Lecture 5

Hypothesis Testing

CSE 445 – Spring 2020

Acknowledgement: Antonio Paiva

What is hypothesis testing?

A statistical hypothesis is an assertion or conjecture concerning one or more populations.

To prove that a hypothesis is true, or false, with *absolute certainty*, we would need *absolute knowledge*. That is, we would have to examine the *entire population*.

Instead, hypothesis testing concerns on how to use a random sample to judge if it is evidence that supports or not the hypothesis.

What is hypothesis testing? (cont.)

Hypothesis testing is formulated in terms of *two* hypotheses:

- H_0: the null hypothesis;
- H_1: the alternate hypothesis.

What is hypothesis testing? (cont.)

The hypothesis we want to test is if H_1 is "likely" true.

So, there are two possible outcomes:

- *Reject* H_0 and accept H_1 because of sufficient evidence in the sample in favor or H_1;
- *Do not reject* H_0 because of insufficient evidence to support H_1.

What is hypothesis testing? (cont.)

Very important!!

Note that failure to reject H_0 does not mean the null hypothesis is true. There is no formal outcome that says "accept H_0." It only means that we do not have sufficient evidence to support H_1.

What is hypothesis testing? (cont.)

Example

In a jury trial the hypotheses are:

- H_0: defendant is innocent;
- H_1: defendant is guilty.

H_0 (innocent) is rejected if H_1 (guilty) is supported by evidence beyond "*reasonable doubt*." Failure to reject H_0 (prove guilty) does not imply innocence, only that the evidence is insufficient to reject it.

Case study

A company manufacturing RAM chips claims the defective rate of the population is 5%. Let p denote the *true* defective probability. We want to test if:

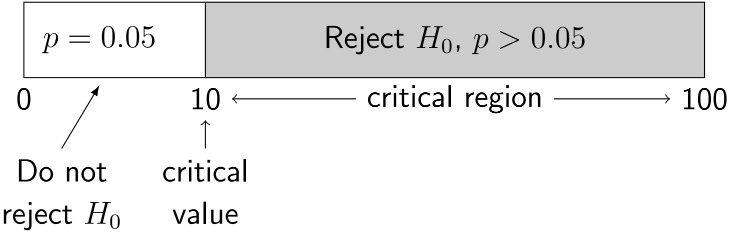
- $H_0 : p = 0.05$
- $H_1 : p > 0.05$

We are going to use a sample of 100 chips from the production to test.

Case study (cont.)

Let X denote the number of defective in the sample of 100.
Reject H_0 if $X \geq 10$ (chosen "arbitrarily" in this case).

X is called the *test statistic*.



Case study (cont.)

Why did we choose a critical value of 10 for this example? Because this is a Bernoulli process, the expected number of defectives in a sample is np. So, if $p = 0.05$ we should expect $100 \times 0.05 = 5$ defectives in a sample of 100 chips. Therefore, 10 defectives would be strong evidence that $p > 0.05$.

The problem of how to find a critical value for a desired level of significance of the hypothesis test will be studied later.

Types of errors

Because we are making a decision based on a finite sample, there is a possibility that we will make mistakes.
The possible outcomes are:

	H_0 is true	H_1 is true
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

Types of errors (cont.)

Definition

The acceptance of H_1 when H_0 is true is called a Type I error. The probability of committing a type I error is called the *level of significance* and is denoted by α.

Example

Convicting the defendant when he is innocent!

The lower significance level α, the less likely we are to commit a type I error. Generally, we would like small values of α; typically, 0.05 or smaller.

Types of errors (cont.)

Case study continued

$$\begin{aligned}
\alpha &= \Pr(\text{Type I error}) = \Pr(\text{reject } H_0 \text{ when } H_0 \text{ is true}) \\
&= \Pr(X \geq 10 \text{ when } p = 0.05) \\
&= \sum_{x=10}^{100} b(x; n = 100, p = 0.05), \qquad \text{binomial distribution} \\
&= \sum_{x=10}^{100} \binom{100}{n} 0.05^x 0.95^{100-x} = 0.0282
\end{aligned}$$

So, the level of significance is $\alpha = 0.0282$.

