STAT011 Statistical Methods I

Lecture 4 Normal Distribution

Lu Chen
Swarthmore College
1/31/2019

Review - Exploratory data analysis

	Summary statistics	Data visualization
Categorical variables	Table of counts `table()` and proportions `prop.table()`	Bar plot `barplot()` Pie chart `pie()`
Quantitative variables	Mean `mean()` Median `median()` SD `sd()` Quartiles `quantile()` IQR `IQR()` Minimum `min()` Maximum `max()` 5-number summary `summary()`	Histogram `hist()` Boxplot `boxplot()`

- The $\mathbf{1.5 \times \textit{IQR}}$ rule for suspected outliers.
- Effect of linear transformations.

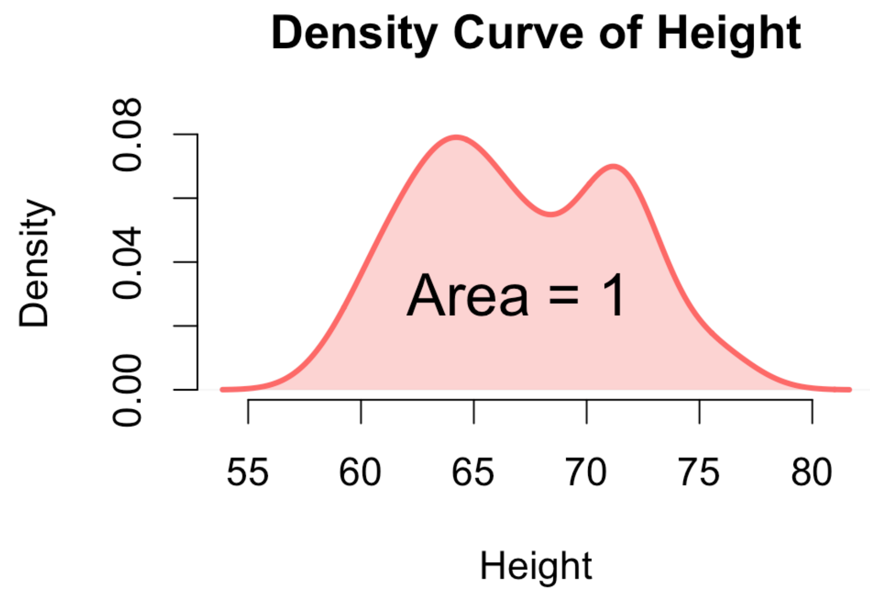
Outline

- Density curve
  - Properties
  - Normal curve
- Normal distribution
  - Density function
  - The 68-95-99.7 rule
  - Standard Normal distribution and standardization
  - Assessing Normality - Normal Quantile-Quantile Plot

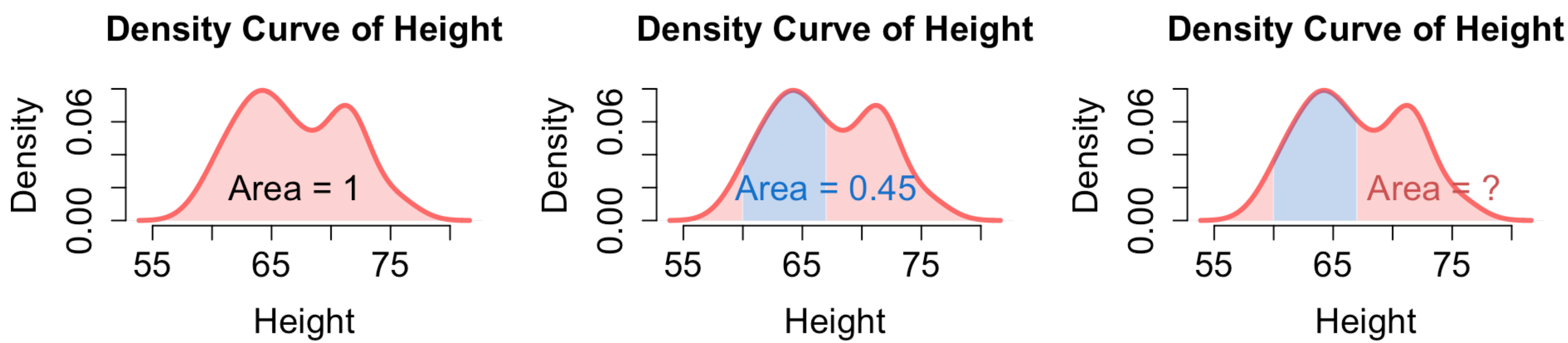
Density curve

A **density curve** describes the overall pattern of a distribution. The **area** under the curve and above any range of values is the **proportion** of all observations that fall in that range.

- It is always on or above the horizontal axis.
- It has area exactly 1 underneath it.

