

Descriptive Statistic, Hypothesis Test with Sale Dataset/ SAS Studio

Didem Bulut Aykurt

MIS540-1 – Introduction to Business Intelligence

Colorado State University-Global Campus

Dr. Alin Tomoiaga

October 2, 2022

Descriptive Statistics

Descriptive analytics helps to know more about raw data using charts, graphs, and tables.

Additionally, descriptive statistics is a part of statistics to describe data. Two ways to measure one is a measure of central tendency helps to find the central position of data as mean, median, and mode. Another one is that the actions of dispersion help to show how-to spread-out data, such as range, standard deviation, quartiles, and scattering, which are all used to find the distribution of data.

Before starting, as descriptive analysts, we should look at the data types and data sets that contain information. I used the Sales.xlsx dataset, which includes 20-month sales and has five variables.

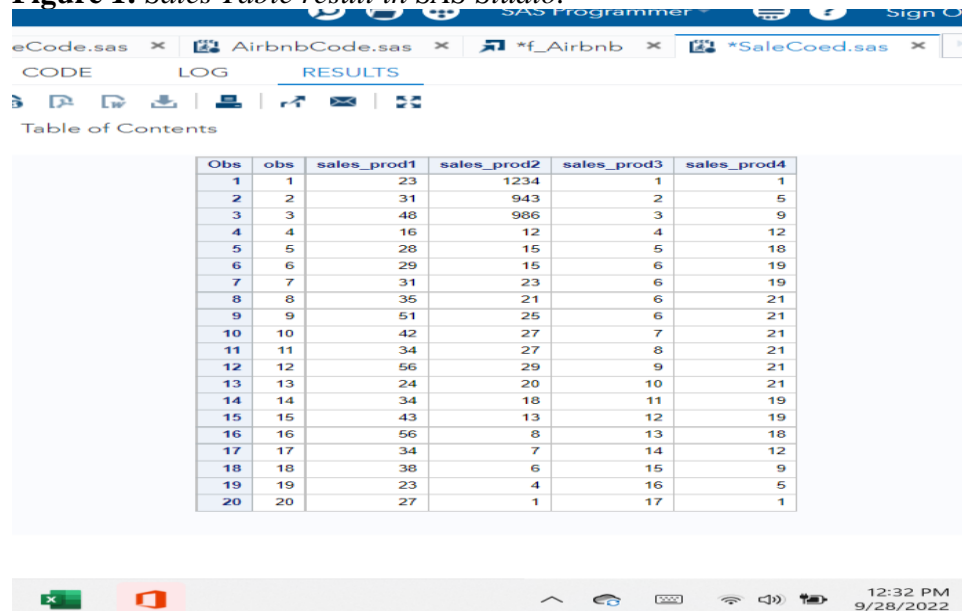
Obs is the observation number.

Sales_prod1 and sales_prod2 are sales of products x and y. Those numerical data are continuous.

Sales_prod3 contains each salesperson's number; this categorical data is ordinal.

Sales_prod4 has the location of the sale. This is also categorical data is ordinal.

Figure 1: Sales Table result in SAS Studio.



Obs	obs	sales_prod1	sales_prod2	sales_prod3	sales_prod4
1	1	23	1234	1	1
2	2	31	943	2	5
3	3	48	986	3	9
4	4	16	12	4	12
5	5	28	15	5	18
6	6	29	15	6	19
7	7	31	23	6	19
8	8	35	21	6	21
9	9	51	25	6	21
10	10	42	27	7	21
11	11	34	27	8	21
12	12	56	29	9	21
13	13	24	20	10	21
14	14	34	18	11	19
15	15	43	13	12	19
16	16	56	8	13	18
17	17	34	7	14	12
18	18	38	6	15	9
19	19	23	4	16	5
20	20	27	1	17	1

Now, I create a business question for what the company is trying to understand whether two products are selling accurately. And study a few critical factors that help to see for business very profitable, such as:

1. What is the highest percentage of sales quantity range for products X and Y?
2. What is the relation between product X and product Y sales?

