Statistics for Computing MA4413

Lecture 15

Hypothesis Testing

Kevin Burke

kevin.burke@ul.ie

Confidence Intervals

Confidence intervals are used mainly for *statistical estimation*.

They provide a *range of plausible values* for the parameter of interest at a given level of *confidence* via

$$\text{statistic} \pm z_{\alpha/2} \times s(\text{statistic})$$

where $s(\text{statistic})$ is the *standard error* of the estimate and $z_{\alpha/2}$ is the z value which puts probability $\alpha/2$ in the lower and upper tails, i.e.,

$$\Pr(\text{interval contains parameter}) = 1 - \alpha.$$

$\Rightarrow (1-\alpha)100\%$ confident that the interval contains the parameter.

Confidence Intervals

We have calculated confidence intervals for the following parameters:

- One mean: μ.
- One proportion: p.
- Difference between means: $\mu_1 - \mu_2$.
- Difference between proportions: $p_1 - p_2$.

In each case there is a formula for calculating the standard error.

For numerical data, we replace the $z_{\alpha/2}$ value with a $t_{\nu,\alpha/2}$ value when the sample size is small and *approximately normal*.

Furthermore, when comparing means in two small samples, there is a different formula for $s(\overline{X}_1 - \overline{X}_2)$ if we assume equal variances.

Confidence Intervals for Hypothesis Testing

A **hypothesis** is a *belief* about the value of a parameter.

We can use confidence intervals to investigate hypotheses.

Recall the gameplay example from Lecture13 where a developer believes there are 16 hours of gameplay; the *hypothesis* is $\mu = 16$.

- Data was collected and a 95% confidence interval was then calculated: $[14.62, 16.18]$.
- The interval contains $\mu = 16$ and, therefore, we *accept* the hypothesis that there are 16 hours of gameplay.

Hypothesis Testing

There is a more *formal approach* to **hypothesis testing**:

1. State hypotheses *before* collecting data.
    - The **null hypothesis**, denoted $\boldsymbol{H_0}$, is the hypothesis to be tested.
    - The **alternative hypothesis**, denoted $\boldsymbol{H_a}$, is the counter-hypothesis, i.e., the *complement* of H_0.
2. Choose the **significance level**: the probability of *error* (denoted α) in rejecting H_0, i.e., the chance of rejecting H_0 when it is true.
3. Collect data and calculate a **test statistic**.
4. Compare this statistic to a **critical value** from statistical tables.
5. Make a decision; either there is:
    a) *Evidence* to **reject** $\boldsymbol{H_0}$ (in favour of H_a) at the α significance level.
    b) *Not enough evidence* to reject H_0 at the α level, i.e., **accept** $\boldsymbol{H_0}$.

Test Statistic

By the central limit theorem,

$$\bar{X} \sim \text{Normal}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

$\Rightarrow$ 95% of the time, the $\bar{x}$ calculated will be in $\mu \pm 1.96\,\frac{\sigma}{\sqrt{n}}$.

Standardising the above gives:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}\,(0, 1)$$

$\Rightarrow$ 95% of the time, the ***z* score** calculated will be in ± 1.96.

(note: usually σ is unknown but when $n > 30$ we can substitute with s)

Test Statistic

We *hypothesise* that the true mean is equal to some value, μ_0. This gives us the following null and alternative hypotheses:

$$\boxed{\begin{aligned} H_0 &: \quad \mu = \mu_0 \\ H_a &: \quad \mu \neq \mu_0 \end{aligned}}$$

Assuming the null hypothesis is true (i.e., $\mu = \mu_0$), we know that, 95% of the time, the calculated **test statistic**

$$\boxed{z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}}$$

will be within ± 1.96, i.e., only 5% of the time will we get a z value outside of this range.

Critical Values

If our calculated z value is outside of ± 1.96 then we **reject H_0 at the 5% level of significance** (note: in 5% of cases we will *wrongly reject* H_0).

