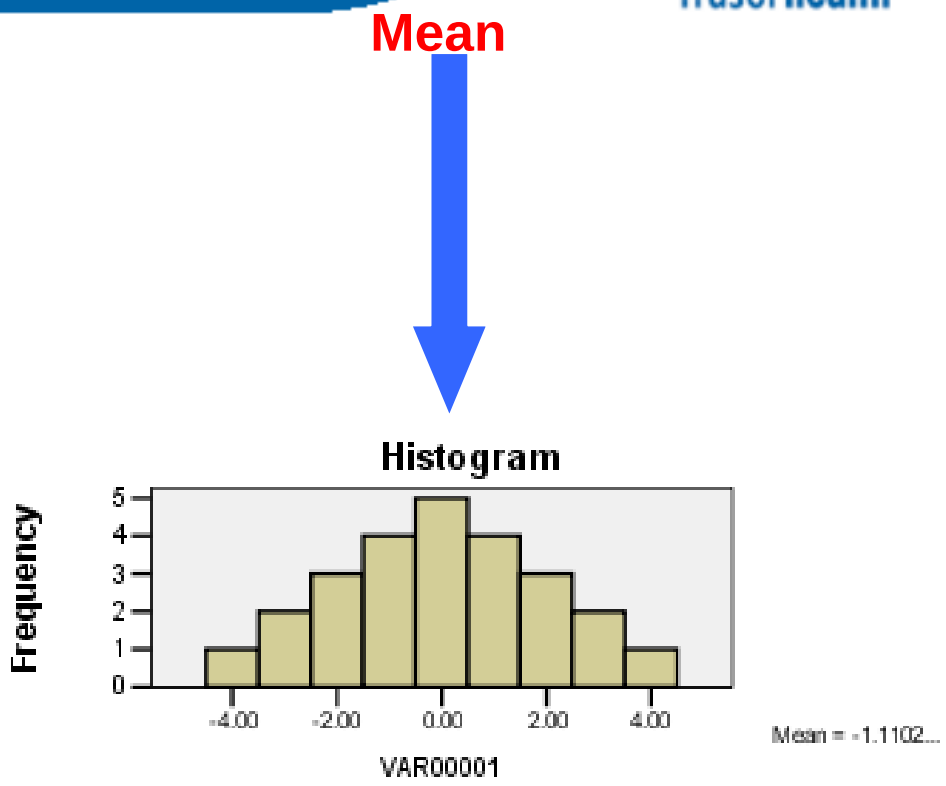


Statistics made simple…

Descriptive Statistics Defined

What is a mean?

- The sum of all the scores divided by the number of scores.
- Often referred to as the average.
- Good measure of central tendency.
- Central tendency is simply the location of the middle in a distribution of scores.



The Mean

"A statistician is someone who can have his head in an oven and his feet in ice, and say that on the average he feels great."

- The mean can be misleading because it can be greatly influenced by extreme scores (very high, or very low scores).
- For example, the average length of stay at a hospital could be greatly influenced by one patient that stays for 5 years.
- Extreme cases or values are called outliers.
- Sometimes the median may yield more information when your distribution contains outliers, or is skewed (not normally distributed).
- What is a median?

Median

A median is the middle of a distribution.

- Half the scores are above the median and half are below the median.
- How do I compute the median?

    - If there is an odd number of numbers, the median is the middle number. For example, the median of 5, 8, and 11 is 8.
    - If there is an even number of numbers, the median is the mean of the two middle numbers. The median of the numbers 4, 8, 9, 13 is (8+9)/2 =8.5.

