Hypothesis tests

Contents

Introduction

Even though it is based on exactly the same theory of the sampling distribution, hypothesis testing takes a different approach to confidence intervals in coping with sampling error and its application for decision making.

We tend to use confidence intervals when we have no idea about the value of a population parameter. On the contrary, hypothesis testing is used when we do have some idea, claim, standard or experience against which we can compare the results from a random sample. Hypothesis testing assumes that the population parameter takes some value, usually determined by the researcher or from historical standards.

We set up two competing hypotheses:

- H_0: the null hypothesis about the population parameter
- H_1: the alternative hypothesis about the population parameter.

The basis for the procedure is to assume H_0 is true until we have highly convincing evidence to the contrary. The sample data are then used, in conjunction with sampling theory, to test whether or not the assumption or hypothesis about the population parameter is consistent with the sample evidence. If the sample is consistent with the null hypothesis, we do not reject it; but if the sample is inconsistent with the null hypothesis, we reject it in favour of the alternative hypothesis.

We all indirectly apply or meet hypothesis testing logic in our personal and professional lives.

- In a court of law, the accused is assumed to be innocent until proven guilty. Throughout the trial, evidence ('sample data') is provided to test the validity of the presumption of innocence. If the 'data' (evidence) is consistent with innocence, the accused is set free; if inconsistent with innocence, and based on the probabilities, then the accused is convicted.
- People frequently say 'I am not going to change my mind until you can convince me otherwise'.
- The government might refuse to approve use of a new cancer drug until research shows there is very little doubt that it does improve recovery from the disease.

Hypothesis testing is a formal statistical procedure for testing one hypothesis against another: innocence v guilt, not change v change, drug has no effect v drug has an effect.

Testing new drugs is an excellent example of how hypothesis testing is used. In testing whether a new drug improves recovery from cancer, our two hypotheses would be:

- The new drug does not improve recovery from cancer

- The new drug does improve recovery from cancer.

Another excellent example where hypothesis testing is used is on production lines in manufacturing companies. For example, consider a machine that is set to produce steel rods with a mean diameter of 120 mm and a standard deviation of 1 mm. Machines are not perfect and can wear out or malfunction and start to make the rods generally thicker (greater mean) or thinner (smaller mean). Thus, from time to time, say, on the hour, every hour, the machine needs to be inspected to see if it is operating to standard. At time of testing at a given hour, our two hypotheses would be:

- the machine is operating to standard, $\mu = 120$ mm.
- the machine is not operating to standard, $\mu \neq 120$ mm.

Say, a sample of 50 rods taken at the end of a particular hour reveals a sample average of 119.99 mm, then we may be confident that the machine was in order, as we would expect some slight variation of individual rods around the population mean. But if the sample revealed a mean of 115 mm, then the evidence would suggest that the hypothesis (that the machine is operating to standard) is not valid. What if the sample mean was 119.0? Should we stop the machine for an overhaul, and risk hours of lost production? Or do we allow the machine to continue to operate, and risk a whole batch of rods being produced at less than the desired diameter? It is a dilemma that the quality control officers would face, and need an objective procedure for deciding one way or the other: 'the machine is operating correctly with a mean of 120 mm and does not need to be stopped' v 'the machine is not operating correctly with a mean different to 120 mm and should be stopped'.

These tests and conclusions depend on sampling distributions. If the population average is 120 mm then it is not unreasonable, given the standard deviation, to expect that the sample mean of 50 elements could be 119.99 mm. It is much harder to accept that, given the machine is operating to standard (that is, 120 mm), a sample from the process could yield an average as low as 115 mm. In the latter case, our statistical calculations (which you can confirm later) would show this sample result is most unlikely if $\mu = 120$. We would conclude that the machine is not operating to standard and conclude that the diameter average has dropped below 120 mm, and proceed to stop the machine for an overhaul.

As in the confidence interval procedure, a high degree of confidence—or low degree of risk—can be associated with conclusions from hypothesis testing.

