Brewing Insights: A Statistical Exploration of Coffee Import and Export Dynamics

Connor Moss // STAT 216
Prof. Hartlaub
Dec 12th, 2023

Introduction

Depending on what site one visits, those searches may yield slightly different results into the world's most popular drink. However, following water, the silver medal always either goes to tea, or coffee; this paper will focus on the latter. How many people ignorantly start their day with a "cup of joe", "java", even "café" and proceed to forget about this magical potion that has provided the recipient with enough energy to function on the five to six hours of sleep that they got the night before… day in, and day out.

If one were to ever reflect for a second and take up an interest in where our coffee is coming from – instead of solely reaping the benefits and carrying on until they need it again –, then that same thoughtful person should keep reading this paper.

Datasets

Before I proceed, let's take a look at how our data is arranged so we have a better sense of the presented material. There are three main datasets that I am using for the following analyses. The datasets measure:

1) Import # by country
2) Export # by country
3) Production # by country [1]

Figure 1.1

	Country	X1990	X1991	X1992	X1993	X1994	X1995	X1996	X1997	X1998
1	Belgium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	France	378.06	393.18	396.72	380.04	382.14	372.84	399.60	402.24	394.56
3	Germany	820.26	793.74	827.34	846.42	814.98	771.12	810.42	834.30	824.40
4	Italy	314.52	277.80	275.70	335.64	333.24	323.28	336.48	344.58	353.34
5	Netherlands	187.68	187.26	192.30	168.24	168.00	174.60	185.04	175.44	171.72

Given the fact that there were 50+ countries in each of the datasets, I decided to focus on the top 12 for each of the three categories for a more feasible, appropriate analysis. For example, there would be no need to test if there is in fact a significant difference between the importing numbers of Cyprus and France during the 90's, given that France's population is almost seven times the population of Cyprus. One initial finding was that coffee can most definitely be characterized as a luxury item given that the stronger global economies are too among the biggest importers of coffee. Is there a correlation between innovation, production, a strong global presence and the amount of coffee a nation consumes? Though

[1] These three datasets are all organized in an identical format. Sorted by country, with value for each year from 1990-2019 for amount of coffee (units are 60kg bags, due to readability I scaled the values, i.e. 378.06 ≈ 378 million bags of coffee in 1990 France)

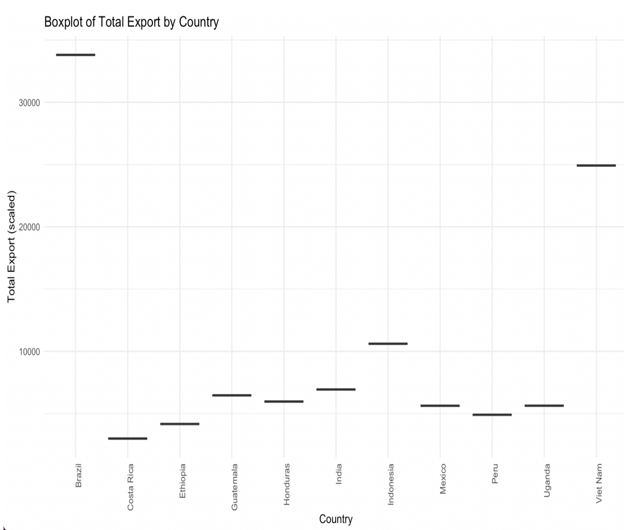
that question may not be answered, here, please enjoy all of my other findings and insights from "one of" the world's most popular drinks.

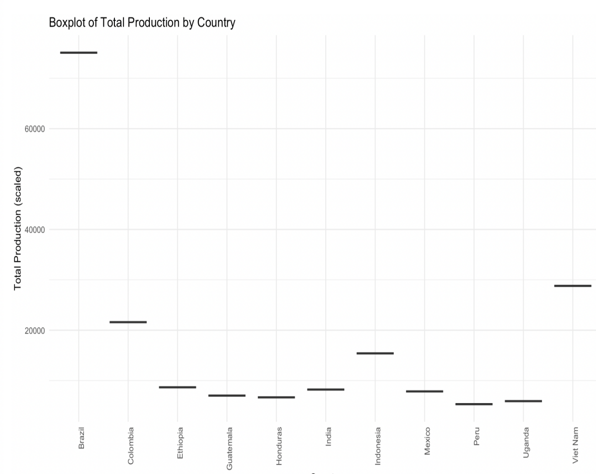
Exploratory Data Analysis (EDA)

In order to get a grasp on our data we can look at numerical summaries and boxplots to see how our datasets stack up. Let's first look at whether the *Total Export by Country* and *Total Production by Country* are comparable. Though the five-number summary seems to suggest completely different data, the boxplots actually show a relatively similar distribution among the two sets. We can also see that outside of Brazil (which exports and produces the most amount of coffee by a large margin) and Vietnam, the rest of the countries are quite close. We will use parametric and nonparametric procedures to test for significant differences amongst these remaining ten countries.

