Introduction to Statistics and Probability

Kevin O'Brien

September 8, 2013

Contents

1 Introduction to Statistics

1.1 Population

As is so often the case in statistics, some words have technical meanings that overlap with their common use but are not the same. Population is one such word. It is often difficult to decide which population should be sampled.

For instance, if we wished to sample 500 listeners to an FM radio station specializing in music should the population be of listeners to that radio station in general, or of listeners to that stations classical music programme, or perhaps just the regular listeners, or any one of many other possible populations that you can construct for yourself? In practice the population is often chosen by finding one that is easy to sample from, and that may not be the population of first choice.

In medical trials (which are an important statistical application) the population may be those patients who arrive for treatment at the hospital carrying out the trial, and this may be very different from one hospital to another. If you look at any collection of official statistics (which are most important for state planning) you will be struck by the great attention that is given to defining the population that is surveyed or sampled, and to the definition of terms.

For instance, it has proved difficult to get consensus on the meaning of unemployed in recent years but a statistician must be prepared to investigate the population of unemployed. Think about this carefully. It should help you with your Sociology and Marketing and Market Research modules as well as this one:

1.2 Bias

In addition to the common-sense meaning of bias, there is also a more technical meaning for the word in statistics. This will be found both in Chapter 10 of this guide and in the work on estimators in Statistics 2. It seems natural enough to wish to avoid bias, but it is not helpful to be swayed by the value judgements inherited from the use of a word outside the limits of academic discussion.

1.2.1 Sampling

Although it seems sensible to sample a population to avoid the cost of a total enumeration (or census) of that population, it is possible to make a strong argument against the practice. One might well consider that sampling is fundamentally unfair because a sample will not accurately represent the whole population, and it allows the units selected for the sample to have more importance than those not selected. This might be thought undemocratic. Many countries continue to take a full census of their population, even though sampling might be cheaper. It is less obvious, but true, that sampling might well be more accurate, because more time can be spent verifying the information collected for a sample.

1.3 Random Sampling

Random sampling Though it may be clear that random sampling should avoid a sample biased by the prejudices of the sampler, you should think carefully whether that is always a good idea. In other times it was thought that a sampler could produce a better sample by using personal judgement than by randomizing, because all the relevant facts could be taken into consideration. Randomization is popular because this belief seems to be contradicted by experience. Remember that a random sample may (though not very often) come out to look very biased just by chance should one then reject it and try again? If the population has both men and women, would you accept a random sample that just by chance included only men? These difficulties are usually dealt with by some form of restricted randomization. One might, for instance in the example above, use a stratified random sample and select half the sample from the men and half from the women. To carry out a random sample one needs a sample frame. This is the list of all the population units. For instance, if you want to investigate UK secondary schools, the sample frame would be a list of all the secondary schools in the UK. The key advantages of random sampling are that it avoids systematic bias and it allows an assessment of the size of sampling error. Random sampling is carried out, for many populations, by choosing at random without replacement a subset of the very large number of units in the population. If the sample is a small proportion of the population, then there is no appreciable difference between the inferences possible for sampling with replacement and sampling without replacement.

1.4 Stratified Sampling

In a stratified sample the sampling frame is divided into non-overlapping groups or strata, e.g. geographical areas, age-groups, genders. A sample is taken from each stratum, and when this sample is a simple random sample it is referred to as stratified random sampling.

Advantages

Stratification will always achieve greater precision provided that the strata have been chosen so that members of the same stratum are as similar as possible in respect of the characteristic of interest.

The bigger the differences between the strata, the greater the gain in precision. For example, if you were interested in Internet usage you might stratify by age, whereas if you were interested in smoking you might stratify by gender or social class.

It is often administratively convenient to stratify a sample. Interviewers can be specifically trained to deal with a particular age-group or ethnic group, or employees in a particular industry. The results from each stratum may be of intrinsic interest and can be analysed separately.

It ensures better coverage of the population than simple random sampling.

Disadvantages

- Difficulty in identifying appropriate strata.
- More complex to organise and analyse results.

