Chapter 8

Introduction to Statistics

The first seven chapters of this book taught us to analyze problems and systems involving uncertainty, to find probabilities, expectations, and other characteristics for a variety of situations, and to produce forecasts that may lead to important decisions.

What was given to us in all these problems? Ultimately, *we needed to know the distribution and its parameters*, in order to compute probabilities or at least to estimate them by means of Monte Carlo. Often the distribution may not be given, and we learned how to fit the suitable model, say, Binomial, Exponential, or Poisson, given the type of variables we deal with. In any case, parameters of the fitted distribution had to be reported to us explicitly, or they had to follow directly from the problem.

This, however, is rarely the case in practice. Only sometimes the situation may be under our control, where, for example, produced items have predetermined specifications, and therefore, one knows parameters of their distribution.

Much more often *parameters are not known.* Then, how can one apply the knowledge of Chapters 1–7 and compute probabilities? The answer is simple: *we need to collect data.* A properly collected sample of data can provide rather sufficient information about parameters of the observed system. In the next sections and chapters, we learn how to use this sample

- to visualize data, understand the patterns, and make quick statements about the system's behavior;
- to characterize this behavior in simple terms and quantities;
- to estimate the distribution parameters;
- to assess reliability of our estimates;

– to test statements about parameters and the entire system;

– to understand relations among variables;

– to fit suitable models and use them to make forecasts.

8.1 Population and sample, parameters and statistics

Data collection is a crucially important step in Statistics. We use the collected and observed *sample* to make statements about a much larger set — the *population*.

DEFINITION 8.1

> A **population** consists of all units of interest. Any numerical characteristic of a population is a **parameter**. A **sample** consists of observed units collected from the population. It is used to make statements about the population. Any function of a sample is called **statistic**.

In real problems, we would like to make statements about the population. To compute probabilities, expectations, and make optimal decisions under uncertainty, we need to know the population *parameters*. However, the only way to know these parameters is to measure the entire population, i.e., to conduct a *census*.

Instead of a census, we may collect data in a form of a random sample from a population (Figure 8.1). This is our data. We can measure them, perform calculations, and *estimate* the unknown parameters of the population up to a certain *measurable* degree of accuracy.

NOTATION

θ = population parameter
$\widehat{\theta}$ = its estimator, obtained from a sample

Example 8.1 (CUSTOMER SATISFACTION). For example, even if 80% of all users are satisfied with their internet connection, it does not mean that exactly 8 out of 10 customers in your observed sample are satisfied. As we can see from Table A2 in the Appendix, with probability 0.0328, only five out of ten sampled customers are satisfied. In other words, there is a 3% chance for a random sample to suggest that contrary to the claimed population parameter, no more than 50% of users are satisfied. ◊

This example shows that a sample may sometimes give a rather misleading information about the population although this happens with a low probability. *Sampling errors cannot be excluded.*

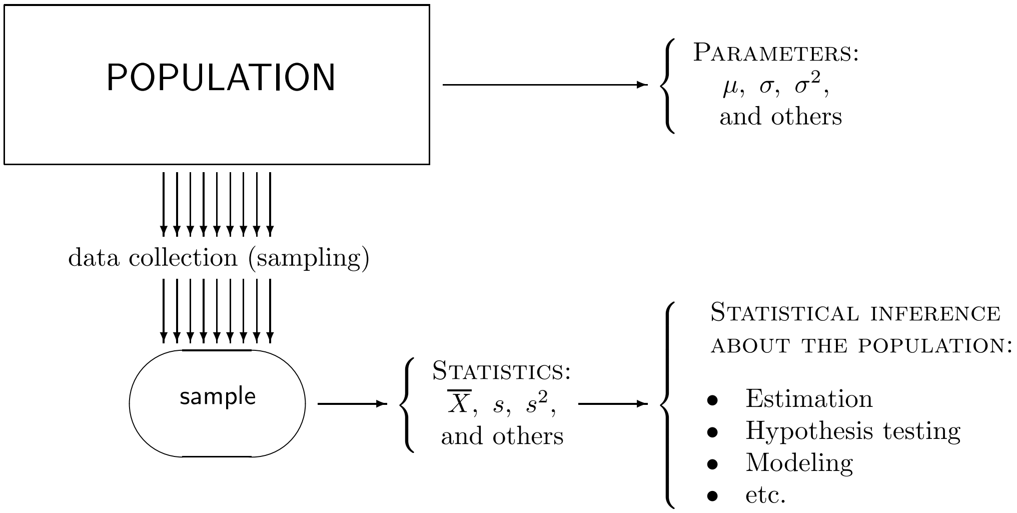

POPULATION
PARAMETERS:
μ, σ, σ^2,
and others
data collection (sampling)
sample
STATISTICS:
$\overline{X}$, s, s^2,
and others
STATISTICAL INFERENCE ABOUT THE POPULATION:
• Estimation
• Hypothesis testing
• Modeling
• etc.

FIGURE 8.1: *Population parameters and sample statistics.*

Sampling and non-sampling errors

Sampling and non-sampling errors refer to any discrepancy between a collected sample and a whole population.

Sampling errors are caused by the mere fact that only a sample, a portion of a population, is observed. For most of reasonable statistical procedures, sampling errors decrease (and converge to zero) as the sample size increases.

Non-sampling errors are caused by inappropriate sampling schemes or wrong statistical techniques. Often no wise statistical techniques can rescue a poorly collected sample of data.

Look at some examples of wrong sampling practices.

Example 8.2 (SAMPLING FROM A WRONG POPULATION). To evaluate the work of a Windows help desk, a survey of *social science students of some university* is conducted. This sample poorly represents the whole population of *all Windows users*. For example, computer science students and especially computer professionals may have a totally different opinion about the Windows help desk. ◊

Example 8.3 (DEPENDENT OBSERVATIONS). Comparing two brands of notebooks, a senior manager asks all employees of her group to state which notebook they like and generalizes the obtained responses to conclude which notebook is better. Again, these employees are not randomly selected from the population of all users of these notebooks. Also, their opinions are likely to be *dependent*. Working together, these people often communicate, and their points of view affect each other. Dependent observations do not necessarily cause non-sampling errors, if they are handled properly. The fact is, in such cases, we cannot assume independence. ◊

Example 8.4 (NOT EQUALLY LIKELY). A survey among passengers of some airline is conducted in the following way. A sample of random flights is selected from a list, and ten passengers on each of these flights are also randomly chosen. Each sampled passenger is asked to fill a questionnaire. Is this a representative sample?

Suppose Mr. X flies only once a year whereas Ms. Y has business trips twice a month. Obviously, Ms. Y has a much higher chance to be sampled than Mr. X. Unequal probabilities have to be taken into account, otherwise a non-sampling error will inevitably occur. ◊

Example 8.5 (PRESIDENTIAL ELECTION OF 1936). A popular weekly magazine The Literary Digest correctly predicted the winners of 1920, 1924, 1928, and 1932 U.S. Presidential Elections. However, it failed to do so in 1936! Based on a survey of ten million people, it predicted an overwhelming victory of Governor Alfred Landon. Instead, Franklin Delano Roosevelt received 98.49% of the electoral vote, won 46 out of 48 states, and was re-elected.

So, what went wrong in that survey? At least two main issues with their sampling practice caused this prediction error. First, the sample was based on the population of subscribers of The Literary Digest that was dominated by Republicans. Second, responses were voluntary, and 77% of mailed questionnaires were not returned, introducing further bias. These are classical examples of non-sampling errors. ◊

In this book, we focus on *simple random sampling*, which is one way to avoid non-sampling errors.

DEFINITION 8.2

Simple random sampling is a sampling design where units are collected from the entire population independently of each other, all being equally likely to be sampled.

Observations collected by means of a simple random sampling design are **iid** (independent, identically distributed) random variables.

Example 8.6. To evaluate its customers' satisfaction, a bank makes a list of all the accounts. A Monte Carlo method is used to choose a random number between 1 and N, where N is the total number of bank accounts. Say, we generate a Uniform(0,N) variable X and sample an account number $\lceil X \rceil$ from the list. Similarly, we choose the second account, uniformly distributed among the remaining $N-1$ accounts, etc., until we get a sample of the desired size n. This is a simple random sample. ◊

Obtaining a good, representative random sample is rather important in Statistics. Although we have only a portion of the population in our hands, a rigorous sampling design followed by a suitable statistical inference allows to estimate parameters and make statements with a certain measurable degree of confidence.

