Module 4: Descriptive Statistics

Introduction to Statistics in Kinesiology

Furtado JR., O

Cal State Northridge

updated: 2022-02-16

```
xaringanExtra::use_webcam()
```

> It is not knowledge, but the act of learning, not possession but the act of getting there, which grants the greatest enjoyment.
>
> — Gauss, C.

Overview

Topics to be covered:

- Measures of central tendency
- Measures of variability
- Skewness and kurtosis
- Standard scores

Measures of central tendency - Mean

- The **mean**[1] is the most popular index of central tendency
- The most sensitive of the central tendency indices
  - affected by every score in the
  - greatly affected by outliers
- Play important role on statistical inference
- Used with interval and ratio data[2]

jamovi considers interval and ration as **continuous**

Equation for the sample mean:

$$\bar{x} = \frac{\sum x}{n}$$

Equation for the population mean:

$$\mu = \frac{\sum x}{N}$$

[1] equal to arithmetic average [2] In `jamovi`, interval and ratio = **continuous** data

Measures of central tendency - Median

- Represents the score at the **50th percentile**
- Divides the data set in two
- Considered the “typical” score because it best represents the majority of other values
- Calculating the median does not take into consideration the value of other scores
- If *N* is odd, **median** is the middle score[1]

Measures of central tendency - Median

- If *N* is even, do one of the following:
  - Use the higher of the two middle scores
  - Compute the average of the two middle scores
  - Use with ordinal data or cases of highly skewed distributions
- Important: not affected by extreme scores

[1] data must be ranked first.

Measures of central tendency - Median

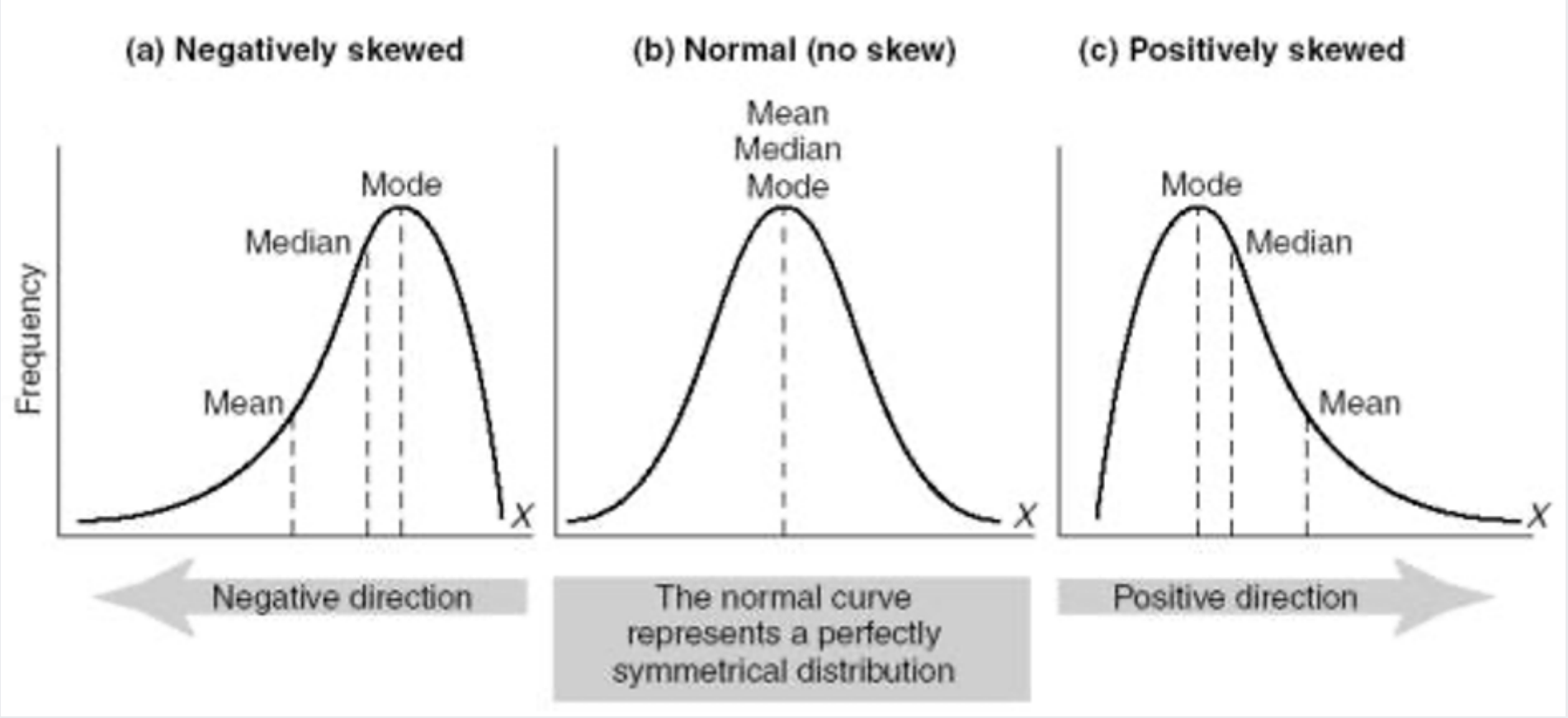
When distribution is EVEN | When distribution is ODD