Types of errors (cont.)

Definition

Failure to reject H_0 when H_1 is true is called a Type II error. The probability of committing a type II error is denoted by β.

Note: It is impossible to compute β unless we have a specific alternate hypothesis.

Types of errors (cont.)

Case study continued

We cannot compute β for $H_1 : p > 0.05$ because the true p is unknown. However, we can compute it for testing $H_0 : p = 0.05$ against the alternative hypothesis that $H_1 : p = 0.1$, for instance.

$$\begin{aligned}
\beta &= \Pr(\text{Type II error}) = \Pr(\text{reject } H_1 \text{ when } H_1 \text{ is true}) \\
&= \Pr(X < 10 \text{ when } p = 0.1) \\
&= \sum_{x=0}^{9} b(x; n = 100, p = 0.1) = 0.4513
\end{aligned}$$

Types of errors (cont.)

Case study continued

What is the probability of a type II error if $p = 0.15$?

$$\begin{aligned}
\beta &= \Pr(\text{Type II error}) \\
&= \Pr(X < 10 \text{ when } p = 0.15) \\
&= \sum_{x=0}^{9} b(x; n = 100, p = 0.15) = 0.0551
\end{aligned}$$

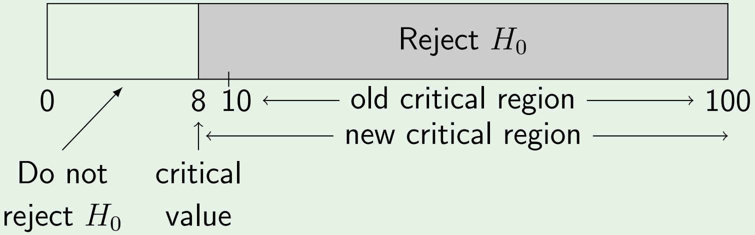
Effect of the critical value

Moving the critical value provides a trade-off between α and β. A reduction in β is always possible by increasing the size of the critical region, but this increases α. Likewise, reducing α is possible by decreasing the critical region.

Effect of the critical value (cont.)

Case study continued

Lets see what happens when we change the critical value from 10 to 8. That is, we reject H_0 if $X \geq 8$.



Effect of the critical value (cont.)

Case study continued

The new significance level is

$$\begin{aligned}\alpha &= \Pr(X \geq 8 \text{ when } p = 0.05) \\ &= \sum_{x=8}^{100} b(x; n = 100, p = 0.05) = 0.128.\end{aligned}$$

As expected, this is a large value than before (it was 0.0282).

Effect of the critical value (cont.)

Case study continued

Testing against the alternate hypothesis $H_1 : p = 0.1$,

$$\beta = \Pr(X < 8 \text{ when } p = 0.1)$$
$$= \sum_{x=0}^{7} b(x; n = 100, p = 0.1) = 0.206,$$

which is lower than before.

Testing against the alternate hypothesis $H_1 : p = 0.15$,

$$\beta = \sum_{x=0}^{7} b(x; n = 100, p = 0.15) = 0.012,$$

again, lower than before.

Effect of the sample size

Both α and β can be reduced *simultaneously* by increasing the sample size.

Case study continued

Consider that now the sample size is $n = 150$ and the critical value is 12. Then, reject H_0 if $X \geq 12$, where X is now the number of defectives in the sample of 150 chips.

Effect of the sample size (cont.)

Case study continued

The significance level is

$$\begin{aligned}\alpha &= \Pr(X \geq 12 \text{ when } p = 0.05) \\ &= \sum_{x=12}^{150} b(x; n = 150, p = 0.05) = 0.074.\end{aligned}$$

Note that this value is lower than 0.128 for $n = 100$ and critical value of 8.

Effect of the sample size (cont.)

Case study continued

Testing against the alternate hypothesis $H_1 : p = 0.1$,

$$\begin{aligned}\beta &= \Pr(X < 12 \text{ when } p = 0.1) \\ &= \sum_{x=0}^{11} b(x; n = 150, p = 0.1) = 0.171,\end{aligned}$$

which is *also* lower than before (it was 0.206).

