Unit #7 (a): Hypothesis Testing

9.1, 9.2, 9.5, 9.7.1, 9.7.2, 9.9.2

Learning Objectives

At the end of this unit, students should be able to:

1. Identify statistical hypotheses in a real-world scenario.
2. Characterize null and alternative hypotheses.
3. Articulate the logic of hypothesis testing.
4. Define a test statistic and describe how it is used in a hypothesis test.
5. Perform hypothesis tests for means and proportions.
6. Define a rejection region, critical value, significance level, type I error, and type II error.
7. Articulate the tradeoff between the rate of type I and type II errors.
8. Describe the relationship between hypothesis testing and confidence intervals.
9. Define and properly interpret p-values. Identify some common misinterpretations.
10. Use p-values to make decisions in hypothesis tests

Statistical Hypotheses

Statistical hypothesis: a claim about the value of a parameter or population characteristic.

Examples:

1.) $\mu = \mu_0$ where μ is a pop. mean and $\mu_0 \in \mathbb{R}$.

2.) $p \leq 0.1$ where p is some proportion (e.g. % of defective parts)

3.) $\mu_1 - \mu_2 = C$, $C \in \mathbb{R}$

Statistical Hypotheses

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

1.) The null hypothesis, H_0 (represents the "status quo")

2.) The alternative hypothesis, H_1 or H_a

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

The sample information is summarized by a **test statistic.** / summary of sample info

Logic of Hypothesis Testing

Analogy: Jury in a criminal trial.

When a defendant is accused of a crime, the jury (is supposed to) presumes that she is not guilty (not guilty; that's the "null hypothesis").

Then, we gather evidence. If the evidence is seems implausible under the assumption of non-guilt, we might reject non-guilt and claim that the defendant is (likely) guilty.

Logic of Hypothesis Testing

research hypothesis

Important Question: Is there strong evidence for the alternative? The burden of proof is placed on those who believe in the alternative claim.

The initially favored claim, the null hypothesis $H_{0,}$ will not be rejected in favor of the alternative hypothesis, H_a or $H_{1,}$ unless the sample evidence provides a lot of support for the alternative.

The two possible conclusions:

1.) We reject the null, H_0 (in favor of H_1)

2.) we fail to reject the null H_0.

Logic of Hypothesis Testing

Why *assume* the null hypothesis?

1.) Sometimes we don't want to accept a particular claim unless data can show support in favor of it.

2.) Reluctant to change (e.g., b/c of cost or time)

Logic of Hypothesis Testing

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.

Logic of Hypothesis Testing

μ = mean wear life for new coating

An appropriate problem formulation would involve testing:

$H_0 : \mu \leq 1000$

$H_1 : \mu > 1000$

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.”

Logic of Hypothesis Testing

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

parameter

$\theta \in \mathbb{R}$

$H_0: \theta \leq \theta_0$

1.) $H_1: \theta > \theta_0$ (upper tailed test)

2.) $H_1: \theta < \theta_0$ (lower tailed test)

} one tailed tests

3.) $H_1: \theta \neq \theta_0$ (two tailed test)

The equality sign is *always* with the null hypothesis.

The alternate hypothesis is the claim for which we are seeking statistical evidence.

Test Statistics—The Evidence

Definition: A *test statistic* is a a quantity derived based on sample data and calculated under the null hypothesis. It is used in a decision about whether to reject H_0.

We can think of a test statistic as our evidence. Next, we need to quantify whether we think our evidence is "rare" under the null hypothesis.

Test Statistics—The Evidence

Example: Company A produces circuit boards, but 10% of them are defective. Company B claims that they produce fewer defective circuit boards. What are the null and alternative hypotheses?

$p_A = 0.1$

p_B

$H_0 : p_B \geq 0.1 = p_A$

$H_1 : p_B < 0.1$

Our data is a random sample of $n = 200$ boards from company B. What test procedure (or rule) could we devise to decide if the null hypothesis should be rejected?

$\hat{p}_B = \frac{\#\text{ of defective for B}}{n}$

$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1-\alpha$

Test Statistics—The Evidence

Which test statistic is “best”?

There are an infinite number of *possible* tests that could be devised, so we have to limit this in some way or total statistical madness will ensue!

In the previous example, we might use:

$$Z = \frac{\hat{P}_B - P_B}{\sqrt{\frac{P_B(1-P_B)}{n}}} \sim N(0,1)$$

Rejection Regions

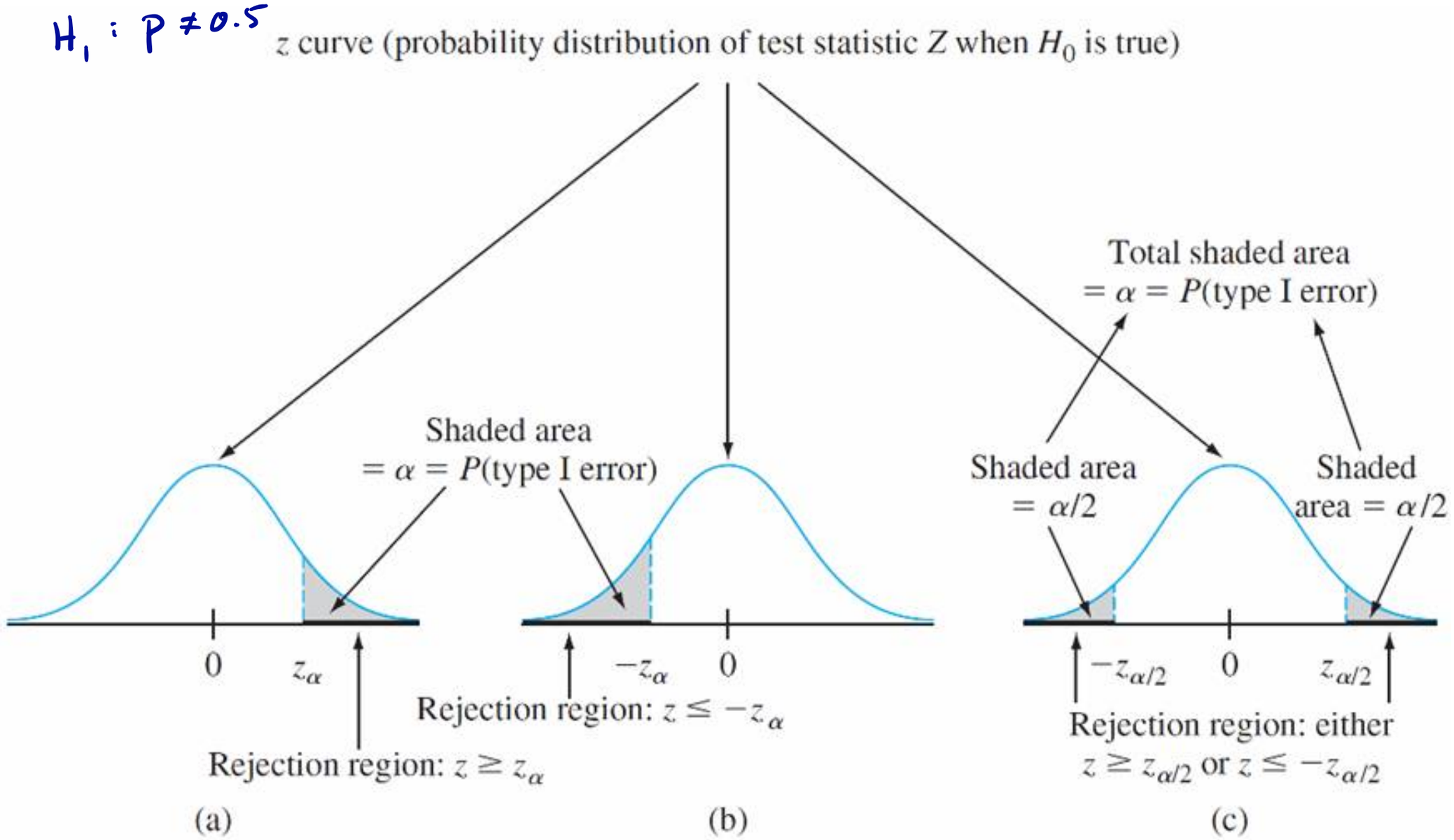
How would we know when the test statistic is "sufficiently rare" under the null hypothesis such that we might regard the null as false?

We could define a *rejection region*—a range of values that leads a researcher to *reject* the null hypothesis.

Rejection Regions

$H_0 : p = 0.5$

$H_1 : p \neq 0.5$



Test for Population Proportion

Example (continued): Company A produces circuit boards, but 10% of them are defective. Company B claims that they produce fewer defective circuit boards. What are the null and alternative hypotheses?

$H_0: p_B \geq 0.1$ lower-tailed test

$H_1: p_B < 0.1$

Suppose that in the sample of n = 200, 50 circuit boards from Company B were defective. Calculate the test statistic, and decide, based on a lower-tailed test with significance level 0.05, whether we should reject the null hypothesis.

$\alpha = 0.05$

$-z_\alpha = -z_{0.05} = \Phi^{-1}(0.05) = -1.64$

Note that $z = 7.1 > -2.57 = -z_\alpha$.

Thus, we fail to reject H_0.

$\hat{p}_B = \frac{50}{200} = \frac{1}{4}$

$$z = \frac{\hat{p}_B - p_B}{\sqrt{\frac{p_B(1-p_B)}{n}}} = 7.1$$

Test for Population Proportion

Null hypothesis: H_0 : $p = p_0$

Test statistic value: $$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0,1)$$

Alternative Hypothesis	Rejection Region for Level α Test
1.) $p > p_0$	$z \geq z_\alpha$
2.) $p < p_0$	$z \leq -z_\alpha$
3.) $p \neq p_0$	$z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$

Test for Population Mean (known variance)

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

Test statistic value : $Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$

Alternative Hypothesis	Rejection Region for Level α Test
$\mu > \mu_0$	$Z \geq Z_\alpha$
$\mu < \mu_0$	$Z \leq -Z_\alpha$
$\mu \neq \mu_0$	$Z \leq -Z_{\alpha/2}$ or $Z \geq Z_{\alpha/2}$

Test for Population Mean (known variance)

Example: Suppose a company is considering putting a new type of coating on bearings that it produces. Let μ denote the true mean life for the new coating. The company would not want to make any (costly) changes unless evidence strongly suggested that exceeds 1000 hours. State the null and alternative hypotheses:

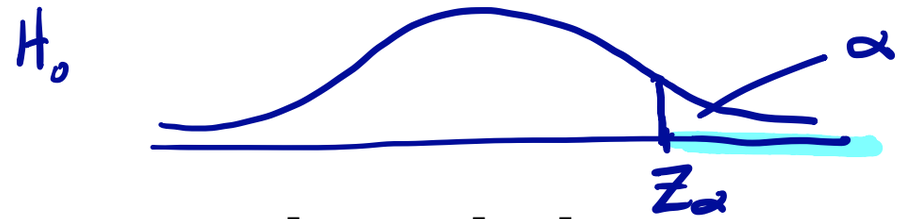
Now, suppose the company tests n = 25 randomly selected bearings with the new coating, and finds that the sample mean for life of the new coating is 1090 hours. Suppose we know that the population standard deviation is 130 hours. Conduct a test for the mean.

