Theme 4. Introduction to biostatistics

Data types in statistics

The purpose of most studies is to collect **data** to obtain information about a particular area of research. Our data comprise **observations** on one or more variables; any quantity that varies is termed a **variable**. For example, we may collect basic clinical and demographic information on patients with a particular illness. The variables of interest may include the sex, age and height of the patients. Our data is usually obtained from a **sample** of individuals which represents the **population** of interest. Our aim is to condense this data in a meaningful way and extract useful information from it.

Statistics encompasses the methods of collecting, summarizing, analysing and drawing conclusions from the data: we use statistical techniques to achieve our aim. Data may take many different forms. We need to know what form every variable takes before we can make a decision regarding the most appropriate statistical methods to use. Each variable and the resulting data will be one of two types: **categorical** or **numerical** (Fig. 4.1).

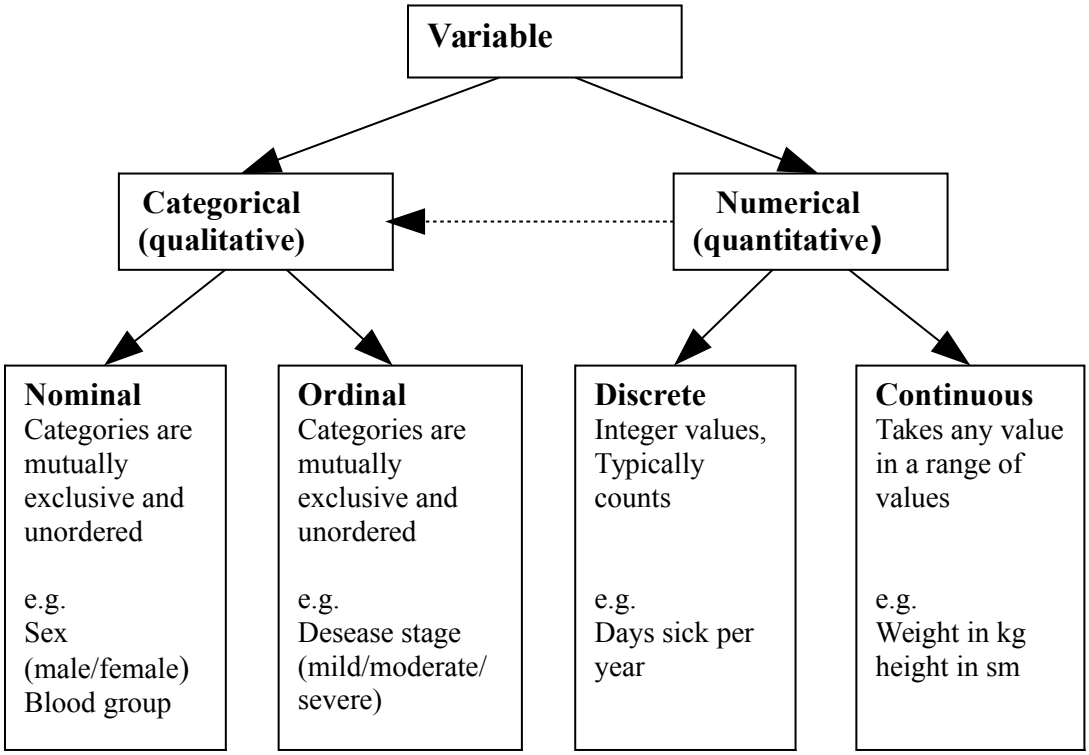

Fig. 4.1. Data types in statistics

Categorical (qualitative) data

These occur when each individual can only belong to one of a number of distinct categories of the variable.

• **Nominal data -** the categories are not ordered but simply have names. Examples include blood group (A, B, AB, and O) and marital status (married/widowed/single etc.). In this case, there is no reason to suspect that being married is any better (or worse) than being single!

• **Ordinal data -** the categories are ordered in some way. Examples include disease staging systems (advanced, moderate, mild, none) and degree of pain (severe, moderate, mild, none).

A categorical variable is **binary** or **dichotomous** when there are only two possible categories. Examples include 'Yes/No', 'Dead/Alive' or 'Patient has disease/Patient does not have disease'.

Numerical (quantitative) data

These occur when the variable takes some numerical value. We can subdivide numerical data into two types.

• **Discrete data** - occur when the variable can only take certain whole numerical values. These are often counts of numbers of events, such as the number of visits to a GP in a year or the number of episodes of illness in an individual over the last five years.

