18.650: Fundamentals of Statistics October 20, 2023

Lecture 16 — p-values

Prof. Philippe Rigollet *Scribe: Anya Katsevich*

Overview. p-values answer the following question:

Can we convey information about how close the call was between H_0 and H_1?

The p-value is a number $p \in [0,1]$. If p is small, this will indicate we have evidence against H_0. If p is large, this will indicate we do *not* have enough evidence to reject H_0.

The evidence scale for the p-value:

$< 1\%$	very strong evidence against H_0
$1\% - 5\%$	strong evidence against H_0
$5\% - 10\%$	weak evidence against H_0
$> 10\%$	little or no evidence against H_0

This is officially accepted wording — use this and not e.g. "so-so" evidence. Before we get into the definition, we note that the p-value is NOT $\mathbb{P}(H_0 \text{ true} \mid \text{data})$ and NOT $\mathbb{P}(H_0 \text{ true})$

1 Definition and Examples

Definition 1.1: p value

The p-value of a test for a given dataset $X_1, \ldots, X_n$ is the smallest level at which the test rejects H_0.

This leads to the following rule:

$$\text{Reject at level } \alpha \iff \text{p-value} \leq \alpha$$

In other words, you reject H_0 for your given dataset whenever the level α of the test is above the p-value. Conversely, if you rejected H_0 for a test of level α, then the p-value must have been smaller than α.

1.1 ER example

Consider the ER waiting time example. Recall that the rejection region was

$$\mathcal{R}_\alpha = \left\{ \frac{\sqrt{n}}{\hat{\sigma}}(\bar{X}_n - 30) > z_\alpha \right\}.$$

Suppose we observed $\bar{X}_n = 33.4$ for a sample size $n = 164$, and suppose $\hat{\sigma} = 12$. Then

$$\frac{\sqrt{n}}{\hat{\sigma}}(\bar{X}_n - 30) = \frac{\sqrt{164}}{12}(33.4 - 30) = 3.62. \tag{1}$$

If we want the test to have level $\alpha = 5\%$, we take the cutoff to be $z_{5\%} = 1.64$. Since $3.62 > z_{5\%} = 1.64$, a test of level 5% would reject H_0. But a test of level 2.5% would also reject H_0 since $3.62 > z_{2.5\%} = 1.96$. More generally, for any α such that $3.62 \geq z_\alpha$, we know the test of level α would reject H_0. See Figure 1. Therefore, by the definition of p-value, we have

$$\text{p-value } = \max\{\alpha \text{ such that } z_\alpha \leq 3.62\}. \tag{2}$$

Clearly, this is given by α such that $z_\alpha = 3.62$ exactly, which is just the area to the right of 3.62 under the standard Gaussian distribution. To summarize, the p value in this example is

$$\text{p-value } = \mathcal{P}(\mathcal{N}(0,1) \geq 3.62) = 1 - \Phi(3.62) \leq 0.0002 = 0.02\%$$

According to our evidence scale, we have very strong evidence to reject H_0!

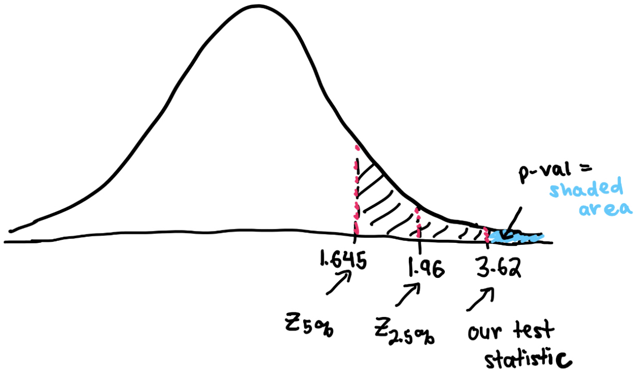

Figure 1: The p value in the ER example.

1.2 Kiss example

Now we return to the kiss example, where we take $H_0 : p = 1/2$ (no head turning preference) and $H_1 : p \neq 1/2$ (head-turning preference in some direction). Our estimator for p is $\hat{p} = \bar{X}_n$, and we have $\hat{\sigma}^2 = \bar{X}_n(1 - \bar{X}_n)$ in this case. Note that the true σ^2 is $\sigma^2 = p(1-p)$, the variance of $\text{Ber}(p)$, and we have replaced p by $\bar{X}_n$ to get $\hat{\sigma}^2$.

To test the hypothesis, consider a test of level α, which rejects the null if $\frac{\sqrt{n}}{\bar{X}_n(1-\bar{X}_n)}|\bar{X}_n - 1/2| \geq z_{\alpha/2}$. In the Nature article, $\bar{X}_n = 0.645$ was observed, for $n = 124$. Therefore,

$$\frac{\sqrt{n}}{\sqrt{\bar{X}_n(1-\bar{X}_n)}}|\bar{X}_n - 0.5| = \frac{\sqrt{124}}{\sqrt{0.645(1-0.645)}}|0.645 - 0.5| = 3.37$$

Therefore, the p-value is the area under the standard normal distribution to the left of -3.37 and to the right of 3.37:

$$\text{p-value} = 2(1 - \Phi(3.37)) = 0.0008 = 0.08\% \implies \text{very strong evidence!}$$

See Figure 2 for a visualization of the p value.

