Exploration Of The Relationship Between Diabetes, Sleep, and Weight

Muhammad Abdul Mannan

23/08/2021

Abstract

This study was done to identify relationships between Diabetes, Sleep and Weight and understand what the potential relationships imply. The dataset that was used was the Canadian Community Health Survey (CCHS) for 2017-2018 which was cleaned and cropped for analysis and used histograms and scatter plots to plot the data. For the analysis, a maximum likelihood estimation and Bayesian estimation along with their respective confidence/credible intervals was used to find estimators for parameters such as μ. Hypothesis testing was also conducted as well as a simple linear regression. From the analysis, we found that Diabetics seem to have less sleep and inconsistent sleep schedules compared to non-diabetics, The average age that diabetes is diagnosed is 52.33 years, On average 9.3% of Canadians are diabetic and, The younger one is when they are diagnosed with diabetes, the greater their weight. From these findings, we concluded that the diagnosis age for diabetes negatively influences weight (causes it to increase) and diabetes it will negatively influence sleep times and schedules (reduces sleep times and variable sleep schedules) as well.

Introduction

According World Health Organization (WHO) (n.d.), one person out of eleven has diabetes in the world and it is also a leading cause of death in the world. Diabetes is a sickness which is caused by high blood sugar and We know that changes in blood sugar levels can impact ones sleep, making it difficult to sleep (Editor, 2019). It is an issue which affects many and it can run in the family as well. In light of this, We are interested in exploring Diabetes and Sleep in Canadians as well as how weight influences diabetes.

We plan on using two specific methods to get estimators for our unknown parameters which are the Maximum Likelihood Estimation (MLE) which is of the Frequentist Framework and Bayesian Estimation from the Bayesian Framework. We will also accompany our estimators with a confidence/credible intervals, which are intervals which denotes an interval which the true unknown parameter can be in with a certain confidence level $(1-\alpha)$ or probability level respectively. Aside from this we will also conduct hypothesis tests to check the estimates against a hypothesis (a claim) and simple linear regression to compare two variables of interest. Through these methods and techniques, we will explore and analyze the relationship between Diabetes, Sleep and Weight and try to better understand the results.

The data we will use for the analysis is the Canadian Community Health Survey (CCHS) for 2017-2018 which was attained from ODESI, which is a repository website which contains thousands of data sets for various topics. We chose this data set as it is fairly recent and it has many observations which will lead to accurate results. The question we wish to analyze is what is the relationship between Diabetes and Sleep and how they influence each other. Based on the prior information, we should also see a difference in people with diabetes and people without it. We should also see a relationship between age that diabetes was diagnosed and sleep and weight.

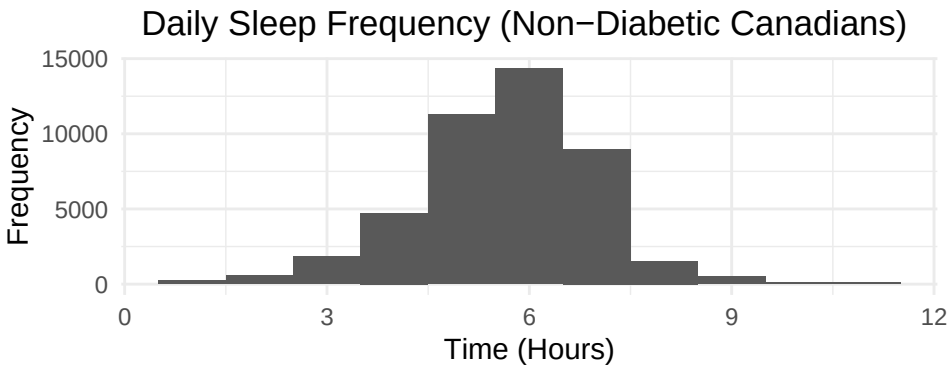
Data

As stated before, the data we will be using is the Canadian Community Health Survey (CCHS) for 2017-2018. This data set has 1051 variables with over 113,290 observations. In order to gather the variables of interest, we had to look over all the variables in the data set to identify the potential variables of interest, which was done over two days. After identifying the potential variables, we then renamed the variables as they were all in code and required a key to know what the code meant which was in a pdf file therefore, to omit the tedious task of finding the variable key, we decided to rename the variables to allow ease of use during the analysis. After we renamed the variables, we removed the observations which had not entered information for the variables of interest which was quickly done as each variable had a code (9, 96 or 99) which indicated that no information was entered. At the end, the dataset had 17 variables and 95,303 observations.

