

Hypothesis testing for single populations

In the lecture on estimation, we have seen how we can guess the value of a parameter (population statistic) with a margin of error or devise an interval of values that includes it, within a level of confidence.

Hypothesis testing is, on the other hand, a process by which one hypothesis about the population parameter is accepted against another, with a probability of error.

1. Hypothesis test for a single population proportion

A London local council recently decided to make all council tax bills paid by direct debit only. This method of payment is cheaper for the council as it will save money on administration, they say. Some local residents were not happy as they feel they will lose control over their expenditure. The chief executive of the council claims that 78% of local council taxpayers are happy to pay by direct debit. Is the claim by the chief executive true?

To answer this question, we need to formulate two hypotheses:

The first is called the **null hypothesis** (denoted by the symbol H_0). It states that the percentage of residents happy to pay by direct debit is 78%. This means that if we survey the residents of the council, a percentage of 78% (or very close to it) will say they happy with the method of payment. In other words, you accept (in fact, you fail to reject) the null hypothesis.

The second is called the **alternative hypothesis** (denoted by the symbol H_1). It states that there is a real difference between the percentage of residents happy with the payment method, claimed by the chief executive, and the one obtained from the survey of residents. If the percentage of happy residents from the survey is very different to 78%, then you will conclude that the null hypothesis must be rejected, and that the alternative hypothesis is upheld.

In short, this is exactly what hypothesis testing is all about. However, the explanation outlined above raises several other questions.

The first one is, how close to 78% must our survey percentage be before we accept the claim by the chief executive (and therefore accept the null hypothesis)? If the survey percentage is 76%, is this considered close? We could argue that, even if the survey is random, it is not an exact replica of the whole population and that some differences are to be expected. In other words, where is the cut-off point beyond which you say this is too far from the claimed percentage?

The second question has to do with the risk of making a wrong conclusion. What if the percentage of local residents happy with the payment method is indeed 78% but our survey was collected impressions from mainly disgruntled residents (by luck), shows a percentage so low compared to 78% and we end up rejecting the null hypothesis. Can we evaluate the risk of making such mistakes?

We will try to answer these questions and show why it makes sense.

We know from sampling distribution, and in the case of proportion, that if we take all possible samples of the same size n , from a population of local residents with whom 78% are happy to pay by direct debit (a proportion of 0.78), then in 95% of such samples, the percentage of happy residents will fall somewhere between:

$$0.78 - 1.96 \times \sqrt{\frac{0.78 \times 0.22}{n}} \quad \text{and} \quad 0.78 + 1.96 \times \sqrt{\frac{0.78 \times 0.22}{n}}$$

Everything in hypothesis testing is based on this result.

Or, 99% of samples lie between: $0.78 - 2.58 \times \sqrt{\frac{0.78 \times 0.22}{n}}$ and $0.78 + 2.58 \times \sqrt{\frac{0.78 \times 0.22}{n}}$

Let us illustrate this example completely. If 78% of residents are happy with the payment method of direct debit, and you choose a sample of 400 residents to survey, the percentage of “happy residents” from this sample has a 95% chance to be $78\% \pm 4\%$, i.e. between 74% and 82%. These results are found by applying the formulae:

$$0.78 - 1.96 \times \sqrt{\frac{0.78 \times 0.22}{400}} = 0.78 - 0.04 \quad \text{and} \quad 0.78 + 1.96 \times \sqrt{\frac{0.78 \times 0.22}{400}} = 0.78 + 0.04$$

They allow us to clarify the meaning of the word “close” – it means within a range of 4%, because this is the margin of error. So, if the true percentage of residents happy with payment method is indeed 78%, then our sample has a 95% chance of returning a percentage between 74% and 82%.

