STAT011 Statistical Methods I

Lecture 3 Exploratory Data Analysis II

Lu Chen
Swarthmore College
1/29/2019

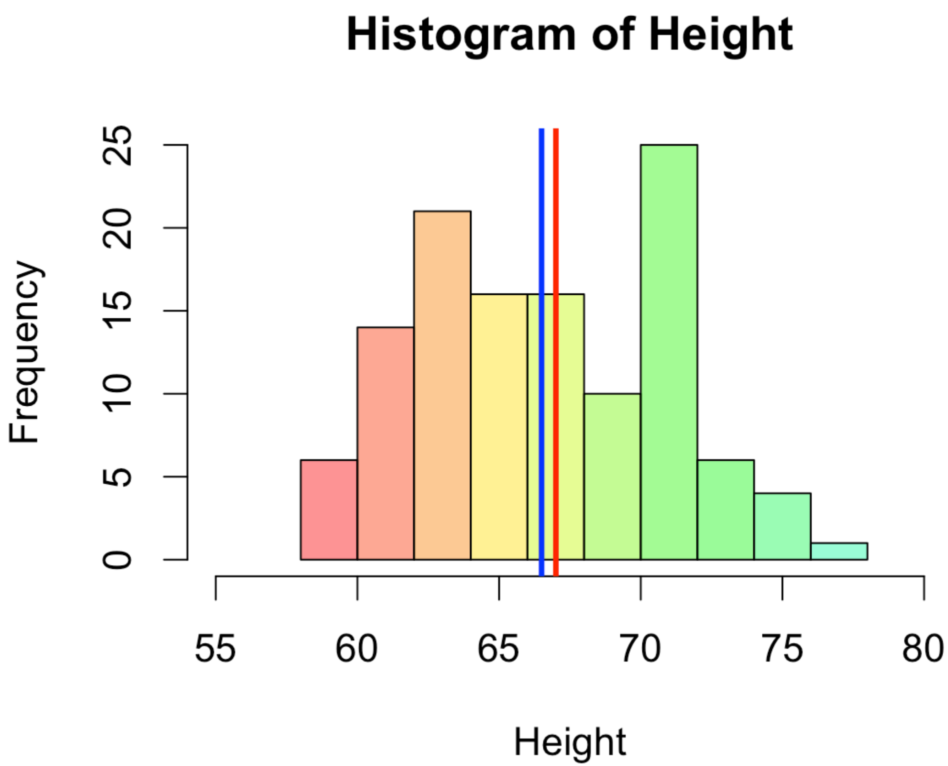
Review

	Summary statistics	**Data visualization**
Categorical variables	Table of counts `table()` and proportions `prop.table()`	Bar plot `barplot()` Pie chart `pie()`
Quantitative variables	Mean `mean()` (center) Median `median()` (center) **Standard deviation (spread)** **Interquartile range (spread)**	Histogram `hist()` **Boxplot**

Outline

- Exploratory data analysis: the spread of a quantitative variable
  - Standard deviation
  - Quartiles and five-number summary
  - Interquartile range (IQR) and range
  - The $1.5 \times \text{IQR}$ rule
  - Boxplot
  - Linear transformation of a quantitative variable
- Density curve

