Applied Statistics

Descriptive statistics

Cesar O. Aguilar
SUNY Geneseo

Portions of these notes were created from *Learning statistics with R* by Danielle Navarro, *Learning statistics with jamovi* by David Foxcroft, and *Introduction to Statistical Thinking* by Benjamin Yakir.

Descriptive statistics

- **Descriptive statistics** are used to summarize data
- Descriptive statistics is solely concerned with properties of the observed data, and it does not rest on the assumption that the data come from a larger population
- Data is summarized from a **population** using numerical values such as the mean, mode, and/or standard deviation
- To many people "statistics" is synomymous with descriptive statistics

Measures of central tendency

- In most situations, the first thing that you'll want to calculate is a measure of **central tendency**
- That is, you'd like to know something about where the "average" or "middle" of your data lies
- The three most commonly used measures are the **mean**, **median** and **mode**

Mean

- The **mean** is another word for the **average**
- To compute the **mean** we sum all the values in a data set divided by the number of values in the data set
- If $X_1, X_2, X_3, \ldots, X_N$ are the values in a data set, then the mean is given by:

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N}$$

- To make the notation clear, using the AFL data set:

the observation	its symbol	the observed value
winning margin, game 1	X_1	56 points
winning margin, game 2	X_2	31 points
$\vdots$	$\vdots$	$\vdots$
winning margin, game 176	X_{176}	10 points

Mean

- A little bit of notation:

$$\begin{aligned}\bar{X} &= \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N} \\ &= \frac{1}{N}(X_1 + X_2 + X_3 + \ldots + X_N) \\ &= \frac{1}{N}\sum_{i=1}^{N} X_i\end{aligned}$$

- Here X_i is the i-th observation in the data set

Median

- The **median** is the value that divides the data set in half, or the middle value
- Suppose we had the data:

$$56, 31, 56, 8, 32$$

- To find the median, sort the data:

$$8, 31, 32, 56, 56$$

- The median is the middle value, which is 32
- When we have an **odd** number of observations, the median is the middle value of the actual data

Median

- However, say we had the data:

$$8, 31, 31, 32, 56, 56$$

- We now have an **even** number of observations
- To find the median, we take the mean of the two middle values:

$$\frac{31 + 32}{2} = 31.5$$

Mean vs Median

- Which one should we choose?
- Suppose Bob (income $50,000), Kate (income $60,000) and Jane (income $65,000) are sitting at a table
- The average income at the table is $58,333 and the median income is $60,000.
- Then Bill sits down with them (income $100,000,000).
- The average income has now jumped to $25,043,750 but the median rises only to $62,500.
- If you're interested in looking at the overall income at the table, the mean might be the right answer
- But if you're interested in what counts as a typical income at the table, the median would be a better choice here.

Mean vs Median

- If your data are ordinal scale, you're more likely to want to use the median than the mean

- The median only makes use of the order information in your data (i.e., which numbers are bigger), but doesn't depend on the precise numbers involved

- That's exactly the situation that applies when your data are ordinal scale.

- The mean, on the other hand, makes use of the precise numeric values assigned to the observations, so it's not really appropriate for ordinal data

Mean vs Median

- For interval and ratio scale data, either one is generally acceptable
- Which one you pick depends a bit on what you're trying to achieve
- The mean has the advantage that it uses all the information in the data (which is useful when you don't have a lot of data)
- But as we saw with the income example, the mean is very sensitive to extreme values

Real life example: Commonwealth Bank of Australia

- This story appeared in ABC News Australia in September 2010

> *Senior Commonwealth Bank executives have travelled the world in the past couple of weeks with a presentation showing how Australian house prices, and the key price to income ratios, compare favourably with similar countries. "Housing affordability has actually been going sideways for the last five to six years," said Craig James, the chief economist of the bank's trading arm, CommSec.*

