Organizing and Graphing Data

Jiří Neubauer

Department of Econometrics FVL UO Brno
office 69a, tel. 973 442029
email:Jiri.Neubauer@unob.cz

Elementary Statistical Terms

Statistics as a subject provides a body of principles and methodology for designing the process of data collection, summarizing and interpreting the data, and drawing conclusions or generalities.

Elementary Statistical Terms

- statistical observation and data finding
- organizing, displaying and describing statistical data sets
- making decision, inferences, predictions and forecasts based on given data sets

Elementary Statistical Terms

Statistics can be divided into two areas:

- **descriptive statistics** – consists of methods for organizing, displaying and describing data using tables, graphs, and summary measures.
- **inferential statistics** – consists of methods that use sample results to help make decisions or predictions about a population.

Elementary Statistical Terms

Definition

Population consists of all elements – individuals, items, or objects – whose characteristics are being studied. The population that is being studied is also called **target population**.

A unit is a single entity (usually a person or an object) whose characteristics are of interest.

Elementary Statistical Terms

The population can be

- **real** – all units really exist (students of FVL, Ford made in 1999, daily production of breads , . . . $\rightarrow$ finite)
- **hypothetical** – is generally defined, but really exists just a particular part of it (physical or chemical measurements, . . . $\rightarrow$ infinite).

Elementary Statistical Terms

Definition

A **sample** from a statistical population is a proportion (a subset) of the population selected for study.

Definition

A survey that includes every member of the population is called **census**. The technique of collecting information from a proportion of the population is called **sample survey**.

A sample that represents the characteristics of the population as closely as possible is called a **representative sample**.

Elementary Statistical Terms

A sample can be

- **random** – A sample drawn in such a way that each element of the population has a chance of being selected. If all samples of the same size selected from a population have the same chance of being selected, we call it **simple random sampling**. Such a sample is called a **simple random sample**.
- **non-random** – The elements of the sample are not selected randomly but with a view of obtaining a representative sample.

Elementary Statistical Terms

Definition

A **variable** is a characteristic under study that assumes different values for different elements.

The value of variable for an element is called an **observation** or **measurement**.

Definition

A **data set** is a collection of observations on one or more variables. The number of observations we call a **sample size** and denote usually n.

Main Types of Data (variables)

We distinguish two basic types of data (variables)

- **qualitative** or **categorical data** – A variable that cannot assume a numerical value but can be classified into two or more non-numeric categories is called a qualitative or categorical variable, the data collected on such a variable are called qualitative data.
- **quantitative** or **numerical data** – A variable that can be measured numerically is called a quantitative variable. The data collected on a quantitative variable are called quantitative data.
  - **discrete variable** – usually integer numbers
  - **continuous variable** – real numbers

Main Types of Data (variables)

- **qualitative** or **categorical variables**: color of cars (black, red, green,. . .), marital status of people (unmarried, married, divorced, widow–widower), sex (male, female), etc.
- **quantitative** or **numerical data – discrete**: number of typographical errors in newspapers, number of persons in a family, number of cars owned by families, etc.
- **quantitative** or **numerical data – continuous**: length of a jump, height, weight, survival time, etc.

Organizing and Graphing Data – Categorical Data

Data are usually organized in the form of a frequency table shows the counts (**frequencies**) of individual categories. Our understanding of the data is further enhanced by calculation of proportion (**relative frequency**) of observations in each category.

$$\text{Relative frequency} = \frac{\text{Frequency in the category}}{\text{Total number of observations}}.$$

Organizing and Graphing Data – Categorical Data

A campus press polled a sample of 280 undergraduate students in the order study student attitude towards a proposed change in the dormitory regulations. Each student was to respond as support, oppose, or neutral in regard to the issue. The numbers were 152 support, 77 neutral, and 51 opposed. Tabulate the results and calculate the relative frequencies for the three response categories.

