Midterm 2

Brandon Keck

2024-11-20

```
library(readr)
diabetes <- read_csv("~/Desktop/diabetes.csv")
```

Section 1: Study Design

1. By focusing on subpopulation it allows researchers to gain valuable insights into predisposed indivuals along with environmental contributors to diabetes within a group known to have high diabetes rates. However, the result of this specific subpopulation may not be generalizable to the general public due to the fact that they may have different diets, and other health risk factors that may play a role in the development of diabetes.

2. The type of women that are most likely to be a part of the study are women who are genetically predisposed to Type 2 diabetes. The result of this may lead to selection bias. For instance, if women who already have diabetes within their family might be more apt to join the study because of awareness in experiencing syptoms and overall health. As opposed to women who may not necessarily have diabetes within their family. In other words what might happen is an overrepresentation of individuals with a heightened awareness of diabetes thereby skewing the results.

Section 2: EDA

```


summary(diabetes)
```

3. Revised

```
##  pregnancies           glucose      blood_pressure      insulin      
##  Length:500         Min.   :  0.0   Min.   :  0.00   Min.   :  0.00  
##  Class :character   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.:  0.00  
##  Mode  :character   Median :117.0   Median : 70.00   Median : 37.50  
##                     Mean   :120.3   Mean   : 68.45   Mean   : 78.41  
##                     3rd Qu.:139.0   3rd Qu.: 80.00   3rd Qu.:126.25  
##                     Max.   :198.0   Max.   :114.00   Max.   :846.00  
##       bmi             age           outcome     
##  Min.   : 0.00   Min.   :21.00   Min.   :0.000  
##  1st Qu.:27.40   1st Qu.:24.00   1st Qu.:0.000  
##  Median :32.40   Median :29.00   Median :0.000  
##  Mean   :32.04   Mean   :32.78   Mean   :0.352  
##  3rd Qu.:36.52   3rd Qu.:40.00   3rd Qu.:1.000  
##  Max.   :67.10   Max.   :70.00   Max.   :1.000  
```

My misunderstanding lies in that I did not notice the high insulin value of 846. I didn't really take a good long look at the dataset and so I didn't notice it until Dr. Moore mentioned it.

Revised The summary statistics reveal some unusual values that could affect our analysis. The minimum BMI is 0, which is not biologically possible and suggests either a data entry error or missing data incorrectly coded as 0. This misrepresentation could skew the dataset and lead to inaccurate conclusions. Similarly, the maximum insulin value of 846 is far above the expected range, with an average of 78.41. This large deviation suggests a potential outlier that could disproportionately affect measures like the mean and standard deviation. Both unusual values need to be addressed—BMI values of 0 should likely be excluded, and insulin outliers should be further investigated.

Justified The reason why I believe this to be a more correct response is because now I have had the time to observe the dataset in more detail. Looking at not only values that are low such as 0.00, but also values that are unusually high.

4. The summary of the dataset shows that there are some variables with zero values. For instance, bmi is a variable that has 0's as data. However, this is impossible to have a bmi of 0. This is likely due to missing data or some sort of error.If we were to do nothing to these values it could bias our summary and lead to a misrepresentation of the data. I would remove these values as they are more than likely misinterpreted data. Either by a data entry error or measurement error. The way that I would remove these values is by using the is.na function in R.

