Tests of Hypotheses Based on a Single Sample

1 Hypotheses and Test Procedures

Hypotheses and Test Procedures

A **statistical hypothesis,** or just *hypothesis*, is a claim or assertion either about the value of a single parameter (population characteristic or characteristic of a probability distribution), about the values of several parameters, or about the form of an entire probability distribution.

One example of a hypothesis is the claim $\mu = .75$, where μ is the true average inside diameter of a certain type of PVC pipe.

Another example is the statement $p < .10$, where p is the proportion of defective circuit boards among all circuit boards produced by a certain manufacturer.

Hypotheses and Test Procedures

In any hypothesis-testing problem, there are two contradictory hypotheses under consideration. One hypothesis might be the claim $\mu = .75$ and the other $\mu \neq .75$, or the two contradictory statements might be $p \geq .10$ and $p < .10$.

Hypotheses and Test Procedures

The objective is to decide, based on sample information, which of the two hypotheses is correct.

There is a familiar analogy to this in a criminal trial. One claim is the assertion that the accused individual is innocent.

In the U.S. judicial system, this is the claim that is initially believed to be true. 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.

Hypotheses and Test Procedures

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 will not be rejected in favor of the alternative claim unless sample evidence contradicts it and provides strong support for the alternative assertion.

Hypotheses and Test Procedures

Definition

The **null hypothesis,** denoted by H_0, is the claim that is initially assumed to be true (the "prior belief" claim). The **alternative hypothesis,** denoted by H_a, is the assertion that is contradictory to H_0.

The null hypothesis will be rejected in favor of the alternative hypothesis only if sample evidence suggests that H_0 is false. 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 from a hypothesis-testing analysis are then *reject* H_0 or *fail to reject* H_0.

Hypotheses and Test Procedures

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

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 quite plausible.

Hypotheses and Test Procedures

A conservative approach is to identify the current theory with H_0 and the researcher's alternative explanation with 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.

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, and so on.

Suppose, for example, that 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.

Hypotheses and Test Procedures

Let p denote the true proportion of defective boards resulting from the changed process.

Then the research hypothesis, on which the burden of proof is placed, is the assertion that $p < .10$. Thus the alternative hypothesis is H_a: $p < .10$.

In our treatment of hypothesis testing, H_0 will generally be stated as an equality claim. If θ denotes the parameter of interest, the null hypothesis will have the form H_0: $\theta = \theta_0$, where θ_0 is a specified number called the *null value* of the parameter (value claimed for θ by the null hypothesis).

Hypotheses and Test Procedures

As an example, consider the circuit board situation just discussed. The suggested alternative hypothesis was H_a: $p < .10$, the claim that the defective rate is reduced by the process modification.

A natural choice of H_0 in this situation is the claim that $p \geq .10$, according to which the new process is either no better *or* worse than the one currently used.

We will instead consider H_0: $p = .10$ versus H_a: $p < .10$.

Hypotheses and Test Procedures

The rationale for using this simplified null hypothesis is that any reasonable decision procedure for deciding between H_0: p = .10 and H_a: p < .10 will also be reasonable for deciding between the claim that $p \geq .10$ and H_a.

The use of a simplified H_0 is preferred because it has certain technical benefits, which will be apparent shortly.

Hypotheses and Test Procedures

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

1. H_a: $\theta > \theta_0$ (in which case the implicit null hypothesis is $\theta \leq \theta_0$),

2. H_a: $\theta < \theta_0$ (in which case the implicit null hypothesis is $\theta \geq \theta_0$), or

3. H_a: $\theta \neq \theta_0$

Hypotheses and Test Procedures

For example, let σ denote the standard deviation of the distribution of inside diameters (inches) for a certain type of metal sleeve.

If the decision was made to use the sleeve unless sample evidence conclusively demonstrated that $\sigma > .001$, the appropriate hypotheses would be H_0: $\sigma = .001$. versus H_a: $\sigma > .001$.

The number θ_0 that appears in both H_0 and H_a (separates the alternative from the null) is called the **null value.**

Test Procedures

Test Procedures

A test procedure is a rule, based on sample data, for deciding whether H_0 should be rejected.

The key issue will be the following: Suppose that H_0 is in fact true. Then how likely is it that a (random) sample at least as contradictory to this hypothesis as our sample would result? Consider the following two scenarios:

1. There is only a .1% chance (a probability of .001) of getting a sample at least as contradictory to H_0 as what we obtained assuming that H_0 is true.

2. There is a 25% chance (a probability of .25) of getting a sample at least as contradictory to H_0 as what we obtained when H_0 is true.

Test Procedures

In the first scenario, something as extreme as our sample is very unlikely to have occurred when H_0 is true—in the long run only 1 in 1000 samples would be at least as contradictory to the null hypothesis as the one we ended up selecting.

In contrast, for the second scenario, in the long run 25 out of every 100 samples would be at least as contradictory to H_0 as what we obtained assuming that the null hypothesis is true. So our sample is quite consistent with H_0, and there is no reason to reject it.

Test Procedures

We must now flesh out this reasoning by being more specific as to what is meant by "at least as contradictory to H_0 as the sample we obtained when H_0 is true."

Before doing so in a general way, let's consider several examples.

Example 8.1

The company that manufactures brand D Greek-style yogurt is anxious to increase its market share, and in particular persuade those who currently prefer brand C to switch brands.

So the marketing department has devised the following blind taste experiment. Each of 100 brand C consumers will be asked to taste yogurt from two bowls, one containing brand C and the other brand D, and then say which one he or she prefers.

The bowls are marked with a code so that the experimenters know which bowl contains which yogurt, but the experimental subjects do not have this information

Example 8.1

Let *p* denote the proportion of all brand C consumers who would prefer C to D in such circumstances. Let's consider testing the hypotheses H_0: $p = .5$ versus H_a: $p < .5$.

The alternative hypothesis says that a majority of brand C consumers actually prefer brand D. Of course the brand D company would like to have H_0 rejected so that *Ha* is judged the more plausible hypothesis.

If the null hypothesis is true, then whether a single randomly selected brand C consumer prefers C or D is analogous to the result of flipping a fair coin.

Examples 8.1

Let X = the number among the 100 selected individuals who prefer C to D. This random variable will serve as our *test statistic*, the function of sample data on which we'll base our conclusion.

Now X is a binomial random variable (the number of successes in an experiment with a fixed number of independent trials having constant success probability p). When H_0 is true, this test statistic has a binomial distribution with p = .5, in which case $E(X) = np = 100(.5) = 50$.

Examples 8.1

Intuitively, a value of *X* "considerably" smaller than 50 argues for rejection of H_0 in favor of H_a.

Suppose the observed value of *X* is *x* = 37. How contradictory is this value to the null hypothesis? To answer this question, let's first identify values of *X* that are even more contradictory to H_0 than is 37 itself.

Clearly 35 is one such value, and 30 is another; in fact, any number smaller than 37 is a value of *X* more contradictory to the null hypothesis than is the value we actually observed.

Example 8.1

Now consider the probability, computed assuming that the null hypothesis is true, of obtaining a value of X at least as contradictory to H_0 as is our observed value:

$$P(X \le 37 \text{ when } H_0 \text{ is true}) = P(X \le 37 \text{ when } X \sim \text{Bin}(100, .5))$$
$$= B(37; 100, .5) = .006$$

(from software). Thus if the null hypothesis is true, there is less than a 1% chance of seeing 37 or fewer successes amongst the 100 trials. This suggests that $x = 37$ is much more consistent with the alternative hypothesis than with the null, and that rejection of H_0 in favor of H_a is a sensible conclusion.

Example 8.1

In addition, note that $\sigma_x = \sqrt{npq} = \sqrt{100(.5).5)} = 5$ when *H*0 is true. It follows that 37 is more than 2.5 standard deviations smaller than what we'd expect to see were H_0 true.

Now suppose that 45 of the 100 individuals in the experiment prefer C (45 successes). Let's again calculate the probability, assuming H_0 true, of getting a test statistic value at least as contradictory to H_0 as this:

$$
\begin{aligned}
P(X \leq 45 \text{ when } H_0 \text{ is true}) &= P(X \leq 45 \text{ when } X \sim \text{Bin}(100, .5)) \\
&= B(45; 100, .5) = .184
\end{aligned}
$$

Example 8.1

So if in fact *p* = .5, it would not be surprising to see 45 or fewer successes.

For this reason, the value 45 does not seem very contradictory to H_0 (it is only one standard deviation smaller than what we'd expect were H_0 true). Rejection of H_0 in this case does not seem sensible.

Test Procedures

The type of probability calculated in Example 8.1 will now provide the basis for obtaining general test procedures.

A **test statistic** is a function of the sample data used as a basis for deciding whether H_0 should be rejected. The selected test statistic should discriminate effectively between the two hypotheses. That is, values of the statistic that tend to result when H_0 is true should be quite different from those typically observed when H_0 is not true.

The ***P*-value** is the probability, calculated assuming that the null hypothesis is true, of obtaining a value of the test statistic at least as contradictory to H_0 as the value calculated from the available sample data. A conclusion is reached in a hypothesis testing analysis by selecting a number α, called the **significance level** (alternatively, *level of significance*) of the test, that is reasonably close to 0. Then H_0 will be rejected in favor of H_a if P-value $\leq \alpha$, whereas H_0 will not be rejected (still considered to be plausible) if P-value $> \alpha$. The significance levels used most frequently in practice are (in order) $\alpha = .05, .01, .001$, and $.10$.

Test Procedures

For example, if we select a significance level of .05 and then compute *P*-value = .0032, H_0 would be rejected because .0032 $\leq$ .05.

