In the previous section, we introduced the t-test, t-statistic, and the t-distribution. The setup is that we consider samples $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$ for some mean μ and variance σ^2. Define the test statistic

$$T_n := \frac{\overline{X_n} - \mu}{\sqrt{\hat{\sigma}^2/n}},$$

where we have

$$\overline{X_n} := \frac{1}{n}\sum_{i=1}^{n} X_i,$$

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

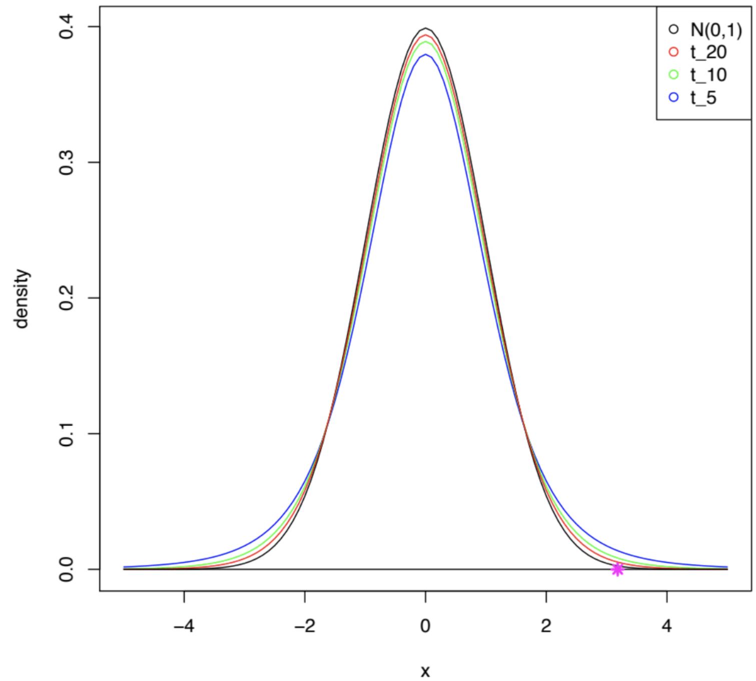
The main result is that the T_n has a t-distribution with $n-1$ degrees of freedom. We discuss this result further.

t distribution

We start by defining the t distribution and its parameter k which specifies the number of **degrees of freedom**. The t distribution with n degrees of freedom is defined as the distribution of $\frac{Y}{\sqrt{Z/n}}$, where

- $Y \sim \mathcal{N}(0, 1)$ is a standard *normal* distribution
- $Z \sim \chi_n^2$ is a *chi-squared* distribution with n degrees of freedom
- Y and Z are *independent*.

As n increases, the distribution has thinner tails; more precisely, the variance of the t_n distribution is $\frac{n}{n-2}$. The t distribution for different values of n are plotted in the figure below.



Intuitively, we can see a rough correspondence from the definition of the t-statistic.

- The sample mean in the numerator of the t statistic is normally distributed, just as the Y in the numerator of the t distribution is.
- The sample variance in the denominator of the t statistic is a sum of squares, which is similar to how the chi-squared distribution in the denominator of the t distribution is defined.

Next, we provide a formal proof that T indeed follows a t distribution with $n-1$ degrees of freedom.

Proof that the t statistic follows a t distribution

To prove that the t statistic follows a t distribution, we specify Y and Z such that $T = \frac{Y}{\sqrt{Z/n}}$ and so that the three conditions for Y and Z given above are satisfied.