Distribution of a quantitative variable

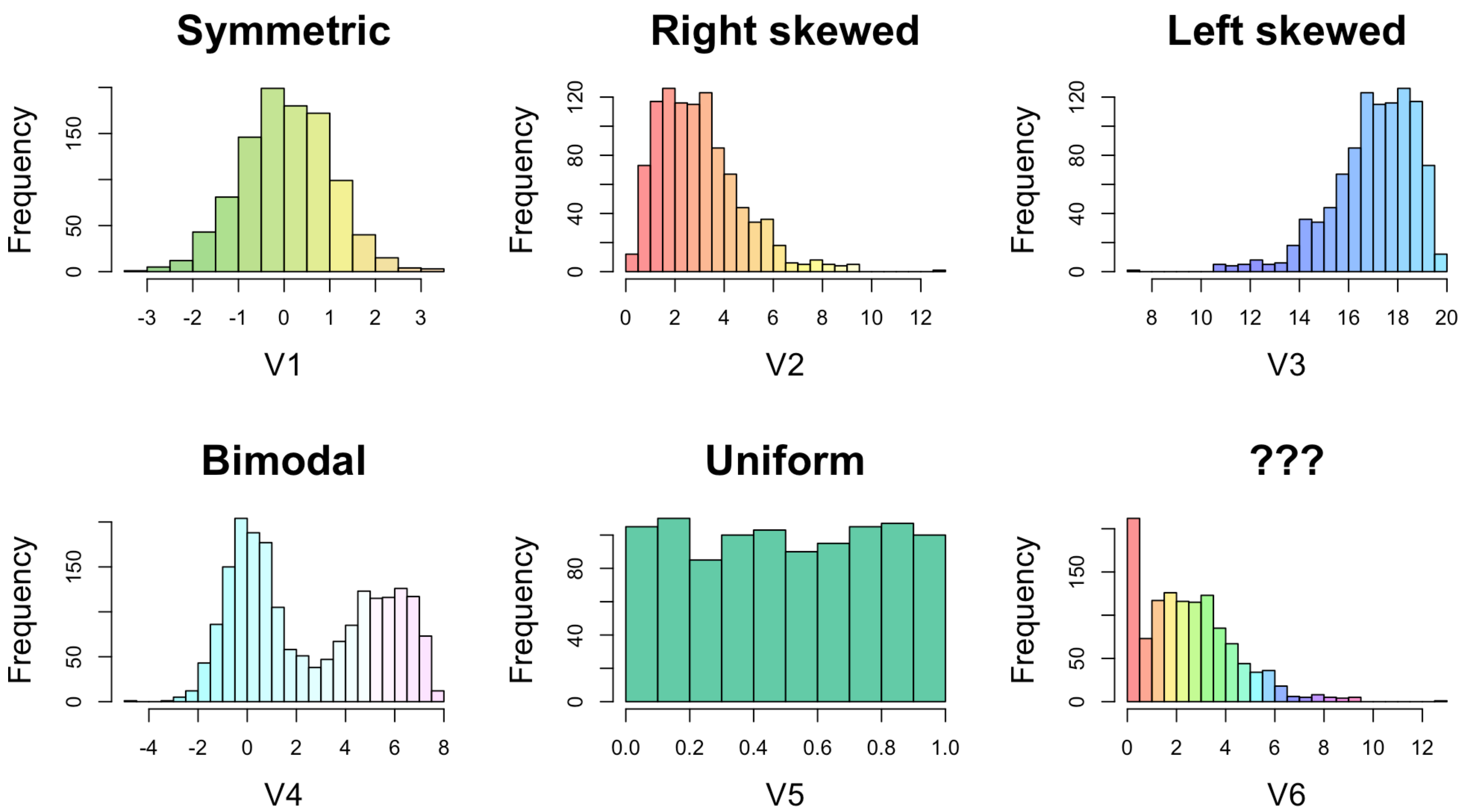


Describe a histogram:

1. **Center**: mean and median
2. **Spread**: standard deviation and interquartile range
3. **Shape**: symmetric, unimodal, bimodal, left/right-skewed, etc.

- **Note**: The histogram intervals are left open and right closed. For example, 60 inch is included in the first not the second interval.

Describing distributions



Quantitative variables - Spread

The **variance s^2** of a set of observations is the average of the squares of the deviations of the observations from their **mean**. In symbols, the variance of n observations $x_1, x_2, \cdots, x_n$ is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

The **standard deviation s** is the square root of the variance s^2:

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}}$$

Quantitative variables - Spread

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}}$$

- **Deviation** $x_i - \bar{x}$: the difference between x_i and their mean $\bar{x}$
  - Positive or negative.
- **Squared deviation** $(x_i - \bar{x})^2$
  - Always positive.
- **Sum of squared deviations** $\sum(x_i - \bar{x})^2$: overall squared deviations of the variable
- **Variance** s^2: *average* squared deviations of the variable
- **Standard deviation** s: *average* deviations of the variable

Quantitative variables - Spread

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}}$$

Three questions:

- Why do we square the deviations?
  - Can we use absolute deviations?
- Why do we emphsize the standard deviation rather than the variance?
- Why do we average by dividing by $n - 1$ rather than n?

Read textbook page 39.

Quantitative variables - Spread

Variance and standard deviation of *Height*

```
var(Survey$Height, na.rm = T) # Variance
```

```
## [1] 19.98091
```

```
sqrt(var(Survey$Height, na.rm = T)) # Squared root of variance
```

```
## [1] 4.470002
```

```
sd(Survey$Height, na.rm = T) # Standard deviation of Height
```

```
## [1] 4.470002
```

```
sd(Survey$Coffee, na.rm = T) # Standard deviation of Coffee
```

```
## [1] 4.711298
```

- Note: Mean of *Height* is 67.0 and mean of *Coffee* is 2.9. The spread of the *Coffee* variable (relative to its mean) is much larger than that of the *Height* variable.

Quantitative variables - Spread

Properties of the Standard Deviation (SD) s

- SD measures spread about the mean and should be used only when the mean is chosen as the measure of center.
- $s = 0$ only when there is *no spread*. This happens when all observations have the same value.
- Similar as mean, SD is NOT resistant to extreme values.

- *Standard Deviation (SD)* is the measure of spread when *mean* is the measure of center.
- *Interquartile range (IQR)* is the measure of spread when *median* is the measure of center.

Quantitative variables - Spread

To calculate the **Quartiles**:

1. Arrange the observations in increasing order and locate the median M in the ordered list of observations.
2. The **first quartile** $\boldsymbol{Q_1}$ is the value that has 25% of the observations fall *at or below* it.
3. The **third quartile** $\boldsymbol{Q_3}$ is the value that has 75% of the observations fall *at or below* it.

The $\boldsymbol{p}$**th percentile** is the value that has p percent of the observations fall *at or below* it.

