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:* $\mu = .75$, 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.

Criminal trials

Analogy to a criminal trial:

1. Assertion that the accused individual is innocent
2. Assertion that the accused individual is guilty

Usually, in the U.S. judicial system, the "innocence" claim is initially accepted as the truth.

Only in the face of strong evidence to the contrary should the jury reject this claim in favor of the alternative assertion that the accused is guilty.

Burden of Proof

In this sense, the claim of innocence is the favored or protected hypothesis, and the **burden of proof** is placed on those who believe in the alternative claim.

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

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

Precautionary principle

Another example is the "precautionary principle" often cited in Environmental Justice

It has to do with the introduction of the new products into the society (chemicals especially). Are they deemed not harmful to kids/pets/adults under most conditions?

"Assumed harmful until shown not harmful" = Precautionary principle

"Assumed non-harmful until enough harm observed" = Regular practice

Many countries (EU) are starting to adopt the Precautionary Principle

Examples:

- AIDS transmitted through blood transfusions (and how good the test of HIV presence should be)
- Nanoparticles – especially TI02 in sunscreens
- Zinc in nasal sprays

To reject or not to reject?

The **null hypothesis,** denoted by H_0, is the claim that is initially assumed to be true (the "status quo belief" claim).

The **alternative hypothesis,** denoted by H_a, is the assertion that is contradictory to H_0 in some way.

The null hypothesis will be rejected in favor of the alternative hypothesis only if sample evidence suggests that H_0 is not likely.

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 H_0: $\mu = .75$ against the alternative H_a: $\mu \neq .75$.

Only if sample data strongly suggests that μ is something other than .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

As an 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

A conservative approach is to think about the current theory as H_0 and the researcher's alternative explanation as H_a.

Rejection of the current theory will then occur only when evidence is much more consistent with the new theory.

In many situations, H_a is referred to as the "researcher's hypothesis," since it is the claim that the researcher would really like to validate.

Forms of Hypotheses

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

Test Procedures

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

Example -- the defective helmet problem:

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

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

Let X denote the number of defective helmets in the sample, a binomial random variable

Test Procedures

If H_0 is indeed true, then:
$E(X) = np = 200(.10) = 20$

However, we can expect fewer than 20 defective helmets if H_a is true.

The number of defective boards we actually observe -- x – is the basis for the test. If x is just a bit below 20, then the data don't contradict H_0 much. It is reasonable to reject H_0 only if x is significantly less than 20.

But what does "significantly" mean? Can we reject H_0 if $x \leq 15$? How about if $x \leq 10$?

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 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 $\{X \leq 15\}$. 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

Thus it is possible that

- H_0 may be rejected when it is true
- H_0 may not be rejected when it is in fact false.

These possible errors are not consequences of a foolishly chosen rejection region.

Either error might result when when any sensible rejection region is used.

Errors in Hypothesis Testing

Definition

- A **type I error (alpha)** consists of rejecting the null hypothesis H_0 when it is true.

  A **type II error (beta)** involves 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

Errors in Hypothesis Testing

The choice of a particular rejection region depends on the probabilities of type I and type II errors we allow the test to make.

These error probabilities are traditionally denoted by α (type I) and β (type II), respectively.

A good test will try to minimize both types of error.

When H_0 specifies a unique value of the parameter, there is a single value of α.

However, there is a different value of β for each value of the parameter consistent with H_a.

Example 1 cont'd

Sample of 200 helmets; X = number of defective helmets

H_0: p = .10 versus H_a: p < .10

Test statistic: X = # defective helmets in the sample

If the null is true, we expect somewhere around 20 defective helmets

Rejection region: R_{15} = {1, 2,…,15}; that is, reject H_0 if the observed number of defective helmets in the sample is 15 or less.

Example 1 cont'd

This rejection region is called *lower-tailed* because it consists only of small values of the test statistic.

When H_0 is true, X has a binomial probability distribution with n = 200 and p = 0.10 Thus:

α = P(type I error) = P(H_0 is rejected when it is true)

= P(X <= 15) = CDF(B(15, 200, 0.10)

= pbinom(15, 200, 0.10)

= 0.14

Example 1 cont'd

That is, when H_0 is actually true, roughly 14% of all experiments would result in H_0 being incorrectly rejected (type I error).

