

# kenna presentation

McKenna Hogan

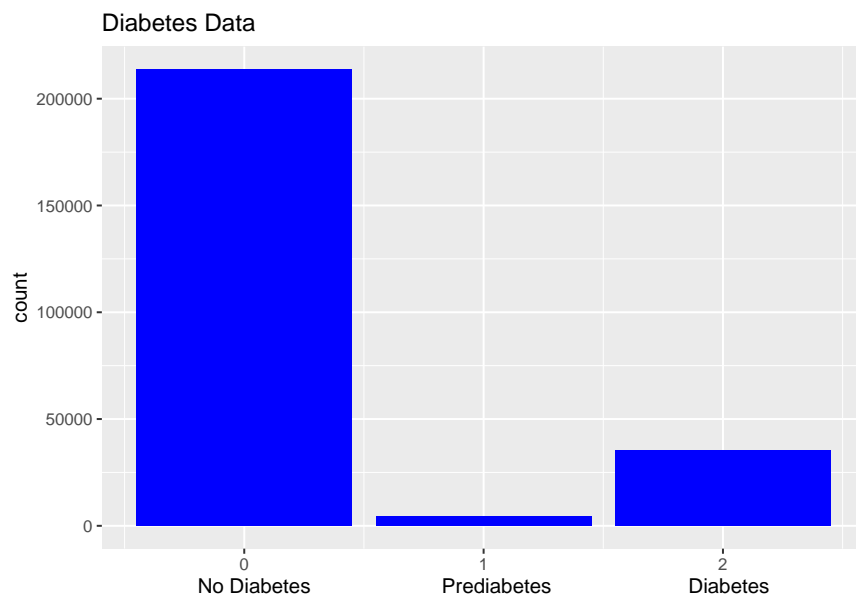
2022-12-14

## Dataset

We got our dataset from kaggle:

The dataset consisted of health indicators for people with no diabetes, prediabetes, and diabetes. We chose to focus on the variables HighBP, HighChol, BMI, AnyHealthcare, GenHlth, and Income out of the 21 factors in the dataset

- High BP, HighChol, and AnyHealthcare are binary
- GenHlth has 5 levels
- BMI and Income are continuous



## Methods

We decided to bootstrap our data because there was a massive difference in sample sizes for the 3 classifications of patients. We utilized all 5 types of bootstrapping.

Table 1: No Diabetes - Prediabetes

	mean	low percentile	high percentile	low basic	high basic
HighBP	-0.258	-0.272	-0.244	-0.272	-0.243
HighChol	-0.241	-0.255	-0.227	-0.257	-0.228
BMI	-2.984	-3.173	-2.783	-3.181	-2.791
AnyHealthcare	0.005	-0.002	0.011	-0.002	0.011
GenHlth	-0.603	-0.632	-0.572	-0.635	-0.575
Income	0.857	0.790	0.920	0.795	0.925

Table 2: Prediabetes - Diabetes

	mean	low percentile	high percentile	low basic	high basic
HighBP	-0.124	-0.272	-0.109	-0.272	-0.243
HighChol	-0.050	-0.064	-0.033	-0.065	-0.035
BMI	-1.210	-1.430	-0.981	-1.459	-1.009
AnyHealthcare	-0.015	-0.021	-0.008	-0.022	0.011
GenHlth	-0.316	-0.632	-0.285	-0.345	-0.281
Income	0.141	0.073	0.209	0.073	0.209

## Percentile and Basic

For the percentile and basic bootstrapping, I decided to focus on the difference in means and seeing if there was any difference between the two types of bootstrapping to check my assumptions about biasness. If there was a large difference between my methods, that would have meant there was bias in the sample. My calculations showed no bias so I wanted to make sure that my bootstrapping methods matched this finding. I used R=1000 for each bootstrap and found the difference in means for each variable for the 3 combinations of no diabetes, prediabetes, and diabetes.

## Normal and Studentized

## BCa

## Results

### Percentile and Basic

In my results, I found that there was a significant difference in means for most of my combinations. The only variable that didn't have consistent significant difference in means was the AnyHealthcare variable. I also found that the percentile and basic bootstraps were very similar which matched both my assumption and my findings for 0 SE.