The procedure of this hypothesis testing is as follows:

Hypothesis

Null hypothesis: H_0 : **Percentage = 78%**

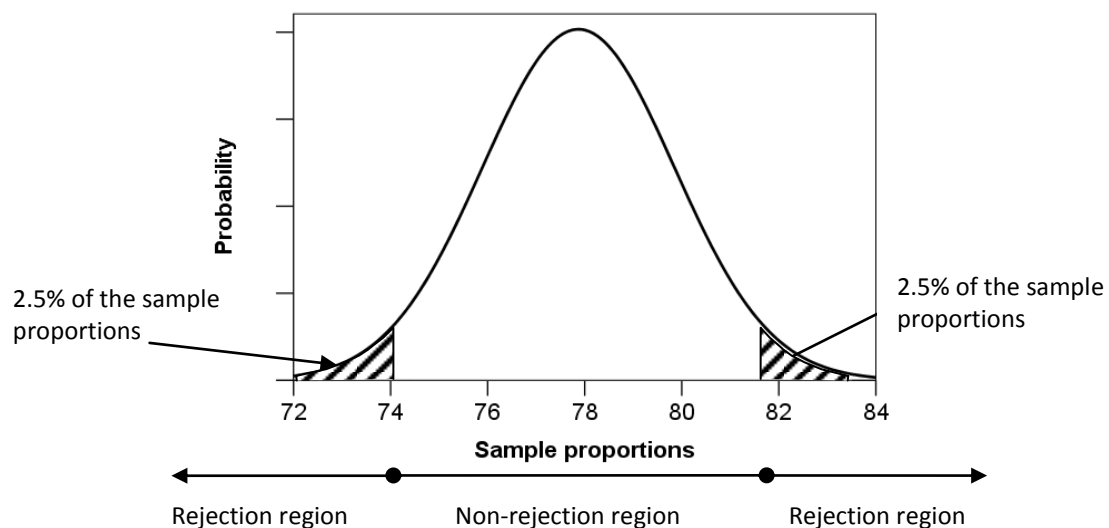
Alternative hypothesis: H_1 : **Percentage $\neq$ 78%**

We determine the risk of error that we are willing to take, to be 5% for example. This is also called the **level of significance**.

We determine the **Non-rejection region** 74% to 82%

We determine the **rejection region**. Less than 74% and more than 82%

These regions are represented graphically:



SCENARIO ONE:

Our sample: Let's say that in our survey of 410 residents, the percentage of them stating that they were happy with direct debit payment is 76%.

Our decision: We should not reject the null hypothesis and conclude that the percentage of residents happy with the direct debit payment is 78% as stated by the chief executive.

SCENARIO TWO:

Our sample: Our survey of 410 residents reveal that the percentage of them stating they were happy with direct debit payment is 73%.

Our decision: We should reject the null hypothesis and conclude that the percentage of residents happy with the direct debit payment is different from 78% as stated by the chief executive.

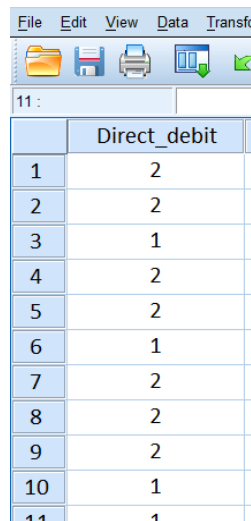
Using SPSS to test a hypothesis for population proportion

To run this test on SPSS, we need to have a sample of responses. Let us take a sample of size 410 similar to our **scenario two**. Let us assume, for simplicity, that the survey question was:

In relation to decision by the council to make council tax payments only by direct debit, are you:

- 1. Unhappy**
- 2. Happy**

The 410 responses are in 1s and 2s, because residents can only answer by either 1 or 2. Statisticians call any distribution that can only two values a **Binomial distribution**. Let us also assume there are no missing data.



	Direct_debit
1	2
2	2
3	1
4	2
5	2
6	1
7	2
8	2
9	2
10	1
11	1

It is always a good idea to produce a frequency table of the data to check the sample percentage for your categories.

Direct_debit					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Unhappy	111	27.1	27.1	27.1
	Happy	299	72.9	72.9	100.0
	Total	410	100.0	100.0	

Because we are interested in the proportion of “happiness” with payment method, **our category of choice here is happiness (code =2)**.

The hypothesised population proportion is 0.78 (78%).

