pcos-analysis-report

aurellaa

2025-03-08

1. Introduction

Polycystic ovary syndrome (PCOS) is a common multisystem disorder affecting the endocrine, reproductive, and metabolic systems in women of reproductive age. It is characterized by hyperandrogenism, chronic anovulation, and polycystic ovaries.[1] Moreover, it is the leading cause of infertility and is associated with a higher risk of developing type 2 diabetes, cardiovascular disease, and mental health disorders such as anxiety and depression.[2]

PCOS affects 8-13% of women in Australia.[3] However, despite its high prevalence, PCOS remains widely misunderstood and underdiagnosed. Many women experience delays in diagnosis and inadequate medical support, leading to long-term health consequences.[4]

This report conducts statistical analysis to answer these three research questions:

1. Does BMI Differ Between Women With and Without PCOS?
2. Are Irregular Periods Associated With Higher PCOS Risk?
3. Do Hormone Levels Differ Between PCOS and Non-PCOS Women?

The dataset used in this report was obtained from Kaggle and originally published by Prasoon Kottarathil in 2020. It contains physical and clinical parameters relevant to PCOS diagnosis and infertility-related issues, collected from 10 hospitals across Kerala, India. The dataset consists of 541 observations and 45 variables, including PCOS diagnosis, age, BMI, hormone levels, menstrual cycle regularity, and other clinical markers.[5] For the purposes of this statistical analysis, a subset of 15 variables was selected based on their relevance to the research questions.

The objective of this report is to help improve early diagnosis, treatment strategies, and screening guidelines for PCOS. Additionally, this report hopes to contribute to better public health awareness about PCOS, helping more women receive timely and effective medical care – ultimately improving their quality of life.

2. Data Preparation and Cleaning

2.1 Loading Required Packages

Before beginning the analysis, we load all necessary R packages to ensure smooth execution of the code.

```
# loading in the necessary packages
library(ggplot2)
library(dplyr)
library(cowplot)
library(readxl)
library(effsize)
```

2.2 Loading the Dataset

```
# loading in the dataset
pcos_data <- read_excel("PCOS_data_without_infertility.xlsx", sheet = 2)
```

2.3 Data Cleaning and Recoding

The following outlines the data preparation steps taken to ensure accuracy and readability:

