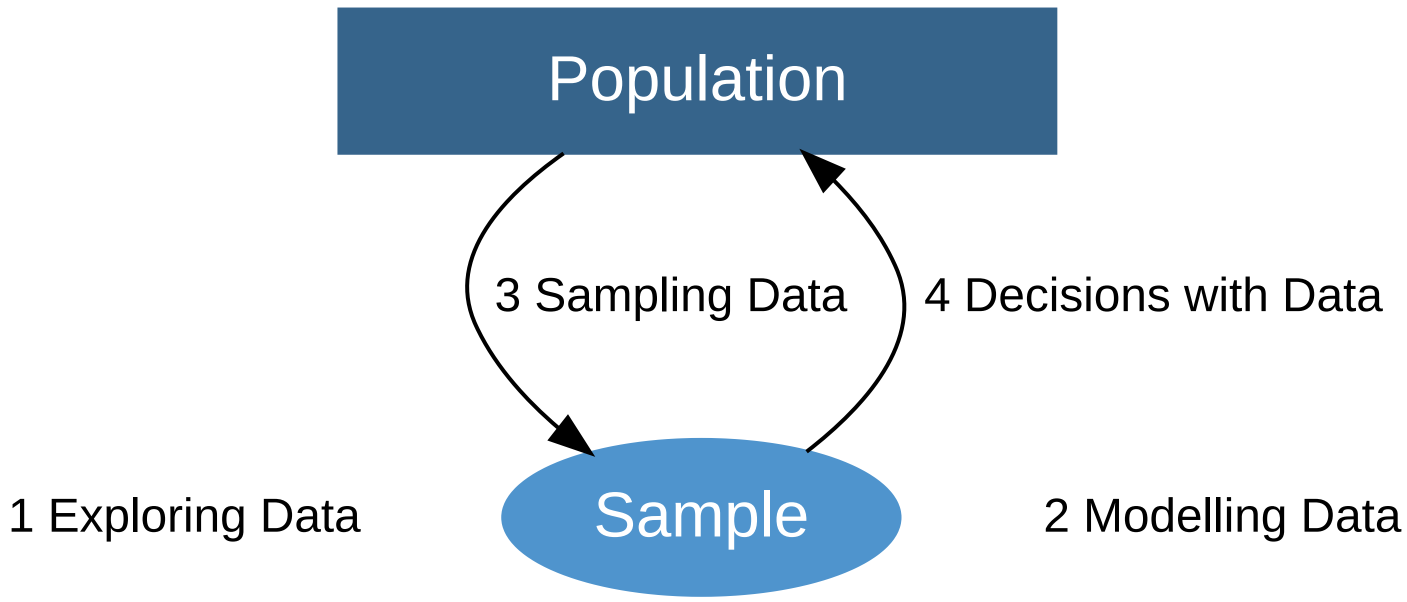
Quantitative Data

Exploring Data | Data & Graphical Summaries

Unit Overview



Population

3 Sampling Data

4 Decisions with Data

1 Exploring Data

Sample

2 Modelling Data

Module1 Exploring Data

Design of Experiments

Where did the data come from & can we make reliable conclusions?

Data & Graphical Summaries

What type of data do we have & how can we visualise it?

Numerical Summaries

What are the main features of the data?

Quantitative Data

Data Story | What causes Australian Road Fatalities?

Histogram

Common Mistakes with Histograms

Box Plot

Scatter Plot

Summary

Data Story

What causes Australian Road Fatalities?