With this same *P*-value, the null hypothesis would also be rejected at the smaller significance level of .01 because .0032 $\leq$ .01. However, at a significance level of .001 we would not be able to reject H_0 since .0032 $\geq$ .001.

Test Procedures

Figure 8.1 illustrates the comparison of the *P*-value with the significance level in order to reach a conclusion.

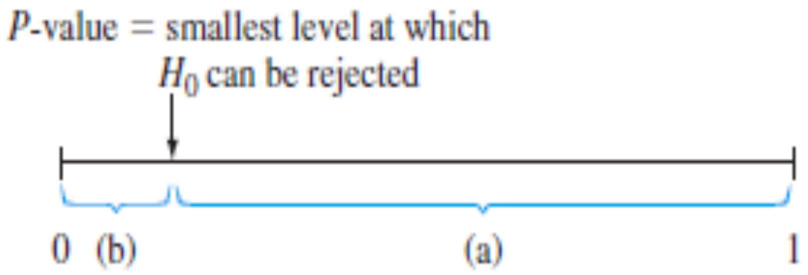


Figure 8.1 Comparing α and the *P*-value: (a) reject H_0 when α lies here; (b) do not reject H_0 when α lies here

We will shortly consider in some detail the consequences of selecting a smaller significance level rather than a larger one. For the moment, note that the smaller the significance level, the more protection is being given to the null hypothesis and the harder it is for that hypothesis to be rejected.

Test Procedures

The definition of a *P*-value is obviously somewhat complicated, and it doesn't roll off the tongue very smoothly without a good deal of practice. In fact, many users of statistical methodology use the specified decision rule repeatedly to test hypotheses, but would be hard put to say what a *P*-value is! Here are some important points:

- The *P*-value is a probability.
- This probability is calculated assuming that the null hypothesis is true.
- To determine the *P*-value, we must first decide which values of the test statistic are at least as contradictory to H_0 as the value obtained from our sample.
- The smaller the *P*-value, the stronger is the evidence against H_0 and in favor of H_a.
- The *P*-value is not the probability that the null hypothesis is true or that it is false, nor is it the probability that an erroneous conclusion is reached.

Errors in Hypothesis Testing

Errors in Hypothesis Testing

The basis for choosing a particular significance level *alpha* lies in consideration of the errors that one might be faced with in drawing a conclusion. Recall the judicial scenario in which the null hypothesis is that the individual accused of committing a crime is in fact innocent.

In rendering a verdict, the jury must consider the possibility of committing one of two different kinds of errors. One of these involves convicting an innocent person, and the other involves letting a guilty person go free. Similarly, there are two different types of errors that might be made in the course of a statistical hypothesis testing analysis.

Errors in Hypothesis Testing

Definition

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

A **type II error** involves not rejecting H_0 when it is false.

Errors in Hypothesis Testing

As an example, a cereal manufacturer claims that a serving of one of its brands provides 100 calories.

Of course the actual calorie content will vary somewhat from serving to serving (of the specified size), so 100 should be interpreted as an average. It could be distressing to consumers of this cereal if the true average calorie content exceeded the asserted value

So an appropriate formulation of hypotheses is to test $H_0 : \mu = 100$ versus $H_a : \mu > 100$. The alternative hypothesis says that consumers are ingesting on average a greater amount of calories than what the company claims.

Errors in Hypothesis Testing

A type I error here consists of rejecting the manufacturer's claim that $\mu = 100$ when it is actually true. A type II error results from not rejecting the manufacturer's claim when it is actually the case that $\mu > 100$.

In the best of all possible worlds, we'd have a judicial system that never convicted an innocent person and never let a guilty person go free. This gold standard for judicial decisions has proven to be extremely elusive.

Errors in Hypothesis Testing

Similarly, we would like to find test procedures for which neither type of error is ever committed. However, this ideal can be achieved only by basing a conclusion on an examination of the entire population.

The difficulty with using a procedure based on sample data is that because of sampling variability, a sample unrepresentative of the population may result.

In the calorie content scenario, even if the manufacturer's assertion is correct, an unusually large value of X may result in a P-value smaller than the chosen significance level and the consequent commission of a type I error.

Errors in Hypothesis Testing

Alternatively, the true average calorie content may exceed what the manufacturer claims, yet a sample of servings may yield a relatively large *P*-value for which the null hypothesis cannot be rejected.

Instead of demanding error-free procedures, we must seek procedures for which either type of error is unlikely to be committed.

That is, a good procedure is one for which the probability of making a type I error is small and the probability of making a type II error is also small.

Example 8.4

An automobile model 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.

The hypotheses to be tested are H_0: $p = .25$ (no improvement) versus H_a: $p > .25$.

The test will be based on an experiment involving $n = 20$ independent crashes with prototypes of the new design.

Example 8.4

cont'd

The natural test statistic here is *X* = the number of crashes with no visible damage.

If H_o is true, $E(X) = np_0 = (20).25) = 5$. Intuition suggests that an observed value *x* much larger than this would provide strong evidence against H_o and in support of H_a.

Consider using a significance level of .10. The *P*-value is $P(X \geq x$ when *X* has a binomial distribution with $n = 20$ and $p = .25) = 1 - B(x - 1; 20, .25)$ for $x > 0$.

Example 8.4

cont'd

Appendix Table A.1 shows that in this case,

$$P(X \geq 7) = 1 - B(6; 20, .25) = 1 - .786 = .214$$
$$P(X \geq 8) = 1 - .898 = .102 \approx .10, P(X \geq 9) = 1 - .959 = .041$$

Thus rejecting H_0 when *P*-value $\leq$ .10 is equivalent to rejecting H_0 when $X \geq 8$. Therefore

$$\begin{aligned} P(\text{committing a type I error}) &= P(\text{rejecting } H_0 \text{ when } H_0 \text{ is true}) \\ &= P(X \geq 8 \text{ when } X \text{ has a binomial distribution with } n = 20 \text{ and } p = .25) \\ &= .102 \\ &\approx .10 \end{aligned}$$

Example 8.4

That is, *the probability of a type I error is just the significance level* α.

If the null hypothesis is true here and the test procedure is used over and over again, each time in conjunction with a group of 20 crashes, in the long run the null hypothesis will be incorrectly rejected in favor of the alternative hypothesis about 10% of the time.

So our test procedure offers reasonably good protection against committing a type I error.

Example 8.4

There is only one type I error probability because there is only one value of the parameter for which H_0 is true (this is one benefit of simplifying the null hypothesis to a claim of equality).

Let β denote the probability of committing a type II error. Unfortunately there is not a single value of β, because there are a multitude of ways for H_0 to be false—it could be false because $p = .30$, because $p = .37$, because $p = .5$, and so on.

There is in fact a different value of β for each different value of p that exceeds .25

Example 8.4

At the chosen significance level .10, H_0 will be rejected if and only if $X \geq 8$, so H_0 will not be rejected if and only if $X \leq 7$. Thus

$$\begin{aligned}\beta(.3) &= P(\text{type II error when } p = .3)\\ &= P(H_0 \text{ is not rejected when } p = .3)\\ &= P[X \leq 7 \text{ when } X \sim \text{Bin}(20, .3)]\\ &= B(7; 20, .3) = .772\end{aligned}$$

When p is actually .3 rather than .25 (a “small” departure from H_o), roughly 77% of all experiments of this type would result in H_o being incorrectly not rejected!

Example 8.4

The accompanying table displays β for selected values of *p* (each calculated as we just did for $\beta(.3)$). 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_o, the more likely it is that such a departure will be detected.

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

The probability of committing a type II error here is quite large when *p* = .3 or .4. This is because those values are quite close to what H_0 asserts and the sample size of 20 is too small to permit accurate discrimination between .25 and those values of *p*.

Example 8.4

The proposed test procedure is still reasonable for testing the more realistic null hypothesis that $p \leq .25$. In this case, there is no longer a single type I error probability α, but instead there is an α for each p that is at most .25: $\alpha(.25)$, $\alpha(.23)$, $\alpha(.20)$, $\alpha(.15)$, and so on.

It is easily verified, though, that *alpha*(*p*) < *alpha*(.25) = .102 if $p < .25$. That is, the largest type I error probability occurs for the boundary value .25 between H_o and H_a.

Thus if α is small for the simplified null hypothesis, it will also be as small as or smaller for the more realistic H_o.

Errors in Test Procedure

The test procedure that rejects H_0 if P-value $\leq \alpha$ and otherwise does not reject H_0 has $P(\text{type I error}) = \alpha$. That is, the significance level employed in the test procedure is the probability of a type I error.

Errors in Hypothesis Testing

The inverse relationship between the significance level α and type II error probabilities can be generalized in the following manner:

Proposition

Suppose an experiment or sampling procedure is selected, a sample size is specified, and a test statistic is chosen. Then increasing the significance level α, i.e., employing a larger type I error probability, results in a smaller value of β for any particular parameter value consistent with H_a.

This result is intuitively obvious because when α is increased, it becomes more likely that we'll have P-value $\leq \alpha$ and therefore less likely that P-value $> \alpha$.

Errors in Hypothesis Testing

This proposition implies that once the test statistic and *n* are fixed, it is not possible to make both α and any values of β that might be of interest arbitrarily small.

A strategy that is sometimes (but perhaps not often enough) used in practice is to specify α and also β for some alternative value of the parameter that is of particular importance to the investigator.

In practice it is usually the case that the hypotheses of interest can be formulated so that a type I error is more serious than a type II error, and then use the largest value of *alpha* that can be tolerated.

