Statistics

Lecture 06: | Summary Statistics

Prof. Dr. Mostafa Elhosseini

https://youtube.com/drmelhosseini

Summary Statistics

- A **sample** is often a **long list of numbers**.
- To help make the **important features** of a sample stand out, we compute **summary statistics**.
- The two **most used** summary statistics are the **sample mean and the sample standard deviation**.

Sample Mean

- The **mean** gives an indication of the **center of the data**
- The **sample mean** is also called the "**arithmetic mean**," or, more simply, the "**average**."
- It is the **sum of the numbers** in the sample, **divided** by how many there are

Let $X_1, \ldots, X_n$ be a sample. The **sample mean** is

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Example

- A simple random sample of five men is chosen from a large population of men, and their heights are measured. The five heights (in inches) are **65.51, 72.30, 68.31, 67.05, and 70.68**. Find the sample **mean**.

$$\boxed{\overline{X} = \frac{1}{5}(65.51 + 72.30 + 68.31 + 67.05 + 70.68) = 68.77 \text{ in.}}$$

The Standard Deviation

- **Standard deviation** gives an **indication of how spread out the data are**
- Here are two lists of numbers: **28, 29, 30, 31, 32** and **10, 20, 30, 40, 50**. Both lists **have the same mean of 30**. But clearly the lists **differ in an important way** that is **not captured by the me**an: the second list is much **more spread out than the first**.
- The **standard deviation** is a quantity that **measures the degree of spread** in a sample
- The **basic idea** behind the standard deviation is that when the **spread is large**, the **sample values will tend to be far from their mean**, but when the spread is small, the values will tend to be close to their mean

The Standard Deviation

- The **sample variance** is the average of the squared deviations, except that we divide by $n - 1$ instead of n. It is customary to denote the sample variance by s^2

Let $X_1, \ldots, X_n$ be a sample. The **sample variance** is the quantity

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

An equivalent formula, which can be easier to compute, is

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)$$

- Its **units** are not the same as the units of the sample values; instead, they are the **squared units**

Sample Standard Deviation

- To obtain a **measure of spread** whose **units** are the same as those of the **sample values**, we simply take the **square root of the variance**. This quantity is known as the **sample standard deviation**

Let $X_1, \ldots, X_n$ be a sample. The **sample standard deviation** is the quantity

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{1.4}$$

An equivalent formula, which can be easier to compute, is

$$s = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2 - n\overline{X}^2\right)}$$

The Standard Deviation

- It is natural to wonder why the sum of the squared deviations is divided by n − 1 rather than n.
- The **purpose** in computing the **sample standard deviation** is to **estimate** the amount of **spread in the population** from which the sample was drawn
- the **population mean** is in general **unknown**, so the **sample mean** is **used** in its place.
- It is a mathematical fact that the **deviations around the sample mean** tend to be a bit **smaller** than the **deviations around the population mean** and that dividing by n − 1 rather than n provides exactly the right correction

The Standard Deviation

- A **simple random sample** of five men is chosen from a large population of men, and their heights are measured. The five heights (in inches) are **65.51, 72.30, 68.31, 67.05, and 70.68**.
- Find the sample variance and the sample standard deviation for the height data

The Standard Deviation

We'll first compute the sample variance by using Equation (1.2). The sample mean is $\overline{X} = 68.77$ (see Example 1.9). The sample variance is therefore

$$s^2 = \frac{1}{4}[(65.51 - 68.77)^2 + (72.30 - 68.77)^2 + (68.31 - 68.77)^2 + (67.05 - 68.77)^2 + (70.68 - 68.77)^2] = 7.47665$$

Alternatively, we can use Equation (1.3):

$$s^2 = \frac{1}{4}[65.51^2 + 72.30^2 + 68.31^2 + 67.05^2 + 70.68^2 - 5(68.77^2)] = 7.47665$$

The sample standard deviation is the square root of the sample variance:

$$s = \sqrt{7.47665} = 2.73$$

The Standard Deviation

- **What would happen** to the sample **mean**, **variance**, and **standard deviation** if the heights were measured in **centimeters** rather than **inches**?
- if we multiply each sample item by a constant, the sample mean is multiplied by the same constant
- if we multiply each sample item by a constant, the sample standard deviation is multiplied by the same constant

The Standard Deviation

- **What would happen** to the sample **mean**, **variance**, and **standard deviation** if the heights were measured in **centimeters** rather than **inches**?
- if we multiply each sample item by a constant, the sample mean is multiplied by the same constant
- if we multiply each sample item by a constant, the sample standard deviation is multiplied by the same constant

The Standard Deviation

- What if each man in the sample put on 2-inch heels? Then each sample height would increase by 2 inches and the sample mean would increase by 2 inches as well. In general, if a constant is added to each sample item, the sample mean increases (or decreases) by the same constant.
- The deviations, however, do not change, so the sample variance and standard deviation are unaffected

The Standard Deviation

Summary

- If $X_1, \ldots, X_n$ is a sample and $Y_i = a + bX_i$, where a and b are constants, then $\overline{Y} = a + b\overline{X}$.
- If $X_1, \ldots, X_n$ is a sample and $Y_i = a + bX_i$, where a and b are constants, then $s_Y^2 = b^2 s_X^2$, and $s_Y = |b| s_X$.