In contrast to α, there is not a single β. Instead, there is a different β for each different alternative value of p.

Thus there is a value of β for p = 0.05, which would imply that X ~ Bin(200, 0.05), another value of β for p = 0.075, which would imply that X ~ Bin(200, 0.075), and so on.

Example 1 cont'd

For example,

$\beta(0.09) = P(\text{type II error when } p = 0.09)$

$= P(H_0 \text{ is } \mathbf{not} \text{ rejected when it is false because } p = .09)$

$= P(X > 15 \text{ when } X \sim \text{Bin}(200, 0.09)) =$
$= 1 - \text{pbinom}(15,200,0.09) = 0.72$

When p is actually 0.09 rather than 0.10 (a "small" departure from H_0), roughly 72% of all experiments of this type would result in **H_0 being not rejected incorrectly!**

Example 1 cont'd

On the other hand,

$\beta(0.05) = P(\text{type II error when } p = 0.05)$

$= P(H_0 \text{ is } \mathbf{not} \text{ rejected when it is false because } p = .05)$

$= P(X > 15 \text{ when } X \sim \text{Bin}(200, 0.05)) =$
$= 1 - \text{pbinom}(15,200,0.05) = 0.044$

When p is actually 0.05 rather than .10 (larger departure from H_0), only 4.4% of all experiments of this type would result in H_0 being not rejected incorrectly.

Power cont'd

β decreases as the value of p moves farther away (in the direction of the alternative hypothesis) from the null value

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"

Composite null hypothesis cont'd

The proposed test procedure is still reasonable for testing the composite null hypothesis H_0: $p >= 0.10$.

In this case, there is no longer a single α, but instead there is an α for each p that is at least 0.10: $\alpha(.10)$, $\alpha(.15)$, $\alpha(.20)$, $\alpha(.25)$, and so on. It is easily verified, though, that

$\alpha(p) = P(X <= 15 \text{ when } X \sim \text{Bin}(200, p)) = \text{pbinom}(15; 200, p)$

$< \alpha(0.10) = .143$ for any $p > 0.10$

That is, the largest value of α occurs for the boundary value 0.10 between H_0 and H_a.

Thus if α is small for the simple null hypothesis, it will also be small for the composite H_0.

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 probability that it will be rejected at all. We do this by decreasing the size of the rejection region.

However, this will also increase the reluctance to reject the null when it is false. Ie, it will result in a larger value of β.

No rejection region will simultaneously make both α and all β's small.

A region must be chosen to strike a compromise between α and β

Historically, rejection of the null is considered to be more serious.

Thus, a type I error is usually views as "more serious" than a type II error.

Type I error in hypothesis testing

Thus, specify **the largest** value of α that can be tolerated, and then find a rejection region that has that α.

The resulting value of α is 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.

Summary: Hypothesis Testing

-- Null hypothesis, H_0 -- the claim initially assumed to be true

-- **Alternative hypothesis,** H_a -- the assertion contradictory to H_0.

The two possible conclusions:

1) reject H_0

2) fail to reject H_0.

Type I error (α) = P(rejecting the null hypothesis H_0 | H_0 is true**)**.

Type II error (β) = P(not rejecting H_0 | a specific *Ha* is true**)**.

Example 2

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 2 cont'd

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

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

Z expresses the distance between $\overline{X}$ and its expected value (when H_0 is true) normalized by the standard error of the sample mean.

For example, $z = 3$ results from an $\overline{x}$ that is 3 standard errors larger than the true mean as postulated by H_0. Rejecting H_0 when $\overline{x}$ "considerably" differs from 1.5 is equivalent to rejecting H_0 when z "considerably" differs from 0.

Example 2 cont'd

As H_a: $\mu > 1.5$, the form of the rejection region is $z \geq c$. Let's now determine c so that $\alpha = 0.05$

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

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

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

The value c must capture upper-tail area .05 under the z curve. So, directly from Appendix Table A.3,

$C = z_{.05} = 1.645$.

Example 2 cont'd

Notice that $z \geq 1.645$ is equivalent to $\overline{x} - 1.5 \geq (.0354)(1.645)$, that is, $\overline{x} \geq 1.56$.

