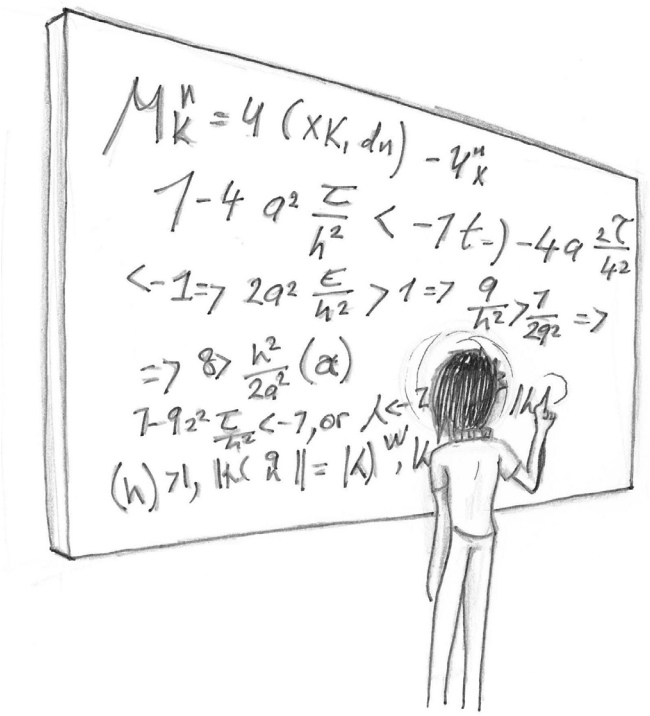
Chapter 11
Statistical Testing



The resolution of revolutions is selection by conflict within the scientific community of the fittest way to practice future science. The net result of a sequence of such revolutionary selections, separated by periods of normal research, is the wonderfully adapted set of instruments we call modern scientific knowledge.

Thomas S. Kuhn

The Structure of Scientific Revolutions (1962), 171.

Any publication on research and statistics cannot exist without a review of the most commonly used statistical tests. The scope here is not to exhaust all statistical testing. There are more than enough books for that dealing with specific disciplines. The scope is to give readers an idea of the basic tests that can be used in their research studies.

One can use this publication in conjunction with specialised publications as per lists below. Always check for new publications as each tackles new methods and case studies. There are two categories of book listed here. The first cover Generic Statistics Publications sorted by Publication Date whilst the second is more targeted. It is thematic in scope and lists publications that are sorted by theme such as: behavioural sciences, criminology, business, and others.

E-Books are also available from the following websites:

HyperStat
http://davidmlane.com/hyperstat/

Probability & Statistics
http://www.e-booksdirectory.com/listing.php?category=15

FreeBookCentre.Net
http://www.freebookcentre.net/SpecialCat/Free-Statistics-Books-Download.html

Generic Statistics Publications sorted by Publication Date

Title	Author/s	Publication Date	Publisher	ISBN-10	ISBN-13
Introductory Statistics, 7th edition	Prem S. Mann	2010	Wiley	0470444665	978-0470444665
Statistics Unplugged, 3rd edition	Sally Caldwell	2009	Wadsworth Publishing	0495602183	978-0495602187
Statistics for People Who (Think They) Hate Statistics: Excel 2007 Edition, 2nd edition	Dr. Neil J. Salkind (Editor)	2009	Sage Publications , Inc	1412971020	978-1412971027
Elementary Statistics: Picturing the World, 4th edition	Ron Larson and Betsy Farber	2008	Prentice Hall	0132424339	978-0132424332
Statistics: The Art and Science of Learning from Data, 2nd Edition	Alan Agresti and Christine Franklin	2008	Prentice Hall	0135131995	978-0135131992
Elementary Statistics: A Step By Step Approach, 7th edition	Allan Bluman	2008	McGraw-Hill Science/ Engineering /Math	0077302354	978-0077302351
Intro Stats, 3rd Edition	Richard D. De Veaux, Paul F. Velleman and David E. Bock	2008	Addison Wesley	0321500458	978-0321500458
Introduction to Statistics and Data Analysis, 3rd edition	Roxy Peck, Chris Olsen and Jay L. Devore	2008	Duxbury Press	0495557838	978-0495557838
Mathematical Statistics with Applications, 7th edition	Dennis Wackerly , William Mendenhall and Richard L. Scheaffer	2007	Duxbury Press	0495110817	978-0495110811

Title	Author/s	Publication Date	Publisher	ISBN-10	ISBN-13
Statistics, 4th edition	David Freedman, Robert Pisani and Roger Purves	2007	W. W. Norton & Company	0393929728	978-0393929720
Elementary Statistics (10th Edition) (MyStatLab Series), 10th edition	Mario F. Triola	2007	Addison Wesley	0321331834	978-0321331830
Statistics For The Terrified, 4th edition	Gerald Kranzler, Janet Moursund and John H. Kranzler	2006	Prentice Hall	0131930117	978-0131930117
Discovering Statistics Using SPSS (Introducing Statistical Methods S.), 2nd edition	Andy Field	2005	Sage Publications Ltd	0761944524	978-0761944522
Fundamentals of Statistics	Michael III Sullivan	2004	Prentice Hall	0131464493	978-0131464490
The Basic Practice of Statistics, 3rd edition	David S. Moore	2003	W.H. Freeman & Company	0716796236	978-0716796237
Statistics for Dummies	Deborah Rumsey	2003	For Dummies	0764554239	978-0764554230
The Craft of Research, 2nd edition (Chicago Guides to Writing, Editing, and Publishing)	Wayne C. Booth, Joseph M. Williams and Gregory G. Colomb	2003	University Of Chicago Press	0226065685	978-0226065687
Statistics Without Tears: A Primer for Non-Mathematicians (Allyn & Bacon Classics Edition)	Derek Rowntree	2003	Allyn & Bacon	0205395090	978-0205395095
The Visual Display of Quantitative Information, 2nd edition	Edward R. Tufte	2001	Graphics Press	0961392142	978-0961392147
Your Statistical Consultant: Answers to Your Data Analysis Questions	Dr. Rae R. Newton and Dr. Kjell E. (Erik) Rudestam	1999	Sage Publications, Inc	0803958234	978-0803958234
How to Lie with Statistics	Darrell Huff and Irving Geis	1993	W. W. Norton & Company	0393310728	978-0393310726
Cartoon Guide to Statistics	Larry Gonick and Woollcott Smith	1993	Collins Reference	0062731025	978-0062731029