Remove missing values Revised

I knew to remove the values but in the moment during the midterm I panicked and didn't know how to do it off the top of my head. I only knew to possibly convert them to NA's instead.

```r
diabetes$bmi <- replace(diabetes$bmi, diabetes$bmi == 0, NA)

summary(diabetes)
```

```
##  pregnancies           glucose      blood_pressure      insulin
##  Length:500         Min.   :  0.0   Min.   :  0.00   Min.   :  0.00
##  Class :character   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.:  0.00
##  Mode  :character   Median :117.0   Median : 70.00   Median : 37.50
##                     Mean   :120.3   Mean   : 68.45   Mean   : 78.41
##                     3rd Qu.:139.0   3rd Qu.: 80.00   3rd Qu.:126.25
##                     Max.   :198.0   Max.   :114.00   Max.   :846.00
##
##       bmi             age           outcome
##  Min.   :18.20   Min.   :21.00   Min.   :0.000
##  1st Qu.:27.60   1st Qu.:24.00   1st Qu.:0.000
##  Median :32.70   Median :29.00   Median :0.000
##  Mean   :32.63   Mean   :32.78   Mean   :0.352
##  3rd Qu.:36.60   3rd Qu.:40.00   3rd Qu.:1.000
##  Max.   :67.10   Max.   :70.00   Max.   :1.000
##  NA's   :9
```

Justified The reason why I believe this to be correct is because rather than converting the values to NA's and then removing them, it would just be easier to remove the 0 values for the bmi variable. This is because this is really the variable of interest of the exam.

Data cleaning:

```
diabetes$outcome <- factor(diabetes$outcome, levels = c(0, 1), labels = c("No Diabetes", "Diabetes"))
```

```
diabetes %>%
  tbl_summary(
    by = outcome,
    missing = "no",
    statistic = list(
      all_continuous() ~ "{mean} ({sd})",
      all_categorical() ~ "{n} ({p}%)"
    ),
    label = list(
      pregnancies = "Number of Pregnancies",
      glucose = "Glucose Level",
      blood_pressure = "Blood Pressure",
      insulin = "Insulin Level",
      bmi = "BMI",
      age = "Age",
      outcome = "Diabetes Outcome"
    )
  )
```

5.

Characteristic	**No Diabetes** N = 324[1]	**Diabetes** N = 176[1]
Number of Pregnancies		
1-2	117 (36%)	30 (17%)
3+	155 (48%)	121 (69%)
None	52 (16%)	25 (14%)
Glucose Level	110 (26)	139 (33)
Blood Pressure	68 (18)	70 (23)
Insulin Level	73 (103)	89 (127)
BMI	31 (7)	35 (7)
Age	30 (11)	37 (11)

[1] n (%); Mean (SD)

