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.

The language of statistics

1. Data Collection:

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

Example 1.1 A typical statistical problem can be described as follows. One or a series of random experiments is performed; some data from the experiment(s) are collected; and our task is to extract information from the data, interpret the results, and draw some conclusions. A descriptive data analysis can be performed to obtain some summary measures of the data, such as the mean, median, range, standard deviation, etc., and some graphical displays, such as the histogram and box-and-whisker diagram (boxplots), etc. Although this kind of analysis is simple and requires almost no assumptions, it may not allow us to gain enough insight into the problem. We also focus on more sophisticated methods of analyzing data: statistical inference and decision theory.

Example 1.2 Let us consider a random 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, appears 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 this example is different from the case when we always get the same result, $x_1 = x_2 = x_3, = \cdots = x_{N-1} = x_N$. This means that we are able to predict the result of the experiment, so it is called **deterministic**.

Example 1.3 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 result we get a sequence of random 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 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.

In statistics we perform **experimental** or only **observational studies**. 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. We are interfering with system.

In contrast, an **observational study** does not involve experimental manipulation.

2. Data organization:

- Individuals and variables

Individuals are the objects described in a set of data. Individuals are sometimes people, animals or other things.

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 tells us what values it takes and how often it takes these values.

Example 2.1

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.

Table 1

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
...	...	...	...	...	...

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.

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

≥ 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.

Table 1a

ID	TotalPoints (W)	Grade (S)
101	899	B
102	866	B
103	780	C
104	962	A
105	861	B
...	...	...

Other 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 in the table:

Table 2

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 **Table 2**, 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. 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.

3. Descriptive statistics

A **descriptive statistic** in the count noun sense is a summary statistic (the mean value, the standard deviation, the median, etc.) 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.

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.

3.1. Examining a distribution

In any graph of data, look for the overall pattern and for striking deviations from that pattern. You can describe the overall pattern of a distribution by its shape, center, and spread. An important kind of deviation is an outlier, an individual value that falls outside the overall pattern.

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

4. 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.

4.1 Graphs for the categorical variables

bar graphs and **pie graphs**

Example 4.1

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

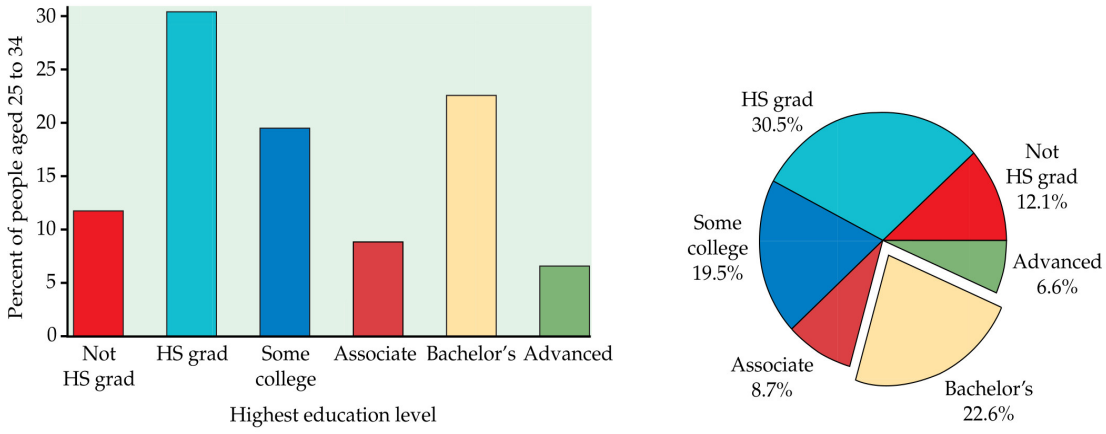


FIGURE 1.3 (a) Bar graph of the educational attainment of people aged 25 to 34 years. (b) Pie chart of the education data, with bachelor's degree holders emphasized.

4.2 Graphs for qualitative variables

Visualization

Stemplot:

To make a stemplot:

1. Separate each observation into a stem consisting of all but the final (rightmost) digit and a leaf, the final digit. Stems may have as many digits as needed, but each leaf contains only a single digit.

2. Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column.

3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.

Example 4.2

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

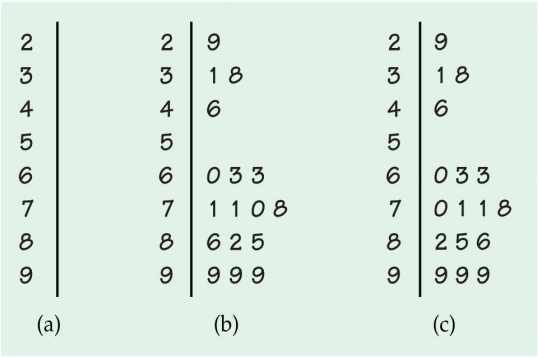


FIGURE 1.5 Making a stemplot of the data in Example 1.7. (a) Write the stems. (b) Go through the data and write each leaf on the proper stem. For example, the values on the 8 stem are 86, 82, and 85 in the order of the table. (c) Arrange the leaves on each stem in order out from the stem. The 8 stem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

Histograms

Example 4.3

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

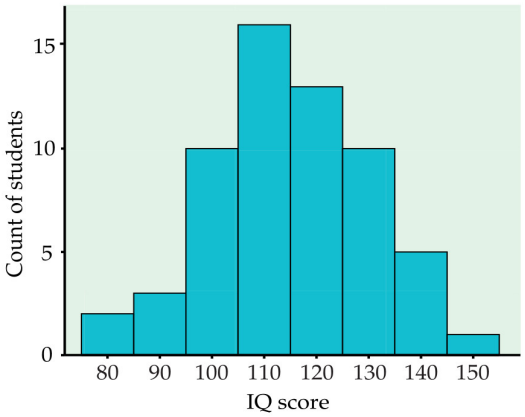


FIGURE 1.7 Histogram of the IQ scores of 60 fifth-grade students, for Example 1.9.

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

Example 4.4

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 4.5

The customer's service call center.
We have data on the length of all 31492 calls made to the customer service center of a small bank in a month. Table 1.1 displays the lengths of the first 80 calls. The numbers are meaningless without some background information. The individuals are calls made to the bank's call center. The variable recorded is the length of each call. The units are seconds. We see that the call lengths vary a great deal. The longest call lasted 2631 seconds, almost 44 minutes. More striking is that 8 of these 80 calls lasted less than 10 seconds. What's going on?

Table 1.1

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 is a histogram of the lengths of all 31492 calls. We did not plot the few lengths greater than 1200 seconds (20 minutes). As expected, the graph shows that most calls last between about a minute and 5 minutes, with some lasting much longer when customers have complicated problems. More striking is the fact that 7.6% of all calls are no more than 10 seconds long. It turned out that the bank penalized representatives whose average call length was too long-so some representatives just hung up on customers in order to bring their average length down. Neither the customers nor the bank were happy about this. The bank changed its policy, and later data showed that calls under 10 seconds had almost disappeared

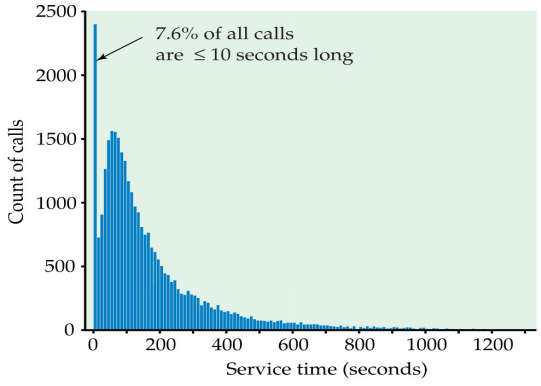


FIGURE 1.4 The distribution of call lengths for 31,492 calls to a bank's customer service center, for Example 1.6. The data show a surprising number of very short calls. These are mostly due to representatives deliberately hanging up in order to bring down their average call length.

The distribution of call lengths collected in Figure 1.4, 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 1.4, 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.

