Introduction to SPSS

By

Dr. Muhammad Usman

Manager (Database)

PASTIC National Center, QAU Campus, Islamabad.

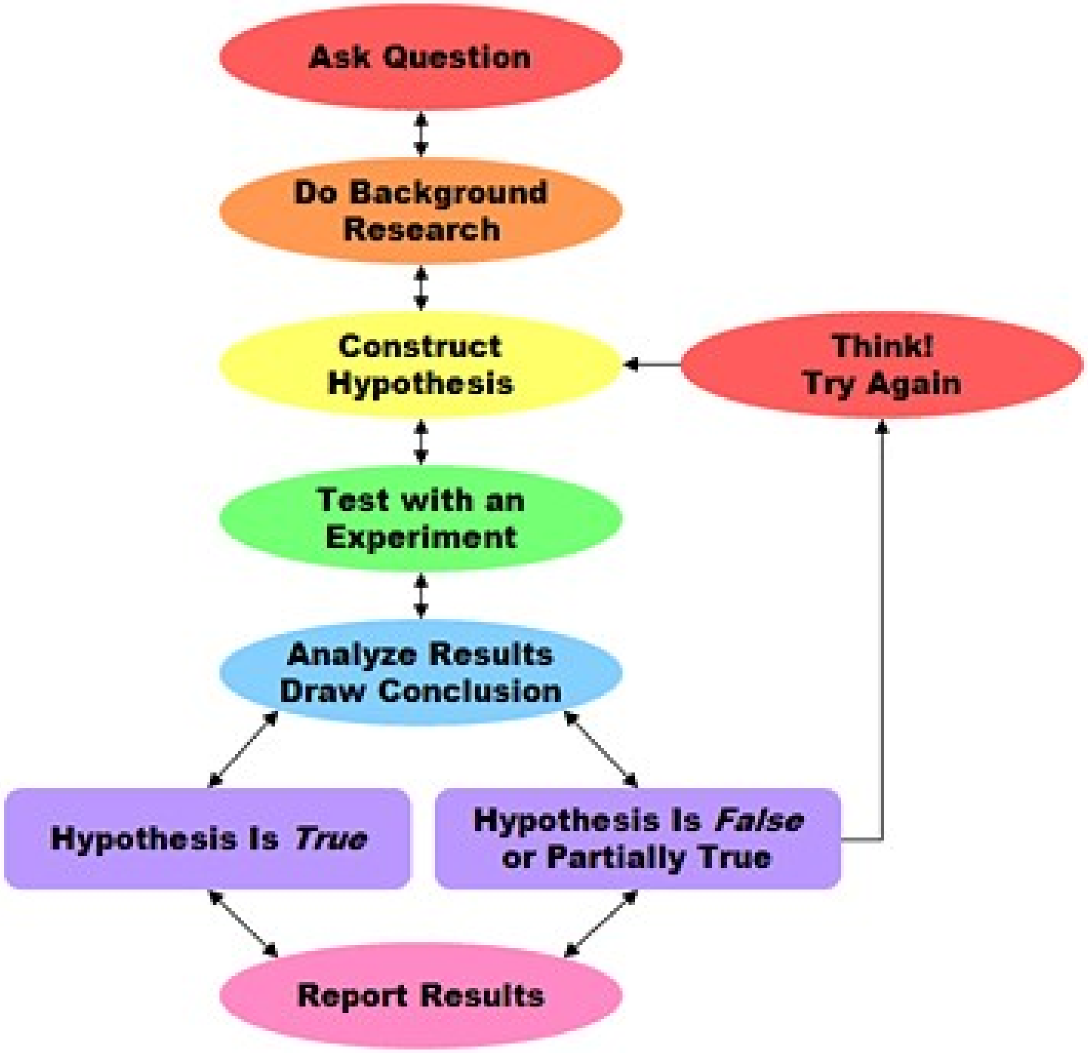
usmiusman@gmail.com

0321-5823532 / 051-9248112

Research Methodology

- Problem Selection
- Review of Existing Research
- Hypothesis
- Methodology
- Data Collection
- Analysis and Interpretation of Data
- Presentation of Results
- Conclusion

Research Process



Data Analysis Softwares

1. SPSS
2. Statistica
3. ADaMSoft
4. R-Language
5. SalStat
6. SOFA
7. Weka
8. Torch
9. PSPP
10. Sage
11. Rapid Miner
12. Stata
13. StatsDirect
14. MAPLE
15. MATLAB
16. SAS
17. Minitab
18. Mathematica

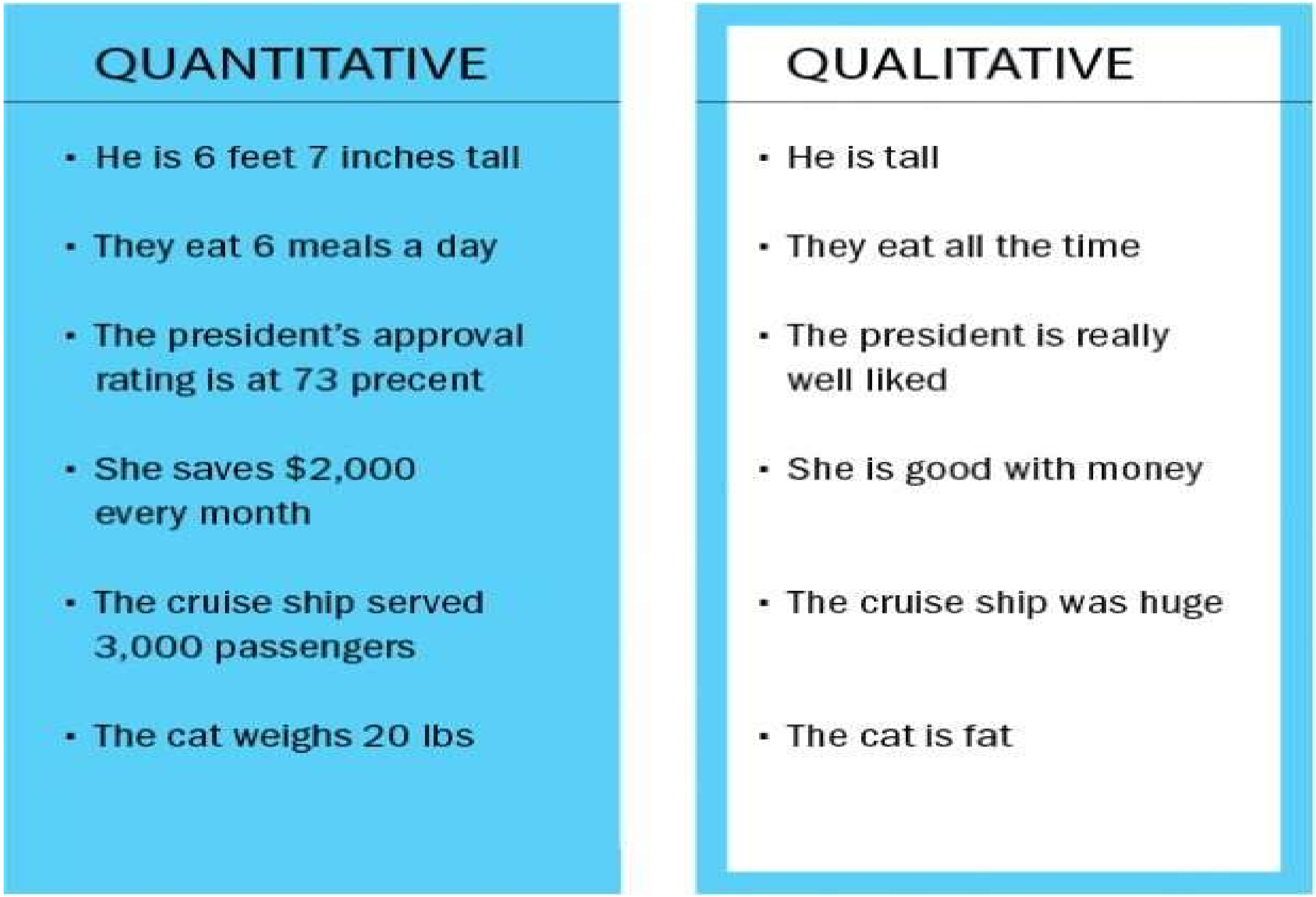
SPSS - Introduction

- Statistical Package for the Social Science
- One of the most popular statistical packages which can perform highly complex data manipulation and analysis with simple instructions
- Data Collection and Organization, Data output, Statistical Tests

SPSS – What will we cover?

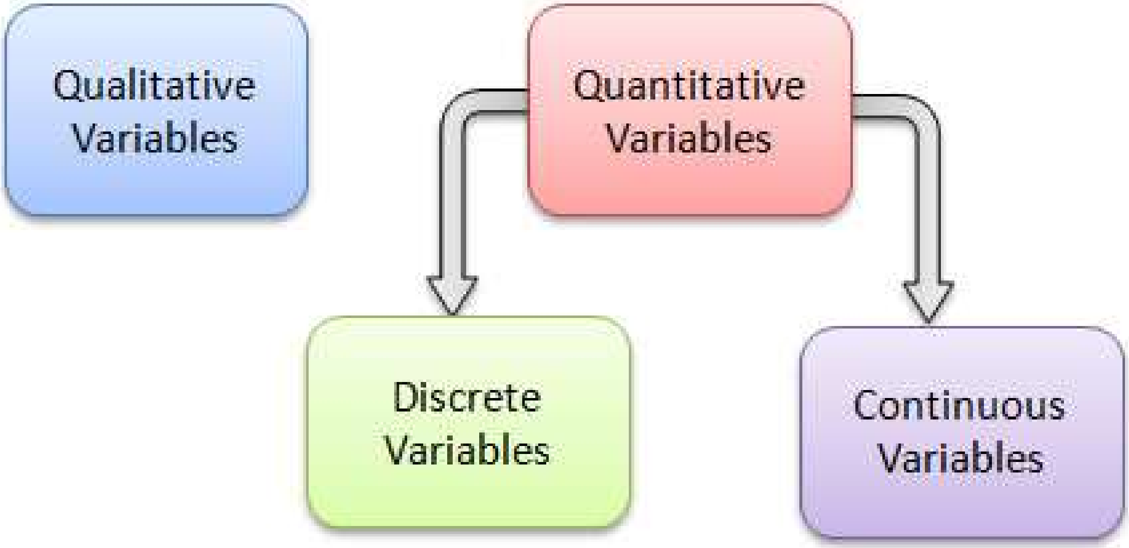
- General Data Manipulation
- Transformation/Recoding
- Handling Missing Values
- Descriptive Statistics
- t-Tests
- ANOVA
- Linear Regression

Qualitative/Quantitative Data



QUANTITATIVE	QUALITATIVE
He is 6 feet 7 inches tall	He is tall
They eat 6 meals a day	They eat all the time
The president's approval rating is at 73 precent	The president is really well liked
She saves $2,000 every month	She is good with money
The cruise ship served 3,000 passengers	The cruise ship was huge
The cat weighs 20 lbs	The cat is fat

Qualitative/Quantitative



Continuous/Discrete Variables

- A discrete variable may have only specific values within a discrete set of values. You may have as example the set of integer numbers. The values the variable may take on are only integer numbers and not the values between(between -1 and 1, the variable may take only the zero value, the rest being excluded).

- The continuous variables have no restraints in taking values within an interval of values and they are good to describe the temperature, the weight and the time.

Continuous/Discrete Variables

Discrete and Continuous Data

Discrete data can only take on certain individual values.

