# R_Script

*Zechen Li, Vaibhavi Gaekwad*

*11/20/2019*

## Contents



## Survey Data

```
data <- data.frame(read.csv(file = './data.csv', header = T))
summary(data)
```

```
##                       Timestamp     gender    nationality country_now
##  2019/11/19 9:21:41 PM PST : 8   Female:42   Chinese:48   China:30
##  2019/11/17 10:29:33 PM PST: 1   Male  :41   Indian :35   India:18
##  2019/11/17 10:53:47 PM PST: 1                            USA  :35
##  2019/11/17 10:55:22 PM PST: 1
##  2019/11/17 11:19:24 PM PST: 1
##  2019/11/17 11:19:35 PM PST: 1
##  (Other)                   :70
##                  child_age   child_height                       exercise
##  Greater than 20 years:83   Min.   :5.083   Attended a sports class: 1
##                             1st Qu.:5.431   No                      :34
##                             Median :5.583   Yes                     :48
##                             Mean   :5.583
##                             3rd Qu.:5.750
##                             Max.   :6.083
##
##            milk          maternal_age mother_height
##  I don't know: 2   > 35 years  : 5    Min.   :4.921
##  No          :32   20-25 years :33    1st Qu.:5.167
##  Yes         :49   26-30 years :34    Median :5.333
##                    31-35 years : 9    Mean   :5.316
##                    I don't know: 2    3rd Qu.:5.417
##                                       Max.   :5.667
##
```

```
##  Is.your.mother.working.
##  No :28
##  Yes:55
##
##
##
##
##
```

```
head(data)
```

```
##                   Timestamp gender nationality country_now
## 1 2019/11/17 3:40:51 PM PST   Male     Chinese         USA
## 2 2019/11/17 3:41:45 PM PST Female     Chinese         USA
## 3 2019/11/17 3:51:30 PM PST   Male     Chinese         USA
## 4 2019/11/17 3:58:14 PM PST   Male      Indian       India
## 5 2019/11/17 4:06:05 PM PST Female     Chinese         USA
## 6 2019/11/17 4:08:45 PM PST   Male      Indian         USA
##               child_age child_height exercise         milk maternal_age
## 1 Greater than 20 years      5.66667      Yes          Yes  26-30 years
## 2 Greater than 20 years      5.16667       No           No  31-35 years
## 3 Greater than 20 years      5.75000      Yes           No  20-25 years
## 4 Greater than 20 years      5.50000      Yes          Yes  31-35 years
## 5 Greater than 20 years      5.58333      Yes I don't know  20-25 years
## 6 Greater than 20 years      5.58333      Yes          Yes  20-25 years
##   mother_height Is.your.mother.working.
## 1       5.33330                    Yes
## 2       5.00000                     No
## 3       5.33330                    Yes
## 4       5.16667                     No
## 5       5.58333                     No
## 6       5.33330                    Yes
```