Then β involves the probability that $\overline{X} < 1.56$ and can be calculated for any alternative μ greater than 1.5.

P(type II error) = P(not rejecting H_0 when H_a is true)

$= P(Z < c$ when Z *standardized with a specific alternative value of* $\mu)$

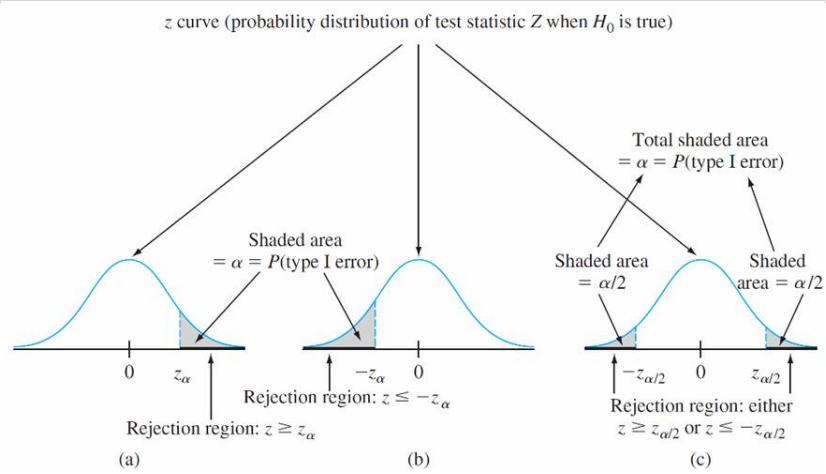
Case I: Testing means of a normal population with known σ

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

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

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

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: $\mu = \mu_0$

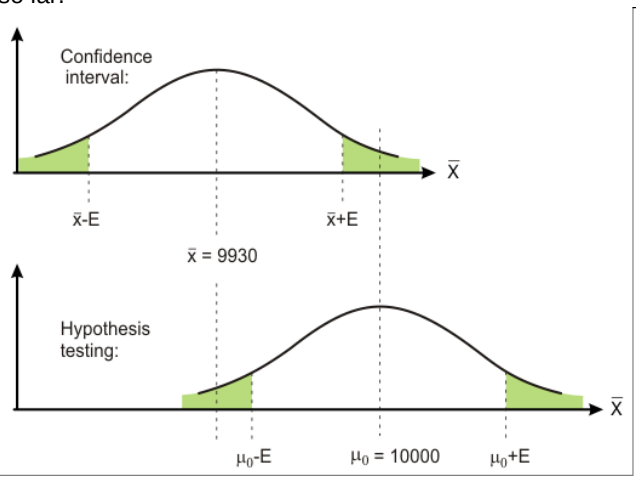
Test statistic value: $t = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$

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 the confidence intervals we've learned about so far.



source: shex.org

Proportions

How does hypothesis testing work for proportions?

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?

Example cont'd

Let's carry out a test of hypotheses using a significance level of .10.

1. p = the true proportion of all commercial chardonnay bottles considered spoiled to some extent

2. The null hypothesis is H_0: $p = .15$.

3. The alternative hypothesis is H_a: $p > .15$, the assertion that the population percentage exceeds 15%.

Example cont'd

4. Since $np_0 = 91(.15) = 13.65 > 10$ and $nq_0 = 91(.85) = 77.35 > 10$, the large-sample z test can be used. The test statistic value is

$$z = (\hat{p} - .15)/\sqrt{(.15)(.85)/n}.$$

5. The form of H_a implies that an upper-tailed test is appropriate: Reject H_0 if $z \geq z_{.10} = 1.28$.

6. $\hat{p} = 16/91 = .1758$, from which

$$z = (.1758 - .15)/\sqrt{(.15)(.85)/91} = .0258/.0374 = .69$$

Example cont'd

7. Since $.69 < 1.28$, z is not in the rejection region. At significance level .10, the null hypothesis cannot be rejected.

Although the percentage of contaminated bottles in the sample somewhat exceeds 15%, the sample percentage is not large enough to conclude that the population percentage exceeds 15%.

The difference between the sample proportion .1758 and the null value .15 can adequately be explained by sampling variability.

