8 Hypothesis Tests for One Sample

Chapter 8 Stat 4570/5570

Material from Devore's book (Ed 8), and Cengage

Statistical Hypotheses

A **statistical hypothesis:** a claim about the value of a parameter, population characteristic (could be a combination of parameters), or about the form of an entire probability distribution.

Examples:

- *H:* μ = 75 cents, where μ is the true population average of daily per-student candy+soda expenses in US high schools
- *H: p* < .10, where *p* is the population proportion of defective helmets for a given manufacturer
- If μ_1 and μ_2 denote the true average breaking strengths of two different types of twine, one hypothesis might be the assertion that $\mu_1 - \mu_2 = 0$, and another is the statement $\mu_1 - \mu_2 > 5$

Null vs Alternative Hypotheses

In any hypothesis-testing problem, there are always two competing hypotheses under consideration:

1. The status quo (null) hypothesis
2. The research (alternative) hypothesis

For example,

$\mu = .75$ versus $\mu \neq .75$

$p \geq .10$ versus $p < .10$

The objective of **hypothesis testing** is to decide, based on sample information, *if the alternative hypotheses is actually supported by the data.*

We usually do new research to challenge the existing (accepted) beliefs.

Burden of Proof

The **burden of proof** is placed on those who believe in the alternative claim.

In testing statistical hypotheses, the problem will be formulated so that one of the claims is initially favored.

This initially favored claim (H_0) will not be rejected in favor of the alternative claim (H_a or H_1) unless the sample evidence contradicts it and provides strong support for the alternative assertion.

If the sample does not strongly contradict H_0, we will continue to believe in the plausibility of the null hypothesis.

The two possible conclusions: ***1) reject H_0***

2) fail to reject H_0.

No proof… only evidence

We can never prove that a hypothesis is true or not true.

We can only conclude that it **is** or **is not** *supported by the data*.

A **test of hypotheses** is a method for *using sample data to decide whether the null hypothesis should be rejected* in favor of the alternative.

Thus we might test the null hypothesis H_0: $\mu = .75$ against the alternative H_a: $\mu \neq .75$. Only if sample data strongly suggests that μ is something other than 0.75 should the null hypothesis be rejected.

In the absence of such evidence, H_0 should not be rejected, since it is still considered plausible.

Why favor the null so much?

Why be so committed to the null hypothesis?

- sometimes we do not want to accept a particular assertion unless (or until) data can show strong support
- reluctance (cost, time) to change

Example: Suppose a company is considering putting a new type of coating on bearings that it produces.

The true average wear life with the current coating is known to be 1000 hours. With μ denoting the true average life for the new coating, the company would not want to make any (costly) changes unless evidence strongly suggested that μ exceeds 1000.

Hypotheses and Test Procedures

An appropriate problem formulation would involve testing H_0: $\mu = 1000$ against H_a: $\mu > 1000$.

The conclusion that a change is justified is identified with H_a, and it would take conclusive evidence to justify rejecting H_0 and switching to the new coating.

Scientific research often involves trying to decide whether a current theory should be replaced, or "elaborated upon."

Hypotheses and Test Procedures

An appropriate problem formulation would involve testing the hypothesis:

$$H_0: \mu = 1000 \text{ against } H_a: \mu > 1000.$$

The conclusion that "a change is justified" is identified with H_a, and it would take conclusive evidence to justify rejecting H_0 and switching to the new coating.

Scientific research often involves trying to decide whether a current theory should be replaced, or "elaborated upon"

Hypotheses and Test Procedures

The word *null* means "of no value, effect, or consequence," which suggests that H_0 should be identified with the hypothesis of no change (from current opinion), no difference, no improvement, etc.

Example: 10% of all circuit boards produced by a certain manufacturer during a recent period were defective.

An engineer has suggested a change in the production process in the belief that it will result in a reduced defective rate. Let p denote the true proportion of defective boards resulting from the changed process. What does the hypothesis look like?

Hypotheses and Test Procedures

The alternative to the null hypothesis H_0: $\theta = \theta_0$ will look like one of the following three assertions:

1. H_a: $\theta \neq \theta_0$
2. H_a: $\theta > \theta_0$ (in which case the null hypothesis is $\theta \leq \theta_0$)
3. H_a: $\theta < \theta_0$ (in which case the null hypothesis is $\theta \geq \theta_0$)

- The equality sign is **always** with the null hypothesis.
- It is typically easier to determine the alternate hypothesis first then the complementary statement is designated as the null hypothesis
- The alternate hypothesis is the claim for which we are seeking statistical proof

Test Procedures

A **test procedure** is a rule, based on sample data, for deciding whether to reject H_0.

Example -- the circuit board problem:

A test of H_0: $p = .10$ versus H_a: $p < .10$

We test this on a random sample of $n = 200$ boards.

How do we use the sample of 200?

Test Procedures

Testing procedure has two constituents:

(1) a *test statistic*, or function of the sample data which will be used to make a decision, and

(2) a *rejection (or critical) region* consisting of those test statistic values for which H_0 will be rejected in favor of H_a.

So if we have decided we can reject H_0 if $x \leq 15$ – then the rejection region consists of {0, 1, 2,…, 15}. Then H_0 will not be rejected if x = 16, 17,. . . ,199, or 200.

Errors in Hypothesis Testing

The basis for choosing a particular rejection region lies in consideration of the errors that one might be faced with in drawing a conclusion.

