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

Introduction:

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. “

Primary Data Analysis: Study and Sample type

In analyzing this data our first deduction is that this is an observational study, as these flights have already occurred. In further analysis, we are able to deduce that this data is from a random sample (a parameter that is given in the study) of a population of flights. Next, what are the population parameters? The population implied here 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 might be a parameter; this is not given so we will not have P (population). Or since the problem says a “small regional airport” we might assume all flight ever from that airport. In any case, the population is not known. 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 a 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. Further, the times are discrete because the measurements are taken to the nearest minute. Data values between 1 min and 2 min are not valid. In table 2 at the bottom of the two columns is the average or mean using equation. The sum and mean without the outlier is listed as sum1 and mean1. The sum and mean including the outlier is giving in sum2 and mean2. The outlier will be addressed later in this paper. 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, 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 (σ) by calculating (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.



Table 1

Examination of Descriptive Statistics: Flight Time

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 in the flight time data. Outliers are data outside the “outlier boundaries”. Outlier data would cause a greater deviation away from the true mean of our data. To check for outliers 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. Median = 835. We then find the median of Q1 and Q2 the same way. Q1= 470, Q3= 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(1.5) from Q1 = 470-720 = -250. Adding IQR \*1.5 to our Q3 = IQR(1.5)+Q3 = 720+955 = 1675. Our outlier range boundaries are -250 thru 1675, as we see all our values for X are well within this, **so there are no outliers (Box Plot 1).** Our other variable will be addressed in the pages that follow.



730

752.5

Box Plot 1

Sample Mean: Flight Time

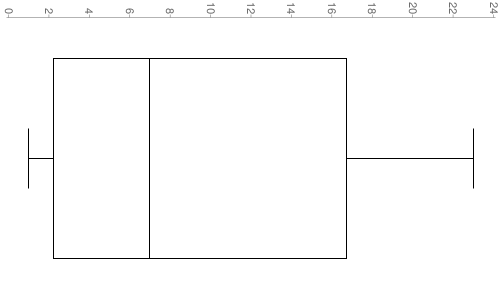
The mean of a sample or x-bar is what we use here as µ is mean for a population. However, since we assumed (µ = µx̄), and (σ = σx̄), then we say that x̄ is an acceptable estimate for µ and since the population mean ==  would be 752.5 is an acceptable estimation. Please refer to column 1 of table 2. Remember that we calculated sum1, mean1, sum2, and mean2. Sum1 and mean1 is without the outlier and in sum2 and mean2 the outlier is included. 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 either right or left. Later we will see that this is true. In fact, one will be left and the other right.

Mode: Flight Time

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.

Examination of Descriptive Statistics: Delay Time

Let us also look at delay time. We may find outliers on the other right-side. The reasons for finding outliers are the same as stated previously in the Flight Time Descriptive Statistics. As a result, this analysis will seem shorter as we have already explained the reasons for these calculations in the previous section. Next, we need the median. This means we need to align the data in ascending order according to flight delay. The low-class limit is 1, and the upper-class limit is 23 (table 3). We will need Q1, Q2, Median, Q3, and Q4. We will start with the median. We can use the equation for the median as we did in the previous section. Median, Median = 7. Q1 = 3.5. Q3 = 10.5. So LL = 1, Q1 = 3.5, Median = 7, Q3 = 10.5, and UL = 23. Using the same process as before Q3-Q1= 10.5-3.5 =7, Lower Boundary is Q1 – IQR(1.5) = 3.5-10.5 = -7 and Upper Boundary is IQR(1.5) plus Q3, 10.5 + 10.5 = 21. So our outlier range is -7 thru 21. It seems our UL of 23 is statistically outside out Upper Boundary of 21.



3.5

10.5

7

7.75

**There is one outlier data value 23 is outside of our range.**

Sample Mean: Delay Time

The mean of a sample or x-bar is what we use here as µ is mean for a population. If we include all data 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 7.75 is an acceptable estimation (table 3, sum2, mean2). Albeit, we found 23 to be an outlier we will be using only 19 pieces of data. Excluding 23 we get mean ==  would be the 6.95 is an acceptable estimation (table 3, sum1, mean1). We have shown these calculations to reveal the effect the outlier has on the mean value. Since the mean is greater than the median we now know this data distribution is not normal but slightly skewed either right or left. Later we will see that this is true. In fact, we will see that flight time and flight delay are skewed in opposite directions. Although, we found an outlier in the delay time this well not void our previous mean in the flight time. Because the common rule for statistics is to use the most conservative value we will be keeping the outlier inside our calculations. We will use sum 2 and mean 2 out of tables 2 & 3. This make our flight time mean = 752.5 and our delay time mean 7.75.

