MAT 271E: PROBABILITY AND STATISTICS

PROF. DR. CANAN SARICAM

WEEK 1

WHAT IS STATISTICS? ROLE OF STATISTICS. STATISTICS IN DAILY LIFE

WHAT IS STATISTICS?

In common usage, the word statistics refers to numerical data.

- Vital statistics are numerical data on births, deaths, marriages, divorces, and communicable diseases;
- Business and economic statistics are numerical data on employment, production, prices, and sales;
- Social statistics are numerical data on housing, crime, education, and social assistance.

WHAT IS STATISTICS?

Data: Facts, especially numerical facts, collected together for reference or information.

Information: Knowledge communicated concerning some particular fact.

Statistics is a ***tool*** for creating ***new understanding*** from a set of numbers.

Statistics is a way to get information from data.

WHAT IS STATISTICS?

The field of statistics deals with the

collection,
presentation,
Analysis
Use of data

make decisions and solve problems.

WEEK 2

DATA SETS, DATA ACQUISITION.
DATA PATTERNS: ARRAYS, STEM AND LEAF DISPLAYS, DOT PLOTS, CUMULATIVE DISTRIBUTIONS
FREQUENCY DISTRIBUTION TABLES

DATA ACQUISITION

- ✓ Statistical analysis requires that the facts of interest in an investigation should be assembled and organized in a useful manner. We refer to such facts as data.
- ✓ Data is a collection of facts, such as numbers, words, measurements, observations or just descriptions of things.
- ✓ If the data are not properly assembled and organized, misleading or incorrect conclusions may be drawn from them.

DATA SETS

A **data set** is a collection of facts assembled for **a particular purpose.**

Data sets are found all around us.

- the financial section of our daily newspaper contains price data for securities, and commodities;
- an economic report shows inflation rates for different countries;
- a computer file contains the names and addresses of members of a professional association.

CHARACTERISTICS OF DATA SET

Sample number (case)	Sample title	Composition	Yarn count Ne	Yarn twist TPI	Tenacity Rkm	Quality class
1	White cone	Cotton	29.8	24.6	11.0	1
2	Yellow cone	Viscose	20.1	18.3	13.2	2
3	Green cone	PET/viscose	13.9	16.5	18.7	1

DATA SETS

Element: A data set provides data about a collection of elements and contains, for each element information about one or more characteristics of interest.

Sample number (case)	Sample title	Composition	Yarn count Ne	Yarn twist TPI	Tenacity Rkm	Quality class
1	White cone	Cotton	29.8	24.6	11.0	1
2	Yellow cone	Viscose	20.1	18.3	13.2	2
3	Green cone	PET/viscose	13.9	16.5	18.7	1

DATA SETS

Element: A data set provides data about a collection of elements and contains, for each element information about one or more characteristics of interest.

In our example, an element of data set is a particular sampling group of bobbins.So there are three elements

Sample number (case)	Sample title	Composition	Yarn count Ne	Yarn twist TPI	Tenacity Rkm	Quality class
1	White cone	Cotton	29.8	24.6	11.0	1
2	Yellow cone	Viscose	20.1	18.3	13.2	2
3	Green cone	PET/viscose	13.9	16.5	18.7	1

DATA SETS

Variable: A characteristic that can take on different possible outcomes is called variable.

The variable is said to be **quantitative** if the outcomes are numbers and **qualitative** if the outcomes are non-numerical qualities or attributes.

Sample number (case)	Sample title	Composition	Yarn count Ne	Yarn twist TPI	Tenacity Rkm	Quality class
1	White cone	Cotton	29.8	24.6	11.0	1
2	Yellow cone	Viscose	20.1	18.3	13.2	2
3	Green cone	PET/viscose	13.9	16.5	18.7	1

CHARACTERISTICS OF DATA SET

Qualitative variable

Quantitative variable

Sample number (case)	Sample title	Composition	Yarn count Ne	Yarn twist TPI	Tenacity Rkm	Quality class
1	White cone	Cotton	29.8	24.6	11.0	1
2	Yellow cone	Viscose	20.1	18.3	13.2	2
3	Green cone	PET/viscose	13.9	16.5	18.7	1

Element

DATA SETS

- In our example, one characteristic of interest is the tenacity. This characteristic takes on different values for different colored cones; hence, tenacity is called a variable.

