Introduction to Statistics and Quantitative Research Methods

Edited from original source by fraserhealth

Statistics Definition

- Statistics (**stat.**) is "the science of using information discovered from studying numbers." (Cambridge dictionary)
- A fact or piece of data obtained from a study of a large quantity of numerical data.(Oxford dictionary)
- The mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling.(American Heritage)

What is statistics?

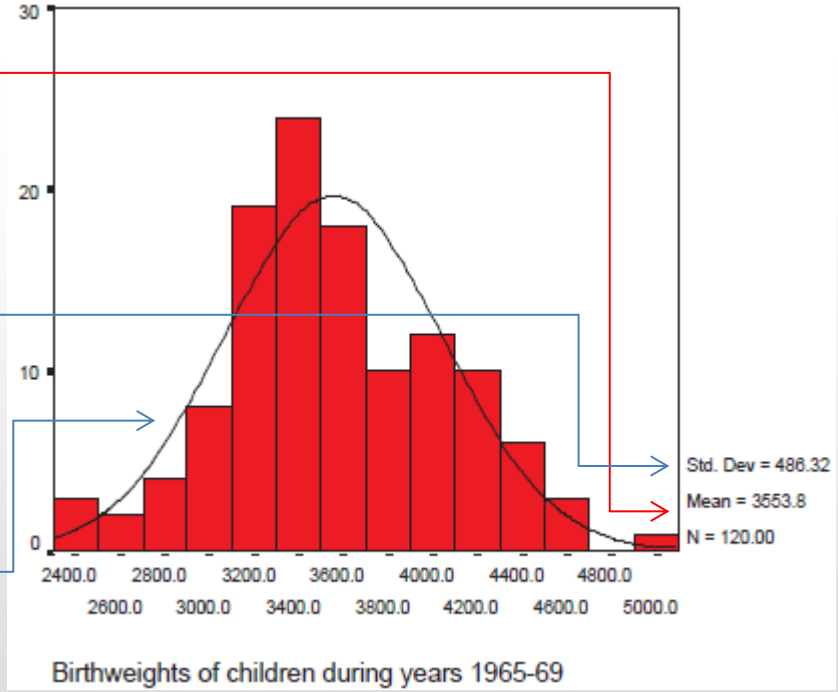
- The science that deals with the collection, classification, analysis and interpretation of numerical data.
- Using mathematical theories of probability.

$$\bar{x} = \frac{1}{N}\sum_{1}^{N} x_i$$

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{1}^{N}(x_i - \bar{x})^2}$$

$$z = \frac{x - \bar{x}}{\sigma}$$

$$f(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{z^2}{2}}\, dz$$



Research Process: Basic steps

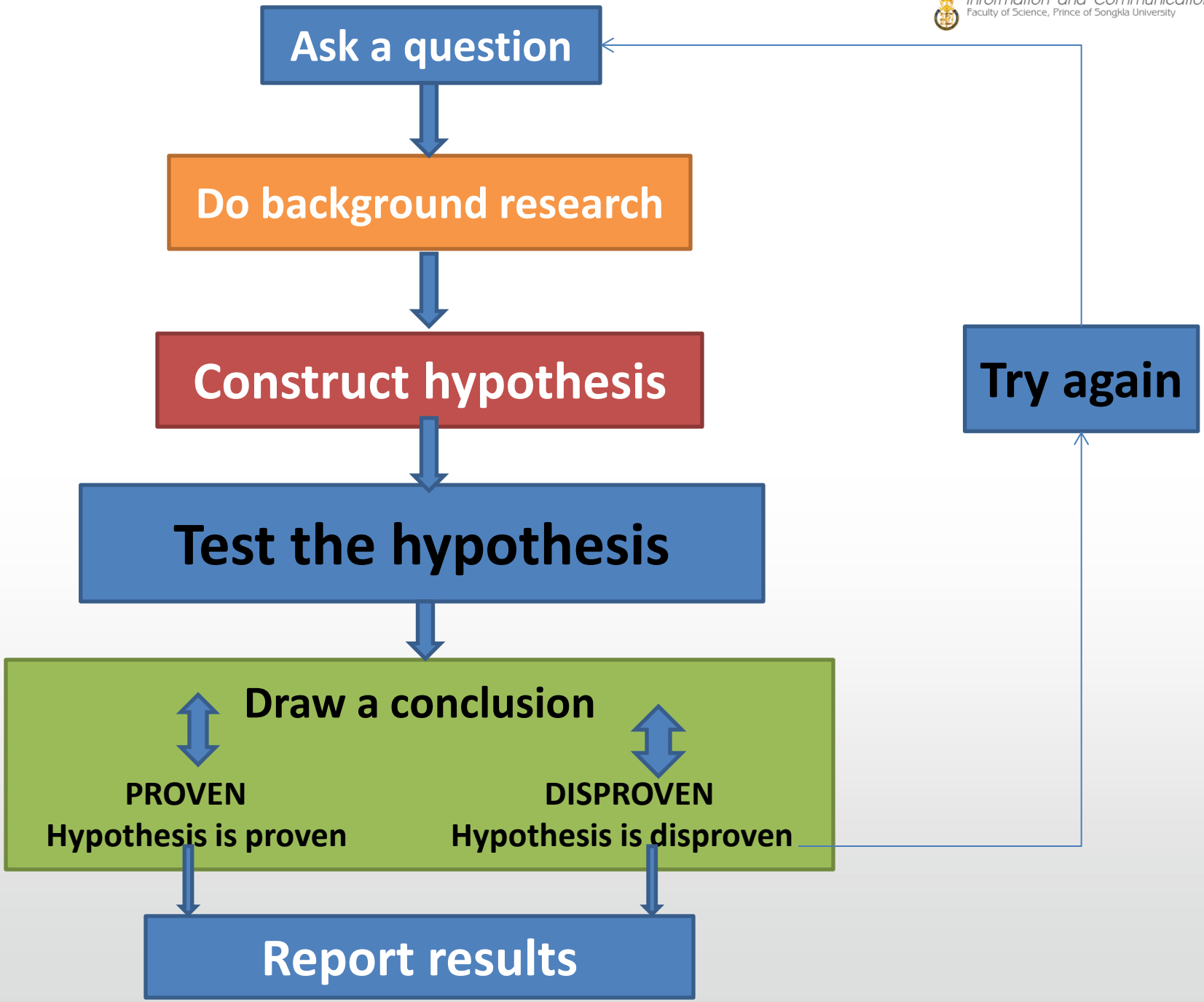
- Develop a research question
  - Based on an initial observation
  - Generated explanations, or theories of those observations