Consider the rejection region $x \leq 15$ in the circuit board problem. Even when H_0: $p = .10$ is true, it might happen that an unusual sample results in $x = 13$, so that H_0 is erroneously rejected.

On the other hand, even when H_a: $p < .10$ is true, an unusual sample might yield $x = 20$, in which case H_0 would not be rejected—again an incorrect conclusion.

Errors in Hypothesis Testing

Definition

- A **type I error** is when the null hypothesis is rejected, but it is true.
- A **type II error** is not rejecting H_0 when H_0 is false.

This is very similar in spirit to our diagnostic test examples

- False negative test = type I error
- False positive test = type II error

Type I error in hypothesis testing

Usually: Specify the largest value of α that can be tolerated, and then find a rejection region with that α.

The resulting value of α is often referred to as the **significance level** of the test.

Traditional levels of significance are .10, .05, and .01, though the level in any particular problem will depend on the seriousness of a type I error—

The more serious the type I error, the smaller the significance level should be.

Example (Type I Error)

Let μ denote the true average nicotine content of brand B cigarettes. The objective is to test

H_o: $\mu = 1.5$ versus H_a: $\mu > 1.5$

based on a random sample $X_1, X_2, \ldots, X_{32}$ of nicotine content.

Suppose the distribution of nicotine content is known to be normal with $\sigma = .20$.

Then $\bar{X}$ is normally distributed with mean value $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_x = .20/\sqrt{32} = .0354$.

Example (Type I Error)

cont'd

Rather than use $\overline{X}$ itself as the test statistic, let's standardize X, assuming that H_0 is true.

$$\text{Test statistic: } Z = \frac{\overline{X} - 1.5}{\sigma/\sqrt{n}} = \frac{\overline{X} - 1.5}{.0354}$$

***Z* expresses the distance between $\overline{X}$ and its expected value (when H_0 is true) as some number of standard deviations of the sample mean.**

Example (Type I Error)

cont'd

As H_a: $\mu > 1.5$, the form of the rejection region is $z \geq c$. What is c so that $\alpha = 0.05$?

When H_0 is true, Z has a standard normal distribution. Thus

$\alpha = P(\text{type I error}) = P(\text{rejecting } H_0 \text{ when } H_0 \text{ is true})$

$= P(Z \geq c \text{ when } Z \sim N(0, 1))$

The value c must capture upper-tail area .05 under the z curve. So, $c = z_{.05} = 1.645$.

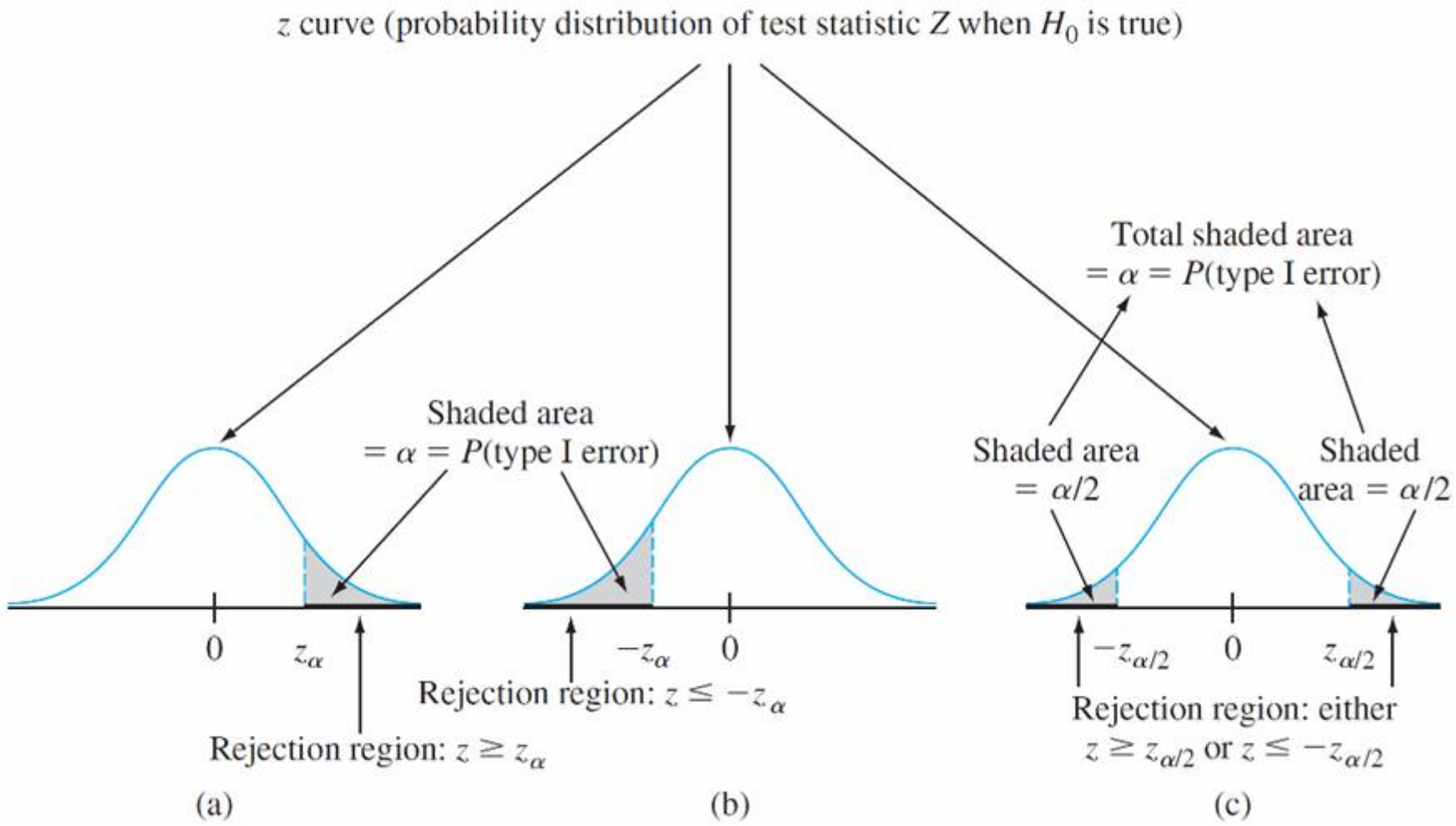
Case I: Testing means of a normal population with known σ

Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic value : $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

Alternative Hypothesis	**Rejection Region for Level α Test**
$H_a: \mu > \mu_0$	$z \geq z_\alpha$ (upper-tailed test)
$H_a: \mu < \mu_0$	$z \leq -z_\alpha$ (lower-tailed test)
$H_a: \mu \neq \mu_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (two-tailed test)

Case I: Testing means of a normal population with known σ



Rejection regions for *z* tests: (a) upper-tailed test; (b) lower-tailed test; (c) two-tailed test

Type II Error Example

A certain type of automobile is known to sustain no visible damage 25% of the time in 10-mph crash tests. A modified bumper design has been proposed in an effort to increase this percentage.

Let *p* denote the proportion of all 10-mph crashes with this new bumper that result in no visible damage.

How do we examine a hypothesis test for n = 20 independent crashes with the new bumper design?

Type II Error Example 1

cont' d

The accompanying table displays β for selected values of p (each calculated for the rejection region R_8).

p	.3	.4	.5	.6	.7	.8
$\beta(p)$	.772	.416	.132	.021	.001	.000

Clearly, β decreases as the value of p moves farther to the right of the null value .25.

Intuitively, the greater the departure from H_0, the less likely it is that such a departure will not be detected.

Thus, 1- β is often called the "power of the test"

Errors in Hypothesis Testing

We can also obtain a smaller value of α -- the probability that the null will be incorrectly rejected – by decreasing the size of the rejection region.

However, this results in a larger value of β for all parameter values consistent with H_a.

No rejection region that will simultaneously make both α and all β's small. A region must be chosen to strike a compromise between α and β.

Case II: Large sample tests for means

When the sample size is large, the *z* tests for case I are easily modified to yield valid test procedures without requiring either a normal population distribution or known σ.

Earlier we used the key result to justify large-sample confidence intervals:

A large *n (>40)* implies that the standardized variable

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

has *approximately* a standard normal distribution.

Case III: Testing means of a **Normal** population with unknown σ, and small n

The One-Sample t Test

Null hypothesis: H_0: $\mu = \mu_0$

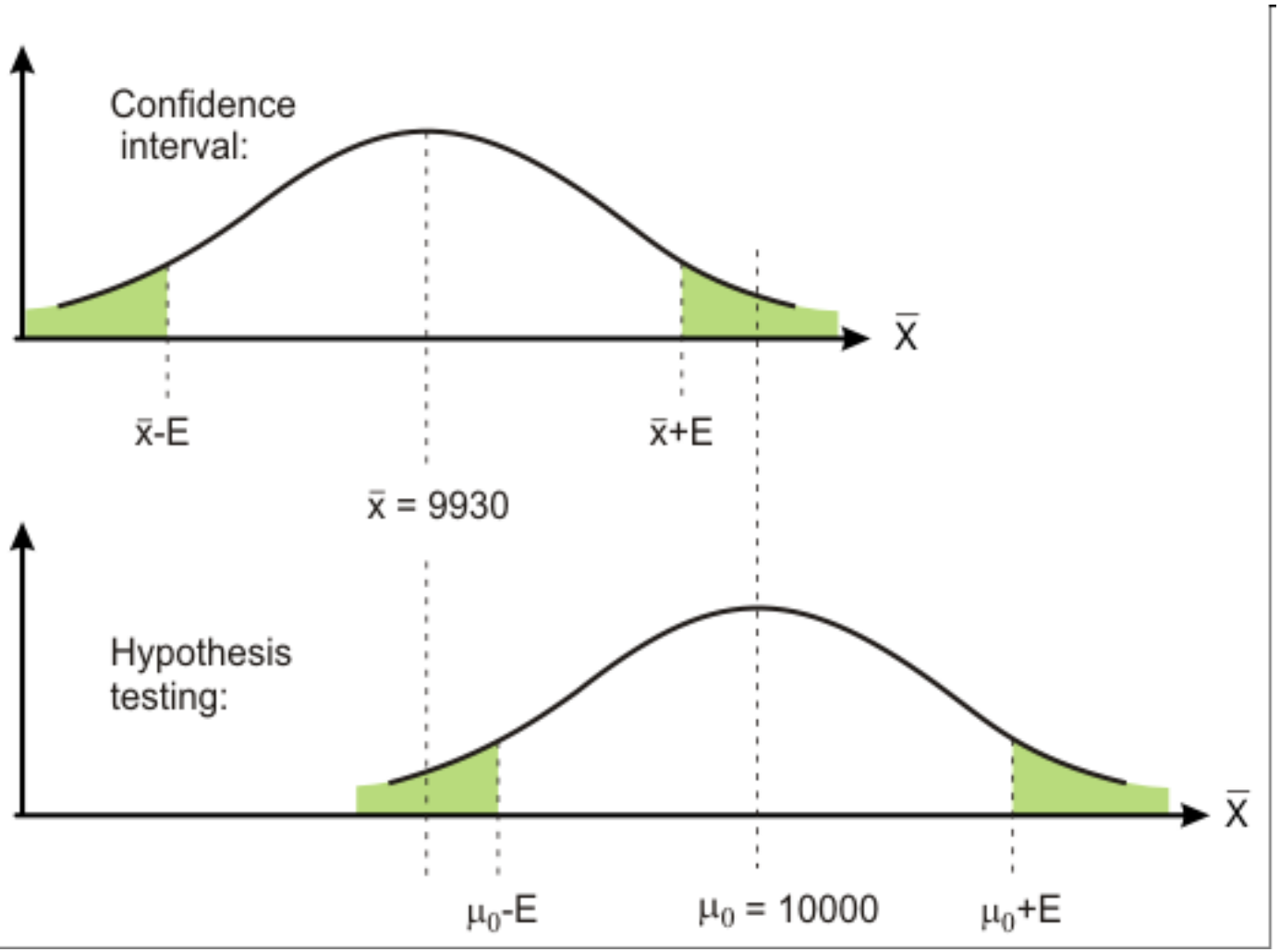
Test statistic value: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Alternative Hypothesis	Rejection Region for a Level α Test
H_a: $\mu > \mu_0$	$t \geq t_{\alpha,n-1}$ (upper-tailed)
H_a: $\mu < \mu_0$	$t \leq -t_{\alpha,n-1}$ (lower-tailed)
H_a: $\mu \neq \mu_0$	either $t \geq t_{\alpha/2,n-1}$ or $t \leq -t_{\alpha/2,n-1}$ (two-tailed)