Of course, we can test H_0 at other significance levels by using different **critical values**, cv, i.e., the values

$$\boxed{\pm z_{\alpha/2}}$$

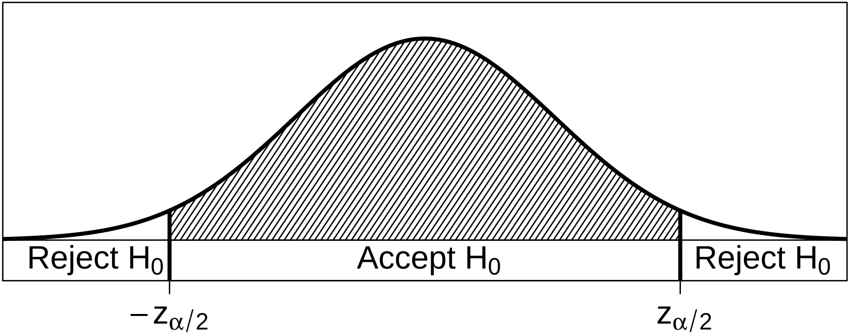
correspond to testing at the α significance level and, furthermore, if $n \leq 30$, the critical values come from the t tables:

$$\boxed{\pm t_{\nu, \alpha/2}}.$$

Hypothesis Testing Procedure

1. When testing the hypothesis that $\mu = \mu_0$ we have:
   - $H_0 : \mu = \mu_0$,
   - $H_a : \mu \neq \mu_0$.
2. Select the α level, e.g., $\alpha = 0.05$.
3. Collect data $\Rightarrow$ calculate $\bar{x}$, s and, hence, the test statistic $z = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$.
4. Compare z to the critical values, cv: either $\pm z_{\alpha/2}$ or $\pm t_{\nu,\alpha/2}$ depending on the sample size.
5. Conclusion:
   a) If z is outside of $\pm cv \Rightarrow$ reject H_0 (in favour of H_a) at the α level.
   b) If z is inside $\pm cv \Rightarrow$ accept H_0 at the α level.

Acceptance and Rejection Regions



- Significance level $\alpha \Rightarrow$ each tail contains probability $\alpha/2$.
- **Acceptance region** lies within $\pm z_{\alpha/2}$.
- **Rejection region** lies outside of $\pm z_{\alpha/2}$.

Note: If n is small use $\pm t_{\nu,\alpha/2}$ instead of $\pm z_{\alpha/2}$.

Proof Vs Evidence

Note that we **never** use the word *proof* since there is always some probability of error.

- *Cannot:* prove H_0 to be true or untrue.
- *Can:* provide evidence that H_0 is reasonable or unreasonable.

If the test statistic falls into the **acceptance region**, we say that the *evidence against* H_0 is **non-significant** $\Rightarrow$ we accept it.

If the test statistic falls into the **rejection region**, we say that the *evidence against* H_0 is **significant** $\Rightarrow$ we reject it.

Example: Tubs of Butter

A machine is programmed to put 250g of butter into each tub. We wish to test this hypothesis at the 5% level. Thus,

$$H_0: \quad \mu = 250,$$
$$H_a: \quad \mu \neq 250,$$

i.e., the hypothesised value is $\mu_0 = 250$.

We are testing at the 5% level $\Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$.

Therefore, the critical values will be $\pm z_{0.025} = \pm 1.96$ or, if the selected sample is small, $\pm t_{\nu, 0.025}$ where $\nu = n - 1$.

Example: Tubs of Butter

We randomly select 35 tubs of butter and measure the contents. The average for this sample is 251.6g and the standard deviation is 3.1g.

Since $n > 30$ the critical values are ± 1.96.

The test statistic is:

$$z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{251.6 - 250}{\frac{3.1}{\sqrt{35}}} = \frac{1.6}{0.524} = 3.05.$$

The calculated $z = 3.05$ lies outside of ± 1.96. Therefore we reject the null hypothesis that $\mu = 250$ at 5% level of significance.

Conclusion: the evidence suggests that the machine is *not* working as programmed.

Question 1

A particular CPU design is expected to have a clock speed of 2.5Ghz. To test this hypothesis, 4 CPUs were selected and the results are as follows:

2.53	2.55	2.54	2.24

a) State the null and alternative hypotheses.

b) What are the critical values (use $\alpha = 0.1$).

c) Calculate $\bar{x}$ and s.

d) Calculate the test statistic.

e) State your conclusion.