# Dataset Graphs

## Graph 1

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

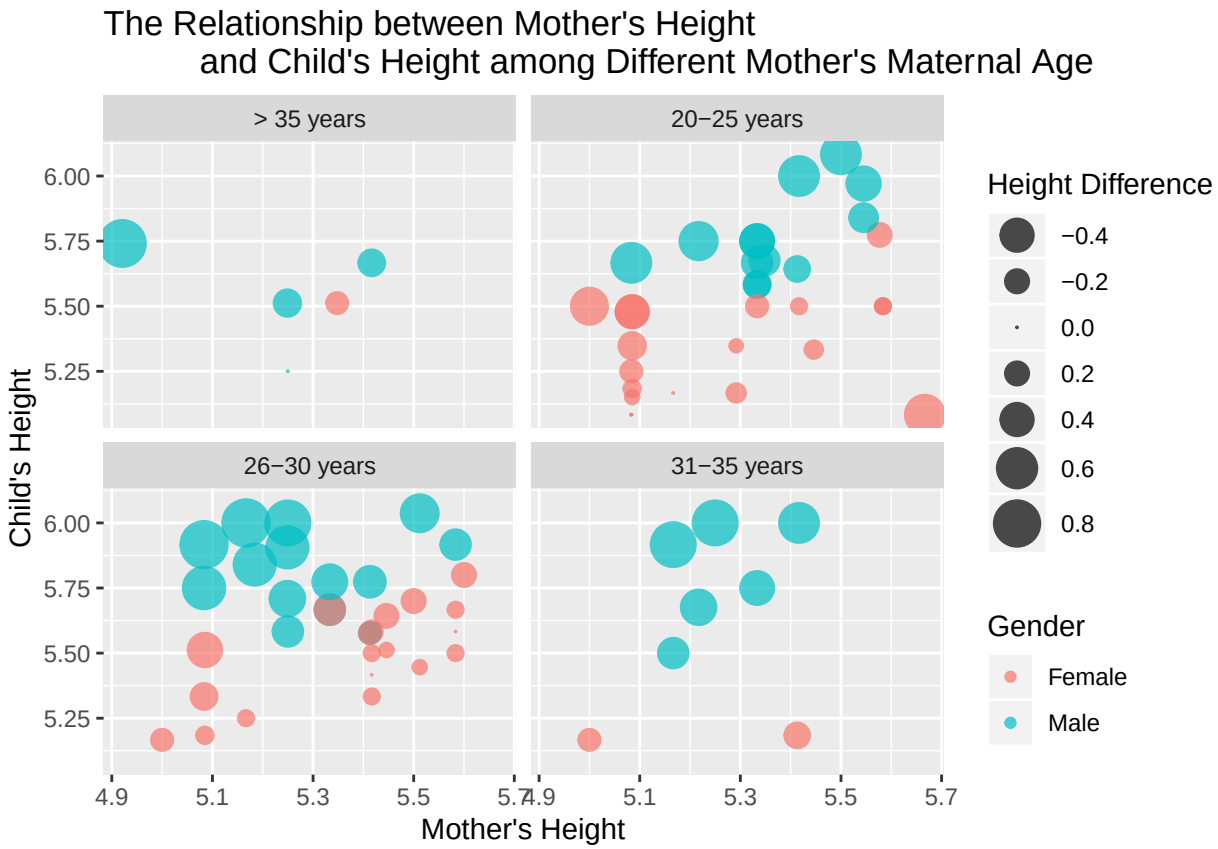
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
useful_data <- data %>%
  mutate(height_difference = child_height - mother_height) %>%
  filter(milk == 'Yes' | milk == 'No',
         exercise == 'Yes' | exercise == 'No',
         maternal_age != "I don't know")
summary(useful_data)
```

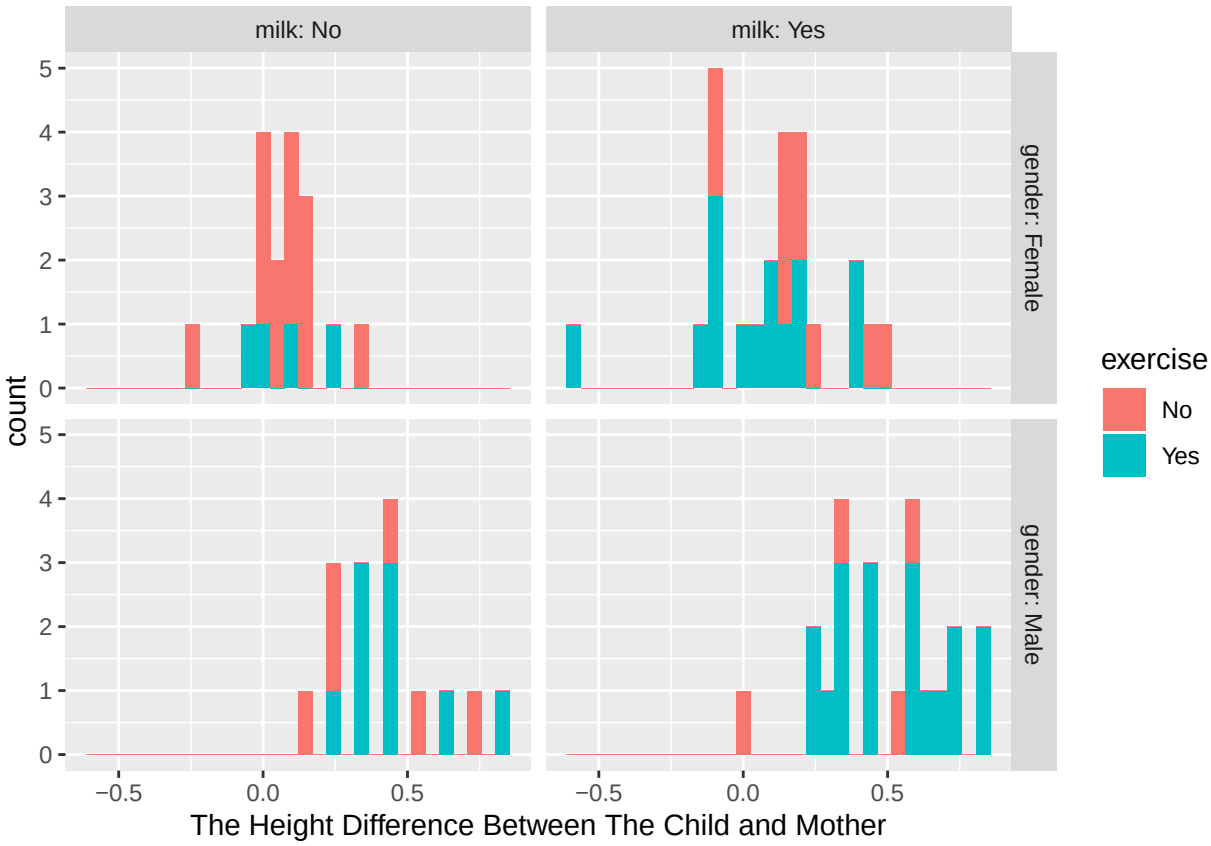
```
##                       Timestamp     gender    nationality country_now
##  2019/11/19 9:21:41 PM PST : 8   Female:41   Chinese:47   China:30
##  2019/11/17 10:29:33 PM PST: 1   Male  :37   Indian :31   India:17
##  2019/11/17 10:53:47 PM PST: 1                            USA  :31
##  2019/11/17 10:55:22 PM PST: 1
##  2019/11/17 11:19:24 PM PST: 1
##  2019/11/17 11:20:45 PM PST: 1
##  (Other)                   :65
##                  child_age   child_height                       exercise
##  Greater than 20 years:78   Min.   :5.083   Attended a sports class: 0
##                             1st Qu.:5.424   No                      :33
##                             Median :5.583   Yes                     :45
##                             Mean   :5.578
##                             3rd Qu.:5.750
##                             Max.   :6.083
##
##            milk          maternal_age mother_height
##  I don't know: 0   > 35 years  : 5    Min.   :4.921
##  No          :32   20-25 years :32    1st Qu.:5.167
##  Yes         :46   26-30 years :33    Median :5.333
##                    31-35 years : 8    Mean   :5.309
##                    I don't know: 0    3rd Qu.:5.417
##                                       Max.   :5.667
##
##  Is.your.mother.working. height_difference
##  No :24                  Min.   :-0.58337
##  Yes:54                  1st Qu.: 0.08333
##                          Median : 0.25001
##                          Mean   : 0.26908
##                          3rd Qu.: 0.42675
##                          Max.   : 0.83334
##
```