Running the hypothesis testing for proportion on SPSS is long but necessary. We need to go step by step..

Step 1:

Go to **Analyze, Nonparametric Tests, One sample**.

You will see three tabs, “Objective”, “Fields” and “Settings”.

Step 2:

From the “Objective” tab, choose **“Automatically compare observed data to hypothesized”**.

Step 3

From the “Fields” tab, transfer the variable of choice to the box called **“Test fields”**.

Step 4

From the “Settings” tab,

- a) Select the “Choose Tests” item, and click on the radio button “Customize tests”.
Tick the test “Compare observed binary probability to hypothesized (Binomial test)”
Click on “Options”. Enter **0.78** in “Hypothesized proportion” and
Click on **“Specify success values”** under **“Define success for categorical fields”** and enter the value **2** in the list.
- b) Select the “Test Options” item, and choose the significance level, say 5% (or 1%).
- c) The “User-Missing Values” is used when we have missing data – We don’t have any here.

Click “Run” to obtain the following output:

Hypothesis Test Summary				
	Null Hypothesis	Test	Sig.	Decision
1	The categories defined by Direct_debit = (Happy) and (Unhappy) occur with probabilities 0.78 and 0.22.	One-Sample Binomial Test	.008	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

This output will be interpreted in class.

Hypothesis test for a single population mean

Let us go back to the example of sugar content in baked beans produced by a food processing company. The manager claims that the average (mean) amount of sugar per kg of baked beans is consistent and is around 23 grams. Is the manager's claim true?

The procedure of this hypothesis testing is similar to the one for proportion except for the margin of error:

Hypothesis

Null hypothesis: H_0 : **Mean** = 23

Alternative hypothesis: H_1 : **Mean** $\neq$ 23

We determine the level of significance to be 5% for example.

Let's say the health and hygiene inspector selects a random sample of 36 separate kilograms of baked beans from different batches and for each kilogram of beans, she measures the amount of sugar content and found a mean of 23.22 grams with a standard deviation of $s = 1.775$

We determine the **Non-rejection region**:

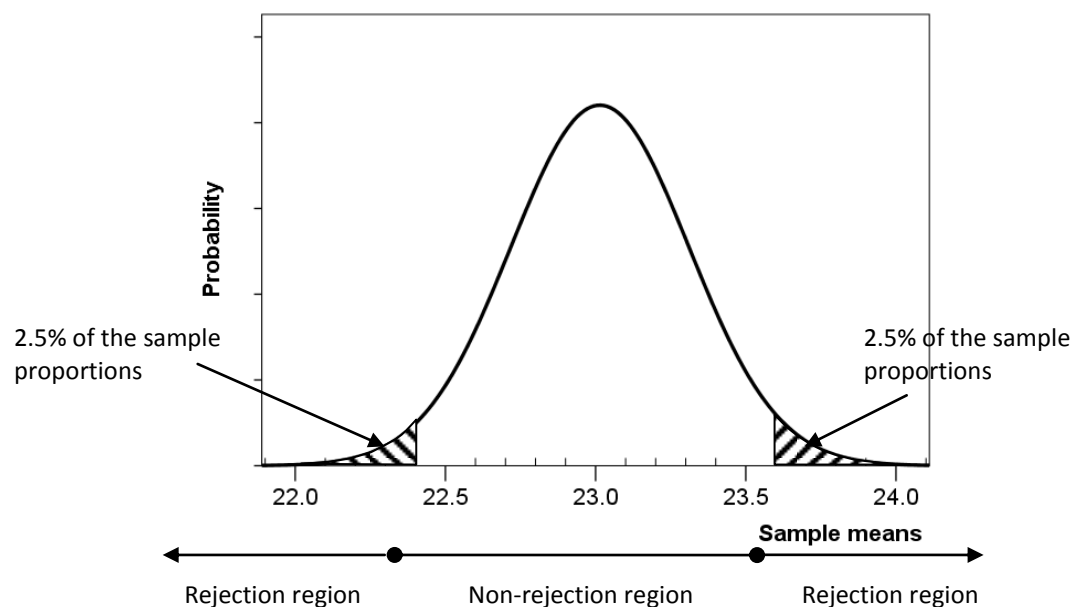
Using the z-scores, this region is: $23 - 1.96 \times \frac{1.775}{\sqrt{36}}$ and $23 + 1.96 \times \frac{1.775}{\sqrt{36}}$