- The first quartile Q_1 is the 25th percentile.
- The third quartile Q_3 is the 75th percentile.

Quantitative variables - Spread

Quartiles and percentiles of *Height*

```r
sort(Survey$Height) # sort() function automatically removes the NA values
```

```
##   [1] 58.5 59.0 59.0 60.0 60.0 60.0 60.5 61.0 61.0 61.0 61.0 61.0 61.0 61.5
##  [15] 61.5 61.5 62.0 62.0 62.0 62.0 63.0 63.0 63.0 63.0 63.0 63.0 63.0 63.0
##  [29] 63.0 63.0 63.5 64.0 64.0 64.0 64.0 64.0 64.0 64.0 64.0 64.0 64.0 64.5
##  [43] 65.0 65.0 65.0 65.0 65.0 65.0 65.0 65.0 65.0 65.0 65.5 66.0 66.0 66.0
##  [57] 66.0 66.5 66.5 66.5 66.5 67.0 67.0 67.0 67.0 67.0 67.0 68.0 68.0 68.0
##  [71] 68.0 68.0 68.0 69.0 69.0 69.0 69.5 69.5 70.0 70.0 70.0 70.0 70.0 70.5
##  [85] 71.0 71.0 71.0 71.0 71.0 71.0 71.0 71.0 71.0 71.5 71.5 71.5 72.0 72.0
##  [99] 72.0 72.0 72.0 72.0 72.0 72.0 72.0 72.0 72.0 72.0 73.0 73.0 73.0 74.0
## [113] 74.0 74.0 75.0 75.5 76.0 76.0 77.0
```

```r
quantile(Survey$Height, prob=c(0, 0.1, 1/4, 0.5, 0.75, 0.95, 1), na.rm=T)
```

```
##    0%   10%   25%   50%   75%   95%  100% 
## 58.50 61.00 63.25 66.50 71.00 74.00 77.00
```

- $Q_1 =?$, $Q_3 =?$ What are the other values?

Quantitative variables - Spread

The **five-number summary** of a set of observations consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is

$$\text{Minimum} \;\; Q_1 \;\; M \;\; Q_3 \;\; \text{Maximum}$$

```
min(Survey$Height, na.rm=T) # Minimum
```

```
## [1] 58.5
```

```
max(Survey$Height, na.rm=T) # Maximum
```

```
## [1] 77
```

Quantitative variables - Spread

The five-number summary

```
# The summary() function in R gives the five-number summary statistics,
# the mean and the number of NA's of a quantitative variable.
summary(Survey$Height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   58.50   63.25   66.50   67.00   71.00   77.00       3
```

To measure the spread of a set of observations when taking median as the center, we calculate

- **Interquartile Range** or **IQR** $= Q_3 - Q_1$

Note: $Maximum - Minimum$is defined as **Range**. IQR is more resistant to extreme values than range or standard deviation.

Quantitative variables - Spread

Range and IQR

```
71 - 63.25 # IQR
```

```
## [1] 7.75
```

```
77 - 58.5 # Range
```

```
## [1] 18.5
```

```
IQR(Survey$Height, na.rm=T) # IQR
```

```
## [1] 7.75
```

```
range(Survey$Height, na.rm=T) # Range
```

```
## [1] 58.5 77.0
```

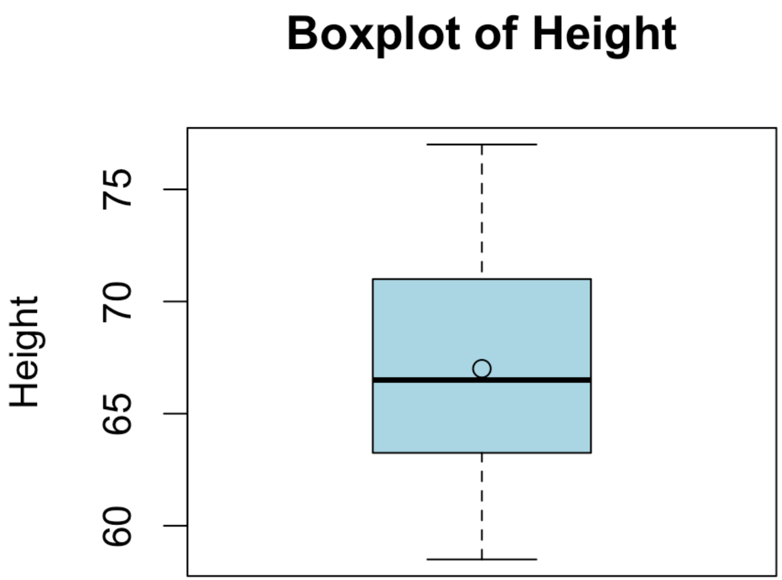
Quantitative variables

Summary statistics

Variable	Height	Coffee
Mean	67.0	2.9
SD	4.5	4.7
Median	66.5	1
Q_1, Q_3	63.25, 71	0, 4
IQR	7.75	4
Min, Max	58.5, 77	0, 30
Range	18.5	30