Figure 2.1, 2.2, 2.3

```
> summary(filtered_export_scaled[ ,32])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   2992    5266    5965   10184    8771   33808 
> summary(filtered_production_scaled[ ,33])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   5318    6847    8222   17324   18502   75083 
```





Testing Spread/Variances (Levene's Test & Jackknifing)

To start, we can address the problem of spread. Despite the numerical summary and the boxplots, we haven't compared the variances of our production and export data. Luckily, we have a parametric and nonparametric procedure to compare that will allow us to quantify our variance comparison. The procedure of Jackknifing is used as a nonparametric procedure instead of Ansari-Bradley procedure due to the strict assumption of equal medians among the two datasets which the numeric summary seems to suggest is not met. The parametric counterpart to the Jackknife procedure is Levene's Test which operates under the assumptions of independence between the groups, and that the groups are approximately normally distributed.

For readability-efficiency sake, X is the export data and Y is the production data. Let's start with the non-parametric approach, Jackknifing. When using the Jackknife procedure we test H_0: $\gamma^2 = 1$ against H_A: $\gamma^2 \neq 1$, or that the population variances of the two groups are equal against the alternative that they are not equal.

Figure 3.1

```
> result <- JackKnife(X, Y, alpha)
The Jackknife estimator is:  0.482057
The  95 % confidence level is: ( 0.301548 , 0.77062 )
The  95 % lower confidence bound is: ( 0.3251717 , inf)
The  95 % upper confidence bound is: (0, 0.7146345 )
> result$Q
[1] -3.048542
```

The results from Figure 3.1 suggest that in fact we have sufficient significant evidence to reject H_0 in favor of H_A. Why is this the case? "I don't see 0 in the interval, shouldn't that suggest evidence towards the null hypothesis?" Great question, however given the nature of the Jackknife procedure and the fact that it lacks the comforting assumption of a symmetric normal distribution, this procedure produces asymmetric confidence intervals based on the distribution of resampled estimates. The Jackknife estimator of 0.482 is the estimate of the ratio of variances, which in the context of this problem implies that the variance in export data is smaller than the variance in production data. The Q-value of -3.049 from the Jackknife procedure indicates that the variance of our export data is smaller than the variance of the production dataset (given by the negative value). The value of the Q signifies the standard deviations away from zero, suggesting that this result is over 3 standard deviations away which would align with the significant result we found. We can feel confident in our conclusion with further evidence based on the 68-95-99 rule that suggests over 99% of the population falls within 3 standard deviations from the mean.

The implications of this test imply that coffee exports might be subject to less variation than coffee production over the years. Factors could include the low volatility of the coffee market, the geographic factors (Brazil and Colombia are consistently optimal climate spots for coffee growth), and predictability of the market over time which could make for interesting regression comparisons.

Now we shift into the classic parametric approach. Levene's Test operates with H_0: $\sigma_X^2 = \sigma_Y^2$ against H_A: $\sigma_X^2 \neq \sigma_Y^2$

Figure 3.2

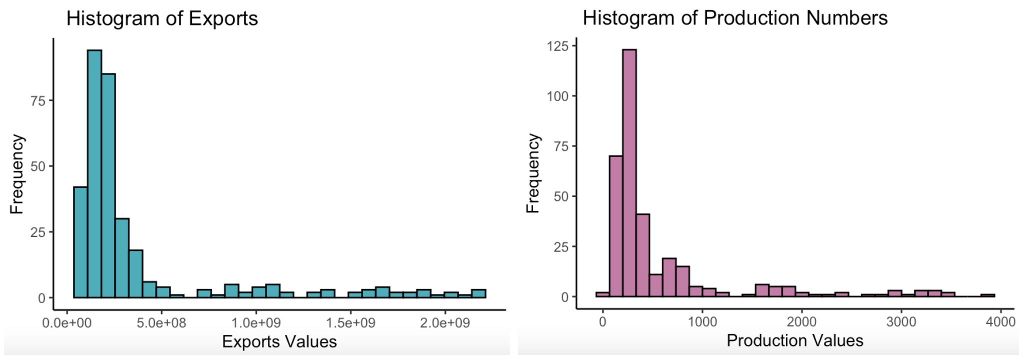
```
Levene's Test for Homogeneity of Variance (center = mean)
       Df F value    Pr(>F)
group   1  22.793 2.225e-06 ***
      658
```

