Lecture 01. Statistics and probability theory.

Statistics is the science of learning from data, involving the collection, organization, analysis, interpretation, and communication of data to make informed decisions and draw conclusions about the world, often in the presence of uncertainty and variation.

Key Aspects of Statistics:

- Data Collection: Gathering raw data from primary or secondary sources using methods like surveys or existing databases.
- Data Organization: Arranging and summarizing data into a structured format, such as tables or graphs, to make it more understandable.
- Data Analysis: Employing statistical methods to explore the characteristics of the data, identify patterns, and understand relationships within it.
- Interpretation and Inference: Using data from samples to make generalizations or draw conclusions about larger populations, a process known as statistical inference.

Why Statistics is Important:

- Informed Decision-Making: Statistics provides the tools to make evidence-based decisions in various fields, from business and healthcare to scientific research.
- Understanding Uncertainty: The field provides methods to quantify and manage uncertainty, a fundamental aspect of real-world data.
- Interdisciplinary Applications: Statistics is a highly interdisciplinary field used in nearly every scientific discipline, helping to solve complex problems.

1. Data Collection:

**Random processes** or **experiments** with random results:

**Example 1** Let us consider an experiment which is repeated $N$ times always under the same conditions. As a result we obtain a variety of values $x_1, x_2, x_3, \ldots, x_{N-1}, x_N$ of some quantity. In this way we cannot predict the result of our experiment, however we can notice that some results let us say

$x_n$, appear with some frequency $f_n$. We will assign the rate $\frac{f_n}{N}$ to $x_n$ and call it a chance of appearing of $x_n$.
The situation in the example is different to the case when we always get the same result , $x_1 = x_2 = x_3, = \cdots = x_{N-1} = x_N$. This mean that we are able to predict the result of the experiment and it is called **deterministic**.

**Example 2** Let us consider a population of $N$ people (we call them individuals). We assign to each individual its height $x$. We draw an individual from the population and record its height $x$. This experiment we repeat $N$ times. Finally as a results we get a sequence of values $x_1, x_2, x_3, \ldots, x_{N-1}, x_N$.

In applying statistics to a scientific, industrial, or social problem, it is conventional to begin with a statistical population or a statistical model to be studied. Populations can be diverse groups of people or objects such as "all people living in a country" or "every atom composing a crystal". Statistics deals with every aspect of data, including the planning of data collection in terms of the design of surveys and experiments. When census data (comprising every member of the target population) cannot be collected, statisticians collect data by developing specific experiment designs and survey samples. Representative sampling assures that inferences and conclusions can reasonably extend from the sample to the population as a whole. An experimental study involves taking measurements of the system under study, manipulating the system, and then taking additional measurements using the same procedure to determine if the manipulation has modified the values of the measurements. In contrast, an observational study does not involve experimental manipulation.

2. Data organization, the language of statistics:

- INDIVIDUALS AND VARIABLES

**Individuals** are the objects described in a set of data. Individuals are sometimes people. When the objects that we want to study are not people, we often call them cases.

A **variable** is any characteristic of an individual. A variable can take different values for different individuals.

CATEGORICAL AND QUANTITATIVE VARIABLES, distribution of variable

A **categorical variable** places an individual into one of two or more groups

or categories.

A **quantitative variable** takes numerical values for which arithmetic operations such as adding and averaging make sense.

The distribution of a variable gives an answer to the question: let $(a, b)$ be an arbitrary interval in $\mathbb{R}$, how many values it takes in the interval $(a, b)$

**Example**
1.2 Data for students in a statistics class. The tables below show part of a data set for students enrolled in an introductory statistics class. Each row gives the data on one student. The values for the different variables are in the columns. This data set has eight variables. ID is an identifier for each student. Exam1, Exam2, Homework, Final, and Project give the points earned, out of a total of 100 possible, for each of these course requirements. Final grades are based on a possible 200 points for each exam and the final, 300 points for Homework, and 100 points for Project. TotalPoints is the variable that gives the composite score. It is computed by adding 2 times Exam1, Exam2, and Final, 3 times Homework plus 1 times Project. Grade is the grade earned in the course. This instructor used cut-offs of 900, 800, 700, etc. for the letter grades. There are no units for ID and grade. The other variables all have "points" as the unit.