8.2 Descriptive statistics

Suppose a good random sample

$$\mathcal{S} = (X_1,\, X_2, \ldots, X_n)$$

has been collected. For example, to evaluate effectiveness of a processor for a certain type of tasks, we recorded the CPU time in seconds for $n = 30$ randomly chosen jobs (data set `CPU`),

$$\begin{array}{rrrrrrrrrr} 70 & 36 & 43 & 69 & 82 & 48 & 34 & 62 & 35 & 15 \\ 59 & 139 & 46 & 37 & 42 & 30 & 55 & 56 & 36 & 82 \\ 38 & 89 & 54 & 25 & 35 & 24 & 22 & 9 & 56 & 19 \end{array} \tag{8.1}$$

What information do we get from this collection of numbers?

We know that X, the CPU time of a random job, is a random variable, and its value does not have to be among the observed thirty. We'll use the collected data to describe the distribution of X.

Simple **descriptive statistics** measuring the location, spread, variability, and other characteristics can be computed immediately. In this section, we discuss the following statistics,

- **mean**, measuring the average value of a sample;
- **median**, measuring the central value;
- **quantiles** and **quartiles**, showing where certain portions of a sample are located;
- **variance**, **standard deviation**, and **interquartile range**, measuring variability and spread of data.

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.

We used similar methods in Section 5.3.2, where we estimated parameters from Monte Carlo samples obtained by computer simulations. Here we estimate parameters and make conclusions based on real, not simulated, data.

8.2.1 Mean

Sample mean $\overline{X}$ estimates the population mean $\mu = \mathbf{E}(X)$.

DEFINITION 8.3

Sample mean $\overline{X}$ is the arithmetic average,

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n}$$

Naturally, being the average of sampled observations, $\overline{X}$ estimates the average value of the whole distribution of X. Computed from random data, $\overline{X}$ does not necessarily equal μ; however, we would expect it to converge to μ when a large sample is collected.

Sample means possess a number of good properties. They are *unbiased*, *consistent*, and *asymptotically Normal.*

Remark: This is true if the population has finite mean and variance, which is the case for almost all the distributions in this book (see, however, Example 3.20 on p. 62).

Unbiasedness

DEFINITION 8.4

An estimator $\widehat{\theta}$ is **unbiased** for a parameter θ if its expectation equals the parameter,

$$\mathbf{E}(\widehat{\theta}) = \theta$$

for all possible values of θ.

Bias of $\widehat{\theta}$ is defined as $\text{Bias}(\widehat{\theta}) = \mathbf{E}(\widehat{\theta} - \theta)$.

Unbiasedness means that in a long run, collecting a large number of samples and computing $\widehat{\theta}$ 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.

Sample mean estimates μ unbiasedly because its expectation is

$$\mathbf{E}(\overline{X}) = \mathbf{E}\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{\mathbf{E}X_1 + \ldots + \mathbf{E}X_n}{n} = \frac{n\mu}{n} = \mu.$$

Consistency

DEFINITION 8.5

An estimator $\widehat{\theta}$ is **consistent** for a parameter θ if the probability of its sampling error of any magnitude converges to 0 as the sample size increases to infinity. Stating it rigorously,

$$\boldsymbol{P}\left\{|\widehat{\theta} - \theta| > \varepsilon\right\} \to 0 \text{ as } n \to \infty$$

for any $\varepsilon > 0$. That is, when we estimate θ from a large sample, the estimation error $|\widehat{\theta} - \theta|$ is unlikely to exceed ε, and it does it with smaller and smaller probabilities as we increase the sample size.

Consistency of $\overline{X}$ follows directly from Chebyshev's inequality on p. 54.

To use this inequality, we find the variance of $\overline{X}$,

$$\operatorname{Var}(\overline{X}) = \operatorname{Var}\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{\operatorname{Var}X_1 + \ldots + \operatorname{Var}X_n}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \tag{8.2}$$

Then, using Chebyshev's inequality for the random variable $\overline{X}$, we get

$$\boldsymbol{P}\left\{|\overline{X} - \mu| > \varepsilon\right\} \le \frac{\operatorname{Var}(\overline{X})}{\varepsilon^2} = \frac{\sigma^2/n}{\varepsilon^2} \to 0,$$

as $n \to \infty$.

Thus, a sample mean is *consistent*. Its sampling error will be small with a higher and higher probability, as we collect larger and larger samples.

Asymptotic Normality

By the Central Limit Theorem, the sum of observations, and therefore, the sample mean have approximately Normal distribution if they are computed from a large sample. That is, the distribution of

$$Z = \frac{\overline{X} - \mathbf{E}\overline{X}}{\operatorname{Std}\overline{X}} = \frac{\overline{X} - \mu}{\sigma\sqrt{n}}$$

converges to Standard Normal as $n \to \infty$. This property is called **Asymptotic Normality**.

Example 8.7 (CPU TIMES). Looking at the CPU data on p. 217 (data set `CPU`), we estimate the average (expected) CPU time μ by

$$\overline{X} = \frac{70 + 36 + \ldots + 56 + 19}{30} = \frac{1447}{30} = 48.2333.$$

We may conclude that the mean CPU time of *all* the jobs is "near" 48.2333 seconds.

◊

NOTATION	μ	=	population mean
	$\overline{X}$	=	sample mean, estimator of μ
	σ	=	population standard deviation
	s	=	sample standard deviation, estimator of σ
	σ^2	=	population variance
	s^2	=	sample variance, estimator of σ

8.2.2 Median

One disadvantage of a sample mean is its *sensitivity to extreme observations*. For example, if the first job in our sample is unusually heavy, and it takes 30 minutes to get processed instead of 70 seconds, this one extremely large observation shifts the sample mean from 48.2333 sec to 105.9 sec. Can we call such an estimator "reliable"?

Another simple measure of location is a *sample median*, which estimates the *population median*. It is much less sensitive than the sample mean.

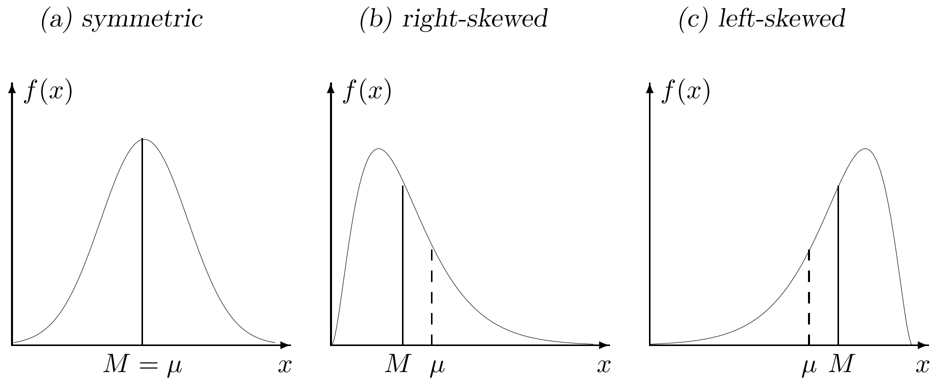


FIGURE 8.2: *A mean μ and a median M for distributions of different shapes.*

DEFINITION 8.6

Median means a "central" value.

Sample median $\widehat{M}$ is a number that is exceeded by at most a half of observations and is preceded by at most a half of observations.

