Let us try to use the interval $[0.7, 1.9]$ for this test too. The null hypothesis in Example 9.34 is rejected at the 10% level in favor of a two-sided alternative, thus

$$|Z| > z_{\alpha/2} = z_{0.05}.$$

Then, either $Z < -z_{0.05}$ or $Z > z_{0.05}$. The first case is ruled out because the interval $[0.7, 1.9]$ consists of positive numbers, hence it cannot possibly support a left-tail alternative.

We conclude that $Z > z_{0.05}$, hence the test (9.22) results in rejection of H_0 at the 5% level of significance.

Conclusion. Our 90% confidence interval for $(\mu_X - \mu_Y)$ shows significant evidence, at the 5% level of significance, that the hardware upgrade was successful. ◊

Similarly, for the case of unknown variance(s).

> A level α T-test of $H_0 : \theta = \theta_0$ vs $H_A : \theta \neq \theta_0$
> accepts the null hypothesis
>
> if and only if
>
> a symmetric $(1 - \alpha)100\%$ confidence T-interval for θ contains θ_0.

Example 9.36 (Unauthorized use of a computer account, continued). A 99% confidence interval for the mean time between keystrokes is

$$[0.24;\ 0.34]$$

(Example 9.19 on p. 267 and data set `Keystrokes`). Example 9.28 on p. 283 tests whether the mean time is 0.2 seconds, which would be consistent with the speed of the account owner. The interval does not contain 0.2. Therefore, at a 1% level of significance, we have significant evidence that *the account was used by a different person.* ◊

9.4.10 P-value

How do we choose α?

So far, we were testing hypotheses by means of acceptance and rejection regions. In the last section, we learned how to use confidence intervals for two-sided tests. Either way, we need to know the *significance level* α in order to conduct a test. Results of our test depend on it.

How do we choose α, the probability of making type I sampling error, rejecting the true

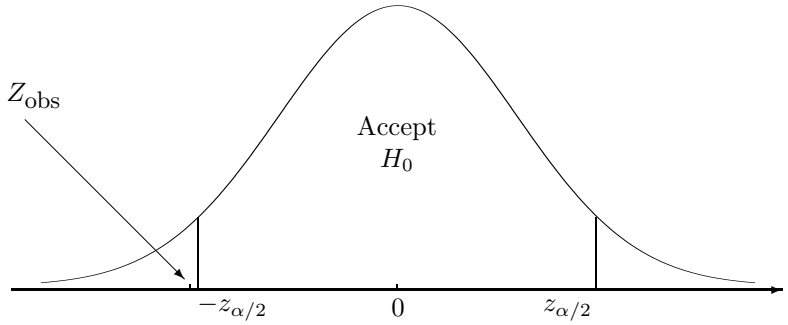

FIGURE 9.9: *This test is "too close to call": formally we reject the null hypothesis although the Z-statistic is almost at the boundary.*

hypothesis? Of course, when it seems too dangerous to reject true H_0, we choose a low significance level. How low? Should we choose $\alpha = 0.01$? Perhaps, 0.001? Or even 0.0001?

Also, if our *observed* test statistic $Z = Z_{\text{obs}}$ belongs to a rejection region but it is "too close to call" (see, for example, Figure 9.9), then how do we report the result? Formally, we should reject the null hypothesis, but practically, we realize that a slightly different significance level α could have expanded the acceptance region just enough to cover Z_{obs} and force us to accept H_0.

Suppose that the result of our test is crucially important. For example, the choice of a business strategy for the next ten years depends on it. In this case, can we rely so heavily on the choice of α? And if we rejected the true hypothesis just because we chose $\alpha = 0.05$ instead of $\alpha = 0.04$, then how do we explain to the chief executive officer that the situation was marginal? What is the statistical term for "too close to call"?

P-value

Using a P-value approach, we try not to rely on the level of significance. In fact, let us try to test a hypothesis using *all levels of significance*!

Considering all levels of significance (between 0 and 1 because α is a probability of Type I error), we notice:

Case 1. If a level of significance is *very low*, we *accept* the null hypothesis (see Figure 9.10a). A low value of

$$\alpha = \boldsymbol{P}\{ \text{ reject the null hypothesis when it is true } \}$$

makes it very unlikely to reject the hypothesis because it yields a very small rejection region. The right-tail area above the rejection region equals α.

