8 Tests of Hypotheses Based on a Single Sample

8.4 *P*-Values

P-Values

Using the rejection region method to test hypotheses entails first selecting a significance level α.

Then after computing the value of the test statistic, the null hypothesis H_0 is rejected if the value falls in the rejection region and is otherwise not rejected.

We now consider another way of reaching a conclusion in a hypothesis testing analysis.

This alternative approach is based on calculation of a certain probability called a *P-value.*

P-Values

One advantage is that the *P*-value provides an intuitive measure of the strength of evidence in the data against H_0.

Definition

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.

P-Values

This definition is quite a mouthful. Here are some key points:

- The *P*-value is a probability.
- 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!
- 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.

5

Example 14

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.

6

Example 14 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.

7

Example 14 cont'd

The test statistic value is

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

Now we must decide which values of *z* are at least as contradictory to H_0.

Let's first consider an easier task:
Which values of $\bar{x}$ are at least as contradictory to the null hypothesis as 2.06, the mean of the observations in our sample?

8

Example 14 cont'd

Because > appears in H_a, it should be clear that 2.10 is at least as contradictory to H_0 as is 2.06, and so in fact is *any* $\bar{x}$ value that exceeds 2.06.

But an $\bar{x}$ value that exceeds 2.06 corresponds to a value of z that exceeds 3.04. Thus the P-value is

$$P\text{-value} = P(Z \geq 3.04 \text{ when } \mu = 2.0)$$

Since the test statistic Z was created by subtracting the null value 2.0 in the numerator, when $\mu = 2.0$—i.e., when H_0 is true—Z has approximately a standard normal distribution.

9

Example 14 cont'd

As a consequence,

P-value = $P(Z \geq 3.04 \text{ when } \mu = 2.0)$

$\approx$ area under the z curve to the right of 3.04

$= 1 - \Phi(3.04)$

$= .0012$

10

P-Values

We will shortly illustrate how to determine the P-value for any z or t test—i.e., any test where the reference distribution is the standard normal distribution (and z curve) or some t distribution (and corresponding t curve).

For the moment, though, let's focus on reaching a conclusion once the P-value is available.

Because it is a probability, the P-value must be between 0 and 1.

11

P-Values

What kinds of P-values provide evidence against the null hypothesis?

Consider two specific instances:

- P-value = .250: In this case, fully 25% of all possible test statistic values are at least as contradictory to H_0 as the one that came out of our sample. So our data is not all that contradictory to the null hypothesis.

12

P-Values

- *P*-value = .0018: Here, only .18% (much less than 1%) of all possible test statistic values are at least as contradictory to H_0 as what we obtained. Thus the sample appears to be highly contradictory to the null hypothesis.

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*. That is, H_0 should be rejected in favor of H_a when the *P*-value is sufficiently small. So what constitutes "sufficiently small"?

13

P-Values

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.

14

P-Values

But if the *P*-value is equal to or less than α, then there is enough evidence to justify rejecting H_0.

In Example 14, 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 $\leq$ .01.

15

P-Values

However, suppose we select a significance level of only .001, which requires more substantial evidence from the data before H_0 can be rejected.
In this case we would not reject H_0 because .0012 $\leq$ .001.

How does the decision rule based on the *P*-value compare to the decision rule employed in the rejection region approach?

The two procedures—the rejection region method and the P-value method—are in fact identical.

16

P-Values

Whatever the conclusion reached by employing the rejection region approach with a particular α, the same conclusion will be reached via the *P*-value approach using that same α.

17

Example 15

The nicotine content problem involved testing H_0: $\mu = 1.5$ versus H_a: $\mu > 1.5$ using a z test (i.e., a test which utilizes the z curve as the reference distribution).

The inequality in H_a implies that the upper-tailed rejection region $z \geq z_\alpha$ is appropriate.

Suppose $z = 2.10$. Then using exactly the same reasoning as in Example 14 gives *P*-value = $1 - \Phi(2.10) = .0179$.

