14 Statistics IV: Inferential Statistics

aims

Inferential statistics is the branch of statistics that uses probability and statistics to draw conclusions from data that are affected by random variation. To work on inferential statistics, we should be able to:

- ☐ Estimate the value of a population proportion
- ☐ Calculate the margin of error for a sample
- ☐ Construct a confidence interval
- ☐ Apply the empirical rule to confidence intervals
- ☐ Test a hypothesis about a population proportion

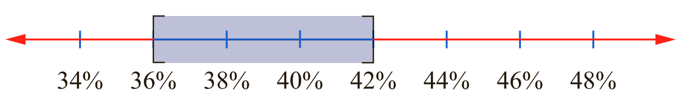
An introduction to confidence intervals

When results of surveys are reported in the media, they often include a statement like: **39% of respondents favour the government in the upcoming election. However, there is a margin of error of 3%.**

What this means is the people who carried out the survey are reasonably confident that in the real election, the percentage of votes for the government will be 39% $\pm$ 3%.

In other words, they are confident that if the election was held now, the government would receive somewhere between 36% and 42% of the vote.

We may show this confidence interval as:



An estimate from a survey should be treated with caution.

Sample proportion

We use $\hat{p}$ (pronounced 'p hat') to denote the sample proportion. $\hat{p}$ is the statistic that will be used to estimate the unknown population proportion, p.

$$p = \text{population proportion} \qquad \hat{p} = \text{sample proportion}$$

The population proportion, p, although unknown, is a fixed number. On the other hand, the sample proportion, $\hat{p}$, is a random variable and its value depends on chance.

Suppose we wanted to know the proportion (percentage) of people in Ireland who are left-handed. We randomly selected 400 people and found that 64 of them are left-handed.

$$\hat{p} = \frac{\text{Number of people in the sample who are left-handed}}{\text{The number of people sampled}} = \frac{64}{400} = 0{\cdot}16\ (16\%)$$

If 72 out of 400 people sampled were left-handed, then:

$$\hat{p} = \frac{72}{400} = 0 \cdot 18 \ (18\%)$$

Notice that the value of $\hat{p}$, the sample proportion, changes depending on the sample chosen. If the sample chosen is a good representation of the population, then $\hat{p}$, the sample proportion, will be a good estimate of the true population proportion, p.

Example

Write down the sample proportion, $\hat{p}$, in each of the following.

(i) In a market research survey, 187 people out of a random sample of 561 from a certain area said that they used a particular brand of toothpaste.

(ii) A survey was carried out on 7,410 randomly selected people and the result was that 6,175 were in favour of holding an election now.

(iii) An insurance company conducted a survey of 845 car crashes. It found that 338 of the crashes occurred within 6 kilometres of the driver's home.

Solution

Remember:

$$\hat{p} = \frac{\text{The number of 'successes' in the sample}}{\text{The sample size}}$$

(i) $\hat{p} = \frac{187}{561} = \frac{1}{3}$

(ii) $\hat{p} = \frac{6,175}{7,410} = \frac{5}{6}$

(iii) $\hat{p} = \frac{338}{845} = \frac{2}{5}$

Margin of error

We now look at the real business of statistics: to save people time and money! None of us want to do unnecessary work and statistics can tell us exactly how lazy we can afford to be. Our problem is that the collections of things in the world are so large, it's very difficult to get the information we want, e.g. voting populations, what percentage favours each candidate, what is the average length of sardines to fit in a can, what proportion of TVs will be defective.

We could answer questions like this by measuring every sardine in the world and doing some calculations. This method is not for statisticians: they want the easy way out.

Statisticians take **samples**. A sample is a relatively small subset of the total population, e.g. pollsters at election time.

An obvious question is: How big a sample do we have to take to get a meaningful result?

The answer turns out to involve $\frac{1}{\sqrt{n}}$, where n is the number of items in the sample.

key point

In statistics, the margin of error is a number that represents the accuracy of a survey.