Thematic Publications sorted by Theme

Title	Author/s	Publication Date	Publisher	ISBN-10	ISBN-13
Essentials of Statistics for the **Behavioral Sciences**, 7th edition	Frederick J Gravetter, Larry B. Wallnau and Jon-David Hague	2010	Wadsworth Publishing	049581220X	978-0495812203
Statistics for the Behavioral Sciences, 8th edition	Frederick J Gravetter and Larry B. Wallnau	2008	Wadsworth Publishing	0495602205	978-0495602200
Comprehending Behavioral Statistics (with CD-ROM), 4th edition	Russell T. Hurlburt	2005	Wadsworth Publishing	053460627X	978-0534606275
Applied Statistics for the Behavioral Sciences, 5th edition	Dennis E. Hinkle, William Wiersma and Stephen G. Jurs	2002	Wadsworth Publishing	0618124055	978-0618124053
Statistics for **Criminal Justice and Criminology**, 3rd edition	Dean J. Champion and Richard D. Hartley	2009	Prentice Hall	0136135854	978-0136135852
Simple Statistics: Applications in Criminology and Criminal Justice	Terance D. Miethe	2006	Oxford University Press, USA	0195330714	978-0195330717
Research Methods for Criminology and Criminal Justice: A Primer (Criminal Justice Illuminated), 2nd edition	Dantzker, M.L., and Hunter, R.D.	2005	Jones & Bartlett Pub	0763736155	978-0763736156
Statistics for **Business and Economics** (with Bind-In Card), 11th edition	David R. Anderson , Dennis J. Sweeney and Thomas A. Williams	2010	South-Western College Pub	0324783248	978-0324783247
The Practice of Business Statistics w/CD, 2nd edition	David S. Moore, George P. McCabe, William M. Duckworth and Layth Alwan	2008	W. H. Freeman	142922150X	978-1429221504
Statistical Techniques in Business and Economics with Student CD, 13th edition	Douglas Lind , William Marchal and Samuel Wathen	2006	McGraw-Hill/Irwin	0073272965	978-0073272962
Research Methods in **Public Administration** and Nonprofit Management: Quantitative and Qualitative Approaches, 2nd edition	David E. McNabb	2008	M.E. Sharpe	0765617676	978-0765617675
Essential Statistics For Public Managers and Policy Analysts, 2nd edition	Evan M Berman	2006	CQ Press	0872893014	978-0872893016
An SPSS Companion to **Political** Analysis, 3rd edition	Philip H. Pollock III	2008	CQ Press	0872896072	978-0872896079
Damned Lies and Statistics: Untangling Numbers from the Media, Politicians, and Activists	Joel Best	2001	University of California Press	0520219783	978-0520219786
Statistics: A Tool for **Social** Research, 8th edition	Joseph F. Healey	2008	Wadsworth Publishing	0495096555	978-0495096559
Statistics Explained: A Guide for Social Science Students, 2nd edition	Perry Hinton	2004	Routledge	0415332850	978-0415332859
Statistics for Social Data Analysis, 4th edition	David Knoke, George W. Bohrnstedt and Alisa Potter Mee	2002	Wadsworth Publishing	0875814484	978-0875814483

Title	Author/s	Publication Date	Publisher	ISBN-10	ISBN-13
Basic Statistics for **Social Workers**, revised edition	Robert A. Schneider	2010	University Press of America	0761849327	978-0761849322
Statistics for Social Workers, 8th edition	Robert W. Weinbach and Richard M. Grinnell	2009	Prentice Hall	0205739873	978-0205739875
Handbook of Research on Civic Engagement in **Youth**	Lonnie R. Sherrod, Judith Torney-Purta and Constance A. Flanagan	2010	Wiley	0470522747	978-0470522745
Quantitative Research in **Education**: A Primer	Wayne K. (Kolter) Hoy	2009	Sage Publications, Inc	1412973260	978-1412973267
Study Guide for Essentials of **Nursing** Research: Appraising Evidence for Nursing Practice, 7th edition	Denise F Polit and Cheryl Tatano Beck	2009	Lippincott Williams & Wilkins	0781785812	978-0781785815
Applied Spatial Statistics for Public Health Data	Lance A. Waller and Carol A. Gotway	2004	Wiley-Interscience	0471387711	978-0471387718
Statistics for **Psychology**, 4th edition	Arthur Aron, Elaine N. Aron and Elliot Coups	2005	Prentice Hall	0131931679	978-0131931671
Understanding Research Methods and Statistics: An Integrated Introduction for Psychology, 2nd edition	Gary Heiman	2000	Wadsworth Publishing	0618043047	978-0618043040
Planning, **Construction**, and Statistical Analysis of Comparative Experiments (Wiley Series in Probability and Statistics)	Francis G. Giesbrecht and Marcia L. Gumpertz	2004	Wiley-Interscience	0471213950	978-0471213956
Elementary Statistics for **Geographers**, 3rd Edition	James E. Burt, Gerald M. Barber and David L. Rigby	2009	The Guilford Press	1572304847	978-1572304840
Practical Statistics for **Environmental** and **Biological** Scientists	John Townend	2002	Wiley	0471496650	978-0471496656
Using Statistics to Understand the Environment (Routledge Introductions to Environment)	Penny A. Cook and C. Phillip Wheater	2000	Routledge	0415198887	978-0415198882
Statistics for the Environment, Pollution Assessment and Control, Volume 3, 3rd edition	Vic Barnett and K. Feridun Turkman (Editors)	1997	Wiley	0471964352	978-0471964353
Environmental Statistics and Data Analysis	Wayne R. Ott	1995	CRC-Press	0873718488	978-0873718486
Environmental Statistics, Assessment, and Forecasting	C. Richard Cothern and N. Phillip Ross	1993	Lewis Publishers	0873719360	978-0873719360
Introduction to **Engineering** Statistics and Six Sigma: Statistical Quality Control and Design of Experiments and Systems	Theodore T. Allen	2006	Springer	1852339551	978-1852339555
Applied Statistics for **Marine** Affairs Professionals	Niels West	1996	Praeger	0275951723	978-0275951726

