Measures of Central Tendency

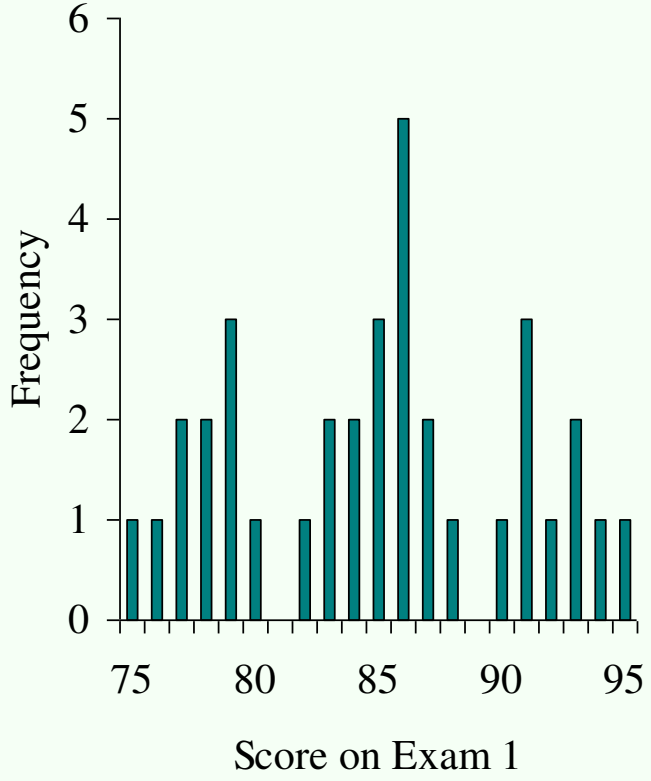
Dr P.SUDARKODI

Measures of Central Tendency

- A *measure of central tendency* is a descriptive statistic that describes the average, or typical value of a set of scores
- There are three common measures of central tendency:
  - the mode
  - the median
  - the mean

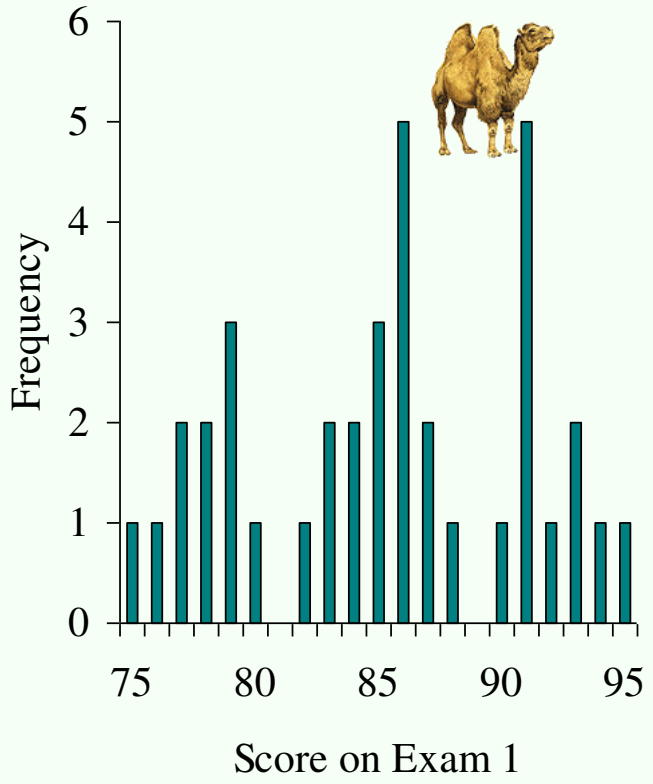
The Mode

- The *mode* is the score that occurs most frequently in a set of data



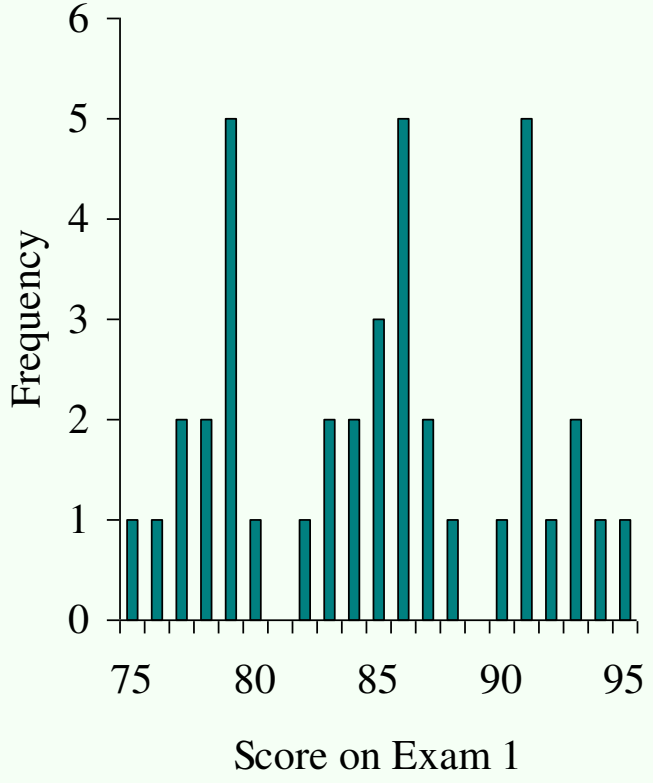
Bimodal Distributions

- When a distribution has two "modes," it is called *bimodal*



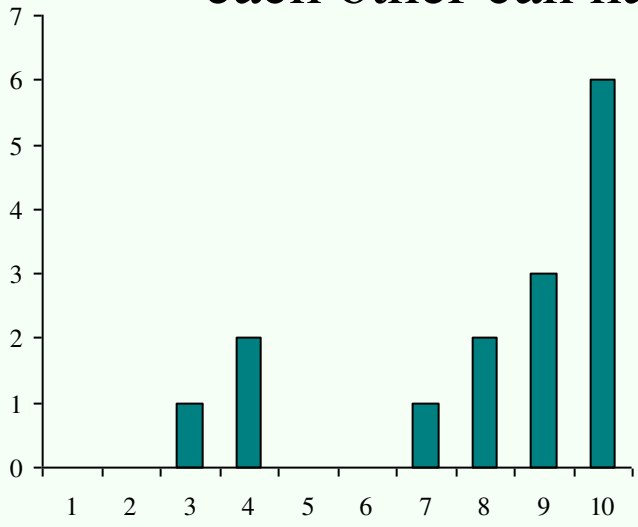
Multimodal Distributions

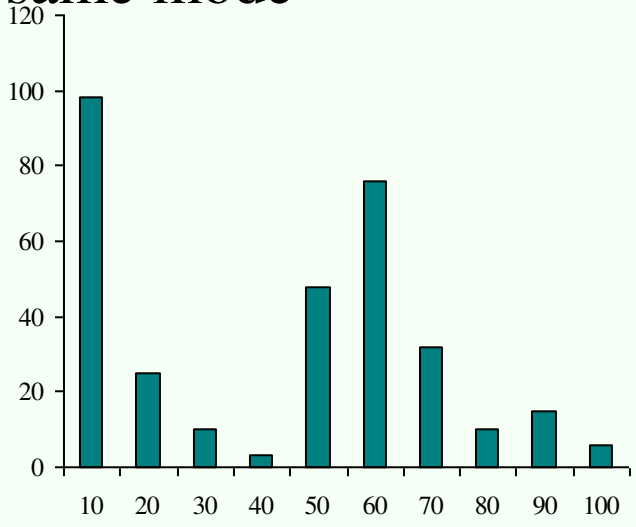
- If a distribution has more than 2 “modes,” it is called *multimodal*



When To Use the Mode

- The mode is not a very useful measure of central tendency
  - It is insensitive to large changes in the data set
    - That is, two data sets that are very different from each other can have the same mode





When To Use the Mode

- The mode is primarily used with nominally scaled data
  - It is the only measure of central tendency that is appropriate for nominally scaled data

The Median

- The *median* is simply another name for the 50th percentile
  - It is the score in the middle; half of the scores are larger than the median and half of the scores are smaller than the median

How To Calculate the Median

- Conceptually, it is easy to calculate the median
  - There are many minor problems that can occur; it is best to let a computer do it
- Sort the data from highest to lowest
- Find the score in the middle
  - middle = (N + 1) / 2
  - If N, the number of scores, is even the median is the average of the middle two scores

Median Example

- What is the median of the following scores:
  10 8 14 15 7 3 3 8 12 10 9
- Sort the scores:
  15 14 12 10 10 9 8 8 7 3 3
- Determine the middle score:
  middle = (N + 1) / 2 = (11 + 1) / 2 = 6
- Middle score = median = 9

Median Example

- What is the median of the following scores:
  24 18 19 42 16 12
- Sort the scores:
  42 24 19 18 16 12
- Determine the middle score:
  middle = (N + 1) / 2 = (6 + 1) / 2 = 3.5
- Median = average of 3rd and 4th scores:
  (19 + 18) / 2 = 18.5

When To Use the Median

- The median is often used when the distribution of scores is either positively or negatively skewed
  - The few really large scores (positively skewed) or really small scores (negatively skewed) will not overly influence the median

The Mean

- The *mean* is:
  - the arithmetic average of all the scores $(\Sigma X)/N$
  - the number, m, that makes $\Sigma(X - m)$ equal to 0
  - the number, m, that makes $\Sigma(X - m)^2$ a minimum
- The mean of a population is represented by the Greek letter μ; the mean of a sample is represented by $\bar{X}$

Calculating the Mean

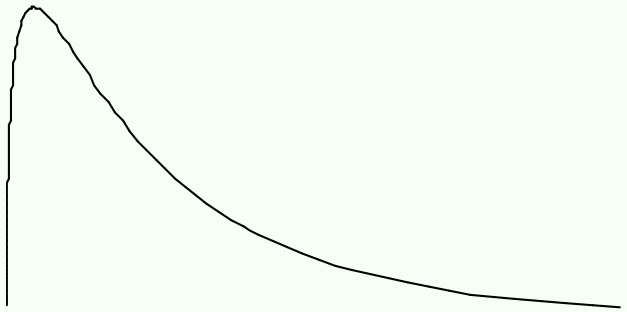
- Calculate the mean of the following data:
  1 5 4 3 2
- Sum the scores (ΣX):
  $1 + 5 + 4 + 3 + 2 = 15$
- Divide the sum ($\Sigma X = 15$) by the number of scores ($N = 5$):
  $15 / 5 = 3$
- Mean = $\overline{X} = 3$

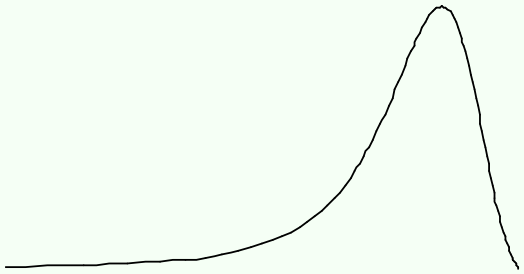
When To Use the Mean

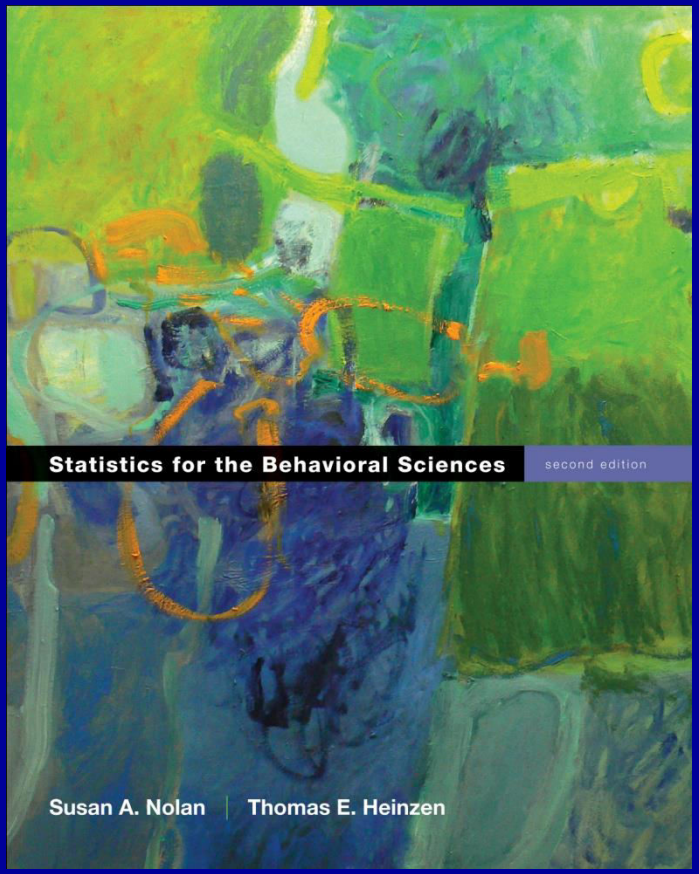
- You should use the mean when
  - the data are interval or ratio scaled
    - Many people will use the mean with ordinally scaled data too
  - and the data are not skewed
- The mean is preferred because it is sensitive to every score
  - If you change one score in the data set, the mean will change

Relations Between the Measures of Central Tendency

- In symmetrical distributions, the median and mean are equal
  - For normal distributions, mean = median = mode
- In positively skewed distributions, the mean is greater than the median
- In negatively skewed distributions, the mean is smaller than the median



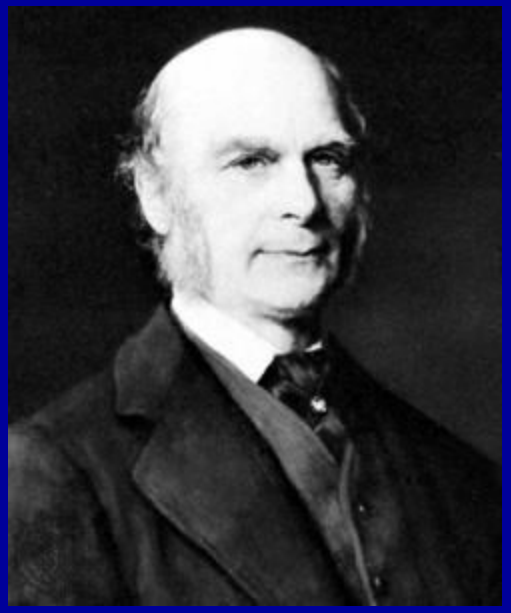




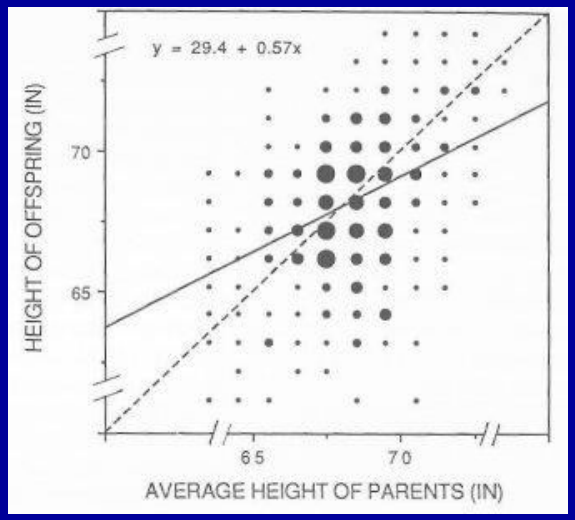
Correlation

Chapter 15

Correlation



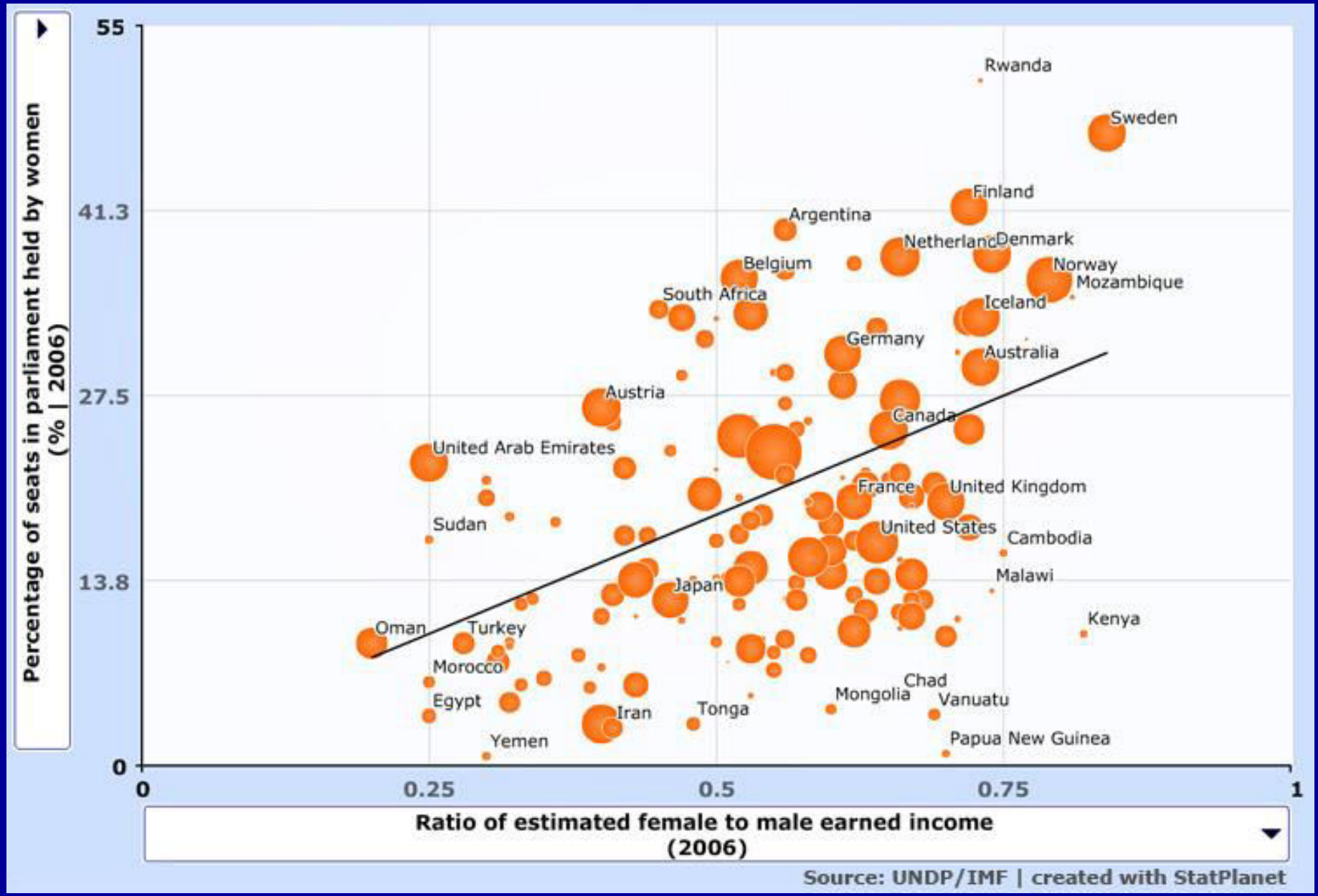
- Sir Francis Galton (Uncle to Darwin
  - Development of behavioral statistics
  - Father of Eugenics
  - Science of fingerprints as unique
  - Retrospective IQ of 200
  - Drove himself mad just to prove you could do it
  - Invented the pocket



