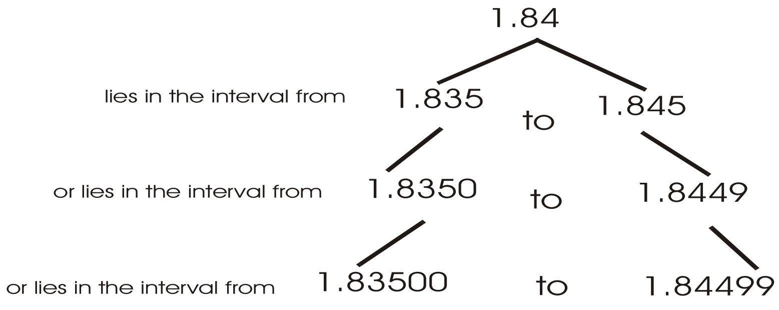
Central Concepts in Statistics

Probability Distribution

A population is an entire set of events characterised by the possibility of possessing some kind of attribute possibly also in varying degrees. For example, every person in the world is a population, and the height of each person is an attribute. The concept of an event is very general. For example, the set of all possible combination of values that arise when a six-sided die is thrown two times is an attribute of each member of the population. We use a variable, say *X*, to describe the population. *X* stands for any member of the population. We use a symbol, say X_i, to stand for the attribute possessed by an element of the population. Thus, $X = X_i$ means the member of the population, *X*, has attribute, or value, X_i

A probability distribution is a measure of the chance or likelihood of a variable *X* taking a value X_i. The symbol $P\left(X = X_1\right)$ means, "the probability that the variable *X* takes the value X_i".

A probability distribution is a description of all the probabilities of a variable taking any of its possible values. Probability distributions are either discrete or continuous. Discrete means that the values that the variable can take are distinct from each other. Between two values it is not possible to find a third value that is some combination of those two values. Eye colour could be discrete variable. For example, the eye colour of a certain animal could be either blue or brown but not somewhere in between. The value of a die throw is a discrete variable. The value can be a 1,2,3,4,5, or 6 but not a value somewhere in between., such as 1·43 or $\sqrt{2}$. A continuous distribution is the opposite of a discrete distribution – that is, between any two halves of a continuous variable. For example, between the heights of 1.2m and 1.3m there are an infinite number of values that can be taken. In fact, the values of a continuous variable are not distinct values but are open ended intervals. To say that a person's height is 1.84m to 3 significant figures is to say that person's height is somewhere in the interval between 1.835M and 1.845M. But this interval is "open" because it is capable of infinite expansion. That is, to represent this diagrammatically


1.84
lies in the interval from
1.835
to
1.845
or lies in the interval from
1.8350
to
1.8449
or lies in the interval from
1.83500
to
1.84499

The ends of these intervals are never capable of sharp definition – they are always open. Thus, strictly, a continuous variable is a variable whose values are open intervals. By contrast a discrete variable is one whose values are closed sets or intervals. A closed set is one whose boundaries are definite. The values of a normal six-sided die are closed sets. The distinct whole numbers 1,2,3,4,5 and 6 are closed off from each other.

Thus, a probability distribution is a measure of the chance or likelihood of a variable X taking a value $X_{i.}$ Probability distributions are either discrete or continuous. Additionally, since the chance of any event cannot be greater than 1 (or 100%) the sum of a probability distribution cannot exceed 1 ; on the other hand, since the chance of every possible event occurring is 100%, the sum of a probability distribution must equal 1. That is, a probability distribution is an assignment of probabilities to the chance of a variable X taking a value X_i such that the sum of all these probabilities is equal to 1. In symbols:

$$\sum P\left(X = X_i\right) = 1$$

Types of Data

A probability distribution is a measure of the chance of a variable X taking a value of X_i. A particular collection of values will be called "data". For example, if I throw a normal six sided die four times and obtain the results 6,1,3,3 then these values will be my data for this "experiment". An experiment is the determination of a set of data.