18

Example 15 cont'd

Consider now testing with several different significance levels:

$\alpha = .10 \Rightarrow z_\alpha = z_{.10} = 1.28 \Rightarrow 2.10 \geq 1.28 \Rightarrow$ reject H_0

$\alpha = .05 \Rightarrow z_\alpha = z_{.05} = 1.645 \Rightarrow 2.10 \geq 1.645 \Rightarrow$ reject H_0

$\alpha = .01 \Rightarrow z_\alpha = z_{.01} = 2.33 \Rightarrow 2.10 < 2.33 \Rightarrow$ do not reject H_0

19

Example 15 cont'd

Because *P-value* = .0179 ≤ .10 and also .0179 ≤ .05, using the *P*-value approach results in rejection of H_0 for the first two significance levels.

However, for $\alpha = .01$, 2.10 is not in the rejection region and .0179 is larger than .01.

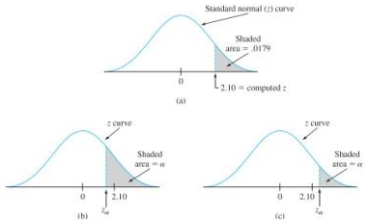
More generally, whenever α is smaller than the *P*-value .0179, the critical value z_α will be beyond the calculated value of z and H_0 cannot be rejected by either method.

20

Example 15

cont'd

This is illustrated in Figure 8.7.



Relationship between α and tail area captured by computed z: (a) tail area captured by computed z; (b) when $\alpha < .0179$, $z_\alpha < 2.10$ and H_0 is rejected; (c) when $\alpha < .0179$, $z_\alpha < 2.10$ and H_0 is not rejected

Figure 8.7

P-Values

Proposition

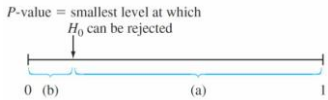
The *P*-value is the smallest significance level α at which the null hypothesis can be rejected.

Because of this, the *P*-value is alternatively referred to as the **observed significance level** (OSL) for the data. It is customary to call the data *significant* when H_0 is rejected and *not significant* otherwise.

The *P*-value is then the smallest level at which the data is significant.

P-Values

An easy way to visualize the comparison of the *P*-value with the chosen a is to draw a picture like that of Figure 8.8.



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

Figure 8.8

The calculation of the *P*-value depends on whether the test is upper-, lower-, or two-tailed. However, once it has been calculated, the comparison with α does not depend on which type of test was used.

Example 16

The true average time to initial relief of pain for a best-selling pain reliever is known to be 10 min.

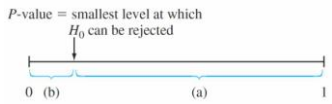
Let μ denote the true average time to relief for a company's newly developed reliever.

Suppose that when data from an experiment involving the new pain reliever is analyzed, the *P*-value for testing H_0: $\mu = 10$ versus H_a: $\mu < 10$ is calculated as .0384.

Example 16

cont'd

Since $\alpha = .05$ is larger than the *P*-value [.05 lies in the interval (a) of Figure 8.8], H_0 would be rejected by anyone carrying out the test at level .05.



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

Figure 8.8

However, at level .01, H_0 would not be rejected because .01 is smaller than the smallest level (.0384) at which H_0 can be rejected.

25

P-Values for *z* Tests

26

P-Values for *z* Tests

The *P*-value for a *z* test (one based on a test statistic whose distribution when H_0 is true is at least approximately standard normal) is easily determined from the information in Appendix Table A.3.

Consider an upper-tailed test and let *z* denote the computed value of the test statistic *Z*.

The null hypothesis is rejected if $z \geq z_\alpha$, and the *P*-value is the smallest a for which this is the case. Since z_α increases as α decreases, the *P*-value is the value of a for which $z = z_\alpha$.

27

P-Values for *z* Tests

That is, the *P*-value is just the area captured by the computed value *z* in the upper tail of the standard normal curve.