Two-Tailed Test

For the butter tub example note that the alternative hypothesis was

$$H_a: \quad \mu \neq 250.$$

If μ is *not equal* to 250 $\Rightarrow$ it is either *less than* 250 or *greater than* 250:

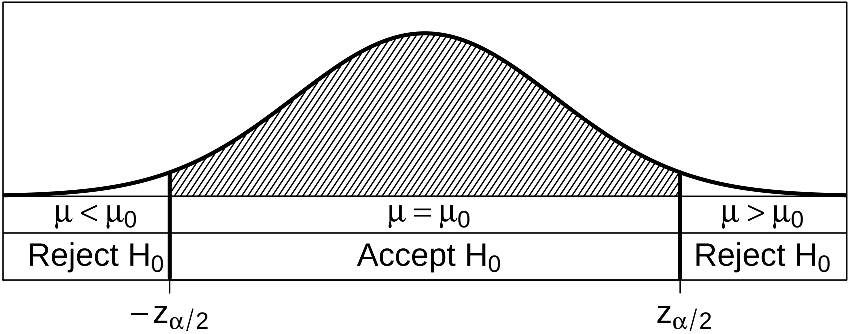
$$\Rightarrow \quad H_a: \quad \mu < 250 \quad \text{or} \quad \mu > 250.$$

The alternative hypothesis points to the rejection region, i.e., we see that the rejection region appears in both the *lower and upper tails* via the "$<$" and "$>$" inequality signs.

Hence, this is known as a **two-tailed** hypothesis test.

Two-Tailed Test

$$\boxed{\begin{aligned} H_0 &: \quad \mu = \mu_0 \\ H_a &: \quad \mu \neq \mu_0 \end{aligned}}$$



Note: If n is small use $\pm t_{\nu,\alpha/2}$ instead of $\pm z_{\alpha/2}$.

One-Tailed Test

In other situations, it makes more sense to carry out a **one-tailed** test.

Consider the gameplay example. Really it is of interest to test the null hypothesis that there are *at least* 16 hours of gameplay:

$$H_0 : \quad \mu \geq 16.$$

Thus, the alternative hypothesis is

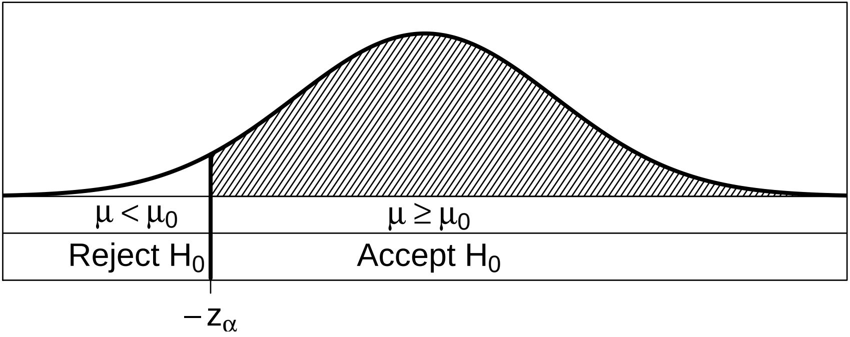
$$H_a : \quad \mu < 16.$$

For a one-tailed test we **do not divide α by two** since all of this probability goes into *one* tail.

The alternative hypothesis points to the rejection region.

One-Tailed Test: Lower Tail

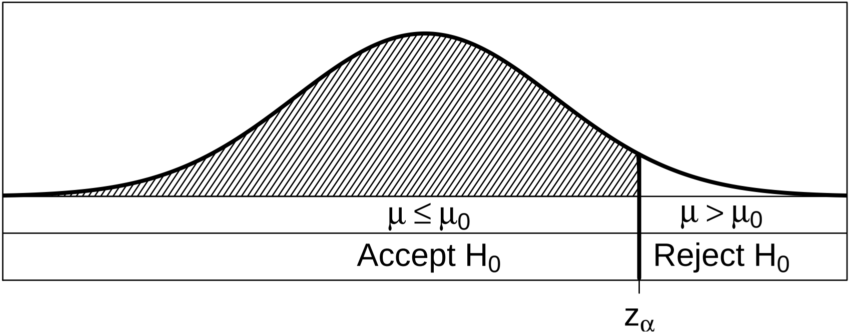
$$\boxed{\begin{aligned} H_0 &: \quad \mu \geq \mu_0 \\ H_a &: \quad \mu < \mu_0 \end{aligned}}$$



Note: If n is small use $-t_{\nu,\alpha}$ instead of $-z_{\alpha}$.

One-Tailed Test: Upper Tail

$$\boxed{\begin{aligned} H_0 &: \quad \mu \leq \mu_0 \\ H_a &: \quad \mu > \mu_0 \end{aligned}}$$



Note: If n is small use $t_{\nu,\alpha}$ instead of z_α.

