VCAA-AMSI maths modules

A guide for teachers – Years 11 and 12

Probability & statistics

Hypothesis testing for means

DRAFT

Years 11 & 12

AMSI
AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Hypothesis testing for means – A guide for teachers (Years 11–12)

Professor Ian Gordon, University of Melbourne
Dr Sue Finch, University of Melbourne

Illustrations and web design: Catherine Tan

Authorised and co-published by the
Victorian Curriculum and Assessment Authority
Level 1, 2 Lonsdale Street
Melbourne VIC 3000

Australian Mathematical Sciences Institute
Building 161
The University of Melbourne
VIC 3010
Email: enquiries@amsi.org.au
Website: www.amsi.org.au

Assumed knowledge

The contents of the modules:

- *Inference for means*, which assumes
  - *Continuous probability distributions*;
  - *Random sampling*;
  - *Exponential and Normal distributions*;
  - *Inference for proportions*.

Motivation

- Why can we rely on random samples to test claims about population means?
- We get a different result every time we take a sample. How consistent are the sample results with a claim about the population mean?
- How can we quantify the consistency of the results from a sample with a claim about a population mean?

Some material from the *Inference for means* "Motivation" is equally applicable here, and in order to create a self-contained module, this material is repeated below (in the next four paragraphs).

In the *Random sampling* module students were introduced to random sampling from a variety of distributions. In that module, the distribution from which the samples were taken was consistently assumed to be known.

In practice, we typically do not know the underlying or parent distribution. We may wish to use a random sample to infer something about this parent distribution. An impression of how this might be possible was given in the *Random sampling* module, using mainly visual techniques.

One important inference in many different contexts is about the unknown population mean.

A random sample can be used to provide a *point estimate* of the unknown population mean μ: the sample mean $\bar{x}$ is an estimate of the population mean μ.

In the *Inference for means* module we discussed quantification of the uncertainty in the estimate of a population mean with a *confidence interval* for the (unknown) population mean. This is one important approach to inference about the unknown population mean.

There is a different approach to making inferences about the unknown population mean. In this second approach, we assert a possible value of the (unknown) population mean. The random sample provides an estimate of the unknown mean, and we evaluate the plausibility of the sample mean we have observed considering the value we postulated for the population mean. If our assertion about the population mean is correct, how likely is the result we have observed?

This provides methods for addressing questions such as these:

- A manufacturer claims the average weight of chocolate bars is 50 grams. How consistent are the bars produced, with this claim?
- The Australian Government's Average Quantity System specifies rules for deciding if a sample of products of a particular type is consistent with the quantity stated on the packaging. If a sample of jars of honey weighs 249 grams on average, what quantity could be stated on the packaging?
- The Australian Government's Department of Health recommends that parents of children aged 5 to 12 years "limit use of electronic media for entertainment to no more than two hours a day". Is there evidence that the average amount of electronic media use is consistent with the upper limit of two hours per day?

Content

- Hypothesis testing for a population mean for a sample drawn from a normal distribution of known variance or for a large sample, including:
  - P-values for hypothesis testing related to the mean.
  - Formulation of a null hypothesis and an alternative hypothesis.
  - Errors in hypothesis testing.

Confidence intervals and hypothesis tests

In the module on *Inference for means*, the idea of confidence intervals is explained. This is one of the ways that we can express uncertainty about an estimated, unknown parameter.

This module deals with the other main way that we express an inference about an unknown parameter: hypothesis testing.

As we shall see, in any of the contexts we consider, we could work out a confidence interval, or we could carry out a hypothesis test. This raises obvious questions of principle and practice. Should we prefer one or the other? Are the two connected somehow, as intuition suggest they must be? If so, what is the connection?

These are all important questions and we will deal them later.

A useful practical observation to make is that in the reporting of statistical inferences, both approaches are common and likely to be seen, often regarding the same parameter in a particular application. That is, a study might report a 95% confidence interval for an unknown population parameter, and also provide the results of a hypothesis test about that parameter.

Since the two approaches are connected, it is not surprising that the underlying ideas and results from probability that support hypothesis testing are the same as those used for confidence intervals. This is an important insight, as it means that there is not a whole new structure to learn about: rather, it is a different way of thinking about inference that uses the same underlying structure. This structure is extensively dealt with in the module on *Inference for means*, and that should be understood thoroughly for the purposes of this module.

For completeness, we summarise these points in the next sub-section.

Review of the theory in inference for means

The theory from *Inference for means* that is required for this module is now briefly reviewed.

Throughout, we are concerned with inference about a population mean. This is only one of many contexts in which inferences are obtained, but it is a very important one.

We consider a random variable X with unknown mean μ; this characterises a population of interest, so that μ is the mean of the population.

A **random sample** "on X" of size n is defined to be n random variables $X_1, X_2, \ldots, X_n$,

which are mutually independent, and have the same distribution as X.

The distribution of X is the underlying or "parent" distribution, producing the random sample.

