quantItative fluency

Jeffery Morse

Brandman University

Author Note

For correspondence concerning this paper, Jeffery Morse can be reached at

jmorse@mail.brandman.edu

quantItative fluency Final

Primary Data Analysis



Table 1

I have selected study number one from the Appendix: Case Studies for Summative Assessment; Case Study #1 Flight Delays.

“As an advisor to the small regional airport, you have been asked to look into the relationship between the delay times for flights and the length of the flight overall (measured in miles). The local division of the Federal Aviation Administration (FAA) has supplied you with a simple random sample of 20 recent flight delays for your analysis. The division would also like to know if the average flight delay times for this year are higher or lower than last year’s average flight delay time of 42 minutes. “

This is an observational study, as these flights have already occurred. Further, the data is from a random sample (a parameter that is given in the study) of a population of flights. The population is all flights ever, of course, this is nearly impossible to attain in reality unless we put a time period parameter on it. For instance, all the flights in one year; this is not given so we will not have P (population). What is given is the sample number (n) which is equal to 20. Both flight length and flight delay are quantitative and discrete data. We will not be looking at any flights that are not complete. We will not be looking at any half flights or quarter flights only fully completed flights. As a result, there is a distinct gap between each whole number flight count. For example there is not number between flight 1 and flight 2 (no flight 1.5). Also, each flight is Independent and mutually exclusive. Each flight is one flight, and each flight only can have one plane. Listed on table 2 the highlighted row is the total and the last row of the first two columns is the average or mean using equation. The sample mean of a frequency is . The sample size is less than the needed amount for being able to assume a normal distribution. The central limit theorem states the sample must be greater than 30 or 5% of the population, or the sample must be stated to be a normal distribution. This has neither, (n) is not greater than 30 nor is it stated to be a normal distribution. Because of this at this point normally we would normally say that we can make no valid assumptions. However, for this paper we will have to assume that this is a normal distribution in order to do any hypothesis testing further on. By assuming this, it will allow us to use the standard normal distribution. It also allows us to assume that (µ = µx̄), and (σ = σx̄). If we did not assume this we could go no further at that point. The independent variable(x) is the Length of Flights as this is the variable that is manipulated or controlled, and the dependent variable is the Flight Delay, as this is the measurement we are measuring. The level of measurement is a ratio, as zero means zero, and the difference between two measurements has significance. To clarify, if a flight has 0 miles in flight, then the flight = 0 and does not exist. Also, if the dependent variable (flight delay) has a value of 0 then time for delay = 0 and there is no delay. We are missing (σ) standard deviation for (µ) population mean, and (S) standard deviation for (x̄) x-bar sample mean. We will have to estimate (S). We further will be using t-score test statistic as a result of the (σ) standard deviation for (µ) being unknown, and sample size being less than 30. The parameter of interest is the length of flight in correlation with delay times. As a result, we will be looking at both.



Sum Total

Mean.

Flight Time

Delay Time

Table

Examination of Descriptive Statistics:

The next step is to examine the data that are provided in the case study. First off, we need to make sure that there are no outliers. Outliers are data outside the “outlier boundaries”. This data would cause a greater deviation away from the true mean of our data. To do this we will make a box and whisker chart. The low-class limit is 190, and the upper-class limit is 1400. Those two areas are easily read from the data. Next, we need to find the median or midpoint. We must first sort the data in order from smallest to greatest as seen in table 2. Since there are 20 classes the median will be the average of classes 10 and 11 which have values 820 and 850. We can use the equation for mean as that is just the average. = 835. We then find the median of Q1 and Q2 the same way. = 470 = 955. So LL = 190, Q1 =470, Q2 or Median= 835, Q3= 955, and UL=1400. We need to next find the IQR (inner quartile range) by subtracting this from our LL, and adding it to our UL we will see if any values of X are considered outliers. Q3-Q1=IQR = 955-470 = 480. Subtracting IQR from Q1 = 470-480 = -10. Adding IQR to our Q3 = IQR+Q3 = 955+480 = 1435. Our outlier range boundaries are -10 thru 1435, as we see all our values for X are well within this, so there are no outliers. Although this range is very wide, the fact that outliers are not present means our data may be more accurate. Our other variable will be addressed in the pages that follow.