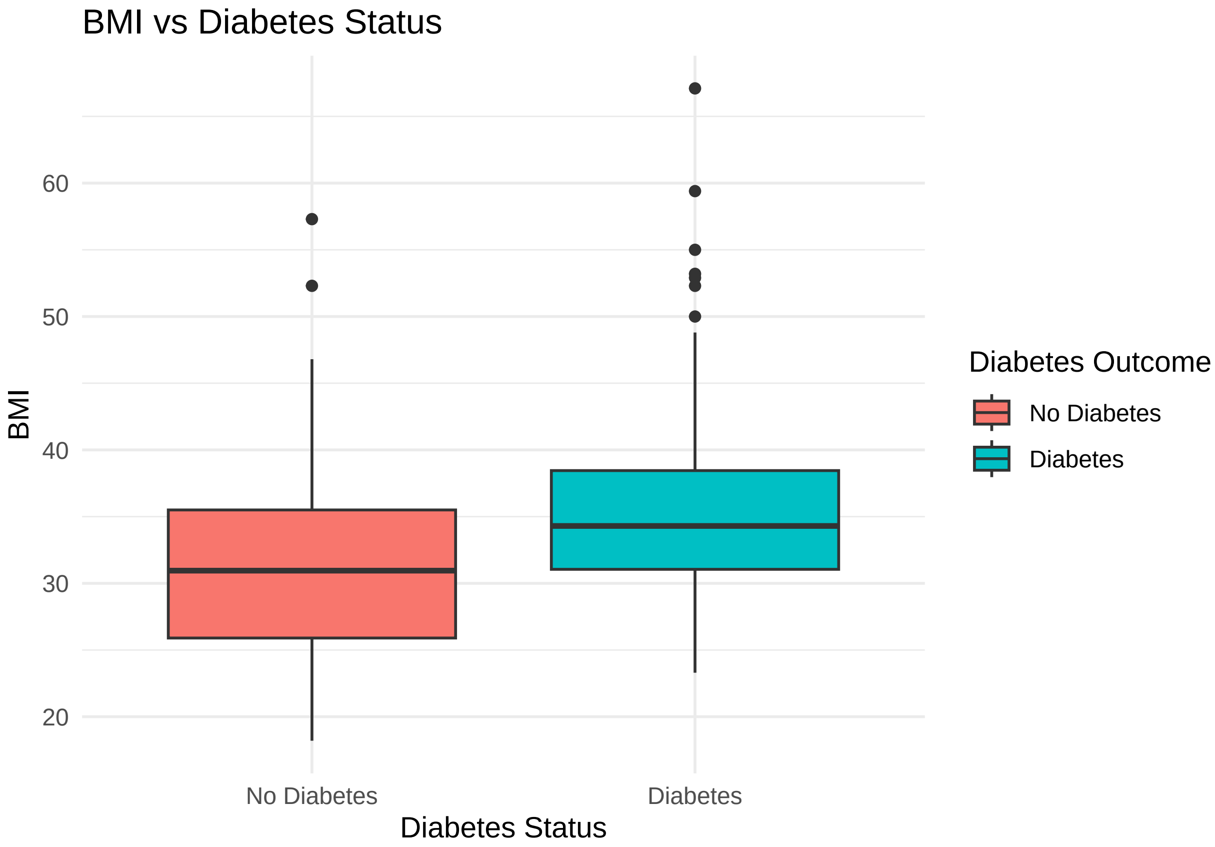
6. Revised From the table summary, the relationship between the number of pregnancies and diabetes status stands out. Among women with 1-2 pregnancies, only 17% have diabetes, compared to 36% of women without diabetes. This difference becomes even more pronounced for women with 3 or more pregnancies: 69% of women with diabetes fall into this category, compared to 48% of those without diabetes. This suggests a possible association between a higher number of pregnancies and increased diabetes prevalence. However, other variables also contribute to the observed differences between groups. For example, BMI tends to be higher among women with diabetes (mean = 35) compared to those without diabetes (mean

= 31), suggesting that higher BMI is an important risk factor. Insulin levels are elevated in the diabetes group (mean = 90) compared to the non-diabetes group (mean = 74), which is consistent with diabetes. Additionally, glucose levels are notably higher in the diabetes group, which is expected given its role as a primary diagnostic marker for diabetes. These patterns indicate that while pregnancies may be associated with diabetes, the interplay of other factors like BMI, insulin, and glucose levels cannot be ignored when understanding diabetes risk.

Misunderstanding I'm not super familiar with any type of diabetes and so I'm not exactly sure what to look out for when looking at this particular dataset. I was able to notice that pregnancies played some sort of role into diabetes however, after researching more into diabetes it is more clear to see what other variables impact diabetes.

Justified The reason why I believe this new response to be more correct is because I have looked more into the other variables and have a better understanding of how they influence diabetes.

```r
ggplot(diabetes, aes(x = outcome, y = bmi, fill = outcome)) +
  geom_boxplot() +
  labs(
    title = "BMI vs Diabetes Status",
    x = "Diabetes Status",
    y = "BMI",
    fill = "Diabetes Outcome"
  ) +
  scale_x_discrete(labels = c("No Diabetes", "Diabetes")) +
  theme_minimal()
```



7.

8. Revised From the created boxplot, the first thing we notice is the presence of outliers, particularly in the Diabetes group. These outliers suggest that some individuals with diabetes have exceptionally high BMIs, contributing to a greater spread in the data. The median BMI is higher for the Diabetes group compared to the No Diabetes group, indicating a difference in tendency. Although the table we created earlier shows that the average BMI for both groups is approximately the same (31 for No Diabetes and 35 for Diabetes), the boxplot shows that the median is in fact different.

Misunderstood I misunderstood that the line in the middle of the boxplot represents the median not the mean. This was just me rushing for time during the exam and not paying close attention to detail.