Quantitative variables - Boxplot

```r
boxplot(Survey$Height, col="lightblue", ylab="Height", main="Boxplot of Height")
# Add mean of Height (as a point) to the boxplot
points(mean(Survey$Height, na.rm=T))
```

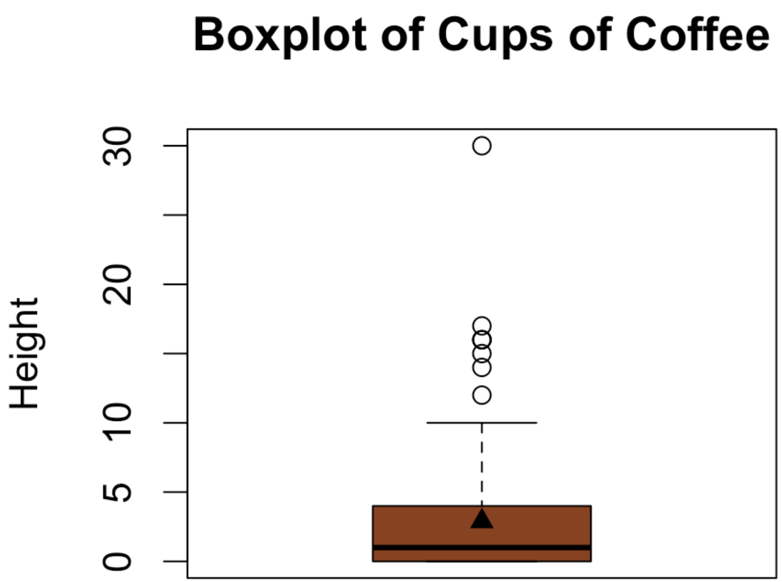


The box plot displays

- Minimum: the line below the box
- Q_1 : the bottom line of the box
- Median: the bold line in the box
- Q_3 : the top line of the box
- Maximum: the line above the box
- Usually the mean is drawn as a point

Quantitative variables - Box plot

```r
boxplot(Survey$Coffee, col="sienna4", ylab="Height",
        main="Boxplot of Cups of Coffee")
# Add mean of Height (as a point) to the boxplot
points(mean(Survey$Coffee, na.rm=T), pch=17) # pch: shape of the point
```



The difference between the boxplot of *Coffee* and that of *Height*

- Mean and median are not close to each other
- Minimum and Q_1 overlap
- A few points lie above the "maximum"
- In fact, the top line is no longer the maximum in this boxplot

Quantitative variables

The 1.5×IQR rule for suspected outliers

> An observation is called a **suspected outlier** if it falls more than $1.5 \times IQR$ above the third quartile or below the first quartile.

```r
summary(Survey$Height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   58.50   63.25   66.50   67.00   71.00   77.00       3
```

```r
Q1 <- 63.25; Q3 <- 71
Q1 - 1.5*(Q3-Q1)
```

```
## [1] 51.625
```

▸ Smaller than the minimum.

```r
Q3 + 1.5*(Q3-Q1)
```

```
## [1] 82.625
```

▸ Greater than the maximum

Quantitative variables

```r
summary(Survey$Coffee)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   2.889   4.000  30.000
```

```r
Q1 <- 0; Q3 <- 4
Q1 - 1.5*(Q3-Q1)
```

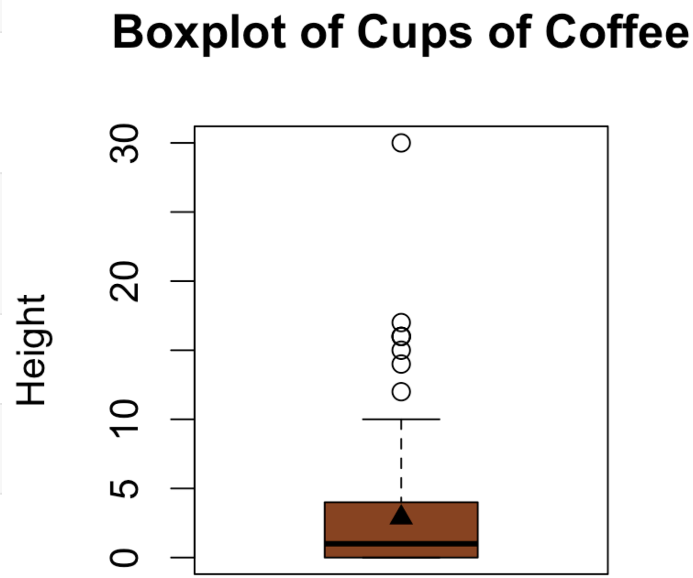
```
## [1] -6
```

- Smaller than the minimum.

```r
Q3 + 1.5*(Q3-Q1)
```

```
## [1] 10
```

- Smaller than the maximum



- The points that are greater than $Q_3 + 1.5 \times IQR = 10$ are suspected outliers.
- The top line in a boxplot with suspected outliers is the line at $Q_3 + 1.5 \times IQR$. Similarly, the bottom line in a boxplot with suspected outliers is the line at $Q_1 - 1.5 \times IQR$.