id	scores
1	2
2	4
3	7
4	11
5	12
6	14

Measures of central tendency - Mode

- Score that occurs most frequently
- No formula to calculate the mode
- A distribution may have more than one mode (bimodal or multimodal)
- Advantages
  - Easy to determine
  - Quick estimate of center
  - With normal distributions provides a fair description of central tendency (all three measures are similar)
- Disadvantages
  - Terminal statistic: lack of info that can be used for further calculations
  - Completely disregards extremes

Relationship between the MCT



Guidelines for which to use

Mode: data are approximately normal & you only need a rough estimate

Median: ordinal data, middle score is needed, most typical score is needed, data is badly skewed by extreme scores

Mean: data are approximately normal, interval or ratio data, all available information needs to be considered, further calculations are to be made

Measures of central tendency - `Practice`

It's time to apply what you have learned and calculate/get the `mean`, `median`, and `mode` for the following data set:

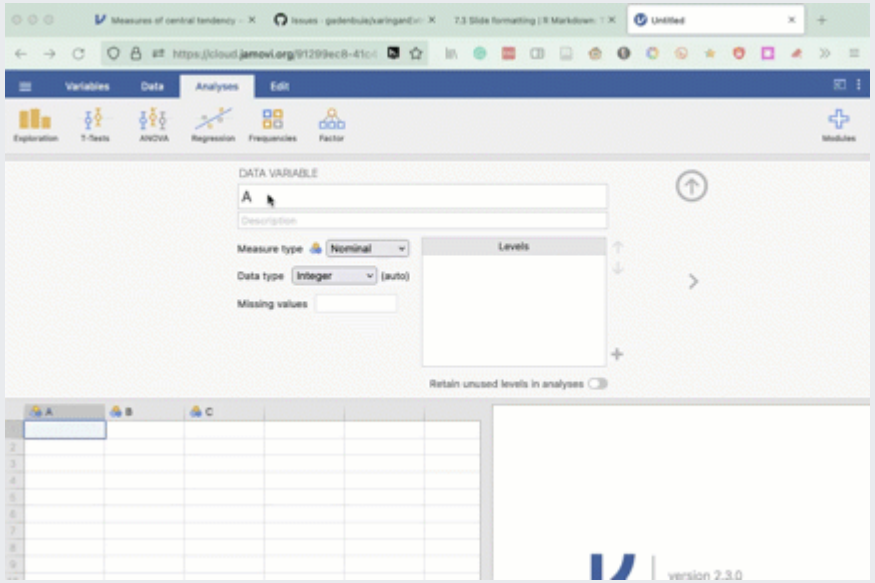
`8, 31, 32, 56, 56`

To do so, open `jamovi` online via https://cloud.jamovi.org

- Then, proceed to type in the scores provided above into column B.
- Next, rename this column to **scores**
- Also rename column A to **ID** and type in values from 1-5
- Finally, delete column C

See an example in the next slide or click here - feel free to zoom in and out

Practice `jamovi`



Follow the instructions on the left to help with this practice exercise

Measure of central tendency - Summary

- Measures of central tendency: Broadly speaking, central tendency measures tell you where the data are. There's three measures that are typically reported in the literature: the mean, median and mode.
- Measures of variability: In contrast, measures of variability tell you about how "spread out" the data are. The key measures are: range, standard deviation, and interquartile range.
- Measures of skewness and kurtosis: We also looked at assymetry in a variable's distribution (skew) and pointness (kurtosis).
- Getting group summaries of variables in jamovi: Since this book focuses on doing data analysis in jamovi, we spent a bit of time talking about how descriptive statistics are computed for different subgroups.
- Standard scores: The z-score is a slightly unusual beast. It's not quite a descriptive statistic, and not quite an inference. Make sure you understand that section. It'll come up again later.