730

752.5

Sample Mean:

The mean of a sample or x-bar is what we use here as µ is mean for a population. However, since we assumed (µ = µx̄), and (σ = σx̄), then we say that x̄ is an acceptable estimate for µ and since the population mean ==  would be 752.5 is an acceptable estimation. Please refer to column 1 of table 2. Notice that the x-bar value is almost at Q3 and is over 80 points below the median. Since the mean is less than the median we now know this data distribution is not normal but skewed to the right.

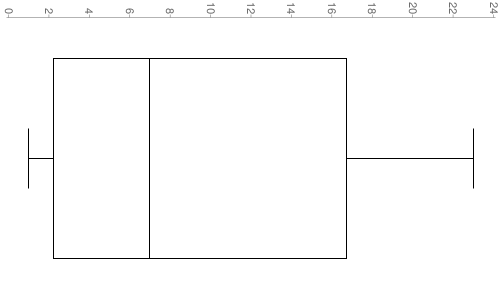
Mode:

The mode is 950 as it is the only value that is entered twice. This means we have a single mode. Commonly distributions with a single mode are often skewed. This is the case for out data distributions. This skew is the result of low sample size, and the presence of the mode at 950. We will see later if this is the case in our distributions. Expectedly, the mode chart here will be something other than a bell curve as this is not a distribution. This chart is listed to highlight the mode, shown in red, with a flat point at 950.

Let us also look at flight delay. We may find outliers on the other side. The reasons for finding outliers are the same as stated previously. The low-class limit is 1, and the upper-class limit is 23. Next we need the median. This means we need to align the data in ascending order according to flight delay in table 3. We can use the equation for mean as that is just the average. = 7. Next, we will plot our Box and Whiskers Chart. Q1 = = 3.5. Q3 = = 10.5. Please also note the these are multi-modal with duplicates at 2, 3,6, 7,9, and 11. The standard mean using  is listed at the bottle of table 2 column 2 and is 7.75. Using the same process as about our IQR is 14.5. Q1 – 14.5 = -11 and Q3 + 14.5 = 25. So our outlier range is -11 thru 25. There are no outliers.



Table 3



3.5

14.5

7

7.75

8.4

Standard Deviation and Variance:

We found x̄ = 752.5, and the variance for a sample is or as *Df= n-1* . Further, the standard deviation is given by . = = 28.092. If we take the square root of that we have 5.300199 this is our Standard Deviation or “S” (appendix 1).

This is right on track as Σ(x-x̄) equals zero according to the table in Appendix 1 zero.

Frequency Sample Mean:

The X̄-bar listed in the frequency distribution x̄f = 730, and x̄d = 8.4 is for frequency sample mean using, we will not be using, as we have determined x̄ by other equations, and is provided just for information purposes only. Albeit, it is nice to see that they are in the same neighborhood.

Flight Time



Delay Time



However, though the distributions are “bell like”, we can defiantly see skew. We were expecting to see skew and both are in fact skewed. What is unexpected is that fact that Delay time seems to be skewed left, while flight time seems to skew right. One probable reason is the small sample size. Since n = 20 is less than the required > 30 or at least 5% of the population (N) our data is skewed heavily. If N = 400 it would be more accurate, but since there is no time constrain we are talking about infinity number of flights. We would have to assume a time constraint of hours for N=400 of which 20 would be 5%. As a result, we must assume that we are looking at a tail. This is possible given the large number of flights compared to our sample size.

Description of Sample Data:

Referring to Table 2. We used the X vs. f column to create the following dot plot. We can see that the data can be centered around a line on the graph. A graph like this in statistics is referred to a regression line. By this we can conclude that a regression line with a positive linear association can be placed. In fact, we can actually find the slope of the line. First let’s find our y intercept using b= ȳ - mx̄. From table 2 ȳ = 7.75 and x̄ = 752.5, further, from plot and regression 2 excel gave us m = 0.0141. b= 7.75-10.61, b= -2.86. Our regression line has the following equation Ŷ = 0.0141X – 2.86. In a regression line, the slope is the heart and soul of the equation because its proportion reveals how much one can expect*Y*to change as*X*increases. The y-intercept or “b” sometimes may reveal stuff as well but through the slope we learn we have a constant rate of change. Why? Since we can draw a regression line then we can say that this measurement is proportional. This means we can expect longer delay times as flight time increases. Also, this proportion is further evidence of a skewed distribution. For such skews we might consider using χ 2 (chi squared) distribution. χ 2 statistics is available in appendix 5 for this reason. χ 2 was not introduced in the competency material so we will be staying with standard T-score statistics.