- This probably comes as a huge surprise to anyone with a mortgage

> *CBA has waged its war against what it believes are housing doomsayers with graphs, numbers and international comparisons. In its presentation, the bank rejects arguments that Australia's housing is relatively expensive compared to incomes. It says Australia's house price to household income ratio of 5.6 in the major cities, and 4.3 nationwide, is comparable to many other developed nations. It says San Francisco and New York have ratios of 7, Auckland's is 6.7, and Vancouver comes in at 9.3.*

Real life example: Commonwealth Bank of Australia

> *Many analysts say that has led the bank to use misleading figures and comparisons. If you go to page four of CBA's presentation and read the source information at the bottom of the graph and table, you would notice there is an additional source on the international comparison – Demographia. However, if the Commonwealth Bank had also used Demographia's analysis of Australia's house price to income ratio, it would have come up with a figure closer to 9 rather than 5.6 or 4.3*

- One group of people say 9, another says 4-5.
- This is a situation where there is a right answer and a wrong answer
- Demographia are correct, and the Commonwealth Bank is incorrect

Real life example: Commonwealth Bank of Australia

> *[An] obvious problem with the Commonwealth Bank's domestic price to income figures is they compare average incomes with median house prices (unlike the Demographia figures that compare median incomes to median prices). The median is the mid-point, effectively cutting out the highs and lows, and that means the average is generally higher when it comes to incomes and asset prices, because it includes the earnings of Australia's wealthiest people. To put it another way: the Commonwealth Bank's figures count Ralph Norris' multi-million dollar pay packet on the income side, but not his (no doubt) very expensive house in the property price figures, thus understating the house price to income ratio for middle-income Australians.*

- The way that Demographia calculated the ratio is the right thing to do
- The way that the Bank did it is incorrect

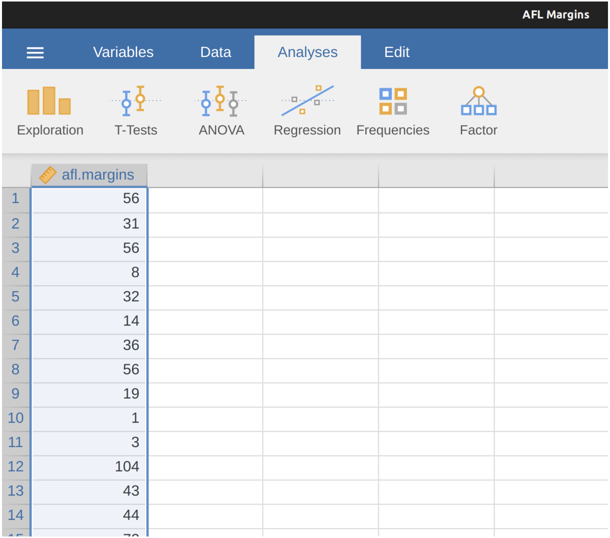
Real life example: Commonwealth Bank of Australia

- How could the bank make such a mistake? Don't smart people work at banks?

> *[As] Australia's largest home lender, the Commonwealth Bank has one of the biggest vested interests in house prices rising. It effectively owns a massive swathe of Australian housing as security for its home loans as well as many small business loans.*

AFL Data

- In jamovi, load the file `aflsmall_margins.csv`



	afl.margins
1	56
2	31
3	56
4	8
5	32
6	14
7	36
8	56
9	19
10	1
11	3
12	104
13	43
14	44

AFL Data

- This file contains data on the margin of victory in the Australian Football League (AFL) for all 2010 games
- Find the mean and mode of the margin of victory data

Mode

- The **mode** is the most frequently occurring value in a data set
- Load the file `afl.finalists.csv` into jamovi
- The file contains the names of all 400 teams that played in all 200 finals matches played during the period 1987 to 2010
- Use the **frequency table** checkbox to produce a frequency table
- What team has played in the most finals matches? How many times?