Errors in Hypothesis Testing

Definitions:

A *type I error* is when the null hypothesis is rejected, but it is true. (False Positive)

A *type II error* is not rejecting H_0 when H_0 is false. (False Negative)



Errors in Hypothesis Testing

sig. level

Typically, we specify the largest value of a type I error, α, that can be tolerated, and then find a rejection region with that α.

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

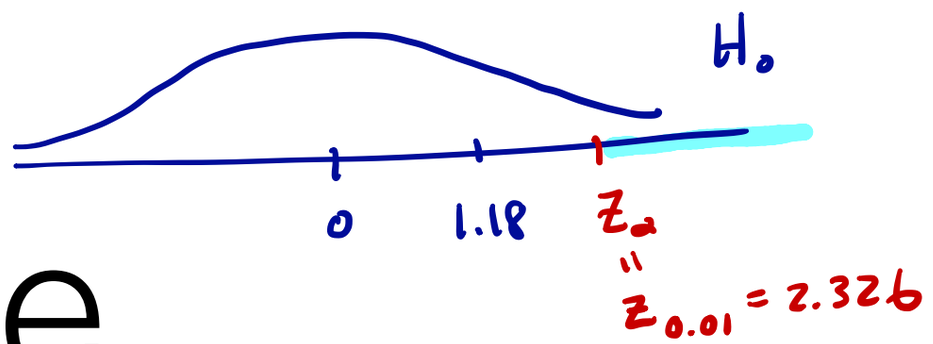
Traditional levels of significance are .10, .05, and .01, though the level in any particular problem will depend on the seriousness of a type I error. The more serious the type I error, the smaller the significance level should be.

Errors in Hypothesis Testing

We can also obtain a smaller value of α—the probability of a type I error—by decreasing the size of the rejection region.

However, this results in a larger value of β—*the* probability of a type II error—for all parameter values consistent with H_a.

No rejection region will simultaneously make α and β small at the same time. A region must be chosen to strike a compromise between these errors.



Practice

Example: An inventor has developed a new, energy-efficient lawn mower engine. He claims that the engine will run continuously for more than 5 hours (300 minutes) on a single gallon of regular gasoline. (The leading brand lawnmower engine runs for 300 minutes on 1 gallon of gasoline.)

From his stock of engines, the inventor selects a simple random sample of 50 engines for testing. The engines run for an average of 305 minutes. The true standard deviation σ is known and is equal to 30 minutes, and the run times of the engines are normally distributed.

Test hypothesis that the mean run time is more than 300 minutes. Use a 0.01 level of significance.

$H_0 : \mu \leq 300$

$H_1 : \mu > 300$

$n = 50$

$\alpha = 0.01$

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{305 - 300}{30/\sqrt{50}} \approx 1.18 < z_\alpha = 2.326$$

Since Z did not fall in the RR, we fail to reject the null.

Testing Means for a Large Sample

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

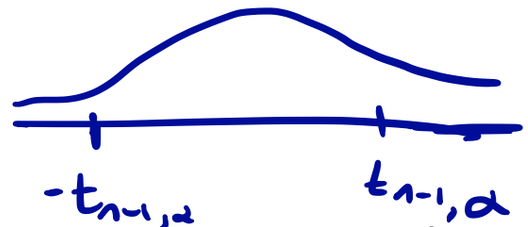
Earlier, we used the key result to justify large-sample confidence intervals:

A large *n (>30)* implies that the standardized variable

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

has *approximately* a standard normal distribution.

Testing Means for a Small Sample



When the sample size is small and the population is normal, we can use a t-test.