Defining Correlation

- Co-variation or co-relation between two variables
- These variables change together
- Usually scale (interval or ratio) variables

- http://www.youtube.com/watch?v=ahp7QhbB8G4


Percentage of seats in parliament held by women (% | 2006)
55
41.3
27.5
13.8
0
0
0.25
0.5
0.75
1
Ratio of estimated female to male earned income (2006)
Rwanda
Sweden
Finland
Argentina
Netherlands
Denmark
Belgium
Norway
Mozambique
South Africa
Iceland
Germany
Australia
Austria
Canada
United Arab Emirates
France
United Kingdom
Sudan
United States
Cambodia
Malawi
Japan
Kenya
Oman
Turkey
Morocco
Chad
Egypt
Mongolia
Vanuatu
Iran
Tonga
Yemen
Papua New Guinea
Source: UNDP/IMF | created with StatPlanet

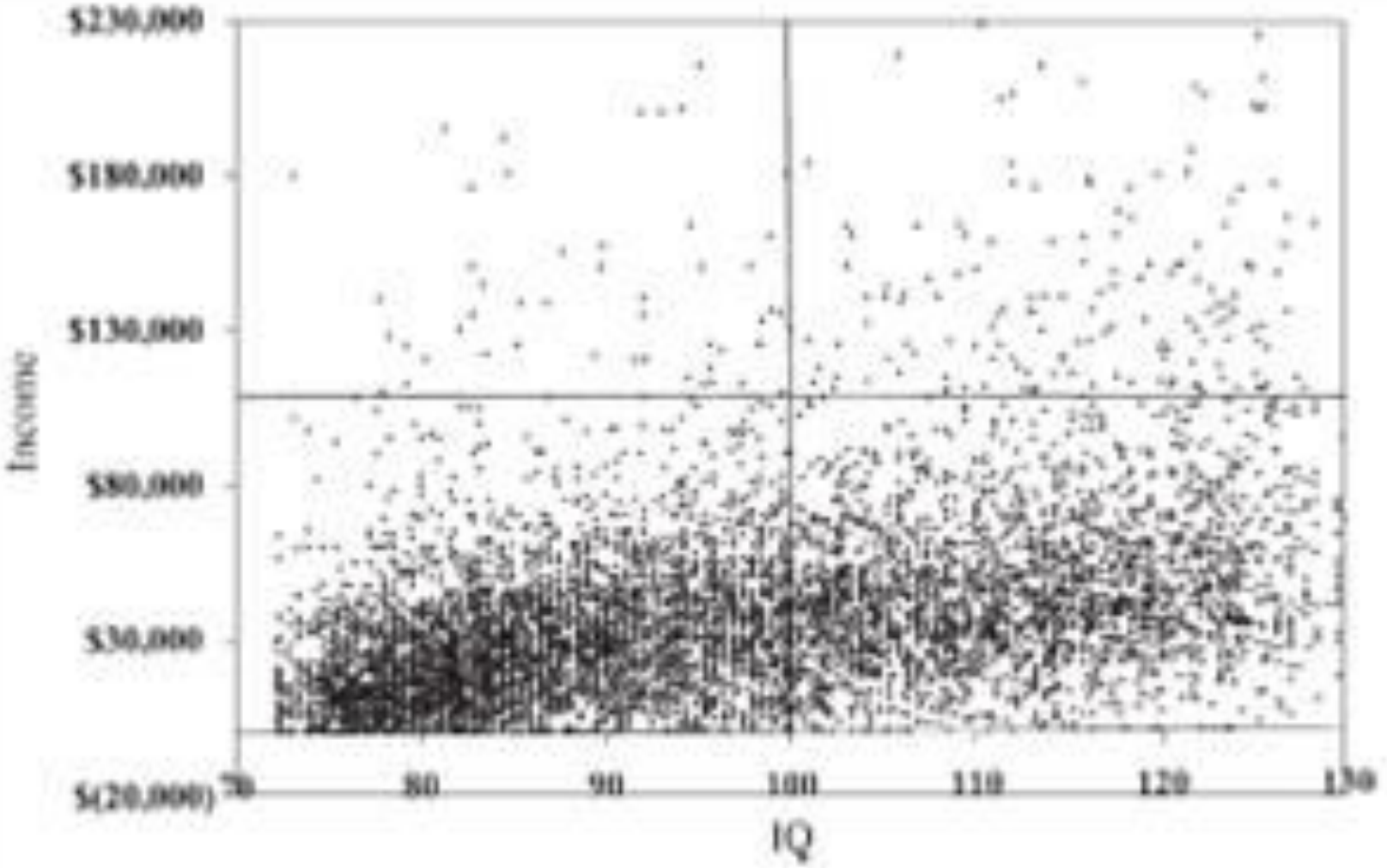

$230,000
$180,000
$130,000
$80,000
$30,000
$(20,000)
INCOME
70
80
90
100
110
120
130
IQ

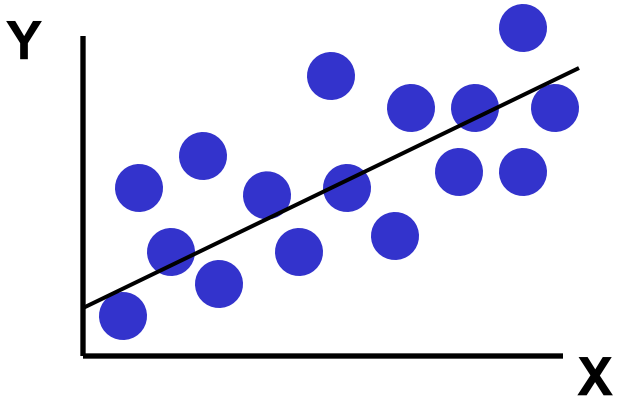
Fig. 1.

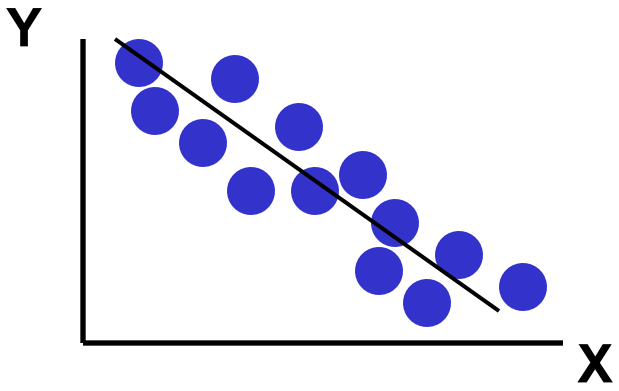
Correlation Coefficient

- A statistic that quantifies a relation between two variables
- Can be either positive or negative
- Falls between -1.00 and 1.00
- The value of the number (not the sign) indicates the strength of the relation

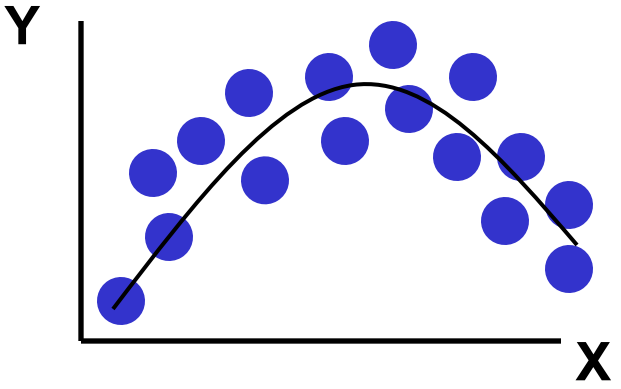
Linear Correlation

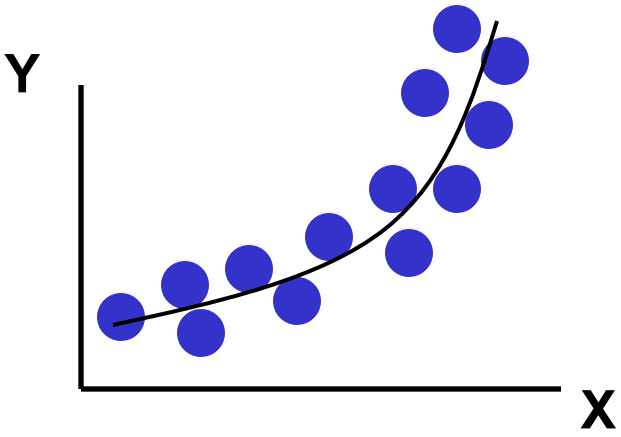
Linear relationships





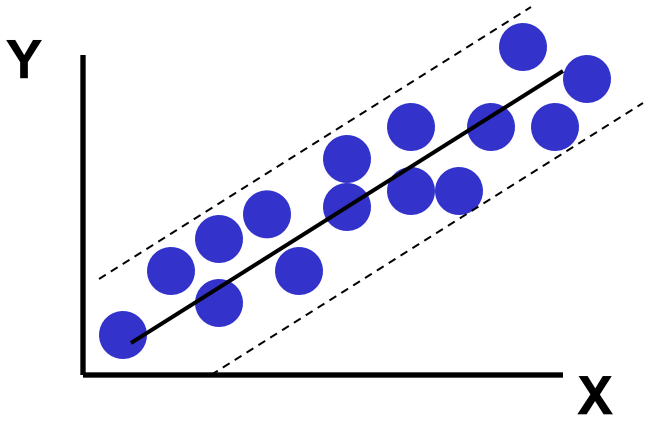
Curvilinear relationships

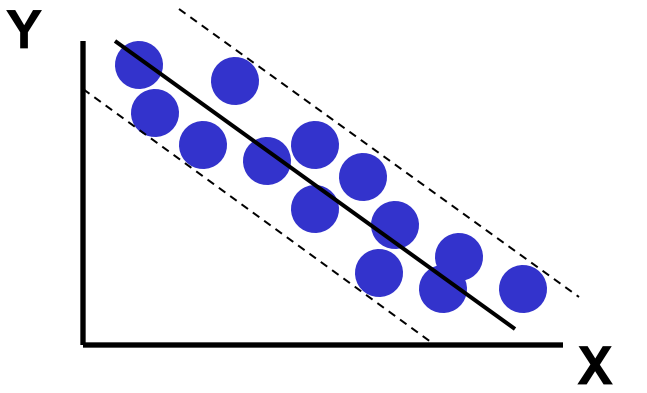




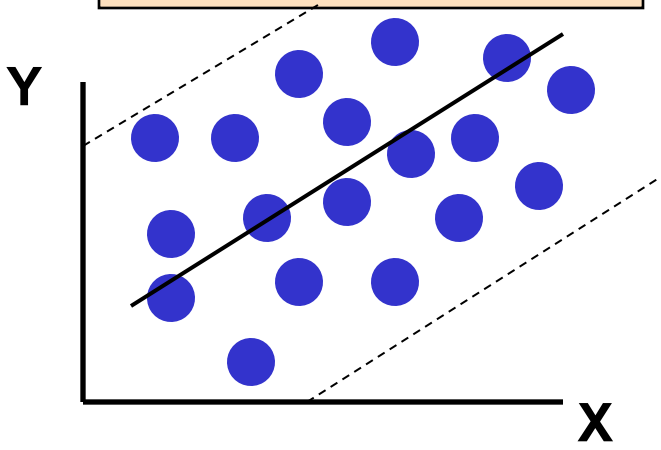
Linear Correlation

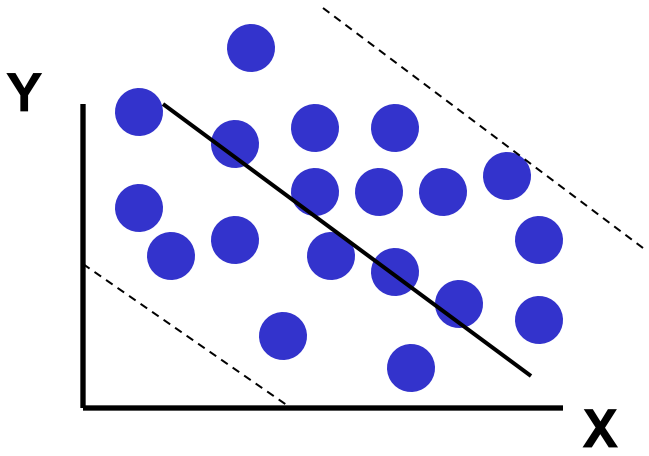
Strong relationships





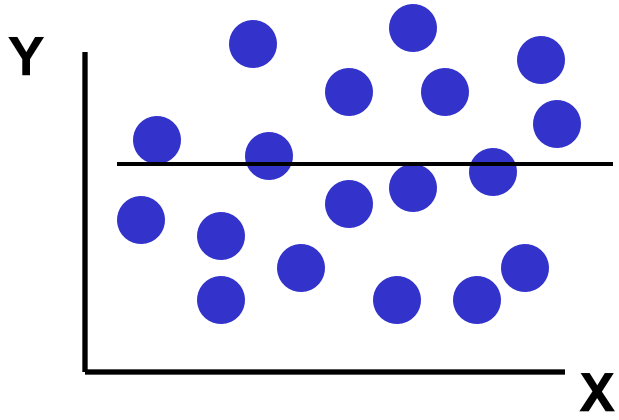
Weak relationships

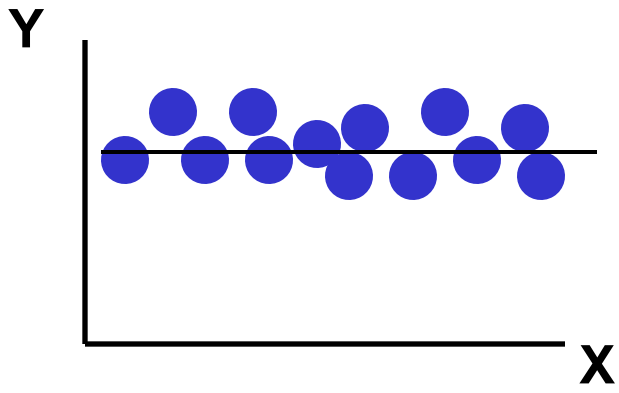




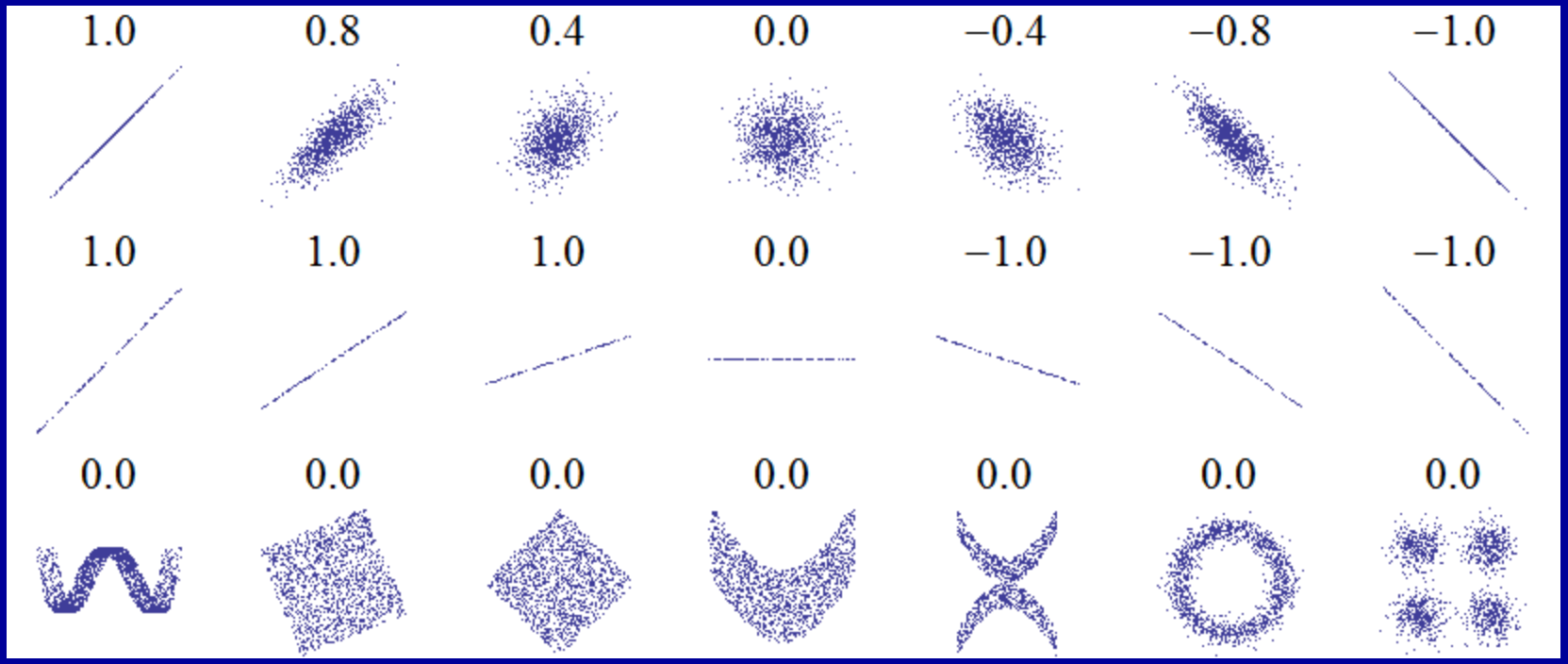
Linear Correlation

No relationship





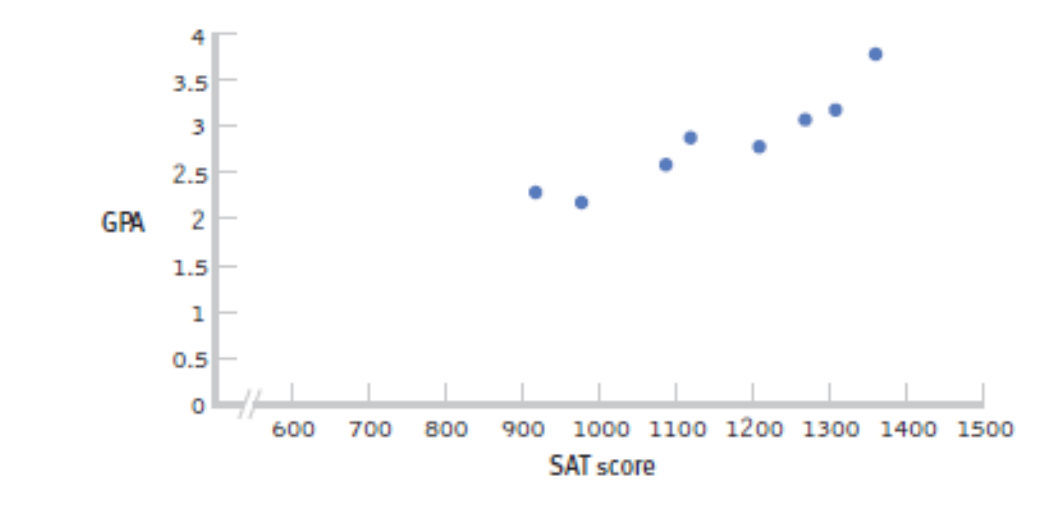
Correlation



Positive Correlation

Association between variables such that high scores on one variable tend to have high scores on the other variable