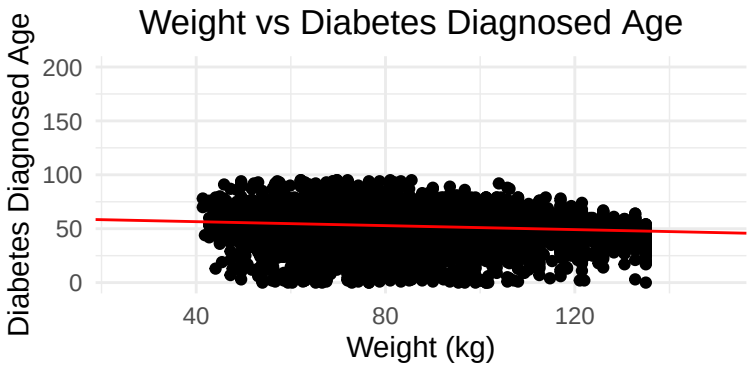
The following is a list of the important variables that were considered:

- Sex (DHH_SEX) - The gender of the individual
- Weight (HWTDGWTK) - The weight of individual in kilograms (kg)
- Diabetes (CCCG100) - The age at which diabetes was diagnosed
- Daily_sleep (SLPG005) - The amount of time an individual sleeps

The following is an example of some plots we will create and use in the upcoming analysis:



This Histogram plot is about the daily sleep frequencies of non-diabetic Canadians. On the x-axis is the time in hours and y-axis is the frequency of individuals. This histogram seems to be normally distributed and we can estimate the mean to be somewhere between 5 and 6 and spread to be 2.

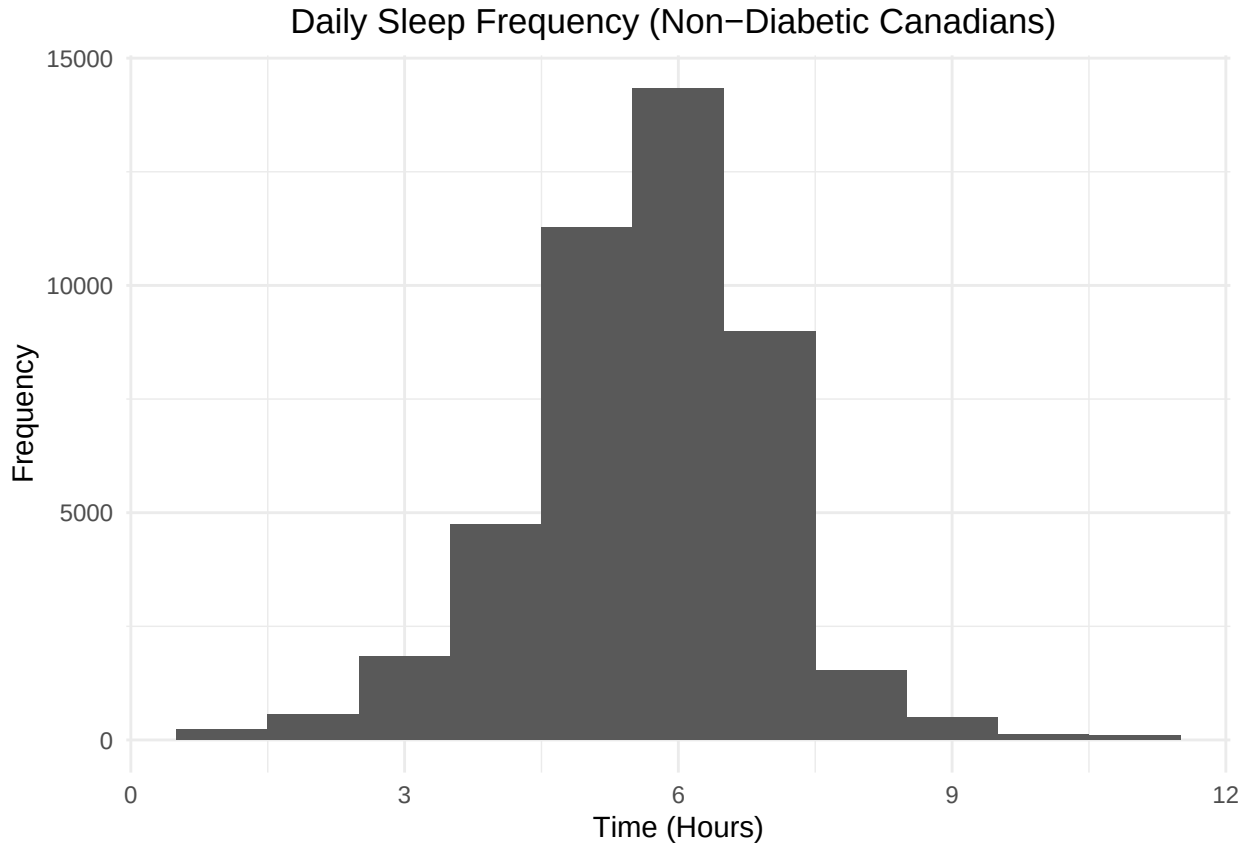


This is a scatter plot we will use in the simple linear regression that we will do. The x-axis is the weight of the individual in kilograms and the y-axis is the age when diabetes was diagnosed. The red line is the regression line and this reveals the relationship between the weight and diabetes diagnosed age.

This analysis will be done using R (programming language) which enables us to clean, manipulate, plot, and calculate various numerical summaries.

Methods - 1

Let $X_1, X_2, ..., X_n \sim N(\mu, \sigma^2)$ be the sleep times of the surveyed individuals without diabetes, which is normally distributed. Note that the data points each represented a range of values eg.(X = 1 -> 0 to 2 hours) so we utilized the impute method to randomly assign a value from the range to each data point. We want to find the Maximum Likelihood estimator for μ and a confidence interval for μ.