How many different types of clouds do you see? How do they differ?

Average birth weight in babies of Thailand is 2.5 kilograms?

Research Process: Basic steps

- Develop a research question
  - Based on an initial observation
  - Generated explanations, or theories of those observations
- Conduct thorough literature review
- Re-define research questions → hypotheses
  - Conducted predictions (hypotheses)
- Design research methodology/study
- Apply for ethics approval
- Collect data
  - Primary (direct from observation)
  - Secondary data (pre-collected by third party)
- Analyse data
- Draw conclusions and relate findings


Ask a question
Do background research
Construct hypothesis
Test the hypothesis
Draw a conclusion
PROVEN
Hypothesis is proven
DISPROVEN
Hypothesis is disproven
Report results
Try again

Research Questions Types

- Research begins when there is a question
- Descriptive type
  - Ex: How many hours a week do employee spend at their desks?
- Inferential type
  - Ex: 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.

Descriptive and Inferential Statistics

Ex. Consider event of tossing dice. The dice is rolled 100 times and the results are forming the sample data.

- Descriptive statistics is used to grouping the sample data to the following table

Outcome of the roll	Frequencies in the sample data
1	10
2	20
3	18
4	16
5	11
6	25

- Inferential statistics can now be used to verify whether the dice is a fair or not.

Variables

- In context of research, variable is the characteristic or phenomenon that can be measured or classified.
- **Quantitative variables** can be classified as either ***discrete*** or ***continuous***.
- **Qualitative (or categorical) variables** consist of 4 types of variables:

  Nominal, Ordinal, Interval, Ratio

Levels for Qualitative Variables

- Nominal
  - Data that is classified into categories and cannot be arranged in any particular order.
  - E.g. Gender, eye colour, ethnicity
- Ordinal
  - Data that can be ordered, but distance between intervals not always equal
  - E.g. rating a brand of soft drink on a scale of 1-5

Levels of Data

- Interval
  - Data that can be ordered and distance between each interval is equal
  - Arbitrary zero point
  - E.g. 1,2,3
- Ratio
  - Similar to interval scale, but has true zero point
  - E.g. weight, salary ($0 = $0)

Types of Variables

- An independent
  - The variable that will influence to outcome measure
- A dependent
  - The variable that is dependent on or influence by the independent variable(s).

Types of Variables

- An intervening variable is the variable that links the independent and dependent variable
  - Independent → Intervening → Dependent

  E.g. Educational level → Occupational level → Income level
- A confounding variable is a variable that has many other variables, or dimensions built into it.
  - For example: GDP (Gross Domestic Product)
    - Expenditure (total spending)
    - Income

Group Exercise

- 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.
- 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.
- 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.

Group Exercise: Discussion

- What is the dependent variable(s) in this study?
- What is the independent variable(s)?
- What is the level of data?
  - Nominal, Ordinal, Interval, Ratio

