Medical Statistics[1]

Basic Concepts

Farhad Pishgar

Defining the data

Population and samples

Except when a full census is taken, we collect data on a **sample** from a much larger group called the **population**. Because of chance, different samples from the population will give different results and this must be taken into account when using a sample to make inferences about the population. This phenomenon, called **sampling variation**, lies at the heart of statistics.

Types of variable

The raw data of an investigation consist of **observations** made on individuals. The number of individuals is called the **sample size**. Any aspect of an individual that is measured, like blood pressure, or recorded, like age or sex, is called a **variable**.

A first step in choosing how best to display and analyze data is to classify the variables into their different types, as different methods pertain to each. The main division is between numerical (or quantitative) variables, categorical (or qualitative) variables and rates.

First six lines of data from a study of outcome after diagnosis of tuberculosis.

Id	Hospital	Sex	Weight (kg)	Smear result	Culture result	Alive after 6 months?
001	1	M	56.3	Positive	Negative	Y
002	1	M	73.5	Positive	Negative	Y
003	1	F	57.6	Positive	Positive	N
004	2	F	65.6	Uncertain	Positive	Y
005	2	M	81.1	Positive	Positive	Y
006	3	M	56.8	Positive	Negative	Y

1. Based on Kirkwood, B. R., & Sterne, J. A. C. (2003). Essential medical statistics (Second edition ed.): Blackwell Scientific Publications.

A **numerical** variable is either **continuous** or **discrete**. A continuous variable, as the name implies, is a measurement on a continuous scale. In contrast, a discrete variable can only take a limited number of discrete values, which are usually whole numbers.

A **categorical** variable is non-numerical, for instance place of birth, ethnic group, or type of drug. A particularly common sort is a **binary** variable (also known as a **dichotomous** variable), which has only two possible values. We should also distinguish **ordered categorical** variables, whose categories, although non-numerical, can be considered to have a natural ordering. A common example of an ordered categorical variable is social class, which has a natural ordering from most deprived to most affluent.

Rates of disease are measured in follow-up studies, and are the fundamental measure of the frequency of occurrence of disease over time. Examples include the survival rates following different treatments for breast cancer, or the number of episodes of diarrhea/person/year among AIDS patients.

Problem

Determine types of each variable.

Types of variables	
Variable	Type
Episodes of diarrhea a child has had	
Education	
Ethnicity	
Height	
Gender	
Smear result	

”

Displaying the data

There is a temptation to jump straight into complex analyses. This should be avoided. The initial displays of data are valuable in identifying **outliers** (unusual values of a variable) and revealing possible errors in the data, which should be checked and, if necessary, corrected.
Diagrams and tables should always be clearly labelled and self-explanatory; it should not be necessary to refer to the text to understand them. At the same time they should not be cluttered with too much detail, and they must not be misleading.

Frequencies (categorical variables)

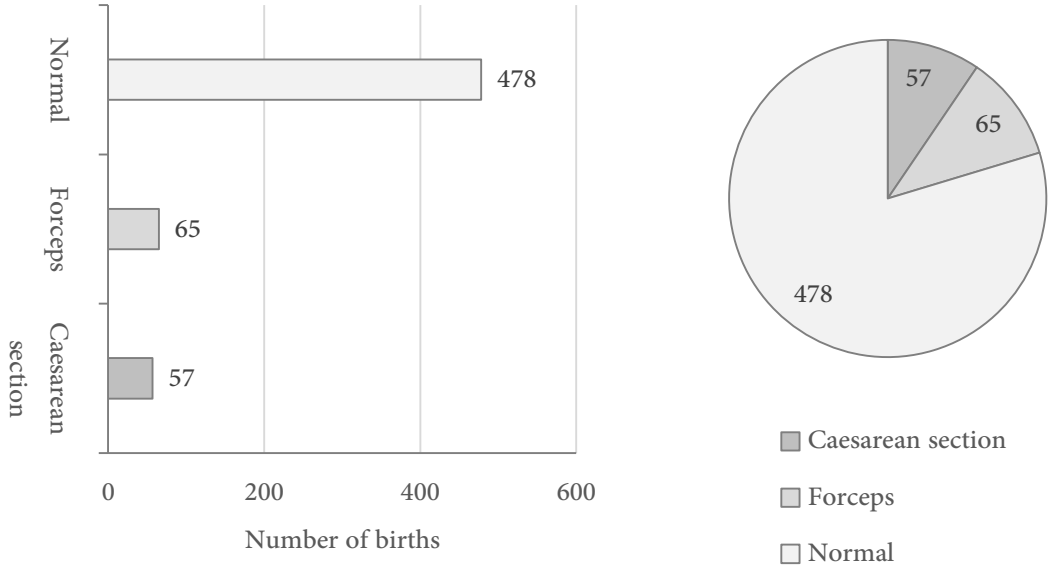
In summarizing categorical variables the main task is to count the number of observations in each category. These counts are called **frequencies**. They are often also presented as **relative frequencies**; that is as proportions or percentages of the total number of individuals.

