


**Describing data**

Samples and populations are often made of lots of individual (observational) units and their associated information (observations, variables).

We need to be able to describe samples by summary statistics (mean, median, variance, etc) so that these summaries can serve as an estimate of the same summaries for their statistical populations.




---

---

---

---

---

---

---

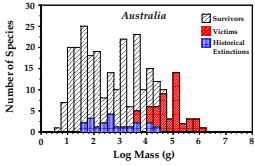
---

1

**Scientific question: Did humans drive mammal extinctions in Australia?**

↓

**Statistical question: Are "victims" bigger than "survivors" and historical extinctions?**



**Survivors (extant species, i.e., alive today).**

**Victims (late Pleistocene, i.e., past 50 000 years, 50 ka).**

Historical extinctions (older than 50 ka) are based on samples (fossils).

**We want to make inferences about all past and present mammals in Australia (i.e., statistical population are all mammal species, past or present, in Australia).**

**Frequency distribution of mammal mass categorized into survivors, "victims" and older (historical) extinctions**

Study by Lyons et al. (2004; Evolutionary Ecology Research 6:339-358)

ka = kiloannus (1000); ~ 50 ka = "behavioural modernity" in humans.

---

---

---

---

---


---

---

---

2

**How measures of location (mean versus median) and spread (standard deviation versus interquartile range) compare**

<p><b>Disarming fish</b> (protection against predation)</p>  <p><small>Threespine stickleback (Gasterosteus aculeatus)</small></p>	<p><b>Plate Genotypes</b> Ectodysplasin (Eda) locus (3<sup>rd</sup> generation)</p> <p><b>MM (marine)</b></p> <p><b>Mm (hybrid)</b></p> <p><b>mm (freshwater)</b></p>
---	---

---

---

---

---

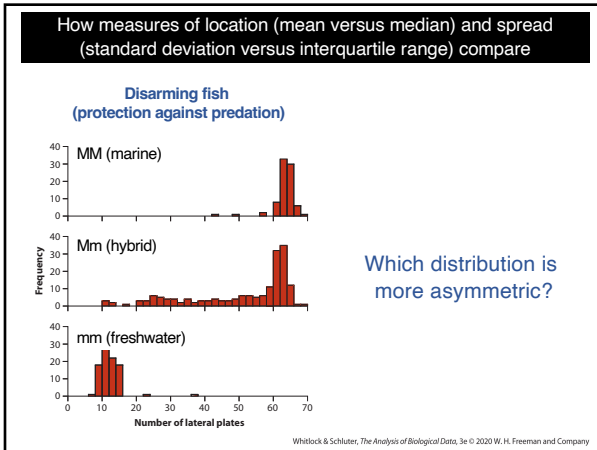
---

---

---

---

3




---

---

---

---

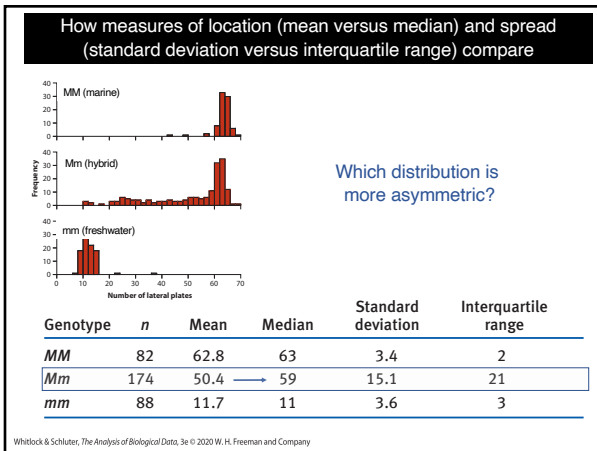
---

---

---

---

4




---

---

---

---

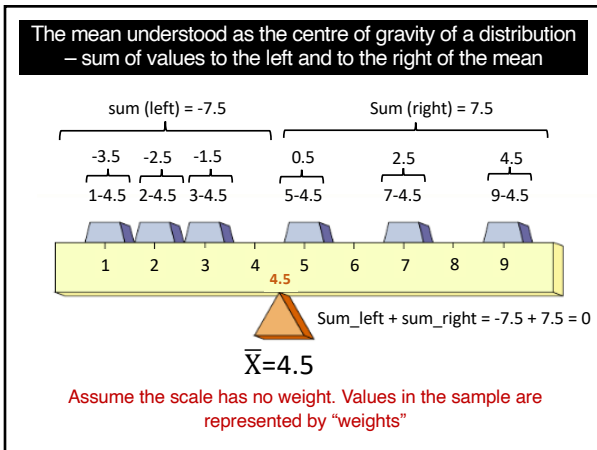
---

---

---

---

5




---

---

---

---

---

---

---

---

6

Remember from lecture 5 – the sum of the deviations from the mean is always zero – the mean is then always the “centre of gravity” of a distribution

Quantities needed to calculate the standard deviation and variance of snake undulation rate ( $\bar{Y} = 1.375 Hz$ ).

