Introduction to Probability and Stats

East Central University

Dr. Nicholas Jacob

January 8, 2021

Deaths and Mortality in the United States

The data that I am using came from the CDC (https://data.cdc.gov/dataset/Deaths-in-122-U-S-cities-1962-2016-122-Cities-Mort/mr8w-325u). This data set has 122 U.S Cities and ranged from years 1962-2016. The vital statistics offices of 122 cities across the United States reported the total number of death certificates processed and the number of those for which pneumonia or influenza was listed as the underlying or contributing cause of death by age group. Mortality data in the table was voluntarily reported. As a Registered Nurse, I found this information very interesting. I like how it lists if pneumonia or influenza was a contributing cause of death.

To collect the information needed I searched through various years and cities. I choose the year 1997 because that's the year I was born. I also choose 25 cities that I have visited. The categorical variables include city and state. Quantitative variables include pneumonia and influenza deaths and all deaths.

Using this data, I hope to find out how much pneumonia and influenza affects the death rate in the United States. I think that finding the relationship will be an eye opener as to just how deadly pneumonia and influenza can be.

YEAR	CITY	STATE	PNEUMONIA & INFLUENZA DEATHS	ALL DEATHS
1997	Batton Rouge	Louisiana	85	2376
1997	Denver	Colorado	495	5578
1997	Dallas	Texas	289	10163
1997	Tulsa	Oklahoma	500	5940
1997	Saint Louis	Missouri	231	5798
1997	Columbus	Ohio	751	9809
1997	Indianapolis	Indiana	411	9744
1997	Austin	Texas	241	3958
1997	Salt Lake City	Utah	512	5522
1997	Las Vegas	Nevada	582	9401
1997	Colorado Springs	Colorado	207	2871
1997	Memphis	Tennessee	767	9479
1997	Birmingham	Alabama	439	7046
1997	Knoxville	Tennessee	468	4613
1997	Miami	Florida	28	5386
1997	Wichita	Kansas	144	3898
1997	Chicago	Illinois	1671	23010
1997	Atlanta	Georgia	269	7929
1997	Savannah	Georgia	278	2785
1997	Mobile	Alabama	59	4043
1997	Nashville	Tennessee	366	7539
1997	Fort Worth	Texas	310	5820
1997	Houston	Texas	1489	19102
1997	Phoenix	Arizona	643	8235
1997	Tampa	Florida	809	9449

Works Cited

CDC, N. I. (2016, October 6). *Deaths in 122 U.S. cities- 1962-2016. 122 Cities Mortality Reporting System*. Retrieved from Centers for Disease Control and Prevention: https://data.cdc.gov/dataset/Deaths-in-122-U-S-cities-1962-2016-122-Cities-Mort/mr8w-325u

Part 2: Analyzing Categorical Data

For the second part of this project, I examined the frequency and relative frequency of one of my categorical variables presented in my project (Table 2.1). I also created a two-way table, to show the correlation between cities and states. (Table 2.2) There are several cities and states in my data set, so I chose to focus on a select few. As the table shows, there is only going to be one Batton Rouge in Louisiana. This goes for the rest of the cities as well. On the states, there is 2 cities on the table that reside in Texas, Dallas and Austin.

State	Frequency	Relative Frequency
Louisiana	1	0.03448276
Colorado	2	0.06896552
Texas	4	0.13793103
Oklahoma	1	0.03448276
Missouri	1	0.03448276
Ohio	1	0.03448276
Indiana	1	0.03448276
Texas	3	0.10344828
Utah	1	0.03448276
Nevada	1	0.03448276
Tennessee	3	0.10344828
Alabama	3	0.10344828
Florida	2	0.06896552
Kansas	1	0.03448276
Illinois	1	0.03448276
Georgia	2	0.06896552
Arizona	1	0.03448276
Total	29	1.00000003

