```
# Install and load packages
install.packages('moments')
install.packages('ggplot2')
install.packages('plyr')
install.packages('car')
library(moments)
library(ggplot2)
library(plyr)
library(car)

# Import and clean babies23 dataset
both_raw <- read.table("babies23.txt", header=TRUE)
both <- babies23_raw[which(babies23['smoke']!=9),
c('smoke','wt','gestation')]

# Add a column indicating whether or not the mother smoked during
pregnancy
smoked <- c()
for(i in 1:nrow(both)) {
  if(both[i,'smoke']== 1)  # Smoked
    smoked <- c(smoked, 'Smoker')
  else
    smoked <- c(smoked, 'Non-Smoker')
}
both <- cbind(babies23, smoked)

# Add a column with standardized weights
std_wt <- (both$wt- mean(both$wt))/sd(both$wt)
both <- cbind(both, std_wt)

# Separate dataset into smoker and non-smokers
nonsmoker <- both[which(babies23$smoked=='Non-Smoker'),]
smoker <- both[which(both$smoked=='Smoker'),]

# Summary statistics
summary(both$wt)
summary(nonsmoker$wt)
summary(smoker$wt)

# Frequency bar graphs
for(threshold in c(88.2,86,87,88,89,90)){
  underweight <- c()

  for(row in 1:nrow(both)){
    if (both[row,'wt'] < threshold)
      underweight <- c(underweight, 'Low Birth Weight')
    else
      underweight <- c(underweight, 'Normal Weight')
  }
```

```
  with_underweight <- cbind(both, underweight)
  counts = table(with_underweight$underweight, both$smoked)
  print(paste(threshold, ":", counts[1,1]/sum(counts[,1]),
counts[1,2]/sum(counts[,2])))
  barplot(counts, main=paste('Low Birth Weights with Threshold',
threshold, 'Ounces'), xlab='Low Birth Weight', ylab = 'Frequency',
col=c('red','green'), legend = rownames(counts), beside=TRUE)
}


# Skewness and kurtosis histograms
normal_skewness <- c()
normal_kurtosis <- c()
nonsmoker_skewness <- c()
nonsmoker_kurtosis <- c()
smoker_skewness <- c()
smoker_kurtosis <- c()

for(i in 1:1000) {
  normal_skewness <- c(normal_skewness,
skewness(rnorm(nrow(nonsmoker))))
  normal_kurtosis <- c(normal_kurtosis,
kurtosis(rnorm(nrow(nonsmoker))))

  nonsmoker_skewness <- c(nonsmoker_skewness,
skewness(sample(nonsmoker$wt,size=nrow(nonsmoker),replace=TRUE)))
  nonsmoker_kurtosis <- c(nonsmoker_kurtosis,
kurtosis(sample(nonsmoker$wt,size=nrow(nonsmoker),replace=TRUE)))

  smoker_skewness <- c(smoker_skewness,
skewness(sample(smoker$wt,size=nrow(smoker),replace=TRUE)))
  smoker_kurtosis <- c(smoker_kurtosis,
kurtosis(sample(smoker$wt,size=nrow(smoker),replace=TRUE)))
}
print(paste('Skewness, kurtosis coefficient of Monte Carlo derived
Normal distribution:', mean(normal_skewness), mean(normal_kurtosis)))
print(paste('Skewness, kurtosis coefficient of distribution of all
birth weights:', skewness(both$wt), kurtosis(both$wt)))
print(paste('Skewness, kurtosis coefficient of distribution of
nonsmoker birth weights:', skewness(nonsmoker$wt),
kurtosis(nonsmoker$wt)))
print(paste('Skewness, kurtosis coefficient of distribution of smoker
birth weights:', skewness(smoker$wt), kurtosis(smoker$wt)))

breaks <- seq(-1,1,by=0.1)
hist(main='Skewness Coefficients of Monte Carlo Derived
Distributions', normal_skewness, col=rgb(1,0,0,1/4), breaks=breaks,
ylim=c(0,400))
hist(nonsmoker_skewness, col=rgb(0,1,0,1/4), breaks=breaks, add=TRUE)
hist(smoker_skewness, col=rgb(0,0,1,1/4), breaks=breaks, add=TRUE)
```

```
legend('topright', c('normal', "smoker", "nonsmoker"),
col=c(rgb(1,0,0,1/4),rgb(0,1,0,1/4),rgb(0,0,1,1/4)), lwd = 4, cex=0.8)

breaks <- seq(1,5,by=0.15)
hist(main='Kurtosis Coefficients of Monte Carlo Derived
Distributions', normal_kurtosis, col=rgb(1,0,0,1/4), breaks=breaks,
ylim=c(0,400))
hist(nonsmoker_kurtosis, col=rgb(0,1,0,1/4), breaks=breaks, add=TRUE)
hist(smoker_kurtosis, col=rgb(0,0,1,1/4), breaks=breaks, add=TRUE)
legend('topright', c('normal', "smoker", "nonsmoker"),
col=c(rgb(1,0,0,1/4),rgb(0,1,0,1/4),rgb(0,0,1,1/4)), lwd = 4, cex=0.8)

# T-tests
t.test(nonsmoker$wt, smoker$wt)
t.test(nonsmoker$gestation, smoker$gestation)