The margin of error is denoted by E. The margin of error, at the 95% level of confidence, is given by:

If $n = 100$: $E = \frac{1}{\sqrt{100}} = 0{\cdot}1 = 10\%$

If $n = 400$: $E = \frac{1}{\sqrt{400}} = 0{\cdot}05 = 5\%$

If $n = 1{,}000$: $E = \frac{1}{\sqrt{1{,}000}} = 0{\cdot}0316227766 = 3{\cdot}16\%$ (correct to two decimal places)

If $n = 10{,}000$: $E = \frac{1}{\sqrt{10{,}000}} = 0{\cdot}01 = 1\%$

key point

Margin of error = $E = \frac{1}{\sqrt{n}}$

where n is the size of the sample.

Some notes on margin of error

- On our course, the margin of error is **always** at the 95% level of confidence.
- As the sample size increases the margin of error decreases.
- At the 95% level of confidence a sample of about
  (i) 80 has a margin of error approximately $\pm 11\%$
  (ii) 1,000 has a margin of error approximately $\pm 3{\cdot}2\%$.
- The size of the (original) population does not matter.
- If the sample size, n, is doubled (say 500 to 1,000) the margin of error, E, is **not** halved.
- The margin of error estimates how accurately the results of a poll reflect the 'true' feelings of the population.

Example

Calculate the margin of error at the 95% level of confidence when the sample size is:

(i) 25 (ii) 1,600 (iii) 2,000

Solution

Margin of error $= E = \dfrac{1}{\sqrt{n}}$

(i) If $n = 25$ then $E = \dfrac{1}{\sqrt{25}} = \dfrac{1}{5} = 0{\cdot}2 = 20\%$

(ii) If $n = 1{,}600$ then $E = \dfrac{1}{\sqrt{1{,}600}} = \dfrac{1}{40} = 0{\cdot}025 = 2{\cdot}5\%$

(iii) If $n = 2{,}000$ then $E = \dfrac{1}{\sqrt{2{,}000}} = 0{\cdot}0223606 = 2{\cdot}24\%$ (correct to two decimal places)

Example

At the 95% confidence level, calculate the sample size, n, to have a margin of error of:

(i) 1·25% (ii) 2·5%

Solution

Margin of error $= E = \dfrac{1}{\sqrt{n}}$

(i) $1{\cdot}25\% = 0{\cdot}0125$

$$\frac{1}{\sqrt{n}} = 0{\cdot}0125$$

$$1 = 0.0125\sqrt{n}$$

(multiply both sides by $\sqrt{n}$)

$$\frac{1}{0{\cdot}0125} = \sqrt{n}$$

(divide both sides by 0·0125)

(ii) $2{\cdot}5\% = 0{\cdot}025$

$$\frac{1}{\sqrt{n}} = 0{\cdot}025$$

$$1 = 0{\cdot}025\sqrt{n}$$

(multiply both sides by $\sqrt{n}$)

$$\frac{1}{0{\cdot}025} = \sqrt{n}$$

(divide both sides by 0·025)

$$\left(\frac{1}{0{\cdot}0125}\right)^n = n$$

(square both sides)

$6{,}400 = n$

$$\left(\frac{1}{0{\cdot}025}\right)^2 = n$$

(square both sides)

$1{,}600 = n$

key point

Notice in this example when we double the margin of error (from 1·25% to 2·5%) the sample size is **not** halved (from 6,400 to 1,600).

Confidence interval

The estimated proportion plus or minus its margin of error is called a confidence interval for the true proportion. The 95% confidence for a proportion is given by:

key point

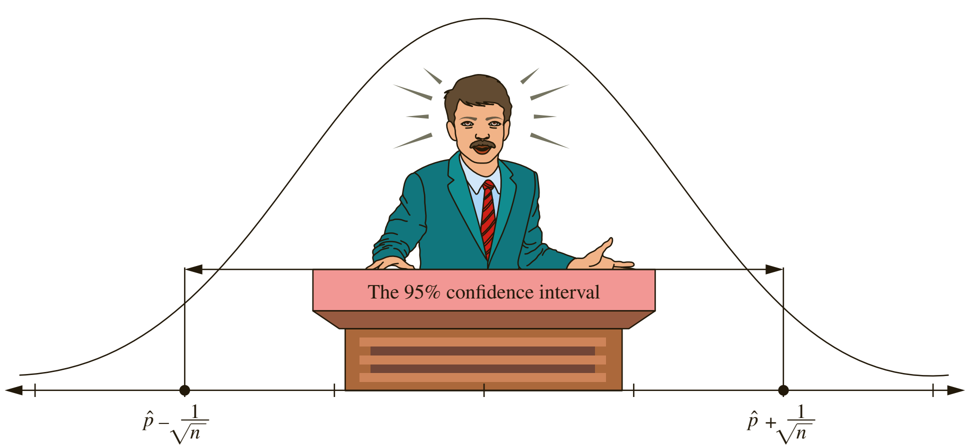
Sample proportion − margin of error ≤ true proportion ≤ sample proportion + margin of error

