MATH & STATISTICS FOR DATA ANALYSIS

Assignment II – Inferential Statistics.

Group Assignment

Daniel Bilitewski, Guillermo Chacón, Nisrine Ferahi, Jonas Hellevang, Mariana Narvaez, Rabiga Shangereyeva

December 1st 2019

Introduction

This report will cover the results of a survey taken to IE Master students about pizza and hamburger preferences and the following statistical analysis of those results. The objective of this report is to show how statistical analysis helps us achieve pertinent conclusions even in simple surveys like the one done for this case.

The Survey:

To show the power of statistical tools, we conducted a survey that contained 2 mean type of questions and 2 proportion type of questions. The questions chosen for this assignment are the following:

Q1: How Many times have you eaten Pizza this Year? (how many do you think).
Q2: How Many times have you eaten hamburgers this year? (how many times do you think)
Q3: Do you like Pineapple in your pizza?
Q4: Do you prefer Hamburgers over Pizza?

Even thought that Q's 1 and 2 are subjective questions, we believe that understanding the perception of the people of these 2 questions and comparing them to objective questions like 3 and 4 will make the analysis more in depth and interesting enough.

In order to take the survey, we used Google drive function, you can find the full survey in the following link:

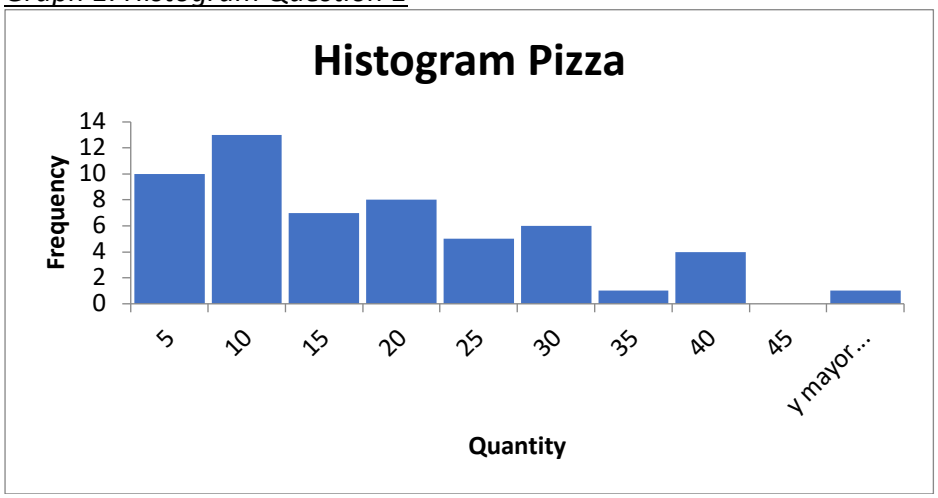
https://docs.google.com/forms/d/e/1FAIpQLScYNMT6aoSGsAPuQouB6vUTOXjtMvY8FOHxgfX22QOJhDaJaQ/viewform?usp=sf_link

Descriptive Analysis:

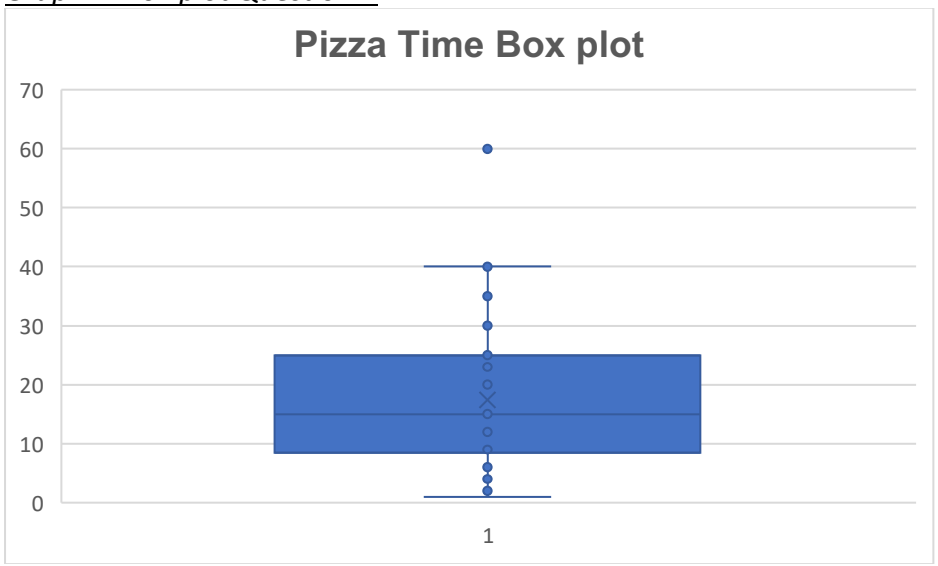
Question 1: How Many times have you eaten Pizza this Year? (how many do you think).

A Total of 56 people answered the survey giving us enough sample size to do statistical analysis. 2 outliers (200 and 7000) were taken out of the data as they were far to big compared to the rest of the data, leaving only 54 answers to the survey.
The following graph 2 shows the histogram of the results without the outliers:

Graph 1: Histogram Question 1



Graph 2: Box plot Question 1



And, the descriptive analysis:

Table 1: Descriptive analysis with and without outliers

Pizza without Outliers	
Mean	17,44
Standard Error	1,71
Median	15,00
Mode	10,00
Standard Deviation	12,57
Variance	157,91
Kurtosis	1,10
Asymmetry	1,04
Range	59,00
Min	1,00
Max	60,00
Sum	942,00
Count	54,00
Q1	8,5
Q2	15
Q3	25
1.5 IQ	25

Pizza with Outliers	
Mean	145,39
Standard Error	124,68
Median	15,00
Mode	10,00
Standard Deviation	933,04
Variance	870563,88
Kurtosis	55,90
Asymmetry	7,47
Range	6999,00
Min	1,00
Max	7000,00
Sum	8142,00
Count	56,00
Q1	9,25
Q2	15,00
Q3	28,75
1.5 IQ	29,00

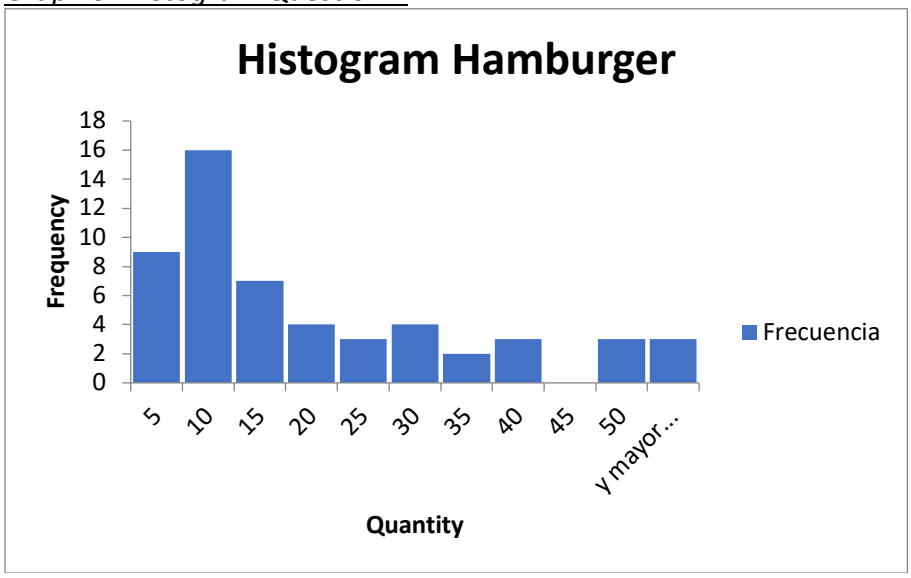
As you can see the data shows high kurtosis and a 5 IQ range of 160, hence sustaining the decision of disregarding outliers (200 and 7000) (1)

Question 2: How Many times have you eaten Hamburgers this Year? (how many do you think).