- The original dataset contains column names with spaces and special characters (e.g., /, ., ()). To prevent coding errors and improve readability, these names are modified by removing spaces and replacing special characters with underscores or meaningful alternatives.
- Since the original dataset contains 45 variables, we select only 15 variables that are relevant to our research questions. This ensures a more focused and efficient analysis.
- Certain variables are stored in incorrect formats. These are converted into the appropriate data types for accurate statistical analysis.
- Any missing values were omitted to prevent errors in analysis.
- Binary variables were converted into descriptive text for clarity (e.g., 0 = No, 1 = Yes for the PCOS variable).

```
# removing the spaces from the variable names to prevent coding errors
colnames(pcos_data) <- gsub(" ", "_", colnames(pcos_data))
colnames(pcos_data)
```

```
##  [1] "Sl._No"                "Patient_File_No."       "PCOS_(Y/N)"
##  [4] "Age_(yrs)"             "Weight_(Kg)"            "Height(Cm)"
##  [7] "BMI"                   "Blood_Group"            "Pulse_rate(bpm)"
## [10] "RR_(breaths/min)"      "Hb(g/dl)"               "Cycle(R/I)"
## [13] "Cycle_length(days)"    "Marraige_Status_(Yrs)"  "Pregnant(Y/N)"
## [16] "No._of_aborptions"     "I___beta-HCG(mIU/mL)"   "II____beta-HCG(mIU/mL)"
## [19] "FSH(mIU/mL)"           "LH(mIU/mL)"             "FSH/LH"
## [22] "Hip(inch)"             "Waist(inch)"            "Waist:Hip_Ratio"
## [25] "TSH_(mIU/L)"           "AMH(ng/mL)"             "PRL(ng/mL)"
## [28] "Vit_D3_(ng/mL)"        "PRG(ng/mL)"             "RBS(mg/dl)"
## [31] "Weight_gain(Y/N)"      "hair_growth(Y/N)"       "Skin_darkening_(Y/N)"
## [34] "Hair_loss(Y/N)"        "Pimples(Y/N)"           "Fast_food_(Y/N)"
## [37] "Reg.Exercise(Y/N)"     "BP__Systolic_(mmHg)"    "BP__Diastolic_(mmHg)"
## [40] "Follicle_No._(L)"      "Follicle_No._(R)"       "Avg._F_size_(L)_(mm)"
## [43] "Avg._F_size_(R)_(mm)"  "Endometrium_(mm)"       "...45"
```

```
# only taking a subset of variables
pcos_data_subset <- pcos_data |>
  select(`PCOS_(Y/N)`,
         `Age_(yrs)`,
         `Weight_(Kg)`,
         `Height(Cm)`,
         BMI,
         `Cycle(R/I)`,
```

```
         `FSH(mIU/mL)`,
         `LH(mIU/mL)`,
         `FSH/LH`,
         `TSH_(mIU/L)`,
         `AMH(ng/mL)`,
         `PRL(ng/mL)`,
         `PRG(ng/mL)`,
         )

colnames(pcos_data_subset)
```

```
##  [1] "PCOS_(Y/N)"  "Age_(yrs)"   "Weight_(Kg)" "Height(Cm)"  "BMI"
##  [6] "Cycle(R/I)"  "FSH(mIU/mL)" "LH(mIU/mL)"  "FSH/LH"      "TSH_(mIU/L)"
## [11] "AMH(ng/mL)"  "PRL(ng/mL)"  "PRG(ng/mL)"
```

```
# recoding variables to suit the analysis
# renaming some of the column names for convenience purposes
pcos_df <- pcos_data_subset |> rename(pcos = `PCOS_(Y/N)`,
                           age = `Age_(yrs)`,
                           weight = `Weight_(Kg)`,
                           height = `Height(Cm)`,
                           bmi = BMI,
                           cycle = `Cycle(R/I)`,
                           fsh = `FSH(mIU/mL)`,
                           lh = `LH(mIU/mL)`,
                           fsh_lh_ratio = `FSH/LH`,
                           tsh = `TSH_(mIU/L)`,
                           amh = `AMH(ng/mL)`,
                           prl = `PRL(ng/mL)`,
                           prg = `PRG(ng/mL)`,
                           )
glimpse(pcos_df)
```

```
## Rows: 541
## Columns: 13
## $ pcos         <dbl> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, ~
## $ age          <dbl> 28, 36, 33, 37, 25, 36, 34, 33, 32, 36, 20, 26, 25, 38, 3~
## $ weight       <dbl> 44.6, 65.0, 68.8, 65.0, 52.0, 74.1, 64.0, 58.5, 40.0, 52.~
## $ height       <dbl> 152.0, 161.5, 165.0, 148.0, 161.0, 165.0, 156.0, 159.0, 1~
## $ bmi          <dbl> 19.30000, 24.92116, 25.27089, 29.67495, 20.06095, 27.2176~
## $ cycle        <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 4, 2, 2, 4, 2, 2, 2, 2, 2, 2, ~
## $ fsh          <dbl> 7.95, 6.73, 5.54, 8.06, 3.98, 3.24, 2.85, 4.86, 3.76, 2.8~
## $ lh           <dbl> 3.68, 1.09, 0.88, 2.36, 0.90, 1.07, 0.31, 3.07, 3.02, 1.5~
## $ fsh_lh_ratio <dbl> 2.160326, 6.174312, 6.295455, 3.415254, 4.422222, 3.02803~
## $ tsh          <dbl> 0.68, 3.16, 2.54, 16.41, 3.57, 1.60, 1.51, 12.18, 1.51, 6~
## $ amh          <chr> "2.07", "1.53", "6.63", "1.22", "2.26", "6.74", "3.05", "~
## $ prl          <dbl> 45.16, 20.09, 10.52, 36.90, 30.09, 16.18, 26.41, 3.97, 19~
## $ prg          <dbl> 0.57, 0.97, 0.36, 0.36, 0.38, 0.30, 0.46, 0.26, 0.30, 0.2~
```

```
# correcting the data types of each variable
str(pcos_df)
```

```
## tibble [541 x 13] (S3: tbl_df/tbl/data.frame)
##  $ pcos        : num [1:541] 0 0 1 0 0 0 0 0 0 0 ...
##  $ age         : num [1:541] 28 36 33 37 25 36 34 33 32 36 ...
##  $ weight      : num [1:541] 44.6 65 68.8 65 52 74.1 64 58.5 40 52 ...
##  $ height      : num [1:541] 152 162 165 148 161 ...
##  $ bmi         : num [1:541] 19.3 24.9 25.3 29.7 20.1 ...
##  $ cycle       : num [1:541] 2 2 2 2 2 2 2 2 2 4 ...
##  $ fsh         : num [1:541] 7.95 6.73 5.54 8.06 3.98 3.24 2.85 4.86 3.76 2.8 ...
##  $ lh          : num [1:541] 3.68 1.09 0.88 2.36 0.9 1.07 0.31 3.07 3.02 1.51 ...
##  $ fsh_lh_ratio: num [1:541] 2.16 6.17 6.3 3.42 4.42 ...
##  $ tsh         : num [1:541] 0.68 3.16 2.54 16.41 3.57 ...
##  $ amh         : chr [1:541] "2.07" "1.53" "6.63" "1.22" ...
##  $ prl         : num [1:541] 45.2 20.1 10.5 36.9 30.1 ...
##  $ prg         : num [1:541] 0.57 0.97 0.36 0.36 0.38 0.3 0.46 0.26 0.3 0.25 ...
```

```
pcos_df <- pcos_df |>
  mutate(
    pcos = as.factor(pcos),  # Convert PCOS (Yes/No) to categorical
    cycle = as.factor(cycle),  # Regular/Irregular as categorical
    age = as.numeric(age),
    weight = as.numeric(weight),
    height = as.numeric(height),
    bmi = as.numeric(bmi),
    fsh = as.numeric(fsh),
    lh = as.numeric(lh),
    fsh_lh_ratio = as.numeric(fsh_lh_ratio),
    tsh = as.numeric(tsh),
    amh = as.numeric(amh),
    prl = as.numeric(prl),
    prg = as.numeric(prg),
  )
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `amh = as.numeric(amh)`.
## Caused by warning:
## ! NAs introduced by coercion
```

```
# an error was thrown. let's see what's happening...

pcos_df |> filter(is.na(as.numeric(amh)))
```

```
## # A tibble: 1 x 13
##   pcos    age weight height   bmi cycle   fsh    lh fsh_lh_ratio   tsh   amh
##   <fct> <dbl>  <dbl>  <dbl> <dbl> <fct> <dbl> <dbl>        <dbl> <dbl> <dbl>
## 1 0        37     56    152  24.2 2      2.91  0.35         8.31    16    NA
## # i 2 more variables: prl <dbl>, prg <dbl>
```

```
# unfortunately there seem to be some typographical errors in the data.
# for instance, row 307 amh is listed as "a". we should omit that.
```

```r
# checking for any other NAs
colSums(is.na(pcos_df))
```

```
##         pcos          age       weight       height          bmi        cycle
##            0            0            0            0            0            0
##          fsh           lh fsh_lh_ratio          tsh          amh          prl
##            0            0            0            0            1            0
##          prg
##            0
```

```r
pcos_df <- na.omit(pcos_df)
```

```r
# let's convert the binary variables values into being words instead
# of numerical values, for clarity.

# changing pcos values such that 0 = No, 1 = Yes
## checking for any possible data input errors
table(pcos_df$pcos)
```

```
##
##   0   1
## 363 177
```

```r
pcos_df <- pcos_df |>
  mutate(pcos = ifelse(pcos == 0, "No", "Yes"))
table(pcos_df$pcos)
```

```
##
##  No Yes
## 363 177
```

```r
# changing cycle values such that 2 = "regular" and 4 = "irregular"
## checking for any possible data input errors
table(pcos_df$cycle)
```

```
##
##   2   4   5
## 389 150   1
```

```r
## it seems that there is mistakenly one "5". we will get rid of this observation
pcos_df <- pcos_df |> filter(cycle != 5)
table(pcos_df$cycle)
```

```
##
##   2   4   5
## 389 150   0
```

```r
# pcos_df will now be down to 539 observations

pcos_df <- pcos_df |>
  mutate(cycle = ifelse(cycle == 2, "Regular", "Irregular"))

glimpse(pcos_df)
```

```
## Rows: 539
## Columns: 13
## $ pcos         <chr> "No", "No", "Yes", "No", "No", "No", "No", "No", "No", "N~
## $ age          <dbl> 28, 36, 33, 37, 25, 36, 34, 33, 32, 36, 20, 26, 25, 38, 3~
## $ weight       <dbl> 44.6, 65.0, 68.8, 65.0, 52.0, 74.1, 64.0, 58.5, 40.0, 52.~
## $ height       <dbl> 152.0, 161.5, 165.0, 148.0, 161.0, 165.0, 156.0, 159.0, 1~
## $ bmi          <dbl> 19.30000, 24.92116, 25.27089, 29.67495, 20.06095, 27.2176~
## $ cycle        <chr> "Regular", "Regular", "Regular", "Regular", "Regular", "R~
## $ fsh          <dbl> 7.95, 6.73, 5.54, 8.06, 3.98, 3.24, 2.85, 4.86, 3.76, 2.8~
## $ lh           <dbl> 3.68, 1.09, 0.88, 2.36, 0.90, 1.07, 0.31, 3.07, 3.02, 1.5~
## $ fsh_lh_ratio <dbl> 2.160326, 6.174312, 6.295455, 3.415254, 4.422222, 3.02803~
## $ tsh          <dbl> 0.68, 3.16, 2.54, 16.41, 3.57, 1.60, 1.51, 12.18, 1.51, 6~
## $ amh          <dbl> 2.07, 1.53, 6.63, 1.22, 2.26, 6.74, 3.05, 1.54, 1.00, 1.6~
## $ prl          <dbl> 45.16, 20.09, 10.52, 36.90, 30.09, 16.18, 26.41, 3.97, 19~
## $ prg          <dbl> 0.57, 0.97, 0.36, 0.36, 0.38, 0.30, 0.46, 0.26, 0.30, 0.2~
```

```r
pcos_df <- pcos_df |>
  mutate(
    pcos = as.factor(pcos),  # Convert PCOS (Yes/No) to categorical
    cycle = as.factor(cycle),  # Regular/Irregular as categorical
    age = as.numeric(age),
    weight = as.numeric(weight),
    height = as.numeric(height),
    bmi = as.numeric(bmi),
    fsh = as.numeric(fsh),
    lh = as.numeric(lh),
    fsh_lh_ratio = as.numeric(fsh_lh_ratio),
    tsh = as.numeric(tsh),
    amh = as.numeric(amh),
    prl = as.numeric(prl),
    prg = as.numeric(prg),
  )

pcos_df # the data is now clean and ready for analysis
```

```
## # A tibble: 539 x 13
##    pcos    age weight height   bmi cycle      fsh    lh fsh_lh_ratio   tsh   amh
##    <fct> <dbl>  <dbl>  <dbl> <dbl> <fct>    <dbl> <dbl>        <dbl> <dbl> <dbl>
##  1 No       28   44.6   152   19.3 Regular   7.95  3.68         2.16  0.68  2.07
##  2 No       36   65     162.  24.9 Regular   6.73  1.09         6.17  3.16  1.53
##  3 Yes      33   68.8   165   25.3 Regular   5.54  0.88         6.30  2.54  6.63
##  4 No       37   65     148   29.7 Regular   8.06  2.36         3.42 16.4   1.22
##  5 No       25   52     161   20.1 Regular   3.98  0.9          4.42  3.57  2.26
##  6 No       36   74.1   165   27.2 Regular   3.24  1.07         3.03  1.6   6.74
##  7 No       34   64     156   26.3 Regular   2.85  0.31         9.19  1.51  3.05
```

```
##  8 No       33   58.5   159   23.1 Regular   4.86  3.07         1.58 12.2   1.54
##  9 No       32   40     158   16.0 Regular   3.76  3.02         1.25  1.51  1
## 10 No       36   52     150   23.1 Irregul~  2.8   1.51         1.85  6.65  1.61
## # i 529 more rows
## # i 2 more variables: prl <dbl>, prg <dbl>
```

