STAT 630 Midterm 2

David Teng

Thursday, Nov 7th

```
library(openintro)
library(dplyr)
library(ggplot2)
library(knitr)
library(gtsummary)
```

```
diabetes <- read.csv("data/diabetes.csv")
```

1.

Focusing on a specific subpopulation, like Pima Indian women aged 21 and older, restricts the ability to apply the findings to a wider population. The outcomes may not be relevant to other age ranges and genders, limiting the generalizability of the study's conclusions on diabetes risk.

2.

Women who are more focused on their health or diabetes may be more inclined to join the study, leading to selection bias. This could cause an overrepresentation of individuals already worried about their well-being, potentially distorting the results and limiting their generalizability of the results.

3.

```
summary(diabetes)
```

```
##  pregnancies          glucose      blood_pressure      insulin
##  Length:500         Min.   :  0.0   Min.   :  0.00   Min.   :  0.00
##  Class :character   1st Qu.:100.0   1st Qu.: 64.00   1st Qu.:  0.00
##  Mode  :character   Median :117.0   Median : 72.00   Median : 22.50
##                     Mean   :121.5   Mean   : 69.75   Mean   : 75.13
##                     3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:122.75
##                     Max.   :199.0   Max.   :122.00   Max.   :680.00
##       bmi              age           outcome
##  Min.   : 0.00   Min.   :21.00   Min.   :0.000
##  1st Qu.:27.50   1st Qu.:24.00   1st Qu.:0.000
##  Median :32.00   Median :29.00   Median :0.000
##  Mean   :32.10   Mean   :33.26   Mean   :0.352
##  3rd Qu.:36.52   3rd Qu.:40.00   3rd Qu.:1.000
##  Max.   :59.40   Max.   :81.00   Max.   :1.000
```

Unusual values, such as extremely high or low glucose, blood pressure, or insulin levels, may indicate data entry errors or outliers. These values can skew summary statistics, leading to misleading conclusions. Finding them is important to ensure accurate analysis, as they might impact variance or introduce bias in estimates of diabetes risk factors.

4.

I will remove the unusual values, as they likely represent data entry errors or extreme outliers that could distort analysis results. Removing these values helps to maintain trustful means and standard deviations, reducing the risk of skewed conclusions. This approach ensures a more accurate estimation of diabetes risk factors.

```
knitr::kable(apply(diabetes,2, function(x) sum(is.na(x))))
```

	x
pregnancies	0
glucose	0
blood_pressure	0
insulin	0
bmi	0
age	0
outcome	0

```
# Remove rows with missing values
dia_clean <- na.omit(diabetes)
```

```
# Convert categorical variables to factors
diabetes <- diabetes %>%
  mutate(
    pregnancies = as.factor(pregnancies),
    outcome = as.factor(outcome)
  )

# Handle unusual values
diabetes_cleaned <- diabetes %>%
  mutate(
    glucose = ifelse(glucose == 0, NA, glucose),
    blood_pressure = ifelse(blood_pressure == 0, NA, blood_pressure),
    insulin = ifelse(insulin == 0, NA, insulin),
    bmi = ifelse(bmi == 0, NA, bmi)
  )
```

5.

```
diabetes %>%
  select(pregnancies, glucose, blood_pressure, insulin, bmi, age, outcome) %>%
  mutate(outcome = ifelse(outcome == "0", "No Diabetes", "Diabetes")) %>%
  tbl_summary(
    by = outcome,
    digits = list(all_continuous() ~ c(2, 2)),
    statistic = all_continuous() ~ "{mean} ({sd})"
  ) %>%
  modify_header(label ~ "**Variable**") %>%
  modify_caption(caption = "Descriptive Statistics of Diabetes Dataset Stratified by Outcome")
```

6.

In the table, those diagnosed with diabetes tend to have higher mean glucose levels, BMI, and insulin values compared to those without diabetes.These factors may have association with increased diabetes risk.