Population median M is a number that is exceeded with probability no greater than 0.5 and is preceded with probability no greater than 0.5. That is, M is such that

$$\left\{\begin{array}{lcl} \boldsymbol{P}\{X > M\} & \leq & 0.5 \\ \boldsymbol{P}\{X < M\} & \leq & 0.5 \end{array}\right.$$

Understanding the shape of a distribution

Comparing the mean μ and the median M, one can tell whether the distribution of X is right-skewed, left-skewed, or symmetric (Figure 8.2):

$$\begin{array}{lcl} \text{Symmetric distribution} & \Rightarrow & M = \mu \\ \text{Right-skewed distribution} & \Rightarrow & M < \mu \\ \text{Left-skewed distribution} & \Rightarrow & M > \mu \end{array}$$

Computation of a population median

For continuous distributions, computing a population median reduces to solving one equation:

$$\left\{\begin{array}{lclcl} \boldsymbol{P}\{X > M\} & = & 1 - F(M) & \leq & 0.5 \\ \boldsymbol{P}\{X < M\} & = & F(M) & \leq & 0.5 \end{array}\right. \quad \Rightarrow \quad F(M) = 0.5.$$

Example 8.8 (Uniform, Figure 8.3a). Uniform(a, b) distribution has a cdf

$$F(x) = \frac{x-a}{b-a} \text{ for } a < x < b.$$

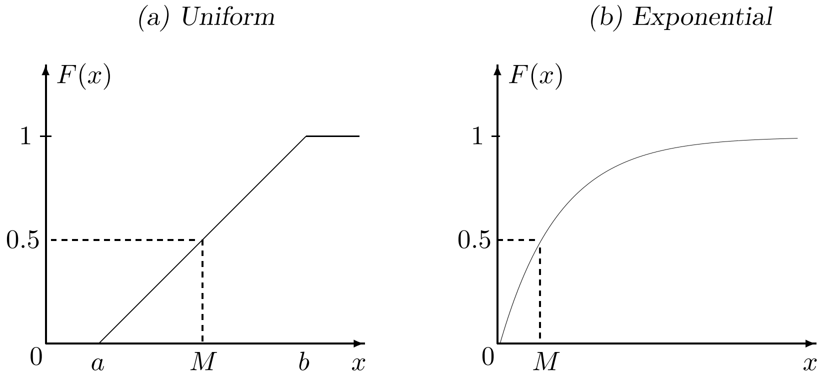


FIGURE 8.3: *Computing medians of continuous distributions.*

Solving the equation $F(M) = (M-a)/(b-a) = 0.5$, we get

$$M = \frac{a+b}{2}.$$

It coincides with the mean because the Uniform distribution is symmetric. ◊

Example 8.9 (Exponential, Figure 8.3b). Exponential(λ) distribution has a cdf

$$F(x) = 1 - e^{-\lambda x} \text{ for } x > 0.$$

Solving the equation $F(M) = 1 - e^{-\lambda x} = 0.5$, we get

$$M = \frac{\ln 2}{\lambda} = \frac{0.6931}{\lambda}.$$

We know that $\mu = 1/\lambda$ for Exponential distribution. Here the median is smaller than the mean because Exponential distribution is right-skewed. ◊

For discrete distributions, equation $F(x) = 0.5$ has either a whole interval of roots or no roots at all (see Figure 8.4).

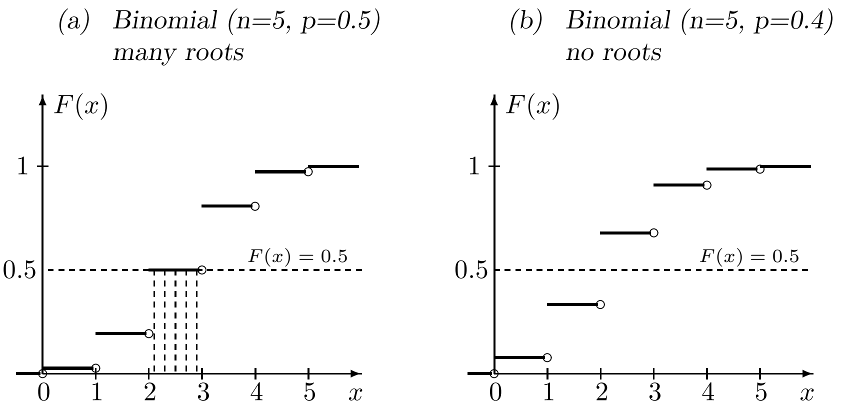
In the first case, any number in this interval, excluding the ends, is a median. Notice that the median in this case is not unique (Figure 8.4a). Often the middle of this interval is reported as the median.

In the second case, the smallest x with $F(x) \geq 0.5$ is the median. It is the value of x where the cdf jumps over 0.5 (Figure 8.4b).

Example 8.10 (Symmetric Binomial, Figure 8.4a). Consider Binomial distribution with $n = 5$ and $p = 0.5$. From Table A2, we see that for all $2 < x < 3$,

$$\begin{cases} \boldsymbol{P}\{X < x\} & = & F(2) & = & 0.5 \\ \boldsymbol{P}\{X > x\} & = & 1 - F(2) & = & 0.5 \end{cases}$$

By Definition 8.6, any number of the interval (2,3) is a median.



(a) Binomial (n=5, p=0.5) many roots

(b) Binomial (n=5, p=0.4) no roots

FIGURE 8.4: *Computing medians of discrete distributions.*

This result agrees with our intuition. With $p = 0.5$, successes and failures are equally likely. Pick, for example, $x = 2.4$ in the interval (2,3). Having fewer than 2.4 successes (i.e., at most two) has the same chance as having fewer than 2.4 failures (i.e., at least 3 successes). Therefore, $X < 2.4$ with the same probability as $X > 2.4$, which makes $x = 2.4$ a central value, a median. We can say that $x = 2.4$ (and any other x between 2 and 3) splits the distribution into two equal parts. Then, it is a median. $\Diamond$

Example 8.11 (Asymmetric Binomial, Figure 8.4b). For the Binomial distribution with $n = 5$ and $p = 0.4$,

$$\begin{array}{lll} F(x) < 0.5 & \text{for} & x < 2 \\ F(x) > 0.5 & \text{for} & x \geq 2 \end{array}$$

but there is no value of x where $F(x) = 0.5$. Then, $M = 2$ is the median.

Seeing a value on either side of $x = 2$ has probability less than 0.5, which makes $x = 2$ a center value. $\Diamond$

Computing sample medians

A sample is always discrete, it consists of a finite number of observations. Then, computing a sample median is similar to the case of discrete distributions.

In simple random sampling, all observations are equally likely, and thus, equal probabilities on each side of a median translate into an equal number of observations.

Again, there are two cases, depending on the sample size n.

Sample median

If n is odd, the $\left(\dfrac{n+1}{2}\right)$-th smallest observation is a median.
If n is even, any number between the $\left(\dfrac{n}{2}\right)$-th smallest and the $\left(\dfrac{n+2}{2}\right)$-th smallest observations is a median.

Example 8.12 (Median CPU time). Let's compute the median of $n = 30$ CPU times from the data on p. 217 (data set `CPU`).

First, order the data,

$$\begin{array}{cccccccccc} 9 & 15 & 19 & 22 & 24 & 25 & 30 & 34 & 35 & 35 \\ 36 & 36 & 37 & 38 & \mathbf{42} & \mathbf{43} & 46 & 48 & 54 & 55 \\ 56 & 56 & 59 & 62 & 69 & 70 & 82 & 82 & 89 & 139 \end{array} \tag{8.3}$$

Next, since $n = 30$ is even, find $n/2 = 15$-th smallest and $(n+2)/2 = 16$-th smallest observations. These are 42 and 43. Any number between them is a sample median (typically reported as 42.5). ◊

We see why medians are not sensitive to extreme observations. If in the previous example, the first CPU time happens to be 30 minutes instead of 70 seconds, it does not affect the sample median at all!

Sample medians are easy to compute. In fact, no computations are needed, only the ordering. If you are driving (and only if you find it safe!), here is a simple experiment that you can conduct yourself.

Example 8.13 (Median speed on a highway). How can you measure the median speed of cars on a multilane road without a radar? It's very simple. Adjust your speed so that a half of cars overtake you, and you overtake the other half. Then you are driving with the median speed! ◊

8.2.3 Quantiles, percentiles, and quartiles

Generalizing the notion of a median, we replace 0.5 in Definition 8.6 by some $0 < p < 1$.

DEFINITION 8.7

A p-**quantile** of a population is such a number x that solves equations

$$\left\{ \begin{array}{rcc} \boldsymbol{P}\{X < x\} & \leq & p \\ \boldsymbol{P}\{X > x\} & \leq & 1-p \end{array} \right.$$

A **sample** p-**quantile** is any number that exceeds at most $100p\%$ of the sample, and is exceeded by at most $100(1-p)\%$ of the sample.