```r
library(ggplot2)
ggplot(useful_data, aes(x= mother_height ,
                        y=child_height,
                        color = gender,
                        size = height_difference)) +
  facet_wrap(~maternal_age) +
  geom_point(alpha = 0.7) +
  scale_size_area(breaks = c(-0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6, 0.8),
                  max_size = 8) +
  labs(x = "Mother's Height",
       y = "Child's Height",
       size = "Height Difference",
       color = 'Gender') +
  ggtitle("The Relationship between Mother's Height
          and Child's Height among Different Mother's Maternal Age")
```



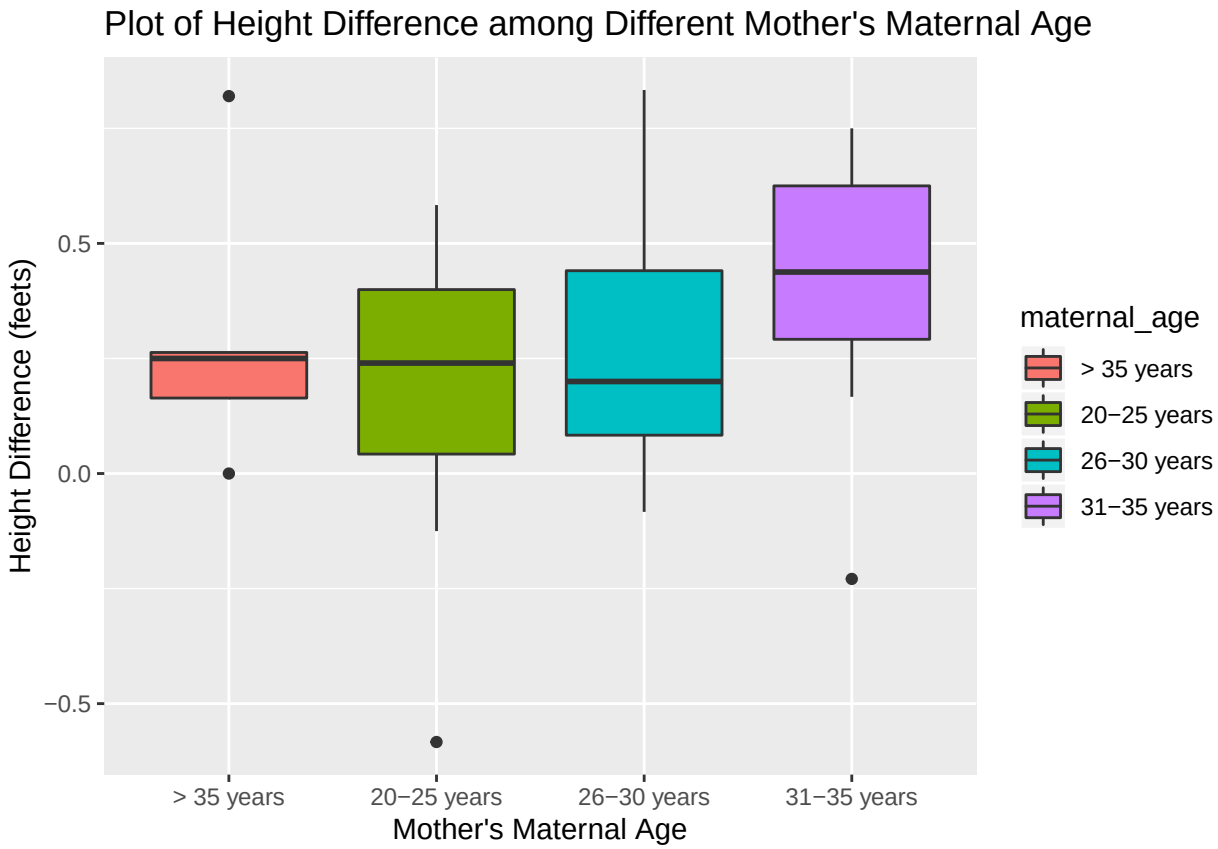

### Graph 2