In order to even look at most of this data we have assumed that this is a normal distribution. Remember we are assuming our sample is < 30 or at least 5%, which we know it is not, to assume that (µ = µx̄), and (σ = σx̄), for acceptable estimates. So if (σ = σx̄), this means (σ = S), which would make our standard deviation for delay time (Sd) = 5.300199. Further, since there seem to be no outliers in either sets of data, we have been assuming (σ = S), we will accept 5.300199 number for (S) and (x̄)= 7.75 for all further calculations. For our flight time the numbers are listed in appendix 1 on the right. (Sf) = 334.95 and (x̄) = 752.5.

Examination of Inferential Statistics:

Here is where we begin to check requirements. First, the text does say that the sample was random. Next, we assume that this random sampling was not biased in any way. Now we will begin our first area which is finding our C.I. (Confidence Interval). There are two different equations for finding the C.I. In the first equation C.I. = x̄ ± zα/2 , we have a small sample size so we will not be using this equation. We will have to go with the second which is not as accurate as we are assuming (S). Although our statistics have shown this to not be normal distribution, one may wish to do a Z-score P-value test anyway for this reason appendix 5 is provided. Here is the equation that we will use. C.I. = x̄ ± tα/2. x̄ = 752.5, S = 334.95. n = 20. We have to first find our T- value. Here is the equation, however, we will use the table in appendix 3. We now need to find our α. We can find this by setting our Significance Level. The three most common are 95%, 97%, and 99.7%. We will choose 95% or .95. We now can take this number and read our T-score from the chart in Appendix 3. We look up .05, for one tail with a Df of 19 this gives 2.093. A second method is using the calculator which gives the answer 2.093024022. Next, comes our error level which is found by E = Tα/2 = (334.95 /√20) = 74.89709691 \* 2.093 = 156.7596238 or rounded 156.760. So now we can find error range of significance.

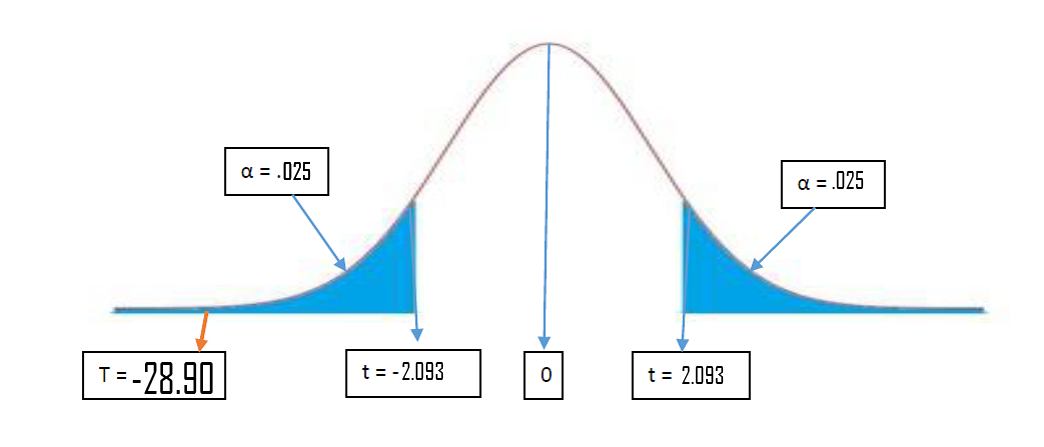
x̄ - E < µ < x̄ + E. 752.5 – 156.760< µ < 752.5 + 156.760= 595.740 < µ < 909.260.

**We do not actually know what the population mean is but we can be 95% certain that it is in the range of 595.740 < µ < 909.260.**

Hypotheses Testing:

“The division would also like to know if the average flight delay times for this year are higher or lower than last year’s average flight delay time of 42 minutes.“ That’s a pretty high number. We will have to work out some info on our delay times. We will start with just a standard sample mean  = 7.75 according to table 2. This is far below the 42 minutes in question. Currently, it is looking as if µ < 42.