CI and Hypotheses

cont' d

Rejection regions have a lot in common with confidence intervals.



Source:

Proportions: Large-Sample Tests

The estimator $\hat{p} = X/n$ is unbiased $(E(\hat{p}) = p)$, has approximately a normal distribution, and its standard deviation is $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$.

When H_0 is true, $E(\hat{p}) = p_0$ and $\sigma_{\hat{p}} = \sqrt{p_0(1-p_0)/n}$, so $\sigma_{\hat{p}}$ does not involve any unknown parameters. It then follows that when n is large and H_0 is true, the test statistic

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

has approximately a standard normal distribution.

Proportions: Large-Sample Tests

Alternative Hypothesis	**Rejection Region**
$H_a: p > p_0$	$z \geq z_\alpha$ (upper-tailed)
$H_a: p < p_0$	$z \leq -z_\alpha$ (lower-tailed)
$H_a: p \neq p_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (two-tailed)

These test procedures are valid provided that $np_0 \geq 10$ and $n(1 - p_0) \geq 10$.

Example

Natural cork in wine bottles is subject to deterioration, and as a result wine in such bottles may experience contamination.

The article "Effects of Bottle Closure Type on Consumer Perceptions of Wine Quality" (*Amer. J. of Enology and Viticulture*, 2007: 182–191) reported that, in a tasting of commercial chardonnays, 16 of 91 bottles were considered spoiled to some extent by cork-associated characteristics.

Does this data provide strong evidence for concluding that more than 15% of all such bottles are contaminated in this way? Use a significance level equal to 0.10.

P-Values

The *P*-value is a probability of observing values of the test statistic that are as contradictory or even more contradictory to H_0 as the test statistic obtained in our sample.

- This probability is calculated assuming that the null hypothesis is true.
- Beware: The *P*-value is not the probability that H_0 is true, nor is it an error probability!
- The *P*-value is between 0 and 1.

Example

Urban storm water can be contaminated by many sources, including discarded batteries. When ruptured, these batteries release metals of environmental significance.

The article "Urban Battery Litter" (*J. of Environ. Engr.*, 2009: 46–57) presented summary data for characteristics of a variety of batteries found in urban areas around Cleveland.

A sample of 51 Panasonic AAA batteries gave a sample mean zinc mass of 2.06g and a sample standard deviation of 0.141g.

Does this data provide compelling evidence for concluding that the population mean zinc mass exceeds 2.0g?

P-Values

More generally, **the smaller the P-value, the more evidence there is in the sample data against the null hypothesis and for the alternative hypothesis.**

The p-value measures the "extremeness" of the sample.

That is, H_0 should be rejected in favor of H_a when the *P*-value is sufficiently small (such large sample statistic is unlikely if the null is true).

So what constitutes "sufficiently small"?

What is "extreme" enough?

Decision rule based on the *P*-value

Select a significance level α (as before, the desired *type I error probability*), then α defines the rejection region.

Then the decision rule is:

reject H_0 if P-value $\leq \alpha$

do not reject H_0 if P-value $> \alpha$

Thus if the P-value exceeds the chosen significance level, the null hypothesis cannot be rejected at that level.

Note, the P-value can be thought of as the smallest significance level at which H_0 can be rejected.

P-Values

In the previous example, we calculated the *P*-value = .0012. Then using a significance level of .01, we would reject the null hypothesis in favor of the alternative hypothesis because .0012 **≤** .01.

However, suppose we select a significance level of 0.001, which requires far more substantial evidence from the data before H_0 can be rejected. In that case we would not reject H_0 because .0012 **>** .001.

This is why we cannot change significance level after we see the data – NOT ALLOWED though tempting!