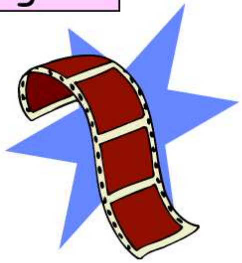
Continuous data can take on any value in a certain range.

Example 1

Number of pages in a book is a **discrete variable.**

Example 2

Length of a film is a **continuous variable.**



Example 3

Shoe size is a **Discrete variable.** E.g. 5, 5½, 6, 6½ etc. <u>Not</u> in between.

Example 4

Temperature is a **continuous variable.**

Example 5

Number of people in a race is a **discrete variable.**

Example 6

Time taken to run a race is a **continuous variable.**

Dependent/Independent Variables

- The "dependent variable" represents the output or effect, or is tested to see if it is the effect.
- A dependent variable is also known as a "response variable", "regressand", "measured variable", "responding variable", "explained variable", "outcome variable", "experimental variable", and "output variable“. (Wiki)

- The "independent variables" represent the inputs or causes, or are tested to see if they are the cause. Other variables may also be observed for various reasons.
- An independent variable is also known as a "predictor variable", "regressor", "controlled variable", "manipulated variable", "explanatory variable", "exposure variable" (see reliability theory), "risk factor" (see medical statistics), "feature" (in machine learning and pattern recognition) or an "input variable. (Wiki)

Examples

- A scientist studies the impact of a drug on cancer. The independent variables are the administration of the drug - the dosage and the timing. The dependent variable is the impact the drug has on cancer.

- A scientist studies the impact of withholding affection on rats. The independent variable is the amount of affection. The dependent variable is the reaction of the rats.

- A scientist studies how many days people can eat soup until they get sick. The independent variable is the number of days of consuming soup. The dependent variable is the onset of illness.

- The number of hours you study, and the score you get.

Data Types - General

- Numbers (Integer, Float)
- Strings (Characters)
- Binary (0,1)
- Dates

Variable Types - SPSS

- Nominal
- Ordinal
- Scale
  - Interval
  - Ratio

Nominal Variable - SPSS

- A categorical variable, also called a nominal variable, is for mutual exclusive, but not ordered, categories.

- Nominal scales are mere codes assigned to objects as labels, but these are not measurements.

- Not a measure of quantity. Measures identity and difference. People either belong to a group or they do not.

- Sometimes the numbers are used to designate category membership.

Nominal Variable Examples

- Eye Color: blue, brown, green, etc.
- Biological Sex: male, female
- Marital Status: Married, Single, Divorced, Widowed
- City: 1=Islamabad, 2=Karachi, 3=Other

Nominal Variable - Which Statistic can be applied?

Statistic	Application
Frequency Distribution and mode	YES
Median and Percentiles	NO
Add or Abstracts	NO
Mean, SD, SE of Mean	NO
Coefficient of Variation	NO

Ordinal Variables - SPSS

- This scale has the ability to rank the individual attributes of two items in the same group but unit of measurement is not available in this scale like student A is taller than student B but their actual heights are not available.

- Designates an ordering: greater than, less than

- Doesn't assume that the intervals between numbers are equal.

Ordinal Variable – Examples

- Rank your food preference where 1= favorite food, and 4 = least favorite:

- ___ pizza _____ Shawarma
- ___ Roll Pratha _____ Baryani

- Final position of horses in a thoroughbred race is an ordinal variable. The horses finish first, second, third, fourth and so on. The difference between first and second is not necessarily equivalent to the difference between second and third, or between third and fourth.

Ordinal Variable – Which Statistic can be applied?

Statistic	Application
Frequency Distribution	YES
Median and Percentiles	YES
Add or Abstracts	NO
Mean, SD, SE of Mean	NO
Coefficient of Variation	NO

Interval Variable - SPSS

- Classifies data into groups or categories
- Designates an equal interval ordering
- Interval data cannot be multiplied or divided.
- The difference in temperature between 20F and 25F is same as the difference between 76F and 81F.

Interval Variable - Examples

- Temperature in Fahrenheit is interval.
- Celsius temperature is an interval variable.
- Common IQ tests are assumed to use interval metric.

Interval Variable – Which Statistic can be applied?

Statistic	Application
Frequency Distribution	YES
Median and Percentiles	YES
Add or Abstracts	YES
Mean, SD, Regression, ANOVA, Correlation	YES
Coefficient of Variation	NO

Ratio Variable - SPSS

- This is highest level of measurement and has the properties of interval scale, coupled with fixed origin and zero point.

- It clearly defines the magnitude or value of difference between two individual items or intervals in same group.

Ratio Variable - Examples

- Measurement of heights of students in this class.

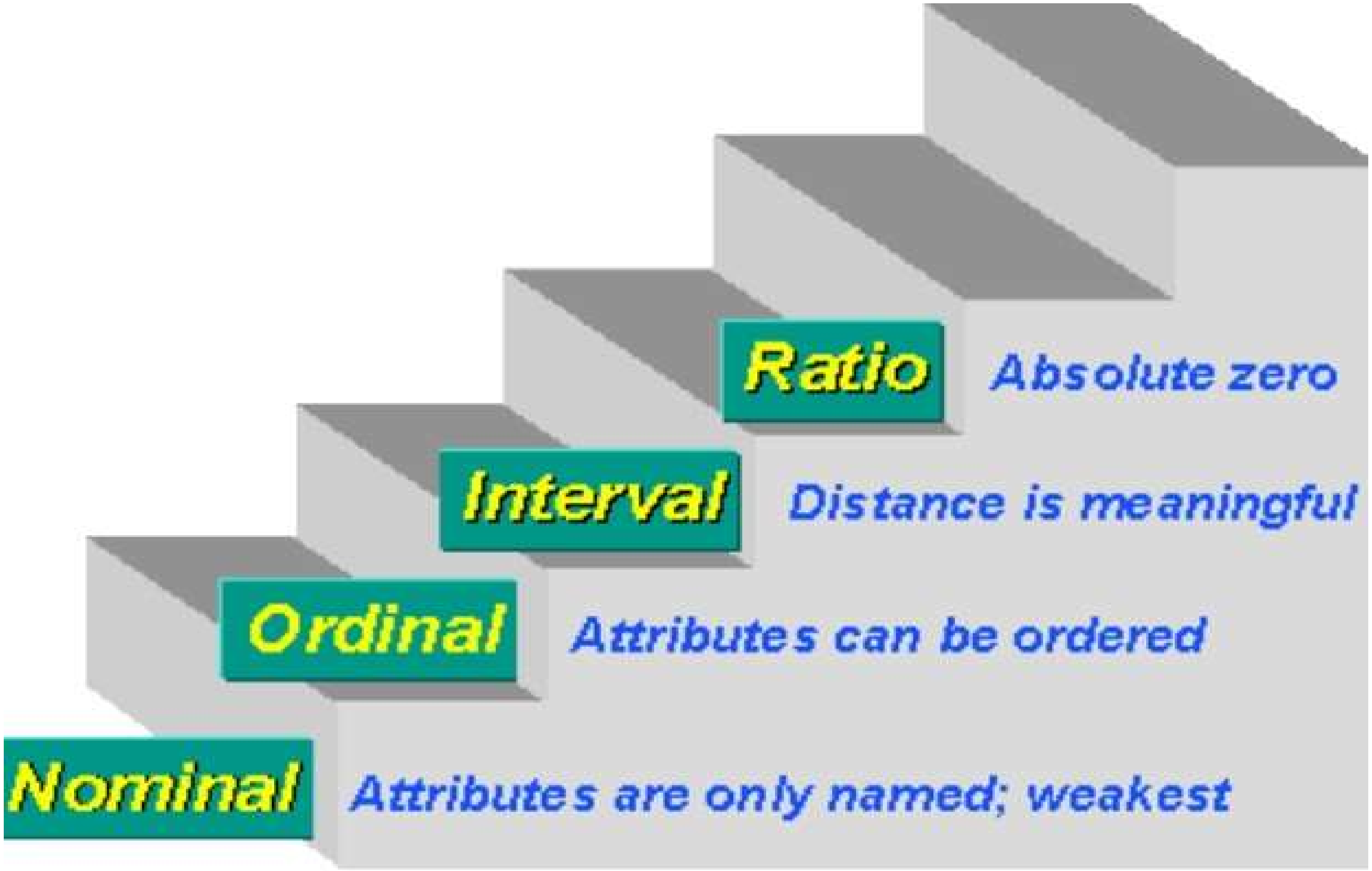
- Someone 6 ft tall is twice as tall as someone 3 feet tall.

- Heart beats per minute has a very natural zero point. Zero means no heart beats.

Ratio Variable – Which statistic can be applied?

Statistic	Application
Frequency Distribution	YES
Median and Percentiles	YES
Add or Abstracts	YES
Mean, SD, Regression, ANOVA, Correlation	YES
Coefficient of Variation	YES

Measurements - Conclusion



Ratio Absolute zero

Interval Distance is meaningful

Ordinal Attributes can be ordered

Nominal Attributes are only named; weakest

Import/Export in SPSS

- dBase (.dbf)
- Excel (.xls, .xlsx)
- Tab Delimited (.dat)
- Comma Delimited (.csv)
- SAS (.sd, .ssd)
- Stata (.dta)
- Etc.

Terms in SPSS

- Variable
- Value
- Case

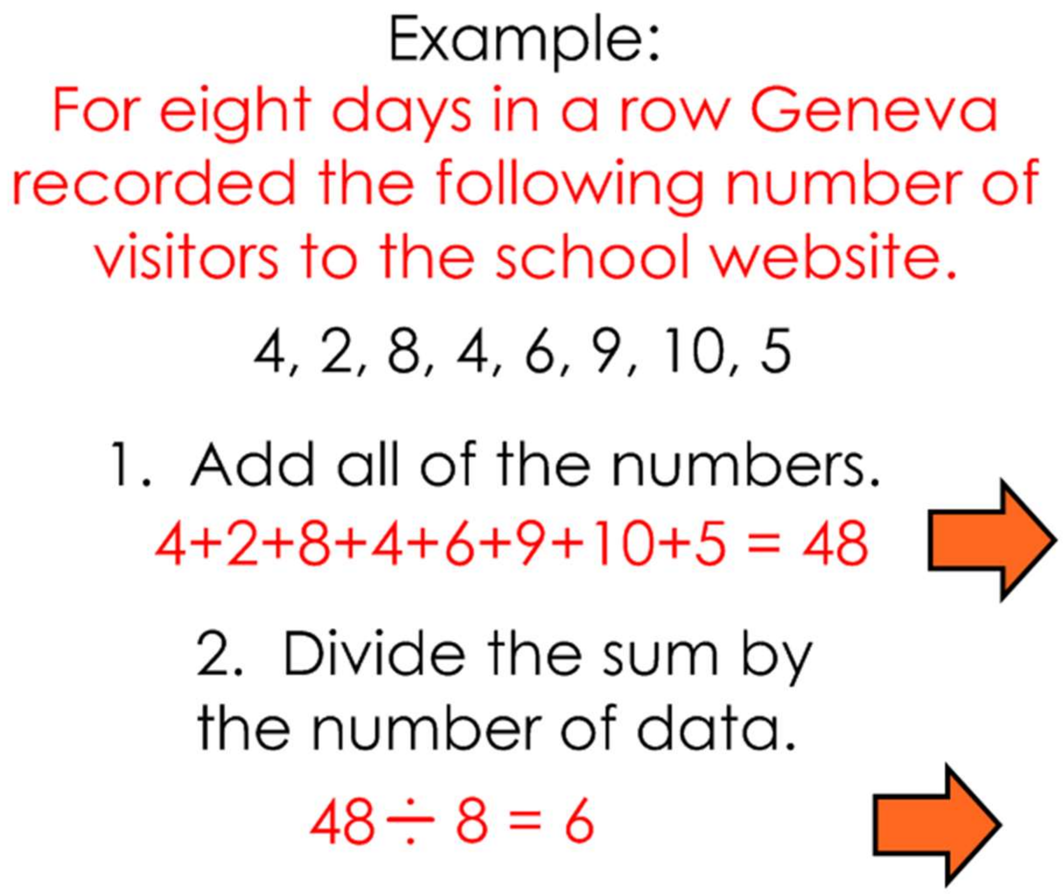
Operations in SPSS

- If-else statements
- Arithmetic Functions (Sin, Cos, Mod)
- Conversion Functions (Numbers, Strings)

Descriptive Statistics - Mean

- Average of Data

$$\bar{X} = \frac{\sum X}{N}$$



Example:

For eight days in a row Geneva recorded the following number of visitors to the school website.

