Matematisk Statistik och Disktret Matematik, MVE055/MSG810, HT19

Föreläsning 7

Nancy Abdallah

Chalmers - Göteborgs Universitet

September 23, 2019

So far we studied random variables with a given distribution. In practice, the exact distribution is unknown. Instead, we study some data and expect the values of the parameters.

Objectives of the rest of the course:

1. **Descriptive Statistics:** Organize and Summarize data.
2. **Inferential Statistics:** Give information about a large body of data by examining only a part of it.

Basic Definitions

- **Population:** Collection of entities of our interest.
- **Random sample:** Part of the population, *chosen randomly*, on which we collect information. Each element in the random sample is a random variable that has the same distribution as the population.
- **Data:** The information we collect from our study (usually defined as numbers)
- **Variable:** The characteristic we are studying.

Example

A study of 250 patients admitted to a hospital during the past year revealed that, on the average, the patients lived 15 miles from the hospital.

The sample in the study is the 250 patients admitted to the hospital during the past year, the population is all the patients admitted to the hospital during the past year, and the variable of interest ”the distance from the hospital”.

- The data of a sample can be summarized graphically or numerically.
  - **Graphical summaries** consist of illustrating data using a histogram, box-plot, polygons, pie chart, etc.
  - **Numerical summaries** consist of computing descriptive measures from the data. These measures are called **statistics** (ex. mean, median, variance, etc.).
    We distinguish two types of descriptive statistics: *measures of central tendency* and *measures of dispersion*.

Measures of central tendency: mean, median

Let $X_1, \cdots, X_n$ be a random sample from the distribution of X.

- The **sample mean** (or **sample average**) is:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + \cdots + X_n}{n}.$$

- The **median** *Me* is the value that divides the set into two equal parts. If n is the total number of values then, if n is odd, the median is the $(\frac{n+1}{2})$th value and if n is even, the median is the average of the $(\frac{n}{2})$th and $(\frac{n}{2}+1)$th values.

Measures of dispersion

- Let $X_1 \cdots, X_n$ be a random sample from the distribution X and let $\overline{X}$ be the mean of $X_1, \cdots, X_n$, then the **sample variance** of $X_1, \cdots, X_n$ is defined by

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

- The variance can also be computed using the formula

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

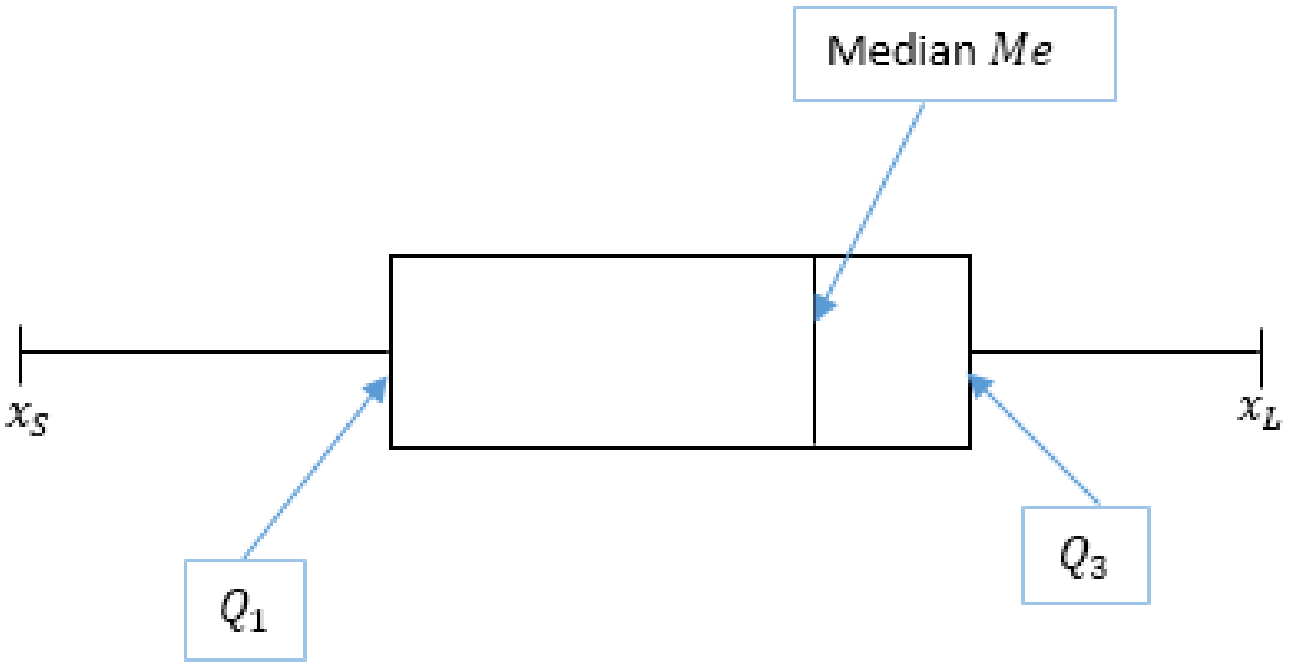
- The variance of a population of N values, $x_1, \cdots, x_N$ with mean μ is given by

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N}$$

- the **Standard Deviation** is the square root of the variance (s for the sample and σ for the population.)
- the **Range** is the difference between the largest value x_L and the smallest value x_S, $R = x_L - x_S = x_n - x_1$ if the values are ordered.

