Properties of the Sample Mean

Valeria Vitelli
Oslo Centre for Biostatistics and Epidemiology
Department of Biostatistics, UiO
valeria.vitelli@medisin.uio.no

MF9130E – Introductory Course in Statistics
10.04.2024

Central Measures

Mean

- The (arithmetic) **sample mean** $\bar{X}$ is the sum of all observations divided by the number of observations:

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

  where n is the sample size
- It is an estimate of the **population mean** μ

Median

- Another central measure is the **sample median** $\tilde{X}$. This is the *middle observation* when all observations are arranged in increasing order:

$$\tilde{X} = \begin{cases} Y_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{1}{2}(Y_{n/2} + Y_{n/2+1}) & \text{if } n \text{ is even} \end{cases}$$

where $Y_{(1)}, \ldots, Y_{(n)}$ are the ascending ordered observations $X_1, \ldots, X_n$, and n is the sample size

Mode

- The **mode** is the most frequently occuring value in the sample

Example: 4.1 in Kirkwood & Sterne

We have measurements of the plasma volumes (in litres) of eight healthy adult males.

Subject	1	2	3	4	5	6	7	8
Plasma volume	2.75	2.86	3.37	2.76	2.62	3.49	3.05	3.12

We find that the **sample mean** is given by

$$\bar{X} = \frac{1}{8}(2.75 + 2.86 + \ldots + 3.12) = 3.00,$$

and the **sample median** is given by

$$\tilde{X} = \frac{1}{2}(2.86 + 3.05) = 2.96$$

Since all the values are different, there is no estimate of the **mode**

Choice of measure

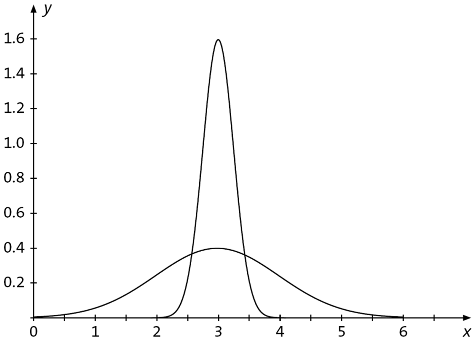
- The choice of measure **depends on the data distribution**

Central measure	**Data distribution**
sample mean	symmetric, normal-like
median	outliers, skewed distribution
mode	seldom used

- The mean, median and mode are equal when the distribution is *symmetrical* and *unimodal*

Measures of Variation

Measures of variation are used to indicate the **spread** of the values in a distribution



Figur 8.1 Den lave kurven viser en normalfordeling med forventning 3 og standardavvik 1. Hvis en tar 16 observasjoner fra denne og beregner gjennomsnittet, vil det ha en normalfordeling med forventning 3 og standardavvik 1/4. Dette er den høye tynne fordelingen

Range and interquartile range

- The **range** is the difference between the *largest* and *smallest* values in the sample:

$$\mathrm{R} = Y_n - Y_1,$$

  where $Y_1 = \min(X)$ and $Y_n = \max(X)$
- The **interquartile range** is the difference between the middle two quartiles:

$$\mathrm{IQR} = Q_3 - Q_1,$$

  where Q_1 and Q_3 are the *lower* and *upper* quartiles respectively. It indicates the spread of the middle 50% of the distribution

Variance

- The **population variance** σ^2 may be estimated by the **empirical variance** s^2. It is found by averaging the *squares of the deviations* of the observations from the sample mean

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

where $(n-1)$ is called the number of **degrees of freedom** (d.f.) of the variance

Standard deviation

- The **population standard deviation** σ is found as the *square root* of the variance. It may be estimated by the **empirical standard deviation** s, which is the square root of the empirical variance:

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

- When the underlying population corresponds to a **normal distribution** we have that:
  - about 70% of the observations lie within *one* standard deviation of their mean
  - about 95% of the observations lie within *two* standard deviations of their mean

Example: 4.2 in Kirkwood & Sterne

We want to calculate the standard deviation of the eight **plasma volume measurements** of Example 4.1 in Kirkwood & Sterne.

	Plasma volume X	Deviation from the mean $X - \bar{X}$	Squared deviation from the mean $(X - \bar{X})^2$	Squared observation X^2
	2.75	-0.25	0.0625	7.5625
	2.86	-0.14	0.0196	8.1796
	3.37	0.37	0.1369	11.3569
	2.76	-0.24	0.0576	7.6176
	2.62	-0.38	0.1444	6.8644
	3.49	0.49	0.2401	12.1801
	3.05	0.05	0.0025	9.3025
	3.12	0.12	0.0144	9.7344
Totals	24.02	0.00	0.6780	72.7980

The sum of squared deviations from the sample mean is $\sum_i (X_i - \bar{X})^2 = 0.6780$, and we have $n - 1 = 7$ degrees of freedom. The **empirical standard deviation** is given by $s = \sqrt{\frac{0.6780}{7}} = 0.31$

Properties of the **Sample Mean** $\bar{X}$

$\bar{X}$ also has a distribution!

- mean equal to the population mean μ
- standard deviation, called the **standard error**, equal to $\sigma/\sqrt{n}$
- The **central limit theorem** says that the distribution is a normal distribution, *whether or not* the underlying population is normal (when the sample size is not too small)

Standard Error of the Mean

The **estimated standard error** of the sample mean $\bar{X}$ is given by

$$\widehat{\text{s.e.}} = s_{\bar{X}} = \frac{s}{\sqrt{n}},$$

where s is the empirical standard deviation, and n is the sample size

Example: 4.3 in Kirkwood & Sterne

Once again, we return to the eight **plasma volumes** of Example 4.1 and Example 4.2 in Kirkwood & Sterne (2003). We found that the sample mean is 3.00 litres, and the empirical standard deviation is 0.31 litres. The **estimated standard error** of the sample mean (in litres) is given by

$$\widehat{\text{s.e.}} = s_{\bar{X}} = \frac{0.31}{\sqrt{8}} = 0.11$$

Standard deviation vs. standard error

Remember that

- the **standard deviation** measures the amount of variability in the *population*
- the **standard error** of the sample mean measures the amount of variability in the *sample mean*

Example: 8.2 in Aalen et al.

We have a sample of 4 independent **measurements of cholesterol** from a population with mean $\mu = 6.5$ mmol/l and **standard deviation** $\sigma = 0.5$ mmol/l

The expected value in the sample equals 6.5 mmol/l, and the **standard error** of the sample mean is $\sigma/\sqrt{n} = 0.5/\sqrt{4} = 0.25$

Summary: properties of the sample mean

- The sample mean: $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
- Expectation of the sample mean: $E(\bar{X}) = \mu$
- Variance of the sample mean: $Var(\bar{X}) = \frac{\sigma^2}{n}$
- Standard deviation of the sample mean = standard error: $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
- The distribution of the sample mean: $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ (the central limit theorem)