Lecture 18 – t-test and p-values

Last Modified: November 10, 2026

Outline:

18.1 Testing a normal mean with unknown variance 1
18.2 The t-test is the unknown-variance LRT 2
18.3 p-values 4

18.1 Testing a normal mean with unknown variance

In Lecture 17, we worked out the level-α LRT for testing $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ in the normal model $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with known σ^2. The test does the following:

$$\text{Reject when } \left| \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2}. \tag{18.1}$$

This is more commonly known as the two-sided *z-test*. Note that the cutoff $z_{\alpha/2}$ comes from the fact that under the null hypothesis,

$$\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1),$$

so

$$\mathbb{P}_{\mu_0,\sigma^2}\left(\left| \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2} \right) = \mathbb{P}(|Z| \geq z_{\alpha/2}) = \alpha. \tag{18.2}$$

We then discussed what to do in the more realistic scenario in which the population variance σ^2 is *unknown*. By analogy with (18.1), we consider the test

$$\text{Reject when } \left| \frac{\bar{X}_n - \mu_0}{\hat{\sigma}/\sqrt{n}} \right| > c_\alpha. \tag{18.3}$$

We need to choose the cutoff c_α to make this have the desired level α, which requires solving

$$\text{Problem: for what } c_\alpha \text{ is } \sup_{\sigma^2>0} \mathbb{P}_{\mu_0,\sigma^2}\left(\left| \frac{\bar{X}_n - \mu_0}{\hat{\sigma}/\sqrt{n}} \right| > c_\alpha \right) = \alpha \text{ ?} \tag{18.4}$$

To solve this problem we introduced the t distribution:

Definition 1 (Student's t-distribution)**.** The Student's t-distribution t_k, with k degrees of freedom, is the distribution with pdf

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}, \qquad x \in \mathbb{R}.$$

Alternatively, t_k is the random variable constructed as follows: if $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_k$ are independent, then

$$\frac{Z}{\sqrt{V/k}} \sim t_k,$$

We write $t_{k,\alpha}$ for the upper-tail quantile of t_k, so that $\mathbb{P}(t_k \geq t_{k,\alpha}) = \alpha$.

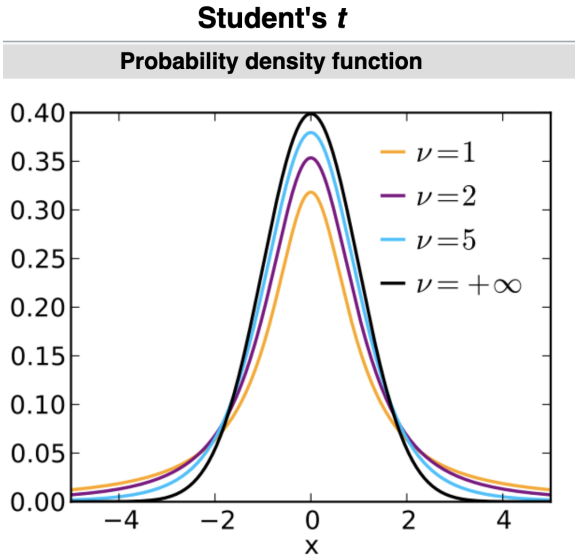


Figure 18.1: The pdf of the student's t distribution with ν degrees of freedom.

See Figure 18.1 for a visualization of the pdf (with degrees of freedom k denoted by ν). The Student's t has heavier tails than the Gaussian distribution, but as k increases, the pdf of Student's t approaches the standard Gaussian pdf.

Proposition 1. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu_0, \sigma^2)$ for any σ, and let $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. Then

$$\frac{\bar{X}_n - \mu_0}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}.$$

This let's us solve the problem posed in (18.4):

$$\alpha = \sup_{\sigma^2>0} \mathbb{P}_{\mu_0,\sigma^2}\left(\left|\frac{\bar{X}_n - \mu_0}{\hat{\sigma}/\sqrt{n}}\right| > c_\alpha\right) = \mathbb{P}(|t_{n-1}| > c_\alpha) \quad \Longrightarrow \quad c_\alpha = t_{n-1,\alpha/2}.$$

This leads us to the t-test:

Definition 2 (Student's t-test)**.** Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where both μ and σ are unknown to us. We want to test $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$. Then the Student's t-test at level α is the test

$$\text{Reject when } \frac{|\bar{X}_n - \mu_0|}{\hat{\sigma}/\sqrt{n}} \geq t_{n-1,\alpha/2},$$

where $t_{n-1,\alpha/2}$ is the $\alpha/2$ quantile of the t_{n-1} distribution, and $\hat{\sigma}^2$ uses the $(n-1)$-normalization.