Errors in Hypothesis Testing

For example, if α = .05 is the largest significance level that can be tolerated, it would be better to use that rather than α =.01, because all β's for the former α will be smaller than those for the latter one.

As previously mentioned, the most frequently employed significance levels are α = .05, .01, .001, and .10.

Some Further Comments on the P-Value

Suppose that the *P*-value is calculated to be .038. The null hypothesis will then be rejected if $.038 \leq \alpha$ and not rejected otherwise. So H_0 can be rejected if $\alpha = .10$ or .05 but not if $\alpha = .01$ or .001.

In fact, H_0 would be rejected for any significance level that is at least .038 but not for any level smaller than .038. For this reason, the *P*-value is often referred to as the **observed significance level** (OSL): it is the smallest value of *a* for which H_0 can be rejected.

Some Further Comments on the P-Value

One very appealing aspect of basing a conclusion from a hypothesis testing analysis on the *P*-value is that all widely used statistical software packages will calculate and output the *P*-value for any of the commonly used test procedures.

Once the *P*-value is available, the investigator need only compare it to the selected significance level to decide whether H_0 should be rejected.

Thus when medical journals report a *P*-value, a significance level is not mandated; instead it is left to the reader to select his or her own level and conclude accordingly.

Some Further Comments on the P-Value

A final point concerning the utility of the *P*-value is that it allows one to distinguish between a close call and a very clear-cut conclusion at any particular significance level. For example, suppose you are told that H_0 was rejected at significance level .05.

This conclusion is consistent with a *P*-value of .0498 and also with a *P*-value of .0003, since in each case *P*-value $\leq \alpha$ = .05. But of course with a *P*-value of .0498, the null hypothesis is barely rejected, whereas with *P*-value = .0003, the null hypothesis is rejected by a country mile.

So it is always preferable to report the *P*-value rather than just stating the conclusion at a particular significance level.

2 Tests About a Population Mean

z Tests for Hypotheses about a Population Mean

Recall from the previous section that a conclusion in a hypothesis testing analysis is reached by proceeding as follows:

i. Compute the value of an appropriate test statistic.

ii. Then determine the P-value—the probability, calculated assuming that the null hypothesis H_0 true, of observing a test statistic value at least as contradictory to H_0 as what resulted from the available data.

iii. Reject the null hypothesis if P-value $\leq \alpha$, where α is the specified or chosen significance level, i.e., the probability of a type I error (rejecting H_0 when it is true); if P-value $> \alpha$, there is not enough evidence to justify rejecting H_0 (it is still deemed plausible).

z Tests for Hypotheses about a Population Mean

Determination of the *P*-value depends on the distribution of the test statistic when H_0 is true. In this section we describe *z* tests for testing hypotheses about a single population mean μ.

By "*z* test," we mean that the test statistic has at least approximately a standard normal distribution when H_0 is true. The *P*-value will then be a *z* curve area which depends on whether the inequality in H_a is $>, <, or \neq$.

z Tests for Hypotheses about a Population Mean

In the development of confidence intervals for μ in the previous topic, we first considered the case in which the population distribution is normal with known σ, then relaxed the normality and known *s* assumptions when the sample size *n* is large, and finally described the one-sample *t* CI for the mean of a normal population.

A Normal Population with Known σ

A Normal Population with Known σ

Although the assumption that the value of σ is known is rarely met in practice, this case provides a good starting point because of the ease with which general procedures and their properties can be developed.

The null hypothesis in all three cases will state that μ has a particular numerical value, the *null value*. We denote by μ_0, so the null hypothesis has the form $H_0\colon \mu = \mu_0$. Let $X_1, \ldots, X_n$ represent a random sample of size *n* from the normal population.

A Normal Population with Known σ

Then the sample mean $\overline{X}$ has a normal distribution with expected value $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \sigma/\sqrt{n}$.

When H_0 is true, $\mu_{\overline{X}} = \mu_0$. Consider now the statistic Z obtained by standardizing $\overline{X}$ under the assumption that H_0 is true:

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

A Normal Population with Known σ

Substitution of the computed sample mean $\bar{x}$ gives z, the distance between $\bar{x}$ and μ_0 expressed in "standard deviation units."

For example, if the null hypothesis is H_0: $\mu = 100$, $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 10/\sqrt{25} = 2.0$, and $\bar{x} = 103$, then the test statistic value is $z = (103 - 100)/2.0 = 1.5$.

That is, the observed value of $\bar{x}$ is 1.5 standard Deviations (of $\bar{X}$) larger than what we expect it to be when H_0 is true.

A Normal Population with Known σ

The statistic *Z* is a natural measure of the distance between $\overline{X}$, the estimator of μ, and its expected value when H_0 is true. If this distance is too great in a direction consistent with H_a, there is substantial evidence that H_0 is false.

Suppose first that the alternative hypothesis has the form $H_a : \mu > \mu_0$. Then an $\bar{x}$ value that considerable exceeds μ_0 provides evidence against H_0 .

Such an $\bar{x}$ value corresponds to a large positive value *z*. This in turn implies that any value *exceeding* the calculated *z* is more contradictory to H_0 than *z* itself.

A Normal Population with Known σ

It follows that

$$P\text{-value} = P(Z \geq z \text{ when } H_0 \text{ is true})$$

Now here is the key point: when H_0 is true, the test statistic *Z* has a standard normal distribution—because we created *Z* by standardizing $\bar{X}$ assuming that H_0 is true (i.e., by subtracting μ_0).

The implication is that in this case, the *P*-value is just the area under the standard normal curve to the right of *z*. Because of this, the test is referred to as *upper-tailed.*

A Normal Population with Known σ

For example, in the previous paragraph we calculated *z* = 1.5. If in the alternative hypothesis there is H_z: μ > 100, then *P*-value = area under the *z* curve to the right of 1.5 = 1 - Φ(1.50) .0668. At significance level .05 we would not be able to reject the null hypothesis because the *P*-value exceeds α.

Now consider an alternative hypothesis of the form $H_a: \mu < \mu_0$. In this case any value of the sample mean smaller than our $\bar{x}$ is even more contradictory to the null hypothesis.

A Normal Population with Known σ

Thus any test statistic value *smaller* than the calculated *z* is more contradictory to H_0 than is *z* itself. It follows that

$$\begin{aligned} P\text{-value} &= P(Z \le z \text{ when } H_0 \text{ is true}) \\ &= \text{area under the standard normal curve to the left of } z = \Phi(z) \end{aligned}$$

The test in this case is customarily referred to as *lower-tailed.* If, for example, the alternative hypothesis is $H_a\colon \mu < 100$ and *z* = -2.75, then *P*-value = Φ(-2.75) = .0030. This is small enough to justify rejection of H_0 at a significance level of either .05 or .01, but not .001.

A Normal Population with Known σ

The third possible alternative, H_a: $\mu \neq \mu_0$, requires a bit more careful thought. Suppose, for example, that the null value is 100 and that *x* = 103 results in *z* = 1.5.

Then any $\bar{x}$ value exceeding 103 is more contradictory to H_0 than is 103 itself.

So any *z* exceeding 1.5 is likewise more contradictory to H_0 than is 1.5. However, 97 is just as contradictory to the null hypothesis as is 103, since it is the same distance below 100 as 103 is above 100. Thus *z* =-1.5 is just as contradictory to H_0 as is *z* = 1.5.

A Normal Population with Known σ

Therefore any *z* smaller than -1.5 is more contradictory to H_0 than is any z greater than 1.5. It follows that

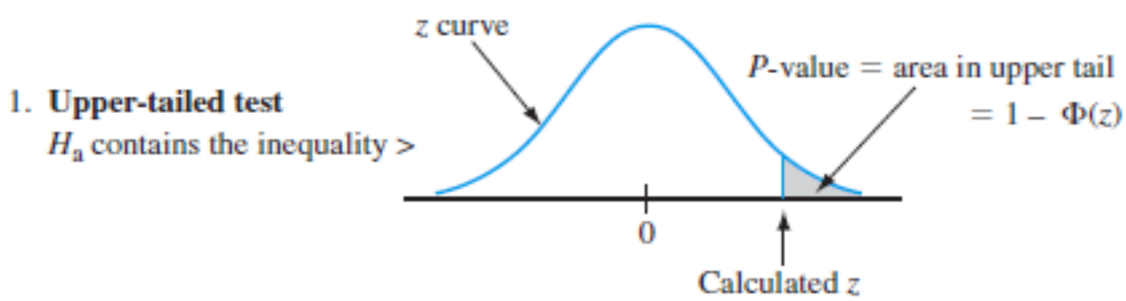
$$\begin{aligned} P\text{-value} &= P(Z \text{ either} \geq 1.5 \text{ or} \leq -1.5 \text{ when } H_0 \text{ is true}) \\ &= (\text{area under the } z \text{ curve to the right of } 1.5) \\ &\quad + (\text{area under the } z \text{ curve to the left of } -1.5) \\ &= 1 - \Phi(1.5) + \Phi(-1.5) = 2[1 - \Phi(1.5)] \\ &= 2(.0668) = .1336 \end{aligned}$$

This would also be the *P*-value if *x* = 97 results in *z* = -1.5. The important point is that because of the inequality $\neq$ in H_a, the *P*-value is the sum of an upper-tail area and a lower-tail area. By symmetry of the standard normal distribution, this becomes twice the area captured in the tail in which *z* falls.

A Normal Population with Known σ

It is natural to refer to this test as being *two-tailed* because z values far out in either tail of the z curve argue for rejection of H_0.

