CENG 222
Probability and Statistics

Introduction to Statistics

Nesli Erdoğmuş & Burak Galip Aslan

Contents

- **8.1 Population and sample, parameters and statistics**
- 8.2 Simple Descriptive statistics
  - 8.2.1 Mean
  - 8.2.2 Median
  - 8.2.3 Quantiles, percentiles and quartiles
  - 8.2.4 Variance and standard deviation
  - 8.2.5 Standard errors of estimates
  - 8.2.6 Interquartile range
- 8.3 Graphical statistics
  - 8.3.1 Histogram
  - 8.3.3 Box plot
  - 8.3.4 Scatter plots and time plots

Introduction to statistics

The previous chapters we covered taught us

- to analyze problems and systems involving uncertainty, and
- to find probabilities, expectations, and other characteristics for a variety of situations.

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.

Introduction to statistics

Much more often parameters are **NOT** known.

Then, how can one apply the knowledge 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.

Introduction to statistics

In the remainder of this course, we will 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.

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**.

Population and sample, parameters and statistics

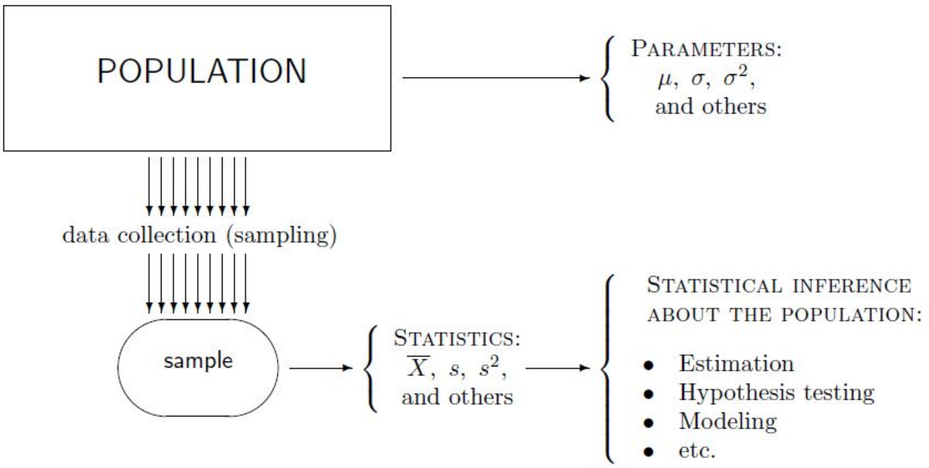
- The only way to know these **parameters** is to measure the entire **population**, i.e., to conduct a *census*.
- Instead of a census, we may collect data in a form of a random **sample** from a **population**. This is our data.
- We can measure them, perform calculations, and estimate the unknown parameters of the population up to a certain measurable degree of accuracy.

NOTATION

θ = population parameter

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

Population and sample, parameters and statistics


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.

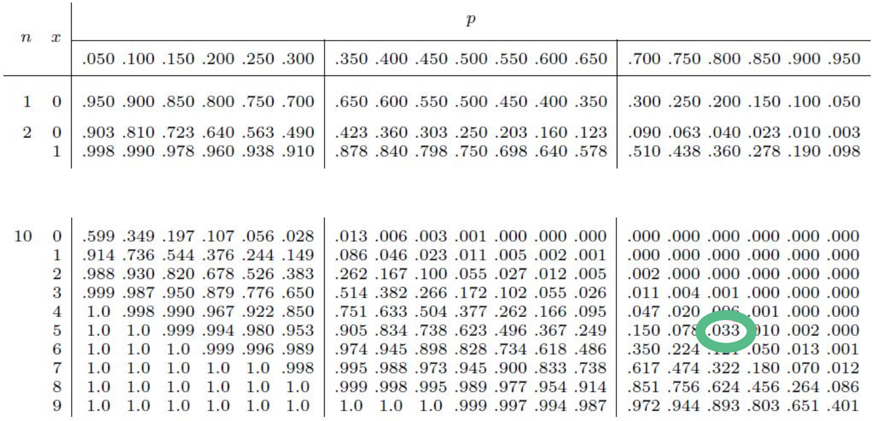
Population and sample, parameters and statistics

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.

For instance, what is the probability of 5 out of 10 sampled customers are satisfied? Is it zero?

Population and sample, parameters and statistics

Example 8.1 (Customer satisfaction). NO! The probability of 5 out of 10 sampled customers are satisfied is 0.033! (Table A2)



n	x	p																		
		.050	.100	.150	.200	.250	.300	.350	.400	.450	.500	.550	.600	.650	.700	.750	.800	.850	.900	.950
1	0	.950	.900	.850	.800	.750	.700	.650	.600	.550	.500	.450	.400	.350	.300	.250	.200	.150	.100	.050
2	0	.903	.810	.723	.640	.563	.490	.423	.360	.303	.250	.203	.160	.123	.090	.063	.040	.023	.010	.003
	1	.998	.990	.978	.960	.938	.910	.878	.840	.798	.750	.698	.640	.578	.510	.438	.360	.278	.190	.098
10	0	.599	.349	.197	.107	.056	.028	.013	.006	.003	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000
	1	.914	.736	.544	.376	.244	.149	.086	.046	.023	.011	.005	.002	.001	.000	.000	.000	.000	.000	.000
	2	.988	.930	.820	.678	.526	.383	.262	.167	.100	.055	.027	.012	.005	.002	.000	.000	.000	.000	.000
	3	.999	.987	.950	.879	.776	.650	.514	.382	.266	.172	.102	.055	.026	.011	.004	.001	.000	.000	.000
	4	1.0	.998	.990	.967	.922	.850	.751	.633	.504	.377	.262	.166	.095	.047	.020	.006	.001	.000	.000
	5	1.0	1.0	.999	.994	.980	.953	.905	.834	.738	.623	.496	.367	.249	.150	.078	.033	.010	.002	.000
	6	1.0	1.0	1.0	.999	.996	.989	.974	.945	.898	.828	.734	.618	.486	.350	.224	.121	.050	.013	.001
	7	1.0	1.0	1.0	1.0	1.0	.998	.995	.988	.973	.945	.900	.833	.738	.617	.474	.322	.180	.070	.012
	8	1.0	1.0	1.0	1.0	1.0	1.0	.999	.998	.995	.989	.977	.954	.914	.851	.756	.624	.456	.264	.086
	9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	.999	.997	.994	.987	.972	.944	.893	.803	.651	.401