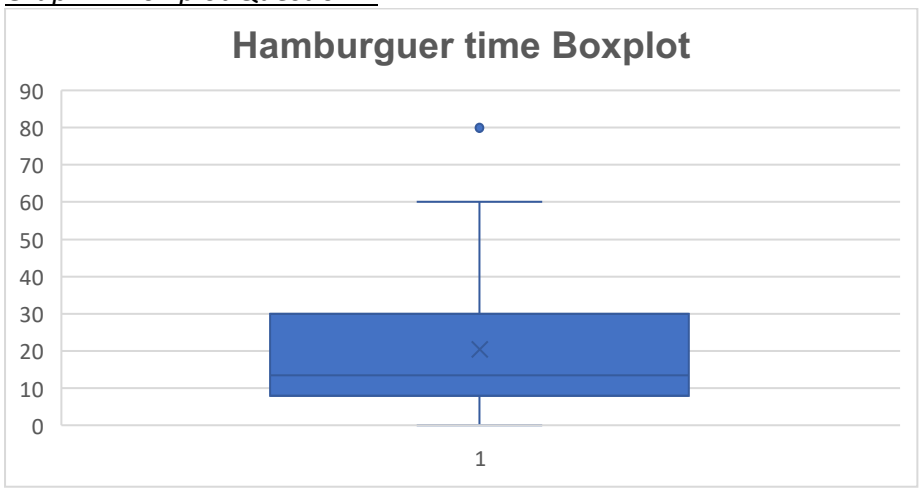
As in Q1, a total of 56 people answered the survey giving us enough sample size to do statistical analysis. Also, like Q1 2 outliers (200 and 7000, coming from the same persons) were taken out of the data as they were far to big compared to the rest of the data, leaving only 54 answers to the survey.

The following graph 2 shows the histogram of the results without the outliers:

Graph 3: Histogram Question 1



Graph 4: Box plot Question 1



(1) Even considering that the 1.5 IQ shows that the upper limit for outliers is 40, we decide to use 5 IQ as the limit in order to not bias our results.

The descriptive analysis is the following:

Table 2: Descriptive analysis with and without outliers

Hamburger without Outliers	
Mean	20,43
Standard Error	2,53
Median	13,50
Mode	10,00
Standard Deviation	18,60
Variance	345,87
Kurtosis	2,41
Asymmetry	1,58
Range	80,00
Min	0,00
Max	80,00
Sum	1103,00
Count	54,00
Q1	8
Q2	13,5
Q3	30
1,5 IQ	33

Hamburger with Outliers	
Mean	148,27
Standard Error	124,64
Median	15,00
Mode	10,00
Standard Deviation	932,74
Variance	869996,24
Kurtosis	55,88
Asymmetry	7,47
Range	7000,00
Min	0,00
Max	7000,00
Sum	8303,00
Count	56,00
Q1	8
Q2	15
Q3	30
1,5 IQ	33

Again, as you can see the data shows high kurtosis and a 5IQ range of 185 more or less, hence sustaining the decision of disregarding outliers (200 and 7000) (1)

