



Chapter S5: Hypothesis Testing

SYLLABUS INCLUDES

- concepts of null hypothesis (H_0) and alternative hypotheses (H_1), test statistic, critical region, critical value, level of significance and p -value
- formulation of hypotheses and testing for a population mean based on:
 - a sample from a normal population of known variance
 - a large sample from any population
- 1-tail and 2-tail tests
- Interpretation of the results of a hypothesis test in the context of the problem

PRE-REQUISITES

- S3 Normal Distribution
- S4 Sampling

CONTENT

- 1 Terminology**
 - 1.1 Null and Alternative Hypotheses
 - 1.2 Level of Significance
 - 1.3 Test Statistic, Critical Value, Critical Region, p -value
 - 1.4 Performing a Hypothesis Test
- 2 Hypothesis Tests on Population Mean**
 - 2.1 Using the z -test

Appendix 1 Type I and Type II Errors

Appendix 2 Performing the z -test using GC

INTRODUCTION

Let's look at the following scenario:

A manufacturer claims that the light bulbs he produces have a mean lifespan of 600 hours, and a standard deviation of 60 hours. A retailer, having received numerous complaints from his customers, suspects that they do not last as long. He contacts the manufacturer and they decide to test a random sample of 50 bulbs. It turns out that the average lifespan of this sample is 586 hours.

Does this constitute proof that the average lifespan of light bulbs is below 600 hours? Well, the manufacturer might argue that this is simply a chance occurrence; after all, we cannot expect the average lifespan of a sample to be exactly the same as the mean lifespan. How might the retailer present his case if he is to argue that this represents objective evidence that the mean lifespan is below 600 hours?

This type of scenario occurs sufficiently commonly in various fields of study that statisticians have devised a procedure, called **hypothesis testing**, to deal with it. In this chapter we will describe how this procedure is carried out in the context of the testing of the population mean.

1 TERMINOLOGY

1.1 Null and Alternative Hypotheses

The manufacturer's stand, that the observed variation from his stated lifespan on the light bulbs is purely due to chance, is what we will call the "**null hypothesis**". The retailer might postulate that the mean is actually lower, and his claim is known as the "**alternative hypothesis**". For the "A" Levels, we perform hypothesis tests on the population mean μ and compare the measured sample mean against a stated value, usually denoted as μ_0 .

In our introductory scenario, $\mu_0 = 600$ and the null hypothesis, denoted as H_0 , is stated formally as $H_0: \mu = 600$.

The alternative hypothesis denoted as H_1 , representing the retailer's claim that the mean is less than 600 hours, is stated formally as $H_1: \mu < 600$.

Thus we are testing $H_0: \mu = 600$ vs $H_1: \mu < 600$.

In general, we can also test for an increase in the mean, or simply a difference of the mean from μ_0 , that is

$H_0: \mu = \mu_0$	vs	$H_1: \mu < \mu_0$ (test for a decrease), or $H_1: \mu > \mu_0$ (test for an increase), or $H_1: \mu \neq \mu_0$ (test for a difference)
--------------------	-----------	--

The tests for a decrease or an increase are known as **1-tail tests**, while the test for a difference is a **2-tail test**.

Example 1

For each of the following cases, formulate appropriate null and alternative hypotheses.

- (a) Flour is sold, on average, at 500 g per bag. Test whether the average mass of the contents of a bag is being understated.

To test $H_0: \mu = 500$ vs $H_1: \mu > 500$

- (b) The company Kookakola produces cans of aerated water. The company claims that the mean content of aerated water is 505 ml. Test whether it is overstating the mean content.

To test $H_0: \mu = 505$ vs $H_1: \mu < 505$

- (c) A machine produces components of mean length 10 cm if it is correctly set up. A random sample of components produced by the machine is taken, and the length of each component is measured. Test whether the machine is correctly set up.

To test $H_0: \mu = 10$ vs $H_1: \mu \neq 10$

When carrying out a hypothesis test, we are looking for significant evidence to reject H_0 in favour of H_1 . But what constitutes significant evidence?

1.2 Level of Significance

When testing hypotheses there is always a possibility of making a wrong decision or error.

The threshold under which we consider H_0 to be sufficiently unlikely to be correct is called the **level of significance** (or **significance level**) of the hypothesis test.

The **level of significance** (or significance level) of a hypothesis test, denoted by α , is defined as the **probability of rejecting H_0 when H_0 is true**.

i.e. Level of significance = $P(\text{reject } H_0 \text{ when } H_0 \text{ is in fact true}) = P(\text{reject } H_0 | H_0 \text{ is true})$

In the context of our introductory scenario,

$\alpha = P(\text{concluding that the **mean** lifespan of a light bulb is shorter than 600 hrs when it actually is 600 hrs}).$

As such, it is tempting to choose α to be as small as possible, since the conclusion is a wrong one; so why not pick $\alpha = 0$? Refer to Appendix 1 (Type I and Type II Errors) for more information.

Appropriate values for α depends on which area of study we are engaged in and are usually chosen by consensus. For the social sciences, it might be chosen as high as $\alpha = 0.3$, the biological and medical fields mostly use $\alpha = 0.05$ (5% level of significance) or $\alpha = 0.01$ (1% level of significance), while the physical sciences might use a value as small as $\alpha = 10^{-6}$.

Note that **the lower the level of significance (i.e. the smaller α is), the stronger the evidence needed to reject H_0** .

