t-tests and p-values

Rebecca Willett, 2016

1 Student's t distribution and t-tests

Consider the following hypothesis testing problem:

$$H_0 : x_i \overset{iid}{\sim} \mathcal{N}(\mu_0, \sigma^2),\ i = 1, \ldots, n$$
$$H_1 : x_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2),\ i = 1, \ldots, n,\ \mu > \mu_0 \text{ but otherwise unknown}$$

We have discussed how to handle this test when σ^2 is known. But how should we proceed if it is unknown?

One option is the GLRT, discussed above. However, (a) we must estimate μ and (b) Wilk's theorem only tells us that the test statistic corresponding to maximum likelihood estimates of σ^2 and μ is *asymptotically* chi-squared. For small n, then, it can be difficult to set a threshold to achieve a desired probability of false positives or type I error.

As alternative is the celebrated t-test. Specifically, let

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$$

and note that under H_0, $\overline{x} \sim \mathcal{N}(\mu_0, \sigma^2/n)$. So if we knew σ^2, we could compute the statistic $\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ and set a threshold as discussed in previous units. Since we do not know σ^2, we can estimate it from our data; specifically, let $s_n := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ be the sample standard deviation. Then $s/\sqrt{n}$ is called the standard error of the mean and is an estimate of $\sigma/\sqrt{n}$. This leads us to the t-statistic:

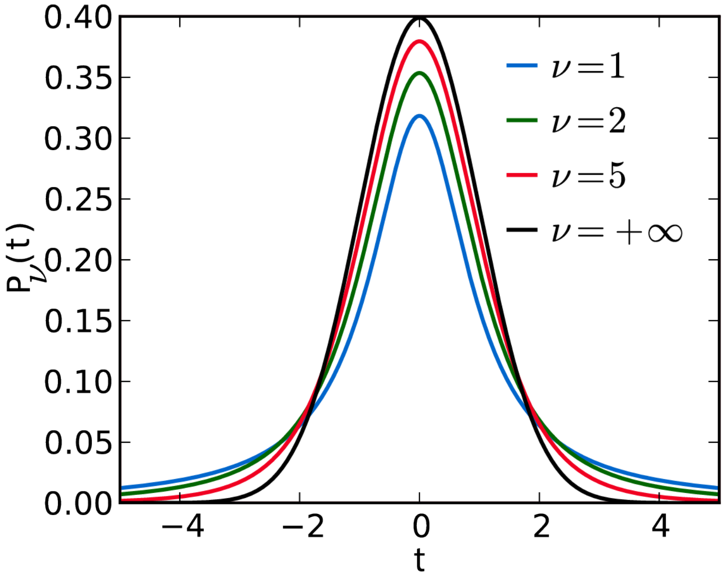
$$t^* = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}.$$

Ultimately we will perform our hypothesis test by thresholding t^*, and to set a threshold guaranteed to yield a certain probability of false positives or type I error we must undertand the distribution of t^*.

In 1908, Guinness statistician William Gosset published a paper characterizing this distribution under the pseudonym "Student", and subsequently the distribution has been dubbed Student's t-distribution. It is parameterized by ν, the number of degrees of freedom in the distribution, and takes the form

$$p_\nu(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}(1 + \frac{t^2}{\nu})^{-\frac{\nu+1}{2}}.$$

The test statistic t^* above is drawn from the t-distribution with $\nu = n - 1$ degrees of freedom.



As $\nu \longrightarrow \infty$, $p_\nu(t) \longrightarrow \mathcal{N}(0,1)$. For smaller ν corresponding to smaller sample sizes, though, the t-distribution has heavier tails, and its tail probabilities can be used to determine appropriate thresholds for t-statistics.

1.1 Two-sample t-tests

In some settings we observe two different sets of data, data $x_1, \ldots, x_{n_x}$ and $y_1, \ldots, y_{n_y}$ and which to perform a test, say to see if they are drawn from distribtuions with the same mean. For instance,

$$\begin{aligned} H_0 : & x_i \overset{iid}{\sim} \mathcal{N}(\mu_0, \sigma_x^2), i = 1, \ldots, n_x \\ & y_i \overset{iid}{\sim} \mathcal{N}(\mu_0, \sigma_y^2), i = 1, \ldots, n_y. \end{aligned}$$

A common approach is to consider a test statistic that is a function of $\overline{x} - \overline{y}$, as under the null hypothesis this difference will have zero mean. We will construct and threshold a t-statistic. This is called a two-sample t-test.