Justified The reason why I believe this to be correct is because now I have dealt with the unusual values and have a better boxplot to interpret. Because of this I believe my analysis to be more correct.

###Section 3: Data Analysis

9. Revised **Hypothesis:** $H_0 : p_{under30} = p_{30+}$

$H_A : p_{under30} \neq p_{30+}$

```r
diabetes <- diabetes %>%
  mutate(age_group = ifelse(age < 30, 0, 1))  # Create age_group variable
diabetes$age_group <- factor(diabetes$age_group,
                             levels = c(0, 1),
                             labels = c("<30", ">=30"))  # Convert to factor

# Generate a contingency table with margins
addmargins(table(diabetes$age_group, diabetes$outcome))
```

```
##
##         No Diabetes Diabetes Sum
##   <30           213       54 267
##   >=30          111      122 233
##   Sum           324      176 500
```

Confidence Interval

```
n1 <- 267
n2 <- 233

phat1 <- 54/176
phat2 <- 122/176

n1*phat1; n1*(1-phat1)
```

```
## [1] 81.92045

## [1] 185.0795
```

```
n2*phat2; n2*(1-phat2)
```

```
## [1] 161.5114

## [1] 71.48864
```

```
se_phat <- sqrt((phat1*(1-phat1))/n1 + (phat2*(1-phat2))/n2)
se_phat
```

```
## [1] 0.04134429
```

```
z_stat <- (phat1 - phat2) / se_phat
z_stat
```

```
## [1] -9.34503
```

```
p_val <- pnorm(z_stat)
p_val
```

```
## [1] 4.593165e-21
```

Decision: With a p-value that is essentially 0 which is less than our significance value of $\alpha = 0.05$ we reject the H_0 that the true proportion between the two groups is the same.

Conclusion: We have enough evidence that the true proportion of diabetes in women under 30 is different than the true proportion of women aged 30 and older.

Misunderstood This is where I was rushing for time in order to complete the exam. I think I just spent more time on the first half rather than the last half. I wanted to make sure I had all hypothesis correct before moving on.

Justified The reason why I believe this problem to be correct now is because I had the time to go over the difference in proportion analysis and conduct the hypothesis analysis correctly. I had the start of the analysis and was now able to complete it.

10. Revised Hypothesis:

$H_0 : \mu_{Diabetes} = \mu_{No\ diabetes}$

$H_A : \mu_{Diabetes} \neq \mu_{No\ diabetes}$

Conditions: 1. Random sample from the population of women near Phoenix, Arizona. Sample size of both groups are larger than 30 observations. Both conditions are satisfied.