4, 2, 8, 4, 6, 9, 10, 5

1. Add all of the numbers.

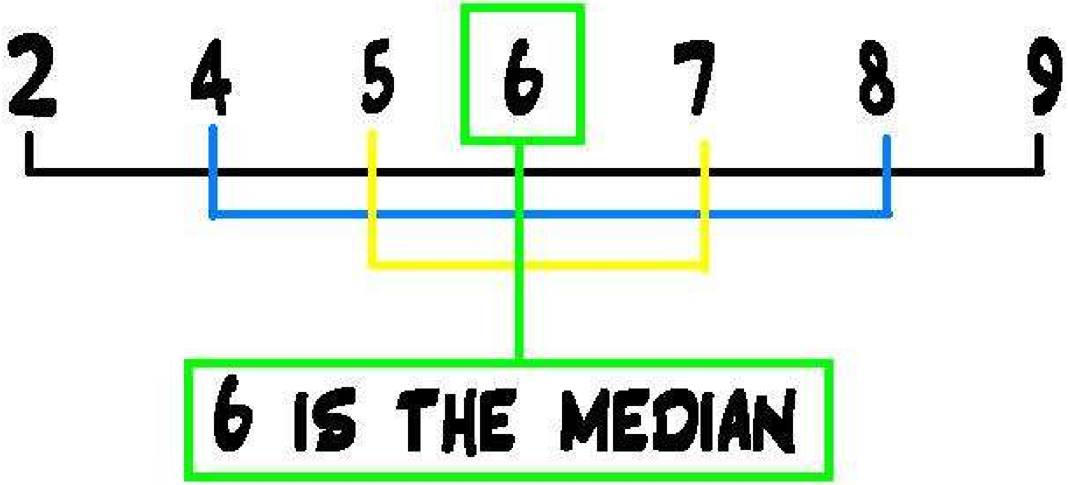
   4+2+8+4+6+9+10+5 = 48

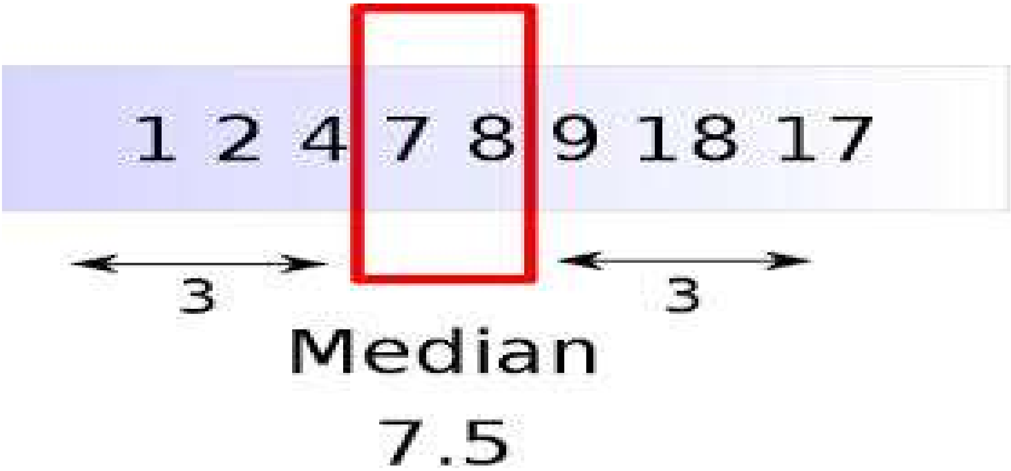
2. Divide the sum by the number of data.

   48 ÷ 8 = 6

Descriptive Statistics - Median

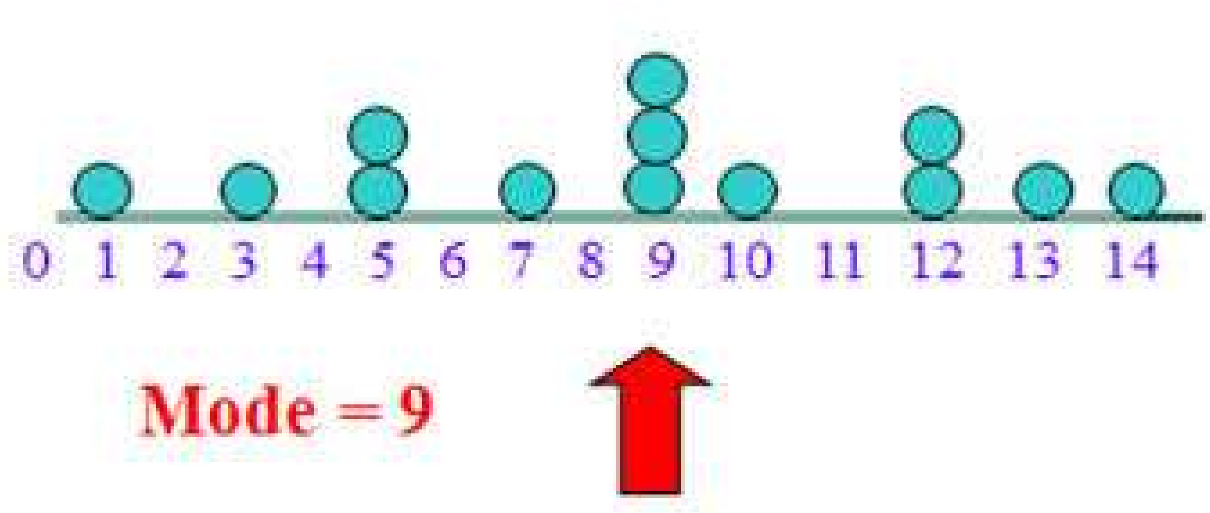
- Middle Most Value



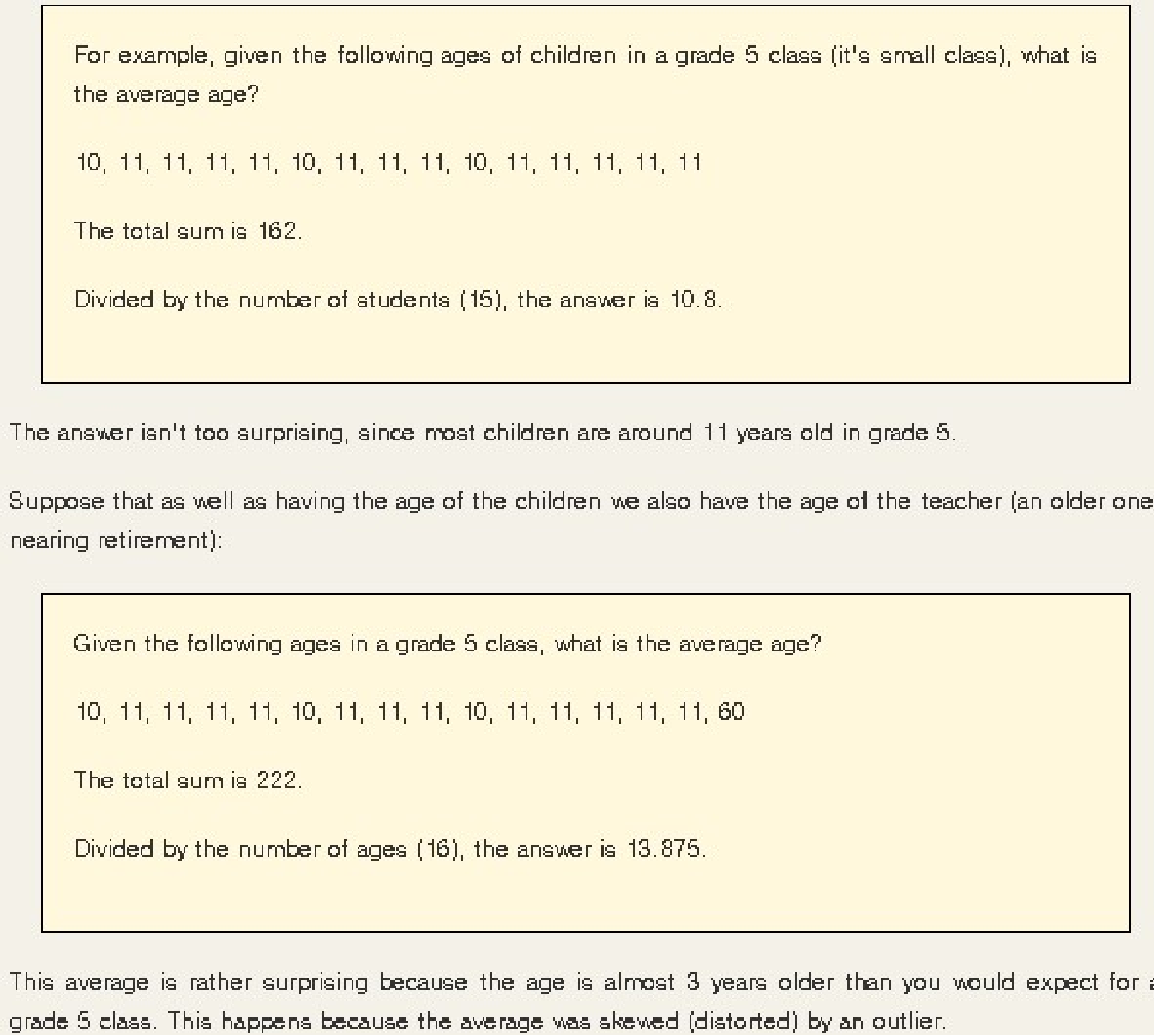


Descriptive Statistics – Mode

- Most repeated term/value



Issues with Mean



For example, given the following ages of children in a grade 5 class (it's small class), what is the average age?

10, 11, 11, 11, 11, 10, 11, 11, 11, 10, 11, 11, 11, 11, 11

The total sum is 162.

Divided by the number of students (15), the answer is 10.8.

The answer isn't too surprising, since most children are around 11 years old in grade 5.

Suppose that as well as having the age of the children we also have the age of the teacher (an older one nearing retirement):

Given the following ages in a grade 5 class, what is the average age?