Measuers of variability

- The mean, median, and mode are all measures of central tendency
- We may also want to summarize the data by describing how spread out the values are, or how much **variability** there is in the data
- Or how far away the values are from the mean
- We will consider the following measures of variability:
  - Range
  - Interquartile range
  - Mean absolute deviation
  - Variance
  - Standard deviation

Range

- The **range** of a data set is the difference between the largest and smallest values
- Although it is the simplest measure of variability, it is not very useful
- Specifically, extreme values in the data set can have a large effect on the range
- For example, consider:

$$-100, 2, 3, 4, 5, 6, 7, 8, 9, 10$$

- The range is

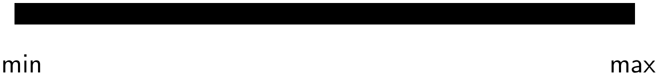
$$10 - (-100) = 110$$

- But if we remove the -100 from the data set, the range is

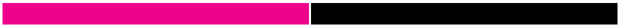
$$10 - 2 = 8$$

Quartiles

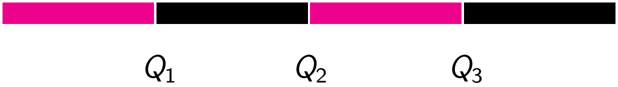
- Before we discuss interquartile range, we first need to discuss **quartiles**
- Quartiles are values that divide a data set into **four** equal parts
- Suppose that our **sorted** data set is represented by this strip:



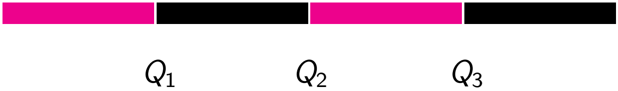
- One way to divide the data set into four equal parts is to first divide it into two equal parts:



- Then divide each of those parts into two equal parts:

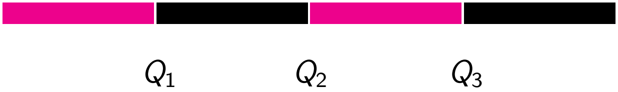


Quartiles



- Q_1 is the **1st quartile**, Q_2 is the **2nd quartile**, and Q_3 is the **3rd quartile**
- Q_1 is the smallest value that is greater than 25% of the values
- Q_2 is the smallest value that is greater than 50% of the values
- Q_3 is the smallest value that is greater than 75% of the values
- These quartiles are special cases of **percentiles**
- For example, the 90th percentile of a data set is the smallest value that is greater than 90% of the values

Quartiles



- Q_1 is the 25th percentile
- Q_2 is the 50th percentile
- Q_3 is the 75th percentile
- Notice that Q_2 is the median, which is found by first sorting the data in ascending order and then finding the middle value
- To find Q_1, we find the median of the lower half of the data set
- To find Q_3, we find the median of the upper half of the data set

Quartiles

Example. Find the quartiles Q_1, Q_2, Q_3 of the following data set:

$$25, 20, 30, 15, 18, 29, 40, 60, 50, 41$$

- First, sort the data set in ascending order:

$$15, 18, 20, 25, 29, 30, 40, 41, 50, 60$$

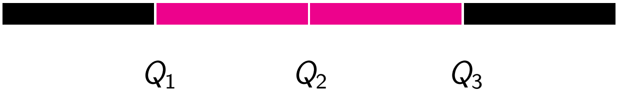
- Next, find the median of the data set: $Q_2 = 29.5$
- Then find the median of the lower half of the data set: $Q_1 = 20$
- Then find the median of the upper half of the data set: $Q_3 = 41$
- The quartiles are $Q_1 = 20$, $Q_2 = 29.5$, and $Q_3 = 41$

Interquartile Range

- Now that we know how to find quartiles, we can define the **interquartile range**:

$$IQR = Q_3 - Q_1$$

- The simplest way to think about IQR is that it is the range spanned by the "middle half" of the data



