Clinical Biostatistics

Data, frequencies, and distributions

Martin Bland

Professor of Health Statistics

University of York

http://martinbland.co.uk/

Types of data

Qualitative data arise when individuals may fall into separate classes. E.g. diagnosis, alive/dead.

A qualitative variable is also termed a **categorical variable** or an **attribute**.

Quantitative data are numerical, arising from counts or measurements.

If the values of the measurements are integers (whole numbers) those data are said to be **discrete**. E.g. family size.

If the values of the measurements can take any number in a range, such as height or weight, the data are said to be **continuous**. E.g. blood pressure, serum cholesterol.

Types of data

Variables are qualities or quantities which vary from one member of a sample to another.

A **statistic** is anything calculated from the data alone.

Frequency distributions

Source of referral of patients in a physiotherapy trial (Frost *et al.*, 2004)

Source of referral:	Frequency	Relative frequency
General practitioner	256	89.8%
Consultant	18	6.3%
Triage *	10	3.5%
Sports centre	1	0.4%
Total	285	100.0%

Source of referral is a qualitative variable.

Frost H, Lamb SE, Doll HA, Carver PT, Stewart-Brown S. (2004) Randomised controlled trial of physiotherapy compared with advice for low back pain. *British Medical Journal* **329**, 708-711.

Frequency distributions

Source of referral of patients in a physiotherapy trial (Frost *et al.*, 2004)

Source of referral:	Frequency	Relative frequency
General practitioner	256	89.8%
Consultant	18	6.3%
Triage *	10	3.5%
Sports centre	1	0.4%
Total	285	100.0%

The count of individuals having a particular quality is called the **frequency** of that quality. The proportion of individuals having the quality is called the **relative frequency** or **proportional frequency**.

The relative frequency of general practitioner referral is 256/285 = 0.898 or 89.8%.

Frequency distributions

Source of referral of patients in a physiotherapy trial (Frost *et al.*, 2004)

Source of referral:	Frequency	Relative frequency
General practitioner	256	89.8%
Consultant	18	6.3%
Triage *	10	3.5%
Sports centre	1	0.4%
Total	285	100.0%

The count of individuals having a particular quality is called the **frequency** of that quality. The proportion of individuals having the quality is called the **relative frequency** or **proportional frequency**.

The set of frequencies of all the possible categories is called the **frequency distribution** of the variable.

Ordered categories

Mobility of patients recruited to the VenUS I trial (data of Nelson *et al.*, 2004).

Mobility	Frequency	Relative frequency	Cumulative frequency	Cumulative relative frequency
Walks freely	238	62.1%	238	62.1%
Walks with difficulty	142	37.1%	380	99.2%
Immobile	3	0.8%	383	100.0%
Total	383	100.0%	383	100.0%

Nelson EA, Iglesias CP, Cullum N, Torgerson DJ. (2004) Randomized clinical trial of four-layer and short-stretch compression bandages for venous leg ulcers (VenUS I). *British Journal of Surgery* **91**, 1292-1299.

Ordered categories

Mobility of patients recruited to the VenUS I trial (data of Nelson *et al.*, 2004).

Mobility	Frequency	Relative frequency	Cumulative frequency	Cumulative relative frequency
Walks freely	238	62.1%	238	62.1%
Walks with difficulty	142	37.1%	380	99.2%
Immobile	3	0.8%	383	100.0%
Total	383	100.0%	383	100.0%

The **cumulative frequency** for a value of a variable is the number of individuals with values less than or equal to that value. The **relative cumulative frequency** for a value is the proportion of individuals in the sample with values less than or equal to that value.

Discrete quantitative variable:

Number of episodes of venous ulcers after first onset for patients recruited to the VenUS I trial

Number of episodes	Frequency	Relative frequency	Relative cumulative frequency
0	11	2.9	2.9
1	145	38.7	41.6
2	101	26.9	68.5
3	39	10.4	78.9
4	23	6.1	85.1
5	14	3.7	88.8
6	9	2.4	91.2
7	4	1.1	92.3
8	6	1.6	93.9
9	1	0.3	94.1
10	9	2.4	96.5
.	.	.	.
.	.	.	.