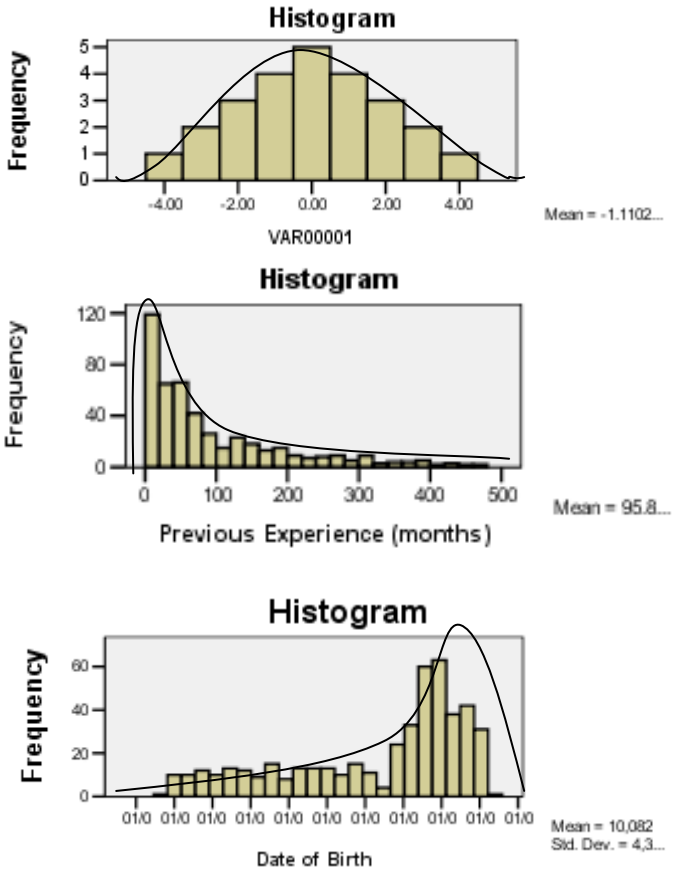
Mode

What is a mode?

- Most frequently occurring score in a distribution.
- Greatly subject to sample fluctuations (statistic takes on different values with different samples).
- Not recommended as the only measure of central tendency.
- Distributions can have more than one mode, called "multimodal."
- Conclusion: Examine your data in order to determine what descriptive statistic is appropriate.

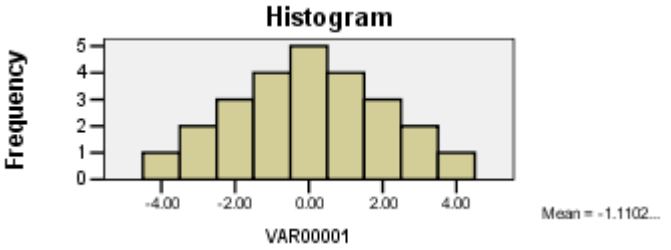
Skewed distributions

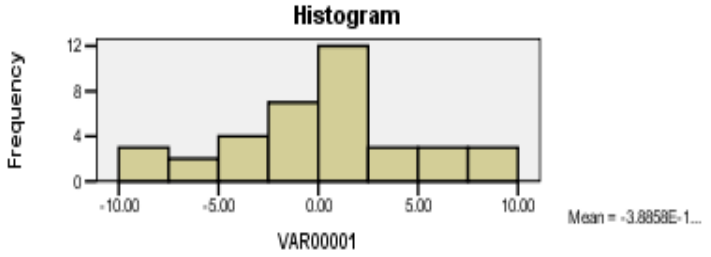
- Normal distribution: Not skewed in any direction.
- Positive skew: The distribution has a long tail in the positive direction, or to the right.
- Negative skew: The distribution has a long tail in the negative direction, or to the left.



More about distributions

- What is a variance?
- The variance is a measure of how spread out a distribution is.
- It is the average squared deviation of the observations from their mean (how the observations 'vary' from the mean).
- The larger the variance, the further spread out the data.





Why square deviations?

To calculate variance, the mean of a group of scores is subtracted from each score to give a group of 'deviations'.

- When we take the average of the deviation scores, the mean is calculated as the average, and the deviance scores total zero (positive and negative scores cancel).
- If you first square the deviation scores and then add them, you avoid this problem.
- The average of these squared deviation scores is called the variance.

$$\frac{\sum(X-M)^2}{n-1}$$

M=mean of all scores
n= number of scores

Example:

80 mean score

5 scores

Individual scores: 90, 90, 70, 70, 80.

