Diabetes 2015-2017 Statistics Analysis

Intro to Probability and Statistics

[redacted]

August 23, 2020

Dr. Nicholas Jacob

Diabetes 2015-2017 Statistics Analysis

The data and variables that were chosen for this statistical analysis interested me due to my passion for health promotion and fitness, as well as how people can possibly improve their health through exercise. As a registered nurse for over twenty-five years, I have seen the devastating effects that diabetes has on patients' health and quality of life. The data that I used for this project originated from the website https://data.austintexas.gov/Health-and-Community-Services/Austin-Public-Health-Diabetes-Self-Management-Educ/48iy-4sbg. Below is an example of the data base that was used in this analysis.

Austin_Public_Health_Diabetes_Self-Management_Education_Participant_Demographics_2015-2017-4

Class Language	Age	Year	Gender	Race/Ethnicity	High Blood Pressure (Yes/No)	Fruits & Vegetable Consumption	Carbohydrate Counting	Exercise	Problem Area in Diabetes (PAID) Scale Score
Spanish	35	2015	F	Hispanic/Latino	Yes	1-2	0 days	0 days	28.75
Spanish	37	2016	F	Hispanic/Latino	No	3-4	I don't know how	1 day	72.5
Chinese/English		2017		Asian					
Spanish	52	2015	F	Hispanic/Latino					
English	60	2015	M	White	Yes	1-2	0 days	5 or more days	11.25
Spanish	27	2015	F	Hispanic/Latino	No				73.75
Spanish	53	2015	F	Hispanic/Latino	No	1-2	0 days	3 days	88.75
Spanish	28	2017	F	Hispanic/Latino					
Spanish	53	2015	M	Hispanic/Latino	No	3-4			15

To determine my sample, I deleted the columns that did not pertain to this particular study. I further decreased the population by deleting rows that did not have complete variable data. This left a population size of 564. I then chose every tenth case to be a part of this study. This systematic sampling resulted in 10% of the population included in this analysis. The sample includes 56 cases. Essential data for this study includes age, the presence or absence of hypertension (high blood pressure), diabetes knowledge, the number of days exercised each week, and their Problem Areas in Diabetes (PAID) Scale Score. This scale measures the degree of difficulty in managing diabetes. It ranges from 0-100. The higher the score indicates more problems managing diabetes.

There were large amounts of variables in the original dataset; however, I chose these particular variables (age, hypertension diagnosis, diabetes knowledge, days of exercise per week, and PAID scale score), because I am wanting to see if there is an association between these variables specifically. The categorical variables include the patient's age, the presence or absence of hypertension as a co-morbidity, and diabetes knowledge. The quantitative variables are the days per week that they exercised and their PAID scale score.

With this data, I will be able to find a relationship between their ability to manage diabetes considering their age, co-morbidity (hypertension), diabetes knowledge, and the number of days of exercise per week. I think it will be interesting to determine if there is a greater chance of hypertension or higher PAID score as age increases. I could also analyze the effect of diabetes knowledge as it relates to their PAID score. More importantly, I will investigate if the days per week of exercise affects the diabetics ability to manage their diabetes any better (lower PAID scale score).

In knowing that as people age, disease process typically worsens without intervention I hope to see the impact that exercise has on their co-morbidity (hypertension) and their ability to lower their PAID scale score. This is important information, because it could suggest that a certain amount of days of exercise per week could potentially make their diabetes more manageable, which, in turn can positively affect their health and quality of life. This study may also find no evidence of benefit in the number of days per week of exercise. In this case, with further research, it would be important to study the effects of other variables, such as smoking history, genetics, diet, and medical compliance as stronger variables that affect the diabetic's

ability to manage their diabetes diagnosis. Below is a sample table including 25 of 56 cases in my population with the data that I will use for this project.