The test procedure is summarized in the accompanying box, and the P-value for each of the possible alternative hypotheses is illustrated in Figure 8.4.



A Normal Population with Known σ

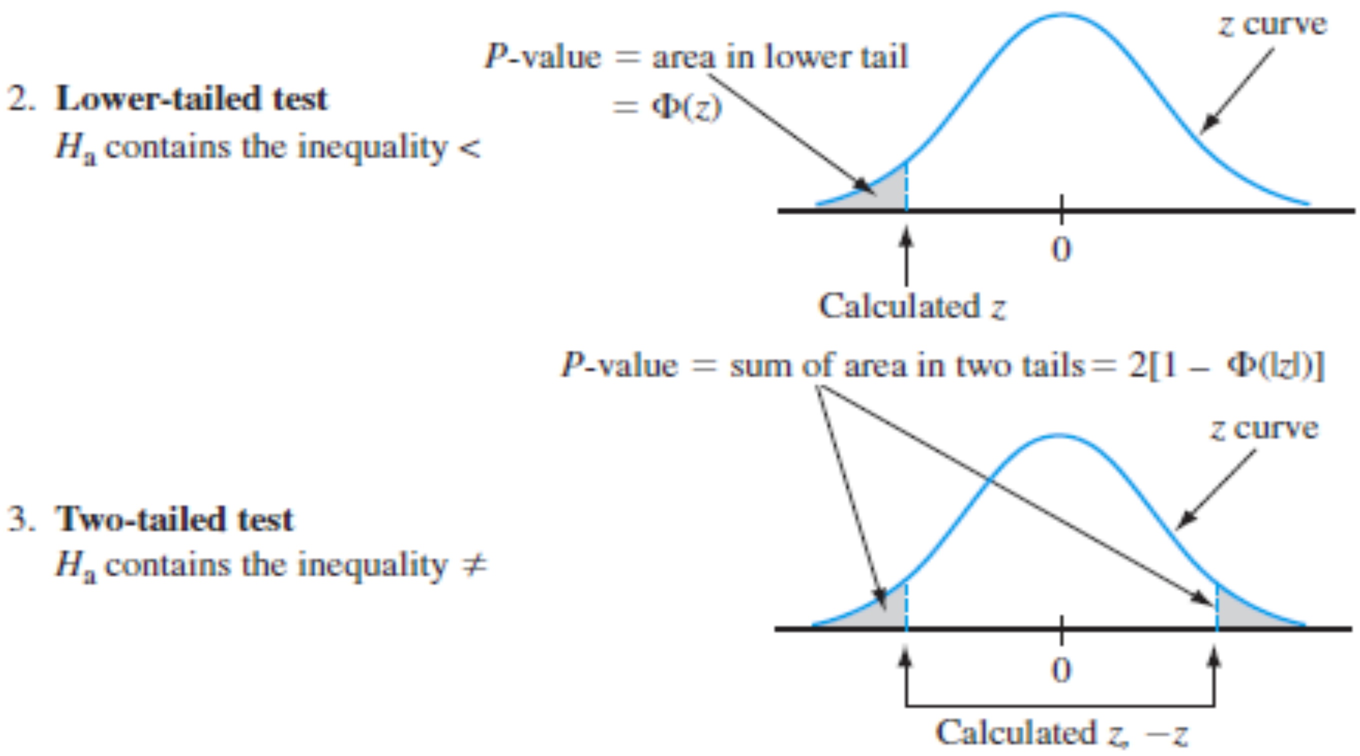


Figure 8.4 Determination of the P-value for a z test

A Normal Population with Known σ

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

Test statistic: $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$

Alternative Hypothesis	*P*-Value Determination
H_a: $\mu > \mu_0$	Area under the standard normal curve to the right of z
H_a: $\mu < \mu_0$	Area under the standard normal curve to the left of z
H_a: $\mu \neq \mu_0$	$2 \cdot$ (area under the standard normal curve to the right of $\lvert z \rvert$)

Assumptions: A normal population distribution with known value of σ.

A Normal Population with Known σ

Use of the following sequence of steps is recommended when testing hypotheses about a parameter. The plausibility of any assumptions underlying use of the selected test procedure should of course be checked before carrying out the test.

1. Identify the parameter of interest and describe it in the context of the problem situation.
2. Determine the null value and state the null hypothesis.
3. State the appropriate alternative hypothesis.
4. Give the formula for the computed value of the test statistic (substituting the null value and the known values of any other parameters, but *not* those of any sample-based quantities).

A Normal Population with Known σ

5. Compute any necessary sample quantities, substitute into the formula for the test statistic value, and compute that value.

6. Determine the *P*-value.

7. Compare the selected or specified significance level to the *P*-value to decide whether H_0 should be rejected, and state this conclusion in the problem context.

The formulation of hypotheses (Steps 2 and 3) should be done before examining the data, and the significance level *alpha* should be chosen prior to determination of the *P*-value.

Example 8.6

A manufacturer of sprinkler systems used for fire protection in office buildings claims that the true average system-activation temperature is 130°.

A sample of $n = 9$ systems, when tested, yields a sample average activation temperature of 131.08°F.

If the distribution of activation times is normal with standard deviation 1.5°F, does the data contradict the manufacturer's claim at significance level $\alpha = .01$?

Example 8.6

cont'd

1. Parameter of interest: μ = true average activation temperature.

2. Null hypothesis: H_0: $\mu = 130$ (null value = $\mu_0 = 130$).

3. Alternative hypothesis: H_a: $\mu \neq 130$ (a departure from the claimed value in *either* direction is of concern).

4. Test statistic value:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 130}{1.5/\sqrt{n}}$$

Example 8.6

cont'd

5. Substituting n = 9 and $\bar{x}$ = 131.08,

$$z = \frac{131.08 - 130}{1.5/\sqrt{9}} = \frac{1.08}{.5} = 2.16$$

That is, the observed sample mean is a bit more than 2 standard deviations above what would have been expected were H_0 true.

6. The inequality in H_a implies that the test is two-tailed, so the P- value results from doubling the captured tail area:

$$P\text{-value} = 2[1 - \Phi(2.16)] = 2(.0154) = .0308$$

Example 8.6

cont'd

7. Because *P*-value = .0308 $>$ .01 = α, H_0 cannot be rejected at significance level .01. The data does not give strong support to the claim that the true average differs from the design value of 130.

A Normal Population with Known σ

β **and Sample Size Determination** The *z* tests with known σ are among the few in statistics for which there are simple formulas available for β, the probability of a type II error.

Consider first the alternative H_a: $\mu > \mu_0$. The null hypothesis is rejected if *P*-value $\leq \alpha$, and the *P*-value is the area under the standard normal curve to the right of *z*. Suppose that $\alpha = .05$. The *z* critical value that captures an upper-tail area of .05 is $z_{.05}$ =1.645

A Normal Population with Known σ

Thus if the calculated test statistic value *z* is smaller than 1.645, the area to the right of *z* will be larger than .05 and the null hypothesis will then *not* be rejected.

Now substitute $(\bar{x} - \mu_0)/\sigma/\sqrt{n}$ in place of *z* in the inequality *z* $<$ 1.645 and manipulate to isolate *x* on the left (multiply both sides by $\sigma/\sqrt{n}$ and then add μ_0 to both sides). This gives the equivalent inequality $\bar{x} < \mu_0 + Z_a \cdot \sigma/\sqrt{n}$.

A Normal Population with Known σ

Now let μ' denote a particular value of μ that exceeds the null value μ_0. Then,

$$
\begin{aligned}
\beta(\mu') &= P(H_0 \text{ is not rejected when } \mu = \mu') \\
&= P(\overline{X} < \mu_0 + z_\alpha \cdot \sigma/\sqrt{n} \text{ when } \mu = \mu') \\
&= P\left(\frac{\overline{X} - \mu'}{\sigma/\sqrt{n}} < z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} \text{ when } \mu = \mu'\right) \\
&= \Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)
\end{aligned}
$$

A Normal Population with Known σ

As μ' increases, $\mu_0 - \mu'$ becomes more negative, so $\beta(\mu')$ will be small when μ' greatly exceeds μ_0

If σ increases, the probability of a type II error also increases.

Finally, if n increases the probability of a type II error decreases.

A Normal Population with Known σ

Suppose we fix α and also specify β for such an alternative value. In the sprinkler example, company officials might view $\mu' = 132$ as a very substantial departure from H_0: $\mu = 130$ and therefore wish $\beta(132) = .10$ in addition to $\alpha = .01$.

More generally, consider the two restrictions $P(\text{type I error}) = \alpha$ and $\beta(\mu') = \beta$ for specified α, μ' and β.

A Normal Population with Known σ

Then for an upper-tailed test, the sample size *n* should be chosen to satisfy

$$\Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) = \beta$$

This implies that

$$-z_\beta = \begin{array}{l} z \text{ critical value that} \\ \text{captures lower-tail area } \beta \end{array} = z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}$$

A Normal Population with Known σ

This equation is easily solved for the desired *n*. A parallel argument yields the necessary sample size for lower- and two-tailed tests as summarized in the next box.

Alternative Hypothesis	Type II Error Probability $\beta(\mu')$ for a Level α Test
H_a: $\mu > \mu_0$	$\Phi\left(z_\alpha + \dfrac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$
H_a: $\mu < \mu_0$	$1 - \Phi\left(-z_\alpha + \dfrac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$
H_a: $\mu \neq \mu_0$	$\Phi\left(z_{\alpha/2} + \dfrac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) - \Phi\left(-z_{\alpha/2} + \dfrac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$

where $\Phi(z) =$ the standard normal cdf.

