Scientific Decision Making
Course Project Part I
Seja Digital – Implementing Hypothesis Tests for Inventory Shortages.

Rodrigo Henrique Correa

Outline

Delivering 15 million TV conversion kits. That is the job.

In order to obtain the concession of the 4G internet, telecommunications companies faced a *quid-pro-quo* situation with the Brazilian government: to universalize the access to digital TV signal, so that the analog signal could be shut down.

In 2014, the Brazilian government auctioned 4G transmission concessions for further developing its telecommunications infrastructure. Among the rules of the auction, it was necessary to shut down the analog transmission signal for TVs, implement the digital signal and then make use of the band frequency for 4G internet.

The shutdown of the analog signal would require that people had full access to digital conversion for television sets, and economically frail people could not go without full access to technology.

From that demand came Seja Digital, a company set up by the auction-winners to provide digital signal converters and antennas. Seja Digital then hired Correios to operate all the logistics of the project.

In order to distribute the conversion kits, Correios had set up distribution points according to Seja Digital's guidelines.

As of July 2017, delivery was completed for the cities and metropolitan areas of Brasilia and São Paulo. The latter is the largest city in the Southern Hemisphere, with its metropolitan area comprising around 20 million inhabitants, and where the business model was really put to the test. Major cities, such as Rio de Janeiro (population 6 Million) are still to begin their distribution.

After the experience in São Paulo, some operational constraints were revealed, and some inventory shortages began to arise from different points of distribution. Even though the operation was successful in delivering the conversion kits to more than 99% of the target population, shortages aforementioned required contingencies and affected the project's profitability.

In this case, as a post-operation analysis, this project is to indicate relevance of those shortages compared to the daily demand of the service. This inference may lead to more accurate resource planning for the following cities.

Therefore, it plans to determine whether those shortages were events of chance or requires distribution processes refinement. In order to do that, there will be made a comparison between two populations: demand and inventory.

Part One – Set Up the Hypothesis Test

The Data Set

The data set comprises of daily inventories and scheduled demands for each of the 60 points of distribution available. In order to transform into shortages, there will be considered demands minus inventory available on that particular point.

This will be illustrated first under the assumption of both populations being normal, under Central Limit Theorem Assumption.

Hypothesis Testing

1. Question: *Is the inventory planning satisfying demand on average?*

2. H_0: *Average Inventory is equal or larger than Average Demand*
3. H_a: *Average Inventory is lower than Average Demand*

Test Statistic

4. Test statistic: *paired t-test (left-tailed)*

This statistic should be used for testing two populations.

To use the correct formula, it is necessary to:

1) Determine whether the variances of the two populations are the same, using F-test.
2) Calculate the t-statistic, using the degrees of freedom indicated by each calculation.

The populations are assumed to be normally distributed. But, the differences between the populations may not be normal, this should be tested as well, with the approaches

such as Shapiro-Wilk or Kolmogorov-Smirnov tests, in order to determine whether this is correct. This leads to another assumption

H_0: Differences ~ Z

H_a: Differences do not follow a normal distribution.

Should this test do not follow a normal distribution, alternatives will be shown in the presentation of the full project.

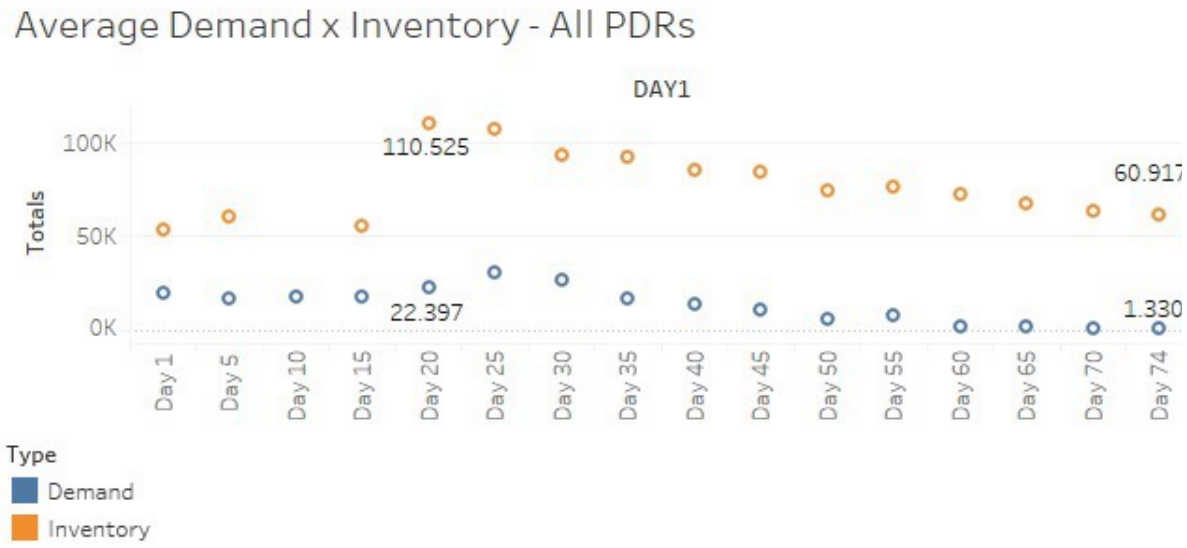
Errors

Error Mitigation		
Error Type	**Downside / risk of this error**	**Current parameters**
Type I (false positive)	*If H_0 is accurate and rejected, efforts and resources will be spent to rectify an operation that works with good results.*	$\alpha = 0,05$
Type II (false negative)	*H_0 is not rejected while it has no statistical support, resources will be spent over the capacity to support a false claim.*	$\Pi > 0.90$

Testing Aggregate Supply and Demand

The operation consists in 75 Distribution Points, called PDRs (Ponto de Relacionamento, or Point of Relationship in English). Those PDRs are fed by 6 Distribution Centers, named Clusters A-F for the purpose of this Project.

Focusing on the claim that shortages affected the operation, it is important to visualize the data as a starting point.