The results from the Levene Test, degrees of freedom = 1, 658 because we were comparing two groups and the within group degrees of freedom (658), our F-value = 22.793 and a p-value ≠ 0 which is smaller than any appropriate alpha ∝ level suggest that we have statistically significant evidence to reject our null hypothesis, thus concluding that the variances of coffee production and coffee export data are significantly different when considering their mean values. Both the nonparametric and parametric procedures reached the same overall conclusion, but which one is right?

Conditions/Normality

The primary difference for nonparametric and parametric is due to the normality condition being met, or not. One way we can go about checking this condition is by viewing histograms and analyzing the shape of our data. Both sets of data can be seen as severely right skewed signaling that the parametric procedures may not be best suited to handle this data. Despite the fact that both Levene's Test and Jackknifing produced the same conclusion of significant differences amongst the variances, we should proceed with caution on the results of Levene's Test due to the fact that the data is not normally distributed. Hence, why nonparametric counterparts, such as Jackknifing exist.

Figure 4.1, 4.2



Another way to check the conditions of normality (linearity and random spread) which are most usually considered the conditions for parametric procedures are analyzed in a QQ Plot and a Residual v. Fitted Plot, which will be addressed in the following section.

ANOVA & Friedman Test (Multiple Comparison Follow-up)

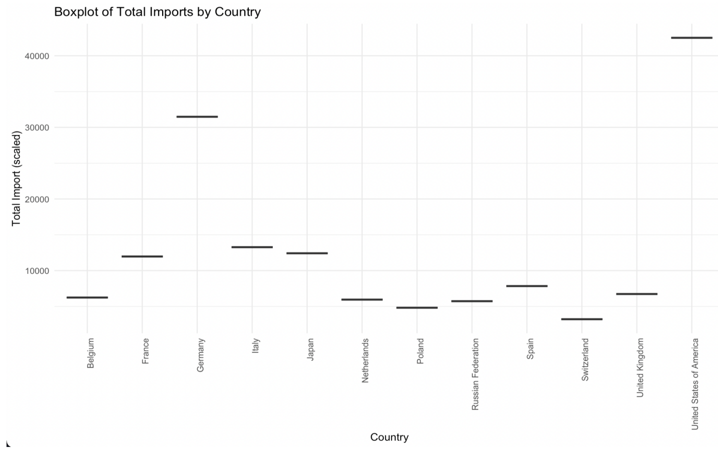
Figure 5.1

$$X_{ijt} = \theta + \beta_i + \tau_j + e_{ijt}, \quad i = 1, \ldots, n; \quad j = 1, \ldots, k; \quad t = 1, \ldots, c_{ij},$$

where θ is the overall median for the nonparametric procedures (replace θ with μ for parametric model), τ_j is the treatment *j* effect (country), β_i is the block effect (year), and the *e*'s form a random sample from a continuous distribution with median 0.

With our model in place, we can move onto our hypotheses for the given tests. But before that, maybe you are wondering why run an Anova test, or what one hopes to gain out of performing such procedures? Figure 5.2 makes for a baseline visualization of what we could speculate in regards to the differences of the top 12 coffee-importing countries. It is quite clear that the United States and Germany are the two biggest importers of coffee by a particularly large margin – with the US importing over 40 billion pounds of coffee over the last 30 years. Because of that, I am more focused on whether or not statistically significant differences exist between countries in the 3rd - 12th spot, such as Italy and Japan. And learning this from our EDA, I will intentionally leave US and Germany out of this test as they are obviously significantly different from the rest (as done in earlier/off stage analysis). Let us turn to statistical approaches.