Method of delivery of 600 babies born in a hospital.

Method of delivery	No. of births	Percentage
Normal	478	79.7
Forceps	65	10.8
Caesarean section	57	9.5
Total	600	100.0

Frequencies and relative frequencies are commonly illustrated by a **bar chart** (also known as a **bar diagram**) or by a **pie chart.**

Bar chart and pie chart showing method of delivery of 600 babies born in a hospital.



Frequencies distributions (numerical variables)

If there are more than about 20 observations, a useful first step in summarizing numerical (quantitative) variable is to form a **frequency distribution**. This is a table showing the number of observations at different values or within certain ranges.

Hemoglobin levels in g/100 ml for 70 women.						
Raw data with the highest and lowest values underlined.						
10.2	13.7	10.4	14.9	11.5	12.0	11.0
13.3	12.9	12.1	9.4	13.2	10.8	11.7
10.6	10.5	13.7	11.8	14.1	10.3	13.6
12.1	12.9	11.4	12.7	10.6	11.4	11.9
9.3	13.5	14.6	11.2	11.7	10.9	10.4
12.0	12.9	11.1	8.8	10.2	11.6	12.5
13.4	12.1	10.9	11.3	14.7	10.8	13.3
11.9	11.4	12.5	13.0	11.6	13.1	9.7
11.2	15.1	10.7	12.9	13.4	12.3	11.0
14.6	11.1	13.5	**10.9**	13.1	11.8	12.2

The first things to do are to count the number of observations and to identify the lowest and highest values. Then decide whether the data should be grouped and, if so, what grouping interval should be used. As a rough guide one should aim for 5–20 groups, depending on the number of observations. The starting points of the groups should be round numbers and, whenever possible, all the intervals should be of the same width. There should be no gaps between groups. The table should be labelled so that it is clear what happens to observations that fall on the boundaries.

Problem

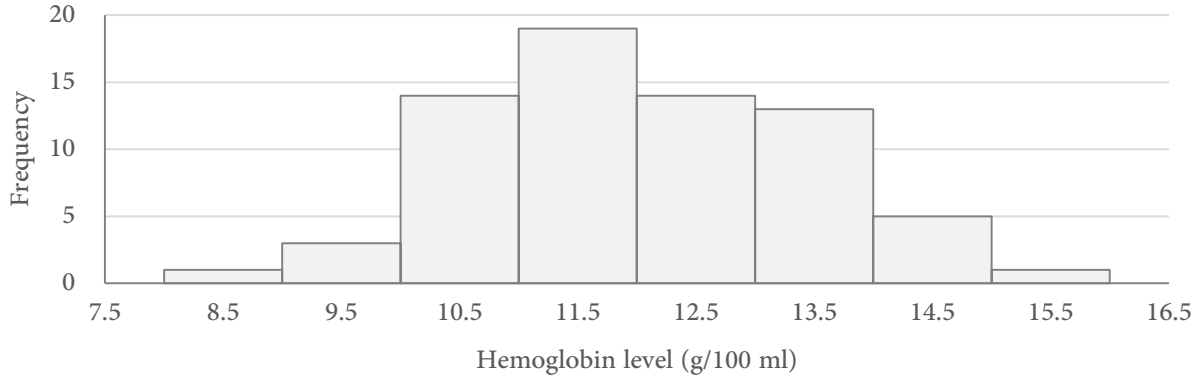
Complete the following frequency distribution table for mentioned hemoglobin levels of 70 women.