From our calculation of the maximum likelihood estimators (MLE) (see Appendix A), we got

$$\hat{\mu}_{mle} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

We know that $\hat{\mu}_{mle} = \bar{x}$ is an unbiased and consistent estimator for μ.

Now we calculate a 95% confidence interval for μ and since the data is normally distributed and σ^2 is unknown, we will use a T-distribution so then:

$$P(a \leq \bar{X} \leq b) = 1 - \alpha)$$

$$P(-t_{n-1,\alpha/2} \leq \frac{\bar{X} - \mu}{s/\sqrt{n}} \leq t_{n-1,\alpha/2}) = 0.95$$

$$\bar{X} \pm t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}$$

Now we calculate $\bar{X}$ and S (sample standard deviation) and $t_{n-1,\alpha/2}$ and find them to be:

$$\bar{X} = 5.6560 \;\mid\; S = 1.3471 \;\mid\; t_{n-1,\ \alpha/2} = -1.6449$$

So then using the formula that we derived, we calculate the 95% confidence interval:

$$5.6560 \;\pm\; 1.6449\ \frac{1.3471}{\sqrt{44230}}$$

So then the estimators $\bar{X}$, $\hat{\sigma}$, and 95% confidence interval is

$$\bar{x} = 5.6560 \;\mid\; S = 1.3471$$

$$95\%\ CI : (5.6455, 5.6665)$$

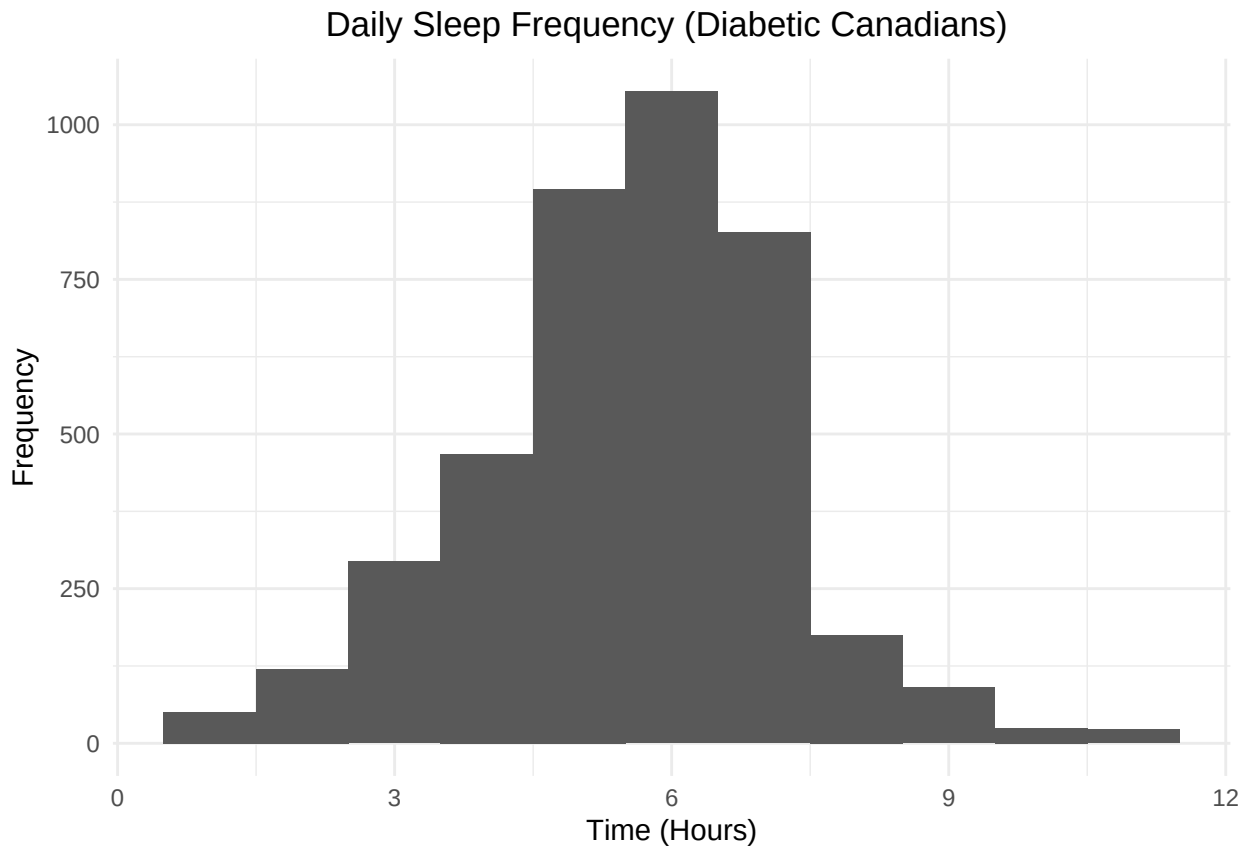
Results - 1

The Maximum Likelihood estimator for μ was $\hat{\mu}_{mle} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$ (see Appendix A). These results are reasonable as they were derived from a normal distribution and we have already noted that $\bar{x}$ is an unbiased and consistent estimator for μ.