Mode: Delay Time

Please note according to table 3 this data is multi-modal with duplicates at 2, 3, 6, 7, 9, and 11. Further, due to the outlier in the delay time, our mode for the flight time now is 10.

Standard Deviation and Variance: Flight Time

We found x̄ = 752.5, and the variance for a sample is or as *Df= n-1* . Further, the standard deviation is given by . = = 113413.29. If we take the square root of that we have 336.7688957; this is our Standard Deviation or “S” (appendix 1). Notice that if we had excluded the outlier (S) = 306.4403723.

Standard Deviation and Variance: Delay Time

We found x̄ = 6.95, and the variance for a sample is or as *Df= n-1* . Further, the standard deviation is given by. = = 28.76579. If we take the square root of that we have 5.363375 this is our Standard Deviation or “S” (appendix 1). Notice that if we had excluded the outlier (S) = 4.006574

Frequency Distribution

Although the distributions are “bell-like”, we can make out a slight 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. We have already shown this to only have 1 mode previously, and the skew seems to be around #7 at 800-1k. One probable reason is the small sample size. Since n = 19 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 constraint we are talking about

an unknown population. 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 due to the large number of flights compared to our sample size.

Delay Time



Flight Time



Description of Sample Data: Plot and Regression

 Referring to Table 2 we used the X vs. f column to create the plot and regression chart. This sample data was given as part of the problem. It is insufficient in that it is not either >30 or stated to be a normal distribution so in all rights the only thing that we can truly say is there is nothing we could say about this data. 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 scatter plot. Here we have included a regression line provided through the Microsoft excel plot and regression function. By this, we might 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 us 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 = 56.473. Our regression line has the following equation Ŷ = 56.473x + 314.83. 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.

Examination of Inferential 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) = 5.363375. Further, since we concluded (σ = S) and (σ = σx̄), we will accept 5.363375 number for (S) and (x̄) = 7.75 for all further calculations. For our flight time the, numbers are listed in appendix 1. (Sf) = 336.7688957.and (x̄) = 752.5.

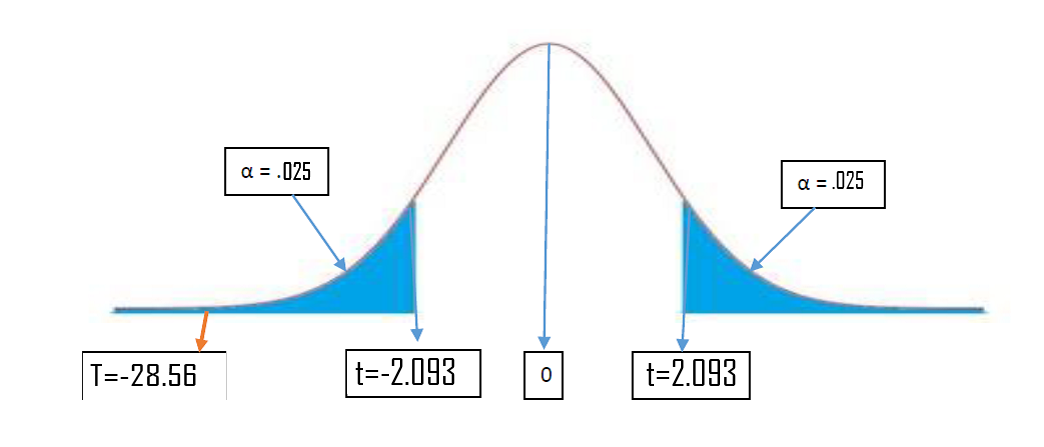
Here is where we begin to check requirements. First, the text does say that the sample was random. Next, we assume that this random sampling was not biased in any way. Now we will begin our first area which is finding our C.I. (Confidence Interval). There are two different equations for finding the C.I. 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 = 336.7688957. 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 and .025 for two tails. With a Df of 19 the chart reads 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 = (336.77 /√20) = 75.3038\* 2.093 = 157.61088 or rounded 157.611. So now we can find error range of significance.

x̄ - E < µ < x̄ + E. 752.5 – 157.611< µ < 752.5 + 157.611 = 594.889 < µ < 910.111.

**We do not actually know what the population mean is but we can be 95% certain that it is in the range of 594.889 < µ < 910.111.**

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 to our previous calculations delay, time is 5.363375. 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 -2.101.  = -28.5586326. 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 95% 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 just to help identify the µ equality and whether it is one tailed or a two tailed test. H0: µ is > or = 42; Ha: µ < 42. Since µ is less than 42 this will be a left tail test. α = .05 for a 95% probability level. Now on to  , x̄ = 7.75, S = 5.363375, and *n* = 20. Also µ is found in the restatement of H0: µ = 42.  = -28.5586326, 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.