3. Summary Statistics

Before conducting further analysis, we examine the summary statistics of the cleaned dataset.

```
# summary stats
summary(pcos_df)
```

```
##   pcos          age             weight           height           bmi
##  No :363   Min.   :20.00   Min.   : 31.00   Min.   :137.0   Min.   :12.42
##  Yes:176   1st Qu.:28.00   1st Qu.: 52.00   1st Qu.:152.0   1st Qu.:21.64
##            Median :31.00   Median : 59.00   Median :156.0   Median :24.22
##            Mean   :31.44   Mean   : 59.62   Mean   :156.5   Mean   :24.31
##            3rd Qu.:35.00   3rd Qu.: 65.00   3rd Qu.:160.0   3rd Qu.:26.65
##            Max.   :48.00   Max.   :108.00   Max.   :180.0   Max.   :38.90
##        cycle          fsh                lh            fsh_lh_ratio
##  Irregular:150   Min.   :   0.210   Min.   :   0.020   Min.   :   0.0021
##  Regular  :389   1st Qu.:   3.335   1st Qu.:   1.025   1st Qu.:   1.4148
##                  Median :   4.860   Median :   2.300   Median :   2.1612
##                  Mean   :  14.647   Mean   :   6.492   Mean   :   6.9081
##                  3rd Qu.:   6.415   3rd Qu.:   3.680   3rd Qu.:   3.9493
##                  Max.   :5052.000   Max.   :2018.000   Max.   :1372.8261
##       tsh              amh              prl              prg
##  Min.   : 0.040   Min.   : 0.100   Min.   :  0.40   Min.   : 0.0470
##  1st Qu.: 1.480   1st Qu.: 2.010   1st Qu.: 14.52   1st Qu.: 0.2500
##  Median : 2.260   Median : 3.700   Median : 21.92   Median : 0.3200
##  Mean   : 2.961   Mean   : 5.624   Mean   : 24.38   Mean   : 0.6122
##  3rd Qu.: 3.570   3rd Qu.: 6.950   3rd Qu.: 29.93   3rd Qu.: 0.4550
##  Max.   :65.000   Max.   :66.000   Max.   :128.24   Max.   :85.0000
```

4. Research Question 1: Does BMI Differ Between Women With and Without PCOS?

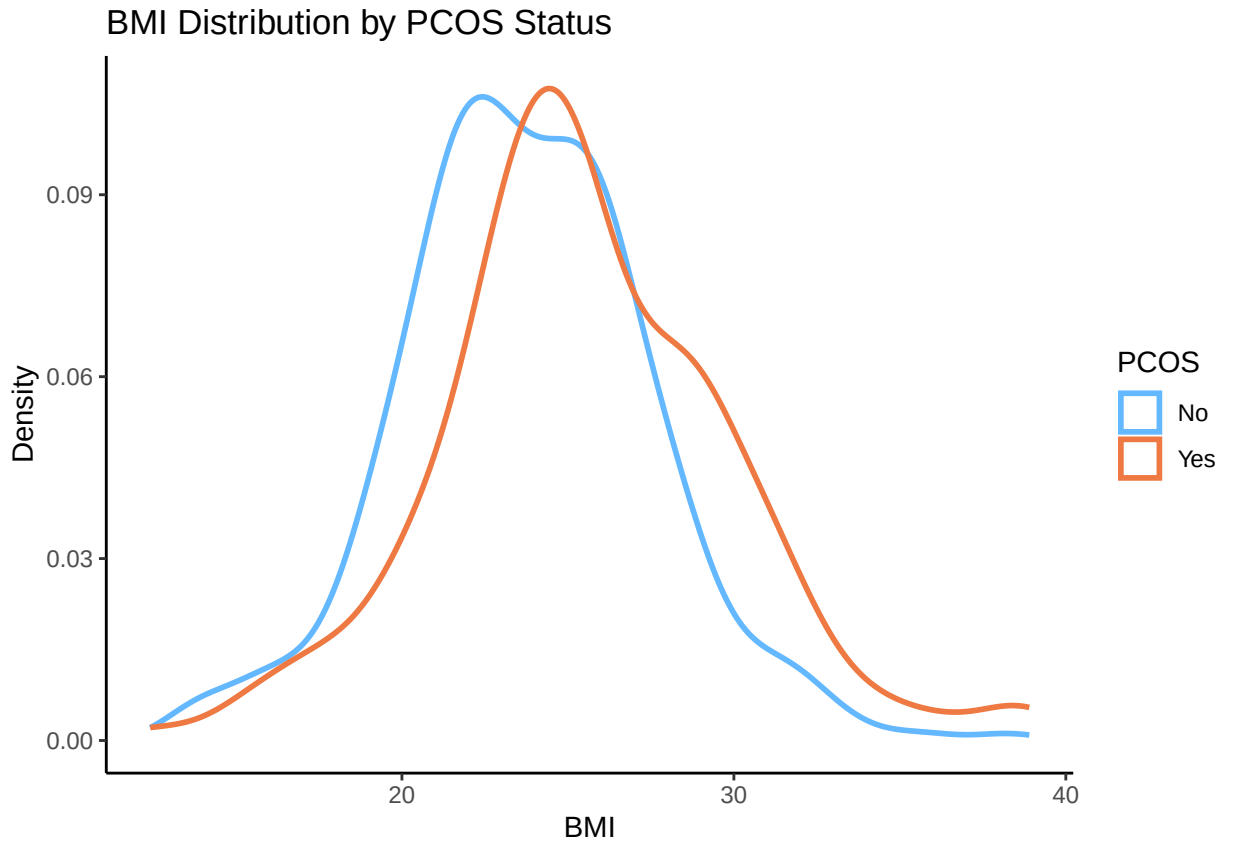
4.1 Setting a Global Theme

```
#### global theme
theme_set(
  theme_classic()
)
```