To calculate the confidence interval for μ (the average time a non-diabetic person sleeps (in hours)), we first found the average and standard deviation from the sample to be $\bar{X} = 5.6560$ hrs (which is also our MLE) and $S = 1.3471$ hrs. Then we found the confidence interval to be $95\%\ CI :\ (5.6455, 5.6665)$. So we are 95% confident that individuals sleep on average 5.6455 hrs to 5.6665 hrs per day. This result is reasonable as our sample size is very large (44230), the sample mean will be a very good approximation of μ and based on our plot, we can intuitively see this as well as the most of the individuals filling the surveys likely have a job which may account for why the average time is that low. Since our sample size is very large, our confidence interval is also quite tight which is also very good as μ has an even smaller range of values it take.

Methods - 2

Let X_i be the amount of sleep the ith individual gets (hours). Now we would like to find the average sleep (in hours) of diabetic Canadians to compare it to the overall average sleep times of non-diabetic patients to see whether people with diabetes sleep less compared to regular people.



Similar to the calculations of the average sleep times of Canadians and the Confidence interval, we compute the 95% confidence interval for μ_D (the average time (hours) a diabetic Canadian sleeps) as well as the point estimate for μ_D, which is also the MLE $\bar{X}_D$:

Since the data is normally distributed and σ^2 is unknown, we will use a T-distribution

$$\bar{X}_D \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$$

Now we calculate $\bar{X}_D$ and S (sample standard deviation) and $t_{n-1,\alpha/2}$ and find them to be:

$$\bar{X}_D = 5.554 \;\mid\; S = 1.6613 \;\mid\; t_{n-1,\ \alpha/2} = -1.6452$$

So then using the formula for the interval, we calculate the 95% confidence interval:

$$5.554 \;\pm\; 1.6452\ \frac{1.6613}{\sqrt{4021}}$$

So then the estimators $\bar{X}_D$, $\hat{\sigma}$, and 95% confidence interval is

