Applied Biostatistics

Frequency distributions

Martin Bland

Professor of Health Statistics
University of York

http://www-users.york.ac.uk/~mb55/

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, weight.

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

Principle diagnosis of patients in Tooting Bec Hospital

Diagnosis	Number of patients
Schizophrenia	474 (32.3%)
Affective disorders	277 (18.9%)
Organic brain syndrome	405 (27.6%)
Subnormality	58 (4.0%)
Alcoholism	57 (3.9%)
Other and not known	196 (13.4%)
Total	1467 (100.0%)

Diagnosis is a qualitative variable.

Frequency distributions

Principle diagnosis of patients in Tooting Bec Hospital

Diagnosis	Number of patients
Schizophrenia	474 (32.3%)
Affective disorders	277 (18.9%)
Organic brain syndrome	405 (27.6%)
Subnormality	58 (4.0%)
Alcoholism	57 (3.9%)
Other and not known	196 (13.4%)
Total	1467 (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 schizophrenia is 474/1467 = 0.323 or 32.3%.

Frequency distributions

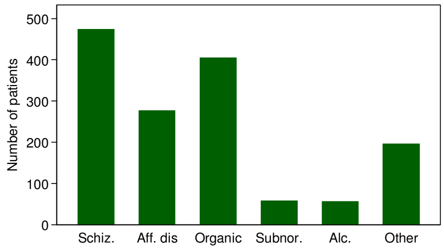
Principle diagnosis of patients in Tooting Bec Hospital

Diagnosis	Number of patients
Schizophrenia	474 (32.3%)
Affective disorders	277 (18.9%)
Organic brain syndrome	405 (27.6%)
Subnormality	58 (4.0%)
Alcoholism	57 (3.9%)
Other and not known	196 (13.4%)
Total	1467 (100.0%)

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

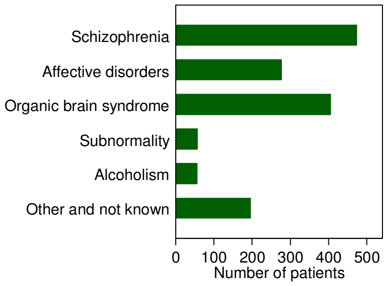
Frequency distributions

We sometimes show this graphically as a bar chart:



Frequency distributions

We can also show this horizontally:



Ordered categories

Likelihood of discharge of patients in Tooting Bec Hospital

Discharge:	Frequency	Relative frequency	Cumulative frequency	Relative cumulative frequency
unlikely	871	0.59	871	0.59
possible	339	0.23	1210	0.82
likely	257	0.18	1467	1.00
Total	1467	1.00	1467	1.00

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:

Parity of 125 women attending antenatal clinics at St. George's Hospital

Parity	Frequency	Relative frequency	Cumulative frequency (per cent)	Relative cumulative frequency (per cent)
0	59	47.2	59	47.2
1	44	35.2	103	82.4
2	14	11.2	117	93.6
3	3	2.4	120	96.0
4	4	3.2	124	99.2
5	1	0.8	125	100.0
Total	125	100.0	125	100.0

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

Continuous variable:

FEV1 (litres) of 57 male medical students

2.85	3.19	3.50	3.69	3.90	4.14	4.32	4.50	4.80	5.20
2.85	3.20	3.54	3.70	3.96	4.16	4.44	4.56	4.80	5.30
2.98	3.30	3.54	3.70	4.05	4.20	4.47	4.68	4.90	5.43
3.04	3.39	3.57	3.75	4.08	4.20	4.47	4.70	5.00	
3.10	3.42	3.60	3.78	4.10	4.30	4.47	4.71	5.10	
3.10	3.48	3.60	3.83	4.14	4.30	4.50	4.78	5.10	

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

To get a useful frequency distribution we need to divide the FEV1 scale into class intervals, e.g. from 3.0 to 3.5, from 3.5 to 4.0, and so on, and count the number of individuals with FEV1s in each class interval.