```
ggplot(useful_data,  aes(height_difference))+
  facet_grid(gender~milk,labeller = label_both) +
  geom_histogram(aes(fill = exercise), bins=30) +
  xlab('The Height Difference Between The Child and Mother')
```



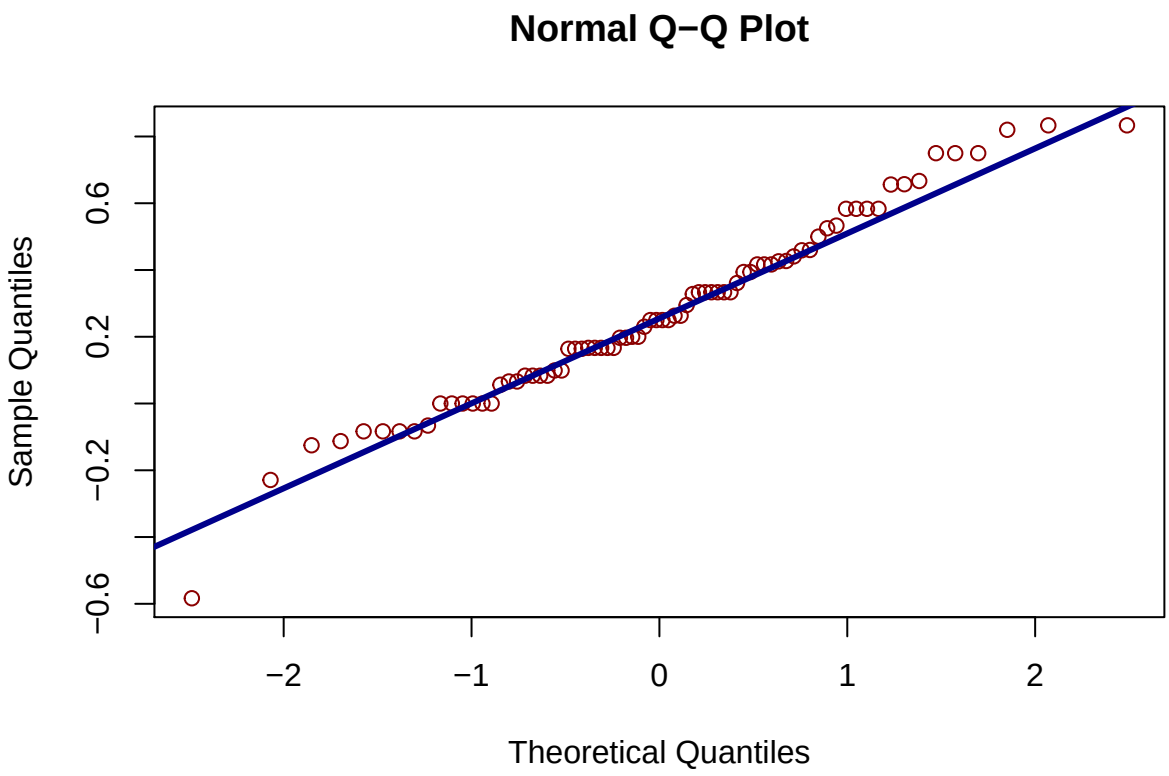

## Box-plot

```
ggplot(useful_data, aes(x=maternal_age, y=height_difference)) +
  geom_boxplot(aes(fill= maternal_age))  +
  labs(title="Plot of Height Difference among Different Mother's Maternal Age",
       x="Mother's Maternal Age",
       y = "Height Difference (feets)")
```




## Whole Population Q-Q Plot