$$\bar{x} = 5.554 \;\mid\; S = 1.6613$$

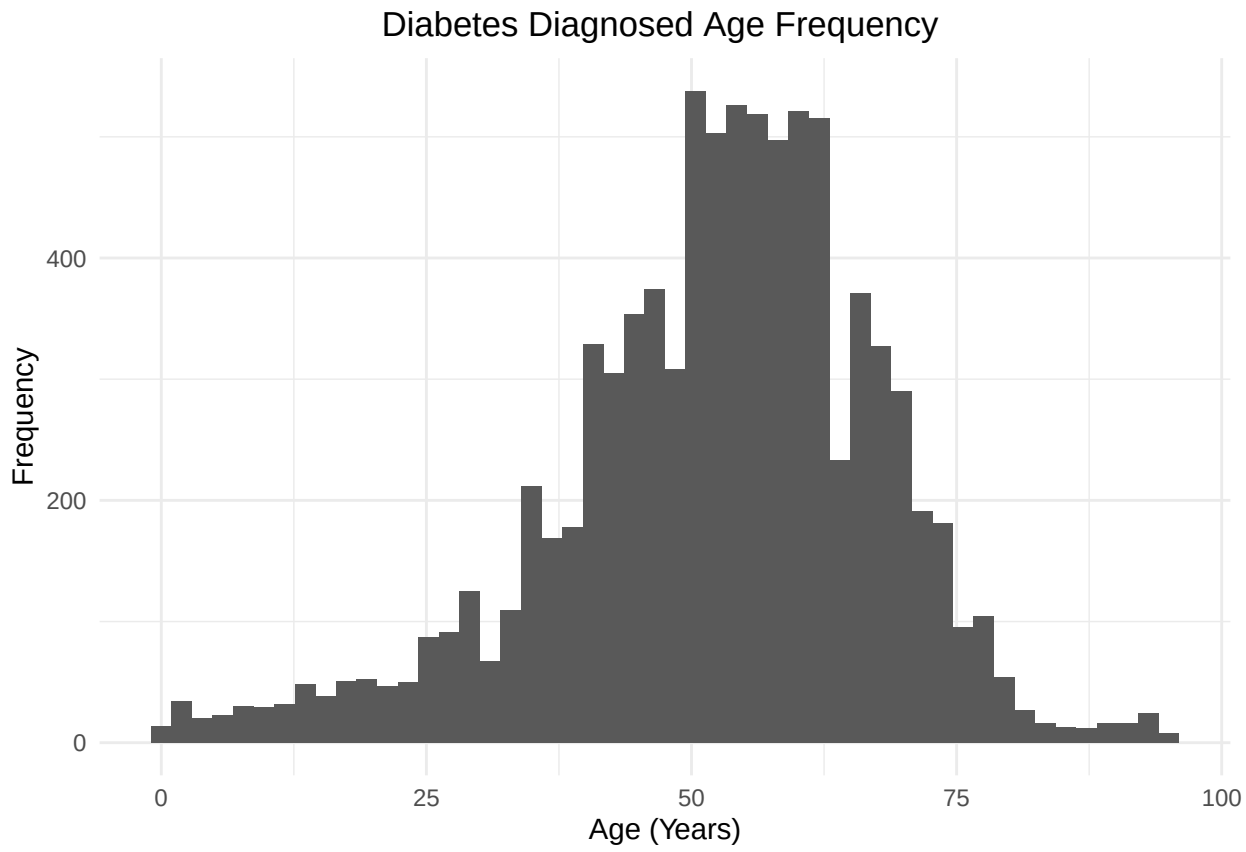
$$95\%\ CI_D : (5.5109, 5.5971)$$

Results - 2

The average sleep time for diabetic Canadians was $\bar{X}_D = 5.554$ and the 95% confidence interval $(5.5109, 5.5971)$. We can see that $\bar{X}_D = 5.554$ is less than $\bar{X} = 5.6560$ and the 95% $CI_D = (5.5109, 5.5971)$ is also more wider than 95% $CI = (5.6455, 5.6665)$ which shows that μ_D be in a larger range of values with 95% confidence. According to an article by Pacheco and Dr.Singh (2020), one quarter of diabetic patients report getting less than 6 hours of sleep per night and also states that they usually have irregular sleep times as well. These results are in-line with our analysis as we have seen from the results we observed.

Methods - 3

Out of our sample of 95,303 Canadians, the proportion who have diabetes is: 8773/95303 = 0.0921 or 9.21%. We want to analyze the age at which these Canadians (who already have diabetes) developed diabetes. We will first find an estimator for the average age μ using the Bayesian Framework and then find the 90% Bayesian Credibility Interval. Since the data values were categorical with a range (eg. X = 1 -> 0 to 11 years), we used the impute method to randomly assign values to the data for each respective range. The following plot depicts the distribution of the age at which people got diabetes.



Based on the plot, the data seems to be normally distributed so we will use a normal distribution and estimate the parameter μ using the Bayesian framework. The Bayesian model is another way to estimate a true parameter (in this case μ), it utilizes two key things to estimate the parameter, the given data and a prior (any information we may have prior to the estimation). Then we combine the 2 pieces of information to get the posterior distribution and from that, a beyesian estimator, which we can use to find an estimate for the parameter.

For the prior, based on the plot, μ seems to be between 6 and 11 so we will use a normal distribution with mean = 54 and variance = 9. We use the sample standard variance ($\hat{\sigma}^2$) for σ^2 as it is assumed to be known.

So then our Bayesian Model is:

$$X_1, ..., X_{8773} \sim N(\mu, \sigma_0^2)$$

$$\mu \sim N(54, 9)$$

Since both the likelihood (function attained based on the data) and prior are normally distributed, we know that the posterior distribution is also normally distributed which is:

$$\mu|x_1, ..., x_n, \sigma_0^2 \sim N(\frac{\mu_0/\tau_0^2 + n\bar{x}/\sigma_0^2}{1/\tau_0^2 + n/\sigma_0^2}, \frac{1}{1/\tau_0^2 + n/\sigma_0^2})$$

To get the beyesian estimator, we use the expectation of the posterior distribution:

$$\hat{\mu}_{\text{bayes}} = E(\mu|X_1, ..., X_n, \sigma_0^2) = \frac{\mu_0/\tau_0^2 + n\bar{x}/\sigma_0^2}{1/\tau_0^2 + n/\sigma_0^2}$$

Using the equation, we get the average to be:

$$\hat{\mu}_{\text{bayes}} = 52.33 \text{ yrs}$$

Now to get the 90% Bayesian Credible Interval, we find the posterior distribution to be:

$$\mu|x_1, ..., x_n, \sigma_0^2 \sim N(52.33, 0.027)$$

Then the 90% credible interval is found using the posterior distribution and $\alpha/2 = 0.05$ which is:

$$90\%\ CI:\ (52.3231, 52.4116)$$

Results - 3

So based on our data and prior, on average Canadians who have diabetes were first diagnosed with it at the age of $\hat{\mu}_{\text{bayes}} = 52.33$ ~53 years. This is a reasonable result and is close to what we had originally believed it to be and the plotted sample distribution seemed to be slightly left-skewed which implied that the mean would be towards the left of the mode (peak of data). One thing to note is that since the data has many observations (8773), the prior will have little effect on the estimator as much as the data would because as n increases, the less influence the prior has on the final estimate.

The Bayesian credible interval is $90\%\ CI:\ (52.3231, 52.4116)$. The true average age is contained between 52.3231 and 52.4116 years with 90% probability. This interval is acceptable and quite accurate and tight, because of the sample size we worked with (8773 observations). Using our estimator for the average age that diabetes is diagnosed, we can use these kinds of statistics and apply them to help individuals (with a family history of diabetes or are nearing this kind of age) take precautionary measures to prevent or reduce the chances of getting diabetes.

Methods - 4

In 2015, the estimated diabetic Canadians were 9.3% or 0.093 of all Canadians (Robyn L, n.d.). We believe that the estimated diabetic Canadians is less than 9.3% as in the previous model regarding the Canadians with diabetes, we found that the number of diabetic Canadians in the sample of 95303 Canadians is 8773. We want to do a one-sided hypothesis test to see how usual/unusual is our outcome being less than 9.3%.

Hypothesis

$$H_o = p = 0.093$$
$$H_a = p < 0.093$$

We will first find the sample statistic and then test it assuming the null hypothesis ($H_o = p = 0.093$) to be true.