1.3 Test Statistic, Critical Value, Critical Region, p -value

Test Statistic

To carry out the test, the focus moves from X , the lifespan of lightbulbs to the distribution of $\bar{X}$, the mean lifespan from a sample of lightbulbs. $\bar{X}$ is the **test statistic** for the population mean μ .

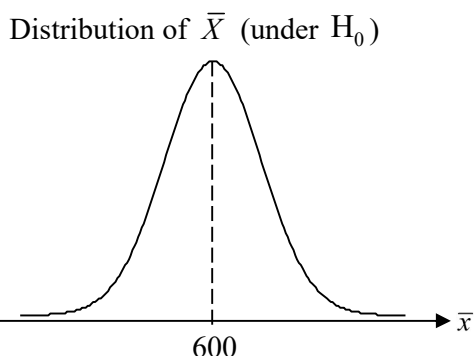
A **test statistic** is a random variable used to make the decision “do not reject H_0 ” or “reject H_0 ”. The test statistic measures the degree of agreement between a sample of data and the null hypothesis. Its observed value changes randomly from one random sample to another.

We will discuss in Section 2, how to make use of the test statistic, $\bar{X}$, and its distribution, when performing a hypothesis test on the population mean, μ .

In our example, $H_0 : \mu = 600$ vs $H_1 : \mu < 600$.

Under the assumption that the null hypothesis is true, the population mean and standard deviation are $\mu = 600$ and $\sigma = 60$ (these values are given by the manufacturer), so, by the

Central Limit Theorem, we have $\bar{X} \sim N\left(600, \frac{60^2}{50}\right)$ approximately, since $n = 50$ is large.

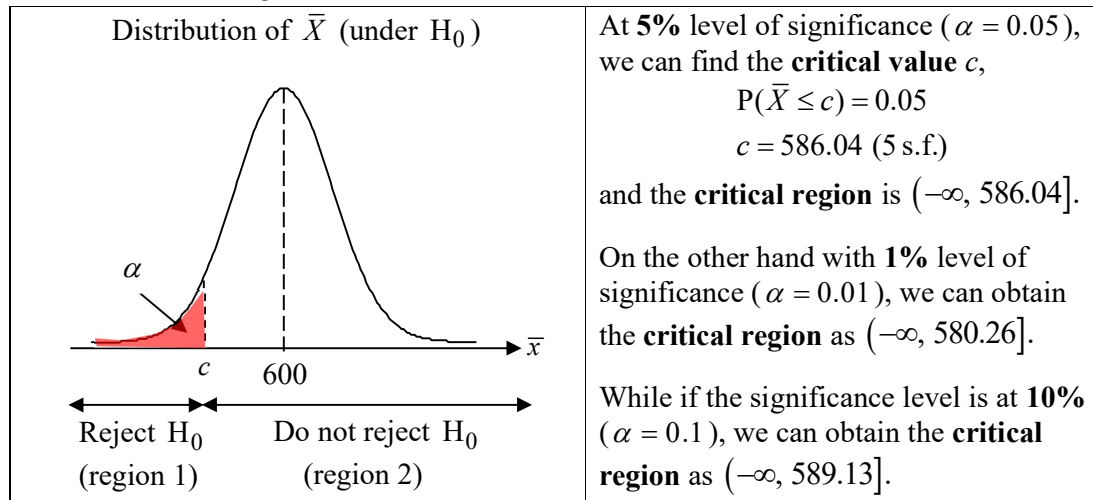


Critical Value, Critical Region

The result of the test depends on the whereabouts of the test value in the sampling distribution, 586 hours (observed mean value of lifespan of a random sample of 50 bulbs). If the value is close to 600 hours then it is likely to have come from a distribution with mean 600 hours. On the other hand, if it is far away from 600 hours, then it is unlikely to have come from a distribution with mean 600 hours.

A decision needs to be taken about the cut-off point, c , known as the **critical value** which indicates the boundary of the region where values would be considered too far away from 600 hours and therefore unlikely to occur. The region is known as the critical region or rejection region. The value of c depends on the stated significance level α .