```
qqnorm(useful_data$height_difference, col = "darkred", main = "Normal Q-Q Plot")
qqline(useful_data$height_difference, col = "darkblue", lwd = 3)
```



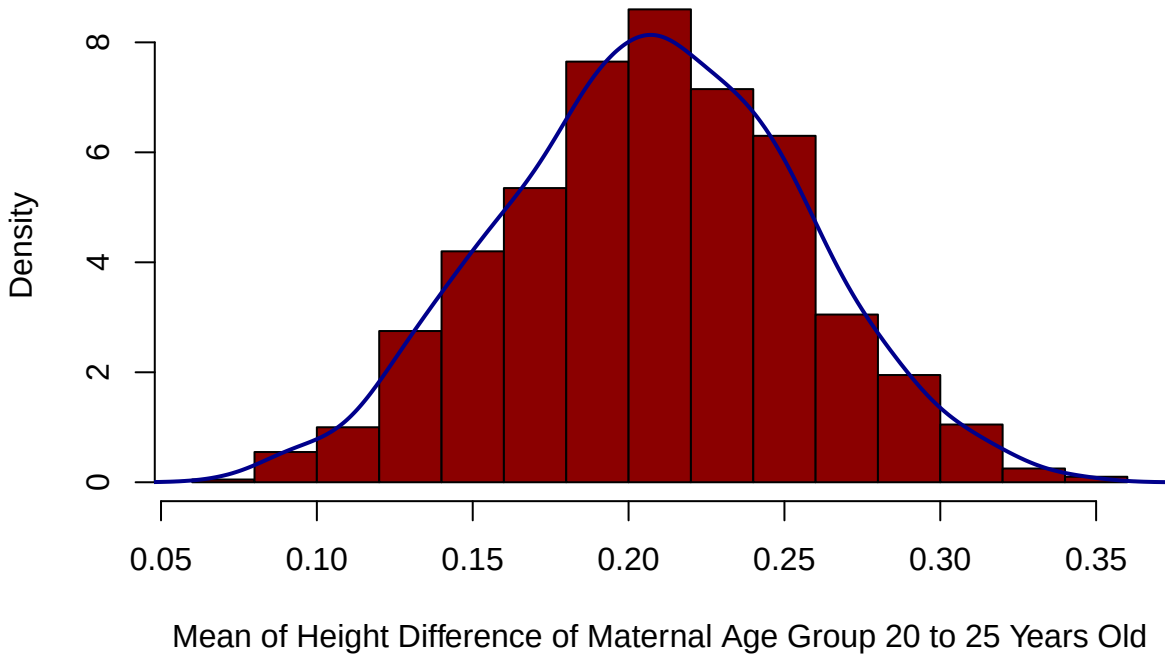

## Histogram of Sampling Distribution

```
require(mosaic)
```

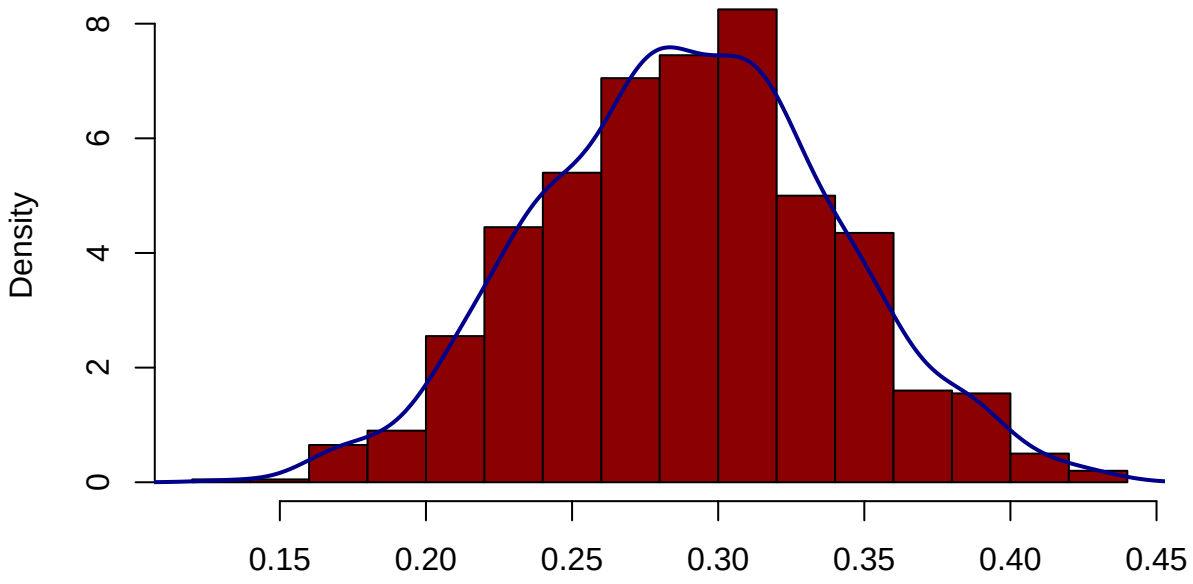
```
## Loading required package: mosaic
## Loading required package: lattice
## Loading required package: ggformula
## Loading required package: ggstance
##
## Attaching package: 'ggstance'
## The following objects are masked from 'package:ggplot2':
##
##     geom_errorbarh, GeomErrorbarh
##
## New to ggformula?  Try the tutorials:
##  learnr::run_tutorial("introduction", package = "ggformula")
##  learnr::run_tutorial("refining", package = "ggformula")
## Loading required package: mosaicData
## Loading required package: Matrix
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
##
```

```
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.
##
## Note: If you use the Matrix package, be sure to load it BEFORE loading mosaic.