- Tenacity is a quantitative variable because its outcomes are numerical in nature. On the other hand, the yarn composition is a qualitative variable because its outcomes are non-numerical.

- Four variables (yam count, twist, tenacity, and quality class) are quantitative and others are qualitative of all six variables.

DATA SETS

- **Sometimes, a quantitative variable is converted into a qualitative one** by grouping the possible numerical outcomes into non-numerical categories. This conversion is done to facilitate reporting, interpreting, or analysing the data.

- As another example, body sizes in apparel products are the quantitative variable (measured in cm) which are converted into a qualitative variable by dividing all possible sizes into categories such as S, M, L, and XL.

- At other times, the reverse is done to convert a qualitative variable into a quantitative one by assigning an arbitrary numerical value or code to each non-numerical category. For instance, the variable quality class, in our example, might be quantified by using the numbers 1 or 2 for different quality levels.

DATA SETS

Case: The information on all variables for one element in the data set is called as case or record (observation vector). In the table, each row of data represents a case.

Sample number (case)	Sample title	Composition	Yarn count Ne	Yarn twist TPI	Tenacity Rkm	Quality class
1	White cone	Cotton	29.8	24.6	11.0	1
2	Yellow cone	Viscose	20.1	18.3	13.2	2
3	Green cone	PET/viscose	13.9	16.5	18.7	1

DATA SETS

Observation: The information about a single variable for an element of data set is called an **observation**, a **measurement**, a **reading**, or an outcome. Thus, 13.9 is the observation on the yarn count variable for Green cone.

Data sets may be distinguished on the basis of the number of variables they contain.

An **univariate data set** contains one variable; a **bivariate data set** contains two variables; and a **multivariate data set** contains three or more variables.

CHARACTERISTICS OF DATA SET

Qualitative variable

Quantitative variable

Sample number (case)	Sample title	Composition	Yarn count Ne	Yarn twist TPI	Tenacity Rkm	Quality class
1	White cone	Cotton	29.8	24.6	11.0	1
2	Yellow cone	Viscose	20.1	18.3	13.2	2
3	Green cone	PET/viscose	13.9	16.5	18.7	1

Case

Element

Observation

TYPES OF DATA

Statistical data are of several different types.

- **Measurement data**
- **Count data**
- **Rank data**
- **Classification data**

TYPES OF DATA

Statistical data are of several different types.

- **Measurement data:** Data that represent measurements of amounts, capacities, or similar characteristics are called measurement data; for example, actual test results of yarn count.
- **Count data**. Data that are counts or frequencies and here, are necessarily whole numbers are called count data; for example, number of defective units in a lot or number of operation breaks in a textile process for a given period of time.
- **Rank data.** Data obtained by ranking or ordering elements are called rank. For example, a published list gives rank data for the worlds longest rivers, with the Nile shown as the longest (rank 1), the Amazon as the second longest (rank 2) , and so on. Likewise, a monthly report on fabric sales gives a list for fabric meters sold according to the types in the descending order.
- **Classification data.** Data where classes or categories are set up and each element is assigned to its appropriate category are called classification. Observations of the variable quality class can be given as an example for classification data.

DATA SOURCES

- Statistics is concerned not only with organizing and analysing data once they are assembled but also with the sources of data and how data are collected for a study.

- The first stage of any investigation involves a specification or a definition of the problem to be studied. The problems are usually concerned with the effects of one or more variables called factors on the variable of interest.

- For instance, what are the effects of age and income on the amount of food expenditures? Here, age and income are the factors and food expenditures is the variable of the interest.

- The specification of the problem leads to an identified need for particular types of data to deal with the problem. The question, then, is where or how to obtain the necessary data.

DATA SOURCES

- ***Primary data*** are collected especially for the purpose of whatever survey is being conducted.
- ***Secondary data*** are those which have already been collected elsewhere, for some other purpose, but which can be used or adapted for the survey being conducted.