Visualization 1: Inventory versus Demand Time Series. Made with Tableau

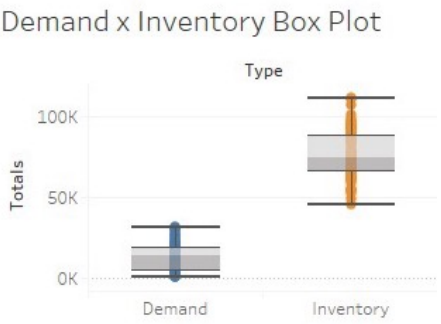
From visualization 1, the compared time series indicate that Inventory in general is much higher than demand in aggregate terms. In this case, the hypothesis tests are probably going to indicate that shortages are not statistically significant for this analysis.

In truth, verifying this information alone indicates that there even might be happening quite the opposite, and that could be derived from the response from the shortage claim and thus creating an effect of constant oversupply. In supply chain this is known as the Bullwhip effect.

From the visualization, it also seems that to assume that inventory in general is lower the demand in any point is unreasonable. For this, the hypothesis will be substituted by its opposite.

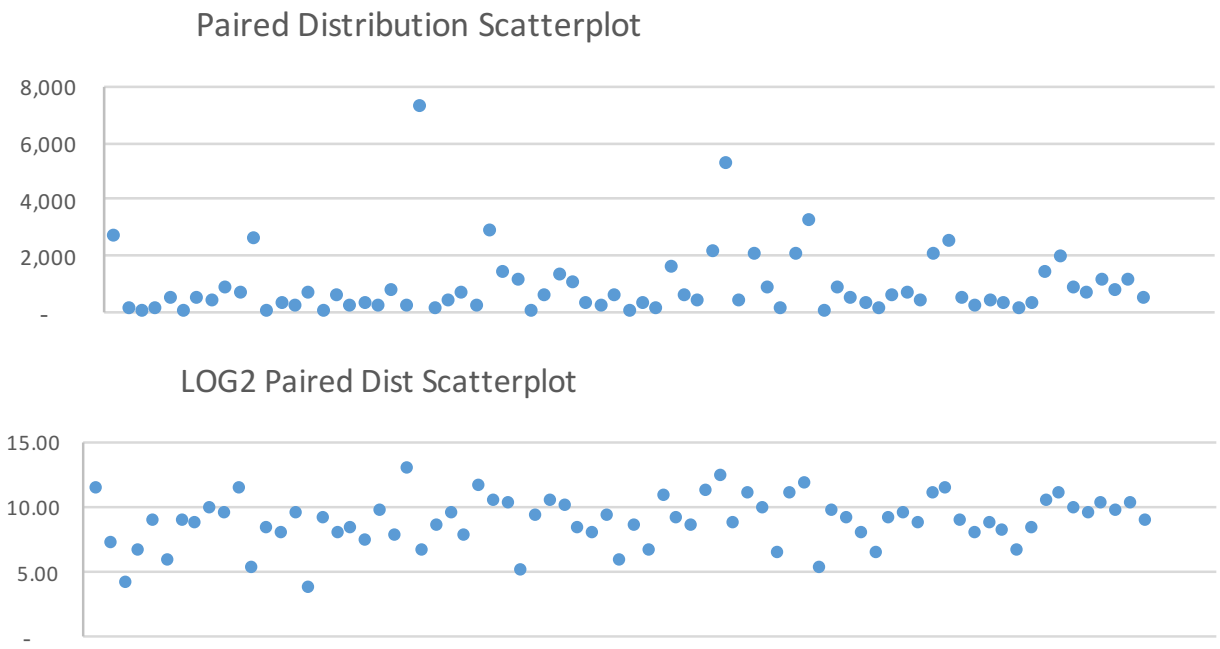
H_0: *Average Inventory is equal than Average Demand*
H_a: *Average Inventory is different than Average Demand*

This creates a two-tailed hypothesis test.



Visualization 2: Box Plot, Demand and Inventory. Made with Tableau

The first step to compare the Demand and Inventory as a whole was to visualize them in pairs, or “net inventory”. This distribution, however, shows very significant outliers and those may interfere in the tests. In this case, for this specific test it was applied the technique of “smoothing” the series through logarithm. This technique conserves the characteristics of the series but outliers distort the result in a lesser magnitude.



Visualization 3: Scatterplots for original paired series and log2 paired series. Data shown by individual PDR.

In this case, the formulation of the first Hypothesis set will be on log2(Inventory) series and log2(Demand) series.

Now comes the decision of using the right hypothesis test. For this comparison, the t-paired test was selected as the appropriate, assuming inventory should depend on demand. For dependent variables, the paired t-test is the statistic to use.

However, there are two types of paired tests: assuming same variance and different variances; this means that before testing pairs, it is necessary to test variances. This is made with F-test, using F-Snedecor distribution. The hypothesis for this set of tests are:

H_0: $\sigma_I^2 = \sigma_D^2$

H_a: $\sigma_I^2 \neq \sigma_D^2$

$H_0: \mu_I - \mu_D = 0$

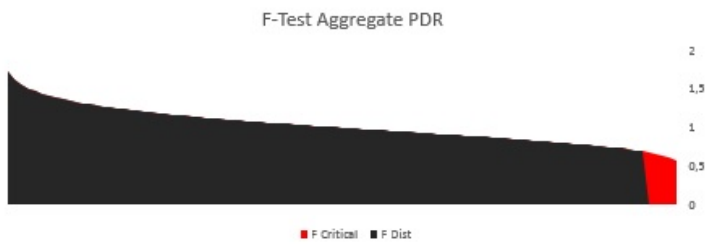
$H_a: \mu_I - \mu_D < 0$

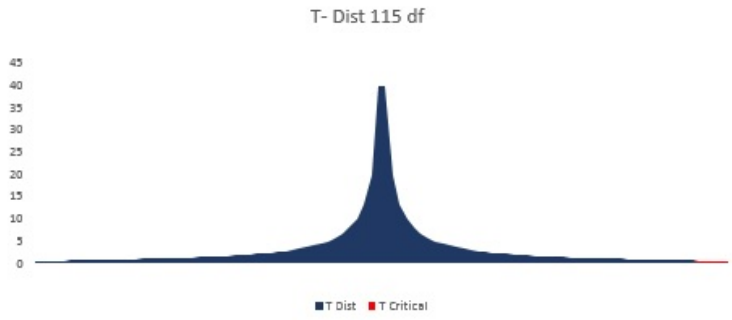
F-test: two samples for variances

	Log2Inventory	*Log2Demand*
Average	9,082279925	7,103381057
Variance	4,153721507	1,262211282
Observations	75	75
df	74	74
F	3,290829012	
P(F<=f) single tailed	3,40808E-07	
F critical single tailed	1,469451006	