Frequency distribution.			
No.	Hemoglobin (g/100 ml)	No. of women	Percentage
1			
2			
3			
4			
5			
6			
7			
8			
-	Total	70	100.0

Histogram

Frequency distributions are usually illustrated by **histograms**. The general rule for drawing a histogram when the intervals are not all the same width is to make the heights of the rectangles proportional to the frequencies divided by the widths, that is to make the areas of the histogram bars proportional to the frequencies.

Histogram of hemoglobin levels of 70 women.



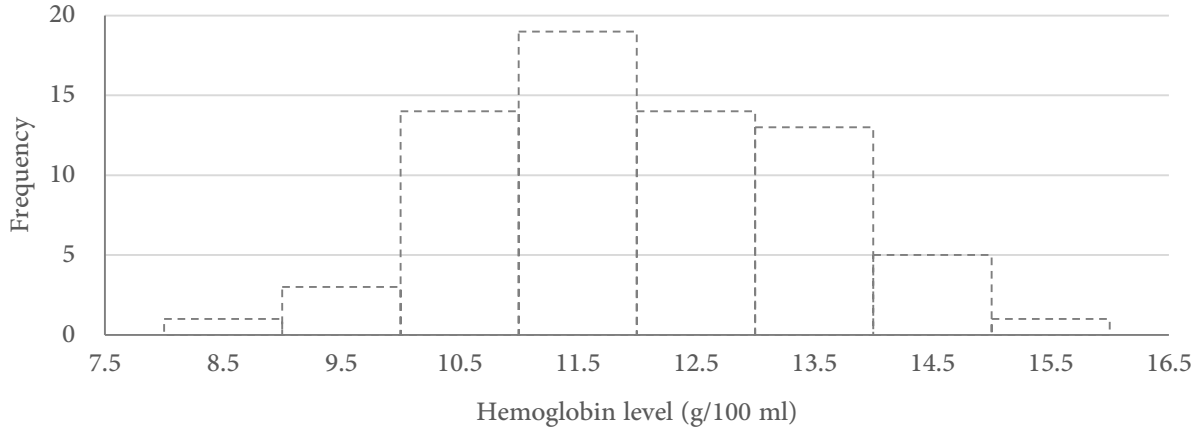
Frequency polygon

An alternative but less common way of illustrating a frequency distribution is a frequency polygon. This is particularly useful when comparing two or more frequency distributions by drawing them on the same diagram.

Problem

Draw the frequency polygon for hemoglobin levels of 70 women.

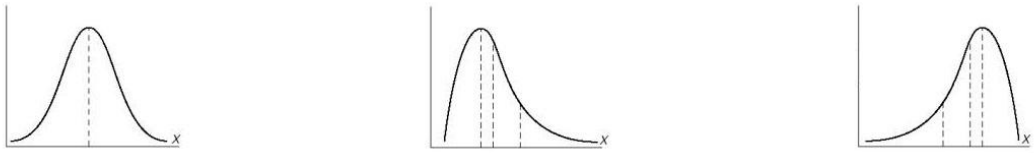
Frequency polygon of hemoglobin levels of 70 women.



Shapes of frequencies distributions

Common shapes of frequency distributions have high frequencies in the center of the distribution and low frequencies at the two extremes, which are called the **upper** and **lower tails** of the distribution. The (a) distribution is also **symmetrical** about the center; this shape of curve is often described as ‘bell-shaped’. The two other distributions are asymmetrical or **skewed**. The upper tail of the distribution in figure (b) is longer than the lower tail; this is called **positively skewed** or skewed to the right. The distribution in figure (c) is **negatively skewed** or skewed to the left.

Three common shapes of frequency distributions with an example of each.