The values that variables can take –the data- come in different types. The three main types of data are:

Nominal data
The values that the variable can take are discrete, distinct qualitative categories. The values cannot be ranked or measured, they can only be classified . For example, the subjects taken by students at a certain school could be Mathematics, English, French and Sport. Students could be classified according to which subjects they take, but the subjects themselves cannot be placed ina ny kind of order.

Ordinal data
The values that the variable can take can be ranked, or ordered in some way. The values are positions. Thus, position in a race is a standard example of ordinal data.

Interval data
The values that the variable can take are measurable quantities such that equal differences between values in the scale genuinely correspond to real differences between the physical quantities that are being measured. The use of a tape measure or a stop watch exemplify internal level data. The difference between 1.0*m* and 2.0*m* is equal to the difference between 2.0*m* and 3.0*m*. The difference between 1*s* and 2*s* is equal to the

difference between 2*s* and 3*s*. Equal differences in the scale correspond to equal differences in the physical quantities they measure. The scale is also continuous.

Statistics

A statistic is any numerical quantity determined from a set of data. The first statistics that a student meets are measures of central tendency and dispersion.

A measure of central tendency is a measure of the most likely value that occurs in a population. A measure of dispersions is some indication of how to spread out the values of an attribute are within a given population.

Corresponding to each different data type there is a different statistic.

With nominal data it makes sense only to talk of the most common category. This is called the node. The only meaningful concept of central tendency is that of the most frequent category. It is not possible to talk meaningfully of a spread of results. For nominal data there is no measure of dispersion.

With ordinal data the central value is that value is that value which occurs exactly in the middle of the data when the data is ranked. This value is called the median. If there is no exact middle value then the median is the average of the two halves closest to the middle. The spread of the data is measured usually by the interquartile range. This is the range of values from that value that divides the data into the first and third quartiles. The first quartile is that value that separates the lower 1/4 of the values from the upper 3/4. The third quartile is that value that separates the lower 3/4 of the values from the upper 1/4.

With internal data the central value is the mean or average. IF $X, X_{2,} X_{3,} ....X_i$ ... represent a collection (set) of data points then their mean, $\overline{X}$, is given by:

$$\overline{X} = \frac{\sum Xi}{n}$$

Where n is the number of data points. The measure of dispersion is called the standard deviation (σ) or variance (σ^2), which is the square of the standard deviation:

$$\sigma = \sqrt{\frac{\sum x^2}{n} - \left(\overline{x}\right)^2}$$

Actually, the standard deviation is defined by:

$$\sigma = \sqrt{\frac{\sum\left(\bar{x}-x\right)^2}{n}}$$

We can show that the two formulae are equivalent.

In summary, the elementary statistics derived from data of different types are: -

	Central tendency	**Dispersion**
Nominal	Mode	Not defined
Ordinal	Median	Inter-quartile range
Interval	Mean	Standard deviation, or variance

Population and Sample

A population is an entire set of events and its attributes. A probability distribution is a description of the probability of an event drawn from this population taking a particular attribute or value of an attribute.

A sample is a small set of events taken from a population. Occasionally, it is possible to sample the entire population. This is called a census. A census is a sample whose size is equal to the population.

It is not always possible to take a census. In such a case, the best evidence for the properties of the population as a whole is a sample, or series of samples, of that population. The standard sample statistics for a sample of interval level data are the central tendency – the mean $\bar{x}$, and the standard deviation, the measure of dispersion σ. When the properties of a population are unknown it makes sense to estimate them from the values drawn in the sample. The sample statistics may then become "estimates" of the population values.

In other words, the properties of a distribution – its central tendency and its dispersion - can be determined in three different ways:

1. They may simply be assumed or given, or already known as the result of a full census.

2. They may be calculated theoretically from knowledge of the probability distribution.

3. They may be estimated from known statistics calculated for a sample.

We need different symbols to designate these properties in each given context:

Context	How Found	Mean	Variance
Sample	Experimentally – by taking a sample	$\overline{x}$	S^2
Population	Assumed, given or determined by a census	μ	σ^2
Population	Estimates derived from sample statistics	$\hat{\mu}$	$\hat{\sigma}^2$
Population	Theoretically determined from the definition of a probability density function	$E(x)$	$\text{var}(x)$

The symbols $\overline{x}$, S^2 (sample), μ, σ^2 (population), $\hat{\mu}$, $\hat{\sigma}^2$ (estimates), are appropriate only to interval level data. However, the symbols $E(x)$, var(x) apply to any kind of data and any type of probability distribution.

