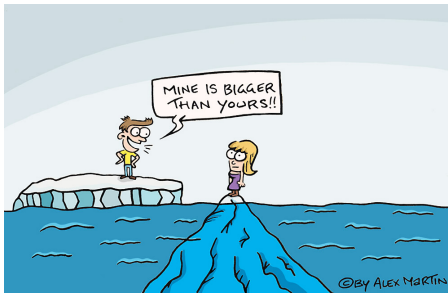


REPORTS - Data generation philosophy

In some reports, you will **generate your own datasets**. This is a deliberate pedagogical choice. This approach helps you think more carefully about how different statistical metrics and methods behave and what they actually measure.

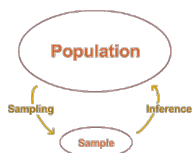
1

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



2

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.

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).

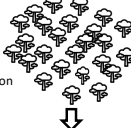
Individual	Weight (kg)	Height (cm)
1	75.5	172
2	55.3	152
3	61.2	164
4	50.3	148
5	99.4	192
6	66.2	171
7	75.3	169
8	74.6	182
9	60.5	162
10	93.5	184
11	73.6	169



4

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

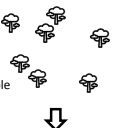
We want to know about a large number of trees



Population

Population mean height
(here the parameter of interest, i.e., unknown quantity)

Selected trees to measure height



Sample

Sample mean height

Inference

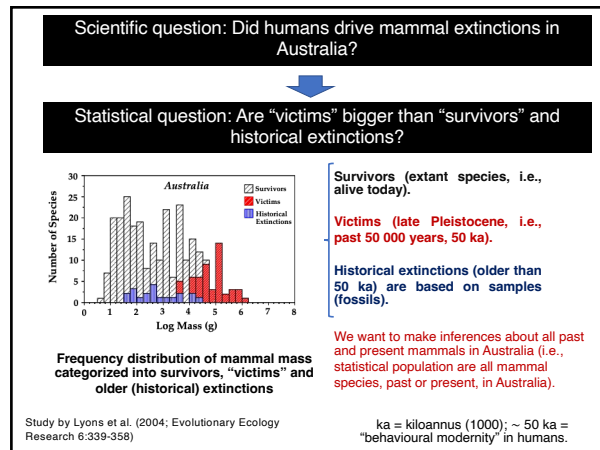
Inspired by <https://www.diffondos.com/study-guides/statistics/sampling/populations-samples-parameters-and-statistics>

5

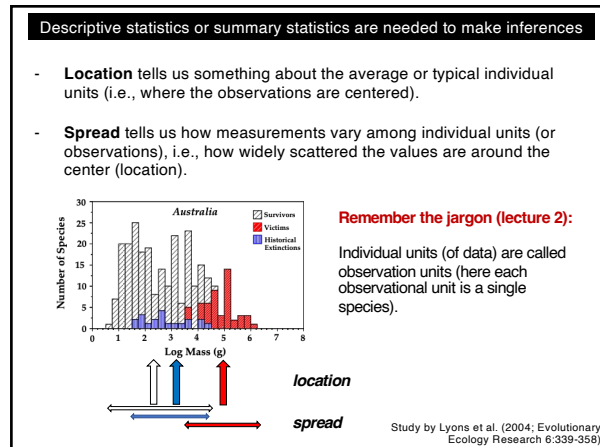
Key Learning Objectives today

1. Differentiate between measures of location and measures of spread (or width).
2. Recognize that variability is not merely noise, but a fundamental parameter that can be estimated and interpreted.
3. Become familiar with the most commonly used descriptive statistics.
4. Understand when the mean or the median provides a more appropriate summary of location.
5. Summarize single variables using both location and spread measures (with extensions to multiple variables later in the course).

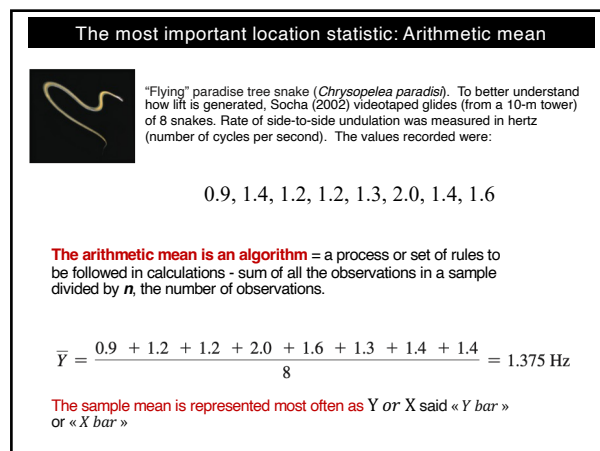
6



7



8

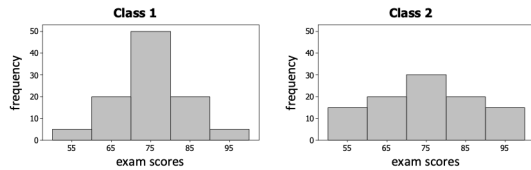


9

The concept of spread around the mean

Deviation (from the mean) measures uncertainty. When deviation is small, values cluster near the mean, making the mean a good guess for a randomly chosen observation. When deviation is large, values are more spread out, increasing uncertainty in any guess.