Or, more like SPSS, using the t-scores, it is:

$$23 - 2.0301 \times \frac{1.775}{\sqrt{36}} = 22.4 \quad \text{and} \quad 23 + 2.0301 \times \frac{1.775}{\sqrt{36}} = 23.6$$

We determine the **rejection region**. Less than 22.4 and more than 23.6

These regions are represented graphically here:



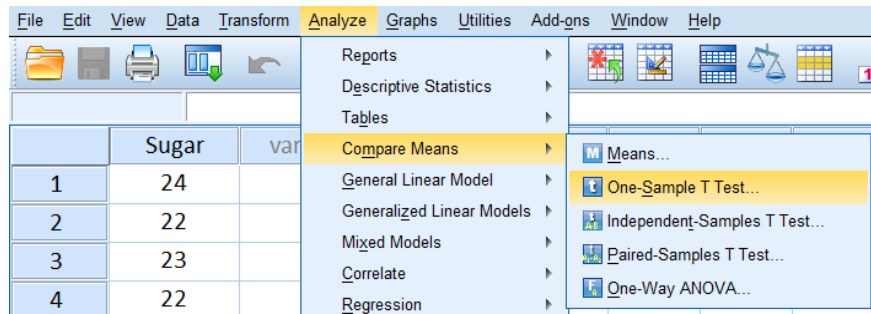
Our decision: As our sample mean of 23.22 grams is within the non-rejection region, we should not reject the null hypothesis and conclude that the average sugar content per kg of baked beans is 23 grams.

Using SPSS to test a hypothesis for population mean

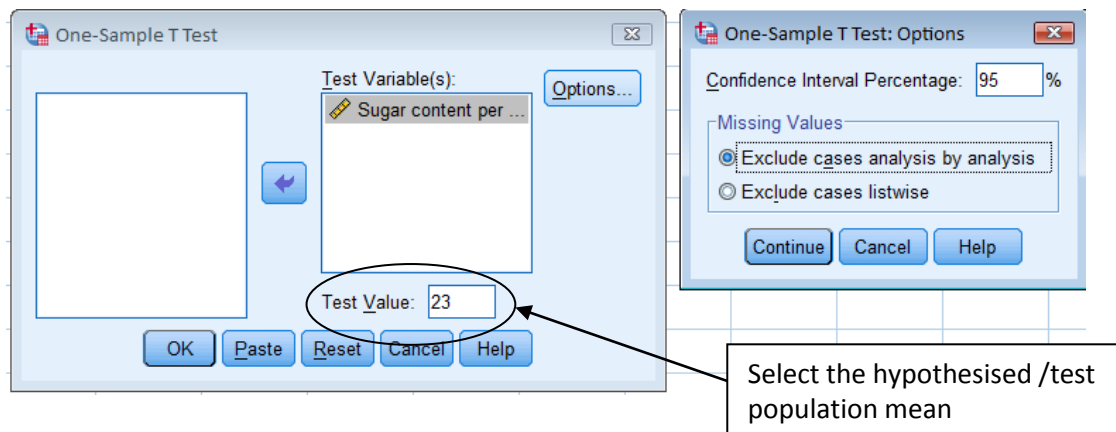
To run this test on SPSS, we need to have a sample of data. Let us use the following sample of sugar content per kg of baked beans from 36 different batches. We enter this on SPSS.

24	22	23	22	23	19	23	20	23	24	26	22	22	23	23	23	26	23
23	27	22	26	23	23	26	24	24	23	24	23	22	21	23	21	27	23

Step 1: Go to **Analyze, Compare Means, One-Sample T test..**



Step 2: Transfer the variable of choice to the “Test Variable(s)” box. Click on “Options” and choose the confidence interval. The default is always 95%.