DATA SOURCES

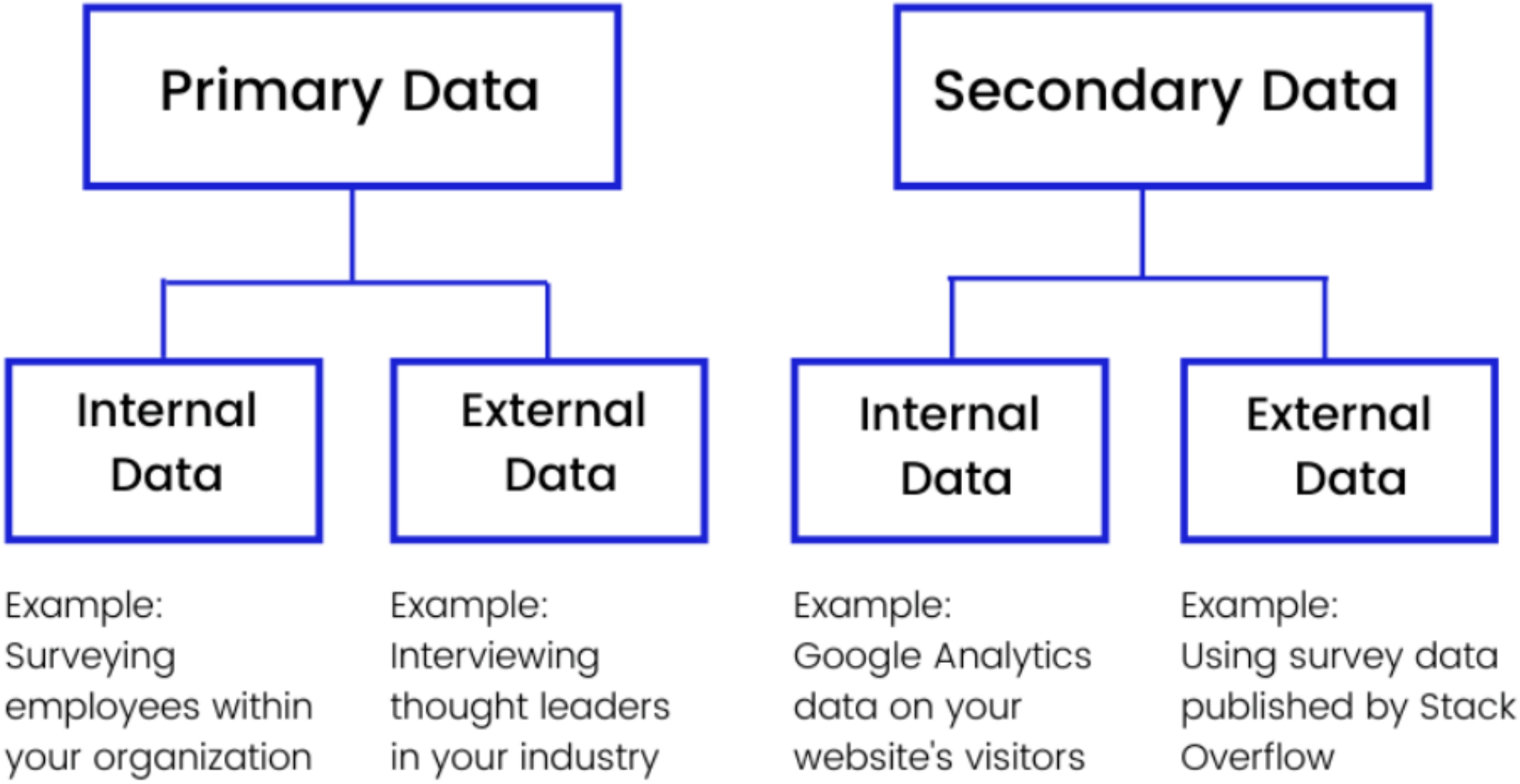
Internal Sources

- Internal data are collected from the organisation itself. It relates to the activities or transactions performed within the organisation e.g. sales, financial, employee, stocks etc.

External Sources

- External data are collected from outside the organisation and relate to the environment in which the organisation operates.

DATA SOURCES



DATA ACQUISITION

When the data needed for an investigation are not available in existing sources, some methods for obtaining them directly must be considered.

- 1) Observational Study: A study in which the researcher merely observes what is happening or what has happened in the past and tries to draw conclusion based on these observations.

- 2) Experimental Study: A study in which the researcher manipulates one of the variables and tries to determine how the manipulation influences other variables.

DATA ACQUISITION

In an **experimental study**, for instance, for the investigation of the effect of production speed on the yarn breakages of the ring machines during the spinning process in a mill, the results might be taken for each different machine speed level under the same circumstances of yarn structural parameters and machine settings; but, more than one machines of the same type should be randomly assigned to eliminate the uncontrolled factors effects.