Organizing and Graphing Data – Categorical Data

Responses	*Frequency* n_i	*Relative frequency* p_i
Support	152	$\frac{152}{280} \doteq 0.543$
Oppose	51	$\frac{51}{280} \doteq 0.182$
Neutral	77	$\frac{77}{280} = 0.275$
Total	280	1

Table: Summary results of an opinion poll

Organizing and Graphing Data – Categorical Data

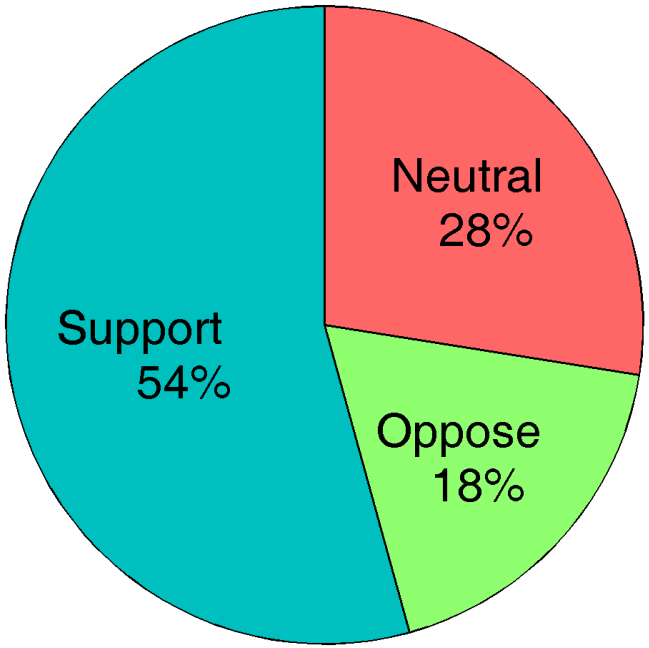

Figure: Pie chart

Organizing and Graphing Data – Categorical Data

Graduate students in a counseling course were asked to choose one of their personal habits that needed improvement.

Activity	*Frequency*
Watching TV	58
Reading newspaper	21
Talking on phone	14
Driving a car	7
Grocery shopping	3
Other	12

Table: Frequency table

Organizing and Graphing Data – Categorical Data

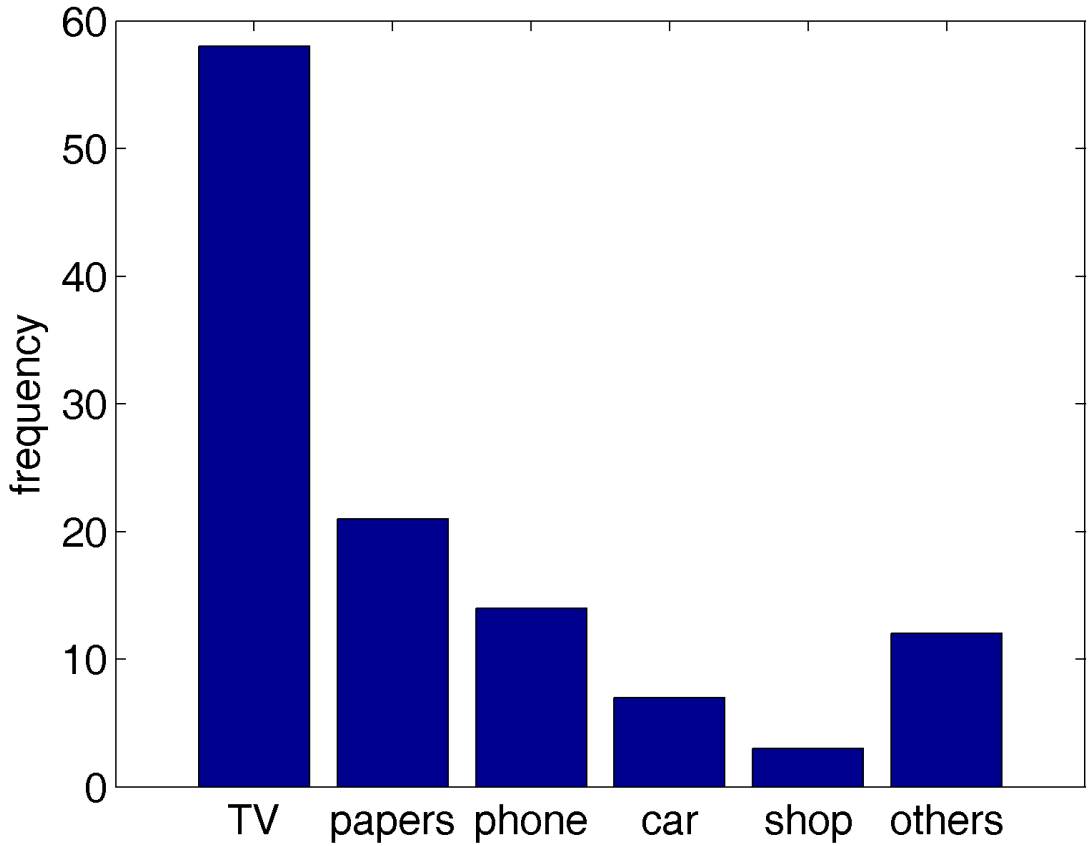

Figure: Pareto diagram

Organizing and Graphing Data – Quantitative Data

Small sample – if the sample size is small ($n < 30$)

