

Introduction to Statistics

Section 1: Summarising data

INTRODUCTION

Statistics is that branch of mathematics which studies natural variation.

Adds objectivity to data analysis (deal with variation in scientist's interpretation)

Variation in:

data

scientist's interpretation of that data.

POPULATIONS AND SAMPLES

In collecting data we wish to describe something about the world. What we measure is the character or variable.

Sample = collection of individual observations selected by a specified procedure.

Population = the totality of individual observations about which inferences are to be made, existing anywhere in the world, or at least within a definitely specified sampling area limited in space and time.

Population = mice

Sample = set of mice in laboratory

Population = Regensburg University students

Sample = students in this room

This sample may or may not be representative depending upon focal question. For instance, it is likely that blood groups from this sample may represent Regensburg students as a whole but possibly sex ratio does not.

Sample must give a good representation if we are to make inferences!

VARIABLE TYPES

A quantitative (or measurement) variable takes numerical values for which arithmetic operations such as differences and averages makes sense. Quantitative variables may be continuous or discrete (discontinuous). Continuous refers to a variable in which fractions of a unit make sense, e.g. fraction of a second, fraction of a centimetre. Discrete variables make only take whole numbers (integers), e.g. number of legs, and do not vary continuously.

Examples:

continuous: size, weight, distance

discrete: number of offspring, number of hairs, number of queens in ant colony

Example:

number of male rats in litter in five:

possible set of values: 0,1,2,3,4,5

sample: 2,3,3,2,0,3,2,3,2,1,2,3,0,5,1,4,2,4,1,3,2,2,3,4,4

A categorical variable simply records into which of several categories a person or thing falls into, e.g. sex, alleles, colour. Also known as attributes.

Examples:

1) possible set of values: Male, Female
sample: M, M, F, F, M, F, M, F, F, M, F

2) possible set of values: Aa, AA, aa, aA
sample: AA, aa, Aa, AA, Aa, Aa, aA, aA

Also ranked variables, e.g. *order* in which pupae eclose: 1st, 2nd, 3rd, 4th, etc.

These variables operate on one of four scales:

nominal	(category, e.g. male versus female; red versus white)
ordinal	(rank, e.g. low medium and high)
interval	(e.g. temperature, date)
ratio	(size, weight etc.)

DISPLAYING DATA

Displaying the raw data is useful when there are relatively few observations but impractical when the data set is large. We want a way to summarise the main characteristics of the data to others without them having to delve into the raw data: convey the distribution. The pattern of variation of a variable is called its distribution.

stem-and-leaf plots

Offers a quick way to picture the shape of a distribution while including the actual numerical values. This method is only really useful for small samples of discrete data. Separate data into stems (e.g. first digit) and leaves (second digit). E.g. the number of home runs hit by Babe Ruth in each of his 15 years with the New York Yankees (1929–1934) was 54, 59, 35, 41, 46, 25, 47, 60, 54, 46, 49, 46, 41, 34, 32.

This gives the following stem and leaf plot

2		25
3		45
4		1166679
5		449
6		0

These plots are useful in giving an overview of the shape of the data and comparing two sets of data:

	Player 1		Player 2
2	00112234489	2	25
3	124456	3	45
4	2579	4	1166679
5	048	5	449
6	1	6	0

Frequency table

When there are multiple occurrences of the same values then a frequency table may be useful.

Example:

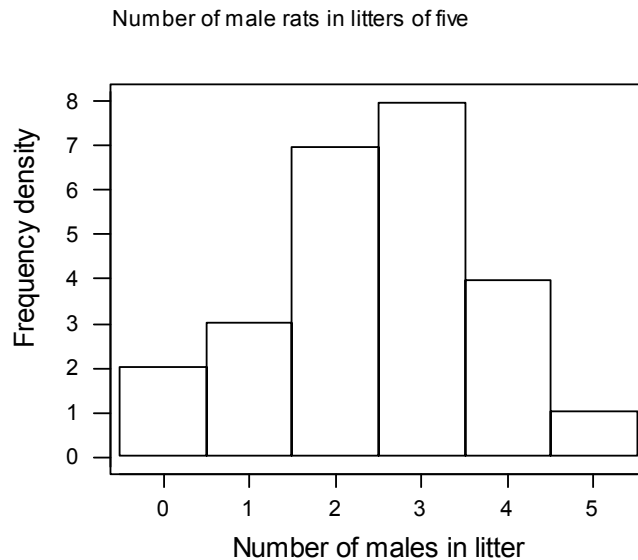
Number of male rats in litter in five. Possible set of values: 0,1,2,3,4