Recall the important features of such a random sample:

- Any single element of the random sample, X_i, has a distribution that is the same as the distribution of X. So the chance that X_i takes any particular value is determined by the shape and pattern of the distribution of X.
- There is variation between different random samples of size n from the same underlying population distribution.
- If we take a very large random sample from X, and draw a histogram of the sample, the shape of the histogram will tend to resemble the shape of the distribution of X.
- If n is small, the sample may be appear to be consistent with a number of different parent distributions.
- Independence between the X_is is a crucial feature: if the X_is are not independent then the features we discuss here may not apply, and often will not apply.

We define the **sample mean** $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$.

From an actual random sample of size of the random variable X, we have an actual observation, $\bar{x} = \frac{\sum x}{n}$, of the sample mean, called a **point estimate** of μ.

We distinguish notationally and conceptually between the the random variable $\bar{X}$ and its corresponding observed value $\bar{x}$, we refer to the first of these as the **estimator** $\bar{X}$, and the observed value as the **estimate** $\bar{x}$.

In summary: the sample mean $\bar{X}$ is a random variable, with its own distribution.

Two general results for the distribution of $\bar{X}$ were proved in the *Inference for means* module. For a sample mean $\bar{X}$ based on a random sample of size n on X,

- $\mathbb{E}(\bar{X}) = \mu$;
- $\text{var}(\bar{X}) = \frac{\sigma^2}{n}$, and $\text{sd}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$.

This means that the distribution of $\bar{X}$ is centred around the unknown μ, and the spread of the distribution gets smaller as the sample size n increases. Both of these are desirable features. And these two results are true regardless of the shape of the parent distribution (of X), and for any sample size n.

What about the *shape* of the distribution of $\bar{X}$?

Firstly, there is a special case. When the parent distribution of X is itself Normal, the

distribution of $\bar{X}$ is itself Normal; specifically, $\bar{X} \overset{\text{d}}{=} \text{N}(\mu, \frac{\sigma^2}{n})$.

1

This exercise revises material in Inference for means.

Suppose that when adults go on a particular weight loss treatment, the amount of weight they lose, X, has a Normal distribution with mean μ kg and standard deviation 4 kg. That is, $X \overset{\text{d}}{=} \text{N}(\mu, 4^2)$; remember that the second figure in the brackets is the variance, not the standard deviation, so we sometimes express it in this way as a reminder.

A random sample of n people receiving this treatment is obtained. Find the probability that the sample mean, $\bar{X}$, is within 1 kg of the true mean, μ, for samples of the following sizes:

1. $n = 5$;
2. $n = 20$;
3. $n = 50$.

When the parent distribution of X is not a Normal distribution, there is no general result for the distribution of the sample mean.

However, due to the remarkable Central Limit Theorem, the distribution of the sample mean from any parent distribution becomes closer and closer to a Normal distribution, as the sample size increases. Readers of this module who are not familiar with this should read the *Inference for means* module, where this is dealt with extensively, considering samples from uniform distributions, exponential distributions and a strange, non-standard distribution. Because of its importance, we restate the Central Limit Theorem here:

> For large samples, the distribution of the sample mean is approximately Normal. If we have a random sample of size n from a parent distribution with mean μ and variance σ^2, then as n grows large the distribution of $\bar{X}$, the sample mean, tends to a Normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

As the averages from any shape of distribution tend to have a Normal distribution, provided the sample size is large enough, we do not need information about the parent distribution of the data to describe the properties of the distribution of sample means.

Finally, we note the useful standardisation that occurs and can be used to find probabilities for sample means.

For a random sample of size n from a Normal distribution,

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \stackrel{\mathrm{d}}{=} \mathrm{N}(0,1).$$

Under the specific conditions of sampling from a Normal distribution, and only then, this result holds for any value of n.

The Central Limit Theorem gives the result when the distribution of X is not a Normal distribution: for a random sample of size n from any distribution with a finite mean μ and finite variance σ^2, for large n,

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \stackrel{\mathrm{d}}{\approx} \mathrm{N}(0,1).$$

In fact, as we saw in the module *Inference for means*, we can even go one step further. For large n, the sample standard deviation can be substituted for σ, and

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \stackrel{\mathrm{d}}{\approx} \mathrm{N}(0,1),$$

where S is the sample standard deviation.

Possible values of μ

In many research contexts, we may be curious to know about μ, the population mean. In *Inference for means*, this was addressed by obtaining an interval which contains μ with a specified level of confidence.

Here we address our curiosity about μ in a different way. We ask: "Could μ be equal to a particular number?" "Are the data consistent with $\mu = 75$?"

You may wonder why this question is not posed much more directly. The question could be asked this way: "Is $\mu = 75$?" Why don't we just use this simple question? The answer is equally simple: since 75 is a very specific value, we know — without any collection of data — that this is almost certainly *not* the value of μ. The value of μ in any example may be in an interval of the real line. Even if the true value of μ is in the vicinity of 75, it would be an utter fluke if it really was exactly that value. It might be 76.1, or 74.9, or 75.03, but it would be bizarre if it turned out to be exactly the queried value, that is, $\mu = 75.000\ldots$.

On the other hand, the data may be consistent with a range of values for μ, and it is possible that the data are consistent with $\mu = 75$, ...and also consistent with $\mu = 76.1$, and $\mu = 74.9$, and .... But, perhaps, *not* consistent with $\mu = 70$.