Approximating the binomial distribution using the normal distribution

Factorials of very large numbers are problematic to compute accurately, even with Matlab. Thankfully, the binomial distribution can be approximated by the normal distribution (see Section 6.5 of the book for details).

Approximating the binomial distribution using the normal distribution (cont.)

Theorem

If X is a binomial random variable with n trials and probability of success of each trial p, then the limiting form of the distribution of

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \qquad n \to \infty$$

is the standard normal distribution.

This approximation is good when n is large and p is not extremely close to 0 or 1.

Approximating the binomial distribution using the normal distribution (cont.)

Case study continued

Lets recompute α with the normal approximation.

$$
\begin{aligned}
\alpha &= \Pr(\text{Type I error}) = \Pr(X \geq 12 \text{ when } p = 0.05) \\
&= \sum_{x=12}^{150} b(x; n = 150, p = 0.05) \\
&\approx \Pr\left(Z \geq \frac{12 - 150 \times 0.05}{\sqrt{150 \times 0.05 \times 0.95}}\right) = \Pr(Z \geq 1.69) \\
&= 1 - \Pr(Z \leq 1.69) = 1 - 0.9545 = 0.0455.
\end{aligned}
$$

Not too bad... (It was 0.074.)

Approximating the binomial distribution using the normal distribution (cont.)

Case study continued

What if we increase the sample size to $n = 500$ and the critical value to 40?

The normal approximation should be better since n is larger.

$$\alpha \approx \Pr\left(Z \geq \frac{40 - 500 \times 0.05}{\sqrt{500 \times 0.05 \times 0.95}}\right) = \Pr(Z \geq 3.08)$$
$$= 1 - \Pr(Z \leq 3.08) = 1 - 0.999 = 0.001.$$

Very unlikely to commit type I error.

Approximating the binomial distribution using the normal distribution (cont.)

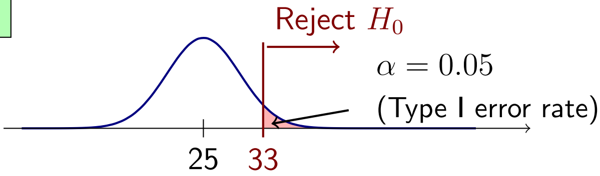
Case study continued

Testing against the alternate hypothesis $H_1 : p = 0.1$,

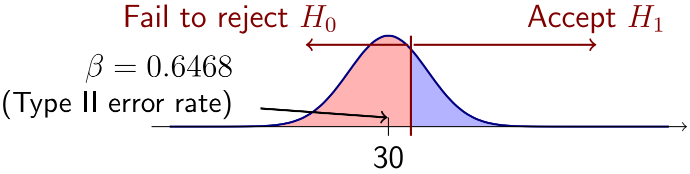
$$\begin{aligned}\beta &= \sum_{x=0}^{39} b(x; n = 500, p = 0.1)\\ &\approx \Pr\left(Z \leq \frac{39 - 500 \times 0.1}{\sqrt{500 \times 0.1 \times 0.9}}\right)\\ &= \Pr(Z \leq -1.69) = 0.0681.\end{aligned}$$

Visual interpretation with normal approximation

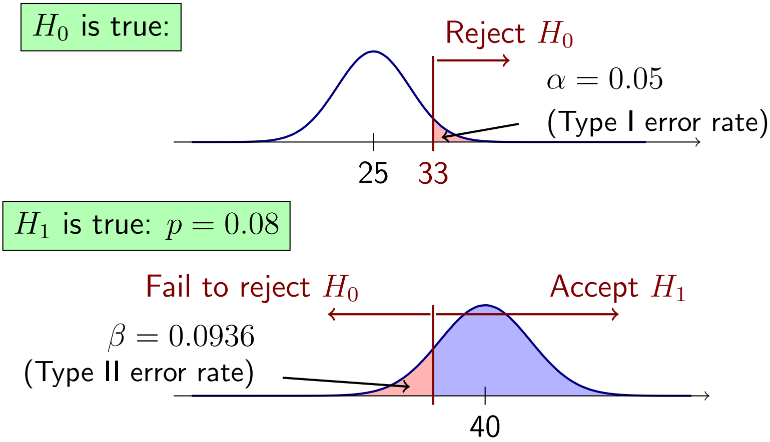
H_0 is true:



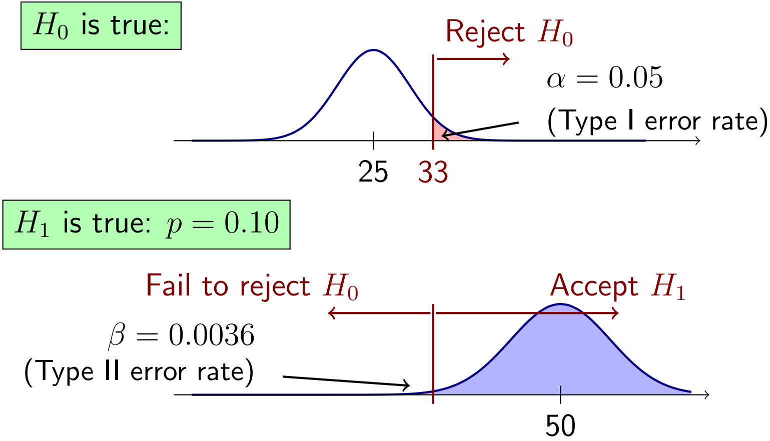
H_1 is true: $p = 0.06$



Visual interpretation with normal approximation



Visual interpretation with normal approximation



Power of a test

Definition

The *power* of a test is the probability of rejecting H_0 given that a specific alternate hypothesis is true. That is,

$$\text{Power} = 1 - \beta.$$

Summary

Properties of hypothesis testing

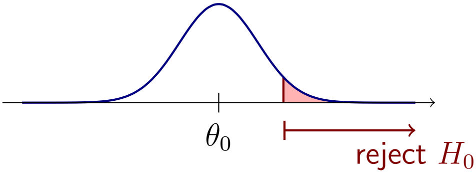
1. α and β are related; decreasing one generally increases the other.
2. α can be set to a desired value by adjusting the critical value. Typically, α is set at 0.05 or 0.01.
3. Increasing n decreases both α and β.
4. β decreases as the distance between the true value and hypothesized value (H_1) increases.

One-tailed vs. two-tailed tests

In our examples so far we have considered:

- H_0: $\theta = \theta_0$
- H_1: $\theta > \theta_0$.

This is a one-tailed test with the critical region in the right-tail of the test statistic X.

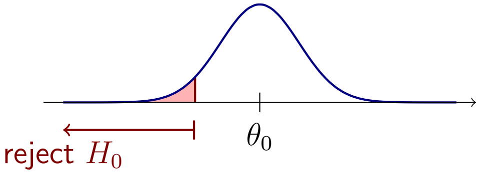


One-tailed vs. two-tailed tests (cont.)

Another one-tailed test could have the form,

- H_0: $\theta = \theta_0$
- H_1: $\theta < \theta_0$,

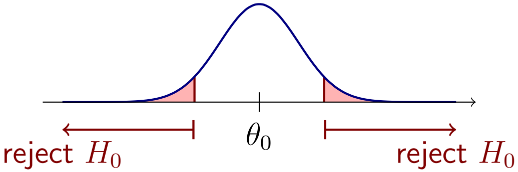
in which the critical region is in the left-tail.



One-tailed vs. two-tailed tests (cont.)

In a two-tailed test check for differences:

- H_0: $\theta = \theta_0$
- H_1: $\theta \neq \theta_0$,



Two-tailed test: example

Consider a production line of resistors that are supposed to be 100 Ohms. Assume $\sigma = 8$. So, the hypotheses are:

- H_0: $\mu = 100$
- H_1: $\mu \neq 100$,

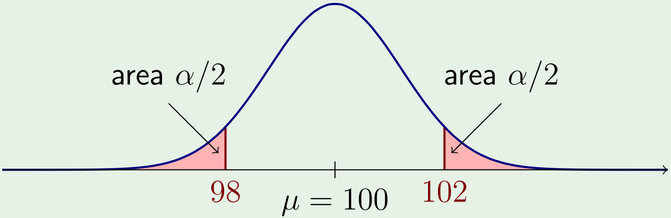
Let $\bar{X}$ be the sample mean for a sample of size $n = 100$.