A γ-**percentile** is (0.01γ)-quantile.

First, second, and third **quartiles** are the 25th, 50th, and 75th percentiles. They split a population or a sample into four equal parts.

A **median** is at the same time a 0.5-quantile, 50th percentile, and 2nd quartile.

NOTATION

q_p	$=$	population p-quantile
$\widehat{q}_p$	$=$	sample p-quantile, estimator of q_p
π_γ	$=$	population γ-percentile
$\widehat{\pi}_\gamma$	$=$	sample γ-percentile, estimator of π_γ
$Q_1,\ Q_2,\ Q_3$	$=$	population quartiles
$\widehat{Q}_1,\ \widehat{Q}_2,\ \widehat{Q}_3$	$=$	sample quartiles, estimators of $Q_1,\ Q_2,$ and Q_3
M	$=$	population median
$\widehat{M}$	$=$	sample median, estimator of M

Quantiles, quartiles, and percentiles are related as follows.

Quantiles, quartiles, percentiles

$$\boxed{\begin{array}{c} q_p = \pi_{100p} \\ Q_1 = \pi_{25} = q_{1/4} \qquad Q_3 = \pi_{75} = q_{3/4} \\ M = Q_2 = \pi_{50} = q_{1/2} \end{array}}$$

Sample statistics are of course in a similar relation.

Computing quantiles is very similar to computing medians.

Example 8.14 (SAMPLE QUARTILES). Let us compute the 1st and 3rd quartiles of CPU times. Again, we look at the ordered sample (data set `CPU`)

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

First quartile $\widehat{Q}_1$. For $p = 0.25$, we find that 25% of the sample equals $np = 7.5$, and 75% of the sample is $n(1-p) = 22.5$ observations. From the ordered sample, we see that only the 8th element, 34, has no more than 7.5 observations to the left and no more than 22.5 observations to the right of it. Hence, $\widehat{Q}_1 = 34$.

Third quartile $\widehat{Q}_3$. Similarly, the third sample quartile is the 23rd smallest element, $\widehat{Q}_3 = 59$.

$\Diamond$

Example 8.15 (CALCULATING FACTORY WARRANTIES FROM POPULATION PERCENTILES). A computer maker sells extended warranty on the produced computers. It agrees to issue a warranty for x years if it knows that only 10% of computers will fail before the warranty expires. It is known from past experience that lifetimes of these computers have Gamma distribution with $\alpha = 60$ and $\lambda = 5$ years^{-1}. Compute x and advise the company on the important decision under uncertainty about possible warranties.

Solution. We just need to find the tenth percentile of the specified Gamma distribution and let

$$x = \pi_{10}.$$

As we know from Section 4.3, being a sum of Exponential variables, a Gamma variable is approximately Normal for large $\alpha = 60$. Using (4.12), compute

$$\begin{aligned} \mu &= \alpha/\lambda = 12, \\ \sigma &= \sqrt{\alpha/\lambda^2} = 1.55. \end{aligned}$$

From Table A4, the 10th percentile of a standardized variable

$$Z = \frac{X - \mu}{\sigma}$$

equals (-1.28) (find the probability closest to 0.10 in the table and read the corresponding value of z). Unstandardizing it, we get

$$x = \mu + (-1.28)\sigma = 12 - (1.28)(1.55) = \underline{10.02}.$$

Thus, the company can issue a 10-year warranty rather safely.

Remark: Of course, one does not have to use Normal approximation in the last example. A number of computer packages have built-in commands for the exact evaluation of probabilities and quantiles. For example, the 10th percentile of Gamma($\alpha = 60, \lambda = 5$) distribution can be obtained by the command `qgamma(0.10,60,5)` in R and by `gaminv(0.10, 60, 1/5)` in MATLAB. $\Diamond$

8.2.4 Variance and standard deviation

Statistics introduced in the previous sections showed where the average value and certain percentages of a population are located. Now we are going to measure *variability* of our variable, how unstable the variable can be, and how much the actual value can differ from its expectation. As a result, we'll be able to assess reliability of our estimates and accuracy of our forecasts.

DEFINITION 8.8

For a sample $(X_1, X_2, \ldots, X_n)$, a **sample variance** is defined as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2 . \tag{8.4}$$

It measures variability among observations and estimates the population variance $\sigma^2 = \text{Var}(X)$.

Sample standard deviation is a square root of a sample variance,

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

It measures variability in the same units as X and estimates the population standard deviation $\sigma = \text{Std}(X)$.

Both population and sample variances are measured in squared units (in^2, sec^2, \$^{2}, etc.).

Therefore, it is convenient to have standard deviations that are comparable with our variable of interest, X.

The formula for s^2 follows the same idea as that for σ^2. It is also the average squared deviation from the mean, this time computed for a sample. Like σ^2, sample variance measures how far the actual values of X are from their average.

Computation

Often it is easier to compute the sample variance using another formula,

Sample variance

$$\boxed{s^2 = \frac{\sum\limits_{i=1}^{n} X_i^2 - n\overline{X}^2}{n-1}.} \tag{8.5}$$

Remark: Expressions (8.4) and (8.5) are equivalent because

$$\sum \left(X_i - \overline{X}\right)^2 = \sum X_i^2 - 2\overline{X}\sum X_i + \sum \overline{X}^2 = \sum X_i^2 - 2\overline{X}\left(n\overline{X}\right) + n\overline{X}^2 = \sum X_i^2 - n\overline{X}^2.$$

When $X_1, \ldots, X_n$ are integers, but $(X_1 - \overline{X}), \ldots, (X_n - \overline{X})$ are fractions, it may be easier to use (8.5). However, $(X_n - \overline{X})$ are generally smaller in magnitude, and thus, we'd rather use (8.4) if $X_1, \ldots, X_n$ are rather large numbers.

Example 8.16 (CPU TIME, CONTINUED). For the data in (8.1) on p. 217 (data set `CPU`), we have computed $\overline{X} = 48.2333$. Following Definition 8.8, we can compute the sample variance as

$$s^2 = \frac{(70 - 48.2333)^2 + \ldots + (19 - 48.2333)^2}{30 - 1} = \frac{20,391}{29} = 703.1506 \text{ (sec}^2\text{)}.$$

Alternatively, using (8.5),

$$s^2 = \frac{70^2 + \ldots + 19^2 - (30)(48.2333)^2}{30 - 1} = \frac{90,185 - 69,794}{29} = 703.1506 \text{ (sec}^2\text{)}.$$

The sample standard deviation is

$$s = \sqrt{703.1506} = 26.1506 \text{ (sec}^2\text{)}.$$

We can use these results, for example, as follows. Since $\overline{X}$ and s estimate the population mean and standard deviation, we can make a claim that at least 8/9 of all tasks require less than

$$\overline{X} + 3s = 127.78 \text{ seconds} \tag{8.6}$$

of CPU time. We used Chebyshev's inequality (3.8) to derive this (also see Exercise 8.3). ♢

A seemingly strange coefficient $\left(\frac{1}{n-1}\right)$ ensures that s^2 is an **unbiased** estimator of σ^2.

PROOF: Let us prove the unbiasedness of s^2.

Case 1. Suppose for a moment that the population mean $\mu = \mathbf{E}(X) = 0$. Then

$$\mathbf{E}X_i^2 = \mathrm{Var}X_i = \sigma^2,$$

and by (8.2),

$$\mathbf{E}\overline{X}^2 = \mathrm{Var}\overline{X} = \sigma^2/n.$$

Then,

$$\mathbf{E}s^2 = \frac{\mathbf{E}\sum X_i^2 - n\,\mathbf{E}\overline{X}^2}{n-1} = \frac{n\sigma^2 - \sigma^2}{n-1} = \sigma^2.$$

Case 2. If $\mu \neq 0$, consider auxiliary variables $Y_i = X_i - \mu$. Variances don't depend on constant shifts (see (3.7), p. 53), therefore, Y_i have the same variance as X_i. Their sample variances are equal too,

$$s_Y^2 = \frac{\sum \left(Y_i - \overline{Y}\right)^2}{n-1} = \frac{\sum \left(X_i + \mu - (\overline{X} - \mu)\right)^2}{n-1} = \frac{\sum \left(X_i - \overline{X}\right)^2}{n-1} = s_X^2.$$

Since $\mathbf{E}(Y_i) = 0$, Case 1 applies to these variables. Thus,

$$\mathbf{E}(s_X^2) = \mathbf{E}(s_Y^2) = \sigma_Y^2 = \sigma_X^2.$$

□

Similarly to $\overline{X}$, it can be shown that under rather mild assumptions, sample variance and sample standard deviation are **consistent** and **asymptotically Normal**.