We first construct Y, which must have a $\mathcal{N}(0,1)$ distribution. We already know that the z-statistic $\dfrac{\overline{X_n}-\mu}{\sigma/\sqrt{n}}$ has a standard normal distribution, so we can let $Y=\dfrac{\overline{X_n}-\mu}{\sigma/\sqrt{n}}$. Then, we can solve for Z by equating the expressions for the t-statistic and the t_{n-1} distribution:

$$T=\frac{Y}{\sqrt{Z/(n-1)}}=\frac{\overline{X_n}-\mu}{\sqrt{\hat{\sigma}^2/n}}.$$

Hence, we derive the corresponding Z as:

$$\sqrt{\frac{Z}{n-1}}=\frac{Y\sqrt{\hat{\sigma}^2/n}}{\overline{X_n}-\mu}=\frac{\sqrt{\hat{\sigma}^2/n}}{\sigma/\sqrt{n}}\implies Z=(n-1)\frac{\hat{\sigma}^2}{\sigma^2}=\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i-\overline{X_n}\right)^2.$$

Note that Y only depends on $\overline{X_n}$. Hence, it suffices to show that

- $\dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}\left(X_i-\overline{X_n}\right)^2$ has a χ^2_{n-1} distribution
- $\overline{X_n}$ and $\displaystyle\sum_{i=1}^{n}\left(X_i-\overline{X_n}\right)^2$ are independent.

A popular approach to show both at the same time is to consider a related quantity which has distribution χ^2_n, as $\frac{X_i-\mu}{\sigma}$ has a $\mathcal{N}(0,1)$ distribution:

$$W:=\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\sim\chi^2_n.$$

By some algebra manipulation (left as an exercise to the reader), we can write

$$W := \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X_n})^2 + \frac{n}{\sigma^2} (\overline{X_n} - \mu)^2.$$

We now reason using multivariate Gaussians, as $X_1, \ldots, X_n$ are i.i.d. Gaussians. Therefore, $\overline{X_n} \sim \mathcal{N}(\mu, \sigma^2/n)$, so $\frac{n}{\sigma^2}(\overline{X_n} - \mu)^2 \sim \chi_1^2$. More generally, we can construct variables out of linear combinations of $X_1, \ldots, X_n$. If we have a pair of such variables, they will be *jointly Gaussian* so they are independent iff they have zero covariance.

We apply this technique to show that $X_i - \overline{X_n}$ and $\overline{X_n}$ are independent. Indeed,

$$\text{Cov}\,(X_i, \overline{X_n}) = \text{Cov}\,(X_i, \frac{1}{n} X_i) = \frac{1}{n}\sigma^2,$$

and

$$\text{Cov}\,(\overline{X_n}, \overline{X_n}) = \sum_{i=1}^{n} \text{Cov}\,(\frac{1}{n} X_i, \frac{1}{n} X_i) = n \left(\frac{1}{n^2} \sigma^2 \right) = \frac{1}{n}\sigma^2.$$

Hence, we get that $\text{Cov}\,(X_i - \overline{X_n}, \overline{X_n}) = 0$, and so $X_i - \overline{X_n}$ and $\overline{X_n}$ are independent.

Using the above fact for $i = 1, \ldots, n$, this proves the claim that $\overline{X_n}$ and $\sum_{i=1}^{n} (X_i - \overline{X_n})^2$ are independent. Hence, the two components of W are also independent.

As the latter component has a χ_1^2 distribution, the former must have a χ_{n-1}^2 distribution. This is based on the additivity property of a χ^2 distribution: the sum of a χ_1^2 and χ_{n-1}^2 distribution, the two independent from each other, is a χ_n^2 distribution.

The uniqueness of this distribution can be shown by considering the uniqueness of the moment generating function. Indeed, write $W = W_1 + W_2$, where W_1 and W_2 are independent. Knowing that W and W_2, as well as that they are independent, we can divide the mgf's of W and W_2 to get the mgf of W_1 from which there is a unique corresponding distribution.

Discussion

Hide Discussion

Topic: Module 1. Review: Statistics, Correlation, Regression, Gradient Descent:Hypothesis Testing / 8. t-statistic and the t-distribution

Add a Post

Show all posts by recent activity

- Y and Z are independent? — 3
  It is kind of surprising since Y and Z is calculated based on the same data set. Can someone give a...
- [STAFF] Typo in t-distribution proof — 5
- Does n need to be large for the test statistic to follow the distribution? — 3
  It seems that the proof assumes that n is large; for instance when we say the z-statistics follows a ...