Caution. You should be aware that the appearance of a histogram can change when you change the classes.

Figure 1.8 is a histogram of the customer service call lengths produced by software with no special instructions from the user. The software's 'default' histogram shows the overall shape of the distribution, but it hides the spike of very short calls by lumping all calls of less than 100 seconds into the first class.

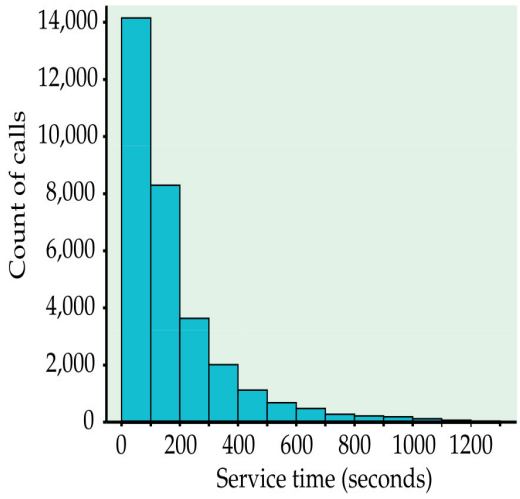


FIGURE 1.8 The "default" histogram produced by software for the call lengths in Example 1.6. This choice of classes hides the large number of very short calls that is revealed by the histogram of the same data in Figure 1.4.

Summary

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.

5. 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.

Example 5.1

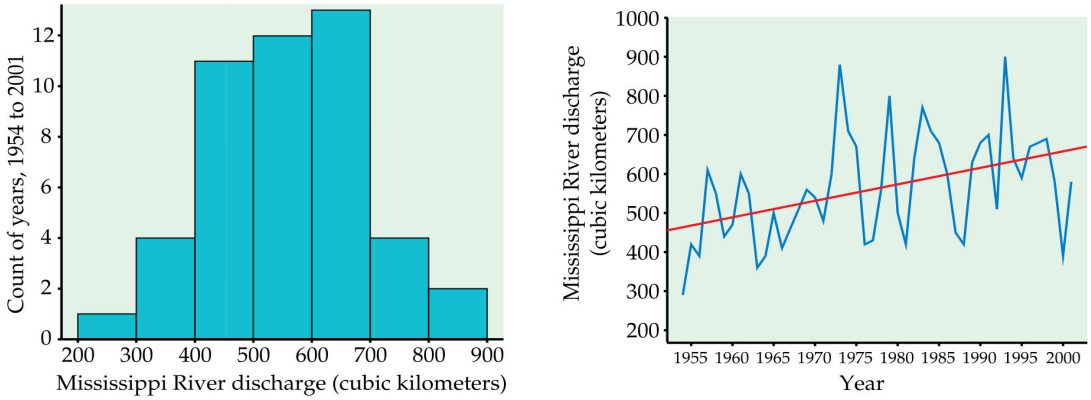


Figure 1.10 (a): Histogram of the volume of water discharged by the Mississippi River over the 48 years from 1954 to 2001. Data are from Table 1.4.

Figure 1.10 (b): Time plot of the volume of water discharged by the Mississippi River for the years 1954 to 2001. The line shows the trend toward increasing river flow, a trend that cannot be seen in the histogram

Table 1.4

Yearly discharge of the Mississippi River (in cubic kilometers of water)							
Year	Discharge	Year	Discharge	Year	Discharge	Year	Discharge
1954	290	1966	410	1978	560	1990	680
1955	420	1967	460	1979	800	1991	700
1956	390	1968	510	1980	500	1992	510
1957	610	1969	560	1981	420	1993	900
1958	550	1970	540	1982	640	1994	640
1959	440	1971	480	1983	770	1995	590
1960	470	1972	600	1984	710	1996	670
1961	600	1973	880	1985	680	1997	680
1962	550	1974	710	1986	600	1998	690
1963	360	1975	670	1987	450	1999	580
1964	390	1976	420	1988	420	2000	390
1965	500	1977	430	1989	630	2001	580

Summary

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 (outliers often result from malfunctions in measuring equipment).