Parametric VS. Non-parametric

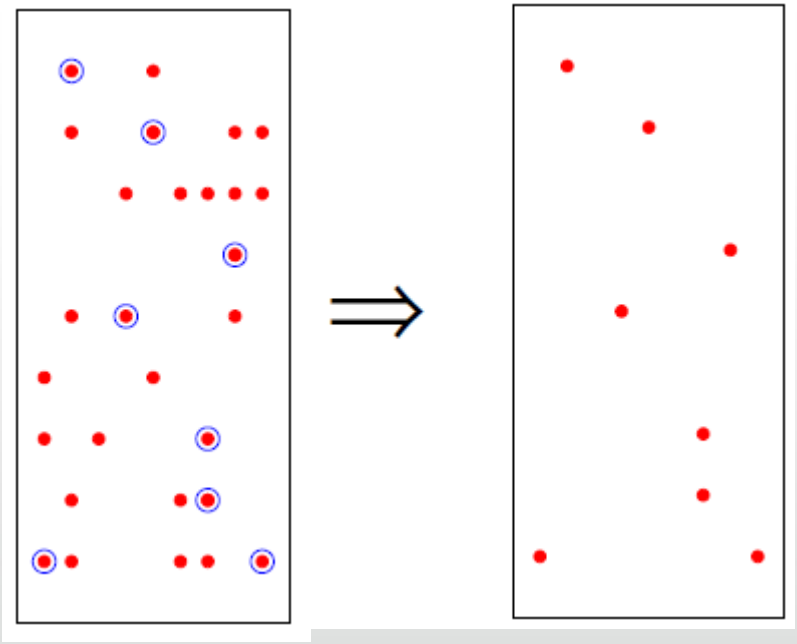
- Parametric tests
  - Based on the normal distribution which have 4 basic assumptions that must be met for the test
    - Normally distributed data
    - Homogeneity of variance
    - Interval data
    - Independence

Parametric VS. Non-parametric

- Non-parametric
  - Do not require the assumption of normality
  - Do not require an interval or ratio level of measurement
  - Can be used with nominal/ordinal level data
    - Data that are not normally distributed
  - Use when all assumptions of parametric statistics cannot be met

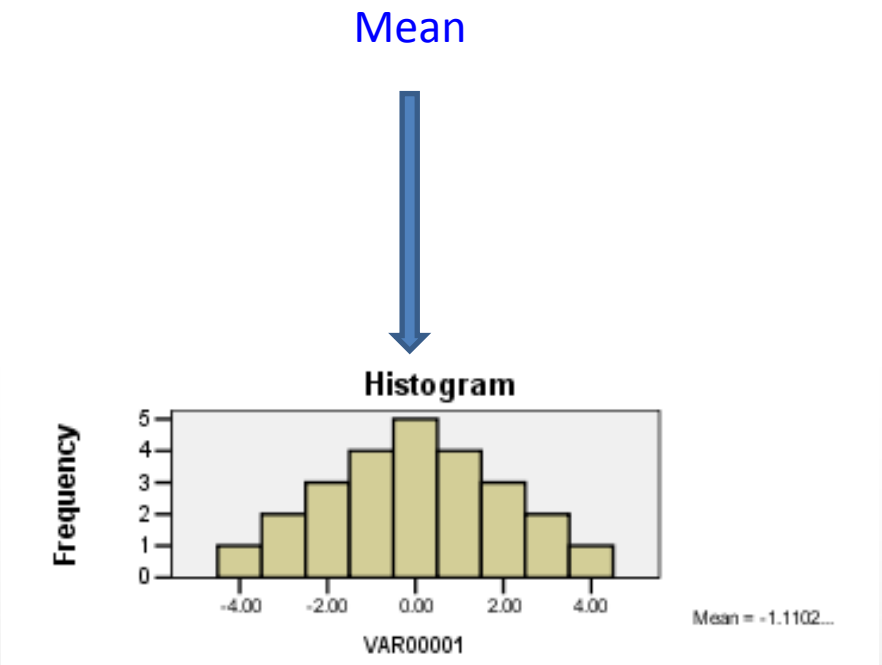
Data Collection: How to measure

- Correlational/Cross-sectional research
  - The researchers observe what naturally goes on in the world without directly interfering with it.



Descriptive Statistics Defined

- Mean
  - The sum of all the scores divided by the total number of scores.
  - Often referred as the average.
  - Good measure of central tendency
    - The location of the middle in a distribution of scores
  - The mean can be misleading by extreme scores (very high, or very low scores)
    - Extreme cases or values are called outliers.

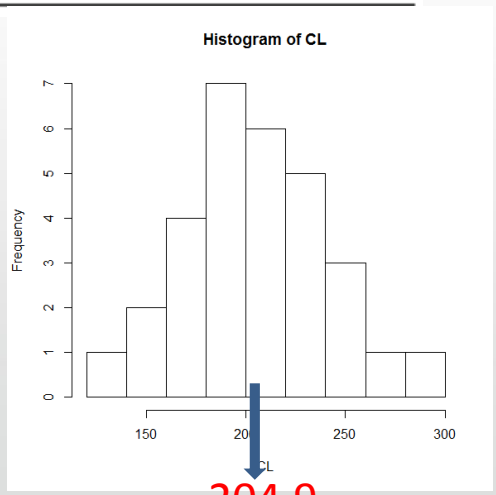


Mean

- Ex. Consider measurements from 30 female participants, serum levels of cholesterol were measured, which are given in the Table.