Australian Road Fatality Data

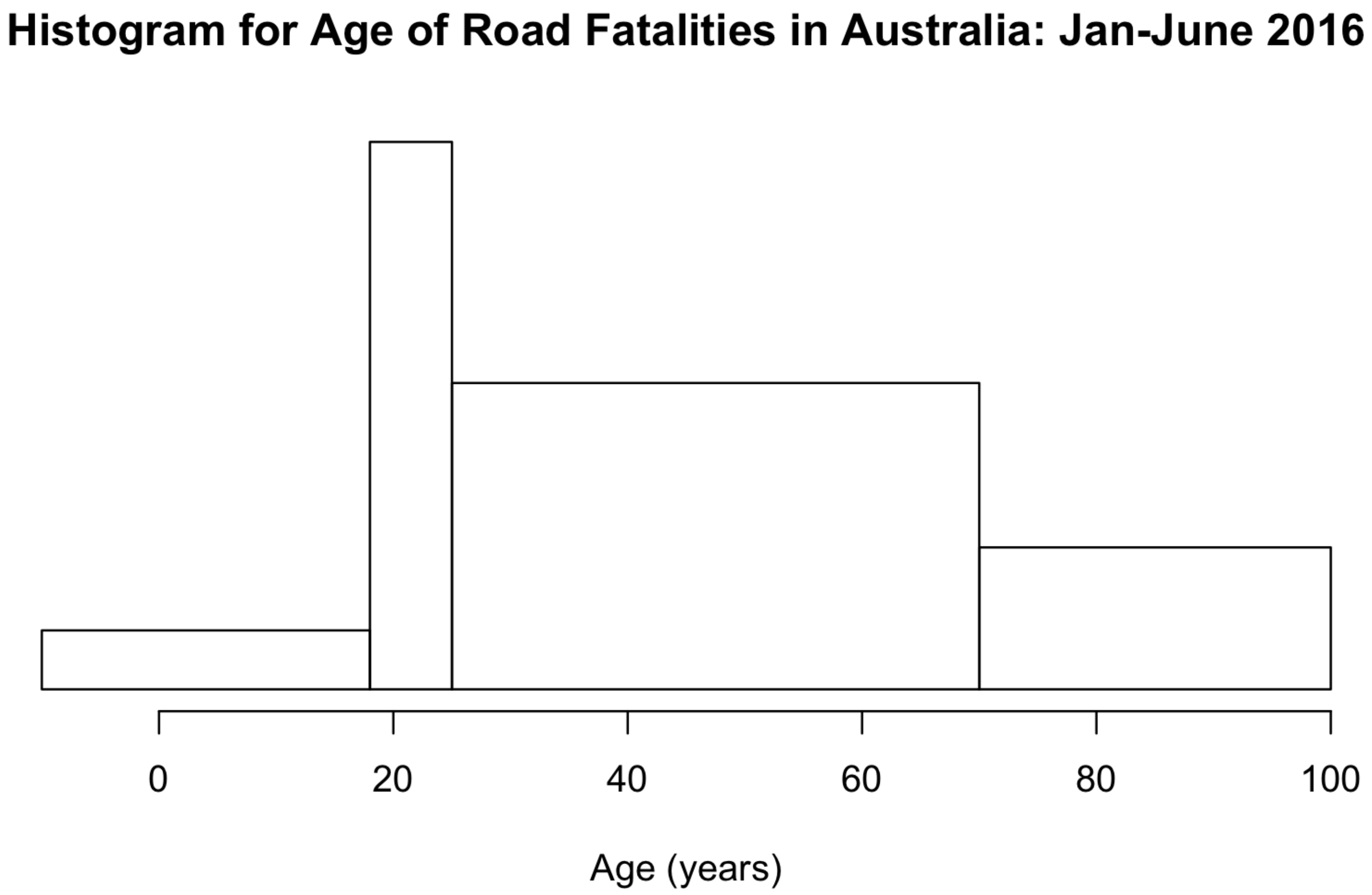
Statistical Thinking

Before we investigate the following questions from the data, what would you expect and why?

- What were the 3 highest ages at which a road fatality occurred?
- Were there any unusual ages at which a road fatality occurred? Is there any difference between the ages of male and female fatalities?
- Were there any unusual speeds at which fatalities occurred?

Histogram

Histogram of Age of Road Fatalities



Statistical Thinking

In approximately what percentage of road fatalities were the people aged under 18?

- Approx 5-10%. We compare the relative area of the 1st block to the rest of the plot.

Why does the 1st block start below 0?

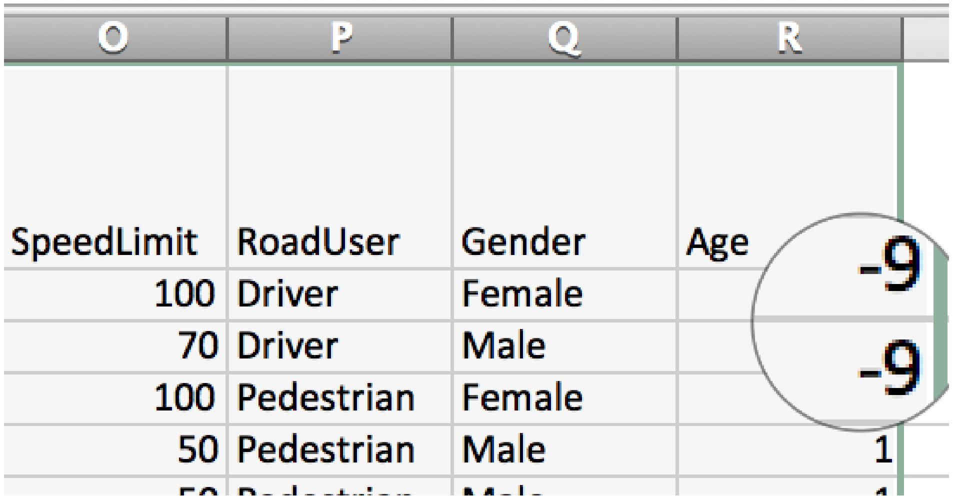
- According to the Data Dictionary, missing values are coded as ‘-9’.