Case 2. On the other extreme end, a *high significance level* α makes it likely to reject the null hypothesis and corresponds to a large rejection region. A sufficiently large α will produce such a large rejection region that will cover our test statistic, forcing us to *reject* H_0 (see Figure 9.10b).

Conclusion: there exists a boundary value between α-to-accept (case 1) and α-to-reject (case 2). This number is a *P-value* (Figure 9.11).

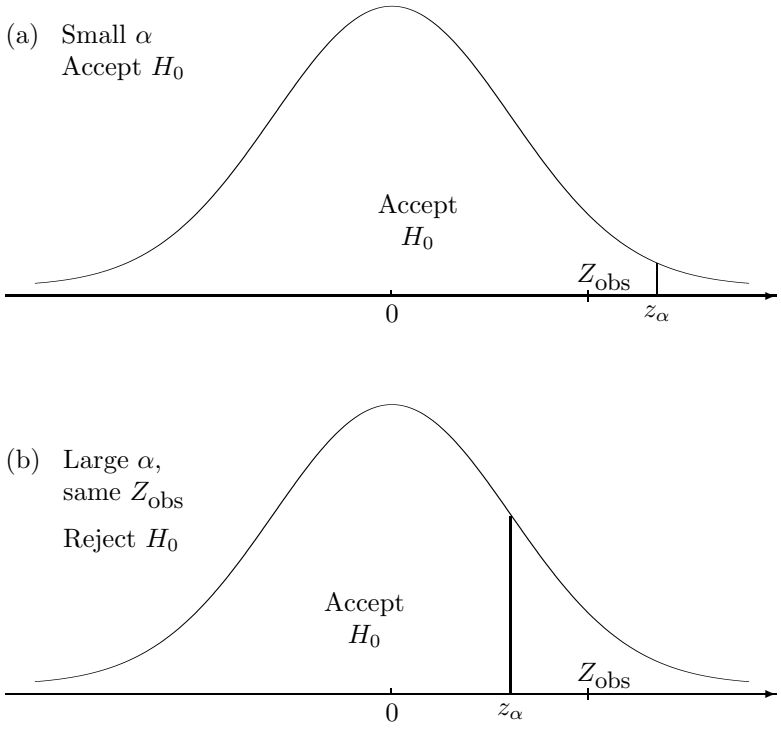

FIGURE 9.10: *(a) Under a low level of significance α, we accept the null hypothesis. (b) Under a high level of significance, we reject it.*

DEFINITION 9.9

P-value is the lowest significance level α that forces rejection of the null hypothesis.

P-value is also the highest significance level α that forces acceptance of the null hypothesis.

Testing hypotheses with a P-value

Once we know a P-value, we can indeed test hypotheses at *all* significance levels. Figure 9.11 clearly shows that for all $\alpha < P$ we accept the null hypothesis, and for all $\alpha > P$, we reject it.

Usual significance levels α lie in the interval [0.01, 0.1] (although there are exceptions). Then, a P-value greater than 0.1 exceeds all natural significance levels, and the null hypothesis should be accepted. Conversely, if a P-value is less than 0.01, then it is smaller than all natural significance levels, and the null hypothesis should be rejected. Notice that we did not even have to specify the level α for these tests!

Only if the P-value happens to fall between 0.01 and 0.1, we really have to think about the level of significance. This is the "marginal case," "too close to call." When we report

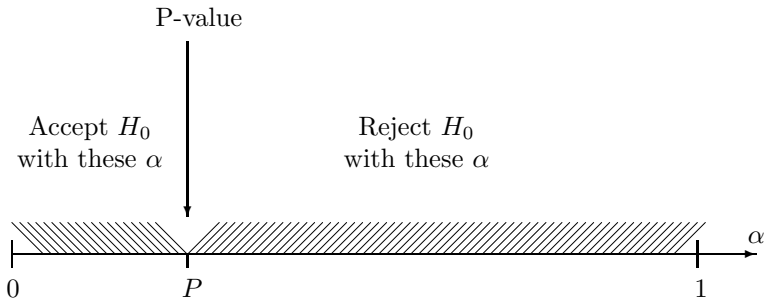

FIGURE 9.11: *P-value separates α-to-accept and α-to-reject.*

the conclusion, accepting or rejecting the hypothesis, we should always remember that with a slightly different α, the decision could have been reverted. When the matter is crucially important, a good decision is to collect more data until a more definitive answer can be obtained.