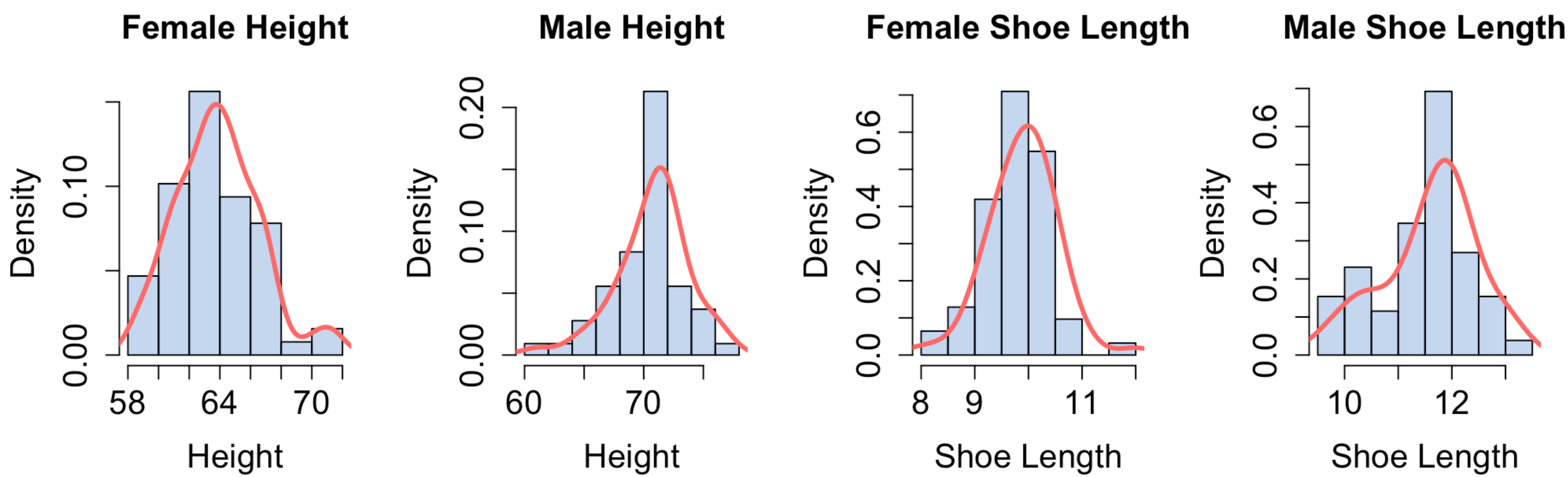


Density curve



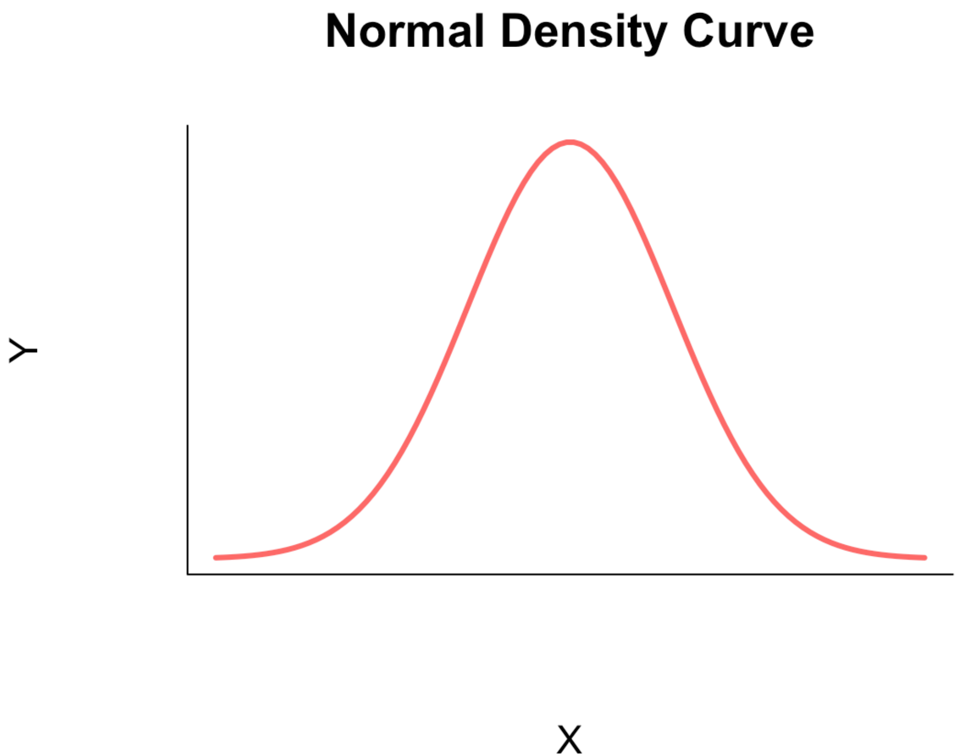
- The total area under the curve is 1.
- Suppose the blue area under the curve and between 60 and 67 inches is 0.45. This means that the proportion of students whose height falls between 60 and 67 inches is 45%.
- What is the proportion of students whose height is below 60 and above 67 inches?
- $1 - 0.45 = 0.55$

Normal curve



- These density curves are **symmetric**, **unimodal** and **bell-shaped**.
- They are called **Normal curves** and used to describe **Normal distributions**.

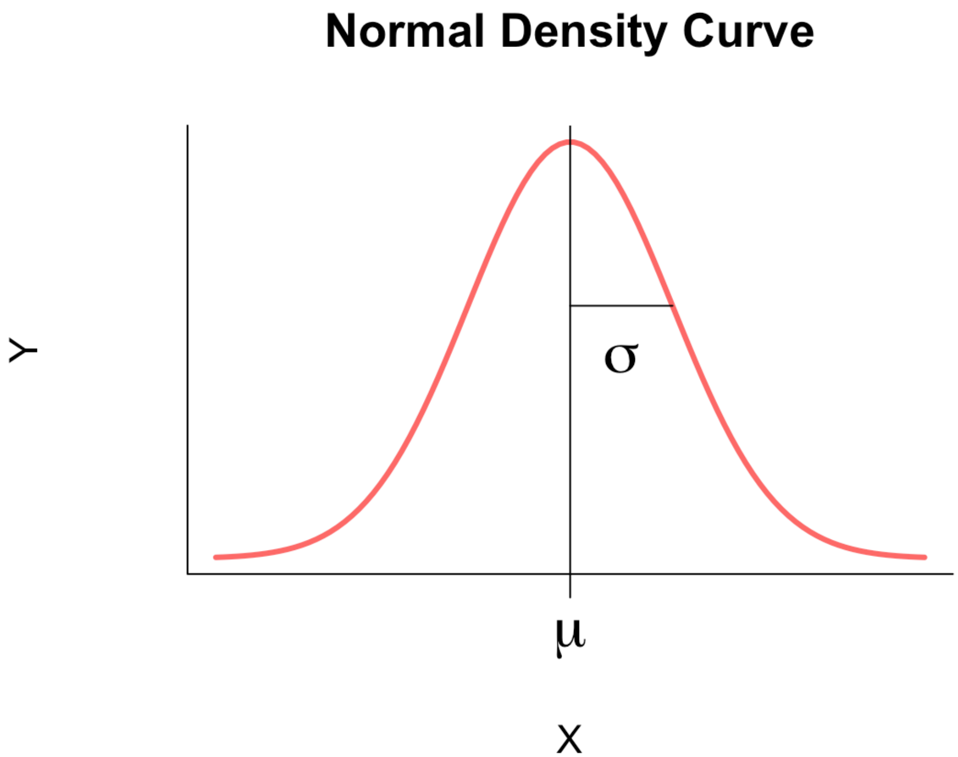
Normal distribution - Density function



The **Normal density curve** is characterized by the following **density function**