With interval estimates we attach a degree of confidence, typically 90%, 95%, 98%, or 99%. With hypothesis tests, we use a concept called the *level of significance*, denoted by α, and typically 10%, 5%, 2% or 1%. This level of significance can be viewed as a measure of risk: you will see it is the maximum probability that we are willing to allow of concluding H_0 is false when in fact it is true. As an example, what risk should society run of convicting someone who was in fact guilty? What risk are we willing to run to stop an important machine in our production process when in fact there is nothing wrong with it? Desirably, this risk should be zero. But ensuring zero risk may well increase the chance of not accepting H_1 when in fact H_1 is true. This balancing act is handled in statistics through probability.

The hypothesis test and confidence interval procedures can be extended to comparisons of two separate population parameters. For example, we may wish to test if the productivity of one worker at Plant A matches that of the workers at Plant B. Or whether there is any difference in the proportions of men and women in favour of a new workplace agreement.

Objectives

At the completion of this topic you should be able to:

- explain the basics of hypothesis testing
- explain how the concept of the sampling distribution can be applied to hypothesis tests
- conduct hypothesis tests on the population mean and population proportion for single sample situations
- analyse and explain the results of hypothesis testing, including how the *p*-value is used and how we draw conclusions and make decisions.

Hypothesis test

As explained in the introduction, we set up two competing hypotheses:

- H_0: the null hypothesis about the population parameter
- H_1: the alternative hypothesis about the population parameter.

The hypothesis test procedure is undertaken on the assumption that the hypothesised value for the population parameter is correct. We use the sampling distribution under the population parameter assumption to test the hypothesis. The basis for the procedure is to assume that H_0 is true until we have evidence to the contrary. The sample data are then used (in conjunction with sampling theory) to test whether or not the assumption or hypothesis about the population parameter is consistent with the sample evidence. We use *probability* as a proxy for *distance*. Thus, if our sample result has a very low chance (probability) of occurring (less than our required level of significance, α) if the null hypothesis were true, then we would say the sample result is 'distant' from the hypothesised

value of the parameter: too distant for the null hypothesis to be true. Instead we would claim that the alternative hypothesis is more likely.

Six steps in solving a hypothesis testing problem

Hypothesis testing is made easier if you adopt a consistent routine for testing.

Step 1: Set up H_0 and H_1

Note that the null hypothesis must have an equal sign in its formulation, and that the null and alternative hypotheses must between them cover all possibilities.

Decide on whether or not you are dealing with a quantitative (numerical) variable or a qualitative (categorical) variable (or attribute or proportion), as this affects the symbols used (μ, $\overline{X}$, etc.).

Decide on whether it is a one population/sample test, a two population/sample test or a three population/sample test.

Use key words or terms like 'equal to', 'different', 'no more than', 'at least', 'greater than' to help find the direction of the test and to help identify which must be H_0 and which must be H_1.

Step 2: Decide on the type of test

From H_0 and H_1, you can now determine what kind of test must be performed, that is:

- two tail test
- upper-tail test
- lower-tail test.

Step 3: Decide on a level of significance, α

This enables you to set critical value(s) and region(s).

The level of α chosen will depend on how serious it would be to commit a Type I error (that is, to reject H_0 when in fact it is true; that is, to accept H_1 when H_1 is in fact false).

The critical region (for a one-tail test) or regions (for a two-tail test) can now be specified in terms of *Z* or *t*. Find critical value(s) of *Z* or *t* and shade in the critical region(s) on the sampling distribution.

Step 4: Formulate the Decision Rule

This is the rule under which you will act once you take your sample.

This can be formulated in different (but equivalent) ways, for example, in terms of *Z* or *t* or the *p*-value.

Step 5: Analyse your sample

Steps 1 to 4 would normally be carried out before this step, so that the sample results do not affect the way in which the test is set up.

When the sample results become available, we calculate appropriate statistics and then the estimated standard error, the *Z* or *t* statistic and the *p*-value.

