Outliers

The interquartile range is also involved in another important aspect of statistical analysis, namely the detection of outliers. An *outlier*, as the name suggests, is basically an atypical value, "far away" from the rest of the data, that does not seem to belong to the distribution of the rest of the values in the data set.

We have seen how the mean is very sensitive to outliers. Other statistical procedures can be gravely affected by the presence of outliers in the data. Thus, the problem of detecting and locating an outlier is an important part of any statistical data analysis process.

How to classify a value as being "extreme"? First, we could use a simple property, known as the "3σ rule". This is an application of Chebyshev's inequality

$$P(|X - E(X)| < \varepsilon) \geq 1 - \frac{V(X)}{\varepsilon^2},\ \forall \varepsilon > 0.$$

If we use the classical notations $E(X) = \mu, V(X) = \sigma^2,\ \text{Std}(X) = \sigma$ for the mean, variance and standard deviation of X and take $\varepsilon = 3\sigma$, we get

$$\begin{aligned} P(|X - \mu| < 3\sigma) &\geq 1 - \frac{\sigma^2}{9\sigma^2} \\ &= \frac{8}{9} \approx .89. \end{aligned}$$

This is saying that it is *very* probable (at least 0.89 probable) that $|X - \mu| < 3\sigma$, or, equivalently, that $\mu - 3\sigma < X < \mu + 3\sigma$. In words, the 3σ rule states that *most of the values that any random variable takes, at least* 89%*, lie within* 3 *standard deviations away from the mean*. This property is true in general, for any distribution, but especially for unimodal and symmetrical ones, where that percentage is even higher.

Based on that, one simple procedure would be to consider an outlier any value that is more than 2.5 standard deviations away from the mean, and an *extreme* outlier a value more than 3 standard deviations away from the mean.

A more general approach, that works well also for skewed data, is to consider an outlier any observation that is outside the range

$$\left[Q_1 - \frac{3}{2}IQR,\ Q_3 + \frac{3}{2}IQR\right] = \left[Q_1 - 3IQD,\ Q_3 + 3IQD\right].$$

Also, the coefficient $3/2$ can be replaced by some other number to decrease or enlarge the interval of “normal” values (or, equivalently, the domain that covers the outliers):

$$[Q_1 - w \cdot IQR,\ Q_3 + w \cdot IQR]\,,\ w = 0.5, 1, 1.5.$$

For our example on CPU times of processors, we have

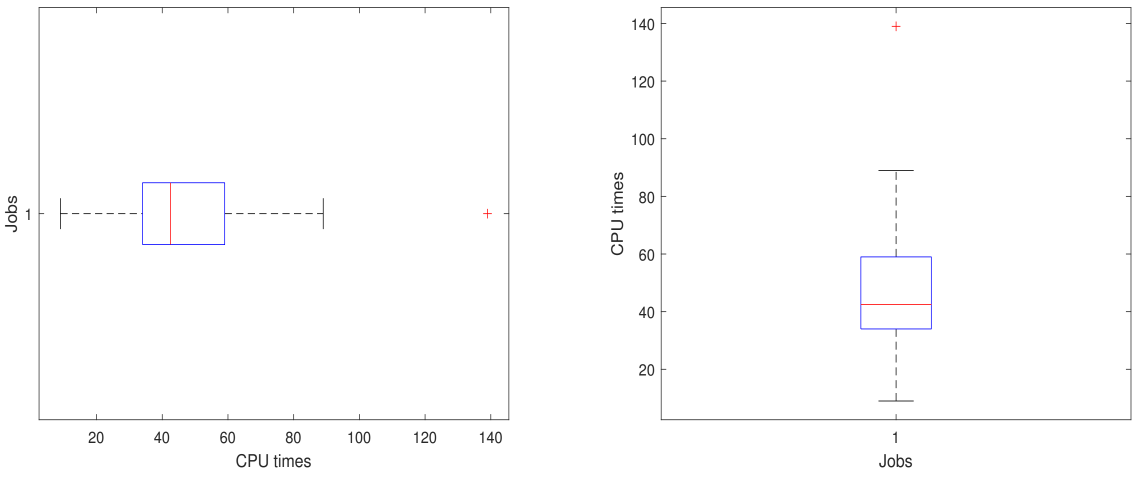
$$\begin{aligned} Q_1 - \frac{3}{2} IQR &= -3.5, \\ Q_3 + \frac{3}{2} IQR &= \ \ 96.5, \end{aligned}$$

so observations outside the interval $[-3.5, 96.5]$ are considered outliers. In this case, there is only one, the value 139.