Why is the histogram tallest above [18,25)?

- There are lots of road fatalities in this age group. We call this **crowding**. However there are overall more fatalities in [25,70), as this age interval is so much longer.

Primitive cleaning of the data

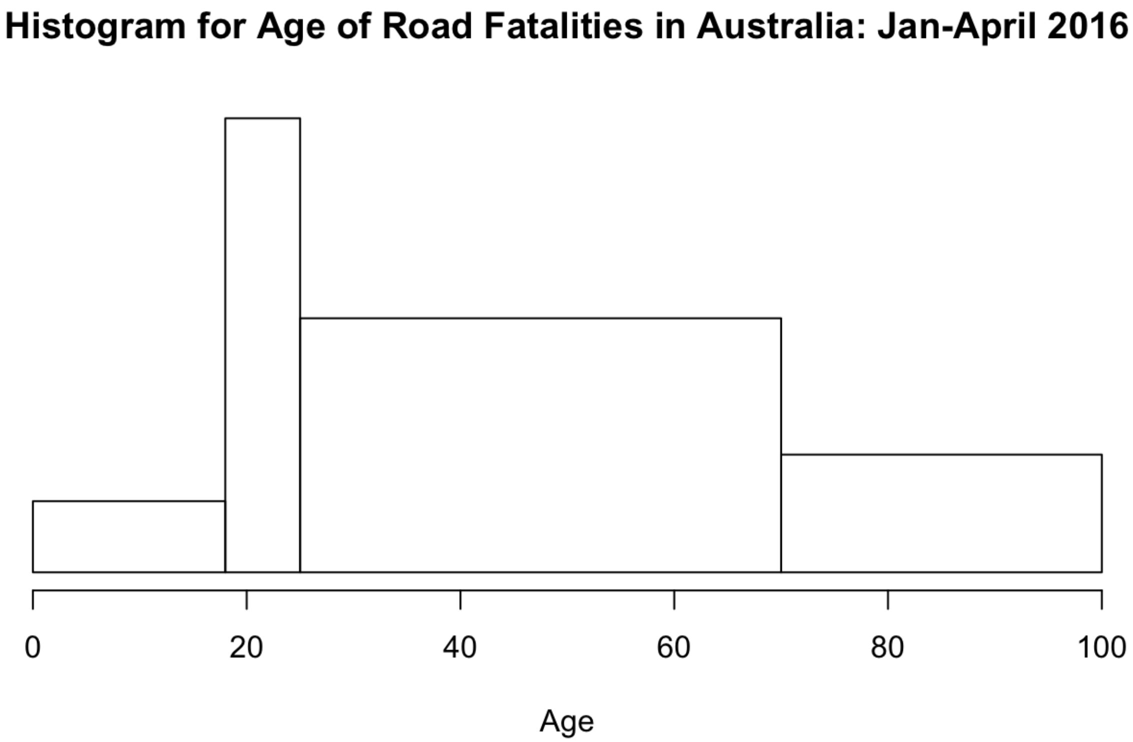
- For now, a primitive way to clean up the Excel data file is to replace the 2 cells with ‘-9’ by a blank.



O	P	Q	R
SpeedLimit	RoadUser	Gender	Age
100	Driver	Female	
70	Driver	Male	
100	Pedestrian	Female	
50	Pedestrian	Male	1

- For a more sophisticated approach, see Data Wrangling.

```r
# C is the cleaned version of data
data = read.csv("data/2016FatalitiesC.csv",header=T)
```