8.2.5 Standard errors of estimates

Besides the population variances and standard deviations, it is helpful to evaluate variability of computed statistics and especially parameter estimators.

DEFINITION 8.9

Standard error of an estimator $\widehat{\theta}$ is its standard deviation, $\sigma(\widehat{\theta}) = \mathrm{Std}(\widehat{\theta})$.

NOTATION

$$\begin{array}{lcl} \sigma(\widehat{\theta}) & = & \text{standard error of estimator } \widehat{\theta} \text{ of parameter } \theta \\ s(\widehat{\theta}) & = & \text{estimated standard error } = \widehat{\sigma}(\widehat{\theta}) \end{array}$$

As a measure of variability, standard errors show precision and reliability of estimators. They show how much estimators of the same 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 (Figure 8.5).

Example 8.17 (STANDARD ERROR OF A SAMPLE MEAN). Parameter $\theta = \mu$, the population mean, is estimated from the sample of size n by the sample mean $\widehat{\theta} = \overline{X}$. We already know that the standard error of this estimator is $\sigma(\overline{X}) = \sigma/\sqrt{n}$, and it can be estimated by $s(\overline{X}) = s/\sqrt{n}$. $\Diamond$

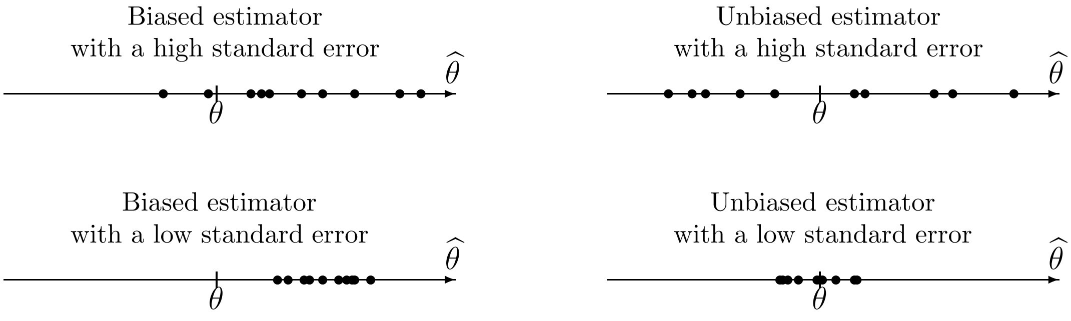


FIGURE 8.5: *Bias and standard error of an estimator. In each case, the dots represent parameter estimators $\widehat{\theta}$ obtained from 10 different random samples.*

8.2.6 Interquartile range

Sample mean, variance, and standard deviation are *sensitive to outliers*. If an extreme observation (an **outlier**) erroneously appears in our data set, it can rather significantly affect the values of $\overline{X}$ and s^2.

In practice, outliers may be a real problem that is hard to avoid. To detect and identify outliers, we need measures of variability that are not very sensitive to them.

One such measure is an interquartile range.

DEFINITION 8.10

An **interquartile range** is defined as the difference between the first and the third quartiles,

$$IQR = Q_3 - Q_1.$$

It measures variability of data. Not much affected by outliers, it is often used to detect them. IQR is estimated by the *sample interquartile range*

$$\widehat{IQR} = \widehat{Q}_3 - \widehat{Q}_1.$$

Detection of outliers

A "rule of thumb" for identifying outliers is the rule of **1.5(IQR)**. Measure $1.5(\widehat{Q}_3 - \widehat{Q}_1)$ down from the first quartile and up from the third quartile. All the data points observed outside of this interval are assumed suspiciously far. They are the first candidates to be handled as outliers.

Remark: The rule of $1.5(IQR)$ originally comes from the assumption that the data are nearly normally distributed. If this is a valid assumption, then 99.3% of the population should appear within 1.5 interquartile ranges from quartiles (Exercise 8.4). It is so unlikely to see a value of X outside of this range that such an observation may be treated as an outlier.

Example 8.18 (Any outlying CPU times?). Can we suspect that sample (8.1) (data set `CPU`) has outliers? Compute

$$\widehat{IQR} = \widehat{Q}_3 - \widehat{Q}_1 = 59 - 34 = 25$$

and measure 1.5 interquartile ranges from each quartile:

$$\begin{array}{rcccl} \widehat{Q}_1 - 1.5(\widehat{IQR}) & = & 34 - 37.5 & = & -3.5; \\ \widehat{Q}_3 + 1.5(\widehat{IQR}) & = & 59 + 37.5 & = & 96.5. \end{array}$$

In our data, one task took 139 seconds, which is well outside of the interval $[-3.5, 96.5]$. This may be an outlier. ♢

Handling of outliers

What should we do if the 1.5(IQR) rule suggests possible outliers in the sample?

Many people simply delete suspicious observations, keeping in mind that one outlier can significantly affect sample mean and standard deviation and therefore spoil our statistical analysis. However, deleting them immediately may not be the best idea.

It is rather important to track the history of outliers and understand the reason they appeared in the data set. There may be a pattern that a practitioner would want to be aware of. It may be a new trend that was not known before. Or, it may be an observation from a very special part of the population. Sometimes important phenomena are discovered by looking at outliers.

If it is confirmed that a suspected observation entered the data set by a mere mistake, it can be deleted.

8.3 Graphical statistics

Despite highly developed theory and methodology of Statistics, when it comes to analysis of real data, experienced statisticians will often follow a very simple advice:

Before you do anything with a data set,
look at it!

A quick look at a sample may clearly suggest

- a probability model, i.e., a family of distributions to be used;

– statistical methods suitable for the given data;

– presence or absence of outliers;

– presence or absence of heterogeneity;

– existence of time trends and other patterns;

– relation between two or several variables.

There is a number of simple and advanced ways to *visualize* data. This section introduces

- histograms,
- stem-and-leaf plots,
- boxplots,
- time plots, and
- scatter plots.

Each graphical method serves a certain purpose and discovers certain information about data.

8.3.1 Histogram

A **histogram** shows the shape of a pmf or a pdf of data, checks for homogeneity, and suggests possible outliers. To construct a histogram, we split the range of data into equal intervals, "bins," and count how many observations fall into each bin.

A **frequency histogram** consists of columns, one for each bin, whose height is determined by the *number* of observations in the bin.

A **relative frequency histogram** has the same shape but a different vertical scale. Its column heights represent the *proportion* of all data that appeared in each bin.

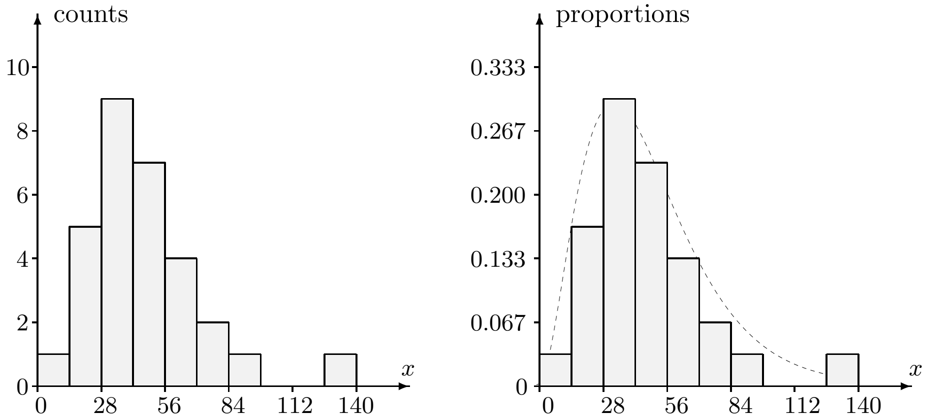
The sample of CPU times on p. 217 (data set `CPU`) stretches from 9 to 139 seconds. Choosing intervals [0,14], [14,28], [28,42], . . . as bins, we count

1	observation	between	0	and	14
5	observations	”	14	”	28
9	”	”	28	”	42
7	”	”	42	”	56
4	”	”	56	”	70
		. . . .			