Next, create the null and alternative hypotheses for each business question as follows:

- The first business problem hypotheses are as follows:
- **Null hypothesis:** The highest percentage of sale quantity range is 40 to 45.
- **Alternative hypothesis:** The highest percentage of sales quantity is not 40 to 50.

Figure 3 on the distribution of sales_prod1 graph shows the highest percentage range is 30 to 40 additional goodness of fit test with a p-value of all three results higher than 0.05, which means this sample dataset is not genuinely representative of the population. Thus, we can't reject the null hypothesis. Also, we can't accept the alternative hypothesis that might add more variables that help to improve the result.

Null hypothesis: Product Y's highest sales quantity percentage significantly differs from 10 to 100.

Alternative hypothesis: Product Y's highest percentage of sales quantity range is 10 to 100.

Figure 4 on the distribution of sales_prod2 graph shows the highest percentage of sales quantity lower than 100. Figure 2 on the measure of dispersion table has an upper quartile of 27 in 75% of all variables that fall below the point of 27. That also tells us that the most variable under the third quartile is 27. If the goodness of fit test three p-value results lower than 0.05, we can say that the sample data is representative of the population. We can reject null hypotheses.

- The second business problem hypothesis is as follows:

Null hypothesis: there is no relationship between product X and Y.

Alternative hypothesis: There is some statistical significance between product X and product Y.

Figure 5 on the matrix scatter plot graph shows a nonrelation between two variables, so the correlation table shows a -6.8% negative relation. The result was close to zero but not zero. Thus, we can't reject the null hypothesis and not accept the alternative view that might add more variables and get a better result. Twenty observations are too low a number to take sample data to decide.

All the hypotheses are framed according to the business question, and the test used for all the ideas is a summary and descriptive statistics. The dataset has two continuous and two categorical variables. I used summary and descriptive statistics analysts as two continuous variables.

Figure 2: *The Sale dataset's summary statistics in SAS Studio.*

9/28/22, 12:34 PM

Results: SaleCoed.sas

Statistic of Measure of Center

The MEANS Procedure

Variable	Mean	Median	Mode	N
sales_prod1	35.1500000	34.0000000	34.0000000	20
sales_prod2	171.7000000	19.0000000	15.0000000	20

Statistic of Measure of Dispersion

The MEANS Procedure

Variable	Range	Std Dev	Skewness	Lower Quartile	Upper Quartile	Minimum	Maximum	N
sales_prod1	40.0000000	11.1980966	0.4904537	27.5000000	42.5000000	16.0000000	56.0000000	20
sales_prod2	1233.00	383.8948732	2.2045044	10.0000000	27.0000000	1.0000000	1234.00	20

Descriptive statistics can apply continuous variables as Products x, and y are continuous variables in the Sales dataset. Let's look at each measurement result.

Statistic of Measure of Center.

Mean helps to find the average sales quantity of product X is 35.15 and product Y 171.7, which means product y sells more than product X, not really because product Y has an outlier very far from the mean that might affect the standard.

The median shows the middle value of product X's data point half variables lower than 34 and a half values greater than 34, and product Y is 19 half product sales point lower than 19 and higher than 19. Product Y's median number is far from the mean, which should be skewed.

Mode is one number that is repeated more than once the occurring number of product X is 34, and for product Y, often the sale quantity is 15 as we can see product sales data point lower number than product X.

Statistic of Measure of Variation.

The range shows the difference between the highest and lowest values. As we can see, product Y has the highest content, 1233, which tells how to spread product Y's data point or quantity of sales far from lowest to highest. Additionally, there might be some issues with lower-sell locations. Product X's range is 40, which is approving to see the range data point.

Std Dev shows the distance between each data point and the mean point. Product X's mean is 11.198 from the data point; for product Y's std, it is 383.895 from the data point, which is a high number far from the mean value.

