Chapter 3:
Describing data

1

2

3

Ex 3.1: Gliding snakes

- Sample of 8 individuals

Ind	Rate
1	0.9
2	1.4
3	1.2
4	1.2
5	1.3
6	2.0
7	1.4
8	1.6

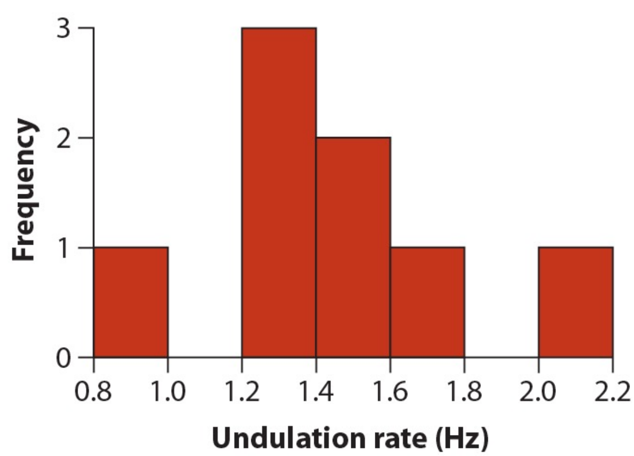


Fig 3.1-1

4

Sample mean

- The **sample mean** is the sum of all the observations in a sample divided by *n*, the number of observations

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

Y = variable
Y_i = measurement for individual i
n = number of observations

$$\bar{Y} = \frac{0.9+1.2+1.2+2.0+1.6+1.3+1.4+1.4}{8} = 1.375$$

5

Spread of the distribution

Rate	$Y_i - \bar{Y}$	$\left(Y_i - \bar{Y}\right)^2$
0.9		
1.2		
1.2		
1.3		
1.4		
1.4		
1.6		
2.0		
Sum		

deviation

square of deviation

6

Spread of the distribution

Rate	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$
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	0.735

sum of squares

7

Spread of the distribution

Rate	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$
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	0.735

- **Variance**

$$s^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}{n-1}$$

8

Spread of the distribution

Rate	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$
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	0.735

- The **standard deviation** is a common measure of the spread of the distribution. It indicates just how different measurements typically are from the mean

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

9

Spread of the distribution

Rate	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$
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	0.735

Variance

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} = \frac{0.735}{7} = 0.11$$

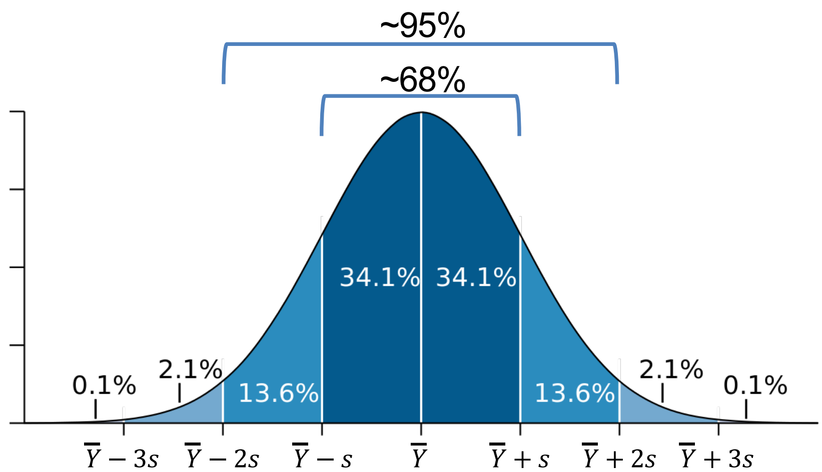
Standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}} = \sqrt{\frac{0.375}{7}} = 0.324037$$

10

Standard deviation

- Never negative, and has same units as the observations
- If the frequency distribution is bell shaped:



11

Comparing spread of distribution in different populations (same units)

- Within a sample, the mean and the standard deviation have the same units (apples to apples)
- But values for the same variable can vary considerably across populations
- Consider weight in elephants (big apples) vs mice (tiny apples)
  - Standard deviation of 2 g in sample of mice might be a bigger relative spread than a standard deviation of 2000 g in elephants

12

Comparing spread of distribution in different populations (diff units)