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Normal distribution

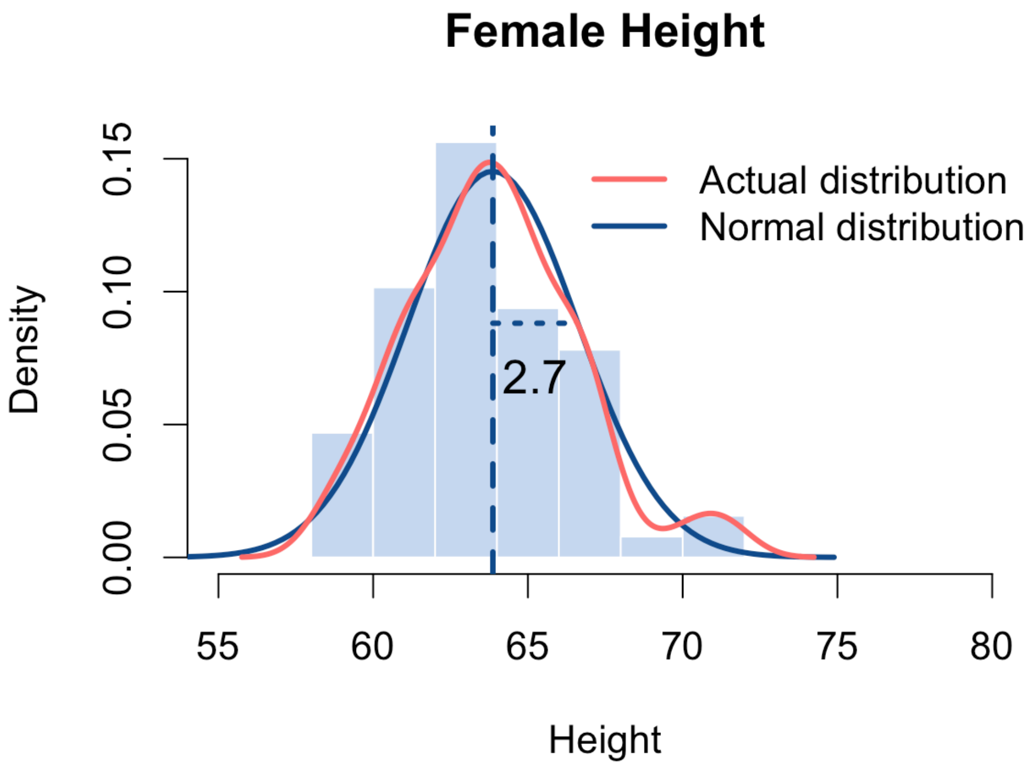


The Normal density curve is characterized by

- Center μ: mean of the distribution
- Spread σ: standard deviation of the distribution

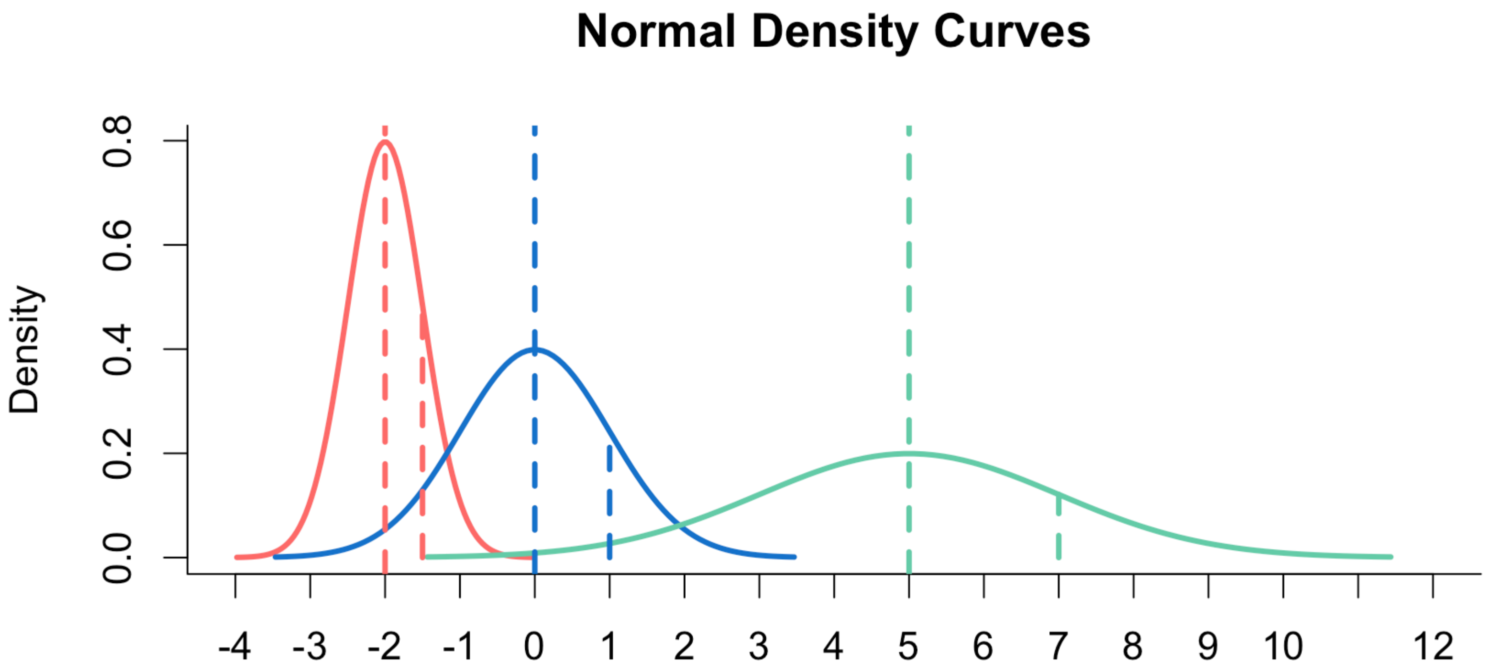
and expressed as $X \sim N(\mu, \sigma)$, "X follows a Normal distribution with mean μ and standard deviation σ".