- The p^{th} **percentile** P is the value of X such that p% or less of the observations are less than P and $(100 - p)$% or less are greater than P.
  In particular, P_{25} is the 25^{th} percentile or the **First Quartile** denoted also by Q_1, P_{50} is the 50^{th} percentile or the **Second Quartile** Q_2 which is also the median, P_{75} is the 75^{th} percentile or the **Third Quartile** denoted by Q_3.
  Note that $Q_1 = \frac{n+1}{4}$th ordered observation,
  $Q_2 = \frac{2(n+1)}{4} = \frac{n+1}{2}$th ordered observation, and
  $Q_3 = \frac{3(n+1)}{4}$th ordered observation.

Box-and-Whisker Plots, or Boxplot



Example

Given the following set of data :

$$18,\ 1,\ 20,\ 15,\ 12,\ 15,\ 14,\ 7,\ 11,\ 9,\ 6,\ 4$$

Order the numbers from the lowest to the highest

$$1,\ 4,\ 6,\ 7,\ 9,\ 11,\ 12,\ 14,\ 15,\ 15,\ 18,\ 20$$

Mean:

$$\overline{x} = \frac{20+18+15+15+14+12+11+9+7+6+4+1}{12} = 11$$

Median: Since $n = 12$ is even, the median is the average of the 6^{th} and the 7^{th} values. $Me = \frac{11+12}{2} = 11.5$.

Example

Range: $R = 20 - 1 = 19$.

Variance and standard deviation:

$$s^2 = \frac{(20-11)^2 + (18-11)^2 + \cdots + (-7)^2 + (-10)^2}{11} \approx 33.3$$

and

$$s = \sqrt{s^2} = 5.77$$

Quartiles:

$$Q_1 = 6.25 \text{ and } Q_3 = 15$$

Inferential Statistics

- There are two common forms of **inferential statistics**: **estimation** and **hypothesis testing**.
- There are two forms of estimation: **point estimation** and **interval estimation** (confidence interval).
- A **parameter** is a statistical constant that describes a feature about a population (for instance, μ, σ, etc.)
- In an estimation problem, there is at least one parameter θ whose value is to be estimated or approximated on the basis of a sample.

Point Estimation

- A **point estimate** is a single value that approximate a population parameter θ.
- The rule that tells us how to compute an estimate is referred to as an **estimator**. Estimators are usually presented as formulas. For example,

$$\overline{X} = \frac{\sum X_i}{n}$$

  is an estimator of the population mean, μ. An estimator is a random variable. The single numerical value that results from evaluating this formula is called an estimate of the parameter μ.
- An estimator of θ is denoted by $\hat{\theta}$.

Unbiasedness

- Some estimators are better than others. One of these criteria is the property **unbiasedness**.
- An estimator $\hat{\theta}$ is said to be **unbiased** if and only if $E[\hat{\theta}] = \theta$, where $E[\hat{\theta}]$ is the expected value of θ and is obtained by taking the average value of θ computed from all possible samples of a given size that may be drawn from the population.

Unbiasedness of the sample mean

Let $X_1, \ldots, X_n$ be a random sample of size n from a distribution with mean μ, and let $\bar{X}$ be the sample mean.

$$\begin{aligned} E[\overline{X}] &= E[\frac{1}{n}(X_1 + \cdots + X_n)] \\ &= \frac{1}{n}(E[X_1] + \cdots + E[X_n]) \\ &= \frac{1}{n}(\mu + \cdots + \mu) \\ &= \frac{1}{n}(n\mu) = \mu \end{aligned}$$

Therefore the sample mean $\overline{X}$ is an unbiased estimator of μ

Variance of the sample mean

- Another criteria for an estimator $\hat{\theta}$ to be good is to have a small variance for large sample sizes.
- Let $\overline{X}$ be the sample mean based on a random sample of size n from a distribution with mean μ and variance σ^2. Then

$$Var[\overline{X}] = \frac{\sigma^2}{n}$$

- If the sample size is sufficiently large, the variance of $\overline{X}$ tends to 0 and $\overline{X}$ becomes close to μ.
- The standard deviation of $\overline{X}$ is equal to $\frac{\sigma}{\sqrt{n}}$ and is called the **standard error of the mean**.

Unbiasedness of the sample variance

Let $X_1, \ldots, X_n$ be a random sample of size n drawn from a distribution with mean μ and variance σ^2

- The sample variance $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$ is an unbiased estimator of σ^2.
- The standard deviation s is a **biased** estimator of σ.
- The reason why we divide by $n-1$ and not by n is for S^2 to be unbiased.

Sampling from a normal distribution

- Let $X_1, \ldots, X_n$ be independent normally distributed random variables with mean $\mu_1, \ldots, \mu_n$ and variance $\sigma_1^2, \ldots, \sigma_n^2$ respectively, then $X_1 + \cdots + X_n$ is normally distributed with $\mu = \sum_{i=1}^{n} \mu_i$ and $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$.
- We know that if X is normally distributed, then kX is normally distributed for all k.
- Let $X_1, \cdots, X_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2, then $\overline{X}$ is normally distributed with mean μ and variance σ^2/n