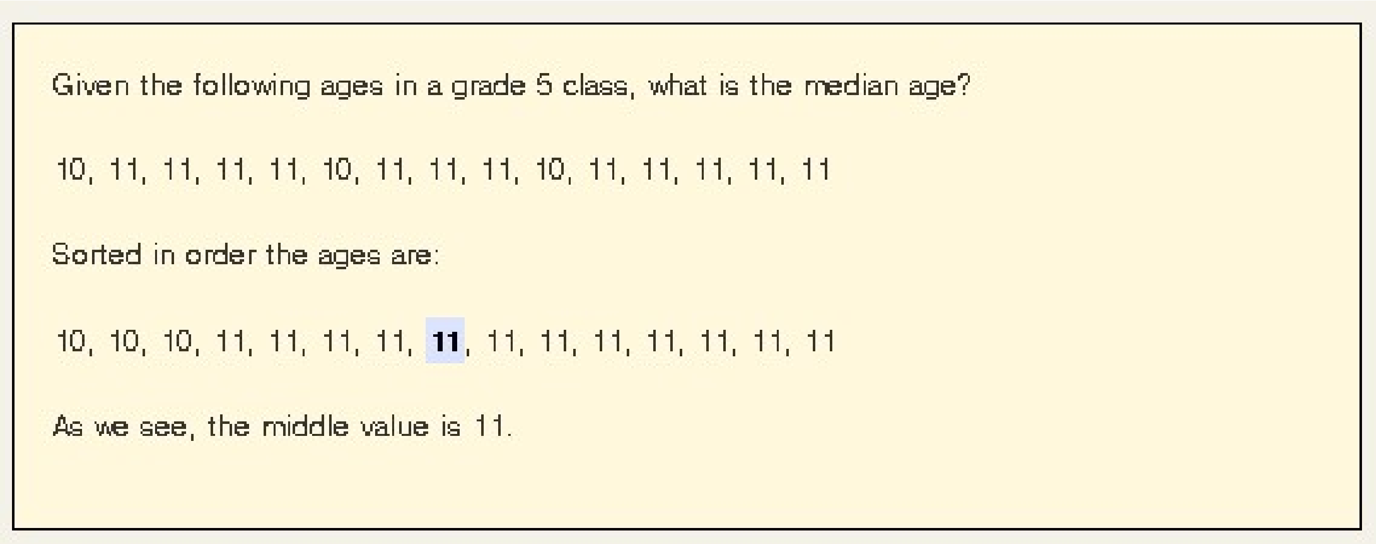
10, 11, 11, 11, 11, 10, 11, 11, 11, 10, 11, 11, 11, 11, 11, 60

The total sum is 222.

Divided by the number of ages (16), the answer is 13.875.

This average is rather surprising because the age is almost 3 years older than you would expect for a grade 5 class. This happens because the average was skewed (distorted) by an outlier.

Issues with Median



Given the following ages in a grade 5 class, what is the median age?

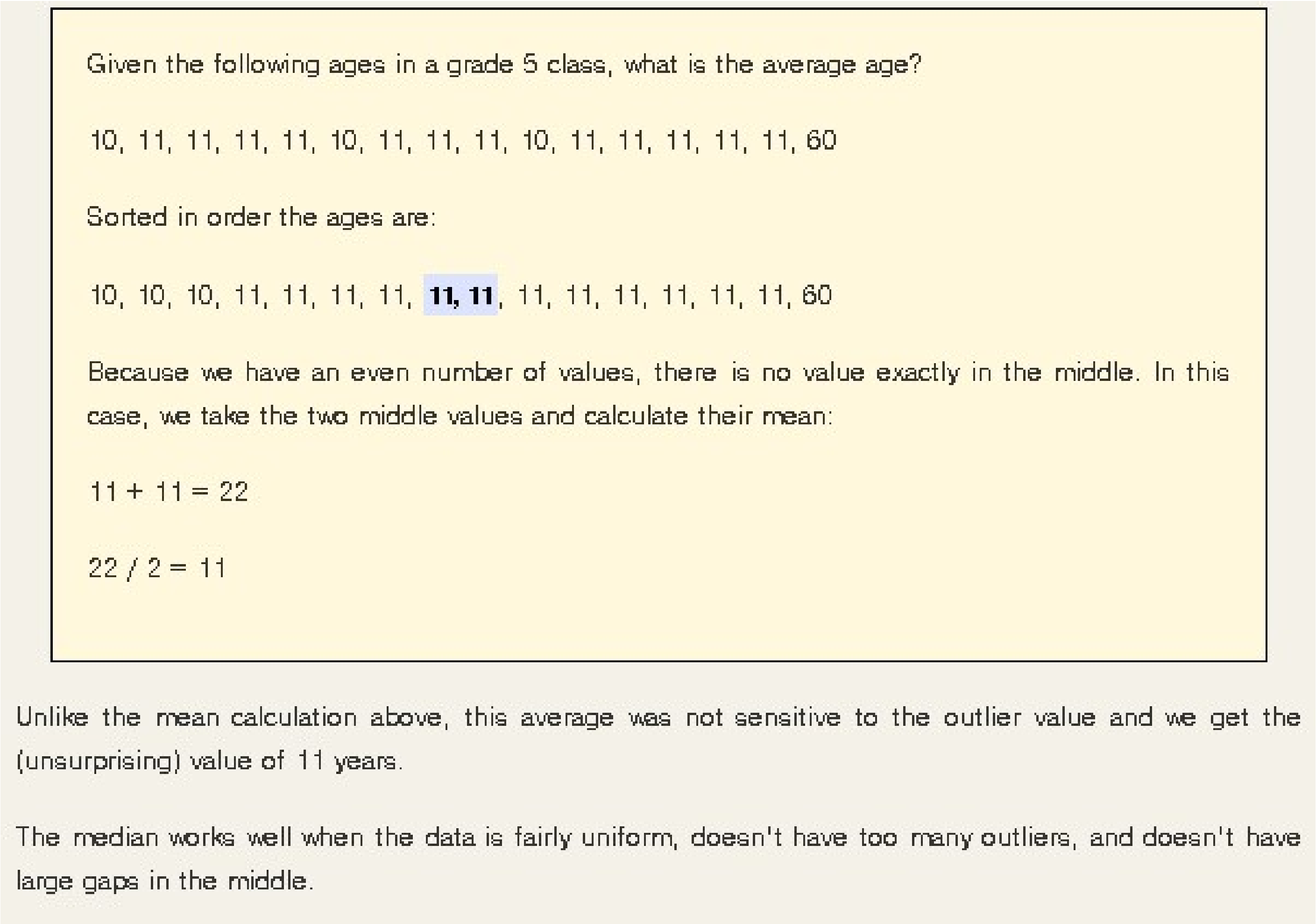
10, 11, 11, 11, 11, 10, 11, 11, 11, 10, 11, 11, 11, 11, 11

Sorted in order the ages are:

10, 10, 10, 11, 11, 11, 11, **11**, 11, 11, 11, 11, 11, 11, 11

As we see, the middle value is 11.

Issues with Median



Given the following ages in a grade 5 class, what is the average age?

10, 11, 11, 11, 11, 10, 11, 11, 11, 10, 11, 11, 11, 11, 11, 60

Sorted in order the ages are:

10, 10, 10, 11, 11, 11, 11, **11, 11**, 11, 11, 11, 11, 11, 11, 60

Because we have an even number of values, there is no value exactly in the middle. In this case, we take the two middle values and calculate their mean:

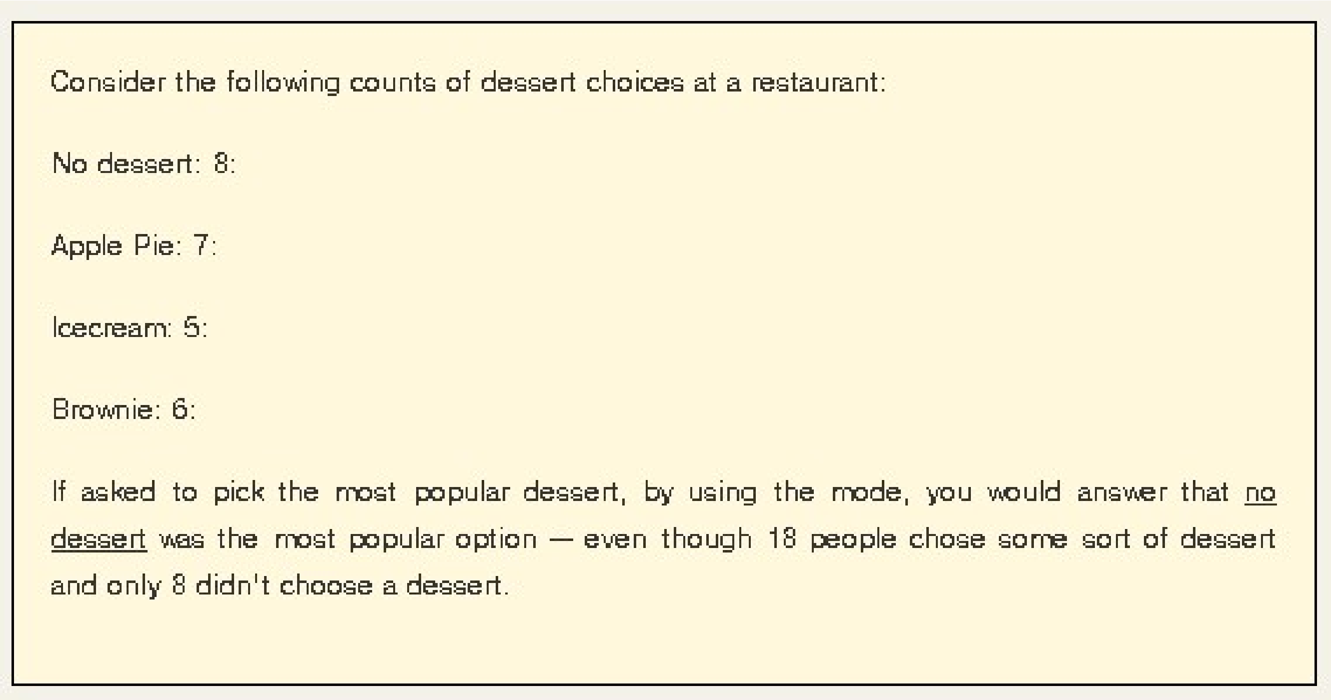
11 + 11 = 22

22 / 2 = 11

Unlike the mean calculation above, this average was not sensitive to the outlier value and we get the (unsurprising) value of 11 years.

The median works well when the data is fairly uniform, doesn't have too many outliers, and doesn't have large gaps in the middle.