The One-Sample t Test

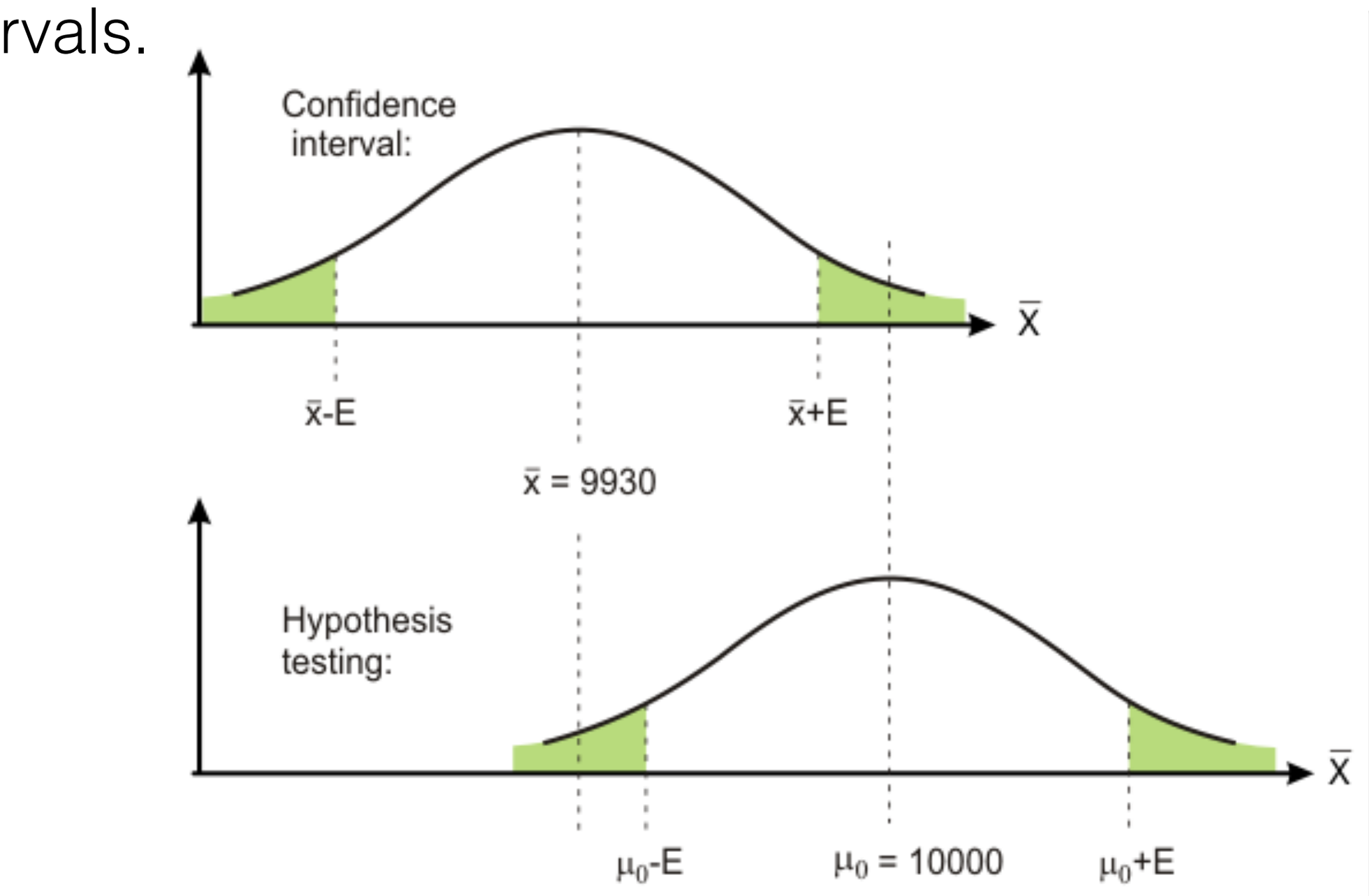
Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic value: $T = \dfrac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}$

Alternative Hypothesis	Rejection Region for a Level α	Test
$\mu > \mu_0$	$T \geq t_{n-1,\alpha}$	
$\mu < \mu_0$	$T \leq -t_{n-1,\alpha}$	
$\mu \neq \mu_0$	$T \leq -t_{n-1,\alpha/2}$ or $T \geq t_{n-1,\alpha/2}$	

CIs and Hypothesis Tests

Rejection regions have a lot in common with confidence intervals.

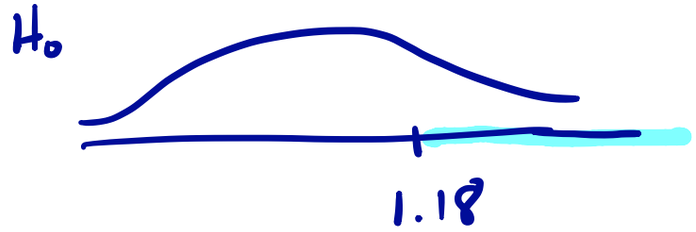


CIs and Hypothesis Tests

Example: The Brinell scale is a measure of how hard a material is. An engineer hypothesizes that the mean Brinell score of all subcritically annealed ductile iron pieces is not equal to 170. It is known that these scores follow a normal distribution.

The engineer measured the Brinell score of 25 pieces of this type of iron and calculated the sample mean to be 174.52 and the sample standard deviation to be 10.31.

Perform a hypothesis test that the true average Brinell score is not equal to 170, as well as the corresponding confidence interval. Set alpha = 0.01.



P-Values

Depends on the type of test

The p-value measures the "extremeness" of the test statistic.

Definition: A *p-value* is the probability**, under the null hypothesis,** that we would get a test statistic *at least as extreme as the one we calculated.*

So, the smaller the p-value, the more evidence there is in the sample data against the null hypothesis (so the story goes…).

So what constitutes "sufficiently small" and "extreme enough" to make a decision about the null hypothesis?

P-Values

Select a significance level (as before, the desired *type I error probability*), then defines the rejection region. Then the decision rule is:

If $p \leq \alpha$ then we reject the null.

If $p > \alpha$ then we fail to reject the null.

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.

(Frequentist)

P-Values

The p-value measures the "extremeness" of the test statistic.