The sample size n for which a level α test also has $\beta(\mu') = \beta$ at the alternative value μ' is

$$n = \begin{cases} \left[\dfrac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'}\right]^2 & \text{for a one-tailed (upper or lower) test} \\ \left[\dfrac{\sigma(z_{\alpha/2} + z_\beta)}{\mu_0 - \mu'}\right]^2 & \text{for a two-tailed test (an approximate solution)} \end{cases}$$

Example 8.7

Let μ denote the true average tread life of a certain type of tire.

Consider testing H_0: $\mu = 30{,}000$ versus H_a: $\mu > 30{,}000$ based on a sample of size $n = 16$ from a normal population distribution with $\sigma = 1500$.

A test with $\alpha = .01$ requires $z_\alpha = z_{.01} = 2.33$.

Example 8.7

cont'd

The probability of making a type II error when μ = 31,000 is

$$\beta(31{,}000) = \Phi\left(2.33 + \frac{30{,}000 - 31{,}000}{1500/\sqrt{16}}\right)$$

$$= \Phi(-.34)$$

$$= .3669$$

Example 8.7

cont'd

Since $z_{.1}$ = 1.28, the requirement that the level .01 test also have β(31,000) = .1 necessitates

$$n = \left[\frac{1500(2.33 + 1.28)}{30{,}000 - 31{,}000}\right]^2$$

$$= (-5.42)^2$$

$$= 29.32$$

The sample size must be an integer, so n = 30 tires should be used.

Large-Sample Tests

Large-Sample Tests

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

The key result to justify large-sample confidence intervals was used in the previous topic to justify large sample confidence intervals:

A large *n* implies that the standardized variable

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

has *approximately* a standard normal distribution.

Large-Sample Tests

Substitution of the null value μ_0 in place of μ yields the test statistic

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

which has approximately a standard normal distribution when H_0 is true.

Large-Sample Tests

The *P*-value is then determined exactly as was previously described in this section (e.g., $\Phi(z)$ when the alternative hypothesis is $H_a: \mu < \mu_0$). Rejecting H_0 when *P*-value $\leq \alpha$ gives a test with *approximate* significance level a.

The rule of thumb *n* > 40 will again be used to characterize a large sample size.

Example 8.8

A dynamic cone penetrometer (DCP) is used for measuring material resistance to penetration (mm/blow) as a cone is driven into pavement or subgrade.

Suppose that for a particular application it is required that the true average DCP value for a certain type of pavement be less than 30.

The pavement will not be used unless there is conclusive evidence that the specification has been met.

Example 8.8

cont'd

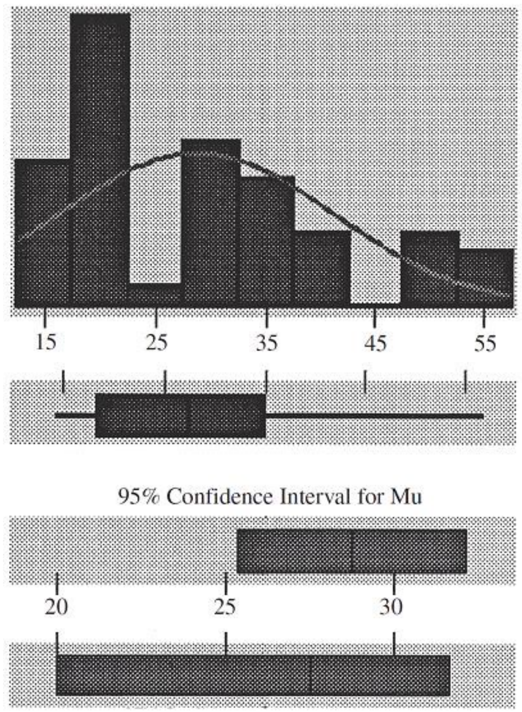
Let's state and test the appropriate hypotheses using the following data ("Probabilistic Model for the Analysis of Dynamic Cone Penetrometer Test Values in Pavement Structure Evaluation," *J. of Testing and Evaluation*, 1999: 7–14):

14.1	14.5	15.5	16.0	16.0	16.7	16.9	17.1	17.5	17.8
17.8	18.1	18.2	18.3	18.3	19.0	19.2	19.4	20.0	20.0
20.8	20.8	21.0	21.5	23.5	27.5	27.5	28.0	28.3	30.0
30.0	31.6	31.7	31.7	32.5	33.5	33.9	35.0	35.0	35.0
36.7	40.0	40.0	41.3	41.7	47.5	50.0	51.0	51.8	54.4
55.0	57.0								

Example 8.8

cont'd

Figure 8.5 shows a descriptive summary obtained from stats software



Variable: DCP

Anderson-Darling Normality Test

A-Squarect	1.902
P-Value:	0.000
Mean	28.7615
StDev	12.2647
Variance	150.423
Skewness	0.808264
Kurtosis	–3.9E–01
N	52
Minimum	14.1000
1st Quartile	18.2250
Median	27.5000
3rd Quartile	35.0000
Maximum	57.0000

95% Confidence Interval for Mu

25.3470 3.21761

95% Confidence Interval for Sigma

10.2784 15.2098

95% Confidence Interval for Median

20.0000 31.7000

Descriptive Statistics

Minitab descriptive summary for the DCP data of Example 8

Figure 8.5

Example 8.8

cont'd

The sample mean DCP is less than 30. However, there is a substantial amount of variation in the data (sample coefficient of variation = $s/\overline{x}$ = . 4265).

The fact that the mean is less than the design specification cutoff may be a consequence just of sampling variability.

Notice that the histogram does not resemble at all a normal curve, but the large-sample *z* tests do not require a normal population distribution.

Example 8.8

cont'd

1. μ = true average DCP value

2. H_0: μ = 30

3. H_a: μ < 30(so the pavement will not be used unless the null hypothesis is rejected)

4. $z = \dfrac{\bar{x} - 30}{s/\sqrt{n}}$

Example 8.8

cont'd

5. A test with significance level .05 rejects H_0 when $z \leq -1.645$ (a lower-tailed test).

6. With n = 52, $\bar{x}$ = 28.76, and s = 12.2647,

$$z = \frac{28.76 - 30}{12.2647/\sqrt{52}} = \frac{-1.24}{1.701} = -.73$$

7. Since –.73 > –1.645, H_0 cannot be rejected. We do not have compelling evidence for concluding that $\mu < 30$; use of the pavement is not justified.

3 The One-Sample *t* Test

The One-Sample *t* Test

When *n* is small, the Central Limit Theorem (CLT) can no longer be invoked to justify the use of a large-sample test.

We faced this same difficulty in obtaining a small-sample confidence interval (CI) for μ

Our approach here will be the same one used there: We will assume that the population distribution is at least approximately normal and describe test procedures whose validity rests on this assumption.

The One-Sample *t* Test

The key result on which tests for a normal population mean are based was used in the previous topic to derive the one-sample *t* CI:

If $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution, the standardized variable

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has a *t* distribution with $n - 1$ degrees of freedom (df).

The One-Sample t Test

Consider testing H_0: $\mu = \mu_0$ against H_a: $\mu > \mu_0$ by using the test statistic $T = (\overline{X} - \mu_0)/(S/\sqrt{n})$.

That is, the test statistic results from standardizing $\overline{X}$ under the assumption that H_0 is true (using $S/\sqrt{n}$, the estimated standard deviation of $\overline{X}$, rather than $\sigma/\sqrt{n}$).

When H_0 is true, this test statistic has a t distribution with n – 1 df.

The One-Sample *t* Test

Knowledge of the test statistic's distribution when H_0 is true (the "null distribution") allows us to determine the P-value.

The test statistic is really the same here as in the large sample case but is labeled T to emphasize that the reference distribution for P-value determination is a *t* distribution with a $n - 1$ df rather than the standard normal (z) distribution. Instead of being a z curve area as was the case for large-sample tests, the P-value will now be an area under the t_{n-1} curve.

The One-Sample t Test

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	P-Value Determination
H_a: $\mu > \mu_0$	Area under the t_{n-1} curve to the right of t
H_a: $\mu < \mu_0$	Area under the t_{n-1} curve to the left of t
H_a: $\mu \neq \mu_0$	$2 \cdot$ (Area under the t_{n-1} curve to the right of $\lvert t \rvert$)

Assumption: The data consists of a random sample from a normal population distribution.