Table 2: Descriptive Statistics of Diabetes Dataset Stratified by Outcome

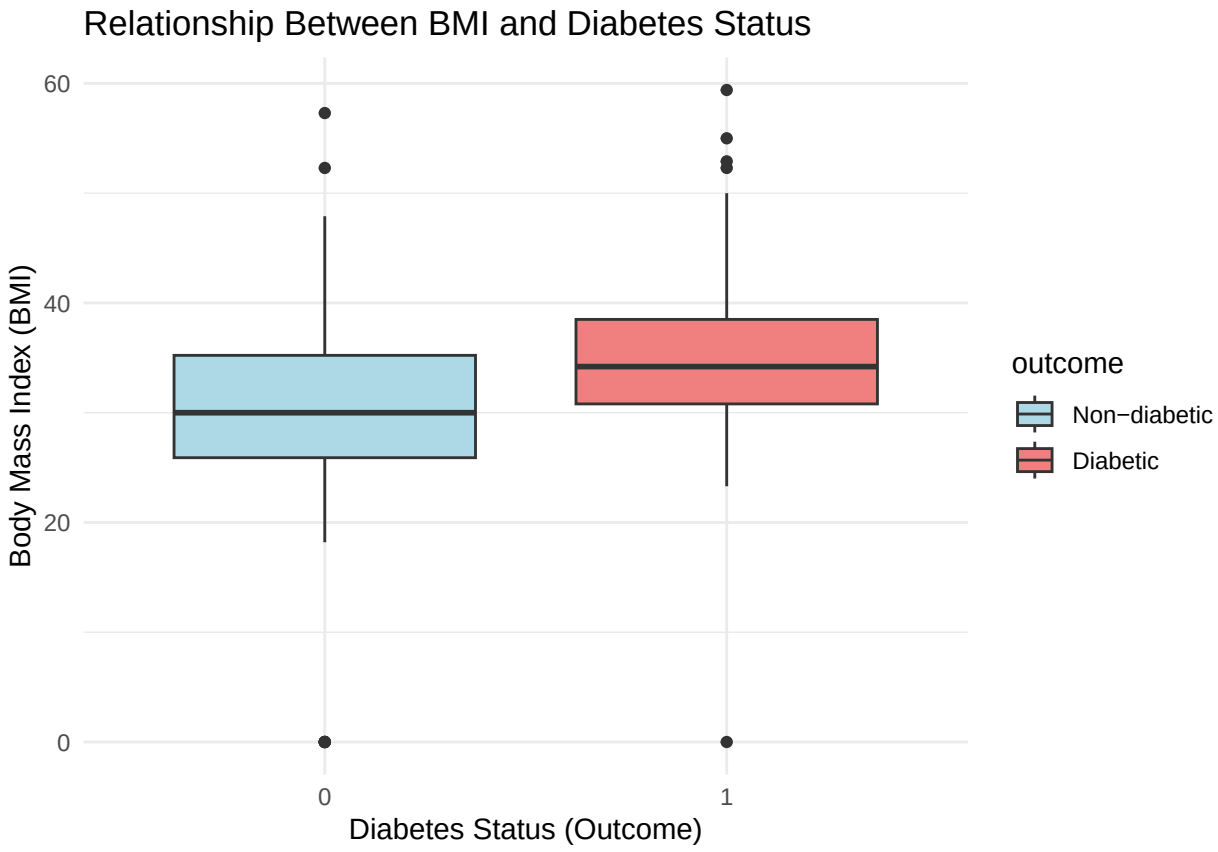
Variable	**Diabetes** N = 176[1]	**No Diabetes** N = 324[1]
pregnancies		
1-2	27 (15%)	127 (39%)
3+	124 (70%)	153 (47%)
None	25 (14%)	44 (14%)
glucose	143.87 (31.33)	109.35 (25.28)
blood_pressure	69.49 (23.03)	69.88 (15.95)
insulin	91.28 (125.68)	66.36 (89.89)
bmi	35.09 (7.09)	30.48 (7.62)
age	36.88 (10.81)	31.30 (11.86)

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

Additionally, women with diabetes appear to have a higher frequency of pregnancies. These patterns indicate potential relationships between these variables and diabetes prevalence.

7.

```
# Create a plot to visualize the relationship between BMI and diabetes status
ggplot(diabetes, aes(x = outcome, y = bmi, fill = outcome)) +
  geom_boxplot() +
  labs(
    title = "Relationship Between BMI and Diabetes Status",
    x = "Diabetes Status (Outcome)",
    y = "Body Mass Index (BMI)"
  ) +
  scale_fill_manual(values = c("lightblue", "lightcoral"), labels = c("Non-diabetic", "Diabetic")) +
  theme_minimal()
```



8.

The boxplot shows BMI distribution for non-diabetic (outcome = 0) and diabetic (outcome = 1) groups. The median BMI is higher in diabetics, with a slightly larger standard deviation, indicating more variability. Overall, higher BMI may be associated with diabetes.