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

where n is the sample size, p is the population proportion and $\hat{p}$ is the sample proportion.

We can state with 95% confidence that the true population, p, lies inside this interval. What this means is that if the same population was surveyed on numerous occasions and the confidence interval was calculated, then about 95% of these confidence intervals would contain the true proportion and about 5% of these confidence intervals would not contain the true proportion.

The end points of the 95% confidence are given by $\hat{p} \pm \frac{1}{\sqrt{n}}$.



exam focus

What does '95% confidence' really mean?

What do we mean when we say we have 95% confidence that our interval contains the randomly selected value? Formally, what we mean is that '95% of randomly selected values will fall into the confidence interval'. This is correct but somewhat long-winded, so we usually say 'we are 95% confident that the (randomly) selected value lies in our interval'.

Our uncertainty is about whether the particular (randomly) selected value is one of the successful ones or one of the 5% that falls outside the interval.

Example

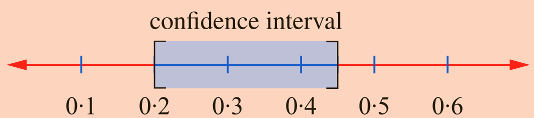
Show on separate diagrams the following confidence intervals.

(i) $0{\cdot}2 \leq p \leq 0{\cdot}45$

(ii) In a clinical study, 68% of patients reported relief after taking a new drug. The margin of error was calculated as 4%.

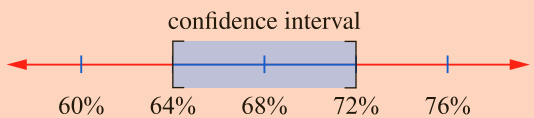
Solution

(i) $0{\cdot}2 \leq p \leq 0{\cdot}45$



(ii) $68\% - 4\% \leq p \leq 68\% + 4\%$

$64\% \leq p \leq 72\%$



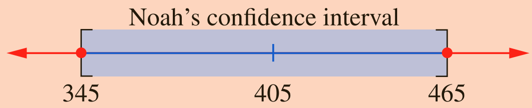
Example

Noah is sitting his Leaving Cert in June. After Christmas he made an estimate of how many CAO points he expected to get.

His estimate was 405 CAO points. Noah was not very confident of his estimate so he allowed each of his six subject grades go up or down by 10 points. Construct a confidence interval for Noah's CAO points estimate.

Solution

6 subjects by 10 points each $= 6 \times 10 = 60$ points
Noah's lowest estimate would be $405 - 60 = 345$
Noah's highest estimate would be $405 + 60 = 465$



Example

In a survey carried out in a large city, 225 households out of a random sample of 625 owned at least one pet. Calculate the 95% confidence interval for the proportion of households that own at least one pet.

Solution

Step 1: Calculate the sample proportion $= \hat{p} = \dfrac{225}{625} = \dfrac{9}{25} = 0{\cdot}36$

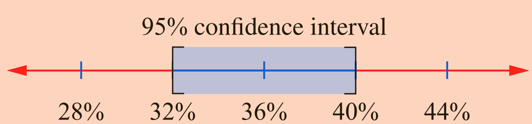
Step 2: At the 95% level of confidence, the margin of error $= \dfrac{1}{\sqrt{n}}$

$$= \frac{1}{\sqrt{625}} = \frac{1}{25} = 0{\cdot}04$$

Step 3: The 95% confidence interval for the proportion of households, p, that own at least one pet is:

$$\hat{p} - \frac{1}{\sqrt{n}} \leq p \leq \hat{p} + \frac{1}{\sqrt{n}}$$
$$0{\cdot}36 - 0{\cdot}04 \leq p \leq 0{\cdot}36 + 0{\cdot}04$$
$$0{\cdot}32 \leq p \leq 0{\cdot}4$$
$$32\% \leq p \leq 40\%$$