18.2 The t-test is the unknown-variance LRT

Let us show the t-test, which we came to through an ad-hoc argument, is precisely the LRT! Recall: $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and σ^2 is unknown, so the hypotheses $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ are actually hypotheses on the full parameter $\theta = (\mu, \sigma^2)$. Thus

$$H_0 : \theta \in \Theta_0 = \{(\mu, \sigma^2) : \mu = \mu_0,\ \sigma^2 > 0\} \quad \text{vs.} \quad H_1 : \theta \in \Theta_1 = \{(\mu, \sigma^2) : \mu \neq \mu_0,\ \sigma^2 > 0\}.$$

This leads to the likelihood ratio statistic

$$\Lambda(\underline{x}) = \frac{\sup_{\sigma^2>0} L_{\underline{x}}(\mu_0, \sigma^2)}{\sup_{\mu\in\mathbb{R},\, \sigma^2>0} L_{\underline{x}}(\mu, \sigma^2)} = \frac{L_{\underline{x}}(\mu_0, \hat{\sigma}^2_{\mathrm{MLE}}(\mu_0))}{L_{\underline{x}}(\hat{\mu}_{\mathrm{MLE}}, \hat{\sigma}^2_{\mathrm{MLE}})}, \tag{18.5}$$

where

$$\hat{\sigma}^2_{\mathrm{MLE}}(\mu_0) = \operatorname*{arg\,max}_{\sigma^2>0} L_{\underline{x}}(\mu_0, \sigma^2), \tag{18.6}$$

$$(\hat{\mu}_{\mathrm{MLE}}, \hat{\sigma}^2_{\mathrm{MLE}}) = \operatorname*{arg\,max}_{\mu\in\mathbb{R},\, \sigma^2>0} L_{\underline{x}}(\mu, \sigma^2). \tag{18.7}$$

I expect you to know the below material, but we will not cover it in lecture.

To find the maximizer in (18.6), write $Q(\mu_0) = \sum_{i=1}^n (x_i - \mu_0)^2$. The likelihood and log likelihood are

$$\begin{aligned} L_{\underline{x}}(\mu_0, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{Q(\mu_0)}{2\sigma^2}\right), \\ \ell_{\underline{x}}(\mu_0, \sigma^2) &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{Q(\mu_0)}{2\sigma^2}. \end{aligned} \tag{18.8}$$

Differentiating with respect to σ^2 gives

$$\frac{\partial}{\partial\sigma^2}\ell_{\underline{x}}(\mu_0, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{Q(\mu_0)}{2(\sigma^2)^2} = \frac{Q(\mu_0) - n\sigma^2}{2(\sigma^2)^2}.$$

Since $Q(\mu_0) > 0$, this derivative is positive for $\sigma^2 < Q(\mu_0)/n$ and negative for $\sigma^2 > Q(\mu_0)/n$. Thus the unique maximizer in (18.6) is

$$\hat{\sigma}^2_{\mathrm{MLE}}(\mu_0) = \frac{Q(\mu_0)}{n} = \frac{1}{n}\sum_{i=1}^n (x_i - \mu_0)^2. \tag{18.9}$$

To work out the unconstrained MLE in (18.7), first define $\hat{\sigma}^2_{\mathrm{MLE}}(\mu)$ as in (18.9) except with μ instead of μ_0: this is just the maximizer of the likelihood with respect to σ^2 for a fixed μ. Next, recall that the MLE for μ when σ^2 was considered known *did not actually depend on* σ^2. It is was always just $\bar{x}$. This means that if we now treat σ^2 as variable/unknown, it still holds

$$L_{\underline{x}}(\bar{x}, \sigma^2) \geq L_{\underline{x}}(\mu, \sigma^2), \qquad \forall\mu,\ \forall\sigma^2.$$

But now if we plug in $\hat{\sigma}^2_{\mathrm{MLE}}(\bar{x})$ for σ^2, we get, by definition,

$$L_{\underline{x}}(\bar{x}, \hat{\sigma}^2_{\mathrm{MLE}}(\bar{x})) \geq L_{\underline{x}}(\bar{x}, \sigma^2) \geq L_{\underline{x}}(\mu, \sigma^2), \qquad \forall\mu,\ \forall\sigma^2.$$

We conclude that the solution to (18.7) is

$$\hat{\mu}_{\mathrm{MLE}} = \bar{x}, \qquad \hat{\sigma}^2_{\mathrm{MLE}} = \hat{\sigma}^2_{\mathrm{MLE}}(\bar{x}) = \frac{Q(\bar{x})}{n} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2.$$