T-test: Two samples, different variances

	Log2Inventory	*Log2Demand*
Average	9,082279925	7,103381057
Variance	4,153721507	1,262211282
Observations	75	75
Hipotesis of difference in average	1	
df	115	
Stat t	3,642770617	
P(T<=t) single tailed	0,000202995	
t critical single tailed	1,65821183	
P(T<=t) dual tailed	0,00040599	
t critical dual tailed	1,980807541	





Visualization 3: F-Test, Rejected Null Hypothesis, Paired T-test: different variances, Rejected Null Hypothesis. Translated from Portuguese by the author. Empirical distributions generated in Excel.

Using the Excel's Data Analysis tool, both F-test and T-test were calculated. F-test indicated rejection of the assumption of same variances, moving the T-test for different variances. This test changes the calculation of the degrees of freedom to an equation depending on sample variances.

The T-test rejection and it p-value of 0.00002, while the significance level alpha is set to 0.05 support the first analysis that distributions are not equal on average, even when inventory has a larger standard deviation.

It is, therefore, not possible to support the claim on general shortages affecting the operation in any significant way. Thus, the original shortage claim holds no statistical value for the operation as a whole.

Nevertheless, what if this data indicated poor performance in six out of seven clusters and a major inventory on the seventh? Wouldn't that misrepresent the analysis, even though results could very well be very similar? And how about the Central Limit Theorem assumption, can it hold true?

Skepticism is a Brazilian thing. So, to solve that, the analysis will be separated by clusters in order to verify the results. Also, there will be implemented a Goodness of Fit Test, using Kolmogorov-Smirnov (K-S) model.

Now, the underlying claim is that if the distributions are not normal, the results of the F-Tests and T-tests are not reliable for breaking the normality assumption.

Testing Clusters

Now, the underlying claim is that if the distributions are not normal, the results of the F-Tests and T-tests are not reliable for breaking the normality assumption. All the original PDRs were divided by their supplying distribution center, and inventories and demands were again calculated.

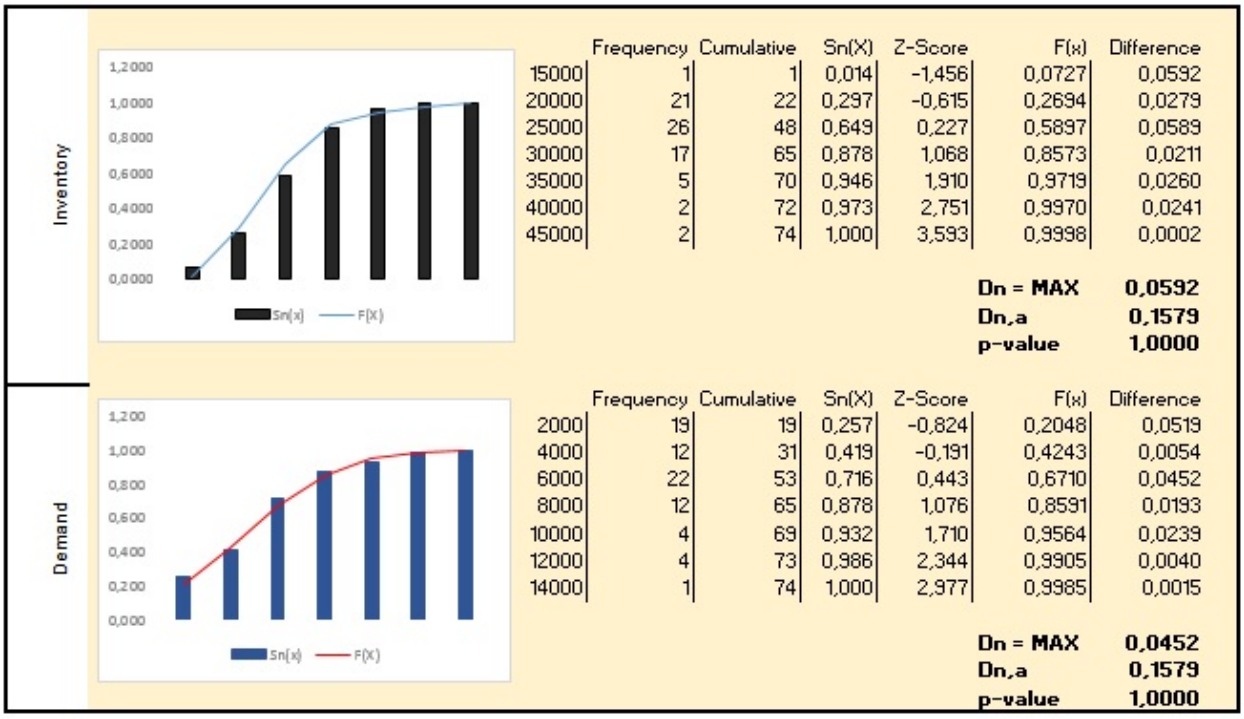
The Kolmogorov-Smirnov Test for Normality

The assumption on normality is the parameter that the classic hypothesis tests rely on. So, in order to prove the possibility of comparison, a goodness of fit test is necessary.

The K-S model was chosen to be widely used and simple. It is not as rigorous as other tests, such as Shapiro-Wilk, but detains a rationale that is simpler to visualize.

The test is based on the maximum distance of the observed distribution relative to the empirical distribution. If the difference is too high, the null hypothesis of normality is rejected.

To calculate the K-S model, I have downloaded a supplement for Excel called RealStats, which calculates the critical and P-value of Kolmogorov-Smirnov distribution.