Measures of variability - Overview

1. Why variability is valuable as a descriptive tool
2. How to compute the range, the standard deviation, and the variance
3. How the standard deviation and variance are alike- and how they are different

Measures of variability - Importance & Symbols

- Variability: reflects how scores differ from one another
- Also called spread or dispersion

Measurement	statistics	parameter
Proportion	p	P
Data points	x	X
Mean	$\bar{x}$	μ
Standard deviation	s	σ
Variance	s^2	σ^2
Number of persons/objects	n	N
Correlation coefficient	r	ρ

Measures of variability - What is it?

- A measure of the spread or dispersion of a set of data.
- Builds on some of the fundamental concepts of parametric statistics.
- When we know both central tendency and variability, we can compare sets of data.
- Four measures of variability:
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation

Measures of variability - Range

- Range is the most general estimate of variability
- There are 2 types of range, although the most commonly used is the exclusive range
- General Formula for Range
  - Also known as the Exclusive Range
  - Range = H – L

Note: H is the highest score, L is the lowest score

Measures of variability - Interquartile range

- Often used to analyze ordinal, or ratio data interval data when highly skewed.
- More interested in the middle scores than in the extremes scores
- The difference between the raw scores at the 75th and 25th percentile.
  - Not affected by highly divergent scores at the extremes (useful for skewed data)
- Presents a typical picture, but it does not consider all information about the data.

Calculation: IQR = $Q_3 - Q_1$

Measures of variability - Variance

- Previous two methods consider only 2 points of data to determine variability
- How about the remaining of the data?
  - Distance of each raw score from the mean of the data (deviation)
- Variance is more useful than range and IQR
- Allows us to compare the overall variability of two data sets.

Equation for the population variance

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N}$$

Equation for the sample variance

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

Calculation of Variance

X	$(X-\bar{X})$	$(X-\bar{X})^2$	Y	$(Y-\bar{Y})$	$(Y-\bar{Y})^2$
27	+2	4	35	+10	100
26	+1	1	30	+5	25
25	0	0	25	0	0
24	−1	1	20	−5	25
23	−2	4	15	−10	100
$\Sigma = 125$	$\Sigma = 0$	$\Sigma = 10$	$\Sigma = 125$	$\Sigma = 0$	$\Sigma = 250$

$V_x = 10/5 = 2$

$V_y = 250/5 = 50.$

- Calculating the distance of a raw score from the mean indicates variability (deviation).
- The sum of deviations around the mean will always equal 0.
- Absolute values show variability but cause a loss of information.
- Squaring the deviations is a better method.
- Taking the average of the squared deviations from the mean = variance.

Measures of variability - Standard Deviation

- Calculating variance alone creates a greater magnitude of variability than the raw data alone allows (squaring deviation scores)
- Bring values for the variance **in line with the unit values of the original raw data**
- Standard Deviation = the square root of the variance
  - Estimate of the population standard deviation
- This allows for a score that is **standardized** with the value of the original raw data.

Measures of variability - Standard Deviation

- Most frequently used measure of variability
- SD = s= average amount of variability in a set of scores
- What do these symbols represent?

See equations in the next slide

Equation for the population standard deviation

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

Equation for the sample standard deviation

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

What happens if you *square* the standard deviation?

Measures of variability - Why $n - 1$?

- Samples rarely include the extreme scores.
- Samples almost alwayts have less variability - adjustment needed
- Standard deviation is an estimate of the POPULATION standard deviation
- To make it an “unbiased estimate” you must subtract 1 from n
  - This artificially inflates the SD (it makes it bigger, unbiased) because it makes the denominator smaller

Biased estimates are appropriate if your intent is only to describe the characteristics of the sample. But if you intend to use the sample as an estimate of a population parameter, then it’s best to calculate the unbiased estimate.

Measures of variability - CoV

- What if one needs to compare standard deviations of two data sets that are in different measurement units?

Equation:

$$V = \frac{s}{\bar{x}} \cdot 100\%$$

Example in the next slide

Example

If students completed the push-up and sit-up tests and you want to verify in which test has less variability (is more homogeneous). Below are the results:

Push-up test

- $\bar{x}$ = 25
- s = 5

5/25 = 0.2 * 100 = 2%

$V = \frac{s}{\bar{x}} \cdot 100\%$

Sit-up test

- $\bar{x}$ = 30
- s = 10

10/30 = 0.333 * 100 = 3.33%

$V = \frac{s}{\bar{x}} \cdot 100\%$

So, which of the two tests is more homogeneous (or consistent)?

