8.5: Critical values, p-values, and significance level

A low probability value casts doubt on the null hypothesis. How low must the probability value be in order to conclude that the null hypothesis is false? Although there is clearly no right or wrong answer to this question, it is conventional to conclude the null hypothesis is false if the probability value is less than 0.05. More conservative researchers conclude the null hypothesis is false only if the probability value is less than 0.01. When a researcher concludes that the null hypothesis is false, the researcher is said to have rejected the null hypothesis. The probability value below which the null hypothesis is rejected is called the α level or simply α ("alpha"). It is also called the significance level. If α is not explicitly specified, assume that $\alpha = 0.05$.

The significance level is a threshold we set before collecting data in order to determine whether or not we should reject the null hypothesis. We set this value beforehand to avoid biasing ourselves by viewing our results and then determining what criteria we should use. If our data produce values that meet or exceed this threshold, then we have sufficient evidence to reject the null hypothesis; if not, we fail to reject the null (we never "accept" the null).

There are two criteria we use to assess whether our data meet the thresholds established by our chosen significance level, and they both have to do with our discussions of probability and distributions. Recall that probability refers to the likelihood of an event, given some situation or set of conditions. In hypothesis testing, that situation is the assumption that the null hypothesis value is the correct value, or that there is no effect. The value laid out in H0 is our condition under which we interpret our results. To reject this assumption, and thereby reject the null hypothesis, we need results that would be very unlikely if the null was true. Now recall that values of z which fall in the tails of the standard normal distribution represent unlikely values. That is, the proportion of the area under the curve as or more extreme than z is very small as we get into the tails of the distribution. Our significance level corresponds to the area under the tail that is exactly equal to α: if we use our normal criterion of $\alpha = .05$, then 5% of the area under the curve becomes what we call the rejection region (also called the critical region) of the distribution. This is illustrated in Figure 8.5.1.

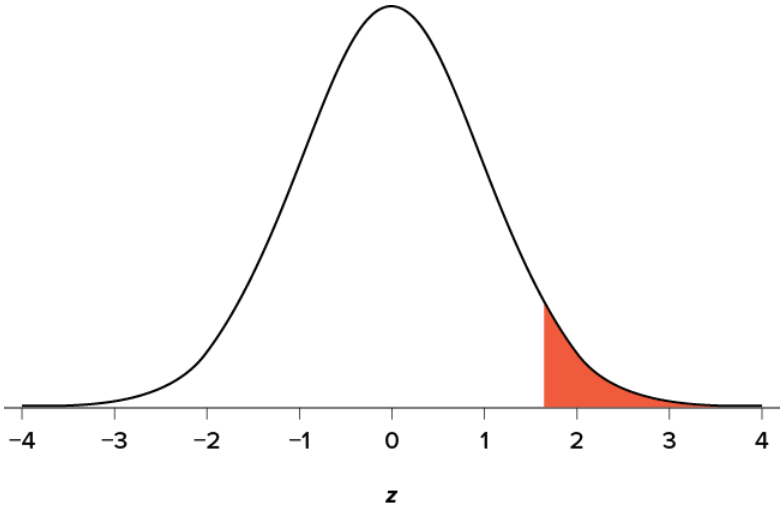

Figure 8.5.1: The rejection region for a one-tailed test.

Image Credit: Judy Schmitt, from Cote et al, 2021

The shaded rejection region takes us 5% of the area under the curve. Any result which falls in that region is sufficient evidence to reject the null hypothesis.

The rejection region is bounded by a specific z-value, as is any area under the curve. In hypothesis testing, the value corresponding to a specific rejection region is called the critical value, z_{crit} ("z-crit") or $z*$ (hence the other name "critical region"). Finding the critical value works exactly the same as finding the z-score corresponding to any area under the curve like we did in Unit 1. If we go to the normal table, we will find that the z-score corresponding to 5% of the area under the curve is equal to 1.645 ($z = 1.64$ corresponds to 0.0405 and $z = 1.65$ corresponds to 0.0495, so .05 is exactly in between them) if we go to the right and -1.645 if we go to the left. The direction must be determined by your alternative hypothesis, and drawing then shading the distribution is helpful for keeping directionality straight.

