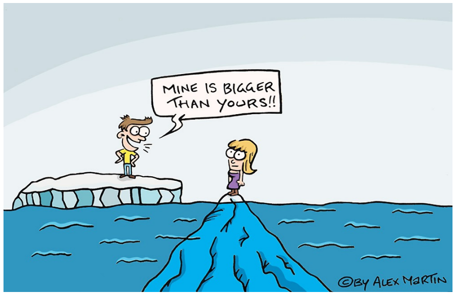
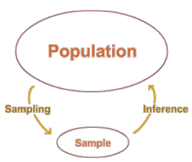
Statistics is about drawing conclusions from incomplete information. While errors can happen, statistical methods help us estimate the associated risks and uncertainties.



1

Statistics is based on samples!



The primary goal of statistics is to estimate (infer) an unknown characteristic of an entire population based on sample data.

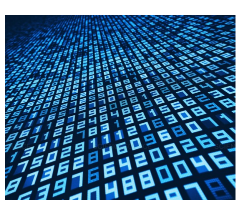
These estimates, often derived from descriptive statistics, are then used to make informed decisions about the population.

2

Describing data

Samples and populations typically consist of numerous individual observational units along with their associated data (observations, variables).

To describe samples effectively, we use summary statistics (mean, median, variance, etc.), which serve as estimates of the corresponding statistics for the entire population.



3

Describing data

Samples and populations typically consist of numerous individual observational units along with their associated data (observations, variables).

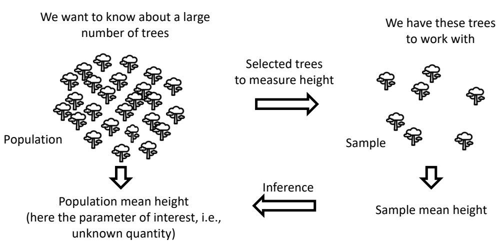
To describe samples effectively, we use summary statistics (mean, median, variance, etc.), which serve as estimates of the corresponding statistics for the entire population.

Today: data summaries for each variable (separately).

Individual	Weight (kg)	Height (cm)
1	75.5	172
2	55.3	152
3	61.2	164
4	50.3	148
5	99.4	192
6	66.2	171
7	75.3	169
8	74.6	182
9	60.5	162
10	93.5	184
11	73.6	169

4

The primary goal of statistics is to **estimate (infer)** an unknown characteristic (e.g., height) of an entire population based on sample data.



Inspired by https://www.cliffsnotes.com/study-guides/statistics/sampling/populations-samples-parameters-and-statistics

5

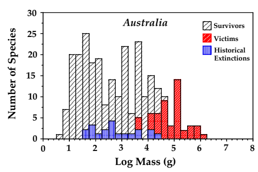
Key Learning Objectives today

1. Differentiate between estimates of location and estimates of spread (or width).
2. Recognize that variability is not simply noise but is a key parameter that can be estimated.
3. Become familiar with the most common descriptive statistics.
4. Know when the mean or median is a more appropriate summary of location.
5. Location and spread summaries of single variables (multiple variables later in the course).

6

Scientific question: Did humans drive mammal extinctions in Australia?

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



Frequency distribution of mammal mass categorized into survivors, "victims" and older (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).

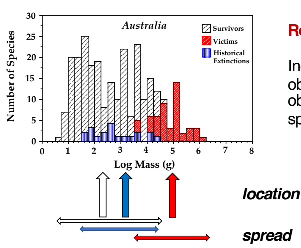
Study by Lyons et al. (2004; Evolutionary Ecology Research 6:339-358)

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

7

Descriptive statistics or summary statistics are needed to make inferences

- **Location** tells us something about the average or typical individual units (i.e., where the observations are centered).
- **Spread** tells us how measurements vary among individual units (or observations), i.e., how widely scattered the values are around the center (location).



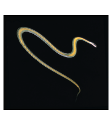
Remember the jargon (lecture 2):

Individual units (of data) are called observation units (here each observational unit is a single species).