Population and sample, parameters and statistics

Example 8.1 (Customer satisfaction). NO! The probability of 5 out of 10 sampled customers are satisfied is 0.033!

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 and sample, parameters and statistics

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.

Population and sample, parameters and statistics

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.

Population and sample, parameters and statistics

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 nonsampling errors, if they are handled properly. The fact is, in such cases, we cannot assume independence.

Population and sample, parameters and statistics

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 <u>higher chance to be sampled</u> than Mr. X. Unequal probabilities have to be taken into account, otherwise a non-sampling error will inevitably occur.

Population and sample, parameters and statistics

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.

Population and sample, parameters and statistics

Example 8.5 (Presidential Election of 1936).

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.

Population and sample, parameters and statistics

In this course, 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.

Population and sample, parameters and statistics

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**.

Contents

- 8.1 Population and sample, parameters and statistics
- **8.2 Simple descriptive statistics**
  - **8.2.1 Mean**
  - **8.2.2 Median**
  - **8.2.3 Quantiles, percentiles and quartiles**
  - **8.2.4 Variance and standard deviation**
  - **8.2.5 Standard errors of estimates**
  - **8.2.6 Interquartile range**
- 8.3 Graphical statistics
  - 8.3.1 Histogram
  - 8.3.3 Box plot
  - 8.3.4 Scatter plots and time plots

Simple descriptive statistics

Suppose a good random sample

$$\mathbf{S} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$$

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 (dataset CPU):

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

Simple descriptive statistics

What information do we get from this collection of numbers?

We know that $\mathcal{X}$, 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 $\mathcal{X}$.

Simple descriptive statistics

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.

Simple descriptive statistics

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 before, 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.

Mean

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

DEFINITION 8.3 Sample mean $\bar{X}$ is the arithmetic average,

$\bar{X} = (\mathbf{X}_1 + \mathbf{X}_2 + \ldots + \mathbf{X}_n) / n$

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

Mean

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

DEFINITION 8.3 Sample mean $\bar{X}$ is the arithmetic average,

$\bar{X} = (\mathbf{X}_1 + \mathbf{X}_2 + \ldots + \mathbf{X}_n) / n$

- 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).

Mean

DEFINITION 8.3 An estimator θ is **unbiased** for a parameter θ if its expectation equals the parameter,

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

for all possible values of θ.

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

Unbiasedness means that in a long run, collecting a large number of samples and computing θ from each of them, on the average we hit the unknown parameter θ exactly.

Mean

DEFINITION 8.3 An estimator θ is **unbiased** for a parameter θ if its expectation equals the parameter,

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

for all possible values of θ.

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

Sample mean estimates μ unbiasedly because its expectation is

$$\mathbf{E}(\bar{X}) = \mathbf{E}((\mathbf{X}_1 + \mathbf{X}_2 + . . . + \mathbf{X}_n) / n) = (\mathbf{E}(\mathbf{X}_1) + \mathbf{E}(\mathbf{X}_2) + ... + \mathbf{E}(\mathbf{X}_n)) / n$$

$$= (\mu + \mu + ... + \mu) / n = n\mu / n = \mu$$

Mean

DEFINITION 8.4 An estimator θ 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,

$$P\{|\theta - \theta| > \varepsilon\} \rightarrow 0 \text{ as } n \rightarrow \infty$$

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

Mean

The variance of $\overline{X}$,

$\mathrm{Var}(\overline{X}) = \mathrm{Var}((\mathbf{X}_1 + \mathbf{X}_2 + . . . + \mathbf{X}_n) / n)$

$= (\mathrm{Var}(\mathbf{X}_1) + \mathrm{Var}(\mathbf{X}_2) + \ldots + \mathrm{Var}(\mathbf{X}_n)) / n^2)$

$= (\sigma^2 + \sigma^2 + \ldots + \sigma^2) / n^2 = n\sigma^2 / n^2 = \sigma^2 / n$

Using Chebyshev’s inequality:

$P \{ |\overline{X} - \mu| > \varepsilon \} \leq \mathrm{Var}(\overline{X}) / \varepsilon^2 = (\sigma^2 / n) / \varepsilon^2 \rightarrow 0$ as $n \rightarrow \infty$

- 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.