Normal distribution



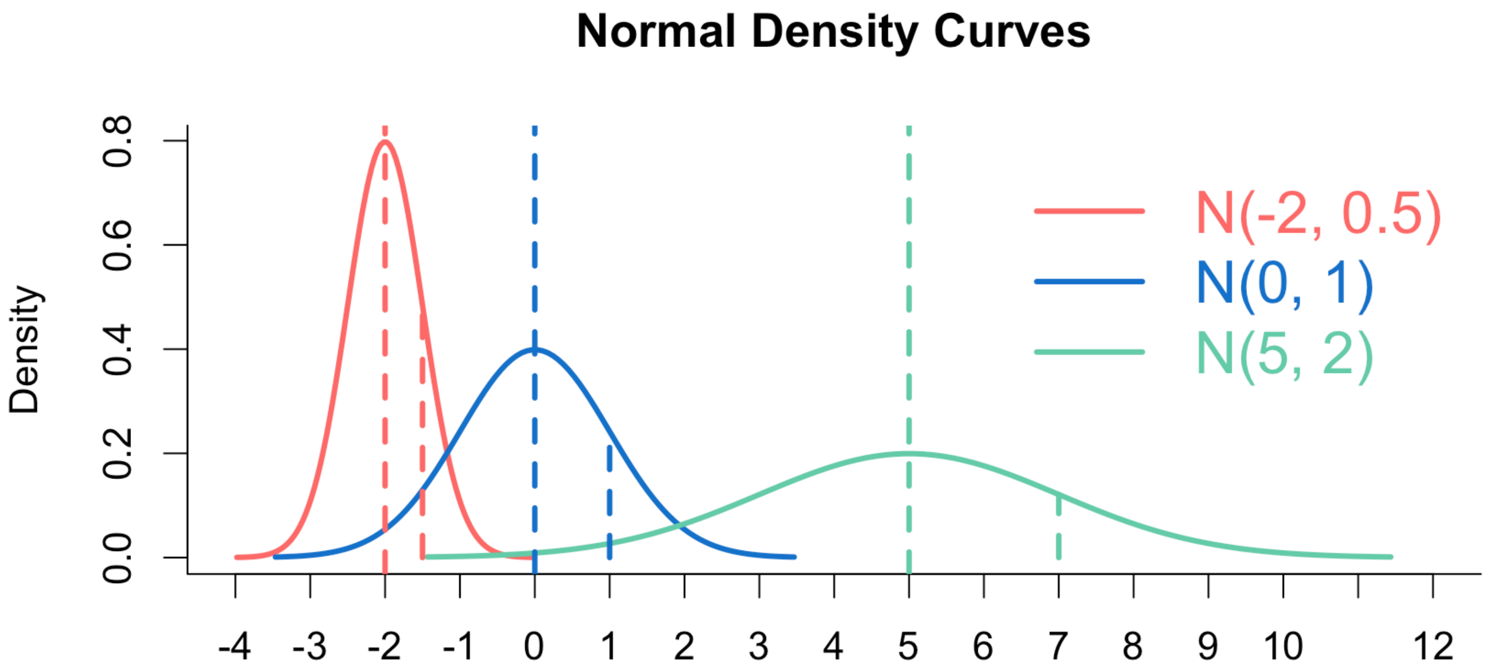
- Female height, mean = 63.9 inches, SD = 2.7 inches.
- Female height $\overset{approx.}{\sim}$ $N(63.9, 2.7)$ "Female height follows an approximately Normal distribution with mean 63.9 and SD 2.7."

Normal distribution



Write down the form $N(\mu, \sigma)$ for each Normal density curve.

Normal distribution

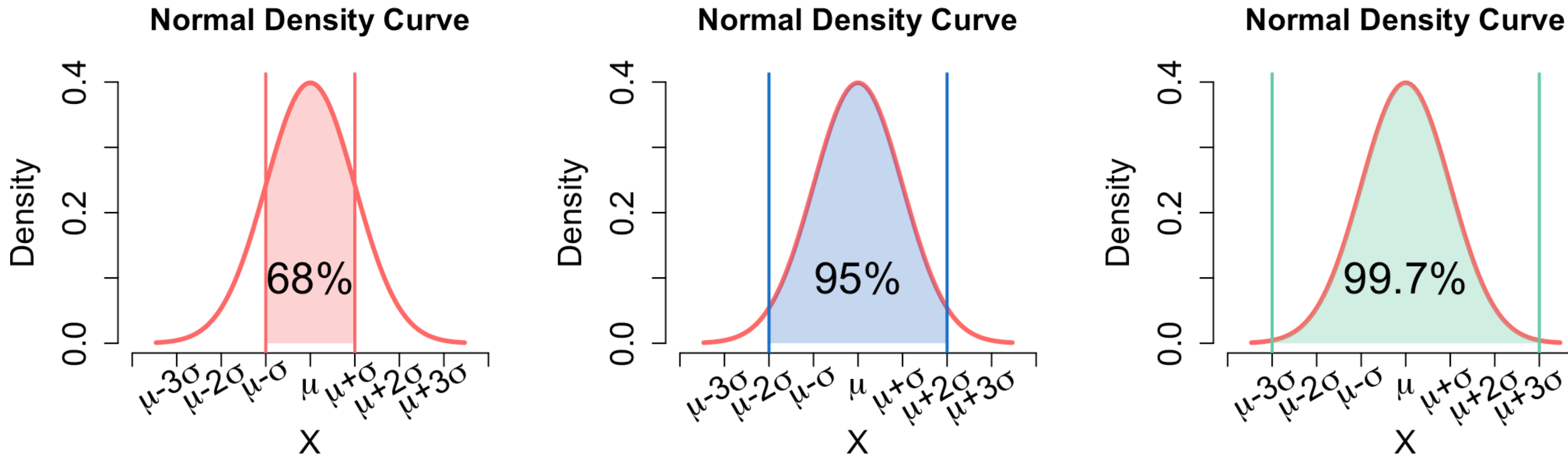


The 68-95-99.7 rule

> In any Normal distribution with mean μ and standard deviation σ:
>
> - Approximately 68% of the observations fall within σ of the mean μ.
> - Approximately 95% of the observations fall within 2σ of μ.
> - Approximately 99.7% of the observations fall within 3σ of μ.

- This is an important characteristic of Normal distribution.
- This is true for any value of μ and σ.

The 68-95-99.7 rule



- The area under the curve within σ of μ is 0.68.
- The area under the curve within 2σ of μ is 0.95.
- The area under the curve within 3σ of μ is 0.997.

The 68-95-99.7 rule - Example 1

Female height $\overset{approx.}{\sim}$ $N(63.9, 2.7)$.

Height of **about** 68% of the female students falls between

- $63.9 - 2.7$ and $63.9 + 2.7$ inches

Height of **about** 95% of the female students falls between

- $63.9 - 2 \times 2.7$ and $63.9 + 2 \times 2.7$ inches

Height of **about** 99.7% of the female students falls between

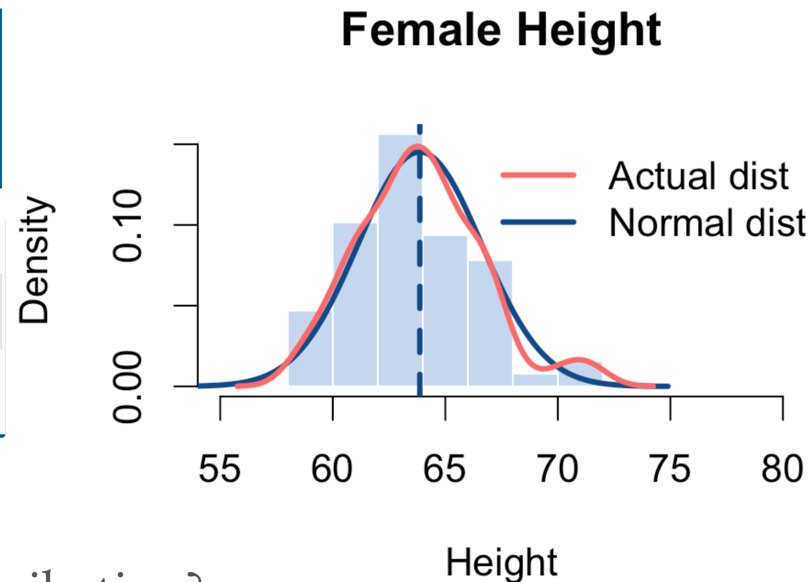
- $63.9 - 3 \times 2.7$ and $63.9 + 3 \times 2.7$ inches

Note: the distribution of female height is NOT exactly Normal and these calculations are approximations.