Sample of Diabetes Statistic Analysis								
Age		**HTN**		**DM Knowledge**		**Exercise**		**PAID scale**
71		Yes		Fair		2		33.75
58		Yes		Fair		1		86.25
44		Yes		Fair		0		91.25
70		No		Fair		3		72.5
47		No		Fair		1		42.5
55		Yes		Fair		3		23.75
64		Yes		Good		5		82.5
54		No		Poor		0		61.25
57		Yes		Poor		2		87.5
64		Yes		Good		1		16.25
23		No		Fair		0		37.5
40		No		Good		2		55
47		No		Good		0		38.75
38		No		Fair		3		78.75
60		Yes		Poor		2		35
53		Yes		Fair		1		55
47		No		Fair		2		82.5
52		No		Fair		0		86.25
45		No		Fair		2		21.25
51		No		Poor		2		90
54		No		Fair		1		10
57		Yes		Fair		1		93.75
56		Yes		Poor		0		72.5
47		No		Poor		3		100
52		No		Poor		1		92.5

Part 2: Analyzing Categorical Data

For the second part of my project, I examined the frequency and relative frequency of several different categorical variables that I was interested in. I was curious to see the frequency and relative frequency of diabetics that had hypertension (Table 2.1), as well as their perceived diabetes knowledge (Table 2.2). Table 2.3 is the frequency and relative frequency of the age groups.

(Table 2.1)

Hypertension	**Frequency**	**Relative Frequency**
No	22	0.60714286
Yes	34	0.39285714
Total	56	1

(Table 2.2)

Diabetes Knowledge	**Frequency**	**Relative Frequency**
Poor	9	0.1607142
Fair	36	0.6428571
Good	11	0.1964285
Total	56	1

In these tables, 61% of the diabetic patients did not have hypertension, while 39% did have hypertension. 16% of the diabetics that admitted to having “poor” diabetes knowledge, while 36% considered themselves as having “fair” knowledge, and 11% admitted to having “good” knowledge of diabetes. Table 2.3 shows that the 50-59 age group was the largest group.

They represented 34% of the sample, while the smallest age group was 20-29 with only 3.5% representation. The other age groups were similar in size ranging from 14-17%.

(Table 2.3)

Age	Frequency	Relative Frequency
20-29	2	0.0357142
30-39	8	0.1428571
40-49	10	0.1785714
50-59	19	0.3392857
60-69	8	0.1428571
70 and over	9	0.1607142
Total	56	1

Table 2.4 is a comparison of the age groups as it relates to a hypertension diagnosis. In this table, it reveals a significant increase in percent of diabetics who have hypertension as they get older. With 0% of diabetic patients having hypertension in their 20's, and 25% in their 30's. It continues to increase to 40% in their 40's and 63% in their 50's. There was a 100% hypertension rate in the patients in their 60's and 89% in the 70 and over age group.

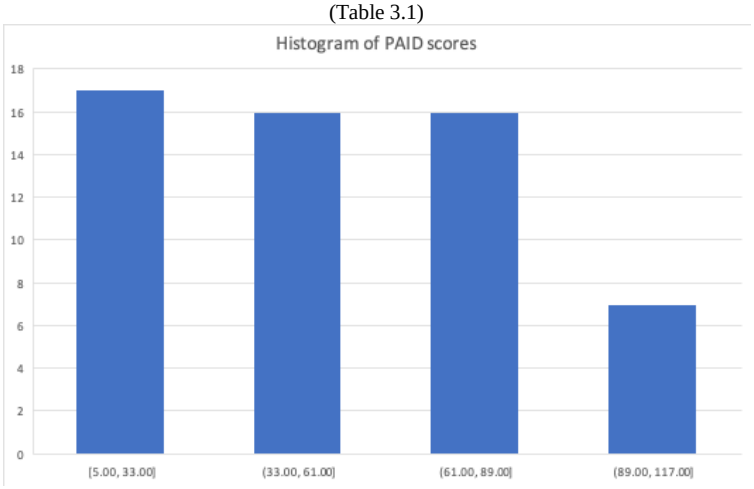
(Table 2.4)

	20-29	30-39	40-49	50-59	60-69	70 and over	Total
With Hypertension	0	2	4	12	8	8	34
Without Hypertension	2	6	6	7	0	1	22
Total	2	8	10	19	8	9	56

Part 3: Quantitative Variables and Graphics

For the third part of my project, I analyzed the PAID scores (difficulty in managing diabetes). Information was obtained from a histogram (Table 3.1) and a box plot (Table 3.2). The histogram is skewed right, which is confirmed in the box plot. The mean is more than the median. When analyzing the histogram, we can see that 17 people had PAID scores between 5-33, 16 people had scores from 33-61, 16 people had scores 61-89, and the least amount of people (7) had the highest paid scores between 89-117. The smallest group had the hardest time managing their diabetes. It is worth mentioning that Table 3.1 is not a normal bell shape. This could be because it is uniform, which means any score between 1-100 are equally likely.