Interest in a particular value of μ in a population arises when there is a specification about the average value of a quantity. For regulatory purposes, for example, we might require that the stated quantity of a mass-produced food item, such as a can of tuna or a bar of chocolate, should be the average value of the population of such items. Another context is an engineering specification, for manufacturing purposes. We want the length of a component of a hydraulic pump to be 12 mm. We know that there will be variation among individual components, so they will not all be exactly 12 mm. But we do want them to be 12 mm on average; that is, we require $\mu = 12$.

It will usually be desirable also to think about the variation in the random variable, and require that to be small. Suppose that the chocolate bar is labelled as 100 g, and we get one which is 90 g. It would be small consolation to be told that the population average is 100 g, and receiving a bar of 90 g was just due to random variation: you were unlucky. Similarly, in the manufacturing context, small variation is important; specification limits define the acceptable range for the dimension of the component.

In statistical terms, this is saying that not only do we want to specify μ, we want to ask for a small value of σ.

In this module, however, the focus is on the first of these: the value of the population mean μ.

Cans of tuna

A particular brand of tuna has a stated amount of 95 g on the outside of the can. Suppose that the weight, X, of tuna in the can has a Normal distribution with mean $\mu = 95$ g and a standard deviation $\sigma = 1.2$ g. A random sample of 25 cans of this product is obtained, and the sample mean $\bar{x}$ is calculated. Are any of the following values for $\bar{x}$ surprising: $\bar{x} = 94.9$, $\bar{x} = 95.4$, $\bar{x} = 94.2$?

Solution

For the setting given, we know that the sample mean, $\bar{X}$, has a Normal distribution with mean 95 and variance $\frac{1.2^2}{25}$. That is, $\bar{X} \overset{\text{d}}{=} \text{N}(95, \frac{1.2^2}{25})$. Note that this means the standard deviation of $\bar{X}$ is $\text{sd}(\bar{X}) = 1.2/5 = 0.24$. A plot of this distribution is shown below, with the positions of the three sample means indicated.

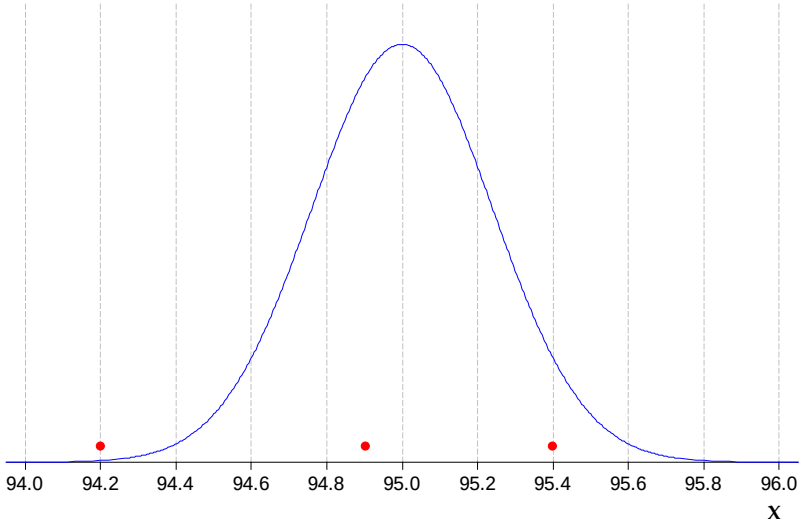

Figure 1: Distribution of the sample mean, $\bar{X}$, based on a random sample of weights of 25 cans of tuna from a Normally distributed population with mean 95 g and standard deviation 1.2 g. Also shown are three possible sample means.

Among the three sample means contemplated:

- $\bar{x} = 94.9$ is close to the expected value of the distribution (95 g) and hence is not surprising at all;
- $\bar{x} = 95.4$ is towards the right hand tail of the distribution of $\bar{X}$ but is not large enough to be very surprising;
- $\bar{x} = 94.2$ is a very unusual observation and surprisingly low.

How have we measured "surprising" in the canned tuna example? What makes us conclude that $\bar{x} = 94.2$ is a surprisingly low sample mean?

Note, firstly, that according to the given distribution of weights for the cans themselves, an *individual* can of weight 94.2 g would be not at all strange. The standard deviation, σ, for the weight distribution is 1.2 g, so a single can weight of 94.2 g is within one standard deviation of the mean (95 g).

But that is for a single can. It is a very different story when we consider the sample mean from a random sample of 25 cans. The distribution of $\bar{X}$, the sample mean, is much narrower; it has a standard deviation of $1.2/5 = 0.24$. Hence an observation of 94.2 g *for the sample mean* is more than three standard deviations below the mean.

In this example, we have implictly used the known distribution of the sample mean to produce a qualititative measure of surprise. We have looked at how far the sample mean

is from the population mean, and taken into account the distribution of $\bar{X}$ in Figure 1.

Can we go further? It is useful to produce a quantitative measure of surprise. Since we know the distribution of $\bar{X}$, it can be used to obtain a formal probability that reflects our level of surprise in the observed sample mean, $\bar{x}$. This probability is based on the distribution of $\bar{X}$, and the observed sample mean, which we will denote by $\bar{x}_{\text{obs}}$. We determine

$$\Pr\left(|\bar{X} - \mu| \geq |\bar{x}_{\text{obs}} - \mu|\right).$$

In words: the probability that the distance between $\bar{X}$ and μ is at least as big as the distance between the observed sample mean and μ.

