Practicum 4: Statistical Inference II

January 1, 2017

Ole J. Forsberg

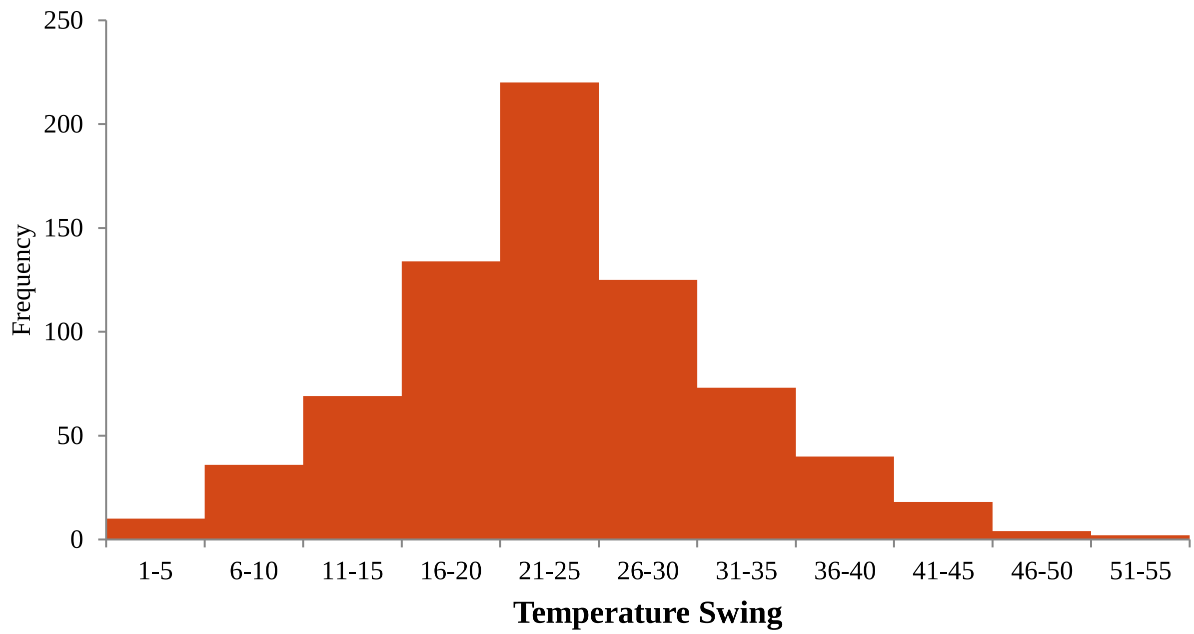
Research Question 1: Temperature Swing

In this first problem, we are to estimate the average temperature swing in a day. To determine the correct procedure, note that we are estimating a population average for a single population. That means we will be using the one-sample t-procedure.

Using all of the data, the mean temperature swing is 23.58, the standard deviation is 8.46, and the sample size is 731. Using the formula for calculating confidence intervals using the t-distribution,

$$\bar{x} \pm t(0.95, n-1)s/\sqrt{n}$$

we get a 95% confidence interval for μ from 22.97 to 24.19°F. Thus, we are 95% confident that the average daily temperature swing for Stillwater is between 22.97 and 24.19°F, with a point estimate of 23.58°F. The following graphic shows the distribution of temperature swings in this data.



Note that the sample size is quite large, so the Central Limit Theorem does not require the swings to have a Normal distribution. However, the distribution of the daily swings is close to Normal. This increases my confidence in the results even more.

