MAKING SENSE OF DATA

Essentials series

PREREQUISITES

Prerequisites
Descriptive statistics
Charts and graphs
The normal distribution
Surveys and sampling
Correlation and regression

This "essentials" series provides guidance on the **essential concepts and techniques used in statistics** for the analyses of administrative or similar numerical data.

Notes used

A few notes and warning are used which are explained here:

WARNING! More advanced formulas or equations are being used. Some readers may choose to skip these sections.

IMPORTANT! Vital information which must be followed to prevent incorrect or inaccurate results.

TIP: Tips for completing tasks.

NOTE: Information for special situations.

What is statistics?

Statistics is the study of the collection, organization, analysis, interpretation, and presentation of data. It uses the scientific method to answer question and make conclusions. This typically involves designing studies, collecting good quality data, describing the data with numbers and graphs, analysing the data, and then making conclusions.

However, this guidance series focuses on the use of administrative or similar data; such as the 2011 Census. As such "good quality data" has already been collected through surveys or other administrative processes and what remains is to describe the data, analyse the data, and make conclusions.

Describing data

Data can be summarised which helps in getting a handle on the big picture, often in conjunction with charts, graphs or maps. *Descriptive statistics* are numbers that describe a data set in terms of its important features.

Numerical data represent measurements or counts, where the actual numbers have meaning (such as height and weight). With numerical data, more features can be summarised besides the number or percentage in each group. Some of these features include measures of center (in other words, where is the "middle" of the data?); measures of spread (how diverse or how concentrated are the data around the center?); and, if appropriate, numbers that measure the relationship between two variables (such as height and weight).

Analysing data

After the data have been described using pictures and numbers, then comes the fun part: navigating through that black box called the *statistical analysis*. If the study has been designed properly, the original questions can be answered using the appropriate analysis, the operative word here being *appropriate*. Many types of analyses exist; choosing the wrong one will lead to wrong results.

Making conclusions

Researchers perform analysis with computers, using formulas. But neither a computer nor a formula knows whether it's being used properly, and they don't warn you when your results are incorrect. At the end of the day, computers and formulas can't tell you what the results mean. One of the most common mistakes made in conclusions is to overstate the results, or to generalise the results to a larger group than was actually represented by the study.

Statistics is about much more than numbers. It's important to understand how to make appropriate conclusions from studying data.

The remainder of this topic reviews basic *operations and symbols* used in mathematics, *probability* and *linear equations*.

Types of number

Integers

…, -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, …

Rational (Fractions)

A rational number is formed by dividing one integer into another.

e.g. $\frac{1}{2}, \ -\frac{3}{4}, \ \frac{14}{57}, \ \frac{1073}{31}, \ -\frac{73}{2}$

Rational numbers are of the form $\frac{p}{q}$ where p and q are integers and $q \neq 0$ (not equal to). p is called the numerator, q is called the denominator.

Irrational

Irrational numbers consists of all numbers which cannot be expressed as a rational number.

For example $\sqrt{2} = 1.414213562373\ldots$ and does not terminate or have any repeating pattern. It is not possible to write $\sqrt{2}$ in the form $\frac{p}{q}$.

The set of all integer, rational and irrational numbers is called the set of all *real numbers* represented by $\mathbb{R}$.

Miscellaneous symbols

$=$	is equal to
$\neq$	is not equal to
$\approx$	is approximately equal to
$<$	is less than
$\leq$	is less than or equal to, is not greater than
$>$	is greater than
$\geq$	is greater than or equal to, is not less than
∞	infinity

Rules of arithmetic

These are the laws applied to real numbers which govern the operations of addition, subtraction, multiplication and division.

In what follows a, b, c, etc. represent real numbers.

Commutative

a) $a + b = b + a$
b) $a\,b = b\,a$ where $a\,b$ means $a \times b$

Associative

a) $(a + b) + c = a + (b + c)$
b) $(a\,b)\,c = a\,(b\,c)$

The brackets indicate the order in which the arithmetic operations are to be carried out. For example $(3 + 5) + 7 = 8 + 7 = 15$ and $3 + (5 + 7) = 3 + 12 = 15$.