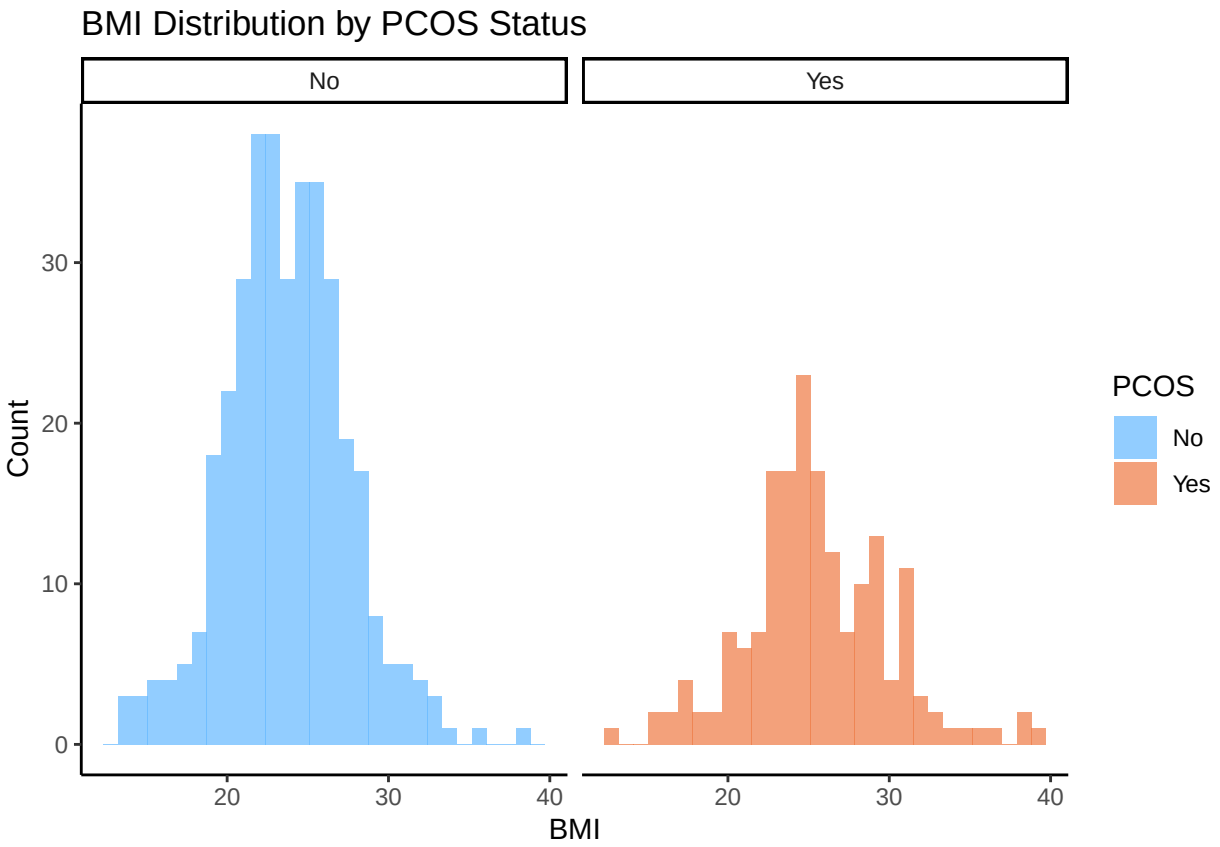
4.2 Visualising the Distribution of the Data

```r
#### density plot

ggplot(pcos_df, aes(x=bmi, color = pcos)) +
  geom_density(lwd = 1) +
  labs(x = "BMI", y = "Density", title = "BMI Distribution by PCOS Status", color = "PCOS") +
  scale_color_manual(values = c("No" = "steelblue1", "Yes" = "sienna2"))
```




```r
#### histogram faceted by pcos
ggplot(pcos_df, aes(x = bmi, fill = pcos)) +
  geom_histogram(alpha = 0.7, position = "identity", bins = 30) +
  facet_wrap(~ pcos) +
  labs(x = "BMI", y = "Count", fill = "PCOS", title = "BMI Distribution by PCOS Status") +
  scale_fill_manual(values = c("No" = "steelblue1", "Yes" = "sienna2"))
```



While the non-PCOS seems to be roughly normally distributed, it seems like the PCOS group could have some skewness.

4.3 Checking if the Assumptions for T-Test are Fulfilled

```
# might be skewness for pcos group
# checking for normality or non-normality with the shapiro wilk test
# if p < 0.05, the group is not normally distributed


shapiro.test(pcos_df$bmi[pcos_df$pcos == "No"])    # no pcos group
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pcos_df$bmi[pcos_df$pcos == "No"]
## W = 0.99272, p-value = 0.07472
```

```
# W = 0.99272, p-value = 0.07472
# it is normally distributed!

shapiro.test(pcos_df$bmi[pcos_df$pcos == "Yes"])  # has pcos group
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  pcos_df$bmi[pcos_df$pcos == "Yes"]
## W = 0.98482, p-value = 0.05312
```

```
# W = 0.9839, p-value = 0.05312
# this is just barely above 0.05, but n is much larger than 30 so its reasonable for us to do a t-test
```

Since the PCOS group may have some skewness but sample size > 30, we proceed with a t-test.

4.4 Conducting a t-test

To examine whether BMI differs between women with and without PCOS, we conduct an independent t-test:

```
t.test(bmi ~ pcos, data = pcos_df, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  bmi by pcos
## t = -4.4482, df = 302.36, p-value = 1.218e-05
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -2.4803244 -0.9588662
## sample estimates:
##  mean in group No mean in group Yes
##          23.74604          25.46564
```

The results indicate that the difference in BMI between the two groups is statistically significant ($p < 0.05$). Therefore, we reject the null hypothesis and conclude that BMI is significantly different between women with and without PCOS.

The mean BMI for women with PCOS is 23.75, and for women without PCOS, it's 25.47. This suggests that on average, women with PCOS have a higher BMI.

The 95% confidence interval (-2.48, -0.96) shows that we are 95% confident that the true difference in mean BMI between the two groups falls within this range. Since the confidence interval is entirely negative and does not include zero, we confirm that BMI is higher in the PCOS group.

However, while statistical significance confirms that a difference exists, we must assess whether this difference is meaningful in practice.

4.5 Cohen's d Test for Effect Size

To determine the magnitude of the difference, we calculate Cohen's d:

```
cohen.d(pcos_df$bmi ~ pcos_df$pcos)
```

```
##
## Cohen's d
##
## d estimate: -0.431446 (small)
```

```
## 95 percent confidence interval:
##      lower      upper
## -0.6137148 -0.2491771
```

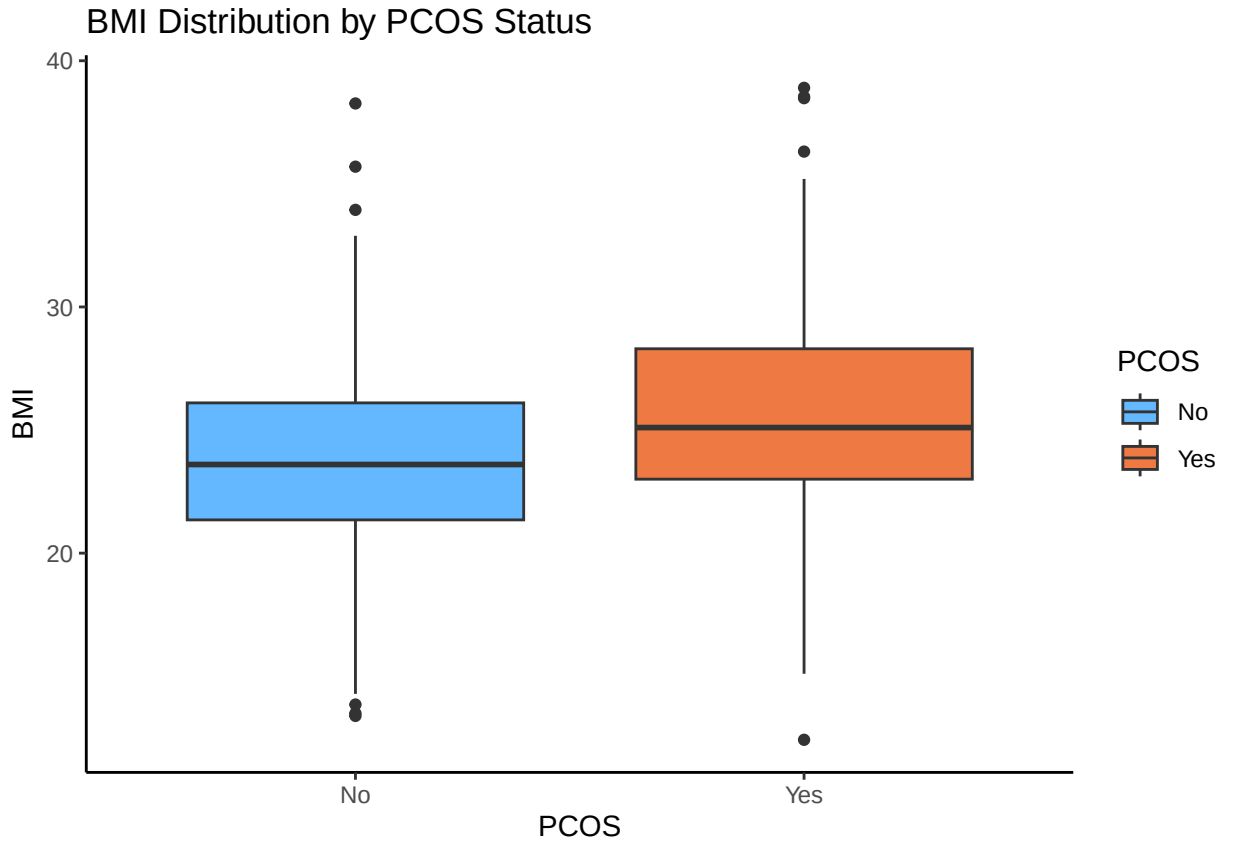
```
# the effect size is small (but almost moderate)
```

The effect size (d = -0.431) suggests that the difference in BMI between the two groups is small to moderate.

