```
## Warning:  package 'latex2exp' was built under R version 4.3.3
```

STAT115: Introduction to Biostatistics

University of Otago

Ōtākou Whakaihu Waka

Lecture 4: Data Summaries

- Long-term goal: fit, and interpret statistical models to real data
- We need some more background information first:
  - What is a statistical model?
  - Introduction to probability and random variables
- Today: look at data summaries
  - You may have seen these summaries before
  - Calculate these in R
  - Introduce 'mathematical notation'
  - Look at how these summaries point toward statistical modelling
    - Data summaries are the starting point, not the finish line
    - Motivate a better understanding of probability

Data: Auckland Heart Attack Patients

- Data introduced in Lecture 2
- Will focus here on variable `Vol`, end-diastolic volume in ml
- Option 1: provide (list) the data
  - Not very enlightening with $n = 32$ observations
  - It might not be possible
    - Privacy concerns
    - Other considerations (ethical or otherwise) which prevent sharing of data
- Option 2: visualize the data
  - Good idea, but hard to summarize quantitatively
- Option 3: numerically summarize the data
- Option 4: approaches we are yet to learn

Into R

- Step 1: call data into R
  - Import using menu (File > Import Dataset)
  - Use commands

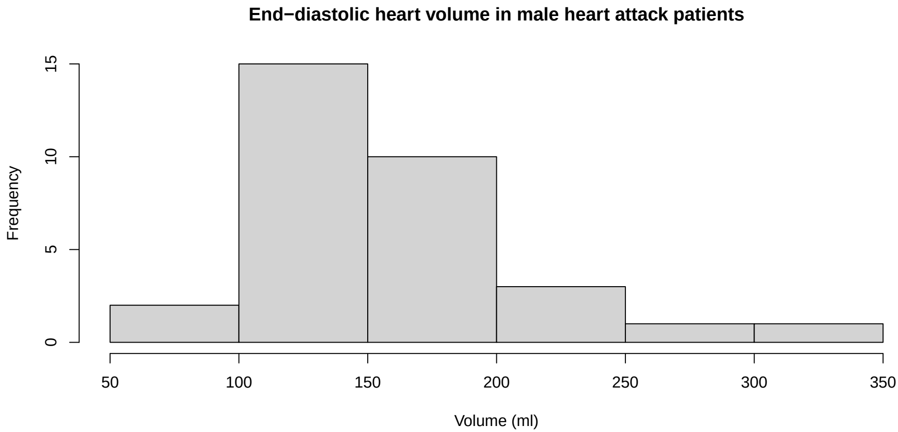
```
nzheart = read.csv('nzheart.csv')
```

  - `nzheart.csv` needs to be in the current working directory in Rstudio
- Step 2: visualize the data

```
hist(nzheart$Vol, xlab = "Volume (ml)",
     main = "End-diastolic heart volume in male heart attack patients")
```

- Remember: `nzheart` has multiple variables (columns)
  - `nzheart$Vol` obtains end-diastolic volume variable)

Histogram


End−diastolic heart volume in male heart attack patients
Frequency
0
5
10
15
50
100
150
200
250
300
350
Volume (ml)

Sample Mean

- The mean is a common summary
  - Often called the average
  - Inherits same units as data
- The sample mean is the sum of the observed values divided by the number of observations

$$\bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n}$$

- Let's unpack:
  - What does $\bar{y}$ represent?[1]
  - What does y_1 represent?
  - What does y_2 represent?
  - What does n represent?

[1] $\bar{y}$ is said: y-bar

Sample Mean

continued

- The sample mean is given as

$$\bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n}$$

- Commonly we will see this written as

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

- Let's unpack:
  - What does y_i represent?
  - What does $\sum_{i=1}^{n}$ represent?
- The two equations say exactly the same thing

Tutorial: what the $\sum$?

- The sample mean is

$$\bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n} = \frac{\sum_{i=1}^{n} y_i}{n}$$

- $\sum$ is the Greek letter Sigma (capital)
  - It represents a sum
  - $\sum_{i=1}^{n} y_i$ says that we:
    - Set $i = 1$ and find y_i: gives y_1
    - Set $i = 2$ and *add* y_i: gives $y_1 + y_2$
    - Set $i = 3$ and *add* y_i: gives $y_1 + y_2 + y_3$
    - Keep going...

Finding the mean

- It is worth knowing how to find a mean 'the old fashioned way'
  - What is the mean of 10, 6, 13, 7?
  - It means you can (in principle) calculate a mean anywhere, anytime
    - In your head (if not exactly, then approximately)
    - On a calculator / phone

Finding the mean

- The majority of the time we use the computer (R or other software)

```
y = c(10, 6, 13, 7) # c() is used to create a vector (or collection) of values
y
## [1] 10  6 13  7
```

- Use the R function `mean()` to find the mean

```
mean(y)
## [1] 9
```

- For the heart attack patient data

```
mean(nzheart$Vol)
## [1] 159.8
```

R: excursion

- You may have noticed that sometimes I have created an R object

```
y = c(10, 6, 13, 7) # c() is used to create a vector (or concatenation) of values
```

