Statistical Inference

Eoghan O Leary

Inferential statistics uses a random sample of data taken from a population to describe and make inferences about that population. Inferential statistics are valuable when examination of each member of an entire population is not convenient or possible. For example, to measure the diameter of each nail that is manufactured in a factory is impractical. You can measure the diameters of a representative random sample of nails. You can use the information from the sample to make generalisations about the diameters of all of the nails.

The two main applications of inferential statistics that we will study involve the use of sample data to:

1. Estimate the value of a population proportion
2. Test some claim (hypothesis) about a population proportion

To begin, you will need to know what statisticians mean by a **population proportion** and a **sample proportion**. Suppose you want to know the percentage of adults in the USA who plan to vote for Donald Trump in an upcoming election. To find out, you contact 100 people in the USA and ask them who they plan to vote for. You then use the result to infer the percentage of adults in the whole of the USA who are likely vote for Donald Trump . Suppose the survey reveals that 49 out of the 100 people are planning on voting for Donald Trump. Then $\frac{49}{100}$, 0.49, 49 % is the sample proportion. We use the symbol $\hat{p}$ to denote the sample proportion. The population proportion is the proportion of the whole population (in this case, all of the voters in the USA) who plan to vote for Donald Trump. We use p to denote the population proportion. We can never find the exact value of p from $\hat{p}$. We can only ever estimate p.

Margin of Error

The formula for margin of error is:

$$E = \frac{1}{\sqrt{n}}$$

The margin of error is the maximum likely difference between the sample proportion $\hat{p}$ and the population proportion p.

Note: While the margin of error is associated with the OL course, students were required to use it in the HL Paper 2 2016.

Suppose for instance that a newspaper surveyed 1111 people and asked them who they would vote for in the next general election.

In this example, $n = 1111$, so

$$E = \frac{1}{\sqrt{n}} = \frac{1}{1111} = 0.03$$

So the margin of error relating to any results of the survey is 3%. But what does that mean?

Lets say the question asked of the 1111 was, "would you vote for Fine Gael in the next general election?". If 444 people say they would vote for Fine Gael, then

$$\hat{p} = \frac{444}{1111} = 0.3996 \approx 40\%$$

There is a 3% margin of error so I feel that in reality somewhere between 37% and 43% of people would actually vote for Fine Gael.

Mathematically, 'I can say with 95% confidence that the proportion of adults in the country who would vote for Fine Gael lies in the interval $0.37 < p < 0.43$.' This interval is called a confidence interval.

The confidence interval for the population proportion is using the margin of error is:

$$\hat{p} - \frac{1}{\sqrt{n}} < p < \hat{p} + \frac{1}{\sqrt{n}}$$

Example 1

A company wishes to estimate the proportion, p, of its employees who were absent for three days or more during the past year. A random sample of 20 employees was taken. Seven of the sample were absent for three days or more during the past year. Using the Margin of Error, construct a 95% confidence interval for p.

Step 1 : Calculate the sample proportion $\hat{p}$.

$$\hat{p} = \frac{7}{20} = 0.35$$

Step 2 : Find E, the margin of error.

$$E = \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{20}} = 0.2236$$

Step 3: Construct the confidence interval:

$$\hat{p} - \frac{1}{\sqrt{n}} < p < \hat{p} + \frac{1}{\sqrt{n}}$$
$$0.35 - 0.2236 < p < 0.35 + 0.2236$$
$$0.1264 < p < 0.5736$$

I can say with 95% confidence that the true proportion of its employees that were absent for three days or more lies between 12.64% and 57.36%.

The margin of error 0.2236 or 22.36% is quite big. It is important that students come to the understanding that margins of error that are this big are of little use. However, we can quite easily reduce the margin of error, simply by increasing the size of our sample.

Difference between hypothesis, hypothesis test and the null hypothesis

A hypothesis is a claim or statement about a property of the population.

A hypothesis test is a procedure for testing a claim about a population.

The null hypothesis H_0 is the **"hypothesis of no change"**.

Procedure for hypothesis testing for a population proportion

Step 1: State clearly the null hypothesis, H_0 , and the alternative hypothesis, H_1 .

Step 2: Calculate $\hat{p}$, the sample proportion.

Step 3: Set up a confidence interval for p, the population proportion.

- If the proportion for the population stated in the null hypothesis is within the confidence interval, then accept H_0, the null hypothesis.
- If the population proportion is outside the confidence interval, then reject the null hypothesis and accept H_1.