Why "at least" as big? Suppose that on one day in a heat wave the maximum temperature is 44.3 degrees Celsius. This is very hot, and the media writes about it, because it is so extreme. There is interest in just how extreme such a day is, perhaps in terms of the historical record for our location. When this is done, we do not just ask how unusual are days with a maximum of exactly 44.3; we ask how rare are daily maximum temperatures of *at least* 44.3 degrees. In a conversation, someone mentions that "Jessica is really tall; there's only 1% of girls her age who are that tall". The same meaning applies: it is implied in this remark that 1% of girls who are Jessica's age are her height *or taller*.

In the same way, we always include more extreme values than that observed, in our assessment of how surprising or extreme we should regard a particular observed value of $\bar{x}$.

2

For the canned tuna example, find the probability that the sample mean $\bar{X}$ is at least as far away from $\mu = 95$ as each of the three observed sample means considered:

1. $\bar{x} = 94.9$;
2. $\bar{x} = 95.4$;
3. $\bar{x} = 94.2$.

So far, we have assumed that the population distribution from which we are sampling is Normal, and that we know the population standard deviation σ. These are two features of the population or model, and in general we are making inferences about the population (notably, the population mean μ). So it is reasonable to ask: can we avoid having to assume a Normal distribution and known σ?

As we saw in the module *Inference for means*, we can avoid making these assumptions, provided that we have a large enough sample size n. Due to the Central Limit Theorem,

we can use, as an approximation for large n, the following result:

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \overset{\text{d}}{\approx} \text{N}(0,1). \tag{1}$$

Cans of tuna (continued)

Suppose that we do not know that the distribution of weights of the individual cans is Normal, and we do not know the standard deviation, σ, of the distribution of weights. This is closer to what would happen in practice. We obtain a random sample of $n = 225$ cans. The average weight is found to be $\bar{x} = 94.8$ g and the standard deviation is $s = 1.4$ g, based on the sample. Is this an unusual sample, if the population mean is $\mu = 95$ g?

We obtain the result as follows. We require the probability that a sample mean from a sample of size $n = 225$ is at least as far from 95 g as the observed sample mean, $\bar{x} = 94.8$ g. The sample size is large, so we can use the result given in equation (1). We find that the required probability is

$$\Pr\left(\left|\frac{\bar{X}-95}{S/\sqrt{225}}\right| \geq \left|\frac{94.8-95}{1.4/\sqrt{225}}\right|\right) \approx \Pr(|Z| \geq 2.1429) = 0.0321,$$

where Z is the standard Normal distribution: $Z \overset{\text{d}}{=} \text{N}(0,1)$. So for a random sample of $n = 225$, when the standard deviation $s = 1.4$, a sample mean of 94.8 g is quite unusual.

In the canned tuna example, we have assumed that the population mean, $\mu = 95$ g, which is the quantity on the can's label. However, there are many situations where we would not want just to take the given population mean for granted. We might prefer to think of the population mean as unknown, which, in practice, it usually is. In that case, we regard assertions about the value of the population mean as *hypotheses* which may or may not be true.

Theories about μ

In the canned tuna example, instead of knowing or assuming that the value of the population mean was $\mu = 95$ g, we could instead test the theory or hypothesis that $\mu = 95$. All of the logic that we have seen so far applies to the method of testing a specific theory about μ.

"Weights and measures" is the area of government regulation that is concerned with fairness in the amounts of products sold, according to their labelled or advertised size or volume. In this context, one approach to testing whether the manufacturer of a particu-

lar product is violating the regulations, is to test the hypothesis that the population mean of their product is equal to the label value.

If cans of tuna have the label "95 g", we may test the hypothesis that $\mu = 95$. This is a very specific hypothesis. We then ask: are the data consistent with this value of μ? Or are they inconsistent with this hypothesised value? The way that these questions are answered is by the P-value.

When a particular hypothesis about a population parameter is tested, the hypothesis is known as the **null hypothesis**. There is a reason for this name. In this module we are looking at a simple context of testing hypotheses about a single population mean μ. We have suggested contexts in which this might be done. In research more generally, it is common to be testing hypothesis that arise from a *comparison* of populations. We might want to compare the average time that people survive on cancer treatments A and B, or the average score of students taught using methods A and B, or the average height of plants grown using treatments A and B.

If we compare two populations and we are especially interested in their means, μ_1 and μ_2, then a natural hypothesis test is that $\mu_1 = \mu_2$, which is equivalent to the hypothesis $\mu_1 - \mu_2 = 0$. This represents the pessimist's view of the world: that there is no difference between the populations means. And now we see where "null" comes from; *nullus* is the Latin word for 'none', or 'not any', and hence associated with zero. Here we are testing that the difference between the means is equal to zero. In other inference settings there are are different parameters but, again, we often test the hypothesis that represents an *absence* of effect, which is therefore aptly named the "null" hypothesis.