Observations ( $Y_i$ )	Deviations ( $Y_i - \bar{Y}$ )	Squared deviations ( $(Y_i - \bar{Y})^2$ )
0.9	-0.475	0.225625
1.2	-0.175	0.030625
1.2	-0.175	0.030625
1.3	-0.075	0.005625
1.4	0.025	0.000625
1.4	0.025	0.000625
1.6	0.225	0.050625
2.0	0.625	0.390625
Sum	0.000	0.735

Whitlock & Schluter, The Analysis of Biological Data, 3e © 2020 W. H. Freeman and Company

---

---

---

---

---

---

---

---

---

---

---

---

7

Mean ( $\bar{Y}$ ) versus Median ( $Q_2$ )

Mm (hybrid) - most asymmetric distribution

Sum (left) = -1104.759      Sum (right) = 1104.759

60 individuals      114 individuals

mean

$\bar{Y} = 50.4$

Mean is the “centre of gravity”

Whitlock & Schluter, The Analysis of Biological Data, 3e © 2020 W. H. Freeman and Company

---

---

---

---

---

---

---

---

---

---

---

---

8

Mean ( $\bar{Y}$ ) versus Median (referred as to  $Q_2$ )

Mm (hybrid) - most asymmetric distribution

86.5 individuals smaller than the  $Q_2$ .

86.5 individuals bigger than the  $Q_2$ .

Median is the middle of the distribution

median = 59.0

Mean is the “centre of gravity”

60 individuals      114 individuals

mean = 50.4

Whitlock & Schluter, The Analysis of Biological Data, 3e © 2020 W. H. Freeman and Company

---

---

---

---

---

---

---

---

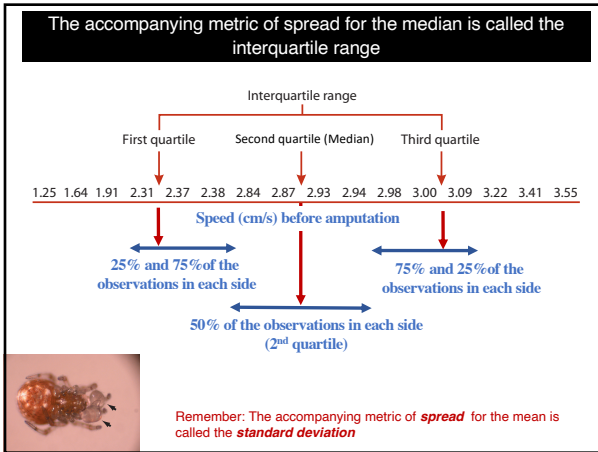
---

---

---

---

9



10

---

---

---

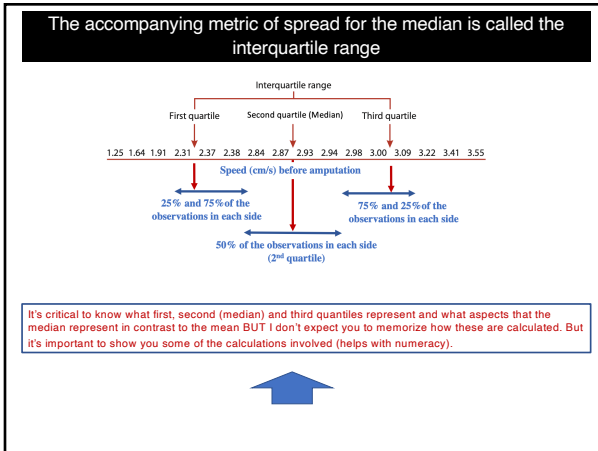
---

---

---

---

---



11

---

---

---

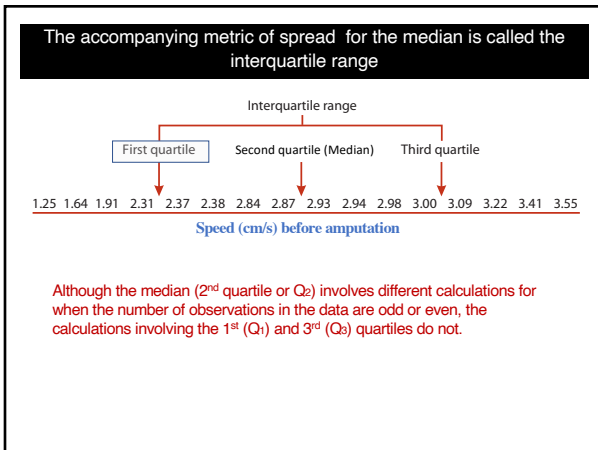
---

---

---

---

---



12

---

---

---

---

---

---

---

---

The accompanying metric of spread for the median is called the interquartile range