The One-Sample *t* Test

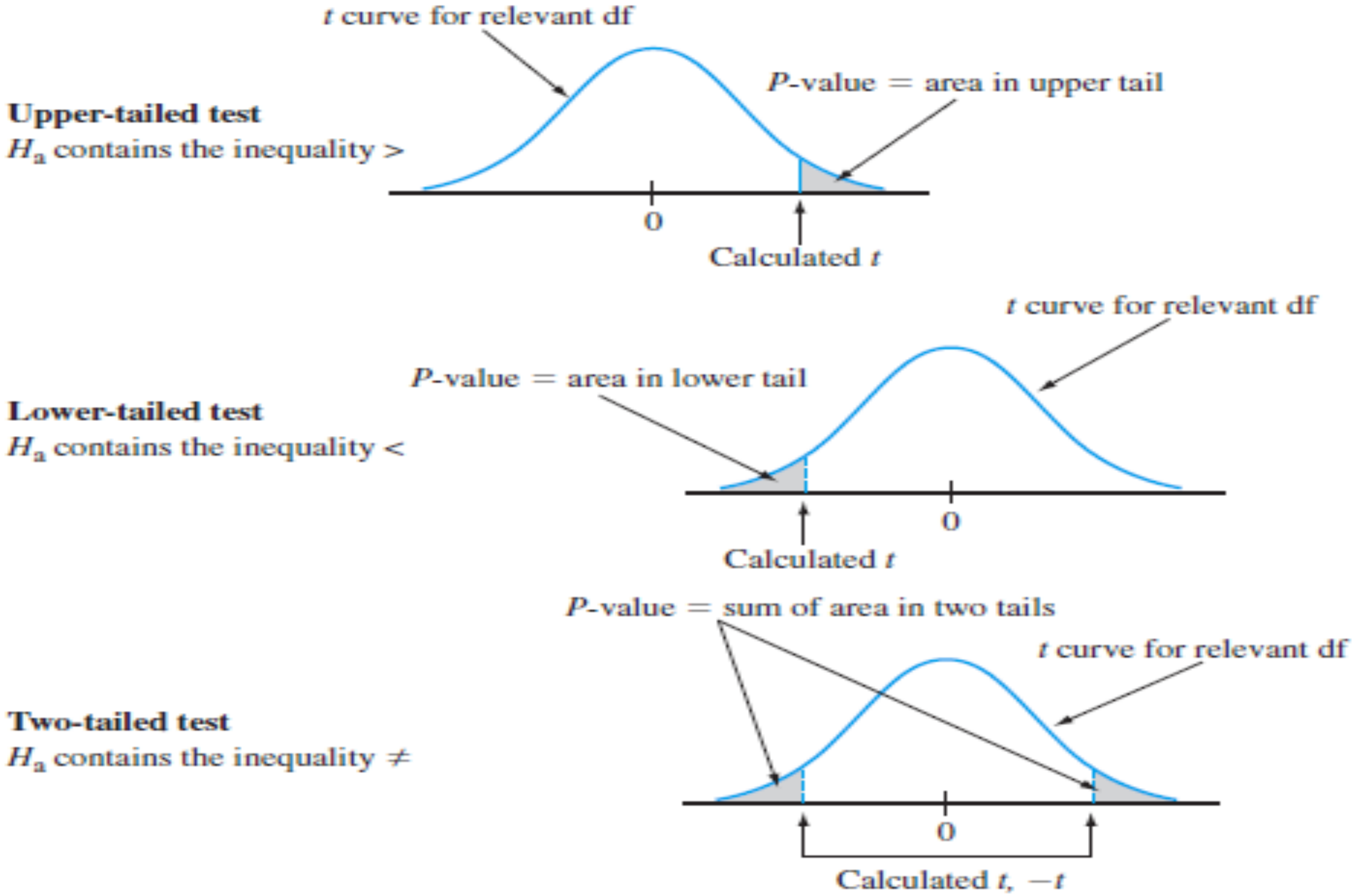


Figure 8.6 *P*-values for *t* tests

The One-Sample *t* Test

Suppose, for example, that a test of H_0: $\mu =$100 versus H_a: $\mu >$ 100 is based on the 8 df *t* distribution.

If the calculated value of the test statistic is *t* = 1.6, then the *P*-value for this upper-tailed test is .074. Because .074 exceeds .05, we would not be able to reject H_0 at a significance level of .05. If the alternative hypothesis is H_a: $\mu < 100$ and a test based on 20 df yields *t* = -3.2, then the *P*-value is the captured lower-tail area .002.

Example 8.9

Carbon nanofibers have potential application as heat-management materials, for composite reinforcement, and as components for nanoelectronics and photonics.

The accompanying data on failure stress (MPa) of fiber specimens was read from a graph in the article "Mechanical and Structural Characterization of Electrospun PAN-Derived Carbon Nanofibers" (*Carbon*, 2005: 2175–2185).

300	312	327	368	400	425	470	556	573	575
580	589	626	637	690	715	757	891	900	

Example 8.9

cont'd

Summary quantities include n = 19, $\bar{x}$ =562.68, s = 180.874, $s/\sqrt{n}$ = 41.495. Does the data provide compelling evidence for concluding that true average failure stress exceeds 500 MPa?

Example 8.9

cont'd

Let's carry out a test of the relevant hypotheses using a significance level of .05.

1. The parameter of interest is $\mu =$the average failure stress
2. The null hypothesis is $H_0: \mu = 500$
3. The Appropriate alternative hypothesis is $H_a: \mu > 500$ (so we'll believe that true average failure stress exceeds 500 only is the null hypothesis can be rejected).
4. The one-sample *t* test statistic is $T = (\bar{X} - 500/(S/\sqrt{n})$. Its value *t* for the given data results from replacing $\bar{X}$ by $\bar{x}$ and S by *s.*

Example 8.9

cont'd

5. The test-statistic value is t = (562.69 – 500)/41.495 = 1.51

6. The test is based on 19-1 = 18 df. Since the test is upper-tailed (because $>$ appears in H_a) it follows that P-value $\approx$.075

7. Because .075 $>$ $.05$, there is not enough evidence to justify rejecting the null hypothesis at significance level .05. Rather than conclude that the true average failure stress exceeds 500, it appears that sampling variability provides a plausible explanation for the fact that the sample mean exceeds 500 by a rather substantial amount.

4 Tests Concerning a Population Proportion

Tests Concerning a Population Proportion

Let *p* denote the proportion of individuals or objects in a population who possess a specified property (e.g., college students who graduate without any debt, or computers that do not need service during the warranty period).

If an individual or object with the property is labeled a success (*S*), then *p* is the population proportion of successes.

Tests concerning *p* will be based on a random sample of size *n* from the population.

Large-Sample Tests

Provided that n is small relative to the population size, X (the number of S's in the sample) has (approximately) a binomial distribution. Furthermore, if n itself is large [$np \geq 10$ and $n(1 - p) \geq 10$], both X and the estimator $\hat{p} = X/n$ are approximately normally distributed.

We first consider large-sample tests based on this latter fact and then turn to the small-sample case that directly uses the binomial distribution

Large-Sample Tests

Large-sample tests concerning p are a special case of the more general large-sample procedures for a parameter θ.

Let $\hat{\theta}$ be an estimator of θ that is (at least approximately) unbiased and has approximately a normal distribution.

The null hypothesis has the form H_0: $\theta = \theta_0$ where θ_0 denotes a number (the null value) appropriate to the problem context.

Large-Sample Tests

Suppose that when H_0 is true, the standard deviation of $\hat{\theta}$, $\sigma_{\hat{\theta}}$, involves no unknown parameters. For example, if $\theta = \mu$ and $\hat{\theta}$ = X, $\sigma_{\hat{\theta}} = \sigma_{\bar{x}} = \sigma/\sqrt{n}$, which involves no unknown parameters only if the value of σ is known.

A large-sample test statistic results from standardizing $\hat{\theta}$ under the assumption that H_0 is true (so that $E(\hat{\theta})$ = θ_0):

$$\text{Test statistic: } Z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$

Large-Sample Tests

If the alternative hypothesis is $H_a: \theta > \theta_0$, an upper-tailed test whose significance level is approximately α has *P*-value = 1 - Φ(*z*).

The other two alternatives, $H_a: \theta < \theta_0$and $H_a: \theta \neq \theta_0$, are tested using a lower-tailed *z* test and a two-tailed *z* test, respectively.

In the case $\theta = p, \sigma_{\hat{\theta}}$ will not involve any unknown parameters when H_0 is true, but this is atypical.

Large-Sample Tests

When $\sigma_{\hat{\theta}}$ does involve unknown parameters, it is often possible to use an estimated standard deviation $S_{\hat{\theta}}$ in place of $\sigma_{\hat{\theta}}$ and still have *Z* approximately normally distributed when H_0 is true (because this substitution does not increase variability in *Z* by very much).

The large-sample test of the previous section furnishes an example of this: Because σ is usually unknown, we use $s_{\hat{\theta}} = s_{\bar{x}} = s/\sqrt{n}$ in place of $\sigma/\sqrt{n}$ in the denominator of *z*.

Large-Sample Tests

The estimator $\hat{p} = X/n$ is unbiased ($E(\hat{p})$ = *p*) and its standard deviation is $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$.

These facts along with approximate normality were used in Section 7.2 to obtain a confidence interval for *p*. 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

Large-Sample Tests

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. The *P*-value for the test is then a *z* curve area, just as it was in the case of large-sample *z* tests concerning *m.*

Its calculation depends on which of the three inequalities in H_a is under consideration.

Large-Sample Tests

Null hypothesis: H_0: $p = p_0$

Test statistic value: $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$

Alternative Hypothesis	P-Value Determination
H_a: $p > p_0$	Area under the standard normal curve to the right of z
H_a: $p < p_0$	Area under the standard normal curve to the left of z
H_a: $p \neq p_0$	2·(Area under the standard normal curve to the right of $\lvert z \rvert$)

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

They are referred to as *upper-tailed*, *lower-tailed*, and *two-tailed*, respectively, for the three different alternative hypotheses.

Example 8.13

Student use of cell phones during class is perceived by many faculty to be an annoying but perhaps harmless distraction.

However, the use of a phone to text during an exam is a serious breach of conduct. The article "The Use and Abuse of Cell Phones and Text Messaging During Class: A Survey of College Students" (*College Teaching*, 2012: 1–9) reported that 27 of the 267 students in a sample admitted to doing this.

Can it be concluded at significance level .001 that more than 5% of all students in the population sampled had texted during an exam?