1.5 Quota Sampling

Quota sampling Quota sampling avoids the need for a sample frame. The interviewer seeks out units to ensure that the sample contains given quotas of units meeting specified criteria (known as quota controls). Quota sampling is cheap, but it may be biased systematically by the choices made by the inter- viewer. For instance, interviewers avoid anyone who looks threatening, or too busy, or too strange. Quota sampling does not allow an accurate assessment of sampling error.

Definitions of Statistical Terms

Statistics is a branch of mathematics in which groups of measurements or observations are studied. The subject is divided into two general categories ***descriptive statistics*** and ***inferential statistics***. In descriptive statistics one deals with methods used to collect, organize and analyze numerical facts. Its primary concern is to describe information gathered through observation in an understandable and usable manner.

Similarities and patterns among people, things and events in the world around us are emphasized. Inferential statistics takes data collected from relatively small groups of a population and uses inductive reasoning to make generalizations, inferences and predictions about a wider population. Throughout the study of statistics certain basic terms occur frequently. Some of the more commonly used terms are defined below:

Population

A population is a complete set of items that is being studied. It includes all members of the set. The set may refer to people, objects or measurements that have a common characteristic. Examples of a population are all high school students, all cats, all scholastic aptitude test scores.

A relatively small group of items selected from a population is a sample. If every member of the population has an equal chance of being selected for the sample, it is called a random sample. Examples of a sample are all algebra students at Central High School, or all Siamese cats.

Data are numbers or measurements that are collected. Data may include numbers of individuals that make up the census of a city, ages of pupils in a certain class, temperatures in a town during a given period of time, sales made by a company, or test scores made by ninth graders on a standardized test.

Variables are characteristics or attributes that enable us to distinguish one individual from another. They take on different values when different individuals are observed. Some variables are height, weight, age and price. Variables are the opposite of constants whose values never change.

Definitions

The key terms used in data collection can be defined as follows:

- A variable is the phenomenon being measured in the experiment or observational study. item A continuous variable takes any value on a range of real numbers. item A discrete variable takes only distinct values, usually often integers (analogous to counting)

The key terms used in data collection can be defined as follows:

- A variable is the phenomenon being measured in the experiment or observational study.
- A continuous variable takes any value on a range of real numbers.
- A discrete variable takes only distinct values, usually often integers (analogous to ‘counting).

2 Descriptive Statistics

Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.

2.1 Univariate Analysis

Univariate analysis involves the examination across cases of one variable at a time. There are three major characteristics of a single variable that we tend to look at:

- the distribution
- the central tendency
- the dispersion

2.2 Measures of Central Tendency

The central tendency of a distribution is an estimate of the ”center” of a distribution of values. There are three major types of estimates of central tendency:

- Mean
- Median
- Mode
- The **Mean** or average is probably the most commonly used method of describing central tendency. To compute the mean all you do is add up all the values and divide by the number of values.

  For example, the mean or average quiz score is determined by summing all the scores and dividing by the number of students taking the exam. For example, consider the test score values:

  $$15, 20, 21, 20, 36, 15, 25, 15$$

  The sum of these 8 values is 167, so the mean is 167/8 = 20.875.

- The **Median** is the score found at the exact middle of the set of values. One way to compute the median is to list all scores in numerical order, and then locate the score in the center of the sample. For example, if there are 500 scores in the list, score no. 250 would be the median. If we order the 8 scores shown above, we would get:

  $$15, 15, 15, 20, 20, 21, 25, 36$$

  There are 8 scores and score no. 4 and no. 5 represent the halfway point. Since both of these scores are 20, the median is 20. If the two middle scores had different values, you would have to interpolate to determine the median.

- The **mode** is the most frequently occurring value in the set of scores. To determine the mode, you might again order the scores as shown above, and then count each one. The most frequently occurring value is the mode. In our example, the value 15 occurs three times and is the model. In some distributions there is more than one modal value. For instance, in a bimodal distribution there are two values that occur most frequently.

2.3 Measures of Dispersion

2.3.1 Standard Deviation

Important :The standard deviation is the square root of the variance.

The standard deviation for the sample is called s and the standard deviation for the population is called σ.

The standard deviation is often preferred to the variance as a descriptive measure because it is in the same units as the raw data e.g. if your data is measured in years, the standard deviation will also be in years whereas the variance will be in years squared.

2.4 The Coefficient of Variation

What happens if you have two sets of data with two different means and two different standard deviations? How do you decide which set is more spread out? Remember the size of the standard deviation is relative to the mean it is associated with.