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

8

The most important location statistic: Arithmetic mean



"Flying" paradise tree snake (*Chrysopelea paradisi*). To better understand how lift is generated, Socha (2002) videotaped glides (from a 10-m tower) of 8 snakes. Rate of side-to-side undulation was measured in hertz (number of cycles per second). The values recorded were:

0.9, 1.4, 1.2, 1.2, 1.3, 2.0, 1.4, 1.6

The arithmetic mean is an algorithm = a process or set of rules to be followed in calculations - sum of all the observations in a sample divided by ***n***, the number of observations.

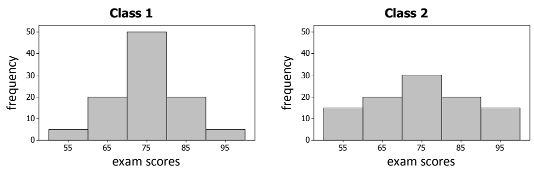
$$\bar{Y} = \frac{0.9 + 1.2 + 1.2 + 2.0 + 1.6 + 1.3 + 1.4 + 1.4}{8} = 1.375 \text{ Hz.}$$

The sample mean is represented most often as Y *or* X said « *Y bar* » or « *X bar* »

9

The concept of spread around the mean

Which class has the most variation in exam scores?



Note: scales (X and Y axis limits) are exactly the same

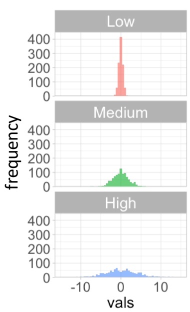
- Source: Cooper & Shore; Journal of Statistics Education (vol. 18, #2)

10

The concept of spread around the mean

Population variability should not be dismissed as mere noise around the mean; it holds biological significance in its own right.

As location values (e.g., mean, median), variation (spread) also has an inherent true value within a population (a parameter) that we aim to estimate through sampling.



Modified from:

11

The most important spread statistics: variance and standard deviation (the accompanying statistics of spread for the mean)

It indicates how far the different measurements typically are from the mean. The standard deviation is large if most observations are far from the mean, and it is small if most measurements lie close to the mean.

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

12

The most important spread statistics: variance and standard deviation (the accompanying statistics of spread for the mean)

Important measure: "Sum of Squared deviations from the mean"

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

variance

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

$$s^2 = \frac{0.735}{7} = 0.11\ \text{Hz}^2.$$

Variance is the "average" squared deviation from the mean (units, here Hz, are squared, i.e., Hz^2.

standard deviation

$$s = \sqrt{\frac{\Sigma(Y_i - \bar{Y})^2}{n-1}} = \sqrt{\frac{0.735}{7}} = 0.324037\ \text{Hz}.$$

Square root of the variance (in the same unit as the original variable).

13

The most important spread statistics: variance and standard deviation

Why is the sum of the squared deviations from the mean divided by *n-1* and not *n?*

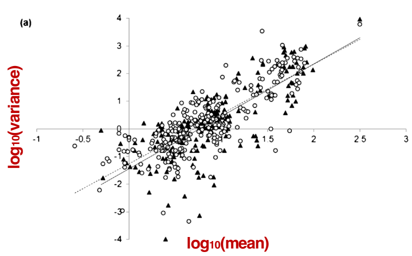
We will understand this in a couple of lectures!

variance

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

14

Variance (and standard deviation) often varies as a function of the mean (i.e., "bigger things tend to vary more")



Relationship between x = $\log_{10}$(mean) and y = $\log_{10}$(variance) for grain yields (biomass) of two lentil genotypes. Each point (circle of triangle) represents a crop field, showing the mean and variance in biomass. The two different symbols (circles and triangles) represent different genotypes. Study by Döring et al. (2015, Field Crops Research 183:294–302)

15

A relative metric of spread: the coefficient of variation (CV)
often important when comparing groups of individuals belonging to different classes or variables with different units.

Elephants have greater mass than mice, but they may vary less in mass than mice relative to their means. When comparing variables that vary in variance and scale (e.g., ºC and F), we may care more about the relative variation among individuals.

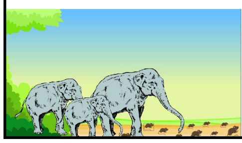
A higher CV means that there is more variability relative to the mean when compared to a lower CV.

coefficient of variation:

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

Snake undulation data:

$$\mathrm{CV} = \frac{0.324}{1.375} 100\% = 24\%$$



16

A relative metric of spread: the coefficient of variation (CV)
often important when comparing groups of individuals belonging to different classes or variables with different units.