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 must be 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 .141g.

Example cont'd

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

With μ denoting the true average zinc mass for such batteries, the relevant hypotheses are H_0: $\mu = 2.0$ versus H_a: $\mu > 2.0$.

The sample size is large enough so that a *z* test can be used without making any specific assumption about the shape of the population distribution.

Example cont'd

The test statistic value is

$$z = \frac{\bar{x} - 2.0}{s/\sqrt{n}} = \frac{2.06 - 2.0}{.141/\sqrt{51}} = 3.04$$

An $\bar{x}$ value that is further away from 2 than 2.06 is from 2, corresponds to a value of *z* that exceeds 3.04.

Thus the *P*-value is

$$P\text{-value} = P(Z \geq 3.04)$$
$$= 1 - \Phi(3.04) = .0012$$

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).

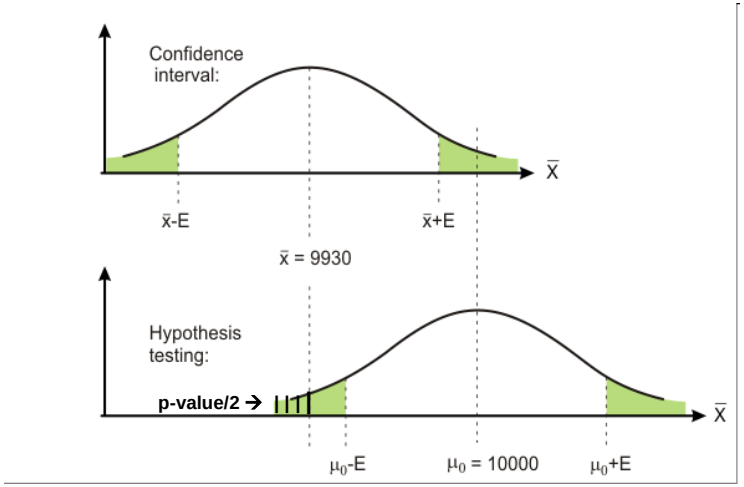
Then

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.

P-Values



P-Values

In the previous Example, we calculated 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 only .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.

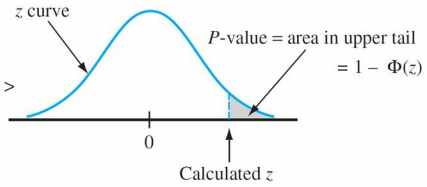
$$P\text{-value: } P = \begin{cases} 1 - \Phi(z) & \text{for an upper-tailed } z \text{ test} \\ \Phi(z) & \text{or an lower-tailed } z \text{ test} \\ 2[1 - \Phi(|z|)] & \text{for a two-tailed } z \text{ test} \end{cases}$$

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

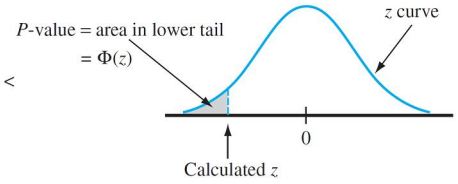
P-Values for *z* Tests

The three cases are illustrated in Figure 8.9.

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



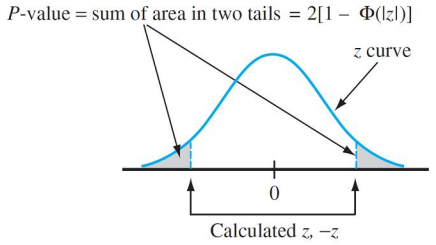
2. **Lower-tailed test**
   H_a contains the inequality <



P-Values for *z* Tests

cont'd

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



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?

Example

cont'd

1. Parameter of interest: μ = true average wafer thickness
1. Null hypothesis: H_0: $\mu = 245$
1. Alternative hypothesis: H_a: $\mu \neq 245$
1. Formula for test statistic value: $z = \dfrac{\bar{x} - 245}{s/\sqrt{n}}$
1. Calculation of test statistic value: $z = \dfrac{246.18 - 245}{3.60/\sqrt{50}} = 2.32$

Example cont'd

6. Determination of *P*-value: Because the test is two-tailed, *P*-value = $2(1 - \Phi(2.32)) = .0204$

7. Conclusion: Using a significance level of .01, H_0 would not be rejected since .0204 > .01.

   At this significance level, there is insufficient evidence to conclude that true average thickness differs from the target value.