Measures of variability - final notes

- Standard deviation is computed as the average distance from the mean
- The larger the standard deviation the more spread out the values are
- Like the mean, the standard deviation is sensitive to extreme scores
- If s = 0, then there is no variability among scores and the scores are essentially identical in value
- While the formulas are quite similar…the two are also quite different.

Interpretation

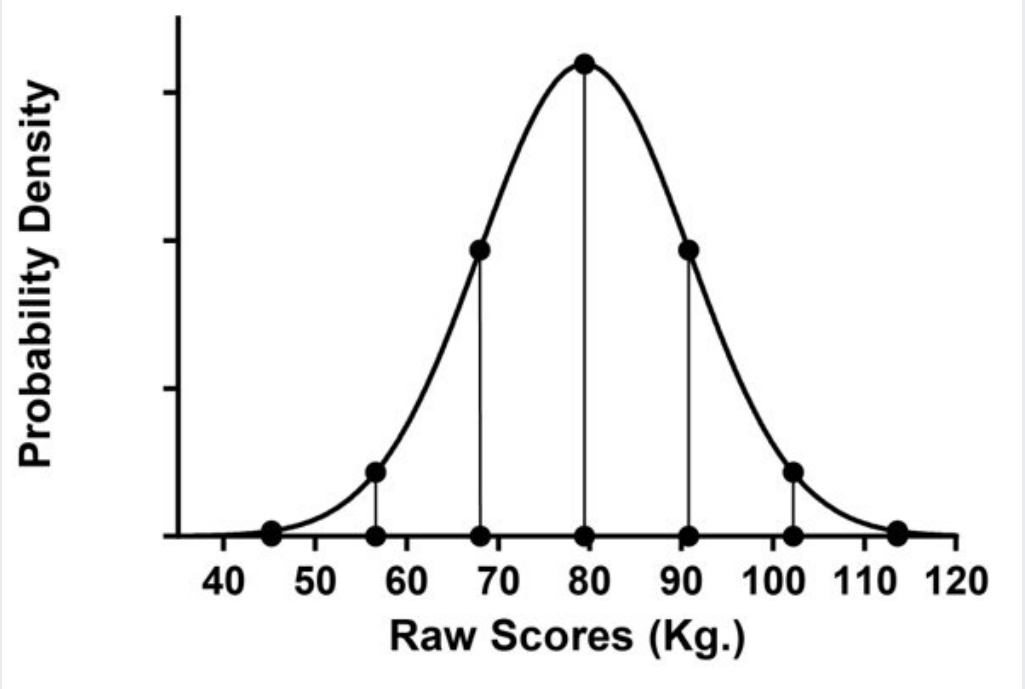
- Standard deviation is stated in original units
- Variance is stated in units that are squared
  - Which do you think is easier to interpret???

Standard scores

- The normal curve plots the score on the X-axis and an index of how common or frequent a score is on the Y-axis
- The standard deviation tells something particularly useful in the context of the normal curve
  - Standardizing scores helps us make sense of scores in relation to others
  - An example of this is Z scores

Z-scores

- Z score—a raw score expressed in standard deviation units
  - 1 SD = 11 kg
  - Mean of data set is 79 kg
  - 91 kg = Z score of +1 since it is 1 standard deviation above the mean
  - 68 kg = Z score of –1



Calculation

Equation:

$$z_i = \frac{X_i - \bar{X}}{\hat{\sigma}}$$

Z-score associated with a raw score of 91

Z = 91-79 / 11 = + 1

Z-score associated with a raw score of 68

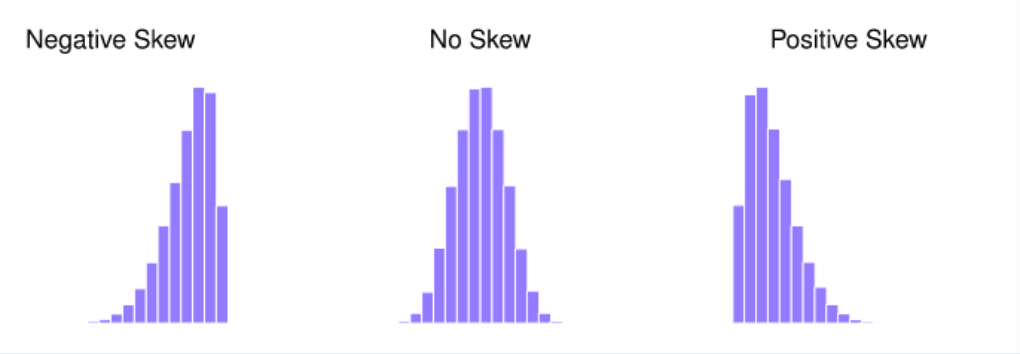
Z = 68-79 / 11 = -1

Z scores and stardard scores

- Z score of 0 is equal to the mean of the distribution
- Z score of 1 is equal to one standard deviation
- The sign of Z score indicates the direction from the mean where – is below the mean and + is above the mean

- Which of the following is a better score in comparison to the population?
  - Z = –1.3 on long jump or
  - Z = –0.50 on gymnastics scores?