Interquartile range

First quartile      Second quartile (Median)      Third quartile

1.25 1.64 1.91 2.31 2.37 2.38 2.84 2.87 2.93 2.94 2.98 3.00 3.09 3.22 3.41 3.55

Speed (cm/s) before amputation

Positioning  $Q_1$   $j = 0.25n = (0.25)(16) = 4$

where  $n$  is the number of observations in the data.

Because here  $j$  is an integer (i.e., whole number, not a fraction), then the 1<sup>st</sup> quartile is the average of  $Y_{(j)}$  and  $Y_{(j+1)} = Y_{(4)}$  and  $Y_{(4+1)} = (2.31 + 2.37) / 2 = 2.340$  cm/s

**First quartile ( $Q_1$ ) = 2.340 cm/s**

13

---

---

---

---

---

---

---

---

The accompanying metric of spread for the median is called the interquartile range

Interquartile range

First quartile      Second quartile (Median)      Third quartile

1.25 1.64 1.91 2.31 2.37 2.38 2.84 2.87 2.93 2.94 2.98 3.00 3.09 3.22 3.41 3.55

Speed (cm/s) before amputation

**THIS IS NOT EXACTLY THE DEFAULT RULE IN R: BUT VALUES ARE VERY SIMILAR (THERE ARE A FEW DIFFERENT RULES FOR CALCULATING QUANTILES)**

**First quartile ( $Q_1$ ) = 2.340 cm/s**

14

---

---

---

---

---

---

---

---

The accompanying metric of spread for the median is called the interquartile range

Interquartile range

First quartile      Median      Third quartile

1.25 1.64 1.91 2.31 2.37 2.38 2.84 2.87 2.93 2.94 2.98 3.00 3.09 3.22 3.41 3.55

Speed (cm/s) before amputation

If  $j$  was not an integer, round  $j$  (e.g., say  $j$  was 4.32 then round  $j = 4$ ). We would then have picked the 4<sup>th</sup> value in the ranked distribution (i.e., 2.31 cm/s)

15

---

---

---

---

---

---

---

---

The accompanying metric of spread for the median is called the interquartile range

Speed (cm/s) before amputation

Positioning  $Q_3, j = 0.75n = (0.75)(16) = 12$

where  $n$  is the number of observations. If  $j$  is an integer (whole number, not a fraction), then the 3<sup>rd</sup> quartile is the average of  $Y_{(j)}$  and  $Y_{(j+1)} = Y_{(12)}$  and  $Y_{(12+1)} = (3.00 + 3.09) / 2 = 3.045$  cm/s

**Third quartile ( $Q_3$ ) = 3.045 cm/s**

16

---

---

---

---

---

---

---

---

The accompanying metric of spread for the median is called the interquartile range

Speed (cm/s) before amputation

If  $j$  was not an integer, round  $j$  (e.g., say  $j$  was 12.72 then  $j = 13$ ). We would then have picked the 13<sup>th</sup> value in the ranked distribution (i.e., 3.09 cm/s)

17

---

---

---

---

---

---

---

---

The accompanying metric of spread for the median is called the interquartile range

Speed (cm/s) before amputation

The **interquartile range (IQR)** for the speed data before amputation is then  $Q_3 - Q_1 = 3.045 - 2.340 = 0.705$  cm/s

18

---

---

---

---

---

---

---

---

Remember: the mean carries information about all values in any given frequency distribution, but it is influenced by extreme values. The median does not characterize frequency distributions as well as the mean (i.e., not influenced by all values), but it is not sensitive to extreme values.

$Y = 53, 58, 62, 64, 68, 72, 73, 77, 86, 87, 88, 92$

$\bar{Y} = 73.3$   
 $Q_2 = 72.5$

$Y = 53, 58, 62, 64, 68, 72, 73, 77, 86, 87, 88, 192$

$\bar{Y} = 81.7$   
 $Q_2 = 72.5$

---

---

---

---

---


---

---

---

19

Let's take a power break – 2 minutes




---

---

---

---

---

---

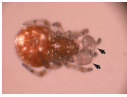
---

---

20

The median is the middle measures of a set of observations (distribution)


If the number of observations ( $n$ ) is **even**, then the median is calculated differently:



It gives an "arm" (or a pedipalp) for a female spider.