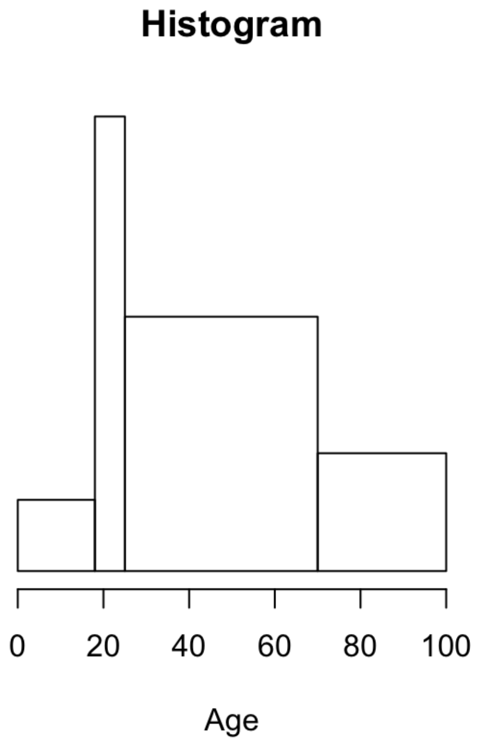
Overview of histogram

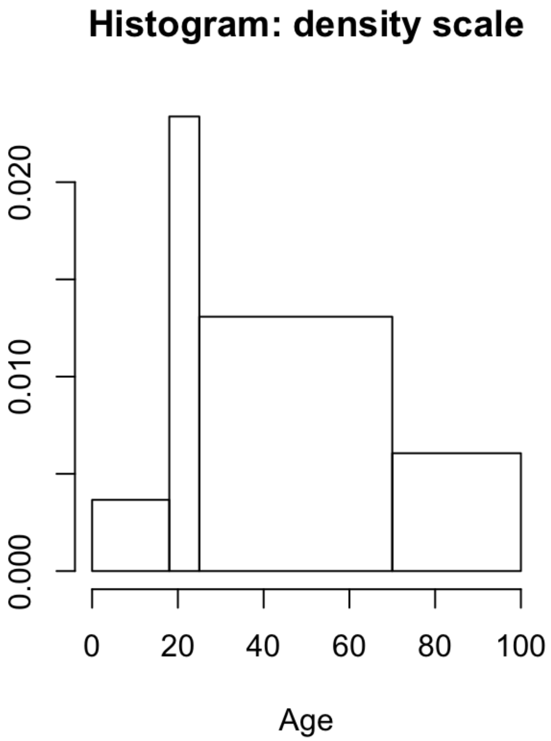
- We use a histogram for quantitative data.

- A histogram highlights the percentage of data in one class interval compared to another.
    - It consists of a set of blocks which represent the percentages by area.
    - The area of the whole histogram is 100%.
    - The horizontal scale is divided into **class intervals**.
    - The **area of each block** represents the percentage of subjects in that particular class interval.
    - The **height** of each block represents **crowding**.

Choices with histogram

1. There is no need for a vertical scale to assess the relative areas.





2. We will mostly use the **density scale**.

The density scale has advantages for later modelling.

Density scale

$$\text{Height of each block} = \frac{\text{\% in the block}}{\text{length of the class interval}}$$

$$\text{ie Height of each block} = \text{\% per horizontal unit}$$

3. For continuous data, we need an **endpoint convention** for data points that fall on the border of two class intervals.

- If an interval contains the left endpoint but excludes the right endpoint, then an 18 year old would be counted in [18,21) not [0,18).
- We call this left-closed and right-open.

How to produce a histogram by hand

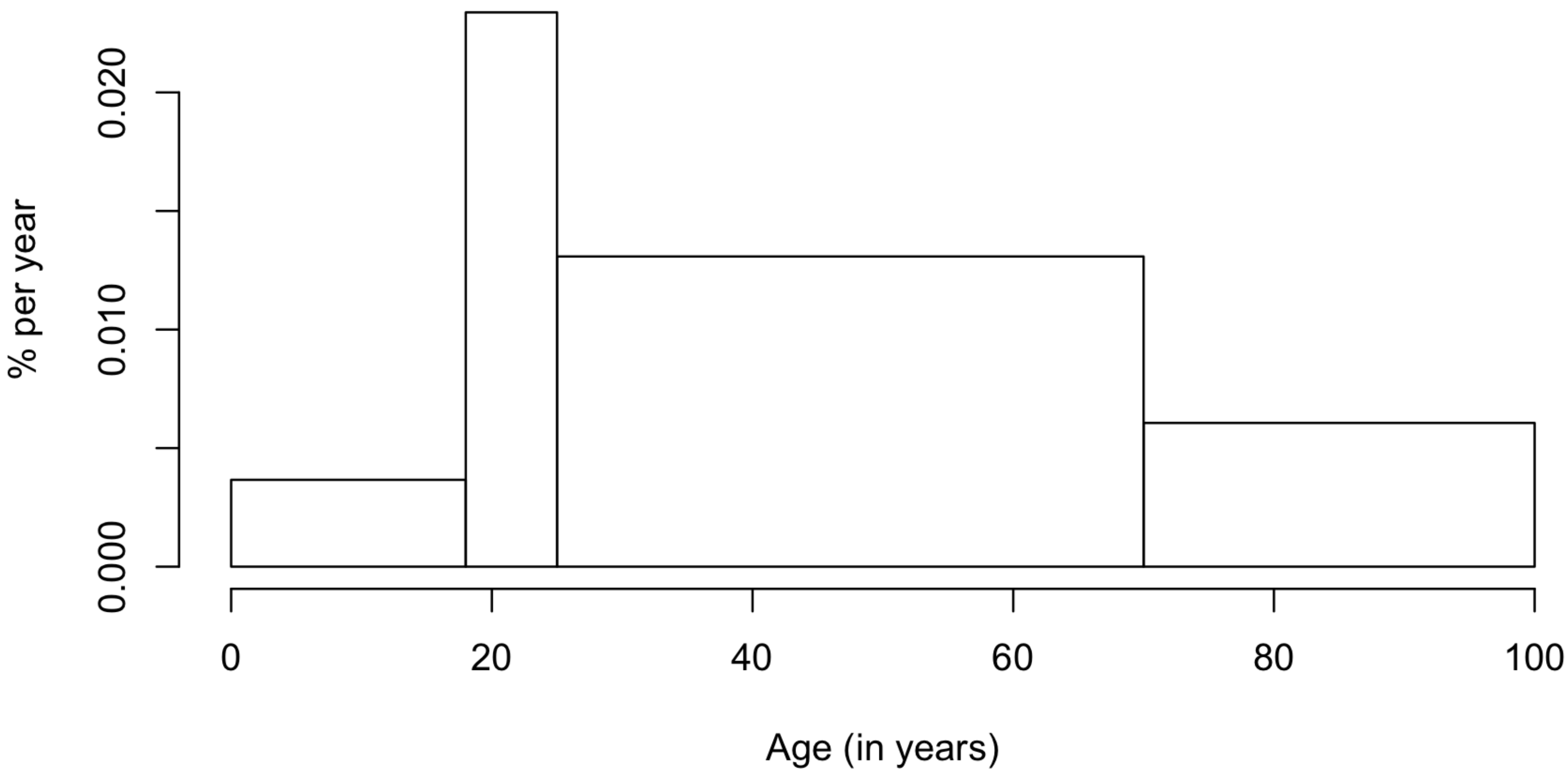
Step1: Construct the **distribution table**.