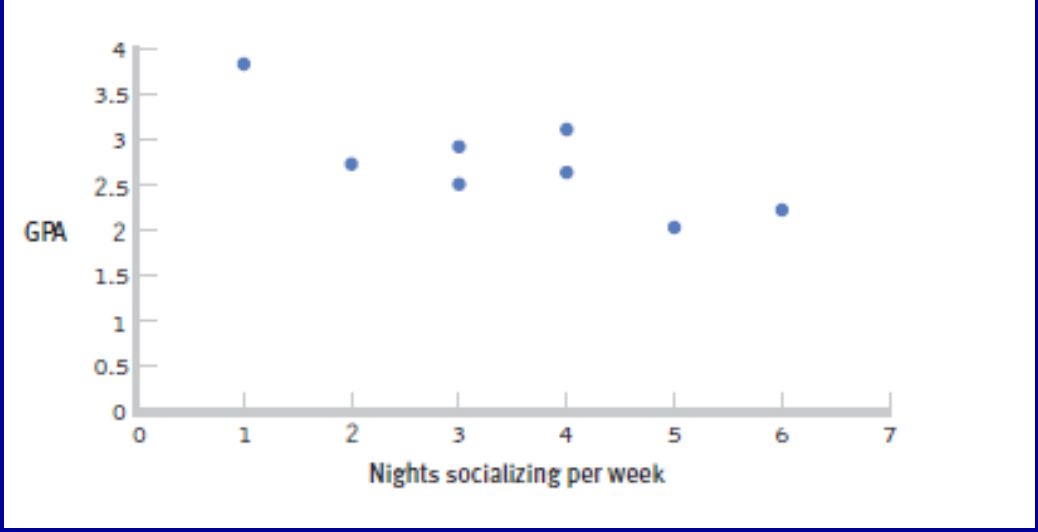
A direct relation between the variables



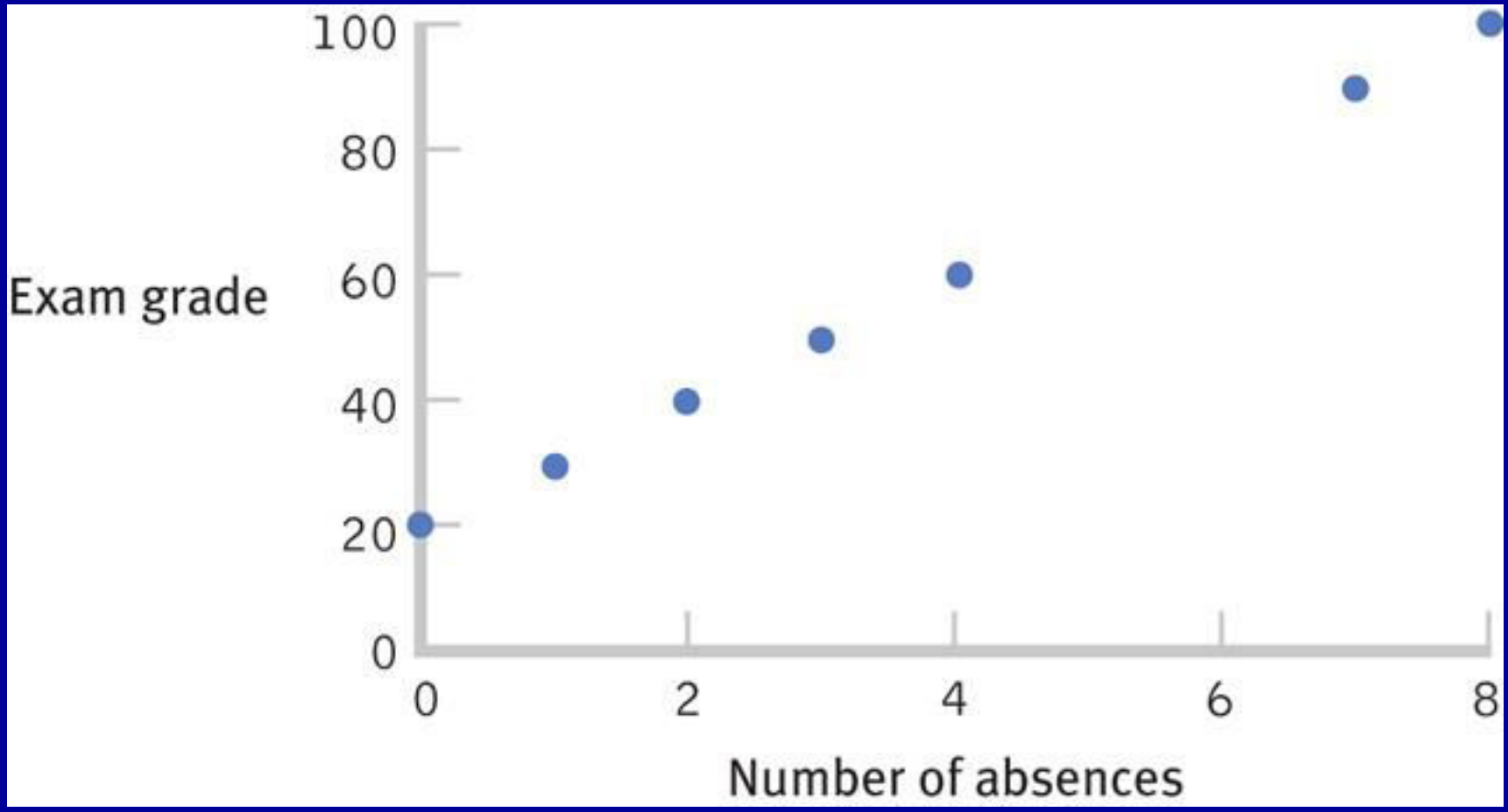
Negative Correlation

Association between variables such that high scores on one variable tend to have low scores on the other variable

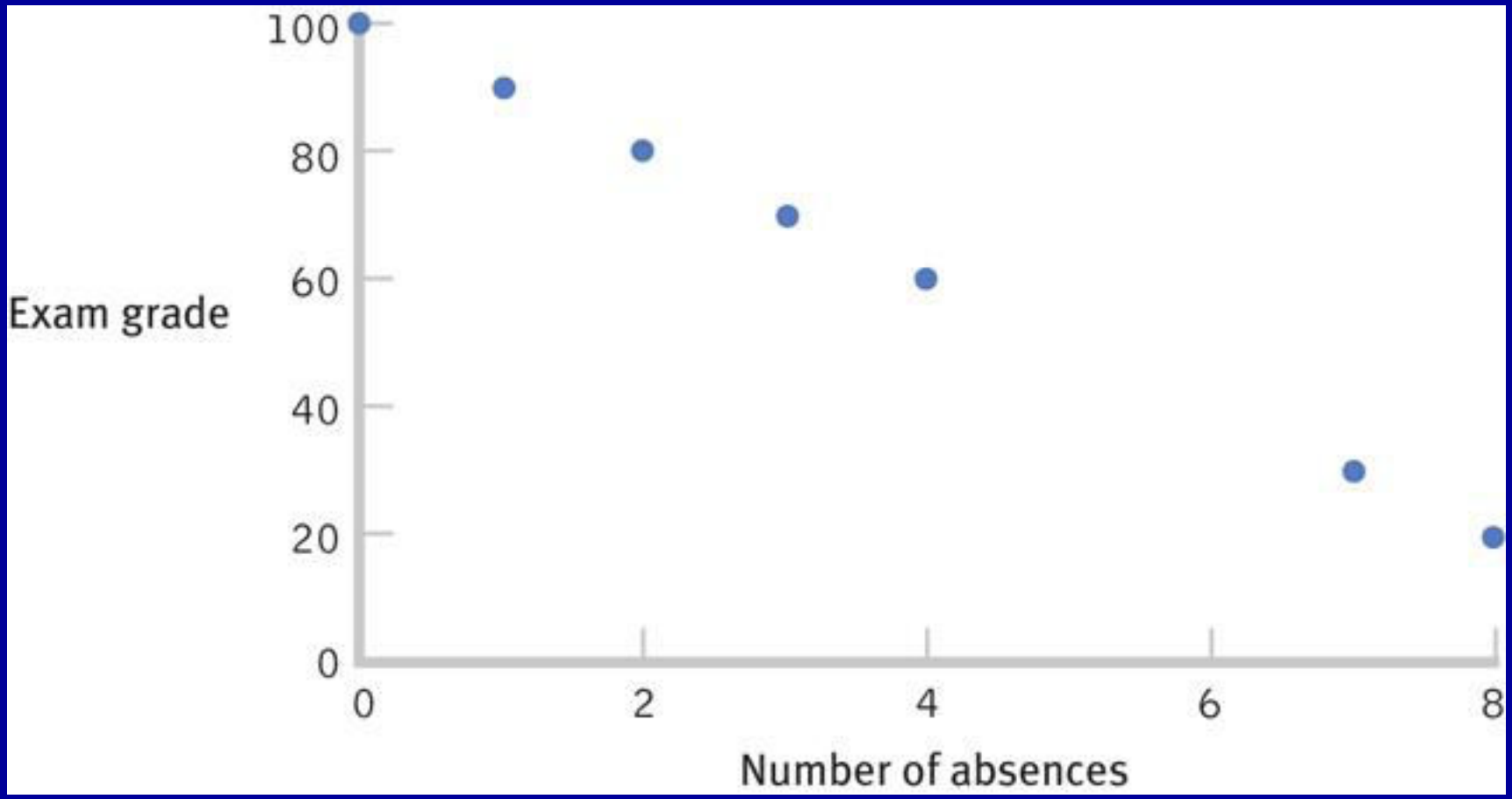
An inverse relation between the variables



A Perfect Positive Correlation

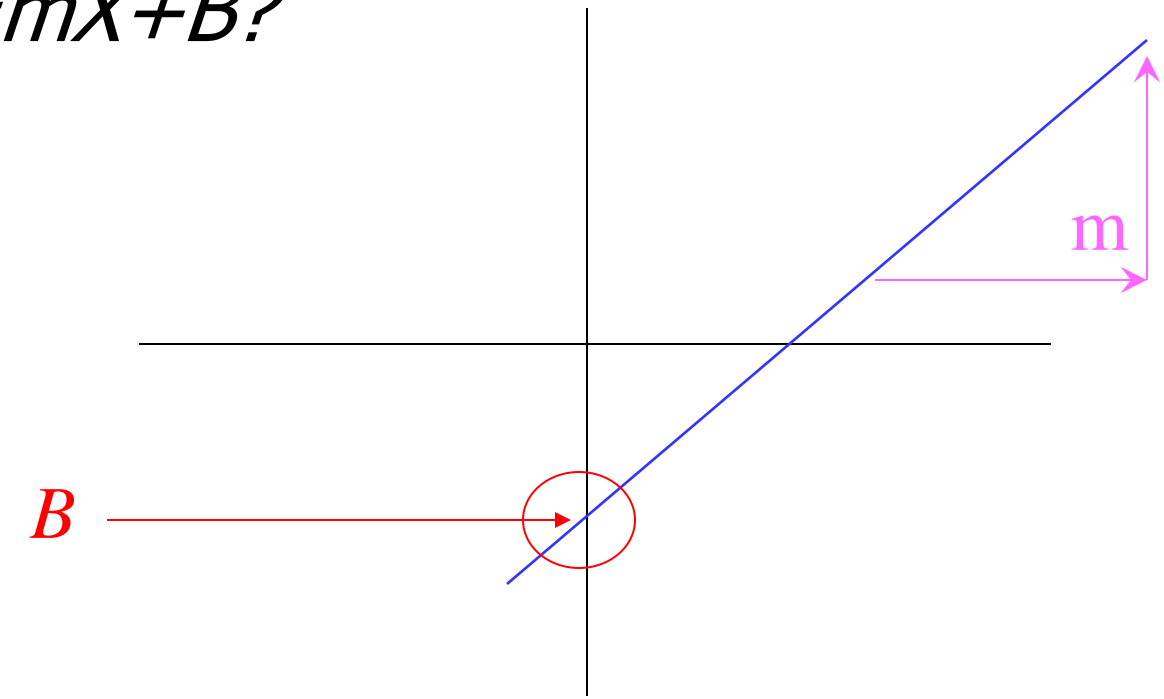


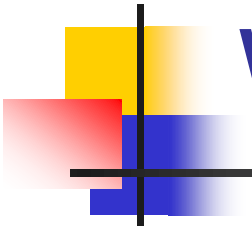
A Perfect Negative Correlation



What is "Linear"?

- Remember this:
- *Y=mX+B?*

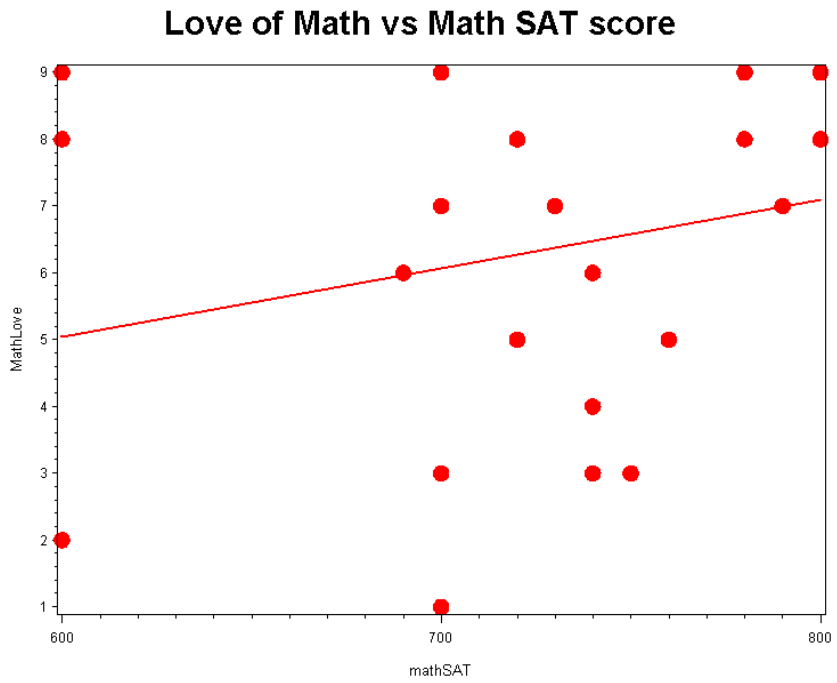




What's Slope?

A slope of 2 means that every 1-unit change in X yields a 2-unit change in Y.

Simple linear regression



P=.22; not significant

The linear regression model:

Love of Math = 5 + .01*math SAT score

intercept

slope

TABLE 15-1. How Strong Is an Association?

Cohen (1988) published guidelines to help researchers determine the strength of a correlation from the correlation coefficient. In social science research, however, it is extremely unusual to have a correlation as high as 0.50, and many have disputed the utility of Cohen's conventions for many social science contexts.

Size of the Correlation	Correlation Coefficient
Small	0.10
Medium	0.30
Large	0.50

Check Your Learning

- Which is stronger?
  - A correlation of 0.25 or -0.74?

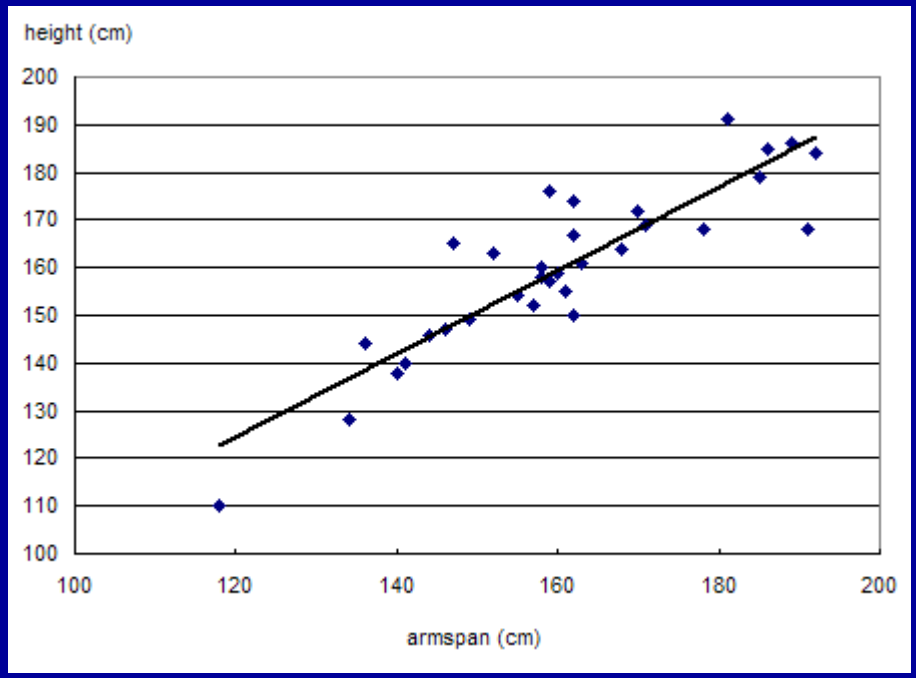
Misleading Correlations

- Something to think about
  - There is a 0.91 correlation between ice cream consumption and drowning deaths.
    - Does eating ice cream cause drowning?
    - Does grief cause us to eat more ice cream?