Figure 5.2



Classical Approach (ANOVA)

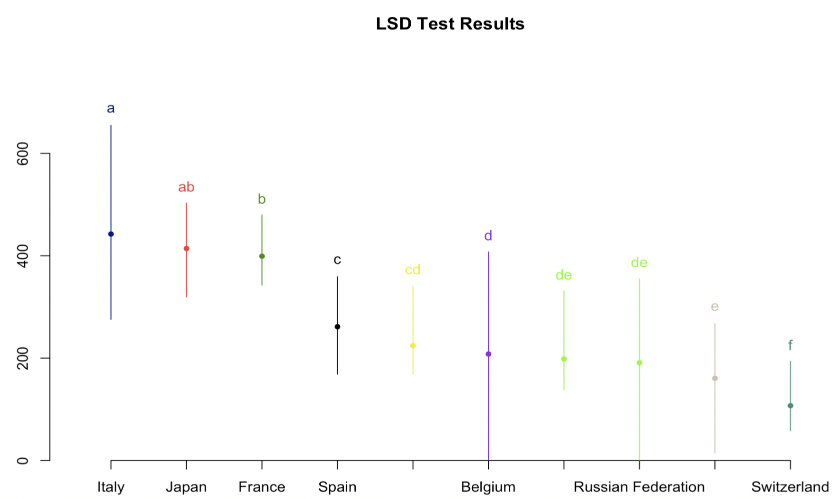
First we start with the classical approach of Anova with blocking, where we test the H_0: $\mu_1 = \ldots = \mu_{10}$ against H_A: at least one of one of the μ are different. Said another way, we are testing if the mean import numbers are the same across all countries against the alternative that this is not the case. When we run *with(filtered_import_long_minus2, aov(Import_Value ~ Country))* we see the following output in which our df = 9, 290 (10 countries - 1 = 9 // 300 observations - 10 groups = 290) and an F value = 66.75 which is derived from dividing the two mean squared terms. We know that 66.75 is a significantly high TS as it correlates to a p-value of approximately 0, which is less than any reasonable value we would use for alpha. With this in mind we find that there does exist at least one country that has a significantly different mean value of total coffee importation than the rest (we reject our null hypothesis). We suspected this result based on the boxplots at the beginning of this section.

Figure 5.3

```
             Df  Sum Sq Mean Sq F value Pr(>F)    
Country       9 3664790  407199   66.75 <2e-16 ***
Residuals   290 1768998    6100                   
```

It isn’t quite useful or interesting to simply state that there exists a difference among the groups. “I don’t know where it is, but I do have evidence that shows that at least one difference exists. Alright, good day." Good luck making a living on delivering those kinds of results. Though, what one could do to provide some valuable insight to a prospective boss who was interested in coffee import numbers, is follow-up this Anova procedure with a multiple comparisons test. This is a way of answering the question of where that difference exists, or where in fact the evidence in favor of our alternative hypothesis exists. Using Tukey HSD provides a lengthy chart of every combination with the respective confidence intervals and p-values to signify whether a difference exists between the two countries or not. Luckily, Fisher's Least Significant Differences provides an easy-to-grasp visual representation of all of the significant differences. But we must be careful with Fishers LSD test as unlike Tukey, the family-wise error rate is not control. This could lead to more Type I (false-positive) errors, but given the non-deadly implications of a false-positive in this study, we can proceed with Fishers LSD; as seen in Figure 5.4

Figure 5.4



From the chart we can see that significant differences exist in the mean number of imported coffee between the likes of Spain and Belgium, but not Belgium and Russia. We can see that Italy is significantly different from France, but Japan is neither significantly different from Italy nor France. If someone were looking to quantify how large the differences were, then the aforementioned confidence intervals Tukey HSD provides could come into play. For example, Japan imports an average of 975 million less pounds of coffee each year than Italy, which may seem like a lot, but given the corresponding p-value of 0.925, we can accuratley say that this difference is not significant and more importantly, the scaling of the amount of coffee that a country imports is completely incomprehensible to us. From an analysis like this, one could use this research as a starting point into understanding why certain countries have higher import or export values – especially geo-economically equivalent countries, if in fact significant differences exist between such countries. One could also use this to look at consumption habits by countries, or consumer behavior in regards to coffee.

Nonparametric Approach (Friedman Test)

The nonparametric approach allows us the freedom to unconcern ourselves with the distribution of the data. With the Friedman Test, we are testing H_0: [$\tau_1 = \ldots = \tau_k$] against H_A: [$\tau_1, \ \ldots, \tau_k$ not all equal] or that the treatment effects are identical across all treatments in regards to the median amount of imported coffee. Whereas the alternative suggests that all of the treatment effects are not identical, and at least one of the treatments (countries) has a significantly higher or lower median than the others – or that the country factor indeed does have an effect.