Issues with Mode



Consider the following counts of dessert choices at a restaurant:

No dessert: 8:

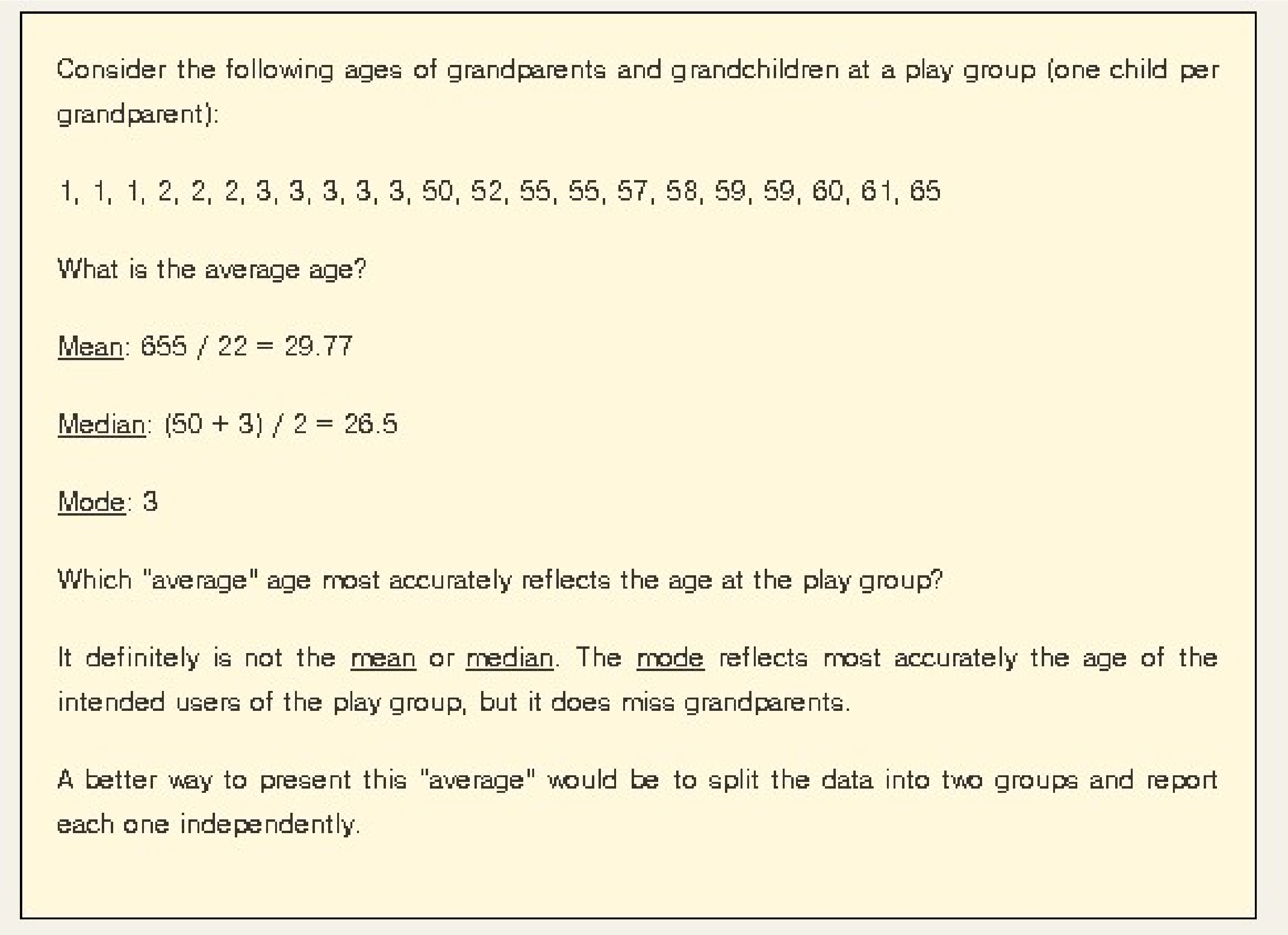
Apple Pie: 7:

Icecream: 5:

Brownie: 6:

If asked to pick the most popular dessert, by using the mode, you would answer that no dessert was the most popular option — even though 18 people chose some sort of dessert and only 8 didn't choose a dessert.

Conclusion



Consider the following ages of grandparents and grandchildren at a play group (one child per grandparent):

1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 3, 50, 52, 55, 55, 57, 58, 59, 59, 60, 61, 65

What is the average age?

Mean: 655 / 22 = 29.77

Median: (50 + 3) / 2 = 26.5

Mode: 3

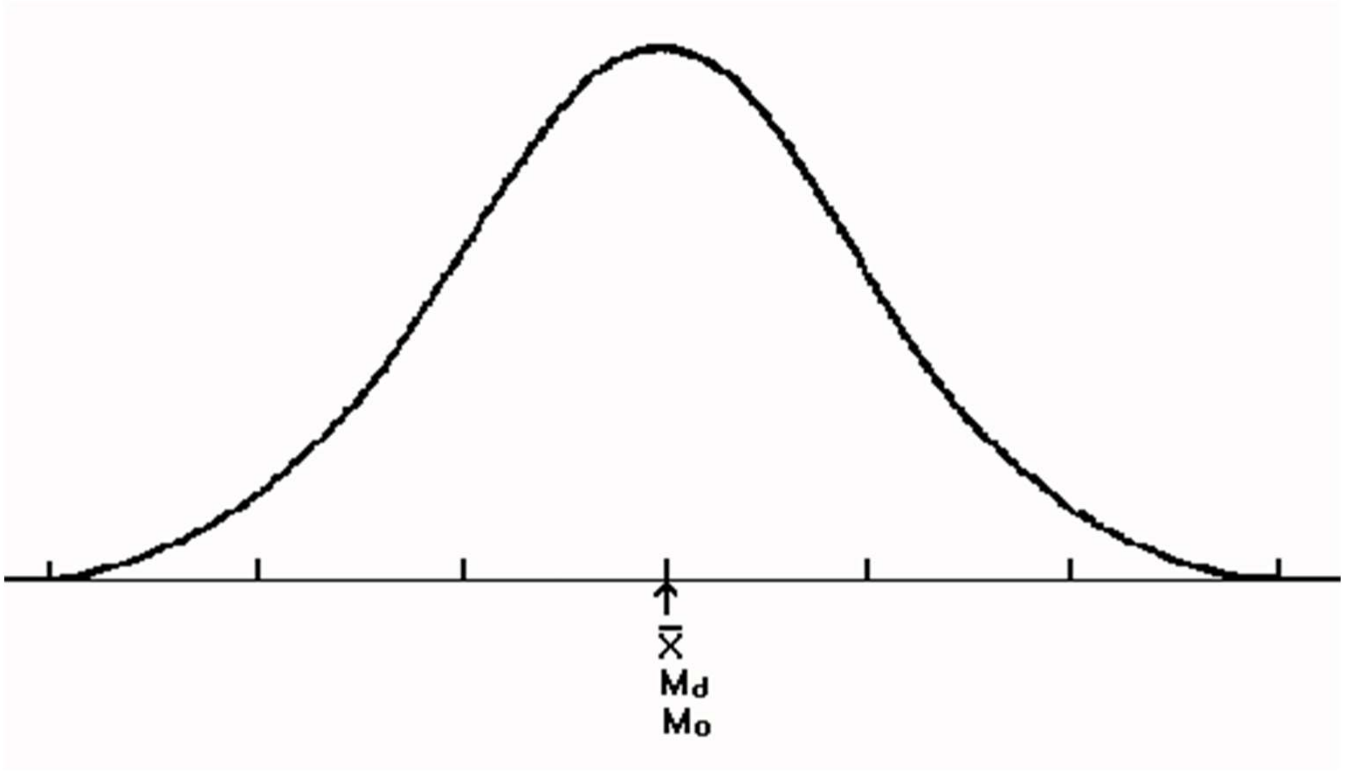
Which "average" age most accurately reflects the age at the play group?

It definitely is not the mean or median. The mode reflects most accurately the age of the intended users of the play group, but it does miss grandparents.

A better way to present this "average" would be to split the data into two groups and report each one independently.

Symmetrical Distribution

- Mean = Median = Mode



Descriptive Statistics – Standard Deviation

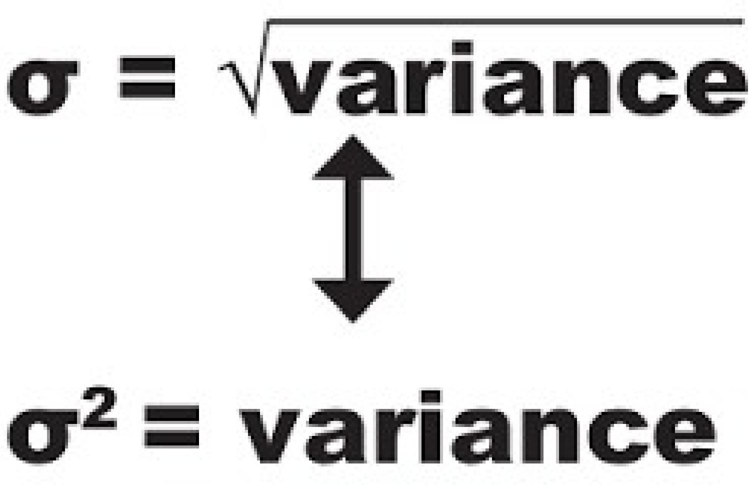
- A measure that is used to quantify the amount of variation or dispersion of a set of data values

$$\sigma = \sqrt{\frac{\Sigma\ (x - \overline{x}\ )^2}{n}}$$

σ	=	standard deviation
Σ	=	sum of
x	=	each value in the data set
$\overline{x}$	=	mean of all values in the data set
n	=	number of value in the data set

Descriptive Statistics - Variance

- The average of the squared differences from the Mean
- A measurement of the spread between numbers in a data set.
- The **variance** measures how far each number in the set is from the mean.