F.T.R

C.V. =-2.093

T =-28.56

P-value Method

Let us double check to make sure using the P – Value test. Our C.I. is .05 > 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.

Comparison and Conclusions:



F.T.R

α = 0

T =-28.56

**There is enough evidence to support the claim that average delay time given in this sample is less than 42 minutes with a 95% degree of confidence.**

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. Balobaba is using a slightly different dependent variable – “buffer”; defines as “SCHEDULED ARRIVAL TIME minus DELAY-FREE ARRIVAL TIME”. This buffer adds a bit of a cushion to his findings as this buffer helps to reduce the hard line delay times per flight. Ours has just the non-buffered times listed. Belobaba found x̄ ≤ 5 (P. Belobaba, 2010). x̅ = 7.75 from the standard form in our data. One reason for the difference is the use of the buffer in his calculations. Needless to say that we are in range of the 5 minute average listed by Belobaba.We found a positive regression that allows us to say there is a positive relationship between flight time and flight delay. Further, we found the distribution to be skewed as well. Further, our average wait time seems to be comparable to Belobaba’s. Ŷ = 0.0141X – 2.86 was the equation we found from our regression line. The difference in the slope values tells us that the ratio of increase in delay time due to increase in flight time is much larger at our airport. This possibly is due to the difference in airports, as the one we studies was a small regional airport, and Belobaba studied major U.S. air ports. Regional airports tend to have smaller slower aircraft. This could also account for the difference in the y-intercepts. The R2 values show a value difference as well. This is more than likely due to the difference in number of data. We were working with only 19 after the outlier was excluded. As you can see by Belobaba’s dot plot and regression line, he had many more than 19.

Recommendations:

Though Belobaba covers many of the reasons for delay times among flights little or no solutions are offered. I am not an expert on aircraft, airports, or even a statistician. As a result, I do not have any solutions other than further study. The correlation between flight time and delay suggests that there may be possible places for improvement as to help with reducing delays that are not the outcome of acts of God. Our stats show possible correlations that there may be a positive linear relationship between flight time and delay time and this does agree with Belobaba. It seems that as flights time increases so does delay time. One may assemble teams to categorize the sources of flight delays. With that information on may be able to find places for improvement. Further, the national average given seems unrealistic according to our data and the average wait time is only 6.95 minutes which is much less than the 42 minutes given. Even if we accepted the outliers the average would be 7.75 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. These findings are inadequate as the care in taking the data was well below the parameters needed. There must be either n > 30 or a known normal distribution and we had neither. The sample size was so low that the frequency distribution was skewed, and almost unreadable. We could have increased the class about but with only 19 pieces of data, it is to no avail. Further, a set population in a set time frame is needed to qualify the finding. Flight surely have different delay times in winter than in summer. Lastly, it is my feeling that the number 42 may be a misprint and needs to be confirmed. Both Belobaba’s and our data showed single digit delay times. Is it possible that this 42 minute time should have been 4.2 minutes. If so this is much closer to our finding.

**List of Works Cited.**

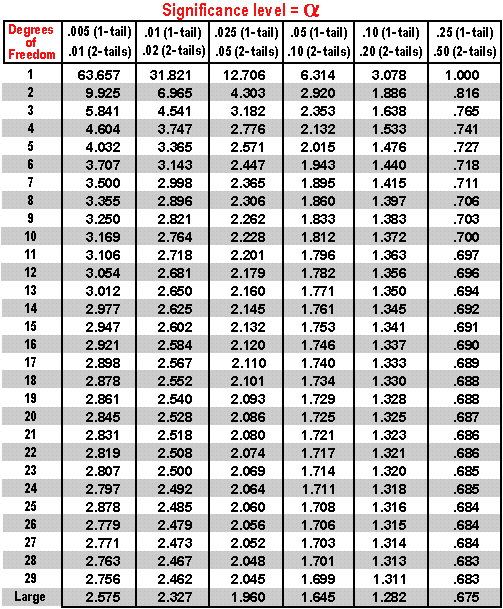
Belobaba, P. (2010) FLIGHT TIME COMPONENTS AND THEIR DELAYS ON US DOMESTIC ROUTES Dr. Prof. Amy Cohn MIT GLOBAL AIRLINE INDUSTRY PROGRAM, *http://web.mit.edu/airlines/industry\_outreach/board\_meeting\_presentation\_files/meeting-nov-2010/Skaltsas%20Schedules%20and%20Delays.pdf*

**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

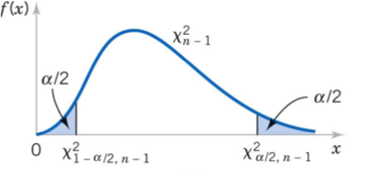
Degrees of Freedom: *Df=n-1*

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