Sample Statistic

$$\hat{p} = 8773/95303 = 0.0921$$
$$n = 95303$$

We know that $\hat{p} = N(p, \frac{p(1-p)}{n})$ if Central Limit Theorem applies which it does. So then assuming H_o is true, $\hat{p}$ has a distribution:

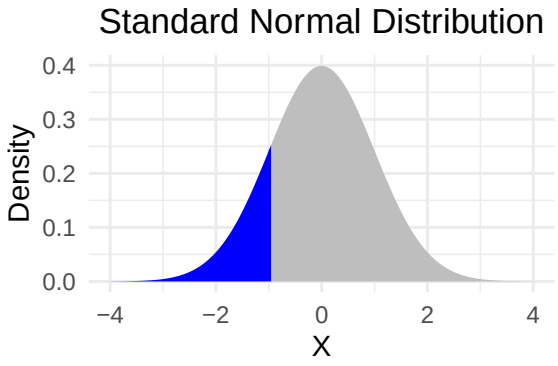
$$\hat{p} \sim N(0.093, \frac{0.093(1-0.093)}{95303})$$

Now we check How unusual is 0.0921 if the average is 9.3%.

Test statistic

To find the p-value, we first find the Z-value using:

$$Z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}} = \frac{0.0921 - 0.093}{\sqrt{\frac{0.093(0.907)}{95303}}} = -0.9566$$



Then the p-value is:

$$\text{p-value} = P(Z < -0.9566) = 0.1694$$

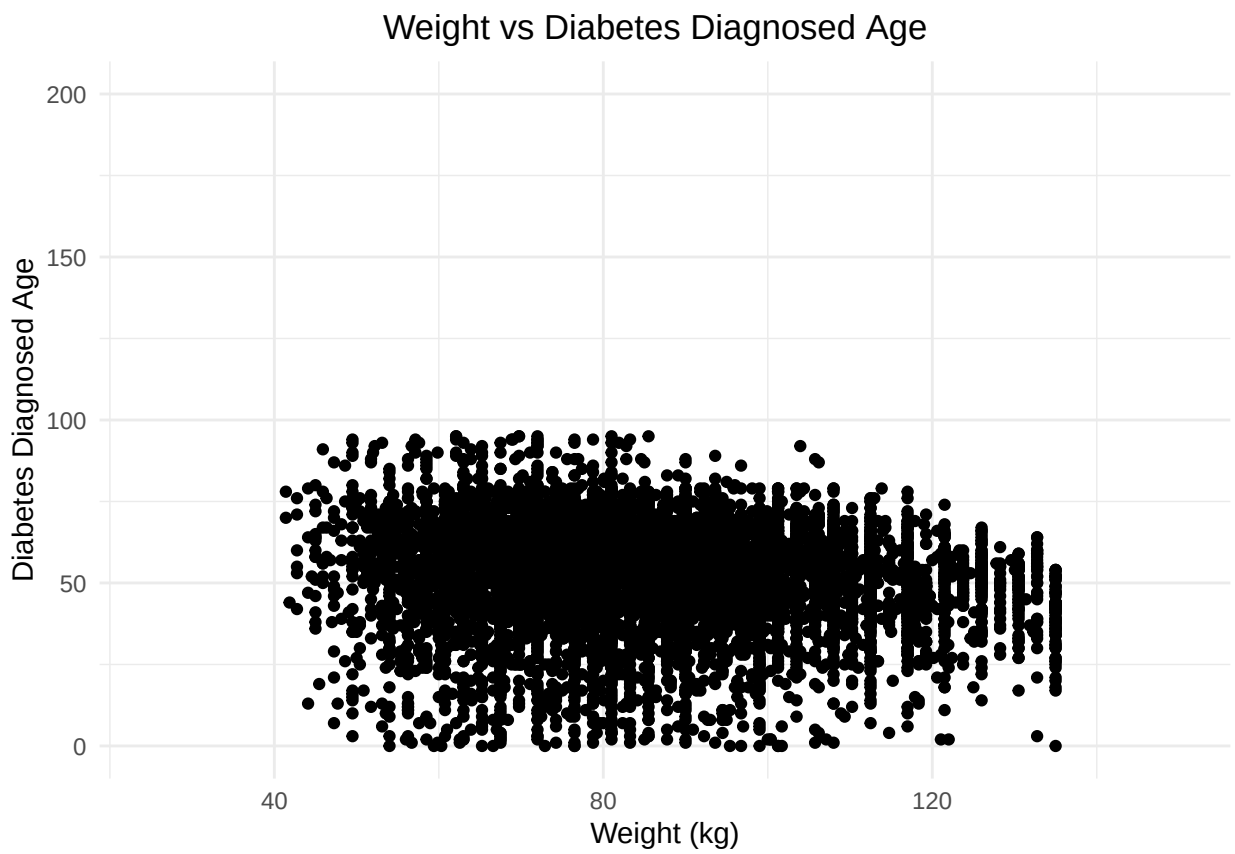
Results - 4

Interpretation & Conclusion

The p value is p-value = 0.1694. So there is a 16.94% chance that we will see a difference of least 0.09% from the average 9.3% in a sample of 95303 Canadians. Since the p-value is quite large, there is an absence of evidence against the null hypothesis H_o so the data is consistent with H_o.