Running speed (cm/s) of male *Tidarren* spiders before and after voluntary amputation of one pedipalp.

Spider	Speed before	Speed after	Spider	Speed before	Speed after
1	1.25	2.40	9	2.98	3.70
2	2.94	3.50	10	3.55	4.70
3	2.38	4.49	11	2.84	4.94
4	3.09	3.17	12	1.64	5.06
5	3.41	5.26	13	3.22	3.22
6	3.00	3.22	14	2.87	3.52
7	2.31	2.32	15	2.37	5.45
8	2.93	3.31	16	1.91	3.40



*Oxyopes salticus*

---

---

---

---

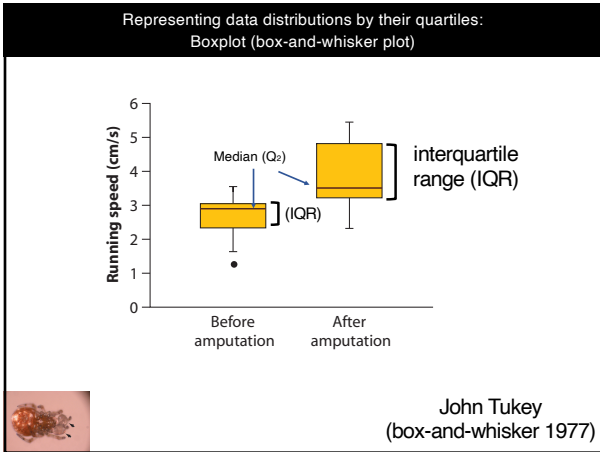
---

---

---

---

21



22

---

---

---

---

---

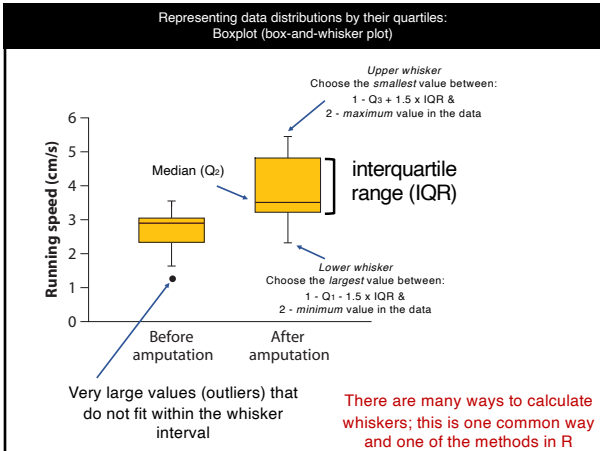
---

---

---

---

---



23

---

---

---

---

---

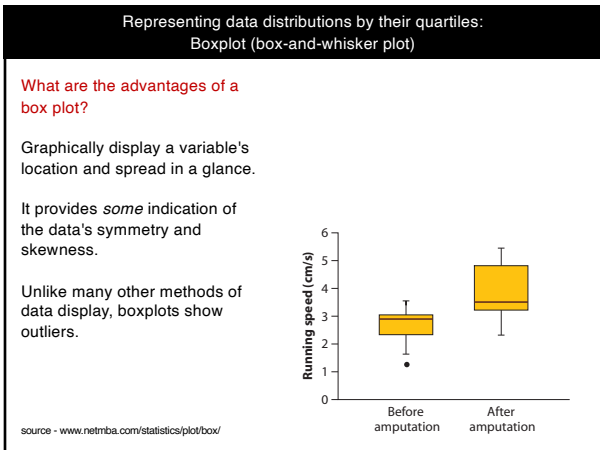
---

---

---

---

---



24

---

---

---

---

---

---

---

---

---

---



Representing data distributions by their quartiles:  
Boxplot (box-and-whisker plot)

**What are the advantages of a box plot?**

Graphically display a variable's location and spread at a glance.

It can (not always) provide **some** indication of the data's symmetry and skewness (not always true but very often the case).

median < mean

right skewed

median = mean

symmetric

median > mean

left skewed

---

---

---

---

---

---

---

---

---

---

25

Representing data distributions by their quartiles:  
Boxplot (box-and-whisker plot)

Three fictional data sets to show calculation and properties of distributions via their boxplots (boxplot in the next slide) – do you see their differences?

Left-skewed distribution: 9,11,31,44,52,58,61,61,63,64,66

---

---

---

---

---

---

---

---

---

---

26

Representing data distributions by their quartiles:  
Boxplot (box-and-whisker plot)

Three fictional data sets to show calculation and properties of distributions via their boxplots (boxplot in the next slide) – do you see their differences?