Linear transformations

A linear transformation changes the original variable X into the new variable X_{new} given by an equation of the form

$$X_{new} = a + bX$$

Adding the constant a shifts all values of X upward or downward by the same amount. In particular, such a shift changes the origin (zero point) of the variable. Multiplying by the positive constant b changes the size of the unit of measurement.

- **Question**: A variable X with values $1, 2, \cdots, 10, 15$ has
  mean $\bar{x} = 6.4$, standard deviation $s = 4.1$,
  median $M = 6$, quartiles $Q_1 = 3.5$ and $Q_3 = 8.5$
  and IQR $Q_3 - Q_1 = 5$.
  What are the mean, SD, median, quartiles and IQR for variable $Y = 3 + 2X$?

Effect of linear transformations

To see the effect of a linear transformation on measures of center and spread, apply these rules:

- Multiplying each observation by a positive number b multiplies both measures of center (mean and median) and measures of spread (standard deviation and interquartile range) by b.
- Adding the same number a (either positive or negative) to each observation adds a to measures of center and to quartiles and other percentiles but does not change measures of spread.

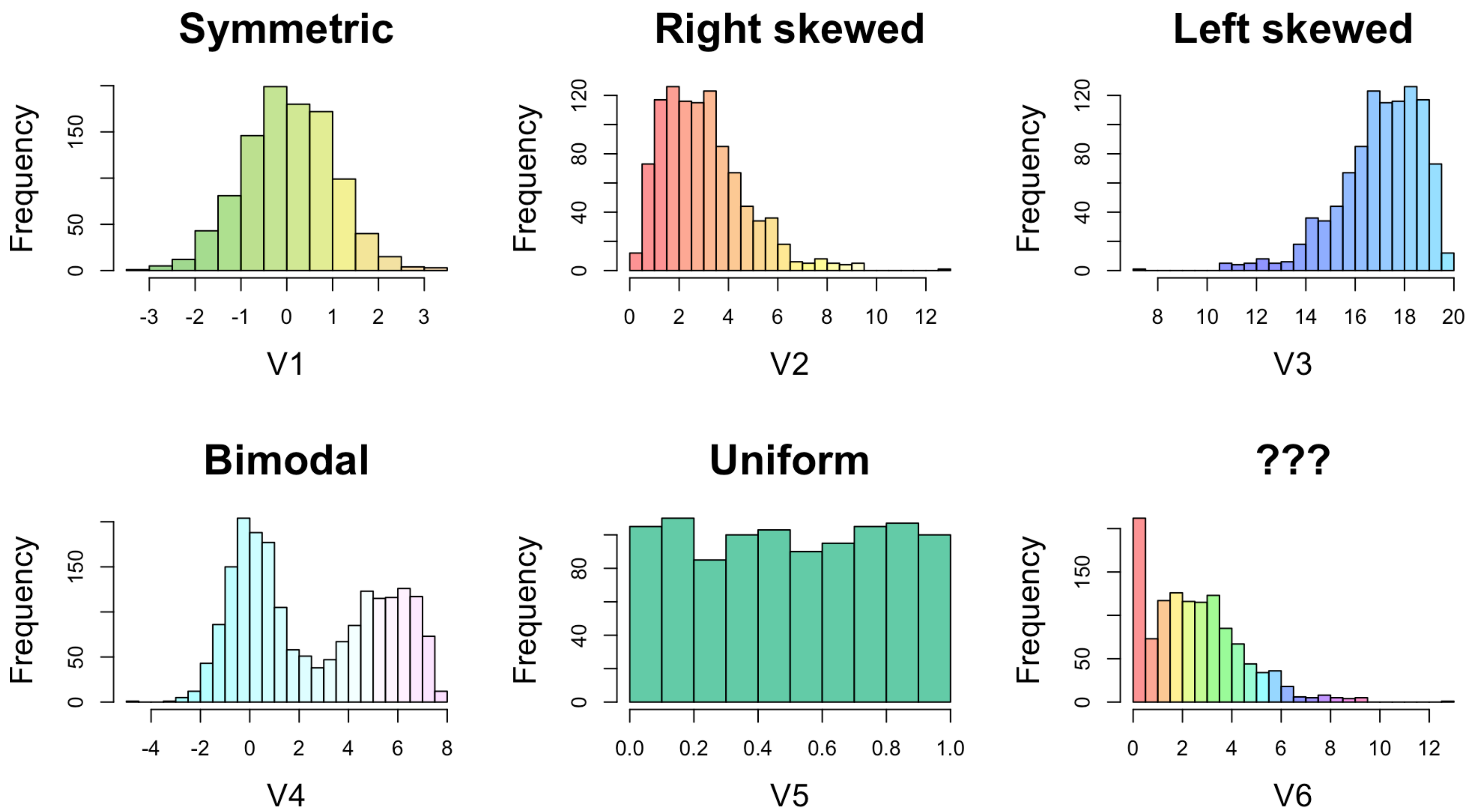
- Therefore, For the variable $Y = 3 + 2X$,
  $\bar{y} = 3 + 2 \times 6.4 = 15.8$, $s = 2 \times 4.1 = 8.2$,
  $M = 3 + 2 \times 6 = 15$, $Q_1 = 3 + 2 \times 3.5 = 10$, $Q_3 = 3 + 2 \times 8.5 = 20$
  and $IQR = 2 \times 5 = 10$.