4.6 Visualising the Differences

We create a boxplot to visualize the distribution of BMI across the two groups:

```
# visualising the difference
ggplot(pcos_df, aes(x = pcos, y = bmi, fill = pcos)) +
  geom_boxplot() +
  labs(title = "BMI Distribution by PCOS Status", x = "PCOS", y = "BMI", fill = "PCOS") +
  scale_fill_manual(values = c("No" = "steelblue1", "Yes" = "sienna2"))
```




```
# pcos group has a higher median BMI compared to the non-PCOS group.
# there are outliers in both groups
```

The boxplot confirms that the median BMI is higher in the PCOS group. Additionally, both groups contain outliers.

4.7 Conclusion

While women with PCOS tend to have higher BMI compared to those without PCOS, the effect size suggests that the difference, though statistically significant, may have limited practical significance.

5. Research Question 2: Are Irregular Periods Associated With Higher PCOS Risk?

Since both variables are categorical, we conduct a chi-square test of independence.

5.1 Creating a contingency table

```
table(pcos_df$pcos, pcos_df$cycle)
```

```
##
##       Irregular Regular
##   No         56     307
##   Yes        94      82
```

The chi-square test assumes that all expected counts are 5, which we confirm is met in this dataset.

5.2 Chi-Square Test

```
chisq.test(pcos_df$pcos, pcos_df$cycle)
```

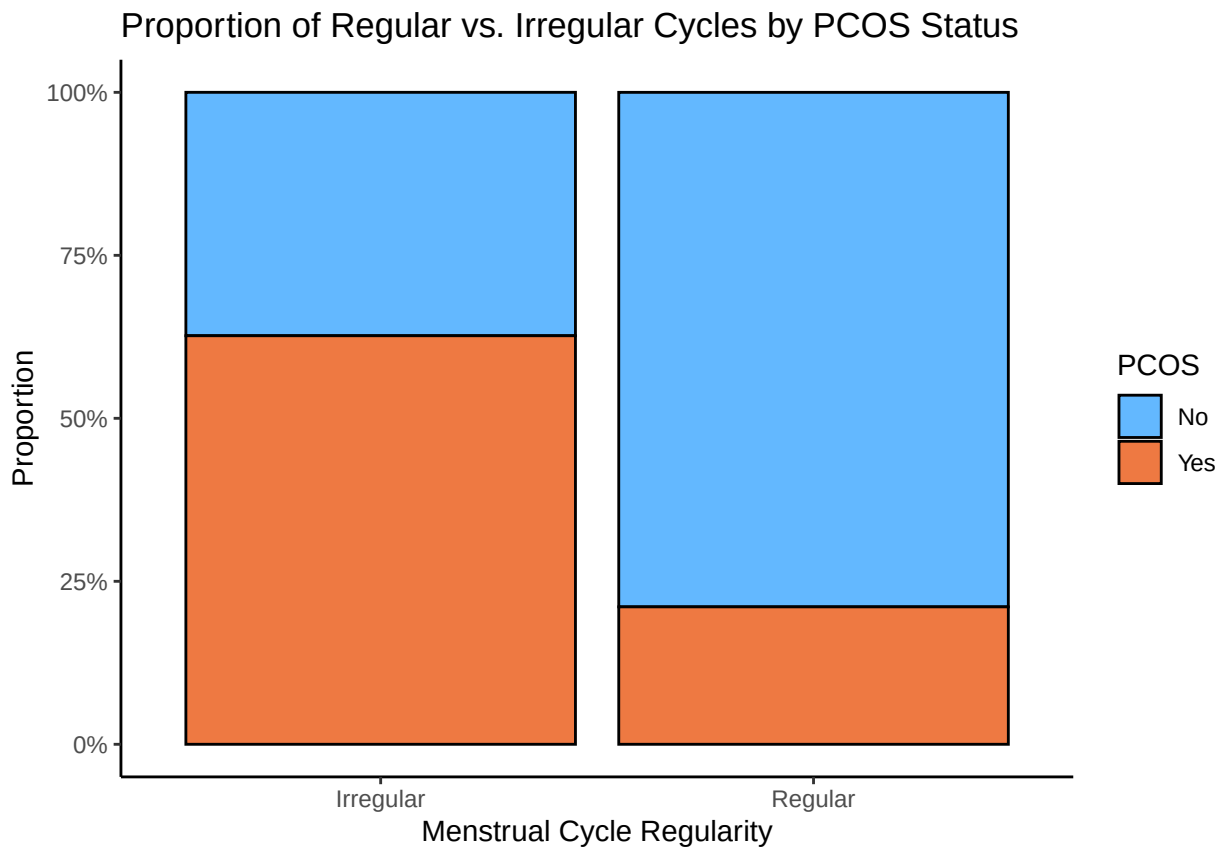
```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  pcos_df$pcos and pcos_df$cycle
## X-squared = 83.258, df = 1, p-value < 2.2e-16
```

The results indicate a highly significant association between menstrual cycle regularity and PCOS status (^{2} = 84.189, df = 1, p < 2.2e-16). This provides strong evidence that women with irregular periods have a significantly higher likelihood of having PCOS compared to those with regular cycles.

5.3 Visualising the Association

We create a bar graph to illustrate the proportion of regular vs. irregular menstrual cycles by PCOS status:

```
ggplot(pcos_df, aes(x = cycle, fill = pcos)) +
  geom_bar(position = "fill", color = "black") +
  labs(x = "Menstrual Cycle Regularity", y = "Proportion", fill = "PCOS",
       title = "Proportion of Regular vs. Irregular Cycles by PCOS Status") +
  scale_y_continuous(labels = scales::percent) +
  scale_fill_manual(values = c("No" = "steelblue1", "Yes" = "sienna2"))
```



5.4 Conclusion

- The majority of women with irregular menstrual cycles have PCOS.
- Most women with regular cycles do not have PCOS, though a small proportion still do.

This analysis confirms a strong association between menstrual cycle irregularity and increased PCOS risk.

6. Research Question 3 : Do Hormone Levels Differ Between PCOS and Non-PCOS Women?

6.1 Checking Normality of Hormone Levels

We first assess whether hormone levels follow a normal distribution using the Shapiro-Wilk test:

```
hormones <- c("fsh", "lh", "fsh_lh_ratio", "tsh", "amh", "prl", "prg")

for (hormone in hormones) {
  print(hormone)
  print(shapiro.test(pcos_df[[hormone]]))
}
```

```
## [1] "fsh"
##
##  Shapiro-Wilk normality test
##
## data:  pcos_df[[hormone]]
## W = 0.02379, p-value < 2.2e-16
##
## [1] "lh"
##
##  Shapiro-Wilk normality test
##
## data:  pcos_df[[hormone]]
## W = 0.027288, p-value < 2.2e-16
##
## [1] "fsh_lh_ratio"
##
##  Shapiro-Wilk normality test
##
## data:  pcos_df[[hormone]]
## W = 0.049997, p-value < 2.2e-16
##
## [1] "tsh"
##
##  Shapiro-Wilk normality test
##
## data:  pcos_df[[hormone]]
## W = 0.4139, p-value < 2.2e-16
##
## [1] "amh"
##
##  Shapiro-Wilk normality test
##
## data:  pcos_df[[hormone]]
## W = 0.72465, p-value < 2.2e-16
##
## [1] "prl"
##
##  Shapiro-Wilk normality test
##
## data:  pcos_df[[hormone]]
## W = 0.82024, p-value < 2.2e-16
##
## [1] "prg"
##
##  Shapiro-Wilk normality test
##
## data:  pcos_df[[hormone]]
## W = 0.051056, p-value < 2.2e-16
```

The results show that all p-values are < 0.05, indicating that none of the hormone levels are normally distributed. Since normality assumptions are violated, we proceed with a non-parametric test.