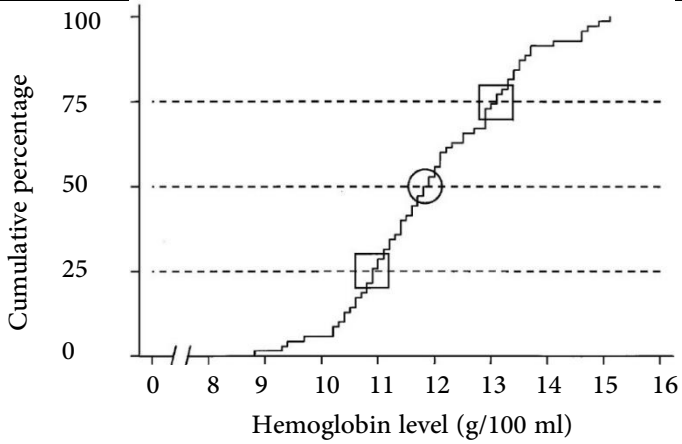


a. Symmetrical and bell-shaped, e.g. height	b. Positively skewed or skewed to the right, e.g. triceps skinfold measurement	c. Negatively skewed or skewed to the left, e.g. period of gestation

Cumulative frequency distribution

Cumulative frequency distributions start from the lowest value and show how the number and percentage of individuals accumulate as the values increase. Cumulative frequency curves are steep where there is a concentration of values, and shallow where values are sparse. An advantage of cumulative frequency distributions is that they display the shape of the distribution without the need for grouping, as required in plotting histogram. However the shape of a distribution is usually more clearly seen in a histogram.

Cumulative frequency distribution of hemoglobin levels of 70 women.



Median and quartiles

Cumulative frequency distributions are useful in recoding a numerical variable into a categorical variable. The median is the midway value; half of the distribution lies below the median and half above it.

Cumulative percentages for different ranges of hemoglobin levels of 70 women.

Observation	Cumulative percentage	Hemoglobin level (g/100 ml)		Quartile
1	1.4	8.8	Minimum = 8.8	1
2	2.9	9.3		1
3	4.3	9.3		1
4	5.7	9.7		1
5	7.1	10.2		1
⋮	⋮	⋮		
15	21.4	10.8		1
16	22.9	10.9		1
17	24.3	10.9		1
18	25.7	10.9	Lower quartile = 10.9	1
19	27.1	11.0		2
20	28.6	11.0		2
⋮	⋮	⋮		
33	47.1	11.7		2
34	48.6	11.8		2
35	50.0	11.8		2
36	51.4	11.9	Median = 11.85	3
37	52.9	11.9		3
38	54.3	12.0		3
⋮	⋮	⋮		
50	71.4	12.9		3
51	72.9	12.9		3
52	74.3	13.0		3
53	75.7	13.1	Upper quartile = 13.1	4
54	77.1	13.1		4
55	78.6	13.2		4
⋮	⋮	⋮		
66	94.3	14.6		4
67	95.7	14.6		4
68	97.1	14.7		4
69	98.6	14.9		4
70	100	15.1	Maximum = 15.1	4

$$Median = \frac{(n+1)^{th}}{2}\ value\ of\ the\ ordered\ observations$$
$$(n = number\ if\ observation)$$

Also there are the two points where the 25% and 75% lines cross the curve. These are called the lower and upper quartiles of the distribution, respectively, and together with the median they divide the distribution into four equally-sized groups.

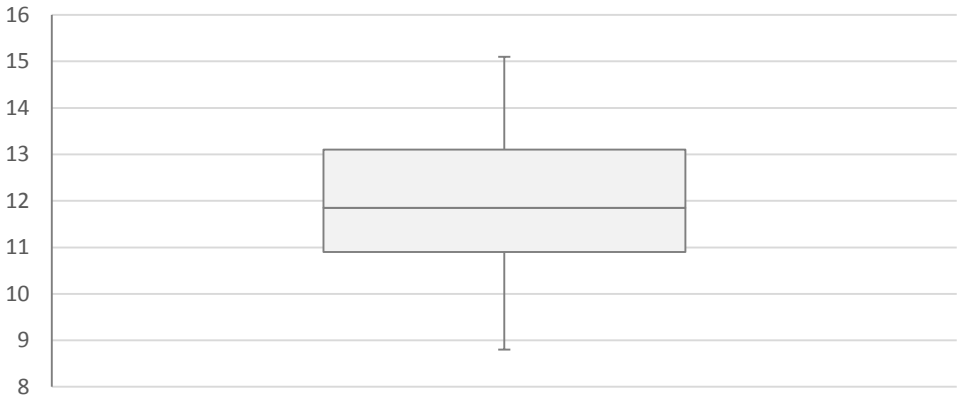
$$Lower\ quartile = \frac{(n+1)^{th}}{4}\ value\ of\ the\ ordered\ observation$$
$$Upper\ quartile = \frac{3 \times (n+1)^{th}}{4}\ value\ of\ the\ ordered\ observation$$