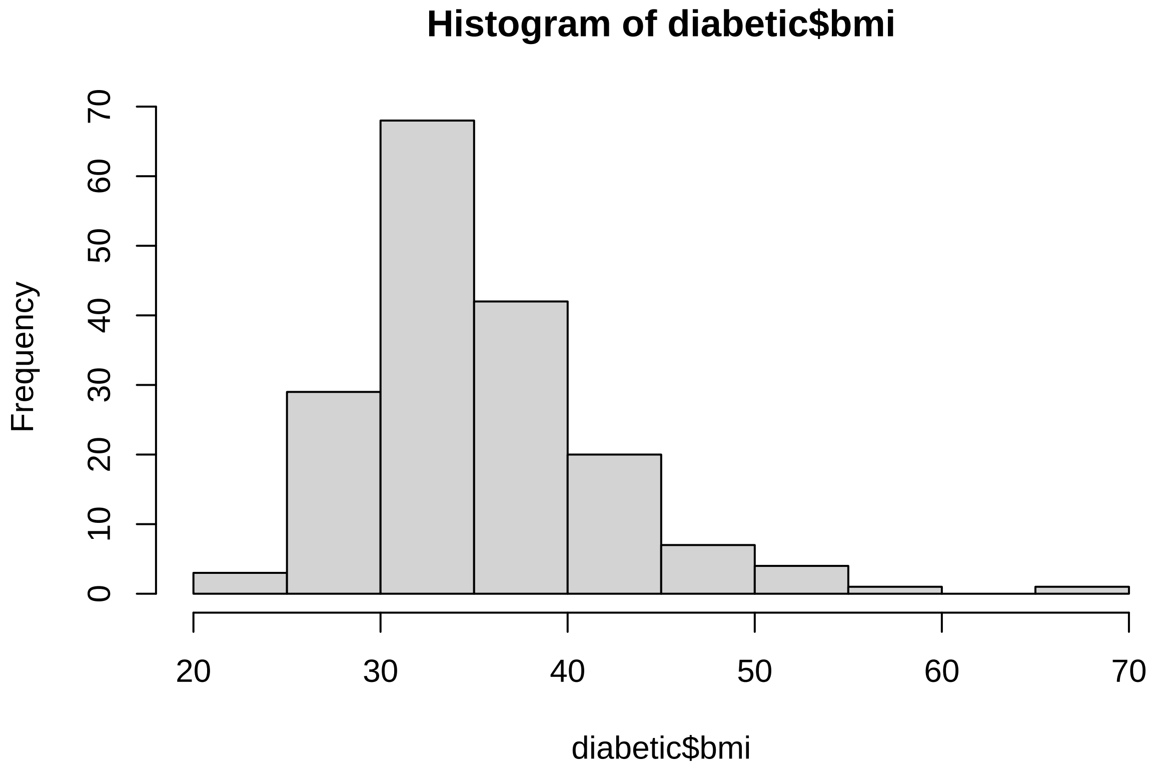
```
table(diabetes$outcome)
```

```
##
## No Diabetes    Diabetes
##         324         176
```

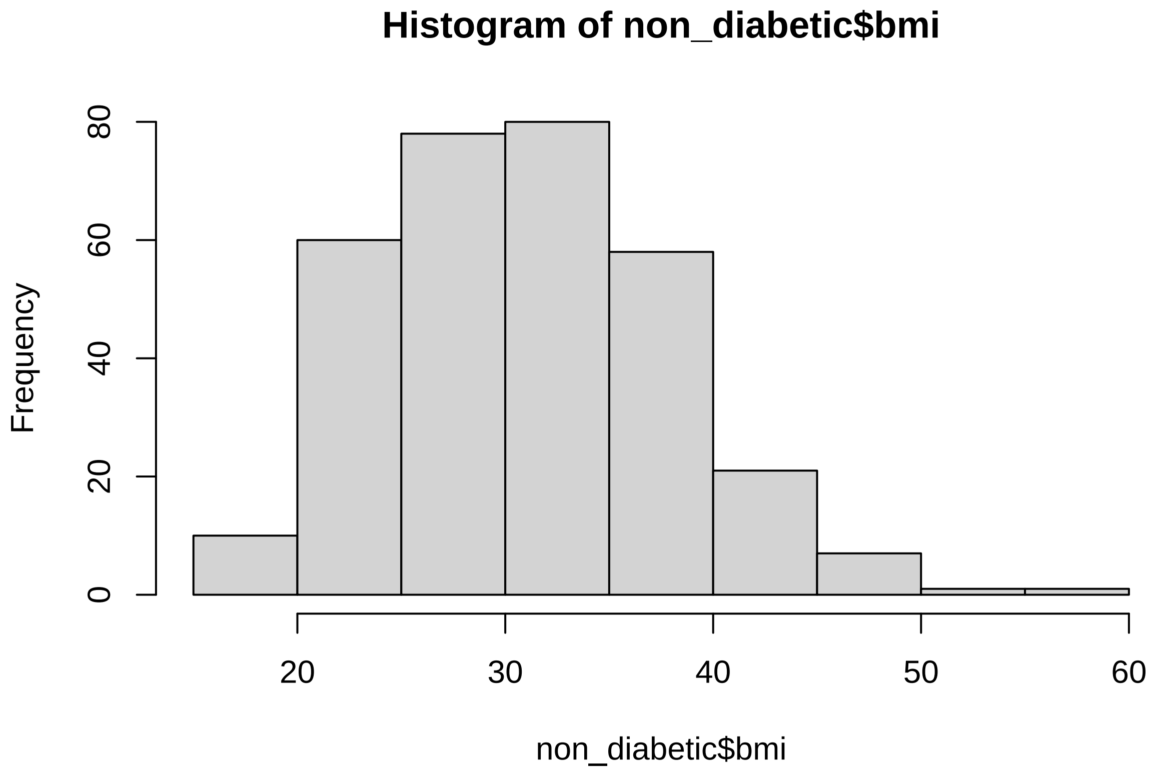
```
diabetic <- diabetes %>%
  filter(outcome == "Diabetes")

non_diabetic <- diabetes %>%
  filter(outcome == "No Diabetes")

hist(diabetic$bmi)
```



```
hist(non_diabetic$bmi)
```



```
mean_diabetic <- mean(diabetic$bmi, na.rm = TRUE)
mean_non_diabetic <- mean(non_diabetic$bmi, na.rm = TRUE)

mean_diabetic - mean_non_diabetic
```

```
## [1] 4.353958
```

```
t.test(diabetic$bmi, non_diabetic$bmi, conf.level = 0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  diabetic$bmi and non_diabetic$bmi
## t = 6.9087, df = 361.84, p-value = 2.212e-11
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  2.722036 5.985881
## sample estimates:
## mean of x mean of y
##  35.43086  31.07690
```

