Worksheet 3; Smoking 1

Smoking and Birth Weight

Preamble

```
# Preamble
## Install Pacman
load.pac <- function() {

  if(require("pacman")){
    library(pacman)
  }else{
    install.packages("pacman")
    library(pacman)
  }

  pacman::p_load(xts, sp, gstat, ggplot2, rmarkdown, reshape2, ggmap,
                 parallel, dplyr, plotly, tidyverse, reticulate, UsingR, Rmpfr,
                 swirl, corrplot, gridExtra, mise, latex2exp, tree, rpart)

}

load.pac()
```

```
## Loading required package: pacman
```

```
mise()
```

Load the Data

```
(birthwt <- as_tibble(read.csv(file = "./0datasets/birthwt.csv", header = TRUE, sep = ",")))
```

```
## # A tibble: 1,226 x 2
##      bwt smoke
##    <int> <fct>
##  1  3429 no
##  2  3229 no
##  3  3657 yes
##  4  3514 no
##  5  3086 yes
##  6  3886 no
##  7  3943 no
##  8  3771 no
##  9  3429 no
## 10  4086 yes
## # ... with 1,216 more rows
```

```
birthwt$smoke <- c(FALSE, TRUE)[birthwt$smoke]
summary(birthwt)
```

```
##       bwt          smoke
##  Min.   :1571   Mode :logical
##  1st Qu.:3114   FALSE:742
##  Median :3429   TRUE :484
##  Mean   :3415
##  3rd Qu.:3743
##  Max.   :5029
```

```
str(birthwt)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1226 obs. of  2 variables:
##  $ bwt  : int  3429 3229 3657 3514 3086 3886 3943 3771 3429 4086 ...
##  $ smoke: logi  FALSE FALSE TRUE FALSE TRUE FALSE ...
```

```
dim(birthwt)
```

```
## [1] 1226    2
```

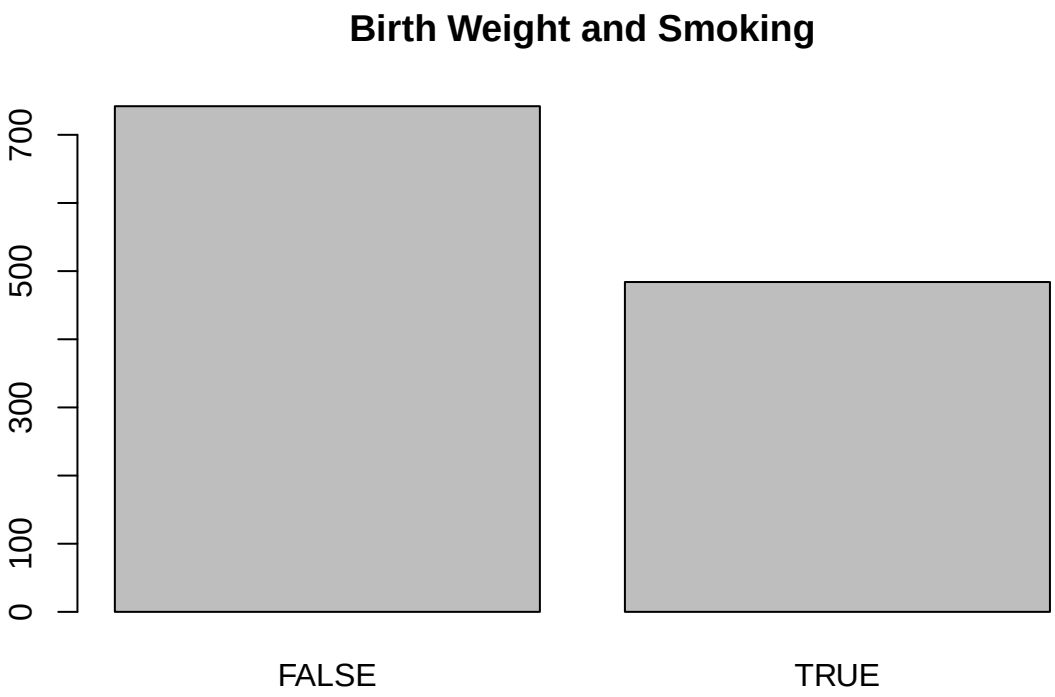
Summaries

Table

```
table(birthwt$smoke)
```

```
##
## FALSE  TRUE
##   742   484
```

```
table(birthwt$smoke) %>% barplot(main = "Birth Weight and Smoking")
```



Barplot

```
desc_stats <- function(x) {
  mean(x)
  median(x)
  var(x)
  sd(x)
}

(desc_stats <- data.frame(
mean = apply(birthwt, 2, mean),
median = apply(birthwt, 2, median),
var = apply(birthwt, 2, var),
sd = apply(birthwt, 2, sd)
))
```

```
##               mean median          var          sd
## bwt   3414.8303426   3429 2.704985e+05 520.0946858
## smoke    0.3947798      0 2.391237e-01   0.4890028
```

```
range(birthwt$bwt)
```