Skewness helps to see how data distribute that positive result positively skewed or skewed right; a negative impact is negatively skewed or skewed left as product X's 0.49 and product Y's 2.20 both products is positively skewed.

Q1/First quartile shows the number halfway between the middle and lowest numbers and 25% of data split probability distribution as data points are above the value. Product X's lowest quartile is 27, and product Y's 10.

Q3/Third quartile helps find the number halfway between the highest and median numbers, and 75% of all values fall below the third quartile. Product X's upper quartile is 42.5, and product Y's is 27.

Minimum is the lowest data point that product X's 16 and product Y's 1. Probably, product Y has an outlier close to the zero point.

Maximum shows the highest point in the dataset as product X's 56 and product Y's 1234. It might be that product Y has an outlier highest per data point.

Analyzing the distribution of temp 'sales_prod1' and 'sales_prod2' variables

Figure 3: The resulting distribution analysis is a histogram of the ‘sales_prod1’ variable on SAS Studio.

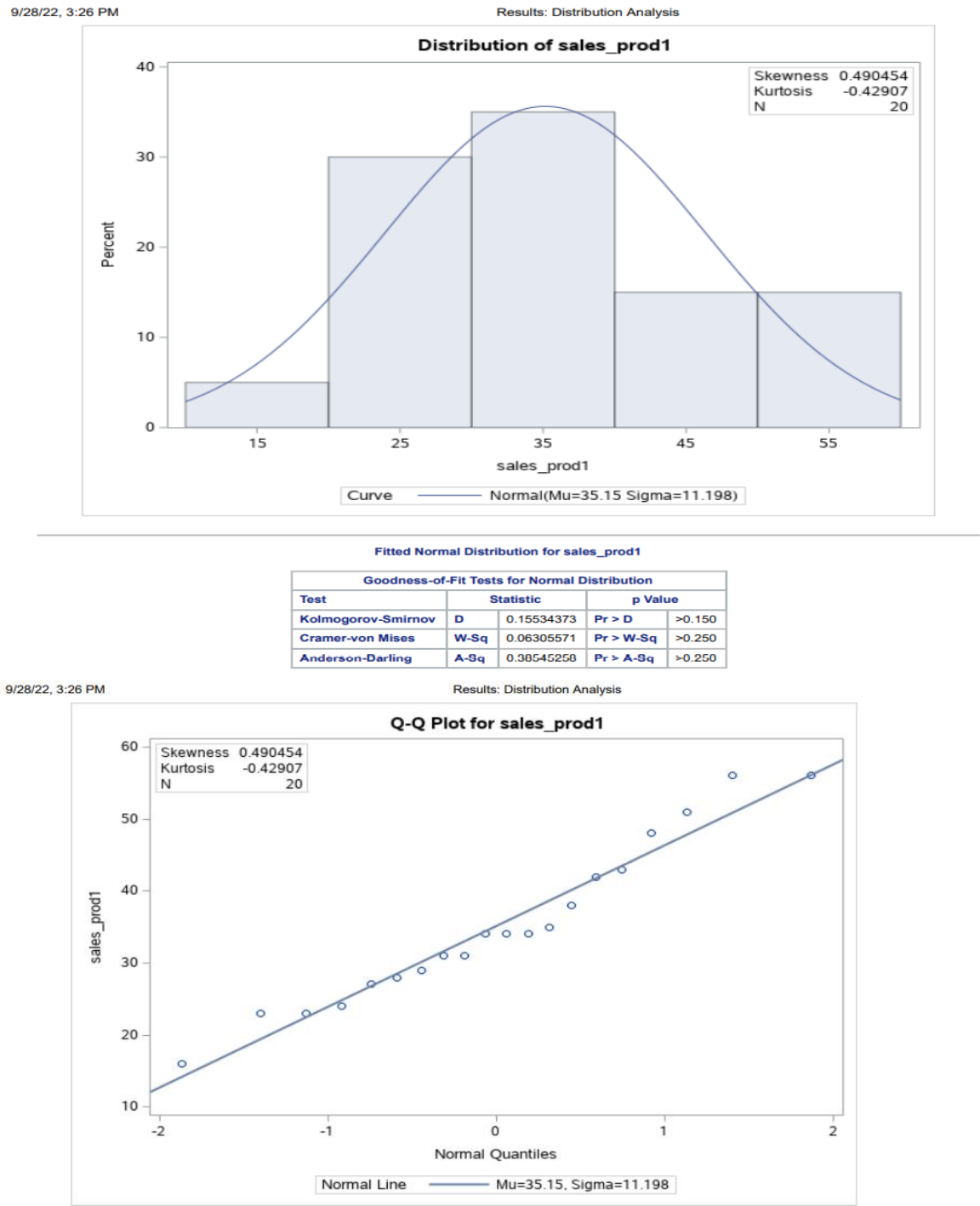
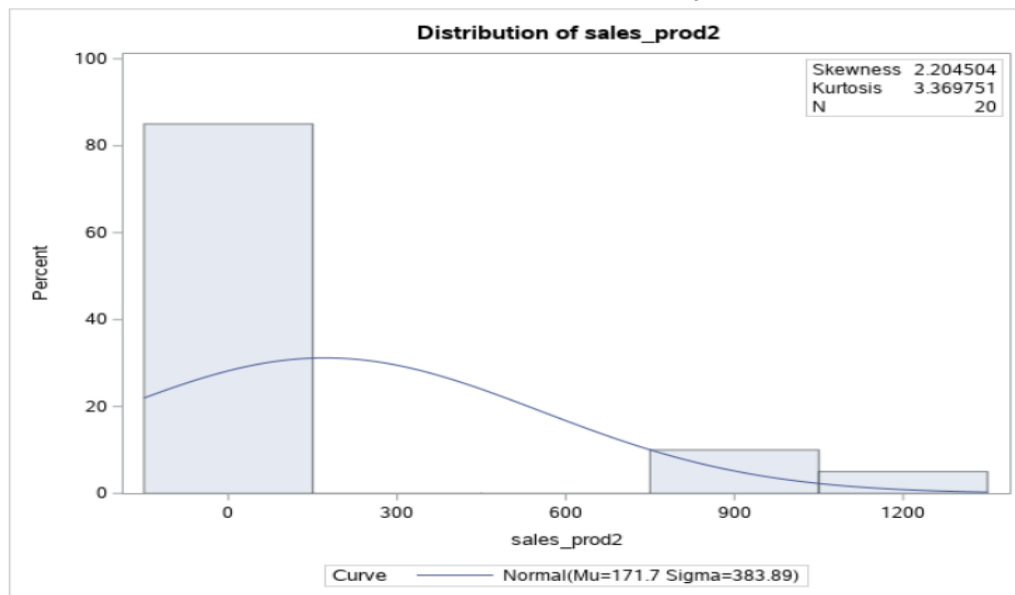


Figure 4: The resulting distribution analysis is a histogram of the 'sales_prod2' variable on SAS Studio.