# Chi-squared test of independence
for(threshold in c(88.2, seq(55,175,by=5))){
  underweight <- c()

  for(row in 1:nrow(both)){
    if (both[row,'wt'] < threshold){
      underweight <- c(underweight, 1)
    } else {
      underweight <- c(underweight, 0)
    }
  }
  print(paste('Percentage of low birth weight babies with low birth
weight threshold of ', threshold, 'ounces: ', sum(underweight==1),
'/', length(underweight)))
  print('Chi-squared test:')
  print(chisq.test(table(both$smoked, underweight)))
}

# Calculate mean
means <- ddply(both, 'smoked', summarise, grp.mean = mean(std_wt))

# Basic histogram
ggplot(nonsmoker, aes(x=bwt)) + geom_histogram(binwidth= 8)
# Change colors
p<-ggplot(smoker, aes(x=wt)) + geom_histogram(color="black",
fill="white")
p

# Histogram with density plot
ggplot(nonsmoker, aes(x=wt)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white")+
  geom_density(alpha=.2, fill="#FF6666")+
geom_vline(aes(xintercept=mean(wt)),
                                                color="blue",
```

```
linetype="dashed", size=1)

ggplot(smoker, aes(x=wt)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white")+
  geom_density(alpha=.2, fill="#FF6666")+
geom_vline(aes(xintercept=mean(wt)),
                                                    color="blue",
linetype="dashed", size=1)

# Comparison Histogram
ggplot(both, aes(x=std_wt, color= status23, fill= status23)) +
  geom_histogram(aes(y=..density..), position="identity", alpha=0.5)+
  geom_density(alpha=0.6)+
  geom_vline(data=mu23, aes(xintercept=grp23.mean),
             linetype="dashed")+
  scale_color_manual(values=c("#999999", "#E69F00", "#56B4E9"))+
  scale_fill_manual(values=c("#999999", "#E69F00", "#56B4E9"))+
  labs(title="Density Comparison - Standardized Baby Birth Weights of
Mothers With Different Smoking Status",x="Standardized Birth Weight",
y = "Density Among Population", size = 0)+
  theme_classic()

ggplot(both, aes(x=std_wt, color=smoked, fill=smoked)) +
  geom_histogram(aes(y=..density..), position="identity", alpha=0.5)+
  geom_density(alpha=0.6)+
  geom_vline(data=means, aes(xintercept=grp.mean),
             linetype="dashed")+
  scale_color_manual(values=c("#999999", "#E69F00", "#56B4E9"))+
  scale_fill_manual(values=c("#999999", "#E69F00", "#56B4E9"))+
  labs(title="Density Comparison - Standardized Baby Birth Weights of
Mothers With Different Smoking Status",x="Standardized Birth Weight",
y = "Density Among Population", size = 0)+
  theme_classic()

## Boxplot (Use to compare median)
boxplot(std_wt~smoke, both, main = "BoxPlot- Standardized Baby Birth
Weights of Mothers With Different Smoking Status", xlab = "Mothers'
smoking status", ylab = "Babies' Birth Weight")
boxplot(wt~smoke_status, both, main = "BoxPlot- Standardized Baby
Birth Weights of Mothers With Different Smoking Status", xlab =
"Mothers' smoking status", ylab = "Babies' Birth Weight")
# IQR is the length of the edge of the box, and anything
# beyond the Whisker will be the outlier cases

## QQPlot
library(car)
qqnorm(nonsmoker$std_wt, pch = 1, frame = FALSE)
qqline(nonsmoker$std_wt, col = "steelblue", lwd = 2)
qqPlot(nonsmoker$std_wt, xlab = "Theoretical Quantiles",
       ylab = "Observed Quantiles", main = "QQ-Plot for Birth Weights
```

```
of Babies Born to non-Smoking Mothers")

qqnorm(smoker$std_wt, pch = 1, frame = FALSE)
qqline(smoker$std_wt, col = "steelblue", lwd = 2)
qqPlot(smoker$std_wt, xlab = "Theoretical Quantiles",
       ylab = "Observed Quantiles", main = "QQ-Plot for Birth Weights
of Babies Born to Smoking Mothers")

qqplot(both$std_wt[both$smoke==1], both$std_wt[both$smoke!= 1],
       xlab = "Quantile of Standardized Birth Weights of Babies born
to smokers",
       ylab = "Quantile of Standardized Birth Weights of Babies born
to non-smokers", main = "QQ-Plot Comparison for Babies' Birth Weights
Based on Mothers' Smoking Status")
qqline(both$std_wt[both$smoke==1], both$std_wt[both$smoke!= 1])
```