In contrast to the null hypothesis, we have the **alternative hypothesis**. Often, this is taken to be the denial of the null hypothesis. For example, if the null hypothesis is taken to be $\mu = 95$, the alternative hypothesis could be $\mu \neq 95$. This is known as a "two-sided" alternative hypothesis, because it allows for either of the two logically possible alternatives to $\mu = 95$, namely, $\mu < 95$ and $\mu > 95$. In this module we do not consider one-sided alternative hypotheses.

A null hypothesis is an assertion about a population or model, and a very specific assertion. In this module, we study the situation in which the null hypothesis is $\mu = \mu_0$, where μ_0 is a specified numerical value. Since μ is a parameter, we will always be uncertain about its value, and therefore unable to conclude, for sure, whether or not the null hypothesis is true. But a random sample from the relevant population does shed some light on the hypothesis, and we can sensibly ask how consistent the sample is, with the null hypothesis. The answer to this question is provided by the **P-value**.

P-value

The P-value is a probability. It is used for testing a hypothesis, often about an unknown population parameter. In this module, we are only considering the context of inference and hypothesis testing about the population mean μ, but there is a general definition for the P-value that applies to the testing of any null hypothesis.

This definition is as follows:

> The P-**value** for testing a null hypothesis is defined to be the probability of a result at least as extreme as that observed, given that the null hypothesis is true.

There are several important aspects to this definition.

- The probability is conditional on the null hypothesis. It assumes that the null hypothesis is true.
- (Therefore) the P-value cannot be construed as the probability that the null hypothesis is true.
- The P-value depends on the data (the "result"). To calculate it, we need a **test statistic**: a function of the data whose distribution is known, at least approximately, when the null hypothesis is true, *and* has a markedly different distribution when the null hypothesis is not true.
- The word "extreme" is important in the definition, since the way we calculate the P-value in a particular case is determined by what constitutes more "extreme" than that observed. The alternative hypothesis plays a key role in determining the correct interpretation of "at least as extreme".
- Small P-values are evidence against the null hypothesis, and the smaller the P-value, the stronger the evidence.
- On the other hand, large P-values indicate that we have data that are consistent with the null hypothesis. That does not mean that large P-values provide strong evidence that the null hypothesis is true. We explore this point further, later.

Interpreting the P-value

Suppose we wish to test the null hypothesis that the true mean weight for cans of tuna is 95 g. Again we assume that the standard deviation of the weight of the cans is known to be 1.2 g, and that the weights of the cans are Normally distributed. We take a random sample of 25 cans, and observe a sample mean $\bar{x}$ of 95.6 g. Figure 2 shows the distribution

of sample means for samples of size 25, centred at the null hypothesis. The probability of means more extreme than the observed mean of 95.6 g is shown in the grey areas of the distribution. Based on combining the tail areas, the P-value is 0.012. How do we think about and interpret this P-value?

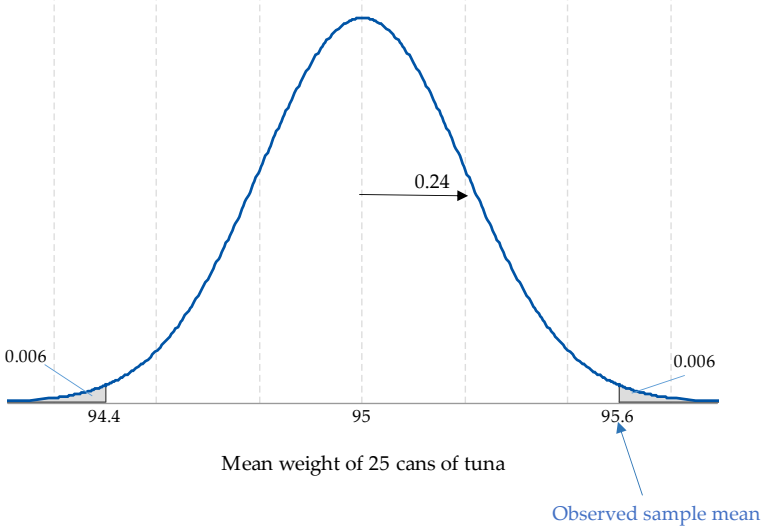

Figure 2: Distribution of the sample mean, $\bar{X}$, based on a random sample of 25 tins of canned tuna from a Normally distributed population with mean 95 g and standard deviation 1.2 g.

The mean weight of 95.6 g is quite surprising, if the true mean weight of the cans is 95 g. That is, a mean weight for 25 cans that is 0.6 g or more different from 95 g is a little unusual; the P-value quantifies this.

Connection between P-values and confidence intervals

We noted above that there are two broad approaches to statistical inference about an unknown parameter: confidence intervals and hypothesis tests. The module *Inference for means* covered the topic of a confidence interval for a population mean μ, and this module covers testing a null hypothesis regarding μ. Surely the two must be somehow connected?

They are indeed. A 95% confidence interval for μ can be expressed as an interval containing all values of μ, which, if tested as null hypotheses, would give a P-value ≥ 0.05. Put more simply, and without the precise details: the 95% confidence interval consists of the values of μ with which the data are consistent.

There are some unstated conditions here: we need to be testing against a two-sided al-

ternative hypothesis, and using a two-sided confidence interval (one-sided confidence intervals exist but have not been considered). Further, there is an important connection between the threshold of "0.05" in the statement above, and the confidence coefficient for the confidence interval, 95%: $0.05 = 1 - 0.95$.