Figure 5.5

```
        Friedman rank sum test

data:  Import_Value, Year and Country
Friedman chi-squared = 272.67, df = 29, p-value < 2.2e-16
```

From this result we see that we have df = 29 (30 treatments - 1), $\chi^2 = 272.67$, and a p-value of approximately zero which signals that we have statistically significant evidence to reject our null hypothesis and conclude that at least one of the treatment effects is in fact different from the others. Now, in order to find where that difference lies, we use a nonparametric multiple comparison procedure by the likes of Wilcoxon, Nemenyi, McDonald-Thompson. Like the Tukey command, the *pWNMT()* command produces an extensive list of every possible treatment combination, so I will highlight the few that were mentioned in the parametric setting in order to compare.

Figure 5.6

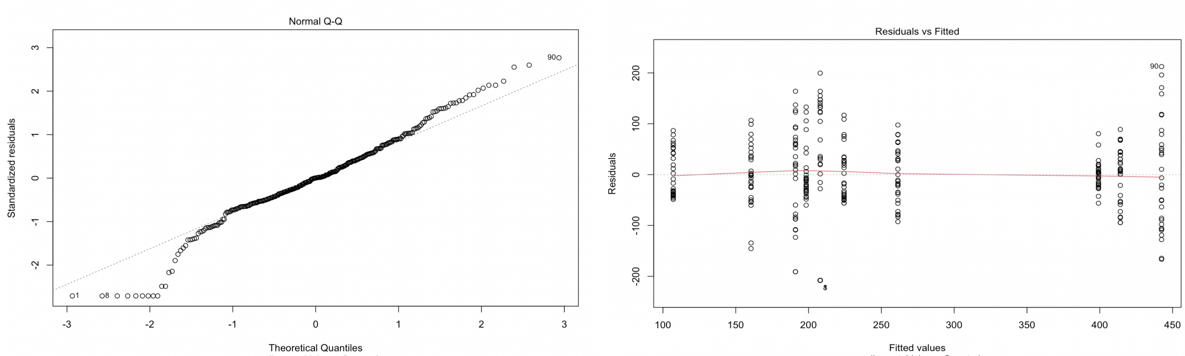
```
For treatments 4 - 5, the Wilcoxon, Nemenyi, McDonald-Thompson R Statistic is 4.
The smallest experimentwise error rate leading to rejection is 1.

For treatments 4 - 6, the Wilcoxon, Nemenyi, McDonald-Thompson R Statistic is 156.
The smallest experimentwise error rate leading to rejection is 0.
```

For the sake of this brief example, 4 = Italy, 5 = Japan, 6 = Netherlands. We can see that R gives a straightforward understanding of the differences that exist or don't exist between groups. In this example we can see that the treatment effect is not significantly different between Italy and Japan (same conclusion as the classical approach), and their does exist a significant difference between Italy and Netherlands based on the "smallest experimentwise error rate leading to rejection is 0" (think of like a p-value).

After utilizing both approaches, again we should ask ourselves which results are more applicable to this appropriate dataset, and to check this time (instead of a histogram) we are going to look at the aforementioned plots from the end of the previous section.

Figure 5.7 and 5.8 present two standard ways to check the conditions of many common parametric procedures. First we start with the QQ Plot, which is our test for normality, we hope to see as many of our observations on the 45-degree line as possible. Notwithstanding the chunk of points that fall on the line in the middle of the graph, we see that the tails are severely misaligned with the rest of the line, signaling that one: our normality condition is not met and a nonparametric procedure may have more reputable results and two: if one wanted to proceed with a parametric procedure, both tails of the QQ Plot failing to be on the line could mean a transformation (exponential, logarithmic) would solve that issue. Then, we move onto the Residulas vs Fitted Plot which is how we test linearity and constant variance. Although it seems to be relatively even on both sides of the red line, which would check off our constant variance condition, our linearity condition is completely violated. We would hope to see no glaring pattern, or random spread amongst our observations in the Residulas vs Fitted Plot. Obviously here, we do not have that, and instead have evident pattern (more specifically, discrete data) which is more evidence to the fact that we should most likely follow the results from the Friedman's Test instead of the classic Anova approach for this specific dataset.

Figure 5.7, 5.8



Kolmogorov-Smirnov

In my deep-dive of geographic coffee analysis, I also wondered if the distribution of countries that exported the most coffee was the same as the distribution of countries that produced the most coffee, and if not, were and why does that disconnect exist. To tackle this question I called on the Russian scientist Andrey Kolmogorov to find out if the underlying distributions of exportation and production were the same. This is because 10 of the 12 countries in the one dataset were also at the top of the other dataset, so after unlisting the data to solely contain the values, I ran *ks.test*() command.