Continuous variable:

FEV1 (litres) of 57 male medical students

2.85	3.19	3.50	3.69	3.90	4.14	4.32	4.50	4.80	5.20
2.85	3.20	3.54	3.70	3.96	4.16	4.44	4.56	4.80	5.30
2.98	3.30	3.54	3.70	4.05	4.20	4.47	4.68	4.90	5.43
3.04	3.39	3.57	3.75	4.08	4.20	4.47	4.70	5.00	
3.10	3.42	3.60	3.78	4.10	4.30	4.47	4.71	5.10	
3.10	3.48	3.60	3.83	4.14	4.30	4.50	4.78	5.10	

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 3.5 contains 3.0 but not 3.5.

We can write this as '3.0 —' or '3.0 — 3.5^-' or '3.0 — 3.499'.

Continuous variable:

FEV1 (litres) of 57 male medical students

2.85	3.19	3.50	3.69	3.90	4.14	4.32	4.50	4.80	5.20
2.85	3.20	3.54	3.70	3.96	4.16	4.44	4.56	4.80	5.30
2.98	3.30	3.54	3.70	4.05	4.20	4.47	4.68	4.90	5.43
3.04	3.39	3.57	3.75	4.08	4.20	4.47	4.70	5.00	
3.10	3.42	3.60	3.78	4.10	4.30	4.47	4.71	5.10	
3.10	3.48	3.60	3.83	4.14	4.30	4.50	4.78	5.10	

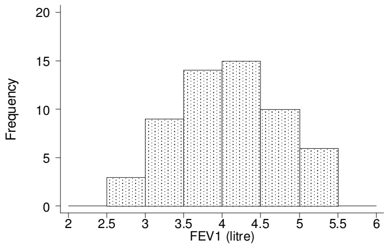
Frequency distribution of FEV1 in 57 male medical students

FEV1	Frequency	Relative frequency
2.0 –	0	0.0
2.5 –	3	5.3
3.0 –	9	15.8
3.5 –	14	24.6
4.0 –	15	26.3
4.5 –	10	17.5
5.0 –	6	10.5
5.5 –	0	0.0
Total	57	100.0

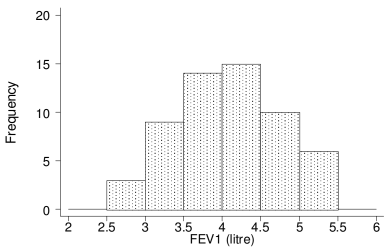
Histograms and other frequency graphs

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

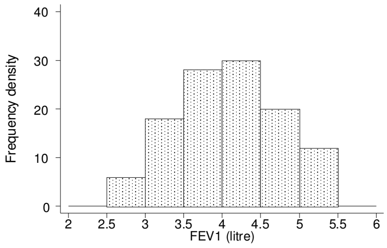
A diagram where the class intervals are on an axis and rectangles with heights or areas proportional to the frequencies erected on them.



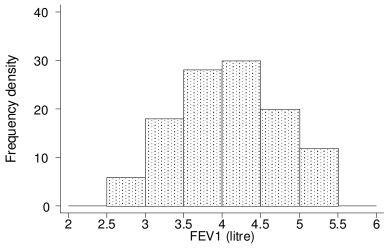
Histogram of FEV1: frequency scale



Histogram of FEV1: frequency per unit FEV1
or frequency density scale



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



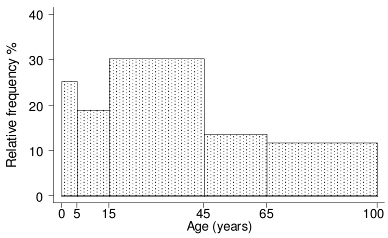
Frequency between 2.5 and 3 = 6 × 0.5 = 3.

Frequency density is particularly useful when we have unequal intervals.

Distribution of age in people suffering accidents in the home

Age group	Relative frequency (per cent)
0 – 4	25.3
5 – 14	18.9
15 – 44	30.3
45 – 64	13.6
65+	11.7

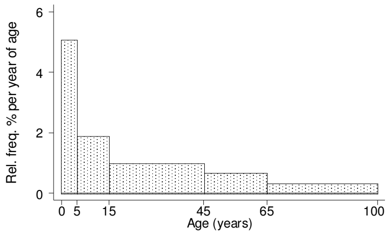
Age distribution of home accident victims: relative frequency scale