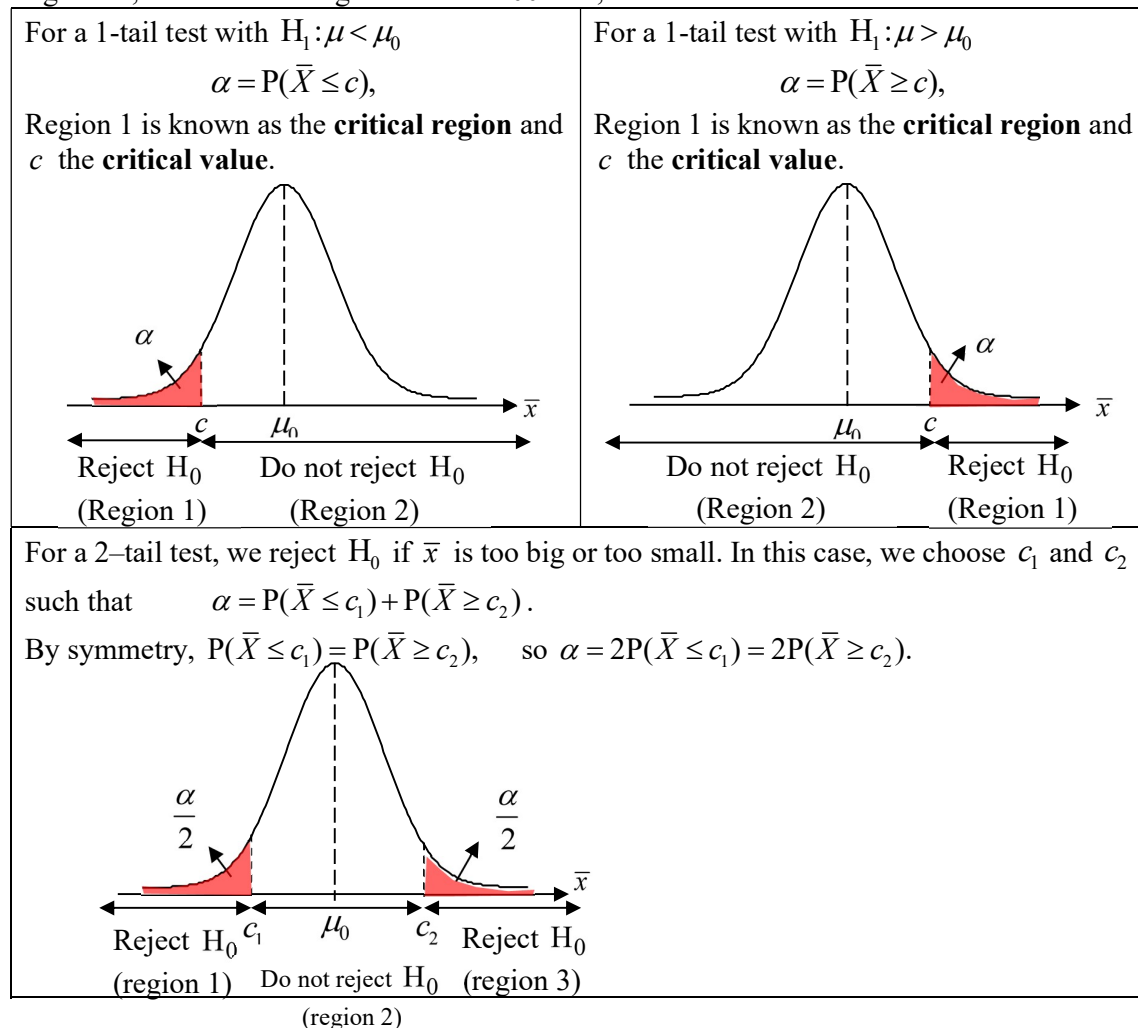
The situation in the original scenario is illustrated below:



Here, $\bar{x} = 586 \in (-\infty, 586.04]$, i.e. $\bar{x}$ lies in the critical region, so we reject the null hypothesis at the 5% (and 10%) level of significance, in favour of the alternative hypothesis.

On the other hand, at 1% level of significance, $\bar{x} = 586 \notin (-\infty, 580.26]$, i.e. $\bar{x}$ does not lie in the critical region, so we fail to reject the null hypothesis.

In general, if the level of significance is $100\alpha\%$,



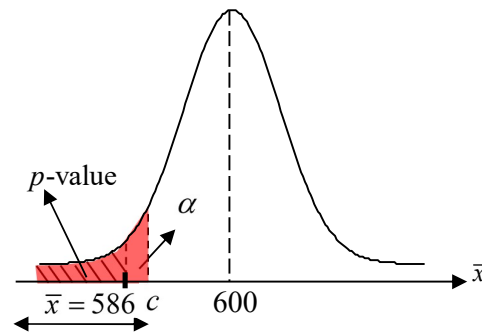
p-value

In the context of our introductory scenario, we know that $\bar{X} \sim N\left(600, \frac{60^2}{50}\right)$ approximately by

Central Limit Theorem, and the observed value, $\bar{x} = 586$. This allows us to compute the probability of obtaining a value as small as, or smaller than the observed value $\bar{x} = 586$, i.e. finding $P(\bar{X} \leq 586)$. This probability is known as the **p-value**.

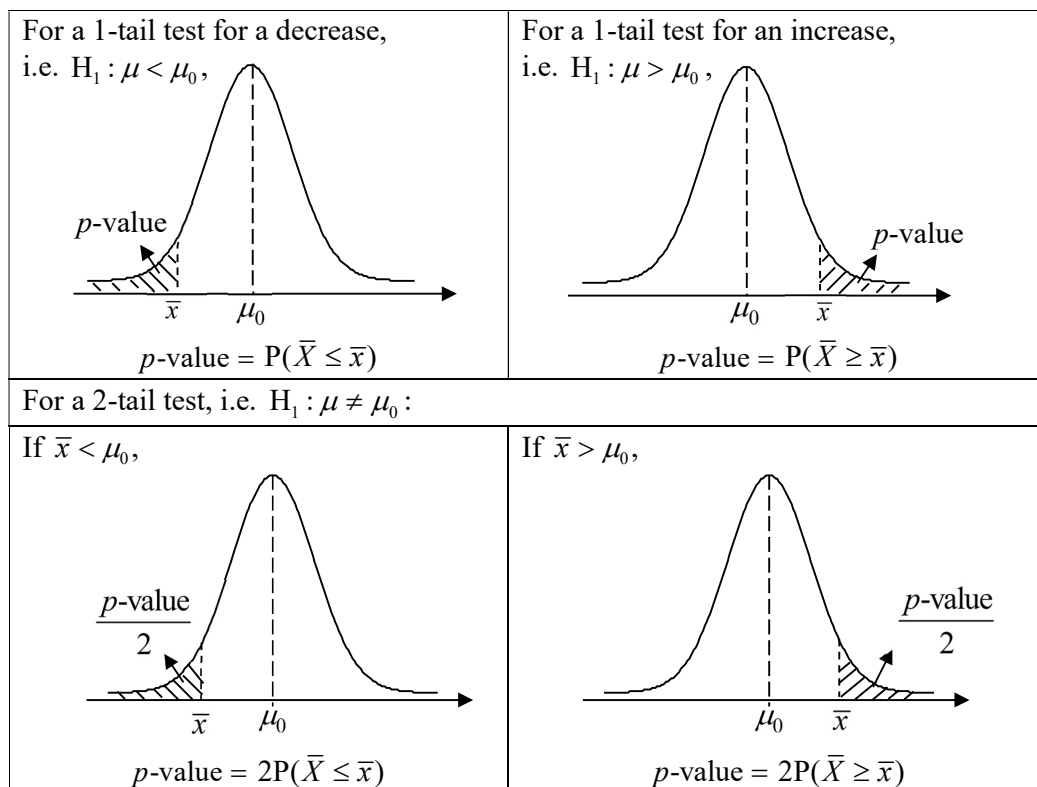
In this context, using the graphing calculator,
 $p\text{-value} = P(\bar{X} \leq 586) = 0.0495$. (3 s.f.)

