

## Lecture 2 and Lecture 3

We can describe distributions using 3 characteristics: shape, center and spread.

- These characteristics have been discussed since the foundation of statistics.
- Shape can be determined from pictures of data (like histograms).

We discuss center and spread next.

**Definition 1.** A **parameter** is a descriptive measure of a population. A **statistic** is a descriptive measure of a sample.

- Choice for the spread drives choice for the center.

**Definition 2.** The **arithmetic mean** of a variable is computed by determining the sum of all the values of the variable in the data set, divided by the number of observations. In other words, if  $x_1, x_2, \dots, x_n$  are  $n$  observations of a variable from a population (sample), then the population (sample) arithmetic mean is denoted by  $\mu$  (respectively  $\bar{x}$  for a sample) and is given by:

$$\mu(\text{or } \bar{x}) = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}.$$

- Strength: Calculable for all data, unique, can combine (knowing the mean of several samples and their size one could compute the mean of the combined sample)
- Weaknesses: Easily contaminated, sensitive to extreme values: not resistant or robust

## The median

**Definition 3.** The **median** of a variable is the value that lies in the middle of the data when arranged in ascending order. Usually we denote the median by  $M$ .

*Remark* Half the data are below the median and half are above the median.

### How do we compute the median?

- a) Arrange the data in ascending order and find out the number of observations  $n$ .
- b)• If  $n$  is odd then there is a middle value and that is the median. It will be on the  $\frac{n+1}{2}$  position.
  - If  $n$  is even there is no middle value and then the median is computed by computing the arithmetic mean of the  $\frac{n}{2}$  and  $\frac{n}{2} + 1$  value.

- Advantage: robustness/resistance, better measure of center for nonsymmetric distributions.
- Disadvantages: More difficult to compute, does not combine, not always unique (if  $n$  is even anything between the 2 middle observations could be a median).

## The mode

**Definition 4.** The **mode** of a variable is the most frequent observation of the variable that occurs in the data set.

*Remark:* If there is no observation that occurs with most frequency, we say the data has **no mode**. If there are two observations that are equally frequent and they are the largest then we say that the data is **bimodal**. The data could be **trimodal**, etc.

- Advantages: usable for qualitative data, no computation required.
- Disadvantages: does not combine, horrible mathematical properties.

Example: Compute the mean, median, and mode for the number of children in the households in my building:

6,2,1,0,3,1,4,1,2,3

In general, answer the questions:

Question 1) If we take a value from the right of the median and move to the right do we change the mode? Do we change the mean?

Question 2) What if we take a value from the left of the median and we move it to the right of the median?

Question 3) What is the position of the mean relative to the median in the different types of distributions? Where is the mode?

## Measures of Spread

### The range

Range =  $R$  = Largest Data Value – Smallest Data Value

- Advantages: easy to compute, useful in absolute terms
- Disadvantages: uses only 2 points, hence there is no intermediate information.

## Percentiles

Let  $k$  be a number between 0 and 100. The  **$k$ -th percentile** of a data set is that value that separates the lower  $k\%$  of the data from the upper  $(100 - k)\%$  of the data.

The  $k$ -th percentile is denoted by  $P_k$ .

### How to determine the $k$ -th percentile, $P_k$ .

1. Sort: Arrange the data in ascending order (smallest to largest).
2. If  $\frac{k}{100}n$  is an integer, take  $i$  to be the next integer; if  $\frac{k}{100}n$  is not an integer, then round **up** to the next integer (round up  $\neq$  round off).

## How to find the percentile that corresponds to a data value

1. Sort: Arrange the data in ascending order.
2. Compute:

$$\text{percentile of } x = \frac{\text{number of values less than } X}{n} \times 100.$$

3. Round off the percentile to the nearest integer.

### Example

This section has 18 students. Suppose only 3 students get better scores than you in the final exam. At what percentile is your final-exam score?

## Quartiles

The Quartiles are three specific quantiles that are used often:

First Quartile:  $Q_1 = P_{25}$ .

Second Quartile:  $Q_2 = P_{50}$  (this is the median).

Third Quartile:  $Q_3 = P_{75}$ .

The quartiles divide the data in 4 pieces with (approximately) the same number of observations.

To compute a quartile, compute the corresponding percentile, or compute the median of the whole data and then the median of the first half of the data and the median of the second half of the data.

## Outliers

An **outlier** is a data value that does not follow the pattern of the other observations. Usually it is too large or too small, compared with the other values.