Which class has the most variation in exam scores?



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

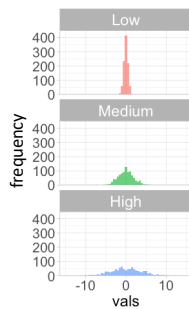
~ Source: Cooper & Shore; Journal of Statistics Education (vol. 18, #2)

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: © 2020 W.H. Freeman and Company

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

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

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

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

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

standard deviation

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

Variance is the average squared deviation of observations from the mean, measuring overall uncertainty or spread (units, here Hz, are squared, i.e., Hz²)

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 (number of observations)?

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 $x = \log_{10}(\text{mean})$ and $y = \log_{10}(\text{variance})$ 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.

The CV tells you how difficult it is to predict a typical value (average values) relative to its magnitude. A small CV means values are tightly clustered around the typical value, whereas a large CV means greater uncertainty relative to the size of that value.

coefficient of variation: $CV = \frac{s}{\bar{Y}} \times 100\%$

Snake undulation data: $CV = \frac{0.324}{1.375} 100\% = 24\%$



CV answers: Are individuals similarly variable relative to what is typical (average) individual is when values differ greatly in magnitude or units?

A 5-gram deviation matters a lot for a mouse but very little for an elephant. CV captures this by measuring variation relative to average size, not in absolute grams.

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.

Making the coefficient of variation (CV) more obvious!

				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

CV answers: Are individuals similarly variable relative to what is typical (average)? even when values differ greatly in magnitude or units.

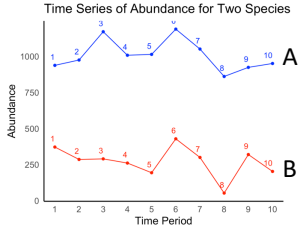
17

When using the coefficient of variation, it is important to be clear about why variation (measured by the standard deviation) is being expressed relative to the mean.

Consider the case of species fluctuations in abundance through time:

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

Time Series of Abundance for Two Species



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

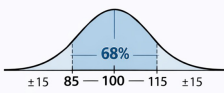
CV_B > CV_A = greater risk of extinction of Species B.

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

18

The CV as a measure of variability around the mean, i.e., the CV tells us how risky it is to use the mean as a predictor of individual observations.

CV = 0.15
Mean = 100
SD = 15
68% within 85 to 115



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

A CV of 0.15 means that the standard deviation is 15% of the mean. For example, if the mean is 100, the standard deviation is 15.

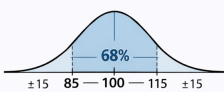
If the data are approximately normally distributed, about two-thirds of the observations (~68%) will fall within one standard deviation of the mean—that is, within 15 units of the mean, or between 85 and 115.

Thus, saying that the standard deviation is 15% of the mean describes the relative scale of variation, not the percentage of observations within a particular range. Assuming a normal distribution allows us to translate that scale of variation into specific frequencies. With a non-normal distribution, the CV still describes the same relative scale of variation, but the proportion of observations within one standard deviation of the mean may differ from 68%.

19

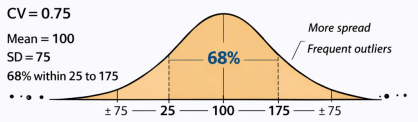
The CV as a measure of variability around the mean, i.e., the CV tells us how risky it is to use the mean as a predictor of individual observations.

CV = 0.15
Mean = 100
SD = 15
68% within 85 to 115



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

CV = 0.75
Mean = 100
SD = 75
68% within 25 to 175



In both examples above the mean is identical, but the relative spread of the data, and therefore the variability associated with using the mean as a predictor, is dramatically different.