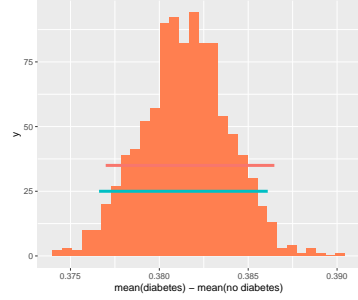
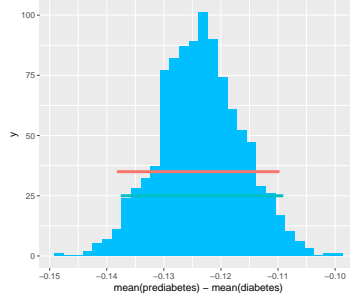
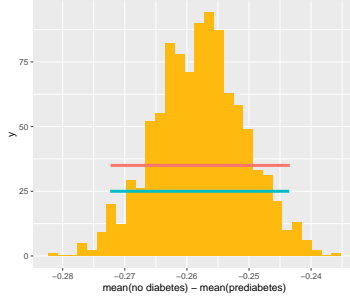
The following are tables showing the difference in means among the combinations of no diabetes, prediabetes, and diabetes for all the variables and for both kinds of bootstrapping:

The following are the histograms for each combination of difference in means for:

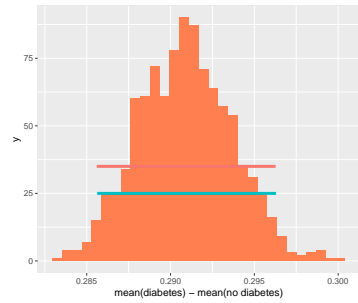
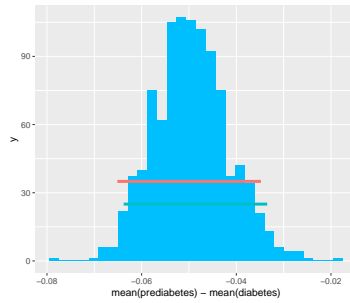
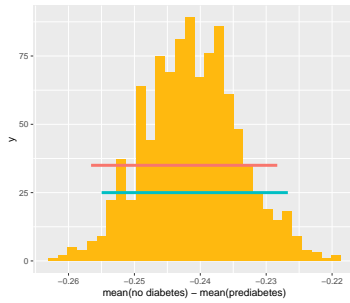
High Blood Pressure:

Table 3: Diabetes - No Diabetes

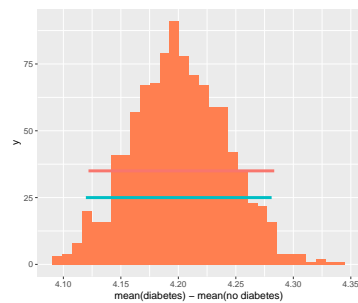
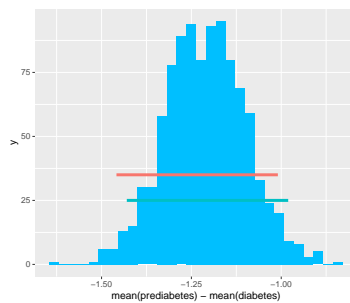
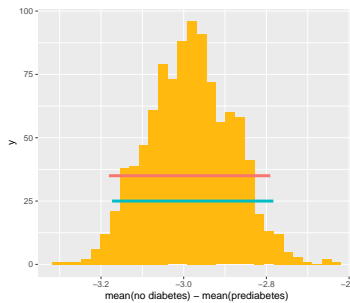
	mean	low percentile	high percentile	low basic	high basic
HighBP	0.381	0.377	0.386	0.377	0.386
HighChol	0.291	0.286	0.296	0.286	0.296
BMI	4.201	4.120	4.281	4.122	4.283
AnyHealthcare	0.010	0.008	0.012	0.008	0.012
GenHlth	0.919	0.907	0.930	0.907	0.930
Income	-0.998	-1.021	0.920	-1.024	-0.976