- What if you have two different measurements in two different samples, and you want to know which one is more "spread out?"
- For example, weight (kg) and resting heart rate (bpm) in sample of elephants
- Now you have apples (kg) and oranges (bpm), so comparing the standard deviations doesn't make sense

13

Coefficient of variation

- The **coefficient of variation** is the standard deviation expressed as a percentage of the mean

$$CV = \frac{s}{\bar{\bar{Y}}} \times 100\%$$

Can be used to compare variability of same trait in different populations (e.g., weight in elephants vs mice) or different traits (e.g., heart rate vs cholesterol)

14

15

Ex 3.2: I'd give my right arm for a female

Tidarren spider. Males are tiny, but have large pedipalps (arrows). Male amputates one when searching for mate

TABLE 3.2-1 Running speed (cm/s) of male *Tidarren* spiders before and after voluntary amputation of a pedipalp.

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

16

Median

- The **median** is the middle measurement of a set of observations

- The "middle" value in a set of **sorted** observations

Odd number of obs:

$$Median = Y_{(n+1/2)}$$

Even number of obs:

$$Median = \left[Y_{(n/2)} + Y_{\left(\frac{n}{2}+1\right)}\right]/2$$

17

Ex 3.2: I'd give my right arm for a female

Even number of obs:

$$Median = \left[Y_{(n/2)} + Y_{\left(\frac{n}{2}+1\right)}\right]/2$$

$$Median = \frac{[2.87 + 2.93]}{2} = 2.90$$

TABLE 3.2-1 Running speed (cm/s) of male *Tidarren* spiders before and after voluntary amputation of a pedipalp.

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

18

Quartiles and interquartile range

- Quartiles are values that partition the data into quarters
- The **interquartile range** (IQR) is the difference between the third and first quartiles of the data. It is the span of the middle 50% of the data

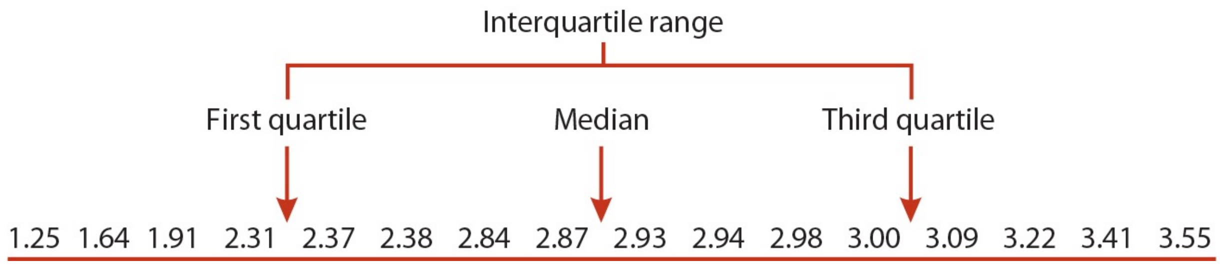


Fig 3.2-1

19

Quartiles and interquartile range

- Quartiles are values that partition the data into quarters
- The **interquartile range** (IQR) is the difference between the third and first quartiles of the data. It is the span of the middle 50% of the data

- The *X*th percentile is the sample below which *X* percent of the observations lie
  - If you're the 84th percentile of GPA for your class then you're GPA is higher than 84% of your classmates