Title	Author/s	Publication Date	Publisher	ISBN-10	ISBN-13
Handbook of **Spatial** Statistics (Chapman & Hall/CRC Handbooks of Modern Statistical Methods)	Alan E. Gelfand, Peter Diggle, Peter Guttorp and Montserrat Fuentes	2010	CRC Press	1420072870	978-1420072877
Spatial Statistics and Modeling (Springer Series in Statistics)	Carlo Gaetan and Xavier Guyon	2009	Springer	0387922563	978-0387922560
Applied Spatial Data Analysis with R (Use R)	Roger S. Bivand , Edzer J. Pebesma and Virgilio Gómez-Rubio	2008	Springer	0387781706	978-0387781709
Statistical Methods for Spatial Data Analysis (Chapman & Hall/CRC Texts in Statistical Science)	Oliver Schabenberger and Carol A. Gotway	2004	Chapman and Hall/CRC	1584883227	978-1584883227
Spatial Statistics through Applications (Advances in Ecological Sciences)	J. Mateu and F. Montes (Editors)	2002	WIT Press / Computational Mechanics	1853126497	978-1853126499
Statistics for Spatial Data (Wiley Series in Probability and Statistics), revised sub edition	Noel A. C. Cressie	1993	Wiley-Interscience	0471002550	978-0471002550

Statistical testing helps researchers to control and validate the analysis carried out in their studies. These tests ensure that errors are not committed during the course of an analytical process. In addition, one should also be able to identify the quantity of errors generated.

There are many tools available for research, some are simple whilst other are quite complex and require increasing levels of tests to ensure precision and accuracy. This chapter will cover a few of the simplest tests ranging from measures of central tendency to regression analysis.

Before reviewing some basic statistics, it is best to define four words that are mostly used in statistical analysis: Descriptive and Inferential Statistics as well as Independent and Dependent Variables.

Descriptive Statistics are used to describe a dataset quantitatively through summarization rather than through the usage of probability analysis. Examples of descriptive statistics include the measures of central tendency (Mean, Median and Mode), standard deviation and variance.

Inferential Statistics, also called inductive statistics, on the other hand, employ probability tests through comparative tests that allow one to infer on a population. Inferential tests include the Z-Score, the T-Tests, the ANOVA and the Chi squared. When researchers present their data employing mainly the inferential test, the submission of descriptive statistics is still deemed necessary as they enhance any study and aid in the understanding of the results.

Basic Statistics

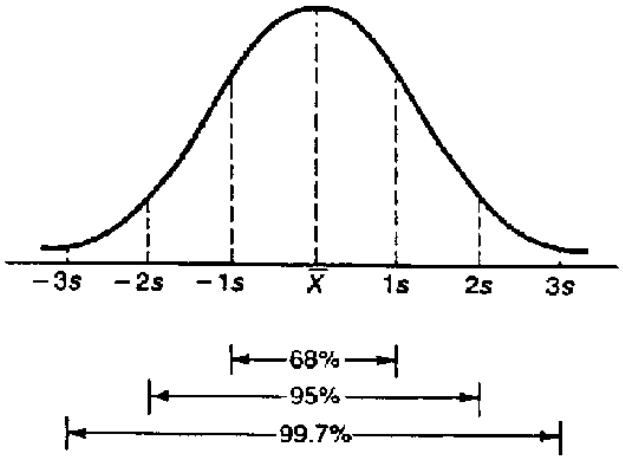
In statistics a basic assumption is made, which revolves around the issue that a set of data has a **Normal** or **Gaussian** distribution. Behavioral sciences' statistical tests assume a normal distribution. These tests can be used even where the distribution is only approximately normal.

Normal distributions are consonant with values that are more concentrated in the middle of the distribution curve than in the tails. They are defined by two parameters: the mean (μ) and the standard deviation (σ).

The Normal assumes that the data has very specific ranges within which that data falls, abiding by the **Empirical Rule** which states that in a **normal distribution:**

- about 68% of the scores are within one standard deviation of the mean ($\mu \pm \sigma$) and

- about 95% of the scores are within two standard deviations of the mean ($\mu \pm 2\sigma$) and
- about 99% are within three standard deviations of the mean ($\mu \pm 3\sigma$)



1. Measures of Central Tendency

As you would recall in Chapter 9 we mention the measures of central tendency. To recapitulate these measures refer to the values that are either at the middle point of a set of data or are typical of that type of data. There are three measures of Central Tendency: the Mean, Median and Mode.

1. **Mean** or "**average**" value: This value computes the central tendency of a frequency distribution ex **Interval** / **Ratio** data;

2. **Median** or **middle** value: Appropriate measure of central tendency for **Ordinal** level data;

3. **Mode** or **most frequent value**: Providing the least precise information about central tendency for **Nominal** (categorical) data.

Now let us go on to working out the Measures of Central Tendency.

1. The Mean

The Mean is the score located at the mathematical centre of a distribution and represents the arithmetic mean which is also called the average: The mean is calculated as the sum of all the scores divided by the number of scores.