From these, apply either the critical value approach or the *p*-value approach.

Step 6: Draw your conclusion

You will have to decide whether to Reject H_0 or whether to Not Reject H_0.

You should be able to give a full conclusion in words (as well as in symbolic form).

The critical value approach and the *p*-value approach

In step 5 above, you will see that it says to apply either the critical value approach or the *p*-value. These are two alternative but interchangeable approaches to drawing your final conclusion: both give the same result. The reasoning behind these approaches is as follows.

From a practical point of view and our knowledge of sampling distributions, it is rare for a sample mean to be more than 3 standard deviations (or standard errors) from its population mean. That is, the standardised score or *Z* value should be within ± 3 in most cases. Thus, if H_0 is true, the sample mean should be within ± 3 standard errors of the mean assumed under H_0. If the sample mean is more than ± 3 standard errors away, it appears (on the basis of the probabilities) that H_0 is not true, and that the population mean is different to that assumed under H_0. Thus, a large *Z* score would be evidence to suggest that the hypothesised mean was incorrect.

As a normal rule in hypothesis testing, we don't use ± 3 standard errors as our cut-off, as this is very conservative in probability terms, corresponding to an α of about 0.27%. Instead we use a cut-off *Z* score of 1.645, 1.96, 2.33 or 2.575, corresponding to α levels of 10%, 5%, 2% or 1%, respectively.

- The critical values of *Z* (or *t*) are derived from the level of significance, α, assigned by the user.
- The *Z*-statistic (or *t*-statistic) is derived from our sample and estimates how many standard errors our sample statistic was from the parameter value assumed under H_0.
- The *p*-value is also derived from the sample, and measures the areas in the tail (or tails) beyond where the sample statistic fell.

With the *critical value approach*:

- if the *Z*-statistic (or *t*-statistic) is less (or equal) in magnitude than the critical value of *Z* or *t*, we do Not Reject H_0
- if the *Z*-statistic (or *t*-statistic) is greater in magnitude than the critical value of *Z* or *t*, we Reject H_0.

With the *p-value approach*:

- if the *p*-value is greater (or equal) to α, we do Not Reject H_0
- if the *p*-value is less than α, we Reject H_0.

Note that the two work in conjunction as follows:

- low *Z* (or *t*) statistic means a high *p*-value (indicating to Not Reject H_0)
- high *Z* (or *t*) statistic means a low *p*-value (indicating to Reject H_0).

Either approach is acceptable, and both give the same conclusion. With computer output we generally prefer to use the *p*-value approach, as it is a very quick and simple matter to check the *p*-value and compare it to your required α.

However, we need to provide some words of caution about the *p*-value approach. Before any hypothesis testing, the level of significance α should be considered and set by the researcher before the test is undertaken. The reason why it must be given some consideration is that it explicitly allows for the probability of making a Type I error, that is, the probability of rejecting H_0 when in fact H_0 is true. In setting up H_0 and H_1, it is important to weigh up the chances of Type I and Type II errors and to examine the consequences and associated risks of such errors.

The value of the *p*-value approach is that it provides a single figure on which to base a decision. However, always review that figure (the *p*-value) in terms of your pre-specified α of 5%, 10%, 2% or 1%.

In some problems, the conclusions of hypothesis tests can be altered by choosing a different level of significance. A small level of significance will lead to more '*Do not Reject* H_0' conclusions whereas large values of significance will lead to more rejections of H_0. For example, if a *p*-value for a particular problem was 0.0675 or 6.75%, we would not reject H_0 at α = 5% but would reject H_0 at α = 10%.

APPLICATION 7.1: LOAN APPLICATIONS

Three years ago, a large bank undertook a detailed study of the time taken to process loan applications. On the basis of the study, the bank determined that in that year the mean time taken to process an application was μ = 50 minutes.