For a probability distribution based on interval level data, the symbols μ (mean) and σ^2 (variance) represent the real, objective properties of the population. Often we do simply assume these values, or calculate them theoretically from knowledge of the probability distribution, but equally important is their estimation from sample statistics.

When a population properly (mean or variance) is estimated from a sample statistic there is always a possibility that the estimate is not correct. It is usual to give a range of values for the estimate and the probability that the real value lies in this range. This is called a confidence interval for a given probability level.

Estimation raises two further issues. The size of a sample, *n*, is usually taken to be less then the size of the population. Additionally, in many cases the process of sampling involves returning the object sampled to the population – this is sampling with replacement. The object sample is replaced into the population and consequently can be sampled again. Consequently, the sample size is potentially infinite, even when the population is finite in size. Estimation involves a process of calculation (an algorithm) that starts with the sample data, or simple statistics calculated from the sample data, and arrives at the estimates. Thus, it is appropriate to ask – what happens to the estimates as the size of the sample grows larger and larger? An estimation process that converges on a single unique value is said to be consistent. An estimation process that arrives at the real value that it is estimating is said to be biased. We use the term estimator to designate the process of arriving at an estimate. An estimator is a function (or algorithm or process) and an estimate is its value. The estimator is a function of sample size. Thus we seek estimators (processes) that are consistent and unbiased.

It turns out that for interval level data that the sample mean $\overline{X}$ is both a consistent and unbiased estimator for the population mean μ.

$$\overline{X} \to \mu \text{ as n} \to \infty$$

But S^2, the sample variance is only consistent. S^2 does not converge on the real variance of the population. It is, consequently, called the biased estimator of the population variance. It turns out that the statistic

$$s^2 = \frac{n}{n-1} S^2$$

where n is the sample size, is an unbiased estimator for the population variance. Thus,

$$s^2 \to \sigma^2 \text{ as } n \to \infty$$

Thus, we need to add to our list of symbols the symbol s^2:

$\bar{X}$	Sample mean, and the unbiased estimate of the population mean.
S^2	Sample variance, and the biased estimate of the population variance.
s^2	Unbiased estimate of the population variance drawn from a sample with variance S^2.
μ	Real population mean.
σ^2	Real population variance.
$\hat{\mu}$	An estimate for the population mean. Usually we take$\hat{\mu} = \bar{X}$.
$\hat{\sigma}^2$	An estimate for the population variance. Usually we take $\hat{\sigma}^2 = s^2$.
$E(X)$	A theoretical value for the central tendency of a given probability distribution.
Var(X)	A theoretical value for the variance of a given probability distribution.

Standard Probability Distributions

The definition of a probability distribution merely states that the sum of its values is equal to 1.

$$\sum_{\mathit{for\ all\ i}} P\left(X = x_i\right) = 1$$

This is a very general definition, and, further, the distribution may be based on discrete or discontinuous variable. Consequently, there are, theoretically, an infinite range of different probability density functions.

However, in practice, only a small number of probability distributions occur in practical distributions, so each distribution can be studied separately. Here we list those principle distributions:

We use the symbol ~ to represent an assignment of a probability density function to a random variable. Thus, for example, $X \sim N\left(\mu, \sigma^2\right)$ states "*X* is a random variable that is normally distributed with mean μ and variance σ^2".

Each probability density function represents a family of similar functions. Each member of each family is basically similar to each other – they differ only in the value of some population parameter, such as the mean or variance.

Principle probability density functions.

1. Binomial Distribution $X \sim B(n,p)$

A discrete finite probability distribution that arises in the context of some trial where there are just two possible outcomes – a "success" with probability p and a failure with probability q=1-p. The trial is repeated n times.

2. Geometric Distribution $X \sim Geo(p)$

A discrete, infinite, probability distribution that arises in the context when a trial is repeated, potentially infinitely, until a "success" is obtained. The particular form of a geometric distribution is determined only by the probability of a "success", p, and this is its sole parameter.