The Greek letter Σ (a capital sigma) is used to designate summation.

The Mean Formula

Mean = sum of elements / number of elements
= a1+a2+a3+.....+an/n

Note that sometimes it is difficult to calculate the mean of a whole population as that would take forever, thus sometimes it is best to use a sample and calculate the mean for that.

The table below gives the formulas for both the whole population and a sample population. In the following examples we are assuming that the sample population is being analysed.

	Sample	Population
Mean	$\bar{X}$ (X-BAR)	μ (mu)
Variable	X	X
Add up all the Scores	ΣX	ΣX
No of Scores	N	N
Formula	$\bar{x} = \frac{\sum x_i}{n}$	$\mu = \frac{\sum x}{N}$

Each section contains a few working examples. Kindly attempt to analyse the figures based on the formula given for each measure.

- Mean: Calculations
    - Calculate the Mean for the following:

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54
Q2 16.5, 18, 63.2, 88.88, 19, 20, 21, 22, 18.5
Q3 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1000000

Worked Example: Q1 - Mean

Formula =

$$\bar{x} = \frac{\sum x_i}{n}$$

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54

Step 1: Add all the scores

Σ^x = 15 + 3 + 48 + 23 + 8 + 18 + 6 + 19 + 54

Σ^x = 194

Step 2: Count the number of Scores (**N**) = **9**

Step 3: Divide the Sum by the Count

$\bar{x} = \Sigma^x / N$

$\bar{x} = 194 / 9$

$\bar{x} = 21.56$

Mean = 21.56

- Answers

Q1 21.56
Q2 31.90
Q3 90914.10 – should we use this or another measure since the figure 1000000 is such a large outlier?

Note that should one outcome be registered as far from the rest of the data, this number is called an outlier as in Q3 above, which would strongly affect the data. One can use an alternate measure called the median.

2. The Median

The Median refers to the score located at the 50th percentile. The median allows researchers to identify that middle value which serves as a divider between the two halves of a dataset. Thus, Median is the middle score.

Symbol is M or Mdn

The Median Formula

Median = position of the value
= number of elements / 2

= (N + 1)/2

There are various sequential steps to calculate the Median:

1. Sort the observations smallest to largest;
2. Compute (n + 1)/2. This gives the *position* of the median (not the median itself) in the ordered data set;
3. Then find the corresponding number in the ordered set;
4. Median = number of units plus 1/2
   = (n + 1)/2 = (5 + 1)/2 = 6/2 = 3;
5. What happens when there are two elements in the middle? (occurs in even number of elements) – both middle ones are chosen.

- Median: Calculations
  - Calculate the Median for the following:

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54
Q2 16.5, 18, 63.2, 88.88, 19, 20, 21, 22, 18.5
Q3 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1000000
Q4 1, 2, 3, 4, 5, 6

Worked Example: Q1 - Median

Formula: Mdn = (n + 1)/2

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54

Step 1: Sort the scores

3, 6, 8, 15, 18, 19, 23, 48, 54

Step 2: Add the number of scores to 1 and divide by 2 (where **n** = **9**)

Mdn = (n + 1)/2

Mdn = (9 + 1)/2

Mdn = (10)/2

Mdn = 5 (the fifth score)

Step 3: Check which score is in the fifth position

3, 6, 8, 15, 18, 19, 23, 48, 54

3, 6, 8, 15, 18, 19, 23, 48, 54

Median = 18

- Median: Calculations
    - Answers

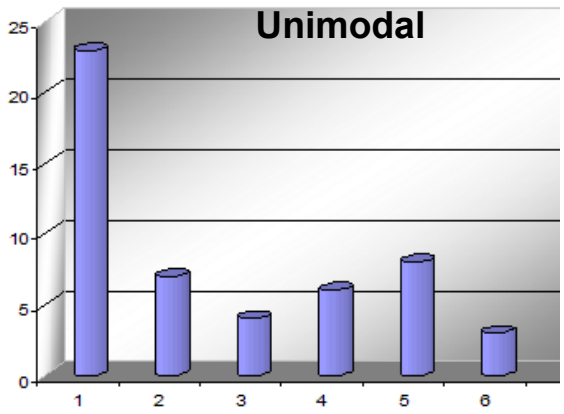
Q1 18
Q2 20
Q3 6
Q4 (3+4)/2 = 3.5

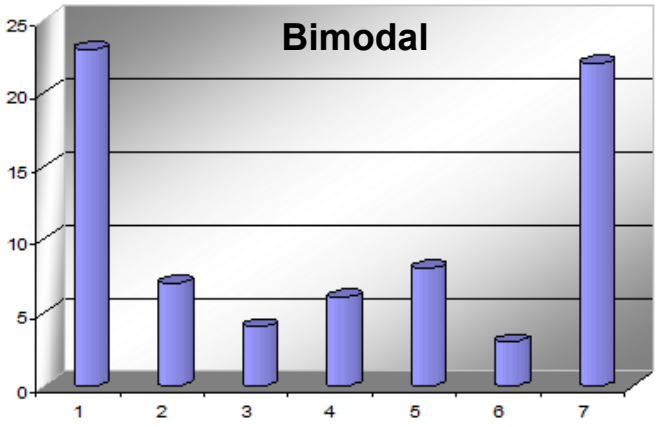
(actually 3rd and 4th Place – the mean of those can be calculated)

3. Mode

The Mode refers to the score that occurs most frequently: example there are more females than males in the elderly age cohorts. There are two types of Mode: the Unimodal and the Bimodal.