(8.7)

Using this for column heights, a (frequency) histogram of CPU times is then constructed (Figure 8.6a). A relative frequency histogram (Figure 8.6b) is only different in the vertical scale. Each count is now divided by the sample size $n = 30$.

What information can we draw from these histograms?



(a) Frequency histogram

(b) Relative frequency histogram

FIGURE 8.6: *Histograms of CPU data.*

> *Histograms have a shape similar to the pmf or pdf of data, especially in large samples.*

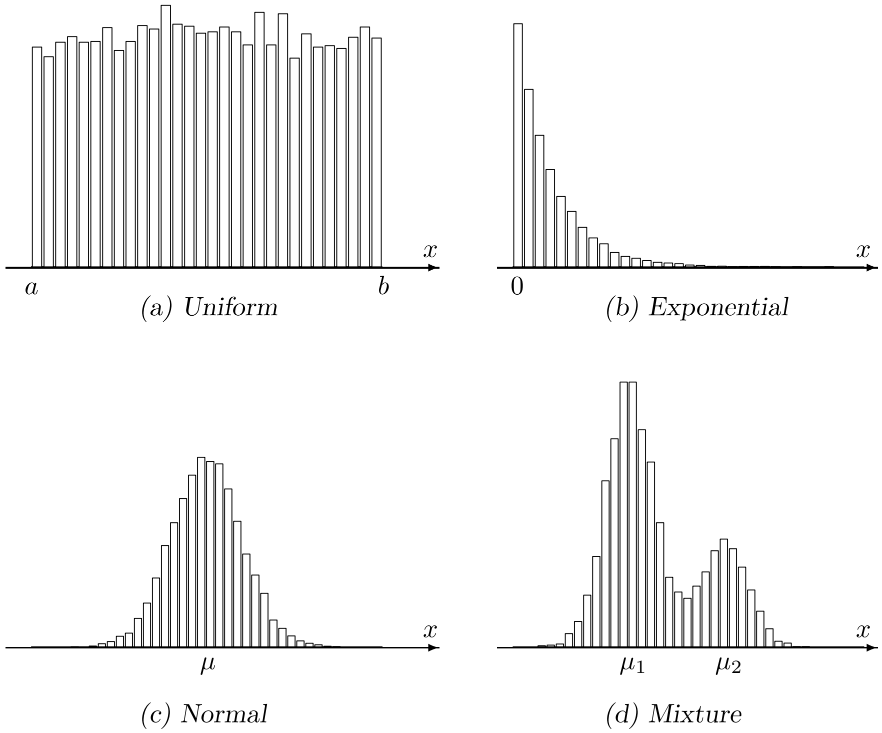
Remark: To understand the last statement, let's imagine for a moment that the data are integers and all columns in a relative frequency histogram have a unit width. Then the height of a column above a number x equals the proportion of x's in a sample, and in large samples it approximates the probability $P(x)$, the pmf (probability is a long-run proportion, Chapter 2).

For continuous distributions, the height of a unit-width column equals its area. Probabilities are areas under the density curve (Chapter 4). Thus, we get approximately the same result either computing sample proportions or integrating the curve that connects the column tops on a relative frequency histogram.

Now, if columns have a non-unit (but still, equal) width, it will only change the horizontal scale but will not alter the shape of a histogram. In each case, this shape will be similar to the graph of the population pmf or pdf.

The following information can be drawn from the histograms shown in Figure 8.6:

- Continuous distribution of CPU times is not symmetric; it is right-skewed as we see 5 columns to the right of the highest column and only 2 columns to the left.
- Among continuous distributions in Chapter 4, only Gamma distribution has a similar shape; a Gamma family seems appropriate for CPU times. We sketched a suitable Gamma pdf with a dashed curve in Figure 8.6b. It is rescaled because our columns don't have a unit width.
- The time of 139 seconds stands alone suggesting that it is in fact an outlier.
- There is no indication of heterogeneity; all data points except $x = 139$ form a rather homogeneous group that fits the sketched Gamma curve.



(a) Uniform

(b) Exponential

(c) Normal

(d) Mixture

FIGURE 8.7: *Histograms of various samples.*

How else may histograms look like?

We saw a rather nice fit of a Gamma distribution in Figure 8.6b, except for one outlier. What other shapes of histograms can we see, and what other conclusions about the population can we make?

Certainly, histograms come in all shapes and sizes. Four examples are shown in Figure 8.7.

In Figure 8.7a, the distribution is almost symmetric, and columns have almost the same height. Slight differences can be attributed to the randomness of our sample, i.e., the *sampling error*. The histogram suggest a Uniform or Discrete Uniform distribution between a and b.

In Figure 8.7b, the distribution is heavily right-skewed, column heights decrease exponentially fast. This sample should come from an Exponential distribution, if variables are continuous, or from Geometric, if they are discrete.

In Figure 8.7c, the distribution is symmetric, with very quickly vanishing "tails." Its bell shape reminds a Normal density that, as we know from Section 4.2.4, decays at a rate of $\sim e^{-cx^2}$. We can locate the center μ of a histogram and conclude that this sample is likely to come from a Normal distribution with a mean close to μ.

Figure 8.7d presents a rather interesting case that deserves special attention.

Mixtures

Let us look at Figure 8.7d . We have not seen a distribution with two "humps" in the previous chapters. Most likely, here we deal with a **mixture of distributions**. Each observation comes from distribution F_1 with some probability p_1 and comes from distribution F_2 with probability $p_2 = 1 - p_1$.

Mixtures typically appear in heterogeneous populations that consist of several groups: females and males, graduate and undergraduate students, daytime and nighttime internet traffic, Windows, Unix, or Macintosh users, etc. In such cases, we can either study each group separately, or use the Law of Total Probability on p. 31, write the (unconditional) cdf as

$$F(x) = p_1 F_1(x) + p_2 F_2(x) + \ldots,$$

and study the whole population at once.

Bell shapes of both humps in Figure 8.7d suggest that the sample came from a mixture of two Normal distributions (with means around μ_1 and μ_2), with a higher probability of having mean μ_1, since the left hump is bigger.

The choice of bins

Experimenting with histograms, you can notice that their shape may depend on the choice of bins. One can hear various rules of thumb about a good choice of bins, but in general,

- there should not be too few or too many bins;
- their number may increase with a sample size;
- they should be chosen to make the histogram informative, so that we can see shapes, outliers, etc.

In Figure 8.6, we simply divided the range of CPU data into 10 equal intervals, 14 sec each, and apparently this was sufficient for drawing important conclusions.

As two extremes, consider histograms in Figure 8.8 constructed from the same CPU data.

The first histogram has too many columns; therefore, each column is short. Most bins have only 1 observation. This tells little about the actual shape of the distribution; however, we can still notice an outlier, $X = 139$.

The second histogram has only 3 columns. It is hard to guess the family of distributions here, although a flat Uniform distribution is already ruled out. The outlier is not seen; it merged with the rightmost bin.

Both histograms in Figure 8.8 can be made more informative by a better choice of bins.

8.3.2 Stem-and-leaf plot

Stem-and-leaf plots are similar to histograms although they carry more information. Namely, they also show how the data are distributed *within* columns.