Plugging in the solutions to (18.5) gives

$$\Lambda(\underline{x}) = \frac{L_{\underline{x}}(\mu_0, \hat{\sigma}^2_{\mathrm{MLE}}(\mu_0))}{L_{\underline{x}}(\bar{x}, \hat{\sigma}^2_{\mathrm{MLE}}(\bar{x}))}$$

Using equation (18.8) for the likelihood, note that when we plug in $\hat{\sigma}^2_{\text{MLE}}(\mu_0)$, the exponent simplifies to $-Q(\mu_0)/(2\hat{\sigma}^2_{\text{MLE}}(\mu_0)) = -n/2$. Therefore,

$$L_{\underline{x}}(\mu_0, \hat{\sigma}^2_{\text{MLE}}(\mu_0)) = (2\pi\hat{\sigma}^2_{\text{MLE}}(\mu_0))^{-n/2}e^{-n/2}.$$

Similarly,

$$L_{\underline{x}}(\bar{x}, \hat{\sigma}^2_{\text{MLE}}(\bar{x})) = (2\pi\hat{\sigma}^2_{\text{MLE}}(\bar{x}))^{-n/2}e^{-n/2}.$$

The likelihood ratio is therefore

$$\Lambda(\underline{x}) = \frac{L_{\underline{x}}(\mu_0, \hat{\sigma}^2_{\text{MLE}}(\mu_0))}{L_{\underline{x}}(\bar{x}, \hat{\sigma}^2_{\text{MLE}}(\bar{x}))} = \left(\frac{\hat{\sigma}^2_{\text{MLE}}(\mu_0)}{\hat{\sigma}^2_{\text{MLE}}}\right)^{-n/2} = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2/\hat{\sigma}^2_{\text{MLE}}\right)^{-n/2}.$$

Using the sum-of-squares identity $\sum_{i=1}^n (x_i - \mu_0)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2$ derived in Lec 17, we obtain $\frac{1}{n}\sum_{i=1}^n (x_i - \mu_0)^2 = \hat{\sigma}^2_{\text{MLE}} + (\bar{x} - \mu_0)^2$, so

$$\Lambda(\underline{x}) = \left(1 + \frac{(\bar{x} - \mu_0)^2}{\hat{\sigma}^2_{\text{MLE}}}\right)^{-n/2}.$$

Thus rejecting when $\Lambda(\underline{x})$ is small is equivalent to rejecting when $|\bar{x} - \mu_0|/\hat{\sigma}_{\text{MLE}}$ is large. Since $\hat{\sigma}_{\text{MLE}}$ is just some constant multiple of $\hat{\sigma}/\sqrt{n}$, where $\hat{\sigma}$ is as in the t-test, rejecting when $|\bar{x} - \mu_0|/\hat{\sigma}_{\text{MLE}}$ is large is the same as rejecting when $|\bar{x} - \mu_0|/(\hat{\sigma}/\sqrt{n})$ is large. To make this a size-α test, we should use precisely the same cutoff as discussed above with the t-test.

18.3 p-values

p-values answer the following question: how close was the call between H_0 and H_1? To understand what this means, let's return to the ER waiting time example: a patient alleges that Duke Hospital falsely advertised its ER waiting times as being under half an hour, so the patient is suing. According to the innocent until proven guilty standard in society, we formulate the null and alternative as

$$H_0 : \mu \leq 30 \quad \text{vs.} \quad H_1 : \mu > 30.$$

We decide we can tolerate at most a 5% Type 1 error level: the probability of falsely rejecting the null, when Duke Hospital is actually innocent, should be no greater than 5%. Using that $z_{5\%} = 1.64$, the level-5% test is

$$\text{Reject when } \frac{\bar{X}_n - 30}{\sigma/\sqrt{n}} > z_{5\%} = 1.64.$$

Now, suppose we observed an average waiting time of $\bar{X}_n = 33.4$ minutes for a sample of $n = 164$ patients in the ER, and suppose $\sigma = 12$. Then

$$\frac{\bar{X}_n - 30}{\sigma/\sqrt{n}} = \frac{33.4 - 30}{12/\sqrt{164}} = 3.62.$$