P-Values for *z* Tests

The calculation of the *P*-value depends on whether the test is upper-, lower-, or two-tailed.

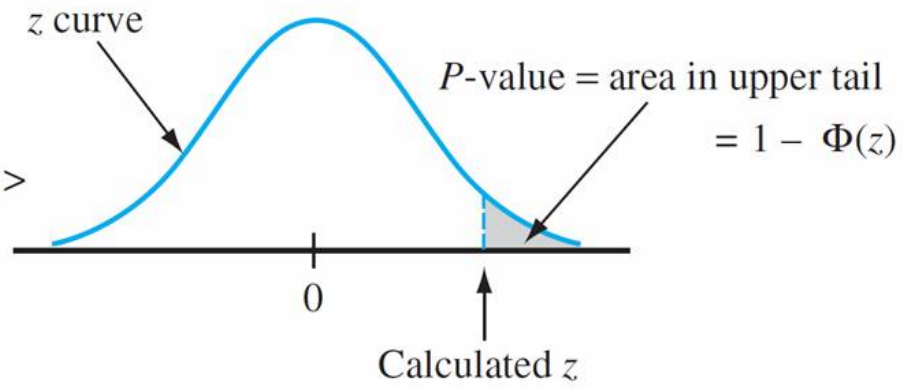
$$P\text{-value: } P = \begin{cases} 1 - \Phi(z) & \text{for an upper-tailed } z \text{ test} \\ \Phi(z) & \text{or an lower-tailed } z \text{ test} \\ 2[1 - \Phi(|z|)] & \text{for a two-tailed } z \text{ test} \end{cases}$$

Each of these is the probability of getting a value at least as extreme as what was obtained (assuming H_0 true).

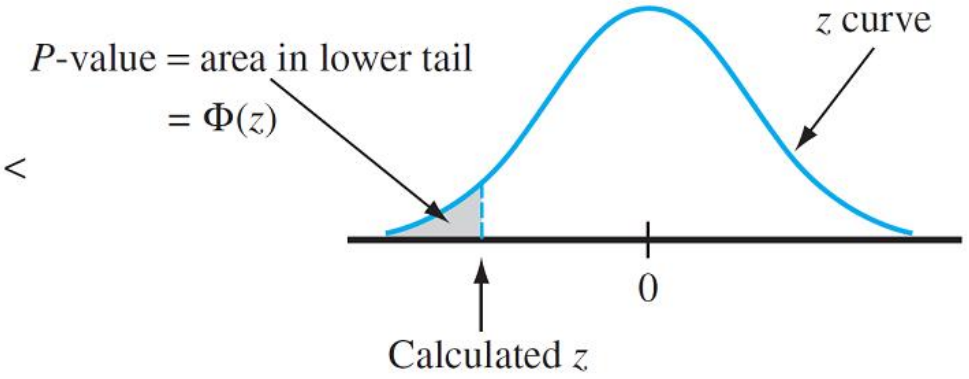
P-Values for *z* Tests

The three cases are illustrated in Figure 8.9.

1. **Upper-tailed test**
   H_a contains the inequality >



2. **Lower-tailed test**
   H_a contains the inequality <

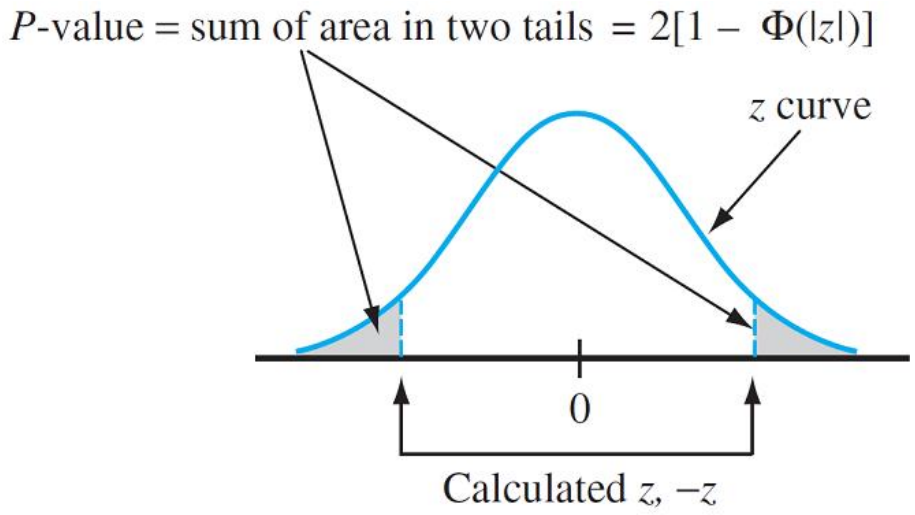


P-Values for *z* Tests

cont' d

3. **Two-tailed test**

   H_a contains the inequality $\neq$



Example

The target thickness for silicon wafers used in a certain type of integrated circuit is 245 μm.

A sample of 50 wafers is obtained and the thickness of each one is determined, resulting in a sample mean thickness of 246.18 μm and a sample standard deviation of 3.60 μm.