- This has created the object y
  - This object is then available to 'use', e.g. when finding the mean

```
mean(y)

## [1] 9
```

- In the code above, the mean value is not assigned to an object
  - It can be – it is then available to 'use' later on
  - For example, might want to compare with meanc value for healthy adult males

```
ybar = mean(y)
ybar

## [1] 9
```

Other Summaries

- The (sample) mean tells us a lot
  - Among our sample of $n = 32$ patients, the mean volume was 159.8 ml.
  - A patient with volume of 200 ml is above average.
- There is a lot the mean does not tell us
  - Is it surprising if we saw a patient with volume 200 ml?
- Another summary that tells us how variable (or dispersed) the data are would be useful.
  - High variability: commonly see a volume less than 70 ml or more than 300 ml
  - Low variability: unlikely to see a volume less than 70 ml or more than 300 ml

Sample Variance and Standard Deviation

- We will focus on two measure of variation (dispersion)
  - Variance
  - Standard deviation
- These are different expressions of the same thing
  - The variance is $(\text{standard deviation})^2$
  - The standard deviation is $\sqrt{\text{variance}}$

Sample Variance

- Sample variance: average squared distance between observations and the mean

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

  - We divide by $n-1$ (and not n)
    - There is some mathematical nuance
    - For our purposes: it gives a more reliable answer
  - It is a difficult calculation to do by hand
    - It is worth doing for a small problem to ensure you understand the formula
    - What is the variance of 10, 6, 13, 7?[2]
- We can find it easily in R

```
var(nzheart$Vol)

## [1] 2453
```

[2] The answer is 10

Sample Variance

- Sample variance: average squared distance between observations and the mean

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

  - If an observation y_i is far from $\bar{y}$
    - $(y_i - \bar{y})^2$ will be large
  - If the observations $y_1, \ldots, y_n$ are spread out
    - Many of the values $(y_i - \bar{y})^2$ will be large
    - s^2 will be large

  - If an observation y_i is close to $\bar{y}$
    - $(y_i - \bar{y})^2$ will be small
  - If the observations $y_1, \ldots, y_n$ are close together
    - Most of the values $(y_i - \bar{y})^2$ will be small
    - s^2 will be small

Sample Standard Deviation

- The sample standard deviation

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

- It represents a kind of average deviation of observations from the mean
  - Useful when considering how far the data are distributed from the mean
  - Easier to interpret than the variance
  - Standard deviation measured in same units as data; variance in squared units
- We can find it easily in R

```
sd(nzheart$Vol)

## [1] 49.53
```

Standard Deviation

Rules of thumb

- To better help us understand what the standard deviation represents
  - Approximately 70% of the data will be within one standard deviation of the mean
  - Approximately 95% of the data will be within two standard deviations of the mean
- These are only rules of thumb.
  - e.g. they do not hold if the data are skewed or multimodal

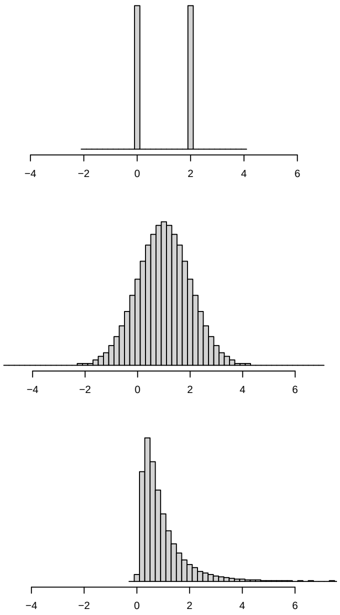
Data Summaries: Big Picture

- On one hand: lost a lot of information
  - $n = 32$ into two numbers
- On the other hand: created order out of chaos
  - It is hard for us to get an understanding of $n = 32$ values [3]
  - Summarized the data to gain an understanding about important features of the data
    - Later we might ask questions like: does the volume change with age? or disease severity?
  - The idea of finding a "simple" description (or model) of complex data will be a theme
- Look into the limitations of data summaries

[3] It is even worse if we have $n = 32,000$ values!

Limitations of Data Summaries

- Data summaries are useful, but...
  - Lose a lot of information: $n = 32$ into two numbers
  - Be careful not to over-interpret
- Three histograms: data with the same sample mean ($\bar{y} = 1$) and variance ($s^2 = 1$)



Limitations of Data Summaries

continued

- Data summaries are useful, but...
  - Samples do not give perfect information about the population
  - If we took a different sample, get a different sample mean (and variance)
- Consider population of all New Zealand males suffering a heart attack
- The mean end-diastolic volume of the population is unlikely to be exactly 159.8 ml
  - The value of 159.8 ml can be thought of as an educated guess (or estimate)
  - Can we quantify how precise (or uncertain) that estimate is?
- We cannot get this information from data summaries alone
  - What we will be working toward
  - Use probability to describe the variation in the data
  - Statistical models

Summary

- Calculate basic data summaries in R
- Understand how to calculate data summaries by hand (if we need to)
- Introduce mathematical notation
- Looked at limitations of data summaries