- The Unimodal type has one peak (highest point in a distribution – indicating the most frequent score)
- The Bimodal type has two peaks (highest point in a distribution – indicating the most frequent score)





There are various sequential steps to calculate the Median:

- Run a test in Microsoft Excel or another spreadsheet;
- Then run a pivot table
- The element with the highest Count renders the Mode

- Mode: Calculations
    - Calculate the Mode for the following:

Q1 15, 3, 3, 23, 8, 8, 6, 48, 23, 8, 18, 6, 19, 54
Q2 16.5, 18, 63.2, 1, 1, 15, 15, 1
Q3 14, 18, 14, 1, 1, 15, 15, 1, 15

Worked Example: Q1 - Mode

Formula: The score with the largest number of instances

Q1 15, 3, 3, 23, 8, 8, 6, 48, 23, 8, 18, 6, 19, 54

Step 1: Sort the scores

3, 3, 6, 6, 8, 8, 8, 15, 18, 19, 48, 54

Step 2: Check how many instances there are for each score

3	**6**	**8**	**15**	**18**	**19**	**48**	**54**
2	**2**	**3**	**1**	**1**	**1**	**1**	**1**

Step 3: Check which score has the largest number of instances

3	**6**	**8**	**15**	**18**	**19**	**48**	**54**
2	**2**	**3**	**1**	**1**	**1**	**1**	**1**

8 has 3 instances

Mode = 8

(since the mode falls on a score and not between scores, it is termed unimodal)

- Mode : Calculations
    - Answers

Q1 8 – unimodal
Q2 1 – unimodal
Q3 1, 15 – bimodal

Proportions and Percentages

These measures are heavily dependent on the issue of proportion. Such refers to the degree that an attribute is found within a population. One can calculate this through defining whether one needs to depict the degree as a fraction or as a percentage.

- **Proportions**

Proportions refer to fractions of the total. As an example one can state that the fraction of Maltese who have brown hair (300,000 of 400,000) equates to a proportion of ¾ or 0.75.

300,000/400,000 = ¾ or 0.75

- **Percentages**

Percentages refer to the same method as proportions but expressed as a figure out of 100. In effect:

Percentage = proportion * 100

Therefore the example above results in a figure of 75 percent:

0.75 * 100 = 75%

2. Measures of Variability

The next step to understand concerns the fact that numbers are rarely found aggregated around a single figure such as the improbable state where all the population of Valletta is aged 35. Since we study real life populations our study group rarely falls under the same number. In fact, all populations range from 0 to 120 in extreme cases. All start at Day 1 and have a somewhat different end Date!

How does one calculate for such a variation in numbers? There must be hundreds of hundreds of thousands individuals in a population which we cannot calculate individually! This is where measures of Variability or of Dispersion come in. There are three parameters that help in understanding such variability: the **Range**, the **Standard Deviation** and the **Variance**.

The next steps analyse each of the parameters in depth.

Each section contains a few working examples. Kindly attempt to analyse the figures based on the formula given for each measure.

1. Min – Max – Range

Whilst the first parameters refer to the Range, it is best to understand Range through its components. The Range is defined as the difference between the two extremes in the data range: the **Minimum** and **Maximum**.

Range = Maximum - Minimum

A. The Minimum (Min)

The Minimum (Min) refers to the smallest number in the dataset.

- Min Calculations

- Calculate the Min for the following:

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54
Q2 16.5, 18, 63.2, 88.88, 19, 20, 21, 22, 18.5
Q3 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1000000
Q4 1, 2, 3, 4, 5, 6

- Min Calculations
    - Answers

Q1 3
Q2 16.5
Q3 1
Q4 1

B. The Maximum (Max)

The Maximum (Max) refers to the largest number in the dataset.

- Max Calculations
    - Calculate the Max for the following:

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54
Q2 16.5, 18, 63.2, 88.88, 19, 20, 21, 22, 18.5
Q3 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1000000
Q4 1, 2, 3, 4, 5, 6

- Max Calculations
    - Answers

Q1 54
Q2 88.88
Q3 1000000
Q4 6

The Range

The Range as already stated refers to the difference between the Minimum and the Maximum values.

Range = Maximum - Minimum

- Range Calculations
    - Calculate the Range for the following:

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54
Q2 16.5, 18, 63.2, 88.88, 19, 20, 21, 22, 18.5
Q3 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1000000
Q4 1, 2, 3, 4, 5, 6

Worked Example: Q1 - Range

Formula: Range = Max - Min

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54

Step 1: Sort the scores

3, 6, 8, 15, 18, 19, 23, 48, 54

Step 2: identify the smallest and the largest number

(smallest) **3, 6, 8, 15, 18, 19, 23, 48, 54** (largest)

Step 2: Deduct the smallest from the largest

Range = **54** – **3**

Range = 51

- Range Calculations
    - Answers

Q1 51
Q2 72.38
Q3 9 (remove the outlier)
Q4 5

2. Standard Deviation

Standard Deviation is a widely used measure to calculate the deviation (dispersion) of the data around the mean. It helps researchers to understand the structure of their data in terms of how the individual observations deviate from or vary around the mean of that variable.

Thus, standard deviation allows for variation and no variation can exist where the standard deviation is marked as 0. The larger the spread of the data, the larger the standard deviation.

Standard Deviation is designated as σ (sigma)

As indicated in the opening section the following table depicts how many values normally fall within each standard deviation.

1 Standard Deviation	**68% of cases within a normal distribution would fall within one standard deviation of the mean**
2 Standard Deviations	**95% of the cases would be catered for**
3 Standard Deviations	**99% of the cases would be catered for**

It is best to understand how one can calculate deviation. The following simple examples depict the differences from the Mean.

18	**18**	**18**	**18**	**19**	**19**	**19**	**20**	**20**	**20**	**20**
-1	**-1**	**-1**	**-1**	**0**	**0**	**0**	**+1**	**+1**	**+1**	**+1**

- 18 deviates -1 from the Mean
- 20 deviates +1 from the Mean

1	1	1	1	1	10	19	19	19	19	19
-9	-9	-9	-9	-9	0	+9	+9	+9	+9	+9

- 1 deviates -9 from the Mean
- 19 deviates +9 from the Mean

Standard Deviation is calculated as follows:

> Subtract the mean from all of the numbers, square the differences, find the average of all of these squared-differences and finally take the square-root; in short one calculates the 'root-mean-square-deviation' about the mean.