Correlation

Correlation is NOT causation

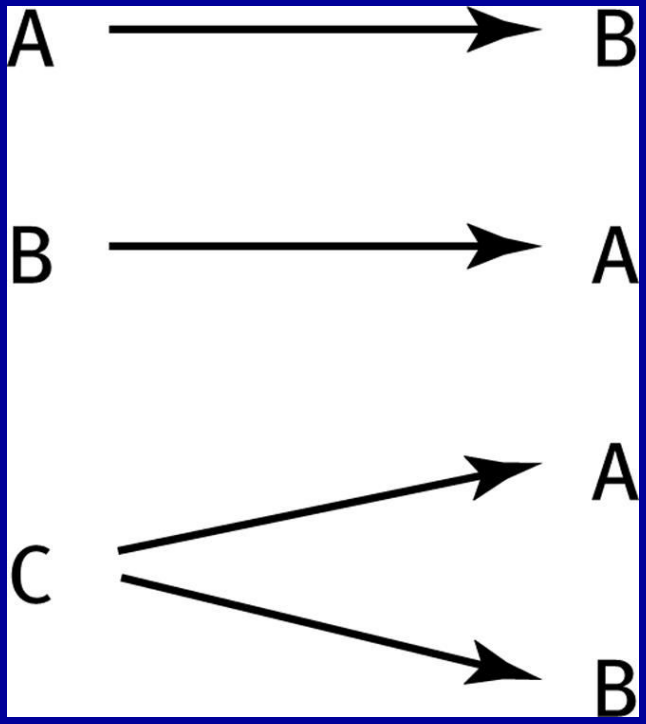
-e.g., armspan and height



The Limitations of Correlation

- Correlation is not causation.
  - Invisible third variables

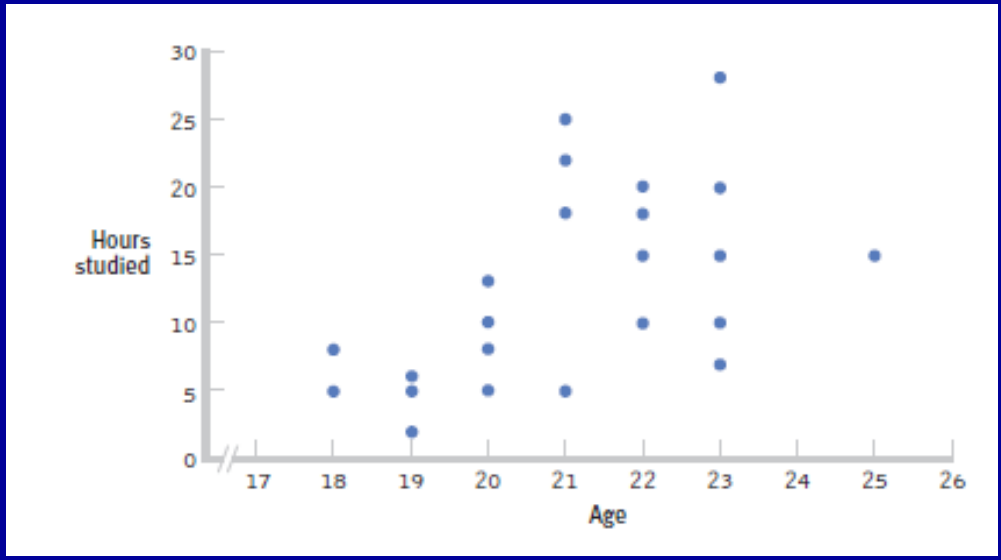
Three Possible Causal Explanations for a Correlation



The Limitations of Correlation, cont.

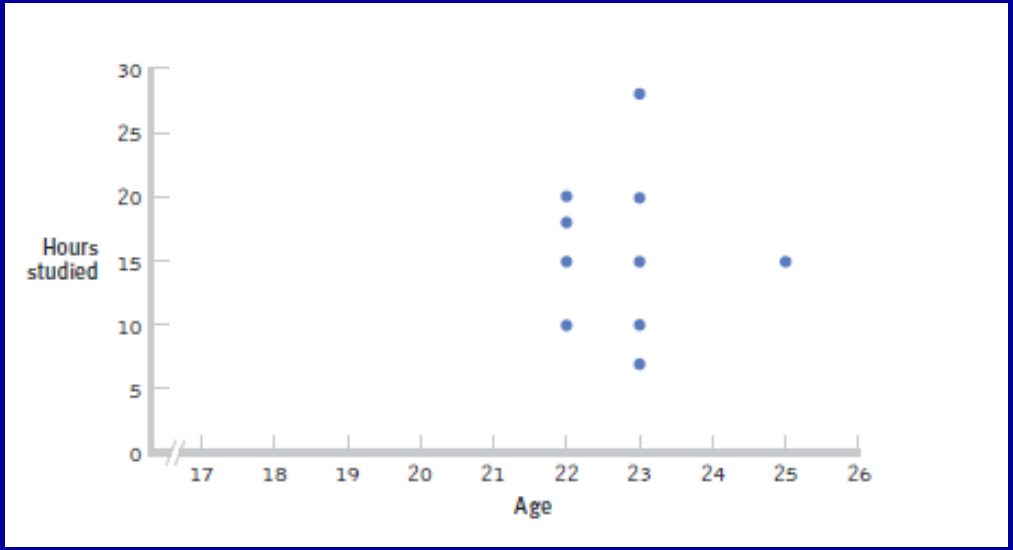
> Restricted Range.

A sample of boys and girls who performed in the top 2% to 3% on standardized tests - a much smaller range than the full population from which the researchers could have drawn their sample.



> Restricted Range, cont.

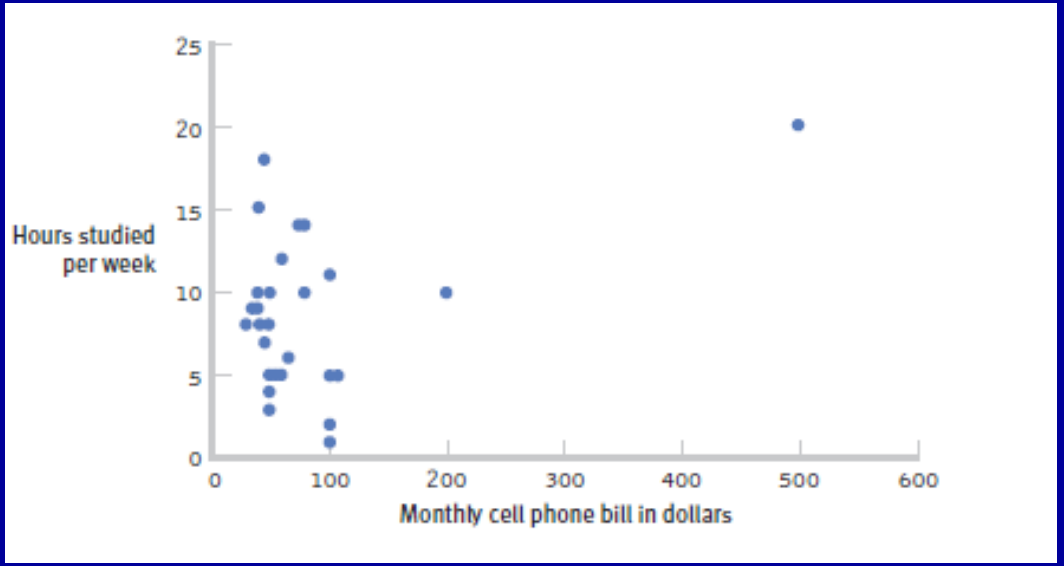
If we only look at the older students between the ages of 22 and 25, the strength of this correlation is now far smaller, just 0.05.



The Limitations of Correlation, cont.

> The effect of an outlier.

One individual who both studies and uses her cell phone more than any other individual in the sample changed the correlation from 0.14, a negative correlation, to 0.39, a much stronger and positive correlation!



The Pearson Correlation Coefficient

- A statistic that quantifies a linear relation between two scale variables.
- Symbolized by the italic letter *r* when it is a statistic based on sample data.
- Symbolized by the italic letter *p* "rho" when it is a population parameter.

- Pearson correlation coefficient
  - *r*
  - Linear relationship

$$r = \frac{\sum[(X - M_X)(Y - M_Y)]}{\sqrt{(SS_X)(SS_Y)}}$$

Correlation Hypothesis Testing

- Step 1. Identify the population, distribution, and assumptions
- Step 2. State the null and research hypotheses.
- Step 3. Determine the characteristics of the comparison distribution.
- Step 4. Determine the critical values.
- Step 5. Calculate the test statistic
- Step 6. Make a decision.

TABLE 15-2. Is Skipping Class Related to Statistics Exam Grades?

Here are the scores for 10 students on two scale variables: number of absences from class in one semester and exam grade.

Student	Absences	Exam Grade
1	4	82
2	2	98
3	2	76
4	3	68
5	1	84
6	0	99
7	4	67
8	8	58
9	7	50
10	3	78

Always Start with a Scatterplot

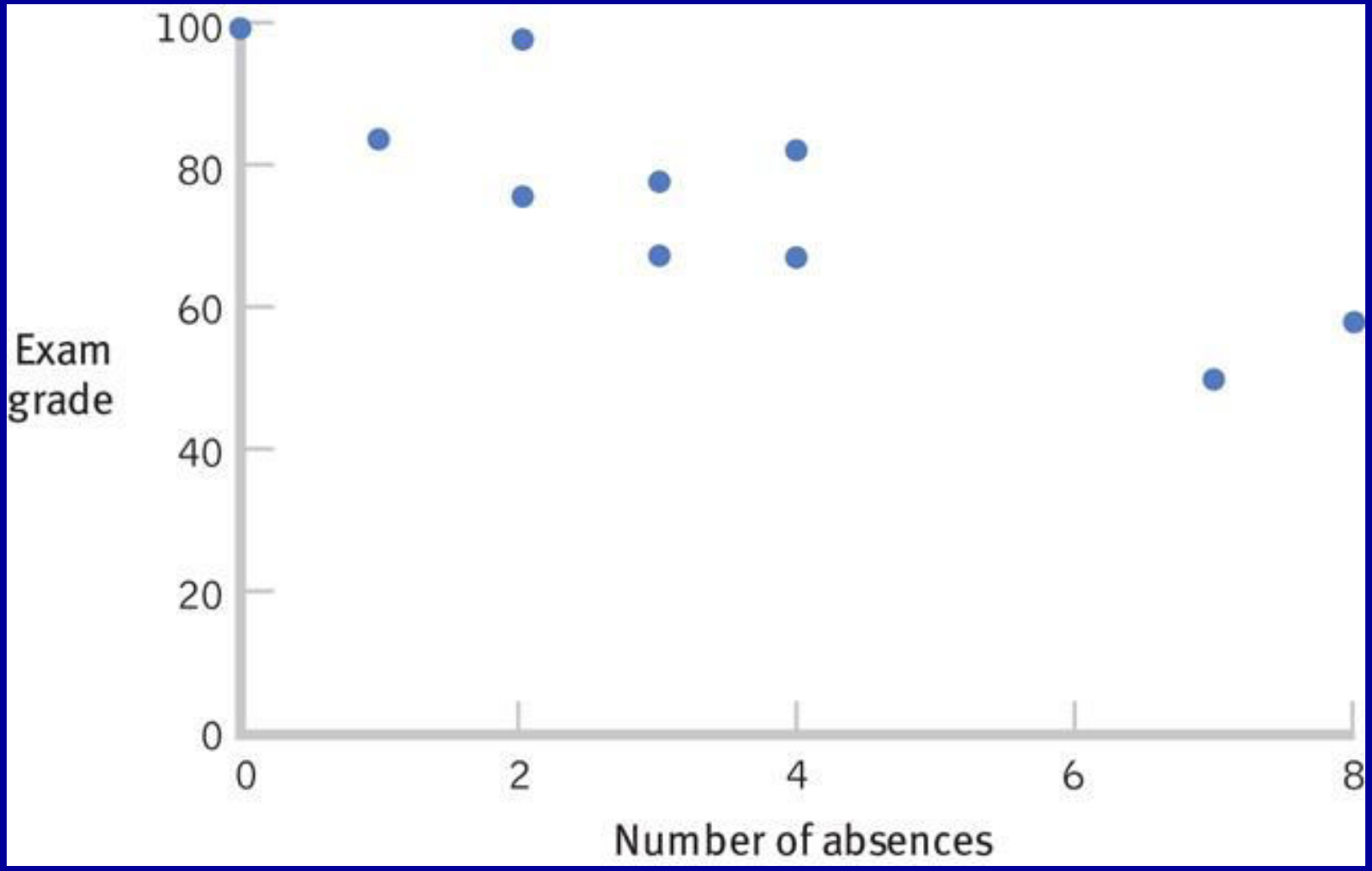


TABLE 15-3. Calculating the Numerator of the Correlation Coefficient

Absences (X)	$(X - M_X)$	Exam Grade (Y)	$(Y - M_Y)$	$(X - M_X)(Y - M_Y)$
4	0.6	82	6	3.6
2	−1.4	98	22	−30.8
2	−1.4	76	0	0.0
3	−0.4	68	−8	3.2
1	−2.4	84	8	−19.2
0	−3.4	99	23	−78.2
4	0.6	67	−9	−5.4
8	4.6	58	−18	−82.8
7	3.6	50	−26	−93.6
3	−0.4	78	2	−0.8
$M_X = 3.400$		$M_Y = 76.000$		$\Sigma[(X - M_X)(Y - M_Y)] = -304.0$

TABLE 15-4. Calculating the Denominator of the Correlation Coefficient

Absences (X)	$(X - M_X)$	$(X - M_X)^2$	Exam Grade (Y)	$(Y - M_Y)$	$(Y - M_Y)^2$
4	0.6	0.36	82	6	36
2	−1.4	1.96	98	22	484
2	−1.4	1.96	76	0	0
3	−0.4	0.16	68	−8	64
1	−2.4	5.76	84	8	64
0	−3.4	11.56	99	23	529
4	0.6	0.36	67	−9	81
8	4.6	21.16	58	−18	324
7	3.6	12.96	50	−26	676
3	−0.4	0.16	78	2	4
		$\Sigma(X - M_X)^2 = 56.4$			$\Sigma(Y - M_Y)^2 = 2262$

Correlation and Psychometrics

- Psychometrics is used in the development of tests and measures.
- Psychometricians use correlation to examine two important aspects of the development of measures—reliability and validity.

Reliability

- A reliable measure is one that is consistent.
- One particular type of reliability is test–retest reliability.
- Correlation is used by psychometricians to help professional sports teams assess the reliability of athletic performance, such as how fast a pitcher can throw a baseball.

Validity

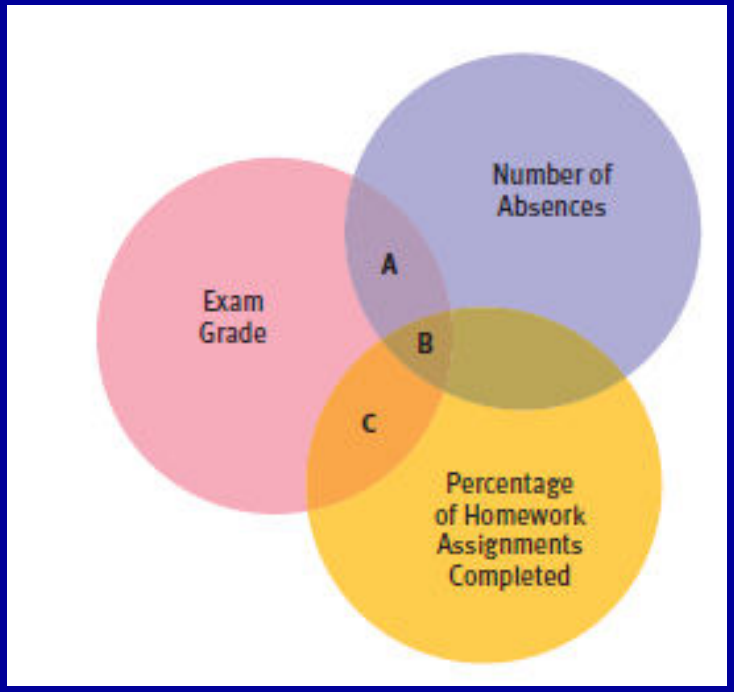
- A valid measure is one that measures what it was designed or intended to measure.
- Correlation is used to calculate validity, often by correlating a new measure with existing measures known to assess the variable of interest.

- Correlation can also be used to establish the validity of a personality test.
- Establishing validity is usually much more difficult than establishing reliability.
- Most magazines and newspapers never examine the psychometric properties of the quizzes that they publish.

Partial Correlation

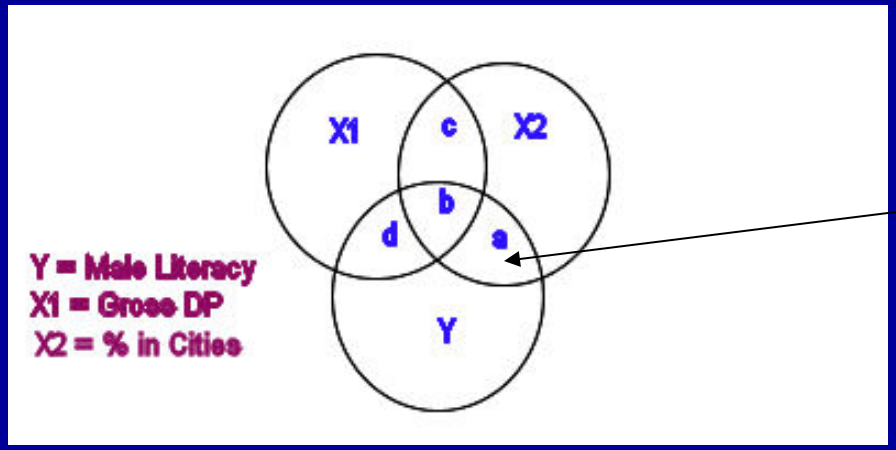
- A technique that quantifies the degree of association between two variables after statistically removing the association of a third variable with both of those two variables.
- Allows us to quantify the relation between two variables, controlling for the correlation of each of these variables with a third related variable.