Boxplots

All the information we discussed above is summarized in a graphical display, called a **boxplot** ($\boxed{\text{boxplot}}$), a plot in which a rectangle is drawn to represent the second and third quartiles (so the interquartile range), with a line inside for the median value and which indicates which values are considered extreme. The “whiskers” of the boxplot are the endpoints of the interval on which normal values lie (so everything outside the whiskers is considered an outlier).

For the data in Example 2.6 from last time (CPU times), the boxplot is displayed in Figure 1.



(a) horizontally

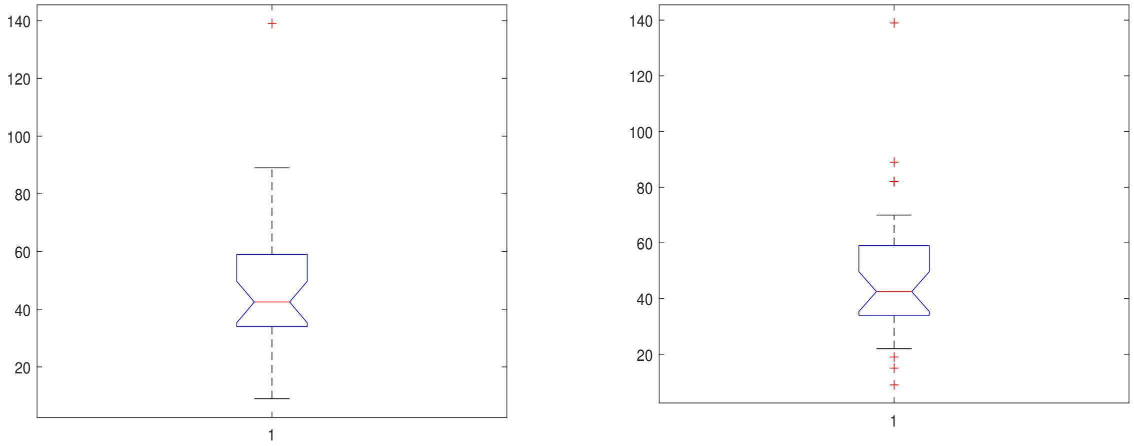
(b) vertically

Fig. 1: Quartiles, Interquartile Range, Outliers

A boxplot can be displayed vertically (default) or horizontally, as in Figure 1. The box can have a “notch” (indentation) at the value of the median, as in Figure 2(a). The width of the interval of the whiskers can be changed. The interval that determines the outliers (i.e., outside of which values are considered too extreme, outliers) is

$$[Q_1 - w \cdot IQR, Q_3 + w \cdot IQR].$$

The default value is $w = 1.5$. With the smaller whiskers, boxplot displays more data points as outliers. In Figure 2(b), the whisker size is set to $w = 0.5$. Then, outliers are all the values outside the interval $[Q_1 - 0.5 \cdot IQR, Q_3 + 0.5 \cdot IQR] = [21.5, 71.5]$. These would be $9, 15, 19$ (too small) and $82, 89, 139$ (too large).



(a) boxplot with a notch

(b) whisker $w = 0.5$

Fig. 2: Boxplots

Boxplots are also very useful when we want to compare data from different samples (see Figure 4). We can compare the interquartile ranges, to examine how the data is dispersed between each sample. The longer the box, the more dispersed the data.

2.1 Moments, variance, standard deviation and coefficient of variation

The idea of the mean can be generalized, by taking various powers of the values in the data.