Skewness & Kurtosis

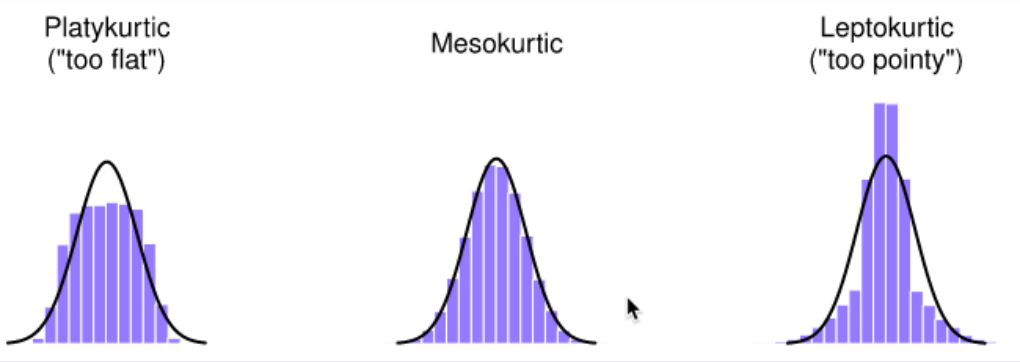


Skewness: is a measure of asymmetry of a distribution curve

$$\text{skewness}(X) = \frac{1}{N\hat{\sigma}^3}\sum_{i=1}^{N}(X_i - \bar{X})^3$$

- Positive SD: more extremely large values than extremely small ones (right panel)
- Negatively SD: data tend to have a lot of extreme small values (i.e., the lower tail is “longer” than the upper tail) and not so many extremely large values (left panel)

Skewness & Kurtosis



Kurtosis: is a measure of the "pointiness" of a data set

informal term	technical name	kurtosis value
"too flat"	platykurtic	negative
"just pointy enough"	mesokurtic	zero
"too pointy"	leptokurtic	positive

$$\text{kurtosis}(X) = \frac{1}{N\hat{\sigma}^4} \sum_{i=1}^{N} \left(X_i - \bar{X}\right)^4 - 3$$

Skewness & Kurtosis - Z-scores

- Z scores used to quantify the amount of skewness and kurtosis in a set of data
- We convert raw skewness and kurtosis scores to Z scores!
- Take the raw skewness and kurtosis scores and divide them by their respective standard errors (a type of standard deviation)
- The numbers **6** in the SE_{skew} equation and **24** in the SE_{kurt} equation, are constants

$$SE_{skew} = \sqrt{\frac{6}{N}} \qquad SE_{kurt} = \sqrt{\frac{24}{N}}$$

Z_{skew} = Skewness / $SE_{skewness}$

Z_{kurt} = Kurtosis / $SE_{kurtosis}$

Skewness & Kurtosis - interpretation

If values of Z_{skew} fall within +- 2.0 z-cores, then the distribution is approximating normality as far as SKEWNESS is concerned.

If values of Z_{kurt} fall within +- 2.0 z-scores, then the distribution of scores is approximating normality as far as KURTOSIS is concerned.

Else, the distribution of scores would be deviating from normality as far as...

> Example in the next slide

Example

$$\text{skewness} = \frac{-9.21}{12} = -0.77,$$

$$SE_{\text{skew}} = \sqrt{\frac{6}{12}} = 0.71,$$

$$Z_{\text{skew}} = \frac{-0.77}{0.71} = -1.08.$$

$$\text{kurtosis} = \frac{29.80}{12} - 3.0 = -0.52,$$

$$SE_{\text{kurt}} = \sqrt{\frac{24}{12}} = 1.41,$$

$$Z_{\text{kurt}} = \frac{-0.52}{1.41} = -0.37.$$

What is the conclusion here?

More jamovi practice

While in class, we will use jamovi to calculate:

- measures of central tendency
- variability
- skewness and kurtosis