##
## Attaching package: 'mosaic'

## The following object is masked from 'package:Matrix':
##
##     mean

## The following object is masked from 'package:ggplot2':
##
##     stat

## The following objects are masked from 'package:dplyr':
##
##     count, do, tally

## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median,
##     prop.test, quantile, sd, t.test, var

## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

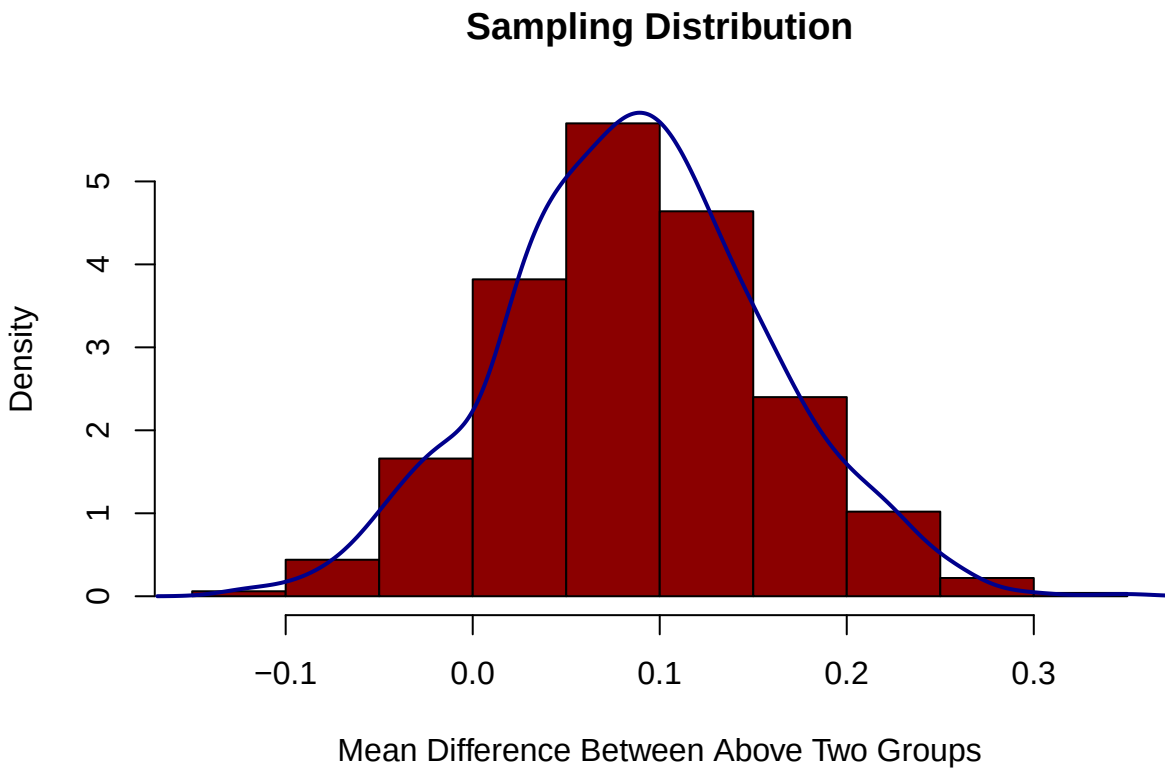
```r
set.seed(1)
group20to25 <- useful_data %>%
  filter(maternal_age == '20-25 years')
group20to25_1000 <- do(1000) * mean(sample(group20to25$height_difference,15))
hist(group20to25_1000$mean,
     main = "Sampling Distribution with Size = 15 and Simulations = 1000",
     xlab = "Mean of Height Difference of Maternal Age Group 20 to 25 Years Old",
     prob = T,
     col = "darkred")
lines(density(group20to25_1000$mean),
      col = "darkblue",
      lwd = 2)
```




```
group26to30 <- useful_data %>%
  filter(maternal_age == '26-30 years')
group26to30_1000 <- do(1000) * mean(sample(group26to30$height_difference,15))
hist(group26to30_1000$mean,
     main = "Sampling Distribution with Size = 15 and Simulations = 1000",
     xlab = "Mean of Height Difference of Maternal Age Group 26 to 30 Years Old",
     prob = T,
     col = "darkred")
lines(density(group26to30_1000$mean),
      col = "darkblue",
      lwd = 2)
```