261	160	259	223	169	127	221	190
224	228	229	294	204	177	199	212
186	207	192	241	162	249	206	210
200	213	185	171	189	159		

$$\bar{x} = \frac{1}{N}\sum_{1}^{N} x_i = \frac{261 + 160 + \cdots + 159}{30} = 204.9$$



The Median

- A median is the middle of a distribution.
  - Half the scores are above the median and half are below the median.
- Computing the median
  - Arrange the scores into ascending order
  - If there is an odd number of numbers, the median is the middle number
    - Ex: 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.
    - Ex: the median of the numbers 4, 8, 9, 13 is (8+9)/2 = 8.5

Median

- Data:

261, 160, 259, 223, 169, 127, 221, 190,224, 228, 229, 294, 204, 177, 199, 212,186, 207, 192, 241, 162, 249, 206, 210,200, 213, 185, 171, 189, 159

- Sort:

127 159 160 162 169 171 177 185 186
189 190 192 199 200 204 206 207 210
212 213 221 223 224 228 229 241 249
259 261 294

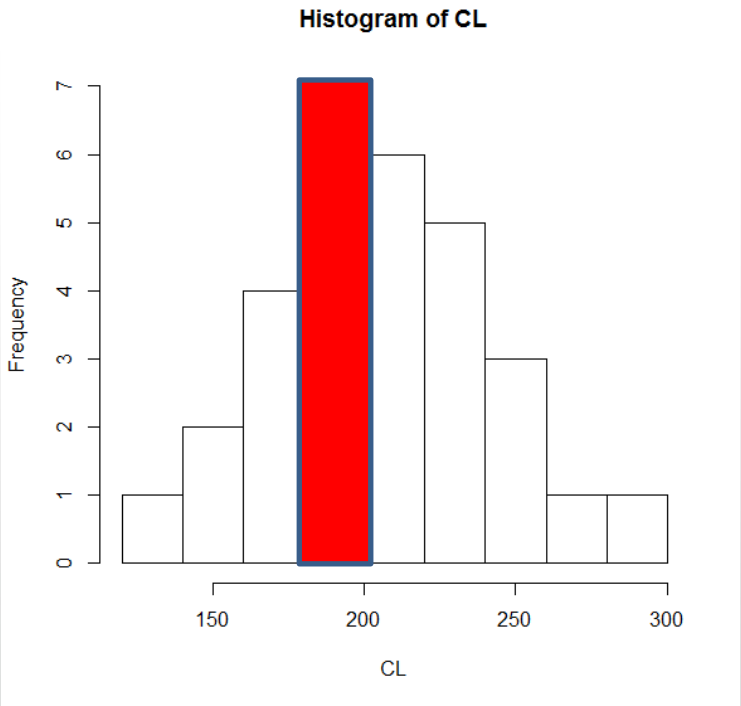
- Median = (204+206)/2=205

The Mode

- Mode is the most frequently occurring score in a distribution/data set.
- Calculating the mode
  - Place the data in ascending order, count how many times each score occurs
  - The score that occurs the most is the mode
- Distributions can have more than one mode, called “multimodal.”

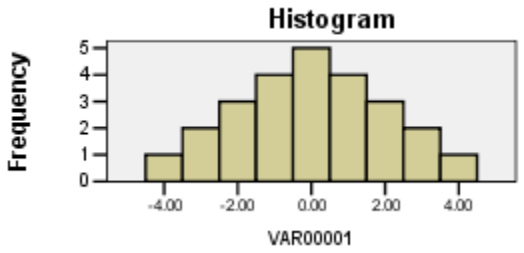
Mode

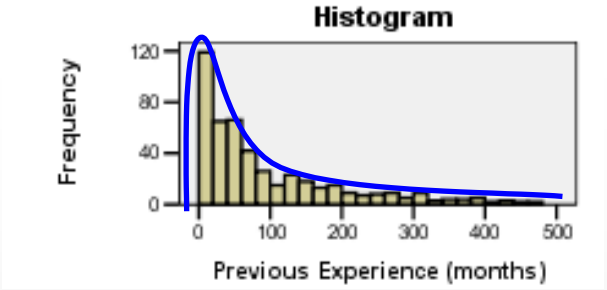
261	160	259	223	169	127	221	190
224	228	229	294	204	177	199	212
186	207	192	241	162	249	206	210
200	213	185	171	189	159		

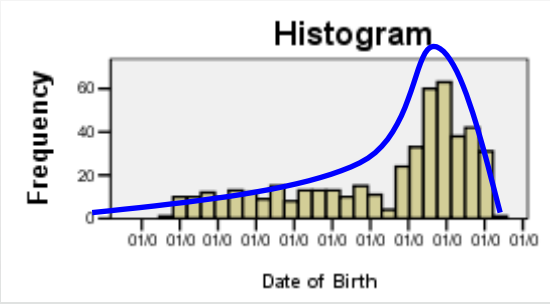


Skewed distributions

- 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.







Variance

- The variance is a measure of how spread out a distribution is.
- The variance shows how the observations 'vary' from the mean.
- The larger the variance, the further spread out the data.