Figure 6.1

```
Warning: p-value will be approximate in the presence of ties
	Asymptotic two-sample Kolmogorov-Smirnov test

data:  kolmogorov_scaled_export and kolmogorov_scaled_production
D = 0.31212, p-value = 2.187e-14
alternative hypothesis: two-sided
```

We find that the maximum difference between the cumulative distribution functions (CDF) for export and production is 0.3121, which indicates a noticeable difference between the two distributions. We can feel quite confident in our results due to the extremely low p-value that is approximately zero. The outcome suggests that the top-producing coffee countries might not align with the top-exporting countries; or that the distribution trends for production and export across different countries do not mirror each other. But is that accurate?

I added this section to show how statistics can be misused. We know from the boxplots and the datasets that the countries involved in both exporting and production are almost identical (aside from Columbia), however, this test seems to offer evidence against that fact – to someone who doesn't understand the implications of this particular test. The Kolmogorov-Smirnov test is not at all concerned with the labels, central tendencies, dispersion, or any other parameters The KS test looks at the shape of the distribution, and from those same boxplots (due to data filtering based on the alphabetized names of the countries) we can see that the two distributions are quite different. See Figures 2.2 & 2.3. Whereas the production data seems to peak and dip multiple times throughout, the exporting data actually seems to follow the umbrella alternative trend quite nicely. [2] Without understanding the interpretation of the KS test, one would be led to believe that statistics and more importantly R is incorrect, in saying that the same countries that export the most coffee aren't the same countries that are producing the most coffee. A better

[2] For this part of the analysis I excluded both endpoints (Brazil and Vietnam) as they were both extremely similar in both distributions and would not have allowed for any accurate shape to form

approach would be to test correlation between coffee exportation and production. Which conveniently leads us to our penultimate section.

Correlation (Pearson & Kendall)

The question that I was implicitly trying to answer (incorrectly) in the last section was whether or not the countries that export the most coffee are correlated to the countries that produce the most. To answer that question, we turn to a correlation comparison between nonparametric and parametric procedures. Here we are testing $H_0 : [F_{X,Y}(x, y) \equiv F_X(x)F_Y(y), \text{ for all}(x, y) \text{ pairs}]$ against the alternative hypothesis (H_A): $\tau \neq 0$, or put another way, we are testing whether the correlation coefficient $\tau = 0$, indicating no relationship. This is being evaluated against the alternative that suggest some level of dependence between X and Y – which for this section will be the total_production values and the total_export values, respectively.

The above hypotheses are for the nonparametric Kendall's test but, the idea is almost identical in the classical sense, however the hypotheses are H_0: $p = 0$ vs. H_A: $p \neq 0$, which is simply replacing the way correlation is calculated. Whereas Pearson's correlation coefficient is calculated by dividing the covariance by the two measured variables standard deviations. Kendall's Tau is calculated in a ranked counted measure by dividing the difference of the concordant and discordant pairs by the sum of the concordant and discordant pairs.

We start with the nonparametric approach as we have seen the shape of the data in earlier sections and can conclude that Kendall's Tau would be more appropriate based on that notion. We see from Figure 7.1 that our τ-hat = 0.6 meaning that we have a positive and fairly strong correlation. We see that our T test statistic = 36 which signifies the amount of concordant pairs between the two datasets, and we know that this TS is significant based on the p-value = 0.0166 < $\propto$ = 0.05. Because of this, we can conclude that in fact we have sufficient evidence to reject the null hypothesis and state that there is significant correlation between total production and total export between countries. Based on the τ-hat value, we should expect that as the total production values increase, the export values should increase as well in the included countries.

Figure 7.1

```
	Kendall's rank correlation tau

data:  x and y
T = 36, p-value = 0.01667
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau
0.6
```

Moving to the classical approach of Pearson's correlation (though knowing that these results may not be applied based on the normality condition not being previously met, though through a transformation, then these results may be applicable) we find through Figure 7.2 that the correlation coefficient is approximately 0.946 indicating a strong, positive linear relationship between the total production and total export of coffee. In a practical sense, this suggests that countries with higher coffee production tend to also have higher coffee export numbers – which isn't too groundbreaking, but not all statistics is about finding those niche, earth-rattling insights. The p-value is approximately 0 and suggest that this observed correlation is statistically significant; or it is highly unlikely that this strong correlation occurred by chance. The t-value = 8.221 is understood via the p-value as a signficant one and provides evidence against the null hypothesis. The degrees of freedom are simply calculated as 10 countries - 2 (because there are 2 variables being compared). And the confidence interval – which provides more insight than the one-shot p-value also provides evidence to the fact that this strong correlation is not some random occurrence. Given the fact that zero is not found in the interval, we can interpret the confidence interval (0.781, 0.987) as the range of values for the correlation coefficient 95% of the time if one were to repeatedly conduct this identical analysis, further providing evidence against the null hypothesis, thusly concluding that coffee exporting numbers and coffee production numbers are strongly correlated.