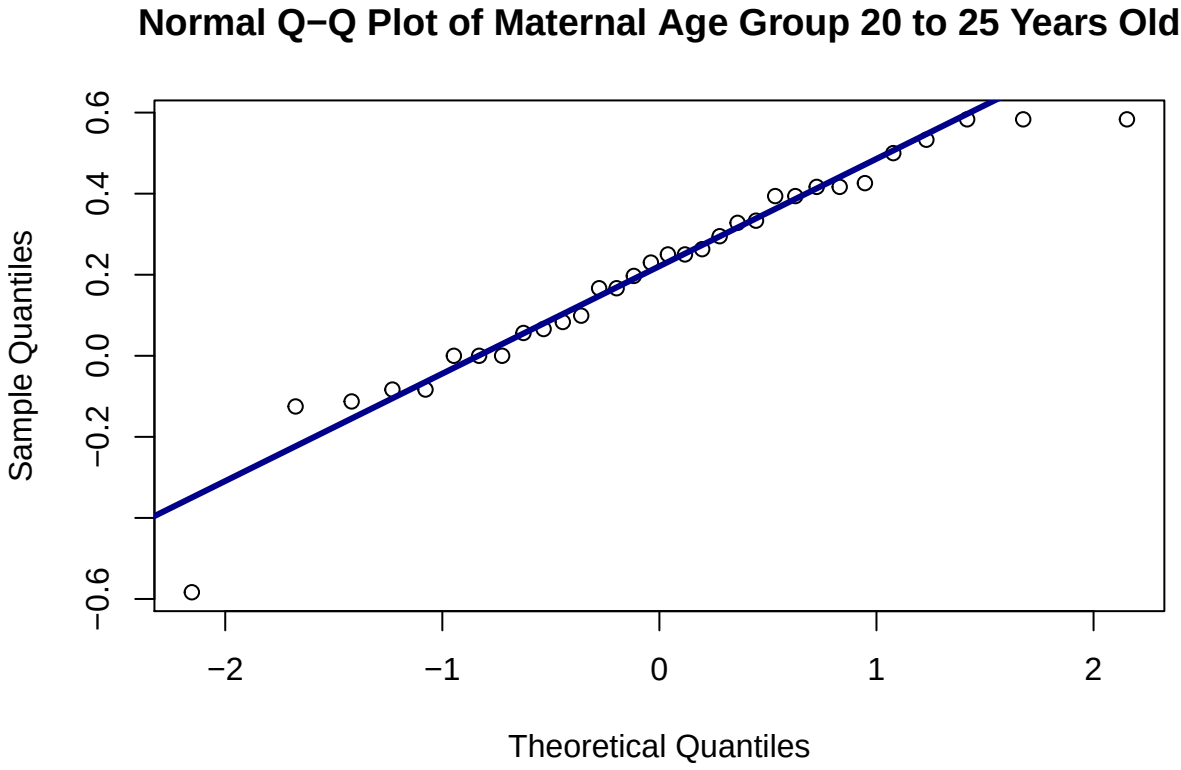



```
twoGroup_1000 <- do(1000) *mean(sample(group26to30$height_difference,15) -
                                sample(group20to25$height_difference,15))
hist(twoGroup_1000$mean,
     main = "Sampling Distribution",
     xlab = "Mean Difference Between Above Two Groups",
     prob = T,
     col = "darkred")
lines(density(twoGroup_1000$mean),
      col = "darkblue",
      lwd = 2)
```




## Two Samples Q-Q Plot Graphs

```
group20to25 <- useful_data %>%
  filter(maternal_age == '20-25 years')
qqnorm(group20to25$height_difference,
       main = "Normal Q-Q Plot of Maternal Age Group 20 to 25 Years Old")
qqline(group20to25$height_difference,
       col = "darkblue",
       lwd = 3)
```