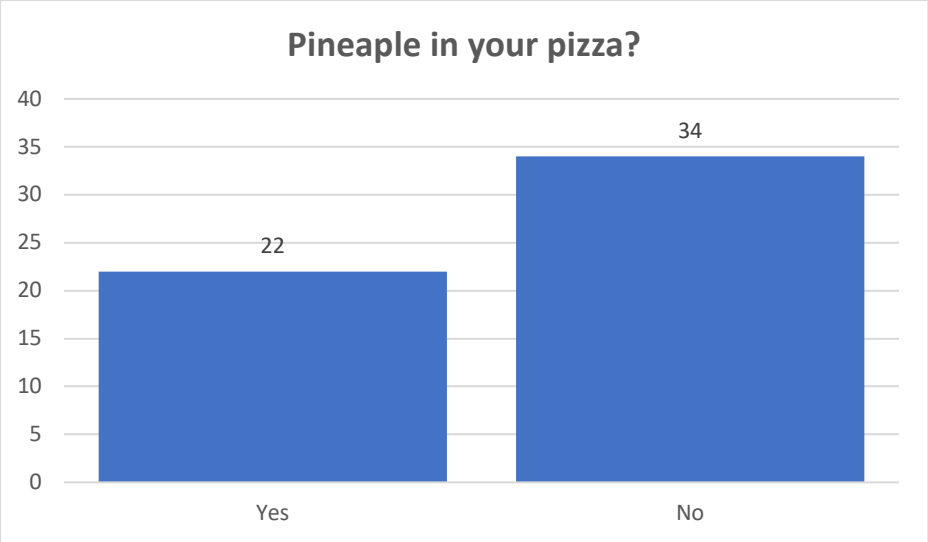
(1) Even considering that the 1.5 IQ shows that the upper limit for outliers is 48, we decide use 5IQ as limit in order to not bias our results.

Question 3: Do you like Pineapple in your pizza?

As Q1 and Q2 56 people answered Q3, in this case no values were filtered due to the nature of the question.

The bar plot of the answers is the following:

Graph 4: Bar plot Question 3.



	Count	Perc
Yes	22	39,29%
No	34	60,71%

For this question is very important to determine if the sample size is enough to assume normality according to central limit theorem, for this we use the following formula:

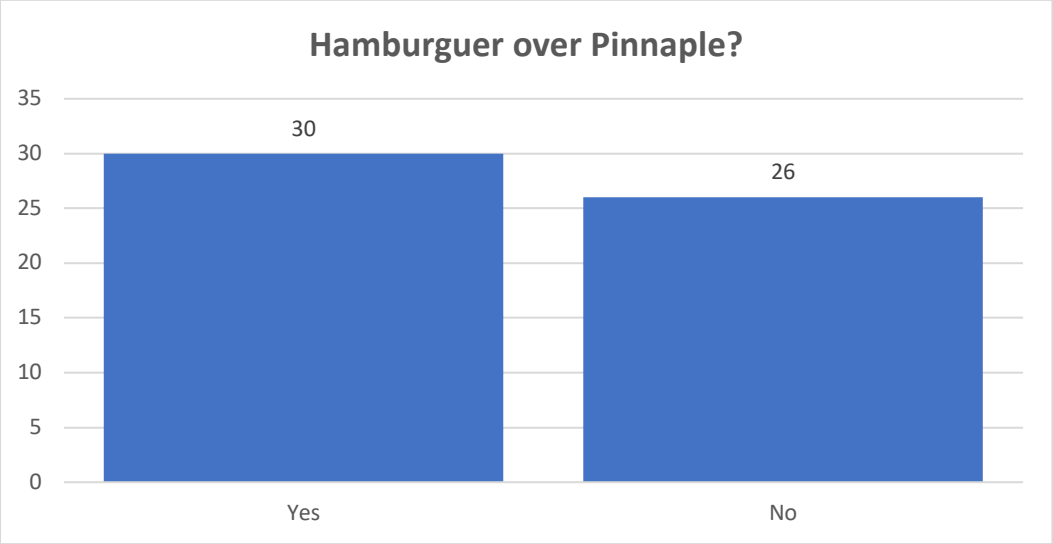
$$n * p * (1 - p) > 5$$

Doing the calculation for this question lead to the value of 13,36, this value allows us to assume normality in the results.

Question 4: Do you prefer hamburger over pizza?

Finally, like Q3 we took the full scope of the survey (56 sample size). The following bar plot shows the results:

Graph 5: Bar plot Question 4.