Class intervals	Number of subjects in the interval	%	Height of block
[0,18)	29	6.6	0.004
[18,25)	72	16.4	0.023
[25,70)	259	58.9	0.013
[70,100)	80	18.2	0.006
	440	100	

where Height of block = % per year.

Step2: Draw the horizontal axis and blocks.

Histogram for Age of Road Fatality in Australia: Jan-June 2016



The speedy way in R

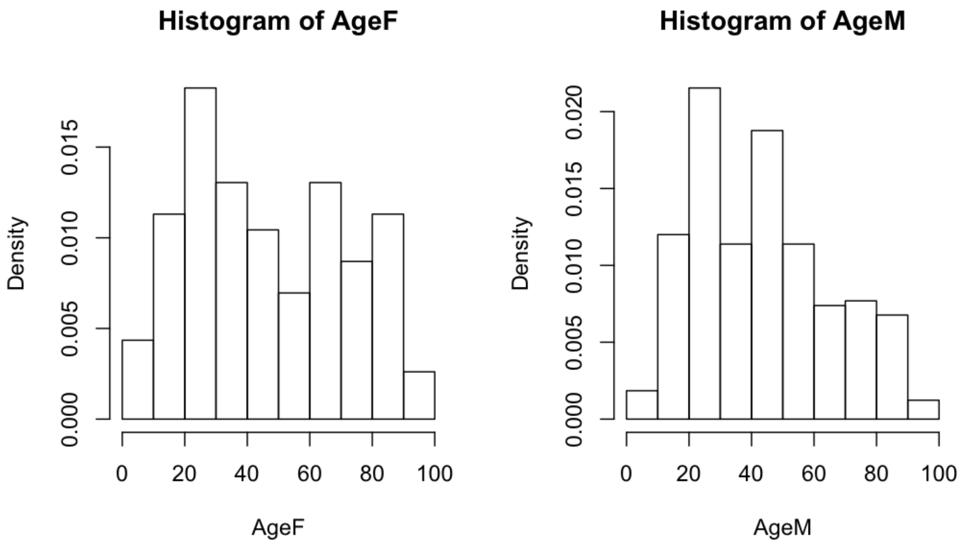
```
#Read in data
data = read.csv("data/2016FatalitiesC.csv",header=T)

# Choose a variable
Age = data$Age

# Choose the class intervals
breaks=c(0,18,25,70,100)

# Produce a distribution table
table(cut(Age,breaks,right=F))

# Produce a histogram
hist(Age,br=breaks,freq=F,right=F,
     xlab="Age (in years)", ylab="% per year",
     main="Histogram for Age of Road Fatality in Australia: Jan-June 2016")
```

Note:

- `freq=F` produces the histogram on the density scale.
- `right=F` makes the intervals right-open.