Example 2 : Testing a claim using a sample

A watchdog group is investigating a claim made by the CEO of a large multinational company. The CEO claimed that 80% of the company's 450,000 customers are satisfied with the service they receive. Using simple random sampling, the group surveyed 200 customers. Among the sampled customers, 146 said they were satisfied with the company's service. Based on these findings, can they reject the CEO's claim that 80% of customers are satisfied with the company's service?

Step 1: State the null and alternative hypothesis.

H_0 : (The null hypothesis) 80% of customers are satisfied with the service.

H_1 : (The alternative hypothesis) the satisfaction rating is not 80%.

Step 2: Calculate sample proportion

$$\hat{p} = \frac{146}{200} = 0.73$$

Step 3: Set up confidence interval

Margin of error $= \frac{1}{\sqrt{200}} = 0.071$

Confidence interval:

$$\hat{p} - \frac{1}{\sqrt{n}} < p < \hat{p} + \frac{1}{\sqrt{n}}$$
$$0.73 - 0.071 < p < 0.73 + 0.071$$
$$0.659 < p < 0.801$$

I can say with 95% confidence that the true proportion of its customers that are happy with the companies service lies between 65.9% and 80.1%.

The population proportion is within the confidence interval. Therefore, we fail to reject the null hypothesis that the satisfaction rating is 80%, and hence, they can not reject the CEO's claim.

Confidence interval for a population proportion

If many samples of the same size are taken from a population, each sample will produce a different (but similar) proportion. All these proportions form their own distribution called the sampling distribution of the proportion.

The standard error of the sampling distribution can be found on pg. 34 of *Formula and Tables*. It is called the standard error of the proportion.

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

The 95% confidence interval for a population proportion is given as:

$$\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}$$

This formula is most likely to be used in Higher Level Leaving Cert Maths.

Example 3

In a sample of 400 shops taken in 2012, it was discovered that 136 of them sold carpets at below the list prices which had been recommended by manufacturers.

(i) Calculate the 95% confidence limits for this estimate, and explain briefly what these mean.

The 95% confidence interval for a population proportion is given as

$$\hat{p} = \frac{136}{400} = 0.34$$

$$\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}$$

$$0.34 - 1.96\sqrt{\frac{0.34(1-0.34)}{400}} < p < 0.34 + 1.96\sqrt{\frac{0.34(1-0.34)}{400}}$$

$$0.34 - 0.046 < p < 0.34 + 0.046$$

$$0.294 < p < 0.386$$

I can say with 95% that the true proportion of shops who sold carpets below list prices lies between 29.4% and 38.6%.

Why use this formula rather that the margin of error? Let's return to our example about the watchdog to find out.

Example 4 : Testing a claim using the standard error of the proportion

A watchdog group is investigating a claim made by the CEO of a large multinational company. The CEO claimed that 80% of the company's 450,000 customers are satisfied with the service they receive. Using simple random sampling, the group surveyed 200 customers. Among the sampled customers, 146 said they were satisfied with the company's service. Based on these findings, can they reject the CEO's claim that 80% of customers are satisfied with the company's service?

Step 1: State the null and alternative hypothesis.

H_0 : (The null hypothesis) 80% of customers are satisfied with the service.

H_1 : (The alternative hypothesis) The satisfaction rating is not 80%.

Step 2: Calculate sample proportion

$$\hat{p} = \frac{146}{200} = 0.73$$

Step 3: Set up confidence interval

$$\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}$$

$$0.73 - 1.96\sqrt{\frac{0.73(1-0.73)}{200}} < p < \hat{p} + 1.96\sqrt{\frac{0.73(1-0.73)}{200}}$$

$$0.668 < p < 0.792$$

I can say with 95% confidence that the true proportion of its customers that are satisfied with the companies service lies between 66.8% and 79.2%

The population proportion is not within the confidence interval. Therefore, we reject the null hypothesis that the satisfaction rating is 80%, and hence, we **CAN** reject the CEO's claim.

So, finally to answer the question, why do we use this formula, rather than the margin of error formula? The answer is that you get a narrower interval but maintain the same degree of confidence. This makes the findings more useful to those who are using them.

Confidence interval for a population mean

In all of the examples above, we were dealing with a proportion, a fraction, decimal or percentage. Sometimes examining proportions would not be appropriate because of the data you are dealing with. In the examples that follow we will be dealing with the mean (average) value of the data. For example, if we are examining the number of words on each page of a novel.