Distributive

a) $a(b + c) = ab + ac$
b) $(a + b)(c + d) = ac + ad + bc + bd$

E.g. $4 \times (5 + 3) = 4 \times 8 = 32$ and
$4 \times (5 + 3) = 4 \times 5 + 4 \times 3 = 20 + 12 = 32$

Laws of signs

a) $a + (-b) = a - b$
b) $a - (-b) = a + b$
c) $(-a) \times (+b) = (+a) \times (-b) = -ab$
d) $(-a) \times (-b) = ab$

e) $\frac{-a}{-b} = \frac{a}{b}$

f) $\frac{a}{-b} = \frac{-a}{b} = -\frac{a}{b}$

Laws of rational numbers

a) $\frac{a}{b} \times \frac{c}{d} = \frac{ac}{bd}$

b) $\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd}$

c) $\frac{a}{b} \Big/ \frac{c}{d} = \frac{a}{b} \times \frac{d}{c} = \frac{ad}{bc}$

Powers and roots

Powers for positive integers are defined by

$$x^n = x \times x \times x \ldots \times x \quad (n \text{ factors of } x)$$

For example $2^5 = 2 \times 2 \times 2 \times 2 \times 2 = 32$

Laws of powers

a) $x^1 = x$ and $\frac{x^m}{x^n} = x^{m-n}$

b) $x^0 = 1$ and $x^{-n} = \frac{1}{x^n}$

c) $x^n \times x^m = x^{n+m}$ and $(xy)^n = x^n y^n$

d) $(x^n)^m = x^{nm}$ and $\left(\frac{x}{y}\right)^n = \frac{x^n}{y^n}$

Roots are the inverse of powers. An nth root "undoes" raising a number to the nth power, and vice-versa.

$$\sqrt[n]{x} = y \; if \; and \; only \; if \; y^n = x$$

E.g. $\sqrt[3]{64} = 4 \; because \; 4^3 = 64$

Summation

The mathematical symbol used to represent the **summation** of many similar terms is the capital Greek letter Sigma Σ and is defined as:

$$\sum_{i=m}^{n} x_i = x_m + x_{m+1} + x_{m+2} + \cdots + x_{n-1} + x_n$$

Where

i	index of summation
x_i	indexed variable
m	lower bound of summation
n	upper bound of summation

An example:

$$\sum_{i=3}^{6} i^2 = 3^2 + 4^2 + 5^2 + 6^2 = 86$$

Factorial

The **factorial** of a non-negative integer n (denoted by $n!$) is the product of all positive integers less than or equal to n.

Example:

$$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$$

The factorial function is formally defined by

$$n! = \prod_{k=1}^{n} k$$

with the empty product 0! = 1

Permutations and combinations

A **permutation** of a set of distinct objects is an *ordered* arrangement of these objects.

E.g. All permutations of 1, 2, 3

123, 132, 213, 231, 312, 321

Given n objects there are $n!$ different permutations.

An ordered arrangement of r elements of a n-set is called an **r-permutation**.

$$P(n,r) = {}^{n!}/_{(n-r)!}$$

E.g. All 3-permutations of 4, 5, 6, 7, 8, 9

456, 457, 458, 459, 546, 547, … , 986, 987

$r = 3$ and $n = 6$ giving

$$P(n,r) = \frac{6!}{(6-3)!} = \frac{720}{6} = 120$$

An **r-combination** of an n-set is an *unordered* selection of r elements from the set.

where $0 \leq r \leq n$

$$C(n,r) = {}^{n!}/_{r!\,(n-r)!}$$

E.g. How many ways can we select two items from the set {a, b, c, d}?

$$C(4,2) = \frac{4!}{2!\,(4-2)!} = \frac{24}{4} = 6$$

Elementary probability

Random experiments

A random experiment is one for which the outcome is uncertain. For example throwing a die with 6 faces several times or drawing multiple cards from a pack of 52.

Experiment A: Throwing a single coin

Consider the experiment of throwing a coin which can land heads up (H) or tails up (T). We list the outcomes as a set $\{H, T\}$ – the order is unimportant. On any particular throw of a coin heads or tails are equally likely to occur; we say for a fair coin H and T are *equally-likely outcomes*.

Experiment B: Throwing two coins

Suppose two coins are thrown and we note how each lands. If we write HT to indicate that the first coin shows heads and the second shows tails and so on then the 4 equally-likely outcomes of the experiment are $\{TT, TH, HT, HH\}$. This set is called the *sample space S* of the experiment, containing all possible outcomes.

Experiment C: Throwing a single coin many times

Suppose we throw a coin 10 times and obtain six heads and four tails; does this suggest that the coin is biased? What about the case when we obtain 9 heads and 1 tail? We conduct an experiment in which a coin is thrown repeatedly and the result recorded as 1 if a head appeared face up and 0 if a tail appears. Figure 1 shows the plot of the average score $\frac{r}{n}$, where r is the number of heads and n is the total number of throws, against n for $n = 1, 2, \ldots, 100$. The quantity $\frac{r}{n}$ is called the *relative frequency* of heads.