1.4 Variables for students in a statistics course. Suppose the data for the students in the introductory statistics class were also to be used to study relationships between student characteristics and success in the course. For this purpose, we might want to use a data set like the Excel spreadsheet in Figure 1.2.

| ID | Exam1 (x) | Exam2 (y) | Homework (z) | Final (u) | Project (v) |
|---|---|---|---|---|---|
| 101 | 89 | 94 | 88 | 87 | 95 |
| 102 | 78 | 84 | 90 | 89 | 94 |
| 103 | 71 | 80 | 75 | 79 | 95 |
| 104 | 95 | 98 | 97 | 96 | 93 |
| 105 | 79 | 88 | 85 | 88 | 96 |
| ... | ... | ... | ... | ... | ... |

Total points policy: $2 \times (\text{Exam } 1 + \text{Exam } 2 + \text{Final}) + 3 \times \text{Homework} + \text{Project}$
Grade policy:

| $\geq 900$ | $800 \leq \cdots < 900$ | $700 \leq \cdots < 800$ | $600 \leq \cdots < 700$ | $500 \leq \cdots < 600$ |
|---|---|---|---|---|
| A | B | C | D | E |

cut off 900, 800, 700, etc. for the letter grades.

| ID | TotalPoints (w) | Grade (s) | Gender | PrevStat | Year |
|---|---|---|---|---|---|
| 101 | 899 | B | F | Yes | 4 |
| 102 | 866 | B | M | Yes | 3 |
| 103 | 780 | C | M | No | 3 |
| 104 | 962 | A | M | No | 1 |
| 105 | 861 | B | M | No | 4 |
| ... | ... | ... | ... | ... | ... |

Comments:
It is a good idea to avoid spaces in names of the variables

In the second table, we have decided to focus on the TotalPoints and Grade as the outcomes of interest.
Other variables of interest have been included: Gender, PrevStat (whether or not the student has taken a statistics course previously), and Year (student classification as first, second, third, or fourth year).
ID is a categorical variable, TotalPoints is a quantitative variable, and the remaining variables are all categorical.
In our example, the possible values for the grade variable are A, B, C, D, and F. When computing grade point averages, many colleges and universities translate these letter grades into numbers using A = 4, B = 3, C = 2, D = 1, and F = 0. The transformed variable with numeric values is considered to be quantitative because we can average the numerical values across different courses to obtain a grade point average. Sometimes, experts argue about numerical scales such as this. They ask whether or not the difference between an A and a B is the same as the difference between a D and an F. Similarly, many questionnaires ask people to respond on a 1 to 5 scale with 1 representing strongly agree, 2 representing agree, etc. Again we could ask about whether or not the five possible values for this scale are equally spaced in some sense. From a practical point of view, the averages that can be computed when we convert categorical scales such as these to numerical values frequently provide a very useful way to summarize data.

Descriptive statistics
A descriptive statistic (in the count noun sense) is a summary statistic that quantitatively describes or summarizes features of a collection of information, while descriptive statistics in the mass noun sense is the process of using and

analyzing those statistics.
1.1 Displaying Distributions with Graphs
Descriptive statistics is distinguished from inferential statistics (or inductive statistics), in that descriptive statistics aims to summarize a sample, rather than use the data to learn about the population that the sample of data is thought to represent.

The **distribution of a variable** tells us what values it takes and how often it takes these values.

## 2. Displaying Distributions with graphs

Statistical tools and ideas help us examine data in order to describe their main features. This examination is called exploratory data analysis. Like an explorer crossing unknown lands, we want first to simply describe what we see.