9.

```r
# Create new variable for age categories
diabetes <- diabetes %>%
  mutate(age = ifelse(age < 30, "Under 30", "30 and older"))
```

```r
# 1: hypotheses
# H0: p1 - p2 = 0 (Proportion of diabetes in women under 30 is equal to those 30 and older)
# Ha: p1 - p2  0 (Proportion of diabetes in women under 30 is different from those 30 and older)

# Step 2: Choose significance level
alpha <- 0.05

# Step 3: Calculate the test statistic
prop_test <- prop.test(
  x = table(diabetes$outcome, diabetes$age)[2, ], # Number of successes (diabetic cases)
  n = table(diabetes$age), # Number of trials per group
  alternative = "two.sided"
)

prop_test
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  table(diabetes$outcome, diabetes$age)[2, ] out of table(diabetes$age)
## X-squared = 39.981, df = 1, p-value = 2.565e-10
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1894081 0.3594404
## sample estimates:
##    prop 1    prop 2
## 0.4957983 0.2213740
```

```
decision <- ifelse(prop_test$p.value < alpha, "Reject H0", "Fail to reject H0")
print(decision)
```

```
## [1] "Reject H0"
```

Decision: Reject H_0 Conclusion: We have enough evidence that proportion of diabetes in women under 30 is significantly different from those 30 and older.

10.

```
# Calculate summary statistics for BMI by diabetes outcome
bmi_summary <- diabetes %>%
  group_by(outcome) %>%
  summarise(
    mean_bmi = mean(bmi, na.rm = TRUE),
    sd_bmi = sd(bmi, na.rm = TRUE),
    n = n()
  )

bmi_summary
```

```
## # A tibble: 2 x 4
##   outcome mean_bmi sd_bmi     n
##   <fct>      <dbl>  <dbl> <int>
## 1 0           30.5   7.62   324
## 2 1           35.1   7.09   176
```

```
# Step 1: hypotheses
# H0:  1 - 2 = 0 (The average BMI for those with diabetes is equal to the average BMI for those without
# Ha:  1 - 2  0 (The average BMI for those with diabetes is different from those without)

# Step 2: Check necessary conditions
# Assume normality due to large sample size (Central Limit Theorem applies)

# Step 3: Calculate confidence interval for the difference in means
t_test_result <- t.test(
  bmi ~ outcome,
  data = diabetes,
  conf.level = 0.99,
  var.equal = FALSE
)
t_test_result
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  bmi by outcome
## t = -6.7661, df = 382.18, p-value = 4.984e-11
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 99 percent confidence interval:
##  -6.379420 -2.848694
## sample estimates:
## mean in group 0 mean in group 1
##        30.47685        35.09091
```

```r
# Calculate 99% confidence interval
ci <- t_test_result$conf.int
ci
```

```
## [1] -6.379420 -2.848694
## attr(,"conf.level")
## [1] 0.99
```

```r
# Step 4: Conclusion
# Interpretation
if (ci[1] > 0 | ci[2] < 0) {
  decision <- "Reject H0"
} else {
  decision <- "Fail to reject H0"
}

decision
```

```
## [1] "Reject H0"
```

Decision: Reject H_0 Conclusion: There is a significant difference in the average BMI between those with diabetes and those without diabetes.

11.

```r
# Step 1: hypotheses
# H0: There is no association between the number of pregnancies and diabetes status (independence)
# Ha: There is an association between the number of pregnancies and diabetes status (dependence)

# Step 2: Choose significance level
alpha <- 0.05

# Step 3: Calculate the test statistic

chisq <- chisq.test(diabetes$pregnancies, diabetes$outcome)

# Step 4: Draw conclusions based on p-value
chisq
```

```
##
##  Pearson's Chi-squared test
##
## data:  diabetes$pregnancies and diabetes$outcome
## X-squared = 32.218, df = 2, p-value = 1.009e-07
```

```r
# Step 5: Make decision
decision <- ifelse(chisq$p.value < alpha, "Reject H0", "Fail to reject H0")
decision
```

```
## [1] "Reject H0"
```

Decision: Reject H_0 Conclusion: There is an association between the number of pregnancies and diabetes status

12.

age: There is a significant difference in diabetes proportion between women under 30 and those aged 30 or older.

bmi: There is a significant difference in the average BMI between individuals with diabetes and those without.

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

13.

age Introduce targeted intervention strategies for women over 30, focusing on regular screenings and lifestyle interventions to lower diabetes risk within this age group.

bmi Implement community health initiatives to promote healthy weight and BMI through exercise programs, nutritional education, and support groups aimed at diabetes prevention.

pregnancies Develop educational programs and prenatal care strategies specifically for pregnant women to address diabetes risks, including monitoring glucose levels during and after pregnancy and providing gestational diabetes care.

14. No.
15. E - Excellent