Sample: 2,3,3,2,0,3,2,3,2,1,2,3,0,5,1,4,2,4,1,3,2,2,3,4,4

Number of male rats	0	1	2	3	4	5	Total
Frequency	2	3	8	7	4	1	25

Histogram

Often a graphical interpretation of a frequency table is more useful as one can clearly see how the frequency is related to the variable of interest. In a histogram the frequencies are represented by the areas of the bars and the height is known as frequency density (= frequency / width of bar). When the interval is 1, as above, then frequency equals frequency density.



Drawing graphs. All graphs need the following: x-axis label, y-axis label, x-axis values, y-axis values, and title. When there is more than one series then these should be clearly distinguished, usually with a legend.

Histograms can be used with continuous data as above if they happen to be integers, e.g. time in seconds. Also, we can produce a histogram with non-integer data by assigning the observations to different intervals (or bins).

Tip: How many intervals? There should be enough intervals to show the pattern but not so many that the frequencies are too low. A good rule is the number of intervals should be about the square root of n . Thus, for 100 observations, 10 intervals is probably useful; for 50 observations, 7 intervals is probably appropriate.

So for the following data:

35,6 39,8 40,6 43,5 43,6 44,5 44,7 44,9 45,0 45,6 ... 68,9 73,3

We might assign the data thus:

Interval	35–	39–	43–	47–	51–	55–	59–	63–	67–	71–	Total
	38	42	46	50	54	58	62	66	70	74	
Frequency	1	2	10	6	4	2	1	2	1	1	30

But, there are too many intervals with low numbers. As $n = 30$, use six intervals of six units each:

Interval	38–43	44–49	50–55	56–61	62–67	68–73	Total
Frequency	6	13	4	3	2	2	30

Convention = round up: intervals are 37.5–43.5; 43.5–49.5 etc.

The intervals need not be the same width. The following data are the survival times of greenfly in a study of the effects of a pesticide:

Survival time (s)	0–20	20–40	40–60	60–100	100–140	140–240	Total
Number of insects	10	51	32	19	9	4	30

Because most of the insects died in the first sixty seconds it was not useful to continue to record the data so accurately for the whole observation period. Here the intervals are not of the same width and bars in the histogram will be of different widths too.

Let one unit width equal 20 seconds. The width of the intervals are 1,1,1,2,2, and 5 units.

Survival time (s)	0–20	20–40	40–60	60–100	100–140	140–240	Total
Number of insects	10	51	32	19	9	4	30
Interval width / units	1	1	1	2	2	5	
frequency density	10	51	32	$19/2 = 9,5$	$9/2 = 4,5$	$4/5 = 0,8$	

What is most important here is not the *absolute* values on the y axis but the *relative* heights of the bars when corrected for width.

MEAN

The mean is way of estimating the centre (location) of a distribution. The mean is also known as the (arithmetic) average, or the “expected” value and is usually denoted as \bar{x} . The mean is affected by extreme values, i.e. is sensitive to outliers. That is, the mean can be increase by an increased in one or a few extreme observations thus not reflecting the true centre of distribution. In a skewed distribution the tail will pull the mean towards its long tail, e.g. exponential. Because the mean cannot resist the influence of extreme observations, we say that it is not a *resistant measure*. A measure that is resistant does more than limit the influence of outliers; its value does not respond strongly to changes in a few observations, no matter how large those changes may be. Rarely does the mean work out as an integer, i.e. whole number. The mean may be used on both continuous and discrete data. With discrete data the mean often results in a number that is only sensible in a mathematical sense, e.g. British couples have an average of 2,1 children. Clearly, couples do not have two whole children plus an extra 0,1 child.