The Formulae

Note that the standard deviation for a population and that for a sample are slightly different.

Population	Sample
$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n - 1}}$	$\sigma = \sqrt{\frac{\Sigma\ [x - \bar{x}]^2}{n}}$

This section contains a few working examples. Kindly attempt to analyse the figures based on the process given above.

- Standard Deviation Calculations
    - Calculate the Standard Deviation for the following:

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54
Q2 16.5, 18, 63.2, 88.88, 19, 20, 21, 22, 18.5
Q3 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Q4 1, 2, 3, 4, 5, 6

Worked Example: Q1 – Standard Deviation

Formula:

$$\sigma = \sqrt{\frac{\Sigma\ [x - \bar{x}]^2}{n}}$$

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54

Step 1: Find the Mean (or the average of the scores) – Refer to workings for Mean above (Section 1)

15, 3, 48, 23, 8, 18, 6, 19, 54

$\bar{x}$ = **21.56**

Step 2: Find the Deviation of each of the scores from the Mean
(**x - $\bar{x}$**) (score minus mean)

15 - **21.56 = - 6.56**
3 - **21.56 = - 18.56**
48 - **21.56 = 26.44**
23 - **21.56 = 1.44**
8 - **21.56 = - 13.56**
18 - **21.56 = - 3.56**
6 - **21.56 = - 15.56**
19 - **21.56 = - 2.56**
54 - **21.56 = 32.44**

Step 3: Square the deviations found in Step 2. This step amplifies the positive numbers and changes the negative results to positive results $(\mathbf{x - \bar{x}})^2$

$-6.56^2 = 43.03$
$-18.56^2 = 344.47$
$26.44^2 = 699.07$
$1.44^2 = 2.07$
$-13.56^2 = 183.87$
$-3.56^2 = 12.67$
$-15.56^2 = 242.11$
$-2.56^2 = 6.55$
$32.44^2 = 1052.35$

Step 4: Sum the obtained squares (as a first step to obtaining an average) $\Sigma(\mathbf{x - \bar{x}})^2$

= 43.03 + 344.47 + 699.07 + 2.07 + 183.87 + 12.67 + 242.11 + 6.55 + 1052.35

= 2586.22

Step 5: Divide the Sum the obtained squares by the number of scores $\Sigma(\mathbf{x - \bar{x}})^2$ **/ n**

= 2586.22 / 9

= 287.36

Step 6: To find the Standard Deviation, run a square root of then result of Step 5 $\sqrt{\Sigma(\mathbf{x - \bar{x}})^2/n}$

$$\sigma = \sqrt{\frac{\Sigma\ [x - \bar{x}]^2}{n}}$$

= $\sqrt{287.36}$

= 16.95

Standard Deviation = 16.95

- Standard Deviation Calculations
    - Answers

Q1 16.95
Q2 24.41
Q3 2.87
Q4 1.71

3. Variance

The variance is defined as the sum of the squared deviations from the mean, divided by n-1. It is computed as the average squared deviation of each number from its mean.

The Variance is designated as σ^2 (sigma squared) (S^2 for a sample)

In other words, the variance is the square of the standard deviation. Thus, vice versa, the standard deviation formula is very simple: it is the square root of the variance. Both variance and standard deviation provide the same information; one can always be obtained from the other.

The Formulae

Note that the standard deviation for a population and that for a sample are slightly different.

Population	Sample
$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$	$S^2 = \frac{\Sigma (X-\bar{X})^2}{n-1}$

- Variance Calculations
    - Calculate the Variance for the following:

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54
Q2 16.5, 18, 63.2, 88.88, 19, 20, 21, 22, 18.5
Q3 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Q4 1, 2, 3, 4, 5, 6

Worked Example: Q1 – Variance

Formula:

$$S^2 = \frac{\Sigma (X-\bar{X})^2}{n-1}$$

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54

Step 1: Find the Mean (or the average of the scores) – Refer to workings for Mean above (Section 1)

15, 3, 48, 23, 8, 18, 6, 19, 54

$\bar{x}$ = 21.56

Step 2: Find the Deviation of each of the scores from the Mean
(**$x - \bar{x}$**) (score minus mean)

15 - **21.56 = - 6.56**
3 - **21.56 = - 18.56**
48 - **21.56 = 26.44**
23 - **21.56 = 1.44**
8 - **21.56 = - 13.56**
18 - **21.56 = - 3.56**
6 - **21.56 = - 15.56**
19 - **21.56 = - 2.56**
54 - **21.56 = 32.44**

Step 3: Square the deviations found in Step 2. This step amplifies the positive numbers and changes the negative results to positive results $(\mathbf{x - \bar{x}})^2$

$-6.56^2 = 43.03$
$-18.56^2 = 344.47$
$26.44^2 = 699.07$
$1.44^2 = 2.07$
$-13.56^2 = 183.87$
$-3.56^2 = 12.67$
$-15.56^2 = 242.11$
$-2.56^2 = 6.55$
$32.44^2 = 1052.35$

Step 4: Sum the obtained squares (as a first step to obtaining an average) $\Sigma(\mathbf{x - \bar{x}})^2$

= 43.03 + 344.47 + 699.07 + 2.07 + 183.87 + 12.67 + 242.11 + 6.55 + 1052.35

= 2586.22

Step 5: To find the Variance, divide the Sum the obtained squares by the number of scores $\Sigma(\mathbf{x - \bar{x}})^2$ **/ n**

$$S^2 = \frac{\Sigma(X - \bar{X})^2}{n-1}$$

= 2586.22 / 9

= 287.36

Variance = 287.36

OR SIMPLER

Standard Deviation Squared

σ^2

$= 16.95^2$

= 287.36

Variance = 287.36

- Variance Calculations
    - Answers

Q1 287.36
Q2 595.69
Q3 8.25
Q4 2.92

4. The Z-Score

All the above begs the question: How does one calculate where a particular value falls within a standard deviation? Does a 30-year old male fall within 1 standard deviation (that is within 68% percent of the population) or within the other deviations?