This year, a review of the workplace arrangements was undertaken. The bank officers' union has requested a review of expected times for some tasks. Consequently, 36 loan applications in March were randomly selected to see if there had been any change over the three-year period. The result was a sample mean of 58 minutes and sample standard deviation of 12 minutes. At a 10% level of significance what do you conclude?

It makes sense to 'assume there has been no change over the three-year period, until we have evidence to the contrary'. The key words 'any change' indicate the direction of the test we need to perform. In this case, 'change' could mean an 'increase' or a 'decrease', hence we have to allow for movement in either direction, that is, we need to conduct a two-tail procedure test. As 'no change

from 50 minutes' has the 'equal' sign in it, and 'change from 50 minutes' does not have 'equal' in it, they become null and alternative hypotheses respectively. We can then set up our hypothesis test in our six steps as follows:

Step 1

H_0: $\mu = 50$. That is, we assume there has been no change in the mean value over three years.

H_1: $\mu \neq 50$. That is, there has been a change over the three years.

Step 2

Two-tail

Step 3

$\alpha = 10\%$, therefore CV of $t = \pm 1.6896$ (df = n – 1 = 35)

Step 4

Decision rule: If the sample mean $\overline{X}$ is more than 1.6896 standard errors away from 50 minutes, reject H_0.

Step 5

From our sample we have:

$$n = 36,\ \overline{X} = 58,\ \mathrm{s} = 12,\ \mathrm{s}_{\bar{x}} = \frac{\mathrm{s}}{\sqrt{n}} = {}^{12}\!/_{6} = 2,\ \alpha = 10\%$$

We apply the *t* test:

$$\text{t statistic} = \frac{\overline{X} - \mu_x}{\frac{\mathrm{s}}{\sqrt{n}}} = \frac{58 - 50}{2} = 4.00$$

Step 6

Conclusion: As the sample t-value is more than the critical t value, we reject the null hypothesis H_0 at the 10% significance level. Thus, we conclude that there has been a change in the mean time to process applications over the three-year period. In fact, we would conclude the mean has increased.

Management and the union can use this information as evidence that workplace arrangements for this task need adjustment and act accordingly.

If we reject H_0, then it makes sense to follow up with a confidence interval. In this case, given that we do reject H_0 and conclude the mean is now bigger than 50 minutes, an obvious question is how much bigger. Thus, we could construct a confidence interval estimate for the true figure.

One-tail tests and t tests for the population mean, μ

The above example demonstrated a two-tail tests, meaning we are testing for H_1 in both directions. For example, we may be interested in whether the average number of hours worked is any *different* now compared to 10 years ago. However, sometimes a test will indicate the direction in which the test should be performed, meaning we conduct what we call an 'upper-tail test' or a 'lower-tail test', for example, a drug manufacturer may wish to determine if a particular drug *lowers* the average cholesterol level

Before we look at these specifically, note the following.

Deciding on, and setting up, H_0 and H_1 can be tricky. Take the following into account:

- Decide on whether or not you are dealing with a numerical (quantitative) variable or a categorical variable (or attribute or proportion or qualitative variable). This will indicate whether you are dealing with μ or with p.
- Then, if you are performing a hypothesis test, decide on whether it is a one population/sample test or a two population/sample test.
- You still have to decide on the direction of the test. Keep the following in mind:
  - H_0 must contain an equal sign, either = $\geq$ or $\leq$
  - H_1 cannot contain an = sign, and must cover all other possibilities, either $\neq$ < or >

Key words/phrases can indicate the direction of a test and can point to either H_0 and H_1.

Examples of key words/phrases you should look for are:

- equal to (points to H_0)
- not equal to (points to H_1)
- more than (points to H_1)
- at least (points to H_0)
- exceeds (points to H_1)
- no change (points to H_0)
- difference (points to H_1)
- no more than (points to H_0)
- no difference (points to H_0).

Many of the above key words will be useful when you come across different types of hypothesis tests in your studies in this subject.