Thus, the 95% confidence level for the proportion of households that own at least one pet is between 32% and 40% inclusive. On a number line, it is:

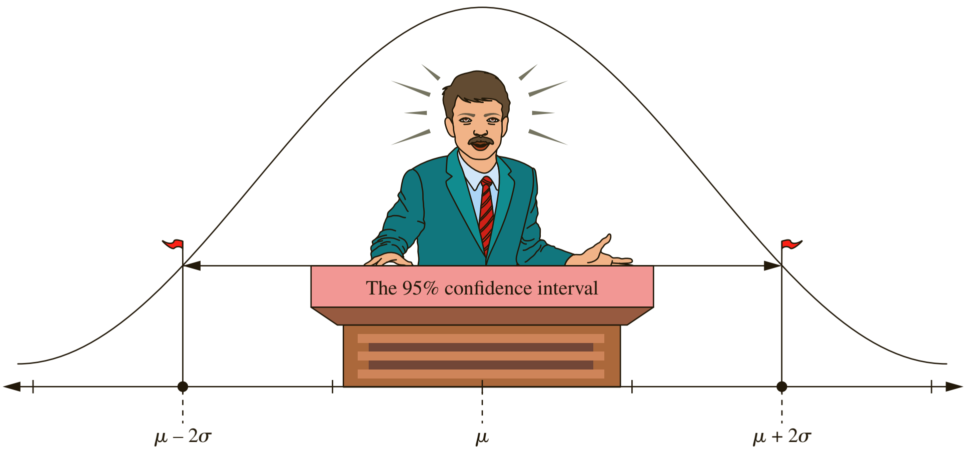


key point

When we show a 95% confidence interval, it means that 95% of the time, the true population proportion is inside the interval we have constructed.

The normal curve and confidence intervals

Given data that is approximately normal and using the mean (μ) and standard deviation (σ), the empirical rule tells us that about 95% of all randomly selected data points will be within $\pm 2\sigma$ from μ. That is to say, if we reach out 2 standard deviations on both sides from the mean, we are sure to 'trap' 95% of the data.



exam Q

A scientific expedition discovers a large colony of birds. The weights x kg of a random sample of 150 of these birds are measured and the following results obtained:

$$\sum x = 169{\cdot}5$$

(i) Calculate the mean, μ, of the weights of these birds.

(ii) Given that the standard deviation, σ, of the weights of these birds was 0.19, find using the empirical rule:

(a) A 68% confidence interval for an individual bird

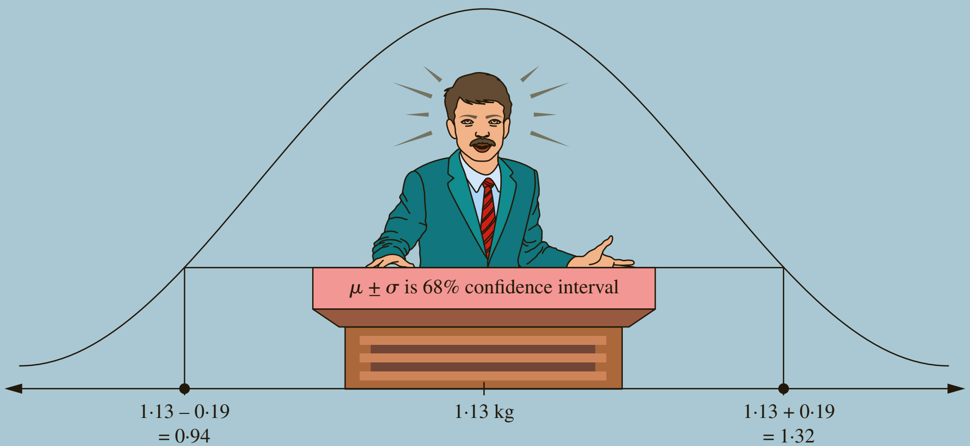
(b) A 95% confidence interval for an individual bird.

