FA9 - GROUP 1

Cobarrubias, Dela Rosa, Quijano, Sigue

2025-05-06

```
# install.packages('moments') if need be
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.4.3
```

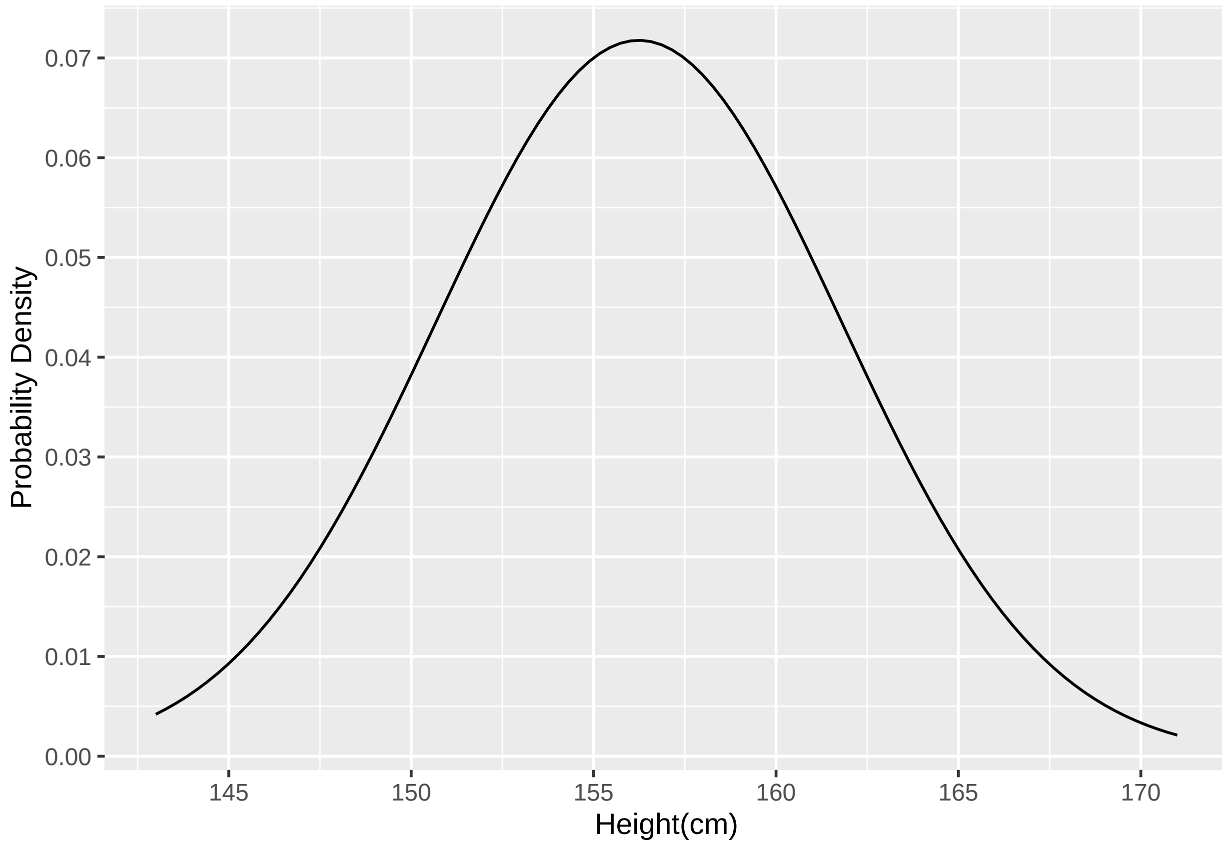
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```
library(moments)

data = read_excel("Data (PROB FA9).xlsx")
```

```
## New names:
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
```

```
height_data = as.numeric(data$`Height (cm)`[1:50])

mean_data = as.numeric(data$`Height (cm)`[51])

sdev_data = as.numeric(data$`Height (cm)`[52])

min_x <- floor(min(height_data)/5) * 5  #For graph visuals only

max_x <- ceiling(max(height_data)/5) * 5  #For graph visuals only

height_distribution = ggplot(data.frame(height_data), aes(x = height_data)) +
    stat_function(fun = dnorm, n = 100, args = list(mean = mean_data, sd = sdev_data)) +
    ylab("Probability Density") + scale_y_continuous(breaks = seq(0, 0.1,
    by = 0.01)) + xlab("Height(cm)") + scale_x_continuous(breaks = seq(min_x,
    max_x, by = 5))

height_distribution
```




```
# Identify the percentage of data that falls within 1sd, 2sd, and 3sd
# from the mean.