Methods - 5

Diabetes also usually entails weight gain as blood sugar levels can become more difficult to control (Dowshen, 2018). We wanted to observe how the age at which diabetes was diagnosed relates to the weight of the individual. So we first plot our collect our data which is the ages at which diabetes was diagnosed and the weights of the individuals and we plot the data:



Since we see a linear relationship between weight and diagnosed age based on the plot, we will use a simple linear regression model which allows us to find and describe the linear relationship between two variables.

The model is defined to be:

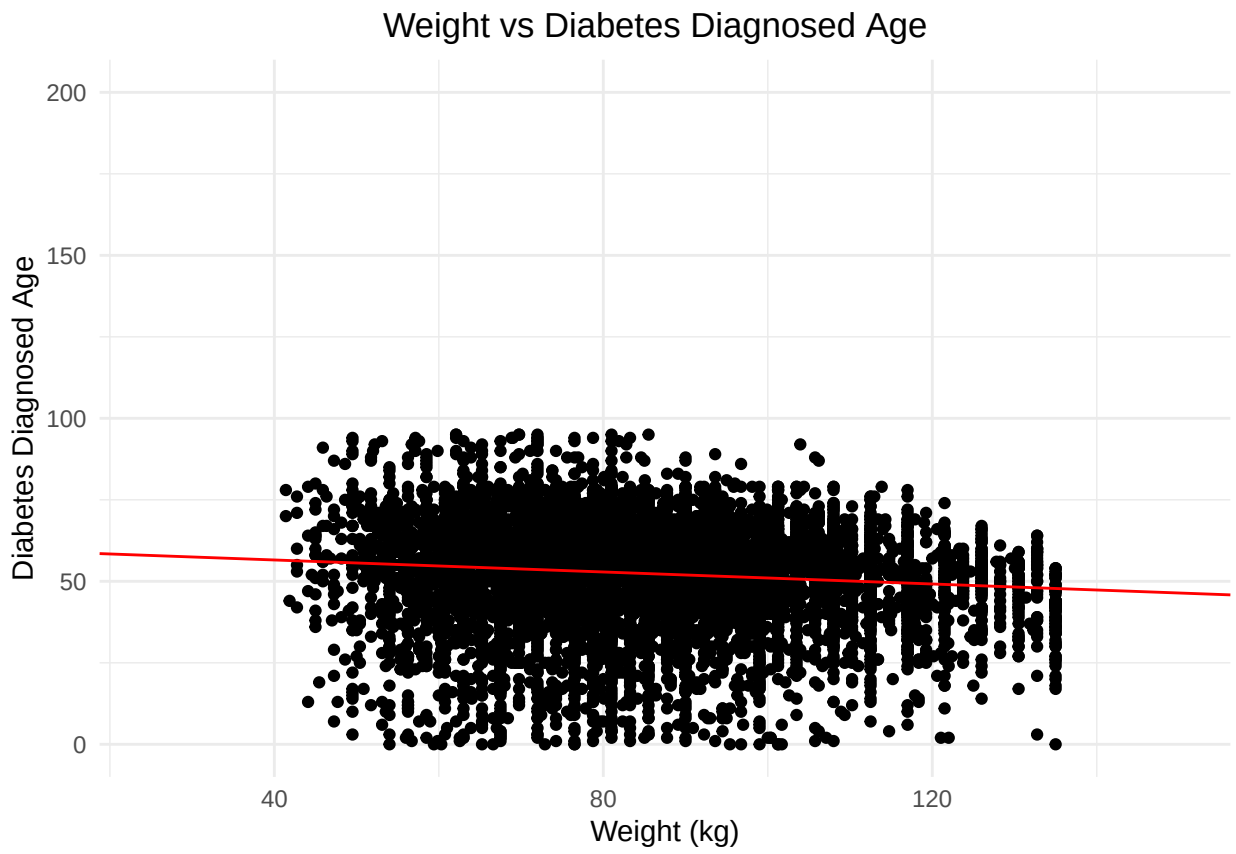
$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

where x_i is the ith weight (not random), and Y_i is the diagnosed age of the ith weight (random variable). We are interested in estimating β_1 which is the slope of the regression line, which also describes the linear relationship of the variables.

The estimate $\hat{\beta}_1$ is:

$$\hat{\beta}_1 = -0.09212$$

The following is the plot with the regression line (red):



Results - 5

The estimate, $\hat{\beta}_1$ for β_1 is equal to -0.09212. From the linear regression we conclude that on average, between 40 and 135 kg, the diagnosed age decreased by -0.09212 each weight increase. This shows is that the younger one was diagnosed with diabetes, the more their weight is and this suggests that an individual's blood sugar level may fluctuate causing one to retain or gain weight if diabetes is diagnosed at an early age however, to attain profound results, we must account for variables such as weight at time of diagnosis as being overweight, poor diet can increase the risks of diabetes as well.

Conclusions

In the first analysis, we analyzed the frequency of non-diabetic Canadians and their sleep times, we calculated the MLE for μ (the average sleep time of non-diabetic Canadians). We also found that the estimator for μ was $\bar{X} = 5.6560$ also found the 95% confidence interval for μ using the t-distribution since the data was normally distributed and noted that it was 95% $CI_D : 5.6455, 5.6665$. Note that this confidence interval is tight so we have a better idea of what μ can be and that the average sleep time of a non-diabetic Canadian is between 5.6455 and 5.6665 hours daily, with 95% confidence.