	Frequency	Cumulative	Sn(X)	Z-Score	F(x)	Difference
15000	1	1	0,014	-1,456	0,0727	0,0592
20000	21	22	0,297	-0,615	0,2694	0,0279
25000	26	48	0,649	0,227	0,5897	0,0589
30000	17	65	0,878	1,068	0,8573	0,0211
35000	5	70	0,946	1,910	0,9719	0,0260
40000	2	72	0,973	2,751	0,9970	0,0241
45000	2	74	1,000	3,593	0,9998	0,0002

Dn = MAX	**0,0592**
Dn,a	**0,1579**
p-value	**1,0000**

	Frequency	Cumulative	Sn(X)	Z-Score	F(x)	Difference
2000	19	19	0,257	-0,824	0,2048	0,0519
4000	12	31	0,419	-0,191	0,4243	0,0054
6000	22	53	0,716	0,443	0,6710	0,0452
8000	12	65	0,878	1,076	0,8591	0,0193
10000	4	69	0,932	1,710	0,9564	0,0239
12000	4	73	0,986	2,344	0,9905	0,0040
14000	1	74	1,000	2,977	0,9985	0,0015

Dn = MAX	**0,0452**
Dn,a	**0,1579**
p-value	**1,0000**

Visualization 4: K-S tests for Inventory and Demand Distributions, Example.

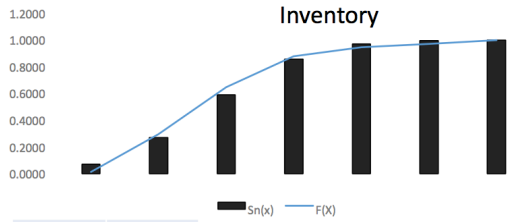
The visual objective of the test is for the observed distribution Sn(x), frequency/n (cumulative), matches the expected values of the normal distribution, F(x). The maximum absolute difference between Sn(x) and F(x) is then compared to the critical value. In the example above, both distributions for cluster A has support for not rejecting the null Hypothesis.The general hypothesis test for the goodness of fit is:

$$H_0: Sn(x) \sim Z(\mu; \sigma)$$

$$H_a: Sn(x) \neg \sim Z(\mu; \sigma)$$

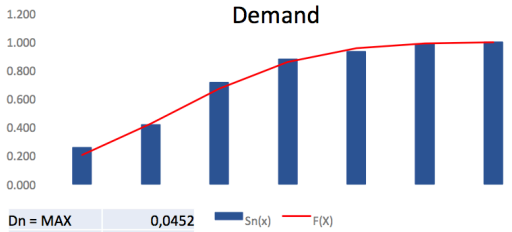
Dividing Clusters A-F, we have the following tests:

K – S Test: Cluster A



Dn = MAX	0,0592
Dn,a	0,1892
p-value	1,0000

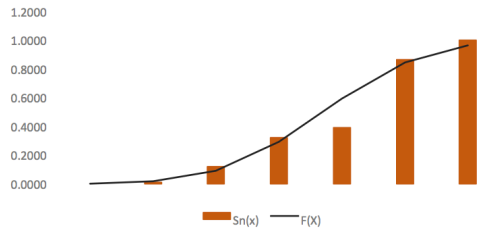
H₀ not rejected



Dn = MAX	0,0452
Dn,a	0,1892
p-value	1,0000

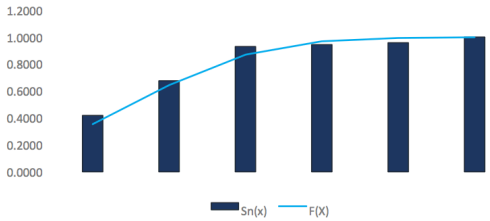
H₀ not rejected

K – S Test: Cluster B



Dn = MAX	0,20144
Dn,a	0,1892
p-value	0,00254

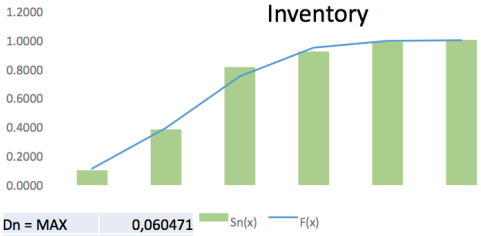
H₀ rejected



Dn = MAX	0,06494
Dn,a	0,1892
p-value	1,00000

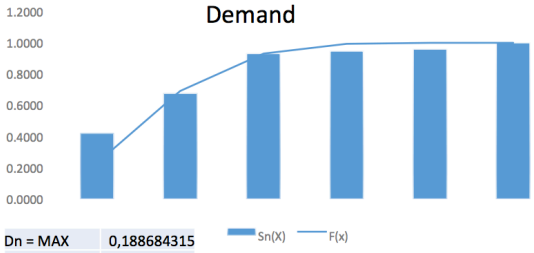
H₀ not rejected

K – S Test: Cluster C



Dn = MAX	0,060471
Dn,a	**0,1892**
p-value	1,000000

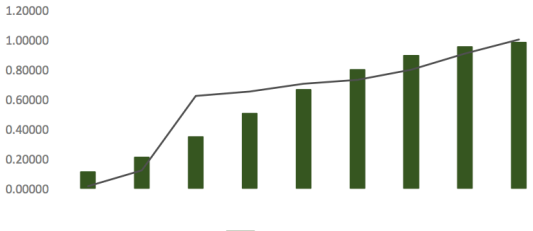
H₀ not rejected



Dn = MAX	0,188684315
Dn,a	**0,1892**
p-value	0,010208267

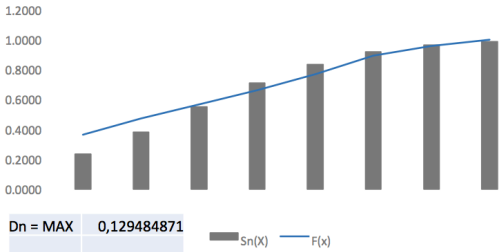
H₀ not rejected

K – S Test: Cluster D



Dn = MAX	0,277351071
Dn,a	**0,1892**
p-value	0,0000000

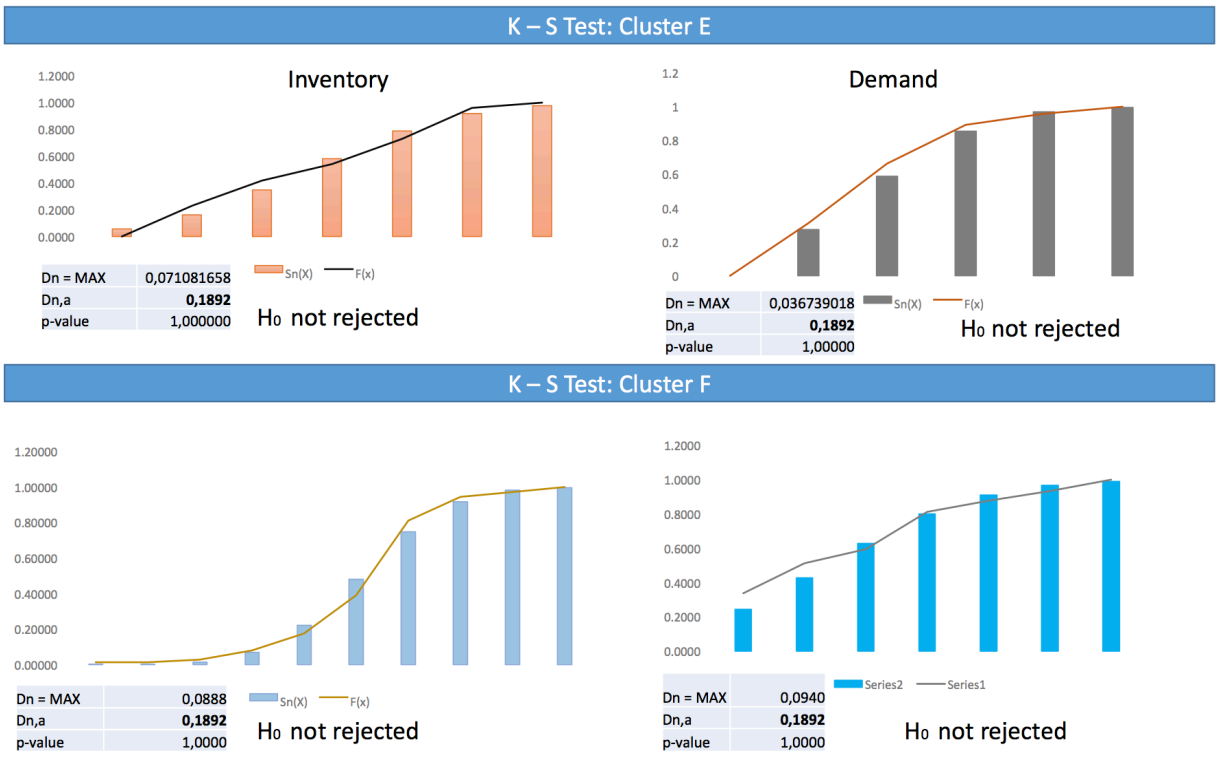
H₀ rejected



Dn = MAX	0,129484871
Dn,a	**0,1892**
p-value	0,165257849

H₀ not rejected

Visualization 5: K-S tests for Inventory and Demand Distributions, Clusters A, B, C and D. Alpha-values: 1%.



Visualization 6: K-S tests for Inventory and Demand Distributions, Clusters E and F.

The goodness of fit tests did not hold the normal distribution assumption for two out of twelve tests performed. This means that clusters B and D will not present conclusive results by the F and T tests alone.

The F and T-Tests: Clusters

Following the same principles of the hypothesis testing for the aggregate Inventory and Demand, the Cluster tests will try to identify shortages as statistically significant.

Cluster A

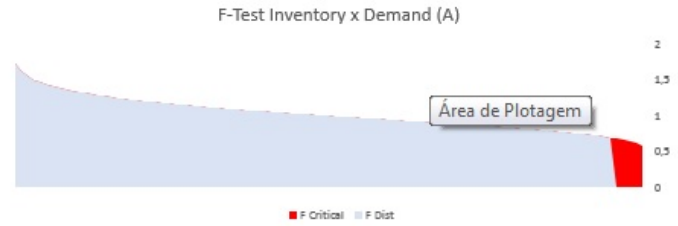
Cluster A

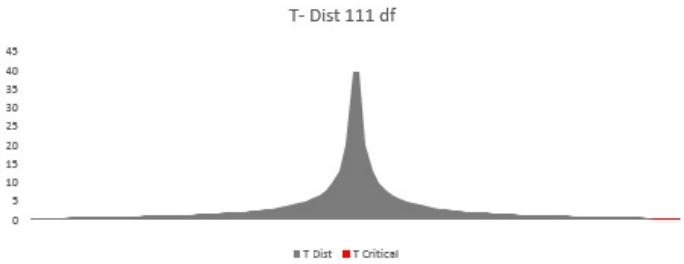
F-test: two samples for variances

	INVENTORY	*DEMAND*
Average	23652,51351	4602,472973
Variance	35304639,13	9963469,376
Observations	74	74
df	73	73
F	3,543408204	
P(F<=f) single tailed	8,87215E-08	
F critical single tailed	1,47336709	

T-test: Two samples, different variances

	Inventory	*Demand*
Average	23652,51351	4602,472973
Variance	35304639,13	9963469,376
Observations	74	74
Hipotesis of difference in average	1	
df	111	
Stat t	24,35526122	
P(T<=t) single tailed	1,21011E-46	
t critical single tailed	1,658697265	
P(T<=t) dual tailed	2,42023E-46	
t critical dual tailed	1,981566757	





Visualization 7: F and T tests for Inventory and Demand Distributions, Clusters A.

By Rejecting the null Hypothesis for variances and for averages, Cluster A shows that demand and inventory are really different. The difference in stat T with the critical value (1.98, dual tailed), shows how extreme the values are in terms of comparison.

As a summary, for cluster A, the tests made were:

Normality for Inventory (K-S): Not Rejected.

Normality for Demand (K-S): Not Rejected.