• **Continuous data -** occur when there is no limitation on the values that the variable can take, e.g. weight or height, other than that which restricts us when we make the measurement.

Frequency distributions

An empirical ***frequency distribution*** of a variable relates each possible observation, class of observations (i.e. range of values) or category, as appropriate, to its observed frequency of occurrence. If we replace each frequency by a *relative frequency* (the percentage of the total frequency), we can compare frequency distributions in two or more groups of individuals. Once the *frequencies* (or *relative frequencies*) have been obtained for categorical or some discrete numerical data, these can be displayed visually by *histogram.*

Histogram - is a bar chart, but there should be no gaps between the bars as the data are continuous (Fig. 4.2).

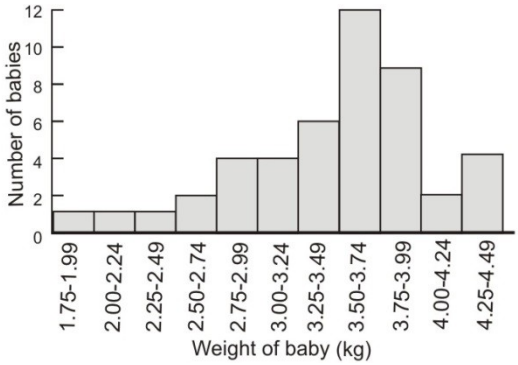

Fig. 4.2. Histogram, showing the weight of the baby at birth

The width of each bar of the histogram relates to a range of values for the variable. For example, the baby's weight (Fig. 5.2) may be categorized into 1.75–1.99 kg, 2.00–2.24kg, . . . , 4.25–4.49 kg. The area of the bar is proportional to the frequency in that range. Therefore, if one of the groups covers a wider range than the others, its base will be wider and height shorter to compensate.

The 'shape' of the frequency distribution

The choice of the most appropriate statistical method will often depend on the shape of the distribution. The distribution of the data is usually **unimodal** in that it has a single 'peak'. Sometimes the distribution is **bimodal** (two peaks) or **uniform** (each value is equally likely and there are no peaks). When the distribution is unimodal, the main aim is to see where the majority of the data values lie, relative to the maximum and minimum values. In particular, it is important to assess whether the distribution is:

• **symmetrical** - centred around some mid-point, with one side being a mirror-image of the other;

• **skewed to the right (positively skewed) -** a long tail to the right with one or a few high values. Such data are common in medical research;

• **skewed to the left (negatively skewed) -** a long tail to the left with one or a few low values (Fig. 5.2).

Describing data

It is very difficult to have any 'feeling' for a set of numerical measurements unless we can summarize the data in a meaningful way. We can also condense the information by providing measures that describe the important characteristics of the data.

The average is a general term for *a measure of location*; it describes a typical measurement.

The arithmetic ***mean***, often simply called ***the mean,*** of a set of values is calculated by adding up all the values and dividing this sum by the number of values in the set.

It is useful to be able to summarize this verbal description by an algebraic formula. Using mathematical notation, we write our set of *n* observations of a variable, x, as x_1, x_2, x_3,..., x_n. For example, x might represent an individual's height (cm), so that x_1 represents the height of the first individual, and x_i the height of the i^{th} individual, etc. We can write the formula for the arithmetic mean of the observations as:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_x}{n}$$

The median

If we arrange our data in order of magnitude, starting with the smallest value and ending with the largest value, then the median is the middle value of this ordered set. The median divides the ordered values into two halves, with an equal number of values both above and below it.

It is easy to calculate the median if the number of observations, ***n, is odd***. It is the $(n + 1)/2^{th}$ observation in the ordered set. So, for example, if n = 11, then the median is the $(11 + 1)/2 = 12/2 = 6^{th}$ observation in the ordered set. If ***n is even*** then, strictly, there is no median. However, we usually calculate it as the arithmetic mean of the two middle observations in the ordered set [i.e. the $n/2^{th}$ and the $(n/2 + 1)^{th}$]. So, for example, if n = 20, the median is the arithmetic mean of the $20/2 = 10^{th}$ and the $(20/2 + 1) = (10 + 1) = 11^{th}$ observations in the ordered set.

The median is similar to the mean if the data are symmetrical, less than the mean if the data are skewed to the right, and greater than the mean if the data are skewed to the left.