A one-tailed test is conducted when the researcher is concerned if the population parameter exceeds some specified value or is less than some specified value. Earlier we gave an example relating to a drug possibly lowering the average cholesterol level

As another example, consider a researcher wishing to know if the waste water from a factory manufacturing process meets pollution standards. In this case it is important to know if the population parameter *exceeds* the standard. We are not concerned if the true parameter is below the standard.

Another situation may be the life of products. If our product is advertised as lasting a certain number of hours, we must ensure that the average life is not lower than that advertised. We may not be overly concerned about exceeding the average to a small degree (from a regulations point of view).

In these situations, we will focus on one side of the parameter range. Accordingly, our hypothesis test is said to be one-sided.

Thus, you should now appreciate that the way to decide whether a test is one tail or two-tail is in the specification of either the null or the alternative hypotheses. If the alternative hypothesis is $\neq$, then the test is a two-tail test. If the alternative is in terms of > or <, then it is a one-tail test. In practical situations, if you wish to test that the population parameter is different to a specified value but do not know in which direction, use a two-tail test. If you have a clear idea of the direction, then use a one-tail test.

To reiterate, remember that:

- the null hypothesis must always contain an = (equals) sign and the alternative can never contain the equal sign
- H_0 and H_1 between them will cover all possibilities for the parameter (that is, they cover the whole range of the horizontal axis).

APPLICATION 7.2: TARGET MARKET INCOME

We are interested to know if the average income of our target market exceeds $35,000. Since our product is a luxury item we will only undertake production if the average income of our target market group exceeds $35,000.

Use a 1% level of significance to conduct a test to see if we should proceed with the product.

Step 1

Since we are mainly concerned if the target market income could exceed $35,000, we set up the hypotheses as below:

- H_0: $\mu \leq$ $35,000 (We assume the population mean does not exceed $35,000 and we do not produce the luxury product.)
- H_1: $\mu >$ $35,000. (The population mean is more than $35,000 and we do produce the product.)

Note that the conservative H_0 is that no production is undertaken. We will assume this until the evidence (our sample) suggests otherwise.

Step 2

Since we are only interested if the value of the population mean exceeds $35,000, the test will be conducted in the upper-tail.

Step 3

Given α = 1%, the critical region must be located in the top 1% of the sampling distribution. The critical value of t = 2.3646 (df = 99)

Step 4

The decision rule is based on the critical value *t* = 2.3646. If the observed *t* statistic for your sample exceeds 2.3646 then we will Reject H_0, that is, suggest that average target market income exceeds $35,000.

Step 5

A sample of 100 target market customers is eventually taken and reveals a sample mean = $\bar{X}$ = $35,875, s = $4500.

The next step is to calculate the observed *t* value which is the standardised value of the sample mean. We use the standardising formula:

$$t \text{ statistic} = \left(\bar{X} - \mu\right) / \left(\frac{s}{\sqrt{n}}\right)$$

Substitution of the relevant values from the data:

$$= (35{,}875 - 35{,}000) / \left(\frac{4{,}500}{\sqrt{100}}\right)$$

$$= 1.94$$

Step 6

In conclusion, on the basis of the test result, the observed *t* does not exceed the critical value of 2.3646. This leads us to 'Not reject H_0', that is, the sample evidence is consistent with the average target market income being $35,000 or less. In terms of a decision it would suggest we not launch the product. As in most situations, however, management should use this information as a guide—not as the sole determinant of actions.

Tests for the population proportion, π

The theoretical concerns for tests of the population proportion are similar in concept to those for the population mean. The basis for a proportion is a categorical variable and in large samples we can use the normal distribution to assess the sampling distribution. This allows us to construct confidence intervals and hypothesis tests in a way similar to the preceding sections.

We assume a value for the population proportion, π, and use the sample proportion, p, and its associated sampling distribution to conduct the test and reach conclusions.

As for numerical variables, tests for categorical variables can be two-tail, lower-tail or upper-tail tests.