(Table 3.1)



(Table 3.2)

According to the box plot, there are no outliers in this study. With this box plot we can analyze the following:
Mean: 50.9821429
Standard Deviation: 30.0848368
Minimum: 5
Q1: 23.4375
Median: 45
Q3: 82.5
Max: 100
Range: 95

Part 4: Writing a Null and Alternative Hypothesis

For the fourth part of this project, I will create a null and alternative hypothesis for one categorical and one quantitative variable. The categorical hypothesis that I will test is the percentage of diabetic patients with hypertension. My null hypothesis states that 50% of the diabetic patients will have hypertension, and my alternative hypothesis states that the percentage

of patients that also have hypertension will not equal 50%. This was my educated guess simply from observing diabetic patients in my care over the years.

Categorical hypothesis:

H_0: $P\_{hypertension} = .50$
H_a: $P\_{hypertension} \neq .50$

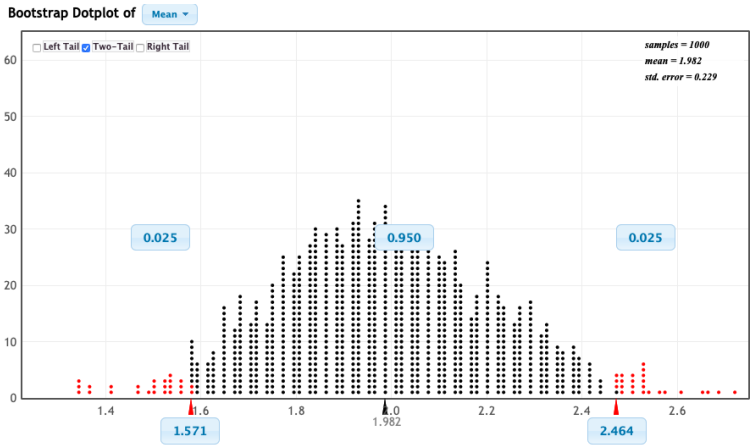
The hypothesis for my quantitative variable is the amount of days per week that the patients exercised. It is recommended that diabetic patients exercise at least 3 days per week or 150 minutes per week; therefore, my hypothesis reflects what the diabetics have been told to do, and it assumes that they are being compliant with recommendations to control their blood sugars. This is significance because the amount of days per week or exercise could potentially affect their PAID scores. My null hypothesis states that the days of exercise is equal to three; whereas, my alternative hypothesis states that that the days of exercise is not equal to 3.

$H_{0:}$ $\mu = 3$
$H_{a:}$ $\mu \neq 3$

Part 5: Hypothesis Test Using Bootstrapping

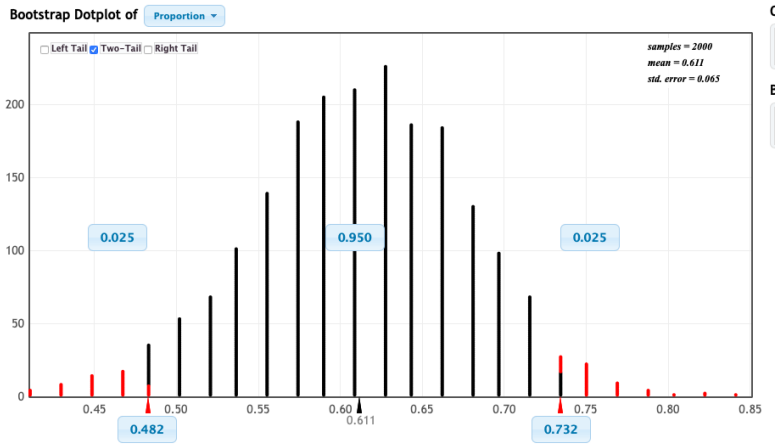
For the 5th part of the project, I will reject or fail to reject my quantitative and qualitative hypotheses. I used Stat Key for my quantitative null hypothesis $H_{0:}$ $\mu = 3$. My bootstrapping sample was n=1000.

Bootstrap Dotplot of Days of Exercise



The standard error is 0.229. The 95% confidence interval is between 1.571 and 2.464. I will have to reject my hypothesis that the days per week of exercise are equal to 3. It is outside of the highest confidence interval.

Bootstrap Dot Plot of Proportion of Diabetics with HTN



My qualitative null hypothesis stated H_0: $P\_{hypertension} = .50$ and this reflects the proportion of diabetic patients that also have hypertension. The standard error for this bootstrap sample of 2000 is 0.065. The 95% confidence interval is .482 (48%) to .732 (73%); therefore, I will fail to reject my hypothesis P_hypertention = .50, because it is within the confidence interval.

