Care in interpretation

How do we interpret the interval given in Example 4.6? It is an estimate of how long we expect people in general to be able to hold their breath – an estimate of the mean time all people, represented by the sample, can hold their breath. So we can be confident that this mean that applies to the general situation is somewhere between 45 and 53 seconds. Note that this does not predict how long an individual can hold their breath because it's estimating an overall mean. Predicting how long an individual can hold their breath will have a much greater margin of error. From Figure 3.13, we see that the values in this sample of data range from about 15 seconds to 109 seconds, with only about 19 people out of 102 with a holding breath time from 45 to 53 seconds. This emphasises that the interval (45, 53) is estimating the overall mean holding breath time – over all possible attempts for all people randomly represented by the sample.

MIND YOUR STEP !

A confidence interval for a mean is not an interval for individual values.

Whenever you see a confidence interval for a mean, no matter what the context or how it is obtained, remember that it is giving a lower estimate for the mean and an upper estimate for the mean, and that it is not providing any predictions for individual cases.

LINK ME

Chapters 8, 9

Figure 3.19 also shows the wide range of ages in the sample and demonstrates that age should be taken into account, with children probably considered separately to adults. Other variables were also observed in the study, and perhaps the investigators might wish to estimate the mean holding breath time for different combinations of these variables as well as for adults or children. Allowing for different combinations of values of other variables is among the topics of Chapters 8 and 9.

We interpret the 95% in the same way as for the confidence intervals in estimating proportions. If we took many samples of the same size in the same circumstances, and obtained a 95% confidence interval for the mean for all of them, we would expect about 95% of these intervals to include the general quantity – the overall mean holding breath time for all people represented by the sample.

4.6 Answering questions

The question posed at the beginning of this chapter is: who blinks more? This could be between male and female subjects, or it could be, as we shall see below, comparing the effects of male and female interviewers. Questions like this crop up all the time in investigations across all disciplines. In Chapter 3, we have seen which plots can be used to explore the data. Earlier in this chapter, we have seen examples of features of data and how to comment on them, and such comments go some way in answering such questions, but they are not really providing the types of answers investigations want. For example, the research question to be investigated might be 'can non-smokers hold their breath longer than smokers?' The researchers may want to allow for age, gender, fitness etc, but what they want to know is, 'in general, can non-smokers hold their breath longer than smokers?'

Statistics answers such questions by looking at how likely we were to get the data we have if the truth is that there is no difference in general between non-smokers and smokers in how long they can hold their breath. There are many different contexts and complications in using data to answer such questions, and many useful situations are covered in this book, especially in Chapters 5, 8, 9, 11, and 14, but we will see just one simple example here to introduce the basic concepts.

EXAMPLE 4.7

Comparing male and female reaction times

In an experiment to investigate the effects of alcohol on reaction times of young adults, the reaction times of nine volunteers, four female and five male, with zero blood alcohol readings, were measured under the same conditions in a pilot experiment. The reaction times were in response to a prompt on a computer screen and were measured precisely by the computer program.

The observed reaction times for the females, in seconds, were 1.96, 1.84, 2.02, 2.26, and for the males were 1.75, 1.78, 1.72, 2.00, 1.84. Figure 4.7 below gives these data.

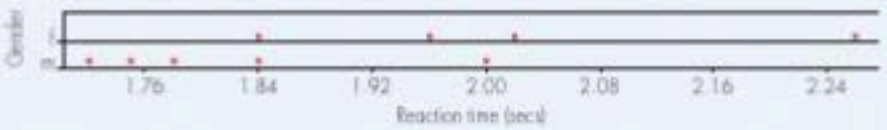

Figure 4.7 Dotplot of reaction times by gender

We see that three of the male times are less than all the female times and two of the female times are greater than all the male times, but there are very few observations and they are quite spread out. The sample means are 2.02 for the females and 1.818 for the males, with a difference of 0.202 seconds. Is 0.202 seconds a big enough difference in comparison with the sizes of the observations to say that female mean reaction times are more than male mean reaction times? Is it possible to say anything sensible in this situation?