Note:

1. This probability is calculated assuming that the null hypothesis is true.

2. Beware: The *p*-value is not the probability that H_0 is true, nor is it an error probability!

3. The *p*-value is between 0 and 1.

P-Values for Z Tests

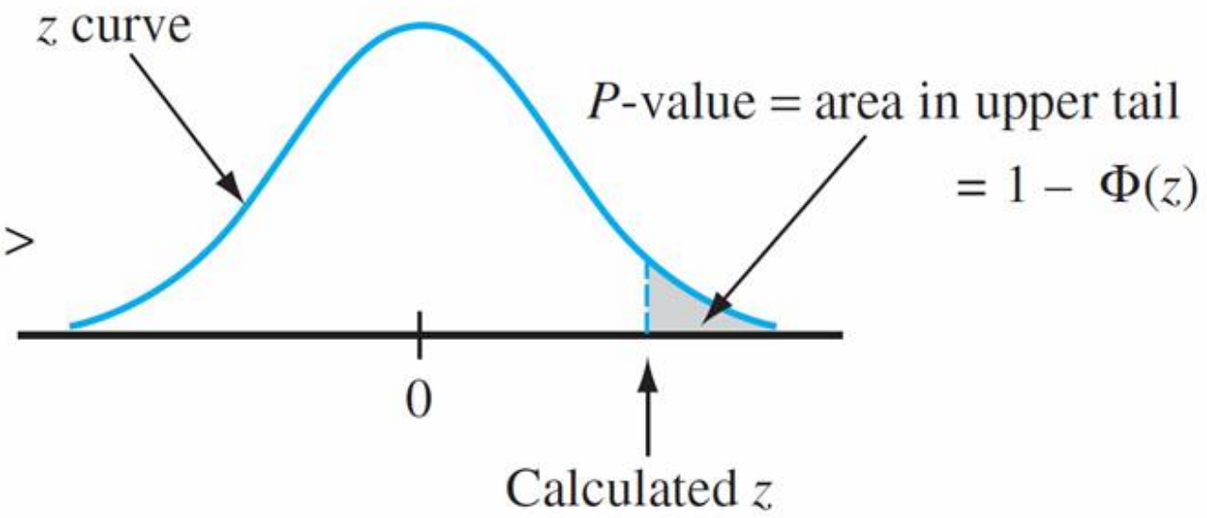
The calculation of the *p*-value depends on whether the test is upper-, lower-, or two-tailed.

$$P = \begin{cases} 1 - \Phi(z) & \text{upper tailed} \\ \Phi(z) & \text{lower tailed} \\ 2(1 - \Phi(z)) & \text{two tailed} \end{cases}$$

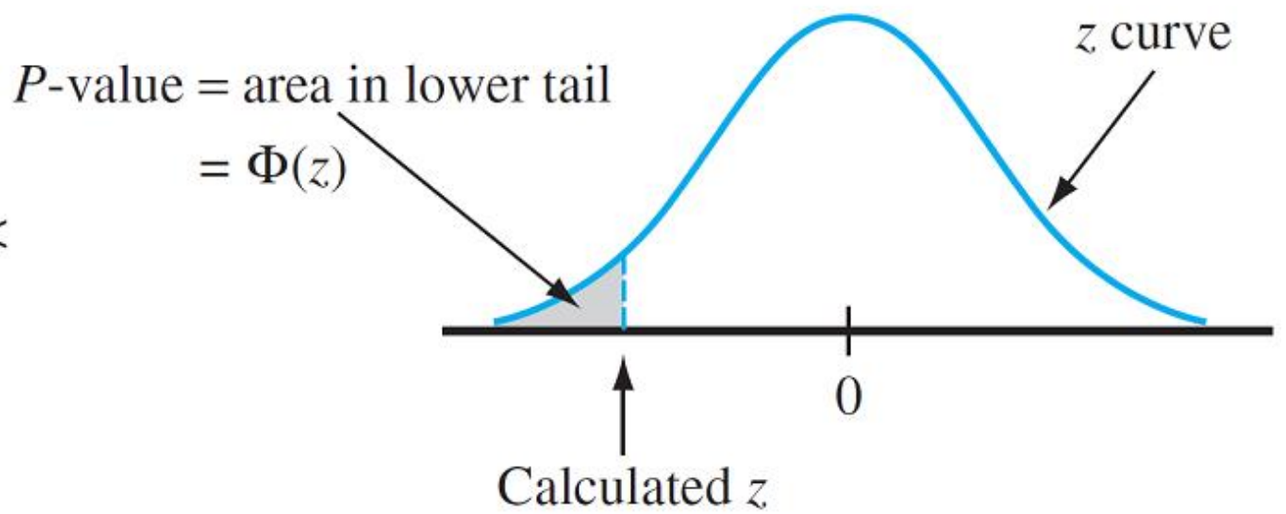
Each of these is the probability of getting a value at least as extreme as what was obtained (assuming H_0 true).