The mode

The mode is the value that occurs most frequently in a data set; if the data are continuous, we usually group the data and calculate the modal group. Some data sets do not have a mode because each value only occurs once. Sometimes, there is more than one mode; this is when two or more values occur the same number of times, and the frequency of occurrence of each of these values is greater than that of any other value. We rarely use the mode as a summary measure.

The range is the difference between the largest and smallest observations in the data set.

Percentiles

Suppose we arrange our data in order of magnitude, starting with the smallest value of the variable, x, and ending with the largest value. The value of x that has 1% of the observations in the ordered set lying below it (and 99% of the observations lying above it) is called ***the first percentile***. The value of x that has 2% of the observations lying below it is called the second percentile, and so on. The values of x that divide the ordered set into 10 equally sized groups, that is the 10th, 20th, 30th,..., 90th percentiles, are called ***deciles***. The values of x that divide the ordered set into four equally sized groups, that is the 25th, 50th, and 75th percentiles, are called ***quartiles***. *The 50th percentile is the median*.

The variance

One way of measuring the spread of the data is to determine the extent to which each observation deviates from the arithmetic mean. Clearly, the larger the deviations, the greater the variability of the observations. However, we cannot use the mean of these deviations as a measure of spread because the positive differences exactly cancel out the negative differences. We overcome this problem by squaring each deviation, and finding the mean of these squared deviation; we call this ***the variance***. If we have a sample of n observations, x1, x2, x3,..., xn, whose mean is , we calculate the variance, usually denoted by s^2, of these observations as:

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

The units of the variance are the square of the units of the original observations, e.g. if the variable is weight measured in kg, the units of the variance are kg^2.

The standard deviation

The standard deviation is the square root of the variance. In a sample of n observations, it is:

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

It is evaluated in the same units as the raw data. If we divide the standard deviation by the mean and express this quotient as a percentage, we obtain ***the coefficient of variation***. It is a measure of spread that is independent of the units of measurement, but it has theoretical disadvantages so is not favoured by statisticians.

Theoretical distributions

A theoretical probability distribution is described by a mathematical model. When our empirical distribution approximates a particular probability distribution, we can use our theoretical knowledge of that distribution to answer questions about the data. This often requires the evaluation of probabilities.

Probability measures uncertainty; it lies at the heart of statistical theory. *A probability measures the chance of a given event occurring*. It is a positive number that lies between zero and one. *If it is equal to zero, then the event cannot occur. If it is equal to one, then the event must occur.*

A probability distribution shows the probabilities of all possible values of the random variable. It is a theoretical distribution that is expressed mathematically, and has a mean and variance that are analogous to those of an empirical distribution. Each probability distribution is defined by certain parameters which are summary measures (e.g. mean, variance) characterizing that distribution (i.e. knowledge of them allows the distribution to be fully described). These parameters are estimated in the sample by relevant statistics. Depending on whether the random variable is discrete or continuous, the probability distribution can be either discrete or continuous.

• ***Discrete*** (e.g. *Binomial, Poisson*) - we can derive probabilities corresponding to every possible value of the random variable. The sum of all such probabilities is one.

• ***Continuous*** (e.g. *Normal, Chi-squared, t and F*) - we can only derive the probability of the random variable, x, taking values in certain ranges (because there are infinitely many values of x). If the horizontal axis represents the values of x, we can draw a curve from the equation of the distribution (***the probability density function***). The total area under the curve is one; this area represents the probability of all possible events. The probability that x lies between two limits is equal to the area under the curve between these values (Fig. 4.3).

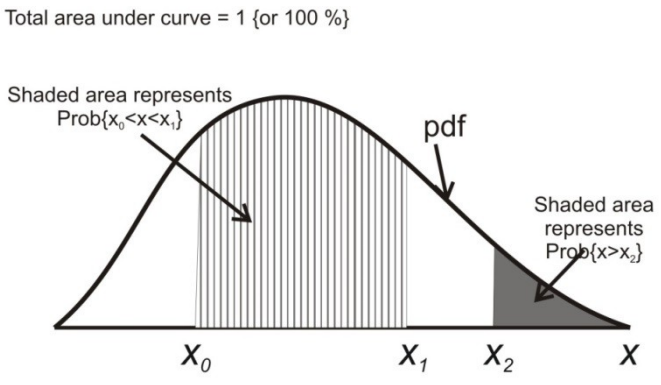

Fig. 4.3. The probability density function.