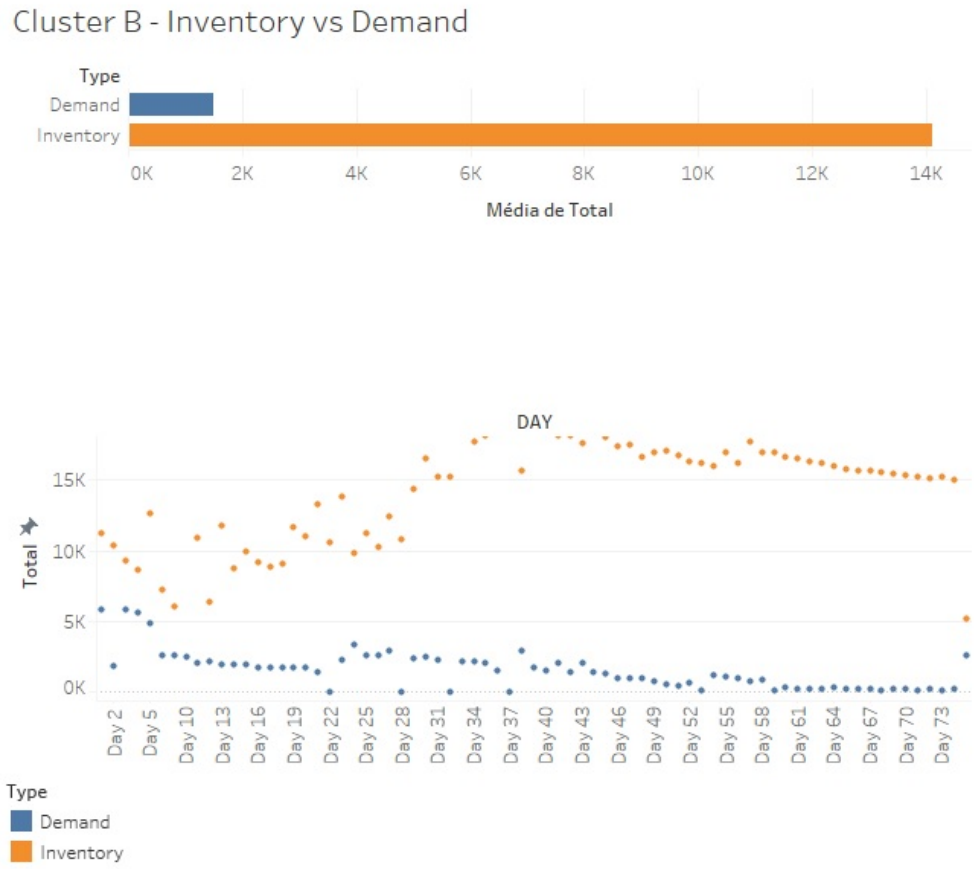
Equal Variances for Inventory x Demand (F-test): Rejected.

Equal Averages For Inventory x Demand (T-Test): Rejected.

Cluster B

Cluster B analysis is one of the clusters in which K-S tests failed for normality, meaning that the tests alone hold little reliability since the underlying assumption for parametric hypothesis testing is that the observed distribution could be approximated to a known distribution.

In this case, one of the simplest ways for statistical inference for non parametric models is simply to visualize them.



Visualization 9: Time Series and Averages Inventory vs Demand, Cluster B.

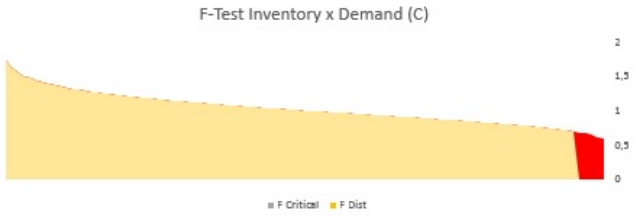
Since in no given point in the observation we witness demand being higher than inventory and averages are so distant, it seems that there are no grounds to support shortages or equality in averages.

Cluster C

Cluster C

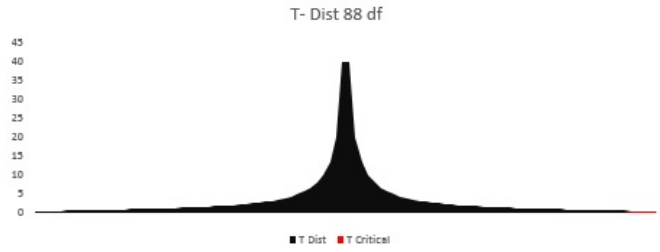
F-test: two samples for variances

	Inventory	*Demand*
Average	6865,378378	1497,851351
Variance	9984213,252	1026668,457
Observations	74	74
df	73	73
F	9,724866079	
P(F<=f) single tailed	3,70525E-19	
F critical single tailed	1,47336709	



T-test: Two samples, different variances

	Inventory	*Demand*
Average	6865,378378	1497,851351
Variance	9984213,252	1026668,457
Observations	74	74
Hipotesis of difference in average	1	
df	88	
Stat t	13,91227431	
P(T<=t) single tailed	3,04333E-24	
t critical single tailed	1,662354029	
P(T<=t) dual tailed	6,08666E-24	
t critical dual tailed	1,987289865	



Visualization 10: F and paired T-tests; F test Rejected for equal variances, T-test rejected for equal averages, Cluster C.

Cluster C shows a very similar behavior to Cluster A. Variances cannot be considered equal and equality on averages is rejected.

As a summary, for cluster C, the tests made were:

Normality for Inventory (K-S): Not Rejected.