6.2 Comparing Hormone Levels Using the Mann-Whitney U Test

We use the Mann-Whitney U test (Wilcoxon rank-sum test) to compare hormone levels between women with and without PCOS:

```
# non-parametric test - the mann whitney u test
wilcoxon_results <- lapply(hormones, function(hormone) {
  test <- wilcox.test(pcos_df[[hormone]] ~ pcos_df$pcos)
  return(data.frame(Hormone = hormone, W = test$statistic, p_value = test$p.value))
})

wilcoxon_results_df <- do.call(rbind, wilcoxon_results)
print(wilcoxon_results_df)
```

```
##         Hormone       W      p_value
## W           fsh 36416.5 8.352508e-03
## W1           lh 30311.5 3.357906e-01
## W2 fsh_lh_ratio 36630.5 5.715480e-03
## W3          tsh 31092.5 6.157303e-01
## W4          amh 22684.5 4.739950e-08
## W5          prl 31053.5 5.996543e-01
## W6          prg 33295.0 4.181215e-01
```

The results indicate that FSH, FSH/LH ratio, and AMH show significant differences between groups, while LH, TSH, PRL, and PRG do not.

6.3 Effect Size: Cliff's Delta

To assess the magnitude of these differences, we compute Cliff's delta for the significant hormones:

```
significant_hormones <- c("fsh", "fsh_lh_ratio", "amh")


# effect size cliffs delta
cliff.delta(pcos_df$fsh ~ pcos_df$pcos)
```

```
##
## Cliff's Delta
##
## delta estimate: 0.1400106 (negligible)
## 95 percent confidence interval:
##      lower      upper
## 0.03874955 0.23842536
```

```
cliff.delta(pcos_df$fsh_lh_ratio ~ pcos_df$pcos)
```

```
##
## Cliff's Delta
##
## delta estimate: 0.1467099 (negligible)
## 95 percent confidence interval:
##      lower      upper
## 0.03925573 0.25080963
```

```r
cliff.delta(pcos_df$amh ~ pcos_df$pcos)
```

```
##
## Cliff's Delta
##
## delta estimate: -0.2898666 (small)
## 95 percent confidence interval:
##      lower      upper
## -0.3916938 -0.1810248
```

The results suggest weak effect sizes, indicating that while the differences are statistically significant, they may not be clinically meaningful.

6.4 Visualizing Hormone Level Differences

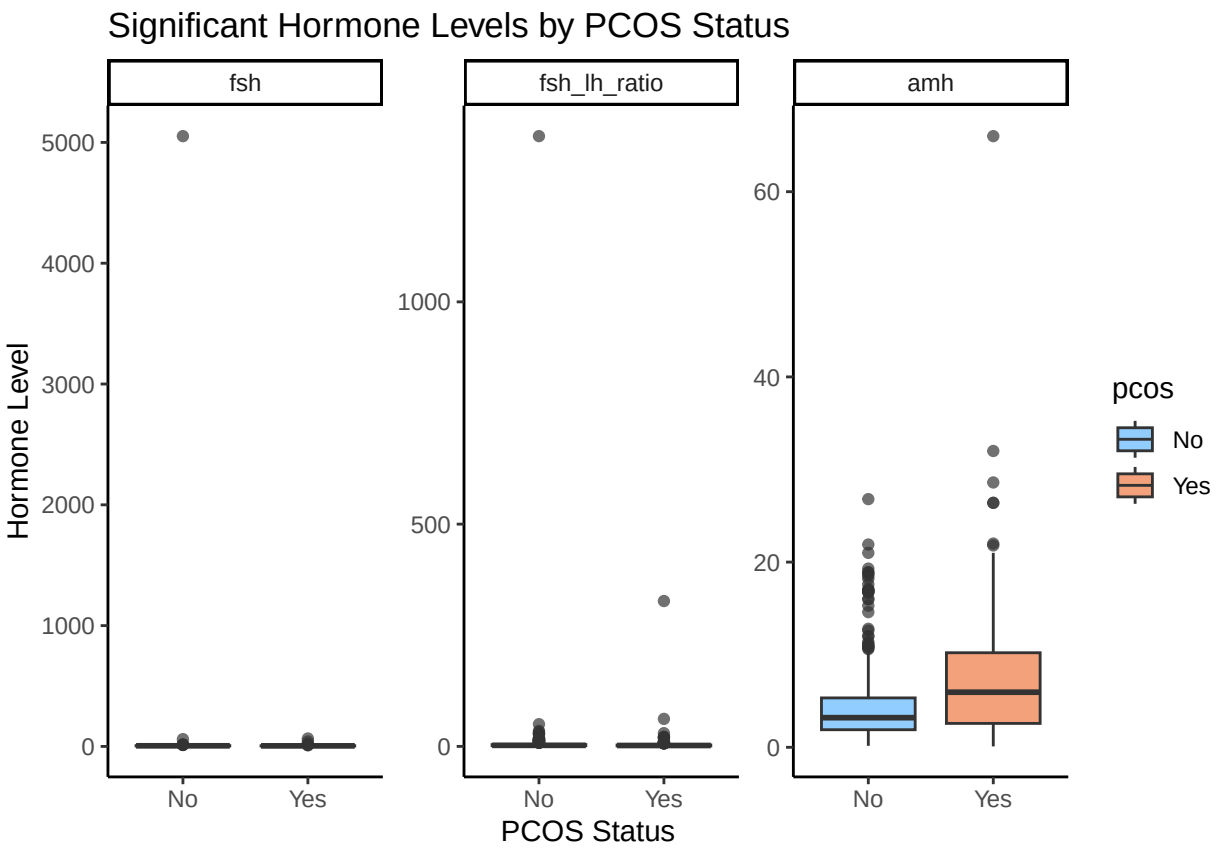
Since hormone levels are in separate columns, we reshape the dataset for visualization:

```r
# reshaping the data for ggplot
pcos_long <- reshape2::melt(pcos_df, id.vars = "pcos", measure.vars = significant_hormones)
head(pcos_long)
```

```
##   pcos variable value
## 1   No      fsh  7.95
## 2   No      fsh  6.73
## 3  Yes      fsh  5.54
## 4   No      fsh  8.06
## 5   No      fsh  3.98
## 6   No      fsh  3.24
```

We then create boxplots to compare hormone levels between groups:

```r
# boxplot
ggplot(pcos_long, aes(x = pcos, y = value, fill = pcos)) +
  geom_boxplot(alpha = 0.7) +
  facet_wrap(~variable, scales = "free") +
  labs(title = "Significant Hormone Levels by PCOS Status",
       x = "PCOS Status", y = "Hormone Level") +
  scale_fill_manual(values = c("No" = "steelblue1", "Yes" = "sienna2"))
```



The boxplots reveal extreme outliers. The presence of extreme outliers is disproportionately stretching the y-axis, causing the boxplots to appear highly compressed. This makes it difficult to visually assess the distribution of hormone levels across groups. Therefore, we remove the outliers for clearer visualization and more reliable interpretation.