Figure 7.2

```
	Pearson's product-moment correlation

data:  x and y
t = 8.2211, df = 8, p-value = 3.587e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7809055 0.9873709
sample estimates:
      cor
0.9456003
```

Smoothing/Line Graphs

In a perfect world, this paper would naturally flow into regression and attempt to predict what the next decade of coffee evaluation would look like as seen through simple linear regression and nonparametric theil regression. But just like working with real world data, unexpected problems arise, and I am down to my final three pages still needing to introduce an uncovered topic from this past semester, so we will discuss smoothing.

I found Friedman's approach of cross-validation (that strongly resembled the idea of jackknifing) interesting. The idea of using local averages of nearby neighbors instead of the actual observations is a conservative way to maintain the optimal span size. Due to this, the concern for the span size being too large and falling victim to oversmoothing or the span size being too small and being highly sensitive to random fluctuations is addressed. In doing this, we can hope to visualize and understand complex patterns in the data without the constraints of the classical approaches. Let us first look at the line graphs to understand how the import numbers for the ten highest countries have changed over the last three decades. We can see that while just about every country has seen some level of increase, two countries that could be specifically of interest for smoothing would be the US and Belguim, as both seem to have odd rapid peaks and drop-offs.

Figure 8.1

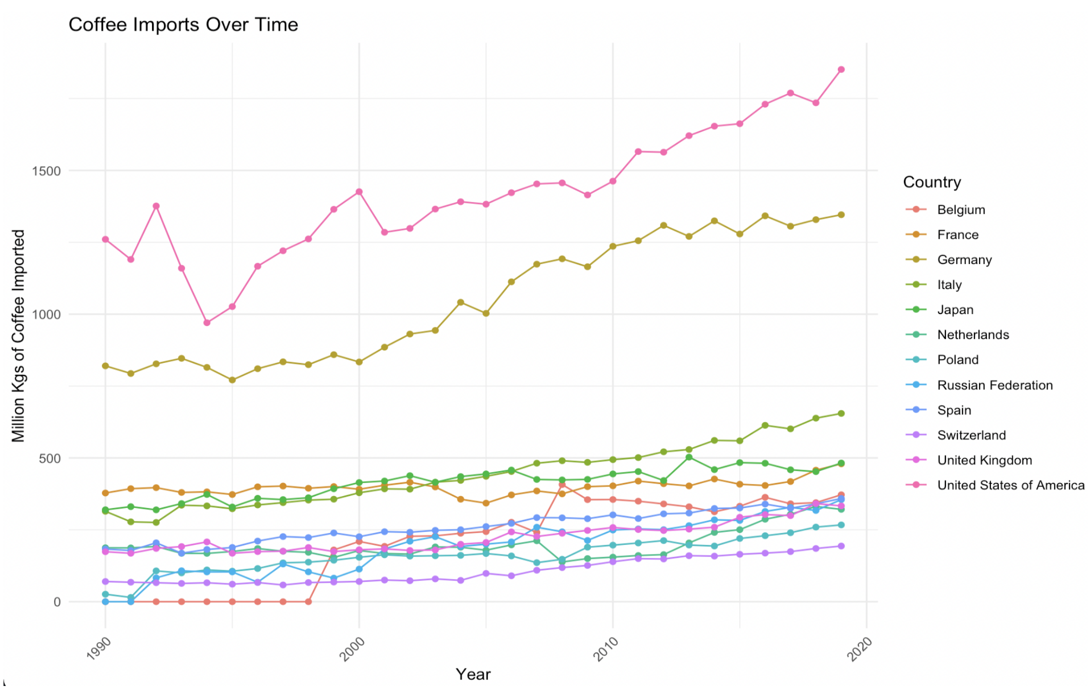
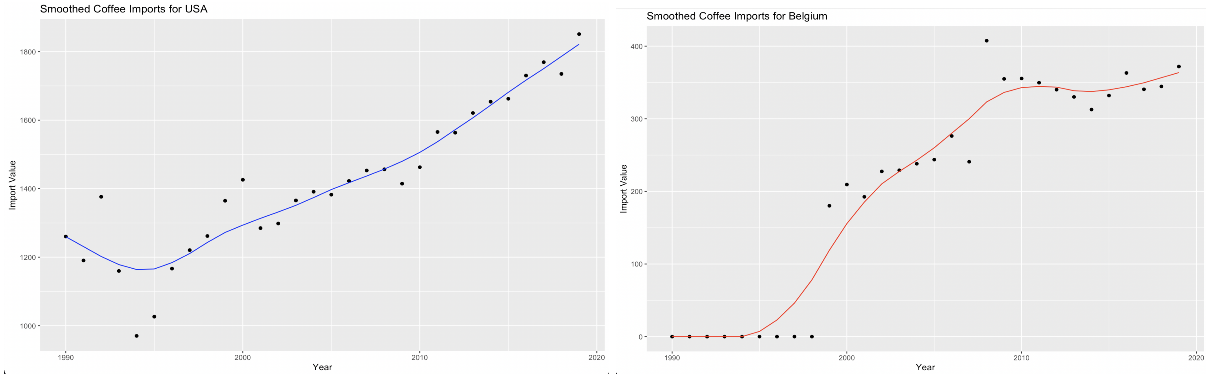