The coefficient of variation (cv) is often used to compare the relative dispersion between two or more sets of data. It is formed by dividing the standard deviation by the mean and is usually expressed as a percentage i.e. (multiplied by 100). Again we distinguish between the population and sample coefficient of variation.

2.5 Arithmetic Mean

The arithmetic mean is what is commonly called the average: When the word "mean" is used without a modifier, it can be assumed that it refers to the arithmetic mean. The mean is the sum of all the scores divided by the number of scores. The formula in summation notation is:

$$\mu = \sum X/N$$

where μ is the population mean and N is the number of scores.

If the scores are from a sample, then the symbol M refers to the mean and N refers to the sample size. The formula for M is the same as the formula for .

$$\mu = \sum X/N$$

The mean is a good measure of central tendency for roughly symmetric distributions but can be misleading in skewed distributions since it can be greatly influenced by scores in the tail. Therefore, other statistics such as the median may be more informative for distributions such as reaction time or family income that are frequently very skewed

2.6 Quantiles and Percentiles

A **percentile** is defined as a point below which a certain per cent of the observations lie e.g. the 50th percentile is the point below which half the observations lie. The percentiles that divide the data into four quarters are called:

Q1 25th percentile or lower quartile

Q2 50th percentile or median

Q3 75th percentile or upper quartile

2.6.1 Interquartile Range (IQR)

The Interquartile Range (Q3 Q1) is a measure of variability commonly used for skewed data. The IQR the difference between the point below which 25% of your data lie and the point below which 75% of your data lie i.e. Q3 - Q1.

2.7 Five number summary and the boxplot

The five number summary consists of the smallest value, the lower quartile, the median, the upper quartile and the largest value in ascending order.

A quarter of the measurements in the data set lie between each of the four pairs of values.

The five-number summary can be used to create a simple graph called a boxplot to visually describe the distribution.

From the boxplot, you can quickly detect any skewing in the shape of the distribution and see whether there are any outliers present in the data set.

An outlier can be caused by human error when entering data or by malfunctioning equipment.

However, outliers can also be valid measurements, and for this reason, it is necessary to isolate them as soon as possible in the analysis. The Box-plot was designed for this purpose.

2.8 Median, and Trimmed Mean

One problem with using the mean, is that it often does not depict the typical outcome. If there is one outcome that is very far from the rest of the data, then the mean will be strongly affected by this outcome. Such an outcome is called an ***outlier***.

An alternative measure is the median. The median is the middle score. If we have an even number of events we take the average of the two middles. The median is better for describing the typical value. It is often used for income and home prices.

Standard Deviation

The standard deviation (σ) for the population or s for the sample is the square root of the variance.

2.9 Calculating Variance and Standard Deviation

Variance and Standard Deviation: Step by Step

1. Calculate the mean, $\bar{x}$.
2. Write a table that subtracts the mean from each observed value.
3. Square each of the differences.
4. Add this column.
5. Divide by $n - 1$ where n is the number of items in the sample This is the ***variance***.
6. To get the standard deviation we take the square root of the variance.

The sample standard deviation will be denoted by s and the population standard deviation will be denoted by the Greek letter σ.

The sample variance will be denoted by s^2 and the population variance will be denoted by $sigma^2$.

The variance and standard deviation describe how spread out the data is. If the data all lies close to the mean, then the standard deviation will be small, while if the data is spread out over a large range of values, s will be large. Having outliers will increase the standard deviation.

One of the flaws involved with the standard deviation, is that it depends on the units that are used. One way of handling this difficulty, is called the coefficient of variation which is the standard deviation divided by the mean times 100%

2.10 Geometric and Harmonic Mean

2.11 Measures of Dispersion

The most common measures of dispersion are:

- Variance
- Standard deviation
- Mean Absolute Deviation (MAD)

- Range

- Inter-quartile range.

2.12 Inter-Quartile Range

The inter-quartile range is used when you have measured location using the median. It has the same advantages and disadvantages as the median.