				X	s	CV
1	2	3	4	2.5	1.29	51.7%
31	32	33	34	32.5	1.29	3.97
204	205	206	207	205.5	1.29	0.63
1300	1301	1302	1303	1301.5	1.29	0.10

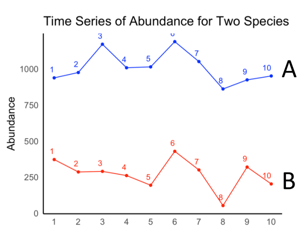
Making the coefficient of variation (CV) more obvious!

17

When using the coefficient of variation, it's crucial to consider why you want to express variation (in this case, standard deviation) relative to the mean.

Consider the case of species fluctuations in abundance through time:

Statistics	Species A	Species B
Mean (X)	1008.2	270.8
Standard deviation (s)	104.9	103.9
Coefficient of variation (CV)	10.4	38.4



Species A: More abundance (mean) but the same variation (*s*) through time as species B.

CV_B > CV_A = greater risk of extinction of Species B.

Species B: Less abundance (mean) but the same variation (*s*) through time as species A.

18

Before we go too far: a word on rounding numerical values

- When recording data, always retain as many significant digits (often involving decimals places) as your calculator or computer can provide.
- When presenting results, however, numbers should be rounded before being presented.
- There are no strict rules on the number of digits that should be retained when rounding.
- A common strategy, is to round descriptive statistics (e.g., means, standard deviations) to one decimal place more than the original measurements themselves.

Example: the mean rate of undulation for the eight snakes (measured with a single decimal place; e.g., 0.9), calculated as 1.375 Hz, would be communicated as:

0.9, 1.4, 1.2, 1.2, 1.3, 2.0, 1.4, 1.6 $\overline{Y} = 1.38$ Hz

19

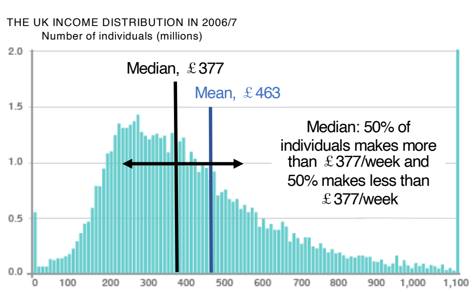
Let's take a break – 1 minute

20

Arithmetic mean *versus* **median** – the second most common statistic to describe the location of a frequency distribution

Arithmetic mean is influenced by how unbalanced (i.e., asymmetric) the distribution is by extreme values.

The median is the middle measure (value) of a set of observations (distribution).

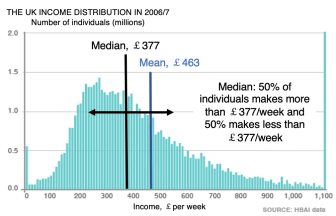


21

Arithmetic mean *versus* **median** – the second most common statistic to describe the location of a frequency distribution

Arithmetic mean is influenced by how unbalanced (i.e., asymmetric) the distribution is by extreme values.

The median is the middle measure (value) of a set of observations (distribution).

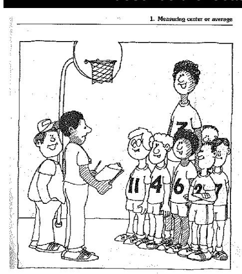


The fewer rich (income on the right of the distribution) drives the mean to be larger than the median.

The **median** reflects the midpoint of salaries, representing what a typical person might earn, while the **mean** provides the overall average but can be influenced by extremely high or low salaries.

22

Arithmetic mean *versus* **median** – the second most common statistic to describe the location of a frequency distribution



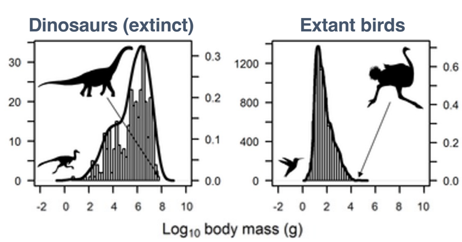
Should we scare the the opposition by announcing our mean or calm them by announcing our median height?