The 68-95-99.7 rule - Example 1

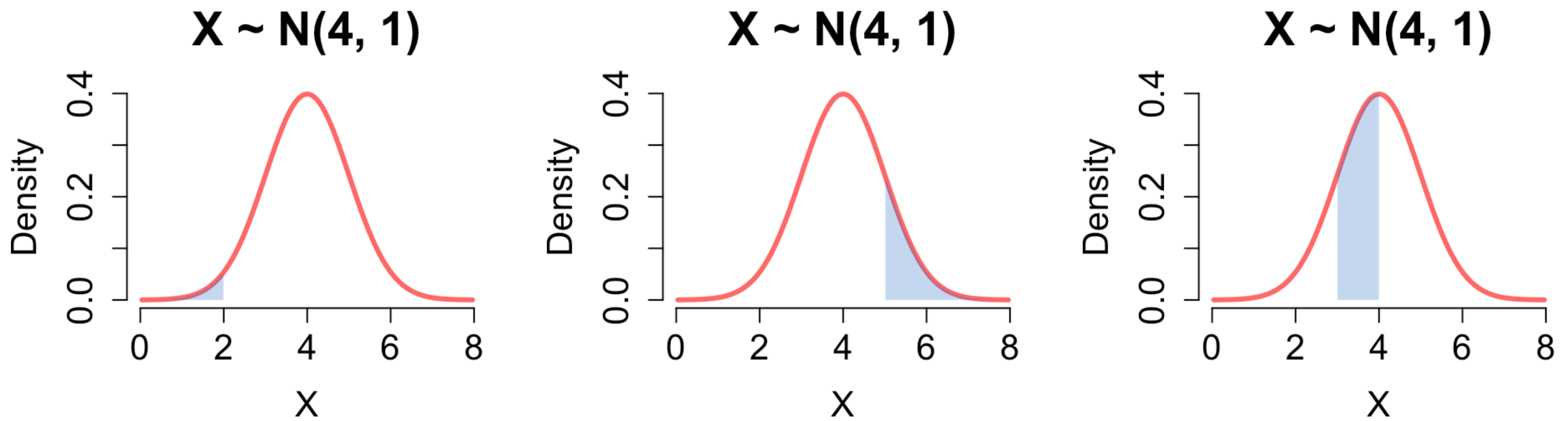
Distribution of female height, comparing the actual distribution to the Normal approximation $N(63.9, 2.7)$.

Proportions	Actual distribution (by percentiles)	Normal distribution (by 68-95-99.7 rule)
68%	[61.0, 66.5]	[61.2, 66.6]
95%	[59.0, 70.4]	[58.5, 69.3]
99.7%	[58.5, 71.5]	[55.8, 72.0]



- How to obtain the intervals from the actual distribution?
- For 68%, find the 16th and 84th percentile using the `quantile()` function.

The 68-95-99.7 rule - Example 2

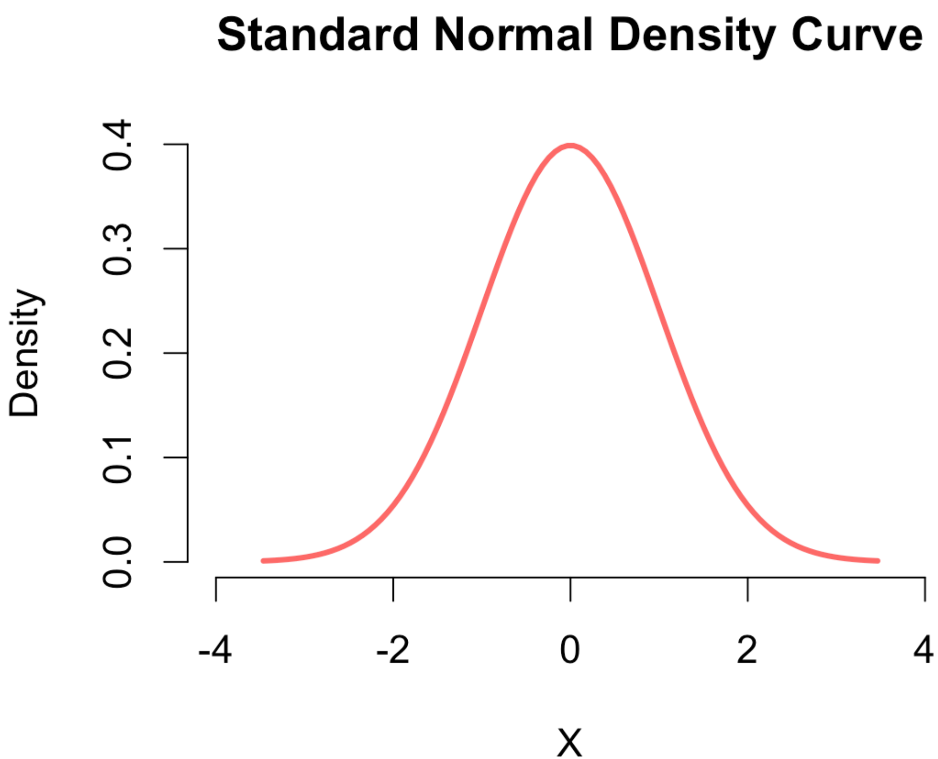


Suppose variable X follows a Normal distribution with mean 4 and SD 1,

- What is the proportion of observations whose x values fall below 2?
- What is the proportion of observations whose x values fall above 5?
- What is the proportion of observations whose x values fall between 3 and 4?