Since $3.62 > z_{5\%} = 1.64$, the level 5% test would indeed reject H_0. But the call between H_0 and H_1 was not even close! Why? Because a test of level 2.5%, for example, would also reject H_0 because $3.62 > z_{2.5\%} = 1.96$. More generally, for any α such that $3.62 \geq z_\alpha$, we know the test of level α would reject H_0. This smallest level of α is precisely what we call the p-value:

$$\text{p-value } = \mathbb{P}(Z \geq 3.62) \approx 0.0002 = 0.02\%.$$

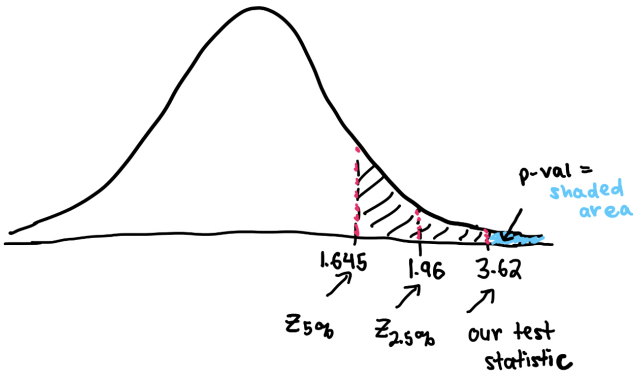


Figure 18.2: The p value in the ER example.

See Figure 18.2. This means that, even if we set up the test to tolerate at most a 0.02% probability of incorrectly concluding Duke Hospital is guilty of false advertising, we would still conclude Duke is guilty based on our evidence (the data we collected). The p-value 0.02% gives us much more information than only the fact that we reject the null at level 5%.

We can interpret the p-value as strength of evidence against H_0. The smaller p is, the more evidence we have. The following evidence scale is typically used, based on the value of p:

$$\begin{aligned} < 1\% &\quad \text{very strong evidence against } H_0 \\ 1\% - 5\% &\quad \text{strong evidence against } H_0 \\ 5\% - 10\% &\quad \text{weak evidence against } H_0 \\ > 10\% &\quad \text{little or no evidence against } H_0 \end{aligned}$$

So in the above ER example, we have very strong evidence against H_0.

Definition 3 (General p-value definition)**.** The p-value is the smallest α for which we would reject the null at level α. Suppose we reject based on the value of a test statistic $T(\underline{x})$. Then the p-value based on an observed T^{obs} is the size of the test which uses T^{obs} as the cutoff. Explicitly, the p-value is given as follows, for the following three types of rejection regions:

$$\mathcal{R} = \{\underline{x} : T(\underline{x}) > c_\alpha\} \implies p = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T > T^{\text{obs}})$$

$$\mathcal{R} = \{\underline{x} : T(\underline{x}) < c_\alpha\} \implies p = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T < T^{\text{obs}})$$

$$\mathcal{R} = \{\underline{x} : |T(\underline{x})| > c_\alpha\} \implies p = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(|T| > |T^{\text{obs}}|)$$

For the hypotheses we have seen, the largest type I probability occurs at the boundary point θ_0 in Θ_0, which is closest to the alternative Θ_1. In such cases, $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta = \mathbb{P}_{\theta_0}$. E.g. in the ER example above,

$$\sup_{\mu \leq 30} \mathbb{P}_\mu(T \geq T^{\text{obs}}) = \mathbb{P}_{\mu=30}(T \geq T^{\text{obs}}) = \mathbb{P}(\mathcal{N}(0,1) > T^{\text{obs}}),$$

since $T \sim \mathcal{N}(0,1)$ under the boundary null value $\mu = 30$.

Remark 1 (From p-value to rejection decision)**.** Suppose we compute a p-value and now want to actually decide whether or not to reject the null at some specified level α. We can easily do this, using the definition of p-value: if $p \leq \alpha$, we reject the null at level α. if $p > \alpha$, we do not reject the null.

Example 1 (Two-sided z-test)**.** A technician checks the calibration of a laboratory scale by weighing a reference mass which is supposed to weigh 50 grams, $n = 9$ times. Suppose the readings are independent, with $X_1, \dots, X_9 \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, and the standard deviation is known to be $\sigma = 1.5$ grams. A correctly calibrated scale has $\mu = 50$, so we test

$$H_0 : \mu = 50 \quad \text{vs.} \quad H_1 : \mu \neq 50.$$

We use the test which rejects when $|T|$ is large, where

$$T = \frac{\bar{X}_9 - 50}{1.5/\sqrt{9}}.$$

Suppose the observed average reading is $\bar{x}_9 = 51$ grams. Then

$$T^{\text{obs}} = \frac{51 - 50}{1.5/\sqrt{9}} = 2.$$

Since $T \sim \mathcal{N}(0, 1)$ under the null, the p-value is

$$p = \mathbb{P}_{\mu=50}(|T| > T^{\text{obs}}) = \mathbb{P}(|Z| \geq 2) = 2(1 - \Phi(2)) \approx 4.6\%.$$

See Figure 18.3.