The range of the distribution is the difference between the minimum and maximum values. The difference between the lower and upper quartiles of the hemoglobin data is known as the interquartile range.

$$Range = highest\ value - lowest\ value$$
$$Interquartile\ range = upper\ quartile - lower\ qaurtile$$

A useful plot, based on these values, is a box and whiskers plot. The box is drawn from the lower quartile to the upper quartile; its length gives the interquartile range. The horizontal line in the middle of the box represents the median. Just as a cat's whiskers mark the full width of its body, the 'whiskers' in this plot mark the full extent of the data. They are drawn on either end of the box to the minimum and maximum values.
Note that equal values should always be placed in the same group, even if the groups are then of slightly different sizes.

Box and whiskers plot of the distribution of the hemoglobin levels of 70 women



Cross tabulation

When both variables are categorical, we can examine their relationship informally by cross-tabulating them in a contingency table. A useful convention is for the rows of the table to

correspond to the exposure values and the columns to the outcomes. The interpretability of contingency tables can be improved by including marginal totals and percentages. A useful guide is that the percentages should correspond to the exposure variable.

Comparison of principal sources of water used by households in three villages in West Africa

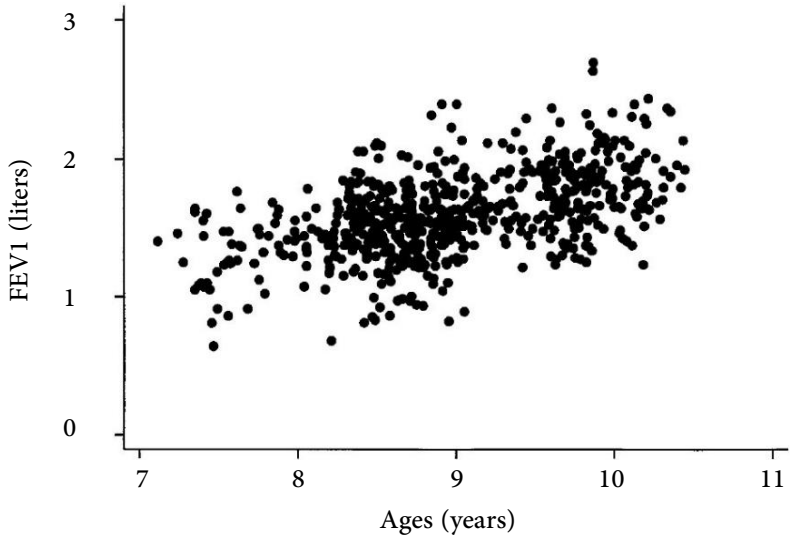
	Water source			
Village	River	Pond	Spring	Total
A	20 (40%)	18 (36%)	12 (24%)	50 (100%)
B	32 (53%)	20 (33%)	8 (13%)	60 (100%)
C	18 (45%)	12 (30%)	10 (25%)	40 (100%)
Total	70 (47%)	50 (33%)	30 (20%)	150 (100%)

Scatter plots

When we wish to examine the relationship between two numerical variables, we should start by drawing a scatter plot. This is a simple graph where each pair of values is represented by a symbol whose horizontal position is determined by the value of the first variable and vertical position is determined by the value of the second variable. By convention, the outcome variable determines vertical position and the exposure variable determines horizontal position.

Scatter plots may also be used to display the relationship between a categorical variable and a continuous variable.