Discrete quantitative variable:

Number of episodes of venous ulcers after first onset for patients recruited to the VenUS I trial

Number of episodes	Frequency	Relative frequency	Relative cumulative frequency
.	.	.	.
.	.	.	.
13	1	0.3	96.8
15	1	0.3	97.1
17	1	0.3	97.3
20	3	0.8	98.1
26	1	0.3	98.4
29	1	0.3	98.7
40	1	0.3	98.9
50	3	0.8	99.7
64	1	0.3	100.0
Total	375	100.0	100.0

Discrete quantitative variable:

Number of episodes of venous ulcers after first onset for patients recruited to the VenUS I trial

Number of episodes	Frequency	Relative frequency	Relative cumulative frequency
0	11	2.9	2.9
1	145	38.7	41.6
2	101	26.9	68.5
3	39	10.4	78.9
4	23	6.1	85.1
5	14	3.7	88.8
6	9	2.4	91.2
.	.	.	.
.	.	.	.

We can count the number of times each possible value occurs to get the frequency distribution.

Continuous variable:

Serum cholesterol (mmol/L) measured on a sample of 86 stroke patients (data of Markus *et al.*, 1995)

3.7	4.8	5.4	5.6	6.1	6.4	7.0	7.6	8.7
3.8	4.9	5.4	5.6	6.1	6.5	7.0	7.6	8.9
3.8	4.9	5.5	5.7	6.1	6.5	7.1	7.6	9.3
4.4	4.9	5.5	5.7	6.2	6.6	7.1	7.7	9.5
4.5	5.0	5.5	5.7	6.3	6.7	7.2	7.8	10.2
4.5	5.1	5.6	5.8	6.3	6.7	7.3	7.8	10.4
4.5	5.1	5.6	5.8	6.4	6.8	7.4	7.8	
4.7	5.2	5.6	5.9	6.4	6.8	7.4	8.2	
4.7	5.3	5.6	6.0	6.4	7.0	7.5	8.3	
4.8	5.3	5.6	6.1	6.4	7.0	7.5	8.6	

Markus HS, Barley J, Lunt R, Bland JM, Jeffery S, Carter ND, Brown MM. (1995) Angiotensin-converting enzyme gene deletion polymorphism: a new risk factor for lacunar stroke but not carotid atheroma. *Stroke* 26, 1329-33.

Continuous variable:

Serum cholesterol (mmol/L) measured on a sample of 86 stroke patients (data of Markus *et al.*, 1995)

3.7	4.8	5.4	5.6	6.1	6.4	7.0	7.6	8.7
3.8	4.9	5.4	5.6	6.1	6.5	7.0	7.6	8.9
3.8	4.9	5.5	5.7	6.1	6.5	7.1	7.6	9.3
4.4	4.9	5.5	5.7	6.2	6.6	7.1	7.7	9.5
4.5	5.0	5.5	5.7	6.3	6.7	7.2	7.8	10.2
4.5	5.1	5.6	5.8	6.3	6.7	7.3	7.8	10.4
4.5	5.1	5.6	5.8	6.4	6.8	7.4	7.8	
4.7	5.2	5.6	5.9	6.4	6.8	7.4	8.2	
4.7	5.3	5.6	6.0	6.4	7.0	7.5	8.3	
4.8	5.3	5.6	6.1	6.4	7.0	7.5	8.6	

As most of the values occur only once, counting the number of occurrences does not help.

Continuous variable:

Serum cholesterol (mmol/L) measured on a sample of 86 stroke patients (data of Markus *et al.*, 1995)

3.7	4.8	5.4	5.6	6.1	6.4	7.0	7.6	8.7
3.8	4.9	5.4	5.6	6.1	6.5	7.0	7.6	8.9
3.8	4.9	5.5	5.7	6.1	6.5	7.1	7.6	9.3
4.4	4.9	5.5	5.7	6.2	6.6	7.1	7.7	9.5
4.5	5.0	5.5	5.7	6.3	6.7	7.2	7.8	10.2
4.5	5.1	5.6	5.8	6.3	6.7	7.3	7.8	10.4
4.5	5.1	5.6	5.8	6.4	6.8	7.4	7.8	
4.7	5.2	5.6	5.9	6.4	6.8	7.4	8.2	
4.7	5.3	5.6	6.0	6.4	7.0	7.5	8.3	
4.8	5.3	5.6	6.1	6.4	7.0	7.5	8.6	

Divide the serum cholesterol scale into class intervals, e.g. from 3.0 to 4.0, from 4.0 to 5.0, and so on.