Mean

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

$$Z = (\overline{X} - \mathbf{E}(\overline{X})) / \mathrm{Std}(\overline{X}) = (\overline{X} - \mu) / (\sigma\sqrt{n})$$

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

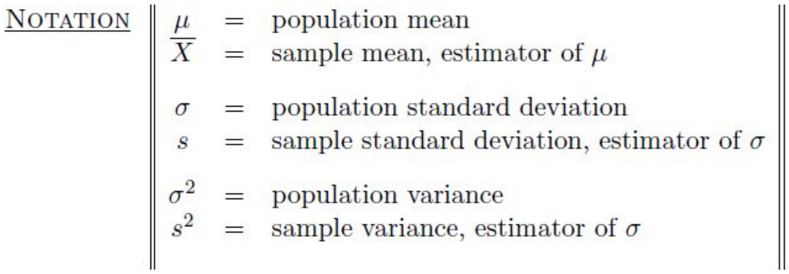
Mean

Example 8.7 (CPU times) Looking at the previously given CPU data, we estimate the average (expected) CPU time μ by

$\overline{X} = (70 + 36 + \ldots + 56 + 19) / 30 = 1447 / 30 = 48.2333$

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

Mean



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

Median

One disadvantage of a sample mean is its s**ensitivity 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.

Median

DEFINITION 8.6

Median means a “central” value.

Sample median $\hat{M}$ 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

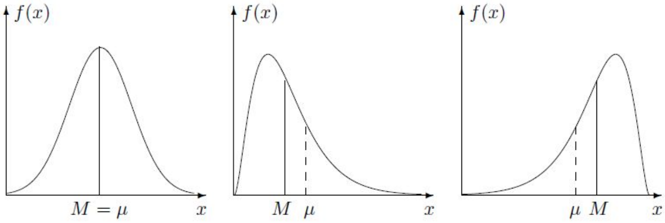
$$\begin{cases} P\{ X > M \} \le 0.5 \\ P\{ X < M \} \le 0.5 \end{cases}$$

8.2.2 Median

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:

- Symmetric distribution ⇒ $M = \mu$
- Right-skewed distribution ⇒ $M < \mu$
- Left-skewed distribution ⇒ $M > \mu$



8.2.2 Median

Computation of a population median

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

$$\begin{cases} P\{ X > M \} = 1 - F(M) \le 0.5 \\ P\{ X < M \} = \quad F(M) \quad \le 0.5 \end{cases}$$

$$\Rightarrow F(M) = 0.5$$

8.2.2 Median

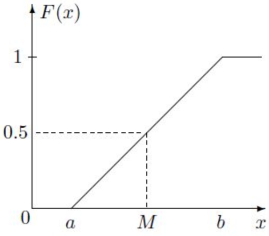
Computation of a population median

Example 8.8 (Uniform) Uniform(a, b) distribution has a cdf

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

Solving the equation

$$F(M) = (M - a) / (b - a) = 0.5,$$



we get

$$M = (a+b) / 2.$$

M = μ because the Uniform distribution is symmetric.

8.2.2 Median

Computation of a population median

Example 8.8 (Exponential) Exponential(λ) distribution has a cdf

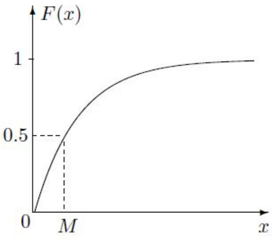
$$F(x) = 1 - e^{-\lambda x} \text{ for } x > 0.$$

Solving the equation

$$F(M) = 1 - e^{-\lambda M} = 0.5$$

we get

$$M = \ln 2 / \lambda = 0.6931 / \lambda.$$

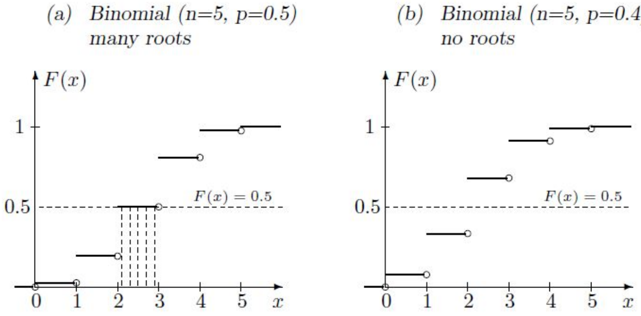


$M < \mu = 1/\lambda$ because Exponential distribution is right-skewed.

8.2.2 Median

Computation of a population median

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

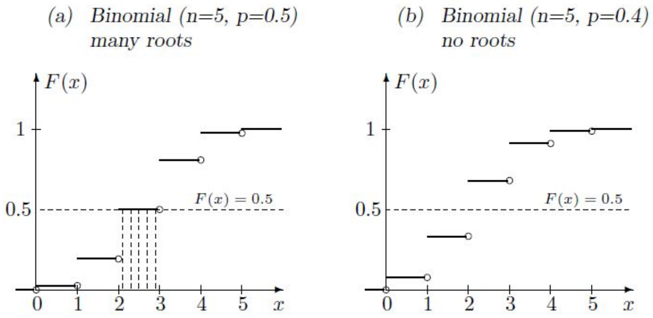

(a) Binomial (n=5, p=0.5) many roots

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

8.2.2 Median

Computation of a population median

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. Often the middle of this interval is reported as the median.

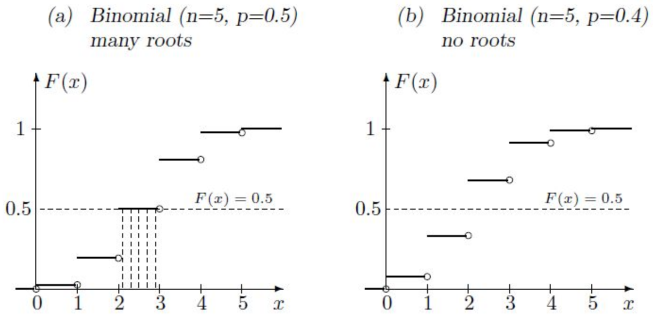

(a) Binomial (n=5, p=0.5) many roots

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

8.2.2 Median

Computation of a population median

In the second case, the smallest x with F(x) ≥ 0.5 is the median. It is the value of x where the cdf jumps over 0.5.