The Normal (Gaussian) distribution

One of the most important distributions in statistics is the *Normal distribution*. Its probability density function (Fig. 4.4) is:

• completely described by two parameters, the mean (m) and the variance (s^2);

• bell-shaped (unimodal);

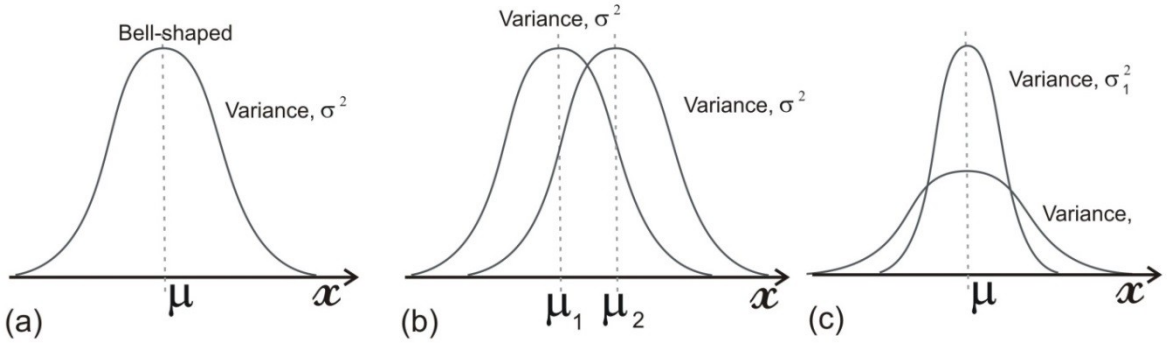

Fig. 4.4. The probability density functions of the *Normal distribution* of the variable x.

• symmetrical about its mean;

• shifted to the right if the mean is increased and to the left if the mean is decreased (assuming constant variance);

• flattened as the variance is increased but becomes more peaked as the variance is decreased (for a fixed mean).

Additional properties are that:

• the mean and median of a *Normal distribution* are equal;

• the probability (Fig. 4.5) that a Normally distributed random variable, x, with mean, m, and standard deviation, s, lies between:

($\mu - \sigma$) and ($\mu + \sigma$) is 0,68

(μ -1,96σ) and (μ + 1,96σ) is 0,95

(μ -2,58σ) and (μ + 2,58σ) is 0,99

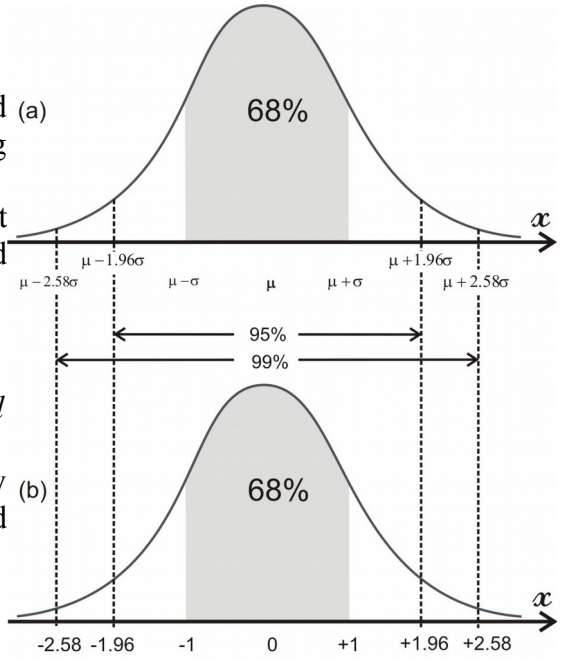

Fig.4.5. Areas (percentages of total probability) under the curve for Normal distribution

Continuous probability distributions are based on continuous random variables.

The t-distribution (Fig.4.6)

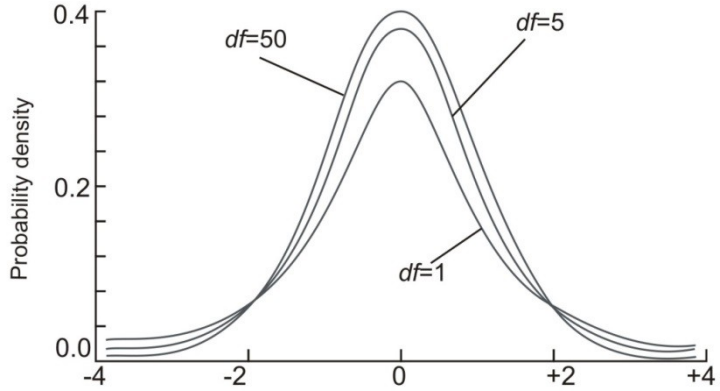

Fig. 4.6. The t-distribution with degree of freedom *df*

• Derived by W.S. Gossett, who published under the pseudonym 'Student', it is often called Student's t-distribution.