This is carried out using the Z-Score test. The test calculates the position where a number on the x axis resides in terms of the standard deviation. In summary, the z-score defines the distance the sample value is from the mean – always in terms of standard deviations.

The larger the values, the further away from the Mean that value resides and into the higher standard deviations. If a Z score is negative, then the value (X) is below the mean. If it is positive, X is above the mean.

The Formula

$$z = \frac{x - \overline{x}}{s}$$

The Z-Score is calculated as follows:

Z = (a given value - mean) / standard deviation

For example, for a young population that is normally distributed with a mean(μ) of 20 and a standard deviation (σ) of 5, you want to find out the Z score for a value of 30 (*x*). This value (X = 30) is 10 units above the mean, with a Z value of:

Z = (30 - 20)/(5) = (10)/(5) = +2

The point is within 2 Standard Deviations of the Mean

This section contains a few working examples. Kindly attempt to analyse the figures based on the process given above. Use the following example as a guide:

- Z-Score
    - Calculate the Z-Score for the following number in brackets: **15, 3, 48, 23, 8, 18, 6, 19, 54 (23)**

i) Firstly, calculate the Mean (μ) of 15, 3, 48, 23, 8, 18, 6, 19, 54. This results in a mean of 21.6
ii) Secondly, calculate the standard deviation(σ) as per relative guide above. This gives a standard deviation of 17.0
iii) Thirdly, deduct the mean from your value - 23 (*x*)and divide by the standard deviation.
iv) The result is that of 0.1 which fall within 1 standard deviation (between 0 and 1 and is a positive number) - (+0.1)

- Z-Score Calculations

- Calculate the Z-Score for the following:

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54 (23)
Q2 16.5, 18, 63.2, 88.88, 19, 20, 21, 22, 18.5 (80)
Q3 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 (2)
Q4 1, 2, 3, 4, 5, 6 (5.5)

Worked Example: Q1 – Z- Score

Formula:

$$Z = x - \mu / \sigma$$

Therefore:

$$\frac{23 - 21.6}{17.0} = 0.1$$

Note that 0.1 fall within the 1st Standard Deviation (less that 1), thus the results can be said to be within 1 standard deviation (+0.1).

- Z-Score Calculations
    - Answers

Q1 within 1 standard deviation (+0.1)
Q2 within 2 standard deviations (+2)
Q3 over -1 standard deviation (-1.2)
Q4 over 1 standard deviation (+1.2)

Statistical Tests

This section will outline some statistical tests that are used to help researchers understand their data as well as carry out comparative studies. The tests include the F-test, the T-tests, Regression Analysis, ANOVA and Chi Squared. The first two employ both the mean and the standard deviation in order to test if two sets of normally distributed data are similar or otherwise. If they are similar then they can be attributed to the same population. Regression analyzes how the independent variables are related to the dependent variable. The ANOVA is used where more than one factor can exert an influence on a value. Chi-Squared is used to check whether a categorical sample represents the population.

The next steps describe each of the tests in summary. Refer to one of the indicated statistical books for a full description and step by step process of how to employ these tests. If one is referring to online tutorial, ensure that they are produced by reliable sources such as Stat Trek[1].

1. F-Test

One of the first inferential tests that one can use in order to establish whether the variances between two populations are equal is the F-Test. This test compares the ratio of the two variances which, if equal, should result in a value of 1.

2. T-tests

The T-tests are employed for testing standard deviations when the population is normally distributed. It is a random interval or ratio sample, where the standard deviation is computed from the sample data.

There are different tests which are given names according to the type of similarity between the datasets.

[1] http://stattrek.com/

Independent datasets that have very similar Standard Deviations: employ the **Student's t-test**. This is applied for small datasets having N (number of data) less than 30 and where the F-test shows that they are similar.

Independent datasets that have significantly differing Standard Deviations: employ the **Cochran t-test.** This is applied for small datasets having N (number of data) less than 30 and where the F-test shows that they are dissimilar.

Highly Dependent datasets employ the **paired t-test**. This can be employed when the same samples are used for two different tests.

3. Regression Analysis

Normally described as the Line of Best Fit, regression is used to establish the existence of a linear relationship.

Regression analysis assumes that a change in dataset X brings about a definite change in dataset Y. In correlation analysis, a change in X brings about a change in Y, which could be anything from an increase to a decrease or even no change at all. Regression is not so 'easy' on the relationship as a change in X must bring about a change in Y.

4. ANOVA

The Analysis of Variance, also known as the ANOVA, determines the existence of differences in datasets that contain two or more sample means. A two-way ANOVA is tested for when two independent variables are chosen.

5. Chi Squared

Chi Squared ($\chi 2$) is a critical test that investigates, looking for the frequencies of category (Nominal) presence in a sample and analyzes whether they represent the predicted frequencies in the total population.

This short summary does not do justice to the beauty that is statistical testing and readers are encouraged to acquire a statistics book that pertains to their particular theme as listed in the Thematic Publication Table.

Spatial Statistics

This section outlines some specialised spatial statistical tests that are used to help researchers understand their data both in normal statistics and in a higher-level mode where statistics are also depicted in visual modes.

Spatial statistical books need to be referenced for a full description and step by step process of how to employ these tests. There are different types of spatial statistics, best clustered in four-groups: Spatial distribution, Distance statistics, 'Hot spot' analysis routines and Interpolation statistics:

1. Spatial Distribution

Spatial distribution refers to the spread of values around a spatial mean. These include the mean centre, centre of minimum distance, standard deviational ellipse and Moran's I spatial autocorrelation index, or angular mean.