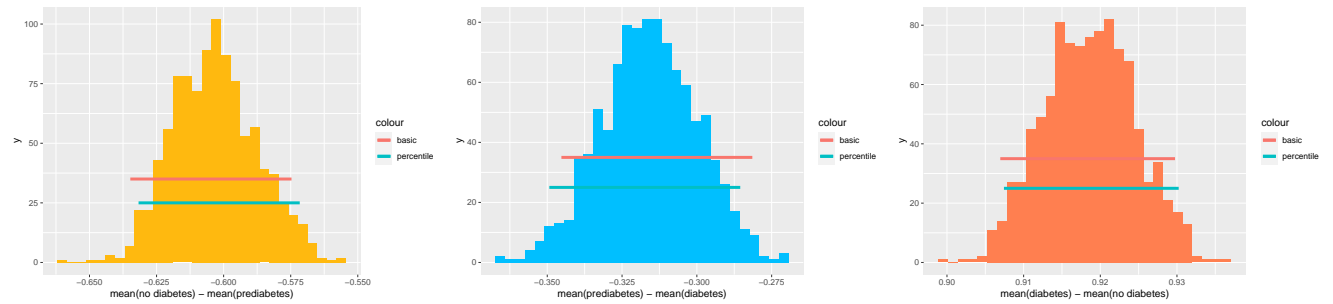
High Cholesterol:



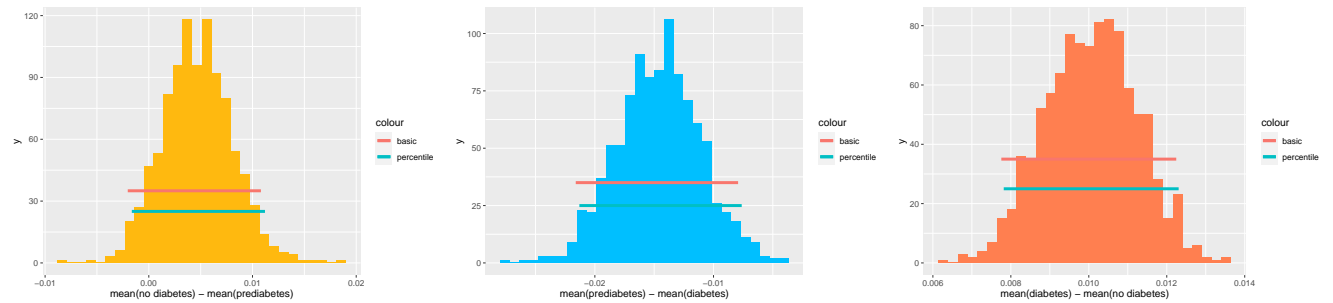
Body Mass Index (BMI):



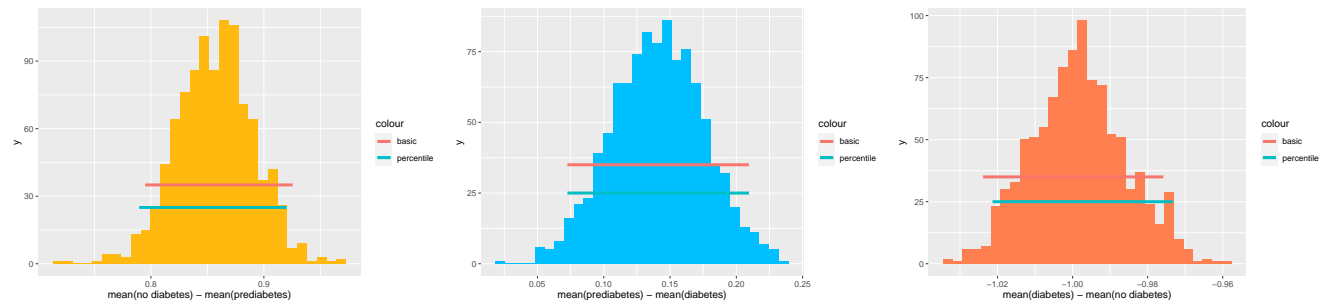
General Health:



Any Healthcare (Indicator):



Income:



As you can see from these histograms, 0 is almost never included in the interval. The only times 0 is in the interval is for AnyHealthcare. This means that the difference in means between no no diabetes and prediabetes, prediabetes and diabetes, and diabetes and no diabetes are significant for every variable but AnyHealthcare. The fact that the basic and percentile CI's are extremely similar show that there is no bias in the samples. This matched the outputted values for SE.

Normal and Studentized

BCa

Overall Results