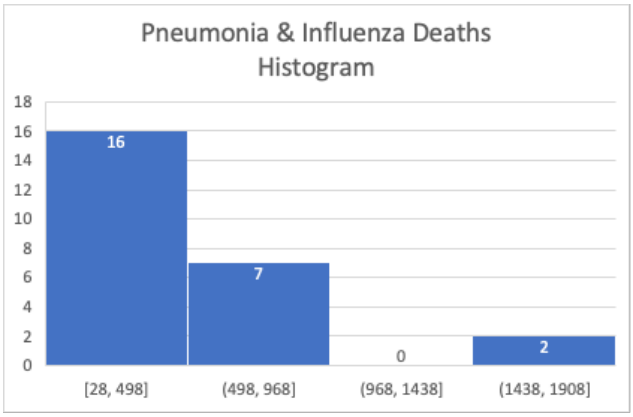
Table 2.1

Two-Way	Batton Rouge	Denver	Dallas	Tulsa	Saint Louis	Columbus	Indianapolis	Austin	Salt Lake City	Las Vegas
Louisiana	1	0	0	0	0	0	0	0	0	0
Colorado	0	1	0	0	0	0	0	0	0	0
Texas	0	0	1	0	0	0	0	1	0	0
Oklahoma	0	0	0	1	0	0	0	0	0	0
Missouri	0	0	0	0	1	0	0	0	0	0
Ohio	0	0	0	0	0	1	0	0	0	0
Indiana	0	0	0	0	0	0	1	0	0	0
Utah	0	0	0	0	0	0	0	0	1	0
Nevada	0	0	0	0	0	0	0	0	0	1

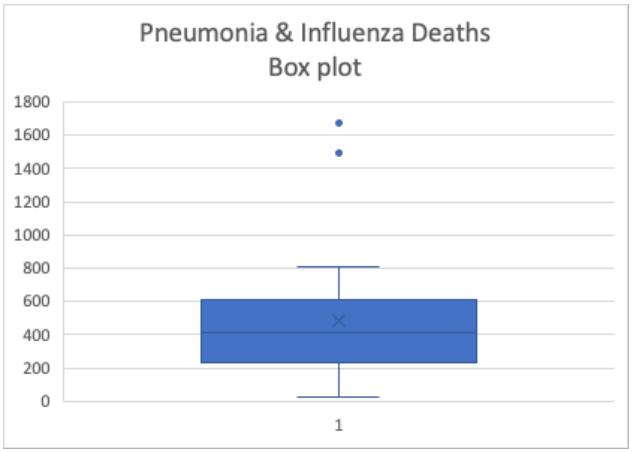
Table 2.2

Part 3: Analyzing Quantitative Data

For the third part of this project, I analyzed the deaths related to pneumonia and influenza. Information was obtained from the histogram (Table 3.1) and the box plot (Table 3.2). (Table 3.3) includes the summary statistics. The histogram is skewed to the right, which is confirmed in the boxplot. The mean is more than the median, as shown in (Table 3.3). When analyzing the histogram, we can see that 16 cities had between 28 and 298 deaths related to pneumonia and influenza, 7 cities had between 498 and 968 deaths related to pneumonia and influenza, 0 cities had between 968 and 1438 deaths related to pneumonia and influenza, and 2 cities had between 1438 and 1908 deaths related to pneumonia and influenza. As shown in (Table 3.2) there are two outliners present with values of 1438 and 1908.



(Table 3.1)



(Table 3.2)

Minimum	28
Q1	241
Q2 (Median)	411
Q3	582
Maximum	1671
Mean	481.76
Range	1643
Standard Deviation	395.472426

(Table 3.3)

Part 4: Writing a Null and Alternative Hypothesis

For the fourth part of this project, I created a null and alternative hypothesis for one categorical and one quantitative variable.

Quantitative Hypothesis:

My Hypothesis is that the average number of all deaths in my selected data set for the year 1997 is 7,000. I made this hypothesis after reviewing the number of all deaths in the data set. My alternative hypothesis would then be that it does not equal 7000. I found this by calculating the mean for all deaths in the data set for the year 1997. This came to 7579.76

$H0: \mu AD = 7{,}000$

$Ha: \mu AD \neq 7{,}000$

Categorical Hypothesis:

My Hypothesis is that out of the 25 states used in the data set 12% of those states were Texas. I made this hypothesis after reviewing the cities and states listed in the data set. My alternative hypothesis would then be that it is greater than 0.12. I used the formula P=x/n. X=4 and N = 25 and came up with P= 0.16. So, out of the 25 states used in the data set 16% of those states were Texas.