(iii) State, with a reason, whether or not your answers in part **(ii)** require the assumption that the weights are normally distributed.

Solution

(i) Mean $= \mu = \dfrac{\sum x}{n} = \dfrac{169{\cdot}5}{150} = 1{\cdot}13$

(ii) (a)



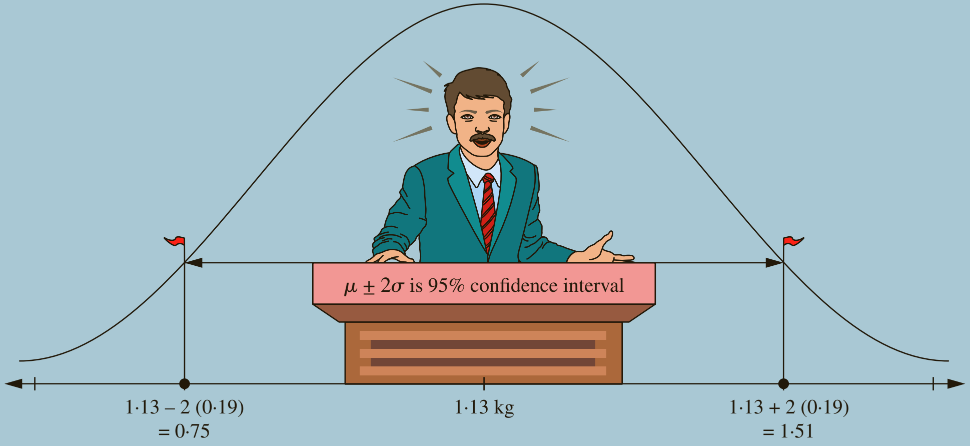
68% confidence interval given by mean ± 1 standard deviation

$= \mu \pm \sigma$

$= 1{\cdot}13 \pm 0{\cdot}19$

$= 0{\cdot}94$ kg to $1{\cdot}32$ kg

(b)



95% confidence interval given by mean ±2 standard deviations

$= \mu \pm 2\sigma$

$= 1{\cdot}13 \pm 2(0{\cdot}19)$

$= 0{\cdot}75$ kg to $1{\cdot}51$ kg

(iii) Yes, the answer requires the assumption that the weights are normally distributed in order to apply the empirical rule.

Example

The table below shows the ages of 40 golfers in a competition.

34	52	39	50	56	45	31	66	60	53
53	28	52	29	47	39	40	41	28	48
62	49	49	37	37	48	48	47	19	38
48	58	38	43	38	23	48	58	35	26

(i) Calculate, correct to one decimal place, the mean and standard deviation of the data.

(ii) Show that the empirical rule holds true for
(a) One standard deviation around the mean
(b) Two standard deviations around the mean.

Solution

(i) Using your calculator

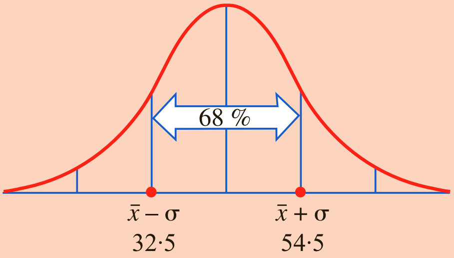
Mean $= \bar{x} = 43{\cdot}5$ and standard deviation $= \sigma = 11$.

(ii) (a) Upper range $=$ mean $+$ one standard deviation
$= 43{\cdot}5 + 11$
$= 54{\cdot}5$

Lower range $=$ mean $-$ one standard deviation
$= 43{\cdot}5 - 11$
$= 32{\cdot}5.$

Of the forty golfers, 27 (count them yourself!) are aged between 32·5 and 54·5 years that is $\frac{27}{40} \times 100 = 67{\cdot}5\%$

Hence, approximately 68% of the golfers are aged between 32·5 and 54·5 years.