- In the previous examples, the quartiles were $Q_1 = 20$, $Q_2 = 29.5$, and $Q_3 = 41$, so:

$$IQR = Q_3 - Q_1 = 41 - 20 = 21$$

Mean absolute deviation

- Another approach describing the spread or variability of a data set is to measure “typical” **deviations** from the mean or median

- Given a data set $X_1, X_2, \ldots, X_N$, with mean $\bar{X}$, the **absolute deviation** of X_i from $\bar{X}$ is the number

$$|X_i - \bar{X}|$$

- Here $|x|$ is the absolute value of the number x, for example:

$$|2.7| = 2.7 \qquad |-3.9| = 3.9$$

- The value $|X_i - \bar{X}|$ is the numerical distance between X_i and $\bar{X}$ irrespective of whether X_i is greater than or less than $\bar{X}$

- For example, if $X_i = 22$ and $\bar{X} = 35$, then $|X_i - \bar{X}| = |22 - 35| = 13$

- Whereas if $X_i = 48$ and $\bar{X} = 35$, then $|X_i - \bar{X}| = |48 - 35| = 13$

Mean absolute deviation

- We can calculate the absolute deviation from the mean of each value:

which observation	**value**	**deviation from mean**	**absolute deviation**
i	X_i	$X_i - \bar{X}$	$\lvert X_i - \bar{X} \rvert$
1	56	19.4	19.4
2	31	-5.6	5.6
3	56	19.4	19.4
4	8	-28.6	28.6
5	32	-4.6	4.6

- The **mean absolute deviation** is the average of the absolute deviations from the mean:

$$MAD = \frac{1}{N}\left(|X_1 - \bar{X}| + |X_2 - \bar{X}| + \cdots + |X_N - \bar{X}|\right)$$

- For the data set above, the mean absolute deviation is:

$$MAD = \frac{1}{5}(19.4 + 5.6 + 19.4 + 28.6 + 4.6) = 16.4$$

Variance

- It turns out that from a mathematical perspective, the mean absolute deviation is not the best measure of variability

- Specifically, dealing with absolute values is not very convenient

- Instead of absolute values, we can **square** a difference of two numbers to obtain a measure of deviation

- For example, if $X_i = 22$ and $\bar{X} = 35$, then the squared deviation from the mean is

$$(X_i - \bar{X})^2 = (-13)^2 = 169$$

which observation	value	deviation from mean	squared deviation
i	X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
1	56	19.4	376.36
2	31	-5.6	31.36
3	56	19.4	376.36
4	8	-28.6	817.96
5	32	-4.6	21.16

Variance

- As with the mean absolute deviation, we compute the average of all squared deviations from the mean $\bar{X}$, which is called the **sample variance** usually denoted by $Var(X)$ but more commonly by s^2

- Thus, the sample variance is defined as:

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

- However, for reasons that will be made clear later when we discuss inferential statistics, instead of dividing by N we divide by $N-1$ to obtain the **unbiased sample variance**:

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

- To a lot of people, this is what the variance is

Variance

- Unfortunately, the variance is not a very intuitive measure of variability
- This is because the units of the variance are the **square** of the units of the original data
- For example, if the original data is in dollars, then the variance is in dollars squared
- Nonetheless, the variance has some elegant mathematical properties that suggest that it really is a fundamental quantity for expressing variation

Example: Variance

- Consider the data below:

which observation	value	deviation from mean	squared deviation
i	X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
1	56	19.4	376.36
2	31	-5.6	31.36
3	56	19.4	376.36
4	8	-28.6	817.96
5	32	-4.6	21.16

- The mean is $\bar{X} = \frac{1}{5}(56 + 31 + 56 + 8 + 32) = 36.6$
- The sample variance is:

$$s^2 = \frac{1}{5}(376.36 + 31.36 + 376.36 + 817.96 + 21.16) = 324.64$$

- And the unbiased sample variance is:

$$\hat{\sigma}^2 = \frac{1}{4}(376.36 + 31.36 + 376.36 + 817.96 + 21.16) = 405.8$$