Step 3

Click Continue, and then OK. You should obtain the following printout:

One-Sample Statistics				
	N	Mean	Std. Deviation	Std. Error Mean
Sugar content per kg of baked beans	36	23.22	1.775	.296

One-Sample Test						
	Test Value = 23					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Sugar content per kg of baked beans	.751	35	.457	.222	-.38	.82

We should interpret this output in class.

The various cases of alternative hypotheses

Let us use the example of sugar content to illustrate the various cases.

If we want to check that the amount of sugar per kilogram of baked beans is consistent throughout the production process and that the average does not vary too much from the claimed 23 grams, then we are inclined to alternatively test that the average is significantly different to 23 grams. Therefore our hypotheses hypothesis read:

Null hypothesis: H_0 : **Mean** = 23

Alternative hypothesis: H_1 : **Mean** $\neq$ 23 (two-tailed)

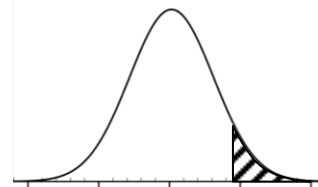


This hypothesis test is called a **two-tailed** test because the rejection regions are on both sides of the mean. It is also due to the alternative hypothesis. A population mean different from 23 means it is either greater (1 tail) or lower (1 tail) than 23, which is 2 tails.

If however, we are in a situation where the manager deliberately announced a lower average amount of sugar per kilogram (23 grams) because he fears the government health officers will be on his neck, for using large sugar content (bad for health), then the alternative to test if this average is higher than 23 grams. The hypotheses of choice in this case will be:

Null hypothesis: H_0 : **Mean** = 23

Alternative hypothesis: H_1 : **Mean** > 23 (one-tailed)

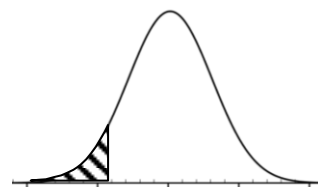


This hypothesis test is called **one-tailed** hypothesis because the rejection region is one side of the mean and forming a tail on the right hand side of the mean.

The third alternative hypothesis is when the manager claims that his baked beans are rich in iron, which is good for consumers' health, and that the average iron content per portion is 30 grams, more than the iron content of his competitors. The hypotheses of choice for this case could be written as:

Null hypothesis: H_0 : **Mean** = 30

Alternative hypothesis: H_1 : **Mean** < 30 (one-tailed)



Understanding the probability of error

Type I error

In our example about the proportion of residents happy with the direct debit method for paying their council tax, for scenario two, we concluded that we should reject the null hypothesis because a sample proportion of 73% was in the “rejection” region. That is to say, by sheer bad luck, our sample included a large number of disgruntled residents, unhappy with the payment method, answered “unhappy” when surveyed. They are a minority among very happy residents. This sample a 5% probability for being selected from a population who is 78% happy. We will therefore reject a true claim that the proportion is 78%. This type of error is conventionally called a **type I error** and can be made with a probability of 5% in the case of our scenario because we chose to test at this level of significance. The same can be said if the level of significance is chosen to be 1%.

Type II error

There is however another type of error, that of accepting the null hypothesis when it should be rejected. This happened in our example of sugar content. We did not reject the hypothesis that claimed that the overall sugar is on average equal to 23, when in reality we should have rejected it. This generally happens when we accept a hypothesis based on a sample that is so spurious its result proves anything. This type of error is called **type II error**. The probability of making type II error is easy to calculate.

Let us capture these possibilities of error in the following table:

		The reality	
		H_0 is true	H_0 is false
Our decision	H_0 is true	Correct decision	Wrong decision. Type II error. No precise probability of error
	H_0 is false	Wrong decision. Type I error. A precise probability of error	Correct decision

It follows from the above table that when we say that we reject H_0 (i.e. accept H_1), this decision is made for two reasons.

- We know exactly the risk of error for making this decision, and
- We can make this risk as small as we wish – 5%, 1% or even 0.5%.