- **We can assess the correlation between number of absences and exam grade, over and above the correlation of percentage of completed homework assignments with these variables.**



Partial Correlation

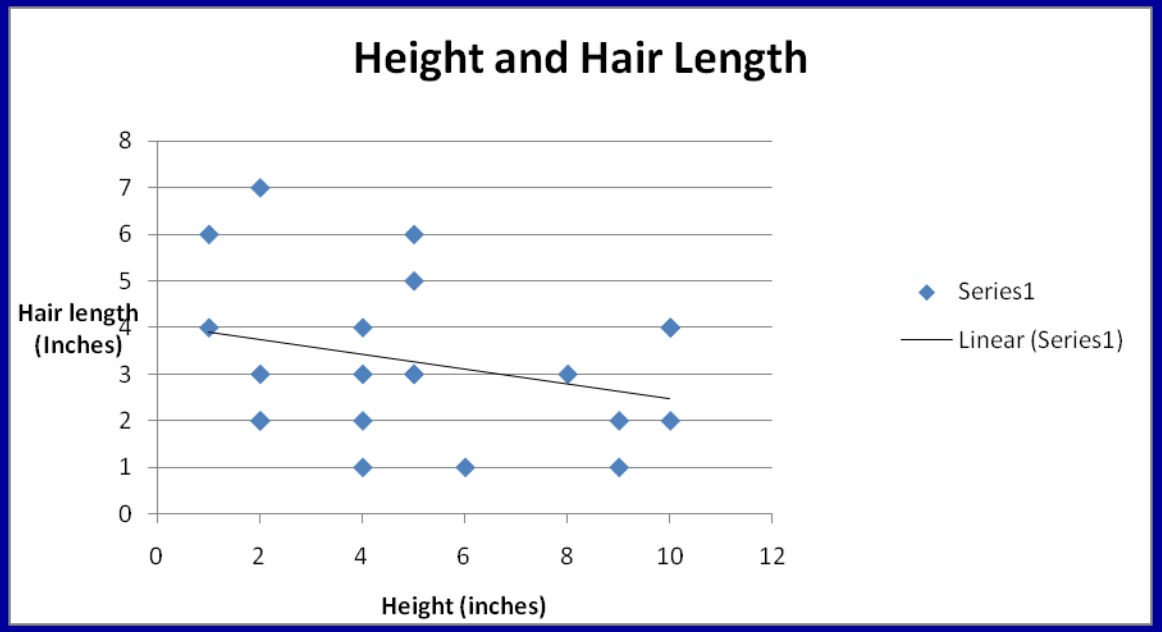
- A partial correlation is the relationship between two variables after removing the overlap with a third variable completely from both variables. In the diagram below, this would be the relationship between male literacy (Y) and percentage living in cities (X2), after removing the influence of gross domestic product (X1) on both literacy and percentage living in cities



In the calculation of the partial correlation coefficient $r_{YX2.X1}$, the area of interest is section a, and the effects removed are those in b, c, and d; partial correlation is the relationship of X2 and Y after the influence of X1 is completely removed from both variables. When only the effect of X1 on X2 is removed, this is called a part correlation; part correlation first removes from X2 all variance which may be accounted for by X1 (sections c and b), then correlates the remaining unique component of the X2 with the dependent variable, Y

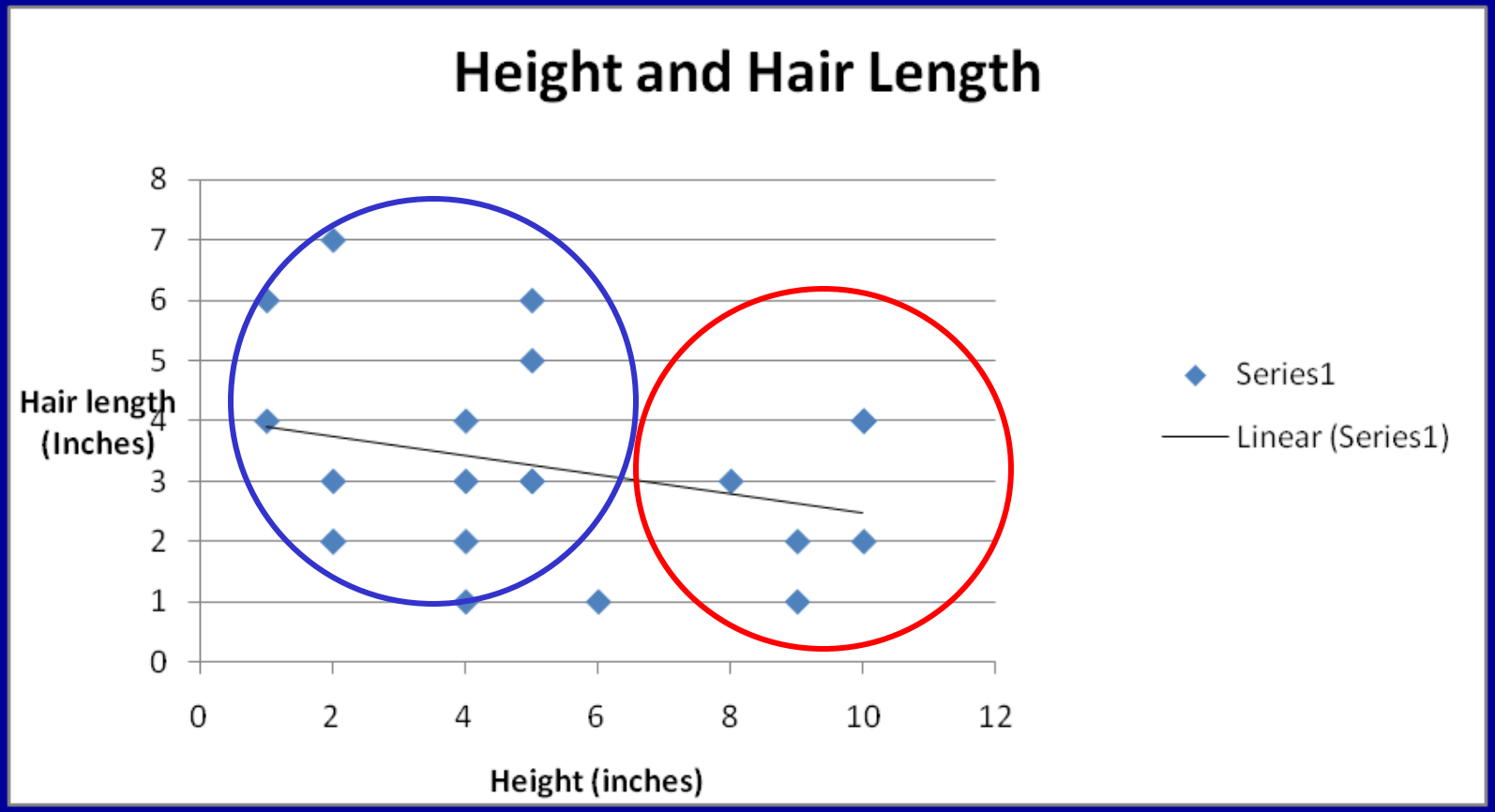
Statistical Control

- Using Multivariate Analysis



Statistical Control

- Using Multivariate Analysis

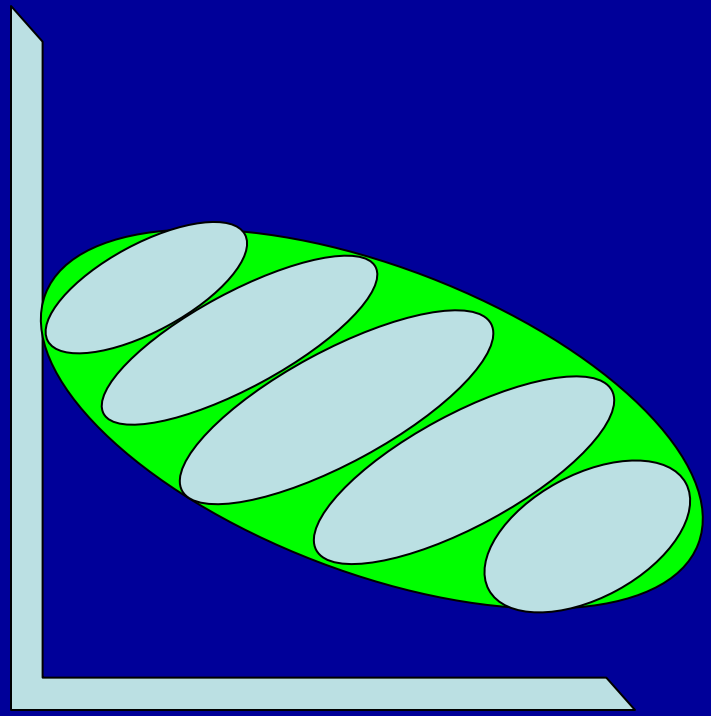


Simpson's Paradox

- In each of these examples, the bivariate analysis (cross-tabulation or correlation) gave misleading results
- Introducing another variable gave a better understanding of the data
  - It even reversed the initial conclusions

Another Example

- A study of graduates' salaries showed negative association between economists' starting salary and the level of the degree
  - i.e. PhDs earned less than Masters degree holders, who in turn earned less than those with just a Bachelor's degree
  - Why?
- The data was split into three employment sectors
  - Teaching, government and private industry
  - Each sector showed a positive relationship
  - Employer type was confounded with degree level



Introduction to Time Series Analysis

Dr P.Sudarkodi

Time Series?

- A set of observations indexed by time t
- Discrete and continuous time series

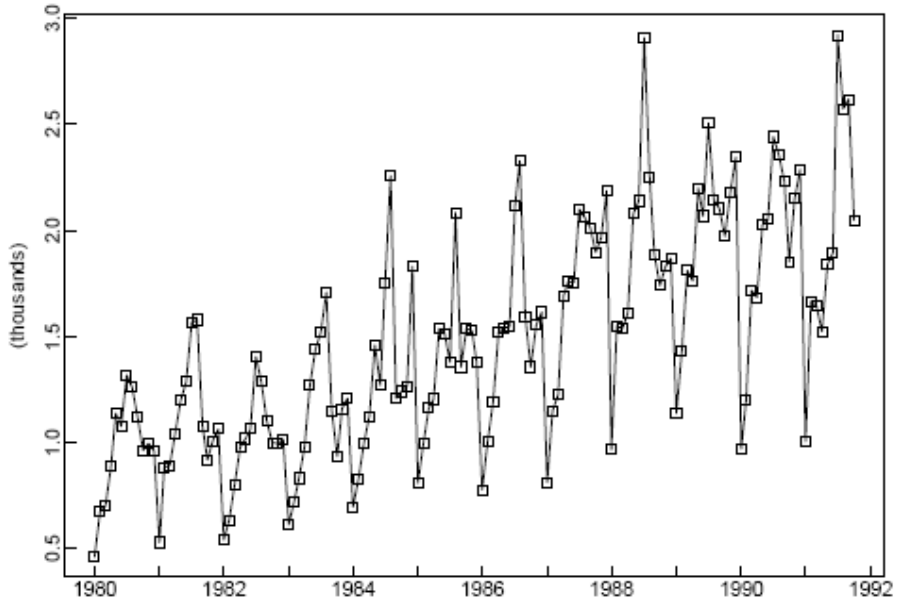


Figure 1-1
The Australian red wine sales, Jan. '80 – Oct. '91.

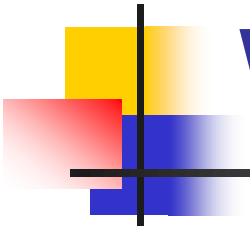
Stationary Time Series

- (Weakly) stationary
  - The covariance is independent of t for each h

$$\gamma_X(X_t, X_{t-h}) \equiv E\left[(X_t - \mu)(X_{t-h} - \mu)\right]$$

  - The mean is independent of t

$$E(X_t) = \mu$$



Why Stationary Time Series?

- Stationary time series have the best linear predictor.
- Nonstationary time series models are usually slower to implement for prediction.

Converting Nonstationary Time Series to Stationary Time Series

- Remove deterministic factors
  - Trends
    - Polynomial regression fitting
    - Exponential smoothing
    - Moving average smoothing
    - Differencing (B is a back shift operator)

$$\nabla = X_t - X_{t-1} = (1 - B)X_t$$

Converting Nonstationary Time Series to Stationary Time Series

- Remove deterministic factors
  - Seasonality (usually combined with trends removal)

$$X_t = m_t + s_t + Y_t$$

  - Differencing

$$\nabla_d = X_t - X_{t-d} = (1 - B^d) X_t$$

Converting Nonstationary Time Series to Stationary Time Series

- Example

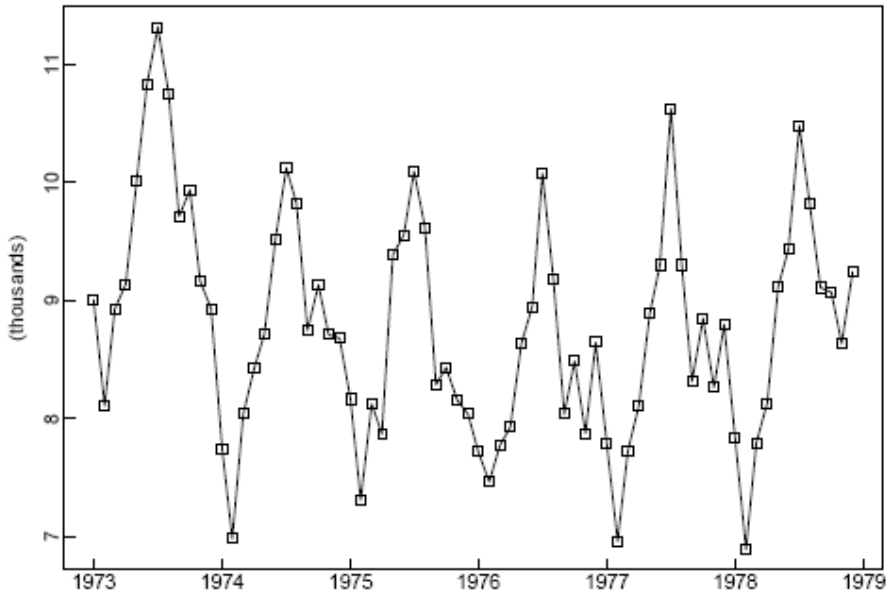


Figure 1-3
The monthly accidental deaths data, 1973–1978.

Converting Nonstationary Time Series to Stationary Time Series

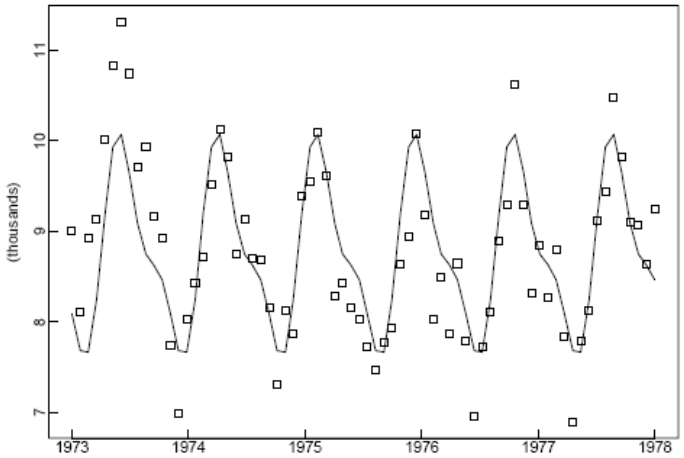


Figure 1-11
The estimated harmonic component of the accidental deaths data from ITSM.

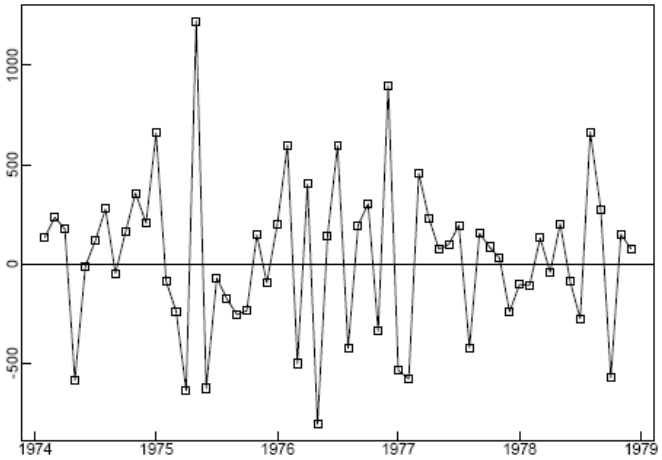
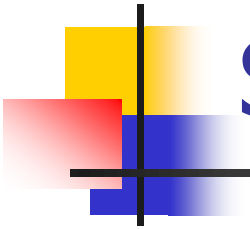


Figure 1-27
The differenced series $\{\nabla\nabla_{12}x_t, t = 14, \ldots, 72\}$ derived from the monthly accidental deaths $\{x_t, t = 1, \ldots, 72\}$.