$H0: p = 0.12$

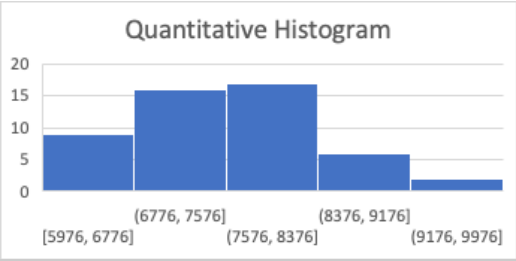
$Ha: p > 0.12$

Part 5: Bootstrap

For the fifth part of this project, I performed a bootstrap to find out if I could reject or fail to reject my quantitative and qualitative hypothesis that was created in part four.

Quantitative Variable:

To create my quantitative hypothesis, I took the average number of all deaths in 1997. My hypothesis was as followed. H0: μ AD = 7,000; Ha: μ AD $\neq$ 7,000. I created a bootstrap distribution with 50 entries. The standard error is 880.295621. The 95% confidence interval for the mean had a lower range of 5863.32858 and a higher range of 9314.08742. I have included a histogram of the bootstrap distribution of the means below (Figure 5.1). As you can see the histogram is bell shaped with the center of the bootstrap distribution being 7588.708. Looking at the information, I can conclude and fail to reject my hypothesis. My hypothesis states that it equals 7,000. This value fell within the 95% confident interval.

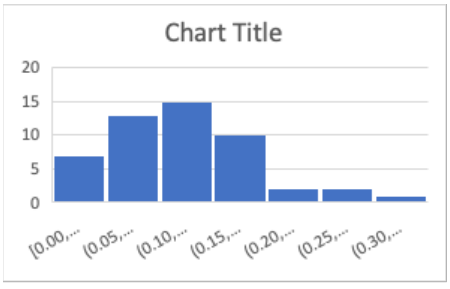


Mean	7588.708
Standard Error	880.295621
95% Confidence Interval	5863.32858-931408742

Figure 5.1

Categorical Hypothesis:

To create my categorial hypothesis, I made it after reviewing the cities and states listed in the data set. My hypothesis was as followed. H0: p = 0.12; Ha: p >0.12. On excel, I created a "IF" statement and I had it put a "1" it is was Texas and a "0" if it wasn't Texas. I then created a bootstrap distribution with 50 entries. The standard error is 007466278. The 95% confidence interval for the mean had a lower range of 0.01366095 and a higher range of 0.306339905. I have included a histogram of the bootstrap distribution of the means below (Figure 5.2). As you can see the histogram is bell shaped with the center of the bootstrap distribution being 0.12. Looking at the information, I can conclude and fail to reject my hypothesis. This value fell within the 95% confident interval.



Mean	0.16
Standard Error	0.07466278
95% Confidence Interval	0.01366095-0.306339905

Table 5.2

Part 6: Categorical Inference with Formulas

For the sixth part of my project, I am going to repeat the hypothesis test on my categorical variable using the correct formulas. Then, I will compare the results to my bootstrapping in part 5. I created (Table 6.1) to show the statistic, formula, and result used to retest my hypothesis. My Hypothesis stated: H0: p = 0.12; Ha: p >0.12.

STATISTIC	FORMULA	RESULT
N	Sample Size	25
X	Number of Successes (in my case number of times Texas appears)	4
Proportion	(from hypothesis)	0.12
P Value	=NORM.S. DIST (Z, True) 1-VALUE (0.73087367)	0.26912633
P-Hat	X/N	0.16
Standard Error	=SQRT(p*(1-p)/n	0.06499231
Z Statistic	= (Phat- p)/SE	0.61545745
Z*	=NORM.S.INV (0.975)	1.95996398
CI (low)	=phat-(2*SE)	0.03001539
CI (high)	=phat+(2*SE)	0.28998461

Table 6.1

My null hypothesis stated that out of all the states used in the data set 12% of those states were Texas. The 95% confidence interval is 0.03001539-0.28998461. Therefore, I will fail to reject my null hypothesis, because 12% falls within the 95% confidence interval.