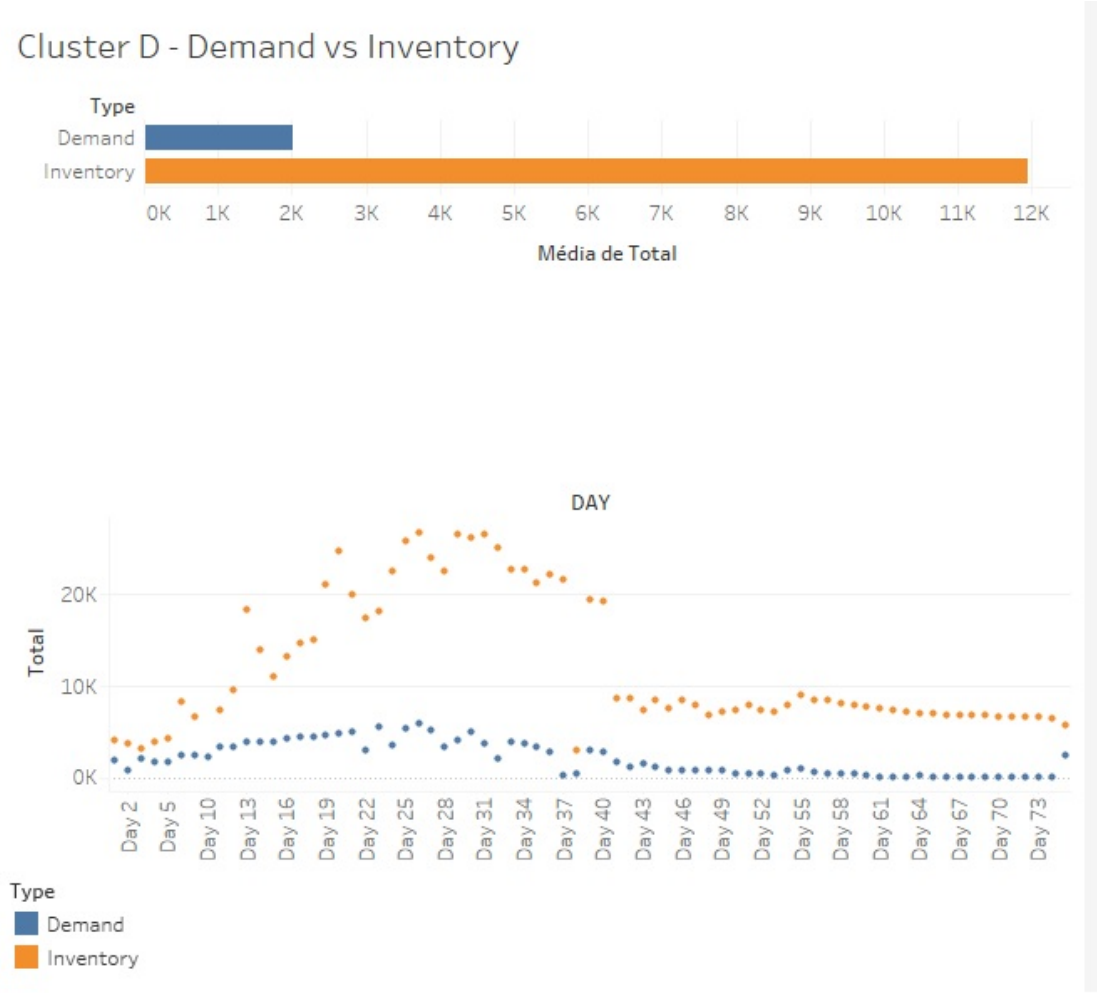
Normality for Demand (K-S): Not Rejected.

Equal Variances for Inventory x Demand (F-test): Rejected.

Equal Averages For Inventory x Demand (T-Test): Rejected.

Cluster D

Cluster D is the second cluster that has a failed test on normality, not being possible to compare through the F and paired T-tests. Through the same approach as cluster B, the inference will be made from the time series.



Visualization 11: Time Series and Averages Inventory vs Demand, Cluster D.

In a very similar behavior to other clusters, average inventory is much higher than demand. What is interesting in this series is that it is possible to see a trend upwards in inventory at day 13 all the way to day 40. After that, there seems to appear a new adjustment, that follows demand in behavior, but at also at much higher values.

Through this analysis of the time series, it doesn't seem correct to assume equality on averages, since inventory reaches six times the average demand.

Cluster E

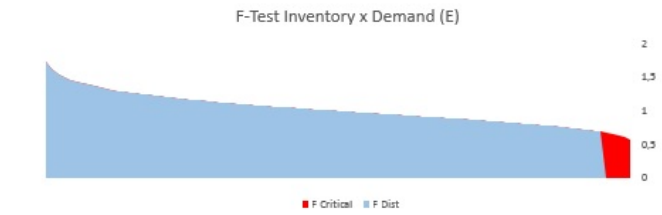
Cluster E

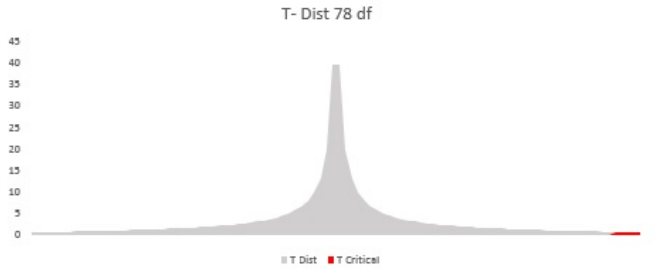
F-test: two samples for variances

	Inventory	*Demand*
Average	7312,5	862,7027027
Variance	11258392,31	364763,6638
Observations	74	74
df	73	73
F	30,86489534	
P(F<=f) single tailed	1,9677E-35	
F critical single tailed	1,47336709	

T-test: Two samples, different variances

Average	7312,5	862,7027027
Variance	11258392,31	364763,6638
Observations	74	74
Hipotesis of difference in average	1	
df	78	
Stat t	16,27168593	
P(T<=t) single tailed	4,31165E-27	
t critical single tailed	1,664624645	
P(T<=t) dual tailed	8,6233E-27	
t critical dual tailed	1,990847069	





Visualization 12: F and paired T-tests; F test Rejected for equal variances, T-test rejected for equal averages, Cluster E.

Just as on the other tests, cluster E shows extreme values for the F-Test and the paired T-test. There is no support for equality on averages.

As a summary, for cluster E, the tests made were:

Normality for Inventory (K-S): Not Rejected.

Normality for Demand (K-S): Not Rejected.

Equal Variances for Inventory x Demand (F-test): Rejected.

Equal Averages For Inventory x Demand (T-Test): Rejected.