If we assume there is no difference between males and females with respect to reaction times under these experimental conditions, then we have nine observations that are just assigned by chance to the categories of male and female. So, under the assumption of no difference, any of the possible groupings of these nine observations into four female and five male observations are equally likely. For example, another possible grouping is females: 1.96, 1.75, 1.78, 2.26; and males: 1.72, 1.84, 1.84, 2.00, 2.02, with a difference between the sample means of 0.0535 seconds.

We could consider all possible groupings of these nine observations into two groups of four and five, and see how many of these give a difference between female and male sample means of 0.202 or larger. The proportion of these would then give us the chance of getting the difference in our data or greater, if the truth is that there is no difference between male and female reaction times. This is because if there is no difference between males and females with respect to reaction times under these conditions, then all possible groupings of the nine values into two groups of four and five are equally likely.

How many groupings are there? This comes from an area of maths called combinations. If you have seen this area, you will know that the number of possible groupings is 9!/(4!5!) which is 126. To work through 126 possible groupings by hand is too much but a computer program can do it easily. For each possible grouping, it needs to calculate the average for the group of four and subtract the average for the group of five values and record this difference. Figure 4.8 below gives the histogram of all 126 possible differences of average female reaction time – average male reaction time.

It turns out that there are five of these differences that are at least as great as 0.202, so the chance of getting a difference between female and male average reaction times (for these sample sizes) at least as great as in our data, if the truth is that there's no difference between male and female reaction times under these conditions, is 5/126 = 0.04. A 4% chance is quite a small chance, so we can justify saying that there is evidence in these data that the mean female reaction time in this experiment is greater than the mean male reaction time.

▶▶

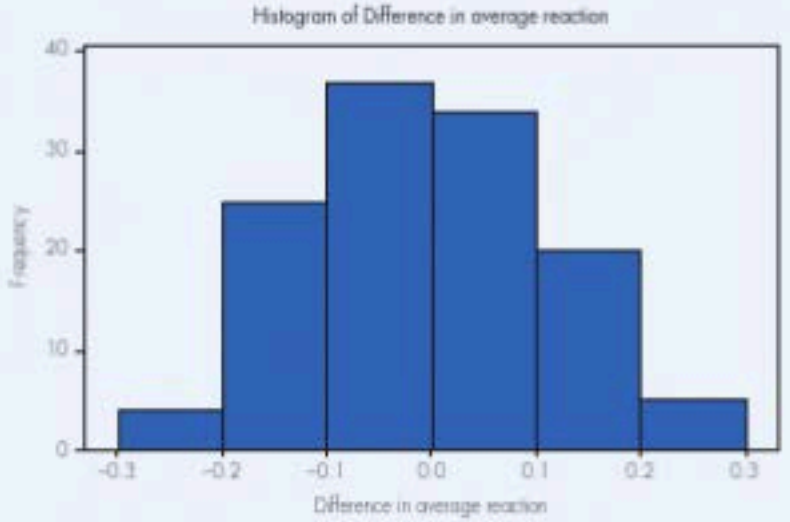

Figure 4.8 Histogram of all possible differences between female and male average reaction times

The general principle used in the above example is to assume that, on the average, there is no difference between female and male reaction times under these experimental conditions – that is, there's no difference in how quickly we expect a male or female to react. Then, under this assumption, we find the probability of getting at least the value of the difference in the sample means in our data. If this probability is small, then we had only a small chance of getting our observed difference or greater under our assumption of no difference, and there is evidence against our assumption. The smaller this probability, the greater the evidence.

The technique in the above example to find this probability makes no other assumptions and simply uses the fact that if males and females belong to the same group with respect to reaction times, then all possible divisions of the nine values into groups of four and five are equally likely, and the chance of getting our data or more extreme is the proportion of all possible divisions that give at least our observed difference. This technique is called a **randomisation test**.

Of course, nine observations is a tiny number of observations. For most practical datasets, the total number of possible divisions into two groups is too large for even comfortable computing. For example, if we have 10 males and five females, there are 3003 possible groupings of the 15 values. If we have 30 males and 20 females, there are 47 129 210 000 000 different possible groupings of the 50 values! It is of course impossible to do all these groupings, but because there are so many, we can take a fairly large random sample of these groupings and use the relative frequency of these groupings, for which the difference in sample means is at least our observed value, to estimate the proportion over all groupings. This procedure is also called a randomisation test. We will see this in the example at the end of this chapter.