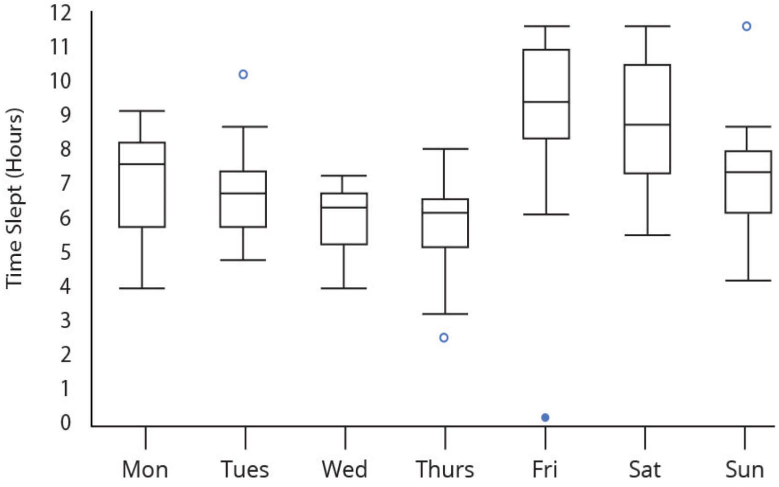


Fig. 3: Multiple boxplots

Definition 2.1.

(1) *The **moment of order k** is the value*

$$\overline{\nu}_k = \frac{1}{N}\sum_{i=1}^{N} x_i^k, \ \ \overline{\nu}_k = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^k, \tag{2.1}$$

for primary and for grouped data, respectively.

(2) *The **central moment of order k*** ($\boxed{\text{moment}}$) *is the value*

$$\overline{\mu}_k = \frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^k, \ \ \overline{\mu}_k = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i-\overline{x})^k \tag{2.2}$$

for primary and for grouped data, respectively.

(3) *The **variance*** ($\boxed{\text{var}}$) *is the value*

$$\overline{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2, \ \ \overline{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i-\overline{x})^2 \tag{2.3}$$

*for primary and for grouped data, respectively. The quantity $\overline{\sigma} = \sqrt{\overline{\sigma}^2}$ is the **standard deviation*** ($\boxed{\text{std}}$).

Remark 2.2.

1. A more efficient computational formula for the variance is

$$\overline{\sigma}^2 = \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\Big(\sum_{i=1}^{N} x_i\Big)^2\right) = \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2 - N\overline{x}^2\right), \tag{2.4}$$

which follows straight from the definition.

2. We will see later that when the data represents a sample (not the entire population), a better formula is

$$\begin{aligned} s^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{x})^2 &&= \frac{1}{N-1}\Big(\sum_{i=1}^{N} x_i^2 - N\overline{x}^2\Big), \\ s^2 &= \frac{1}{N-1}\sum_{i=1}^{n} f_i(x_i-\overline{x})^2 &&= \frac{1}{N-1}\Big(\sum_{i=1}^{N} f_i x_i^2 - N\overline{x}^2\Big), \end{aligned} \tag{2.5}$$

for the *sample* variance for primary or grouped data. The reason the sum is divided by $N-1$ instead of N will have to do with the “bias” of an estimator and will be explained later on in the next chapter. To fully explain why using N leads to a biased estimate involves the notion of *degrees of freedom*, which takes into account the number of constraints in computing an estimate. The sample observations $x_1, \ldots, x_N$ are independent (by the definition of a random sample), but when computing the variance, we use the variables $x_1 - \overline{x}, \ldots, x_N - \overline{x}$. Notice that by subtracting the sample mean $\overline{x}$ from each observation, there exists a linear relation among the elements, namely

$$\sum_{k=1}^{N}(x_k - \overline{x}) = 0$$

and, thus, we lose 1 degree of freedom due to this constraint. Hence, there are only $N-1$ degrees of freedom. So, we will use (2.4) to compute the variance of a set of data that represents a population and (2.5) for the variance of a sample.