Converting Nonstationary Time Series to Stationary Time Series



- After conversion, remaining data points are called residuals
- If residuals are IID, then no more analysis is necessary since its mean value will be the best predictor

Wold Decomposition

- Stationary time series can be represented as the following

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} + V_t,$$

$$\{Z_t\} \sim WN(0, \sigma),$$

$$Cov(Z_t, V_t) = 0,$$

$$V_t : \text{Determinis tic},$$

$$\sum \psi_j^{\,2} < \infty,$$

Stationary Time Series Prediction

- P_n is a prediction function of X_{n+h} with forward lag h from X_n.

$$P_n X_{n+h} = a_0 + a_1 X_n + \cdots + a_n X_1$$

- The prediction error is measured in the minimum mean square

$$S(a_0, \cdots, a_n) = E(X_{n+h} - (a_0 + a_1 X_n + \cdots + a_n X_1))^2$$

Stationary Time Series Prediction

- Since S is a quadratic function, the minimum value will be obtained when all the partial derivatives are 0.

$$\frac{\partial S(a_0,\cdots,a_n)}{\partial a_j}=0,\ \ j=0,\cdots n$$

$$E\left[X_{n+h}-a_0-\sum_{i=1}^{n}a_i X_{n+1-i}\right]=0,$$

$$E\left[(X_{n+h}-a_0-\sum_{i=1}^{n}a_i X_{n+1-i})X_{n+1-j}\right]=0,\quad j=1,\ldots,n.$$

Stationary Time Series Prediction

- In another form

$$a_0 = \mu\left(1 - \sum_{i=1}^{n} a_i\right)$$

and

$$\Gamma_n \mathbf{a}_n = \gamma_n(h),$$

where

$$\mathbf{a}_n = (a_1, \ldots, a_n)', \qquad \Gamma_n = [\gamma(i-j)]_{i,j=1}^{n},$$

and

$$\gamma_n(h) = (\gamma(h), \gamma(h+1), \ldots, \gamma(h+n-1))'.$$

Stationary Models

- AR (AutoRegressive)

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{1-p} + Z_t,$$

$$Z_t \sim WN(0, \sigma^2),$$

$$\gamma(X_s, Z_t) = 0, \quad s < t.$$

- AR's predictor

$$P_n X_{n+1} = \phi_1 X_n + \cdots + \phi_p X_{n+1-p}$$

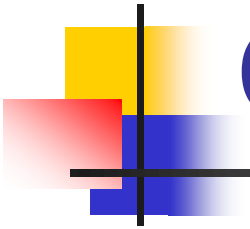
Stationary Models

- ARMA
  - Reduces large autocovariance functions
  - A transformed linear predictor is used

$\{X_t\}$ is an **ARMA(p, q) process** if $\{X_t\}$ is stationary and if for every t,

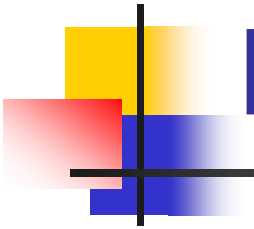
$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \qquad (3.1.1)$$

where $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$ and the polynomials $(1 - \phi_1 z - \ldots - \phi_p z^p)$ and $(1 + \theta_1 z + \ldots + \theta_q z^q)$ have no common factors.



Other Models

- Mutivariate Cointegration
- ARIMA
- SARIMA
- FARIMA
- GARCH



References

- Introduction to Time Series and Forecasting 2nd ed., P. Brockwell and R. Davis, Springer Verlag
- Adaptive Filter Theory 4th ed., Simon Haykin, Prentice Hall
- Time Series Analysis, James Douglas Hamilton, Princeton University Press

Sampling Methods

Dr P.Sudarkodi

Defining the Target Population

- It is critical to the success of the research project to clearly define the target population.
- Rely on logic and judgment.
- The population should be defined in connection with the objectives of the study.

Technical Terminology

- An element is an object on which a measurement is taken.
- A population is a collection of elements about which we wish to make an inference.
- Sampling units are nonoverlapping collections of elements from the population that cover the entire population.

Technical Terms

- A sampling frame is a list of sampling units.
- A sample is a collection of sampling units drawn from a sampling frame.
- Parameter: numerical characteristic of a population
- Statistic: numerical characteristic of a sample

Errors of nonobservation

- The deviation between an estimate from an ideal sample and the true population value is the sampling error.
- Almost always, the sampling frame does not match up perfectly with the target population, leading to errors of coverage.

Errors of nonobservation

- Nonresponse is probably the most serious of these errors.
  - Arises in three ways:
    - Inability of the person responding to come up with the answer
    - Refusal to answer
    - Inability to contact the sampled elements

Errors of observation

- These errors can be classified as due to the interviewer, respondent, instrument, or method of data collection.

Interviewers

- Interviewers have a direct and dramatic effect on the way a person responds to a question.
  - Most people tend to side with the view apparently favored by the interviewer, especially if they are neutral.
  - Friendly interviewers are more successful.
  - In general, interviewers of the same gender, racial, and ethnic groups as those being interviewed are slightly more successful.

Respondents

- Respondents differ greatly in motivation to answer correctly and in ability to do so.
- Obtaining an honest response to sensitive questions is difficult.
- Basic errors
  - Recall bias: simply does not remember
  - Prestige bias: exaggerates to ‘look’ better
  - Intentional deception: lying
  - Incorrect measurement: does not understand the units or definition

Census Sample

- A census study occurs if the entire population is very small or it is reasonable to include the entire population (for other reasons).

- It is called a census sample because data is gathered on every member of the population.

Why sample?

- The population of interest is usually too large to attempt to survey all of its members.
- A carefully chosen sample can be used to represent the population.
  - The sample reflects the characteristics of the population from which it is drawn.

Probability versus Nonprobability

- **Probability Samples:** each member of the population has a known non-zero probability of being selected
  - Methods include random sampling, systematic sampling, and stratified sampling.
- **Nonprobability Samples:** members are selected from the population in some nonrandom manner
  - Methods include convenience sampling, judgment sampling, quota sampling, and snowball sampling

Random Sampling

Random sampling is the purest form of probability sampling.

- Each member of the population has an equal and known chance of being selected.
- When there are very large populations, it is often 'difficult' to identify every member of the population, so the pool of available subjects becomes biased.
  - You can use software, such as minitab to generate random numbers or to draw directly from the columns

Systematic Sampling

- **Systematic sampling** is often used instead of random sampling. It is also called an Nth name selection technique.
- After the required sample size has been calculated, every Nth record is selected from a list of population members.
- As long as the list does not contain any hidden order, this sampling method is as good as the random sampling method.
- Its only advantage over the random sampling technique is simplicity (and possibly cost effectiveness).

Stratified Sampling

- **Stratified sampling** is commonly used probability method that is superior to random sampling because it reduces sampling error.
- A stratum is a subset of the population that share at least one common characteristic; such as males and females.
  - Identify relevant stratums and their actual representation in the population.
  - Random sampling is then used to select a *sufficient* number of subjects from each stratum.
  - Stratified sampling is often used when one or more of the stratums in the population have a low incidence relative to the other stratums.

Cluster Sampling

- Cluster Sample: a probability sample in which each sampling unit is a collection of elements.
- Effective under the following conditions:
  - A good sampling frame is not available or costly, while a frame listing clusters is easily obtained
  - The cost of obtaining observations increases as the distance separating the elements increases
- Examples of clusters:
  - City blocks – political or geographical
  - Housing units – college students
  - Hospitals – illnesses
  - Automobile – set of four tires

Convenience Sampling

- **Convenience sampling** is used in exploratory research where the researcher is interested in getting an inexpensive approximation.
- The sample is selected because they are convenient.
- It is a nonprobability method.
  - Often used during preliminary research efforts to get an estimate without incurring the cost or time required to select a random sample

Judgment Sampling

- **Judgment sampling** is a common nonprobability method.
- The sample is selected based upon judgment.
  - an extension of convenience sampling
- When using this method, the researcher must be confident that the chosen sample is truly representative of the entire population.

Quota Sampling

- **Quota sampling** is the nonprobability equivalent of stratified sampling.
  - First identify the stratums and their proportions as they are represented in the population
  - Then convenience or judgment sampling is used to select the required number of subjects from each stratum.

Snowball Sampling

- **Snowball sampling** is a special nonprobability method used when the desired sample characteristic is rare.
- It may be extremely difficult or cost prohibitive to locate respondents in these situations.
- This technique relies on referrals from initial subjects to generate additional subjects.
- It lowers search costs; however, it introduces bias because the technique itself reduces the likelihood that the sample will represent a good cross section from the population.

Sample Size?

- The more heterogeneous a population is, the larger the sample needs to be.
- Depends on topic – frequently it occurs?
- For probability sampling, the larger the sample size, the better.
- With nonprobability samples, not generalizable regardless – still consider stability of results

Response Rates

- About 20 – 30% usually return a questionnaire
- Follow up techniques could bring it up to about 50%
- Still, response rates under 60 – 70% challenge the integrity of the random sample
- How the survey is distributed can affect the quality of sampling

ANOVA:
Analysis of Variation

The basic ANOVA situation

Two variables: 1 Categorical, 1 Quantitative

Main Question: Do the (means of) the quantitative variables depend on which group (given by categorical variable) the individual is in?

If categorical variable has only 2 values:

- 2-sample t-test

ANOVA allows for 3 or more groups

An example ANOVA situation

Subjects: 25 patients with blisters
Treatments: Treatment A, Treatment B, Placebo
Measurement: # of days until blisters heal

Data [and means]:

- A: 5,6,6,7,7,8,9,10 [7.25]
- B: 7,7,8,9,9,10,10,11 [8.875]
- P: 7,9,9,10,10,10,11,12,13 [10.11]

Are these differences significant?

Informal Investigation

Graphical investigation:

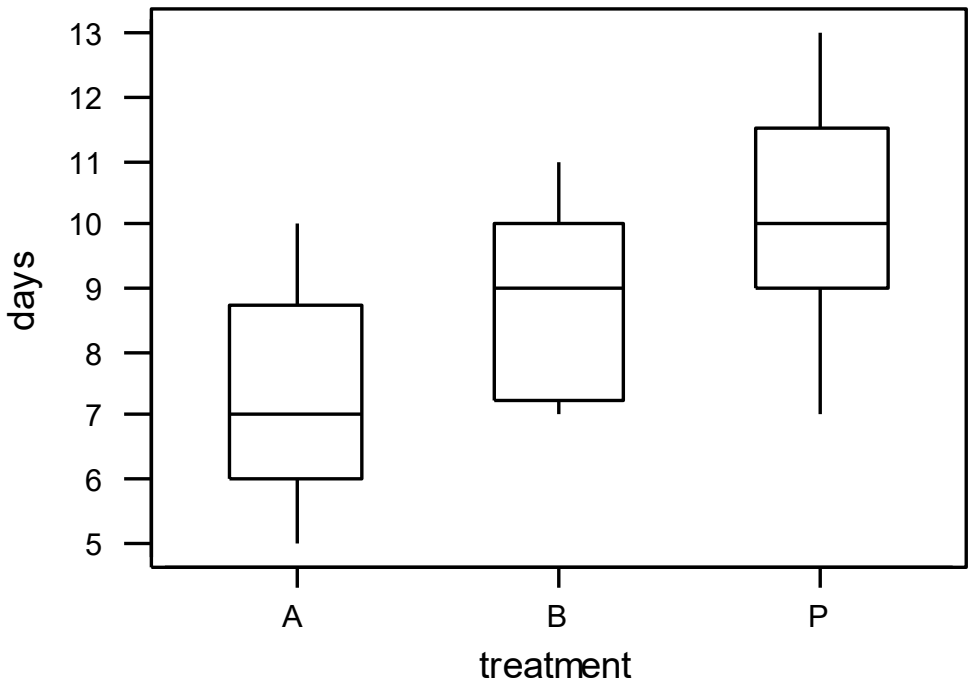
- side-by-side box plots
- multiple histograms

Whether the differences between the groups are significant depends on

- the difference in the means
- the standard deviations of each group
- the sample sizes

ANOVA determines P-value from the F statistic

Side by Side Boxplots



What does ANOVA do?

At its simplest (there are extensions) ANOVA tests the following hypotheses:

H_0: The means of all the groups are equal.

H_a: Not all the means are equal

- doesn't say how or which ones differ.
- Can follow up with "multiple comparisons"

Note: we usually refer to the sub-populations as "groups" when doing ANOVA.

Assumptions of ANOVA

- each group is approximately normal
  - check this by looking at histograms and/or normal quantile plots, or use assumptions
  - can handle some nonnormality, but not severe outliers
- standard deviations of each group are approximately equal
  - rule of thumb: ratio of largest to smallest sample st. dev. must be less than 2:1

Normality Check

We should check for normality using:

- assumptions about population
- histograms for each group
- normal quantile plot for each group

With such small data sets, there really isn't a really good way to check normality from data, but we make the common assumption that physical measurements of people tend to be normally distributed.

Standard Deviation Check

Variable	treatment	N	Mean	Median	StDev
days	A	8	7.250	7.000	1.669
	B	8	8.875	9.000	1.458
	P	9	10.111	10.000	1.764

Compare largest and smallest standard deviations:

- largest: 1.764
- smallest: 1.458
- 1.458 x 2 = 2.916 > 1.764

Note: variance ratio of 4:1 is equivalent.

Notation for ANOVA

- n = number of individuals all together
- I = number of groups
- $\overline{X}$ = mean for entire data set is

Group i has

- n_i = # of individuals in group i
- x_{ij} = value for individual j in group i
- $\overline{x}_i$ = mean for group i
- s_i = standard deviation for group i

How ANOVA works (outline)

ANOVA measures two sources of variation in the data and compares their relative sizes

- variation BETWEEN groups
  - for each data value look at the difference between its group mean and the overall mean

$$\left(\overline{x}_i - \overline{x}\right)^2$$

- variation WITHIN groups
  - for each data value we look at the difference between that value and the mean of its group

$$\left(x_{ij} - \overline{x}_i\right)^2$$

The ANOVA F-statistic is a ratio of the Between Group Variaton divided by the Within Group Variation:

$$F = \frac{Between}{Within} = \frac{MSG}{MSE}$$

A large F is evidence *against* H_0, since it indicates that there is more difference between groups than within groups.

Minitab ANOVA Output

Analysis of Variance for days

Source	DF	SS	MS	F	P
treatment	2	34.74	17.37	6.45	0.006
Error	22	59.26	2.69		
Total	24	94.00			

How are these computations made?

We want to measure the amount of variation due to BETWEEN group variation and WITHIN group variation

For each data value, we calculate its contribution to:

- BETWEEN group variation: $\left(\overline{x}_i - \overline{x}\right)^2$
- WITHIN group variation: $\left(x_{ij} - \overline{x}_i\right)^2$

An even smaller example

Suppose we have three groups

- Group 1: 5.3, 6.0, 6.7
- Group 2: 5.5, 6.2, 6.4, 5.7
- Group 3: 7.5, 7.2, 7.9

We get the following statistics:

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	3	18	6	0.49
Column 2	4	23.8	5.95	0.176667
Column 3	3	22.6	7.533333	0.123333

Excel ANOVA Output

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	5.127333	2	2.563667	10.21575	0.008394	4.737416
Within Groups	1.756667	7	0.250952			
Total	6.884	9				

1 less than number of groups

1 less than number of individuals (just like other situations)

number of data values - number of groups (equals df for each group added together)

Computing ANOVA F statistic

			WITHIN		**BETWEEN**	
			difference:		difference	
		group	data - group mean		group mean - overall mean	
data	group	mean	plain	squared	plain	squared
5.3	1	6.00	-0.70	0.490	-0.4	0.194
6.0	1	6.00	0.00	0.000	-0.4	0.194
6.7	1	6.00	0.70	0.490	-0.4	0.194
5.5	2	5.95	-0.45	0.203	-0.5	0.240
6.2	2	5.95	0.25	0.063	-0.5	0.240
6.4	2	5.95	0.45	0.203	-0.5	0.240
5.7	2	5.95	-0.25	0.063	-0.5	0.240
7.5	3	7.53	-0.03	0.001	1.1	1.188
7.2	3	7.53	-0.33	0.109	1.1	1.188
7.9	3	7.53	0.37	0.137	1.1	1.188
TOTAL				1.757		5.106
TOTAL/df				**0.25095714**		**2.55275**

overall mean: 6.44 F = 2.5528/0.25025 = 10.21575

Minitab ANOVA Output

Analysis of Variance for days

Source	DF	SS	MS	F	P
treatment	2	34.74	17.37	6.45	0.006
Error	22	59.26	2.69		
Total	24	94.00			

1 less than # of groups

# of data values - # of groups

(equals df for each group added together)

1 less than # of individuals (just like other situations)

Minitab ANOVA Output

Analysis of Variance for days

Source	DF	SS	MS	F	P
treatment	2	34.74	17.37	6.45	0.006
Error	22	59.26	2.69		
Total	24	94.00			

$$\sum_{obs} (x_{ij} - \bar{x}_i)^2$$

$$\sum_{obs} (x_{ij} - \bar{x})^2$$

$$\sum_{obs} (\bar{x}_i - \bar{x})^2$$

SS stands for sum of squares

- ANOVA splits this into 3 parts

Minitab ANOVA Output

Analysis of Variance for days

Source	DF	SS	MS	F	P
treatment	2	34.74	17.37	6.45	0.006
Error	22	59.26	2.69		
Total	24	94.00			

MSG = SSG / DFG
MSE = SSE / DFE

F = MSG / MSE

P-value comes from F(DFG,DFE)

(P-values for the F statistic are in Table E)

So How big is F?

Since F is

Mean Square Between / Mean Square Within

= MSG / MSE

A large value of F indicates relatively more
difference between groups than within groups
(evidence against H_0)

To get the P-value, we compare to *F(I-1,n-I)*-distribution

- *I-1* degrees of freedom in numerator (# groups -1)
- *n - I* degrees of freedom in denominator (rest of df)

Connections between SST, MST, and standard deviation

If ignore the groups for a moment and just compute the standard deviation of the entire data set, we see

$$s^2 = \frac{\sum \left(x_{ij} - \bar{x}\right)^2}{n-1} = \frac{SST}{DFT} = MST$$

So $SST = (n-1)\, s^2$, and $MST = s^2$. That is, *SST* and *MST* measure the TOTAL variation in the data set.

Connections between SSE, MSE, and standard deviation

Remember: $s_i^2 = \dfrac{\sum \left(x_{ij} - \bar{x}_i\right)^2}{n_i - 1} = \dfrac{SS[\text{Within Group } i]}{df_i}$

So $SS[\text{Within Group } i] = (s_i^2)\,(df_i)$

This means that we can compute SSE from the standard deviations and sizes (df) of each group:

$$SSE = SS[Within] = \sum SS[Within\,Group\ i]$$

$$= \sum s_i^2 (n_i - 1) = \sum s_i^2 (df_i)$$

Pooled estimate for st. dev

One of the ANOVA assumptions is that all groups have the same standard deviation. We can estimate this with a weighted average:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_I - 1)s_I^2}{n - I}$$

$$s_p^2 = \frac{(df_1)s_1^2 + (df_2)s_2^2 + \cdots + (df_I)s_I^2}{df_1 + df_2 + \cdots + df_I}$$

$$s_p^2 = \frac{SSE}{DFE} = MSE$$

so MSE is the pooled estimate of variance

In Summary

$$SST = \sum_{obs} (x_{ij} - \bar{x})^2 = s^2(DFT)$$

$$SSE = \sum_{obs} (x_{ij} - \bar{x}_i)^2 = \sum_{groups} s_i^{\,2}(df_i)$$

$$SSG = \sum_{obs} (\bar{x}_i - \bar{x})^2 = \sum_{groups} n_i(\bar{x}_i - \bar{x})^2$$

$$SSE + SSG = SST; \quad MS = \frac{SS}{DF}; \quad F = \frac{MSG}{MSE}$$

R^2 Statistic

R^2 gives the percent of variance due to between group variation

$$R^2 = \frac{SS[Between]}{SS[Total]} = \frac{SSG}{SST}$$

This is very much like the R^2 statistic that we computed back when we did regression.