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

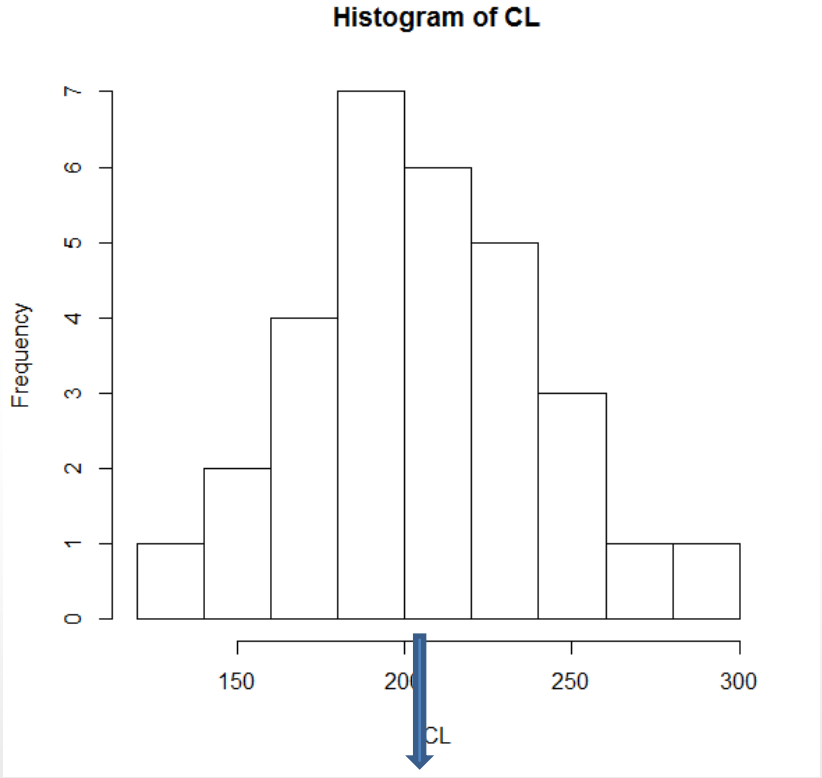
Square deviations

- 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 can avoid this problem
- The average of these squared deviation scores is called the variance.

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

X = individual score
M = mean of all scores
n = number of scores

Variance



$\bar{X} = 204.9$

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

$$= \frac{(127 - 204.9)^2 + \cdots + (294 - 204.9)^2}{29}$$

$$= 1247.472$$

Square deviations

Example:
80 mean score
5 scores

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

$$(90-80)+(90-80)+(70-80)+(70-80)+(80-80)$$
$$=10+10+(-10)+(-10)+0=0$$

NEED TO SQUARE!

$$(90-80\ )^2+(90-80)^2+(70-80)^2+(70-80)^2+(80-80)^2$$

$$=\frac{100+100+100+100+0}{5-1}=\frac{400}{4}=100$$

Variance=100

Standard Deviation

- Variance hard to interpret because when the values are squared, so are the units.
- To get back to original units, it need to be taken the square root of the variance.
- This is the standard deviation.

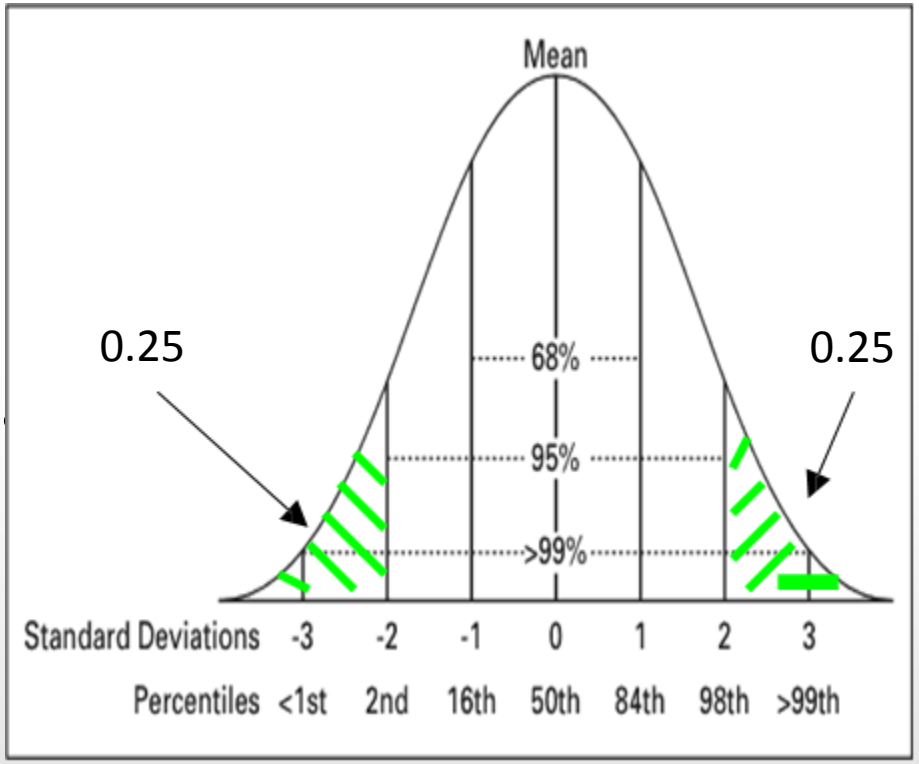
$$s = \sqrt{\frac{\sum(X - M)^2}{n - 1}}$$

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.

Standard deviation in Normal Distribution

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



Inferential Statistics

- Inference statistics are used to draw inferences about a population from a sample.
  - Population: A group of thing that the researchers wishes to study.
  - Sample: A group of individuals selected from the population.
  - Census: Gathering data from units of a population, no sampling.
- Generally, inferential statistics require that data come from a random sample.
  - A random sample: each person/object/item of the population has an equal chance of being chosen.