9/28/22, 3:31 PM

Results: Distribution Analysis

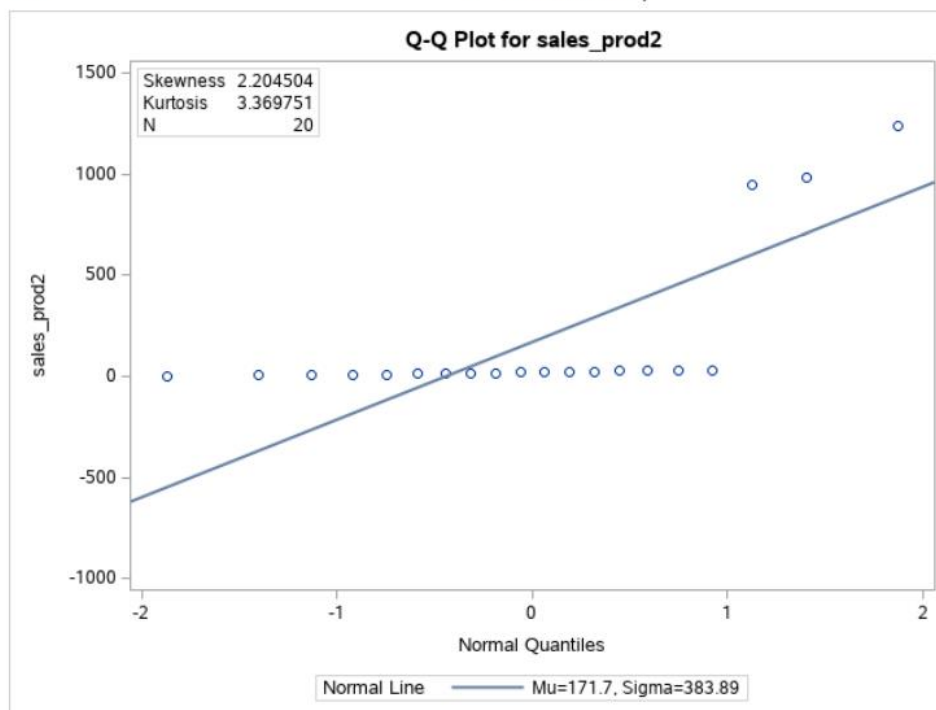


Fitted Normal Distribution for sales_prod2

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.49494797	Pr > D	<0.010
Cramer-von Mises	W-Sq	1.08794975	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	5.31906172	Pr > A-Sq	<0.005

9/28/22, 3:31 PM

Results: Distribution Analysis



We discussed **Skewness** in summary statistics as it tells whether a dataset has an asymmetric distribution that measures three different distributions. One zero skew means the distribution is symmetrical. Others negative skew when the number is negative and positively skew when the number is positive. Product X's skewness shows a positive number of 0.42, which tells the right skew, and product Y has a positive result of 2.20, which is a pretty high right skew.