Since $p\text{-value}$ is smaller than our level of significance of 5% ($\alpha = 0.05$), we conclude that the observed value of $\bar{x} = 586$ is sufficiently unlikely to be due to chance alone. Hence we reject the null hypothesis H_0 .



The **p-value** is the probability of observing from a sample, a test statistic value “*at least as extreme as*” the observed value computed under the assumption that the null hypothesis is true.

In general, the **p-value** is computed, depending on the particular type of test being performed, as illustrated below:



There are 2 possible outcomes of testing:

- (1) If $p\text{-value} \leq \alpha$, we **reject** H_0 and conclude that there is **sufficient** evidence against the claim of H_0 .
- (2) If $p\text{-value} > \alpha$, we **do not reject** H_0 and conclude that there is **insufficient** evidence against the claim of H_0 .

Since we compare the p -value against the significance level to decide whether to reject H_0 , the p -value is the **smallest level of significance at which H_0 can be rejected**.

Note:

“Do not reject H_0 ” is **not** equivalent to “accept H_0 ”. In fact, under the framework of hypothesis testing, there is no way to prove that H_0 is true; all we can assess is whether there is sufficient evidence against it.

The p -value, together with the value of the associated test statistic, is a convenient summary of the evidence found in experiments and is commonly reported in journals. This leaves it open for the reader to judge the significance of the result. (i.e. choose their own α .) The p -value is often referred to as an **observed** level of significance.

1.4 Performing a Hypothesis Test

Now that we have learnt all the essential elements of a test of hypothesis, below is a framework for carrying out a hypothesis test:

Step 1	Understand the given question and <u>write down</u> the null hypothesis H_0 and the alternative hypothesis H_1
Step 2	<u>Write down</u> the level of significance α (usually given in the question)
Step 3	Decide on the test statistic to be used and determine its distribution
Step 4	<u>Use the GC</u> to calculate the p-value
Step 5	Reject H_0 if $p\text{-value} \leq \alpha$, OR Do not reject H_0 if $p\text{-value} > \alpha$ Write down the <u>conclusion</u> in the <u>context of the question</u>

Note that aside from Step 3, where we need to decide which test statistic is to be used, the other steps require us to only “write down” certain information or to “use the GC”. In essence, once we have decided on the test statistic in Step 3, we would have solved the problem, provided the test statistic is the correct one.

2 HYPOTHESIS TESTS ON POPULATION MEAN

As mentioned in Section 1.4, we need to decide which test statistic to use. In the A level H2 Math context, we will focus on performing hypothesis tests on population mean, μ , with the sample mean, $\bar{X}$, as the test statistic. To work out the distribution of the test statistic, we assume that H_0 is true.

Test Statistic and its distribution

	Characteristic of population	Distribution of test statistic, $\bar{X}$
Population Variance σ^2 known	(I) <i>Normal Population</i>	$\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$
	(II) <i>Non-normal population with <u>large</u> sample size</i>	$\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$ approximately by Central Limit Theorem
1) Population Variance σ^2 unknown and 2) <u>large sample size</u> Note: s^2 , an unbiased estimate of σ^2 , is used instead.	(III) <i>Normal Population</i>	$\bar{X} \sim N\left(\mu_0, \frac{s^2}{n}\right)$ approximately
	(IV) <i>Non-normal population</i>	$\bar{X} \sim N\left(\mu_0, \frac{s^2}{n}\right)$ approximately by Central Limit Theorem

2.1 Using the z-test

When the test statistic follows normal distribution, we use the **z-test**. Thus in all cases above, we use the **z-test** unless sample size is small and population variance is unknown, which is not included in the H2 Math syllabus. If we know the population variance σ^2 (Cases I and II as above), we will use it. Otherwise we will estimate it with s^2 (Cases III and IV as above).

When the sample size is small and the population variance is known, we can only proceed if we know that the population is normally distributed (Case I as above). For examination questions, this condition will either be given, or you will have to state it as an assumption where applicable.

Example 2

From records, the IQ of university students is normally distributed with a mean of 118. Students from ABC University claim that the mean IQ of the students in their university is greater than 118. A sample of 50 students was tested and the mean IQ was found to be 121.