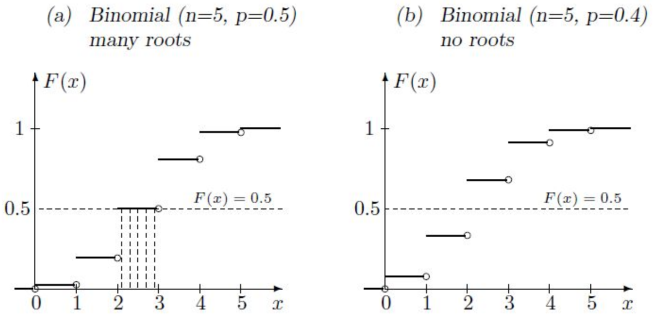

(a) Binomial (n=5, p=0.5) many roots

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

8.2.2 Median

Computation of a population median

- For the Binomial distribution with n = 5 and p = 0.5, any number of the interval (2,3) is a median.
- For the Binomial distribution with n = 5 and p = 0.4, there is no value of $\boldsymbol{x}$ where F ($\boldsymbol{x}$) = 0.5. Then, M = 2 is the median.



8.2.2 Median

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.

8.2.2 Median

Computing sample medians

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

- If n is odd, the ((n+1)/2)-th smallest observation is a median.
- If n is even, any number between the (n/2)-th smallest and ((n+2)/2)-th smallest observations is a median.

8.2.2 Median

Computing sample medians

Example 8.12 (Median CPU time). Let's compute the median of n = 30 CPU times from the previously given data. First, order the data,

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

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

8.2.2 Median

Computing sample medians

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.

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

Sample medians are easy to compute. In fact, no computations are needed, only the ordering.

8.2.2 Median

Computing sample medians

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

DEFINITION 8.7 A **p-quantile** of a population is such a number x that solves equations

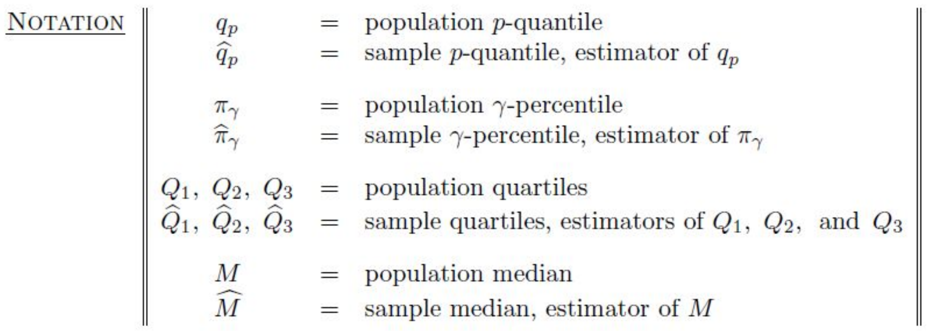
$$\begin{cases} P\{ X > M \} \leq \ p \\ P\{ X < M \} \leq 1\text{-}p \end{cases}$$

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.

8.2.3 Quantiles, percentiles and quartiles



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

8.2.3 Quantiles, percentiles and quartiles

- Quartiles 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.
- Quantiles, quartiles, and percentiles are related as follows.

$$q_p = \pi_{100p}$$

$$Q_1 = \pi_{25} = q_{1/4} \qquad Q_3 = \pi_{75} = q_{3/4}$$

$$M = Q_2 = \pi_{50} = q_{1/2}$$

- Sample statistics are of course in a similar relation.
- Computing quantiles is very similar to computing medians.

8.2.3 Quantiles, percentiles and quartiles

Example 8.14 (Sample quartiles). Let us compute the 1st and 3rd quartiles of CPU times. We look at the ordered sample

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

Q1. 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, Q1 = 34.

Q3. Similarly, Q3 is the 23rd smallest element, Q3 = 59.

8.2.4 Variance and standard deviation

Statistics introduced so far 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.

8.2.4 Variance and standard deviation

DEFINITION 8.8 For a sample ($\mathbf{X}_1$, $\mathbf{X}_2$, . . . , $\mathbf{X}_n$), a **sample variance** is defined as

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

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

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 $\sigma = \mathrm{Std}(X)$.

8.2.4 Variance and standard deviation

- 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.

8.2.4 Variance and standard deviation

Computation

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

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

Can you prove that this is equivalent to the sample variance formulation in DEFINITION 8.8?

8.2.4 Variance and standard deviation

Example 8.16 (CPU time) For the previously given CPU data, we have computed $\bar{X} = 48.2333$. Now, we can compute the sample variance as:

$s^2 = ((70 - 48.2333)^2 + \ldots + (19 - 48.2333)^2) / (30 - 1)$

$\quad = 20391 / 29 = 703.1506$ (sec^2)

or

$s^2 = 70^2 + \ldots + 19^2 - (30)(48.2333)^2 / (30 - 1)$

$\quad = (90185 - 69794) / 29 = 703.1506$ (sec^2)

Hence, $s = \sqrt{703.1506} = 26.1506$ (sec)

8.2.4 Variance and standard deviation

A seemingly strange coefficient $1/(n-1)$ ensures that s^2 is an **unbiased** estimator of σ^2.