The corresponding cumulative area is $\Phi(z)$, so in this case *P*-value = $1 - \Phi(z)$.

An analogous argument for a lower-tailed test shows that the *P*-value is the area captured by the computed value *z* in the lower tail of the standard normal curve.

28

P-Values for *z* Tests

More care must be exercised in the case of a two-tailed test. Suppose first that z is positive. Then the *P*-value is the value of α satisfying $z = z_{\alpha/2}$ (i.e., computed z = upper-tail critical value).

This says that the area captured in the upper tail is half the *P*-value, so that *P*-value = $2[1 - \Phi(z)]$.

If z is negative, the *P*-value is the α for which $z = -z_{\alpha/2}$, or, equivalently, $-z = z_{\alpha/2}$ so that *P*-value = $2[1 - \Phi(-z)]$.

P-Values for *z* Tests

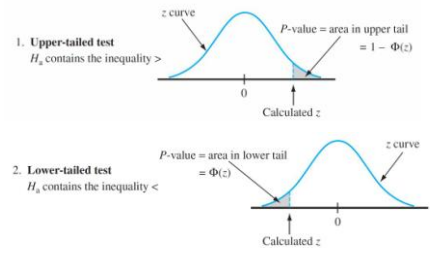
Since $-z = |z|$ when z is negative, *P*-value = $2[1 - \Phi(|z|)]$ for either positive or negative z.

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

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.

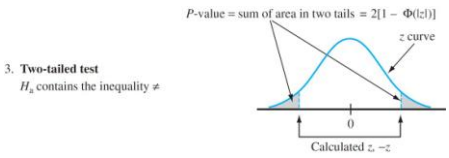


Determination of the P-value for a *z* test

Figure 8.9

P-Values for *z* Tests

cont'd



Determination of the P-value for a z test

Figure 8.9

Example 17

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 17

cont'd

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

Example 17

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.

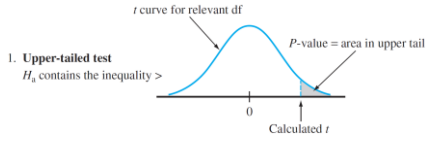
   At this significance level, there is insufficient evidence to conclude that true average thickness differs from the target value.

P-Values for *t* Tests

P-Values for *t* Tests

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

Figure 8.10 illustrates the three different cases. The number of df for the one-sample *t* test is $n - 1$.

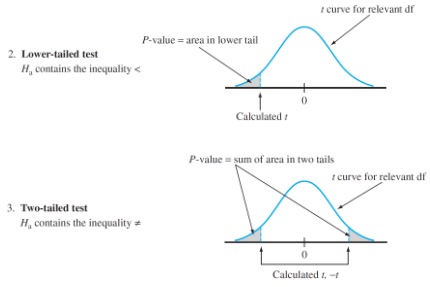


P-values for *t* tests

Figure 8.10

P-Values for *t* Tests

cont'd



P-values for *t* tests

Figure 8.10

P-Values for *t* Tests

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

So we have included another *t* table in Appendix Table A.8, one that contains a tabulation of upper-tail *t*-curve areas.

Each different column of the table is for a different number of df, and the rows are for calculated values of the test statistic *t* ranging from 0.0 to 4.0 in increments of .1.

P-Values for *t* Tests

For example, the number .074 appears at the intersection of the 1.6 row and the 8 df column, so the area under the 8 df curve to the right of 1.6 (an upper-tail area) is .074.

Because *t* curves are symmetric, .074 is also the area under the 8 df curve to the left of –1.6 (a lower-tail area).

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.

P-Values for *t* Tests

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 Appendix Table A.8 shows that the *P*-value is the captured lower-tail area .002. The null hypothesis can be rejected at either level .05 or .01.

Consider testing H_0: $\mu_1 - \mu_2 = 0$ versus H_a: $\mu_1 - \mu_2 \neq 0$; the null hypothesis states that the means of the two populations are identical, whereas the alternative hypothesis states that they are different without specifying a direction of departure from H_0.