3 Graphical Methods for Descriptive Statistics

The main representations we use in this subject are histograms, stem and leaf diagrams, and boxplots. We also use scatter plots for two variables

Histograms

A frequency distribution can be represented graphically on a histogram. A histogram is a bar graph on which the bars are adjacent to each other with no space between them. To construct a histogram, arrange the data in equal intervals. Represent the frequencies along the vertical axis and the scores along the horizontal axis.

3.1 The shape of a Frequency Distribution

The shape of a distribution refers to

- its symmetry (or lack of it). If a distribution is not symmetric, it is said to be **skewed**
- its peakedness, formally known as **kurtosis**.

3.2 Constructing a Box-plot

To construct a boxplot

1. Calculate Q1, the median, Q3 and the IQR.
2. Draw a horizontal line to represent the scale of measurement.
3. Draw a box just above the line with the right and left ends at Q1 and Q3.
4. Draw a line through the box at the location of the median.
5. To detect outliers you need to determine a lower fence and an upper fence.

   a. Lower fence is

b. Upper fence is

6. Any values below the lower fence or above the upper fence are classes as outliers.

7. To finish the boxplot

   a. Mark any outliers with an asterisk ($*$) on the graph.

   b. Extend horizontal lines , called whiskers, from the ends of the box to the smallest and largest values that are not outliers.

   c. (Remark a variation is to extend to the lower and upper fences)

3.3 Interpreting Box-plots

- **Outliers**
  The first feature that you look for when analysing a boxplot is the presence of outliers.

  Outliers are extreme values and can greatly influence your analysis. For that reason, you should check your data and make sure you have entered it correctly.

  You also have the option of removing outliers, making a note that you have removed them, and presenting your analysis without them.

- **Skewed Data**
  The second feature is the degree of skewness. As you learned earlier, the quartiles divide the data into four sections, each containg 25% of the measurements.

  You are interested in how spread out or tightly packed the data are. The length of the whiskers and the position of the median in the box tell you this. Notice that 25% of the values in the boxplot are less than Q1 and this includes the outliers.

- **IQR**
  The third feature is the variation/dispersion around the median. The IQR is the middle 50% of the data. When you are dealing with skewed data, the IQR is the most reliable measure of variation. Outliers affect the mean, making it an unrealistic measure of centrality (for symmetric data).

The most common use of box plots is for comparing two data sets on the same scale.

For now, it is important that you are clear what a box-plot tells you about a distribution of data and what measure of centrality and variability are most appropriate based on the distribution.

4 Probability

4.1 Introduction to Probability

There are many situations in everyday life where the outcome is not known with certainty. For example; applying for a job or sitting an examination.

We use words like "Chance", "the odds", "likelihood" etc but the most effective way of dealing with uncertainty is based on the concept of probability.

Probability can be thought of as a number which measures the chance or likelihood that a particular event will occur.

An example of the use of probability is in decision making. Decision making usually involves uncertainty. For example, should we invest in a company if there is a chance it will fail?

Should we start production of a product even though there is a likelihood that the raw materials will arrive on time in poor? Having a number which measures the chances of these events occurring helps us to make a decision.

Why are we interested in probability in this module? Many statistical methods use the idea of a probability distribution for this data.

We have already looked at relative frequency distribution in Section 2. Probability distributions are based on the same concepts as relative frequency distributions. They are used to calculate probabilities of different values occurring in the data collected.

We will examine probability distributions in more detail in Section 4. First we need to learn about the basic concepts of probability.

4.2 Random experiment

- **Sample Space**, S. For a given experiment the sample space, S, is the set of all possible outcomes.

- **Event**, E. This is a subset of S. If an event E occurs, the outcome of the experiment is contained in E.

Probability concerns itself with random phenomena or probability experiments. These experiments are all different in nature, and can concern things as diverse as rolling dice or flipping coins. The common thread that runs throughout these probability experiments is that there are observable outcomes. If we collect all of the possible outcomes together, then this forms a set that is known as the sample space.

In this set theory formulation of probability the sample space for a problem corresponds to an important set. Since the sample space contains every outcome that is possible, it forms a setting of everything that we can consider. So the sample space becomes the universal set in use for a particular probability experiment.

A probability distribution is a table of values showing the probabilities of various outcomes of an experiment.