Testing H_0 with a P-value

For	$\alpha < P$,	accept H_0
For	$\alpha > P$,	reject H_0
Practically,		
If	$P < 0.01$,	reject H_0
If	$P > 0.1$,	accept H_0

Computing P-values

Here is how a P-value can be computed from data.

Let us look at Figure 9.10 again. Start from Figure 9.10a, gradually increase α, and keep your eye at the vertical bar separating the acceptance and rejection region. It will move to the left until it hits the observed test statistic Z_{obs}. At this point, our decision changes, and we switch from case 1 (Figure 9.10a) to case 2 (Figure 9.10b). Increasing α further, we pass the Z-statistic and start accepting the null hypothesis.

What happens at the border of α-to-accept and α-to-reject? Definition 9.9 says that this borderline α is the **P-value**,

$$P = \alpha.$$

Also, at this border our observed Z-statistic coincides with the critical value z_α,

$$Z_{\text{obs}} = z_\alpha,$$

and thus,

$$P = \alpha = \boldsymbol{P}\{Z \geq z_\alpha\} = \boldsymbol{P}\{Z \geq Z_{\text{obs}}\}.$$

In this formula, Z is any Standard Normal random variable, and Z_{obs} is our observed test

statistic, which is a concrete number, computed from data. First, we compute Z_{obs}, then use Table A4 to calculate

$$\boldsymbol{P}\{Z \geq Z_{\text{obs}}\} = 1 - \Phi(Z_{\text{obs}}).$$

P-values for the left-tail and for the two-sided alternatives are computed similarly, as given in Table 9.3.

This table applies to all the Z-tests in this chapter. It can be directly extended to the case of unknown standard deviations and T-tests (Table 9.4).

Understanding P-values

Looking at Tables 9.3 and 9.4, we see that *P-value* is the probability of observing a test statistic *at least as extreme as* Z_{obs} or t_{obs}. Being "extreme" is determined by the alternative. For a right-tail alternative, large numbers are extreme; for a left-tail alternative, small numbers are extreme; and for a two-sided alternative, both large and small numbers are extreme. In general, the more extreme test statistic we observe, the stronger support of the alternative it provides.

This creates another interesting definition of a P-value.

DEFINITION 9.10

P-value is the probability of observing a test statistic that is as extreme as or more extreme than the test statistic computed from a given sample.

The following philosophy can be used when we test hypotheses by means of a P-value.

We are deciding between the null hypothesis H_0 and the alternative H_A. Observed is a test statistic Z_{obs}. If H_0 were true, how likely would it be to observe such a statistic? In other words, are the observed data consistent with H_0?

A high P-value tells that this or even more extreme value of Z_{obs} is quite possible under H_0, and therefore, we see no contradiction with H_0. The null hypothesis is not rejected.

Conversely, a low P-value signals that such an extreme test statistic is unlikely if H_0 is true.

Hypothesis H_0	Alternative H_A	P-value	Computation
$\theta = \theta_0$	right-tail $\theta > \theta_0$	$\boldsymbol{P}\{Z \geq Z_{\text{obs}}\}$	$1 - \Phi(Z_{\text{obs}})$
	left-tail $\theta < \theta_0$	$\boldsymbol{P}\{Z \leq Z_{\text{obs}}\}$	$\Phi(Z_{\text{obs}})$
	two-sided $\theta \neq \theta_0$	$\boldsymbol{P}\{\lvert Z\rvert \geq \lvert Z_{\text{obs}}\rvert\}$	$2(1 - \Phi(\lvert Z_{\text{obs}}\rvert))$

TABLE 9.3: *P-values for Z-tests.*

Hypothesis H_0	Alternative H_A	P-value	Computation
$\theta = \theta_0$	right-tail $\theta > \theta_0$	$\boldsymbol{P}\{t \geq t_{\text{obs}}\}$	$1 - F_\nu(t_{\text{obs}})$
	left-tail $\theta < \theta_0$	$\boldsymbol{P}\{t \leq t_{\text{obs}}\}$	$F_\nu(t_{\text{obs}})$
	two-sided $\theta \neq \theta_0$	$\boldsymbol{P}\{\|t\| \geq \|t_{\text{obs}}\|\}$	$2(1 - F_\nu(\|t_{\text{obs}}\|))$

TABLE 9.4: *P-values for T-tests (F_ν is the cdf of T-distribution with the suitable number ν of degrees of freedom).*

However, we really observed it. Then, our data are not consistent with the hypothesis, and we should reject H_0.