Reject H_0	Do not reject H_0	Reject H_0

98 | 102

In this case the test statistic is the sample mean because this is a continuous random variable.

Two-tailed test: example (cont.)



We know the sampling distribution of $\bar{X}$ is a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n} = 0.8$ due to the central limit theorem.

Two-tailed test: example (cont.)

Therefore we can compute the probability of a type I error as

$$\alpha = \Pr(\bar{X} < 98 \text{ when } \mu = 100) + \Pr(\bar{X} > 102 \text{ when } \mu = 100)$$
$$= \Pr\left(Z < \frac{98 - 100}{8/\sqrt{100}}) + \Pr(Z > \frac{102 - 100}{8/\sqrt{100}}\right)$$
$$= \Pr(Z < -2.5) + \Pr(Z > 2.5)$$
$$= 2 \times \Pr(Z < -2.5) = 2 \times 0.0062 = 0.0124.$$

Confidence interval

Testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ at a significance level α is equivalent to computing a $100 \times (1 - \alpha)\%$ *confidence interval* for μ and H_0 if μ_0 is outside this interval.

Example

For the previous example the confidence interval at a significance level of $98.76\% = 100 \times (1 - 0.0124)$ is $[98, 102]$.

To be continued ...

Tests concerning sample mean
(variance known)

As in the previous example, we are often interested in testing

- H_0: $\mu = \mu_0$
- H_1: $\mu \neq \mu_0$,

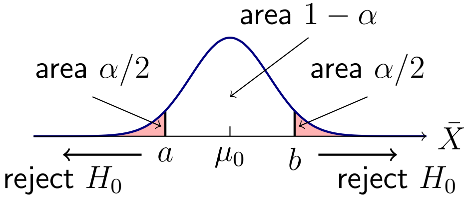
based on the sample mean $\bar{X}$ from samples $X_1, X_2, \ldots, X_n$, with *known* population variance σ^2.

Under $H_0 : \mu = \mu_0$, the probability of a type I error is computed using the sampling distribution of $\bar{X}$, which, due to the central limit theorem, is normal distributed with mean μ and standard deviation $\sigma/\sqrt{n}$.

Tests concerning sample mean (cont.)
(variance known)

From confidence intervals we know that

$$\Pr\left(-z_{\alpha/2} < \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$



Tests concerning sample mean (cont.)
(variance known)

Therefore, to design a test at the level of significance α we choose the critical values a and b as

$$a = \mu_0 - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$
$$b = \mu_0 + z_{\alpha/2} \frac{\sigma}{\sqrt{n}},$$

then we collect the sample, compute the sample mean $\bar{X}$ and reject H_0 if $\bar{X} < a$ **or** $\bar{X} > b$.

Tests concerning sample mean (cont.)
(variance known)

Steps in hypothesis testing

1. State the null and alternate hypothesis
2. Choose a significance level α
3. Choose the test statistic and establish the critical region
4. Collect the sample and compute the test statistic.
   If the test statistic is in the critical region, reject H_0.
   Otherwise, do not reject H_0.

Tests concerning sample mean (cont.)
(variance known)

Example

A batch of 100 resistors have an average of 102 Ohms. Assuming a population standard deviation of 8 Ohms, test whether the population mean is 100 Ohms at a significance level of $\alpha = 0.05$.

Step 1:

$$H_0 : \mu = 100$$
$$H_1 : \mu \neq 100,$$

Note: Unless stated otherwise, we use a two-tailed test.

Step 2: $\alpha = 0.05$

Tests concerning sample mean (cont.)

(variance known)

Example continued

Step 3: In this case, the test statistic is specified by the problem to be the sample mean $\bar{X}$.