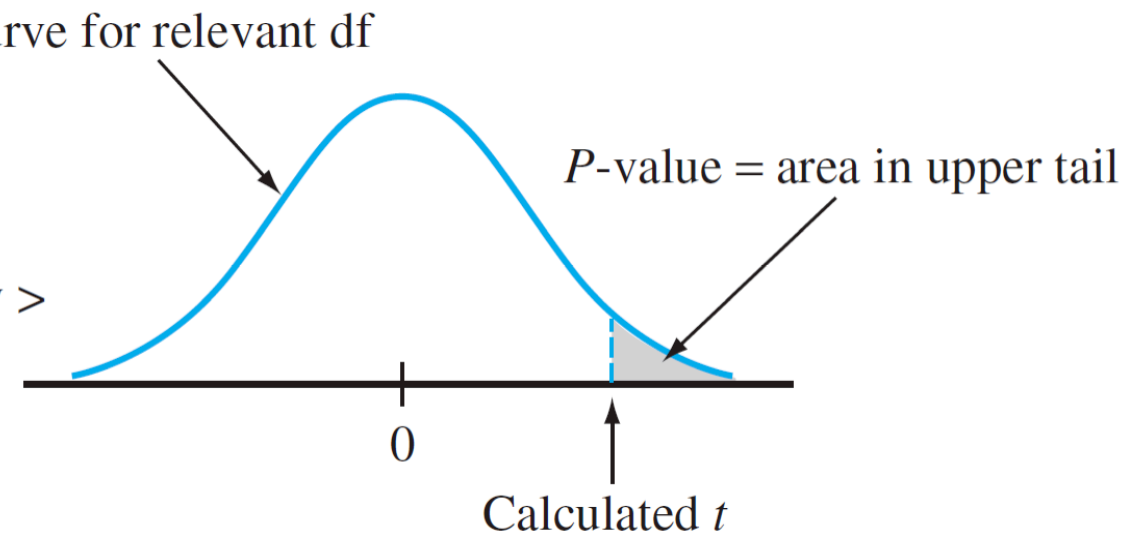
Does this data suggest that true average wafer thickness is something other than the target value? Use a significance level of .01.

P-Values for *t* Tests

Just as the *P*-value for a *z* test is the area under the *z* curve, the *P*-value for a *t* test will be the area under the *t*-curve.

The number of df for the one-sample *t* test is $n - 1$.

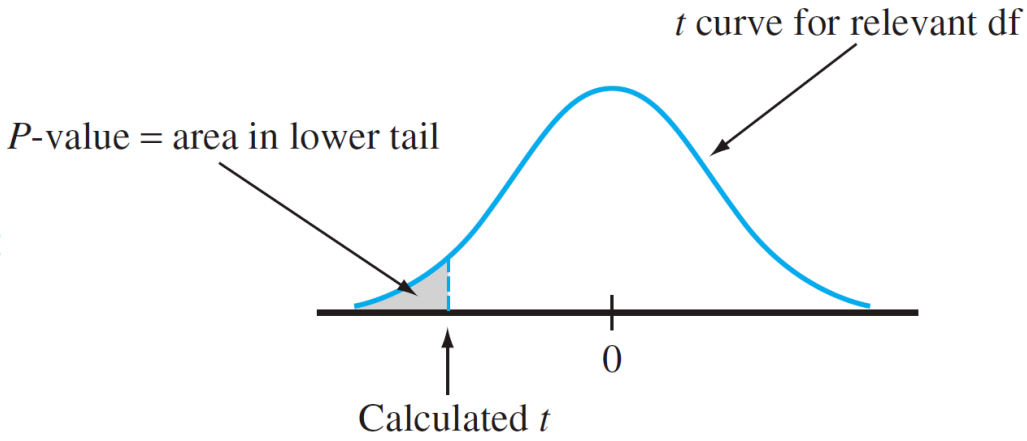
1. **Upper-tailed test**
   H_a contains the inequality >



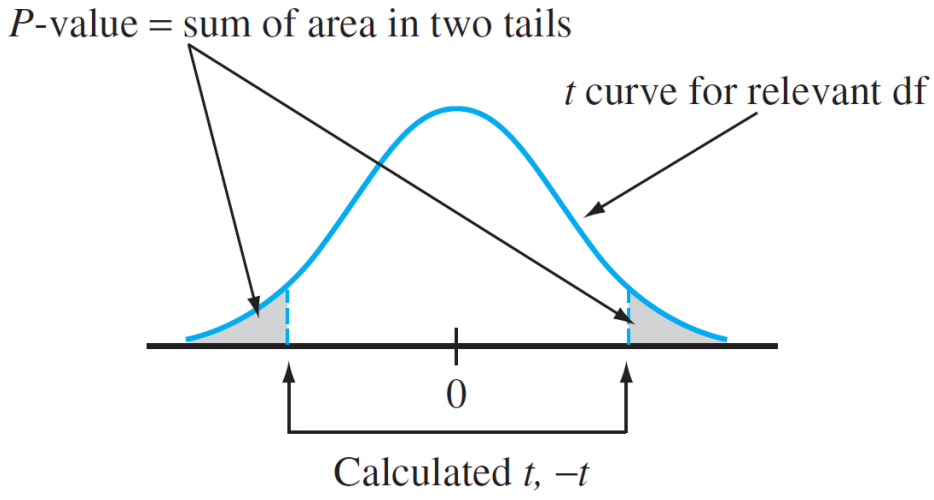
P-Values for *t* Tests

cont' d

2. **Lower-tailed test**
   H_a contains the inequality <



3. **Two-tailed test**
   H_a contains the inequality ≠



P-Values for *t* Tests

The table of *t* critical values used previously for confidence and prediction intervals doesn't contain enough information about any particular *t* distribution to allow for accurate determination of desired areas.

There another *t* table in Appendix Table A.8, one that contains a tabulation of upper-tail *t*-curve areas. But we can also use other tables to get an approximation of the p-value (software is the best).