To construct a stem-and-leaf plot, we need to draw a stem and a leaf. The first one or

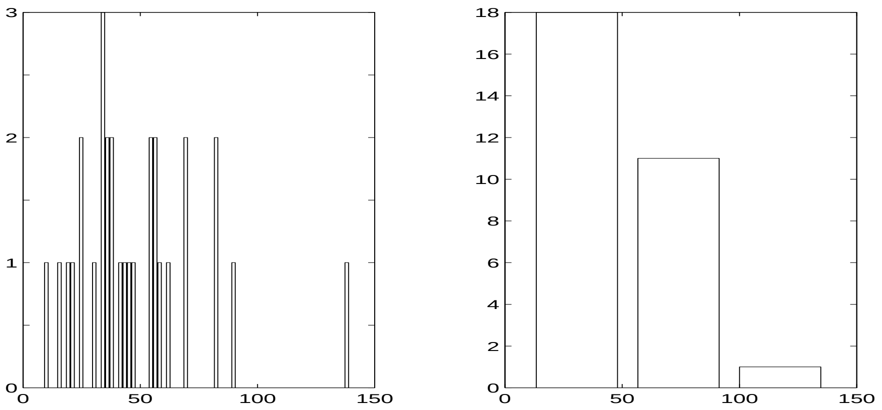


FIGURE 8.8: *Wrong choice of bins for CPU data: too many bins, too few bins.*

several digits form a stem, and the next digit forms a leaf. Other digits are dropped; in other words, the numbers get rounded. For example, a number 239 can be written as

$$23 \,\big|\, 9$$

with 23 going to the stem and 9 to the leaf, or as

$$2 \,\big|\, 3$$

with 2 joining the stem, 3 joining the leaf, and digit 9 being dropped. In the first case, the *leaf unit* equals 1 while in the second case, the leaf unit is 10, showing that the (rounded) number is not 23 but 230.

For the CPU data on p. 217 (data set `CPU`), let the last digits form a leaf. The remaining digits go to the stem. Each CPU time is then written as

$$10 \text{ “stem”} + \text{“leaf”},$$

making the following stem-and-leaf plot,

Leaf unit = 1

$$\begin{array}{r|llllllll}
0 & 9 \\
1 & 5 & 9 \\
2 & 2 & 4 & 5 \\
3 & 0 & 4 & 5 & 5 & 6 & 6 & 7 & 8 \\
4 & 2 & 3 & 6 & 8 \\
5 & 4 & 5 & 6 & 6 & 9 \\
6 & 2 & 9 \\
7 & 0 \\
8 & 2 & 2 & 9 \\
9 & \\
10 & \\
11 & \\
12 & \\
13 & 9
\end{array}$$

Turning this plot by 90 degrees counterclockwise, we get a *histogram* with 10-unit bins (because each stem unit equals 10). Thus, all the information seen on a histogram can be obtained here too. In addition, now we can see individual values within each column. We have the entire sample sorted and written in the form of a stem-and-leaf plot. If needed, we can even compute sample mean, median, quartiles, and other statistics from it.

Example 8.19 (Comparison). Sometimes stem-and-leaf plots are used to compare two samples. For this purpose, one can put two leaves on the same stem. Consider, for example, samples of round-trip transit times (known as pings) received from two locations (data set `Pings`).

Location I: 0.0156, 0.0396, 0.0355, 0.0480, 0.0419, 0.0335, 0.0543, 0.0350, 0.0280, 0.0210, 0.0308, 0.0327, 0.0215, 0.0437, 0.0483 seconds

Location II: 0.0298, 0.0674, 0.0387, 0.0787, 0.0467, 0.0712, 0.0045, 0.0167, 0.0661, 0.0109, 0.0198, 0.0039 seconds

Choosing a leaf unit of 0.001, a stem unit of 0.01, and dropping the last digit, we construct the following two stem-and-leaf plots, one to the left and one to the right of the stem.

Leaf unit = 0.001

						0	3	4	
					5	1	0	6	9
			1	1	8	2			
0	2	3	5	5	9	3	8		
		1	3	8	8	4	6		
					4	5			
						6	1	6	7
						7	8		

Looking at these two plots, one will see about the same average ping from the two locations. One will also realize that the first location has a more stable connection because its pings have lower variability and lower variance. For the second location, the fastest ping will be understood as

$$\{10(\text{leaf } 0) + \text{stem } 3\}\,(\text{leaf unit } 0.001) = 0.003,$$

and the slowest ping as

$$\{10(\text{leaf } 7) + \text{stem } 8\}\,(\text{leaf unit } 0.001) = 0.078.$$

◊

8.3.3 Boxplot

The main descriptive statistics of a sample can be represented graphically by a **boxplot**. To construct a boxplot, we draw a box between the first and the third quartiles, a line inside a box for a median, and extend whiskers to the smallest and the largest observations, thus representing a so-called *five-point summary*:

$$\text{five-point summary} = \left(\min X_i,\ \widehat{Q}_1,\ \widehat{M},\ \widehat{Q}_3,\ \max X_i\right).$$

Often a sample mean $\overline{X}$ is also depicted with a dot or a cross. Observations further than 1.5 interquartile ranges are usually drawn separately from whiskers, indicating the possibility of outliers. This is in accordance with the 1.5(IQR) rule (see Section 8.2.6).

The mean and five-point summary of CPU times were found in Examples 8.7, 8.12, and 8.14,

$$\overline{X} = 48.2333;\ \min X_i = 9,\ \widehat{Q}_1 = 34,\ \widehat{M} = 42.5,\ \widehat{Q}_3 = 59,\ \max X_i = 139.$$

We also know that $X = 139$ is more than $1.5(\widehat{IQR})$ away from the third quartile, and we suspect that it may be an outlier.

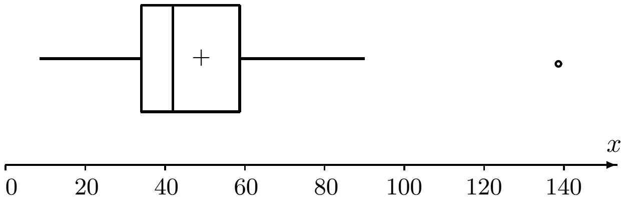


FIGURE 8.9: *Boxplot of CPU time data.*

A boxplot is drawn in Figure 8.9. The mean is depicted with a "+," and the right whisker extends till the second largest observation $X = 89$ because $X = 139$ is a suspected outlier (depicted with a little circle).

From this boxplot, one can conclude:

- The distribution of CPU times is right skewed because (1) the mean exceeds the median, and (2) the right half of the box is larger than the left half.
- Each half of a box and each whisker represents approximately 25% of the population. For example, we expect about 25% of all CPU times to fall between 42.5 and 59 seconds.

Parallel boxplots

Boxplots are often used to compare different populations or parts of the same population. For such a comparison, samples of data are collected from each part, and their boxplots are drawn on the same scale next to each other.

For example, seven parallel boxplots in Figure 8.10 represent the amount of internet traffic handled by a certain center during a week. We can see the following general patterns:

- The heaviest internet traffic occurs on Fridays.
- Fridays also have the highest variability.
- The lightest traffic is seen on weekends, with an increasing trend from Saturday to Monday.
- Each day, the distribution is right-skewed, with a few outliers on each day except Saturday. Outliers indicate occurrences of unusually heavy internet traffic.

Trends can also be seen on scatter plots and time plots.

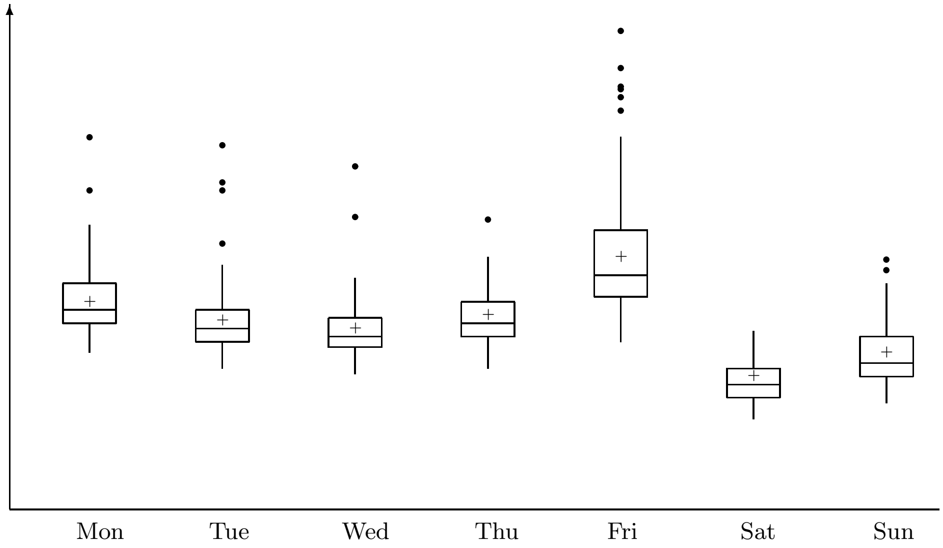


FIGURE 8.10: *Parallel boxplots of internet traffic.*

8.3.4 Scatter plots and time plots

Scatter plots are used to see and understand a relationship between two variables. These can be temperature and humidity, experience and salary, age of a network and its speed, number of servers and the expected response time, and so on.