Decision: With a difference in means between the groups BMI being 4.353958 and since this value is in our confidence interval of 2.722036 and 5.985881 we reject H_0

Conclusion: We have enough evidence that the true mean of women with diabetes is different from the mean for women without diabetes.

Misunderstood Again I was rushing towards the end of the midterm. Here I was more focused on making sure I had the right hypothesis test before moving forward.

Justified The reason why this is now correct is because I was able to spend the time conducting the hypothsis test for a difference in means. Along with the helpful suggestion from Dr. Moore I believe this to be correct.

11. Revised **Hypothesis:** H_0 : There is no association between the number of pregnancies and diabetes status

H_A : There is an association between the number of pregnancies and diabetes status.

Check conditions:

```
pregnancy_table <- table(diabetes$pregnancies, diabetes$outcome)

pregnancy_table
```

```
##
##        No Diabetes Diabetes
##   1-2          117       30
##   3+           155      121
##   None          52       25
```

```
test1 <- chisq.test(pregnancy_table)

test1
```

```
##
##  Pearson's Chi-squared test
##
## data:  pregnancy_table
## X-squared = 23.387, df = 2, p-value = 8.349e-06
```

Conclusion: We have enough evidence that the number of pregnancies and diabetes status are not independent.

12. Revise Individuals with diabetes are generally older than those without diabetes. What this means is that age has a large influence on the outcome of diabetes.

When it comes to BMI and Diabetes individuals with diabetes have larger BMI's than of those who do not have diabetes. This makes sense as BMI is a risk factor for when it comes to diabetes.

Finally higher pregnancies are associated with higher prevelance of diabetes. When comparing women with no pregnancies or 1-2 pregnancies have a lower prevelance of diabetes compared to women who have been pregnant 3+ times. What this suggests is that multiple pregnancies may contribute to diabetes risk.

13. Revised From personal experience having a medical condition what I believe to be the best policy to help reduce the rate of diabetes would be to promote a healthy lifestyle. Implementing nutritional programs that encourage healthier habits can help to reduce BMI and manage weight which is correlated with diabetes. Along with a healthy diet and exercise this can help to reduce the effect of diabetes and keep insulin levels in check.

14. I forgot how to deal with 0's in a dataset. I realized that I needed to convert them to NA's in order to easily deal with them but I couldn't figure it out in time.

15. E