3

For a random sample from a Normal population with mean μ and known standard deviation σ, show that if a test of the null hypothesis $\mu = \mu_0$ gives a P-value that is ≥ 0.05, then μ_0 must be in the 95% confidence interval.

It is also true that if $P < 0.05$, μ_0 is not in the 95% confidence interval.

In the canned tuna example with $n = 25$, $\sigma = 1.2$, Normality assumed, and an observed mean of $\bar{x} = 95.6$ g, we found that a test of $\mu = 95$ g gave a P-value of 0.012; hence $P < 0.05$. The 95% confidence interval for the true mean weight, based on the data, is from 95.1 g to 96.1 g, which does not include 95; the mean weight we have observed is not consistent with the true mean weight of 95 g.

Errors in hypothesis testing

The structure and logic of testing a null hypothesis needs to be kept in mind when interpreting the P-value. There are some common errors in interpretation and a potential error in the inference made that we discuss here.

An error in inference

Consider interpreting the scenario discussed in "Interpreting the P-value" (see Figure 2) as indicating that the true mean weight is not 95 g. Figure 2 reminds us of an important point. Observing a mean weight of 95.6 g or more extreme is relatively unlikely if the true mean is 95 g, but it is not impossible. Unusual sample means will arise from time to time. So interpreting extreme results (relative to the null hypothesis) as indicating that the true mean weight is *not* 95 g runs a (small) risk of being wrong. This is an error in the inference made. This might seem a little worrying, and of course, we do not know we've made this error; the true population mean is unknown to us. We simply know that this error is a possibility if we choose to believe the true mean weight is not 95 g. However in making inferences about population parameters we should not simply rely on interpreting the P-value in this way. Inference is best based on considering both the P-value and the confidence interval.

Here, for example, is how we might interpret the hypothesis test and confidence interval for the mean weight of the cans of tuna, for the data considered in Figure **??**. A mean weight of 95.6 g for a random sample of 25 cans of tuna was observed. The P-value for a hypothesis test of a population mean weight of 95 g was 0.012. This is the probability of observing a mean weight at a distance 0.6 g or more from a population mean of $\mu = 95$ g, if that is the true value of μ. The observation is surprising, if the true mean is 95 g. A 95% confidence interval suggests that plausible values of the true mean weight are between 95.1 g and 96.1 g.

P is not the probability that the null hypothesis is true

The P-value is a probability; it is probability that the distance between $\bar{X}$ and μ is at least as big as the distance between the observed sample mean and μ, if the μ corresponds to the null hypothesis. To find the probability we have assumed that the value of the population parameter corresponds to the null hypothesis. Sometimes the P-value is interpreted as quantifying the probability that the null hypothesis is correct or true. This is wrong. In our tuna example, it's wrong to say that "the probability that the true mean weight of the cans is 95 is 0.012". We don't assign a probability to the null hypothesis. The P-value arises from asking how unusual our observed sample mean is, given we assumed it arose from a distribution centred at the null hypothesis. Figure 2 reminds us of this.

$1 - P$ is not the probability that the null hypothesis is false

The mistake of interpreting the P-value as indicating the probability that the null hypothesis is true has a corollary, which is also an error in interpretation. Sometimes $1 - P$ is interpreted as measuring the probability that the null hypothesis is incorrect or false. Again, this is also wrong.

Large P-values do not prove the null hypothesis

50 g chocolate bars

A manufacturer wishes to claim that the average weight of chocolate bars is 50 grams. In order to make a check, a random sample of 40 bars is taken and weighed; the mean is 49.5 g, and the standard deviation is 2.5 g. A test of the null hypothesis that μ, the true mean weight, is 50 g is carried out. The P-value is 0.21. How can we interpret this P-value?

Is the manufacturer justified in claiming that the large P-value proves that the true mean weight is 50 g? Clearly not, but large P-values are sometimes interpreted as providing

'proof' that the null hypothesis is correct or true. The data observed are consistent with the null hypothesis, $\mu = 50g$, but the data are also consistent with other values of the population parameter. If, for example, the manufacturer had tested the null hypothesis that the true mean weight was 49.8 , the P-value is 0.45. If the null hypothesis was that the true mean weight was 50.1 g, the P-value is 0.13. These are also relatively large P-values, and it would be illogical to claim that there was 'proof' that these hypotheses were also true.

Again, there is value of interpreting the P-value with the confidence interval; the 95% confidence interval for the true mean weight was 48.7 g to 50.3 g. The null hypothesis of interest, $\mu = 50g$ is just one of the plausible values for the true mean that is included in the confidence interval.

P-values do not measure the importance of a result

A small P-value suggests we have found a surprising result, given we had assumed that the null hypothesis was true. Sometimes a result that is surprising (or statistically improbable) is interpreted as being important, and the P-value is interpreted as quantifying the importance of a result. Typically, small P-values are taken to reflect important findings and large P-values are not. This is (very) wrong. The importance of a statistical finding depends on many things. One obvious consideration is the point estimate. In our chocolate bar example, the estimated mean weight was 49.5 g. We need to ask the question: would it matter if the bars were underweight by an average of 0.5 g? This is an average of 1% underweight — such a mean difference might be not be important to the manufacturer but could be important to the consumer! Another important consideration is the confidence interval: 48.7 g to 50.3 g. The confidence interval quantifies the uncertainty in the estimate of the true population mean.