- Suppose for a moment that the population mean $\mu = \mathbf{E}(X) = 0$.
  $\mathbf{E}(X_i^2) = \mathrm{Var}(X_i) = \sigma^2$ and $E(\bar{X}^2) = \mathrm{Var}(\bar{X}) = \sigma^2/n$.
  $\mathbf{E}(s^2) = (\mathbf{E}(\sum X_i^2) - nE(\bar{X}^2)) / (n-1) = (n\sigma^2 - \sigma^2 / (n-1) = \sigma^2$
- If $\mu \neq 0$, consider auxiliary variables $Y_i = X_i - \mu$. Variances don't depend on constant shifts, therefore, Y_i have the same variance as X_i. Their sample variances are equal too.
  $s_Y^2 = \sum(Y_i - \bar{Y})^2/(n-1) = \sum(X_i - \mu - (\bar{X} - \mu))^2/(n-1) = \sum(X_i - \bar{X})^2/(n-1) = s_x^2$
  $\mathbf{E}(s_Y^2) = \mathbf{E}(s_X^2) = \sigma_X^2 = \sigma_Y^2$

8.2.4 Variance and standard deviation

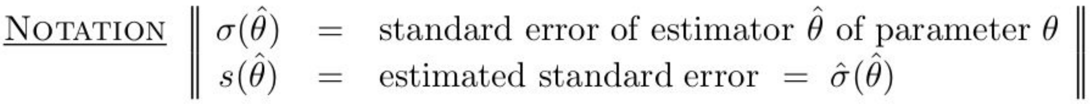
Similarly to $\bar{X}$, it can be shown that under rather mild assumptions, sample variance and sample standard deviation are also **consistent** and **asymptotically Normal**.

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 θ is its standard deviation, $\sigma(\theta) = \mathrm{Std}(\theta)$.

As a measure of variability, standard errors show precision and reliability of estimators.



NOTATION

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

8.2.5 Standard errors of estimates

- 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.

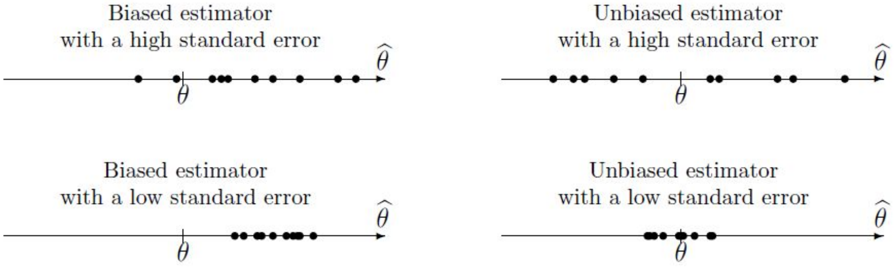


FIGURE 8.5: *Bias and standard error of an estimator. In each case, the dots represent parameter estimators $\widehat{\theta}$ obtained from 10 different random samples.*

8.2.5 Standard errors of estimates

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 $\theta = \bar{X}$. We already know that the standard error of this estimator is $\sigma(\bar{X}) = \sigma/\sqrt{n}$, and it can be estimated by $s(\bar{X}) = s/\sqrt{n}$.

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 $\bar{X}$ 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.

8.2.6 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

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

8.2.6 Interquartile range

Detection of outliers

A “rule of thumb” for identifying outliers is the rule of **1.5(IQR)**. Measure 1.5(Q_3 − 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.

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. It is so unlikely to see a value of x outside of this range that such an observation may be treated as an outlier.

8.2.6 Interquartile range

Example 8.18 (Any outlying CPU times?). Can we suspect that the CPU data set has any outliers? Compute

$\widehat{IQR} = Q3 - Q1 = 59 - 34 = 25$

and measure 1.5 interquartile ranges from each quartile:

$Q1 - 1.5(\widehat{IQR}) = 34 - 37.5 = -3.5$

$Q3 + 1.5(\widehat{IQR}) = 59 + 37.5 = 96.5.$

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

8.2.6 Interquartile range

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.

8.2.6 Interquartile range

Handling of outliers

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

- 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.

Contents

- 8.1 Population and sample, parameters and statistics
- 8.2 Simple Descriptive statistics
  - 8.2.1 Mean
  - 8.2.2 Median
  - 8.2.3 Quantiles, percentiles and quartiles
  - 8.2.4 Variance and standard deviation
  - 8.2.5 Standard errors of estimates
  - 8.2.6 Interquartile range
- **8.3 Graphical statistics**
  - **8.3.1 Histogram**
  - **8.3.3 Box plot**
  - **8.3.4 Scatter plots and time plots**

8.3 Graphical statistics

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.

8.3 Graphical statistics

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

- histograms,
- 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.

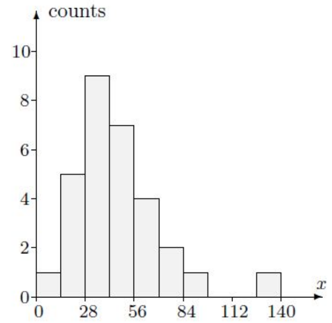
A **frequency histogram** consists of columns, one for each bin, whose height is determined by the *number* of observations in the bin.

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.

8.3.1 Histogram

The sample of CPU times 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 between 14 and 28
- 9 observations between 28 and 42
- 7 observations between 42 and 56
- 4 observations between 56 and 70
- …

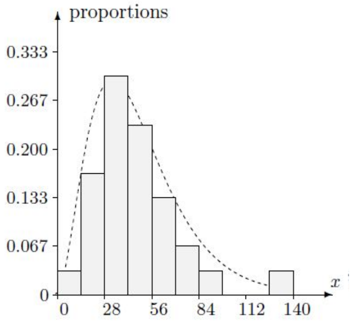


Using this for column heights, a (frequency) histogram of CPU times is then constructed.