41

P-Values for *t* Tests

If a *t* test is based on 20 df and $t = 3.2$, then the *P*-value for this two-tailed test is 2(.002) = .004.

This would also be the *P*-value for $t = -3.2$. The tail area is doubled because values both larger than 3.2 and smaller than –3.2 are more contradictory to H_0 than what was calculated (values farther out in *either* tail of the *t* curve).

42

Example 18

We considered a test of H_0: $\mu = 4$ versus H_a: $\mu \neq 4$ based on a sample of $n = 5$ observations from a normal population distribution.

The test statistic value was $-.594 \approx -.6$.

Looking to the 4 (= 5 – 1) df column of Appendix Table A.8 and then down to the .6 row, the entry is .290.

Because the test is two-tailed, this upper-tail area must be doubled to obtain the *P*-value.

The result is *P*-value ≈ .580.

43

Example 18

cont'd

This *P*-value is clearly larger than any reasonable significance level α(.01, .05, and even .10), so there is no reason to reject the null hypothesis.

The Minitab output as shown below has *P*-value = .594.

```
Test of mu = 4 vs not = 4
Variable   N   Mean  StDev  SE Mean  95% CI              T      P
glyc conc  5  3.814  0.718    0.321  (2.922, 4.706)  −0.58  0.594
```

P-values from software packages will be more accurate than what results from Appendix Table A.8 since values of *t* in our table are accurate only to the tenths digit.

44

More on Interpreting *P*-values

More on Interpreting *P*-values

The *P*-value resulting from carrying out a test on a selected sample is *not* the probability that H_0 is true, nor is it the probability of rejecting the null hypothesis.

Once again, it is the probability, calculated assuming that H_0 is true, of obtaining a test statistic value at least as contradictory to the null hypothesis as the value that actually resulted.

For example,
consider testing H_0: $\mu = 50$ against H_0: $\mu < 50$ using a lower-tailed z test.

More on Interpreting *P*-values

If the calculated value of the test statistic is $z = -2.00$, then
P-value = $P(Z < -2.00 \text{ when } \mu = 50)$

= area under the z curve to the left of –2.00

= 0.228

But if a second sample is selected, the resulting value of z will almost surely be different from –2.00, so the corresponding *P*-value will also likely differ from .0228.

More on Interpreting *P*-values

Because the test statistic value itself varies from one sample to another, the *P*-value will also vary from one sample to another.

That is, the test statistic is a random variable, and so the *P*-value will also be a random variable.

A first sample may give a *P*-value of .0228, a second sample may result in a *P*-value of .1175, a third may yield .0606 as the *P*-value, and so on.

More on Interpreting *P*-values

If H_0 is false, we hope the *P*-value will be close to 0 so that the null hypothesis can be rejected.

On the other hand, when H_0 is true, we'd like the *P*-value to exceed the selected significance level so that the correct decision to not reject H_0 is made.

The next example presents simulations to show how the *P*-value behaves both when the null hypothesis is true and when it is false.

49

Example 19

The fuel efficiency (mpg) of any particular new vehicle under specified driving conditions may not be identical to the EPA figure that appears on the vehicle's sticker.

Suppose that four different vehicles of a particular type are to be selected and driven over a certain course, after which the fuel efficiency of each one is to be determined.

Let μ denote the true average fuel efficiency under these conditions. Consider testing H_0: $\mu = 20$ versus H_0: $\mu > 20$ using the one-sample *t* test based on the resulting sample.

50

Example 19 cont'd

Since the test is based on $n - 1 = 3$ degrees of freedom, the *P*-value for an upper-tailed test is the area under the *t* curve with 3 df to the right of the calculated *t*.

Let's first suppose that the null hypothesis is true. We asked Minitab to generate 10,000 different samples, each containing 4 observations, from a normal population distribution with mean value $\mu = 20$ and standard deviation $\sigma = 2$.