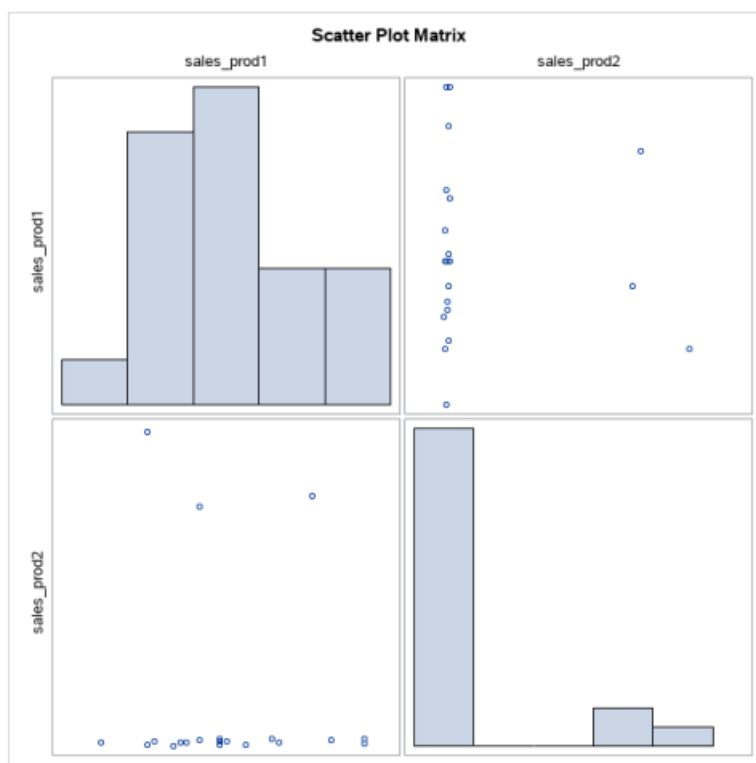
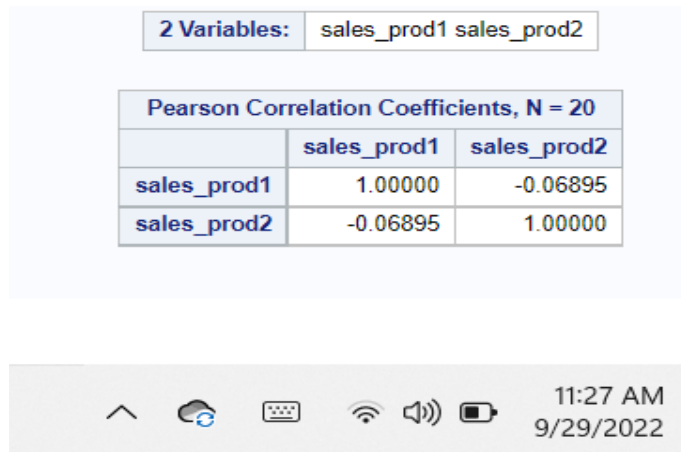
Kurtosis shows the variable's probability or frequency, which also helps to compare which variable has a heavy distribution tail with three kurtosis types. The result of zeros for the normal distribution of kurtosis. Medium tails are **mesokurtic(kurtosis=0)**, low kurtosis is **platykurtic(kurtosis<0)**, and high kurtosis is **leptokurtic(kurtosis>0)**. The product X variable has platykurtic because the product X's kurtosis is -0.42 as negative kurtosis lower than zero and effect Y's positive 3.36 that leptokurtic higher than zero as a normal distribution kurtosis's number is zero or a close zero. Thus, both product distributions might have outliers; the product X skewness result is close to zero, which can be accepted for normal distribution if outliers exceed what we expect.

A **goodness-of-fit test** is how well the sample data set fits an entire population with normal distribution and another way to say how target values are related to the independent values in a model. Kolmogorov-Smirnov test applies a large sample of over 2000. The other two tests also have the same reason to use, which helps to know whether the example of the normal distribution. Product X's three p-values are higher than 0.05, which means the data set is not statistically significant. Product Y's results are lower than 0.05, which means they are statistically significant.

The Q – Q plot shows how data points fall on a straight line. We discussed that the previous kurtosis result shows the dataset might have outliers, and the Q-Q chart proves it. Product X has outliers data points close to the line, and product Y's data does not fit straight.

Correlation Analysis

Figure 5: *The result of correlation analysis in SAS Studio.*



The first table shows the negative number and the scatter plot shows nonrelation because the first table result is -0.068, a pretty small number, which means 6.8% negative relation as we can say nonrelation between product X and Y, as a result, is close to zero.

Conclusion

I used the sample data set to create descriptive analytics and hypothesis testing with the Sales.xlsx data set, which has 20 intakes representing aggregated data and five attributions. I made descriptive statistics with continuous variables sales_prod1(product X) and sales_prod2(product Y). Product X sample data set did not fit a distribution because of the p-value higher than 0.05 on the goodness-of-fit test table. Product Y's means more than product X's mean, as discussed above in the mean explanation; however, we can't say product Y sells better than product X because the outlier affects product Y's mean. Max sales of portion for product X 56 and product Y's 1234. Both products have a positive skew. There is a -6.8% correlation between the two variables.

Reference

Sharda, R., Delen, D., & Turban, E. (2022). Business intelligence, analytics, and data science: A managerial perspective (4th ed.). Pearson.

Elliott, A. C., & Woodward, W. A. (2023). SAS Essentials: Mastering SAS for data analytics (3rd ed.). John Wiley & Sons.