Controlling for a variable

```
AgeF = data$Age[data$Gender=="Female"]   # This selects just the female ages.
AgeM = data$Age[data$Gender=="Male"]
par(mfrow=c(1,2))  # This puts the graphic output in 1 row with 2 columns
hist(AgeF,freq = F)
hist(AgeM,freq = F)
```

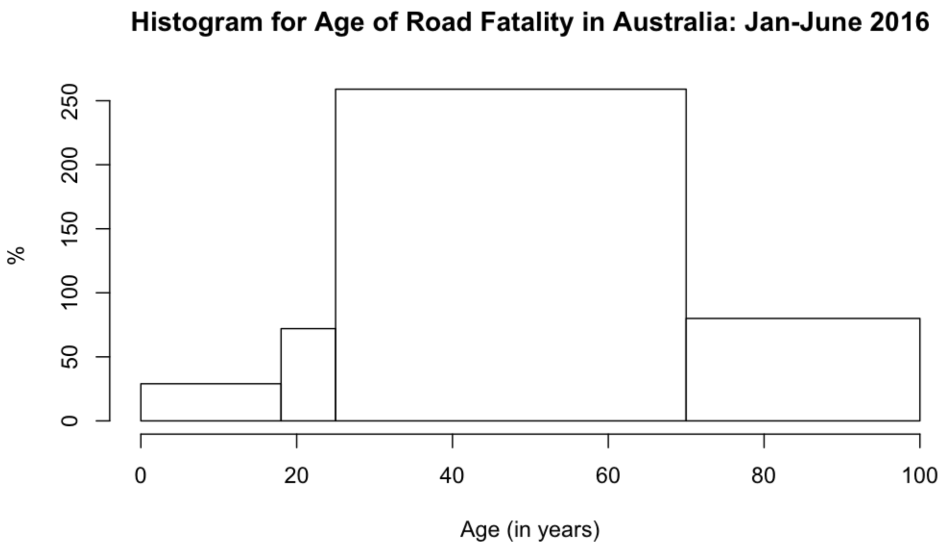


Do you notice any differences between men and women?

Common mistakes with histograms

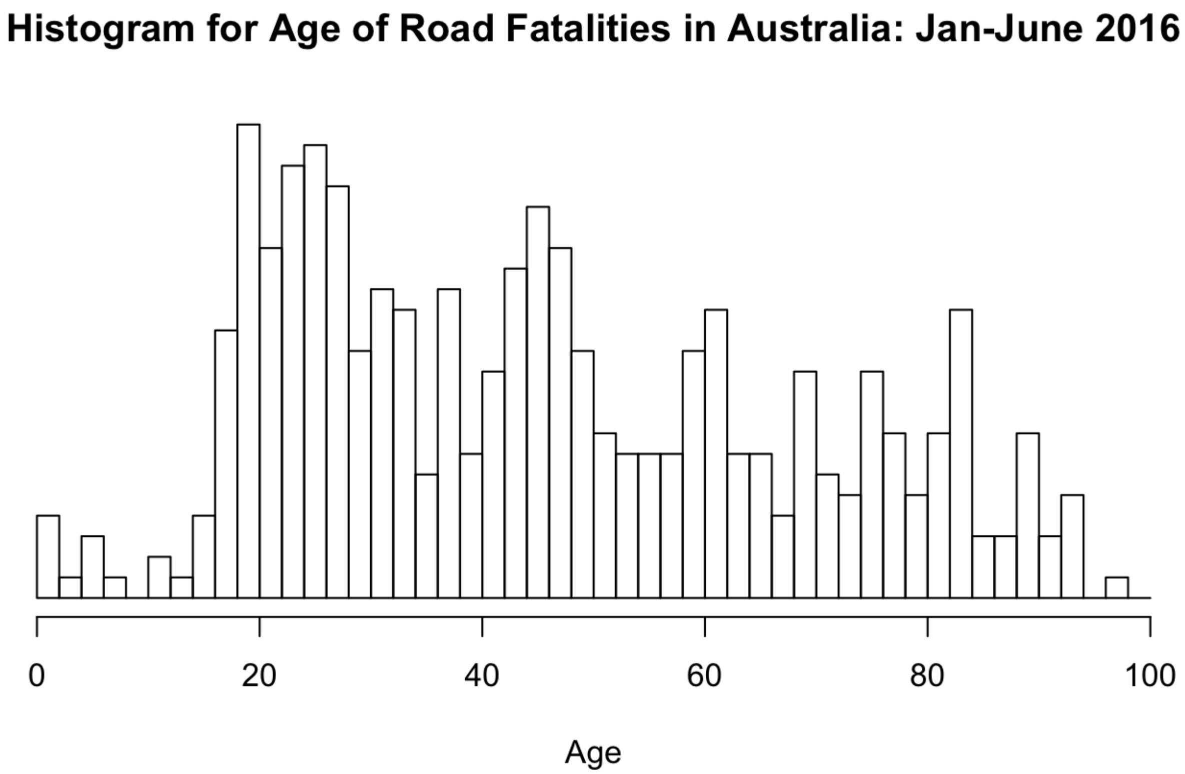
Mistake 1: Make the block heights equal to the percentages

- Here we wrongly use the % from p16 as the heights.
- Unless the class intervals are the same size, this will makes larger class intervals look like a larger overall %.



Mistake 2: Use too many class intervals

This can overcondense the data. As a rule of thumb, use between 10-15 maximum.