Convention:

Upper case letters, such as X, represent variables from some theoretical distribution

Lowercase letters such as x stand for particular numerical values of the variable

e.g. $X \sim N(\mu, \sigma^2)$; $P(x < 10)$

ARITHMETIC MEAN

mean = $\frac{\text{sum of all observations}}{\text{number of observations}}$

If n observations are denoted by x_1, x_2, \dots, x_n , their mean is

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

or in more compact notation

$$\bar{x} = \frac{1}{n} \sum x_i$$

Example:

Number of male rats in litters of five: 2,3,3,2,0,3,2,3,2,1,2,3,0,5,1,4,2,4,1,3,2,2,3,4,4

Sum of observations = $2+3+3+2+0+3+2+3+2+1+2+3+0+5+1+4+2+4+1+3+2+2+3+4+4 = 61$

Number of observations = 25

Mean = $61 / 25 = 2.44$

Tip: How many decimal places should the mean be quoted to?

In general, quote the mean to one more decimal place than the original data was measured. E.g. if you measure in cm then the mean can be quoted to a tenth of a cm. The above example breaks this rule because the mean happens to be exact to two decimal places.

Calculating the mean from a frequency table

This is most easily and accurately achieved with discrete data:

Number of male rats (x_i)	0	1	2	3	4	5	Total
Frequency (f_i)	2	3	8	7	4	1	25

If x_i is the number of males and f_i is the associated frequency then

mean = $\frac{\text{sum of } x_i f_i}{\text{number of observations (= sum of } f_i)}$

ie

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

For the data above:

$$\text{mean} = \frac{(0 \times 2) + (1 \times 3) + (2 \times 8) + (3 \times 7) + (4 \times 4) + (5 \times 1)}{25} = 2.44$$

This method can also be used to calculate the mean for continuous data once it has been collated into a frequency table. However, there is a loss of information during the data collation and so the mean from this method will always be less accurate than calculating the mean from the original data—i.e. we do not know how the original data are placed within an interval.

Interval	38–43	44–49	50–55	56–61	62–67	68–73	Total
Midpoint (x_i)	40.5	46.5	52.5	58.5	64.5	70.5	
Frequency(f_i)	6	13	4	3	2	2	30

Thus, the mean of this data is:

$$\frac{40.5 \times 6 + 46.5 \times 13 + 52.5 \times 4 + 58.5 \times 3 + 64.5 \times 2 + 70.5 \times 2}{30} = \frac{1503}{30} = 50.1$$

(From the raw data, the true mean is $1507 / 30 = 50.23$)

MEDIAN and QUARTILES

Arrange the data in increasing order of size, i.e. rank the data. The median is the middle number. For example, the number of micro-organisms in unit volumes of water from a pond are:

10, 16, 12, 5, 22, 14, 19

Arranged in order these are: 5, 10, 12, **14**, 16, 19, 22

The median is the middle one in this ranked order, i.e. 14. This value has three smaller values (5, 10, 12) and three larger values (16, 19, 22) and so is the middle of the data and is a measure of *location* or a measure of the central tendency of the data. If there are n observations then the median is the $(n+1)/2$ 'th observation, i.e. if $n = 7$, as above, then the median is the 4th observation in the ranked list.

What if there are an odd number of observations? In this case there will be a middle pair so take the average of these two values. For instance, if there was an additional datum 38, then our new data is: 5, 10, 12, **14, 16**, 19, 22, 38, the middle pair is 14 and 16, the average of which is 15. (This is still the $(n+1)/2$ 'th observation, if $n = 8$, as above, the median is the 4.5th observation.)

Note that the addition of the extra and larger value did not greatly affect the median. Medians are not greatly affected by extreme values but chiefly affected by central values and not values at the end. For instance, had our data been:

1, 10, 12, 14, 16, 195, 256

the median is still 14.