Direction of Test

Hypotheses	Tails	Critical Values	Rejection Region
$H_0: \quad \mu = \mu_0$ $H_a: \quad \mu \neq \mu_0$	Two	$\pm z_{\alpha/2}$ or $\pm t_{\nu,\alpha/2}$	Outside critical values
$H_0: \quad \mu \geq \mu_0$ $H_a: \quad \mu < \mu_0$	One	$-z_{\alpha}$ or $-t_{\nu,\alpha}$	Below critical value
$H_0: \quad \mu \leq \mu_0$ $H_a: \quad \mu > \mu_0$	One	z_{α} or $t_{\nu,\alpha}$	Above critical value

Note: the null hypothesis **always** contains the equals.

Example: Gameplay

We return to the gameplay example.

A games developer wishes to test the null hypothesis that their game has *at least* 16 hours of gameplay using the 5% level of significance.

Thus the hypotheses are:

$$H_0: \quad \mu \geq 16$$

$$H_a: \quad \mu < 16$$

- One-sided test $\Rightarrow$ use $\alpha = 0.05$ (do *not* divide by two).
- Alternative hypothesis points to the *lower tail* $\Rightarrow -z_{\alpha}$ or $-t_{\nu,\alpha}$.

Example: Gameplay

From a sample of $n = 33$, $\bar{x} = 15.4$ and $s = 2.3$.

Note that $n > 30 \Rightarrow$ the critical value is $-z_{0.05} = -1.64$.

The test statistic is:

$$z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{15.4 - 16}{\frac{2.3}{\sqrt{33}}} = \frac{-0.6}{0.4} = -1.5.$$

Since $z = -1.5$ is *not* in the rejection region (i.e., the region below -1.64), we accept $H_0 : \mu \geq 16$ at the 5% level.

Conclusion: the number of gameplay hours is sufficient (at least 16) and, therefore, we do not have to add extra content to the game.

Question 2

Assume that a particular CPU is designed to have a normal operating temperature of 30 degrees or less. A sample of 40 CPUs were tested: the average for this sample was 32 degrees and the variance was 10 degrees-squared.

We wish to test the hypothesis that the temperature is less than or equal to 30 degrees at the 1% level.

a) State the null and alternative hypotheses.

b) Define the rejection region.

c) Calculate the test statistic.

d) State your conclusion.

Decisions

When we carry out a hypothesis test, there are two possible decisions:

a) Reject H_0.

b) Accept H_0.

This leads to four different scenarios:

i) Reject H_0 when H_0 is true $\Rightarrow$ **type 1 error**.

ii) Reject H_0 when H_0 is false.

iii) Accept H_0 when H_0 is true.

iv) Accept H_0 when H_0 is false $\Rightarrow$ **type 2 error**.

Errors

We fix the probability of wrongly *rejecting* H_0 at the α level:

$$\text{Pr}(\text{Reject } H_0 \mid H_0 \text{ is true}) = \alpha,$$

i.e., we control the probability of a **type 1 error**.

The probability of wrongly *accepting* H_0 is

$$\text{Pr}(\text{Accept } H_0 \mid H_0 \text{ is false}) = \beta,$$

i.e., a **type 2 error**.

Relationship Between α and β

The value of β **depends on** α. It can be *calculated* once we have chosen the α level.

We will not calculate β in this course but just know the relationship:

Decreasing α leads to an increase in β.

If we reduce α, the acceptance region increases $\Rightarrow$ rejection of H_0 is less likely $\Rightarrow$ we are more open to a type 2 error.

In practice, these values will be chosen to reflect the costs (often monetary) associated with making the two types of error.

Example: Tubs of Butter

Recall that in the butter tub example we had the following hypotheses:

$$H_0: \quad \mu = 250$$
$$H_a: \quad \mu \neq 250$$

We tested H_0 at the 5% level and, since the test is two-tailed and n was large, the critical values were $z_{0.025} = \pm 1.96$.

The test statistic $z = 3.05$ and, since this is outside of ± 1.96, we rejected H_0 at the 5% level.

However, with $z = 3.05$, the evidence is stronger than the 5% level. In fact we would reject H_0 at the 1% level also (since $z_{0.005} = \pm 2.58$).

Example: Tubs of Butter

The question that arises is, just how strong is the evidence against H_0?

What is the lowest α level at which we would reject H_0 if $z = 3.05$? We require the α level that leads to the critical values $z_{\alpha/2} = \pm 3.05$.

From the normal tables we find that $\Pr(Z > 3.05) = 0.00114$. This is the $\alpha/2$ value $\Rightarrow \alpha = 2(0.00114) = 0.00228$ is the lowest α level at which H_0 would be rejected.

Thus, we can see that the evidence against H_0 is very strong.

P-Value

The quantity we have just calculated is called a **p-value**.