- (i) Assuming that $\sigma = 12$, test, at 5% significance level, whether the sample provides significant evidence to support the students' claim.
- (ii) Comment on whether there is any change in the conclusion of the hypothesis test if the mean IQ of the sample of 50 students was 120 instead.

Solution:

- (i) Let X be the IQ of a student in ABC University and μ be the population mean IQ of a student.

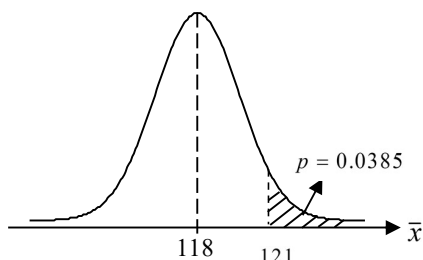
Step 1: To test $H_0: \mu = 118$ vs $H_1: \mu > 118$

Step 2: Perform a 1-tail test at 5% significance level.

Step 3: Under H_0 , $\bar{X} \sim N\left(118, \frac{12^2}{50}\right)$

From the sample, $\bar{x} = 121$

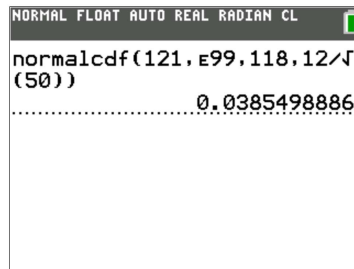
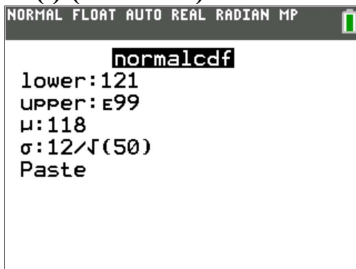
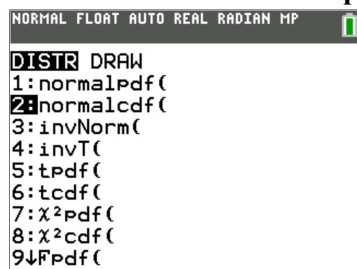
Step 4: Using a z-test, $p\text{-value} = P(\bar{X} \geq 121) = 0.0385$ (3 s.f.)



Step 5: Since $p\text{-value} = 0.0385 \leq 0.05$, we reject H_0 and conclude that there is sufficient evidence, at 5% level of significance, to support the claim that the mean IQ of students in ABC University is greater than 118.

- (ii) If $\bar{x} = 120$, $p\text{-value} = P(\bar{X} \geq 120) = 0.119$ (3 s.f.)

Since $p\text{-value} = 0.119 > 0.05$, we do not reject H_0 and conclude that there is insufficient evidence, at 5% level of significance, to support the claim that the mean IQ of students in ABC University is greater than 118.

GC screen shots for Example 2(i) (Normcdf):

Example 3

The random variable X is thought to have a mean of 50 but it is known that the standard deviation is 14.5. A random sample of 100 gives a mean of 52.6.

Is there evidence that the population mean has increased

- (i) at 5% level of significance, (ii) at 1% level of significance?

Find the least value of the sample mean such that there is sufficient evidence at the 1% level of significance that the population mean has increased, giving your answer correct to 1 decimal place.

State, giving a reason, if any assumption needs to be made about the distribution of X ?

Solution

Let μ be the population mean of X .

To test $H_0: \mu = 50$ vs $H_1: \mu > 50$

Under H_0 , since $n = 100$ is large, $\bar{X} \sim N\left(50, \frac{14.5^2}{100}\right)$ approximately, by Central Limit Theorem.

From the sample, $\bar{x} = 52.6$

Using a 1-tail z-test, $p\text{-value} = P(\bar{X} \geq 52.6) = 0.0365$ (3 s.f.)

- (i) At 5% level of significance

Since $p\text{-value} = 0.0365 \leq 0.05$, we **reject** H_0 and conclude that there is **sufficient** evidence, at 5% level of significance, that the population mean has increased.

- (ii) At 1% level of significance

Since $p\text{-value} = 0.0365 > 0.01$, we **do not reject** H_0 and conclude that there is **insufficient** evidence, at 1% level of significance, that the population mean has increased.

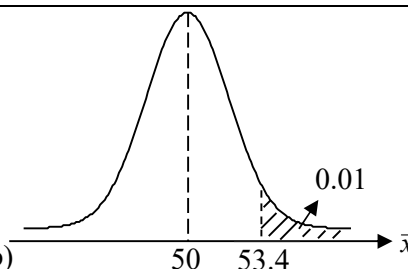
To reject H_0 at 1% level of significance,

$$p\text{-value} \leq 0.01$$

$$P(\bar{X} \geq k) \leq 0.01$$

Since $P(\bar{X} \geq 53.4) = 0.01$, $k \geq 53.4$

Hence, the least value of the sample mean is 53.4 (1 d.p)



No assumption about the distribution of X is needed. Since the sample size 100 is large, by Central Limit Theorem, the sample mean $\bar{X}$ follows a normal distribution approximately.

GC screen shots for Example 3 (invNorm):

```
NORMAL FLOAT AUTO REAL RADIAN MP
DISTR DRAW
1:normalpdf(
2:normalcdf(
3:invNorm(
4:invT(
5:tpdf(
6:tcdf(
7:χ²pdf(
8:χ²cdf(
9:Fpdf(
```

```
NORMAL FLOAT AUTO REAL RADIAN CL
invNorm
area:.01
μ:50
σ:14.5/√(100)
Tail: LEFT CENTER RIGHT
Paste
```

```
NORMAL FLOAT AUTO REAL RADIAN CL
invNorm(.01,50,14.5/√(100),RIGHT)
53.37320442
```

Example 4

A teacher sets an examination paper for students. Let X be the time, in minutes, taken by a student to complete the examination paper.