4

A researcher wished to investigate whether children's habits were consistent with the Australian Government's Department of Health recommendation to "limit use of electronic media for entertainment to no more than two hours a day". She asked parents of sixty 11-year old children, randomly chosen, to record the amount of time per day that their child spends using electronic media. She wanted to know: is there evidence that the average amount of electronic media use for 11-year olds is consistent with the upper limit of two hours per day?

The sample mean $\bar{x}$ is 3.1 hours per day, and the standard deviation s is 2.5 hours per day. The researcher carries out a test of the null hypothesis that true mean is 2 hours per day: $\mu = 2$. The P-value is 0.001; the confidence interval is 2.5 hours to 3.7 hours.

Based on this, which of the following are reasonable claims for the researcher to make about 11-year old children?

1 The probability that the average electronic media use of 11-year olds meets the Government recommendations is 0.001.
2 The probability that the study is wrong is very small.
3 The average electronic media use of 11-year olds is estimated to be more than one hour above the Government recommendation.
4 The result is not consistent with the Government recommendation, as $P = 0.001$.
5 The confidence interval suggests that average electronic media use of 11-year olds could be between 0.5 and 1.7 hours above the recommendation.

5

Consider the study above, but now the researcher asks parents of forty-five 6-year old children, randomly sampled. The sample mean $\bar{x}$ is 2.5 hours per day, and the standard deviation is 2.5 hours per day. The P-value is 0.18 (testing $\mu = 2$). The confidence interval is 1.8 hours to 3.2 hours.

Based on this, which of the following are reasonable claims for the researcher to make about 6-year old children?

1 The large P-value shows that average electronic media use of 6-year olds meets the Government recommendations.
2 The average electronic media use of 6-year olds is estimated to be half an hour above the Government recommendation.
3 The result is consistent with the Government recommendation, as $P = 0.18$.
4 The confidence interval suggests that average electronic media use of 6-year olds could be as much as 1.2 hours above the recommendation.
5 The Government need not be concerned about the electronic media use of 6-year olds as the P-value is 0.18.

Answers to exercises

Question 1

As the distribution of the weight gain is assumed to be Normal, the distribution of the sample mean, $\bar{X}$, is also Normal, and, specifically, $\bar{X} \overset{\text{d}}{=} \text{N}(\mu, \frac{4^2}{n})$. We want the probability that $\bar{X}$ is within 1 kg of μ, or, equivalently, the chance that $\bar{X}$ is betweeen $\mu - 1$ and $\mu + 1$.

That is, we want $\Pr(\mu-1<\bar{X}<\mu+1)$. We use the standardising steps to obtain the result:

$$\begin{aligned}
\Pr(\mu-1<\bar{X}<\mu+1) &= \Pr(-1<\bar{X}-\mu<1) && \text{(subtracting } \mu \text{ throughout)}\\
&= \Pr\left(\frac{-1}{4/\sqrt{n}}<\frac{\bar{X}-\mu}{4/\sqrt{n}}<\frac{1}{4/\sqrt{n}}\right) && \text{(dividing through by } 4/\sqrt{n})\\
&= \Pr\left(\frac{-1}{4/\sqrt{n}}<Z<\frac{1}{4/\sqrt{n}}\right) && \text{(where } Z \overset{\text{d}}{=} \text{N}(0,1)).
\end{aligned}$$

Using calculators or software, the probabilities are:

a $n=5:\Pr(-0.5590<Z<0.5590)=0.4238.$

b $n=20:\Pr(-1.1180<Z<1.1180)=0.7364.$

c $n=50:\Pr(-1.7678<Z<1.7678)=0.9229.$

What does this mean in practice? From a random sample of $n=5$, there is only a 42% chance that the sample mean is within 1 kg of the true mean μ. By increasing the sample size to 50, this becomes about 92%, much higher.

Question 2

For each value of $\bar{x}_{\text{obs}}$ considered, we need to find $\Pr\left(|\bar{X}-95|\geq|\bar{x}_{\text{obs}}-95|\right)$.

a For $\bar{x}_{\text{obs}}=94.9$, we obtain the following:

$$\begin{aligned}
\Pr(|\bar{X}-95|\geq|94.9-95|) &= \Pr(|\bar{X}-95|\geq 0.1)\\
&= \Pr\left(\left|\frac{\bar{X}-95}{1.2/5}\right|\geq\frac{0.1}{1.2/5}\right) && \text{(standardising)}\\
&= \Pr(|Z|>0.4167) && \text{(where } Z \overset{\text{d}}{=} \text{N}(0,1))\\
&= 0.6722.
\end{aligned}$$