Right-skewed distribution: 1,2,3,4,5,6,7,10,20,30,40

---

---

---

---

---

---

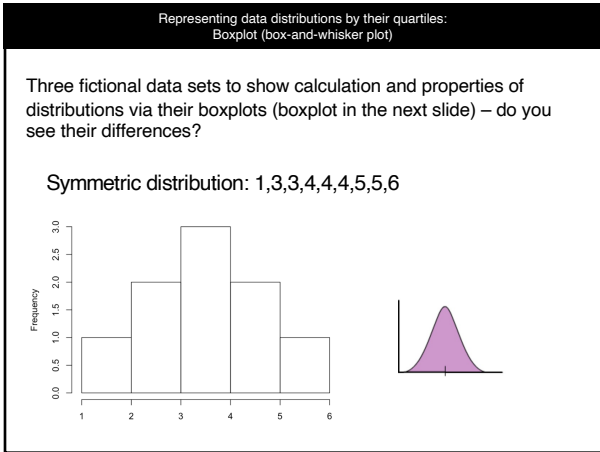
---

---

---

---

27



28

---

---

---

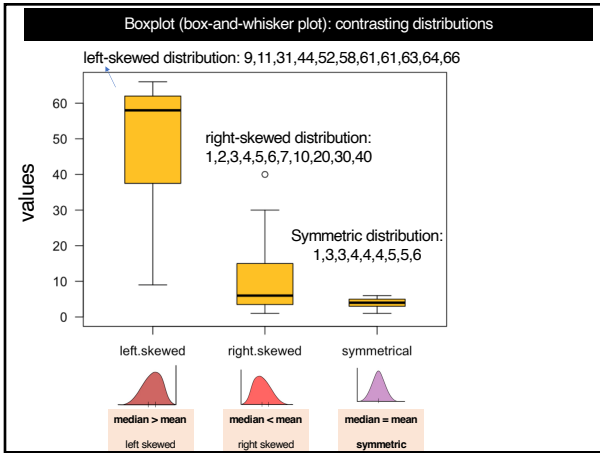
---

---

---

---

---



29

---

---

---

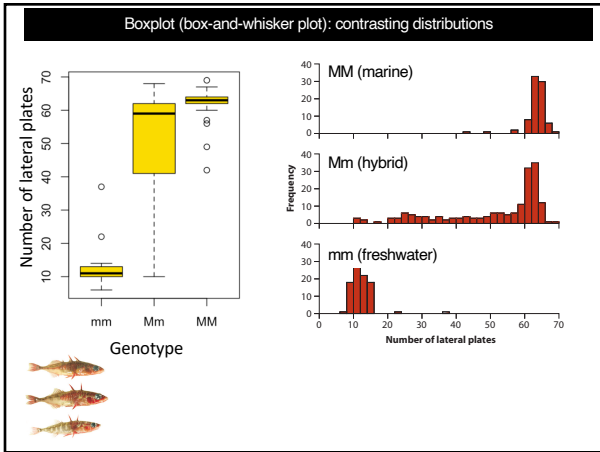
---

---

---

---

---



30

---

---

---

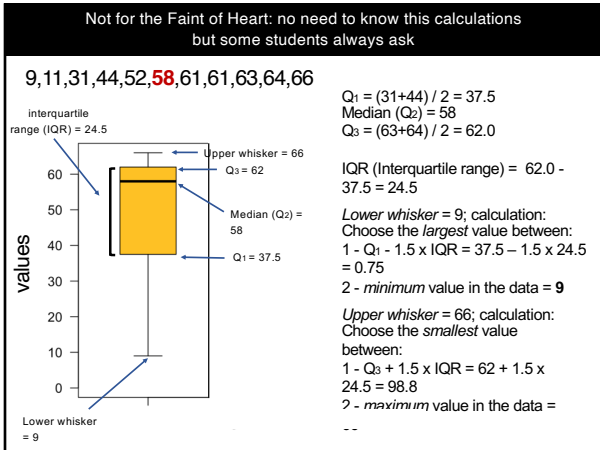
---

---

---

---

---




---

---

---

---

---

---

---

---

31

**Statistics is based on samples!**

The most important goal of statistics is to estimate (infer) an unknown quantity of an entire population based on sample data.

Statistics is the science of making decisions with incomplete knowledge (i.e., based on samples) based on populations that too often have unknown sizes.

But sample quantities (mean, median, standard deviation, etc) vary from sample to sample (i.e., they have some level of uncertainty).

**Next lecture - Estimating with uncertainty**

---

---

---

---

---

---

---

---

32