Example 8.13

1. The parameter of interest is the proportion *p* of the sampled population that has texted during an exam.

2. The null hypothesis is $H_0: p = .05$

3. The alternative hypothesis is $H_a: p > .05$

4. Since $np_0 = 267(.05) = 13.35 \geq 10$ and $nq_0 = 267(.95) = 253.65 \geq 10$, the large-sample *z* test can be used. The test statistic value is $z = (\hat{p} - .05)/\sqrt{(.05)(.95)/n}$.

Example 8.13

5. $\hat{p} = \frac{27}{267} = .1011$, from which $z = (.1011 - .05/\sqrt{(.05).95)/267} = .0511/.0133 = 3.84$

6. The *P*-value for this upper-tailed *z* test is $1 - \Phi(3.84) < 1 - \Phi(3.84) = .0003$

Example 8.13

7. The null hypothesis is rejected because *P*-value =.0003 $\leq .001 = \alpha$. The evidence for concluding that the population percentage of students who text during an exam exceeds 5% is very compelling. The cited article's abstract contained the following comment: "The majority of the students surveyed believe instructors are largely unaware of the extent to which texting and other cell phone activities engage students in the classroom".

β and Sample Size Determination

When H_0 is true, the test statistic *Z* has approximately a standard normal distribution. Now suppose that H_0 is *not* true and that $p = p'$.

Then *Z* still has approximately a normal distribution (because it is a linear function of $\hat{p}$), but its mean value and variance are no longer 0 and 1, respectively. Instead,

$$E(Z) = \frac{p' - p_0}{\sqrt{p_0(1 - p_0)/n}} \qquad V(Z) = \frac{p'(1 - p')/n}{p_0(1 - p_0)/n}$$

β and Sample Size Determination

The null hypothesis will not be rejected if *P*-value $> \alpha$. For an upper-tailed *z* test (inequality $>$ in H_a), we argued previously that this is equivalent to $z < z_\alpha$.

The probability of a type II error (not rejecting H_0 when it is false) is $\beta(p') = P(Z < z_\alpha when\, p = p')$. This can be computed by using the given mean and variance to standardize and then referring to the standard normal cdf. In addition, if it is desired that the level α test also have $\beta(p') = \beta$ for a specified value of β, this equation can be solved for the necessary *n* as in Section 8.2.

β and Sample Size Determination

General expressions for $\beta(p')$ and n are given in the accompanying box.

Alternative Hypothesis	$\beta(p')$
H_a: $p > p_0$	$\Phi\left[\dfrac{p_0 - p' + z_\alpha\sqrt{p_0(1-p_0)/n}}{\sqrt{p'(1-p')/n}}\right]$
H_a: $p < p_0$	$1 - \Phi\left[\dfrac{p_0 - p' - z_\alpha\sqrt{p_0(1-p_0)/n}}{\sqrt{p'(1-p')/n}}\right]$
H_a: $p \neq p_0$	$\Phi\left[\dfrac{p_0 - p' + z_{\alpha/2}\sqrt{p_0(1-p_0)/n}}{\sqrt{p'(1-p')/n}}\right] - \Phi\left[\dfrac{p_0 - p' - z_{\alpha/2}\sqrt{p_0(1-p_0)/n}}{\sqrt{p'(1-p')/n}}\right]$

The sample size n for which the level α test also satisfies $\beta(p') = \beta$ is

$$n = \begin{cases} \left[\dfrac{z_\alpha\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p'(1-p')}}{p' - p_0}\right]^2 & \text{one-tailed test} \\ \left[\dfrac{z_{\alpha/2}\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p'(1-p')}}{p' - p_0}\right]^2 & \text{two-tailed test (an approximate solution)} \end{cases}$$

Example 8.14

A package-delivery service advertises that at least 90% of all packages brought to its office by 9 a.m. for delivery in the same city are delivered by noon that day.

Let *p* denote the true proportion of such packages that are delivered as advertised and consider the hypotheses H_0: $p = .9$ versus H_a: $p < .9$.

If only 80% of the packages are delivered as advertised, how likely is it that a level .01 test based on n = 225 packages will detect such a departure from H_0?

Example 8.14

What should the sample size be to ensure that $\beta(.8) = .01$? With $\alpha = .01$, $p_0 = .9$, $p' = .8$, and $n = 225$,

$$\beta(.8) = 1 - \Phi\left(\frac{.9 - .8 - 2.33\sqrt{(.9)(.1)/225}}{\sqrt{(.8)(.2)/225}}\right)$$
$$= 1 - \Phi(2.00) = .0228$$

Thus the probability that H_0 will be rejected using the test when $p = .8$ is .9772; roughly 98% of all samples will result in correct rejection of H_0.

Example 8.14

Using $z_a = z_\beta = 2.33$ in the sample size formula yields

$$n = \left[\frac{2.33\sqrt{(.9)(.1)} + 2.33\sqrt{(.8)(.2)}}{.8 - .9}\right]^2 \approx 266$$

Small-Sample Tests

Test procedures when the sample size n is small are based directly on the binomial distribution rather than the normal approximation.

Consider the alternative hypothesis $H_a: p > p_0$ and again let X be the number of successes in the sample.

Then X is the test statistic.

Small-Sample Tests

When H_0 is true, *X* has a binomial distribution with parameters *n* and p_0, so

$$\begin{aligned} P\text{-value} &= P(X \geq x \text{ when } H_0 \text{ is true}) \\ &= P(X \geq x \text{ when } X \sim \text{Bin}(n, p_0)) \\ &= 1 - P(X \leq x - 1 \text{ when } X \sim \text{Bin}(n, p_0)) \\ &= 1 - B(x - 1; n, p_0) \end{aligned}$$

Because *X* has a discrete probability distribution, it is usually not possible to obtain a test for which *P*(type I error) is exactly the desired significance level *alpha* (e.g., .05 or .01; refer back to middle of page 323 for an example).

Small-Sample Tests

Let p' denote an alternative value of *p* ($p' > p_0$). When $p = p', X \sim Bin(n, p')$.

The probability of a type II error is then calculated by expressing the condition *P*-value $> \alpha$ in the equivalent form *x* $< c_\alpha$. Then

$$\begin{aligned}\beta(p') &= P(\text{type II error when } p = p') \\ &= P(X < c_\alpha \text{ when } X \sim \text{Bin}(n, p')) = B(c_\alpha - 1; n, p')\end{aligned}$$

Small-Sample Tests

That is, $\beta(p')$ is the result of a straightforward binomial probability calculation.

The sample size *n* necessary to ensure that a level α test also has specified β at a particular alternative value *p'* must be determined by trial and error using the binomial cdf.

Test procedures for $H_a\!: p < p_0$ and for $H_a\!: p \neq p_0$ are constructed in a similar manner. In the former case, the *P*-value is $\beta(x; n, p_0)$. The *P*-value when the alternative hypothesis is $H_a\!: p \neq p_0$ is twice the smaller of the two probabilities $\beta(x; n, p_0)$ and 1 - $B(x - 1; n, p_0)$.

Example 8.15

A plastics manufacturer has developed a new type of plastic trash can and proposes to sell them with an unconditional 6-year warranty.

To see whether this is economically feasible, 20 prototype cans are subjected to in accelerated life test to simulate 6 years of use.

The proposed warranty will be modified only if the sample data strongly suggests that fewer than 90% of such cans would survive the 6-year period.

Example 8.15

Let *p* denote the proportion of all cans that survive the accelerated test. The relevant hypotheses are $H_0: p = 9$ versus $H_a: p < .9$.

A decision will be based on the test statistic *X*, the number among the 20 that survive.

Because of the inequality in H_a, any value smaller than the observed value *x* is more contradictory to H_0 than is *x* itself. Therefore

$$P\text{-value} = P(X \leq x \text{ when } H_0 \text{ is true}) = B(x; 20, .9)$$

Example 8.15

From Appendix Table A.1, $B(15; 20, .9) = .043$, whereas $B(16; 20, .9) = .133$. The closest achievable significance level to .05 is therefore .043.

Since $B(14; 20, .9) = .011$, $H0$ would be rejected at this significance level if the accelerated test results in $x = 14$.

It would then be appropriate to modify the proposed warranty.

Example 8.15

Because P-value $\leq .043$ is equivalent to $x \leq 15$, the probability of a type II error for the alternative value $p' = .8$ is

$$\begin{aligned}\beta(.8) &= P(H_0 \text{ is not rejected when } X \sim \text{Bin}(20, .8)) \\ &= P(X \geq 16 \text{ when } X \sim \text{Bin}(20, .8)) \\ &= 1 - B(15; 20, .8) = 1 - .370 = .630\end{aligned}$$

That is, when $p = .8$, 63% of all samples consisting of n 5 20 cans would result in H_0 being incorrectly not rejected. This error probability is high because 20 is a small sample size and $p' = .8$ is close to the null value $p_0 = .9$.

5 Further Aspects of Hypothesis Testing

Statistical Versus Practical Significance

Statistical Versus Practical Significance

Statistical significance means simply that the null hypothesis was rejected at the selected significance level.

That is, in the judgment of the investigator, any observed discrepancy between the data and what would be expected were H_0 true cannot be explained solely by chance variation.

However, a small *P*-value, which would ordinarily indicate statistical significance, may be the result of a large sample size in combination with a departure from *H*0 that has little **practical significance**.

Statistical Versus Practical Significance

In many experimental situations, only departures from H_0 of large magnitude would be worthy of detection, whereas a small departure from H_0 would have little practical significance.

As an example, let μ denote the true average IQ of all children in the very large city of Euphoria. Consider testing H_0: $\mu = 100$ versus H_a: $\mu > 100$ where μ is the mean of a normal population with $\sigma = 15$.

But one IQ point is no big deal so the value $\mu = 101$ certainly does not represent a departure from H_0 that has practical significance.