Count the number of individuals with serum cholesterols in each class interval.

Continuous variable:

The class intervals should not overlap, so we must decide which interval contains the boundary point to avoid it being counted twice.

It is usual to put the lower boundary of an interval into that interval and the higher boundary into the next interval.

Thus the interval starting at 3.0 and ending at 4.0 contains 3.0 but not 4.0.

We can write this as ‘3.0 —’ or ‘3.0 — 4.0$^-$’ or ‘3.0 — 3.999’.

Continuous variable:

Serum cholesterol (mmol/L)

3.7	4.8	5.4	5.6	6.1	6.4	7.0	7.6	8.7
3.8	4.9	5.4	5.6	6.1	6.5	7.0	7.6	8.9
3.8	4.9	5.5	5.7	6.1	6.5	7.1	7.6	9.3
4.4	4.9	5.5	5.7	6.2	6.6	7.1	7.7	9.5
4.5	5.0	5.5	5.7	6.3	6.7	7.2	7.8	10.2
4.5	5.1	5.6	5.8	6.3	6.7	7.3	7.8	10.4
4.5	5.1	5.6	5.8	6.4	6.8	7.4	7.8	
4.7	5.2	5.6	5.9	6.4	6.8	7.4	8.2	
4.7	5.3	5.6	6.0	6.4	7.0	7.5	8.3	
4.8	5.3	5.6	6.1	6.4	7.0	7.5	8.6	

Cholesterol	Frequency	Cholesterol	Frequency
3.0 –	3	7.0 –	19
4.0 –	11	8.0 –	5
5.0 –	24	9.0 –	2
6.0 –	20	10.0 –	2
		Total	86

Continuous variable:

Frequency distribution of serum cholesterol (mmol/L)

Cholesterol	Frequency	Relative frequency
3.0 –	3	0.035
4.0 –	11	0.128
5.0 –	24	0.279
6.0 –	20	0.233
7.0 –	19	0.221
8.0 –	5	0.058
9.0 –	2	0.023
10.0 –	2	0.023
Total	86	1.000

Depends on choice of interval width.

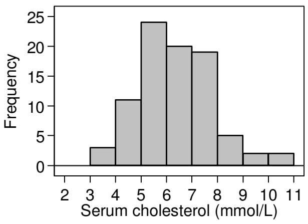
Shape is the important thing.

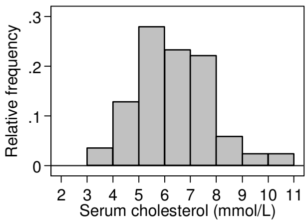
Graphical presentation.

Histograms and other frequency graphs

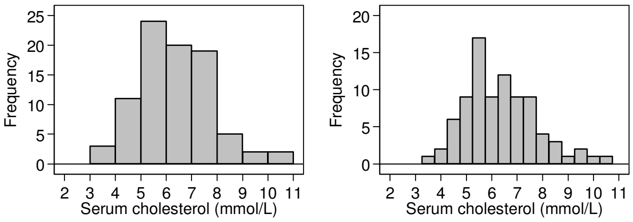
The most common way of depicting a frequency distribution is by a **histogram**.

A diagram where the class intervals are on an axis and rectangles with heights or areas proportional to the frequencies erected on them.



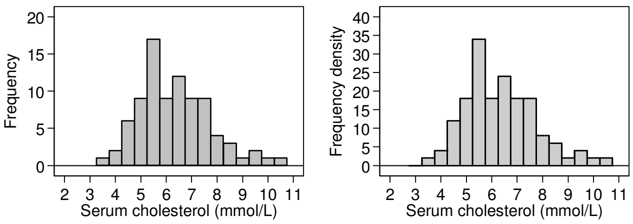


Histograms of cholesterol: frequency scale



Different starting points and interval width, same shape.

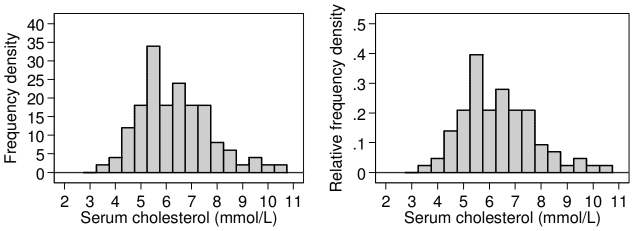
Histograms of cholesterol: frequency density scale



In this case it the area under the histogram which represents the frequency.