She gave the paper to 60 randomly chosen students. The results obtained are summarized as follows:

$$\sum x = 3048 \quad \text{and} \quad \sum (x - \bar{x})^2 = 465.$$

- (i) Calculate unbiased estimates of the population mean and population variance.

The teacher wishes to test at the 2% significance level, whether the population mean time taken by a student to complete the examination paper is 50 minutes.

- (ii) State appropriate hypotheses for the test, defining any symbols you use.

- (iii) Find the critical region for this test and explain what it means in this context.

Solution

X is the time, in minutes, taken by a student to complete the examination paper.

- (i) An unbiased estimate of the population mean is $\bar{x} = \frac{1}{n} \sum x = \frac{3048}{60} = 50.8$

An unbiased estimate of the population variance is $s^2 = \frac{1}{n-1} \sum (x - \bar{x})^2 = \frac{465}{59}$

- (ii) Let μ minutes be the population mean time taken by a student to complete the examination paper.

To test null hypothesis, $H_0: \mu = 50$ versus alternative hypothesis, $H_1: \mu \neq 50$

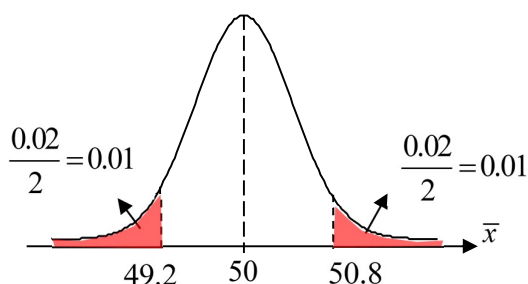
- (iii) Perform a 2-tail test at 2% significance level.

Under H_0 , since the sample size $n = 60$ is large, by Central Limit Theorem,

$\bar{X} \sim N\left(50, \frac{\frac{465}{59}}{60}\right)$ approximately.

Use $s^2 = \frac{465}{59}$. Not allowed to use $s^2 = 7.88$ (3sf) as it might lead to loss of accuracy

Using GC, critical region is $(0, 49.2) \cup (50.8, \infty)$.



Critical region refers to the range of sample mean time, taken by a student in a sample of 60 randomly chosen students, to complete the examination paper such that there is sufficient evidence at 2% significance level to conclude that the mean time taken by a student to complete the paper is not 50 minutes.

Example 5

An electronic device is advertised as being able to retain information stored in it 'for 70 to 90 hours' after power has been switched off. In experiments carried out to test this claim, the retention time in hours, X , was measured on 250 occasions, and the data obtained is summarized by $\sum (x - 76) = 683$ and $\sum (x - 76)^2 = 26132$. The population mean and variance of X are denoted by μ and σ^2 respectively.

- (i) Show that, correct to one decimal place, an unbiased estimate of σ^2 is 97.5.
- (ii) Test the hypothesis that $\mu = 80$ against the alternative hypothesis that $\mu < 80$, at 5% significance level.

Solution

- (i) Let $Y = X - 76$. Then we have $\sum y = 683$, $\sum y^2 = 26132$.

$$s_y^2 = \frac{1}{n-1} \left[\sum y^2 - \frac{(\sum y)^2}{n} \right]$$

$$= \frac{1}{249} \left(26132 - \frac{683^2}{250} \right) \approx 97.454$$

$$s_x^2 = s_y^2 = 97.5 \text{ (1 d.p.)}$$

$\therefore$ An unbiased estimate of σ^2 is 97.5 (1dp)

- (ii) To test $H_0: \mu = 80$ vs $H_1: \mu < 80$
Perform a 1-tail test at 5% significance level.

Under H_0 , since $n = 250$ is large, by Central

Limit Theorem, $\bar{X} \sim N\left(80, \frac{97.5}{250}\right)$ approximately.

We can use $s^2 = 97.5$ here since we were asked to show its value in (i). That is, taking $s^2 = 97.454$ is not necessary here.

From the sample,
 $\bar{y} = \bar{x} - 76$

$$\bar{x} = \bar{y} + 76 = \frac{683}{250} + 76 = 78.732$$

Using a z -test,

$$p\text{-value} = P(\bar{X} \leq 78.732) = 0.0212 \text{ (3 s.f.)}$$

Since $p\text{-value} = 0.0212 \leq 0.05$, we reject H_0 and conclude that there is sufficient evidence, at 5% significance level, that the mean retention time is less than 80 hours.

Example 6 [DHS/9758/2019/Prelim/02/Q9a & c]

The time taken, T (in minutes), for a 17-year-old student to complete a 5-km run is a random variable with mean 30. After a new training programme is introduced for these students, a random sample of n students is taken. The mean time and standard deviation for the sample are found to be 28.9 minutes and 4.0 minutes respectively.

- (a) Find an unbiased estimate of the population variance in terms of n .
- (b) The trainer claims instead that the new training programme is able to improve the mean of T , 30 minutes, by at least 5%. The school wants to test his claim.
- (i) Write down the null and alternative hypothesis.
- (ii) Using the existing sample, the school carried out a test at 1% significance level and found that there was sufficient evidence to reject the trainer's claim. Find the set of values that n can take, stating any necessary assumption(s) needed to carry out the test.