Summary

	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()`

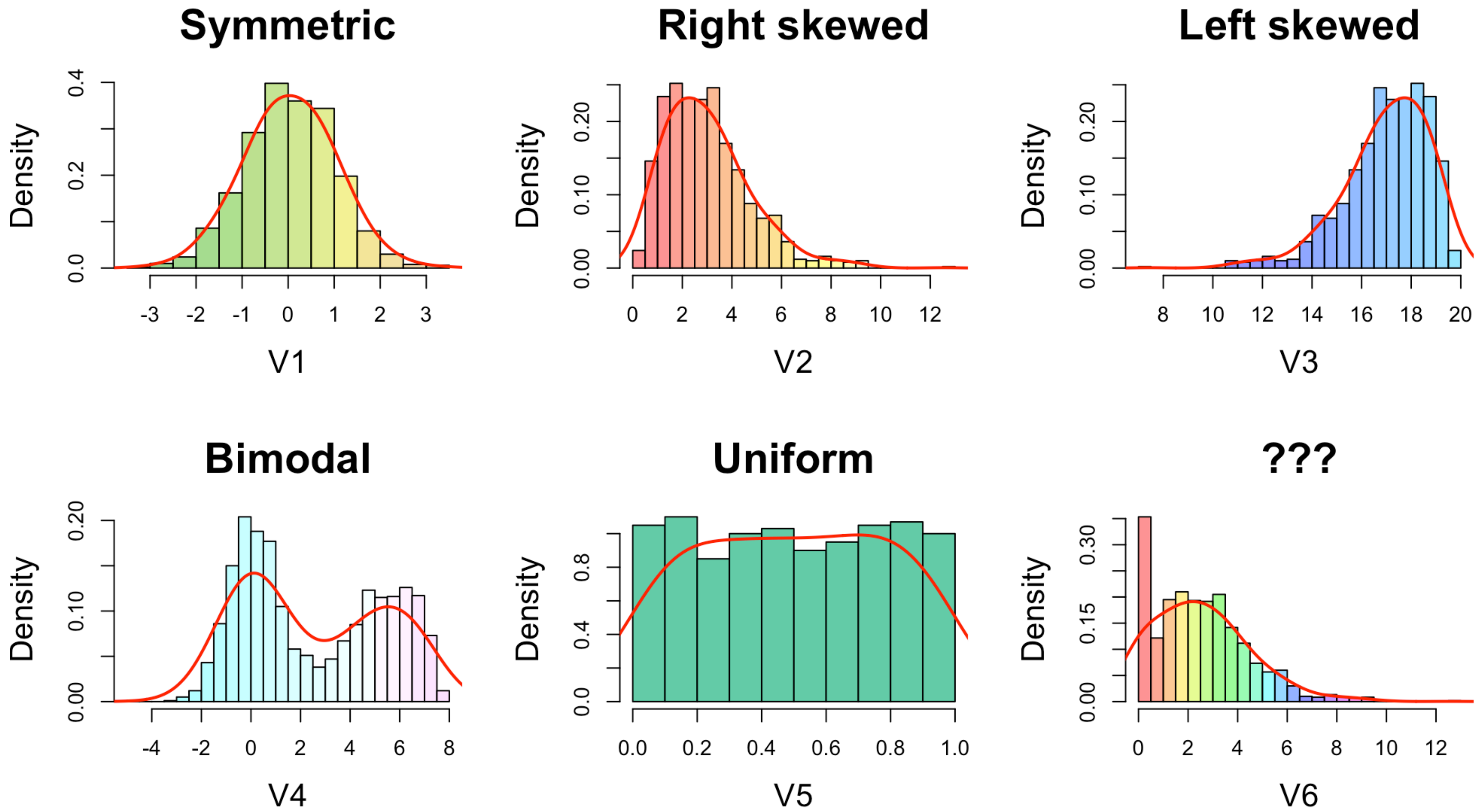
- The $\mathbf{1.5 \times \mathit{IQR}}$ rule for suspected outliers.
- Effect of linear transformations.