An alternative test. Recall that if the true distribution is $\text{Ber}(p)$, then the true variance is $\sigma^2 = p(1-p)$. Typically, we run into the issue that we don't know the variance. But since H_0 is $p = 1/2$, we can compute the Type I error (and thus get the size of the test) by plugging in $p = 1/2$! In other words, the test which rejects if $\left|\frac{\bar{X}_n - 1/2}{\sqrt{(1/2)(1/2)}}\right| > z_{\alpha/2}$ is valid because the Type I error is

$$\mathbb{P}_{p=1/2}\left(\sqrt{n}\left|\frac{\bar{X}_n - 1/2}{\sqrt{(1/2)(1/2)}}\right| > z_{\alpha/2}\right) = \mathbb{P}_{p=1/2}\left(\sqrt{n}\left|\frac{\bar{X}_n - 1/2}{\sqrt{p(1-p)}}\right| > z_{\alpha/2}\right) = \alpha.$$

For this new test statistic, let's compute the p-value when $n = 164$ and $\bar{X}_n = 0.645$. First of all, we compute

$$\sqrt{n}\frac{\bar{X}_n - 1/2}{\sqrt{(1/2)(1/2)}} = 2\sqrt{124}(0.645 - 0.5) = 3.22.$$

Therefore, the p-value is

$$\text{p-value} = \mathbb{P}(|\mathcal{N}(0,1)| \geq 3.22) = 2(1 - \Phi(3.22)) = 0.0014 = 0.14\%,$$

which is slightly different than the 0.08% we got before. However, we still have very strong evidence to reject the null when using this test.

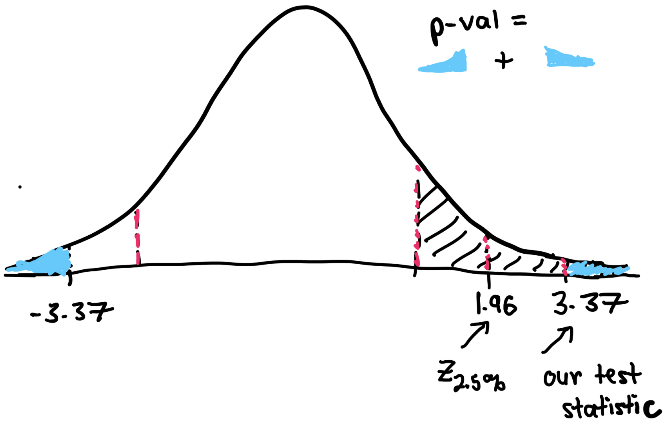

Figure 2: The p value in the kiss example for the first test, with $\hat{\sigma} = \bar{X}_n$

1.3 The general mechanics of p-values

Suppose $T_n = T_n(X_1, \ldots, X_n)$ is our test-statistic and T_n^{obs} is our given observed value of T_n. For example, in the first version of the hypothesis test for the kiss example, we had

$$T_n = \frac{\sqrt{n}}{\bar{X}_n(1 - \bar{X}_n)}|\bar{X}_n - 1/2|, \quad T_n^{\text{obs}} = \frac{\sqrt{124}}{\sqrt{0.645(1 - 0.645)}}|0.645 - 0.5| = 3.37.$$

The p-value is given as follows, for the following three types of rejection regions:

$$\mathcal{R} = \{T_n > c_\alpha\} \implies \text{p-value} = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T_n > T_n^{\text{obs}})$$

$$\mathcal{R} = \{T_n < c_\alpha\} \implies \text{p-value} = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T_n < T_n^{\text{obs}})$$

$$\mathcal{R} = \{|T_n| > c_\alpha\} \implies \text{p-value} = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(|T_n| > T_n^{\text{obs}})$$

In other words, the p-value is just the size of the test we get when we take our cutoff to be $c_\alpha = T_n^{\text{obs}}$. (Recall from Lecture 14 that the size of a test is the largest Type I error, i.e. the largest probability to reject H_0 over all $\theta \in \Theta_0$.)

In practice, often (but not always!) T_n is a standardized test statistic which has distribution $\mathcal{N}(0, 1)$ under $\mathbb{P}_\theta$, where $\theta \in \Theta_0$ is the boundary point maximizing the Type I error. In this case we have e.g. for the first kind of rejection region,

$$\text{p-value} = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T_n > T_n^{\text{obs}}) = \mathbb{P}(\mathcal{N}(0, 1) > T_n^{\text{obs}}).$$