The median (**M**) divides the data in two parts. Quartiles divide data into four parts. One quarter of all the observations have values less than or equal to the lower quartile (**q**), one quarter between **q** and **M**, one quarter between **M** and the upper quartile **Q** and one quarter equal to or greater than **Q**.

Thus, for our original data above: 5, 10, 12, 14, 16, 19, 22

M = 14
q = middle of 5, 10, 12 = 10
Q = middle of 16, 19, 22 = 19

With an even number of observations then we need to take average of middle pairs:

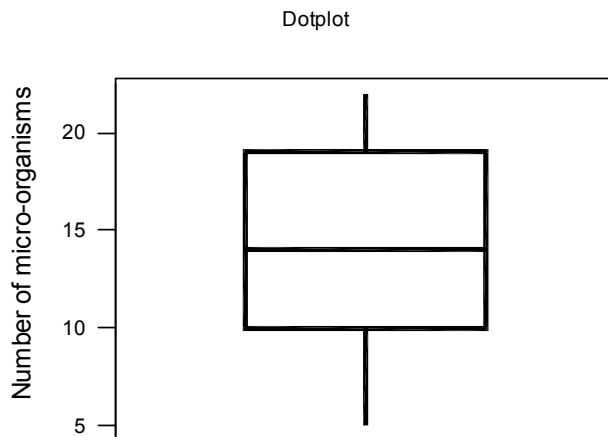
5, 10, 12, 14, 16, 19, 22, 38

M = average of middle pair 14, 16 = 15
 lower half of data = 5, 10, 12, 14. Middle pair = 10, 12 so **q** = 11
 upper half of data = 16, 19, 22, 38. Middle pair = 19, 22 so **Q** = 20,5.

The median and quartiles tell us whether data is symmetrical. If **M-q** is roughly equal to **Q-M** then the middle part of the data at least is fairly symmetrical.

The range from **q** to **Q** is called the interquartile range.

These values can then be used to draw a dotplot (also known as a box-and-whisker plot):
 This shows a box ranging from **q** to **Q** with line (or cross) at **M**. Lines extend to maximum and



minimum (sometimes shown as a star: *). Thus, the combination of minimum, **q**, **M**, **Q**, and maximum give a good overview of the distribution of data. This gives a good way to initially compare two data sets.

MAXIMUM

The maximum, although defined by extreme data, is important and useful in many situations. For instance, a nest site, daily food ration etc designed for the maximum will be sufficiently large for all the individuals. Consider designing nest boxes for blue tits. You measure the size of a sample of blue tits and need to decide what size entrance hole you should drill. The mean, an entrance hole designed for the typical “average” bird, will not suffice because any larger-than-average bird will be too big to fit in the hole, similarly for the median of the data. A hole large enough for the largest bird will be large enough for all the birds.

maximum river height useful for dyke construction

MODE

A third measure of central tendency is the mode. This is the value or interval in which the frequency is highest:

Number of male rats (x_i)	0	1	2	3	4	5	Total
Frequency (f_i)	2	3	8	7	4	1	25

For this data set, the mode is 2 because that gives the highest frequency (8). The mode is rarely used. Note that there may be more than one mode in a dataset, i.e. it will not necessarily give a unique value.

MEAN VERSUS MEDIANS

The mean is best used for symmetrical non-skewed data. Better for theoretical purposes such as significant tests. It is easily calculated but is affected by extreme values. However, outliers should not be discarded without good reason. The median is less affected by extreme values and so is better for skewed data.

In a perfectly symmetrical distribution the mean, mode and median coincide.

In a skewed distribution the mean is nearest the long tail, the mode is the farthest and the median is in between.

GEOMETRIC MEAN

The arithmetic mean gives equal weighting to each point in the set of observations, i.e. each datum contributes equally to the final value. This is fine when the distribution of points is symmetric but not when the data is skewed. With skewed data, such as has a lognormal or exponential distribution, then a different type of mean, the geometric mean, is more appropriate:

GEOMETRIC MEAN

Geometric mean = n^{th} root of the product of n observations.

$$\bar{x}_g = \sqrt[n]{\prod x_i} = \exp\left(\frac{\sum \ln x_i}{n}\right) = \exp\left(\text{mean}\left(\frac{1}{x}\right)\right)$$

[Q; what is problem of first formulation of this mean, i.e product. Answer=overflow]

(Note that if a single value, x_i , is zero then the mean is zero. Note also that this calculation cannot be performed with negative values.)