For 3.75 to 4.25- mmol/L, frequency density = 4 obs per mmol/L. Width of the interval = 0.5, frequency = $4 \times 0.5 = 2$.

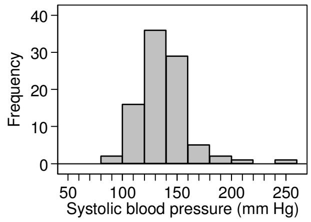
Histograms of cholesterol: frequency density and relative frequency density scales

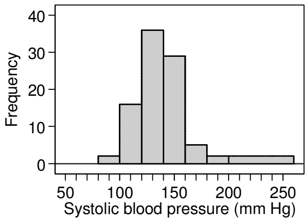


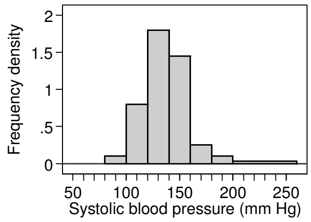
If we plot the relative frequency density, the proportion of observations per unit of the variable, the total area under the histogram is 1.0.

Frequency density enables us to smooth histograms.

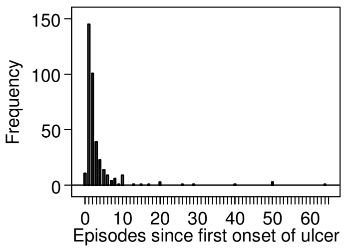
On the frequency scale, combining intervals gives a misleading impression.





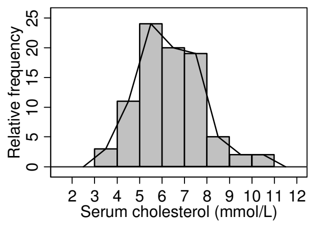


For a discrete variable we can separate the bars:

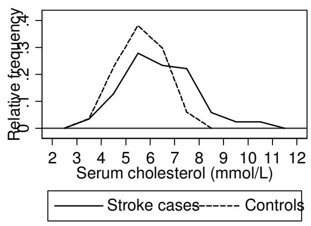


This emphasises the discreteness.

Frequency polygon: join the tops of the bars in the histogram.



Good for showing more than one distribution on the same axes.

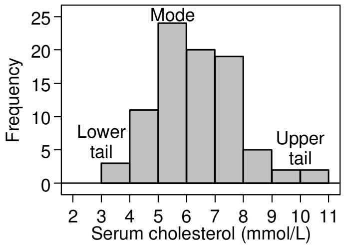


The mode

The most frequently occurring value is called the **mode** of the distribution.

The outer areas are the **tails**.

Unimodal distributions have one mode.

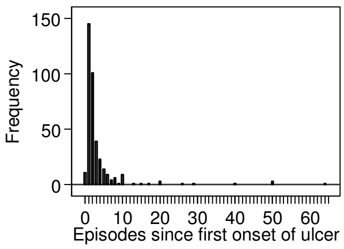


The mode

The most frequently occurring value is called the **mode** of the distribution.

The outer areas are the **tails**.

Unimodal distributions have one mode.

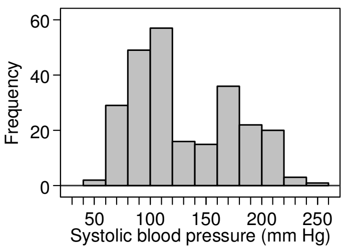


The mode

The most frequently occurring value is called the **mode** of the distribution.

The outer areas are the **tails**.

Bimodal distributions have two modes.

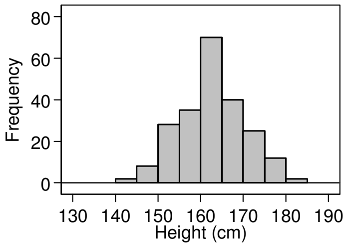


Systolic blood pressure in 251 patients admitted to an intensive therapy unit.

There are two populations.

The parts of the histogram near the extremes are called the **tails** of the distribution.