To get the confidence interval for a population mean, we use the following formula.

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

Where $\bar{x}$ is the sample mean,
μ is the population mean,
σ is the standard deviation of the population,
and n is the sample size.

$\frac{\sigma}{\sqrt{n}}$ is referred to as the standard error of the mean.

Example 5

A random sample of 400 footballs was taken from a large consignment with unknown mean and standard deviation 15 grams.

The mean weight of the random sample was 81.4 grams.

Find a 95% confidence interval for the mean weight of the footballs in the consignment.

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

$$81.4 - 1.96\frac{15}{\sqrt{400}} < \mu < 81.4 + 1.96\frac{15}{\sqrt{400}}$$

$$79.93 < \mu < 82.97$$

I can say with 95% confidence that the true mean weight of the footballs in this consignment lies between 79.93g and 82.87g.

Note: We can use confidence intervals in a similar way as previously to complete hypothesis testing for a population mean.

Hypothesis testing for a population mean

However, we can be asked to find the "Test Statistic" instead. In this test, we speak of rejecting the null hypothesis 'at a certain level'. This 'certain level' is called the level of significance. The 5% level of significance is the one we deal with in our course.

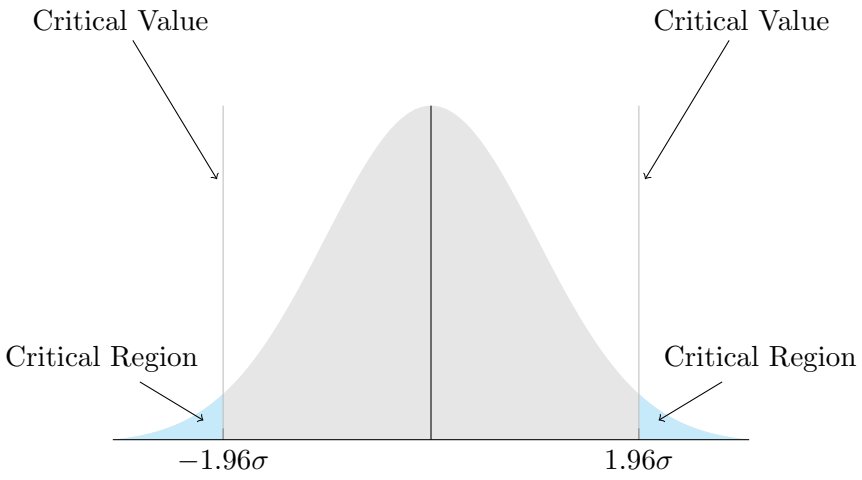
The 5% level of significance means that the result obtained is likely to occur on only 5 occasions out of 100. At the 5% level of significance, the set of values, $z > 1.96$ or $z < -1.96$, is known as the critical region and the boundaries of the critical region are called the critical values.

If the value of z is in the critical region we reject the null hypothesis and conclude that factors other than chance are involved.

The value of the z is called the **Test Statistic** and is given by:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Where $\bar{x}$ is the sample mean,
μ is the population mean,
σ is the standard deviation,
and n is the sample size.



In the normal distribution, 95% of the population lies within 1.96 standard deviations of the mean.

Example 6

Over the years, a horticulturist found that the mean yield from his plants was 1.83 kg per plant with a standard deviation of 0.35 kg per plant.

One year he planted 600 of a new variety and these yielded 1.87 kg per plant. At the 5% level of significance, test whether the mean yield from the new plants is different from his normal variety.

Step 1: State the null and alternative hypothesis.

H_0 : (The null hypothesis) The mean weight of the new variety is the same as the old variety.

H_1 : (The alternative hypothesis) The mean weight of the new variety is the different than that of the old variety.

Step 2: Construct the confidence interval

Or Step 2: We get the **test statistic** by converting it into standard units.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$z = \frac{1.87 - 1.83}{\frac{0.35}{\sqrt{600}}} = 2.797$$

$$2.797 > 1.96$$

$\implies$ The result is significant.

We reject the null hypothesis and accept the alternate hypothesis. We conclude that the new variety is different from the normal variety.