- Sort the data in ascending order: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$
- Graph the data
- Calculate measures (see next lecture)

Organizing and Graphing Data – Quantitative Data

Example. We measured the quantity of fat in 15 sample of milk (in g/l):

14.85	14.68	15.27	14.77	14.83	14,95	15,08	15,02
15.07	14.98	15.15	15.49	14.83	14.95	14.78	

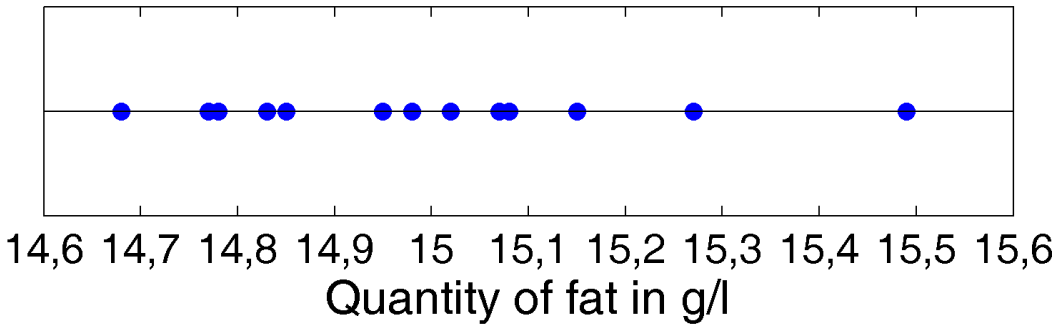

Figure: The quantity of fat

Organizing and Graphing Data – Quantitative Data

Discrete data – $n > 30$ with small number of variants

- Frequency table ($n_i, p_i, N_i, F_i, i = 1, 2, \ldots, k, k$ is the number of variants)
- Graph the data – line plot, histogram, box plot, empirical distribution function
- Calculate measures (see next lecture)

Organizing and Graphing Data – Quantitative Data

Example. We have data set containing the heights of 50 randomly chosen 15 months old boys (in cm):

83	85	81	82	84	82	79	84	80	81
82	82	80	82	80	82	83	84	82	79
83	82	83	82	82	82	81	80	82	82
83	80	82	85	81	83	81	81	83	82
81	85	83	79	81	81	81	84	81	82

Create a frequency table and plot the data.

Organizing and Graphing Data – Quantitative Data

Height x_i	*Freq.* n_i	*Rel. freq.* p_i	*Cumulative frequency* N_i	*Rel. cum. frequency* F_i
79	3	0.06	3	0.06
80	5	0.10	8	0.16
81	11	0.22	19	0.38
82	16	0.32	35	0.70
83	8	0.16	43	0.86
84	4	0.08	47	0.94
85	3	0.06	50	1.00
Σ	50	1.00	—	—