Reject H_0 if $\bar{X} < a$ or $\bar{X} > b$, with

$$\begin{aligned} a &= \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \mu_0 - z_{0.025}\frac{\sigma}{\sqrt{100}} \\ &= 100 - 1.96\frac{8}{10} = 98.432 \\ b &= \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 100 + 1.96\frac{8}{10} = 101.568. \end{aligned}$$

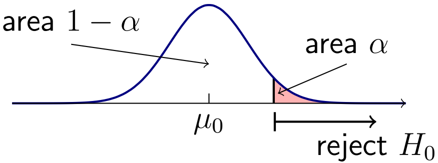
Step 4: We are told that the test statistic on a sample is $\bar{X} = 102 > b$. Therefore, reject H_0.

One-sided sample mean test
(variance known)

Case A:

In this case, we are interested in testing,

- H_0: $\mu = \mu_0$
- H_1: $\mu > \mu_0$.



One-sided sample mean test (cont.)

(variance known)

Under $H_0 : \mu = \mu_0$, the probability of a type I error is

$$\Pr\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_\alpha\right) = 1 - \alpha.$$

Thus, our decision becomes: reject H_0 at significance level α if

$$\bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}.$$

Note that we use z_α instead of $z_{\alpha/2}$, just as in one-tailed confidence intervals.

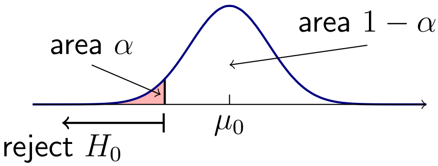
One-sided sample mean test (cont.)

(variance known)

Case B:

In this case, we are interested in testing,

- H_0: $\mu = \mu_0$
- H_1: $\mu < \mu_0$.



One-sided sample mean test (cont.)

(variance known)

Under $H_0 : \mu = \mu_0$, the probability of a type I error is

$$\Pr\left(-z_\alpha < \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right) = 1 - \alpha.$$

The decision becomes: reject H_0 at significance level α if

$$\bar{X} < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}.$$

One-sided sample mean test (cont.)

(variance known)

Example

A quality control engineer finds that a sample of 100 light bulbs had an average life-time of 470 hours. Assuming a population standard deviation of $\sigma = 25$ hours, test whether the population mean is 480 hours vs. the alternative hypothesis $\mu < 480$ at a significance level of $\alpha = 0.05$.

Step 1:

$$H_0 : \mu = 480$$
$$H_1 : \mu < 480,$$

Step 2: $\alpha = 0.05$

One-sided sample mean test (cont.)

(variance known)

Example continued

Step 3: The test statistic is the sample mean $\bar{X}$. Reject H_0 if

$$\bar{X} < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}} = 480 - 1.645 \frac{25}{10} = 475.9$$

Step 4: Since $\bar{X} = 470 < 475.9$, we reject H_0.

Tests concerning sample mean
(variance unknown)

As before, we are often interested in testing

- H_0: $\mu = \mu_0$
- H_1: $\mu \neq \mu_0$,

based a sample $X_1, X_2, \ldots, X_n$, but now with *unknown* variance σ^2. For our decision we use the sample mean $\bar{X}$ and the sample variance s^2.

We know that in this case the sampling distribution for $\bar{X}$ is the t-distribution.

Tests concerning sample mean (cont.)
(variance unknown)

Critical region at significance level α is, $\bar{X} < a$ or $\bar{X} > b$, where

$$a = \mu_0 - t_{\alpha/2} \frac{s}{\sqrt{n}}$$
$$b = \mu_0 + t_{\alpha/2} \frac{s}{\sqrt{n}},$$

where $t_{\alpha/2}$ had $v = n - 1$ degrees of freedom.

Equivalently, let $T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$. Reject H_0 if $T < -t_{\alpha/2}$ or $T > t_{\alpha/2}$, for $v = n - 1$ degrees of freedom.

For one-sided tests, $t_{\alpha/2}$ is replaced by t_α as usual.