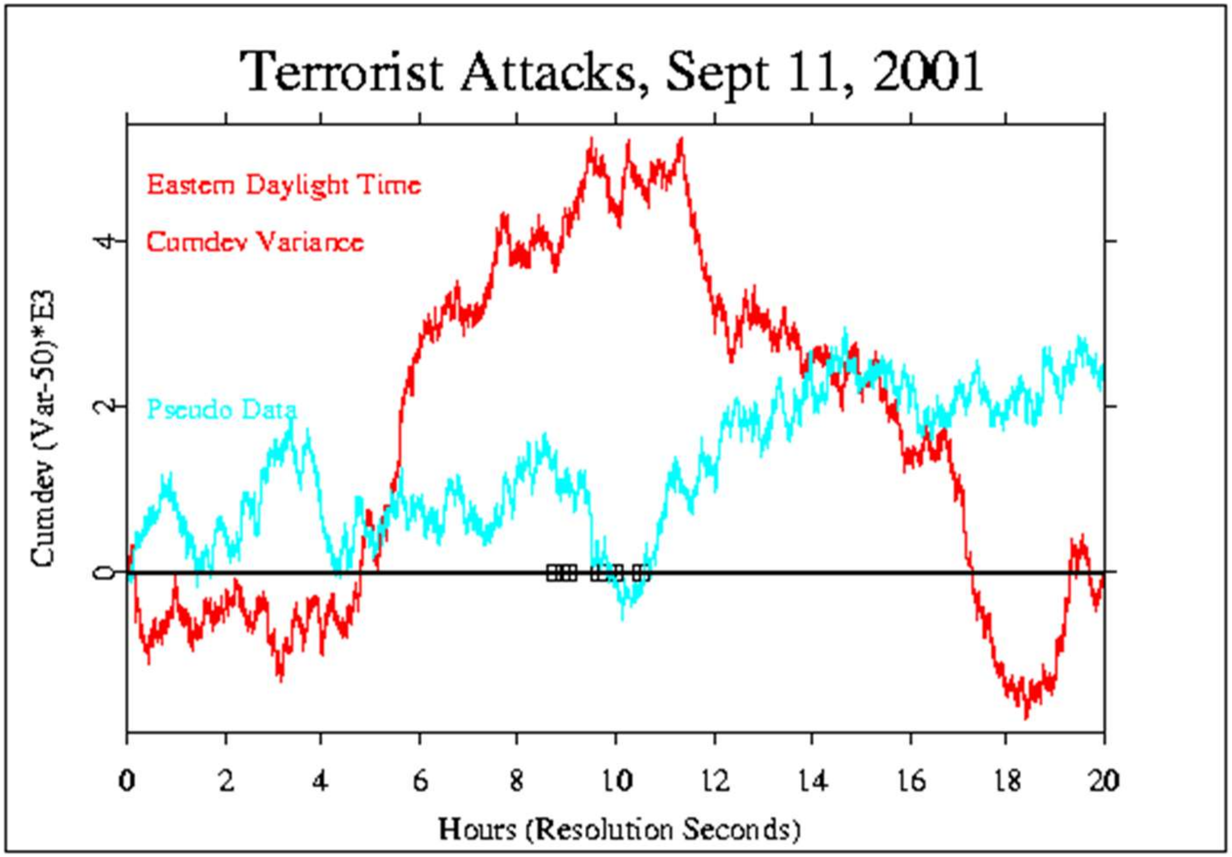
$$\sigma = \sqrt{\text{variance}}$$

$$\updownarrow$$

$$\sigma^2 = \text{variance}$$

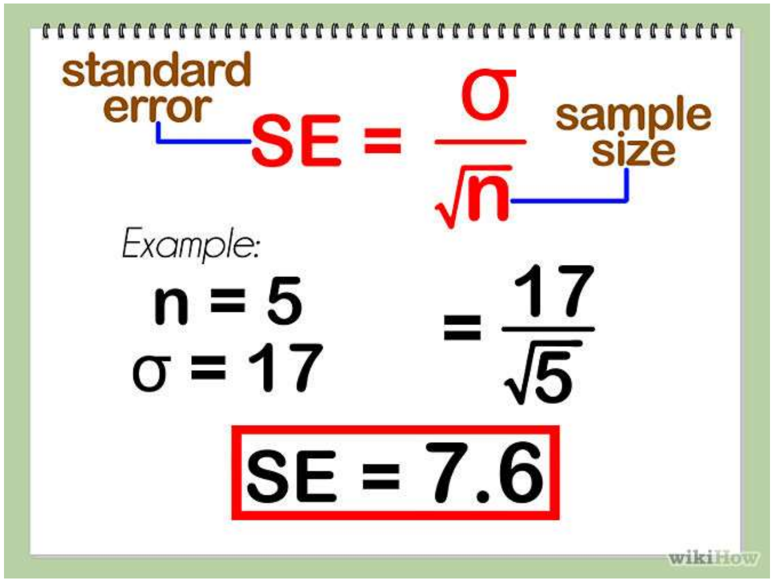
Descriptive Statistics - Variance

- Variation in data / Consistency

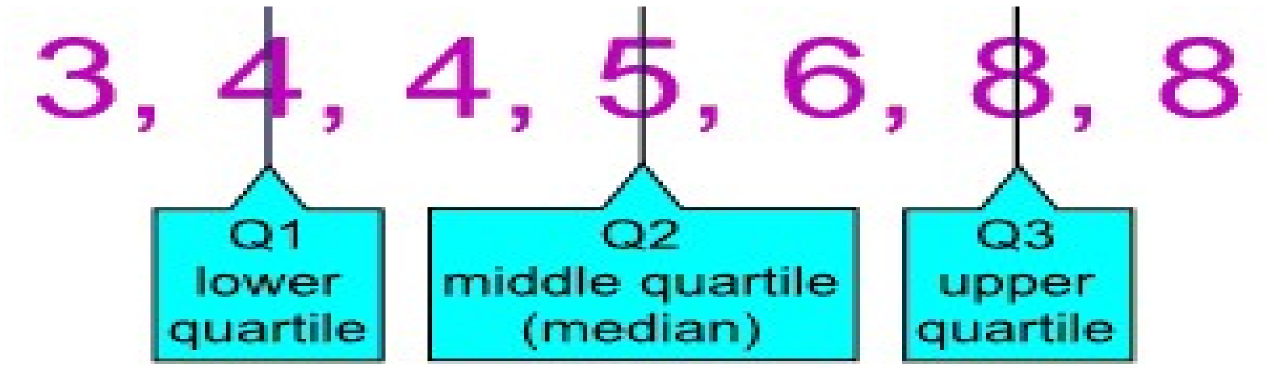


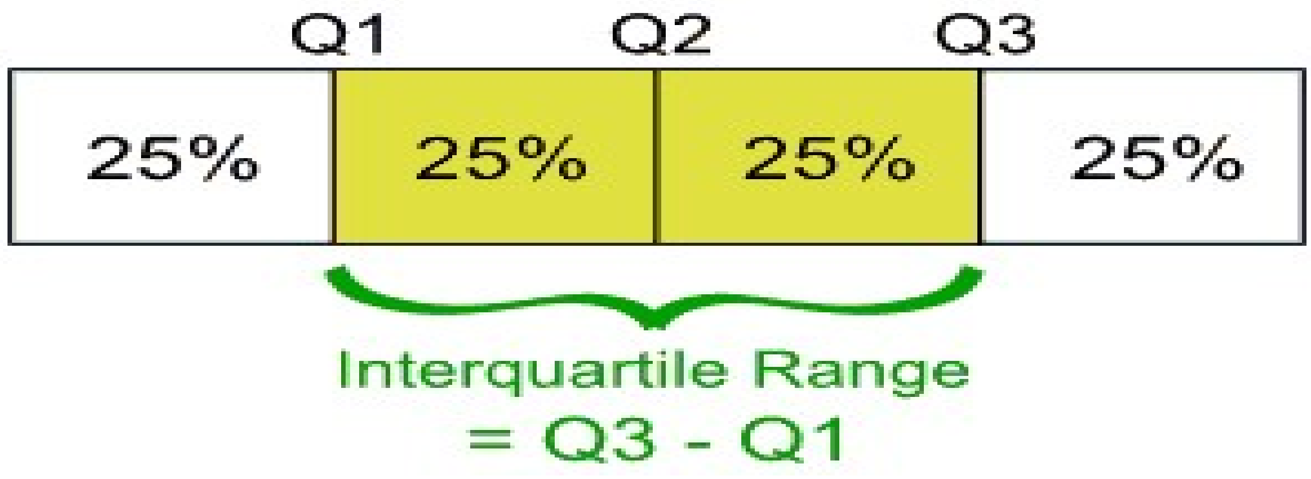
Descriptive Statistics – SE Mean

- How precisely you know the true mean of the population?
- The SEM gets smaller as your samples get larger



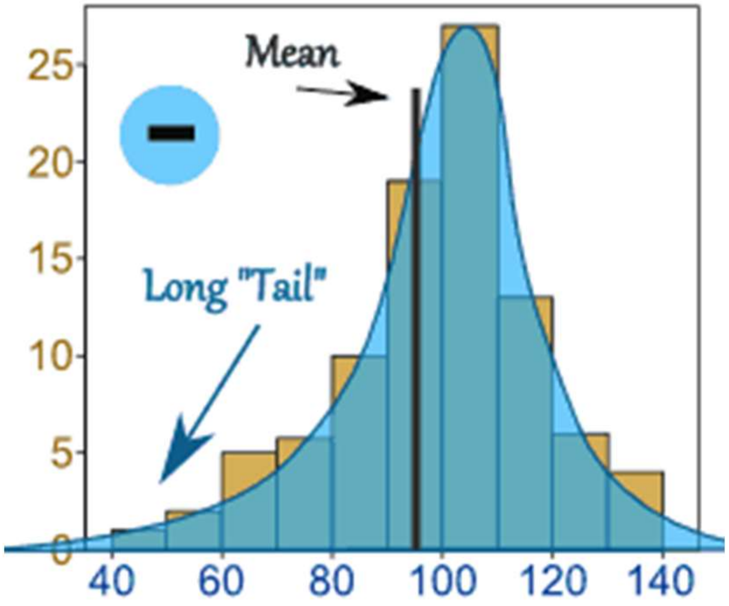
Descriptive Statistics - Quartiles

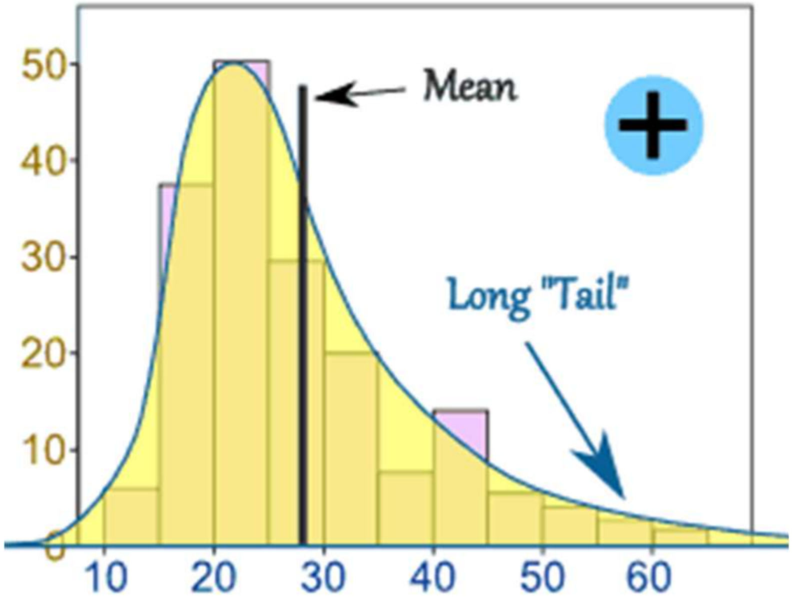




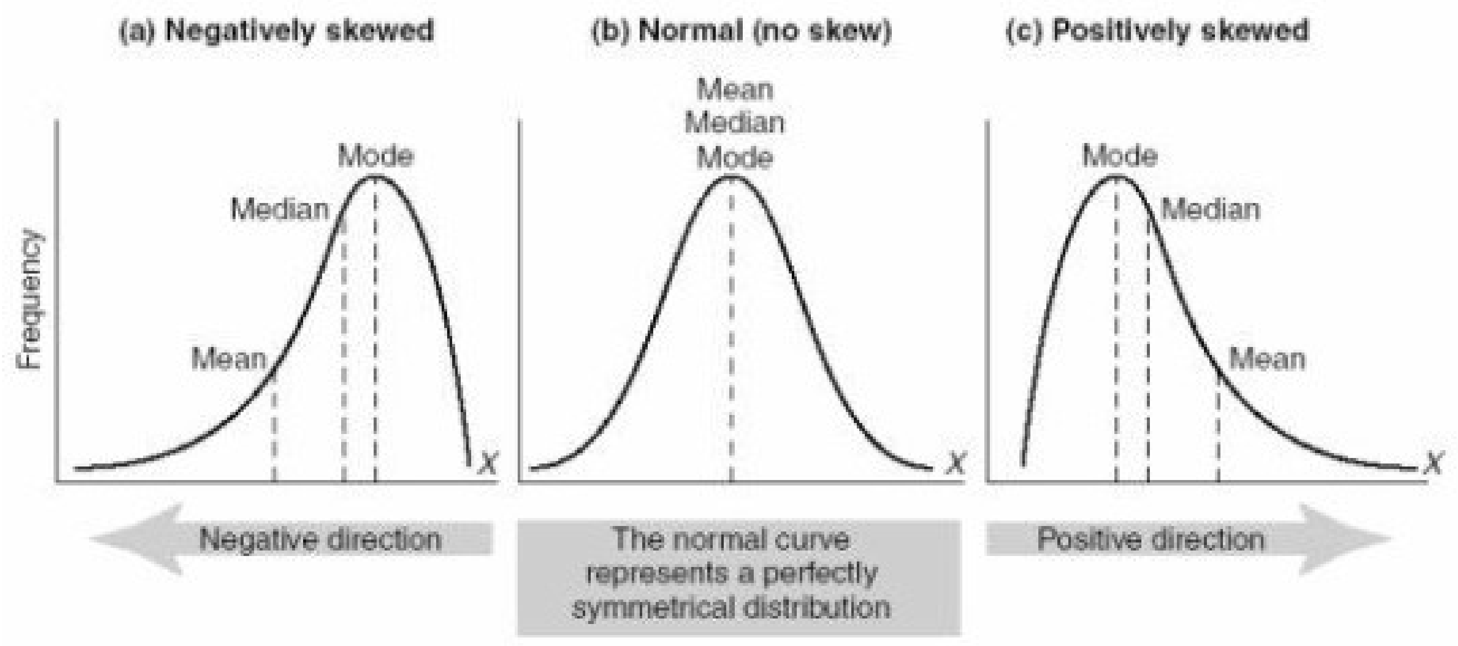
Descriptive Statistics - Skewness

A measure of asymmetry. Zero indicates perfect symmetry; the normal distribution has a skewness of zero. Positive skewness indicates that the "tail" of the distribution is more stretched on the side above the mean. Negative skewness indicates that the tail of the distribution is more stretched on the side below the mean.