P-Values for Z Tests

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



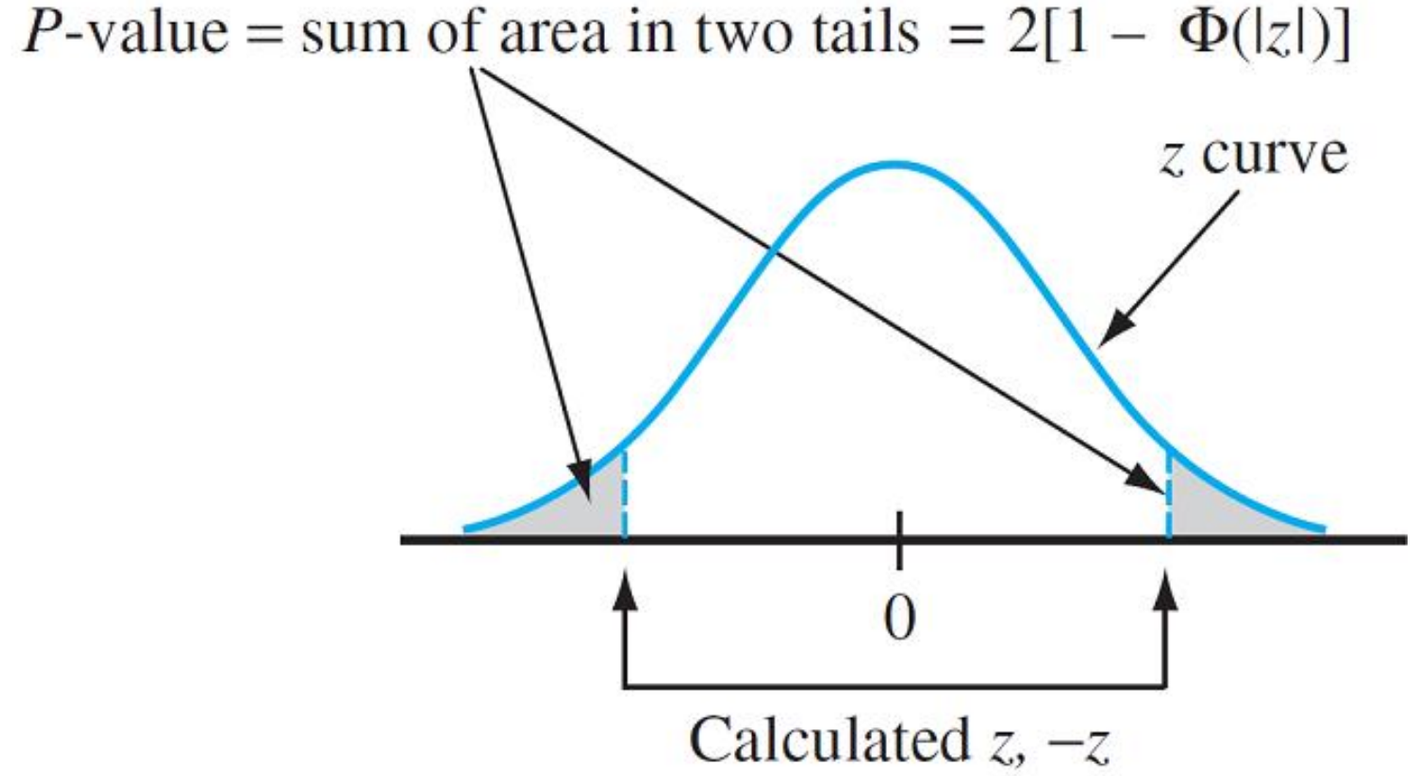
2. **Lower-tailed test**
   H_a contains the inequality <



P-Values for Z Tests

3. **Two-tailed test**

   H_a contains the inequality $\neq$



P-Values for Z Tests

Back to the lawnmower engine example: There, we had

$$H_o: \mu = 300 \qquad \text{vs} \qquad H_a: \mu > 300$$

and Z = 1.18. What is the p-value for this result?