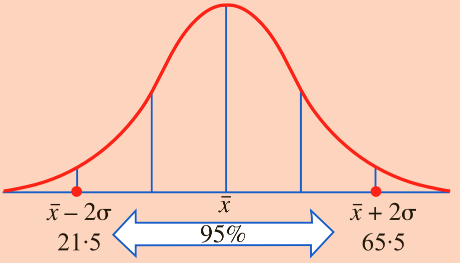


(b) Upper range $=$ mean $+$ two standard deviations
$= 43{\cdot}5 + 2(11)$
$= 65{\cdot}5$

Lower range $=$ mean $-$ two standard deviations
$= 43{\cdot}5 - 2(11)$
$= 21{\cdot}5$

Of the forty golfers, 38 (count them yourself!) are aged between 21·5 and 65·5 that is $\frac{38}{40} \times 100 = 95\%$

Hence approximately 95% of the golfers are aged between 21·5 and 65·5 years.



Hypothesis testing

A hypothesis is a statement (or theory) whose truth has yet to be proven or disproven. Examples of hypotheses:

- More than half the population is satisfied with EU membership
- Drinking fizzy drinks causes tooth decay
- The age of marriage has increased over the past 20 years.

NULL HYPOTHESIS

The statement being tested in a test of significance is called the **null hypothesis**. The test of significance is designed to assess the strength of the evidence against the null hypothesis. Usually the null hypothesis is a statement of no effect or no difference.
We abbreviate 'null hypothesis' as H_0.

Statistics help to make decisions

We can use statistics to reject or fail to reject claims.

1. Is global temperature increasing?
The null hypothesis, H_0, is that global temperature is not increasing, i.e. there is no difference in temperature. The alternative hypothesis, H_A, is that global temperature is increasing.
2. Is a new drug effective at treating HIV/AIDS?
The null hypothesis, H_0, is that the new drug is not effective. The alternative hypothesis, H_A, is that the new drug is effective.
3. Is a survey on left-handed people biased if it indicates that 24% of people are left-handed? The null hypothesis, H_0, is that 24% of people are left-handed, i.e. the survey is not biased. The alternative hypothesis, H_A, is that the survey is biased.

key point

Often the people investigating the data hope to reject H_0. They hope:

(i) their new drug is better than the old one

or

(ii) the new ad campaign is better than the original

or

(iii) the new machine is better than the existing one.

However, in statistics, it is essential that our attitude is one of skepticism. Until we are convinced otherwise, we accept H_0. In other words, we cling to the idea that there is no change, no improvement, no deterioration, no effect.

The reasoning behind hypothesis testing is that we usually prefer to think about getting things right rather than getting them wrong. A similar logic applies in trials by jury, where the defendant is considered innocent until it is shown otherwise.

In a courtroom, the null hypothesis is that the defendant did **not** commit a crime. A verdict of guilty means we reject the hypothesis, that is to say, the defendant committed a crime. However, a verdict of not guilty does not mean the defendant did not commit a crime, but simply that the case has not been proven.

Procedure for carrying out a hypothesis test

The procedure for carrying out a hypothesis test will involve the following steps:

1. Write down H_0, the null hypothesis, and H_A, the alternative hypothesis.

   For example, to test if a coin is biased if we get 6 heads in 10 tosses, we could formulate the following hypothesis:

   H_0: The coin is not biased

   H_A: The coin is biased.
2. Write down or calculate the sample proportion, $\hat{p}$.
3. Find the margin of error.
4. Write down the confidence interval for p, using

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

   In addition, we may illustrate the confidence interval with a diagram.
5. (i) If the value of the population proportion stated is within the confidence interval, we do not challenge H_0.

   (ii) If the value of the population proportion is outside the confidence interval, reject the null hypothesis, H_0, and accept H_A.

Example

RTÉ claims that 55% of all viewers watch the All-Ireland football final each year. An independent survey was carried out on 400 randomly selected viewers to see if the claim was true. The result of the survey was that 192 people were watching the All-Ireland football final.

(i) Calculate the margin of error.

(ii) State the null and alternative hypothesis.

(iii) Would you reject the null hypothesis according to this survey at the 95% level of confidence? Give a reason for your conclusion.