Standard Normal distribution

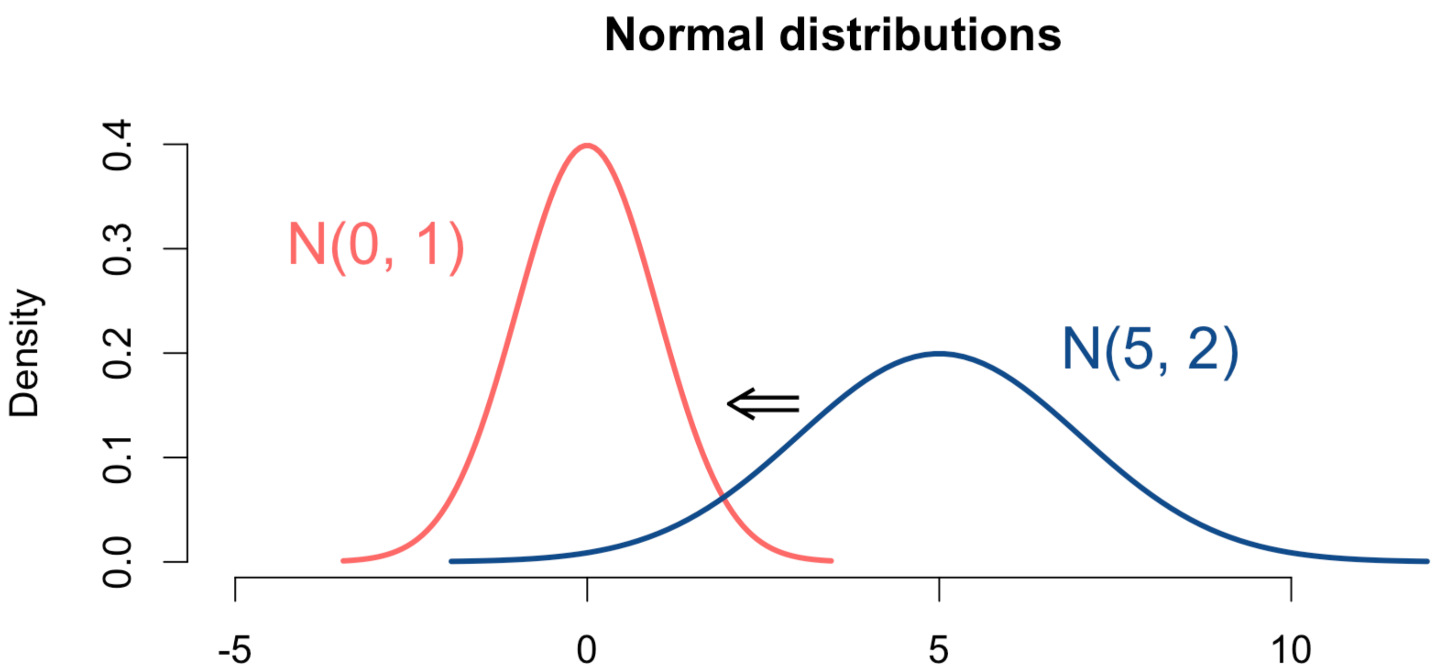
The **standard Normal distribution** is the Normal distribution $N(0, 1)$ with mean 0 and standard deviation 1.



- How does the 68-95-99.7 rule apply on the standard Normal distribution?
- 68% observations fall within $[-1, 1]$
- 95% observations fall within $[-2, 2]$
- 99.7% observations fall within $[-3, 3]$

Standard Normal distribution

Any Normal distribution can be transformed to the standard Normal distribution.



Standardization

If a variable X has **any** Normal distribution $N(\mu, \sigma)$ with mean μ and standard deviation σ, then the **standardized** variable

$$Z = \frac{X - \mu}{\sigma}$$

has the standard Normal distribution.

The tranformation from X to Z is called **standardization**.

- If $X \sim N(\mu, \sigma)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$.
- For example, $X \sim N(5, 2)$, then $Z = \frac{X-5}{2} \sim N(0, 1)$.

Standardization

- Upper case X and Z usually denote variables.
- Lower case x and z usually denote specific values from the variables.

If x is an observation from $X \sim N(\mu, \sigma)$, the **standardized value** of x is

$$z = \frac{x - \mu}{\sigma}$$

It is called the **standardized score of** x, or z**-score**.

- For $X \sim N(5, 2)$, suppose $x = 3$ is an observation from X. Then $z = (x - 5)/2 = (3 - 5)/2 = -1$ is the z-score of x.
- What about $x = 10$? If $z = 0.5$, what's the corresponding x value?
- $x = 10 \Rightarrow z = (10 - 5)/2 = 2.5$;
  For $z = 0.5$, $z = \frac{x-5}{2} = 0.5 \Rightarrow x = 2 \times 0.5 + 5 = 6$

Normal distribution in R

Function `dnorm()` calculates the density value on the curve for a given x value.

```r
dnorm(x=0) # N(0, 1)
```

```
## [1] 0.3989423
```

```r
dnorm(x=1) # N(0, 1)
```

```
## [1] 0.2419707
```

```r
dnorm(x=1, mean=5, sd=2) # N(5, 2)
```

```
## [1] 0.02699548
```

```r
dnorm((1-5)/2) # N(0, 1)
```

```
## [1] 0.05399097
```

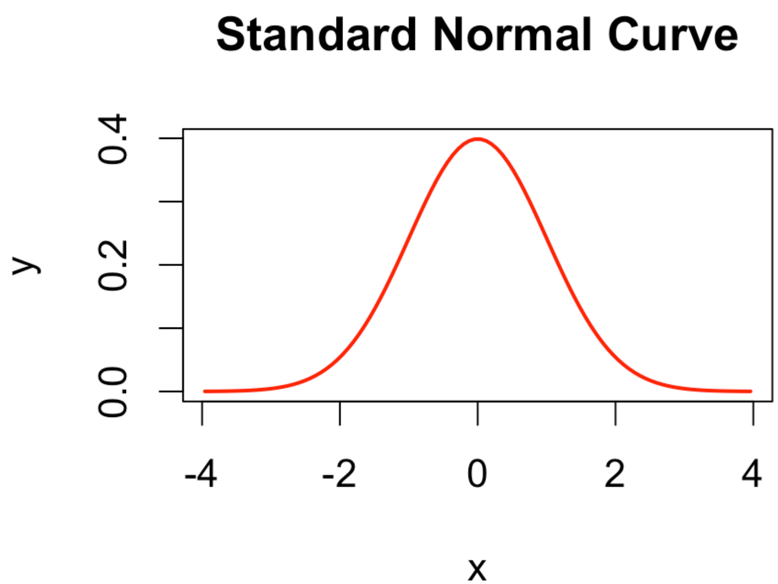
```r
dnorm(4) # N(0, 1)
```

```
## [1] 0.0001338302
```