Descriptive Statistics - Skewness

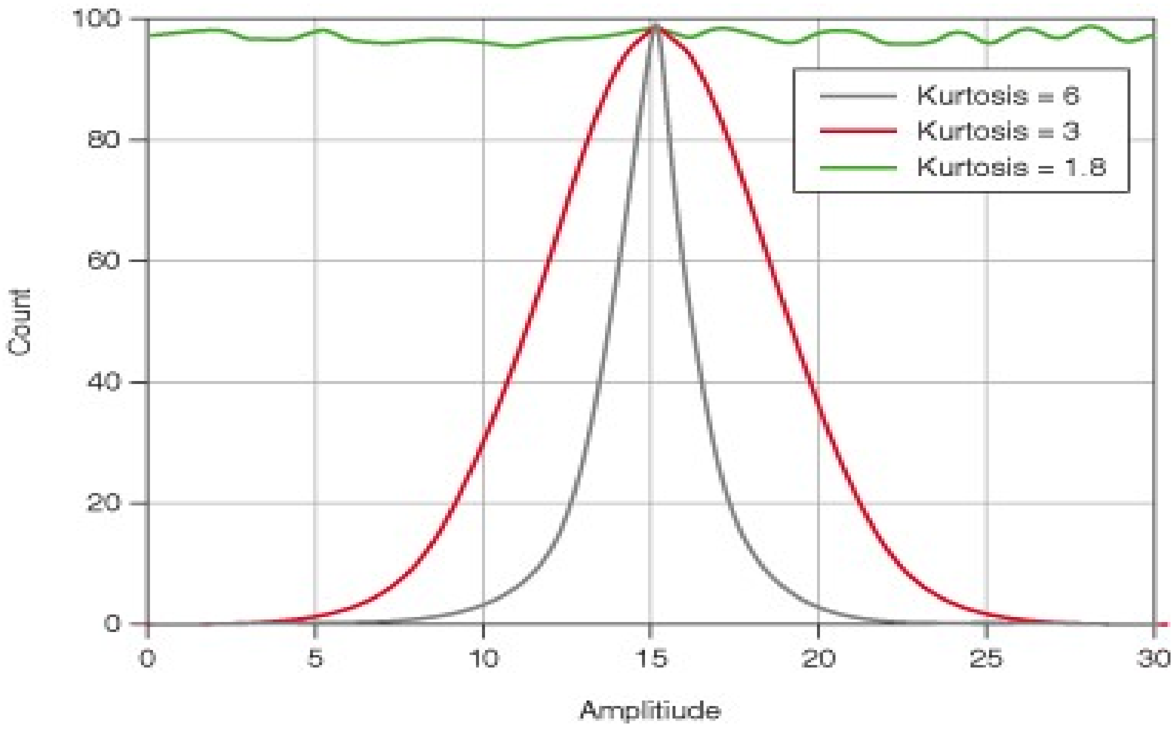


Descriptive Statistics -Kurtosis

Kurtosis is a measure of peakedness.

Distribution with Too Much Peak (Kurtosis > 0)

Too Flat Distribution (Kurtosis < 0)



More on Descriptive Statistics

- http://onlinestatbook.com/2/introduction/**descriptive**.html

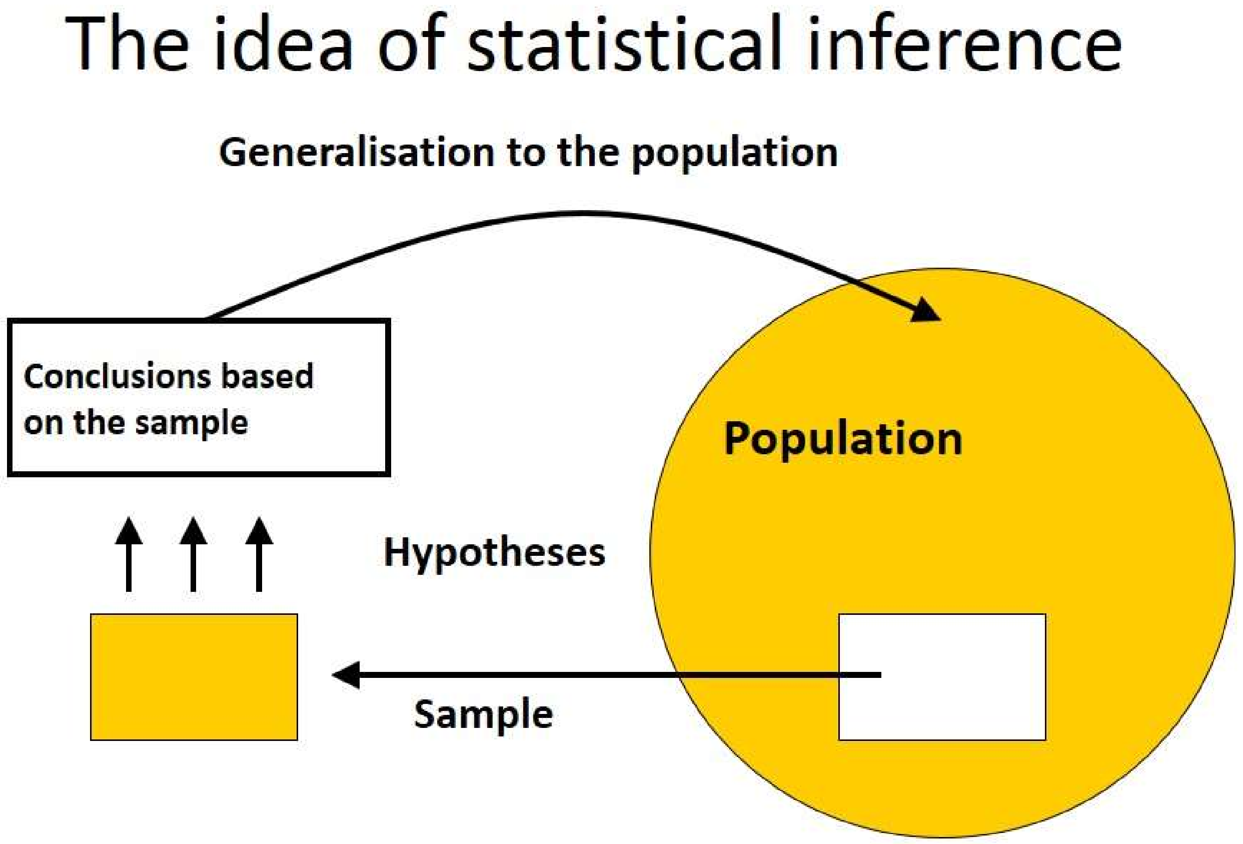
- http://mste.illinois.edu/hill/dstat/dstat.html

- http://www.ats.ucla.edu/stat/spss/modules/descript.htm (SPSS)

- Many other text books available online.

- Workshop CD

Statistical Inference



The idea of statistical inference

Generalisation to the population

Conclusions based on the sample

Hypotheses

Population

Sample

Statistics - Sample Selection

- Probability Sampling
  Random, Stratified, Systematic, Clusters

- Non-Probability Sampling
  (Quota Sampling)

- Sample Size depends upon availability of resources, manpower, budget, ethics

- More on Sample Selection : http://www.roadsafetyevaluation.com/evaluationtopics/info/tecniques-for-selecting-samples.pdf + Workshop CD

Hypothesis

- A procedure, based on sample evidence and probability theory, used to determine whether the hypothesis is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.

- Examples of hypothesis, or statements, made about a population parameter are:

1) The mean monthly income from all sources for Systems Analysts is Rs. 50,000/- in Public Sector.

2) Medicine A is more effective than Medicine B for patients of Hypertension.

Hypothesis Steps

- State the null hypothesis and the alternate hypothesis
- Select the appropriate test statistic and level of significance
- Compute the appropriate test statistic and make the decision
- Interpret the decision

t-Test

- A statistical examination of two population means.

- **One Sample -** Used to test whether the sample mean is different from the mean value of population.

- The paired t-test is used to **compare the means of two dependent samples**. OR whether there is a significant difference between the average values of the **same measurement made under two different conditions**.

  (A researcher would like to determine whether a fitness program increases flexibility. The researcher measures the flexibility (in inches) of 12 randomly selected participants before and after the fitness program.)

- **Independent Sample t-test -** that uses separate samples for each treatment condition. Use this test when the population mean and standard deviation are unknown, and 2 separate groups are being compared. For example, Do males and females differ in terms of their exam scores?

One-Sample t-Test

- The mean daily wages salary in Pakistan is Rs.500.00.

- The mean age of Class 1 is 4 years.

- The average score of first innings at MCG is 254 runs in one-day matches.

- On average students score 70 marks in Fundamentals of Chemistry every year at QAU. (out of 100)

Paired Sample t-Test

- Where subjects are tested prior to a treatment, say for high blood pressure, and the same subjects are tested again after treatment with a blood-pressure lowering medication.

- A farmer decides to try out a new fertilizer on a test plot containing 10 stalks of corn. Before applying the fertilizer, he measures the height of each stalk. Two weeks later, he measures the stalks again, being careful to match each stalk's new height to its previous one.

- We want to compare the mean for user-created videos and the mean for company generated videos.

- Did Math students score better on the second exam than the first?

Independent Samples

- **Whether first year graduate salaries differed based on gender**

  Your dependent variable would be "first year graduate salaries" and your independent variable would be "gender", which has two groups: "male" and "female“.

- **Whether there is a difference in test anxiety based on educational level**

  Your dependent variable would be "test anxiety" and your independent variable would be "educational level", which has two groups: "undergraduates" and "postgraduates“.

- **You wish to know whether the average height for women is significantly (statistically) different from the average height for men.**

  This involves testing whether the sample means for height among female and male subjects in your sample are statistically different (and by extension, inferring whether the means for height in the population are significantly different). You can use an Independent Samples *t* Test to compare the mean heights for males and females.

ANOVA

- Analysis Of Variance
- Analyze the differences between more than two means
- ANOVA of two group means equals t-Test

ANOVA

- Susan Sound predicts that students will learn most effectively with a constant background sound, as opposed to an unpredictable sound or no sound at all. She randomly divides twenty-four students into three groups of eight.
- All students study a passage of text for 30 minutes.
- Those in group 1 study with background sound at a constant volume in the background.
- Those in group 2 study with noise that changes volume periodically.
- Those in group 3 study with no sound at all.
- After studying, all students take a 10 point multiple choice test over the material.

group	test scores							
1) constant sound	7	4	6	8	6	6	2	9
2) random sound	5	5	3	4	4	7	2	2
3) no sound	2	4	7	1	2	1	5	5

Another Example

- Imagine that you are running an experiment to see if there is a relationship between people's religion and what they consider the ideal family size to be. You would likely do this by recruiting individuals from different religious groups and asking them to report what they consider the ideal amount of children in a family should be. Let us further say that you ended up recruiting 10 Catholics, 10 Protestants, and 10 Jewish individuals to answer this question.

- In this case, you have one **independent variable**, which is *religion*, that is thought to have an effect on the opinion of ideal family size, which is the **dependent variable** in this scenario. Religion should affect the ideal family size and *since it is the factor thought to influence the difference, it is the independent variable*. Additionally, this experiment includes three different levels of the independent variable. In this case, the three levels are the three different groups of religions in which one is Catholic, one is Protestant, and one is Jewish.