It is often used when dealing with size and growth rates of an economy or population, i.e. a situation in which values are multiplicative (think compound interest), i.e. ratios.

Consider a population that doubles in size in one year and then trebles the next year. That is, it shows a growth rate of 2 in year 1 and growth rate of 3 in year 2. What is the average growth rate? Although it is tempting to say 2,5, the arithmetic mean, this is not correct. This is because a population that grows 2,5 times per year for two years shows $2,5 * 2,5 (= 6,25)$ times total growth. This is more than the $2*3$ times growth observed. The correct answer is the geometric mean of 2 and 3 = square root of 6 = 2,45.

Example 1: gross domestic product of UK economy

Year	Gross domestic product (£s, billions)	Indices 1979=100	% growth rate on previous year
1979	196.706	100	—
1980	230.603	117,2	17,2
1981	254.103	129,2	10,2
1982	276.409	140,5	8,8
1983	300.973	153,0	8,9
1984	320.120	162,7	6,4

The arithmetic mean growth rate is 10,29%. This estimate overstates the growth rate of the economy and will give ever more and more biased results with the longer length of calculation made.

The geometric mean growth rate is 10,23% (= 5th root of 17,2 × 10,2 × 8,8 × 8,9 × 6,4) which matches data better.

Year	Calculation	Result
1979	196.706	—
1980	196,706*1,1023	216,829
1981	216,829*1,1023	239,010
1982	239,010*1,1023	263,460
1983	263,460*1,1023	290,411
1984	290,411*1,1023	320,120

Example 2: Bird population index from daily bird migration counts at Alaskan bird observatory during autumn migration 1992–1997.

Species	1992	1993	1994	1995	1996	1997	Geomean
Alder flycatcher	2,60	3,26	3,03	4,92	3,39	4,27	3,44
Hammond's flycatcher	1,42	1,47	2,00	2,21	3,35	2,54	2,09
Black-capped chickadee	2,18	1,58	1,53	3,39	1,79	1,84	2,10

3,44 = 6th root of 2,6 × 3,26 × 3,03 × 4,92 × 3,39 × 4,27

HARMONIC MEAN

Whereas the geometric mean is useful for skewed data, the harmonic mean is more useful for highly skewed data such as rates and prices.

HARMONIC MEAN

The harmonic mean is the reciprocal of the arithmetic mean of reciprocals.

$$\bar{x}_h = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}} = \frac{1}{\text{mean}\left(\frac{1}{x}\right)}$$

Example:

Consider an ant that runs at 2 cm/s from the nest to a food source and 3 cm/s from the food source back to the nest. What is its average speed? The arithmetic mean is not correct because the two

trips are not equally weighted in terms of time. The faster the ant runs on the return trip, the smaller the proportion of the whole trip time is taken on that leg of the trip.

1) average velocity (v) = total distance (d) / total time (t)

2) $d = v_1 t_1 + v_2 t_2$

3) $t = t_1 + t_2$

give:

$$v = \frac{v_1 t_1 + v_2 t_2}{t_1 + t_2}$$

Considering outward journey: $v_1 = d/2 / t_1 \Rightarrow t_1 = d / 2v_1$

Considering outward journey: $v_2 = d/2 / t_2 \Rightarrow t_2 = d / 2v_2$

Substitute into equation for v and rearrange:

$$v = 2 \frac{1}{\frac{1}{v_1} + \frac{1}{v_2}} = 2 \frac{1}{\frac{1}{2} + \frac{1}{3}} = \frac{12}{5} = 2,4 \text{ cm / s}$$

Unless the values do not vary (i.e. $x_i = \text{constant}$), arithmetic mean > geometric mean > harmonic mean.