A requirement for hypotheses testing is that the sample must be random and it is. Next, we need to know if σ or S is known. Here “S” is. Finally, we need “n > 30” or the population distribution to be normal. Neither are true in this case, however, for this test, we will assume the population distribution to be normal. Again we will be using a T-Score =  and . Let us work out S first. According our previous calculations delay time is 5.300199. Step one the claim. For a two tailed test we use = or ≠. And since we are given the number 42 in our question we will set µ = 42 as our Ho and µ ≠ 42 as our Ha. In appendix 3 the t-score for a two tailed test with 95% probability is -2.093.  = -28.89903878. We can see that the results of this test show that Ho is rejected, thus Ha is not rejected.



**There is enough evidence to support the claim that average delay time given with a 95% degree of confidence is either higher or lower than the 42 minute average of last year.**

To find out if it is higher or lower than the average let us perform a one tailed test. In this test since it is one tailed we will switch to a 99% probability. The claim is that µ < 42, this makes the opposite claim µ ≥ 42. So to restate the claim to identify H0 and Ha. So the H0 is µ ≥ 42 as this statement has the equality and is the null hypothesis. This makes Ha = µ < 42 the alternative hypothesis. Next, let us restate these again to help identify the µ equality and whether it is one tail or a two tail test. H0 : µ = 42; Ha: µ < 42. Since µ is less than 42 this will be a left tail test. α = .01 for a 99% probability level. Now on to  , x̄ = 7.75, S = 5.300199, and *n* = 20. Also µ is found in the restatement of H0: µ = 42.  = -28.89903878, and Degrees of freedom is 19. Looking at the table in

Appendix 3. We see our Critical Value is -2.54. T falls into the reject region of the distribution using the Traditional Method.

Let us double check to make sure using the P – Value test. Our C.I. is .01 > 0 so we reject again. Therefore, H0 : µ ≥ 42; Ha: µ < 42. So we reject H0 µ ≥ 42. This means that the alternative is true.



F.T.R

α = 0

T =-28.90

**There is enough evidence to support the claim that average delay time given in this sample is less than 42 minutes with a 99% degree of confidence.**



F.T.R

C.V. =-2.54

T =-28.90

Comparison and Conclusions:

The last order of business is to see whether the data and decisions we have come to can be supported by another study. Professor Peter Belobaba of M.I.T. published such a study called FLIGHT TIME COMPONENTS AND THEIR DELAYS ON US DOMESTIC ROUTES” (P. Belobaba, 2010). Belobaba found there is a regression line has a positive linear association between flight time and delay time – though not as pronounced to what we found.

Belobaba found that of the studied route 74% had delays, 20% had average block delays between zero and five minutes and only 6% had delays that were longer than 5 minutes. 92% of all flights arrived on time. Belobaba found x̄ ≤ 5 (P. Belobaba, 2010). x̅ = 5.30 from the standard form in our data. Needless to say that we are in range easily of the 5 minute average listed by Belobaba.

Recommendations:

Though Belobaba covers many of the reasons for delay times among flights little or no solutions are offered. The correlation between flight time and delay suggests that there may be possible places for improvement. Our stats show possible correlations that there may be a positive linear relationship between flight time and delay time. It seems that as flights time increases so does delay time. Further, the national average given seems unrealistic according to our data and the average wait time is much less than 42 minutes. The concern, however, is all the assuming that had to be made in order to perform our hypothesis testing. Not only was the sample size too low but also the distribution was not normal so we had to assume that it was. We calculated our σ assuming that we had a sample that represented n > 30 or at least 5% of our population this allowed us to set (σ = S). Recommendations for further study are warranted. In addition, it is recommended that more care be taken in sampling the population as to make sure to have n > 30.