Suppose, however, that we want to do a non-directional test. We need to put the critical region in both tails, but we don't want to increase the overall size of the rejection region (for reasons we will see later). To do this, we simply split it in half so that an equal proportion of the area under the curve falls in each tail's rejection region. For α = .05, this means 2.5% of the area is in each tail, which, based on the z-table, corresponds to critical values of $z* = \pm 1.96$. This is shown in Figure 8.5.2.

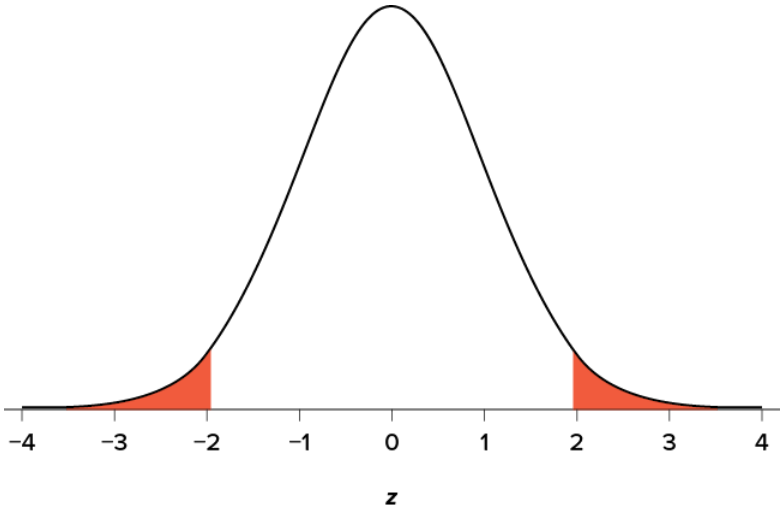

Figure 8.5.2: Two-tailed rejection region.

Image Credit: Judy Schmitt, from Cote et al, 2021

Thus, any z-score falling outside ±1.96 (greater than 1.96 in absolute value) falls in the rejection region. When we use z-scores in this way, the obtained value of z (sometimes called z-obtained) is something known as a test statistic, which is simply an inferential statistic used to test a null hypothesis. The formula for our z-statistic has not changed:

$$z = \frac{\overline{\mathrm{X}} - \mu}{\bar{\sigma}/\sqrt{\mathrm{n}}} \tag{8.5.1}$$

To formally test our hypothesis, we compare our obtained z-statistic to our critical z-value. If $\mathrm{Z}_{\mathrm{obt}} > \mathrm{Z}_{\mathrm{crit}}$, that means it falls in the rejection region (to see why, draw a line for z = 2.5 on Figure 8.5.1 or Figure 8.5.2) and so we reject H_0. If $\mathrm{Z}_{\mathrm{obt}} < \mathrm{Z}_{\mathrm{crit}}$, we fail to reject. Remember that as z gets larger, the corresponding area under the curve beyond z gets smaller. Thus, the proportion, or p-value, will be smaller than the area for α, and if the area is smaller, the probability gets smaller. Specifically, the probability of obtaining that result, or a more extreme result, under the condition that the null hypothesis is true gets smaller.

The z-statistic is very useful when we are doing our calculations by hand. However, when we use computer software, it will report to us a p-value, which is simply the proportion of the area under the curve in the tails beyond our obtained z-statistic. We can directly compare this p-value to α to test our null hypothesis: if $p < \alpha$, we reject H_0, but if $p > \alpha$, we fail to reject. Note also that the reverse is always true: if we use critical values to test our hypothesis, we will always know if p is greater than or less than α. If we reject, we know that $p < \alpha$ because the obtained z-statistic falls farther out into the tail than the critical z-value that corresponds to α, so the proportion (p-value) for that z-statistic will be smaller. Conversely, if we fail to reject, we know that the proportion will be larger than α because the z-statistic will not be as far into the tail. This is illustrated for a one-tailed test in Figure 8.5.3.