8.3.1 Histogram

The sample of CPU times 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 between 14 and 28
- 9 observations between 28 and 42
- 7 observations between 42 and 56
- 4 observations between 56 and 70
- …



A relative frequency histogram is only different in the vertical scale. Each count is now divided by the sample size n = 30.

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

8.3.1 Histogram

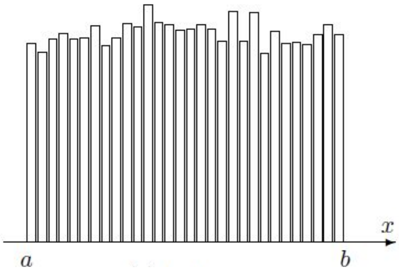
The following information can be drawn from the histograms:

- 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, only Gamma distribution has a similar shape (sketched with a dashed curve).
- 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.

8.3.1 Histogram

How else may histograms look like?

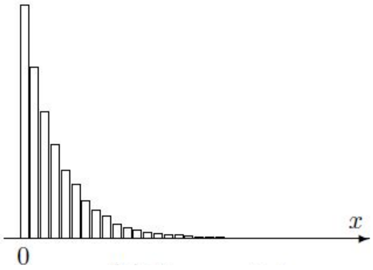
The distribution is almost symmetric, and columns have almost the same height. Slight differences can be attributed to the randomness of our sample, i.e., the *sampling error*. The histogram suggest a Uniform or Discrete Uniform distribution between a and b.



8.3.1 Histogram

How else may histograms look like?

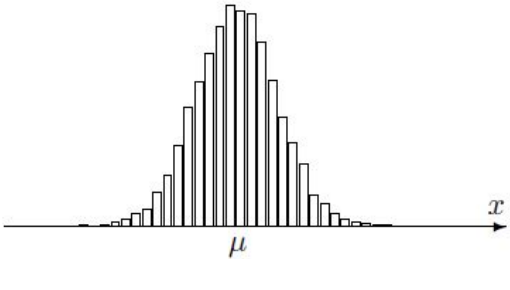
The distribution is heavily right-skewed, column heights decrease exponentially fast. This sample should come from an Exponential distribution, if variables are continuous, or from Geometric, if they are discrete.



8.3.1 Histogram

How else may histograms look like?

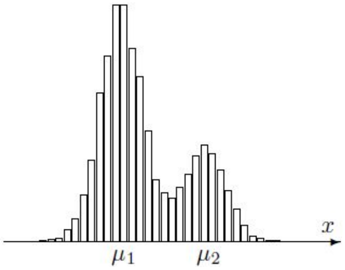
The distribution is symmetric, with very quickly vanishing “tails.” Its bell shape reminds a Normal density. We can locate the center μ of a histogram and conclude that this sample is likely to come from a Normal distribution with a mean close to μ.



8.3.1 Histogram

Mixtures

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 .



8.3.1 Histogram

Mixtures

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, write the (unconditional) cdf as

$$F(\boldsymbol{x}) = p_1F_1(\boldsymbol{x}) + p_2F_2(\boldsymbol{x}) + . . . ,$$

and study the whole population at once.

8.3.1 Histogram

The choice of bins

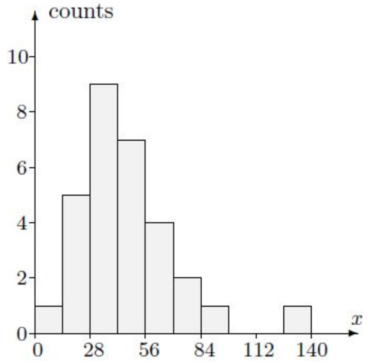
Experimenting with histograms, you can notice that their shape may depend on the choice of bins. One can hear various rules of thumb about a good choice of bins, but in general,

- 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.

8.3.1 Histogram

The choice of bins

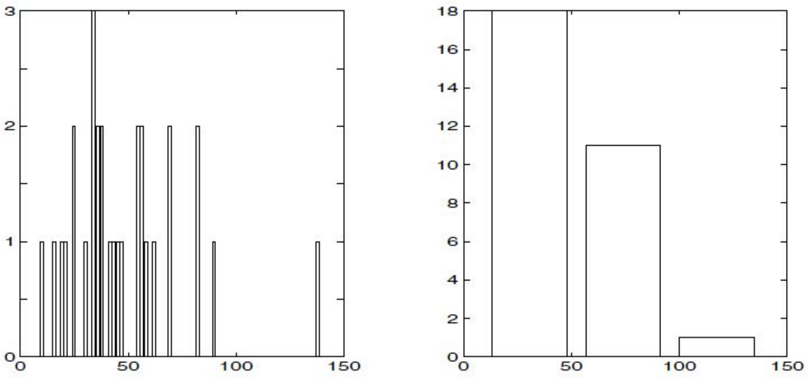
We simply divided the range of CPU data into 10 equal intervals, 14 sec each, and apparently this was sufficient for drawing important conclusions.

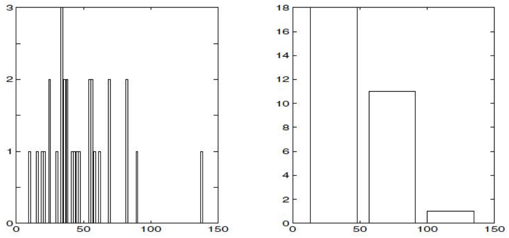


8.3.1 Histogram

The choice of bins

As two extremes, consider the following histograms constructed from the same CPU data.