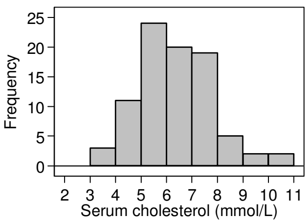
If the tail on the right is of similar length to the tail on the left, the distribution is **symmetrical**:

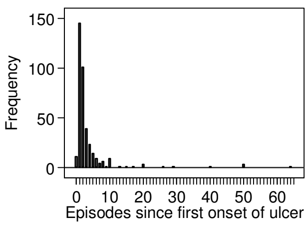


Heights of 222 women admitted to the VenUS I trial.

The parts of the histogram near the extremes are called the **tails** of the distribution.

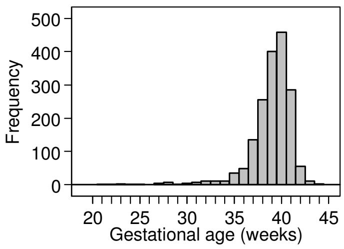
If the tail on the right is longer than the tail on the left, the distribution is **skew to the right** or **positively skew**:





The parts of the histogram near the extremes are called the **tails** of the distribution.

If the tail on the right is longer than the tail on the left, the distribution is **skew to the left** or **negatively skew**:



Gestational age at birth.

Most medical data have unimodal distributions.

Most medical data follow either a symmetrical or positively skew distribution.

Medians and quantiles

The **quantiles** are values which divide the distribution such that there is a given proportion of observations below the quantile.

The **median** is the central value of the distribution, such that half the points are less than or equal to it and half are greater than or equal to it.

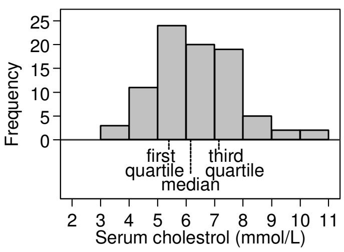
For the cholesterol data the median is 6.15, midway between the 43rd and 44th of the 86 observations.

If we have an odd number of points, the central value is an actual observation, if we have an even number of points, we choose a value midway between the two central values.

Medians and quantiles

The three **quartiles** divide the distribution into four equal parts. The second quartile is the median.

The first quartile has 25% of observations below it, the third quartile has 25% of observations above it.



Note that the quartile is the dividing point, ***not*** the area below it. We should call this a **quarter**.

You will often see this misuse of the term.

Medians and quantiles

We often divide the distribution into 100 **centiles** or **percentiles**.

The median is thus the 50th centile.

The mean

The **arithmetic mean** or **average**, usually referred to simply as the **mean** is found by taking the sum of the observations and dividing by their number.

The mean is often denoted by a little bar over the symbol for the variable, e.g. $\bar{x}$.

The sample mean has much nicer mathematical properties than the median and is thus more useful for the comparison methods described later.

The median is a very useful descriptive statistic, but not much used for other purposes.

Median, mean and skewness:

Mean cholesterol = 6.34, median cholesterol = 6.15.

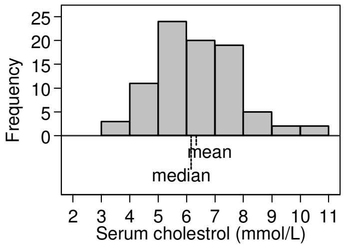
Mean height = 162.2, median height = 162.6.

Mean ulcer episodes = 3.4, median episodes = 2.

If the distribution is symmetrical the sample mean and median will be about the same, but in a skew distribution they will usually be different.

If the distribution is skew to the right, as for serum cholesterol, the mean will usually be greater, if it is skew to the left the median will usually be greater.

This is because the values in the tails affect the mean but not the median.



Increasing the largest observation will pull the mean higher.

It will not affect the median.

Variability

The mean and median are measures of the central tendency or position of the middle of the distribution. We shall also need a measure of the spread, dispersion or variability of the distribution.

The **range** is the difference between the highest and lowest values. This is a useful descriptive measure, but has two disadvantages. Firstly, it depends only on the extreme values and so can vary a lot from sample to sample. Secondly, it depends on the sample size. The larger the sample is, the further apart the extremes are likely to be.

Variability

The range depends on the sample size. The larger the sample is, the further apart the extremes are likely to be.

We can get round this problem by using the **interquartile range** or **IQR**, the difference between the first and third quartiles, a useful descriptive measure.