20

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:

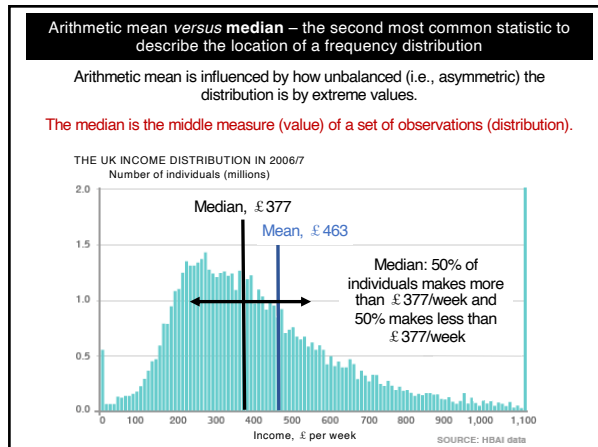
0.9, 1.4, 1.2, 1.2, 1.3, 2.0, 1.4, 1.6

$$\bar{Y} = 1.38 \text{ Hz}$$

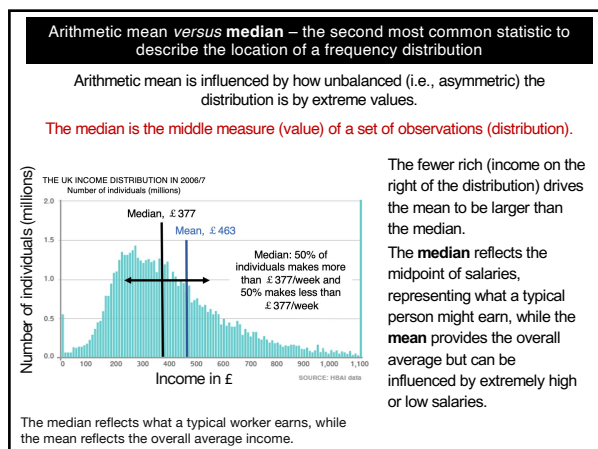
21



22



23



24

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

1. Measuring water or average



Should we intimidate the opposition by announcing our **mean height**, or “reassure” them by announcing our **median height**?

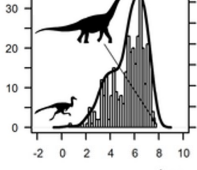
The mean is more sensitive to extreme (large or small) values than the median, which is good for inference. But depending on the distribution, however, the mean is too influenced by extreme values.

25

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

Frequency distributions of species body size

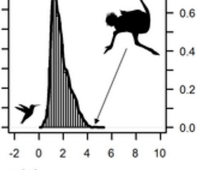
Dinosaurs (extinct)



Median > Mean

The typical dinosaur is relatively large

Extant birds



Median < Mean

The typical bird is relatively small

Log₁₀ body mass (g)

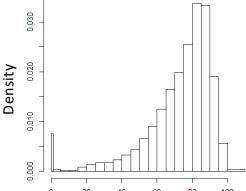
Study by Gorman and Hone (2012, PLOS ONE 8: 10.1371)

26

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

Histogram of age at death of Australian males, 2012

The height of a bar (*density*) reflects the probability that a death occurs within a given age interval



Age of death of Australian Males, 2012

Median > Mean

Most Australian males die at old ages, but a small number of deaths at very young ages pull the mean age at death downward, making the mean smaller than the median.

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

27

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)

28

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:

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)

29

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:

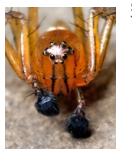


Tidarren (spider)

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.

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

30

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	2.99	3.70
2	2.94	3.50	10	3.51	4.70
3	2.58	4.09	11	2.80	4.26
4	3.09	3.17	12	1.64	5.09
5	3.41	3.36	13	3.22	3.22
6	3.00	3.22	14	2.67	3.12
7	2.31	2.32	15	2.37	3.45
8	2.97	3.71	16	3.01	3.60

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

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

Median = $(2.87 + 2.93)/2 = 2.900$ cm/s

31

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!

32

Arithmetic mean *versus* median

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

Uniform

Mean equals the median

Bell-shaped

Mean equals the median

Asymmetric (skewed)

Mean smaller than the median

Bimodal

Mean greater than the median

Symmetric distributions

Asymmetric distributions

33

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 or positive skewed

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

Left (or Negative) skewed

Mean Median

Right (or Positive skewed)

Median Mean

34

Mean or Median? Consider Skewness

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

- As many large as small values.
- $\sim 1/2$ of values exceed the mean.

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

Left Skewed (asymmetric)

Not Skewed (symmetric)

Right Skewed (asymmetric)

For measures like income, **medians** are generally preferable to **means**. This is because income distributions are right-skewed (a few individuals are extremely wealthy), and we are usually more interested in what a typical ("average") person earns than in the arithmetic average of all incomes.

© 2020 W.H. Freeman and Company

35