8.3.1 Histogram

The choice of bins

As two extremes, consider the following histograms 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.

8.3.3 Boxplot

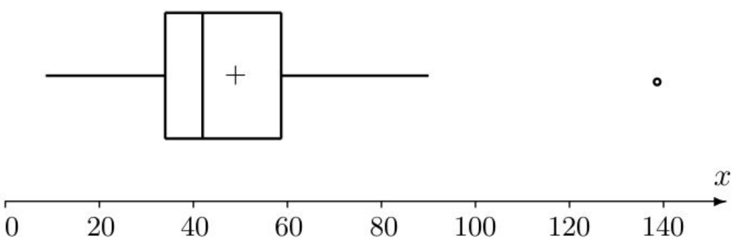
The main descriptive statistics of a sample can be represented graphically by a **boxplot**. To construct a boxplot, we draw a box between the first and the third quartiles, a line inside a box for a median, and extend whiskers to the smallest and the largest observations, thus representing a so-called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

$$\textbf{five-point summary} = (\min X_i\ ,\ \hat{Q}_1\ ,\ \hat{M}\ ,\ \hat{Q}_3\ ,\ \max X_i)$$

Often a sample mean 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.

8.3.3 Boxplot

- The mean and five-point summary of CPU times were found as, X = 48.2333, minX_i=9, $\hat{Q}_1$=34, $\hat{M}$=42.5, $\hat{Q}_3$=59, maxX_i=139.
- We also know that X=139 is more than 1.5$(\widehat{IQR})$ suspect that it may be an outlier.



8.3.3 Boxplot

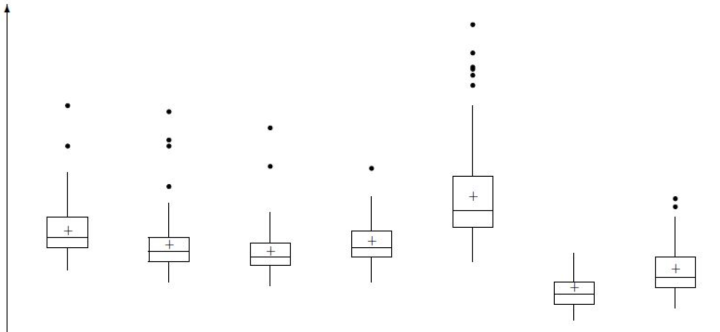
Parallel boxplots

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.

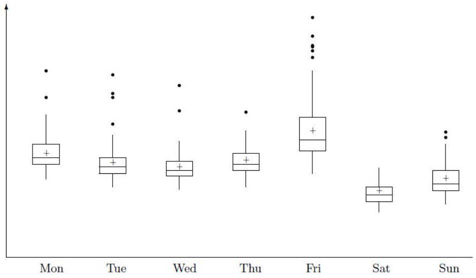
8.3.3 Boxplot

Parallel boxplots

The amount of internet traffic handled by a certain center during a week.



8.3.3 Boxplot



Parallel boxplots

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.

8.3.3 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, etc.

To study the relationship, both variables are measured on each sampled item. Then, a scatter plot consists of n points on an (x,y)-plane, with x- and y-coordinates representing the two recorded variables.

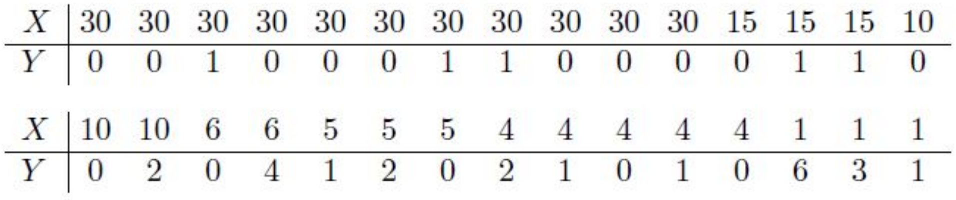
8.3.3 Scatter plots and time plots

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 $\boldsymbol{\mathcal{X}}$) and the number of detected worms (variable $\boldsymbol{\mathcal{Y}}$).

8.3.3 Scatter plots and time plots

Example 8.20 (Antivirus maintenance). 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

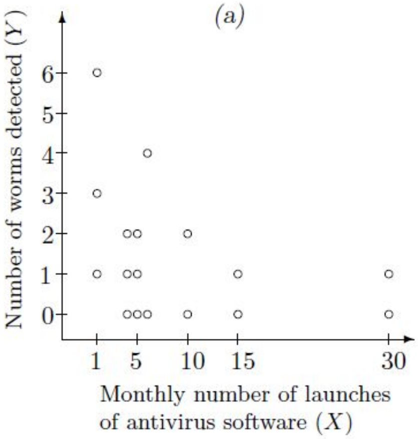
X	10	10	6	6	5	5	5	4	4	4	4	4	1	1	1
Y	0	2	0	4	1	2	0	2	1	0	1	0	6	3	1

- Is there a connection between the frequency of running antivirus software and the number of worms in the system?

8.3.3 Scatter plots and time plots

Example 8.20 (Antivirus maintenance).

A scatter plot of these data is given.



8.3.3 Scatter plots and time plots

Example 8.20 (Antivirus maintenance).

- 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.