Solution

(a) An unbiased estimate for the population variance, $s^2 = \frac{n}{n-1}(4^2) = \frac{16n}{n-1}$ minutes²

(b)(i) New population mean timing = $0.95 \times 30 = 28.5$ minutes

Let μ minutes be the population mean time taken by a student to complete the 5-km run.

Null hypothesis, $H_0 : \mu = 28.5$

Alternative hypothesis, $H_1 : \mu > 28.5$

(b)(ii) Assumption: n is large for Central Limit Theorem to apply.

Under H_0 , $\bar{T} \sim N\left(28.5, \frac{4^2}{n-1}\right)$ approximately by Central Limit Theorem

For H_0 to be rejected, we need

$$p\text{-value} = P(\bar{T} \geq 28.9) \leq 0.01$$

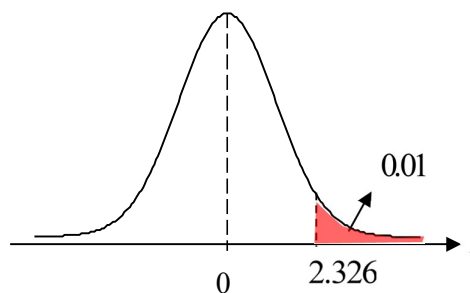
$$P\left(Z \geq \frac{28.9 - 28.5}{\frac{4}{\sqrt{n-1}}}\right) \leq 0.01$$

$$P\left(Z \geq \frac{\sqrt{n-1}}{10}\right) \leq 0.01$$

$$\frac{\sqrt{n-1}}{10} \geq 2.3263$$

$$n \geq 542.2$$

Thus required set = $\{n \in \mathbb{Z} : n \geq 543\}$



APPENDIX 1 Type I and Type II Errors

In the practice of statistics, when we perform hypothesis testing, the null hypothesis is presumed to be true, unless there is sufficient evidence against it.

In conducting a hypothesis test, there is of course a probability of arriving at a wrong conclusion about the validity of the alternative hypothesis. This can be very serious, as a wrong conclusion will lead to wrong decisions and courses of action.

There are **two types of error in hypothesis testing**, namely *Type I* and *Type II Errors*.

Decision \ Truth	Null Hypothesis	Alternative Hypothesis
Null Hypothesis	✓	Type II Error (do not reject H_0 when it is false)
Alternative Hypothesis	Type I Error (reject H_0 when it is true)	✓

Table 1

A *Type I Error* is committed when one wrongly rejects a true null hypothesis and a *Type II Error* is committed when one fails to reject a false null hypothesis. In practice, a good scientific study will minimize the chance of making such errors.

Power of a Test

When performing hypothesis testing, it seems that we should aim to choose α (the probability of rejecting H_0 when it is true) to be as small as possible, since we want to keep the rate of *Type I Error* fairly low.

So why don't we choose α to be 0, that is, $P(\text{Type I Error}) = 0$?

Note that if we do not reject H_0 regardless of whether it is true or false, then we will never wrongly reject H_0 , so $\alpha = 0$. But such a decision rule is meaningless! Moreover, if we do not reject H_0 but H_0 happens to be false, then we are committing Type II Error!

Let $\beta = P(\text{Type II Error}) = P(\text{do not reject } H_0 \text{ when } H_0 \text{ is false})$.

Then the **power** of a hypothesis test $= 1 - \beta$, i.e. the probability of correctly rejecting H_0 when H_0 is false.

Thus, $0 \leq \text{power} \leq 1$ and other considerations being equal, *the larger the power the better*.

Ideally, we want both α and β to be as small as possible. However, it turns out that the smaller the α , the larger the β and vice versa. The challenge is to find a balance between the two and to decide which type of error is a more critical mistake! Also, an increase in the sample size will usually reduce both *Type I and Type II Errors*.

(I) Criminal Trial Analogy – “innocent until proven guilty”

The criminal trial analogy works like hypothesis testing. Some criminal justice system assumes "**the defendant is innocent until proven guilty**." That is, the defendant is presumed to be innocent. In statistics, the hypotheses will be as follows:

- H_0 : Defendant is not guilty (innocent)
- H_1 : Defendant is guilty

The prosecution team then tries to collect evidence (e.g. hair samples, finger prints etc), hoping to find "sufficient evidence" to “reject the null hypothesis”, that is to refute the initial assumption of innocence.

If the evidence presented by the prosecution does not prove that the defendant is guilty beyond a reasonable doubt (do not have sufficient evidence at $\alpha\%$ level of significance that the alternative hypothesis is true), then based on the given evidence, the jury cannot reject that the *possibility* that the defendant *is* innocent (do not reject the null hypothesis). In other words, it has not been proven that the defendant is *innocent* but it cannot be proven legally that the defendant is *guilty*.

On the other hand, if the evidence presented by the prosecution is sufficient beyond a reasonable doubt to convict the defendant (have sufficient evidence at $\alpha\%$ level of significance that the alternative hypothesis is true), then the jury refute the initial assumption that the defendant is innocent (reject the null hypothesis) and deems the defendant guilty.

Criminal trials like statistical tests are susceptible to error. Let's review the two types of errors that can be made in criminal trials:

Jury Decision \ Truth	Not Guilty	Guilty
Not Guilty	✓	Error (guilty going scot - free)
Guilty	Error (innocent being convicted)	✓

Table 2

The above correspond to the two types of errors in hypothesis testing discussed in Table 1.

