Worksheet 03: One-sample T-Test

Today we turn to studying the sample variance S_X^2. To start, let's do a little bit of review from probability theory. Assume that $Z \sim N(0,1)$, a standard normal. Then $Z^2 \sim \chi^2(1)$, a chi-squared distribution with one degree of freedom. If we add together k independent chi-squared random variables, we will get a random variable with a chi-squared with k degrees of freedom ($\chi^2(k)$). Equivalently, assume that we have a random sample of n standard normals: $Z_1, \ldots, Z_n \overset{iid}{\sim} N(0,1)$. Then, we have $\sum_i Z_i \sim \chi^2(k)$.

Q01. With this information, what is the distribution of the following quantity from the sample mean $\bar{X}$ from a random sample with n observations?

$$\left[\frac{\bar{X} - \mu_X}{\sigma_X/\sqrt{n}}\right]^2.$$

The random variable inside of the square is a standard normal, so the squared quantity has a $\chi^2(1)$ distribution.

Now, let's consider the following quantity, which we will temporarily give a name of Y (it's not a quantity we need often, so there is not a standard symbol for it):

$$Y = \frac{1}{\sigma_X^2} \times \sum_i [X_i - \mu_X]^2$$

Q02. What is the distribution of Y? Moving the constant inside, we see that:

$$Y = \sum_i \left[\frac{X_i - \mu_X}{\sigma_X}\right]^2$$

Each of the components of the sum have a distribution of $N(0,1)$ and every component is independent. So, $Y \sim \chi^2(n)$.

The quantity Y looks similar to S_X^2. We will use a common trick to get Y in terms of S_X^2: adding and subtracting the quantity $\bar{X}$ inside of the terms inside the sum. We can put the constant factor in later, and so let's start with the following equality:

$$\begin{aligned}\sum_i [X_i - \mu_X]^2 &= \sum_i \left[X_i - \bar{X} + (\bar{X} - \mu_X)\right]^2 \\ &= \sum_i \left[(X_i - \bar{X}) + (\bar{X} - \mu_X)\right]^2\end{aligned}$$

Make sure that you see why this is valid! **Q03.** Starting with the formula above, distribute the square. You should have three different summation terms. Simplify by showing that the cross-term (the one with the 2 in it) is zero and another one of the terms is a constant in terms of the index i. The third term should look similar to S_X^2. This is somewhat tricky. Make sure you check the

anwer before moving on. We start by the straightforward distribution of the squared term:

$$\begin{aligned}\sum_i [X_i - \mu_X]^2 &= \sum_i \left[(X_i - \bar{X}) + (\bar{X} - \mu_X)\right]^2 \\ &= \sum_i \left[(X_i - \bar{X})^2 + (\bar{X} - \mu_X)^2 + 2(X_i - \bar{X}) \cdot (\bar{X} - \mu_X)\right] \\ &= \sum_i (X_i - \bar{X})^2 + \sum_i (\bar{X} - \mu_X)^2 + \sum_i 2(X_i - \bar{X}) \cdot (\bar{X} - \mu_X)\end{aligned}$$

Terms that do not have an i index can come outside of the summation. The middle term is all constant, so we just remove the sum by multiplying by n (the number of terms in the series):

$$\begin{aligned}\sum_i [X_i - \mu_X]^2 &= \sum_i (X_i - \bar{X})^2 + n \cdot (\bar{X} - \mu_X)^2 + 2 \cdot (\bar{X} - \mu_X) \cdot \sum_i (X_i - \bar{X}) \\ &= \sum_i (X_i - \bar{X})^2 + n \cdot (\bar{X} - \mu_X)^2\end{aligned}$$

The last step comes from the fact that $\sum_i (X_i - \bar{X})$ must be zero. If you do not believe that, distribute the summation and work out the details to see why.

Q04. Divide both sides of your previous answer by σ_X^2. You should have one term on the left and two on the right. Make one of the terms on the right look like quantity in question 1.

$$\begin{aligned}\sum_i \left[\frac{X_i - \mu_X}{\sigma_X}\right]^2 &= \sum_i (X_i - \bar{X})^2 + n \cdot (\bar{X} - \mu_X)^2 + 2 \cdot (\bar{X} - \mu_X) \cdot \sum_i (X_i - \bar{X}) \\ &= \frac{1}{\sigma_X^2} \sum_i (X_i - \bar{X})^2 + \frac{n}{\sigma_X^2} \cdot (\bar{X} - \mu_X)^2 \\ &= \frac{1}{\sigma_X^2} \sum_i (X_i - \bar{X})^2 + \left[\frac{\bar{X} - \mu_X}{\sigma_X / \sqrt{n}}\right]^2.\end{aligned}$$

We will take it as a fact that $\bar{X}$ and S_X^2 are independent random variables. It takes a lot of work to show this and I don't think it helps in understanding the result. However, if you want to see the proof just let me know! **Q05.** Using the previous set of results, what is the distribution of the following quantity?

$$\frac{1}{\sigma_X^2} \sum_i (X_i - \bar{X})^2 = \frac{(n-1)S_X^2}{\sigma_X^2}$$

The left-hand side of the previous answer is $\chi^2(n)$ and the right-hand side is the sum of two independent terms: a $\chi^2(1)$ and the value above. Therefore, the term above must be a $\chi^2(n-1)$ (because then there sum would be a $\chi^2(n)$, as required).