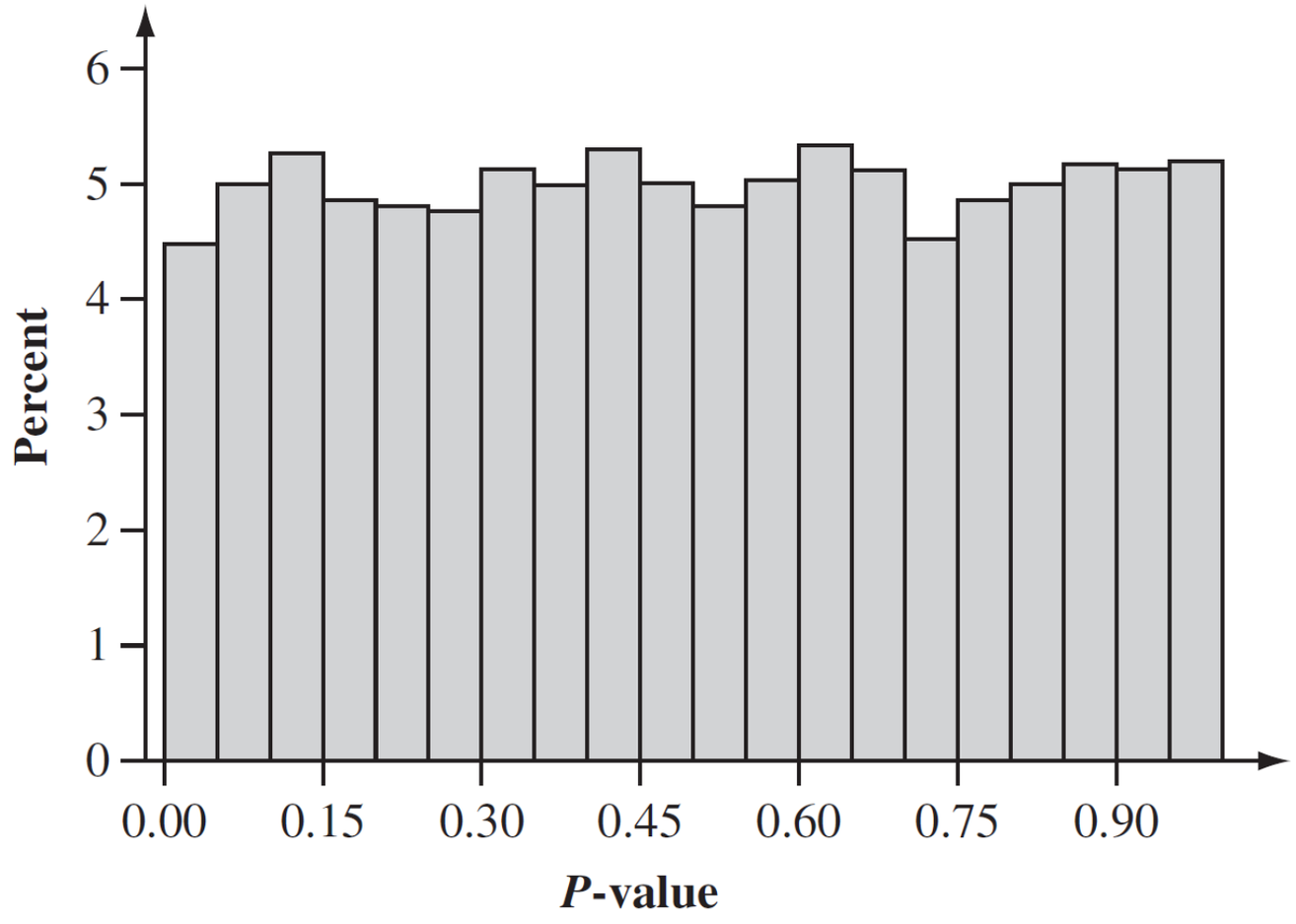
$H_0: \mu \leq 300$
$H_1: \mu > 300$
} upper tailed test $\Rightarrow$ p-value $= 1 - \Phi(1.18) = 0.119 > \alpha$

$\Rightarrow$ fail to reject.

Distribution of P-Values

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

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



Distribution of P-Values

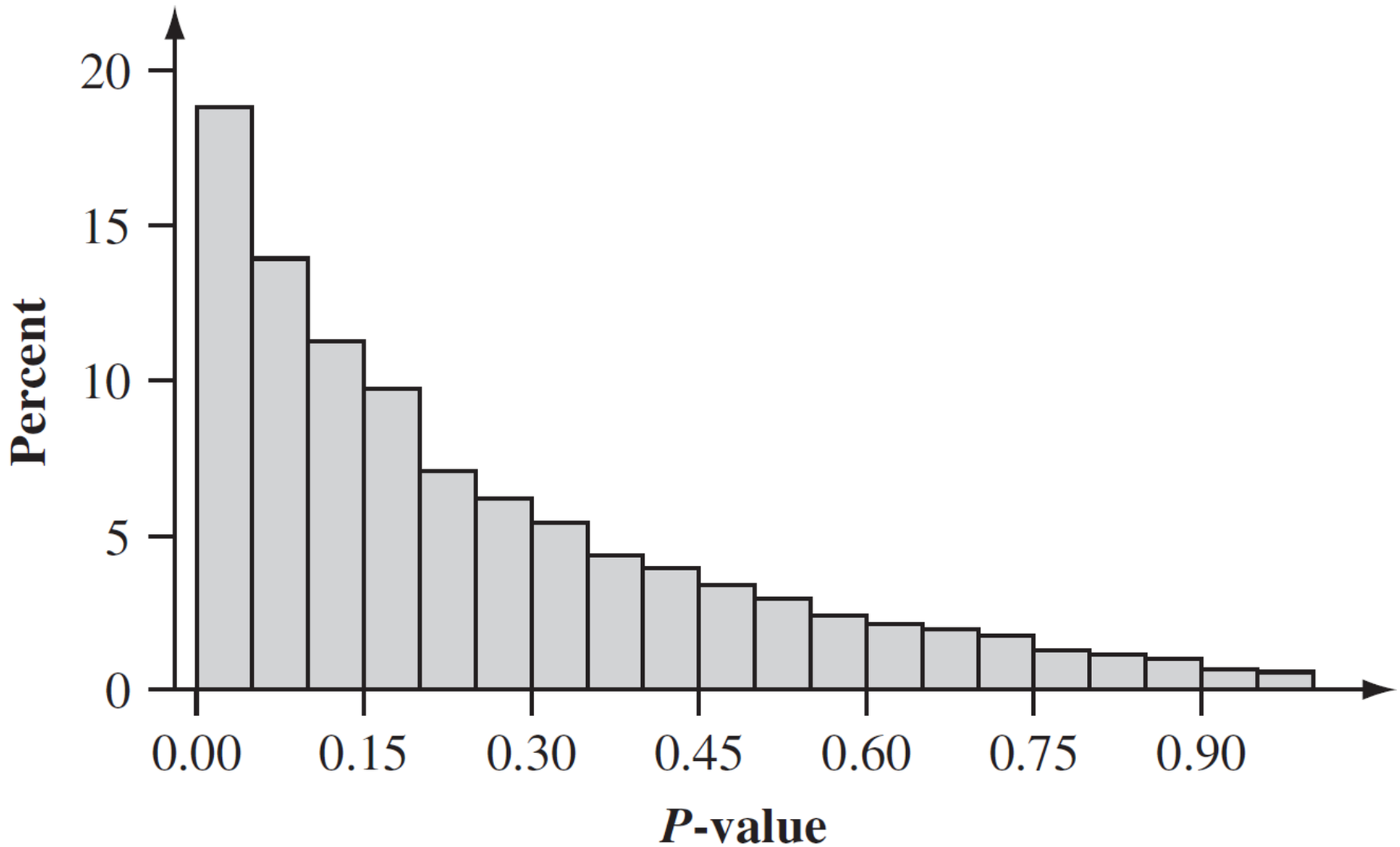
About 4.5% of these *p*-values are in the interval from 0 to .05.