	Count	Perc
Yes	30	53,57%
No	26	46,43%

To ensure that we can assume normality we use the same equation than Q3, leading to a result of 13,93, so again we can assume normality in the distribution of this question.

Inferential Analysis:

Description:

To make the right Inferential analysis first we need to understand the nature of our data for us to make the right decision about the inferential method:

For Questions 1 and 2 we will make Mean inferential analysis assuming that the standard deviation of the population is unknown, this is because our population (IE master Students) is unknown for us. This leads us to a Student's t analysis with 53 degrees of freedom (54 surveys without outliers)

For Questions 3 and 4 we will conduct proportion inferential analysis using Z table, for the calculations for Mean and Standard deviation we will use the following formulas:

$$E(\hat{p}) = P$$

$$\sigma_{\hat{p}} = \sqrt{\frac{P(1-P)}{n}}$$

Given that this data has no outliers we will use the full 56 survey results.

Results:

Q1 and Q2:

Table 3: Inferential Statistics Questions 1 and 2

Q1: Pizza per Year				
Confidence Level	T Value (α/2)	Lower Interval	Mean	Upper Interval
90%	1,67	14,58	17,44	20,31
95%	2,01	14,01	17,44	20,87
99%	2,67	12,88	17,44	22,01
Q2: Hamburger per Year				
Confidence Level	T Value (α/2)	Lower Interval	Mean	Upper Interval
90%	1,67	16,19	20,43	24,66
95%	2,01	15,35	20,43	25,50
99%	2,67	13,66	20,43	27,19

Q3 and Q4:

	Question 3	Question 4
Media Proportion	39,29%	53,57%
Standard dev Proportion	6,53%	6,66%

Table 4: Inferential Statistics Questions 3 and 4

Q3: Pineapple in your pizza?				
Confidence Level	Z Value (α/2)	Lower Interval	Mean	Upper Interval
90%	1,64	28,55%	39,29%	50,02%
95%	1,96	26,49%	39,29%	52,08%
99%	2,58	22,48%	39,29%	56,10%
Q4: Hamburger over Pizzas?				
Confidence Level	Z Value (α/2)	Lower Interval	Mean	Upper Interval
90%	1,64	42,61%	53,57%	64,53%
95%	1,96	40,51%	53,57%	66,63%
99%	2,58	36,40%	53,57%	70,74%

Analyzing the results:

From the results of the inferential analysis we can state the following using the 95% confidence level:

- Q1: We are 95% sure that people think that on average they have eaten between 14,01 and 20,87 pizzas this year.
- Q2: We are 95% sure that people think that on average they have eaten between 15,35 and 25,05 hamburgers this year.
- We cannot assure with these 2 questions that there is a significant statistical difference between pizzas and hamburgers consumption without doing hypothesis testing
- Q3: We know with a 95% confidence level that between 26,49% and 52,08% of our population like pineapple in their pizzas.
- Q4: We know with a 95% confidence level that between 40,51% and 66,63% of our population prefer Hamburger over pizza.

On further analysis it is interesting analyzing how Q1, Q2, Q4 results connect themselves:

We know from the inferential analysis that there is a tendency to eat more hamburgers than pizzas, but we don´t know if this difference is statistically significant (we would need a hypothesis testing for that). The interesting fact is that Q4 proves what we said before, as it's 95% confidence interval is between 40,51% and 66,63% we cannot ensure that there is a preference over pizza or hamburger because if the real value is 45% for example, people prefer pizza, but If it's 60%, people prefer hamburgers. If the confidence interval would've been fully over or under the 50% mark, we could've concluded with a 95% confidence level that people preferred hamburgers over pizzas.

Finally analyzing Q3 as itself provides some interesting results, as pineapple is regarded as a controversial ingredient in pizzas, one would've expected that a low proportion of the population liked it. Actually, the results shows that we can´t conclude that people don't like pineapple in their pizzas, as the interval with a 95% confidence interval is between 26,49% and 52,08%, there is a possibility that more than 50% of the population likes pineapple in their pizzas.