Example 2.3. Consider again our previous example on CPU times (in seconds) for $N = 30$ randomly chosen jobs:

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19

Recall that for this data the sample mean was $\overline{x} = 48.2333$ seconds. The sample variance is

$$s^2 = \frac{(70-48.2333)^2 + \ldots + (19-48.2333)^2}{30-1} = \frac{20391}{29} \approx 703.1506 \text{ sec}^2.$$

Alternatively,

$$s^2 \;=\; \frac{70^2 + \ldots + 19^2 - 30 \cdot 48.2333^2}{30-1} \;=\; \frac{90185 - 69794}{29} \;\approx\; 703.1506 \text{ sec}^2.$$

The sample standard deviation is

$$s \;=\; \sqrt{703.1506} \;\approx\; 26.1506 \text{ sec}.$$

By the 3σ rule, using $\overline{x}$ and s as estimates for the population mean μ and population standard deviation σ, we may infer that at least 89% of the tasks performed by this processor require between $\overline{x} - 3s = -30.2185$ and $\overline{x} + 3s = 126.6851$ (so less than 126.6851) seconds of CPU time.

Definition 2.4. *The **coefficient of variation** is the value*

$$CV = \frac{s}{\overline{x}}.$$

Remark 2.5.
1. The coefficient of variation is also known as the **relative standard deviation (RSD)**.
2. It can be expressed as a ratio or as a percentage. It is useful in comparing the degrees of variation of two sets of data, even when their means are different.
2. The coefficient of variation is used in fields such as Analytical Chemistry, Engineering or Physics when doing quality assurance studies. It is also widely used in Business Statistics. For example, in the investing world, the coefficient of variation helps brokers determine how much volatility (risk) they are assuming in comparison to the amount of return they can expect from a certain investment. The lower the value of the CV, the better the risk-return trade off.

3 Sample Theory

In inferential Statistics, we will have the following situation: we are interested in studying a characteristic (a random variable) X, relative to a population P of (known or unknown) size N. The difficulty or even the impossibility of studying the entire population, as well as the merits of choosing and studying a random sample from which to make inferences about the population of interest, have already been discussed in the previous chapter. Now, we want to give a more rigorous and precise definition of a random sample, in the framework of random variables, one that can then employ probability theory techniques for making inferences.

3.1 Random Samples and Sample Functions

We choose n objects from the population and actually study $X_i,\ i = \overline{1,n}$, the characteristic of interest *for the i^{th} object selected.* Since the n objects were randomly selected, it makes sense that for $i = \overline{1,n}$, X_i is a random variable, one that has *the same* distribution (pdf) as X, the characteristic relative to the entire population. Furthermore, these random variables are independent, since the value assumed by one of them has no effect on the values assumed by the others. Once the n objects have been selected, we will have n numerical values available, $x_1, \dots, x_n$, the observed values of the sample variables $X_1, \dots, X_n$.

Definition 3.1. *A **random sample of size n** from the distribution of X, a characteristic relative to a population P, is a collection of n independent random variables $X_1, \dots, X_n$, having the same distribution as X. The variables $X_1, \dots, X_n$, are called **sample variables** and their observed values $x_1, \dots, x_n$, are called **sample data**.*

Remark 3.2. The term *random sample* may refer to the objects selected, to the sample variables, or to the sample data. It is usually clear from the context which meaning is intended. In general, we use capital letters to denote sample variables and corresponding lowercase letters for their observed values, the sample data.

We are able now to define sample functions, or statistics, in the more precise context of random variables.

Definition 3.3. *A **sample function** or **statistic** is a random variable*

$$Y_n = h_n(X_1, \dots, X_n),$$

where $h_n : \mathbb{R}^n \to \mathbb{R}$ is a measurable function. The value of the sample function Y_n is $y_n = h_n(x_1, \dots, x_n)$.

We will revisit now some sample numerical characteristics discussed in the previous sections and define them as sample functions. That means they will have a pdf, a cdf, a mean value, variance, standard deviation, etc. A sample function will, in general, be an approximation for the corresponding population characteristic. In that context, the standard deviation of the sample function is usually referred to as the **standard error**.

In what follows, $\{X_1, \dots, X_n\}$ denotes a sample of size n drawn from the distribution of some population characteristic X.

3.2 Sample Mean

Definition 3.4. *The* ***sample mean*** *is the sample function defined by*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.1}$$

and its value is $\overline{x}_n = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i$.

Now that the sample mean is defined as a random variable, we can discuss its numerical characteristics.

