Statistics for Health Care Professionals

Distributions and Probabilities

Contributors: Ian Scott & Debbie Mazhindu
Editors: Ian Scott & Debbie Mazhindu
Book Title: Statistics for Health Care Professionals
Chapter Title: "Distributions and Probabilities"
Pub. Date: 2005
Access Date: October 17, 2013
Publishing Company: SAGE Publications, Ltd
City: London
Print ISBN: 9780761974765
Online ISBN: 9781849209960
DOI: http://dx.doi.org/10.4135/9781849209960.n9
Print pages: 81-93

http://dx.doi.org/10.4135/9781849209960.n9

[p. 81 ↓]

Distributions and Probabilities

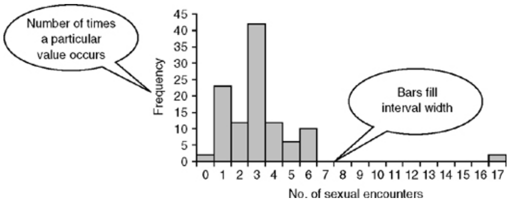
Areas of learning covered in this chapter

One of the more important concepts in statistics is the idea that numbers can be distributed in certain ways. What we mean by 'distributed' is the frequency of occurrence of particular numbers. For example, a data set of the number of sexual partners that each individual has during a lifetime could contain just the values 4 or 3; it's much more likely that it will be a mixture of different numbers, from high to low. The mixture is very important, because the way your numbers are mixed or distributed will largely determine the type of statistical test that you use. The easiest way to see the way in which data combinations are assembled is to plot them in a **frequency histogram** (Figure 9.1).

Frequency histograms

The frequency histogram is really a type of bar chart where the *y* axis is the frequency of occurrence of a particular case. On the *x* axis we have **[p. 82 ↓]** a scale that is bounded by the values of the lowest and the highest of the cases. In between are placed the values of the scale, using suitable intervals. A bar is drawn that fills the whole of each of the intervals being measured; the sides of the bars are parallel and the width of the bar is held constant.

Figure 9.1 Histogram showing the frequency distribution of the number of sexual partners for 109 women aged thirty who responded to a questionnaire distributed in the London borough of Southwark



This type of figure is normally used for variables that are recorded on an **interval** or **ratio** scale. If your data are interval or ratio scale, data plotting them in this manner must be one of your very first steps. This is because the distributions of data and numbers form the basis of many statistical tests. You will find that numbers are distributed in many ways. Some of the distributions have characteristics that can be exploited by researchers. One such distribution that we shall go on to explore is the **normal distribution**. This distribution forms the basis of many statistical tests, but first we need to discuss **probability**.

Using the data given in Chapter 5 for the walk-in clinic, make a histogram of the number of sexual encounters reported during a three-month period.

Probability and statistics

In Chapter 8 we introduced the notion of a statistical test and discussed the idea that statistical tests are performed because we want to know if the results we obtain are due to the experimental treatment or to chance. ‘Chance’ is a word that has the same meaning as the word ‘probability’. **[p. 83 ↓]** When we say, ‘What is the chance of patient x catching malaria?’ we could also say, ‘What is the probability?’ We could also say, ‘What is the likelihood?” These phrases all have the same meaning: we want to know if something is likely to happen or not. Of these terms, probability is used more by statisticians. It is given the symbol *P*. *P* is normally recorded as values between zero (not possible) and 1 (certainty). The probability that you will die is 1; the probability that you will meet Florence Nightingale is 0.

Often in the media and elsewhere you will see probabilities reported as percentages. In this case, the probability that you will die would be recorded as 100 per cent and of you meeting Florence Nightingale 0 per cent.

Box 9.1

Expressing probabilities

Probabilities can be expressed as:

To help put probability more into context, imagine that you are walking down a busy street on a weekend with your eyes closed, then you suddenly open your eyes. The probability of seeing a man will be 0.5 or 50 per cent. The probability of seeing a woman will also be 0.5 (or half of 100 per cent).

What does this probability mean? Half the time when you open your eyes you will see a women and the other half a man. How do we get from a value of 0.5 to a more meaningful fraction? Well, 0.5 can also be written as 0.5/1 (try putting 0.5 in your calculator and dividing it by 1). 0.5/1 is the same as 1/2 (try putting 1 in your calculator and dividing it by 2). In statistics we record a probability of 1/2 as *P* = 0.5. In statistical testing we need to be at least 95 per cent certain that the result we obtained is true (that is, not caused by sampling error). In statistics, however, we normally express this probability in terms of doubt. Thus rather than say we want to be 95 per cent certain we would say that we are 5 per cent uncertain that the result is true. This 5 per cent value is normally expressed as *P* = 0.05. Remember, *P* stands for probability.