Hypotheses

- Hypothesis is a prediction about the state of event/theory/effect.
  - Null hypothesis (H_0): Statement set to argue that a relationship or pattern does not exist.
  - Alternative Hypothesis (H_1): Statement of what study is set to be present/establish.
- Cholesterol study example: In a Randomized Control Trial, the control group (A) and the treatment group (B) have equal levels of cholesterol at the end of a study.
  - H_0: Groups A and B are equal levels of cholesterol.
  - H_1: Groups A and B have different levels of cholesterol.

Hypotheses

- The null hypothesis will be true if the findings are insignificant.
- The null hypothesis will be false (reject) if the findings are significant.

Alpha (α): Type I error

- 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.
- Alpha indicates the probability that the null hypothesis will be rejected when it is true (in other words, the null hypothesis is wrongly rejected).
  - Suggested value is 0.05 or 5% (Fisher, 1925)
  - if we replicated our data collection 100 times we could expect that on five occasions we would obtain a test statistic large enough to make us think that there was a genuine effect in the population even though there isn't.

Alpha (α): Type I error Example

- 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.
  - H_0: 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 (β): Type II error

- Type 2 error is failing to detect an association when one exists, or failing to reject the null hypothesis when it is actually false.
  - Keep 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.
- An ideal world, the probability of this error should be very small.
  - If there is an effect in the population then it's important that we can detect it.
  - Cohen (1992) suggests the maximum acceptable probability of a Type II error would be 0.20 (20%)
  - if we took 100 samples of data from a population in which an effect exists, we would fail to detect that effect in 20 of those samples (so we'd miss 1 in 5 genuine effects).

Type I and Type II error

Truth about the population

Decision based on Sample	H_0 True	H_1 True
Reject H_0	**TYPE I ERROR**	Correct Decision
Fail to Reject H_0	Correct Decision	**TYPE II ERROR**

Hypothesis testing

- Null and alternative hypotheses
  - Convert research question into null and alternative hypotheses
  - H_0 is a claim of "no difference"
  - H_1 is a claim of "a difference in the population"
- Test statistic
  - Calculate a test statistic from the data
- P-value and conclusion
  - The P-value answers the question "If the null hypothesis were true, what is the probability of observing the current data or data that is more extreme?"
  - Small *p* values provide evidence against the null hypothesis
- Decision (optional)
  - If $P \leq \alpha$, we will reject the null hypothesis, otherwise it will be retained for want of evidence.

Testing research questions with statistical models

- Recall a research process:
  - Generate a research question through an initial observation
  - Generate a theory to explain the initial observation
  - Generate hypotheses: break theory down into a set of testable predictions
  - Collect data to test the theory: decide on what variables need to measure to test predictions
  - Analyse the data: fit a statistical model to the data – assess this model to see whether or not it supports the initial predictions

Statistical Significance

- Fisher (1925) suggested that 95% is a useful threshold for confidence: only when we are 95% certain that a result is genuine (i.e. not a chance finding) should we accept it as being true
- The opposite way can say if there is only a 5% chance (a probability of .05) of something occurring by chance then we can accept that it is a genuine effect
- we say it is a statistically significant finding
- "Significant" in this context means "the observed difference is not likely due to chance". It does not mean of "important" or "meaningful"

Effect size

- Even if a test statistic is significant, it does not mean that the effect it measures is meaningful or important.
- An effect size is a measurement of the size of an effect (be that an experimental manipulation or the strength of a relationship between variables).
- An effect size is an objective and standardised measure of the magnitude of observed effect. It can show the importance of an effect.

Confidence Intervals

- Confidence intervals is a boundary (a range of values) which we believe that the true value of the population mean will fall within these limits.
- Typically, the value will set as 95% or 99% confidence intervals
  - If we'd collected 100 samples, calculated the mean and then calculated a confidence interval for that mean then 95/99 of these samples, the confidence intervals we constructed would contain the true value of the mean in the population.

Statistical Power

- Statistical Power is the ability of a test to detect an effect that might have existed.
- Power is the probability that a test or study will detect a statistically significant result.

Power

- Determining power depends on several factors:
  - Sample size: how big was a sample?
  - Effect size: what size of an effect are you looking for? E.g. How large of a difference (association, correlation) are you looking for?
  - Standard deviation: how scattered was your data?
- A power of 80% - 95% is desirable.
  - we can be confident that we achieved sufficient power to detect any effects that might have existed.
  - One of the best ways to increase the power of your study is to increase your sample size.

Types of Analyses

- Univariate analysis: the analysis of one independent variable.
  - Mean, Median, Mode, Standard deviation.
  - Example: How many women have heart disease in Thailand?
- Bivariate analysis: the analysis that explore the association between two variables.
  - Pearson's correlation, *t*-test, Mann-Whitney test.
  - Example: Are height and weight correlated?

Types of Analyses

- Multivariate analysis: the analysis of several dependent variables (outcomes) simultaneously.
  - Multiple regression, multiple logistic regression.
  - Example: Do age, diet, exercise, and diabetes predict heart disease?

Assumptions

- There are various assumptions for each test.
  - Check the assumptions of each test before select a test.
- Some examples of common assumptions are:
  - The dependent variable(s) 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.

