Mathematics and Biostatistics

Distribution of Values - 2

1st Semester 2021

Lecture 4

Review

- Distribution of Values
- Frequency table
- Frequency distribution
- Relative Frequency distribution
- Things we look for in a distribution
    1. Central Tendency
        a. MEAN -- average
        b. MEDIAN -- middle value
        c. MODE -- most frequently observed value.
    2. Variability of values and their spread
    3. Shape of the distribution.
    4. Gaps and clumping of values

Variability of values and their spread

- The concept of dispersion or variability has to do with how spread out and different the values are.
- If the values are close in value or exactly the same there is little dispersion. If there are lots of different values and they are spread over a wide interval, then we have greater levels of dispersion or variability.
- The simplest useful numerical description of a distribution consists of both a measure of center and a measure of spread

Measuring spread: the quartiles

- We can describe the spread or variability of a distribution by giving several

percentiles.

- The median divides the data in two; half of the observations are above the median and half are below the median.
- We could call the median the quartile 50th percentile.
- The upper quartile is the median of the upper half of the data.
- Similarly, the lower quartile is the median of the lower half of the data.
- With the median, the quartiles divide the data into four equal parts; 25% of the data are in each part.

The quartiles Q1 and Q3

To calculate the quartiles:

1. Arrange the observations in increasing order and locate the median M in the ordered list of observations.

2. The first quartile Q1 is the median of the observations whose position in the ordered list is to the left of the location of the overall median.

3. The third quartile Q3 is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

Find the quartiles.

- Here are the scores on the first-exam in an introductory statistics course for 10 students:

80 73 92 85 75 98 93 55 80 90

- Find the quartiles for these first-exam scores.

1. M=? $\longrightarrow$ $55\ 73\ 75\ 80\ \boxed{80\ 85}\ 90\ 92\ 93\ 98 \longrightarrow (80+85)/2 = 82.5$

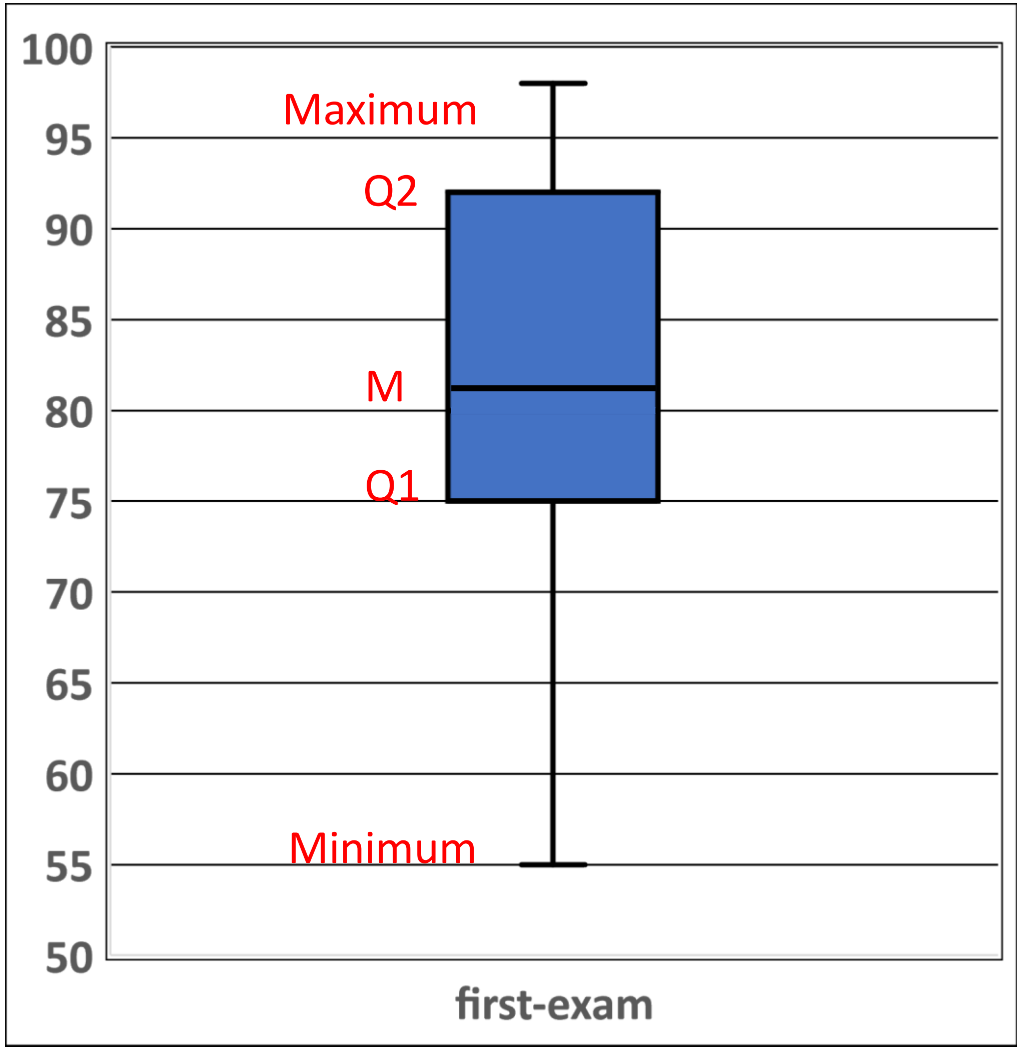
2. Q1 , Q3 $\longrightarrow$ $\underbrace{55\ 73\ 75\ 80\ 80}_{75} \mid \underbrace{85\ 90\ 92\ 93\ 98}_{92}$

The five-number summary and boxplots

The five-number summary of a set of observations consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest.