In an **observational study**, no experimental controls are exercised over factors influencing the variable of interest. A study was undertaken by a large corporation with five plants to obtain information about plant productivity (the variable of interest) and about several explanatory factors, including type and age of machinery used in the plant, education, and age characteristics of the production employees in the plant, and the type of wage-incentive program in the plant.

DATA ACQUISITION

- Both experimental and observational studies can be extremely useful for investigating the effects of one or more factors on the variable of interest. Experimental studies provide stronger evidence of these effects than observational studies. Experiments are especially advantageous in investigating cause-and-effect patterns.
- Despite the advantages of experimental studies, much of the statistical analysis in manufacturing, business, economics, and the social sciences is based on. observational studies.
- One reason is that most available data, such as internal data on company operations and external data on the economy and consumer behavior, are observational data.
- Another reason is that it is often not feasible or may not be desirable to exercise the experimental controls required in the experiment al studies.

DATA ACQUISITION

A variety of procedures for acquiring data are employed in experimental and observational studies.

Three commonly used ones are:

1. observation
2. interview
3. self-enumeration

DATA ACQUISITION

1. Observation. Data acquisition by observation entails direct examination and recording of an ongoing activity. In an engineering study, data about the internal pressure were obtained by reading an instrument inserted in a particular section of the machine.

The limitation of the method is to be dependent on observer's skill, training, behavior, etc.

2. Interview. In an interview, an interviewer asks questions from a questionnaire and records the respondents answers. Interviews may be conducted in person or over the telephone.

3. Self-enumeration. With self-enumeration, the respondent answers questions printed on a questionnaire or displayed on a computer monitor.

DATA PATTERNS

Simple displays of data

1. **Arrays**
2. **Stem and leaf displays**
3. **Dot diagram**
4. **Cumulative distribution**

DATA PATTERNS

Simple displays of data

1. Arrays

A useful first step for discerning a pattern in quantitative data **when the number of observations is not too large** is to list the observations **in increasing or decreasing order of magnitude.** Such an ordering, called an array, can greatly facilitate inspection of the data.

An industrial engineer examined failure data for the bond between a wire and a semi-conductor wafer. An array of the breaking stresses (in milligrams) of 17 bonds, arranged in increasing order of magnitude, follows:

1, 43, 51, 59, 88, 113, 47, 62, 71, 31, 63, 51, 4, 66,

58, 50 ,75

DATA PATTERNS

Simple displays of data

1. Arrays

1, 4, 31, 43, 47, 50, 51, 51, 58, 59, 62, 63, 66, 71, 75, 88, 113

This array quickly tells us that there is a large range of breaking stresses, from 1 to 113 milligrams.

DATA PATTERNS

Simple displays of data

2. Stem-and-leaf displays

A stem-and-leaf display provides more information about the pattern of data than an array does. it assists in the discovery of

(1) **concentrations** of particular values,
(2) **outlying or extreme observations**,
(3) the **extent of symmetry or lack thereof** in the distribution of observations.

To construct a stem-and-leaf diagram, we divide each number into two parts: a **stem,** consisting of one or more of the leading digits, and a **leaf,** consisting of the remaining digits. The digits to the left of the vertical line are the stems and the digits to the right are the leaves.

DATA PATTERNS

Simple displays of data

2. Stem-and-leaf displays

EXAMPLE:

Show, the stem-and-leaf display for the 17 breaking stresses in the semiconductor failure example