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.

Distribution of P-Values

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 $\mu = 20$.



Distribution of P-Values

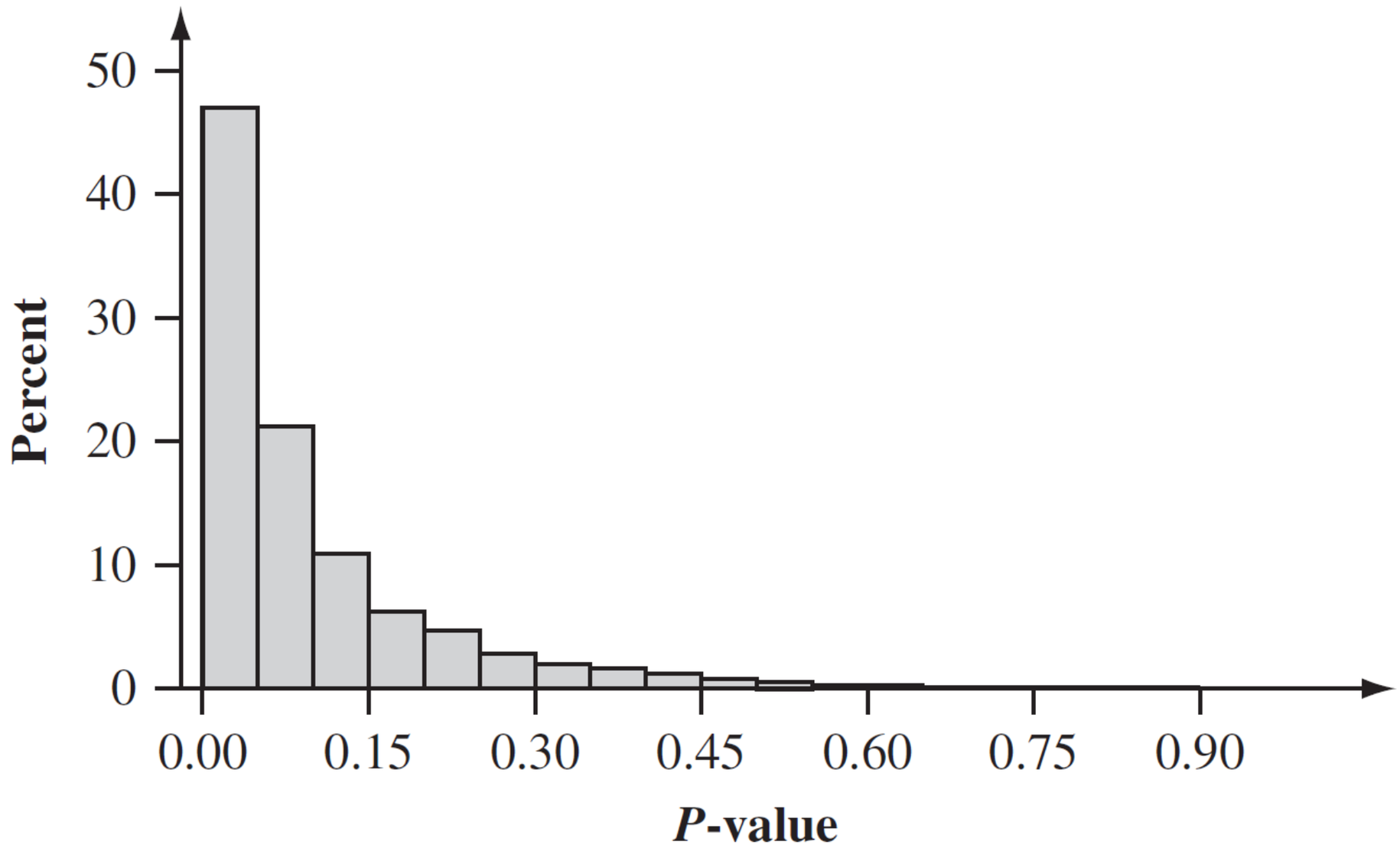
Again H_0 is rejected at significance level .05 whenever the *p*-value is at most .05 (in the first bin).

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 (the “effect size” is small).

Distribution of P-Values

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



Distribution of P-Values

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

In general, as 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 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.

Distribution of P-Values

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.