LECTURE: SAMPLING DISTRIBUTIONS

1. Chi-Square distribution (continued)

Recall:

Let $Y_1, \dots, Y_n$ be iid $N(\mu, \sigma^2)$ then

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma} \right)^2$$

Has a **chi-square distribution** with n degrees of freedom (df)

Recall:

The **sample variance** is

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Fact:

Let $Y_1, \dots, Y_n$ be iid $N(\mu, \sigma^2)$ then

$$\left(\frac{n-1}{\sigma^2} \right) S^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Has a chi-square distribution with $n - 1$ deg of freedom

Note: The $n-1$ is not a typo. This is because $\overline{Y}$ depends on $Y_1, \cdots, Y_n$ so the sum above really just depends on $n-1$ independent random variables and not n

Example 1:

A ball bearing machine produces ball bearings whose diameters are normally distributed with mean μ mm and standard deviation $\sigma = 0.1$ mm.

Suppose we select $n = 10$ samples and compute the sample variance S^2

Find an interval $[a, b]$ such that

$$P(a \leq S^2 \leq b) = 0.90$$

STEP 1: Let $X = \left(\frac{n-1}{\sigma^2}\right) S^2$

From the above, we know that X has a chi-square distribution with $(n-1) = 10 - 1 = 9$ df. Then

$$\begin{aligned} 0.90 =& P(a \leq S^2 \leq b) \\ =& P\left(\frac{(n-1)a}{\sigma^2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \frac{(n-1)b}{\sigma^2}\right) \\ =& P\left(\frac{(n-1)a}{\sigma^2} \leq X \leq \frac{(n-1)b}{\sigma^2}\right) \end{aligned}$$

Where X has a chi-square distribution with 9 df.

STEP 2: **Splitting the Difference:**

There are many ways we can choose a and b to make this happen, but the easiest way is to "split the difference"

Just like for the normal distribution, one way is to choose a and b so

$$P\left(X \leq \frac{(n-1)a}{\sigma^2}\right) = 0.05 \text{ and } P\left(X \geq \frac{(n-1)b}{\sigma^2}\right) = 0.05$$

This is equivalent to:

$$P\left(X \geq \frac{(n-1)a}{\sigma^2}\right) = 0.95 \text{ and } P\left(X \geq \frac{(n-1)b}{\sigma^2}\right) = 0.05$$

From the chi-squared table with row 9 df, the values are 3.325 and 16.919.

STEP 3: Since $n = 10$ and $\sigma = 0.1$, we can solve for a and b:

$$\frac{(n-1)a}{\sigma^2} = 3.325 \Rightarrow a = \left(\frac{\sigma^2}{n-1}\right)(3.325) = \left(\frac{(0.1)^2}{9}\right)(3.325) = 0.0037$$

$$\frac{(n-1)b}{\sigma^2} = 16.919 \Rightarrow b = \left(\frac{\sigma^2}{n-1}\right)(16.919) = \left(\frac{(0.1)^2}{9}\right)(16.919) = 0.0187$$

STEP 4: Thus our interval $[a, b] = [0.0037, 0.0187]$ which has 90% probability of including our sample variance S^2

Note that the true population variance $\sigma^2 = 0.1^2 = 0.01$ lies in that interval. If our sample variance is not in that interval, perhaps we should be suspicious of either our sampling technique or that something is going wrong with the machine!

2. Student's t-distribution

Question: What happens when we don't know the standard dev. σ?

Recall:

If we know the mean μ and the standard deviation σ then

$$Z = \sqrt{n}\left(\frac{\bar{Y}-\mu}{\sigma}\right) \sim N(0,1)$$

If σ is not known then we can replace σ by our sample standard deviation S to get

$$\sqrt{n}\left(\frac{\bar{Y}-\mu}{S}\right)$$

The probability distribution of this quantity is known as the **Student's t-distribution**

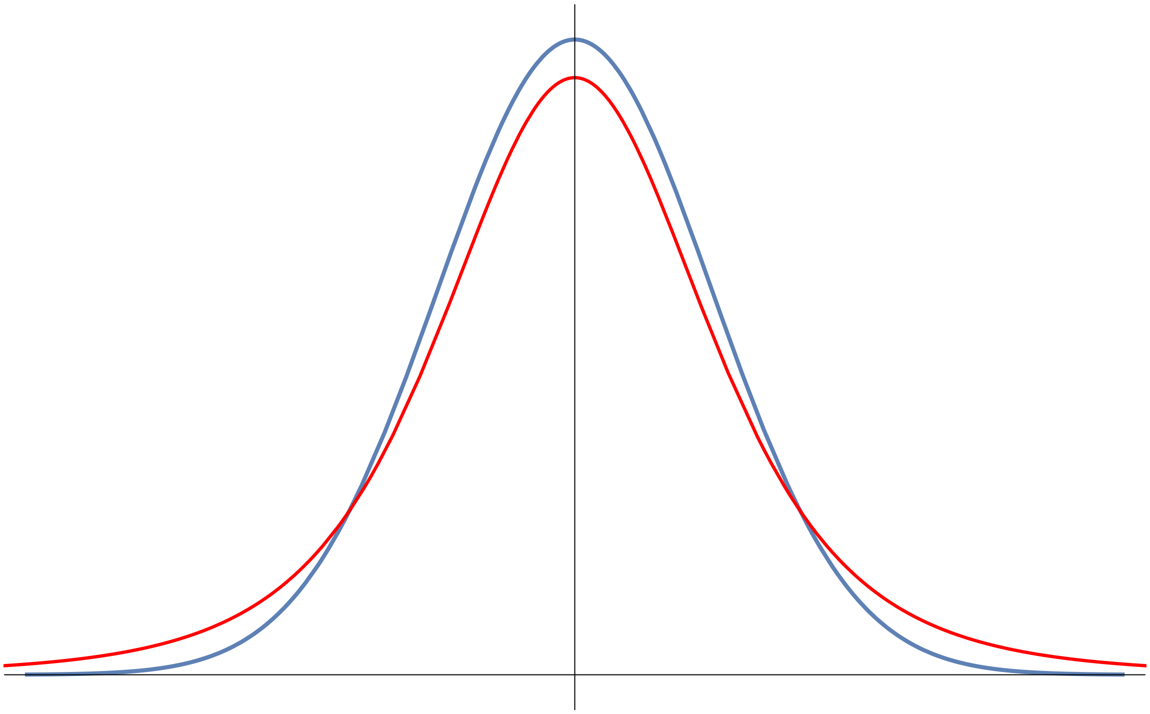
Definition:

Let $Y_1, \dots, Y_n$ be iid $N(\mu, \sigma^2)$ where σ is unknown. Let S^2 be the sample variance as given above. Then

$$T = \sqrt{n}\left(\frac{\bar{Y}-\mu}{S}\right)$$

Has a **Student's t-distribution** with $n-1$ deg of freedom (df).

The t distribution, like the standard normal distribution, is bell-shaped, has a mean of 0, and is symmetric about its mean. However, it has thicker tails, so outlier values are more common. The following plot shows the standard normal distribution in blue and the t distribution with 4 df in red.



Cultural Note: The distribution comes from William Sealy Gosset's 1908 paper under the pseudonym "Student." One version of the origin of the pseudonym is that Gosset's employer preferred staff to use pen names when publishing scientific papers, so he used the name "Student" to hide his identity. Another version is that Guinness, the beer company he was working for, did not want their competitors to know that they were using the t-test to determine the quality of raw material.

Like the Z distribution and chi-square distribution, we use tables to work with the t-distribution (see table on course website)

Example 2:

Once again, we have a bearing machine produces ball bearings whose diameters are normally distributed with mean μ mm and standard deviation σ mm.

We not only have lost the manual for the machine, but the company has gone bankrupt so they cannot tell us anything about the machine!

We take a sample of $n = 16$ ball bearings from the machine and compute the sample mean $\bar{Y}$. We also compute the sample standard deviation S and find that $S = 0.1$.

(a) Find a range $[a, b]$ such that the probability that the difference between $\overline{Y}$ and the population mean μ falls within a and b with a probability of 0.90

STEP 1: We want to find a and b such that $P(a \leq (\bar{Y} - \mu) \leq b) = 0.90$

Multiplying by $\sqrt{n}$ and dividing by S, we get:

$$\begin{aligned} P\left(\frac{a\sqrt{n}}{S} \leq \sqrt{n}\left(\frac{\bar{Y}-\mu}{S}\right) \leq \frac{b\sqrt{n}}{S}\right) &= 0.90 \\ P\left(\frac{a\sqrt{16}}{0.1} \leq T \leq \frac{b\sqrt{16}}{0.1}\right) &= 0.90 \\ P\left(40a \leq T \leq 40b\right) &= 0.90 \end{aligned}$$

STEP 2: Look at the t-table

We have a sample of $n = 16$ so we want $n - 1 = 15$ df

Since the $t-$distribution is symmetric we want

$$P(T \leq 40a) = 0.05 \text{ and } P(T \geq 40b) = 0.05$$

For $P(T \geq 40b)$ we look at the row for 15 df and go across until we get to 0.05 which gives $40b = 1.753 \Rightarrow b = 0.0438$

By symmetry, $40a = -1.753 \Rightarrow a = -0.0438$

STEP 3: Answer

We are 90% confident that the difference between the sample and population means falls within the interval $[-0.0438, 0.0438]$

(b) Same problem but this time you know $\sigma = 0.1$

In that case, since we know σ, convert to the standard normal distribution instead of the $t-$distribution. We do expect the interval to be narrower since we have more precise information

$$\begin{aligned} P\left(\frac{a\sqrt{n}}{\sigma} \leq \sqrt{n}\left(\frac{\bar{Y}-\mu}{\sigma}\right) \leq \frac{b\sqrt{n}}{\sigma}\right) &= 0.90 \\ P\left(\frac{a\sqrt{16}}{0.1} \leq Z \leq \frac{b\sqrt{16}}{0.1}\right) &= 0.90 \\ P\left(40a \leq Z \leq 40b\right) &= 0.90 \end{aligned}$$

Again, we will look for a symmetric interval about the mean, so we want $P(Z \leq 40a) = 0.05$ and $P(Z \geq 40b) = 0.05$

For $P(Z \leq 4a)$ we look at the Z-table and can choose $z = -1.64$ (corresponding to a probability of 0.0505). Choosing $z = -1.64$ we get $40a = -1.64$, so $a = -0.041$. By symmetry, $b = 0.041$.

Answer: We are 90% confident that the difference between the sample and population means falls within the interval $[-0.041, 0.041]$

As predicted, this is a narrower interval than in (a) which was $[-0.0438, 0.0438]$