Statistical Versus Practical Significance

When using

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

one must be especially careful – with large n, ***z* will get large!!!**

So **<u>any small departure from H_0</u>** will almost surely be detected by a test – and the null rejected -- but such a departure may have **<u>little practical significance</u>**.

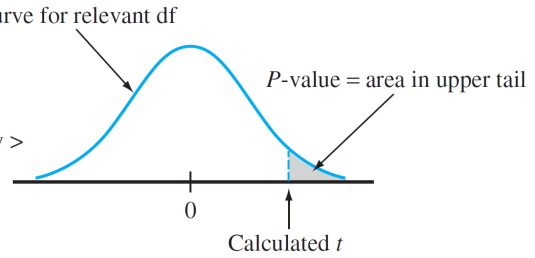
P-Values for *t* Tests

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 $n - 1$.

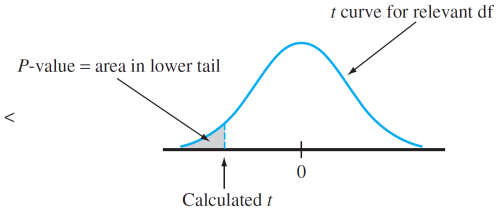
1. **Upper-tailed test**
   H_a contains the inequality >



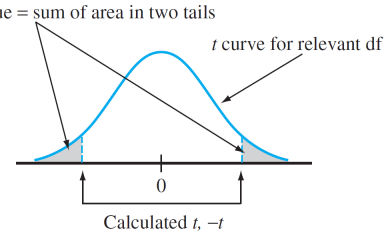
P-Values for *t* Tests

cont'd

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



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

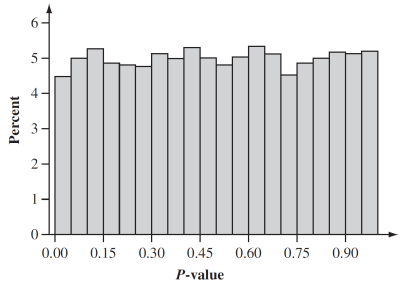


How are p-values distributed?

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 bin 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.

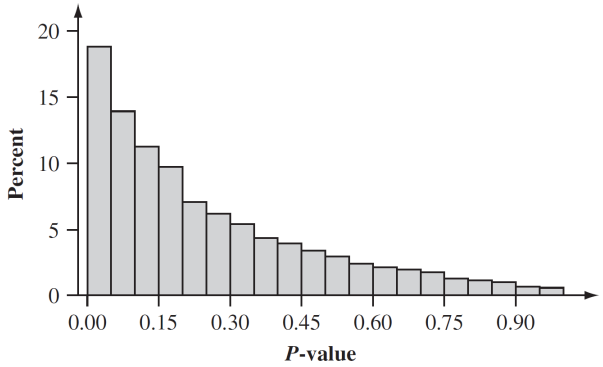
If we continued to generate samples and carry out the test for each sample at significance level .05, in the long run 5% of the *P*-values would be in the first bin.

(the theory works…)

Example

cont'd

A histogram of the *P*-values when we simulate under an alternative $\mu = 21$. 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).

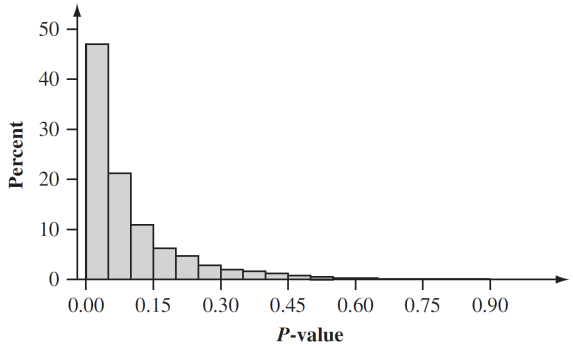
Unfortunately, this is the case for only about 19% of the *P*-values. So only about 19% of the 10,000 tests correctly reject the null hypothesis; for the other 81%, a type II error is committed.

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.