2. Distance Statistics

Distance statistics calculate the values based on proximity and statistical tests include the nearest neighbour analysis, linear nearest neighbour analysis, and Ripley's K statistic.

3. Hotspot Analysis Routines

'Hot spot' analysis routines are some of the most interesting and used spatial statistics as they depict data based on the concentration of values are a spatial location. Tests include the hierarchical nearest neighbour clustering, K-means clustering and local Moran statistics. There are alternative measures of hotspot analysis such as Kernel Density Estimate, Getis-Ord GI* and also multiple regression as an alternative to bi-variate analysis.

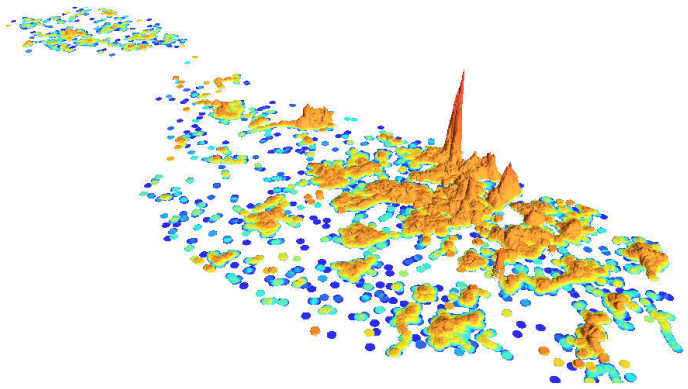
4. Interpolation Statistics

Interpolation statistics use the primary data and interpolate it to predict the probable value of areas within a boundary relative the location of the primary data. These tests include single-variable kernel density estimation resulting in an output of incident density (e.g. burglaries) and a dual-variable kernel density estimation that compares once variable to another baseline variable (e.g., burglaries analysed in relation to dwelling units in the area).

Of all the above, one of the best methods of analysing behavioural patterns is to use clustering methodology. Formosa (2007, Pg149-152) describes that due to the large number of behavioural patterns (such as crime, recreation) occurring in a particular area, analysis may concentrate on the aggregation of these data into specific areas rather than spread them all over the town/city. Clustering helps in identifying areas that are hotspots for specific behaviour types.

ii) Another method that can be employed is the Nearest Neighbour Analysis (NNA) which helps to aggregate data based on the proximity of a crime to the nearest location of another crime (Craglia *et al*, 2000). If an activity occurs within a specific parameter of say '20m' from that being analysed, then these two activities are aggregated, before searching for other activities within the next specific boundary[2]. Once there are no activities left within the recurrent buffers then the hotspot intensity dies out and stops. Where a large number of activities occur in a small area the hotspot is very pronounced and cluster densities can be calculated. Figure 11.1 depicts an example of such an NNA interpolation based on non-serious offences in Malta between 1998-2003 transposed in 3D (Formosa 2007, 150). The shape of the Maltese Islands is easily discernable, particularly the conurbation area. High offence counts are depicted as with red peaks in the main leisure and recreation areas and very few if anything in the rural and rural-urban boundary areas (blue and white respectively). The same methodology can be used to elicit statistical results as well as for visualization purposes.

Figure 11.1: Interpolation of Non-Serious Offences – 1998-2003



Source: Formosa, 2007, Pg 150

[2] Note that variance in the boundary width can produce different results.

Note that each of these methods necessitates knowledge of the limitations in using that specific method which limitations are dependent on a number of factors. These include the sample size taken, the number of minimal points set as the threshold for identifying the least hotspot size, amongst others. The limitations of the methodologies used such as the Nearest Neighbour Hierarchical Analysis Method (NNH) include differing hotspot locations for different spatial aggregations employed, such as a minimal 25-point hotspot cut-off, which signifies where an ellipse boundary should be drawn once no more points falling within those thresholds are encountered. Consistency in the results is ensured as the analysis in this study employ the same threshold limits. Another limitation relates to the issue of cross-comparison of two data-layers that may have widely-differing counts, such as a 10,000 point offender data layer and a 1,000 point poverty layer. Using the same standard-deviation levels and thresholds, error generation can be reduced to a minimum.

In addition, NNH as well-as K-Means employed in the study show their results through ellipsoids, which in effect can cover areas that may not be prone to high incidences being investigated but still fall within the ellipsoid since such a tool cannot eliminate areas within its boundary without compromising the ellipsoid integrity. Also, some ellipsoids might show areas that have high concentrations of incidences when the base data might show few data points, which result is mainly due to a multiplicity of overlapping points found within the base data layer and weighted for in the ellipsoid. Knowledge of the base data layer is required in order to interpret the results of such methodologies.

In summary statistical tests are many and varied, they cover simple descriptive statistics to inferential statistical tests to spatial statistics. This chapter sought to introduce the new initiate to the idea that, whilst appearing 'scary', statistical tools are easy to understand. However, one really needs to consult specialist books in the field which employ theme-specific examples in order to understand how each of the tests is carried out. The initial section outlining basic statistics gave a walkthrough with examples of how to carry out the calculation required, whilst the other two sections gave a summary of the types of statistical tests that exist for inferential statistics and for spatial statistics respectively.

Questions (refer to Appendix for the answers)

1. Why is statistical testing important?
2. Briefly explain what descriptive statistics are, providing examples.
3. Briefly explain what inferential statistics are, providing examples.
4. What do you understand by independent variables?
5. Why are dependent variables also known as criterion variables?
6. List the three measures of central tendency.
7. There are two types of Mode. Name them.
8. How would you define "the range" in statistics?
9. Briefly describe standard deviation, explaining its function in statistics.
10. What is the variance (in statistics)?
11. What does the Z-score do?
12. Mention five statistical tests and very briefly describe each of them.