An outlier can occur because of an error in measurement or recording, or it can be a true (but unusually extreme) value. In either case, it can affect our inferences:

- If you use the mean and standard deviation: **Danger**, these are affected by outliers.
- If you use the median and the **interquartile range**: Less danger, these are **resistant** to outliers (or **robust**).

The **interquartile range** or **IQR** is computed by:

$$\text{IQR} = Q_3 - Q_1.$$

The IQR is an alternative to the standard deviation as a measure of spread.

- If the data is approx. normal (symmetric, no outliers): the standard deviation is better (more information from less data).
- If the data is not normal (asymmetric, outliers), the IQR is better (more **robust**).

## Checking for Outliers

1. Determine  $Q_1, Q_3$ .
2. Compute  $IQR = Q_3 - Q_1$ .
3. Find the fences:

$$\text{Lower Fence} = Q_1 - 1.5(IQR)$$

$$\text{Upper Fence} = Q_3 + 1.5(IQR)$$

4. A value below the lower fence or above the upper fence is considered an outlier.

## The Five-Number Summary

A quick way to get an idea of how the data values are distributed is to use the **five-number summary**:

Minimum	$Q_1$	$M$	$Q_3$	Maximum
---------	-------	-----	-------	---------

With it, we can quickly assess:

- Center ( $M$ )
- Spread (IRQ, range)
- Symmetry
- Existence of outliers.

## Boxplots

These are the graphical counterpart of the five-number summary, and are equally useful.

### Drawing a boxplot (horizontally)

1. Find the five-number summary
2. Compute the lower and upper fences:

$$\text{Lower Fence} = Q_1 - 1.5(\text{IQR})$$

$$\text{Upper Fence} = Q_3 + 1.5(\text{IQR})$$

3. Draw short vertical lines at  $Q_1$ ,  $M$ , and  $Q_3$ ; enclose them in a box.
4. Draw shorter vertical lines (or brackets) at the fences.
5. Draw a line from  $Q_1$  to the smallest data value inside the fences, and a line from  $Q_3$  to the largest data value inside the fences.
6. Mark with an asterisk ( $\star$ ) any values outside the fences.

Example: Construct the boxplot of the heights  
64, 67, 68, 69, 70, 70, 71, 71, 73

## The Population Variance

The *population variance*, denoted  $\sigma^2$ , is the sum of the squared deviations about the population mean divided by the number of observations in the population,  $N$ :

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}.$$

Another alternative formula is:

$$\sigma^2 = \frac{\sum x_i^2}{N} - \left( \frac{\sum x_i}{N} \right)^2 = \frac{\sum x_i^2}{N} - \mu^2.$$

REMARK: To avoid round-off errors, which accumulate quickly in these formulas, do not round until the last computation, and use as many decimal places as allowed in your calculator.

## The Sample Variance

When the population is large, we approximate the population mean  $\mu$  with the sample mean,  $\bar{x}$ .

Similarly, we approximate the population variance  $\sigma^2$  by the **sample variance**, denoted  $s^2$ :

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

The alternative form is:

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

REMARK: Notice that we divide by the sample size **minus one** (this is different from the formula for the population variance).

Informally, we say: a sample of size  $n$  has  $n$  degrees of freedom; one degree of freedom is “used up” in computing  $\bar{x}$ , so there are only  $n - 1$  degrees of freedom available for the sample variance.

## The Standard Deviation

For both cases (the population or the sample), the **standard deviation** is the square root of the corresponding variance:

The **population standard deviation** is denoted by  $\sigma$ :

$$\sigma = \sqrt{\sigma^2}.$$

The **sample standard deviation** is denoted by  $s$ :

$$s = \sqrt{s^2}.$$

Advantage of the (population or sample) standard deviation: it is given in the same units as the observations.

Advantage of the (population or sample) variance: it is easier to manipulate algebraically, in some cases.

Both the standard deviations and variances are interpreted as follows: the larger they are, the more spread is the distribution (if they equal 0, the smallest possible value, then all observations must be equal).

*Remark 5.* Standard deviation measures spread about the mean and should be used only when the mean is chosen as the measure of center.

*Remark 6.* Standard deviation is not robust.

*Remark 7.* The sum of the deviations of the observations from their mean will always be zero.

## Linear Transformations

- A linear transformation shifts the intercept and/or the slope of the original variable  $x$  into the new variable  $x_{new}$  by:

$$x_{new} = a + bx$$

*Remark 8.* Adding a constant changes the measure of center, but not the spread.

*Remark 9.* Multiplying by a constant changes both, measures of center tendency, and measure of spread.

- $\text{Mean}(a+bx) = a + b\text{Mean}(x)$ .
- $\text{Variance}(a+bx) = b^2 \text{Var}(x)$