Figure 8.2, 8.3



After performing the necessary smoothing techniques on these two countries, we can see that for the US, the middle of the 90s rapid decline and immediate incline is less accounted for by the smoothing, but in return a more understood trend becomes apparent. This could be two sided. If an event caused this unexplained break from the trend, then this is something that may want to be represented in an analysis. If the US took a year off from importing coffee for some odd reason, then this would obviously want to be discussed an not simply smoothed over as it adds to the overall story. However, if one were simply trying to introduce smoothing methods into their paper for the first time, than it would be appropriate to smooth over this unsettling break in order to display the consistent theme that occured for United States coffee importing over the last three decades. In the case of Belgium, smoothing may not be appropriate as it is clear that Belgium did not start importing coffee until the 2000s. Because of this, we have an overrepresented portion of the smoothed line that looks like the 2000s started the boom of coffee in Belgium. This would be an example of where Figure 8.1 may more appropriately tell the story of coffee importing numbers in Belgium over the last 30 years.

Conclusion

This paper acts as a baseline and a stepping stone into the world of global coffee dynamics and nonparametric statistics. Whereas an ignorant, younger version of myself would have resorted to endless transformations and hours of data wrangling, – ultimately resulting in me switching topics with easier analyzed data – I reveled in the opportunity to apply nonparametric procedures onto "messy," real-world data; also comparing the classical methods to these newfound nonparametric procedures.

We observed that countries with robust economics are among the top importers of coffee, highlighting the implication of this luxury commodity perhaps being responsible for innovation and productivity. Notably, we discovered the overall stability of the coffee industry and the predictability based on geographical factors through testing the variances. Though the Jackkinfe and Levene's Test produced the same conclusions (export data is subject to slightly less variation than production data), how they came to that result is through different means, ultimately lending the Jackknife procedure more useful due to the shape of the data.

We observed a misused case (or rather a false conclusion reached on the grounds) of the Kolmogorov-Smirnov case in which we incorrectly suggested that export and production data do not come from the same distribution. But we fixed this idea thanks to Kendall and Pearson by finding out that there exists a strong positive correlation between the export value and production value for those countries. But further analysis should be conducted as correlations does not imply causation.

We finally ended this analysis with the idea of smoothing and how and when smoothing should be applied. We analyzed how smoothing can be used to reconstruct abrupt changes, offering a more consistent trend of coffee imports. On the other hand, we saw how Belgium did not start importing coffee until the early 2000s and how smoothing inadvertently obscured the rapid emergence of the country as a coffee importer

In summation, statistics come in all shapes and sizes (parametric and nonparametric), kind of like coffee beans. And though it is efficient to read about how coffee is "linked to a lower likelihood of type 2 diabetes, heart disease, liver and endometrial cances, Parkinson's disease, and depression" (Putka 2). One should understand that the idea of gaining insight from data is not built on these "headliner" insights, but rather the slow, meticulously crafted statistical procedures, just like the best cups of joe.

Works Cited

Cannon, Ann R., et al. STAT 2: Modeling with Regression and ANOVA. W.H. Freeman and Company, 2013.

Hollander, Myles, Douglas A. Wolfe, and Eric Chicken. Nonparametric Statistical Methods. 3rd ed., John Wiley & Sons, Inc., 2014.

Putka, Sophie. "Is Coffee Good for You or Not?" Discover Magazine, 5 Apr. 2021, www.discovermagazine.com/health/is-coffee-good-for-you-or-not.