Basic Research Design

- The level of variable is a major component that effect to the decision what statistical test to use.
- Statistical test selection should be based on:
  - What is a goal?
    - Description/Comparison/Prediction
  - What kind of data have been collected?
  - Is data normally distributed?
    - Use a parametric or non-parametric test.
  - What are the assumptions of the statistical test that will be used?

Example of Statistical Tests

- *t*-test
- Analysis of Variance (ANOVA)
- Correlation

t-test

- Comparison of the mean of two groups.
- Expressed as the standard deviation of the difference between the means.
- Example: 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
  - H_0: Drug A and Drug B will have equal blood sugar lowering times (no difference).
  - H_1: Drug A and Drug B will have different blood sugar lowering times (difference).

Steps involved in the t-test

- Calculate the means and standard deviations of the groups' outcomes
- Calculate the t-ratio
- Check to see if the calculated t is statistically significant using a t-table
- It is significant if t is greater than the critical value of t at the 0.05 level
- If so, the group means are considered different

The *t*-score

- The t-score (a.k.a., t-ratio)
  - The t-distribution and a t-table are used
  - This is because the SD of the population is estimated from the sample
- P values are found using the calculated t-score and a t-table
- t-tables consider the number of subjects in the groups

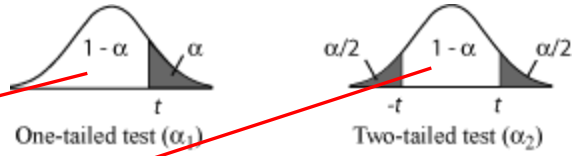
The *t*-score(cont.)

- Referred to as degrees of freedom (df)
  - Signifies the number of subjects in each group minus 1
  - Minus 2 when there are two groups
- Thus, a study that compares the means of 2 groups that involve 30 subjects has 28 df

The *t*-table

- t-distributions eventually become nearly normal when many subjects are included
  - As a result, t-tables usually only go to 100 df
- Alpha levels are shown for
  - When α is all in one tail (α1 or one-tailed test)
  - When α is spit between the tails (α2 or two-tailed test)

t-table showing critical values for *t*. (only to 15 *df*)



	α_1	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
df	α_2	0.20	0.10	0.05	0.02	0.01	0.002	0.001
1		3.078	6.314	12.71	31.82	63.66	318.30	636.62
2		1.886	2.920	4.303	6.965	9.925	22.330	31.60
3		1.638	2.353	3.182	4.541	5.841	10.210	12.92
4		1.533	2.132	2.776	3.747	4.604	7.173	8.610
5		1.476	2.015	2.571	3.365	4.032	5.893	6.869
6		1.440	1.943	2.447	3.143	3.707	5.208	5.959
7		1.415	1.895	2.365	2.998	3.499	4.785	5.408
8		1.397	1.860	2.306	2			5.041
9		1.383	1.833	2.262	2			4.781
10		1.372	1.812	2.228	2.764	3.169	4.144	4.587
11		1.363	1.796	2.201	2.718	3.106	4.025	4.437
12		1.356	1.782	2.179	2.681	3.055	3.930	4.318
13		1.350	1.771	2.160	2.650	3.012	3.852	4.221
14		1.345	1.761	2.145	2.624	2.977	3.787	4.140
15		1.341	1.753	2.131	2.602	2.947	3.733	4.073
…								
Etc.								

Critical value for 10 df and $\alpha_2 = 0.05$

To 100

Remember

Big *t*-value

↓

Small *P* value

↓

Statistical significance

ANOVA

- Allows the comparison of 3 or more groups.
- ANOVA tells us whether three or more means are the same.
- ANOVA is an omnibus test. It tests for an overall experimental effect. It does not provide specific information about with groups were affected.
- Example: Are psychotherapy, family therapy and behaviour therapy equally effective in treating alcoholism?

Correlation

- Allows an examination of the relationship between variables (Covariance).
- If we standardise covariance value, we get Pearson's correlation coefficient, *r*.
- The correlation coefficient has to lie between -1 and +1.
  - A correlation coefficient of 0 indicates there is no relationship between the variables (one variable changes, the other stays the same).
  - A correlation coefficient of -1 indicates a prefect negative relationship (one variable increases, the other decreases by a proportionate amount).
  - A correlation coefficient of +1 indicates a perfect positive relationship (one variables increases, the other increases by a proportionate amount).

Research Examples

- Scenario: Study relationship between arachnophobia (fear of spider) and spider.
- Research Question: Is arachnophobia (fear of spiders) specific to real spiders or is a picture enough?
- Hypothesis
  - H_0: anxiety level when exposes with real spider = anxiety level when exposes with a picture of the same spider.

Research Examples (cont)

- Participants
  - 12 spider phobia individuals
- Manipulation
  - Each participant was exposed to a real spider and a picture of the same spider at two points in time.
- Outcome
  - Anxiety level: heart beat

Collected data

Participant	Picture of spider	Real spider
P1	30	40
P2	35	35
P3	45	50
P4	40	55
P5	50	65
P6	35	55
P7	55	50
P8	25	35
P9	30	30
P10	45	50
P11	40	60
P12	50	39

What kind of test should be used?