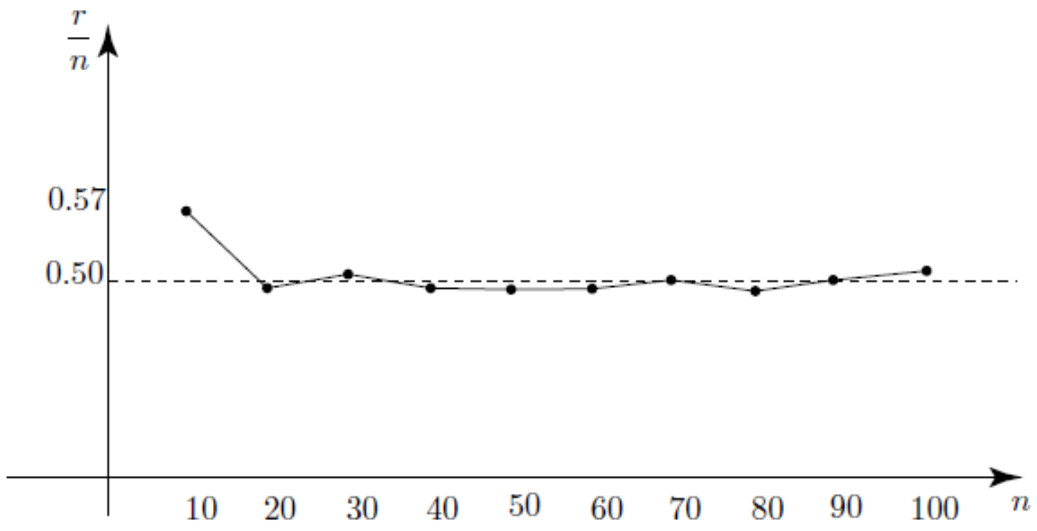

Figure 1

As n increases the relative frequency settles down near the value $\frac{1}{2}$. This is an experimental estimate of the likelihood of throwing a head with a particular coin. The problem is that when $n = 50$ this estimate is 0.49 and when $n = 100$ is 0.46. Hence we use of the word *estimate*.

Theoretically, the likelihood of obtaining a head when a fair coin is thrown is $\frac{1}{2}$. Experimentally, we *expect* the relative frequency to approach $\frac{1}{2}$ as n increases.

Events

A collection of some or all of the outcomes of an experiment is called an *event*. An event is a subset of the sample space S. For example, if a die is thrown then S is {1, 2, 3, 4, 5, 6} and two possible events are:

a) A score of 3 or more, with the set {3, 4, 5, 6}

b) A score which is even, with the set {2, 4, 6}

The *complement* of an event is the set of outcomes which are <u>not</u> members of the event. For example, the complement of the event "score 3 or more is obtained" is the set {1, 2}.

If a capital letter A represents an event, the complementary event is denoted A'.

Probability

The experiment "throwing a six-faced die" has six equally likely outcomes; either a one, a two, … , or a six will appear. For this type of experiment we can write the probability of an event A occurring as $P(A)$ and define it as

$$P(A) = \frac{number\ of\ outcomes\ in\ A}{total\ number\ of\ outcomes}$$

It follows that $0 \leq P(A) \leq 1$

- If $P(A) = 1$ the event is *certain*
- If $P(A) = 0$ the event is *impossible*

The set with no outcomes in it is called the *empty set* and written θ; therefore $P(\theta) = 0$.

The probabilities of an event A and its complement are related. The probability of the event A' is found using this identity known as the *complement rule*

$$P(A') = 1 - P(A)$$

Linear equations and graphs

The linear equation of a straight line is given as

$$y = mx + b$$

Where

m = gradient
b = intercept on the y-axis

E.g. consider the straight line with equation $y = 2x + 1$

x	y
0	1
1	3
2	5

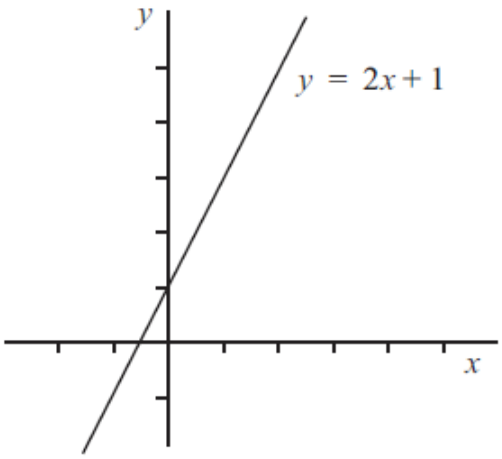

Figure 2

This line cuts the y-axis at $y = 1$, when $x = 0$

The gradient m is calculated by

$$m = \frac{\Delta y}{\Delta x}$$

Where

Δy = the change in y
Δx = the change in x