Proposition 3.5. *Let* X *be a population characteristic with mean* $E(X) = \mu$ *and variance* $V(X) = \sigma^2$*. Then*

$$E\left(\overline{X}\right) = \mu \ \textit{ and } \ V\left(\overline{X}\right) = \frac{\sigma^2}{n}. \tag{3.2}$$

Proof. Since $X_1, \dots, X_n$ are identically distributed, with the same distribution as X, $E(X_i) = E(X) = \mu$ and $V(X_i) = V(X) = \sigma^2$, $\forall i = \overline{1,n}$. Then, by the usual properties of expectation, we have

$$E\left(\overline{X}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\, n\mu = \mu.$$

Further, since $X_1, \dots, X_n$ are also independent, by the properties of variance, it follows that

$$V\left(\overline{X}\right) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2}\, n\sigma^2 = \frac{\sigma^2}{n}.$$

□

Remark 3.6. As a consequence, the standard deviation of $\overline{X}$ is

$$\mathrm{Std}(\overline{X}) \ = \ \sqrt{V(\overline{X})} \ = \ \frac{\sigma}{\sqrt{n}}.$$

So, when estimating the population mean μ from a sample of size n by the sample mean $\overline{X}$, the *standard error* of the estimate is $\sigma/\sqrt{n}$, which oftentimes is estimated by $s/\sqrt{n}$. Either way, notice that as n increases and tends to ∞, the standard error decreases and approaches 0. That means that the larger the sample on which we base our estimate, the more accurate the approximation.

3.3 Sample Moments and Sample Variance

Definition 3.7. *The statistic*

$$\overline{\nu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \tag{3.3}$$

is called the ***sample moment of order k*** *and its value is* $\frac{1}{n}\sum\limits_{i=1}^{n} x_i^k$.

The statistic

$$\overline{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k \tag{3.4}$$

is called the ***sample central moment of order k*** *and its value is* $\frac{1}{n}\sum\limits_{i=1}^{n}(x_i - \overline{x})^k$.

Remark 3.8. Just like for theoretical (population) moments, we have

$$\begin{aligned} \overline{\nu}_1 &= \overline{X}, \\ \overline{\mu}_1 &= 0, \\ \overline{\mu}_2 &= \overline{\nu}_2 - \overline{\nu}_1^2. \end{aligned}$$

Next we discuss the characteristics of these new sample functions.

Proposition 3.9. *Let X be a characteristic with the property that for $k \in \mathbb{N}$, the theoretical moment $\nu_{2k} = \nu_{2k}(X) = E\left(X^{2k}\right)$ exists. Then*

$$E\left(\overline{\nu}_k\right) = \nu_k \text{ and } V\left(\overline{\nu}_k\right) = \frac{1}{n}\left(\nu_{2k} - \nu_k^2\right). \tag{3.5}$$

Proof. First off, the condition that ν_{2k} exists for X ensures the fact that all theoretical moments of X of order up to k also exist. The rest follows as before. We have

$$E\left(\overline{\nu}_k\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^k) = \frac{1}{n}\sum_{i=1}^{n} E(X^k) = \frac{1}{n}\, n\nu_k = \nu_k$$

and

$$\begin{aligned} V\left(\overline{\nu}_k\right) &= \frac{1}{n^2}\sum_{i=1}^{n} V(X_i^k) &&= \frac{1}{n^2}\sum_{i=1}^{n} V(X^k) \\ &= \frac{1}{n^2}\, n\left(\nu_{2k} - \nu_k^2\right) &&= \frac{1}{n}\left(\nu_{2k} - \nu_k^2\right). \end{aligned}$$

□

Proposition 3.10. *Let X be a characteristic with variance $V(X) = \mu_2 = \sigma^2$ and for which the theoretical moment $\nu_4 = E\left(X^4\right)$ exists. Then*

$$\begin{aligned} E\left(\overline{\mu}_2\right) &= \frac{n-1}{n}\,\sigma^2, \\ V\left(\overline{\mu}_2\right) &= \frac{n-1}{n^3}\Big[(n-1)\mu_4 - (n-3)\sigma^4\Big]. \end{aligned} \tag{3.6}$$

Proof. The proof is computational as before, but a bit more complicated and we omit it.

□

Remark 3.11. Notice that the sample central moment of order 2 is the first statistic whose expected value *is not* the corresponding population function, in this case the theoretical variance. This is the motivation for the next definition.