Cluster F

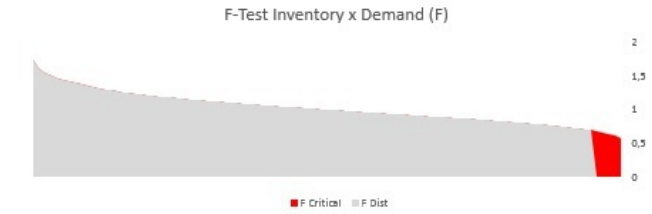
Cluster F

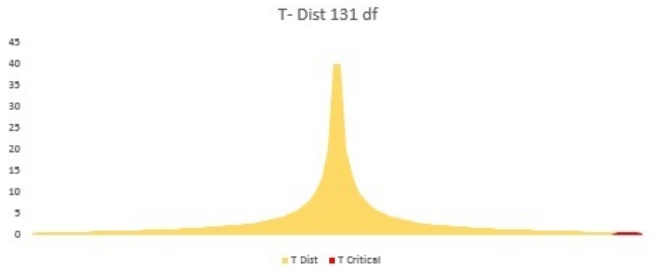
F-test: two samples for variances

	Inventory	*Demand*
Average	12134,59459	2417,121622
Variance	7749800,135	3820577,56
Observations	74	74
df	73	73
F	2,028436804	
P(F<=f) single tailed	0,001438005	
F critical single tailed	1,47336709	

T-test: Two samples, different variances

		TOTALS
Average	12134,59459	2417,121622
Variance	7749800,135	3820577,56
Observations	74	74
Hipotesis of difference in average	1	
df	131	
Stat t	24,57257894	
P(T<=t) single tailed	3,38936E-51	
t critical single tailed	1,656568649	
P(T<=t) dual tailed	6,77872E-51	
t critical dual tailed	1,978238539	





Visualization 13: F and paired T-tests; F test Rejected for equal variances, T-test rejected for equal averages, Cluster F.

The last cluster verified shows the same trend as the others: extreme values, rejection for both null hypothesis.

Thus, there seems to be no support for shortages or equality, given the extreme value located at the right tail of the distribution.

As a summary, for cluster F, the tests made were:

Normality for Inventory (K-S): Not Rejected.

Normality for Demand (K-S): Not Rejected.

Equal Variances for Inventory x Demand (F-test): Rejected.

Equal Averages For Inventory x Demand (T-Test): Rejected.

Conclusions

The Shortage Claim and its Possible Effects

At first glance, operational failures that require quick corrections may incur in some failure in judgement. The analysis shows that at the distribution side, it didn't actually happened with statistical significance.

Possibly, when the isolated shortage events appeared, there was an overly proportional shift in supply, changing parameters of surplus inventory for the sake of safety. This also means "money left on the table", due to unnecessary movements in cargo.

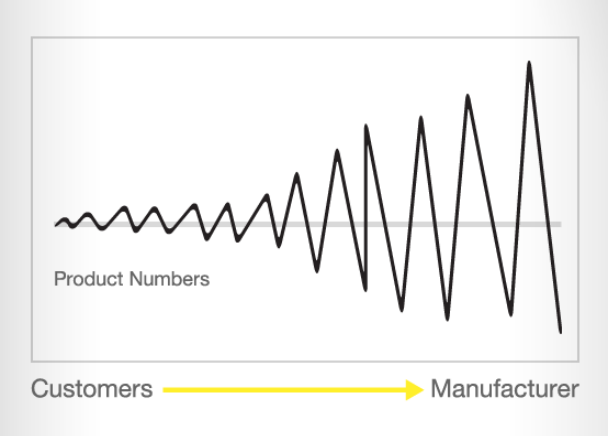
However, also on the safety side, the hypothesis testing reassures and tackles possible claims on shortages as being relevant to the bottom line of the project's profitability.

As a final note on shortages, there is the possibility to be explored when analyzing the PDRs individually for the shortages. This analysis does not make claims on the causes for that shortage since it narrows its scope to distribution; there could be other causes for that.

Also, what seems interesting to note is that shortages, interpreted as a failure or "traumatic event" for the operation, may turn decision-making from holistic to micromanagement; an unsupported claim, holistically speaking, generates a disproportional reaction. And by that, the other side of shortage appears: oversupply.

The Bullwhip Effect

The Bullwhip Effect, also known as the Forrester Effect, happens when customer demand shifts and there is a less than proportional response in inventory.



Visualization 14: Illustration of the Bullwhip Effect. Source: Wikipedia.

Even though the tests were not made directly to verify this phenomenon, their results seem to reveal the possibility of its occurrence. This “side-effect” of the analysis is a very meaningful topic to be explored further, in order to optimize the use of resources for the following cities where it will be implemented.

In this case, evidence, while not directed towards this subject, cannot determine with accuracy the causes of the significant differences in inventory relative to demand of products. As an alternative, demand could have been estimated to a much larger distribution than the actual demand observed. What can be properly affirmed is that demands are far lower than the actual inventory could distribute and that, as a whole, there had been an overestimation on the distribution side for the units.

The Kolmogorov - Smirnov Tests on Normality

The K-S model for normality, just as goodness of fit test tries to adjust a real observation has it flaws. Mainly, it is claimed that it is more relaxed than some other tests, such as Shapiro-Wilk or the Lilliefors correction for K-S.

However, the test selected presents simplicity and serves as confirmation of the normality assumption. For that, it sound reasonable to test the assumption under some strictness, however not being attached to testing if the distributions follow the empirical gaussian by the decimals.

Therefore, the test is considered good enough for the underlying assumptions, with the advantage that it is widely used and the results can be easily understood and confronted.

Error mitigation for T-Tests.

All of the T-tests for comparing the averages for demand and inventory came to the rejection of the null hypothesis: this means that there is no support to the assumption that average inventory and demand are equal.

In order to produce reliable results, the first filter was the Kolmogorov-Smirnov test itself: T-tests should only be performed in proven normal samples.

All p-values show very extreme numbers, giving confidence that the decisions of rejection for the null hypothesis are correct.The K-S model for normality, just as goodness of fit test tries to adjust a real observation has it flaws. Mainly, it is claimed that it is more relaxed than some other tests, such as Shapiro-Wilk or the Lilliefors correction for K-S.

Final Considerations

The concept of using statistics to prove “gut decisions” is a very powerful one. This does not mean that there all decisions must be data-oriented at all times.

However, as humans we tend to rely on our ancestral instincts and react on danger signs to a much larger extent than the real danger itself. In terms of business, that could cost money.

This basic concept of comparing how distribution plays a role in the operation compared to demand cold support decisions of adjustments in the following cities, providing a leaner, more profitable operation without cutting corners in quality. However, to support this, it is necessary to further expand the analysis with operational data and strategic planning.

This project, therefore, serves its purposes of presenting statistics to future decision making. However, other than directed to present only answers, the objective is to help asking better questions about how to improve services.