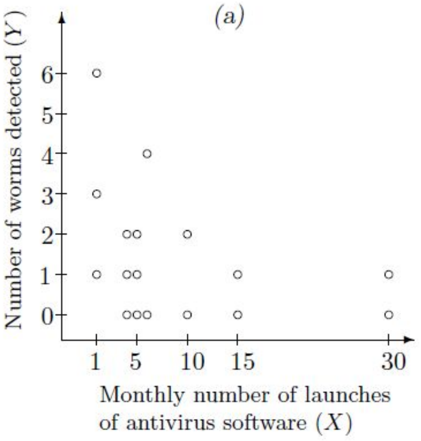
X	30	30	30	30	30	30	30	30	30	30	30	15	15	15	10
Y	0	0	1	0	0	0	1	1	0	0	0	0	1	1	0

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

8.3.3 Scatter plots and time plots

Example 8.21 (Plotting identical points).

Looking at this scatter plot, the manager 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, this figure may be misleading.



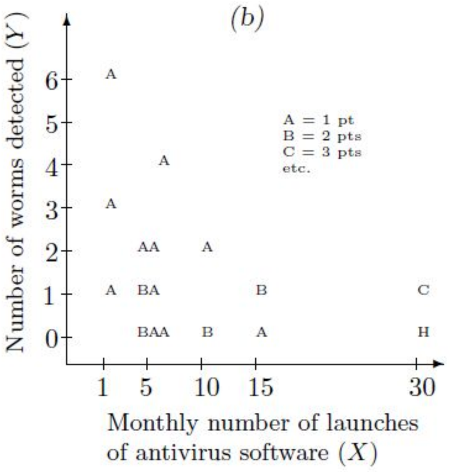
X	30	30	30	30	30	30	30	30	30	30	30	15	15	15	10
Y	0	0	1	0	0	0	1	1	0	0	0	0	1	1	0

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

8.3.3 Scatter plots and time plots

Example 8.21 (Plotting identical points).

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.



8.3.3 Scatter plots and time plots

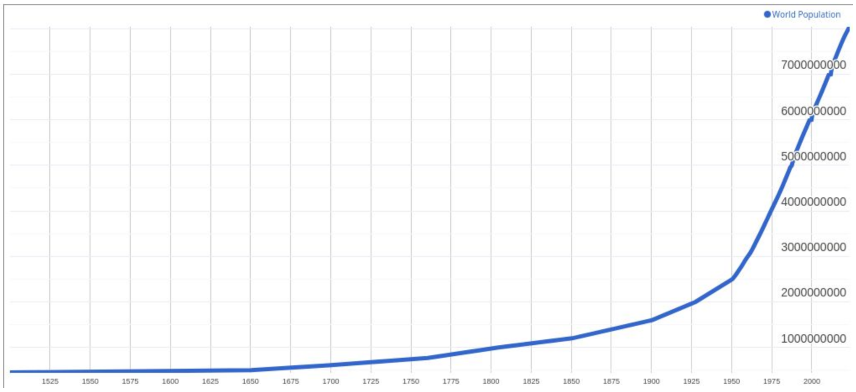
Time plots

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

8.3.3 Scatter plots and time plots

Example 8.22 (World population).

For example, here is how the world population increased between 1500 and 2023.

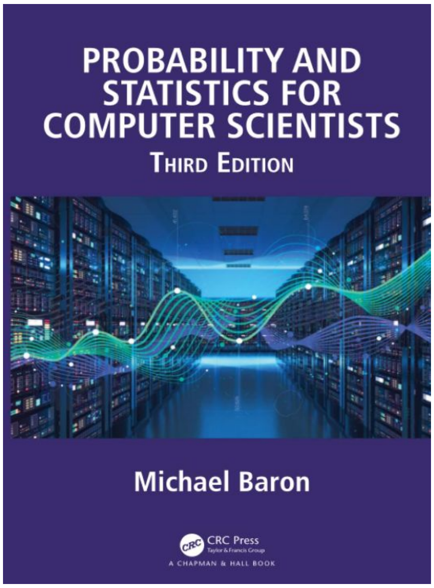


https://www.worldometers.info/world-population/

References

WIKIPEDIA

Google images



PROBABILITY AND STATISTICS FOR COMPUTER SCIENTISTS

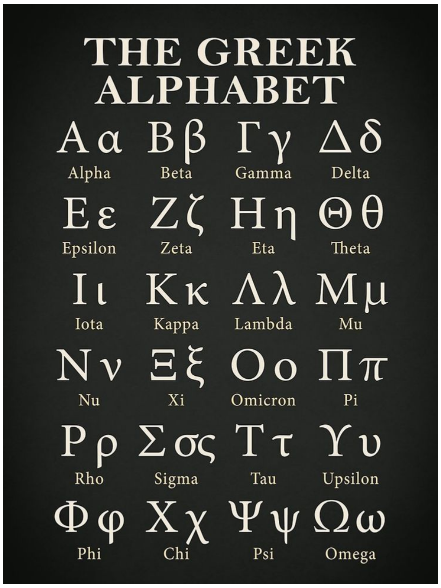
THIRD EDITION

Michael Baron

CRC Press Taylor & Francis Group

A CHAPMAN & HALL BOOK

Appendix



THE GREEK ALPHABET

Α α	Β β	Γ γ	Δ δ
Alpha	Beta	Gamma	Delta
Ε ε	Ζ ζ	Η η	Θ θ
Epsilon	Zeta	Eta	Theta
Ι ι	Κ κ	Λ λ	Μ μ
Iota	Kappa	Lambda	Mu
Ν ν	Ξ ξ	Ο ο	Π π
Nu	Xi	Omicron	Pi
Ρ ρ	Σ σς	Τ τ	Υ υ
Rho	Sigma	Tau	Upsilon
Φ φ	Χ χ	Ψ ψ	Ω ω
Phi	Chi	Psi	Omega