1, 4, 31, 43, 47, 50, 51, 51, 58, 59, 62, 63, 66, 71, 75, 88, 113

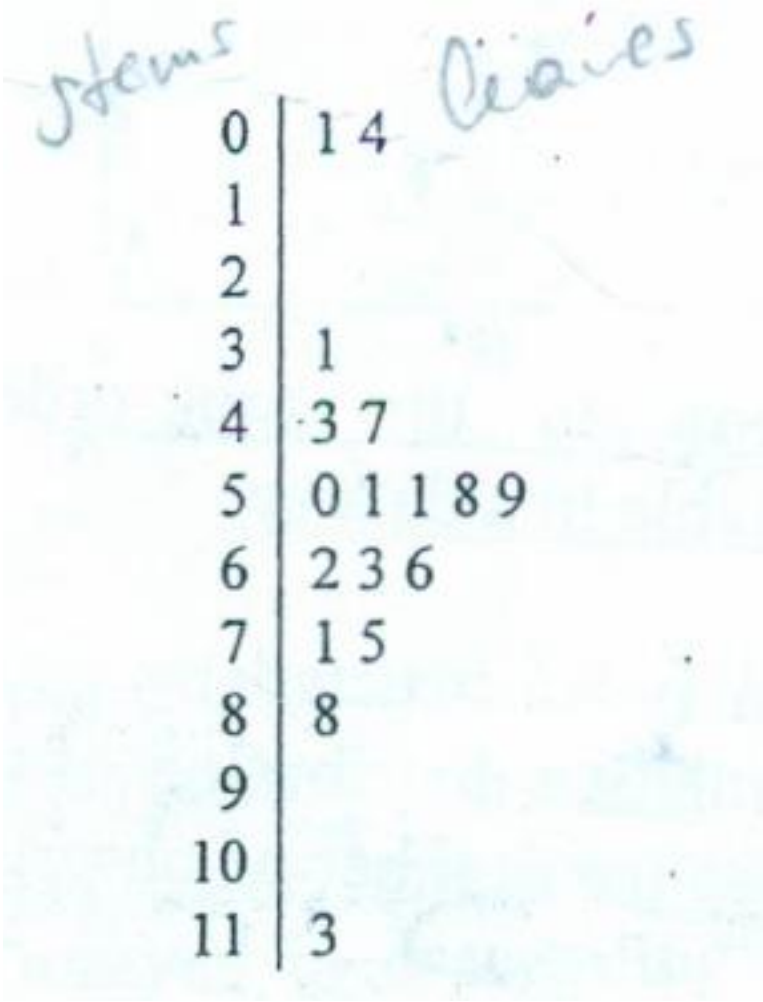
DATA PATTERNS

Simple displays of data

2. Stem-and-leaf displays

EXAMPLE:

The stem-and-leaf display for the 17 breaking stresses in the semiconductor failure example is given below. The digit 0 constitutes the first stem. The breaking stress values 1 and 4 milligrams are recorded on this stern in ascending order of magnitude, and all breaking stress values from 50 to 59 milligrams are recorded on the sixth stern of 5 in ascending order of magnitude.



STEM-AND-LEAF DISPLAY EXAMPLE

Complete a stem-and-leaf plot for the following list of values:

23.25, 24.13, 24.76, 24.81, 24.98, 25.31, 25.57, 25.89, 26.28, 26.34, 27.09

STEM-AND-LEAF DISPLAY EXAMPLE

Complete a stem-and-leaf plot for the following list of values:

23.25, 24.13, 24.76, 24.81, 24.98, 25.31, 25.57, 25.89, 26.28, 26.34, 27.09

If I try to use the last digit, the hundredths digit, for these numbers, the stem-and-leaf plot will be enormously long, because these values are so spread out. (With the numbers' first three digits ranging from 232 to 270, I'd have thirty-nine leaves, most of which would be empty.) So instead of working with the given numbers, I'll round each of the numbers to the nearest tenth, and then use those new values for my plot.

Rounding gives me the following list:

23.3, 24.1, 24.8, 24.8, 25.0, 25.3, 25.6, 25.9, 26.3, 26.3, 27.1

STEM-AND-LEAF DISPLAY EXAMPLE

Complete a stem-and-leaf plot for the following list of values:

23.25, 24.13, 24.76, 24.81, 24.98, 25.31, 25.57, 25.89, 26.28, 26.34, 27.09

If I try to use the last digit, the hundredths digit, for these numbers, the stem-and-leaf plot will be enormously long, because these values are so spread out. (With the numbers' first three digits ranging from 232 to 270, I'd have thirty-nine leaves, most of which would be empty.) So instead of working with the given numbers, I'll round each of the numbers to the nearest tenth, and then use those new values for my plot.

Rounding gives me the following list:

23.3, 24.1, 24.8, 24.8, 25.0, 25.3, 25.6, 25.9, 26.3, 26.3, 27.1

stem	leaf
27	1
26	3 3
25	0 3 6 9
24	1 8 8
23	3

key: "23 | 3" means "23.3"