Part 6: Categorical Inference with Formulas

For the sixth part of my project, I am going to retest the hypothesis above for my categorical value using formulas, and then I will compare it to my bootstrapping test in part 5.

Statistic	Formula	Result
p-hat	35/56	.60714
Standard error	= SQRT(p*(1-p)/n	0.0668
Z statistic	= p hat-p/SE	1.6036
Z*	From the table	1.96
CI (low)	= p hat – (2*SE)	.47
CI (high)	= p hat + (2*SE)	.76

My null hypothesis states that 50% of the diabetic patients have hypertension. The 95% confidence interval is 47% and 76%; therefore, I will fail to reject my null hypothesis, because 50 % falls within this range.

Comparison of Bootstrap from Part 5

Bootstrap from part 5	Retest of Hypothesis using Excel Formulas
95% Confidence interval = 48%-73%	95% Confidence interval = 47% - 76*

It is very interesting to see how 2 different tests can render such similar results. Both categorical tests resulted in failing to reject the null hypothesis. This tells me that hypertension is quite prevalent in diabetic patients in this sample population.

I chose this the hypertension categorical variable for this test, because it is significant in determining other factors that might contribute to a diabetic having difficulty in managing their diabetes (PAID score) other than exercise. Diabetes can damage kidneys, which can lead to hypertension. Hypertension can make diabetes worse, and diabetes can make hypertension worse; therefore, it is congruent to say that patients that have hypertension may have a harder time controlling their diabetes (PAID score). My hope continues to be to see if there is statistical data that links exercise to lower PAID scores in this sample, knowing the proportion of patients with hypertension is relatively high.

Part 7: Quantitative Inference with Formulas

For the seventh part of the project, I will retest the days of exercise hypothesis by using formulas. I will then compare them to the bootstrapping results from part 5. My null hypothesis states that the mean days of exercise in the population sample is 3.

$H_{0:}\ \mu = 3$
$H_{a:}\ \mu \neq 3$

Statistic	Formula	Result
x bar	Add all of the days/n (56)	1.98
Standard deviation	= STEV.S(range)	1.7059
Standard error	= sigma / SQRT (n)	.228
Z*	Per z score table	1.96
T statistic	(xbar- μ) / (sigma/SQRT of n)	-4.47
95% Confidence Interval	M +or- z*SE	1.53, 2.43

Comparison of Bootstrap from Part 5

Bootstrap from part 5	Retest of Hypothesis using Excel Formulas
95% Confidence interval = 1.571, 2.464	95% Confidence interval = 1.53, 2.43

The 95% confidence interval from formulas is 1.53, 2.43, and my mu is 3; therefore, as with the bootstrapping sample, I will have to reject my null hypothesis. The two intervals are very similar. I chose this variable, because I wanted to retest the hypothesis since it is such a vital piece of information that will help me answer the question…. Does the number of days of exercise that a diabetic patient has affect his or her ability to manage their diabetes (PAID score)?

Part 8: Probabilities

For the final part of this project, I wanted to find the probability that 5 days of exercise would help the diabetic control their diabetes (have a lower paid score).

	0	1	2	3	4	5
PAID score 0-25 (controlled)	2	2	4	6	0	4
PAID score 26-50 (moderately controlled)	7	1	1	1	0	2
PAID score 51-75 (Uncontrolled)	3	2	1	3	0	1
PAID score 76-100 (Severely Uncontrolled)	4	4	4	2	0	2

Above is a 2-way table in which A = number of days of exercise per week and B = PAID scores. In this probability calculation, I will be using the intersection of exercise days (5) and PAID score (0-25), which is 4. To calculate the probability, I will be using the following formula:

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

P(A|B) = 1/6 because there are 6 options for days of exercise 0-5.
P (B) = ¼ because there are 4 PAID score ranges.

To compute the conditional probability, I will use (4/56)/.25. This gives me the probability that 5 days of exercise will place a diabetic in the "controlled" PAID score range at 29% of the time! As a nurse, this tell me that, from this sample population, more exercise did not significantly lower their PAID score. There are other variables involved. These could be compliance to diet, other health issues, accessibility to supplies and medication. Another variable that this study did not consider is how long and how intense the workouts are each day of the week.