Definition 3.12. *The statistic*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{3.7}$$

is called the ***sample variance*** *and its value is* $\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2$.
The statistic $s = \sqrt{s^2}$ is called the ***sample standard deviation***.

Remark 3.13. Notice that the sample central moment of order 2 is no longer equal to the sample variance, as we are used. In fact, we have

$$s^2 = \frac{n}{n-1}\,\overline{\mu}_2.$$

Then, by Proposition 3.10, we have for the sample variance

$$\begin{aligned} E\left(s^2\right) &= \mu_2 = \sigma^2, \\ V\left(s^2\right) &= \frac{1}{n(n-1)}\Big[(n-1)\mu_4 - (n-3)\sigma^4\Big] \end{aligned} \tag{3.8}$$

and, again, the estimation of σ^2 by s^2 (or of σ by s) has a standard error that decreases as the sample size increases:

$$\mathrm{Std}(s^2) = \sqrt{\frac{1}{n(n-1)}\Big((n-1)\mu_4 - (n-3)\sigma^4\Big)} \longrightarrow 0, \text{ as } n \to \infty.$$

3.4 Sample Proportions

Definition 3.14. *Assume a subpopulation A of a population consists of items that have a certain attribute. The* ***population proportion*** *is then the probability*

$$p \;=\; P(i \in A), \tag{3.9}$$

i.e. the probability for a randomly selected item i to have this attribute.
The ***sample proportion*** *is*

$$\overline{p} \;=\; \frac{\text{number of sampled items from } A}{n}, \tag{3.10}$$

where n is the sample size.

Proposition 3.15. *Let p be a population proportion. Then*

$$E\left(\overline{p}\right) = p,\;\; V\left(\overline{p}\right) = \frac{p(1-p)}{n} = \frac{pq}{n} \;\textit{ and }\; \sigma\left(\overline{p}\right) = \sqrt{\frac{pq}{n}}, \tag{3.11}$$

where $q = 1 - p$.

Proof. We use the indicator random variable

$$X_i \;=\; \begin{cases} 1, & i \in A \\ 0, & i \notin A \end{cases}.$$

Then $X_i \in Bern(p)$ and, so, we know that $E(X_i) = p$ and $V(X_i) = pq$, for every $i = 1, \dots, n$. But notice that $\overline{p} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$, i.e. the sample mean of the sample $X_1, \dots, X_n$. Thus, by Proposition 3.5, we have

$$\begin{aligned} E\left(\overline{p}\right) &= p, \\ V\left(\overline{p}\right) &= \frac{pq}{n}, \\ \sigma\left(\overline{p}\right) &= \sqrt{\frac{pq}{n}}. \end{aligned}$$

□

3.5 Sample Functions for Comparing Two Populations

It will be necessary sometimes to compare characteristics of two populations. For that, we will need results on sample functions referring to both collections. Assume we have two characteristics $X_{(1)}$ and $X_{(2)}$, relative to two populations. We draw from both populations independent random samples of sizes n_1 and n_2, respectively. Denote the two sets of random variables by

$$X_{11}, \dots, X_{1n_1} \text{ and } X_{21}, \dots, X_{2n_2}.$$

Then we have two sample means and two sample variances, given by

$$\overline{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i}, \quad \overline{X}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} X_{2j}$$

and

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1} \left(X_{1i} - \overline{X}_1\right)^2, \;\; s_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2} \left(X_{2j} - \overline{X}_2\right)^2,$$

respectively. In addition, denote by

$$s_p^2 = \frac{\sum_{i=1}^{n_1} \left(X_{1i} - \overline{X}_1\right)^2 + \sum_{j=1}^{n_2} \left(X_{2j} - \overline{X}_2\right)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

the **pooled variance** of the two samples, i.e. a variance that considers (pools) the data from both samples.

In inferential Statistics, when comparing the means of two populations, we will look at their difference and try to estimate it. Regarding that, we have the following result.