Box Plot

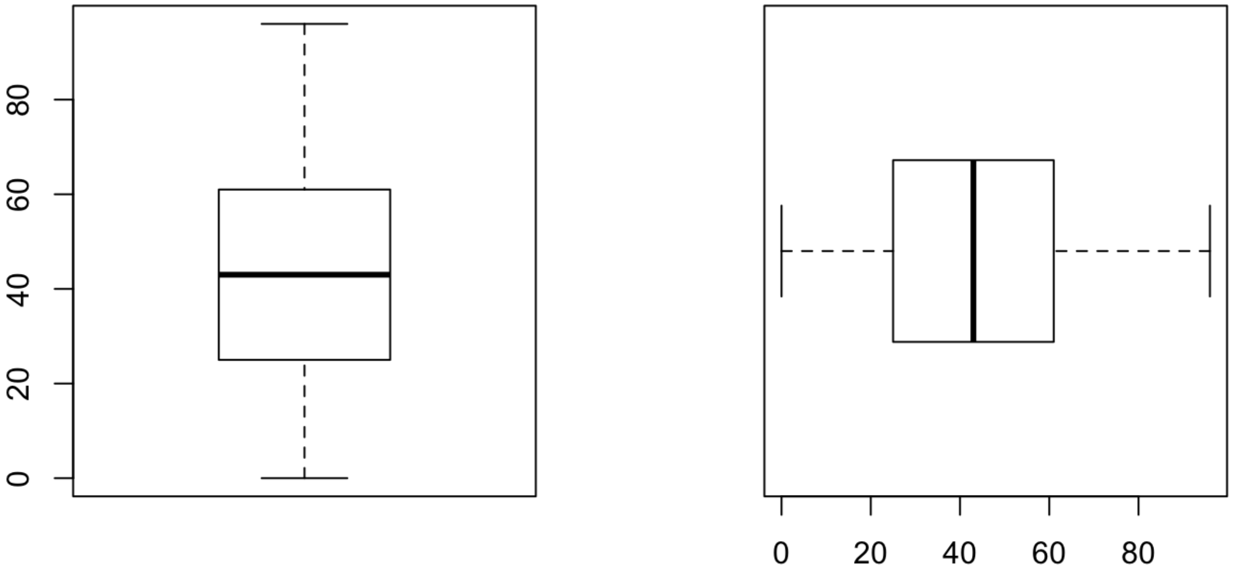
Simple box plot

- The boxplot is useful for comparing multiple data sets.
- The boxplot plots the median (‘middle’ data point), the middle 50% of the data in a box, and determines any outliers.
- We will consider how to draw the box plot when we learn about the interquartile range (IQR) in Spread.

```
Age = data$Age
summary(Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   25.00   43.00   44.85   61.00   96.00       2
```

```
par(mfrow=c(1,2))
boxplot(Age)
boxplot(Age, horizontal=T)
```



Statistical Thinking

What does the simple boxplot reveal about the age of fatalities?

- The box plot is fairly symmetric with no outliers.
- There does not seem to be any extreme ages for fatalities.

Comparative box plots

A comparative boxplot splits up a quantitative variable by a qualitative variable.

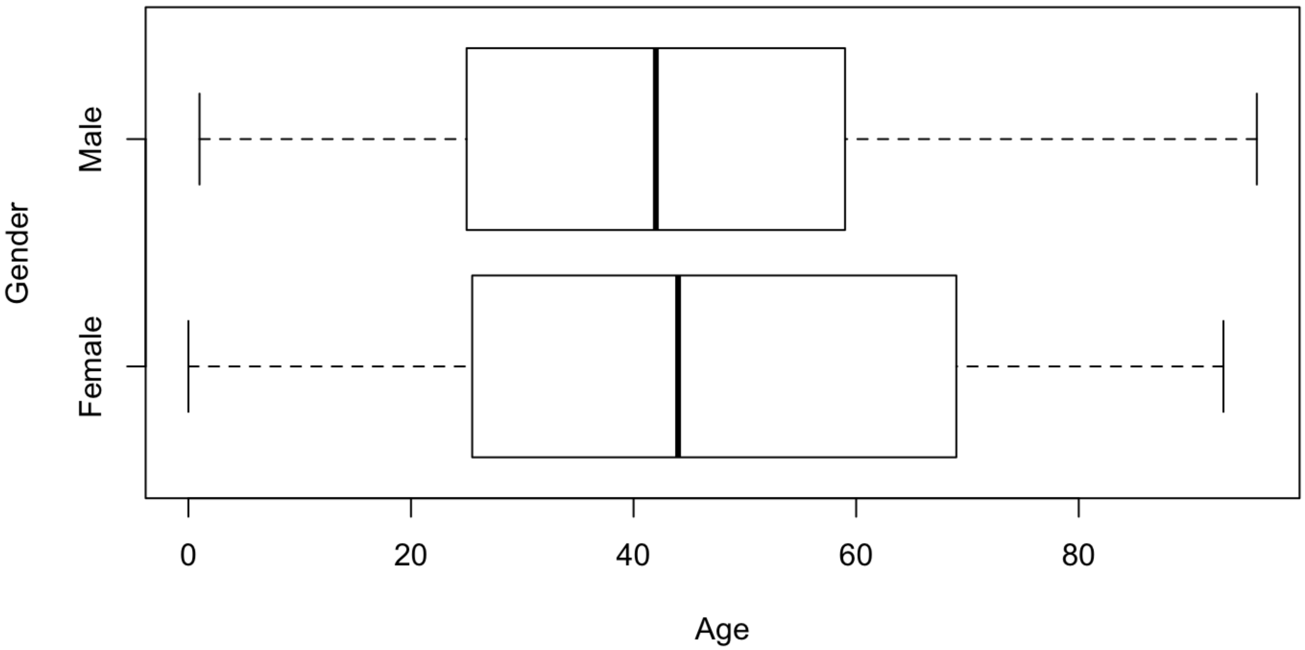
```
# Divide Age into 2 genders
Gender = data$Gender
summary(Age[Gender=="Female"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   25.50   44.00   47.19   69.00   93.00       1
```

```
summary(Age[Gender=="Male"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   25.00   42.00   44.02   59.00   96.00       1
```

```
boxplot(Age~Gender, horizontal = T)
```