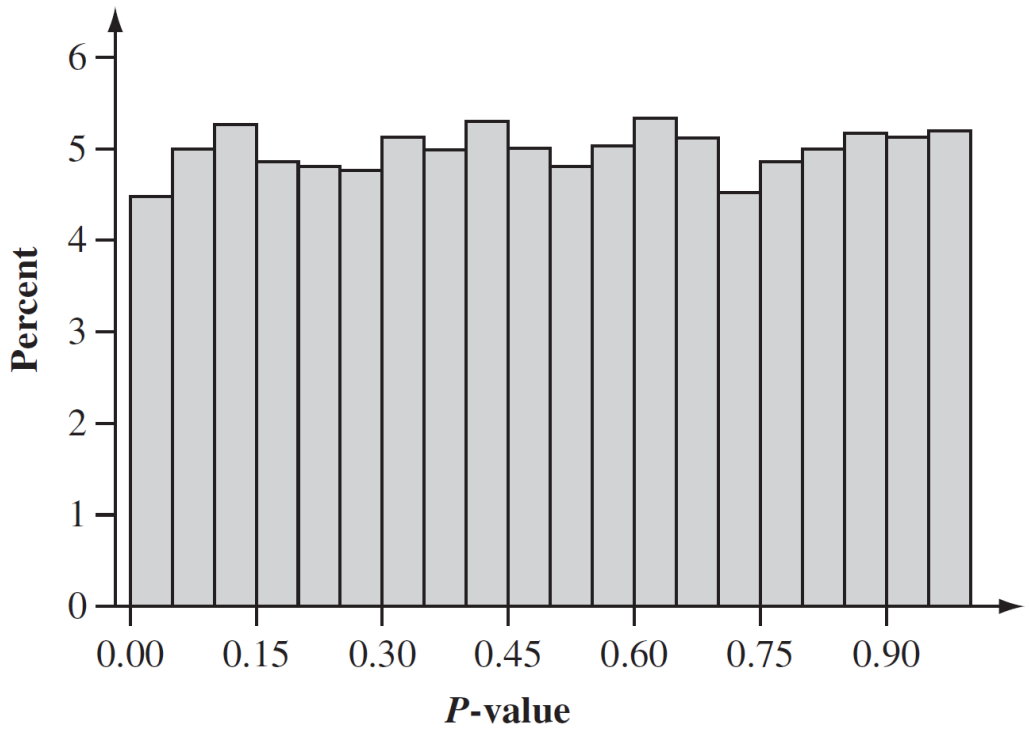
More on Interpreting *P*-values

How are p-values distributed?

cont' d

Figure below shows a histogram of the 10,000 *P*-values from a simulation experiment under a null $\mu = 20$ (with $n = 4$ and $\sigma = 2$).

When H_0 is true, the probability distribution of the *P*-value is a uniform distribution on the interval from 0 to 1.



Example

cont'd

About 4.5% of these *P*-values are in the first class interval from 0 to .05.

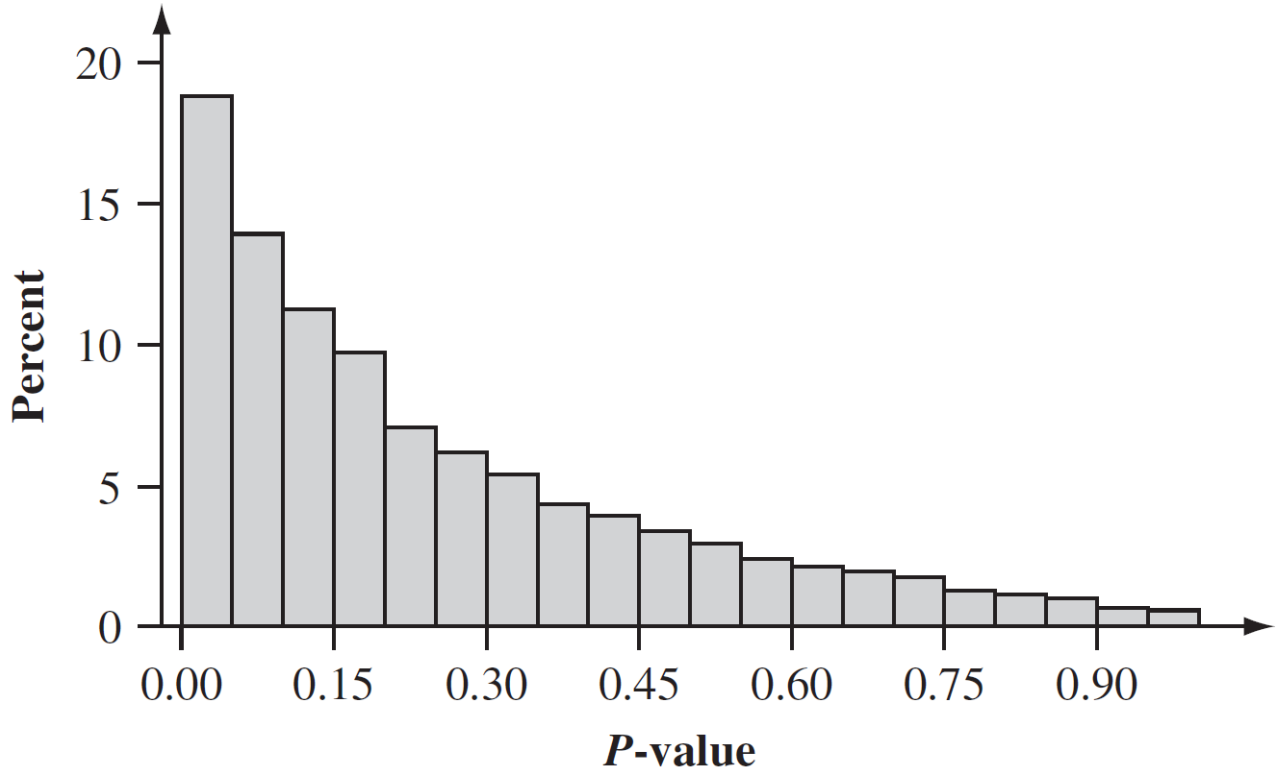
Thus when using a significance level of .05, the null hypothesis is rejected in roughly 4.5% of these 10,000 tests.