Proposition 3.16. *Let $X_{(1)}, X_{(2)}$ be two population characteristics with means $E(X_{(i)}) = \mu_i$ and variances $V(X_{(i)}) = \sigma_i^2, i = 1, 2$. Then*

$$\begin{aligned} E\left(\overline{X}_1 - \overline{X}_2\right) &= \mu_1 - \mu_2, \\ V\left(\overline{X}_1 - \overline{X}_2\right) &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \end{aligned} \tag{3.12}$$

In a similar fashion, we can compare two population proportions. Again, the random variable of interest is their difference.

Proposition 3.17. *Assume we have two population proportions* p_1 *and* p_2*. From each population we draw independent samples of size* n_1 *and* n_2*, respectively, which yield the population proportions* $\overline{p}_1$ *and* $\overline{p}_2$*. Then*

$$\begin{aligned} E\left(\overline{p}_1 - \overline{p}_2\right) &= p_1 - p_2, \\ V\left(\overline{p}_1 - \overline{p}_2\right) &= \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}, \end{aligned} \tag{3.13}$$

with $q_i = 1 - p_1, i = 1, 2.$

Summary of Notations

Notations of the sample functions and their corresponding population characteristics.

Function	Population (theoretical)	Sample
Mean	$\mu = E(X)$	$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
Variance	$\sigma^2 = V(X)$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
Standard deviation	$\sigma = \sqrt{V(X)}$	$s = \sqrt{s^2}$
Moment of order k	$\nu_k = E\left(X^k\right)$	$\overline{\nu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$
Central moment of order k	$\mu_k = E\left[(X - E(X))^k\right]$	$\overline{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$
Proportion	$p = P(i \in A)$	$\overline{p} = \frac{\text{number of } X_i \text{ from } A}{n}$

Table 1: Notations

Chapter 3. Inferential Statistics

Populations are characterized by parameters. The goal of Inferential Statistics is to make inferences (estimates) about one or more population parameters on the basis of a sample.

1 Estimation; Basic Notions

We will refer to the parameter to be estimated as the **target parameter** and denote it by θ.
Two types of estimation will be considered: **point estimate**, when the result of the estimation is one single value and **interval estimate**, when the estimate is an interval enclosing the value of the target parameter. In either case, the actual estimation is accomplished by an **estimator**, a rule, a formula, or a procedure that leads us to the value of an estimate, based on the data from a sample.

In this chapter, we discuss how

- *to estimate parameters of the distribution.* The methods in the previous chapter mostly concern measure of location (mean, median, quantiles) and variability (variance, standard deviation, interquartile range). As we know, this does not cover all possible parameters, and thus, we still lack a general methodology of estimation.

- *to construct confidence intervals.* Any estimator, computed from a collected random sample instead of the whole population, is understood as only an approximation of the corresponding parameter. Instead of one estimator that is subject to a sampling error, it is often more reasonable to produce an interval that will contain the true population parameter with a certain known high probability.

- *to test hypotheses.* That is, we shall use the collected sample to verify statements and claims about the population. As a result of each test, a statement is either rejected on basis of the observed data or accepted (not rejected). Sampling error in this analysis results in a possibility of wrongfully accepting or rejecting the hypothesis; however, we can design tests to control the probability of such errors.

Results of such statistical analysis are used for making decisions under uncertainty, developing optimal strategies, forecasting, evaluating and controlling performance and so on.

Throughout this chapter, we consider a characteristic X (relative to a population), whose pdf $f(x;\theta)$ depends on the parameter θ, which is to be estimated. If X is discrete, then f represents the probability distribution function, while if X is continuous, f is the probability density function.

As before, we consider a random sample of size n, i.e. sample variables $X_1, \dots, X_n$, which are **independent and identically distributed (iid)**, having the same pdf as X. The notations introduced in the previous chapter for some sample functions still stand.

Definition 1.1. *A **point estimator** for (the estimation of) the target parameter θ is a sample function (statistic)*

$$\overline{\theta} = \overline{\theta}(X_1, X_2, \dots, X_n).$$

Other notations may be used, such as $\hat{\theta}$ or $\tilde{\theta}$.

Each statistic is a random variable because it is computed from random data. It has a so-called *sampling distribution*. Each statistic estimates the corresponding population parameter and adds certain information about the distribution of X, the variable of interest. The value of the point estimator, the **point estimate**, is the actual approximation of the unknown parameter.