6.5 Removing Outliers

We define a function to remove values beyond 1.5 times the interquartile range (IQR):

```
# function for removing outliers
remove_outliers <- function(x) {
  q1 <- quantile(x, 0.25, na.rm = TRUE)
  q3 <- quantile(x, 0.75, na.rm = TRUE)
  iqr <- q3 - q1
  x[x > (q3 + 1.5 * iqr)] <- NA  # Remove extreme values
  x[x < (q1 - 1.5 * iqr)] <- NA
  return(x)
}
```

We apply this function to the significant hormones:

```
# applying that function to our significant hormones
pcos_df$filtered_fsh <- remove_outliers(pcos_df$fsh)
pcos_df$filtered_fsh_lh_ratio <- remove_outliers(pcos_df$fsh_lh_ratio)
pcos_df$filtered_amh <- remove_outliers(pcos_df$amh)
```

```r
glimpse(pcos_df)
```

```
## Rows: 539
## Columns: 16
## $ pcos                 <fct> No, No, Yes, No, No, No, No, No, No, No, No, No,~
## $ age                  <dbl> 28, 36, 33, 37, 25, 36, 34, 33, 32, 36, 20, 26, ~
## $ weight               <dbl> 44.6, 65.0, 68.8, 65.0, 52.0, 74.1, 64.0, 58.5, ~
## $ height               <dbl> 152.0, 161.5, 165.0, 148.0, 161.0, 165.0, 156.0,~
## $ bmi                  <dbl> 19.30000, 24.92116, 25.27089, 29.67495, 20.06095~
## $ cycle                <fct> Regular, Regular, Regular, Regular, Regular, Reg~
## $ fsh                  <dbl> 7.95, 6.73, 5.54, 8.06, 3.98, 3.24, 2.85, 4.86, ~
## $ lh                   <dbl> 3.68, 1.09, 0.88, 2.36, 0.90, 1.07, 0.31, 3.07, ~
## $ fsh_lh_ratio         <dbl> 2.160326, 6.174312, 6.295455, 3.415254, 4.422222~
## $ tsh                  <dbl> 0.68, 3.16, 2.54, 16.41, 3.57, 1.60, 1.51, 12.18~
## $ amh                  <dbl> 2.07, 1.53, 6.63, 1.22, 2.26, 6.74, 3.05, 1.54, ~
## $ prl                  <dbl> 45.16, 20.09, 10.52, 36.90, 30.09, 16.18, 26.41,~
## $ prg                  <dbl> 0.57, 0.97, 0.36, 0.36, 0.38, 0.30, 0.46, 0.26, ~
## $ filtered_fsh         <dbl> 7.95, 6.73, 5.54, 8.06, 3.98, 3.24, 2.85, 4.86, ~
## $ filtered_fsh_lh_ratio <dbl> 2.160326, 6.174312, 6.295455, 3.415254, 4.422222~
## $ filtered_amh         <dbl> 2.07, 1.53, 6.63, 1.22, 2.26, 6.74, 3.05, 1.54, ~
```

We reshape the data again:

```r
pcos_long_reshape <- reshape2::melt(pcos_df, id.vars = "pcos",
                            measure.vars = c("filtered_fsh", "filtered_fsh_lh_ratio", "filtered_amh"))

head(pcos_long_reshape)
```

```
##   pcos     variable value
## 1   No filtered_fsh  7.95
## 2   No filtered_fsh  6.73
## 3  Yes filtered_fsh  5.54
## 4   No filtered_fsh  8.06
## 5   No filtered_fsh  3.98
## 6   No filtered_fsh  3.24
```

6.6 Reassessing the Wilcoxon Test and Effect Sizes

We verify whether the results remain consistent after removing outliers:

```r
# just making sure results of the wilcoxon are similar..
wilcox.test(pcos_df$filtered_fsh ~ pcos_df$pcos)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  pcos_df$filtered_fsh by pcos_df$pcos
## W = 35029, p-value = 0.006066
## alternative hypothesis: true location shift is not equal to 0
```

```r
wilcox.test(pcos_df$filtered_fsh_lh_ratio ~ pcos_df$pcos)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  pcos_df$filtered_fsh_lh_ratio by pcos_df$pcos
## W = 30714, p-value = 0.01023
## alternative hypothesis: true location shift is not equal to 0
```

```r
wilcox.test(pcos_df$filtered_amh ~ pcos_df$pcos)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  pcos_df$filtered_amh by pcos_df$pcos
## W = 19375, p-value = 0.0001352
## alternative hypothesis: true location shift is not equal to 0
```

The tests remain statistically significant (p-values < 0.05).

We also check Cliff's delta again:

```r
cliff.delta(pcos_df$filtered_fsh ~ pcos_df$pcos)
```

```
##
## Cliff's Delta
##
## delta estimate: 0.1473469 (small)
## 95 percent confidence interval:
##      lower      upper
## 0.04653326 0.24518917
```

```r
cliff.delta(pcos_df$filtered_fsh_lh_ratio ~ pcos_df$pcos)
```

```
##
## Cliff's Delta
##
## delta estimate: 0.1419542 (negligible)
## 95 percent confidence interval:
##      lower      upper
## 0.03026959 0.25013794
```

```r
cliff.delta(pcos_df$filtered_amh ~ pcos_df$pcos)
```

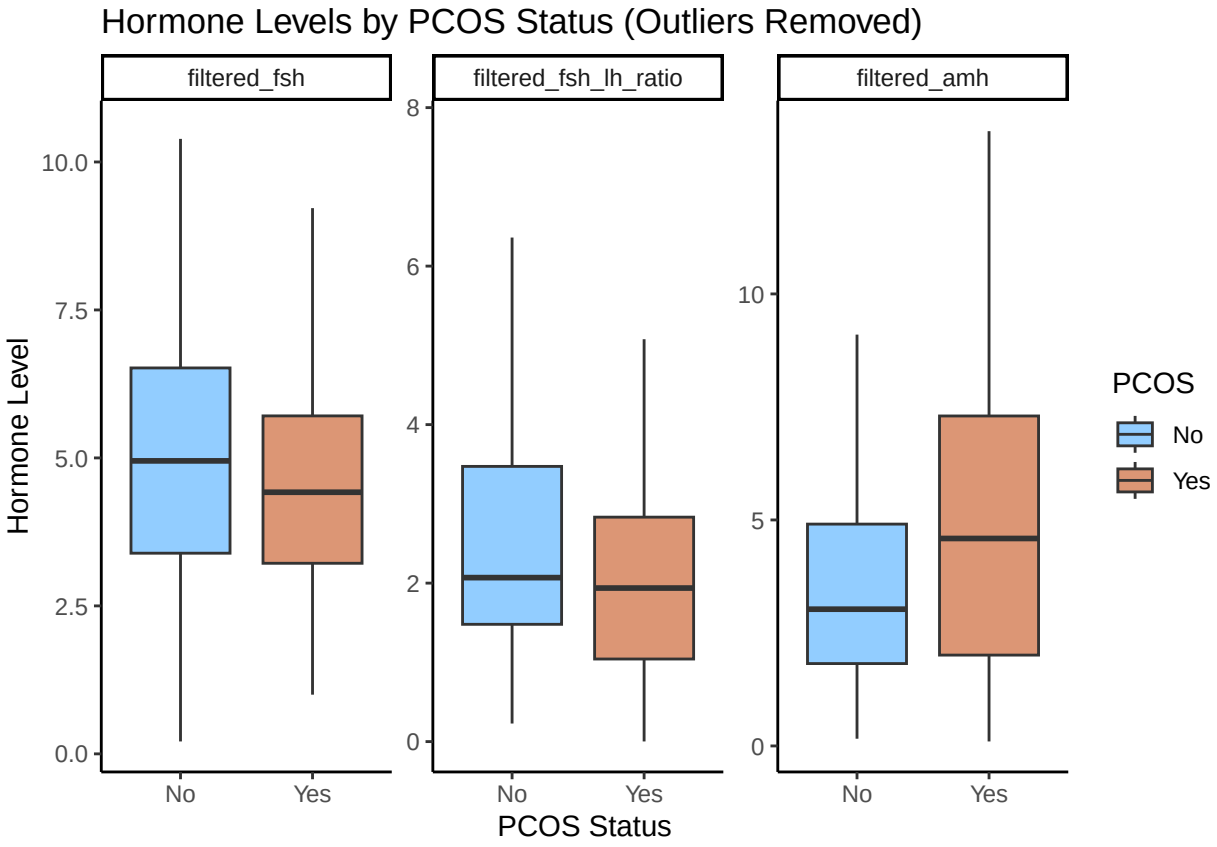
```
##
## Cliff's Delta
##
## delta estimate: -0.2186126 (small)
## 95 percent confidence interval:
##       lower       upper
## -0.33326922 -0.09758426
```

After removing extreme outliers, the effect size for FSH increased from negligible to small. However, the overall effect sizes remain small, indicating that while the differences are statistically significant, their practical significance may be minimal.