[p. 84 ↓]

How are probabilities and distributions linked?

Say you have a bag of laundry with equal numbers of blue and pink towels. You cannot see into the bag. When you reach in and pull out a towel there are two possible outcomes: the towel will be pink or the towel will be blue.

Box 9.2

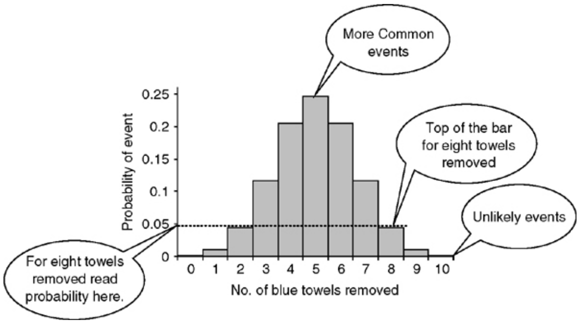
Practising your maths

Now if you pull out two towels the number of possible outcomes increases. It could be two blue towels in a row (BB) or a pink towel twice in a row (PP), or pink and then blue, or blue and then pink. In fact, we have four possible outcomes. So with just two events the number of outcomes and complexity are increasing. However, we can still predict what the probable combination of pink or blue towels might be.

If there are four outcomes (BB, PP, PB and BP) the chance of any one of these combinations occurring is 0.25 or 1/4. Two of these outcomes give you essentially the same combination of towels (PB, BP). Thus the chance of ending up with two pink towels in your hand is 1/4 and of two blue towels 1/4 and of one blue and one pink 1/2, that is, 1/4 +1/4.

Let's work through the obvious next step. You draw out three towels from your bag. The possible outcomes are PPP, BBB, PBB, BBP, PPB, BPP, BPB or PBP. There are eight of them. The probability of each outcome occurring is thus 1/8. We have four combinations: all blue, all pink, one pink and two blue, or two blue and one pink. So what is the probability of obtaining each of these combinations? Well, for PPP and BBB it is straightforward, as we have already said the probability of these outcomes is 1/8. There are three outcomes that give us one pink and two blue towels, so the probability of this combination is 1/8 +1/8 +1/8 = 3/8. There are **[p. 85 ↓]** also three outcomes that give us one blue and two pink towels, so the probability of this combination is 1/8 +1/8 +1/8 = 3/8.

Figure 9.2 Histogram showing the probability of removing various combinations of blue and pink towels from a bag containing of a number of blue and pink towels in equal proportion. Remember, because ten towels were pulled out in total, if eight blue towels are removed the other two in the set must be pink



You can see that as the number of events increases so the probability of each of the outcomes changes. As the number of events increases we need to use graphics – see Figure 9.2. It is possible to draw a histogram that shows the probability of obtaining certain combinations, given a certain number of events. Let's say you were pulling out ten of the towels. Now if you work through the maths you will find that here there are eleven combinations. Using Figure 9.2, a frequency histogram, it is possible to say just how unlucky you had been if, whilst searching for ten pink towels, you actually pulled out eight blue towels. (Find eight blue towels on the *x* axis, go up to the top of the bar and read off the value on the *y* axis.)

Box 9.3

Flipping a coin

If you flip a coin ten times what is the probability that a head will be turned up just three times? – use Figure 9.2 to help.

[p. 86 ↓]

The type of distribution shown here is called the **binomial distribution**. We have seen how it can be used to predict how rare or unusual certain events will be. This is the basis of statistical testing – asking the question ‘What is the probability (likelihood) of obtaining a result by chance?’ Clearly, in the example above, to pull out ten blue towels represents a rare event. We can also see that distributions of numbers and probabilities are linked.

Now the important thing about certain distributions is that they allow us to make predictions, and fortunately it just so happens that natural phenomena produce data sets that have a distribution that is similar to the one above. This distribution is known as the **normal** or **gausian** distribution. This distribution forms the basis of many of the most commonly used statistics. The type of statistics that relies on numbers being distributed in a certain way is called **parametric statistics**. We will now explore the normal distribution.