Range

```
## [1] 1571 5029
```

```
range(birthwt$bwt) %>% diff()
```

```
## [1] 3458
```

```
max(birthwt$bwt) - min(birthwt$bwt)
```

```
## [1] 3458
```

Quantile The quantile function returns x -axis values corresponding to a what proportion of the data is specified, so for example, for a standard normal distribution $\mathcal{N}\ (0, 1)$, 2.5% of the observations lie below 2 and another 2.5% lie above 2.

```
quantile(rnorm(1000), 0.025)
```

```
##      2.5%
## -1.831019
```

```
quantile(birthwt$bwt, 0.25)
```

```
##  25%
## 3114
```

```
quantile(birthwt$bwt, 0.75)
```

```
##  75%
## 3743
```

Inter-Quartile Range This can be calculated thusly:

```
IQR(birthwt$bwt)
```

```
## [1] 629
```

For normally distributed data we would expect:

$$\mathsf{IQR} = 1.349 \times \sigma$$

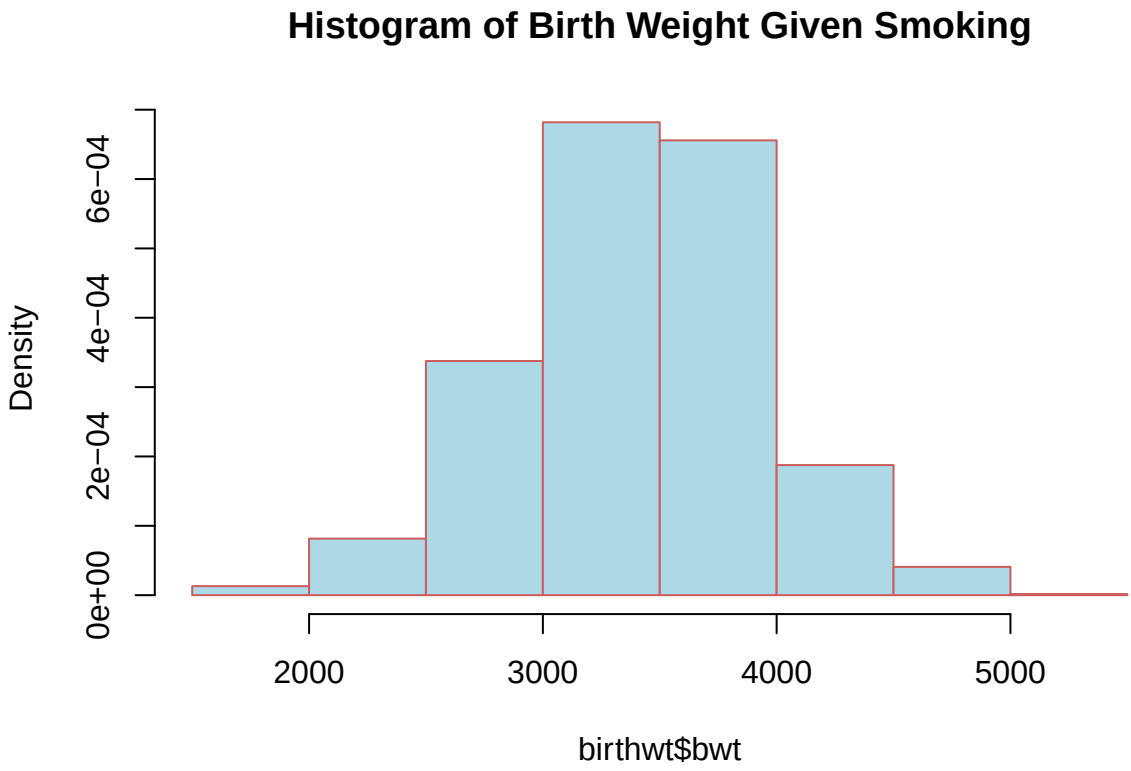
Remember tha the normal distribution is modelled using calculus:

$$
\begin{aligned}
f(x) &= -\sqrt{\frac{k}{2\pi}} \cdot e^{k \cdot \frac{(x-\mu)^2}{2}} \\
f\left(x\right) &= \sqrt{\frac{1}{2\pi}} \cdot \sum_{n=0}^{\infty} \left[\frac{\left(-\frac{1}{2} z^2\right)^n}{n!} \right] \\
\int f\left(x\right) \mathrm{d}x &= \frac{1}{\sqrt{2\pi}} \int \sum_{n=0}^{\infty} \left[\frac{\left(-\frac{1}{2} z^2\right)^n}{n!} \right] \mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}} \cdot \sum_{n=0}^{\infty} \left[\int \frac{(-1)^{-1} z^{2n}}{2^n \cdot n!} \mathrm{d}z \right] \\
&= \frac{1}{\sqrt{2\pi}} \cdot \sum_{n=0}^{\infty} \left[\frac{(-1)^n \cdot z^{2n+1}}{2^n \left(2n+1\right) n!} \right]
\end{aligned}
$$

Histograms

A histogram would offer a better understanding of the data:

```
x <- birthwt$bwt
hist(birthwt$bwt, main = "Histogram of Birth Weight Given Smoking", col = "lightblue", border = "indian
```