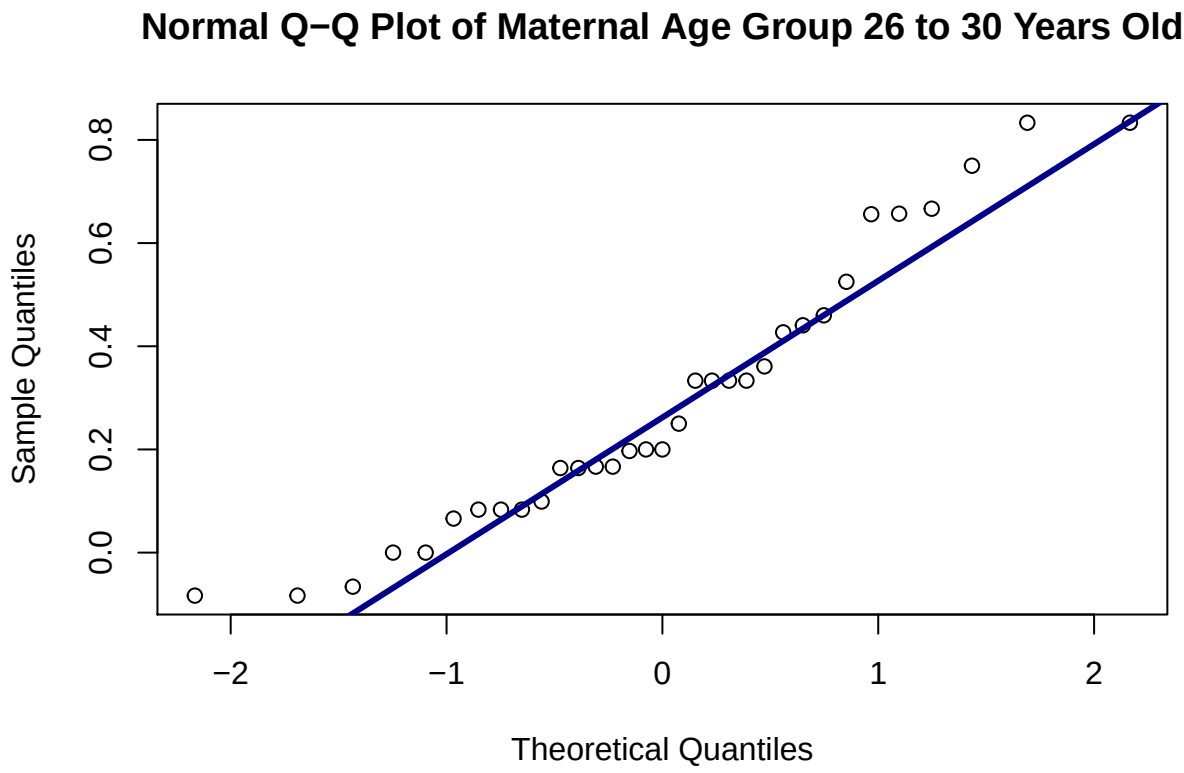



```
summary(group20to25)
```

```
##                       Timestamp     gender    nationality country_now
##  2019/11/19 9:21:41 PM PST : 6   Female:19   Chinese:17   China:11
##  2019/11/17 10:53:47 PM PST: 1   Male  :13   Indian :15   India: 9
##  2019/11/17 11:19:24 PM PST: 1                            USA  :12
##  2019/11/17 11:20:57 PM PST: 1
##  2019/11/17 3:51:30 PM PST : 1
##  2019/11/17 4:08:45 PM PST : 1
##  (Other)                   :21
##                  child_age   child_height                       exercise
##  Greater than 20 years:32   Min.   :5.083   Attended a sports class: 0
##                             1st Qu.:5.312   No                      :12
##                             Median :5.500   Yes                     :20
##                             Mean   :5.512
##                             3rd Qu.:5.694
##                             Max.   :6.083
##
##            milk          maternal_age mother_height
##  I don't know: 0   > 35 years  : 0    Min.   :5.000
##  No          :10   20-25 years :32    1st Qu.:5.085
##  Yes         :22   26-30 years : 0    Median :5.333
##                    31-35 years : 0    Mean   :5.305
##                    I don't know: 0    3rd Qu.:5.424
##                                       Max.   :5.667
##
##  Is.your.mother.working. height_difference
##  No : 9                  Min.   :-0.58337
##  Yes:23                  1st Qu.: 0.04223
##                          Median : 0.24002
```

```
##                          Mean   : 0.20710
##                          3rd Qu.: 0.39968
##                          Max.   : 0.58334
##
```

```
group26to30 <- useful_data %>%
  filter(maternal_age == '26-30 years')
qqnorm(group26to30$height_difference,
       main = "Normal Q-Q Plot of Maternal Age Group 26 to 30 Years Old")
qqline(group26to30$height_difference,
       col = "darkblue",
       lwd = 3)
```




```
summary(group26to30)
```

```
##                      Timestamp     gender    nationality country_now
##  2019/11/19 9:21:41 PM PST : 2   Female:19   Chinese:22   China:13
##  2019/11/17 10:29:33 PM PST: 1   Male  :14   Indian :11   India: 5
##  2019/11/17 3:40:51 PM PST : 1                            USA  :15
##  2019/11/17 4:25:38 PM PST : 1
##  2019/11/17 4:59:53 PM PST : 1
##  2019/11/17 5:36:10 PM PST : 1
##  (Other)                   :26
##                  child_age   child_height                      exercise
##  Greater than 20 years:33   Min.   :5.167   Attended a sports class: 0
##                             1st Qu.:5.500   No                      :15
##                             Median :5.643   Yes                     :18
##                             Mean   :5.631
##                             3rd Qu.:5.774
##                             Max.   :6.037
##
```

```
##            milk          maternal_age mother_height
##  I don't know: 0   > 35 years  : 0    Min.   :5.000
##  No          :15   20-25 years : 0    1st Qu.:5.184
##  Yes         :18   26-30 years :33    Median :5.413
##                    31-35 years : 0    Mean   :5.339
##                    I don't know: 0    3rd Qu.:5.446
##                                       Max.   :5.600
##
##  Is.your.mother.working. height_difference
##  No :10                  Min.   :-0.08334
##  Yes:23                  1st Qu.: 0.08334
##                          Median : 0.20000
##                          Mean   : 0.29194
##                          3rd Qu.: 0.44070
##                          Max.   : 0.83334
##
```

## Two Samples T-Test

```
t.test(group20to25$height_difference, group26to30$height_difference, var.equal = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  group20to25$height_difference and group26to30$height_difference
## t = -1.3089, df = 62.991, p-value = 0.1953
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.21438259  0.04469463
## sample estimates:
## mean of x mean of y
## 0.2070987 0.2919427
```

```
# sample means
x_bar_20to25 <- mean(group20to25$height_difference)
x_bar_26to30 <- mean(group26to30$height_difference)

# null hypothesized population mean difference between the two groups
mu_0 <- 0

# sample variances
s_20to25_sq <- sd(group20to25$height_difference) ** 2
s_26to30_sq <- sd(group26to30$height_difference) ** 2

# sample size
n_20to25 <- length(group20to25$height_difference)
n_26to30 <- length(group26to30$height_difference)

# t-test test statistic
t <- (x_bar_20to25 - x_bar_26to30 - mu_0)/sqrt((s_20to25_sq/n_20to25) +
                                               (s_26to30_sq/n_26to30))

# one sided upper p-value
two_sided_diff_t_pval <- pt(q = t, df = min(n_20to25, n_26to30)-1, lower.tail = TRUE)*2
```

```
two_sided_diff_t_pval
```

```
## [1] 0.2002027
```