Solution

(i) Margin of error $= E = \dfrac{1}{\sqrt{n}} = \dfrac{1}{\sqrt{400}} = \dfrac{1}{20} = 0{\cdot}05 = 5\%.$

(ii) Null hypothesis, H_0: 55% of viewers watch the All-Ireland football final or $H_0: p = 0{\cdot}55$.
Alternative hypothesis, H_A: 55% of viewers do not watch the All-Ireland football final

or $H_A: p \neq 0{\cdot}55$.

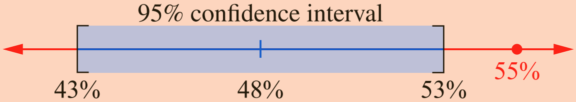
(iii) Find $\hat{p}$.

$$\hat{p} = \frac{192}{400} = 0{\cdot}48 = 48\%$$

Now use $\quad \hat{p} - \dfrac{1}{\sqrt{n}} \leq p \leq \hat{p} + \dfrac{1}{\sqrt{n}}$

$$48\% - 5\% \leq p \leq 48\% + 5\%$$

$$43\% \leq p \leq 53\%$$



Since the claim of 55% viewing the All-Ireland football final is outside the 95% confidence interval, we reject the null hypothesis, H_0.
The above work is the reason for my conclusion.

exam Q

FlyinAir airlines provides flights in Europe. Each month the company carries out a survey among 1,100 randomly selected passengers. The company repeatedly advertises that 75% of their customers are satisfied with their overall service. 803 of the sample stated they were satisfied with the overall service.

(i) State the null hypothesis and the alternative hypothesis.

(ii) Investigate at the 5% level of significance if the company was correct in saying that 75% of their customers were satisfied.

(iii) Clearly state your conclusion.

Solution

(i) Null hypothesis: The proportion of passengers who are satisfied with the service is unchanged at 75%

or $H_0 : p = 0{\cdot}75$.

Alternative hypothesis: The proportion of passengers who are satisfied with the service is changed, not 75%

or $H_A: p \neq 0{\cdot}75$.

(ii) Sample proportion $= \hat{p} = \dfrac{803}{1{,}100} = 0{\cdot}73 = 73\%$

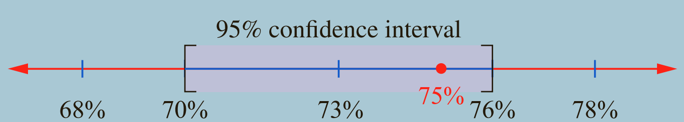
Margin of error $= E = \dfrac{1}{\sqrt{n}} = \dfrac{1}{\sqrt{1{,}100}} = 0{\cdot}03015 \doteqdot 3\%$

(iii) Confidence interval

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

$$73\% - 3\% \leq p \leq 73\% + 3\%$$

$$70\% \leq p \leq 76\%$$

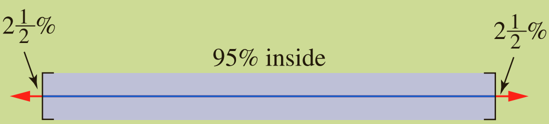


Since the claim that 75% of passengers are satisfied is inside the 95% confidence interval, we do not reject the null hypothesis.

Remember: Unless we have sufficient evidence to the contrary, we do not reject the null hypothesis. We do not challenge H_0. This is **not** the same as saying that we accept the claim.

key point

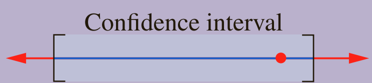
When working with levels of confidence (or levels of significance), statisticians can use percentages ambiguously. In particular, the 5% level of significance and the 95% level of confidence mean the same thing. That is to say, 5% of the time outside the confidence interval or 95% of the time inside the confidence interval.



exam focus

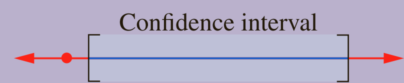
In the final analysis, testing the null hypothesis, H_0 simply involves a confidence interval and a red dot

Either



When the red dot is inside the confidence interval we fail to reject H_0

Or



When the red dot is outside the confidence interval we reject H_0