For example, if a coin is tossed three times, the number of heads obtained can be 0, 1, 2 or 3. The probabilities of each of these possibilities can be tabulated as shown:

Number of Heads	0	1	2	3
Probability	1/8	3/8	3/8	1/8

A discrete variable is a variable which can only take a countable number of values. In this example, the number of heads can only take 4 values (0, 1, 2, 3) and so the variable is discrete. The variable is said to be random if the sum of the probabilities is one.

4.2.1 Common Sample Spaces

Sample spaces abound and are infinite in number. But there are a few that are frequently used for examples in introductory statistics. Below are the experiments and their corresponding sample spaces:

- For the experiment of flipping a coin, the sample space is Heads, Tails and has two elements.

- For the experiment of flipping two coins, the sample space is (Heads, Heads), (Heads, Tails), (Tails, Heads), (Tails, Tails) and has four elements.

- For the experiment of flipping three coins, the sample space is (Heads, Heads, Heads), (Heads, Heads, Tails), (Heads, Tails, Heads), (Heads, Tails, Tails), (Tails, Heads, Heads), (Tails, Heads, Tails), (Tails, Tails, Heads), (Tails, Tails, Tails) and has eight elements.

- For the experiment of flipping n coins, where n is a positive whole number, the sample space consists of 2n elements. There are a total of

$C(n, k)$ ways to obtain k heads and $n - k$ tails for each number k from 0 to n.

- For the experiment consisting of rolling a single six-sided die, the sample space is
$$\{1, 2, 3, 4, 5, 6\}$$

- For the experiment of rolling two six-sided dice, the sample space consists of the set of the 36 possible pairings of the numbers 1, 2, 3, 4, 5 and 6.

- For the experiment of rolling three six-sided dice, the sample space consists of the set of the 216 possible triples of the numbers 1, 2, 3, 4, 5 and 6.

- For an experiment of drawing from a standard deck of cards, the sample space is the set that lists all 52 cards in a deck. For this example the sample space could only consider certain features of the cards, such as rank or suit.

4.2.2 Forming Other Sample Spaces

These are the basic sample spaces. Others are out there for different experiments. It is also possible to combine several of the above experiments. When this is done, we end up with a sample space that is the Cartesian product of our individual sample spaces. We can also use a tree diagram to form these sample spaces.

What is a contingency table?

A contingency table is essentially a display format used to analyse and record the relationship between two or more categorical variables. It is the categorical equivalent of the scatterplot used to analyse the relationship between two continuous variables.

4.3 Combining Probabilities

Events rarely occur in isolation. Usually we are interested in a combination or compound of events; for example

- The probability that two sections of a factory will be understaffed on the same day
- The probability of having a car accident today, given that you have had a car accident in the last five years.

We will look at two laws of probability for combining events

- The Addition Law
- The multiplication Law

4.4 Conditional Probability

The conditional probability of an event is the probability that an event A occurs given that another event B has already occurred. This type of probability is calculated by restricting the sample space that were working with to only the set B.

The formula for conditional probability can be rewritten using some basic algebra. Instead of the formula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

4.5 Probability trees

The setting out of solutions to problems requiring the manipulation of the probabilities of mutually exclusive and independent events can sometimes be helped by the use of probability tree diagrams. These have useful applications in decision theory.

The best choice of probability tree structure often depends upon the question and the natural order in which events like A and B above occur.

4.6 Histograms

A histogram is constructed from a frequency table. The intervals are shown on the X-axis and the number of scores in each interval is represented by the height of a rectangle located above the interval. A histogram of the response times from the dataset Target RT is shown below.

4.7 Cumulative Distribution Function

The cumulative distribution function (c.d.f.) of a discrete random variable X is the function F(t) which tells you the probability that X is less than or equal to t. So if X has p.d.f. P(X = x), we have:

$$F(t) = P(X \leq 1)$$

In other words, for each value that X can be which is less than or equal to t, work out the probability that X is that value and add up all such results.

Example

In the above example where the die is thrown repeatedly, lets work out $P(X \leq t)$ for some values of t.

P(X $\leq$ 1) is the probability that the number of throws until we get a 6 is less than or equal to 1. So it is either 0 or 1.