From probability theory, we have that the expected value of a random variable with a chi-squared distribution with k degrees of freedom is k. Its variance is $2k$. **Q06.** Take the expected value of

the quantity from the previous question and simplify to get the expected value of S_X^2. We have:

$$\begin{aligned}
\mathbb{E}\left[\frac{(n-1)S_X^2}{\sigma_X^2}\right] &= n-1 \\
\frac{(n-1)}{\sigma_X^2} \cdot \mathbb{E}\left[S_X^2\right] &= n-1 \\
\mathbb{E}\left[S_X^2\right] &= \sigma_X^2
\end{aligned}$$

So the expected value of the sample variance is equal to the population variance. We say that this is an unbiased estimator of the variance, a concept that we will return to in the next unit.

Q07. Take the variance of the quantity you started with in the previous question and simplify to get the variance of S_X^2.

$$\begin{aligned}
\text{Var}\left[\frac{(n-1)S_X^2}{\sigma_X^2}\right] &= 2(n-1) \\
\frac{(n-1)^2}{\sigma_X^4} \cdot \text{Var}\left[S_X^2\right] &= 2(n-1) \\
\text{Var}\left[S_X^2\right] &= \frac{\sigma_X^4}{n-1}
\end{aligned}$$

Let's step back and apply to this our example dataset of potato-diet weight loss. Recall that we had 16 observations with the following sample mean and variance:

$$\bar{x} = 2.28, \quad s_X^2 = 8.7.$$

In order to build a confidence interval for the mean last time, we had to cheat and pretend that we knew the variance of the unknown distribution (this almost never happens). Now, what if we instead replace the population variance with the sample variance? We would get the following quantity, which I will give a forward-looking name:

$$T = \frac{\bar{X} - \mu_X}{\sqrt{S_X^2/n}}$$

From our previous statement, we can see that T has a well-defined distribution: It is the ratio of two independent random variables, and since we know the distribution of $\bar{X}$ and S_X, in theory we can work out the distribution of T. Let's do a little bit of the work to uncover the form of this distribution. We can re-write T:

$$T = \frac{\frac{\bar{X}-\mu_X}{\sigma_X/\sqrt{n}}}{\sqrt{\frac{s_X^2(n-1)}{\sigma_X^2}/(n-1)}}$$

And we see that T is equivalent to the a standard normal Z divided by the square-root of a chi-squared with $n-1$ degrees of freedom divided by $n-1$. The name of this quantity is called **Student-T's Distribution** with $n-1$ degrees of freedom. It has a mean of zero and (quickly) converges to a standard normal for large values of n.

Like the normal, there is no closed form function of the cdf of a T-distribution, but we can compute its values numerically. As with the normal, it is helpful to have a nice symbol for the following:

$$\mathbb{P}\left[T > t_{k,\alpha}\right] = \alpha.$$

Where the k represents the degrees of freedom. For our problem, we can compute that $t_{15,1-0.01/2} = 2.947$.

Q08. What is the value of the T statistic for our potato data for an hypothesis test with $H_0 : \mu_X = 0$ and $H_A : \mu_X \neq 0$? We have:

$$t = \frac{2.28 - 0}{\sqrt{8.7/16}} = 3.092.$$

Q09. Is the p-value less than or greater than 0.01? A test statistic of 2.947 would be exactly 0.01, and this statistic is even larger, so the p-value should be smaller.

We can also use the T-distribution to do confidence intervals. Note that the following should hold for any random variable T with a T-distribution having k degrees of freedom:

$$\mathbb{P}\left[t_{k,\alpha/2} < T < t_{k,1-\alpha/2}\right] = \alpha$$

Plugging in the form of our T statistic, we see that:

$$\mathbb{P}\left[t_{k,\alpha/2} < \frac{\bar{X} - \mu_X}{\sqrt{S_X^2/n}} < t_{k,1-\alpha/2}\right] = \alpha$$

$$\mathbb{P}\left[\sqrt{S_X^2/n} \cdot t_{k,\alpha/2} < (\bar{X} - \mu_X) < \sqrt{S_X^2/n} \cdot t_{k,1-\alpha/2}\right] = \alpha$$

$$\mathbb{P}\left[\bar{X} - \sqrt{S_X^2/n} \cdot t_{k,\alpha/2} < \mu_X < \bar{X} + \sqrt{S_X^2/n} \cdot t_{k,1-\alpha/2}\right] = \alpha$$

The T-distribution is also symetric, and so we have the following confidence interval form for μ_X:

$$\bar{X} \pm \sqrt{S_X^2/n} \cdot t_{k,\alpha/2}$$

Q10. What is a 99% confidence interval for the average amount of weight lost on the potato diet based on our data? Plugging in, we have:

$$2.28 \pm \sqrt{8.7/16} \cdot 2.947 \to [0.107, 4.45]$$