Having read this section, you should be aware of what is meant by the term ‘probability’ and the ways in which probabilities can be expressed.

You should be aware that it is possible, using a knowledge of how numbers are distributed, to make predictions.

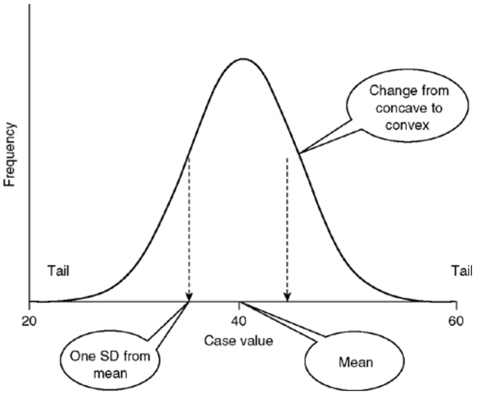
The normal distribution curve

Imagine that the intervals on the *x* axis were infinitely small. Instead of a bar chart with steps we would produce a curve, particularly if we didn't shade in the bars. The normal distribution would look like such a curve (Figure 9.3). The normal distribution has mathematical properties that allow us to make predictions, just like the histogram. Note also how it is drawn, in very much the same way as Figure 9.2, as if we had connected the tops of the bars with a line and then removed the bars.

As a defined distribution curve of numbers the normal distribution has certain properties. The first is very obvious, the curve is symmetrical – you could almost say it had a certain beauty; it is sometimes referred to as ‘bell-shaped’. The exact shape of the curve will depend on the standard deviation of the data. In fact it's worth remembering that the normal distribution is the construction of a mathematician's mind, so few data sets are likely to give a frequency distribution that exactly matches the normal curve.

[p. 87 ↓]

Figure 9.3 Normal distribution curve, shown here with a mean of 40 and a standard deviation of 8



The tails of a normally distributed curve (the rare values) tend to be short. Nevertheless, probably the most important feature of the normal distribution curve is that the point where the curve changes from being concave to convex (the point of inflection) is always one standard deviation (SD) away from the mean. The mean is always in the middle of the *x* axis. What this tells us is that the area enclosed by the boundaries of the mean plus one standard deviation and the mean minus one standard deviation is always a constant proportion of the total area, namely 68.27 per cent.

Box 9.4

The normal distribution

In normally distributed data one standard deviation either side of the mean always encloses 68.27 per cent of the data set.

[p. 88 ↓] If we were to move two standard deviations away from either side of the mean then we would encapsulate 95.44 per cent of the total area. If you were to take a large sample of patients’ arm lengths, you would expect that 68.27 per cent of your results would lie within ± 1 SD of the mean and that 95.44 per cent would lie within ± 2 SD of the mean.

Box 9.5

Making a prediction

You are interested in the number of Opsite dressings used on the average medical ward. You collect data from 102 wards. The data are normally distributed. How many wards will lie within ± 1 SD of the mean? *Hint* in normally distributed data 68.27 per cent of the data lie within ± 1 SD of the mean.

We have now introduced a means by which, if we know the mean and the standard deviation of a set of data, and we know that it is normally distributed, we can make predictions. We use this knowledge as the basis of what are often called **parametric statistics**.

Deviations from the normal distribution

Sometimes we find that the data we have collected do not fit the normal distribution. The best way to get a rough idea whether your data fit the distribution is to plot a frequency histogram. Some deviations have a particular shape and are given special names. The distribution shown in Figure 9.4 is called negatively skewed. This is because the mean lies to the left of the median (as you look at it). The distribution shown in Figure 9.5 is called positively skewed. This is because the mean lies to the right of the median.

Skewed data sets tend to occur when there are values which are much greater or lower than the rest. Thus the frequency histogram is not symmetrical, it's skewed. In these distributions the greater the difference between the mean and the median the greater the skew, or skewness.

It is also possible to have symmetrical distributions that do not conform to the normal distribution. The most common are random distributions and the regular, or under-dispersed, distribution. Examples of which are given below.