If we continued to generate samples and carry out the test for each sample at significance level .05, in the long run 5% of the *P*-values would be in the first class interval.

Example

cont' d

A histogram of the *P*-values when we simulate under an alternative hypothesis. There is a much greater tendency for the *P*-value to be small (closer to 0) when $\mu = 21$ than when $\mu = 20$.



(b) $\mu = 21$

Example

cont'd

Again H_0 is rejected at significance level .05 whenever the *P*-value is at most .05 (in the first bin).

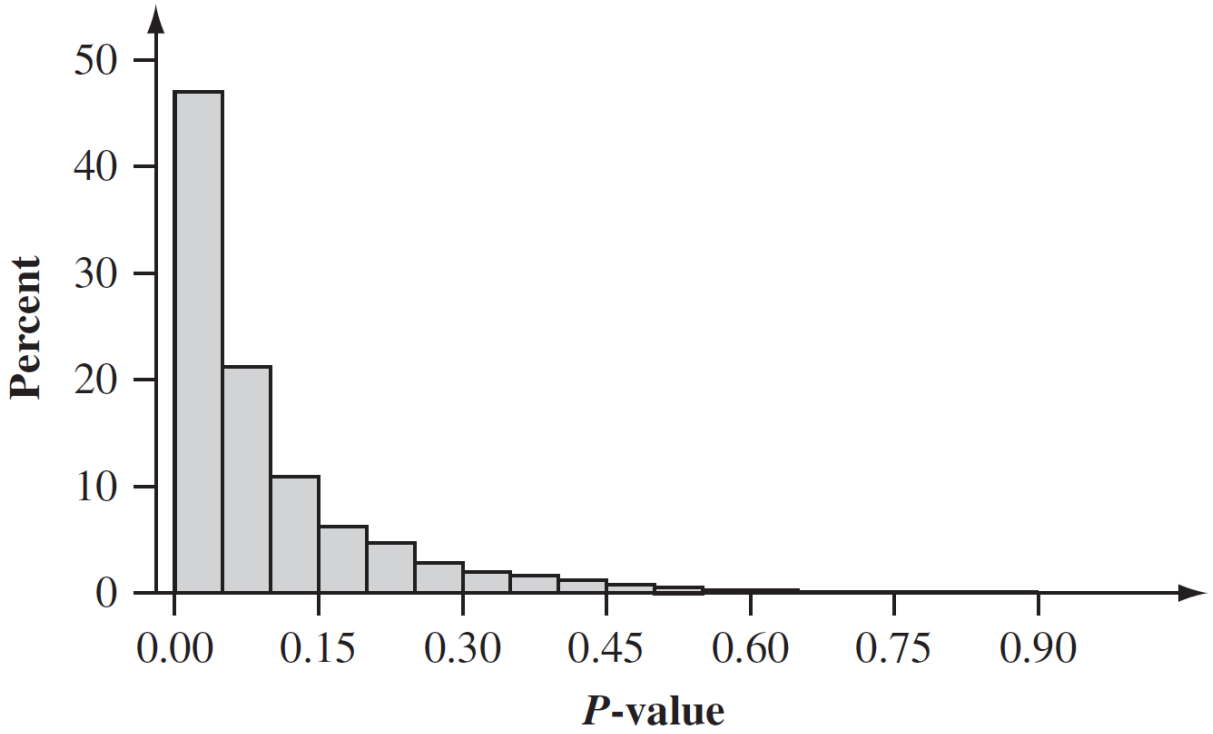
Unfortunately, this is the case for only about 19% of the *P*-values. So only about 19% of the 10,000 tests correctly reject the null hypothesis; for the other 81%, a type II error is committed.

The difficulty is that the sample size is quite small and 21 is not very different from the value asserted by the null hypothesis.

Example

cont'd

Figure below illustrates what happens to the *P*-value when H_0 is false because $\mu = 22$.



(c) $\mu = 22$

Example

cont'd

The histogram is even more concentrated toward values close to 0 than was the case when $\mu = 21$.

In general, as μ moves further to the right of the null value 20, the distribution of the *P*-value will become more and more concentrated on values close to 0.

Even here a bit fewer than 50% of the *P*-values are smaller than .05. So it is still slightly more likely than not that the null hypothesis is incorrectly not rejected. Only for values of μ much larger than 20 (e.g., at least 24 or 25) is it highly likely that the *P*-value will be smaller than .05 and thus give the correct conclusion.

Statistical Versus Practical Significance

When using

$$z = \frac{\bar{x} - 245}{s/\sqrt{n}}$$

one must be especially careful – with large n, what happens to z? How does this affect hypothesis testing?

R code