• The parameter that characterizes the t-distribution is the degrees of freedom, so we can draw the probability density function if we know the equation of the t-distribution and its degrees of freedom.

• Its shape is similar to that of the Standard Normal distribution, but it is more spread out with longer tails. Its shape approaches Normality as the degrees of freedom increase.

• It is particularly useful for calculating confidence intervals for and testing hypotheses about one or two means.

Sampling and sampling distributions

In statistics, a ***population*** represents the entire group of individuals in whom we are interested. Generally it is costly and labour intensive to study the entire population and, in some cases, may be impossible because the population may be hypothetical (e.g. patients who may receive a treatment in the future). Therefore we collect data on a sample of individuals who we believe are representative of this population (i.e. they have similar characteristics to the individuals in the population), and use them to draw conclusions (i.e. assume inferences) about the population.

When we take ***a sample*** of the population, we have to recognize that the information in the sample may not fully reflect what is true of the population. We have introduced ***sampling error*** by studying only some of the population.

Ideally, we aim for a random sample. Although the statistical tests assume that individuals are selected for the sample randomly, the methods are generally reasonable as long as the sample is ***representative*** of the population.

Point estimates

We are often interested in the value of a parameter in the population, e.g. a mean or a proportion. We estimate the value of the parameter using the data collected from the sample. This estimate is referred to as the sample statistic and is a ***point estimate*** of the parameter (i.e. it takes a single value) as opposed to an ***interval estimate*** which takes a range of values.

If we take repeated samples of the same size from a population, it is unlikely that the estimates of the population parameter would be exactly the same in each sample. However, our estimates should all be close to the true value of the parameter in the population, and the estimates themselves should be similar to each other. By quantifying the variability of these estimates, we obtain information on the precision of our estimate and can thereby assess the sampling error. In reality, we usually only take one sample from the population. However, we still make use of our knowledge of the theoretical distribution of sample estimates to draw inferences about the population parameter.

Sampling distribution of the mean

Suppose we are interested in estimating the population mean; we could take many repeated samples of size n from the population, and estimate the mean in each sample. A histogram of the estimates of these means would show their distribution; this is the sampling distribution of the mean. We can show that:

• If the sample size is reasonably large, the estimates of the mean follow a Normal distribution, whatever the distribution of the original data in the population (this comes from a theorem known as the Central Limit Theorem).

• If the sample size is small, the estimates of the mean follow a Normal distribution provided the data in the population follow a Normal distribution.

• The mean of the estimates is an unbiased estimate of the true mean in the population, i.e. the mean of the estimates equals the true population mean.

• The variability of the distribution is measured by the standard deviation of the estimates; this is known as the ***standard error j*** (often denoted by **SEM**). If we know the population standard deviation (σ), then the standard error of the mean is given by:

$$\text{SEM} = \frac{\sigma}{\sqrt{\text{n}}}$$

Interpreting standard errors

•A large standard error indicates that the estimate is imprecise.

•A small standard error indicates that the estimate is precise.

The standard error is reduced, i.e. we obtain a more precise estimate, if:

• the size of the sample is increased;

• the data are less variable.

SD or SEM?

Although these two parameters seem to be similar, they are used for different purposes. The standard deviation describes the variation in the data values and should be quoted if you wish to illustrate variability in the data. In contrast, the standard error describes the precision of the sample mean, and should be quoted if you are interested in the mean of a set of data values.

Confidence intervals

Once we have taken a sample from our population, we obtain a point estimate of the parameter of interest, and calculate its standard error to indicate the precision of the estimate. However, to most people the standard error is not, by itself, particularly useful. It is more helpful to incorporate this measure of precision into *an interval estimate* for the population parameter. We do this by making use of our knowledge of the theoretical probability distribution of the sample statistic to calculate a ***confidence interval*** for the parameter. Generally the *confidence interval* extends either side of the estimate by some multiple of the standard error; the two values (***the confidence limits***) which define the interval are generally separated by a comma, a dash or the word 'to' and are contained in brackets.