### 2.1 Graphs for the categorical variables

bar graphs, pie graphs

**Example**

The educational attainment of people aged 25 to 34 years

| Education | Count (millions) | Percent |
|---|---|---|
| Less than high school | 4.6 | 12.1 |
| High school graduate | 11.6 | 30.5 |
| Some college | 7.4 | 19.5 |
| Associate degree | 3.3 | 8.7 |
| Bachelor's degree | 8.6 | 22.6 |
| Advanced degree | 2.5 | 6.6 |

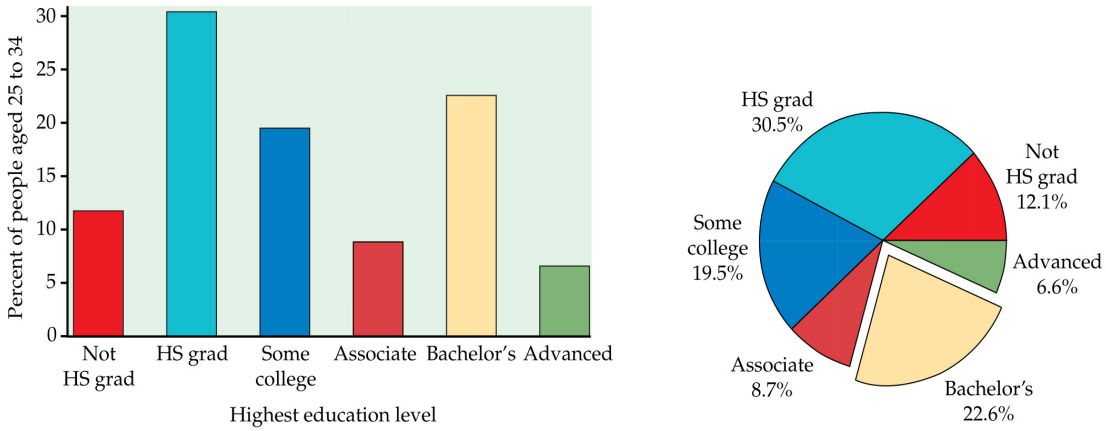


**FIGURE 1.3** (a) Bar graph of the educational attainment of people aged 25 to 34 years. (b) Pie chart of the education data, with bachelor's degree holders emphasized.

## 2.2 Graphs for qualitative variables

Visualization

**Stemplot:**

**Example**

Literacy rates (percent) in Islamic nations

| Country | Female (%) | Male (%) | Country | Female(%) | Male (%) |
|---|---|---|---|---|---|
| Algeria | 60 | 78 | Morocco | 38 | 68 |
| Bangladesh | 31 | 50 | Saudi Arabia | 70 | 84 |
| Egypt | 46 | 68 | Syria | 63 | 89 |
| Iran | 71 | 85 | Tajikistan | 99 | 100 |
| Jordan | 86 | 96 | Tunisia | 63 | 83 |
| Kazakhstan | 99 | 100 | Turkey | 78 | 94 |
| Lebanon | 82 | 95 | Uzbekistan | 99 | 100 |
| Libya | 71 | 92 | Yemen | 29 | 70 |
| Malaysia | 85 | 92 | | | |

Making the stemplots for the literacy rates among females in islamic nations

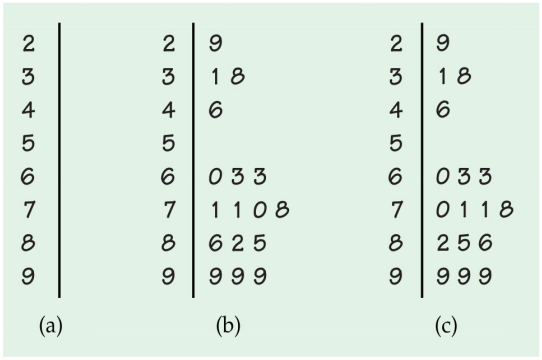


**FIGURE 1.5** Making a stemplot of the data in Example 1.7. (a) Write the stems. (b) Go through the data and write each leaf on the proper stem. For example, the values on the 8 stem are 86, 82, and 85 in the order of the table. (c) Arrange the leaves on each stem in order out from the stem. The 8 stem now has leaves 2 5 6.