**List of Works Cited.**

Belobaba, P. (2010) FLIGHT TIME COMPONENTS AND THEIR DELAYS ON US DOMESTIC ROUTES Dr. Prof. Amy Cohn MIT GLOBAL AIRLINE INDUSTRY PROGRAM, *http://web.mit.edu/airlines/industry\_outreach/board\_meeting\_presentation\_files/meeting-nov-2010/Skaltsas%20Schedules%20and%20Delays.pdf*

**Appendix 1** Spreadsheet 1

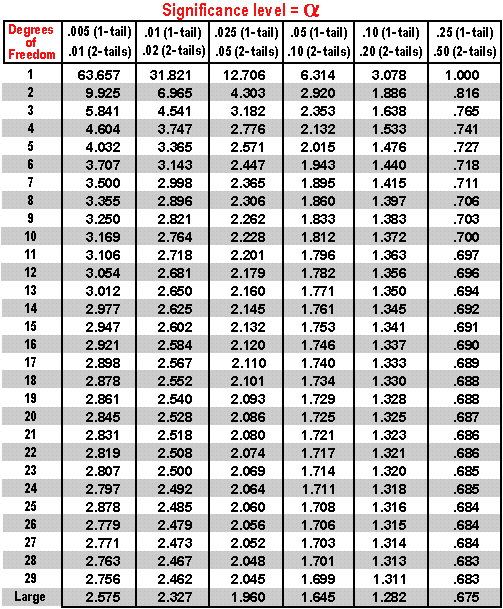


**Appendix 2:** Spreadsheet 2





**Appendix 3:** T - Tables



**Appendix 4:** List of Equations.

Outlier Equation. Median. The Median of the upper is Q1 and Lower is Q2. The IQR is Q2-Q1

Population Mean 

Sample Mean 

Sample Mean of a frequency 

Range = R = Xmax – Xmin

Sample Variance 

Sample Standard Deviation. 

Mean and Standard Deviation of the Sample Mean: Mean µ̄̄̄x̄ = µ , Deviation σx̄

Mean and Standard Deviation of the Population Proportion: Mean µ̄̄̄p̂= µ, Deviation σp̂, *q* = 1-*p*

Population Variance 

Population Standard Deviation 

Independent Theorem. *P (A∩ B)= P(A) \* P(B)*

Variance of a Discrete Random Variable X σ2 = Σ(*x-µ)P(x)*

Standard Deviation of a Discrete Random Variable 

Bothe Z-Scores and T-Scores can be found on tables as well as graphing calculators.

Z- Scores: Population  Sample 

Use T-Score if σ is not known and sample is less than 30 T-Score 

**T-score is sample is more than 30** 

Confidence Intervals. For Z If σ is known: C.I. = x̄ ± zα/2 , If σ is not known: C.I. = = x̄ ± zα/2

Confidence Intervals. for T If σ is known: C.I. = x̄ ± zα/2 , If σ is not known: C.I. = = x̄ ± tα/2

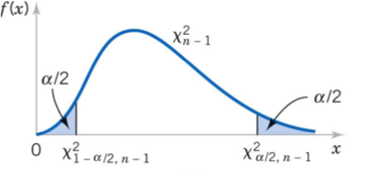
Degrees of Freedom: *Df=n-1*

Measure of error for T. E = Tα/2 

Chi values. Χ 2 = 

**Appendix 5:** Chi Statistics

Chi statistics is for skewed data distribution. . Χ 2 =  . n = sample number, s2 is sample variance, and σ 2 is population variance. Further things to note are: 1. Distribution is no symmetrical. 2. Values are never negative. 3. As df (n-1) goes up the distribution gets more symmetrical. 4. The χ chart reads values to the right. So n = 20 and the confidence level is 95%. With a two sided test we have α = .05/2 that gives a χr score of right 32.852 and χL left 8.907 according to appendix 6. S = 5.300199 from calculations previously make in the paper.



32.952

8.907

Finding the confidence interval for delay time and σ 2 using Χ 2 = . < σ 2 <  X2 left and right are read from the table in appendix 6.

Finding the confidence interval for delay time and σ 2 using Χ 2 = . < σ 2 <  < σ 2 <  = 16.19780527 < σ 2 < 59.92478718 = 4.024649708< σ < 7.741110203 average of this is 5.882879956

We do not know the exact standard deviation for the delay time, but we can be 95% sure that the standard deviation will be in this range 4.025< σ < 7.741.

Finding the confidence interval for fight time and σ 2 using Χ 2 = . < σ 2 < 

< σ 2 <  = 64688.776 < σ 2 < 239320.1467

or 254.340 < σ < 489.204 average of 371.772

We do not know the exact standard deviation for the fight time, but we can be 95% sure that the standard deviation will be in this range 254.340 < σ < 489.204

**Appendix 6:** Chi table