- P(X = 0) = 0
- $P(X = 1) = 1/6$.
- Hence $P(X \leq 1) = 1/6$

Similarly, $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$
= 0 + 1/6 + 5/36 = 11/36

5 Techniques for Counting

- Combinations
- Permutations
- Permutations with constraints

5.1 Permuations of subsets

The number of permutations of subsets of k elements selected from a set of n different elements is

$$P(n, r) = \frac{n!}{(n-k)!}$$

5.2 Combinations of subsets

The number of combinations that can be selected from n items is

$$\binom{n}{k} = \frac{n!}{k! \times (n-k)!}$$

6 Discrete Random Variables

A random variable is a numerical description of the outcome of an experiment.

Random variables can be classified as discrete or continuous, depending on the numerical values they may take.

A ranom variable that may assume any numerical value in an interval or collection of intervals is called a continuous random variable.

7 Discrete Probability Distributions

- Poisson
- Binomial
- Geometric

7.1 What Is a Probability Distribution?

If you spend much time at all dealing with statistics, pretty soon you run into the phrase probability distribution. It is here that we really get to see how much the areas of probability and statistics overlap. Although this may sound like something technical, the phrase probability distribution is really just a way to talk about organizing a list of probabilities. A probability distribution is a function or rule that assigns probabilities to each value of a random variable. The distribution may in some cases be listed. In other cases it is presented as a graph.

7.1.1 Graph of a Probability Distribution

A probability distribution can be graphed, and sometimes this helps to show us features of the distribution that were not apparent from just reading the list of probabilities. The random variable is plotted along the x-axis, and the corresponding probability is plotted along the y - axis.

- For a discrete random variable, we will have a histogram
- For a continuous random variable, we will have the inside of a smooth curve

The rules of probability are still in effect, and they manifest themselves in a few ways. Since probabilities are greater than or equal to zero, the graph of a probability distribution must have y-coordinates that are nonnegative. Another feature of probabilities, namely that one is the maximum that the probability of an event can be, shows up in another way.

$$\text{Area} = \text{Probability}$$

8 Binomial Probability Distribution

The binomial distribution is a particular example of a probability distribution involving a discrete random variable. It is important that you can identify situations which can be modelled using the binomial distribution.

- There are n independent trials
- There are just two possible outcomes to each trial, success and failure, with fixed probabilities of p and q respectively, where q = 1 p.

The discrete random variable X is the number of successes in the n trials. X is modelled by the binomial distribution $B(n, p)$. You can write $X \sim B(n, p)$.

8.1 Poisson probability distribution

A discrete random variable that is often used is one which estimates the number of occurrences over a specified time period or space.

(remark : a specified space can be a specified length , a specified area, or a specified volume.)

If the following two properties are satisfied, the number of occurrences is a random variable described by the Poisson probability distribution

Properties

1) The probability of an occurrence is the same for any two intervals of equal length.

2) The occurrence or non-occurrence in any interval is independent of the occurrence or non-occurrence in any other interval.

The Poisson probability function is given by

- f(x) the probability of x occurrences in an interval.
- λ is the expected value of the mean number of occurrences in any interval. (We often call this the Poisson mean)
- e=2.71828284

8.2 Poisson Approximation of the Binomial Probability Distribution

The Poisson distribution can be used as an approximation of the binomial probability distribution when p, the probability of success is small and n, the number of trials is large. We set (other notation) and use the Poisson tables.

As a rule of thumb, the approximation will be good wherever both and

9 Normal Probability Distribution

9.1 Bell Curve

Bell curves show up throughout statistics. Diverse measurements such as diameters of seeds, lengths of fish fins, scores on the SAT and weights of individual sheets of a ream of paper all form bell curves when they are graphed. The general shape of all of these curves is the same. But all of these curves are different, because it is highly unlikely that any of them share the same mean or standard deviation. Bell curves with large standard deviations are wide, and bell curves with small standard deviations are skinny. Bell curves with larger means are shifted more to the right than those with smaller means.

Characteristics of the Normal probability distribution

1. The highest point on the normal curve is at the mean, which is also the median and mode of the distribution.

2. The normal probability curve is bell-shaped and symmetric, with the shape of the curve to the left of the mean a mirror image of the shape of the curve to the right of the mean.

3. The standard deviation determines the width of the curve. Larger values of the the standard deviation result in wider flatter curves, showing more dispersion in data.

4. The total area under the curve for the normal probability distribution is 1.