Sum of (90-80), (90-80), (70-80), (70-80), (80-80)= **0**

NEED TO SQUARE!

$(90-80)^2 + (90-80)^2 + (70-80)^2 + (70-80)^2 + (80-80)^2 = 100+100+100+100+0=400$

Variance=100

Variance and Standard Deviation

- Variance- hard to interpret because when the values are squared, so are the units.
- To get back to our original units, you need to take the square root of the variance.
- This is the standard deviation.
- Standard deviation is a measure of the spread or dispersion of a set of data.
    - given in the same units as the indicator.
    - indicates the typical distance between the scores of a distribution and the mean.
    - the higher the standard deviation, the greater the spread of data.

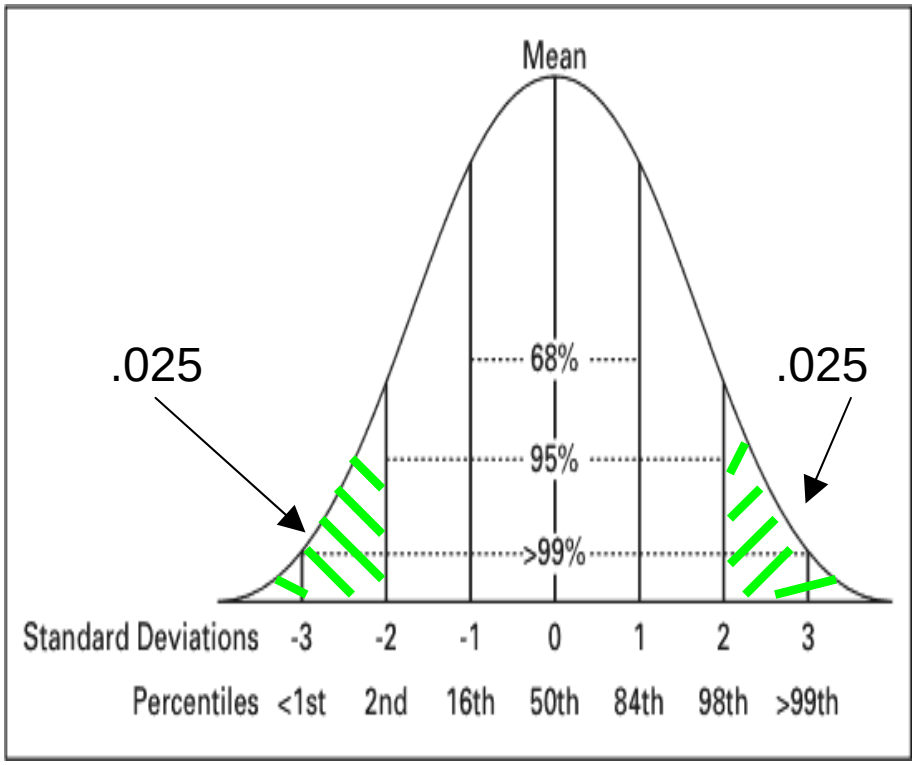
$$S = \sqrt{\frac{\sum (X-M)^2}{n-1}}$$

S = standard deviation
Σ = sum of
X = individual score
M = mean of all scores
n = sample size (number of scores)

Standard Deviation =10

Normal Distribution

- In a normal distribution, about 68% of the scores are within one standard deviation of the mean.
- 95% of the scores are within two standard deviations of the mean.



Inferential Statistics

- Inferential statistics are used to draw inferences about a population from a sample.
- Population: Group that the researcher wishes to study.
- Sample: A group of individuals selected from the population.
- Census: Gathering data from all units of a population, no sampling.

Inferential Statistics

- Inferential statistics generally require that data come from a random sample.

- In a random sample each person/object/item of the population has an equal chance of being chosen.