one_sd = (pnorm(mean_data + sdev_data, mean = mean_data, sd = sdev_data) -
    pnorm(mean_data - sdev_data, mean = mean_data, sd = sdev_data)) * 100

two_sd = (pnorm(mean_data + 2 * sdev_data, mean = mean_data, sd = sdev_data) -
    pnorm(mean_data - 2 * sdev_data, mean = mean_data, sd = sdev_data)) *
    100

three_sd = (pnorm(mean_data + 3 * sdev_data, mean = mean_data, sd = sdev_data) -
    pnorm(mean_data - 3 * sdev_data, mean = mean_data, sd = sdev_data)) *
    100

cat("Percentage of data within 1 standard deviation:", one_sd, "%\n")
```

```
## Percentage of data within 1 standard deviation: 68.26895 %
```

```
cat("Percentage of data within 2 standard deviations:", two_sd, "%\n")
```

```
## Percentage of data within 2 standard deviations: 95.44997 %
```

```
cat("Percentage of data within 3 standard deviations:", three_sd, "%\n")
```

```
## Percentage of data within 3 standard deviations: 99.73002 %
```

Interpret what the distribution tells you about your campus variable.

Is the distribution symmetric?

Skewness Check

```r
skew <- skewness(height_data)
cat("Skewness of the height data:", skew, "\n")
```

```
## Skewness of the height data: 0.3977994
```

Since the skewness of the data is approximately 0.3978, we can assume the graph would be fairly symmetrical. With a slightly longer right tail (positively skewed).

Visual Check

By visually checking the graph is nearly symmetric, with a slight shift to the left.

Are there outliers?

Using the Standard Deviation

```r
# Check for outliers above mean + 2 SD
outliers_above <- height_data[height_data > (mean_data + 2 * sdev_data)]

if (length(outliers_above) > 0) {
    cat("Height values greater than", round(mean_data + 2 * sdev_data,
        2), "cm:\n")
    print(outliers_above)
} else {
    cat("No outliers above", round(mean_data + 2 * sdev_data, 2), "cm\n")
}
```

Using 2SD (Mild Outliers)

```
## Height values greater than 167.36 cm:
## [1] 171 168
```

```r
# Check for outliers below mean - 2 SD
outliers_below <- height_data[height_data < (mean_data - 2 * sdev_data)]

if (length(outliers_below) > 0) {
    cat("Height values less than", round(mean_data - 2 * sdev_data, 2),
        "cm:\n")
    print(outliers_below)
} else {
    cat("No outliers below", round(mean_data - 2 * sdev_data, 2), "cm\n")
}
```

```
## Height values less than 145.12 cm:
## [1] 143
```

```
# Check for outliers above mean + 3 SD
outliers_above <- height_data[height_data > (mean_data + 3 * sdev_data)]

if (length(outliers_above) > 0) {
    cat("Height values greater than", round(mean_data + 3 * sdev_data,
        2), "cm:\n")
    print(outliers_above)
} else {
    cat("No outliers above", round(mean_data + 3 * sdev_data, 2), "cm\n")
}
```

Using 3SD (Extreme Outliers)

```
## No outliers above 172.92 cm
```

```
# Check for outliers below mean - 3 SD
outliers_below <- height_data[height_data < (mean_data - 3 * sdev_data)]

if (length(outliers_below) > 0) {
    cat("Height values less than", round(mean_data - 3 * sdev_data, 2),
        "cm:\n")
    print(outliers_below)
} else {
    cat("No outliers below", round(mean_data - 3 * sdev_data, 2), "cm\n")
}
```

```
## No outliers below 139.56 cm
```

Using IQR

```
# Calculate Q1, Q3, and IQR
Q1 <- quantile(height_data, 0.25)
Q3 <- quantile(height_data, 0.75)
IQR <- Q3 - Q1