3. Normal Distribution $X \sim N(\mu,\sigma^2)$

A continuous probability distribution that arises as the limit of a binomial distribution as the number of trials increases and tends to infinity.

If $X \sim B(n,p)$ then $X \rightarrow N(\mu,\sigma^2)$ as $n \rightarrow \infty$

we can show that $\mu = np$ and $\sigma^2 = np(1-p)$.

The normal distribution is represented by a bell-shaped curve. The central tendency of a normal distribution is its mean μ; its degree of dispersion – how spread out the distribution is – is described by its variance σ^2.

4. Poisson Distribution $X \sim Po(\mu)$

A special form of the binomial distribution that is discrete and potentially infinite. Thus it is used when the number of trials is potentially infinite, but when the probability of "success" is very small, that is "successes" occur infrequently as a proportion of the total number of events. A Poisson distribution is described by a single parameter, μ. This parameter represents the average number of trials required before a success is obtained. It is usually any number less than 5.

These are the four main distributions used to describe populations that we meet in the real world.

However, just as we cannot often be sure what the central tendency and variance of a population are, so we cannot be sure that a given population genuinely does fit a given probability distribution. Thus we require some way of determining the likelihood that a given population does conform to a chosen probability distribution.

Hypothesis testing

Thus throughout statistics – and throughout science – we make frequent assumptions about populations.

1. We may assume that a population conforms to a particular distribution.

2. We may assume that a population has a particular measure of central tendency – mean, median or mode – depending on whether the values are interval, ordinal or nominal values.

3. We may assume that a population has a particular measure of dispersion. Most usually, for interval level data, we may make the assumption that the variance takes a particular value.

4. We may assume that a sample is drawn from a population with given parameters, or that two samples are drawn fro the same population.

In each case we will want to test the assumption. A specific assumption is called a hypothesis. Thus we require a statistical process for testing each and every kind of hypothesis that we make.

Hypothesis testing is a form of decision-making. We make the decision whether or not to accept the hypothesis. Before we make the test we require a clear definition of the hypothesis under consideration. The test is based on a decision between two different alternatives – one is called the null hypothesis and the other the alternative hypothesis.

H_0: null hypothesis

H_1: alternative hypothesis.

For ordinal and interval level data the alternative hypothesis can depart from the null hypothesis in one or two directions. For example, if the null hypothesis is that the mean of a given population takes some specific value, $\mu = \mu_1$, then the alternative hypothesis may take one of two forms:

1. Either that the real mean is greater than this given value: $\mu > \mu_1$
   or that the real mean is less than this given value: $\mu < \mu_1$
   but not both $\mu < \mu_1$ or $\mu > \mu_1$.

2. That the real mean is not equal to the given value, or, which is the same thing, the real value is either less than or greater than the given value.

   $\mu \neq \mu_1$, that is either $\mu > \mu_1$ or $\mu < \mu_1$, where BOTH alternatives are possible.

Hypothesis tests of the first type are called one-tailed; tests of the second type are called two-tailed.

Since a hypothesis test is a form of estimating that which is unknown, it follows that in performing a hypothesis test we can always make mistake. Thus, every hypothesis test involves the possibility of error. In fact, there are two types of error.

Type I error: rejecting the null hypothesis when the null hypothesis is true.

Type II error: accepting the null hypothesis when the null hypothesis is false.

The more we exclude error of type I the more possible errors of type II become. The possibility of error cannot be excluded from the process of hypothesis testing.

So when making a hypothesis test we need to specify the degree of probability, called a significance level, at which we will set the limit of type I error. Thus, a test at the 5% level specifies that we will allow a 5% probability that a valid null hypothesis will be rejected even though it is true. That is, a 1 in 20 chance of getting it wrong in this way. Setting the level at 1% reduces the chance of a type I error to 1 in 100, but at the same time increases the likelihood of accepting a null hypothesis that is not true (type II error).

Hypothesis testing requires:

1. A specific statement of the hypothesis under test, together with a statement of the alternative under consideration.

2. A statement as to whether the test is one or two-tailed.

3. A statement of the significance level.