How should we compute a t-statistic in such a case? Generally we use the formula

$$t^* = \frac{\overline{x} - \overline{y}}{\text{s.e.}}$$

where s.e. is the standard error of the mean, as before. How should this standard error be computed? There are two possibilities:

1. We assume the two distributions have equal variance ($\sigma := \sigma_x = \sigma_y$). In this case, $\overline{x} - \overline{y} \sim \mathcal{N}(0, \sigma^2/n_x + \sigma^2/n_y)$, and we estimate σ^2 via

$$s^2 = \frac{\sum_{i=1}^{n_x}(x_i - \overline{x})^2 + \sum_{i=1}^{n_y}(y_i - \overline{y})^2}{n_x + n_y - 2}$$

and then s.e. $= \sqrt{s^2(1/n_x + 1/n_y)}$. The resulting t-statistic has $\nu = n_x + n_y - 2$ degrees of freedom.

2. We do NOT assume the two distributions have equal variance. In this case, $\overline{x} - \overline{y} \sim \mathcal{N}(0, \sigma_x^2/n_x + \sigma_y^2/n_y)$. We estimate σ_x^2 via

$$s_x^2 = \frac{1}{n_x - 1} \sum_{i=1}^{n_x} (x_i - \overline{x})^2$$

and similarly for σ_y^2. Then the standard error is s.e. $= \sqrt{s_x^2/n_x + s_y^2/n_y}$. The distribution of the resulting statistic is *approximately* a t-distribution with

$$\nu = \frac{(s_x^2/n_x + s_y^2/n_y)^2}{(s_x^2/n_x)^2/(n_x - 1) + (s_y^2/n_y)^2/(n_y - 1)}$$

degrees of freedom. (This is known as the Welch-Satterthwaite equation.)

2 p-values

So far we have considered making decisions or performing hypothesis testing by computing a test statistic and thresholding it. Our aim is the answer the key question

Note: Does our data provide enough evidence for us to reject the null hypothesis H_0?

We saw that we can choose a threshold to minimize the probability of error or probability of false positives or other measures of error. However, the result of such a test is always a binary decision (H_0 or H_1) and not a measure of how strong our evidence is again H_0. p-values bridge this gap.

Specifically, for a given test statistic t^*, we could perform the test

$$t^* \underset{H_0}{\overset{H_1}{\gtrless}} \tau_\alpha$$

where τ_α is a threshold associated with a type I error or false positive rate of α (the value of τ_α depends on the distribution of t^* under the null hypothesis). One can easily imagine that there is a range of values of α which would all lead us to reject H_0. The p-value is essentially the smallest α (corresponding to the largest threshold τ_α) for which we would reject H_0 with our test statistic. More formally

Definition: p-value

The p-value is the smallest level at which we can reject H_0:

$$\text{p-value} = \inf\{\alpha : t^* > \tau_\alpha\}.$$

More generally, if R_α is the rejection region associated with a test at level α, then

$$\text{p-value} = \inf\{\alpha : t^* \in R_\alpha\}.$$

Note: Notes on the p-value

- it measures the strength of the evidence against H_0: a small p-value (e.g., below 0.05, ideally below 0.01) indicates strong evidence against H_0.
- a large p-value is NOT evidence in favor of H_1 (it's possible we just have a low-power test)
- the p-value should NOT be thought of as $\mathbb{P}(H_0|\text{data})$.

Theorem: Computation of the p-value

Let p_0 denote the distribution of the test statistic under H_0. If we have a test of the form *reject H_0 if and only if $t^* \geq \tau_\alpha$*, then

$$\text{p-value} = \mathbb{P}(T \geq t^* | T \sim p_0).$$

In other words, the p-value is the probability under H_0 of observing a test statistic at least as extreme as what was observed.

Distribution of the p-value

If the test statistic has a continuous distribution, then under H_0 the p-value is uniformly distributed between 0 and 1. Thus if we reject H_0 whenever a p-value is less than α, that test as a type I error or probability of false positives of α.

Example: GPA distributions

We sample $n = 15$ students and look at their GPAs. The sample mean GPA among these students was $\overline{x} = 3.15$, and the sample standard deviation was

$$s = \sqrt{\frac{1}{14}\sum_{i=1}^{n}(x_i - \overline{x})^2} = 0.3.$$

We want to test whether the mean GPA is $\mu_0 = 3$ or $\mu > 3$; that is

$$H_0 : x_i \overset{iid}{\sim} \mathcal{N}(\mu_0, \sigma^2),\ i = 1, \ldots, n$$
$$H_1 : x_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2),\ i = 1, \ldots, n,\ \mu > \mu_0 \text{ but otherwise unknown.}$$

We can compute a t-statistic of $t^* = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = 1.94$, which follows a t-distribution with $\nu = n - 1 = 14$ degrees of freedom.

What is the p-value for this statistic? We must compute

$$\text{p-value} = \mathbb{P}(T \geq t^*|T \sim p_{14}(t)) = 1 - \mathbb{P}(T < t^*|T \sim p_{14}(t)); \tag{1}$$

the last expression can be computed by evaluating the CDF of the t-distribution at t^* (e.g. using tcdf in matlab), yielding a p-value of 0.037 – thus we have strong (though not very strong) evidence for rejecting the null hypothesis that the mean GPA is 3.