Confidence interval for the mean using the Normal distribution

The sample mean follows a normal distribution if the sample size is large. Therefore we can make use of the properties of the normal distribution when considering the sample mean. In particular, 95% of the distribution of sample means lies within *1.96 standard deviations (SD)* of the population mean. We call this SD the standard error of the mean (SEM), and when we have a single sample, the 95% confidence interval (CI) for the mean is:

(Sample mean - (1.96 × SEM) to Sample mean + (1.96 × SEM))

We usually interpret this confidence interval as the range of values within which we are 95% confident that the true population mean lies.

Confidence interval for the mean using the t-distribution

If the data are not normally distributed, and/or we do not know the population variance, the sample mean follows a t-distribution. We calculate the 95% confidence interval for the mean as:

(Sample mean - ($t_{0.05}$ ×SEM) to Sample mean + ($t_{0.05}$× SEM))

i.e. it is

$$\text{Sample mean } \pm t_{0.05} \times \frac{\sigma}{\sqrt{n}}$$

where $t_{0.05}$ is the percentage point (percentile) of the t-distribution with (n - 1) degrees of freedom which gives a two-tailed probability of 0.05. This generally provides a slightly wider confidence interval than that using the normal distribution to allow for the extra uncertainty that we have introduced by estimating the population standard deviation and/or because of the small sample size. When the sample size is large, the difference between the two distributions is negligible. *Therefore, we always use the t-distribution when calculating a confidence interval for the mean even if the sample size is large.*

By convention we usually quote 95% confidence intervals. We could calculate other confidence intervals e.g. a 99% confidence interval for the mean. Instead of multiplying the standard error by the tabulated value of the t-distribution corresponding to a two-tailed probability of 0.05, we multiply it by that corresponding to a two-tailed probability of 0.01. The 99% confidence interval is wider than a 95% confidence interval, to reflect our increased confidence that the range includes the true population mean.

Example 1. *Confidence interval for the mean*

We are interested in determining the mean age at first birth in women who have bleeding disorders. In a sample of 49 such women:

Mean age at birth of child, $\overline{X}$ = 27.01 years

Standard deviation, s = 5.1282 years

Standard error,

$$\text{SEM} = \frac{5.1282}{\sqrt{49}} = 0.7326 \text{ years}$$

The variable is approximately Normally distributed but, because the population variance is unknown, we use the t-distribution to calculate the confidence interval. The 95% confidence interval for the mean is:

$$27.01 \pm (2.011 \times 0.7326) = (25.54, 28.48) \text{ years}$$

where 2.011 is the percentage point of the t-distribution with (49 - 1) = 48 degrees of freedom giving a two-tailed probability of 0.05.

We are 95% certain that the true mean age at first birth in women with bleeding disorders in the population lies between 25.54 and 28.48 years. This range is fairly narrow, reflecting a precise estimate.

In the general population, the mean age at first birth in 1997 was 26.8 years. As 26.8 falls into our confidence interval, there is no evidence that women with bleeding disorders tend to give birth at an older age than other women.

Note that the 99% confidence interval (25.05, 28.97 years), is slightly wider than the 95% CI, reflecting our increased confidence that the population mean lies in the interval.

Confidence interval for proportion

Our sample of individuals is selected from the population of interest. Each individual either has or does not have the particular characteristic.

r individuals in our sample of size n have the characteristic. The estimated proportion with the characteristic is p = r/n. The proportion of individuals with the characteristic in the population is π.

The number of individuals with the characteristic follows the Binomial distribution, but this can be approximated by the normal distribution. Then p is approximately normally distributed with an estimated mean = p and an estimated standard deviation = $\sqrt{\frac{p(1-p)}{n}}$.

The 95% confidence interval for p is:

$$p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

We can use this confidence interval to assess the clinical or biological importance of the results. A wide confidence interval is an indication that our estimate has poor precision.

Control questions

1. Which data types are used in statistics?
2. What is frequency distribution? How is it representrd?
3. Which measures of location do you know?
4. Which advantages and disadvantages of averages do you know?
5. How do you measure the spread of the data?
6. What does a probability distribution show?
7. What are point and interval estimates?
8. What is standard error of the mean?
9. What is confidence interval of the mean? How to calculate it?
10. What is confidence interval for proportion?

References

1. Medical statistics at a glance / Aviva Petrie, Caroline Sabin.—2nd ed., 2005. – 157 p.
2. Using and understanding medical statistics / David E. Matthews Vernon T. Farewell. – 4th, completely rev. and enl. ed., 2007. – 322 p.
3. Handbook of statistics. Epidemiology and Medical Statistics / C.R. Rao, J.P. Miller, D.C. Rao, 2008. – 852 p.