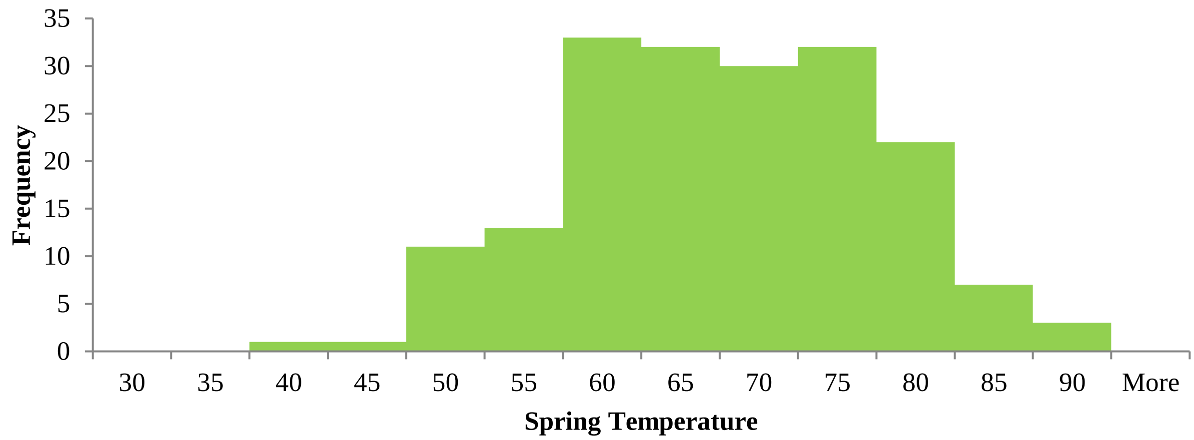
Research Question 2: The Warmer Season

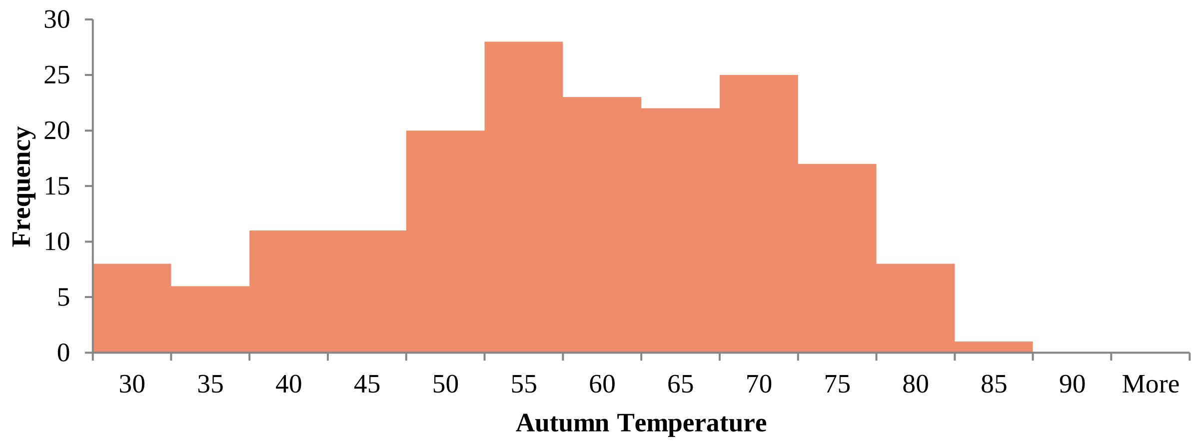
In this part, we are to determine which season was hotter on average, spring or fall. To do this, we will use the independent two-sample t-test. There are two populations (spring and fall), we are comparing means, and we do not know the population standard deviations.

The first step is to define spring and fall. Without prejudice, I use the astronomical definitions. Spring consists of the days March 20 – June 20, 2015, and March 20 – June 19, 2016. Fall consists of the days September 23 – December 21, 2015, and September 22 – December 20, 2016.

According to the independent two-samples t-test (heteroskedastic) in Excel, the p-value is essentially 0. As this is less than $\alpha=0.05$, we reject the null hypothesis that the two means are equal. They are not equal. The sample means tell us that spring is warmer than fall by an average of 9.87°F, with a 95% confidence interval from 7.41 to 12.34°F.