Standard deviation

- If we take the square root of the variance we obtain the **standard deviation** which is usually denoted by s

- Thus, the standard deviation is defined as:

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

- s is also commonly called the **root mean square deviation** (RMSD)

- However, as with the variance, we divide by $N-1$ to obtain the **unbiased standard deviation**:

$$\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

- When we refer to the standard deviation, we are usually referring to the unbiased standard deviation $\hat{\sigma}$

Standard deviation

- Although the standard deviation is a measure of variability, it doesn't have a simple interpretation like say mean absolute deviation
- However, in many situations, we can expect that
    - 68% of the data to fall within 1 standard deviation of the mean
    - 95% of the data to fall within 2 standard deviation of the mean
    - 99.7% of the data to fall within 3 standard deviations of the mean

Example: Standard deviation

which observation	value	deviation from mean	squared deviation
i	X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
1	56	19.4	376.36
2	31	-5.6	31.36
3	56	19.4	376.36
4	8	-28.6	817.96
5	32	-4.6	21.16

- We computed the sample variance to be $s^2 = 324.64$ and thus the sample standard deviation is

$$s = \sqrt{324.64} = 18.02$$

- We computed the unbiased sample variance to be $\hat{\sigma}^2 = 405.8$ and thus the unbiased sample standard deviation is

$$\hat{\sigma} = \sqrt{405.8} = 20.14$$

Which measure of variability should we use?

- **Range**:
  - Gives you the full spread of the data
  - It's very vulnerable to outliers
  - It isn't often used unless you have good reasons to care about the extremes in the data

- **Interquartile range** (IQR):
  - Tells you where the "middle half" of the data sits
  - It's pretty robust, and complements the median nicely
  - This is used a lot

- **Mean absolute deviation**.
  - Tells you how far "on average" the observations are from the mean
  - It's very interpretable, but has a few minor issues that make it less attractive to statisticians than the standard deviation
  - Not used often

Which measure of variability should we use?

- **Variance**:
  - Tells you the average squared deviation from the mean
  - It's mathematically elegant (linear operator), and is probably the "right" way to describe variation around the mean
  - It's completely uninterpretable because it doesn't use the same units as the data
  - Almost never used except as a mathematical tool
- **Standard deviation**:
  - This is the square root of the variance
  - It's fairly elegant mathematically, and it's expressed in the same units as the data so it can be interpreted pretty well
  - In situations where the mean is the measure of central tendency, this is the default
  - This is by far the most popular measure of variation.

Skew

- This is another descriptive statistic and is used to describe the shape of a data set
- We'll talk about this in more detail in a later lecture after we have covered histograms

Standard scores

- Suppose we have a data set $X_1, X_2, \ldots, X_N$, with mean $\bar{X}$ and standard deviation $\hat{\sigma}$

- The **standard score** (commonly called *z*-**score**) of X_i is the number of standard deviations that X_i is from the mean:

$$\text{standard score} = \frac{\text{raw value} - \text{mean}}{\text{standard deviation}}$$

- In mathematical notation:

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

- Who cares?

- It is a way to standardize the data and eliminate the units

- When two (or more) data sets are measured on different scales, they may be converted to z-scores to aid comparison

Example: Standard scores

- Suppose that two groups of students take distinct tests, the SAT and ACT
- The SAT scores have mean $\bar{X} = 1500$ and standard deviation $\hat{\sigma}_X = 300$
- The ACT scores have mean $\bar{Y} = 21$ and standard deviation $\hat{\sigma}_Y = 5$
- Suppose that student A scored 1700 on the SAT, and student B scored 25 on the ACT
- Which student performed better relative to other test-takers?
- To compare, we need to convert the scores to z-scores
- The z-scores are

$$z_A = \frac{1700 - 1500}{300} = 0.\overline{666} \qquad z_B = \frac{25 - 21}{5} = 0.8$$

- Student B performed better compared to other test-takers than did student A