For example, if $P = 0.0001$, there is only 1 chance in 10,000 to observe what we really observed. The evidence supporting the alternative is highly significant in this case.

Example 9.37 (How significant was the upgrade?). Refer to Examples 9.14 and 9.34. At the 5% level of significance, we know that the hardware upgrade was successful. Was it marginally successful or very highly successful? Let us compute the P-value.

Start with computing a Z-statistic,

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} = \frac{8.5 - 7.2}{\sqrt{\frac{1.8^2}{50} + \frac{1.8^2}{50}}} = 3.61.$$

From Table A4, we find that the P-value for the right-tail alternative is

$$P = \boldsymbol{P}\{Z \geq Z_{\text{obs}}\} = \boldsymbol{P}\{Z \geq 3.61\} = 1 - \Phi(3.61) = 0.0002.$$

The P-value is very low; therefore, we can reject the null hypothesis not only at the 5%, but also at the 1% and even 0.05% level of significance! We see now that the hardware upgrade was extremely successful. ◊

Example 9.38 (Quality inspection). In Example 9.26, we compared the quality of parts produced by two manufacturers by a two-sided test. We obtained a test statistic

$$Z_{\text{obs}} = -0.94.$$

The P-value for this test equals

$$P = \boldsymbol{P}\{|Z| \geq |-0.94|\} = 2(1 - \Phi(0.94)) = 2(1 - 0.8264) = 0.3472.$$

This is a rather high P-value (greater than 0.1), and the null hypothesis is not rejected. Given H_0, there is a 34% chance of observing what we really observed. No contradiction with H_0, and therefore, no evidence that the quality of parts is not the same. ◊

Table A5 is not as detailed as Table A4. Often we can only use it to bound the P-value from below and from above. Typically, it suffices for hypothesis testing.

Example 9.39 (Unauthorized use of a computer account, continued). How significant is the evidence in Examples 9.28 and 9.36 on pp. 283, 287 that the account was used by an unauthorized person?

Under the null hypothesis, our T-statistic has T-distribution with 17 degrees of freedom. In the previous examples, we rejected H_0 first at the 5% level, then at the 1% level. Now, comparing $t = 5.16$ from Example 9.28 with the entire row 17 of Table A5, we find that it exceeds all the critical values given in the table until $t_{0.0001}$. Therefore, a two-sided test rejects the null hypothesis at a very low level $\alpha = 0.0002$, and the P-value is $P < 0.0002$. *The evidence of an unauthorized use is very strong!*

$\Diamond$

9.5 Inference about variances

In this section, we'll derive confidence intervals and tests for the population variance $\sigma^2 = \text{Var}(X)$ and for the comparison of two variances $\sigma_X^2 = \text{Var}(X)$ and $\sigma_Y^2 = \text{Var}(Y)$. This will be *a new type of inference* for us because
(a) variance is a scale and not a location parameter,
(b) the distribution of its estimator, the sample variance, is not symmetric.

Variance often needs to be estimated or tested for quality control, in order to assess stability and accuracy, evaluate various risks, and also, for tests and confidence intervals for the population means when variance is unknown.

Recall that comparing two means in Section 9.3.5, we had to distinguish between the cases of equal and unequal variances. We no longer have to guess! In this section, we'll see how to test the null hypothesis $H_0 : \sigma_X^2 = \sigma_Y^2$ against the alternative $H_A : \sigma_X^2 \neq \sigma_Y^2$ and decide whether we should use the pooled variance (9.11) or the Satterthwaite approximation (9.12).

9.5.1 Variance estimator and Chi-square distribution

We start by estimating the population variance $\sigma^2 = \text{Var}(X)$ from an observed sample $\boldsymbol{X} = (X_1, \dots, X_n)$. Recall from Section 8.2.4 that σ^2 is estimated *unbiasedly* and *consistently* by the sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 .$$

The summands $\left(X_i - \overline{X}\right)^2$ are not quite independent, as the Central Limit Theorem on p. 93 requires, because they all depend on $\overline{X}$. Nevertheless, the distribution of s^2 is approximately Normal, under mild conditions, when the sample is large.

For small to moderate samples, the distribution of s^2 is not Normal at all. It is not even symmetric. Indeed, why should it be symmetric if s^2 is always non-negative!