Goal of Statistical Analysis

The goal of statistical analysis is to answer 2 questions:

1) Is there a significant effect/association/difference between the variables of interest ? (i.e., can we reject the null hypothesis?)

2) If there is an effect/association/difference – how big is it?

Hypotheses

Null hypothesis: A hypothesis put forward to argue that a relationship or pattern does not exist.

Cholesterol study example: In a Randomized Control Trial, the control group and the treatment group have equal levels of cholesterol at the end of a study.

Null hypothesis: Groups A and B are equal.

Denoted by Ho:

Hypotheses

Alternative Hypothesis: Statement of what study is set to establish.

Alternative Hypothesis: Groups A and B have different levels of cholesterol.

The null hypothesis will be true if the findings are insignificant.

The null hypothesis will be false if the findings are significant.

Denoted by H1:

Alpha

Alpha level, or significance level, is the value that is determined by the researcher in order to reject or retain the null hypothesis. It is a pre-determined value, not calculated.

- In other words, if we select a value of .05, findings would be deemed statistically significant if they were found to be .05 or less.

What does this mean?

- Alpha indicates the probability that the null hypothesis will be rejected when it is true (in other words, the null hypothesis is wrongly rejected).
- This is called Type 1 error or alpha error.

Type 1 Error

- E.g. in a trial of new Drug X, the null hypothesis might be that the new Drug X is no better than the current Drug Y.
  - H0: there is no difference between Drug X and Drug Y.
- A Type 1 error would occur if we concluded that the two drugs produced different effects when there was no difference between them.

Beta

- Beta is the probability of making a Type 2 error when testing a hypothesis.
- Type 2 error is failing to detect an association when one exists, or failing to reject the null hypothesis when it is actually false.
- You kept the null hypothesis when you should not have.

    - If Drug X and Drug Y produced different effects, and it was concluded that they produce the same effects.

Type 1 and Type 2 Error

Truth	Decision: Reject Ho	Decision: Don't reject Ho
Ho:	**TYPE 1 ERROR**	Correct Decision
H1:	Correct Decision	**TYPE 2 ERROR**

Statistical Significance

What is statistical significance?

- Statistically significant findings mean that the probability of obtaining such findings by chance only is less than 5%.
  - findings would occur no more than 5 out of 100 times by chance alone.

What if your study finds there is an effect?

- You will need to measure how big the effect is, you can do this by using a measure of association (odds ratio, relative risk, absolute risk, attributable risk etc.).

What if there is an effect?

- Absolute Risk is the chance that a person will develop a certain disease over a period of time.
- E.g. Out of 20,000 people, 1600 developed lung cancer over 10 years, therefore the absolute risk of developing lung cancer is 8%.

- Relative Risk (RR) is a measure of association between the presence or absence of an exposure and the occurrence of an event.
- RR is when we compare one group of people to another to see if there is an increased risk from being exposed.
- RR is the measure of risk for those exposed compared to those unexposed.
- Used in randomized control trials and cohort studies- Can't use RR unless looking forward in time.
- E.g. The 20 year risk of lung cancer for smokers is 15%
  The 20 year risk of lung cancer among non-smokers is 1%
- Most commonly reported.

What if there is an effect?

- Odds Ratio (OR) is a way of comparing whether the probability of a certain event is the same for two groups.
- Used for cross-sectional studies, case control trials, and retrospective trials.
- In case control studies you can't estimate the rate of disease among study subjects because subjects selected according to disease/no disease. So, you can't take the rate of disease in both populations (in order to calculate RR).
- OR is the comparison between the odds of exposure among cases to the odds of exposure among controls.
- Odds are same as betting odds. Example: if you have a 1 in 3 chance of winning a draw, your odds are 1:2.
- To calculate OR, take the odds of exposure (cases)/odds of exposure (controls).
- E.g. Smokers are 2.3 times more likely to develop lung cancer than non-smokers.