Comparison of Results from Bootstrap:

RESULTS FROM PART 5 BOOTSTRAP	RESULTS FROM RETEST USING FORMULAS
95% Confidence Interval: 0.01366095-0.306339905	95% Confidence Interval: 0.03001539-0.28998461

Table 6.2

As you can see in (Table 6.2), both methods revealed very similar results. Both tests resulted in failing to reject the null hypothesis. The results that were computed by using formulas have a smaller range between the low and high confidence intervals. This leads me to think that using the formulas will result in a more accurate confidence interval.

Part 7: Quantitative Inference with Formulas

For the seventh part of my project, I am going to repeat my hypothesis test on my quantitative variable using the correct formulas. Then I will compare the results to my bootstrapping in part 5. I created (Table 7.1) to show the statistic, formula, and result used to retest my hypothesis. My hypothesis stated: H0: μ AD = 7,000; Ha: μ AD $\neq$ 7,000.

STATISTIC	FORMULA	RESULT
MU	(From Hypothesis)	7000
Xbar	Average (sum of the numbers/n)	7579.76
Sigma	(standard deviation)	4746.38104
N	Sample Size	25
Standard Error	Sigma/SQRT(n)	949.276207
T Statistic	Xbar-mu/SE	7572.38596
Alpha	For 95% CI (1-.95)	0.05
T* (T-Critical Value)	=T.INV (alpha, n-1)	1.7108821
95% CI (Low)	Xbar-T-Critical Value*SE	5719.17863
95% CI (High)	Xbar+T-Critical value*SE	9440.34137

Table 7.1

My null hypothesis states that the average number of deaths in 1997 was 7,000. The 95% confidence interval is 5719.17863-9440.34137. Therefore, I will fail to reject my hypothesis, because 7,000 falls within the 95% confidence interval.

Comparison of Results from Bootstrap:

RESULTS FROM PART 5 BOOTSTRAP	RESULTS FROM RETEST USING FORMULAS
95% Confidence Interval: 5863.32858- 9314.08742	95% Confidence Interval: 5719.17863-9440.34137

Table 7.2

As you can see in (Table 7.2), both methods revealed very similar results. Both tests resulted in failing to reject the null hypothesis. Because 7000 falls within both confidence intervals.

Part 8: Conditional Probability from Two-Way Table

For the final part of this project, I will create two conditional probabilities using the two-way table from part 2 (Table 2.2).

Two-Way	Batton Rouge	Denver	Dallas	Tulsa	Saint Louis	Columbus	Indianapolis	Austin	Salt Lake City	Las Vegas	**TOTAL**
Louisiana	1	0	0	0	0	0	0	0	0	0	**1**
Colorado	0	1	0	0	0	0	0	0	0	0	**1**
Texas	0	0	1	0	0	0	0	1	0	0	**2**
Oklahoma	0	0	0	1	0	0	0	0	0	0	**1**
Missouri	0	0	0	0	1	0	0	0	0	0	**1**
Ohio	0	0	0	0	0	1	0	0	0	0	**1**
Indiana	0	0	0	0	0	0	1	0	0	0	**1**
Utah	0	0	0	0	0	0	0	0	1	0	**1**
Nevada	0	0	0	0	0	0	0	0	0	1	**1**
Total	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**10**

Table 2.2

The first probability I will create is if one city is randomly selected. What is the probability that the city is Tulsa if the given state is Oklahoma? To calculate the probability, I will be using the following formula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

I am going to repopulate the formula with new variables so that it makes sense for the question. This is optional, but it helps clarify what you're looking at. I'm going to say T is Tulsa and O is Oklahoma, so the formula becomes: **P (T|O) = P (T∩O) /P(O).**
Figure out P (T∩O) from (Table 2.2). The intersection of Tulsa/Oklahoma (the intersection on the table of these two factors) is 1. Figure out P(O) from the table, which is 1. Therefore, P (T|O) = 1/1=1, or in other words there is a 100% chance that Tulsa will be selected if the given state is Oklahoma.

The second probability I will create is if one city is randomly selected. What is the probability that the city is Dallas if the given state is Texas? Following the same steps as above, I repopulated the formula. I am going to say that D is Dallas and T is Texas, so the formula is now: **P(D|T) = P(D∩T)/P(T)**.
Figure out P(D∩T) from the table. The intersection of Dallas/Texas which equals 1. Figure out P(T) from the table, which is 2. Therefore, P(D|T) =1/2=0.5, or in other words there is a 50% chance that Dallas will be selected if the given state is Texas.