(II) False Positives and False Negatives in Health Screening

In medical statistics, **false positives** and **false negatives** are concepts that also correspond to *Type I and Type II errors* in statistical hypothesis testing (Table 1).

Test Result \ Truth	Does not have disease	Has disease
Negative	✓	False Negative
Positive	False Positive	✓

Table 3

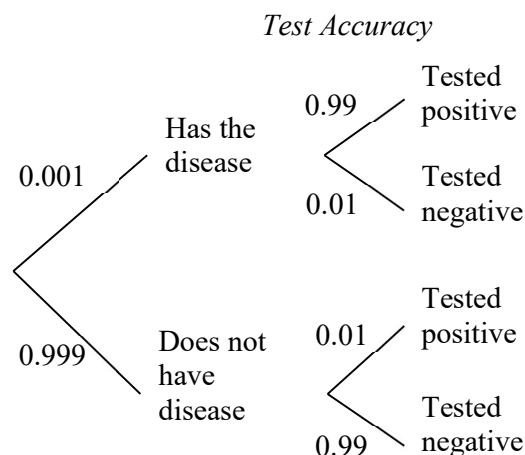
A false negative is a test result that indicates a person does not have a disease or condition when the person actually does have it. Correspondingly, a false positive is a test result that indicates a person has a specific disease or condition when the person actually does not have it.

The False Positive Paradox

If a test for a disease is 99% accurate and you receive a positive result, what are the odds that you actually have the disease?

Many people *would think* that since the test is 99% accurate, a person who received positive results has about a 99% chance of having the disease. This might be wrong! If the disease is very common, your odds of actually having the disease might approach 99%. However, the rarer the disease is, the less accurate the test and the lower the odds that you actually have the disease. The reason involves *conditional probability*.

Suppose the disease affects one out of every 1000 people in the population. Assume that the test accuracy is constant for patients with or without the disease. Using a tree diagram,



$P(\text{test positive} \mid \text{has disease}) = 0.99$, which is usually what the “accuracy” of the test means.

$$\Rightarrow P(\text{test positive \& has disease}) = 0.99 \times 0.001 = 0.00099$$

$$\begin{aligned} P(\text{has disease} \mid \text{test positive}) &= \frac{P(\text{test positive \& has disease})}{P(\text{test positive})} \\ &= \frac{0.00099}{0.99 \times 0.001 + 0.01 \times 0.999} \\ &= 0.09016 \end{aligned}$$

Thus, for any person who tests positive, the odds of them having the disease is about 9%.

Can we then believe a negative test result?

$$\begin{aligned} P(\text{test negative} \mid \text{no disease}) &= 0.99 \\ \Rightarrow P(\text{test negative \& no disease}) &= 0.99 \times 0.999 = 0.98901 \\ P(\text{no disease} \mid \text{test negative}) &= \frac{P(\text{test negative \& no disease})}{P(\text{test negative})} \\ &= \frac{0.98901}{0.01 \times 0.001 + 0.99 \times 0.999} \\ &= 0.99989 \end{aligned}$$

Thus, a negative result is *almost certainly correct*.

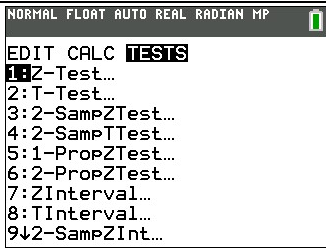
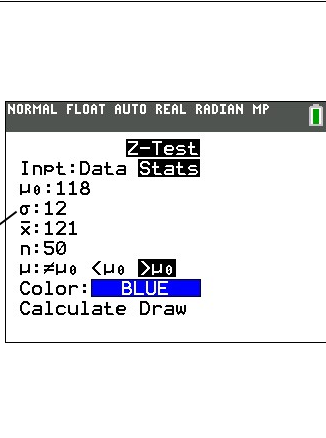
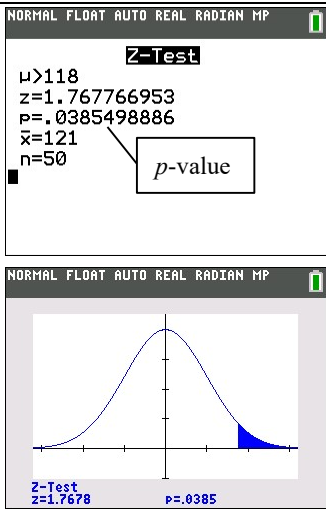
It is important and necessary to look at the conditional probabilities in the interpretation of test results. This could be a real problem in the field of medicine, as patients or even doctors might not fully understand how to interpret such information from health screening statistics.

False Negatives

False negatives create two problems. The first is a false sense of security. For example, if your manufacturing line is unable to detect defective items, it might be thought that the process is running more effectively than it actually is. The second, potentially more serious issue, is that potentially dangerous situations may be missed. For example, a crippling computer virus can wreak havoc if not detected, or an individual with cancer may not receive timely treatment.

APPENDIX 2 Performing the z-test using GC

With reference to Example 2, here are the GC screenshots when performing a z-test.

1 Press [stat] and scroll right to select TESTS Select 1: Z – Test	
2 Select Stats Key in the values of μ_0, σ, $\bar{x}$, n and select the required 1–tail or 2–tail test Take note that unlike “normalcdf”, we key in σ here and not $\frac{\sigma}{\sqrt{n}}$.	
3 Select Calculate and press [enter] to obtain the p-value. OR You may also select Draw and press [enter] for the area associated with the p-value.	

SUMMARY