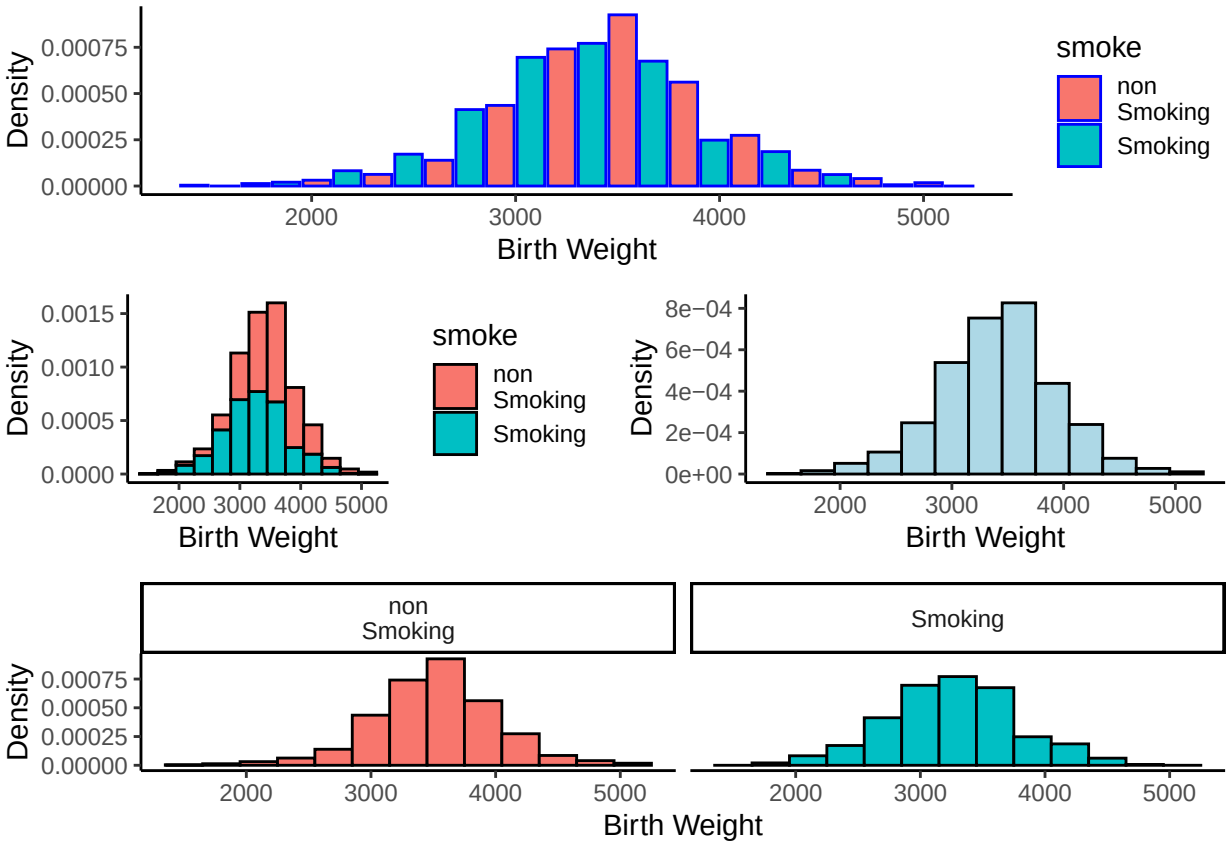



```
# curve(dnorm(x, mean(x), sd(x)), add = TRUE)
```

Adding a Density urve is extremely difficult in base plot, it's so much easier to use ggplot2:

```r
birthwt_pretty <- birthwt
birthwt_pretty$smoke <- ifelse(birthwt$smoke, "Smoking", "non\nSmoking")

hist <- ggplot(birthwt_pretty, aes(x = bwt, fill = smoke, col = "black", y = ..density..)) +
         theme_classic() +
         labs(x = "Birth Weight", y = "Density")

plots <- list()

# Dodge
plots[[1]] <- hist + geom_histogram(position = "dodge2", col = "blue", binwidth = 300)

# Overlay
plots[[2]] <-  hist + geom_histogram(binwidth = 300, col = "black")

# Single Histogram
plots[[3]] <- hist + geom_histogram(binwidth = 300, col = "black", aes(group = 1), fill = "lightblue")

# Facet Grid
 plots[[4]] <- hist + geom_histogram(binwidth = 300, col = "black") +
         facet_grid(. ~ smoke) +
         guides(fill = FALSE)


  # Colour it
  # Make a Facet Grid
  # Add a Density Curve
```

```r
layout <- matrix(c(1, 1, 2, 3, 4, 4), byrow = TRUE, nrow = 3)
# arrangeGrob(grobs = plots, layout_matrix = layout)
grid.arrange(grobs = plots, layout_matrix = layout)
```



Splitting the Data Up

Split Charts

Box Plots

Difference in Means

Challenge Data

East and West

Spiders