Chapter 4: Average and standard deviation

Context

- We are looking at summarizing data
  - We saw that the histogram is a nice graph to summarize data
  - Sometimes we want to summarize even further, by just giving two numbers that describe the center and spread of the data
    - for the center of the data we use:
      average (=mean), median
    - for the spread of the data we use:
      standard deviation, interquartile range, range
- Other new terminology: r.m.s. size, cross-sectional studies, longitudinal studies

2 / 21

Average vs. median

3 / 21

Average

- $$\text{Average of a list} = \frac{\text{sum of entries}}{\text{number of entries}}$$
- Average of a histogram: imagine that the histogram consists of blocks on stiff weightless board. This average is the balance point of the histogram.
- See examples on overhead (Figure 6 on page 63)

4 / 21

Median

- Median is a number such that 50% of the data is smaller than this number, and 50% of the data is larger than this number
  - Median of a list: order the entries from small to large, and take the middle entry. (If the list has an even number of entries, you can take the average of the middle two entries)
  - Median of a histogram: find the point on the horizontal axis, so that half the area of the histogram is to the left, and half the area is to the right of that point. (You often have to guess this)
- See examples on overhead

5 / 21

Average vs. median

- The average and median are both measures of the center of a distribution
- What are the differences?
- The average is more affected by the tail of the distribution. The average is pulled towards the tail. See overhead
- Three cases:
  - Symmetric: average is about the same as the median
  - Long left tail: average is smaller than the median
  - Long right tail: average is larger than the median
- When to use which one?
  - If the distribution is symmetric, it does not matter much
  - If the distribution is asymmetric, the median is often a better description of the center. Sometimes people give both.

6 / 21

r.m.s. size

7 / 21

Determine the 'size' of numbers?

- List: 1, -3, 5, -6, 3
- We want to know the size of these numbers - some measure of how far they are from zero
- $$\text{Average} = \frac{1 + (-3) + 5 + (-6) + 3}{5} = \frac{0}{5} = 0$$
- So the average is not a good measure of the size of these numbers. Why? The positives cancel the negatives.
- So we want to get rid of the negative signs. How?

8 / 21

Getting rid of the signs

- List: 1, -3, 5, -6, 3
- Option (a): We wipe out the signs and then take the average:

$$\frac{1+3+5+6+3}{5} = \frac{18}{5} = 3.6$$

- Option (b): We square the entries, then take the average, and then take the square root again:

$$\sqrt{\frac{1^2+(-3)^2+5^2+(-6)^2+3^2}{5}} = \sqrt{\frac{80}{5}} = \sqrt{16} = 4$$

9 / 21

Getting rid of the signs

- Both options give about the same answer, and are reasonable
- Option (a) seems easier to compute
- For option (b) the math works out nicer. So we will use option (b).
- Option (b) is called r.m.s. size, and r.m.s. stands for "root-mean-square": root of the mean (=average) of the squares

10 / 21

Examples

- Is the r.m.s. size of the following lists around 1, 10 or 20?
  - List: 1, 5, -7, 8, 10, 9, -6, 5, 12, -17
  - List: 22, -18, -33, 7, 31, -12, 1, 24, -6, 16
  - List: 1, 2, 0, 0, -1, 0, 0, -3, 0, 1
- Find the r.m.s. size of the following lists:
  - List: 7, 7, 7, 7
  - List: 7, -7, 7, -7

11 / 21

Interpretation

- The standard deviation (SD) is a measure of spread:
  - If the data are spread out widely, then the SD is large
  - If the data are close together, then the SD is small
- The SD is a measure for how far numbers are from their average.
- Most entries on the list are about 1 SD away from the average. Very few are more than 2 or 3 SDs away.

13 / 21

68% and 95% rules

- Rule of thumb that holds for many lists (but not all):
  - About 68% (2 out of 3) of the entries on a list are within 1 SD from the average. The other 32% are farther away.
  - About 95% (19 out of 20) of the entries on a list are within 2 SDs from the average. The other 5% are farther away.
  - See overhead for pictures

14 / 21

Examples

- Take a group of 11 year old boys. The average height is 146 cm, and the SD is 8 cm.
  - One boy was 170 cm tall. He was above average by ... SDs.
  - Another boy was 1.5 SDs below average in height. He was ... cm tall.
  - Here are the heights of four boys:
    150cm, 130cm, 165cm, 140cm.
    For each boy, say whether his height was unusually short, about average, or unusually high.
  - About what percentage of the boys had heights between 138cm and 154cm?
  - About what percentage of the boys had heights between 130cm and 162cm?

15 / 21

Computing the SD

- SD = r.m.s. size of deviations from average:
  - Step 1: compute the average of the list
  - Step 2: compute the deviations from the average (deviation from average = entry - average)
  - Step 3: compute the r.m.s. size of the deviations

16 / 21

Example

- Compute the SD of the following list: 1, 3, 4, 5, 7
  - Step 1: compute the average of the list:

$$Ave = \frac{1+3+4+5+7}{5} = \frac{20}{5} = 4$$

  - Step 2: compute the deviations from the average:
    -3, -1, 0, 1, 3
  - Step 3: compute the r.m.s. size of the deviations (take the root of the mean of the squares:

$$\sqrt{\frac{(-3)^2+(-1)^2+0^2+1^2+3^2}{5}} = \sqrt{\frac{20}{5}} = \sqrt{4} = 2$$

- The SD of the list is 2.

17 / 21

Cross-sectional vs. longitudinal

18 / 21

Example: HANES

- HANES: Health And Nutrition Examination Survey (1976-1980)
- They examined a representative cross-section of 20,322 Americans age 1 to 74.
- They got data about:
  - demographic variables: age, education, income
  - physiological variables: height, weight, blood pressure, cholesterol levels
  - dietary habits
  - levels of lead and pesticides in the blood
  - prevalence of diseases
- We look at height and weight, see Figure 3 on page 59

19 / 21

Definitions

- Cross-sectional: different subjects are compared to each other at one point in time
- Longitudinal: subjects are followed over time, and compared with themselves at different points in time
- When interpreting effect of age, be careful to know what type of study it is!

20 / 21

Statistical calculator

- We compute the SD as

$$\sqrt{\frac{\text{sum of the squared deviations from the average}}{\text{nr. of entries}}}$$

- Some calculators compute the SD as

$$\sqrt{\frac{\text{sum of the squared deviations from the average}}{\text{nr. of entries} - 1}}$$

- The difference between these two methods is usually small.

21 / 21