In symbols, the five-number summary is

1. Minimum
2. Q1
3. M
4. Q3
5. Maximum



Measuring spread: the standard deviation

- The standard deviation measures spread by looking at how far the observations are from their mean.
- The standard deviation s is the square root of the variance s^2:
- $s = \sqrt{\frac{1}{n-1}\sum (x_i - \overline{x})^2}$, $s^2 = \frac{1}{n-1}\sum (x_i - \overline{x})^2$
- Less compact notation

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

Measuring spread: the standard deviation

- The idea behind the variance and the standard deviation as measures of spread is as follows:

1. The deviations $x_i - \bar{x}$ display the spread of the values x_i about their mean $\bar{x}$.
2. Some of these deviations will be positive and some negative because some of the observations fall on each side of the mean.
3. In fact, the sum of the deviations of the observations from their mean will always be zero.
4. Squaring the deviations makes them all positive, so that observations far from the mean in either direction have large positive squared deviations.
5. The variance is the average squared deviation.
6. Therefore, s^2 and s will be large if the observations are widely spread about their mean, and small if the observations are all close to the mean.

Measuring spread: the standard deviation

- A person’s metabolic rate is the rate at which the body consumes energy. Metabolic rate is important in studies of weight gain, dieting, and exercise. Here are the metabolic rates of 7 men who took part in a study of dieting. (The units are calories per 24 hours. These are the same calories used to describe the energy content of foods.)

  1792 1666 1362 1614 1460 1867 1439

- Find the variance and the standard deviation?

Measuring spread: the standard deviation

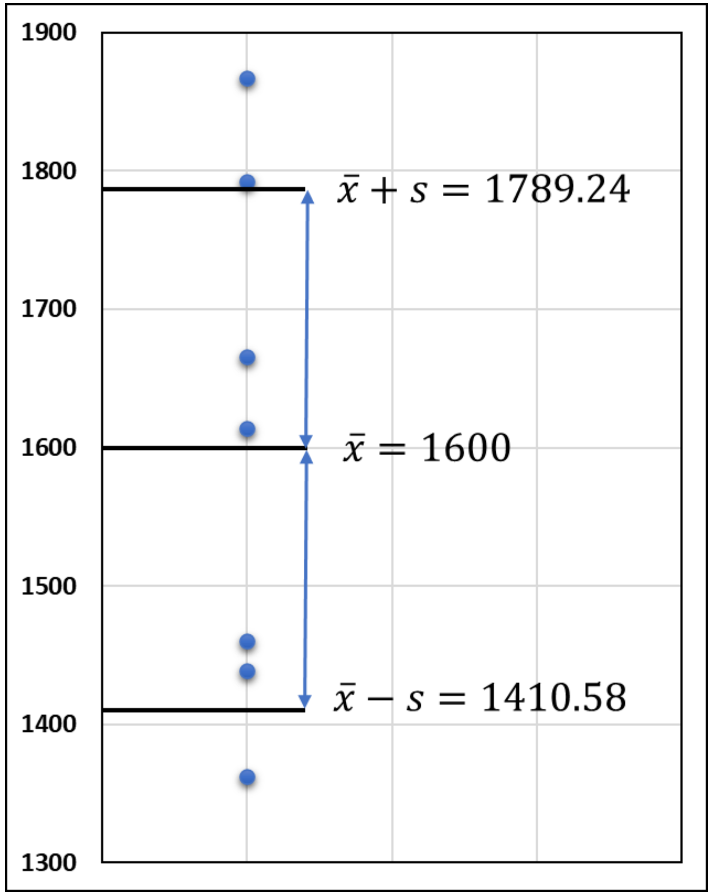
- Variance $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2, \quad n = 7$
- Mean $\bar{x} = \frac{1792+1666+1362+1614+1460+1867+1439}{7} = \frac{11200}{7} = 1600$

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1792	192	36864
1666	66	4356
1362	-238	56644
1614	14	196
1460	-140	19600
1867	267	71289
1439	-161	25921

$$\sum(x_i - \bar{x})^2 = 214870$$

$$s^2 = \frac{214870}{6} = 35811.67$$

$$s = \sqrt{35811.67} = 189.24$$



Properties of the standard deviation

- s measures spread about the mean and should be used only when the mean is chosen as the measure of center.
- $s = 0$ only when there is no spread. This happens only when all observations have the same value.
- Otherwise, $s > 0$. As the observations become more spread out about their mean, s gets larger.
- s, like the mean $\bar{x}$, is not resistant. A few outliers can make s very large.

The End of Lecture