Normal distribution in R

Function `dnorm()` calculates the density value on the curve for a given x value.

```
# Plot standard Normal curve
x <- ppoints(100)*8-4 # ppoints(100): 100 values between 0 and 1
y <- dnorm(x)
plot(x, y, type="l", lwd=2, col="red", main="Standard Normal Curve")
```



Normal distribution in R

Function `pnorm()` calculates the area under the curve for values below a given x, i.e. proportion of observations with values below a given x.

```
pnorm(0) # N(0, 1)
```

```
## [1] 0.5
```

```
pnorm(1) # N(0, 1)
```

```
## [1] 0.8413447
```

```
pnorm(1, mean=5, sd=2) # N(5, 2)
```

```
## [1] 0.02275013
```

```
pnorm((1-5)/2) # N(0,1)
```

```
## [1] 0.02275013
```

Normal distribution in R

`pnorm(x1) - pnorm(x2)` calculates the area under the curve for values between `x1` and `x2`.

```
pnorm(1) - pnorm(-1)
```

```
## [1] 0.6826895
```

```
pnorm(1.5) - pnorm(-2.5)
```

```
## [1] 0.9269831
```

```
pnorm(6, mean=5, sd=2) - pnorm(4, mean=5, sd=2)
```

```
## [1] 0.3829249
```

```
pnorm((6-5)/2) - pnorm((4-5)/2)
```

```
## [1] 0.3829249
```

Normal distribution in R

Function `qnorm()` calculates the quantile of a variable for a given percentage. It is the inverse function of `pnorm()`.

```r
qnorm(0.5) # N(0, 1)
```

```
## [1] 0
```

```r
qnorm(0.025) # N(0, 1)
```

```
## [1] -1.959964
```

```r
qnorm(0.975) # N(0, 1)
```

```
## [1] 1.959964
```

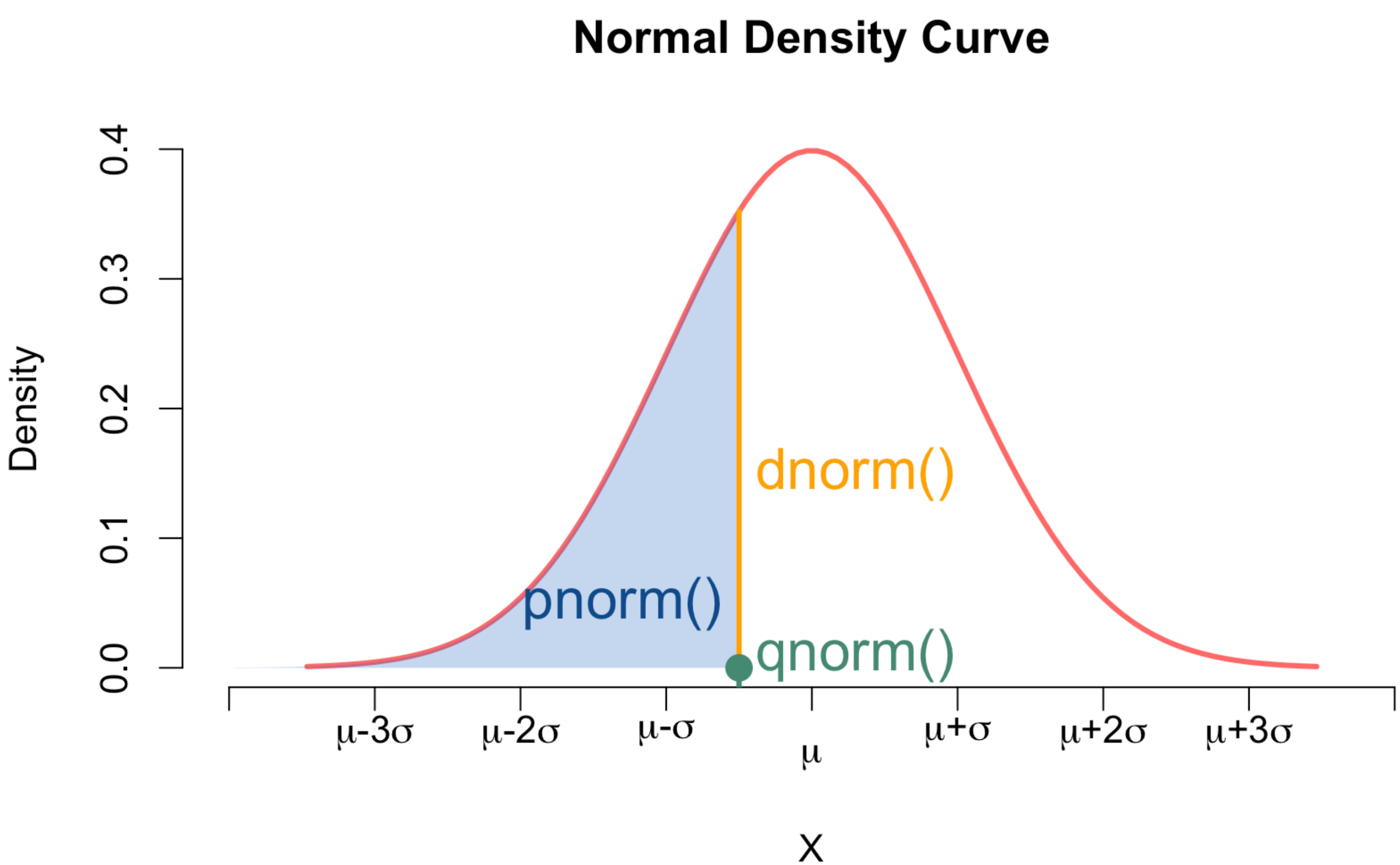
```r
qnorm(0.1, 5, 2) # N(5, 2)
```

```
## [1] 2.436897
```

```r
qnorm(0.1)*2 + 5 # N(0, 1)
```