6.7 Final Visualization (Outliers Removed)

```
ggplot(pcos_long_reshape, aes(x = pcos, y = value, fill = pcos)) +
  geom_boxplot(alpha = 0.7, outlier.shape = NA) +  # Hide outliers
  facet_wrap(~variable, scales = "free") +
  labs(title = "Hormone Levels by PCOS Status (Outliers Removed)",
       x = "PCOS Status", y = "Hormone Level", fill = "PCOS") +
  scale_fill_manual(values = c("No" = "steelblue1", "Yes" = "sienna3"))
```

```
## Warning: Removed 111 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



6.8 Conclusion

FSH is lower in women with PCOS compared to those without PCOS.

FSH/LH ratio is lower in women with PCOS.

AMH is higher in women with PCOS.

However small-to-negligible effect sizes suggest that these differences may have limited clinical relevance.

7. Discussion & Conclusions

From the statistical tests conducted, it may be concluded that:

BMI is significantly higher in PCOS individuals ($p < 0.05$), however the effect size is small to moderate. Higher BMI in PCOS patients is consistent with established research.[6] Studies indicate that women with PCOS have a greater risk of overweight, obesity, and central obesity, emphasizing the importance of weight management as part of clinical treatment.[7] Furthermore, BMI management in PCOS patients enhances glucose metabolism, thus decreasing the risk for developing gestational diabetes during pregnancy.[8]

However, it should be noted that PCOS cases in which the patient's BMI is within the normal range is also prevalent (although lower) in clinical practice.[9]

Irregular periods are strongly associated with PCOS ($p < 0.001$). It was found that 75-85% of women with PCOS experience oligomenorrhea (infrequent periods) or amenorrhea (absent periods).[10]

According to existing literature, this irregularity is mainly due to excess androgen levels, which disrupt normal ovulation. This excess androgen interferes with ovulation, preventing eggs from developing and being regularly released from follicles.[11]

In addition, PCOS is characterised by polycystic ovaries, where multiple small, fluid-filled follicles containing immature eggs accumulate along the ovarian edges. These ovaries may not function properly, further contributing to menstrual irregularities.[12]

Visualisation of the data also showed that not all PCOS cases have irregular cycles, which is also consistent with current statistics.[13]

Some hormone levels, including FSH, FSH/LH ratio, and AMH, significantly differed between groups. This is consistent with prior medical literature:

- FSH levels tend to be lower in PCOS individuals.[14]
- PCOS is typically associated with a low FSH and high LH, leading to a decreased FSH/LH ratio.[15]
- AMH levels are higher in PCOS women, as they have a greater number of antral follicles.[16]

Despite the statistical significance, however, the effect sizes were small, suggesting limited clinical impact. This aligns with research indicating that while hormonal differences are associated with PCOS, their variability within populations may reduce their effectiveness as standalone diagnostic markers.[17]

Interestingly, LH levels did not significantly differ between PCOS and non-PCOS individuals. This contrasts with prior research stating that PCOS involves altered gonadotropin regulation, leading to a relative increase in LH compared to FSH.[18] This could be explained by hormone fluctuations across the menstrual cycle, which were not accounted for in this dataset.

While this study provides meaningful insights, several limitations must be acknowledged. Firstly, the dataset was relatively small, with only 539 observations. Moreover, the data was only collected from hospitals in Kerala, India – thus introducing geographical bias. Hormone fluctuations across different phases of the menstrual cycle were also not accounted for in the dataset. Finally, PCOS is a heterogenous condition, and the dataset used may not fully capture all phenotypic variations.

In the future, research should analyze larger, more diverse datasets, incorporate additional clinical factors, such as insulin resistance and metabolic markers, to enhance PCOS classification, and investigate hormonal patterns across different menstrual phases for a better understanding.

8. Bibliography

1. National Center for Biotechnology Information. Polycystic Ovary Syndrome (PCOS). [Internet]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK459251/

2. Dennett CC, Simon J. The role of polycystic ovary syndrome in reproductive and metabolic health: overview and approaches for treatment. Diabetes Spectr. 2015 May;28(2):116-20. doi: 10.2337/diaspect.28.2.116. PMID: 25987810; PMCID: PMC4433074.
3. The Medical Journal of Australia. Australian-led PCOS guideline an international first. [Internet]. 2018. Available from: https://www.mja.com.au/journal/2018/australian-led-pcos-guideline-international-first
4. Gibson-Helm M, Teede H, Dunaif A, Dokras A. Delayed diagnosis and a lack of information associated with dissatisfaction in women with polycystic ovary syndrome. J Clin Endocrinol Metab. 2017 Feb 1;102(2):604-612. doi: 10.1210/jc.2016-2963. PMID: 27906550; PMCID: PMC6283441.
5. Kottarathil P. Polycystic ovary syndrome (PCOS) [dataset]. Kaggle. 2020. Available from: https://www.kaggle.com/prasoonkottarathil/polycystic-ovary-syndrome-pcos
6. Barber TM, McCarthy MI, Wass JA, Franks S. Obesity and polycystic ovary syndrome: implications for pathogenesis and treatment. J Clin Endocrinol Metab. 2006;91(1):7-13. PMCID: PMC2744370. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC2744370/
7. Lim SS, Davies MJ, Norman RJ, Moran LJ. Overweight, obesity and central obesity in women with polycystic ovary syndrome: a systematic review and meta-analysis. Hum Reprod Update. 2012;18(6):618–637. doi: 10.1093/humupd/dms030.
8. Ma N, Zhou J, Lu W. The normal body mass index (BMI) of women with polycystic ovary syndrome (PCOS) was associated with IVF/ICSI assisted conception outcomes. Clin Exp Obstet Gynecol. 2023;50(11):228. doi: 10.31083/j.ceog5011228.
9. National Library of Medicine. PCOS and obesity. [Internet]. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10986768/
10. Harris HR, Babic A, Webb PM, Nagle CM, Jordan SJ, Risch HA, et al. Polycystic ovary syndrome, oligomenorrhea, and risk of ovarian cancer histotypes: Evidence from the Ovarian Cancer Association Consortium. Cancer Epidemiol Biomarkers Prev. 2018 Feb;27(2):174-182. doi: 10.1158/1055-9965.EPI-17-0655. PMID: 29141849; PMCID: PMC5877463.
11. Cleveland Clinic. Polycystic ovary syndrome (PCOS). [Internet]. Available from: https://my.clevelandclinic.org/health/diseases/8316-polycystic-ovary-syndrome-pcos
12. Mayo Clinic. Polycystic ovary syndrome (PCOS) - Symptoms & causes. 2023. Available from: https://www.mayoclinic.org/diseases-conditions/pcos/symptoms-causes/syc-20353439
13. Carmina E, Lobo RA. Do hyperandrogenic women with normal menses have polycystic ovary syndrome? Fertil Steril. 1999 Feb;71(2):319-22. doi: 10.1016/s0015-0282(98)00455-5. PMID: 9988405.
14. Emanuel RHK, Roberts J, Docherty PD, Lunt H, Campbell RE, Möller K. A review of the hormones involved in the endocrine dysfunctions of polycystic ovary syndrome and their interactions. Front Endocrinol (Lausanne). 2022 Nov 15;13:1017468. doi: 10.3389/fendo.2022.1017468. PMID: 36457554; PMCID: PMC9705998.
15. Emanuel RHK, Roberts J, Docherty PD, Lunt H, Campbell RE, Möller K. A review of the hormones involved in the endocrine dysfunctions of polycystic ovary syndrome and their interactions. Front Endocrinol (Lausanne). 2022 Nov 15;13:1017468. doi: 10.3389/fendo.2022.1017468. PMID: 36457554; PMCID: PMC9705998.
16. National Library of Medicine. Anti-Müllerian hormone and PCOS. [Internet]. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8006968/
17. Skiba MA, Islam RM, Bell RJ, Davis SR. Understanding variation in prevalence estimates of polycystic ovary syndrome: A systematic review and meta-analysis. Hum Reprod Update. 2018;24(6):694–709. doi: 10.1093/humupd/dmy022.
18. Taylor AE, McCourt B, Martin KA, Anderson EJ, Adams JM, Schoenfeld D, Hall JE. Determinants of abnormal gonadotropin secretion in clinically defined women with polycystic ovary syndrome. J Clin Endocrinol Metab. 1997 Jul;82(7):2248-56. doi: 10.1210/jcem.82.7.4105. PMID: 9215302.