20

Boxplot

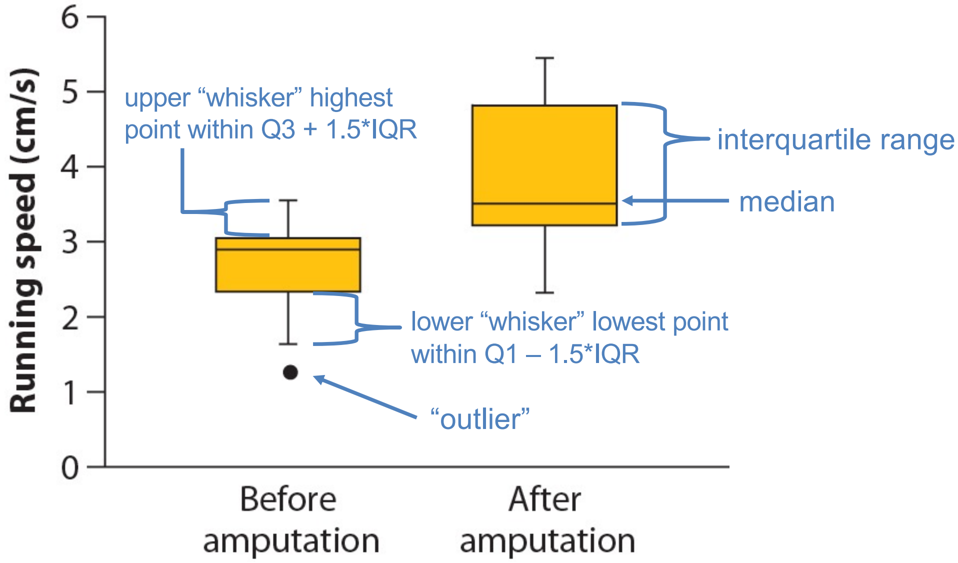


Fig 3.2-2

21

Cumulative frequency

- **Cumulative relative frequency** at a given measurement is the fraction of observations less than or equal to that measurement

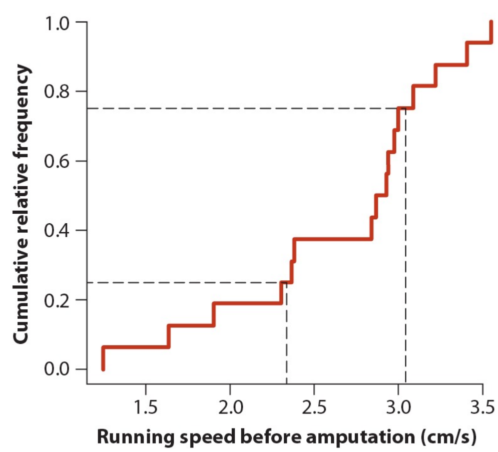


Fig 3.4-1

22

23

Mean vs median

- Which gives more insight of the distribution of measurements?
- Depends on the *shape* of the frequency distribution

24

Ex 3.3: Disarming fish

- Threespine stickleback
- Spines and bony plates reduce mortality
- Number of bony plats determined by single gene *Ectodysplasin*
  - *MM* = many
  - *Mm* = variable
  - *mm* = few

https://www.youtube.com/watch?v=Pv4Ca-f4W9Q

25

Ex 3.3: Disarming fish

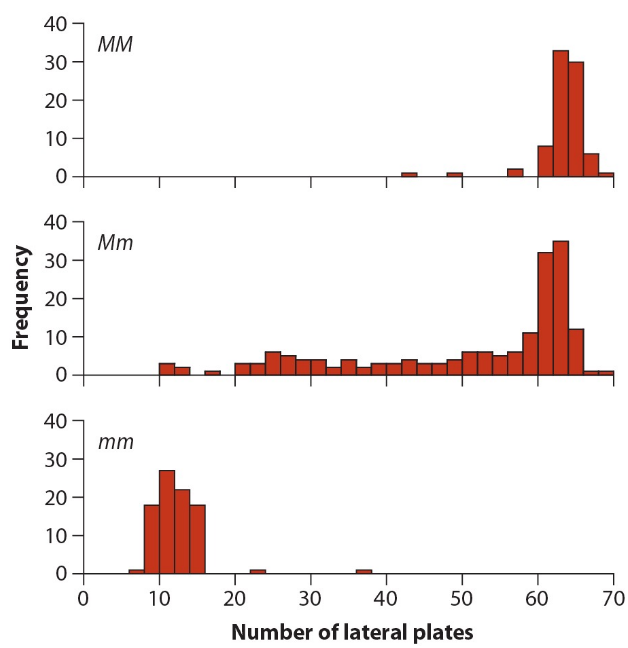


Fig 3.3-1

26

Mean vs median

- When the frequency distribution is symmetric and bell-shaped the mean and median will be similar
  - e.g., *MM* and *mm*