The mean is more sensitive to extreme (large or small) values than the median, which is good for inference. But depending on the distribution, however, the mean is too influenced by extreme values.

23

Arithmetic mean *versus* **median** – symmetry of the frequency distribution

Frequency distributions of species body size



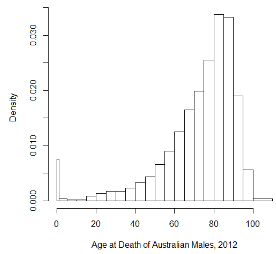
Median > Mean **Median < Mean**

Study by Gorman and Hone (2012, PLOS ONE 8: 10.1371)

24

Arithmetic mean *versus* median – symmetry of the frequency distribution

Frequency distribution of age at death of Australian males, 2012



Median > Mean

Histogram from https://stats.stackexchange.com/questions/89179/real-life-examples-of-distributions-with-negative-skewness

25

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

If the number of observations (n) is ***odd***, then the median is the middle observation:

Values in the distribution (mm):

7, 12, 5, 9, 8, 5, 15, 13, 3

Order values (mm):

3, 5, 5, 7, 8, 9, 12, 13, 15

Median = 8.0 mm

(4 observations in each side of the distribution)

26

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

If the number of observations (n) is ***odd***, then the median is the middle observation:

$$\text{Median} = Y_{([n+1]/2)} = Y_{([9+1]/2)} = Y_5$$

Ordered values (mm):

3, 5, 5, 7, 8, 9, 12, 13, 15

Median = Y_5 = 8.0 mm

(4 observations in each side of the distribution)

27

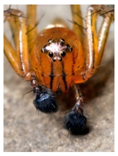
The median is the middle measures of a set of observations (distribution)

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

Tidarren (spider)

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

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



Oxyopes salticus

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

28

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

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

For an ***even*** number of observations, the median is the average of the two central numbers. n = 16 in this study.

Median (speed before) = M = 2.90 cm/s

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

$$\text{Median} = [Y_{(n/2)} + Y_{(n/2+1)}]/2.$$

$$\text{Median} = (2.87 + 2.93)/2 = 2.900 \text{ cm/s}$$

29

Arithmetic mean *versus* median – the second most common statistic to describe the location of a frequency distribution

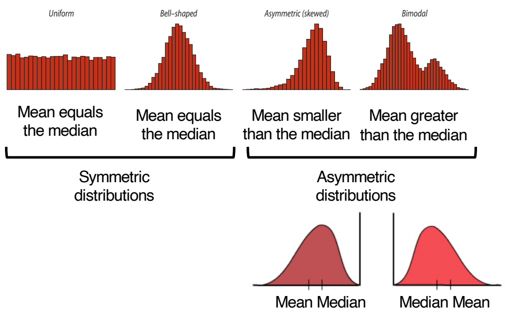
					X	Median
1	2	3	4	5	3	3
1	2	3	4	489	99.8	3
1	2	3	4	6	3.2	3

Highlighting how extreme values have a greater impact on the mean than on the median!

30

Arithmetic mean *versus* median

The arithmetic mean is affected by the imbalance (asymmetry) in the distribution caused by the presence of extreme values.



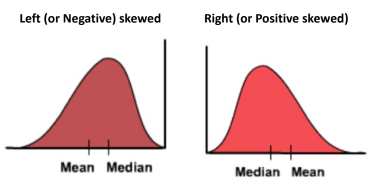
31

Arithmetic mean *versus* median

The arithmetic mean is affected by the imbalance (asymmetry) in the distribution caused by the presence of extreme values.

Asymmetric distributions (skewed) can be either left of positive skewed

The rule comparing the mean to the median is particularly effective for large datasets (more than 30 observations).



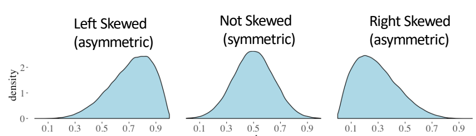
32

Mean or Median? Consider Skewness

- Few small values.
- > 1/2 of values exceed the mean.

- As many large as small values.
- ~ 1/2 of values exceed the mean.

- Few large values.
- > 1/2 of values are less than the mean.



For measures like income, **medians** are generally preferable to **means**.

This is because populations are right-skewed (there are few very rich people), and we care more about the "average person than the average of people".

33