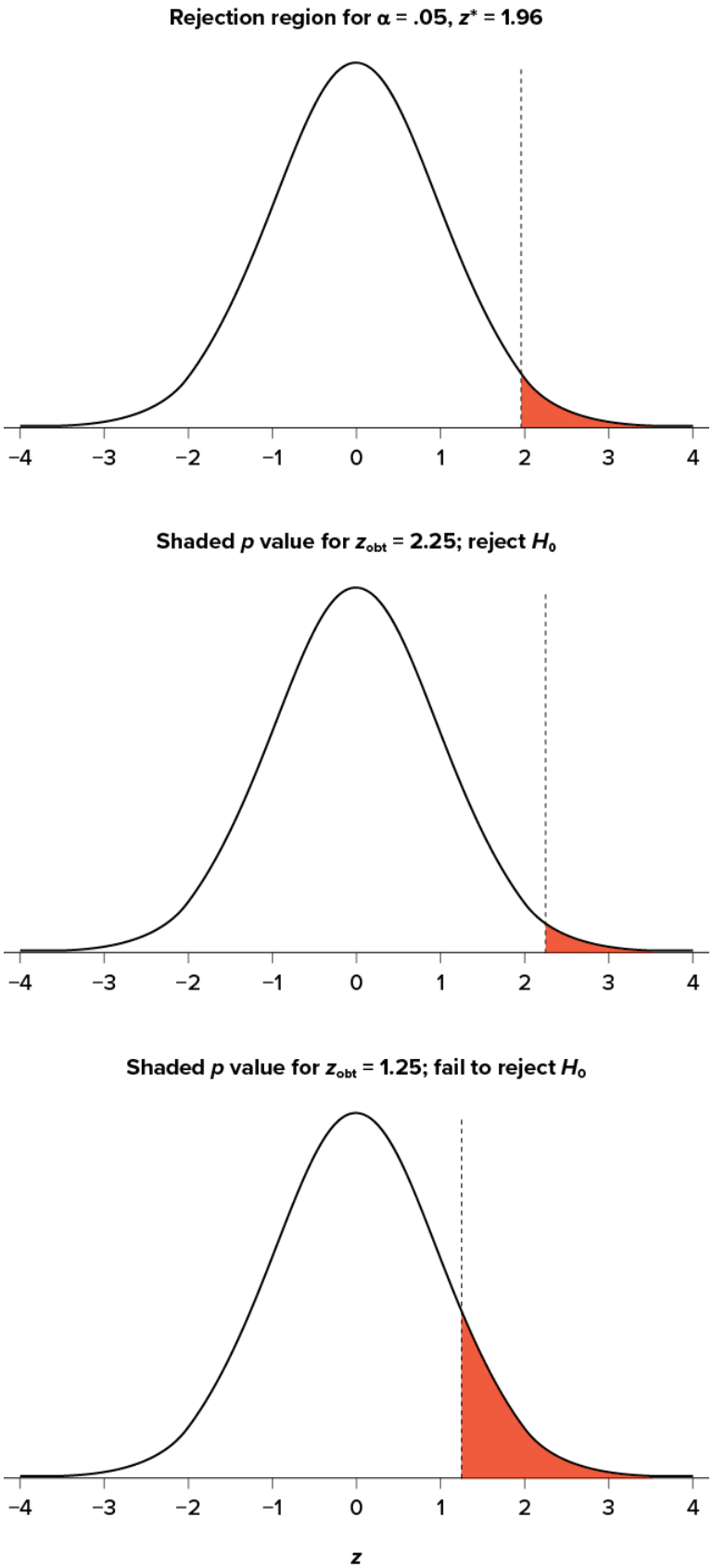

Figure 8.5.3: Relation between α, z_{obt}, and p

Image Credit: Judy Schmitt, from Cote et al, 2021

When the null hypothesis is rejected, the effect is said to be statistically significant. For example, in the Physicians Reactions case study, the probability value is 0.0057. Therefore, the effect of obesity is statistically significant and the null hypothesis that obesity makes no difference is rejected. It is very important to keep in mind that statistical significance means only that the null hypothesis of exactly no effect is rejected; it does not mean that the effect is important, which is what "significant" usually means. When an effect is significant, you can have confidence the effect is not exactly zero. Finding that an effect is significant does not tell you about how large or important the effect is. Do not confuse statistical significance with practical significance. A small effect can be highly significant if the sample size is large enough. Why does the word "significant" in the phrase "statistically significant" mean something so different from other uses of the word? Interestingly, this is because the meaning of "significant" in everyday language has changed. It turns out that when the procedures for hypothesis testing were developed, something was "significant" if it signified something. Thus, finding that an effect is statistically significant signifies that the effect is real and not due to chance. Over the years, the meaning of "significant" changed, leading to the potential misinterpretation.