The two histograms illustrate the distributions of temperatures in these two seasons:

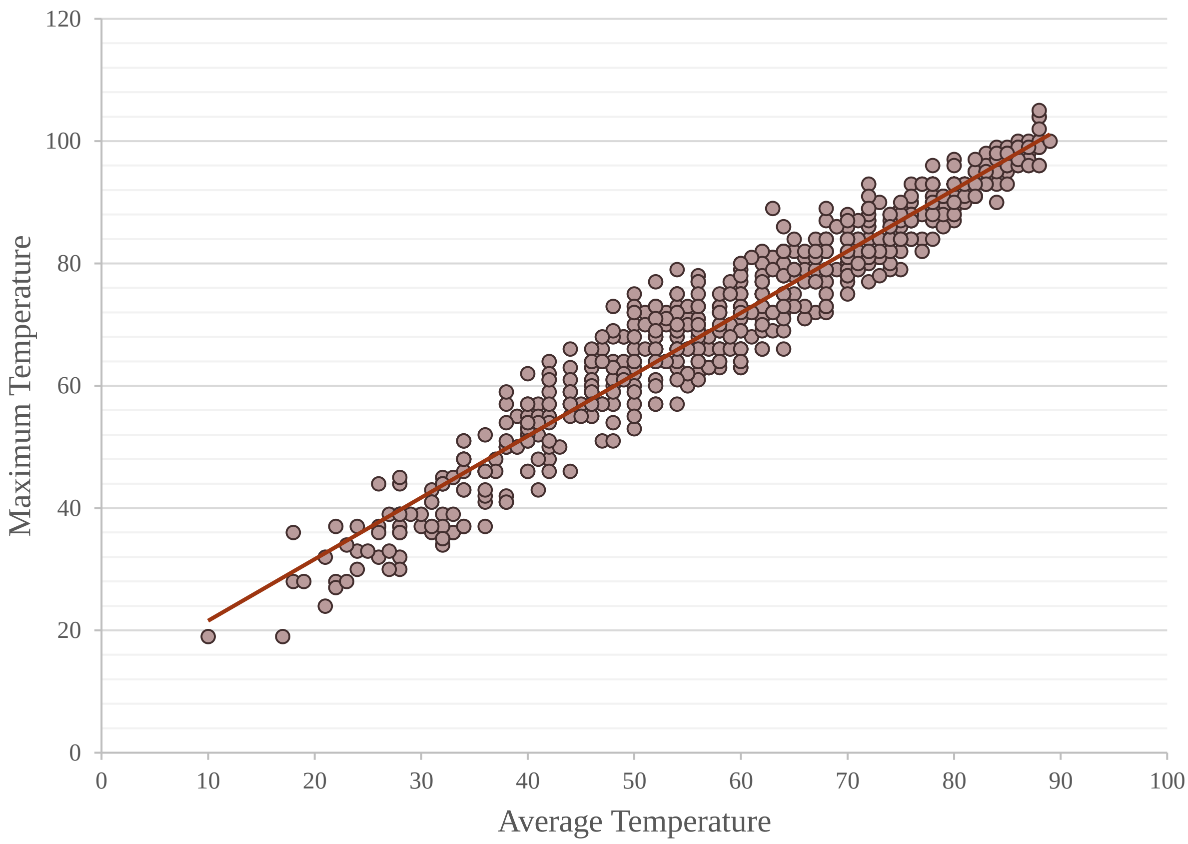




Research Question 3: What is the High?

In this final part, we are to estimate the high temperature on July 25, 2017, given its average temperature. As we need to estimate the value of a numeric variable given the value of another numeric variable, we should use linear regression.

I am using all of the data provided, as the *relationship* between high temperature and average temperature is most likely constant throughout the year, even though the actual highs and averages are not. The scatter plot of the data (below) strongly supports this point. If the relationship differed by time of year, then the scatter plot would be more disjointed.



According to Excel, the equation for that line of best fit is

Maximum Temperature = 11.507 + 1.10069 × Average Temperature

Because the coefficient of determination is so large ($R^2 = 0.9444$), this is a very good fit.

Using this equation and substituting 80 for Average Temperature, we estimate the *expected* maximum temperature on July 25, 2017, to be 92.06°F. Using the formulas in the book gives a 95% confidence interval for the maximum temperature to be 91.60 to 92.51°F.

Using a prediction interval tells us that we are 95% confident the *actual* temperature will be between 83.65 and 100.46°F.