51

Example 19 cont'd

The first sample and resulting summary quantities were

$x_1 = 20.830$, $x_2 = 22.232$, $x_3 = 20.276$, $x_4 = 17.718$

$\bar{x} = 20.264$ $s = 1.8864$ $t = \dfrac{20.264 - 20}{.1.8864/\sqrt{4}} = .2799$

The *P*-value is the area under the 3-df *t* curve to the right of .2799, which according to Minitab is .3989.

Using a significance level of .05, the null hypothesis would of course not be rejected.

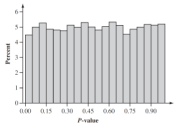
52

Example 19

cont'd

The values of t for the next four samples were –1.7591, .6082 –.7020 and 3.1053, with corresponding P-values .912, .293, .733, and .0265.

Figure 8.11(a) shows a histogram of the 10,000 P-values from this simulation experiment.



(a) $\mu = 20$

P-value simulation results for Example 19

Figure 8.11

Example 19

cont'd

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

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

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

Example 19

cont'd

This is because when H_0 is true and a test with significance level .05 is used, by definition the probability of rejecting H_0 is .05.

Looking at the histogram, it appears that the distribution of P-values is relatively flat. In fact, it can be shown that when H_0 is true, the probability distribution of the P-value is a uniform distribution on the interval from 0 to 1.

That is, the density curve is completely flat on this interval, and thus must have a height of 1 if the total area under the curve is to be 1.

Example 19

cont'd

Since the area under such a curve to the left of .05 is (.05)(1) = .05, we again have that the probability of rejecting H_0 when it is true that it is .05, the chosen significance level.

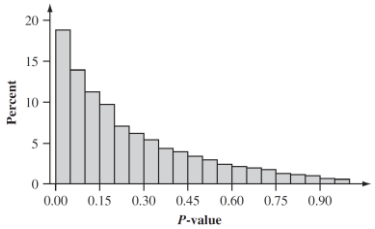
Now consider what happens when H_0 is false because $\mu = 21$. We again had Minitab generate 10,000 different samples of size 4 (each from a normal distribution with $\mu = 21$ and $\sigma = 2$), calculate $t = (\bar{x} - 20)(s/\sqrt{4})$ for each one, and then determine the P-value.

The first such sample resulted in $\bar{x} = 20.6411$ $s = .49637$ $t = 2.5832$, P-value = .0408.

Example 19

cont'd

Figure 8.11(b) gives a histogram of the resulting *P*-values.



(b) $\mu = 21$
P-value simulation results

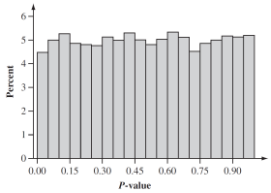
Figure 8.11

57

Example 19

cont'd

The shape of this histogram is quite different from that of Figure 8.11(a)— there is a much greater tendency for the *P*-value to be small (closer to 0) when $\mu = 21$ than when $\mu = 20$.



(a) $\mu = 20$
P-value simulation results

Figure 8.11

58

Example 19

cont'd

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

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.

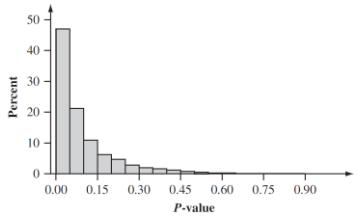
The difficulty is that the sample size is quite small and 21 is not very different from the value asserted by the null hypothesis.

59

Example 19

cont'd

Figure 8.11(c) illustrates what happens to the *P*-value when H_0 is false because $\mu = 22$ (still with $n = 4$ and $\sigma = 2$).



(c) $\mu = 22$
P-value simulation results

Figure 8.11

60

Example 19 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.

61

Example 19 cont'd

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.

The big idea of this example is that because the value of any test statistic is random, the *P*-value will also be a random variable and thus have a distribution.

The further the actual value of the parameter is from the value specified by the null hypothesis, the more the distribution of the *P*-value will be concentrated on values close to 0 and the greater the chance that the test will correctly reject H_0 (corresponding to smaller β).

62