- Independent (Between subjects)
- Dependent (Within subjects/Paired-sample)
  - repeated measure

Variables

- The purpose of this experiment
  - determine the differences in anxiety level between real spiders and pictures of spiders.
- Twelve subjects were shown a picture of a spider and a real spider.
  - After viewing the picture or spider their anxiety level was measured.
  - The order of treatment was counterbalanced (6 shown picture first, 6 shown real spider first).
- In the research question, we did not specify which type of spider should cause more anxiety level. Therefore, the test should be “2-tailed”

Independent or Dependent variables?

- Dependent variable
  - Anxiety level (i.e. Anx_level)
- Independent variable
  - An exposure of spider
- Note: the independent variable (an exposure of spider) causes changes in the dependent variable (anxiety level).

Analysis?

- *t*-test
- ANOVA
- Correlation

Output

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Anxiety_picture	12	30.00	25.00	55.00	40.0000	2.68272	9.29320	86.364	.000	.637	-.996	1.232
Anxiety_real	12	35.00	30.00	65.00	47.0000	3.18377	11.02889	121.636	-.007	.637	-1.121	1.232
Valid N (listwise)	12											

Output

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Picture of Spider	40.00	12	9.293	2.683
	Real Spider	47.00	12	11.029	3.184

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Picture of Spider - Real Spider	-7.000	9.807	2.831	-13.231	-.769	-2.473	11	.031

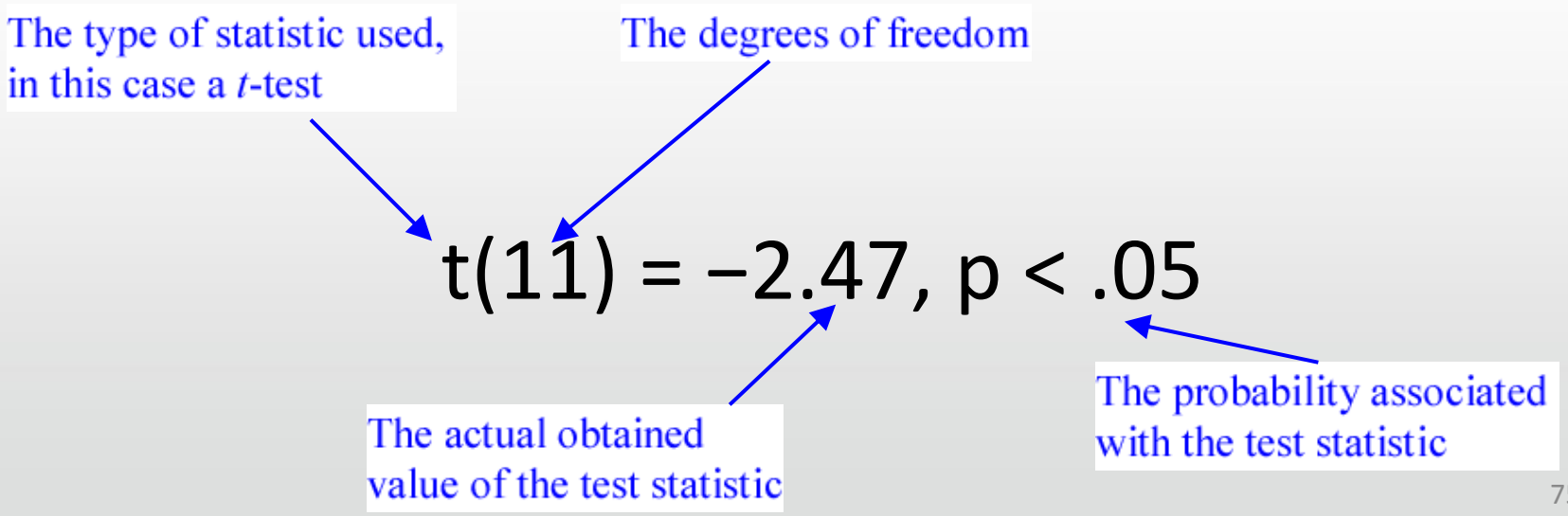
t(11) = -2.473, p =.031 (p is an exact probability)

p < .05

there is a significant difference between the means

Report the results (conclusion)

- On average, participants experienced significantly greater anxiety to real spiders (M = 47.00, SE = 3.18) than to pictures of spiders (M = 40.00, SE = 2.68), t(11) = −2.47, p < .05



Research begin when there is a question

References

- Fraser Health, “Introduction to statistics and Quantitative Research Method”, http://research.fraserhealth.ca/media/Introduction-to-Statistics-and-Quantitative-Research-Methods.pdf [online; accessed September 2014]
- Parina Patel, “Introduction to Quantitative Methods”, http://www.law.harvard.edu/library/research/empirical/quantitative_methods.pdf, 2009 [online; accessed September 2014]
- Michael T. Haneline, Inferential statistics, 2008, http://w3.palmer.edu/michael.haneline/4_EBC_PP_Chap4b.ppt.
- Andy Field, “Discovering statistics using SPSS”, SAGE Publications Ltd, 3rd edition, 2009. ISBN 9781847879066.