Tests concerning sample mean (cont.)
(variance unknown)

Example 10.5 from the textbook

It is claimed that a vacuum cleaner expends 46 kWh per year. A random sample of 12 homes indicates that vacuum cleaners expend an average of 42 kWh per year with (sample) standard deviation 11.9 kWh. At a 0.05 level of significance, does this suggest that, on average, vacuum cleaner expend less than 46 kWh per year? Assume the population to be normally distributed.

Tests concerning sample mean (cont.)
(variance unknown)

Example solution:

Step 1:

$$H_0 : \mu = 46 \text{ kWh}$$
$$H_1 : \mu < 46 \text{ kWh},$$

Step 2: $\alpha = 0.05$

Tests concerning sample mean (cont.)

(variance unknown)

Example solution: continued

Step 3: The test statistic is $T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$.
Reject H_0 if $T < -t_{0.05}$ for $v = n - 1 = 11$ degrees of freedom; that is, reject H_0 if $T < -1.796$.

Step 4: We have that $\bar{X} = 42$, $s = 11.9$ and $n = 12$. So,

$$T = \frac{42 - 46}{11.9/\sqrt{12}} = -1.16 > -1.796.$$

Do not reject H_0.

Hypothesis testing using the p-value

In the approach we have taken so far, the significance level is pre-selected up front, either by choosing a given value or setting the critical region explicitly. In this case, the final outcome is the decision.

Now suppose a hypothesis test is performed at a significance level of 0.05, but someone else wants to test with a stricter significance level of 0.01. This requires recomputing the critical region.

The p-value aims to provide more information about the test statistic with regards to the hypothesis test.

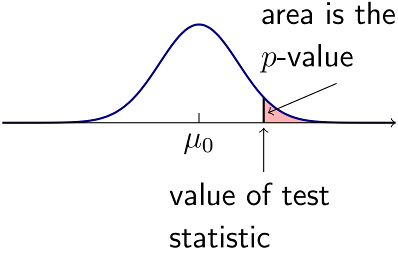
Hypothesis testing using the p-value (cont.)

Definition

The p-value is the lowest level of significance at which the observed value of a test statistic is significant (i.e., one rejects H_0).

Hypothesis testing using the p-value (cont.)

Alternative interpretation: the p-value is the minimum probability of a type I error with which H_0 can still be rejected.



Hypothesis testing using the p-value (cont.)

Example

Suppose that, for a given hypothesis test, the p-value is 0.09. Can H_0 be rejected?

Depends! At a significance level of 0.05, we cannot reject H_0 because $p = 0.09 > 0.05$. However, for significance levels greater or equal to 0.09, we can reject H_0.

Hypothesis testing using the p-value (cont.)

Example

A batch of 100 resistors have an average of 101.5 Ohms. Assuming a population standard deviation of 5 Ohms:

(a) Test whether the population mean is 100 Ohms at a level of significance 0.05.

(b) Compute the p-value.

Hypothesis testing using the p-value (cont.)

Example continued

(a) $H_0 : \mu = 100, \quad H_1 : \mu \neq 100$

Test statistic is $\bar{X}$. Reject H_0 if

$$\bar{X} < 100 - z_{0.025}\frac{\sigma}{\sqrt{n}} = 100 - 1.96 \times \frac{5}{10} = 99.02$$

or

$$\bar{X} > 100 + z_{0.025}\frac{\sigma}{\sqrt{n}} = 100 + 1.96 \times \frac{5}{10} = 100.98$$

$\bar{X} = 101.5$ therefore, reject H_0.

Hypothesis testing using the p-value (cont.)

Example continued

(b) The *observed* z-value is

$$Z = \frac{\bar{X} - 100}{\sigma/\sqrt{n}} = \frac{101.5 - 100}{5/10} = 3.$$

Then, the p-value is

$$p = 2\,\Pr(Z > 3) = 2 \times 0.0013 = 0.0026.$$

This means that H_0 could have been rejected at significance level $\alpha = 0.0026$ which is much stronger than rejecting it a 0.05.

Hypothesis testing using the p-value (cont.)

Example continued