So ..

- The ANOVA can come in handy in a large number of real life situations.
- For instance, in the social sciences there is much research devoted to figuring out what factors influence people's opinions and behaviors.
- The previous example involving religion and number of children fits into this category.

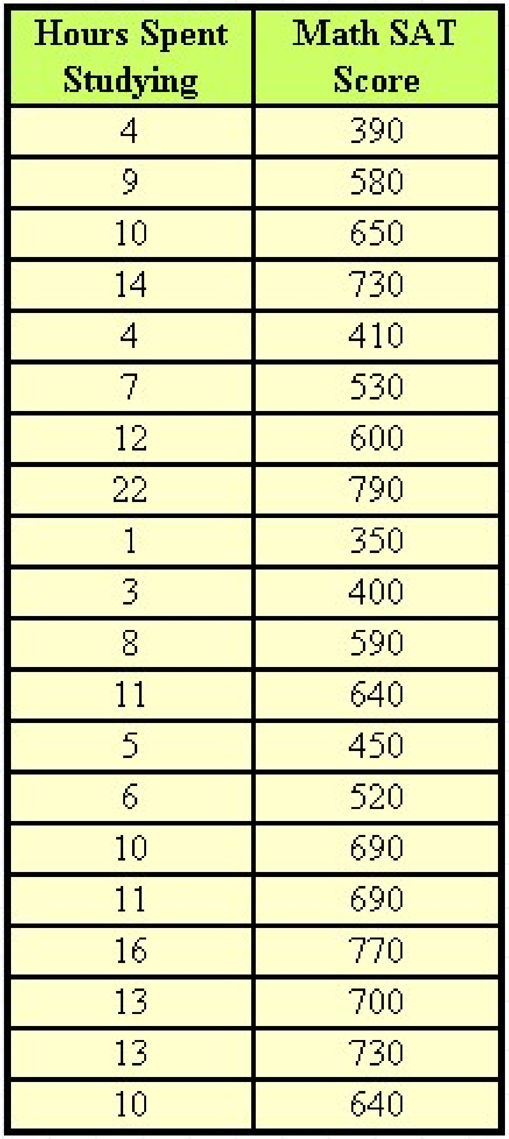
Linear Regression

- We will examine the relationship between quantitative variables x and y via a mathematical equation.
- The motivation for using the technique:
  - Forecast the value of a dependent variable (y) from the value of independent variables (x_1, x_2,…x_k.).
  - Analyze the specific relationships between the independent variables and the dependent variable.

Linear Regression

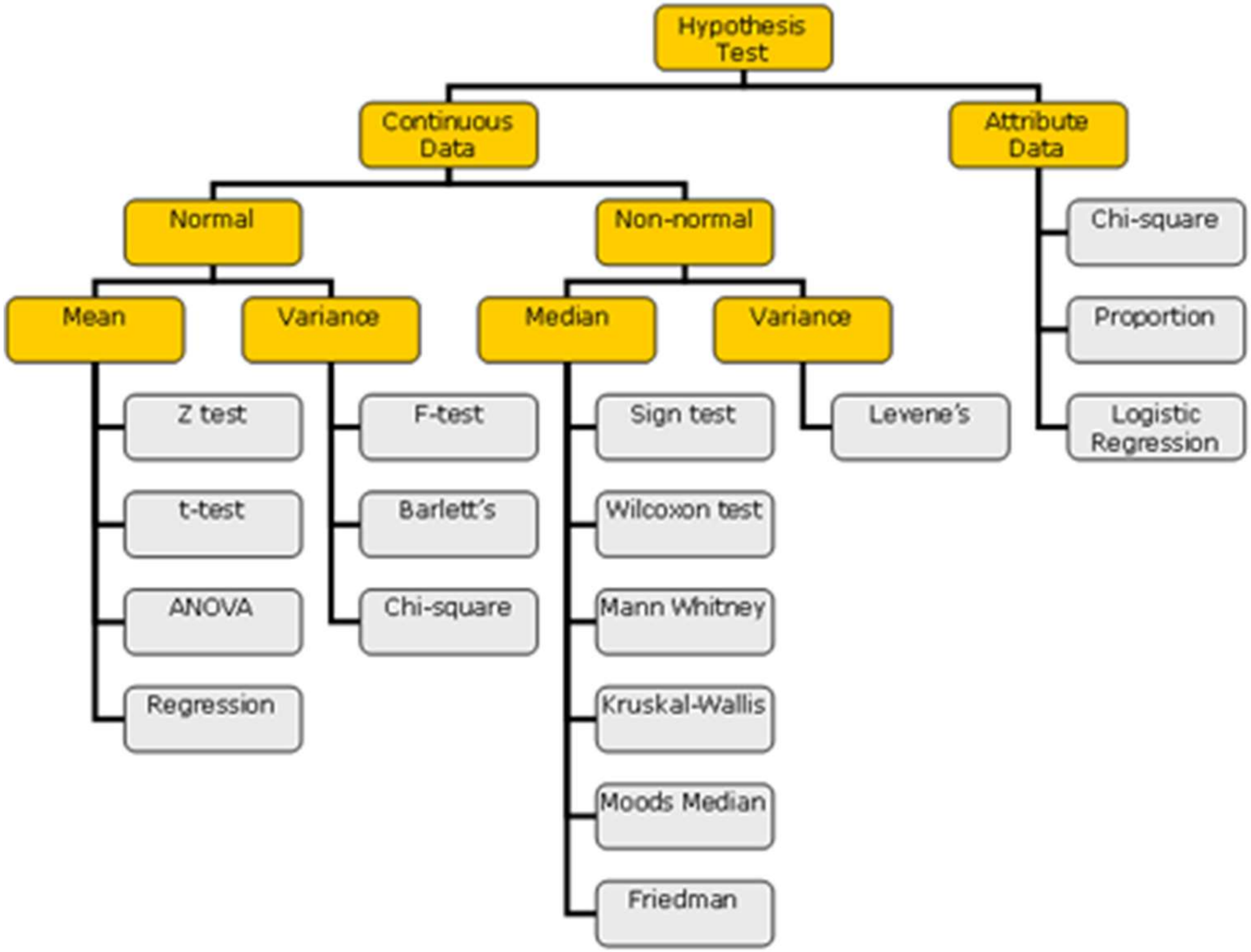
Is there a relationship between Math SAT scores and the number of hours spent studying for the test?

A study was conducted involving 20 students as they prepared for and took the Math section of the SAT Examination.

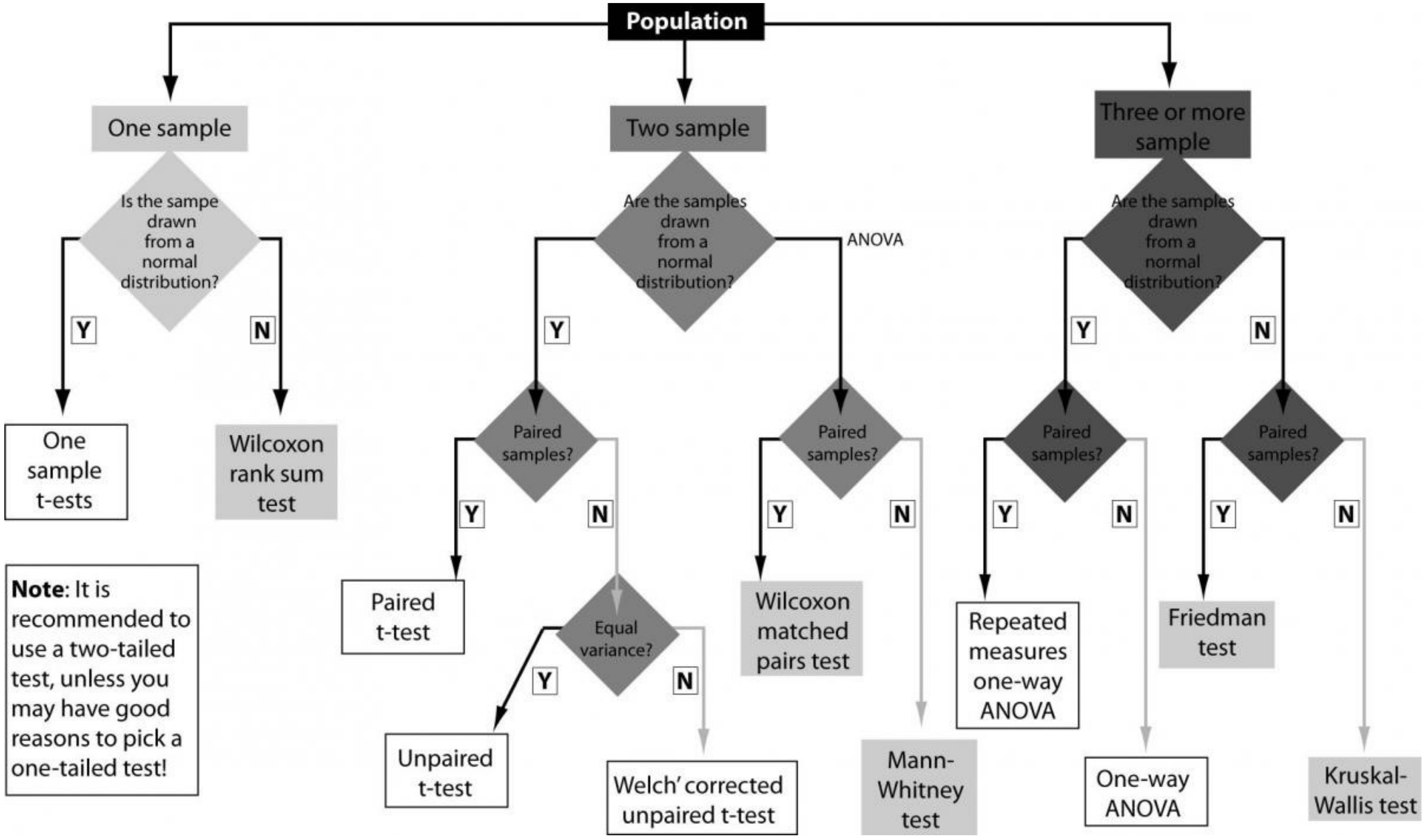


Hours Spent Studying	Math SAT Score
4	390
9	580
10	650
14	730
4	410
7	530
12	600
22	790
1	350
3	400
8	590
11	640
5	450
6	520
10	690
11	690
16	770
13	700
13	730
10	640

Test Selection



Test Selection



Population

One sample

Is the sampe drawn from a normal distribution?

Y

One sample t-ests

N

Wilcoxon rank sum test

Two sample

Are the samples drawn from a normal distribution?

Y

Paired samples?

Y

Paired t-test

N

Equal variance?

Y

Unpaired t-test

N

Welch’ corrected unpaired t-test

ANOVA

Paired samples?

Y

Wilcoxon matched pairs test

N

Mann-Whitney test

Three or more sample

Are the samples drawn from a normal distribution?

Y

Paired samples?

Y

Repeated measures one-way ANOVA

N

One-way ANOVA

N

Paired samples?

Y

Friedman test

N

Kruskal-Wallis test

Note: It is recommended to use a two-tailed test, unless you may have good reasons to pick a one-tailed test!

More on Test Selection

- www.cios.org/readbook/rmcs/ch19.pdf
- www.youtube.com/watch?v=INKjTdXbvXA
- http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3116565/
- http://www.ats.ucla.edu/stat/mult_pkg/whatstat/

Followed By..

- Hands on SPSS/Weka

Thank You !