It tells us the how likely the data is if H_0 were true. Thus, a p-value is a measure of **evidence against $\boldsymbol{H_0}$**:

- Small p-value $\Rightarrow$ data is unlikely under H_0 $\Rightarrow$ evidence to reject H_0.
- Large p-value $\Rightarrow$ data is likely under H_0 $\Rightarrow$ evidence to accept H_0.

We can calculate p-values as follows:

$$\boxed{\text{p-value} \;=\; \begin{cases} 2 \times \Pr(Z > |z|) & \text{if } H_a : \mu \neq \mu_0 \\ \Pr(Z < z) & \text{if } H_a : \mu < \mu_0 \\ \Pr(Z > z) & \text{if } H_a : \mu > \mu_0 \end{cases}}$$

(note: $|z|$ is the *absolute value* of z)

P-Value

Since 1%, 5% and 10% are commonly used significance levels, we can use the following as a rough guide to interpreting p-values:

- $0.00 <$ p-value $< 0.01 \Rightarrow$ strong evidence against H_0.
- $0.01 <$ p-value $< 0.05 \Rightarrow$ evidence against H_0.
- $0.05 <$ p-value $< 0.10 \Rightarrow$ some evidence against H_0 (not strong).
- $0.10 <$ p-value $< 1.00 \Rightarrow$ no evidence against H_0.

Example: Gameplay

Recall that in the gameplay example:

$$H_0: \quad \mu \geq 16$$
$$H_a: \quad \mu < 16$$

This is a one-sided test (with rejection region in the lower tail) and, therefore, the p-value is $\Pr(Z < z)$ where z is the test statistic.

Since the calculated test statistic was $z = -1.5$, we have

$$\text{p-value} = \Pr(Z < -1.5) = \Pr(Z > 1.5) = 0.0668.$$

Thus, although there is some evidence against H_0, it is not strong.

Furthermore, since the stated significance level was $\alpha = 0.05$ in this example, we would not reject H_0.

Other Parameters

We have introduced the idea of hypothesis testing using μ.

The general procedure is much the same for other parameters.

That is, we calculate:

$$z = \frac{\text{statistic} - \text{hypothesised value}}{\text{standard error}}$$

and then compare this to a critical value.

Proportions

Recall that for a proportion, the *true* standard error is $\sigma(\widehat{P}) = \sqrt{\frac{p\,(1-p)}{n}}$.

For the calculation of confidence intervals, we replaced p with $\hat{p}$ (since p is unknown) $\Rightarrow s(\widehat{P}) = \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$.

For hypothesis testing, we use the **hypothesised value** $p = p_0$:

$$\boxed{s(\widehat{P}) = \sqrt{\frac{p_0\,(1-p_0)}{n}}}.$$

Apart from this difference, everything is the same as in the case of hypothesis testing for μ as covered already.

Example: User Age

A company believes that 70% of users of their product are teenagers; this impacts advertising campaigns. They wish to test this hypothesis. Therefore, we have the following null and alternative hypotheses:

$$H_0: \quad p = 0.7$$

$$H_a: \quad p \neq 0.7$$

A random sample of users were selected and it was found that 40 out of 75 were teenagers, i.e., $\hat{p} = \frac{40}{75} = 0.533$.

$$\Rightarrow z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0\,(1-p_0)}{n}}} = \frac{0.533 - 0.7}{\sqrt{\frac{0.7\,(0.3)}{75}}} = \frac{-0.167}{0.0529} = -3.16.$$

Note that p_0 is used in the standard error.

Example: User Age

We will view the evidence against H_0 using the p-value approach. Since this is a two tailed test we have:

$$\begin{aligned}
\text{p-value} = 2\ \Pr(Z > |z|) &= 2\ \Pr(Z > |-3.16|) \\
&= 2\ \Pr(Z > 3.16) \\
&= 2\,(0.00079) \\
&= 0.00158.
\end{aligned}$$

Thus, it appears that H_0 is unlikely. The proportion of teenagers who use the product is not as much as they thought.

Question 3

A new version of an operating system is being developed. A beta version is released to some randomly selected individuals who are then asked: “Do you prefer the new system?”.

By default, the company will assume that the old system is preferred, i.e., the hypothesis to be tested is $p \leq 0.5$.

It was found that 38 out of 65 people prefer the new system.

a) State the null and alternative hypotheses.
b) What is the critical value if $\alpha = 0.05$?
c) Calculate the test statistic.
d) Provide your conclusion.
e) Calculate a p-value also.