b 0.0956

c 0.0009

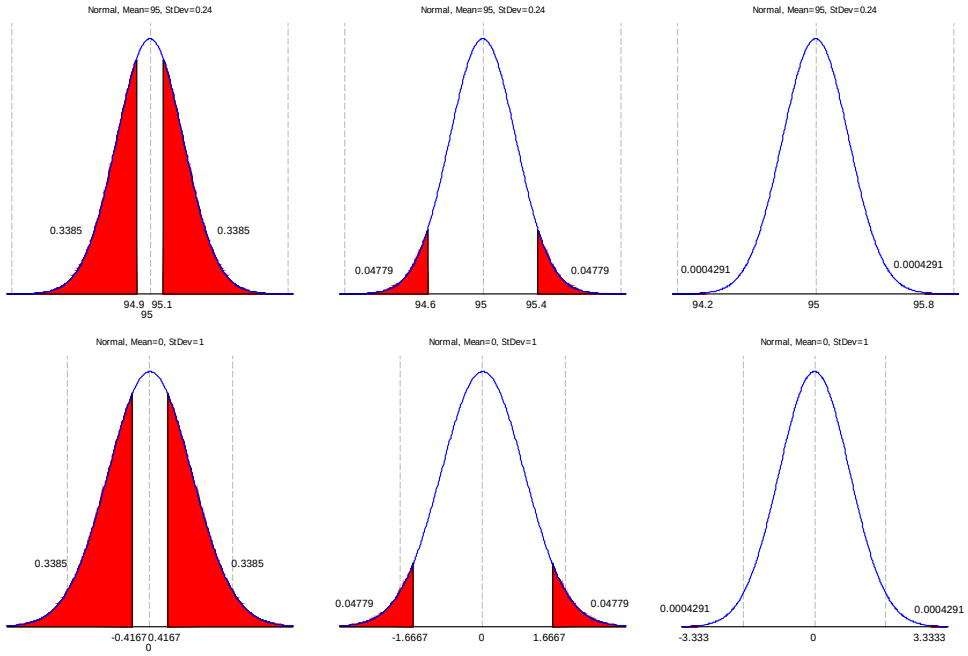

Figure 3: Distribution of the sample mean for Exercise 2, $\bar{X}$, shown in top row; corresponding distribution of the standard Normal, Z, in the bottom row. Probabilities for (a), (b) and (c) from left to right.

The probabilities required in each case correspond to the total of the two shaded areas in Figure 3. The two areas are equal, by symmetry, and the numbers shown on the probability density functions are the (equal) probabilities of each of these areas. For example, for (c), $\Pr(|Z| > 3.3333) = \Pr(Z < -3.3333) + \Pr(Z > 3.3333) = 0.0004291 + 0.0004291 = 0.0009$ (to four decimal places).

These answers confirm the visual impression gained from Figure 1. An observed sample mean of 94.9 g from a random sample of 25 cans is not unusual, while $\bar{x} = 95.4$ is slightly unusual. However, $\bar{x} = 94.2$ is very unusual indeed: samples such as this would occur only 9 times in 10,000 (in a repeated sampling sense).

Question 3

For the given context, $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \overset{\text{d}}{=} N(0,1)$ if the null hypothesis $\mu = \mu_0$ is true. If $P \geq 0.05$

for testing this null hypothesis, then $|z| \leq 1.96$.

$$\begin{aligned} |z| \leq 1.96 \quad &\Rightarrow \quad \left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| \leq 1.96 \\ &\Rightarrow \quad |\bar{x} - \mu_0| \leq 1.96 \frac{\sigma}{\sqrt{n}} \\ &\Rightarrow \quad \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \end{aligned}$$

Since the 95% confidence interval is $(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$, it follows that μ_0 is in this interval.

Further, the implications in this argument are all "if and only if", so the result applies in reverse. If a particular value of μ, μ_0, say, is in the 95% confidence interval for μ, then the P-value for testing $\mu = \mu_0$ must be at least 0.05.

Question 4 **a** *The probability that the average electronic media use of 11-year olds meets the Government recommendations is 0.001.*
This statement is incorrect; the P-value is not the probability that the null hypothesis is true.

b *The probability that the study is wrong is very small.*
This statement is incorrect; the P-value does not provide information about whether the results are right or wrong.

c *The average electronic media use of 11-year olds is estimated to be more than one hour above the Government recommendation.*
This is a correct statement about the observed mean number of hours of electronic media use.

d *The result is not consistent with the Government recommendation, as $P = 0.001$.*
This is a correct statement about the P-value.

e *The confidence interval suggests that average electronic media use of 11-year olds could be between 0.5 and 1.7 hours above the recommendation*
This statement provides an appropriate interpretation of the bounds of the confidence interval.

Question 5 **a** *The large P-value shows that average electronic media use of 6-year olds meets the Government recommendations.*
This statement is vague; large P-values do not mean that the null hypothesis is true.

b *The average electronic media use of 6-year olds is estimated to be half an hour above the Government recommendation.*

This is a correct statement about the observed mean number of hours of electronic media use.

c *The result is consistent with the Government recommendation, as* $P = 0.18$.
This is a correct statement about the P-value.

d *The confidence interval suggests that average electronic media use of 6-year olds could be as much as 1.2 hours above the recommendation.*
This statement provides an appropriate interpretation of the upper bound of the confidence interval.

e *The Government need not be concerned about the electronic media use of 6-year olds as the P-value is 0.18.*
This statement is ambiguous. It could be interpreted as implying that the null hypothesis is true because the P-value is large; this is incorrect. The P-value does not tell us about the 'truth' of the null hypothesis.


Years
11
&
12