Variability

For use in the analysis of data, range and IQR are not satisfactory. Instead we use two other measures of variability: variance and standard deviation.

These both measure how far observations are from the mean of the distribution.

Variance is the average squared difference from the mean.

Standard deviation is the square root of the variance.

Variance

Variance is an average squared difference from the mean.

Note that if we have only one observation, we cannot do this. The mean is the observation and the difference is zero. We need at least two observations.

The sum of the squared differences from the mean is proportional to the number of observations minus one, called the **degrees of freedom**.

Variance is estimated as the sum of the squared differences from the mean divided by the degrees of freedom.

Variance

Height: variance = 49.7 cm^2

Cholesterol: variance = 1.96 $mmol/L^2$.

Episodes of ulceration: variance = 42.3 $episodes^2$

Gestational age: variance = 5.24 $weeks^2$

Variance is based on the squares of the observations and so is in squared units.

This makes it difficult to interpret.

Standard deviation

The variance is calculated from the squares of the observations. This means that it is not in the same units as the observations.

We take the square root, which will then have the same units as the observations and the mean.

The square root of the variance is called the standard deviation, usually denoted by s.

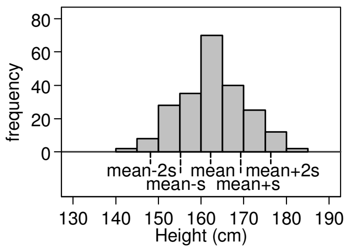
Height: $s = \sqrt{49.7} = 7.1$ cm.

Cholesterol: $s = \sqrt{1.96} = 1.40$ mmol/L.

Episodes of ulceration: $s = \sqrt{42.3} = 6.5$ episodes.

Standard deviation

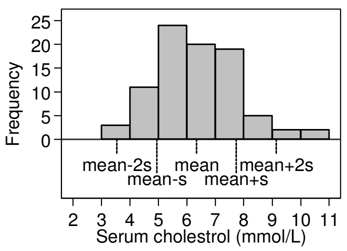
Height: $s = \sqrt{49.7} = 7.1$ cm.



Majority of observations within one SD of mean (usually 2/3 or more). Almost all within about two SD of mean (usually about 95%).

Standard deviation

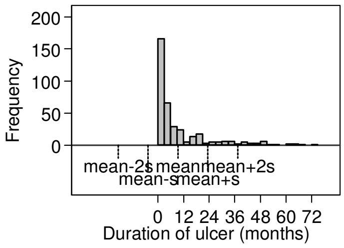
Cholesterol: $s = \sqrt{1.96} = 1.40$ mmol/L.



Majority of observations within one SD of mean (usually 2/3 or more). Almost all within about two SD of mean (usually about 95%), but those outside may be all at one end.

Standard deviation

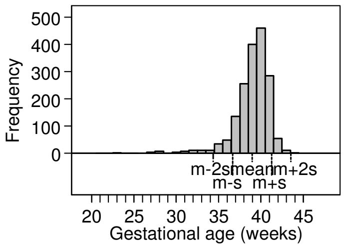
Duration of venous ulcer: $s = \sqrt{189.3} = 13.8$ months.



Majority of observations within one SD of mean (usually 2/3 or more). Almost all within about two SD of mean (usually about 95%), but those outside may be all at one end.

Standard deviation

Gestational age: $s = \sqrt{5.242} = 2.29$ weeks.



Majority of observations within one SD of mean (usually 2/3 or more). Almost all within about two SD of mean (usually about 95%), but those outside may be all at one end.

Spotting skewness

If the mean is less than two standard deviations, two standard deviations below the mean will be negative.

For any variable which cannot be negative, this tells us that the distribution must be positively skew.

If the mean or the median is near to one end of the range or interquartile range, this tells us that the distribution must be skew. If the mean or median is near the lower limit it will be positively skew, if near the upper limit it will be negatively skew.

Spotting skewness

Duration of ulcer: median = 3.0, mean = 9.4, SD = 14.0,
range = 0 to 75, IQR = 1 to 10 months.

These rules of thumb only work one way, e.g. mean may exceed two SD and distribution may still be skew.

Gestational age: median = 39, mean = 38.95, SD = 2.29,
range = 21 to 44, IQR = 38 to 40 weeks.