Histograms



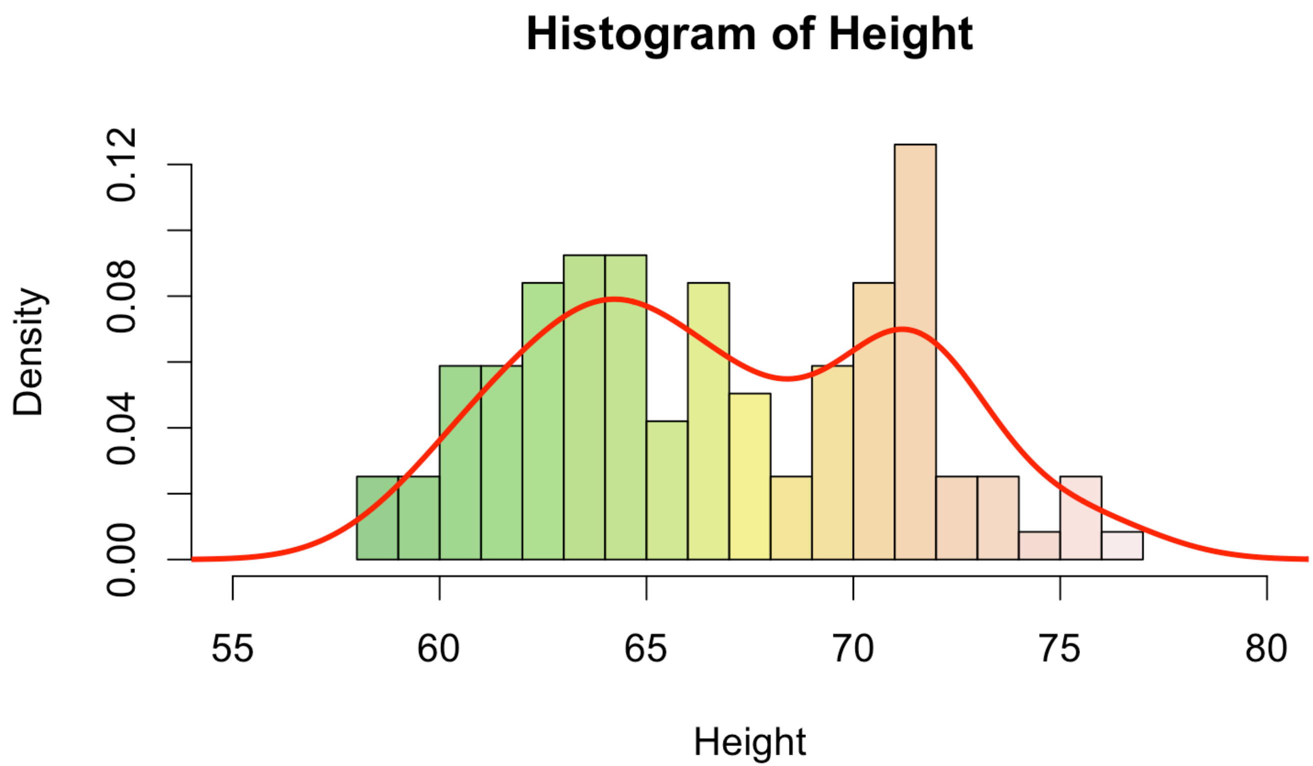
Symmetric
Frequency
0
50
150
-3
-2
-1
0
1
2
3
V1
Right skewed
Frequency
0
40
80
120
0
2
4
6
8
10
12
V2
Left skewed
Frequency
0
40
80
120
8
10
12
14
16
18
20
V3
Bimodal
Frequency
0
50
150
-4
-2
0
2
4
6
8
V4
Uniform
Frequency
0
40
80
0.0
0.2
0.4
0.6
0.8
1.0
V5
???
Frequency
0
50
150
0
2
4
6
8
10
12
V6