Where's the Difference?

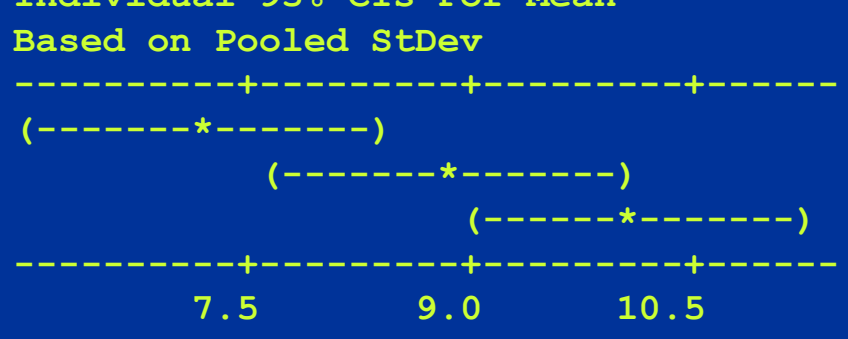
Once ANOVA indicates that the groups do not all appear to have the same means, what do we do?

Analysis of Variance for days

Source	DF	SS	MS	F	P
treatment	2	34.74	17.37	6.45	0.006
Error	22	59.26	2.69		
Total	24	94.00			

Level	N	Mean	StDev
A	8	7.250	1.669
B	8	8.875	1.458
P	9	10.111	1.764

Pooled StDev = 1.641



Clearest difference: P is worse than A (CI's don't overlap)

Multiple Comparisons

Once ANOVA indicates that the groups do not all have the same means, we can compare them two by two using the 2-sample t test

- We need to adjust our p-value threshold because we are doing multiple tests with the same data.
- There are several methods for doing this.
- If we really just want to test the difference between one pair of treatments, we should set the study up that way.

Tuckey's Pairwise Comparisons

```
Tukey's pairwise comparisons

    Family error rate = 0.0500
Individual error rate = 0.0199

Critical value = 3.55

Intervals for (column level mean) - (row level mean)

                A           B

      B      -3.685
              0.435

      P      -4.863      -3.238
             -0.859       0.766
```

95% confidence

Use alpha = 0.0199 for each test.

These give 98.01% CI's for each pairwise difference.

Only P vs A is significant (both values have same sign)

95% CI for A-P is (-0.86,-4.86)

Fisher's Pairwise Comparisons

Now we set the individual error rate (alpha) and see the overall error rate. 95% confidence on each corresponds to 88.1% confidence overall

```
Fisher's pairwise comparisons

    Family error rate = 0.119
Individual error rate = 0.0500

Critical value = 2.074

Intervals for (column level mean) - (row level mean)

                A           B

      B      -3.327
              0.077

      P      -4.515      -2.890
             -1.207       0.418
```

Content

Non-parametric Tests?

JAMES COOK UNIVERSITY SINGAPORE

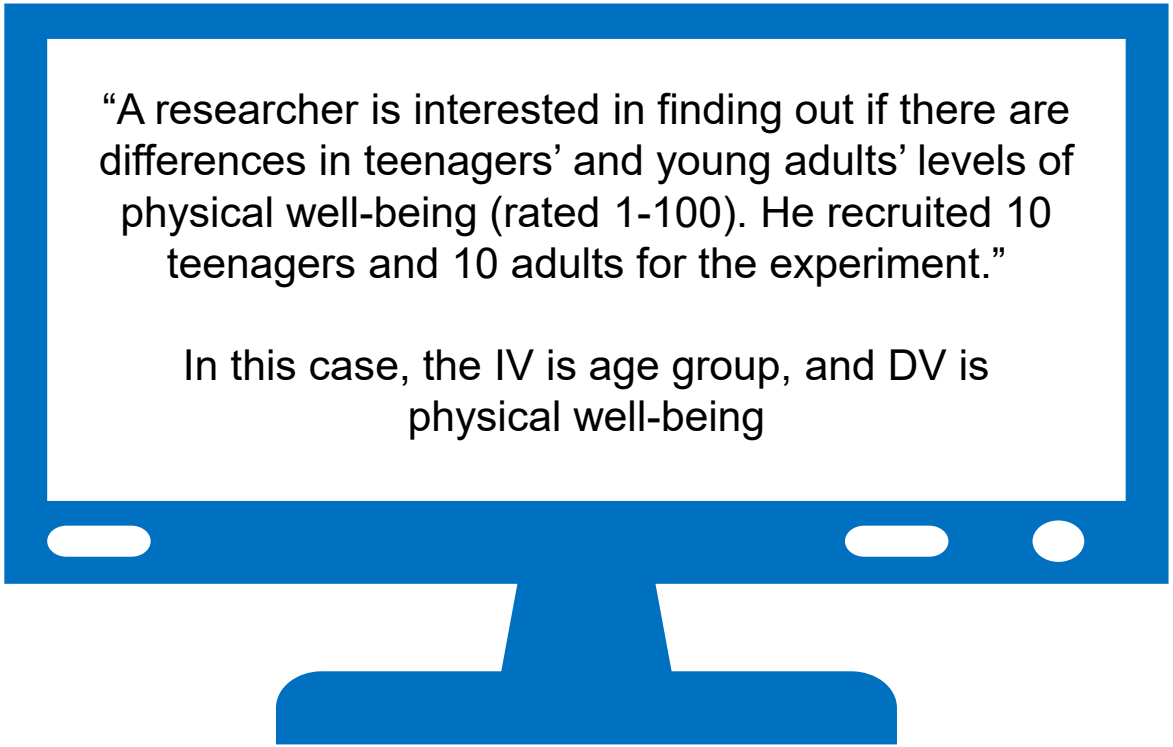
- While most common statistical analyses (e.g., t-tests, ANOVA) are parametric, they need to fulfil a number of criteria before we use them
- These criteria include satisfying the assumptions of outliers, linearity, normality, homoscedasticity, to name a few
- If the data do not fulfil the criteria to conduct the parametric tests, we can opt for non-parametric tests, which do not require those assumptions
- Do note that non-parametric tests make *less* assumptions, not *no* assumptions!
- The trade-off is that non-parametric tests are generally lower in power

Types of Non-parametric Tests

- In this set of slides, the focus is on 4 non-parametric tests
- Each of these 4 tests is a non-parametric version of *t*-tests and ANOVAs

Parametric Test		**Non-parametric Test**
Between Subjects t-test	→	Mann-Whitney *U* Test
Within Subjects t-test	→	Wilcoxon Signed Ranked Test
One-way Between Subjects ANOVA	→	Kruskal-Wallis One-way ANOVA
One-way Within Subjects ANOVA	→	Friedman's ANOVA

Mann-Whitney *U* Test



Location of SPSS Data Files for Practice

JAMES COOK UNIVERSITY SINGAPORE

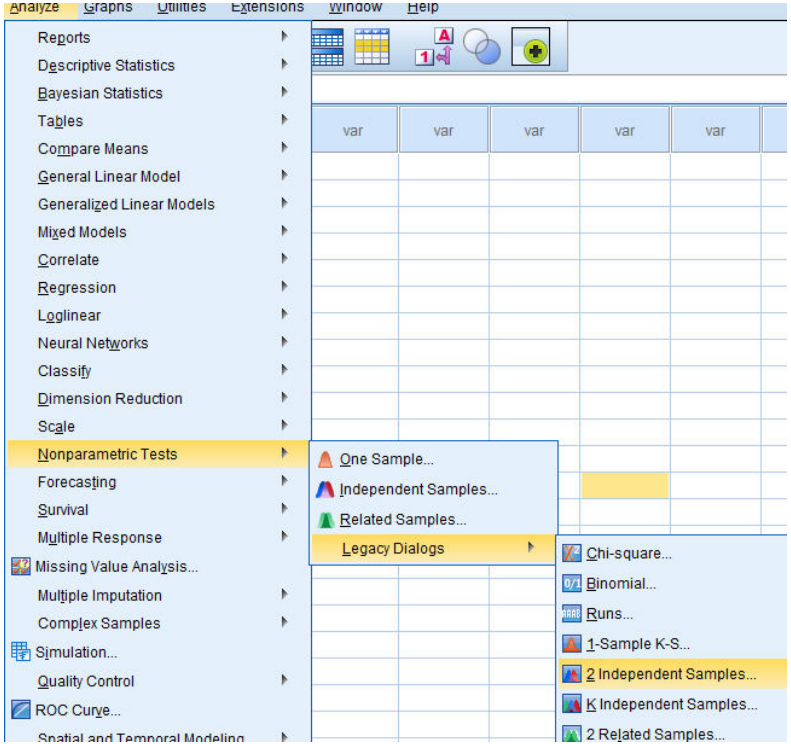
Example SPSS data for practice are available on **LearnJCU**:

Log in to LearnJCU -> Organisations -> Learning Centre JCU Singapore -> Learning Centre -> Statistics and Maths -> SPSS Data for Practice

Mann-Whitney *U* Test - SPSS

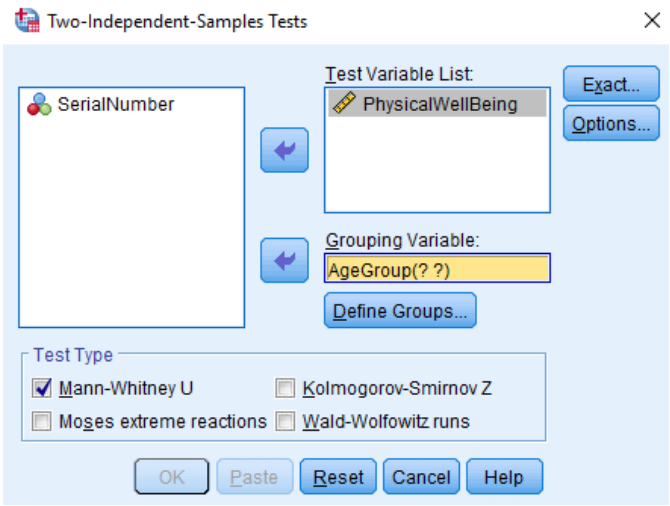
Assume that the data has multiple outliers, which is why the researcher opted to conduct a Mann-Whitney *U* test, rather than a t-test.

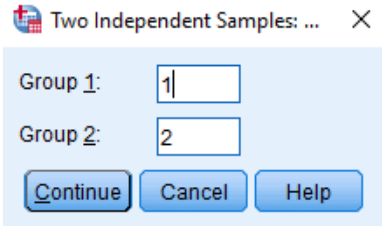
Analyze -> Nonparametrics Tests -> Legacy Dialogs -> 2 Independent Samples…



Mann-Whitney *U* Test - SPSS

1. Move *PhysicalWellBeing* (DV) to the right under Test Variable List
2. Move *AgeGroup* (IV) as our Grouping Variable
3. Then define groups by clicking on **Define Groups**
4. Input ‘1’ and ‘2’ as groups 1 and 2 respectively
5. Continue and OK!





Mann-Whitney *U* Test - SPSS

In a Mann-Whitney test, SPSS ranks the data (e.g., the lowest score of physical wellbeing gets a rank of 1, the next lowest score gets a rank of 2.

The value here displays the average of the rankings

This is the sum of all rankings in each group of the IV

Mann-Whitney Test

Ranks

	AgeGroup	N	Mean Rank	Sum of Ranks
PhysicalWellBeing	Teenager	10	13.45	134.50
	Adult	10	7.55	75.50
	Total	20		

Test Statistics[a]

	PhysicalWell Being
Mann-Whitney U	20.500
Wilcoxon W	75.500
Z	-2.238
Asymp. Sig. (2-tailed)	.025
Exact Sig. [2*(1-tailed Sig.)]	.023[b]

a. Grouping Variable: AgeGroup

b. Not corrected for ties.

Mann-Whitney *U* score = 20.5, $p = .03$

Given an alpha value of .05, there is a significant difference in teenagers' and adults' self reported physical wellbeing

Looking at the mean ranks, on average, teenagers reported higher physical wellbeing than adults

Write-Up

JAMES COOK UNIVERSITY SINGAPORE

An example write-up can be found on:

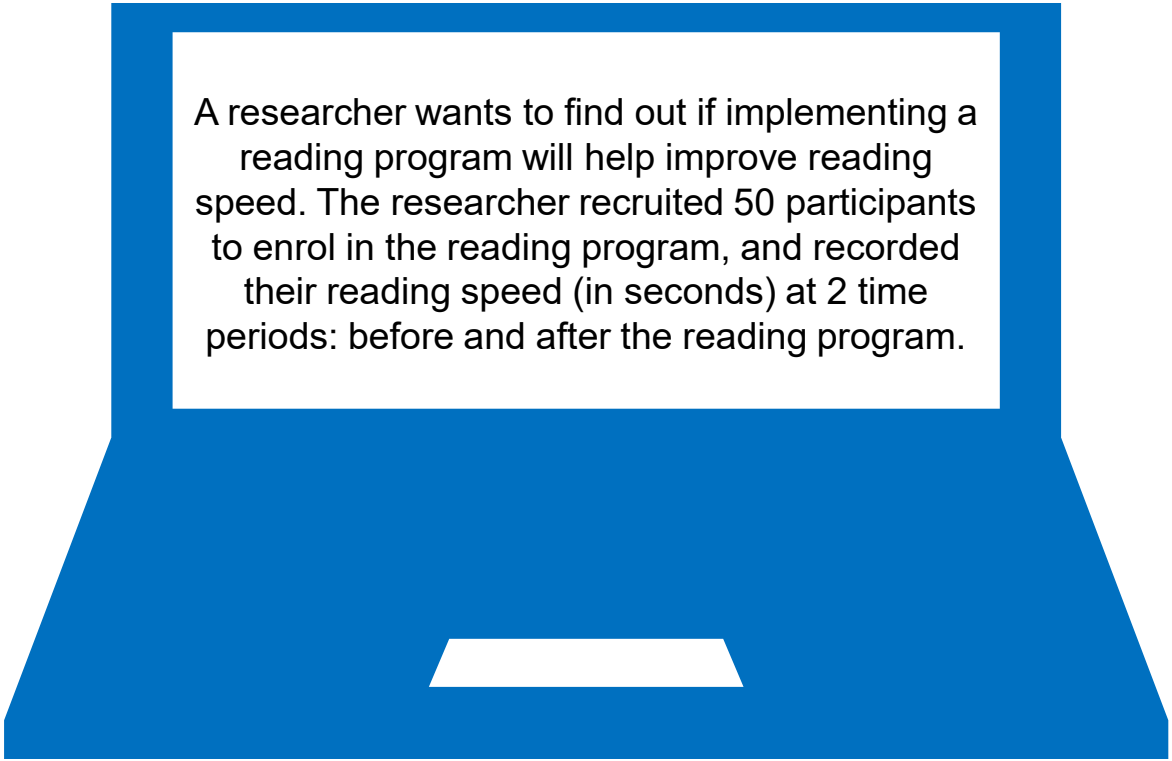
JCUS Learning Centre website -> Statistics and Mathematics Support

Types of Non-parametric Tests

JAMES COOK UNIVERSITY SINGAPORE

Parametric Test		Non-parametric Version
Between Subjects t-test	→	Mann-Whitney *U* Test
Within Subjects t-test	→	Wilcoxon Signed Ranked Test
One-way Between Subjects ANOVA	→	Kruskal-Wallis One-way ANOVA
One-way Within Subjects ANOVA	→	Friedman's ANOVA

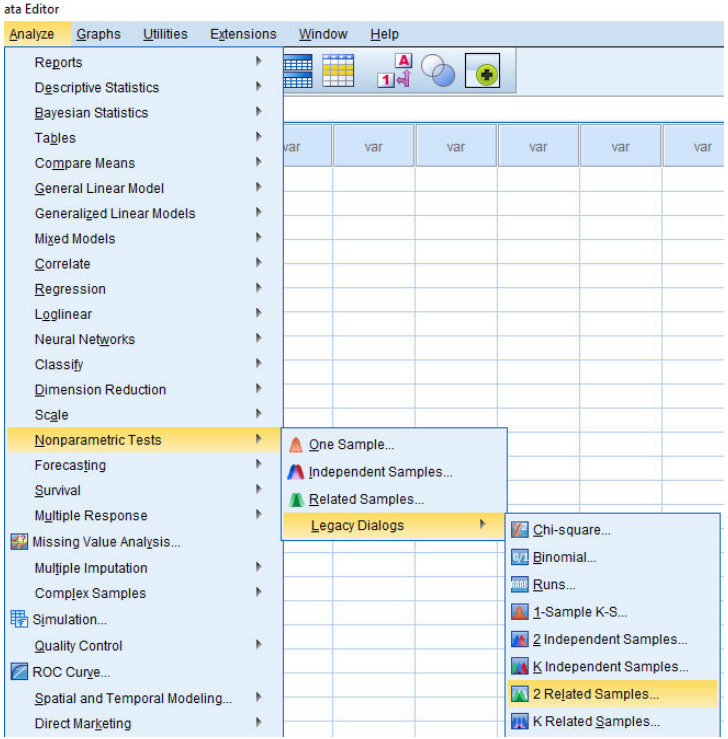
Wilcoxon Signed-Ranks Test



Wilcoxon Signed-Ranks Test - SPSS

Assume that the researcher only managed to recruit 10 participants, and opted to conduct a Wilcoxon signed ranked test, rather than a within subjects *t*-test.