[p. 89 ↓]

Figure 9.4 Negatively skewed distribution

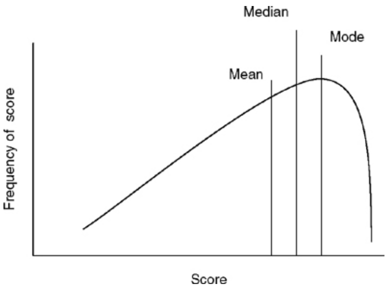
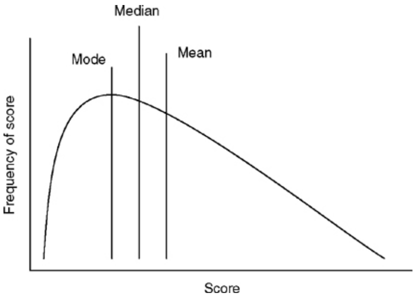


Figure 9.5 Positively skewed distribution

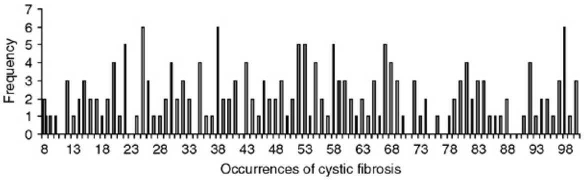


Random and clumped distributions

Data sets where the variance is roughly equal to the mean are referred to as *randomly distributed*. Random distribution tends to be uncommon. An example of a random distribution could be the number of occurrences of certain diseases within defined geographical areas. Such a distribution is shown in Figure 9.6 for the disease cystic fibrosis.

[p. 90 ↓]

Figure 9.6 Incidence of Cystic Fibrosis in the United Kingdom, by parish.



It should be noted that true randomness is comparatively uncommon and that the geographical distribution of many disease phenomena tends to have a clumped, or over-dispersed, distribution. We talk of disease outbreaks where we recognise that particular areas have a high incidence of a certain disease. In random phenomena we are saying that each event (an occurrence of cystic fibrosis) is unrelated to any other occurrence. If the distribution is clumped it suggests that the events are related, for example in the case of a contagious disease, or a disease that is triggered by some environmental factor. Clumped distributions tend to show a strong positive skew (the mean lies to the right of the median). Such a distribution is shown for the occurrence of AIDS cases across the metropolitan districts of the United States (Figure 9.7).

The last distribution to be aware of is the **regular distribution**. The regular distribution is really an extreme form of the normal distribution. In regular distributions the standard deviation is small in relation to the **[p. 91 ↓]** mean, that is, there is very little spread in the data set. An example could be records of the numbers of fingers and toes within a population. Obviously, people with less than eight fingers and ten toes are unusual, and so the distribution would be regular. If a normal distribution is shaped like that shown in Figure 9.8 it is said to show **kurtosis**. It can also be said to show kurtosis if the point of the curve is flattened.

Figure 9.7 AIDS cases per 100,000 population reported in 1999, by metropolitan area. Note that the point after the value of 2,080 has a value of 6,316 and so the x-axis has been truncated. This distribution shows a strong positive skew, and so the data are clumped. The mean is 359, median 162 and the mode 47.

Source: http://www.cdc.gov/hiv/stats/hasrlink.htm

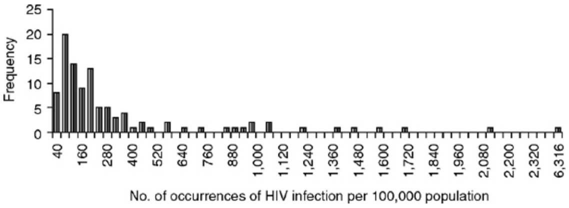
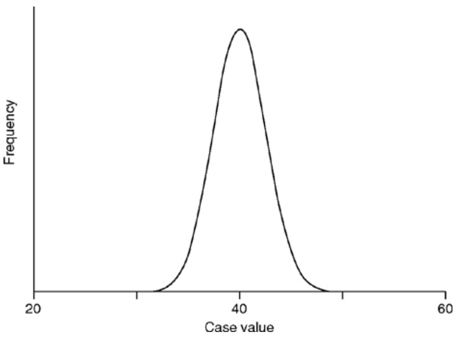


Figure 9.8 Regular distribution with a mean of 40 and a standard deviation of 2



It is important to distinguish between clumped and random distributions. The manner in which data are distributed is important, as it tells us about the fundamental properties we are studying and, as we have seen here, is very relevant to studies of the distribution and spread of disease (epidemiology). We also need to know how data are distributed before we embark on many statistical tests. We can distinguish between the different types of distribution using a statistical test that will be explained later.

Having read this chapter and completed the exercises, you should be familiar with the following ideas and words:

Exercises

http://dx.doi.org/10.4135/9781849209960.n9