Using p-values

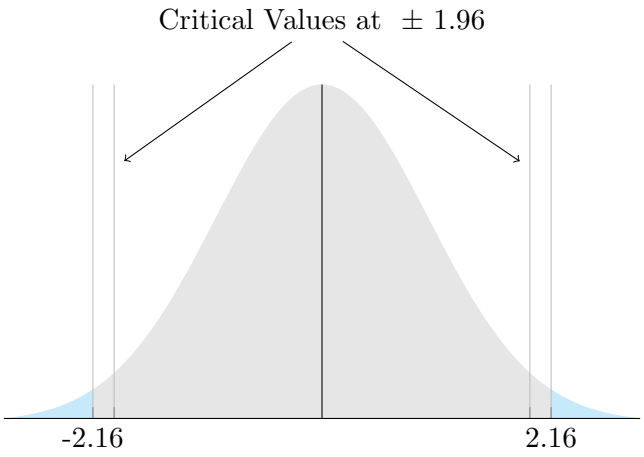
We have one more method of hypothesis testing on our course. That involves finding the p-value.

The p-value, or probability value, tells you how likely it is that your data could have occurred under the null hypothesis. It does this by calculating the likelihood of your **test statistic**.
The p-value tells you how often you would expect to see a test statistic as large or larger than the one calculated by your statistical test if the null hypothesis of that test was true.

The p-value is a proportion: if your p-value is 0.05, that means that 5% of the time you would see a test statistic at least as large as the one you found if the null hypothesis was true.
If the p-value is smaller than the significance level, we reject the null hypothesis and accept the alternative hypothesis.
Lets look at an example: Suppose we carry out a hypothesis test and find the test statistic to

be $z = 2.16$. Since 2.16 is greater than 1.96, we reject the null hypothesis at the 5% level of significance($\alpha = 0.05$).

Instead of comparing $z = 2.16$ with $z = 1.96$(and $z = -1.96$), we compare the total area of the two coloured regions below with the specific level of significance, $\alpha = 0.05$ (5%).



We use pages 36 and 37 of *Formula and Tables* to find the probability that $z \leq -2.16$ or $z \geq 2.16$. These are equal to the highlighted areas in the graph above.

$$
\begin{aligned}
&P(z \leq -2.16) + (z \geq 2.16) \\
&= 2P(z \geq 2.16) \\
&= 2(1 - P(z \leq 2.16)) \\
&= 2(1 - 0.9846) \\
&= 2(0.0154) \\
&= 0.0308
\end{aligned}
$$

The shaded areas above are referred to as the **p-value**, or probability value corresponding to the observed value of the test statistic.

The value 0.0308 found above is the p-value that corresponds to the test statistic $z = |2.16|$.

The p-value 0.0308 is interpreted as the **lowest level of significance** at which the null hypothesis could have been rejected.

With a test statistic of $z = 2.16$, we would certainly have rejected the null hypothesis at the specified level of significance ($\alpha = 0.05$). The p-value of 0.0308 gives us a **specific** or more precise level of significance. The **smaller** the p-value is, the **stronger** is the evidence against the H_0 provided by the data.

Test of significance using a p-value

1. Write down the null hypothesis and the alternative hypothesis.
2. Calculate the test statistic
3. Find the p-value that corresponds to the test statistic
4. If the p-value is greater that 0.05, the result is not significant and we do not reject the null hypothesis.

   If the p-value is less than or equal to 0.05, we reject the null hypothesis in favour of the alternative hypothesis.

Example 7

A random sample of 36 observations is to be taken from a distribution with standard deviation 10. In the past, the distribution has had a mean of 83, but it is believed that the mean may have changed.

When the sample was taken it was found to have a mean of 86.2.

(i) State the null and alternative hypothesis

(ii) Calculate the value of the test statistic

(iii) Calculate the p-value for the test statistic

(iv) Use the p-value to state if the result is significant at the 5% level of significance.

Explain your conclusion.

(i)

Null Hypothesis (H_0) : The mean is 83

Alternative Hypothesis (H_1) : The mean is not 83

(ii)

$$\text{Test Statistic } = z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{86.2 - 83}{\frac{10}{\sqrt{36}}} = 1.92$$

(iii)

Our p-value is:

$$\begin{aligned} 2P(z > 1.92) &= 2(1 - P(z < 1.92)) \\ &= 2(1 - 0.9726) \\ &= 2(0.0274) \\ &= 0.0548 \end{aligned}$$

(iv) p-value $= 0.0548 > 0.05$.

As the p value is not less than or equal to 0.05, the result is not significant, we do not reject the null hypothesis.