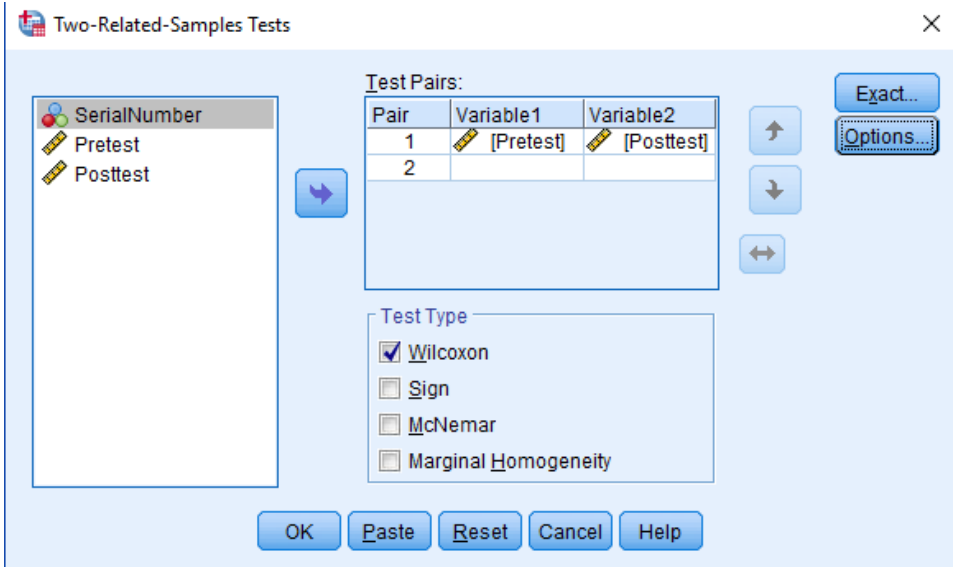
Analyze -> Nonparametrics Tests -> Legacy Dialogs -> 2 Related Samples….



Wilcoxon Signed-Ranks Test - SPSS

JAMES COOK UNIVERSITY SINGAPORE

1. Move *Pretest* and *Posttest* as Pair 1
2. Tick **Wilcoxon** in Test type
3. OK!



Wilcoxon Signed-Ranks Test - SPSS

The legend shows how negative, positive, and tied ranks are calculated. For example, there are 9 cases where a posttest score is lower than a pretest score. This means that in 9 of the 10 participants, reading speed improved after intervention

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
Posttest - Pretest	Negative Ranks	9[a]	6.00	54.00
	Positive Ranks	1[b]	1.00	1.00
	Ties	0[c]		
	Total	10		

a. Posttest < Pretest
b. Posttest > Pretest
c. Posttest = Pretest

Test Statistics[a]

	Posttest - Pretest
Z	-2.701[b]
Asymp. Sig. (2-tailed)	.007

a. Wilcoxon Signed Ranks Test
b. Based on positive ranks.

We are interested in the test statistic, which is -2.70 (Do note that in this case, this value is based on positive ranks)

p value is .007

Given an alpha value of .05, there is a significant difference between pre-test and posttest scores

Based on mean ranks, participants' reading speed improved after the reading program

Write-Up

JAMES COOK UNIVERSITY SINGAPORE

An example write-up can be found on:

JCUS Learning Centre website -> Statistics and Mathematics Support

Types of Non-parametric Tests

JAMES COOK UNIVERSITY SINGAPORE

Parametric Test		Non-parametric Version
Between Subjects t-test	→	Mann-Whitney *U* Test
Within Subjects t-test	→	Wilcoxon Signed Ranked Test
One-way Between Subjects ANOVA	→	Kruskal-Wallis One-way ANOVA
One-way Within Subjects ANOVA	→	Friedman's ANOVA

Kruskal-Wallis One-Way ANOVA

JAMES COOK UNIVERSITY SINGAPORE

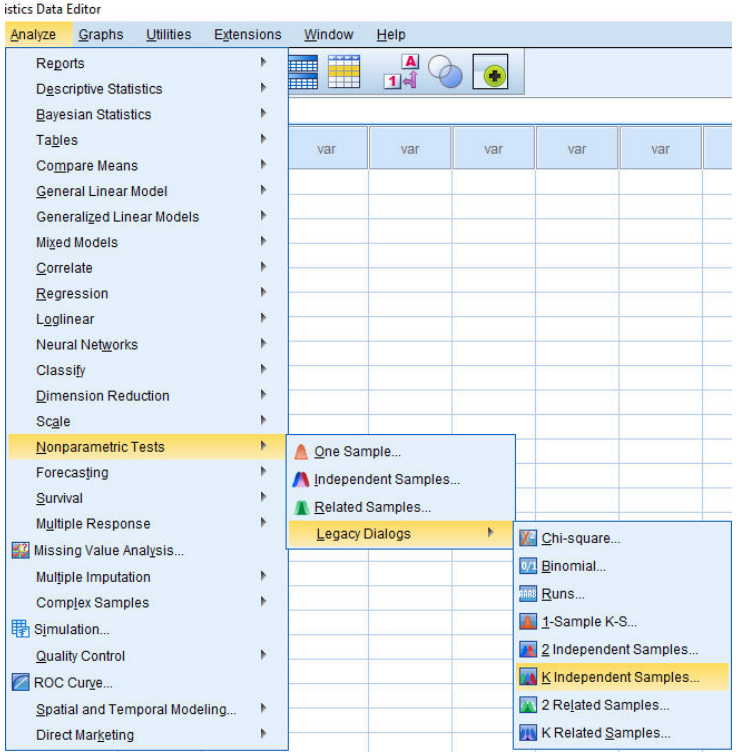
A researcher is interested in finding out if there is a difference in physical well-being (rated 1-100) among teenagers, young adults, and seniors. He recruited 10 teenagers, 10 adults, and 10 seniors for the experiment.

In this case, the IV is age group, and DV is physical well-being

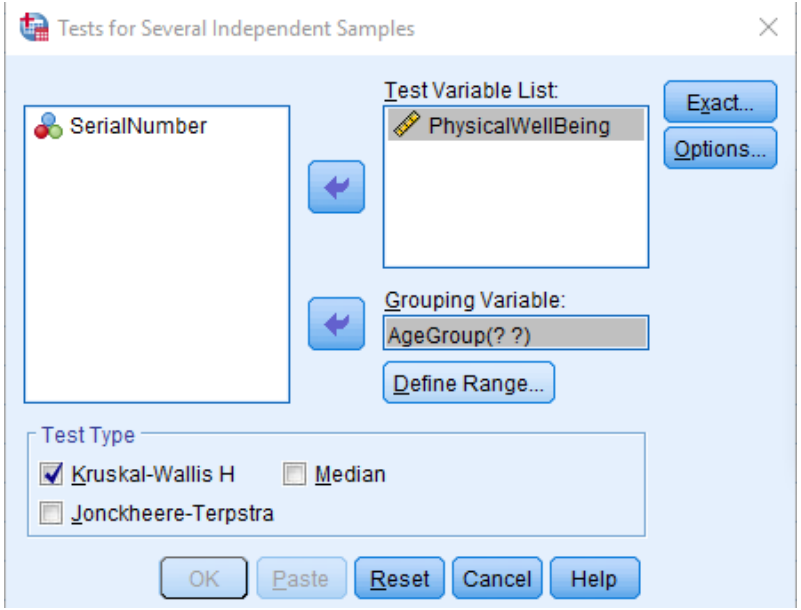
Kruskal-Wallis One-Way ANOVA

Assume that the data did not meet the criteria of parametric tests, thus the researcher opted to conduct a Kruskal-Wallis test.

Analyze -> Nonparametrics Tests -> Legacy Dialogs -> K Independent Samples….

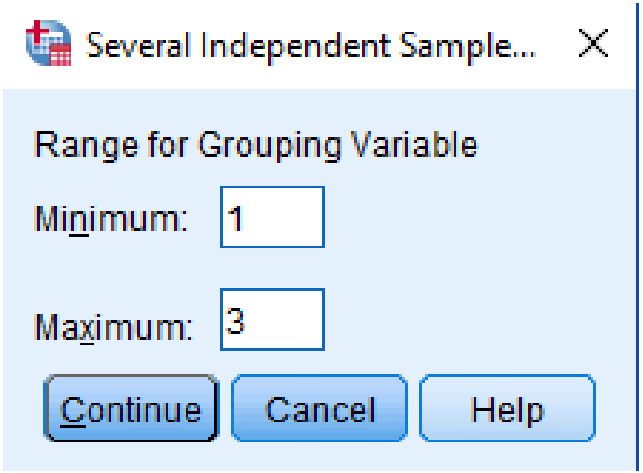


Kruskal-Wallis One-Way ANOVA



1. Move *PhysicalWellBeing* into the test variable list box, and *AgeGroup* into the grouping variable box
2. Tick Kruskal-Wallis H under Test type
3. Then define the grouping variable (**Define Range**)
4. Go to **Options** and select Descriptives

Kruskal-Wallis One-Way ANOVA



To define groups:

5. In our dataset, Teenagers were coded as ‘1’, Adults as ‘2’, and Seniors as ‘3’
6. Hence, the range for our grouping variable is 1-3; with a minimum of 1 and maximum of 3
7. Click Continue, and OK

Kruskal-Wallis One-Way ANOVA

Similar to Mann-Whitney *U* tests, SPSS ranks the data (e.g., the lowest score of physical wellbeing gets a rank of 1, the next lowest score gets a rank of 2.

The value here displays the average of the rankings

Kruskal-Wallis Test

Ranks

	AgeGroup	N	Mean Rank
PhysicalWellBeing	Teenager	10	21.70
	Adult	10	12.65
	Senior	10	12.15
	Total	30	

Test Statistics[a,b]

	PhysicalWell Being
Kruskal-Wallis H	7.501
df	2
Asymp. Sig.	.024

a. Kruskal Wallis Test

b. Grouping Variable: AgeGroup

Kruskal-Wallis H score = 7.50, p = .024

Given an alpha value of .05, there is a significant difference between teenagers', adults', and seniors' self reported physical wellbeing

However

- Although we now know that there is a significant difference between the 3 groups, we do not know exactly where the difference(s) lie
- It could lie between teenagers and adults, adults and seniors, teenagers and seniors, or even all of the above
- To test this, we conduct a post-hoc series of Mann-Whitney *U* tests to find out the answer (you can find out more on Mann-Whitney *U* tests in the earlier example)

Write-Up

JAMES COOK UNIVERSITY SINGAPORE

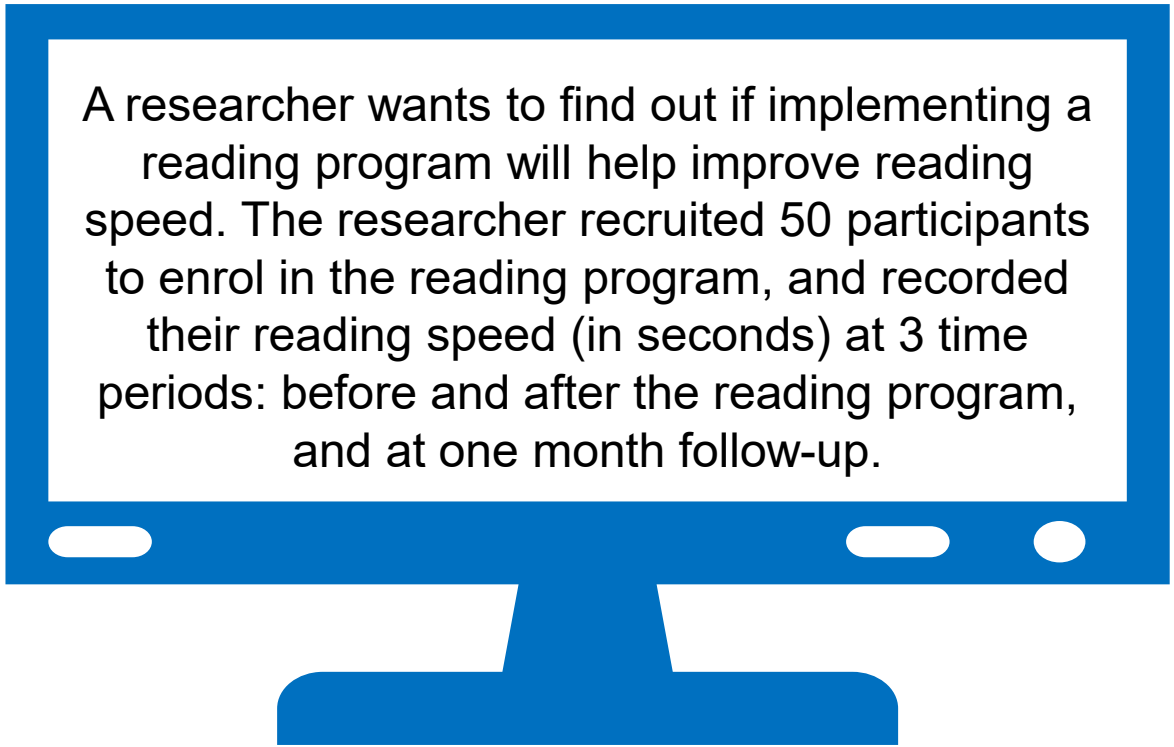
An example write-up can be found on page 294 in

Allen, P., Bennett, K., & Heritage, B. (2019). *SPSS Statistics: A Practical Guide* (4th ed.). Cengage Learning.

Types of Non-parametric Tests

Parametric Test		Non-parametric Version
Between Subjects t-test	→	Mann-Whitney *U* Test
Within Subjects t-test	→	Wilcoxon Signed Ranked Test
One-way Between Subjects ANOVA	→	Kruskal-Wallis One-way ANOVA
One-way Within Subjects ANOVA	→	Friedman's ANOVA

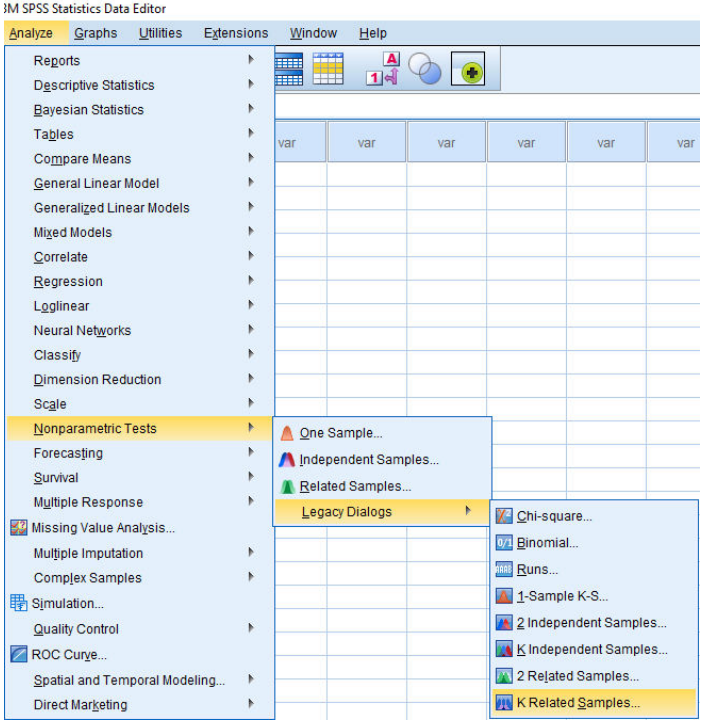
Friedman's ANOVA



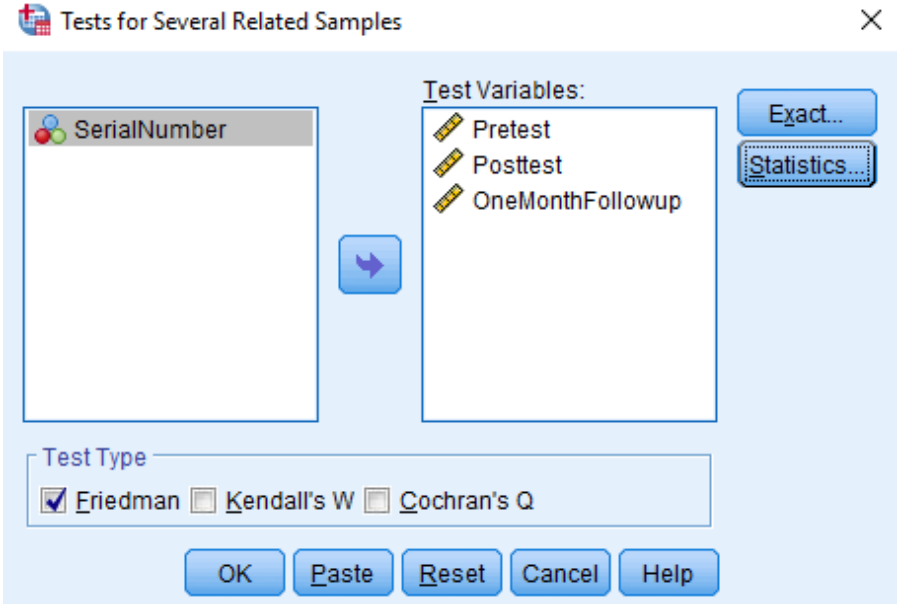
Friedman's ANOVA - SPSS

Assume that the data did not meet the criteria of parametric tests, thus the researcher opted to conduct a Friedman's ANOVA.

Analyze -> Nonparametrics Tests -> Legacy Dialogs -> K Related Samples….



Friedman's ANOVA - SPSS



1. Move *Pretest*, *Posttest*, and *OneMonthFollowup* inot the test variables box
2. Tick Friedman in Test type
3. Go to **Statistics** and select Descriptives
4. OK!

Friedman's ANOVA - SPSS

Friedman Test

Ranks

	Mean Rank
Pretest	2.90
Posttest	1.60
OneMonthFollowup	1.50

Test Statistics[a]

N	10
Chi-Square	12.200
df	2
Asymp. Sig.	.002

a. Friedman Test

Chi-square statistic = 12.2, *p* = .002

Given an alpha value of .05, there is a significant difference between pre-test, postttest, and the one month follow up

However

- Just like the Kruskal-Wallis test, although we now know that there is a significant difference between the three groups, we do not know exactly where the difference(s) lie
- Simply by eyeballing the mean ranks, we can probably guess that the difference comes from the improvement from pre-test to post-test (2.9 vs 1.6), but not so much from the post-test to one month follow-up (1.6 vs 1.5)
- To confirm this, we can conduct a series of post-hoc Wilcoxon Signed Ranks tests (you can find out more in the earlier example on Wilcoxon)

Write-Up

JAMES COOK UNIVERSITY SINGAPORE

An example write-up can be found on page 305 in

Allen, P., Bennett, K., & Heritage, B. (2019). *SPSS Statistics: A Practical Guide* (4th ed.). Cengage Learning.

JAMES COOK UNIVERSITY SINGAPORE

Questions?

learningcentre-singapore@jcu.edu.au

JAMES COOK UNIVERSITY AUSTRALIA

Cairns
Singapore
Townsville

www.jcu.edu.sg