Histograms with density curves

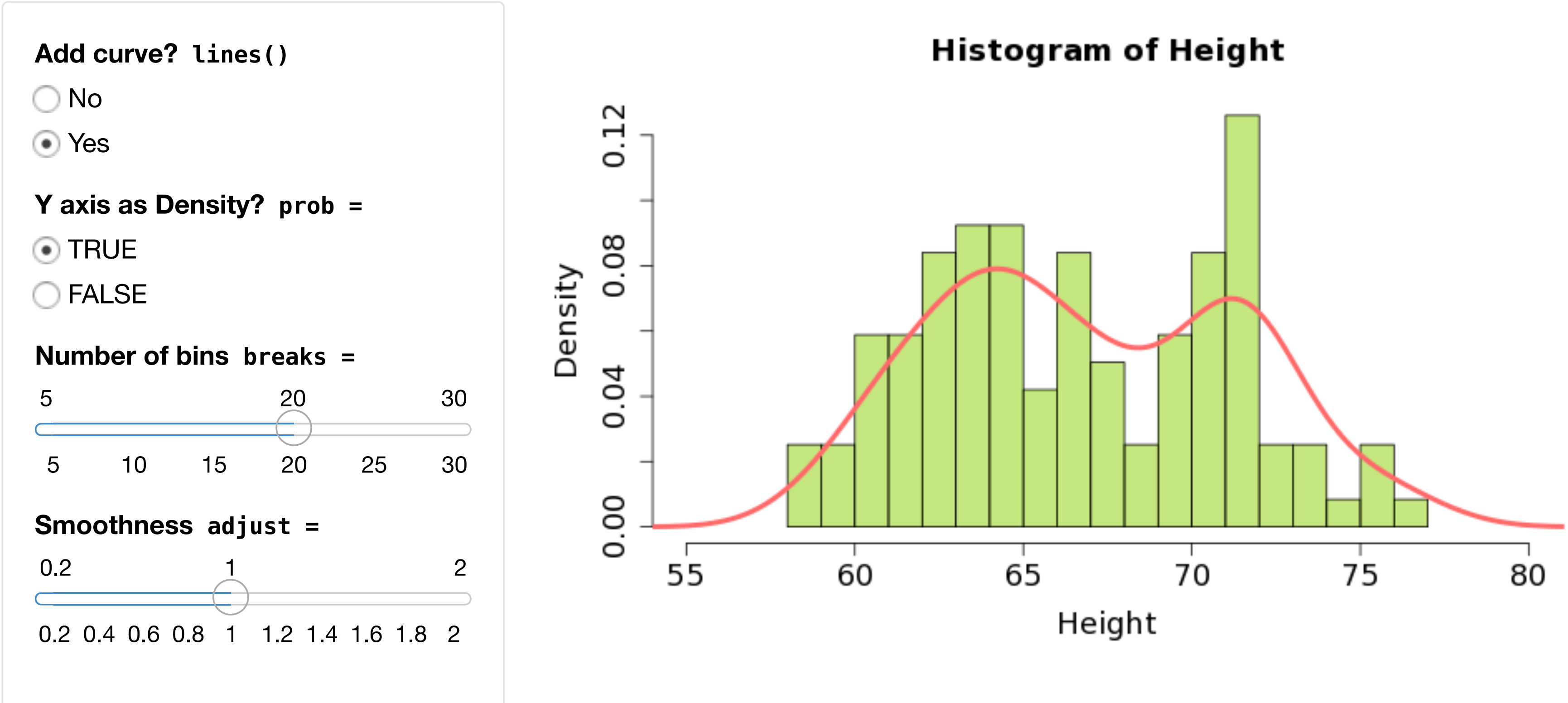


Histograms with density curves

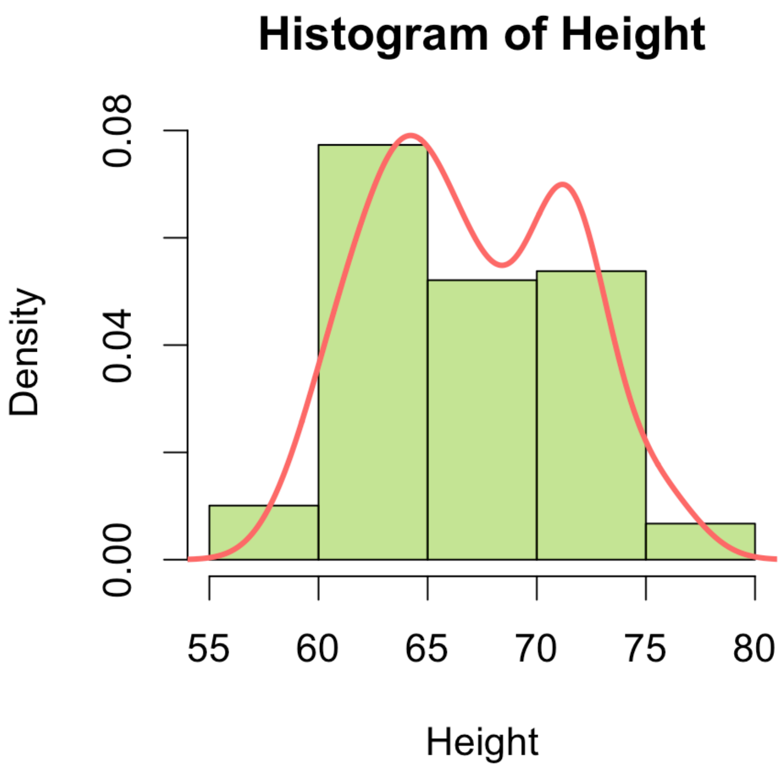
```
hist(Survey$Height, breaks = 20, xlab="Height", main = "Histogram of Height",
     col=terrain.colors(20,alpha=0.5), xlim=c(55,80), prob=TRUE)
lines(density(Survey$Height,na.rm=T, adjust = 1), lwd=3, col="red")
```



Histograms and density curves



Histograms and density curves



- The y-axis value of the histogram is NOT probabilities or proportions.
- Instead, the **area** of each bar (y-axis value $\times$ width) is the proportion of observations in that interval.
- For example, the first bar has y-axis value around 0.01 and width 5. The area is about $0.05 \times 5 = 0.05$. About 5% of the students have height values between 55 and 60 inches.

Histograms and density curves

Interval	Density (Height)	Area (Height $\times$ Width)
[55, 60]	0.01	$0.01 \times 5 = 0.05$
(60, 65]	0.08	$0.08 \times 5 = 0.40$
(65, 70]	0.05	$0.05 \times 5 = 0.25$
(70, 75]	0.05	$0.05 \times 5 = 0.25$
(75, 80]	0.01	$0.01 \times 5 = 0.05$
Total	0.20	1

- The area of the histogram is 1.
- The area under the density curve of the histogram is 1.

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.