In the second analysis, this time, we analyzed the frequency of diabetic Canadians and their sleep times and we used $\bar{X}_D$ to be 5.554 and 95% Confidence interval, 95% $CI_D : 5.5109, 5.5971$. We noted that $\bar{X}_D$ was less than $\bar{X}$ and 95%CI_D was much wider compared to 95%CIm which depicted the results were in line with the article that we had previously mentioned (Pacheco et al., 2020). For both, the first and second analysis, we used the Frequentist framework.

In the third analysis, we wanted to analyze the age (in years) at which Diabetic Canadians were diagnosed with diabetes. We plotted a histogram to see the distribution of the various ages and found it to be normally distributed. We used a Bayesian Model and found that the beyesian point estimate for μ (the average age of diagnosis), $\hat{\mu}_{beyes} = 52.33$ years and got a 90% credible interval of (52.3231,52.4116).

In the fourth analysis, we conducted a hypothesis test to check the estimated proportion of diabetic Canadians, the claim (H_o) was that 9.3% of Canadians were diabetic, from our sample we got that H_a was less that 9.3% and attained a p-value of 0.1694 and deemed that there is no evidence against the null hypothesis and the data was consistent with H_o.

In the last analysis, we used a linear regression model to find the relationship between weight and the age of diagnosis for diabetes. We seemed to depict that the younger one was diagnosed with diabetes, the more their weight is. However, we stated that more research and analysis must be done to attain more profound results as we must take into account variable such as weight at the time of diagnosis which may have influenced our results.

Our initial hypothesis before the analysis was that we will see a relationship between age at which diabetes was diagnosed, weight and sleep. Based on the results, our hypothesis was indeed correct based on our sample. Based on our sample, We conclude that:

- Diabetics do seem to have less sleep daily and don't have a formal sleep schedule therefore Diabetes negatively influences sleep times.
- The estimated average age that diabetes is diagnosed is 52.33 years with a 90% credible interval (52.3231,52.4116).
- On average, 9.3% of Canadians are diabetic and from our sample, we got a similar average of 9.21% which is inline with the observations of the article by Robyn L (n.d)
- The younger one was diagnosed with diabetes, the more their weight is, therefore the age of diabetes diagnosis negatively affects one's weight.

Bibliography

Dowshen, S. (Ed.). (2018, February). Weight and diabetes (for parents) - nemours kidshealth. KidsHealth. https://kidshealth.org/en/parents/weight-diabetes.html#:~:text=Weight%20and%20Type%202%20Diabetes&text=Also%2C%20weight%20gain%20in%20people,move%20 glucose%20into%20the%20cells.

Editor, & Editor. (2019, January 15). Sleep can affect your blood sugar levels and your blood glucose control can also affect your sleep, which results in trouble sleeping. Diabetes. https://www.diabetes.co.uk/diabetes-and-sleep.html.

Robyn L., H. (n.d.). The Challenge of Diabetes. DiabetesCanadaWebsite. https://www.diabetes.ca/health-care-providers/ clinical-practice-guidelines/chapter-1#panel-tab_FullText.

Pacheco, D., & Singh, A. (2020, November 20). Diabetes and sleep: Sleep disturbances & coping. Sleep Foundation. https://www.sleepfoundation.org/physical-health/lack-of-sleep-and-diabetes#:~:text=Researchers%20believe%20that%20sleep%20restriction,having%20elevated% 20blood%20sugar12.

WHO. (n.d.). Diabetes. World Health Organization. https://www.who.int/health-topics/diabetes#tab=tab_1.

Appendix A

Since the data is normally distributed, we will find the MLE of μ.

$f(x_1, ..., x_n|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-1/2 \frac{x_i - \mu}{\sigma}}$

Likelihood Function:

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} f(x_1, ..., x_n|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-1/2 \frac{x_i - \mu}{\sigma}}$$

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} e^{-1/2 \frac{(x_i - \mu)^2}{\sigma}}$$

$$L(\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} e^{-1/2\sigma^2 \sum_{i=1}^{n} (x_i - \mu)^2}$$

Now we maximize:

$$l(\mu, \sigma^2) = \frac{-n}{2}\ log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$l(\mu, \sigma^2) = \frac{-n}{2}\ log(2\pi) - \frac{n}{2}\ log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$\frac{\partial\ l(\mu, \sigma^2)}{\partial\ \mu} = \frac{2}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0$$

Now we use partial derivatives to find μ

$$\frac{\partial\ l(\mu, \sigma^2)}{\partial\ \mu} = \sum_{i=1}^{n} (x_i) - n\mu = 0$$

$$\frac{\partial\ l(\mu, \sigma^2)}{\partial\ \mu} = \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

2nd derivative Test (with respect to μ)

$$\frac{\partial^2\ l(\mu, \sigma^2)}{\partial\ \mu^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = \frac{-\ n}{\sigma^2} < 0 \quad (\text{since } n \geq 0)$$

The second derivative passes therefore the Maximum Likelihood Estimator for μ is:

$$\hat{\mu}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$