Confidence Intervals

- When we measure the size of the effect we use confidence intervals.
- The odds ratio we found from our sample (E.g. Smokers are 2.3 times more likely to develop cancer than non-smokers) is only true for the sample we are using.
- This exact number is only true for the sample we have examined; it might be slightly different if we used another sample.
- For this reason we calculate a confidence interval- which is the range in risk we would expect to see in this population.

- A 95% confidence interval of 2.1 to 3.4 tells us that while smokers in our study were 2.3 times more likely to develop cancer, in the general population, smokers are between 2.1 and 3.4 times more likely to develop cancer. We are 95% confident this is true.

Power

- If findings are statistically significant, then conclusions can be easily drawn, but what if findings are insignificant?
- Power is the probability that a test or study will detect a statistically significant result.
- Did the independent variables or treatment have zero effect?
- If an effect really occurs, what is the chance that the experiment will find a "statistically significant" result?

Power

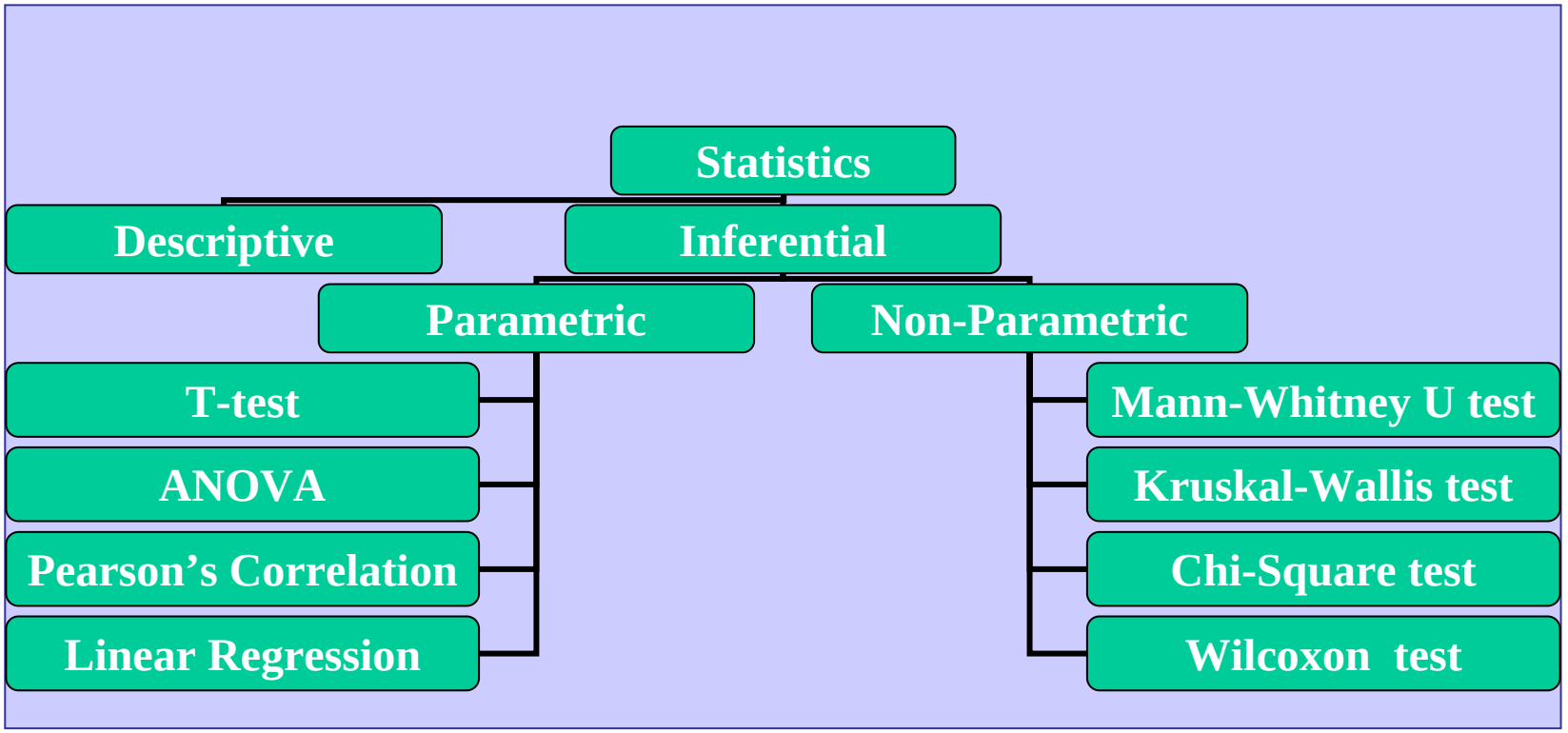
Determining power depends on several factors:

- 1) Sample size: how big was your sample?
- 2) Effect size: what size of an effect are you looking for? E.g. How large of a difference (association, correlation) are you looking for? What would be the most scientifically interesting?
- 3) Standard deviation: how scattered was your data?

  For example, a large sample, with a large effect, and a small standard deviation would be very likely to have found a statistically significant finding, if it existed.
- A power of 80%-95% is desirable.
- One of the best ways to increase the power of your study is to increase your sample size.

Parametric vs. Non-Parametric

Some Advanced Statistics...



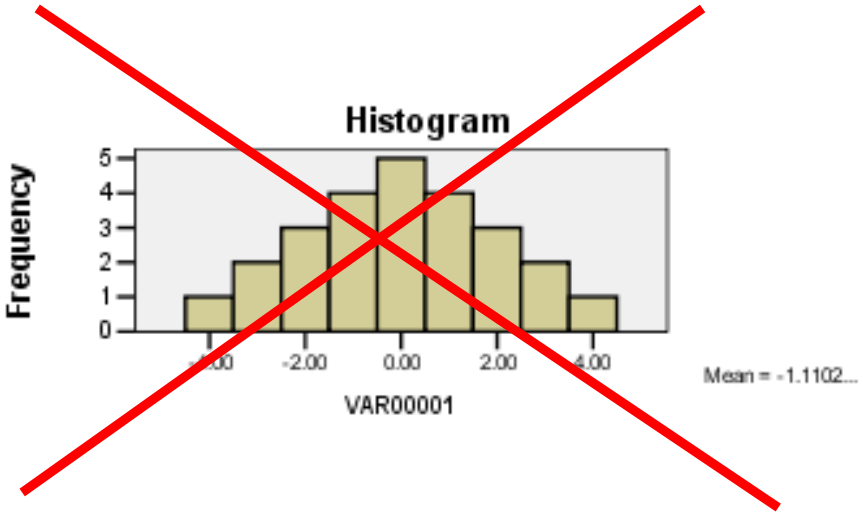
Some examples of parametric and non-parametric tests.

Parametric Tests

- Parametric tests assume that the variable in question is from a normal distribution.
- Non-parametric tests do not require the assumption of normality.
- Most non-parametric tests do not require an interval or ratio level of measurement; can be used with nominal/ordinal level data.

Non-Parametric Tests

- Use when all assumptions of parametric statistics cannot be met.
- Can be used with data that are not normally distributed.



Normality

How do you know if data are normally distributed?

- Run test or histogram in SPSS to determine if data meet the normality assumption required for parametric tests.

Types of Analyses

- Univariate analysis- the analysis of one variable.
  - Mean
  - Median
  - Mode
  - Range- equal to the difference between the largest and the smallest values.
  - Standard deviation

Types of Analyses

- Bivariate analysis is a kind of data analysis that explores the association between two variables.
- Some examples of bivariate analysis include:

  -Pearson's correlation

  -T-Test

  -Spearman's Rho

  -Mann-Whitney Test

  -Linear regression (not multiple regression)

Types of Analyses

- Multivariate analysis: the analysis of more than two variables.
- Some examples of multivariate analysis include:

  -Multiple regression

  -Multiple logistic regression

Research Examples

- Example: Univariate Analysis
  How many women have heart disease in BC?