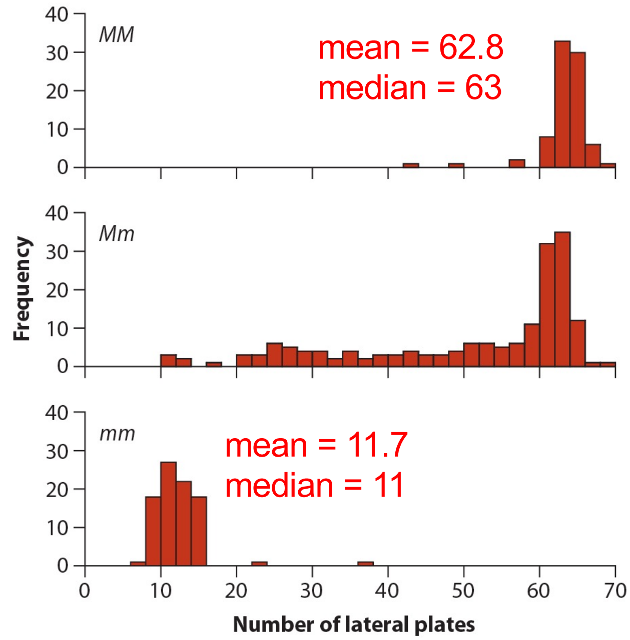


Fig 3.3-1

27

Mean vs median

- BUT when the frequency distribution is asymmetric the values will differ

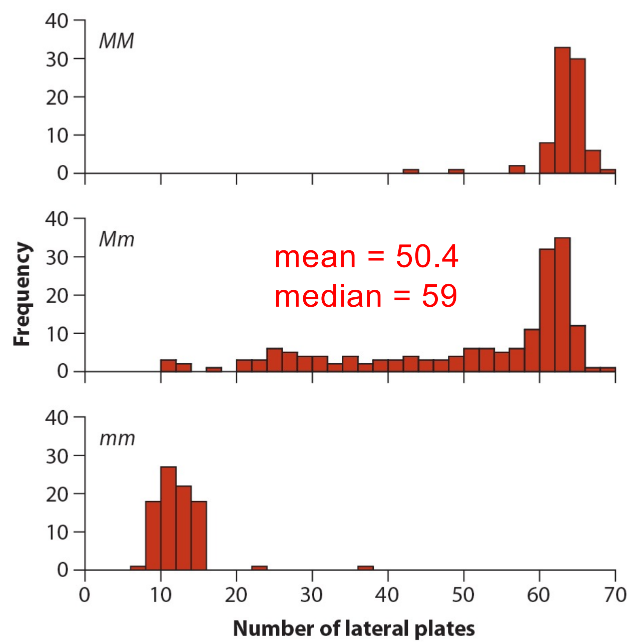
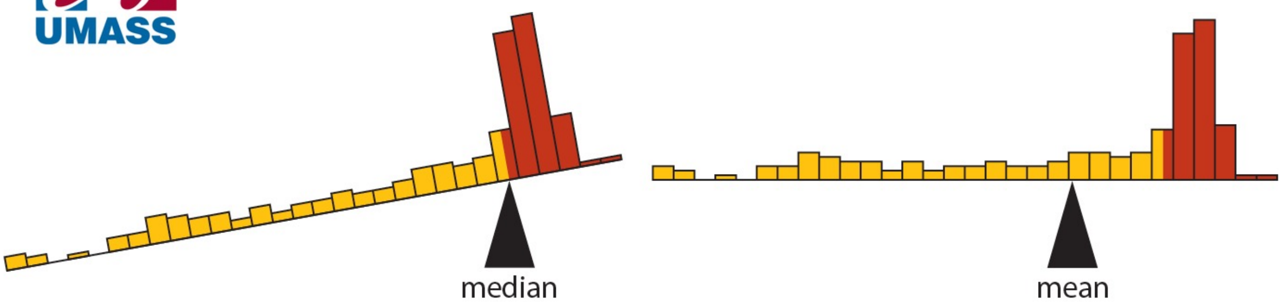


Fig 3.3-1

28

Mean vs median



The median is the middle value, whereas the mean is the “center of gravity”

Fig 3.3-2

29

Proportion

- Proportion of observations in a given category

$$\hat{p} = \frac{num.\, in\ category}{n}$$

TABLE 3.5-1 The number of fish of each genotype from a cross between a marine stickleback and a freshwater stickleback (Example 3.3). As written, the sum of the proportions does not add precisely to one because of rounding.

Genotype	Frequency	Proportion
MM	82	0.24
Mm	174	0.51
mm	88	0.26
Total	344	1.00

30