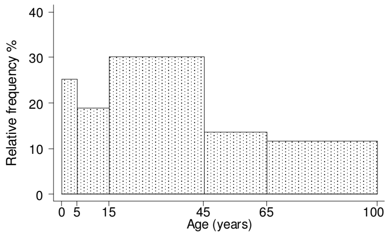
Distribution of age in people suffering accidents in the home

Age group	Relative frequency (per cent)
0 – 4	25.3
5 – 14	18.9
15 – 44	30.3
45 – 64	13.6
65+	11.7

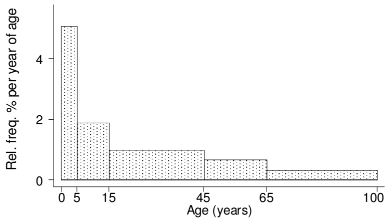
Age distribution of home accident victims: relative frequency **density** scale



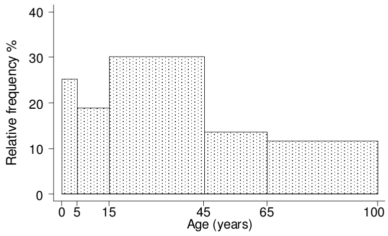
Age distribution of home accident victims: relative frequency scale

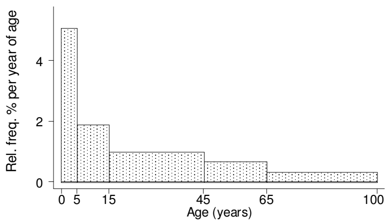


Age distribution of home accident victims: relative frequency density scale

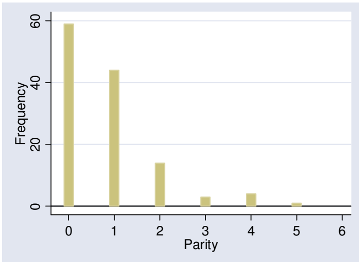


The frequency density scale gives a fair representation of the shape of the distribution when intervals have different widths.



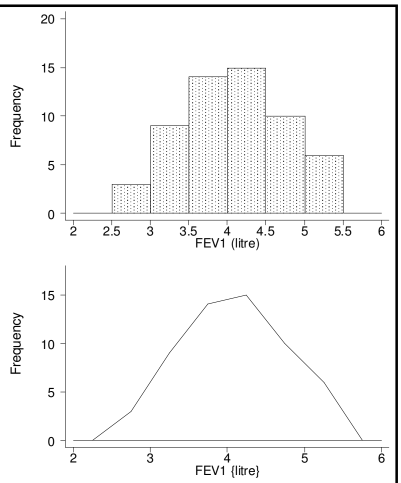


For a discrete variable we can separate the bars:

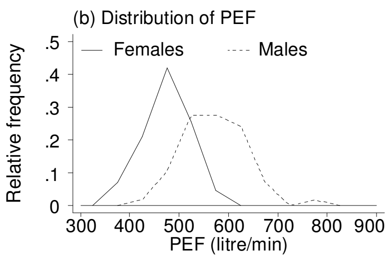


This emphasises the discreteness.

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



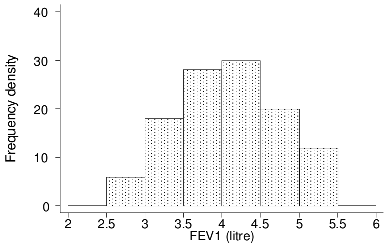
Frequency polygon:
good for showing more than one distribution on the same axes.



The mode

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

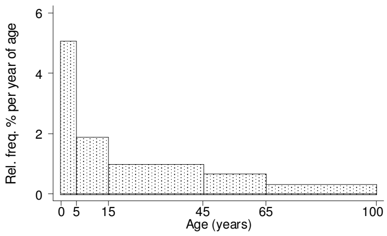
Unimodal:



The mode

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

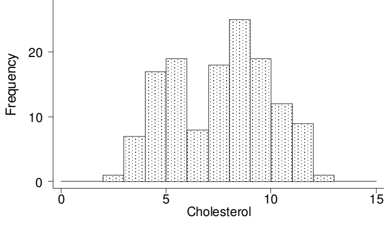
Unimodal:



The mode

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

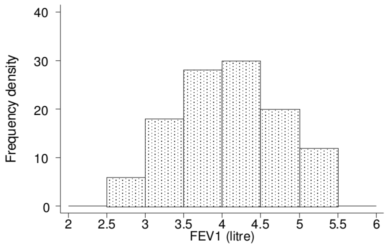
Bimodal:



Serum cholesterol in children from kinships with familial hypercholesterolaemia

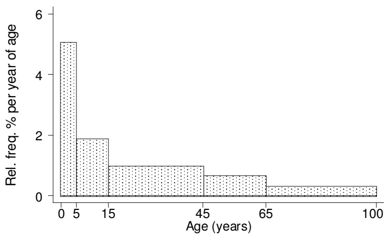
The parts of the histogram near the extremes are called the **tails** of the distribution.