To study the relationship, both variables are measured on each sampled item. For example, temperature and humidity during each of n days, age and speed of n networks, or experience and salary of n randomly chosen computer scientists are recorded. Then, a **scatter plot** consists of n points on an (x, y)-plane, with x- and y-coordinates representing the two recorded variables.

Example 8.20 (Antivirus maintenance). Protection of a personal computer largely depends on the frequency of running antivirus software on it. One can set to run it every day, once a week, once a month, etc.

During a scheduled maintenance of computer facilities, a computer manager records the number of times the antivirus software was launched on each computer during 1 month (variable X) and the number of detected worms (variable Y). The data for 30 computers are in the table (data set `Antivirus`).

X	30	30	30	30	30	30	30	30	30	30	30	15	15	15	10
Y	0	0	1	0	0	0	1	1	0	0	0	0	1	1	0

X	10	10	6	6	5	5	5	4	4	4	4	4	1	1	1
Y	0	2	0	4	1	2	0	2	1	0	1	0	6	3	1

Is there a connection between the frequency of running antivirus software and the number of worms in the system? A scatter plot of these data is given in Figure 8.11a. It clearly shows that the number of worms reduces, in general, when the antivirus is employed more frequently. This relationship, however, is not certain because no worm was detected on some "lucky" computers although the antivirus software was launched only once a week on them. ◊

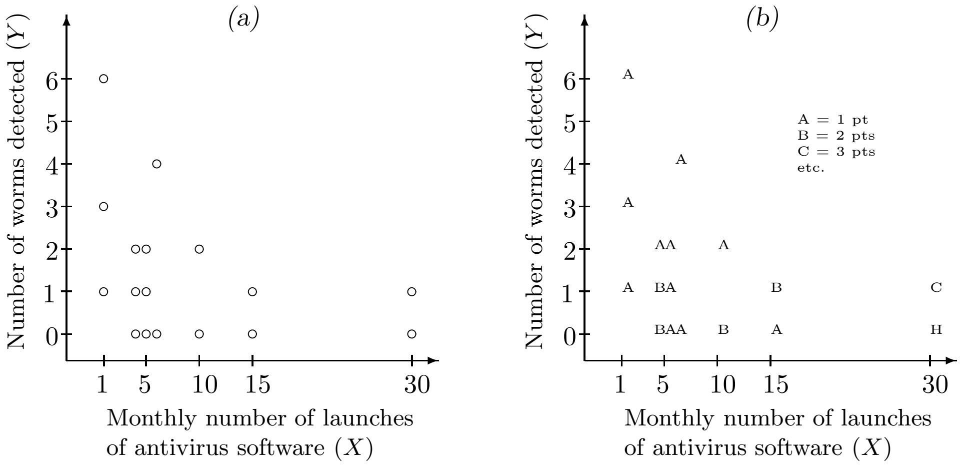


FIGURE 8.11: *Scatter plots for Examples 8.20 and 8.21.*

Example 8.21 (Plotting identical points). Looking at the scatter plot in Figure 8.11a, the manager in Example 8.20 realized that a portion of data is hidden there because there are identical observations. For example, no worms were detected on 8 computers where the antivirus software is used daily (30 times a month). Then, Figure 8.11a may be misleading.

When the data contain identical pairs of observations, the points on a scatter plot are often depicted with either numbers or letters ("A" for 1 point, "B" for two identical points, "C" for three, etc.). You can see the result in Figure 8.11b. ◊

When we study time trends and development of variables over time, we use **time plots**. These are scatter plots with x-variable representing time.

Example 8.22 (World population). For example, here is how the world population increased between 1950 and 2019 (Figure 8.12). We can clearly see that the population increases at an almost steady rate. The actual data are given in Table 11.1 on p. 376 and in data set `PopulationWorld`. Later, in Chapter 11, we'll learn how to estimate trends seen on time plots and scatter plots and even make forecasts for the future. ◊

R notes

In R, `mean`, `median`, `var`, `sd`, `min`, `max`, `range`, `quantile`, and `length` are used to determine the sample mean, median, variance, standard deviation, minimum, maximum, range, quantiles, and the sample size of an observed variable. A `summary` command calculates the whole five-point summary for a boxplot – minimum, maximum, three quartiles, and additionally, the mean. The boxplot itself can be drawn simply by typing `boxplot(X)` for an observed variable X. Try this for the whole data set, `boxplot(Dataset)`, and you will obtain beautiful parallel boxplots (but notice that they will appear all on the same scale).

Similarly, `summary` can also be applied to the entire data set, and this is often used to gain quick information, as the first acquaintance with data.

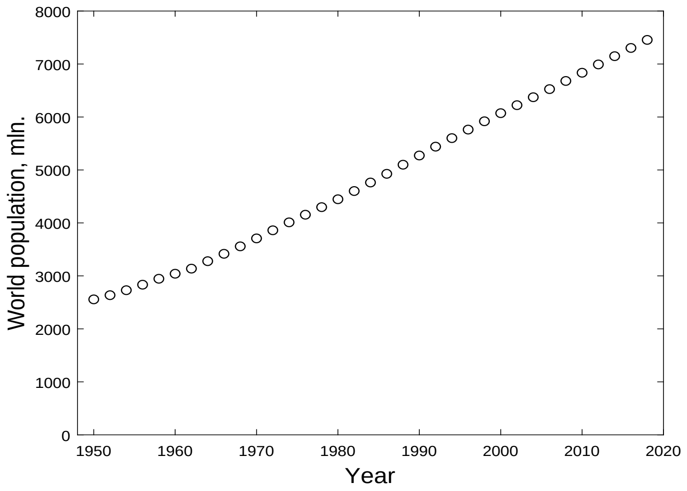


FIGURE 8.12: *Time plot of the world population in 1950–2019.*

For graphical statistics, `hist` is used for histograms, `stem` for stem-and-leaf plots, and `plot` for scatter plots. In these plots, you can optionally set a color and thickness. For example, `plot(X,Y,col="blue",lwd=3)` will produce a plot of Y vs X in blue points that are three times thicker (and therefore, will look brighter) than those without an `lwd` option. To plot several variables on the same scatter plot, you can add plots to an existing one by a command `points`. For example, `points(X,Z,col="green")` will superimpose a green scatter plot of Z vs X on an earlier blue scatter plot of Y vs X. Applying this command to the whole data set, `plot(Dataset)`, will produce a *scatter plot matrix*. That is, R will draw a square matrix of scatter plots of all the variables against each other variable.

MATLAB notes

For the standard descriptive statistics, MATLAB has ready-to-go commands `mean, median, std, var, min`, and `max`. Also, `quantile(X,p)` and `prctile(X,100*p)` give sample p-quantile $\widehat{q}_p$ which is also the $100p$-percentile $\widehat{\pi}_{100p}$, and `iqr(X)` returns the interquartile range.

Graphing data sets is also possible in one command. To see a histogram, write `hist(x)` or `hist(x,n)` where x is the data, and n is the desired number of bins. For a boxplot, enter `boxplot(X)`. By the way, if X is a matrix, then you will obtain parallel boxplots. To draw a scatter plot or a time plot, write `scatter(X,Y)` or just `plot(X,Y)`. You can nicely choose the color and symbol in this command. For example, `plot(X,Y,'r*',X,Z,'bd')` will plot variables Y and Z against variable X, with Y-points marked by red stars and Z-points marked by blue diamonds.