Statistical Versus Practical Significance

For a reasonably large sample size n, this μ would lead to an $\overline{x}$ value near 101, so we would not want this sample evidence to argue strongly for rejection of H_0 when $\overline{x}$ = 101 is observed.

For various sample sizes, Table 8.1 records both the P-value when $\overline{x}$ = 101 and also the probability of not rejecting H_0 at level .01 when μ = 101.

n	P-Value When $\overline{x} = 101$	$\beta(101)$ for Level .01 Test
25	.3085	.9664
100	.1587	.9082
400	.0228	.6293
900	.0013	.2514
1600	.0000335	.0475
2500	.000000297	.0038
10,000	7.69×10^{-24}	.0000

An Illustration of the Effect of Sample Size on P-values and β

Table 8.1

Statistical Versus Practical Significance

The second column in Table 8.1 shows that even for moderately large sample sizes, the *P*-value of $\bar{x} = 101$ argues very strongly for rejection of H_0, whereas the observed $\bar{x}$ itself suggests that in practical terms the true value of μ differs little from the null value $\mu_0 = 100$.

The third column points out that even when there is little practical difference between the true μ and the null value, for a fixed level of significance a large sample size will almost always lead to rejection of the null hypothesis at that level.

The Relationship between Confidence Intervals and Hypothesis Tests

Suppose the standardized variable $Z = (\hat{\theta} - \theta / \hat{\sigma}_{\hat{\theta}}$ has (at least approximately) a standard normal distribution. The central *z* curve area captured between -1.96 and 1.96 is .95 (and the remaining area .05 is split equally between the two tails, giving area .025 in each one).

This implies that a confidence interval for θ with confidence level 95% is $\hat{\theta} \pm 1.96\hat{\sigma}_{\hat{\theta}}$.

Now consider testing $H_0: \theta = \theta_0$ versus $H_a: \theta \neq \theta_0$ at significance level .05 using the test statistic $Z = (\hat{\theta} - \theta_0)/\hat{\sigma}_{\hat{\theta}}$.

The phrase "*z* test" implies that when the null hypothesis is true, Z has (at least approximately) a standard normal distribution. So the *P*-value will be twice the area under the z curve to the right of $|z|$

This *P*-value will be less than or equal to .05, allowing for rejection of the null hypothesis, if and only if either $z \geq 1.96$ or $z \leq -1.96$. The null hypothesis will therefore not be rejected if $-1.96 < z < 1.96$.

Substituting the formula for *z* into this latter system of inequalities and manipulating them to isolate θ_0 gives the equivalent system $\hat{\theta} - 1.96\hat{\sigma}_{\hat{\theta}} < \theta_0 < \hat{\theta} + 1.96\hat{\sigma}_{\hat{\theta}}$.

The lower limit in this system is just the left endpoint of the 95% confidence interval, and the upper limit is the right endpoint of the interval.

What this says is that the null hypothesis will not be rejected if and only if the null value $\theta_0$lies in the confidence interval

Suppose, for example, that sample data yields the 95% CI (68.6, 72.0). Then the null hypothesis $H_0: \theta = 70$ cannot be rejected at significance level .05 because 70 lies in the CI.

But the null hypothesis $H_0: \theta = 65$ can be rejected because 65 does not lie in the CI.

There is an analogous relationship between a 99% CI and a test with significance level .01— the null hypothesis cannot be rejected if the null value lies in the CI and should be rejected if the null value is outside the CI.

The Relationship between Confidence Intervals and Hypothesis Tests

There is a duality between a two-sided confidence interval with confidence level 100(1 - α)% and the conclusion from a two-tailed test with significance level α.

Now consider testing $H_0: \theta = \theta_0$ against the alternative $H_a: \theta > \theta_0$ at significance level .01. Because of the inequality in H_a, the *P*-value is the area under the *z* curve to the right of the calculated *z*.

The *z* critical value 2.33 captures upper-tail area .01.

The Relationship between Confidence Intervals and Hypothesis Tests

Therefore the *P*-value (captured upper-tail area) will be at most .01 if and only if $z \geq 2.33$; we will not be able to reject the null hypothesis if and only if $z < 2.33$.

Again substituting the formula for *z* into this inequality and manipulating to isolate θ_0 gives the equivalent inequality $\hat{\theta} - 2.33\hat{\sigma}_{\hat{\theta}} < \theta_0$.

The lower limit of this inequality is the lower confidence bound for θ with a confidence level of 99%. So the null hypothesis won't be rejected at significance level .01 if and only if the null value exceeds the lower confidence bound.

The Relationship between Confidence Intervals and Hypothesis Tests

Thus there is a duality between a lower confidence bound and the conclusion from an upper-tailed test. This is why a stats software package will output a lower confidence bound when an upper-tailed test is performed.

If, for example, the 90% lower confidence bound is 25.3, i.e., $25.3 < \theta$ with confidence level 90%, then we would not be able to reject H_0: $\theta = 26$ versus H_a: $\theta > 26$ at significance level .10 but would be able to reject H_0: $\theta = 24$ in favor of H_a: $\theta > 26$.

There is an analogous duality between an upper confidence bound and the conclusion from a lower-tailed test. And there are analogous relationships for *t* tests and *t* confidence intervals or bounds.

Proposition

Let $(\hat{\theta}_L, \hat{\theta}_U)$ be a confidence interval for θ with confidence level $100(1-\alpha)\%$. Then a test of H_0: $\theta = \theta_0$ versus H_a: $\theta \neq \theta_0$ with significance level α rejects the null hypothesis if the null value θ_0 is not included in the CI and does not reject H_0 if the null value does lie in the CI. There is an analogous relationship between a lower confidence bound and an upper-tailed test, and also between an upper confidence bound and a lower-tailed test.

In light of these relationships, it is tempting to carry out a test of hypotheses by calculating the corresponding CI or CB. Don't yield to temptation!

Instead carry out a more informative analysis by determining and reporting the *P*-value.

Simultaneous Testing of Several Hypotheses

Simultaneous Testing of Several Hypotheses

Many published articles report the results of more than just a single test of hypotheses.

For example, the article "Distributions of Compressive Strength Obtained from Various Diameter Cores" (*ACI Materials J.*, 2012: 597– 606) considered the plausibility of Weibull, normal, and lognormal distributions as models for compressive strength distributions under various experimental conditions.

Table 3 of the cited article reported exact *P*-values for a total of 71 different tests

Simultaneous Testing of Several Hypotheses

Consider two different tests, one for a pair of hypotheses about a population mean and another for a pair of hypotheses about a population proportion—e.g., the mean wing length for adult Monarch butterflies and the proportion of schoolchildren in a particular state who are obese.

Assume that the sample used to test the first pair of hypotheses is selected independently of that used to test the second pair. Then if each test is carried out at significance level .05 (type I error probability .05),

$$
\begin{aligned}
P(\text{at least one type I error is committed}) &= 1 - P(\text{no type I errors are committed}) \\
&= 1 - P(\text{no type I error in the 1st test}) \cdot P(\text{no type I error in the 2nd test}) \\
&= 1 - (.95)^2 = 1 - .9025 = .0975
\end{aligned}
$$

Simultaneous Testing of Several Hypotheses

Thus the probability of committing at least one type I error when two independent tests are carried out is much higher than the probability that a type I error will result from a single test.

If three tests are independently carried out, each at significance level .05, then the probability that at least one type I error is committed is 1 - $(.95)^3$ = .1426.

Clearly as the number of tests increases, the probability of committing at least one type I error gets larger and in fact will approach 1.

Simultaneous Testing of Several Hypotheses

Suppose we want the probability of committing at least one type I error in two independent tests to be .05—an *experimentwise* error rate of .05. Then the significance level α for each test must be smaller than .05:

$$.05 = 1 - (1 - \alpha)^2 \Rightarrow 1 - \alpha = \sqrt{.95} = .975 \Rightarrow \alpha = .025$$

If the probability of committing at least one type I error in three independent tests is to be .05, the significance level for each one must be .017 (replace the square root by the cube root in the foregoing argument).

Simultaneous Testing of Several Hypotheses

As the number of tests increases, the significance level for each one must decrease to 0 in order to maintain an experimentwise error rate of .05.

Often it is not reasonable to assume that the various tests are independent of one another.

In the example cited at the beginning of this subsection, four different tests were carried out based on the same sample involving one particular type of concrete in combination with a specified core diameter and length-to-diameter ratio.

Simultaneous Testing of Several Hypotheses

It is then no longer clear how the experimentwise error rate relates to the significance level for each individual test. Let *Ai* denote the event that the *ith* test results in a type I error. Then in the case of *k* tests,

$$P(\text{at least one type I error}) = P(A_1 \cup A_2 \cup \ldots \cup A_k) \leq P(A_1) + \cdots + P(A_k) = k\alpha$$

(the inequality in the last line is called the *Bonferroni inequality*; it can be proved by induction on *k*).

Thus a significance level of .05/*k* for each test will ensure that the experimentwise significance level is at most .05.

Simultaneous Testing of Several Hypotheses

Again, the central idea here is that in order for the probability of at least one type I error among *k* tests to be small, the significance level for each individual test must be quite small.

If the significance level for each individual test is .05, for even a moderate number of tests it is rather likely that at least one type I error will be committed.

That is, with *alpha*=.05 for each test, when each null hypothesis is actually true, it is rather likely that at least one of the tests will yield a statistically significant result.

Simultaneous Testing of Several Hypotheses

This is why one should view a statistically significant result with skepticism when many tests are carried out using one of the traditional significance levels.