Comparison of principal sources of water used by households in three villages in West Africa



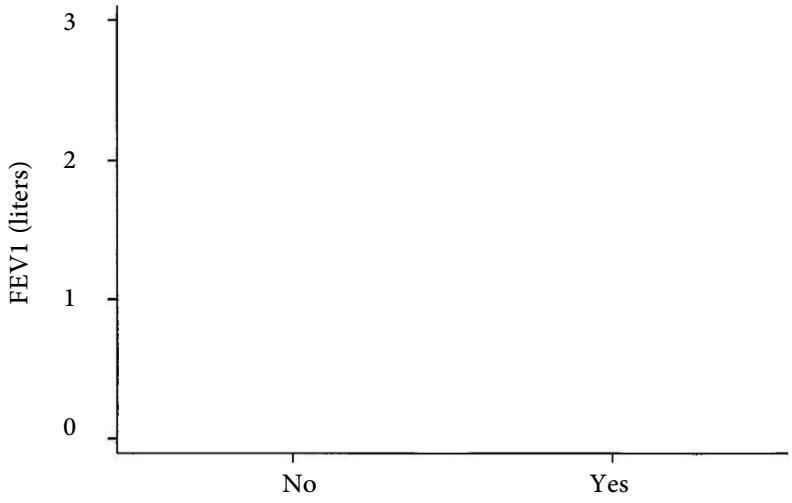
Problem

Draw a chart which shows the distribution of FEV1 in 636 children living in a deprived suburb of Lima, Peru, according to whether they reported respiratory symptoms in the previous 12 months.

Median, interquartile range, and range of FEV1 measurements on 636 children living in a deprived suburb of Lima, Peru, according to whether the child reported respiratory symptoms in the previous 12 months.

Respiratory symptoms in the previous 12 months	n	Lowest FEV1 Value	Lower quartile (25th centile)	Median	Upper quartile (75th centile)	Highest FEV1 value
No	491	0.81	1.44	1.61	1.82	2.69
Yes	145	0.64	1.28	1.46	1.65	2.39
Totals	636	0.64	1.40	1.58	1.79	2.69

Distribution of FEV1 measurements on 636 children living in a deprived suburb of Lima, Peru



Problem

Please complete the following table.

Charts appropriate for each variable

Variable	Chart
Pie chart	Qualitative variables
Bar chart	Qualitative variables
Histogram	Quantitative variables
Area (polygon)	Quantitative variables
Clustered bar chart	Two variables
Box plot	Two variables
Scatter plot	Two variables

”

Means and standard deviation

It is often convenient, to summarize a numerical variable by giving just two measurements, one indicating the average value and the other the spread of the values.

Mean, median and mode

The average value is usually represented by the arithmetic mean, customarily just called the **mean**. This is simply the sum of the values divided by the number of values.

$$Mean, \bar{x} = \frac{\sum x}{n}$$

Other measures of the average value are the **median** and the **mode**. The median was defined before as the value that divides the distribution in half. If the observations are arranged in increasing order, the median is the middle observation. If there is an even number of observations, there is no middle one and the average of the two 'middle' ones is taken. The **mode** is the value which occurs most often.

The mean is usually the preferred measure since it takes into account each individual observation and is most amenable to statistical analysis. The median is a useful descriptive measure if there are one or two extremely high or low values, which would make the mean unrepresentative of the majority of the data. The mode is seldom used.

Range and interquartile range

Two measures of the amount of variation in a data set, the range and the interquartile range, were introduced earlier.

Variance

For most statistical analyses the preferred measure of variation is the **variance** (or the standard deviation, which is derived from the variance, see below).

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

Standard deviation

A disadvantage of the variance is that it is measured in the square of the units used for the observations. For example, if the observations are weights in grams, the variance is in grams squared. For many purposes it is more convenient to express the variation in the original units by taking the square root of the variance. This is called the **standard deviation** (s.d.).

$$s.d., s = \sqrt{\frac{\sum(x-\bar{x})^2}{(n-1)}} \ \ or \ \ s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{(n-1)}}$$

Interpretation of the standard deviation

Usually about 70% of the observations lie within one standard deviation of their mean, and about 95% lie within two standard deviations. These figures are based on a theoretical frequency distribution, called the normal distribution, which is described in later. They may be used to derive reference ranges for the distribution of values in the population.

Coefficient of variation

$$cv = \frac{s}{\bar{x}} \times 100\%$$

The **coefficient of variation** expresses the standard deviation as a percentage of the sample mean. This is useful when interest is in the size of the variation relative to the size of the observation, and it has the advantage that the coefficient of variation is independent of the units of observation.