Histograms

Stemplots display the actual values of the observations. This feature makes stemplots awkward for large data sets. Moreover, the picture presented by a stemplot divides the observations into groups (stems) determined by the number system rather than by judgment. Histograms do not have these limitations. A histogram breaks the range of values of a variable into classes and displays only the count or percent of the observations that fall into each class. You can choose any convenient number of classes, but you should always choose classes of equal width. Histograms are slower to construct by hand than stemplots and do not display the actual values observed. For these reasons we prefer stemplots for small data sets. The construction of a histogram is best shown by example. Any statistical software package will of course make a histogram for you.

**Example 1.9**

Distribution of IQ scores. You have probably heard that the distribution of scores on IQ tests is supposed to be roughly "bell-shaped." Let's look at some actual IQ scores. Table 1.3

| IQ test scores for 60 randomly chosen fifth-grade students | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 145 | 139 | 126 | 122 | 125 | 130 | 96 | 110 | 118 | 118 |
| 101 | 142 | 134 | 124 | 112 | 109 | 134 | 113 | 81 | 113 |
| 123 | 94 | 100 | 136 | 109 | 131 | 117 | 110 | 127 | 124 |
| 106 | 124 | 115 | 133 | 116 | 102 | 127 | 117 | 109 | 137 |
| 117 | 90 | 103 | 114 | 139 | 101 | 122 | 105 | 97 | 89 |
| 102 | 108 | 110 | 128 | 114 | 112 | 114 | 102 | 82 | 101 |

displays the IQ scores of 60 fifth-grade students chosen at random from one school.
1. Divide the range of the data into classes of equal width. The scores in Table 1.3 range from 81 to 145, so we choose as our classes

$$\begin{gathered} 75 \leq \text{IQ score} < 85 \\ 85 \leq \text{IQ score} < 95 \\ \vdots \\ 145 \leq \text{IQ score} < 155 \end{gathered}$$

Be sure to specify the classes precisely so that each individual falls into exactly one class. A student with IQ 84 would fall into the first class, but IQ 85 falls into the second.

2. Count the number of individuals in each class. These counts are called frequencies, and a table of frequencies for all classes is a frequency table.

| Class | Count | Class | Count |
|---|---|---|---|
| 75 to 84 | 2 | 115 to 124 | 13 |
| 85 to 94 | 3 | 125 to 134 | 10 |
| 95 to 104 | 10 | 135 to 144 | 5 |
| 105 to 114 | 16 | 145 to 154 | 1 |

3. Draw the histogram. First, on the horizontal axis mark the scale for the variable whose distribution you are displaying. That's IQ score. The scale runs from 75 to 155 because that is the span of the classes we chose. The vertical axis contains the scale of counts. Each bar represents a class. The base of the bar covers the class, and the bar height is the class count. There is no horizontal space between the bars unless a class is empty, so that its bar has height zero. Figure 1.7

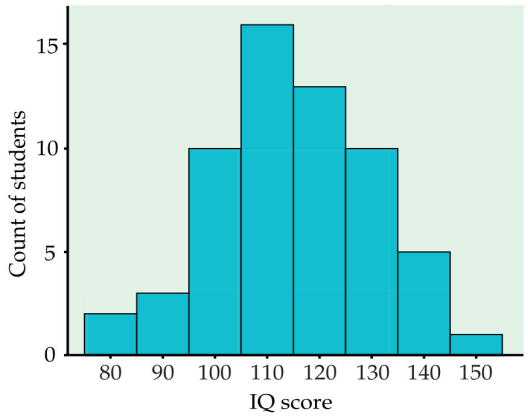


**FIGURE 1.7** Histogram of the IQ scores of 60 fifth-grade students, for Example 1.9.

is our histogram. It does look roughly "bell- shaped."

### Example 1.10

Examine the histogram. What does the histogram of IQ scores (Figure 1.7) tell us? Shape: The distribution is roughly symmetric with a single peak in the center. We don't expect real data to be perfectly symmetric, so we are satisfied if the two sides of the histogram are roughly similar in shape and extent. Center: You can see from the histogram that the midpoint is not far from 110. Looking at the actual data shows that the midpoint is 114. Spread: The spread is from 81 to 145. There are no outliers or other strong deviations from the symmetric, unimodal pattern.