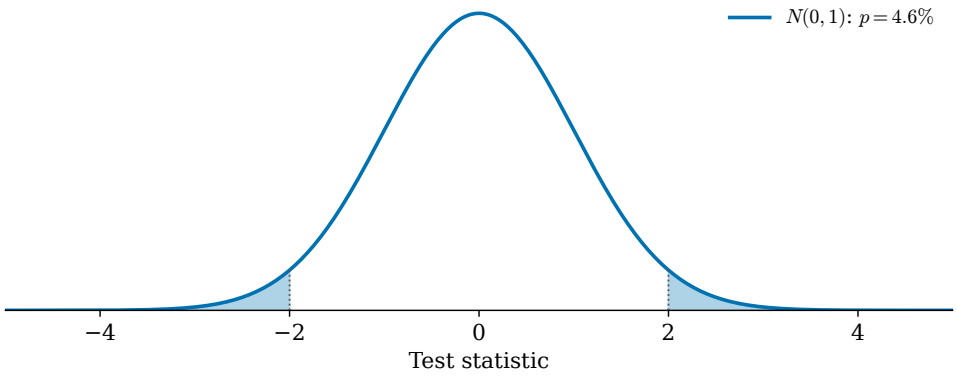


Figure 18.3: The two-sided z-test p-value: the shaded tails beyond ± 2 have total area $p \approx 4.6\%$.

Example 2 (Two-sided t-test)**.** Consider the exact same weight-calibration problem, but now suppose the population variance σ^2 is unknown. We again test $H_0 : \mu = 50$ vs $H_1 : \mu \neq 50$. Suppose the observed sample mean is $\bar{x}_9 = 51$ grams (as before), and the observed sample standard deviation is $\hat{\sigma} = 1.5$ grams, where $\hat{\sigma}$ uses the $n - 1$ normalization:

$$\hat{\sigma}^2 = \frac{1}{8} \sum_{i=1}^{9} (X_i - \bar{X}_9)^2.$$

The test statistic and its observed value are

$$T = \frac{\bar{X}_9 - 50}{\hat{\sigma}/\sqrt{9}}, \qquad T^{\text{obs}} = \frac{51 - 50}{1.5/\sqrt{9}} = 2.$$

Under H_0, we know that $T \sim t_8$, regardless of what the true σ is (recall Proposition 1). Thus the p-value is

$$p = \sup_{\sigma^2 > 0} \mathbb{P}_{\mu=50, \sigma^2>0}(|T| > T^{\text{obs}}) = \mathbb{P}(|t_8| \geq 2) = 2(1 - F_{t_8}(2)) \approx 8.1\%,$$

where F_{t_8} is the cdf of the t_8 distribution.

The observed statistic is 2 *in both examples, but the t-test has a larger p-value!* This is because the t_8 distribution has heavier tails than the standard normal distribution. Estimating the population standard deviation introduces additional uncertainty. In particular, at level 5%, the z-test rejects the null while the t-test does not. See Figure 18.4 for a comparison of the two p-values.

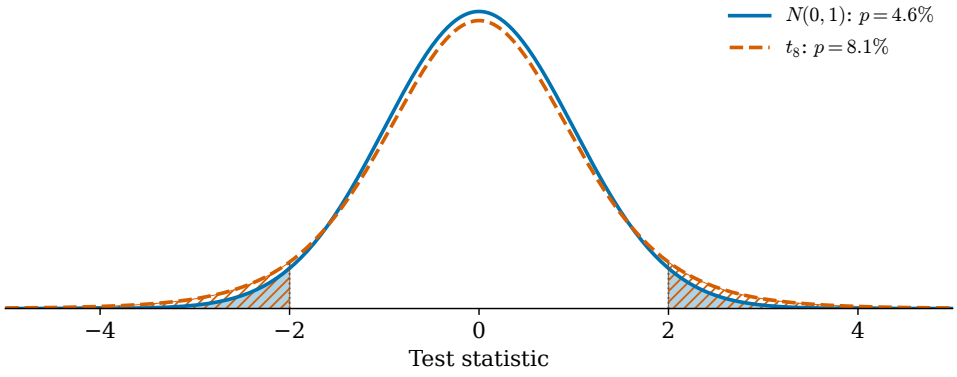


Figure 18.4: The standard normal and t_8 densities on the same axes. Beyond ± 2, the blue shaded tails have total area $p_z \approx 4.6\%$, while the orange hatched tails have total area $p_t \approx 8.1\%$.