2 Properties of Point Estimators

Many different point estimators may be obtained for the same target parameter. Some are considered "good", others "bad", some "better" than others. We need some criteria to decide on one estimator versus another.

2.1 Unbiased Estimators

For one thing, it is highly desirable that the sampling distribution of an estimator $\overline{\theta}$ is "clustered" around the target parameter. In simple terms, we *expect* that the value the point estimator provides to be the actual value of the parameter it estimates. This justifies the following notion.

Definition 2.1. *A point estimator $\overline{\theta}$ is called an **unbiased** estimator for θ if*

$$E(\overline{\theta}) \quad = \quad \theta. \tag{2.1}$$

*The **bias** of $\overline{\theta}$ is the value $B = E(\overline{\theta}) - \theta$.*

Unbiasedness means that in the long-run, collecting a large number of samples and computing $\overline{\theta}$ from each of them, on the average we hit the unknown parameter θ exactly. In other words, in a long run, unbiased estimators neither underestimate nor overestimate the parameter.

Example 2.2.
1. Recall from Proposition 3.5 that for the sample mean, as a random variable, we have $E(\overline{X}) = \mu$. Thus, the sample mean is an *unbiased* estimator for the population mean.
2. More generally, the sample moment of order $k, \overline{\nu}_k$, is an unbiased estimator for the population moment of order $k, \nu_k = E(X^k)$, since $E(\overline{\nu}_k) = \nu_k$ (Proposition 3.9).
3. By Proposition 3.10, the sample central moment of order 2 *is not* an unbiased estimator for the population central moment of order 2 (or it is a *biased* estimator), since

$$E(\overline{\mu}_2) = \frac{n-1}{n}\mu_2 \neq \mu_2 = \sigma^2.$$

4. However, the sample variance

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

is an unbiased estimator for the population variance, since $E(s^2) = \sigma^2$ (see Remark 3.13). That was the main reason for the way the sample variance was defined.

Another desirable trait for a point estimator is that its values do not vary too much from the value of the target parameter. So we need to evaluate variability of computed statistics and especially parameter estimators. That can be accomplished by computing the following statistic.

Definition 2.3. *The* ***standard error*** *of an estimator $\overline{\theta}$, denoted by $\sigma_{\overline{\theta}}$, is its standard deviation*

$$\sigma_{\overline{\theta}} = \sigma(\overline{\theta}) = \mathrm{Std}(\overline{\theta}) = \sqrt{V(\overline{\theta})}.$$

Both population and sample variances are measured in squared units. Therefore, it is convenient to have standard deviations that are comparable with our variable of interest, X. As a measure of variability, standard errors show precision and reliability of estimators. They show how much estimators of the same target parameter θ can vary if they are computed from different samples. Ideally, we would like to deal with unbiased or nearly unbiased estimators that have *low* standard error.

In Table 2 we present some common unbiased estimators, their means and their standard errors.

Remark 2.4.
1. The expected values and the standard errors in Table 2 are valid regardless of the form of the density function of the underlying population.

Target Param. θ	Sample Size	Pt. Estimator $\overline{\theta}$	Mean $E(\overline{\theta})$	St. Error $\sigma_{\overline{\theta}}$
μ	n	$\overline{X}$	μ	$\dfrac{\sigma}{\sqrt{n}}$
ν_k	n	$\overline{\nu}_k$	ν_k	$\sqrt{\dfrac{\nu_{2k} - \nu_k^2}{n}}$
p	n	$\overline{p}$	p	$\sqrt{\dfrac{pq}{n}}$
$\mu_1 - \mu_2$	n_1, n_2	$\overline{X}_1 - \overline{X}_2$	$\mu_1 - \mu_2$	$\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$
$p_1 - p_2$	n_1, n_2	$\overline{p}_1 - \overline{p}_2$	$p_1 - p_2$	$\sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}$

Table 2: Common Unbiased Estimators

2. For large samples (as $n, n_1, n_2 \to \infty$), all these estimators have probability densities that are approximately Normal. The Central Limit Theorem and similar theorems justify these statements. In practice, it was determined that "large" means $n > 30$ for one sample and $n_1 + n_2 > 40$ for two samples.