### Example 1.11

The distribution of call lengths in Figure 1.8,

| Service times (seconds) for calls to a customer service center | | | | | | | |
|---|---|---|---|---|---|---|---|
| 77 | 289 | 128 | 59 | 19 | 148 | 157 | 203 |
| 126 | 118 | 104 | 141 | 290 | 48 | 3 | 2 |
| 372 | 140 | 438 | 56 | 44 | 274 | 479 | 211 |
| 179 | 1 | 68 | 386 | 2631 | 90 | 30 | 57 |
| 89 | 116 | 225 | 700 | 40 | 73 | 75 | 51 |
| 148 | 9 | 115 | 19 | 76 | 138 | 178 | 76 |
| 67 | 102 | 35 | 80 | 143 | 951 | 106 | 55 |
| 4 | 54 | 137 | 367 | 277 | 201 | 52 | 9 |
| 700 | 182 | 73 | 199 | 325 | 75 | 103 | 64 |
| 121 | 11 | 9 | 88 | 1148 | 2 | 465 | 25 |

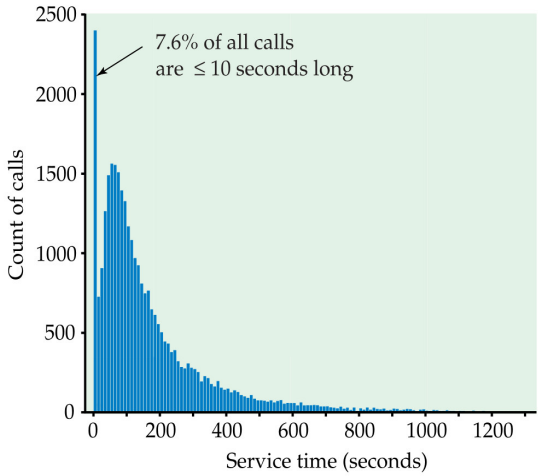


**FIGURE 1.4** The distribution of call lengths for 31,492 calls to a bank's customer service center, for Example 1.6. The data show a surprising number of very short calls. These are mostly due to representatives deliberately hanging up in order to bring down their average call length.

on the other hand, is strongly skewed to the right. The midpoint, the length of a typical call, is about 115 seconds, or just under 2 minutes. The spread is very large, from 1 second to 28,739 seconds. The longest few calls are outliers. They stand apart from the long right tail of the distribution, though we can't see this from Figure 1.8, which omits the largest observations. The longest call lasted almost 8 hours-that may well be due to equipment failure rather than an actual customer call.

A data set contains information on a collection of individuals. Individuals may be people, animals, or things. The data for one individual make up a case. For each individual, the data give values for one or more variables. A variable describes some characteristic of an individual, such as a person's height, gender, or salary.

Some variables are categorical and others are quantitative. A categorical variable places each individual into a category, such as male or female. A quantitative variable has numerical values that measure some characteristic of each individual, such as height in centimeters or annual salary in dollars.

Exploratory data analysis uses graphs and numerical summaries to describe the variables in a data set and the relations among them. The distribution of a variable tells us what values it takes and how often it takes these values.

Bar graphs and pie charts display the distributions of categorical variables.

These graphs use the counts or percents of the categories.

Stemplots and histograms display the distributions of quantitative variables. Stemplots separate each observation into a stem and a one-digit leaf. Histograms plot the frequencies (counts) or the percents of equal-width classes of values.

When examining a distribution, look for shape, center, and spread and for clear deviations from the overall shape.

Some distributions have simple shapes, such as symmetric or skewed. The number of modes (major peaks) is another aspect of overall shape. Not all distributions have a simple overall shape, especially when there are few observations.

Outliers are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them.

**Time plot**

When observations on a variable are taken over time, make a time plot that graphs time horizontally and the values of the variable vertically. A time plot can reveal trends or other changes over time.