Statistical Thinking

What do the summaries and boxplots reveal about the age of fatalities for male and females?

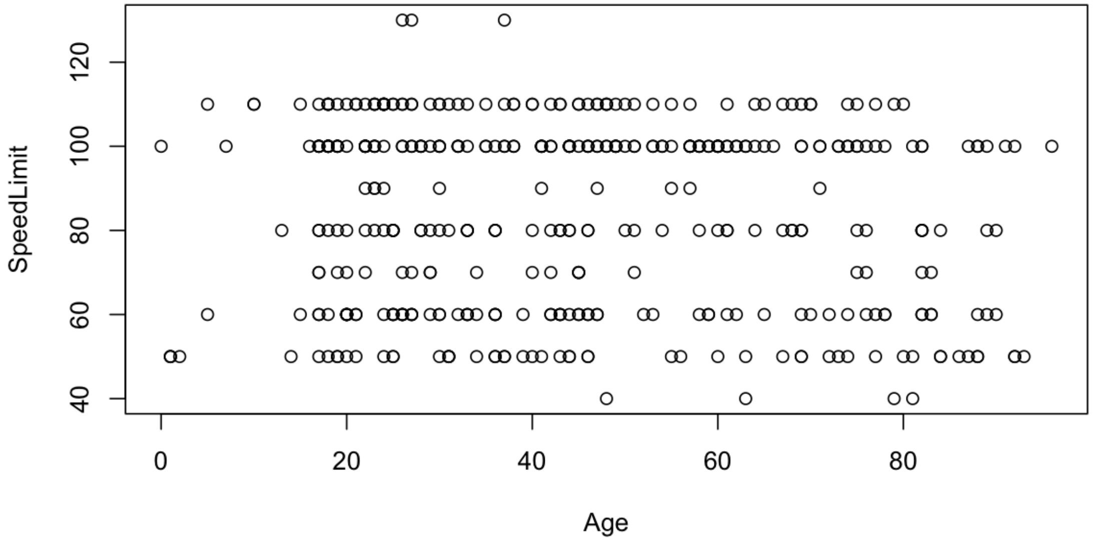
- The median (‘middle’) age is fairly similar.
- In the middle 50%, the age of women is higher than men.

Scatter Plot

Scatter plot

The scatter plot examines the relationship between 2 quantitative variables.

```
SpeedLimit = data$SpeedLimit
plot(Age,SpeedLimit)
```



- As the scatter plot looks random, there does not appear to be a relationship between age and speed limit in fatalities.
- However, this is treating speed limit as a quantitative variable, when strictly it is qualitative.

Summary

The histogram is a graphical summary for quantitative data which shows the percentage of subjects per class interval. The boxplot shows the middle 50% of the data and it's spread and outliers. The scatterplot shows the relationship between 2 variables.

1 Qual	2 Qual	1 Quant	2 Quant	1 Quant + 1 Qual
Simple Bar Plot	Double Bar Plot	Histogram Box Plot	Scatter Plot	Comparative Box Plot

Key Words

data, subject, population, sample, initial data analysis, descriptive statistics, graphical summaries, numerical summaries, variable, dimension, multivariate, bivariate, univariate, quantitaive, qualitative, discrete, continuous, histogram, crowding, density scale, end point convention

R - Histogram by Factor Variable (lesson 2)