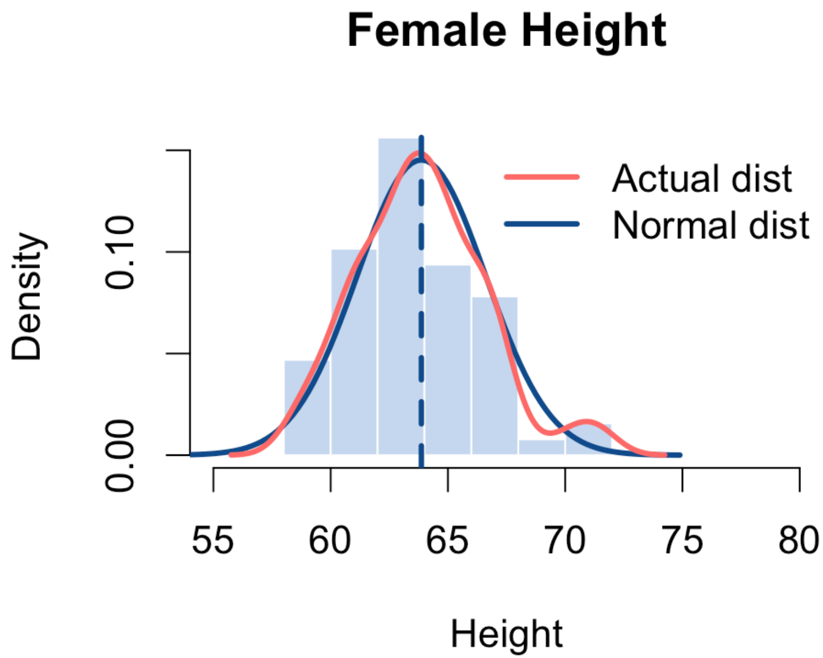
```
## [1] 2.436897
```

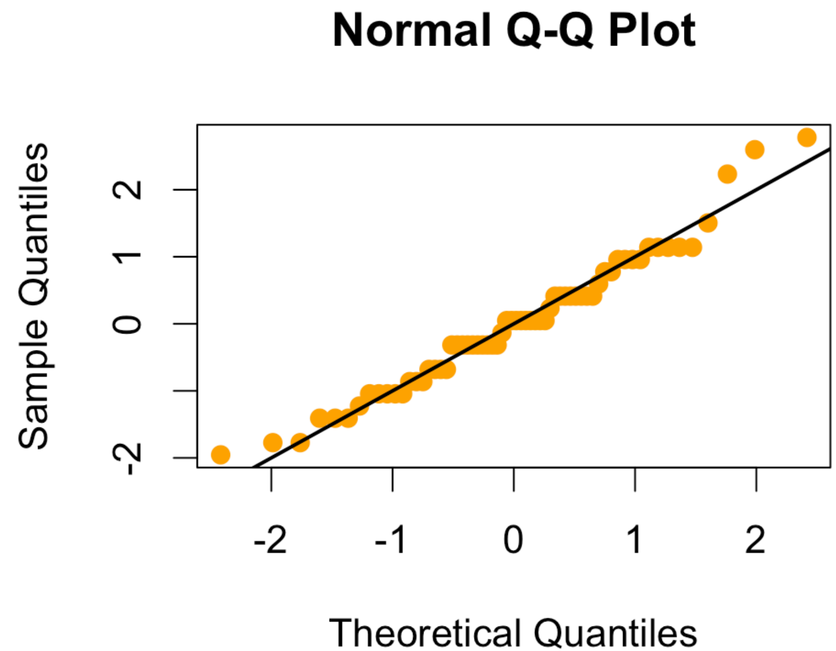
Normal distribution in R



Assessing Normality

Normal Quantile-Quantile Plot

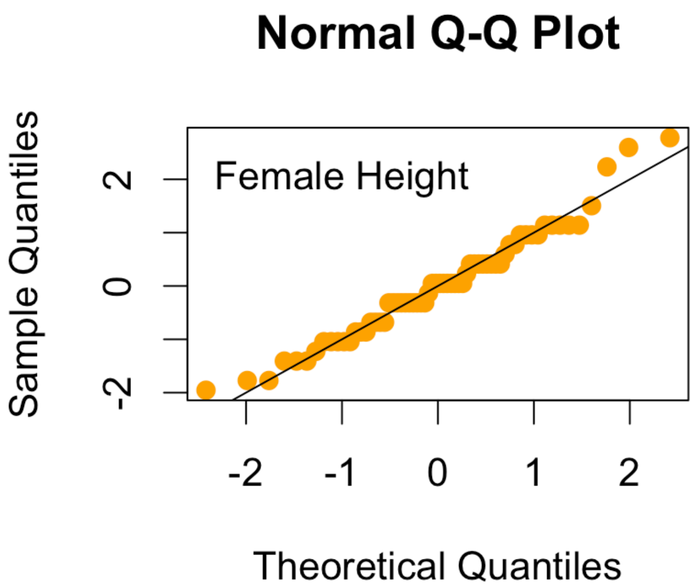


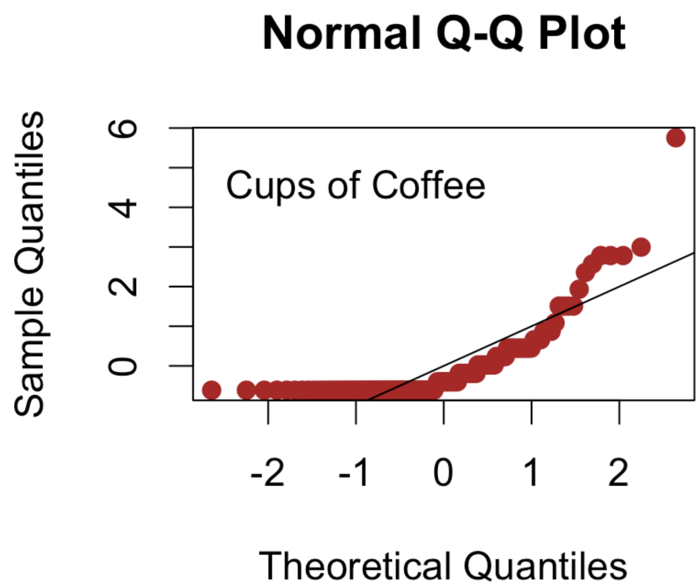


- Normal Q-Q plot compares the distribution of interest (usually after standardization) to the standard Normal distribution.
- If the distribution of interest is close to a Normal distribution, the points on the Q-Q plot should **lie close to the $y = x$ line**.

Assessing Normality

```
female_height <- Survey$Height[Survey$Gender == "F"]
mean_fh <- mean(female_height, na.rm=T) # Mean of female height
sd_fh <- sd(female_height, na.rm=T) # SD of female height
s_female_height <- (female_height - mean_fh)/sd_fh # Standardization
qqnorm(s_female_height, pch=19, col="orange") # Q-Q plot
abline(a=0, b=1) # Add the y=x line (intercept=0 and slope=1)
```





Summary

- Density curve
  - Properties: area under the curve $= 1$; area = proportion;
  - Normal curve: symmetric, unimodal, bell-shaped
- Normal distribution: $N(\mu, \sigma)$
  - Density function
  - The 68-95-99.7 rule
  - Standard Normal distribution $N(0, 1)$ and standardization
    - `dnorm()`, `pnorm()`, `qnorm()`
  - Assessing Normality: Normal Q-Q plot
    - `qqnorm()`, `abline()`