# Calculate outlier thresholds
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

# Print IQR and related statistics
cat("Interquartile Range (IQR):", IQR, "cm\n")
```

```
## Interquartile Range (IQR): 8 cm
```

```r
cat("Q1 (25th percentile):", Q1, "cm\n")
```

```
## Q1 (25th percentile): 152 cm
```

```r
cat("Q3 (75th percentile):", Q3, "cm\n")
```

```
## Q3 (75th percentile): 160 cm
```

```r
cat("Lower bound (Q1 - 1.5 * IQR):", lower_bound, "cm\n")
```

```
## Lower bound (Q1 - 1.5 * IQR): 140 cm
```

```r
cat("Upper bound (Q3 + 1.5 * IQR):", upper_bound, "cm\n")
```

```
## Upper bound (Q3 + 1.5 * IQR): 172 cm
```

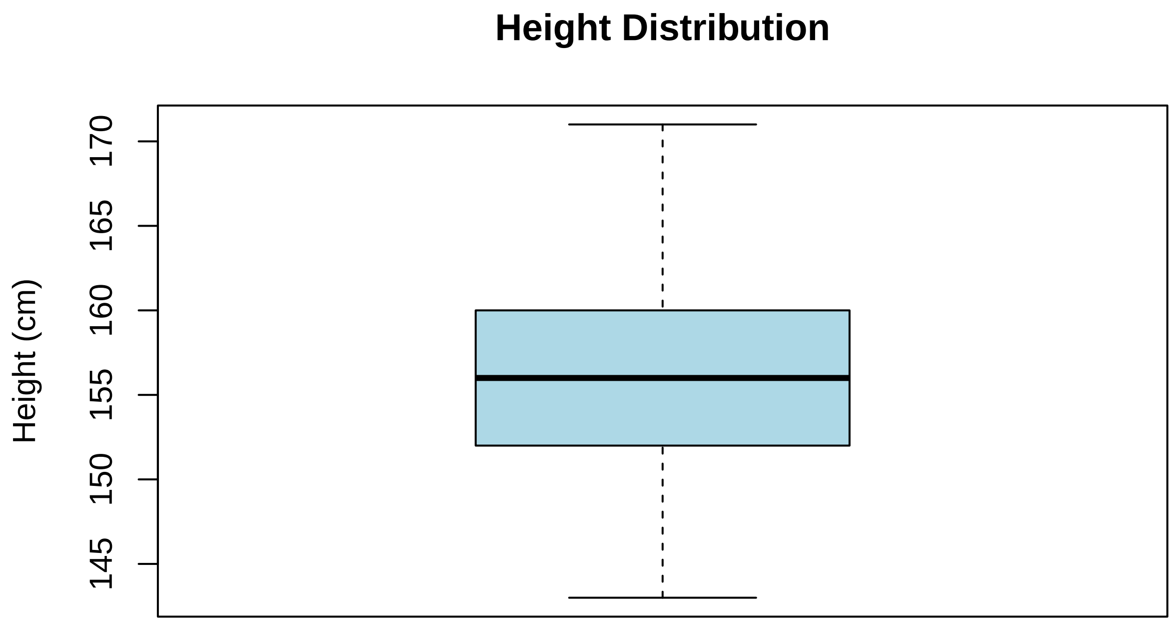
```r
# Find outliers
outliers_below <- height_data[height_data < lower_bound]
outliers_above <- height_data[height_data > upper_bound]

# Check for outliers below lower bound
if (length(outliers_below) > 0) {
    cat("Height values less than", lower_bound, "cm:\n")
    print(outliers_below)
} else {
    cat("No outliers below", lower_bound, "cm\n")
}
```

```
## No outliers below 140 cm
```

```r
# Check for outliers above upper bound
if (length(outliers_above) > 0) {
    cat("Height values greater than", upper_bound, "cm:\n")
    print(outliers_above)
} else {
    cat("No outliers above", upper_bound, "cm\n")
}
```

```
## No outliers above 172 cm
```

```r
boxplot(height_data, main = "Height Distribution", ylab = "Height (cm)",
    col = "lightblue", border = "black")
```



What does the shape of the distribution imply?

The shape of the distribution shows that most of the height of students are close to the average. It looks like a bell curve, so it means it is almost normal distribution. There is a small skew to the right side, which means there are few students that are taller than the others. But overall, the graph still looks symmetrical. This kind of distribution is common when it comes to height or other natural data.

How can this data be useful?

This data can help the FEU in many ways. For example, they can use it for designing chairs, tables, or other facilities that fits the average height of students. It can also help in sports, like if they want to check which students are more fit for some activities. It's also useful in research and maybe even in health checking, as it gives information about the physical characteristics of the students in campus.