STANDARD DEVIATION and VARIANCE

In most cases one uses the arithmetic mean. Whereas interquartile range measures dispersion or spread of points and is associated with the median, standard deviation and variance also measure spread of points and are associated with the arithmetic mean. Basically, it would be wrong to quote median \pm s.d, or mean + q and Q.

To measure the spread of points we can consider the set of deviations from the mean, i.e. $x_i - \bar{x}$. The greater the spread of points the greater the magnitude of the individual deviations. However,

$$\sum (x_i - \bar{x}) = 0$$

One way around this is to square the deviations so they are all positive:

$$\sum (x_i - \bar{x})^2$$

and if we are interested in average deviation:

$$\frac{\sum (x_i - \bar{x})^2}{n}$$

However for a sample we should divided by $n-1$ not n . (This is to do with degrees of freedom.)

VARIANCE (defining formula)

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

However, there is a potential problem with round off error her and so we use an alternative but equivalent formulation. As $\sum_{i=1}^n (x_i - \bar{x})^2 = \sum x_i^2 - \frac{1}{n} (\sum x_i)^2$, then:

VARIANCE (computing formula)

$$s^2 = \frac{1}{n - 1} \left(\sum x_i^2 - \frac{1}{n} (\sum x_i)^2 \right)$$

Standard deviation (s) = square root of variance (s^2)

Example 1:

Number of micro-organisms: 10, 16, 12, 5, 22, 14, 19

$$\sum x_i^2 = 10^2 + 16^2 + 12^2 + 5^2 + 22^2 + 14^2 + 19^2 = 1566$$

$$\sum x_i = 98$$

$$s^2 = \frac{1}{6} \left(1566 - \frac{98^2}{7} \right) = 32,33$$

Example 2:

Number of male rats (x_i)	0	1	2	3	4	5	Total
Frequency (f_i)	2	3	8	7	4	1	25

$$\sum x_i^2 = (2 \times 0^2) + (3 \times 1^2) + (8 \times 2^2) + (7 \times 3^2) + (4 \times 4^2) + (1 \times 5^2) = 187$$

$$\sum x_i = 61$$

$$s^2 = \frac{1}{24} \left(187 - \frac{61 \times 61}{25} \right) = 1,59$$

or alternatively,

x_i	0	1	2	3	4	5	Total
f_i	2	3	8	7	4	1	25
$f_i x_i$	0	3	16	21	16	5	61
$f_i x_i^2$	0	3	32	63	64	25	187

$$s^2 = \frac{1}{n-1} \left(\sum f_i x_i^2 - \frac{(\sum f_i x_i)^2}{\sum f_i} \right) = \frac{1}{24} \left(187 - \frac{61^2}{25} \right) = 1,59$$

Example 3:

A situation in which one is not able to work out the original data:

Interval	Frequency (f_i)	Midpoint (x_i)	$f_i x_i$	$f_i x_i^2$
38–43	6	40,5	243,0	9841,50
44–49	13	46,5	604,5	28109,25
50–55	4	52,5	210,0	11025,00
56–61	3	58,5	175,5	10266,75
62–67	2	64,5	129,0	8320,50
68–73	2	70,5	141,0	9940,00
Total	30	–	1503,0	77503,50

$$s^2 = \frac{1}{29} \left(77503,5 - \frac{(1503,0)^2}{30} \right) = 75,9724$$

Why bother with the defining formula at all? First it makes more intuitive sense of what we are interested in, i.e. the variations from the mean. Second, this approach is part of a family of similar measures (i.e. “moments”):

$\sum (x_i - \bar{x})^3$ basis of measure of skewness of data (3rd moment). (Symmetric distribution gives 0.)

$\sum (x_i - \bar{x})^4$ basis of measure of kurtosis of data (4th moment), i.e. is it sharply peaked or flat?

SUMMARY

Variety of ways of describing and summarising data:

- 1) arithmetic mean (average)
- 2) geometric mean
- 3) harmonic mean
- 4) lower quartile
- 5) upper quartile
- 6) interquartile range
- 7) median
- 8) maximum
- 9) minimum
- 10) range
- 11) mode
- 12) standard deviation
- 13) variance
- 14) coefficient of variation