For these reasons, when we want to demonstrate that a hypothesis is true, we are more convincing if we build this demonstration on the alternative hypothesis. This is why the alternative hypothesis is usually called the **research hypothesis**. This is a very important conclusion to remember.

The P-value

The most crucial point about hypothesis testing is the probability of accepting the alternative hypothesis as true or, if you like, the probability of rejecting the null hypothesis as false. If this probability is high, this means that the sample statistics (mean of proportion) is not a rare statistics from within the hypothesised population and therefore we should accept that it belongs to the population and should not reject H_0 .

If however, this probability is small, then the test statistics is very rare within the hypothesised population and we should reject H_0 .

When using SPSS to test a hypothesis, all you need to do is look for this probability, commonly called the **p-value**, and is presented on SPSS as the “**Sig**” value (see SPSS output on page 42 and 44).

By small, we mean smaller than the chosen level of significance (either 5% or 1% or whatever).

Exercises

- The manager of a local supermarket claimed that the take up on loyalty card in his locality is around 44%. A local activist surveyed a number of customers about the loyalty card take up outside the supermarket in order to test whether the claim by the manager is true. He publishes his results as an SPSS output.

SPSS output

Loyalty card take up.

		Frequency	Percent	Valid Percent
Valid	Have	144	40.0	40.0
	Don't have	216	60.0	60.0
	Total	360	100.0	100.0

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The categories defined by Loyalty card take up. = (Have) and (Don't have) occur with probabilities 0.44 and 0.56.	One-Sample Binomial Test	.070	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Required:

- How many customers he surveyed?
 - What is the percentage of loyalty card take up from his sample?
 - What percentage difference is there between the sample and hypothesised percentage of loyalty card take up?
 - Write down the two hypotheses for this test.
 - What level of significance did the activist used in the test?
 - Give one reason that the activist should accept the manager's claim.
- A survey on the supermarket shelf life of a new variety of strawberries was conducted during a weekend. It is claimed these strawberries should last no more than 65 minutes on shelves during weekend trading. A test based on the survey results was conducted at 1% level of significance and the results are presented as an SPSS output below:

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Shelf life of strawberries (in mins)	102	67.80	10.904	1.080

One-Sample Test

	Test Value = 65					
	t	df	Sig. (2-tailed)	Mean Difference	99% Confidence Interval of the ...	
					Lower	Upper
Shelf life of strawberries (in mins)	2.597	101	.011	2.804	-.03	5.64

Required:

- What is the sample size, the mean and standard deviation of shelf life in the sample?
- Write down the two hypotheses for this test.
- What decision should we make, based on this test output?

Computing session

Activity one:

Download the data file "Sugar content.sav" from Moodle.

- 1) Run a similar hypothesis test to the one presented on page 42 against a hypothesized population mean (sugar content) of 24 grams per kg of baked beans.
You should allow a 5% level of significance.
- 2) What decision should be made?
- 3) Do you think changing the level of significance to 1% will change anything in the output? If so, what is? You should try it.

Activity two:

Download the data file "customer satisfaction.sav" from Moodle. This file was used the first two weeks of this module.

- 1) Produce a frequency table for the rating of "Our departure location is convenient".
1=Completely disagree to 7= Completely agree
- 2) Run a hypothesis test against a hypothesised population proportion of those who rated as 7= Completely agree.
- 3) What decision can be made?



Model answers to seminar exercises

1.
 - a) 360
 - b) 40%
 - c) 4%
 - d) Null: proportion = 0.44 Alternative: proportion $\neq$ 0.44 (two-tailed test)
 - e) 0.05
 - f) The p-value is greater than 0.025, which is half the level of significance, so we do not reject the null hypothesis and conclude that the population proportion of customers having a loyalty card is not significantly different to 44%.
2.
 - a) sample size = 102, sample mean = 67.80, standard deviation = 10.904
 - b) Null: Mean = 65 min Alternative: proportion $>$ 65 min (one-tailed test)
 - c) The p-value (0.011) is greater than 0.01, so we do not reject the null hypothesis and conclude that average shelf life of these new strawberries is not significantly different to 65 minutes.