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



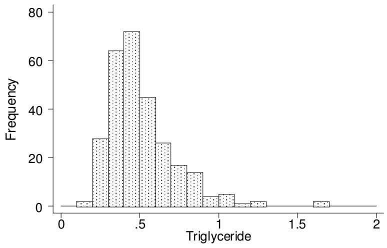
The parts of the histogram near the extremes are called the **tails** of the distribution.

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.

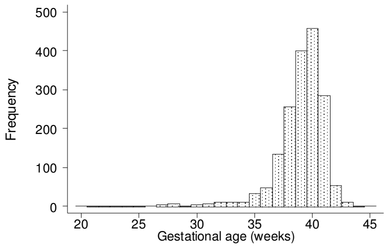
If the tail on the right is longer than the tail on the left, the distribution is **skew to the right** or **positively skew**:



Serum triglyceride in cord blood from 282 babies

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

If the tail on the left is longer than the tail on the right, 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 FEV1 data the median is 4.1, the 29th of the 57 observations.

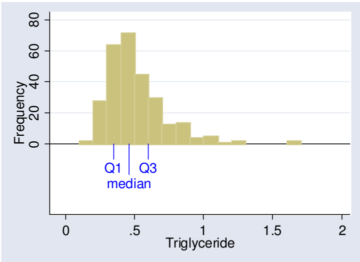
If we have an even number of points, we choose a value midway between the two central values.

Hence the median may not be an actual observation.

Medians and quantiles

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

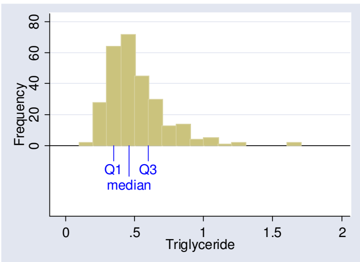
The first quartile has 25% of observations below it, the third quartile has 25% of observations above it.



Medians and quantiles

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.



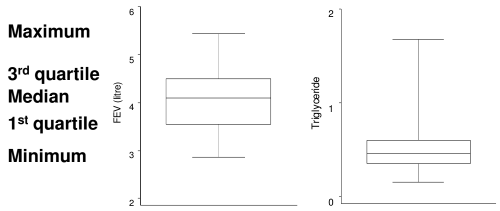
Medians and quantiles

We often divide the distribution into 100 parts at 99 **centiles** or **percentiles**.

The median is thus the 50th centile.

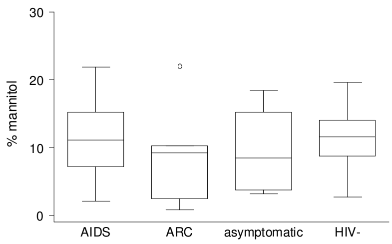
Box and whisker plot

A different way to show a distribution.



Box and whisker plot

Good for comparing several groups.



Gut fat absorption by HIV status

Points more than 1.5 box heights from the top or bottom of the box are often shown separately, as outlying points.

Variability

The median is a measure 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.

Variability

The median is a measure 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.

Two which we often see are the range and the interquartile range.

Variability

The **range** is the difference between the highest and lowest values. This is a useful descriptive measure, but has three disadvantages:

1. It depends only on the extreme values and so can vary a lot from sample to sample.
2. It depends on the sample size. The larger the sample is, the further apart the extremes are likely to be.
3. It is difficult to deal with mathematically and is not useful for use in analysis.

The range is often presented as the minimum and maximum, rather than the difference between them.

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.

The IQR is less variable than the range, but is also difficult to use in analysis.

The interquartile range is often presented as the first quartile and third quartile, rather than the difference between them.