- Example: Bivariate Analysis
  Are height and weight correlated?

- Example: Multivariate Analysis
  Do age, diet, exercise, and diabetes predict heart disease?

Basic Research Design

Selecting the appropriate statistical test requires several steps.

- The level of variable is a major component in deciding what test to use.

- Test selection should be based on:
  1) What is your goal: Description? Comparison? Prediction? Quantify association? Prove effectiveness? Prove causality?
  2) What kind of data have you collected?
  3) Is your data normally distributed? Can you use a parametric or non-parametric test?
  4) What are the assumptions of the statistical test you would like to use? Does the data meet these assumptions?

Assumptions

- There are various assumptions for each test.
- Before you select a test, be sure to check the assumptions of each test.
- You will need to contact a consultant, or review statistical/research methods resources to find this information.
- Some examples of common assumptions are:
  - The dependent variable will need to be measured on a certain level E.g. Interval level.
  - The independent variable(s) will need to be measured on a certain level E.g. Ordinal level.
  - The population is normally distributed (not skewed).

- If your data do not meet the assumptions for a specific test, you may be able to use a non-parametric test instead.

Examples of Tests

- **T-Test**
- Allows the comparison of the mean of 2 groups.
- Compares actual difference between two means in relation to the variation in the data (expressed as the standard deviation of the difference between the means).

- Ex. A doctor gives two different drugs to a group of diabetics to see if blood sugar lowering times differ, and if the difference between times are in fact significant.

- Null hypothesis: Drug A and Drug B will have equal blood sugar lowering times (no difference).
- Alternative hypothesis: Drug A and B will have different blood sugar lowering times (difference).

Examples of Tests

- **Analysis of Variance (ANOVA)**
- Allows the comparison of 3 or more groups.
- Looks at the variation within groups, then determines how that variation would translate into variation between groups (considering number of participants).
- If observed differences are larger than what would be expected by chance, the findings are statistically significant.
- Ex. Are psychotherapy, family therapy and behaviour therapy equally effective in treating alcoholism?

Examples of Tests

- **Correlation**
- Allows an examination of the relationship between variables; is there a relationship between these variables? Are they positively or negatively related?
- A correlation coefficient of 0 means that there is no relationship between the variables, -1 negative relationship, 1 positive relationship.
- Important: Correlation is not causation.

- Ex. What is the relationship between exercise and depression?
- Does depression increase when exercise increases?
- Does depression decrease when exercise increases?
- Is there no significant correlation between exercise and depression?

Examples of Tests

- **Linear Regression**
- Focuses on prediction. Involves discovering the equation for a line that is the best fit for the given data. That linear equation is then used to predict values for the data.

- Do variables a and b predict event c?

- Ex. Does age predict income?

What have you learned?

1) Basic statistical terms and concepts
2) Basic and intermediate research methodology
3) Different types of research questions
4) Different levels of data and variables
5) Descriptive and inferential statistics
6) Parametric and non-parametric tests
7) Variety of popular statistical tests

Conclusions

- Statistics are vital to producing quality research. Appreciate statistics for the important role they play, but do not let statistics prevent you from posing new and exciting research questions. After all, research begins when there is a question…

Resources

Great resources:

- Statistics without tears: An introduction for non-mathematicians.

  Author: Derek Rowntree

- Statsoft Online Textbook (T. Hill and P. Lewicki)
- http://www.statsoftinc.com/textbook/stathome.html
- HyperStat Online Statistics Textbook (D.M. Lane)
- http://davidmlane.com/hyperstat/index.html
- StatNotes (D.D. Garson)
- http://www2.chass.ncsu.edu/garson/pa765/statnote.htm
- Statistics at Square One (T. Swinscow, revised by M. Campbell)
- http://bmj.bmjjournals.com/collections/statsbk/index.shtml

Introduction to Statistics and Quantitative Research Methods