As a final point, don't confuse the proportion '*p*' with a *p*-value. They are not the same. It may be even more confusing because conclusions of hypothesis tests for p can be determined by *p*-values.

APPLICATION 7.3: NEW PROCESS

The manager of the service department for a white goods manufacturer knows from past records that the proportion of units sold that were defective and thus brought in for warranty service was 5%. A new plant production process has been implemented and the service manager wishes to determine if the proportion of defective output has altered, either improved (decreased) or deteriorated (increased). Thus, the manager wishes to know if the proportion of defectives for the new process is different from 5%. A random sample of 400 items will be tested.

Step 1

Write down the two hypotheses in symbols and words.

The hypotheses are:

H_0: $\pi = 0.05$. The defective rate in the new process is also 5%

H_1: $\pi \neq 0.05$. The defective rate in the new process is not 5%.

Step 2

Here, we assume conservatively that the new process has the same defect rate as before. The test is a two-tail test.

Step 2

The hypothesis test can be conducted only if the normal distribution applies. Since $n\pi = 400 \times 0.05 = 20 \geq 5$ and $n(1 - \pi) = 400 \times 0.95 = 380 \geq 5$ (where π is the hypothesised proportion), we may use the normal distribution.

We will use $\alpha = 5\%$ to test the relevant hypothesis. Thus, the critical value of Z will be ±1.96.

Step 4

The decision rule proceeds as discussed previously. Thus, if our sample proportion falls more than 1.96 standard errors away from 5% we reject H_0. Otherwise do not reject H_0.

Step 5

A sample of 400 units of output is taken and reveals 30 defectives. The data given is: $n = 400$, $p = 30/400 = 0.075$ (or 7.5%).

The first step is to calculate the standard error, SE.

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.05x0.95}{400}} = \sqrt{0.0001188} = 0.0109 \text{ or } 1.09\%$$

The next step is to calculate the observed value of Z.

$$\text{Z statistic} = = \frac{\text{p - } \pi}{\text{SE}} = \frac{\text{p - } \pi}{\sigma_p} = \frac{\text{p - } \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.075 - 0.05}{0.0109} = \frac{0.025}{0.019} = 2.29$$

Step 6

Conclusion: As the *Z* statistic = 2.29 > 1.96 we will reject H_0. It appears that the percentage of defectives from the new process exceeds the historical level from the old process of 5%. Management may need to take remedial action as the proportion of defectives from the new process may be unacceptable.

Summary

Hypothesis testing is one of the two key inferential procedures that are used in statistical analysis. We investigated ways in which hypothesis tests enable us to draw conclusions about the population parameters; the mean, proportion, median and variance (standard deviation) for a single sample.

With hypothesis testing, we assume that the population parameter takes on a particular value. Using the theory of sampling distributions, we then check the sample evidence (sample statistic) to see if the evidence is consistent with our assumed hypothesis about the population parameter.

In general, you should examine the assumptions underlying the tests to see if they are satisfied. Most of these are satisfied approximately for large sample sizes. You should be careful when the sample sizes are small. The underlying assumptions may not be valid and the results meaningless, and possibly highly misleading. If you found the assumptions for testing the mean were not appropriate, you could resort to a test for the median as a test for central tendency.

Hypothesis tests are widely used in many fields. Typically, the results are obtained via computer software, which completes a test quickly and easily. You may come across instances where the results of hypothesis testing are summarised in reports and computer output. There, you may find appropriate *Z* or *t* values, or the *p*-values from the tests. You should be able to interpret the results and decide on their validity and the appropriate conclusions.

Further resources

Black, K 2008, *Business statistics for contemporary decision making*, 5th edn, Wiley, NJ.

Anderson, DR, Sweeney, DJ & Williams, TA 2008, *Statistics for business and economics*, 10th edn, South-Western Thomson Learning, Cincinnati.

Selvanathan, A, Selvanathan, S, Keller, G & Warrack, B 2006, *Australian business statistics*, 4th edn, Nelson Thomson Learning, Melbourne.