Table: Frequency table – height of 15 months old boys

Organizing and Graphing Data – Quantitative Data

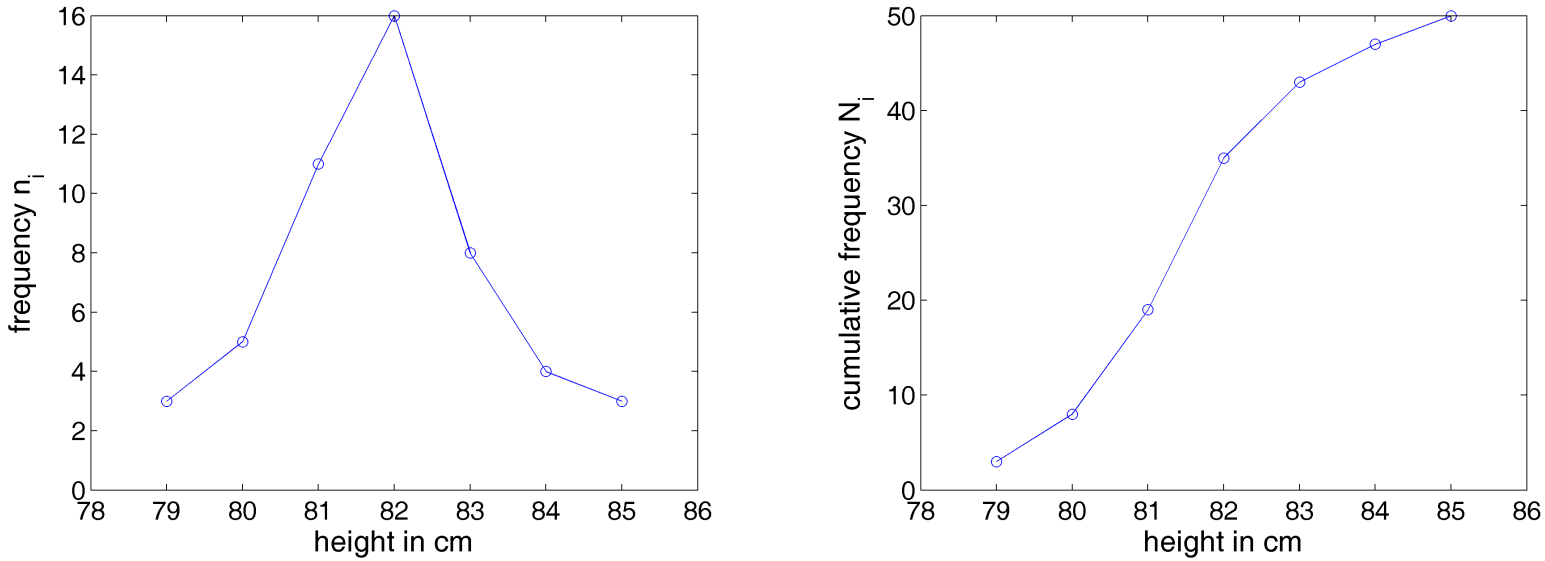

Figure: Frequency distribution

Organizing and Graphing Data – Quantitative Data

Continuous data – $n > 30$, also possible to use for description of discrete data set with large number of variants

- Construct classes (the number, the width and the begin)
- Frequency table
- Graph the data – histogram, box plot, empirical distribution function
- Calculate measures (see next lecture)

Organizing and Graphing Data – Quantitative Data

Calculation of classes

- find n, $x_{\min}$, $x_{\max}$ and calculate the range $R = x_{\max} - x_{\min}$
- the number of classes k we can determinate by following rules
  - Sturges' rule $k \approx 1 + 3.32 \log n$
  - Yule's pravidlo $k \approx 2.5\sqrt[4]{n}$
  - other rules $k \approx \sqrt{n}$, $k \approx 5 \log n$
- calculation of class width $h \approx R/k$ or $h \approx$ from $0.08 \cdot R$ till $0.12 \cdot R$

Organizing and Graphing Data – Quantitative Data

Example. We have data set containing the quantity of the dust particles (in $\mu g/m^3$):

1.23	1.10	1.54	1.34	1.06	1.09	1.41	1.48	1.52	1.37	1.37	1.63
1.51	1.53	1.31	1.23	1.31	1.27	1.17	1.27	1.34	1.27	1.09	1.01
1.41	1.22	1.27	1.37	1.14	1.22	1.43	1.40	1.41	1.51	1.51	1.47
1.14	1.34	1.16	1.51	1.58	1.33	1.31	1.04	1.58	1.12	1.19	1.17
1.47	1.24	1.45	1.29	1.17	1.63	1.39	1.02	1.38	1.39	1.43	1.28

Create a frequency table and plot the data.

Organizing and Graphing Data – Quantitative Data

Class	*Middle* x_j	*Freq.* n_j	*Rel. freq.* p_j	*Cum. freq.* N_j	*Rel. cum. Freq.* F_j
$(1.00; 1.10\rangle$	1.05	7	0.177	7	0.117
$(1.10; 1.20\rangle$	1.15	8	0.133	15	0.250
$(1.20; 1.30\rangle$	1.25	11	0.183	26	0.433
$(1.30; 1.40\rangle$	1.35	14	0.233	40	0.667
$(1.40; 1.50\rangle$	1.45	9	0.150	49	0.817
$(1.50; 1.60\rangle$	1.55	9	0.150	58	0.967
$(1.60; 1.70\rangle$	1.65	2	0.033	60	1.000
Σ	—	60	1	—	—

Table: Frequency table – quantity of dust particles in $\mu g/m^3$

Organizing and Graphing Data – Quantitative Data

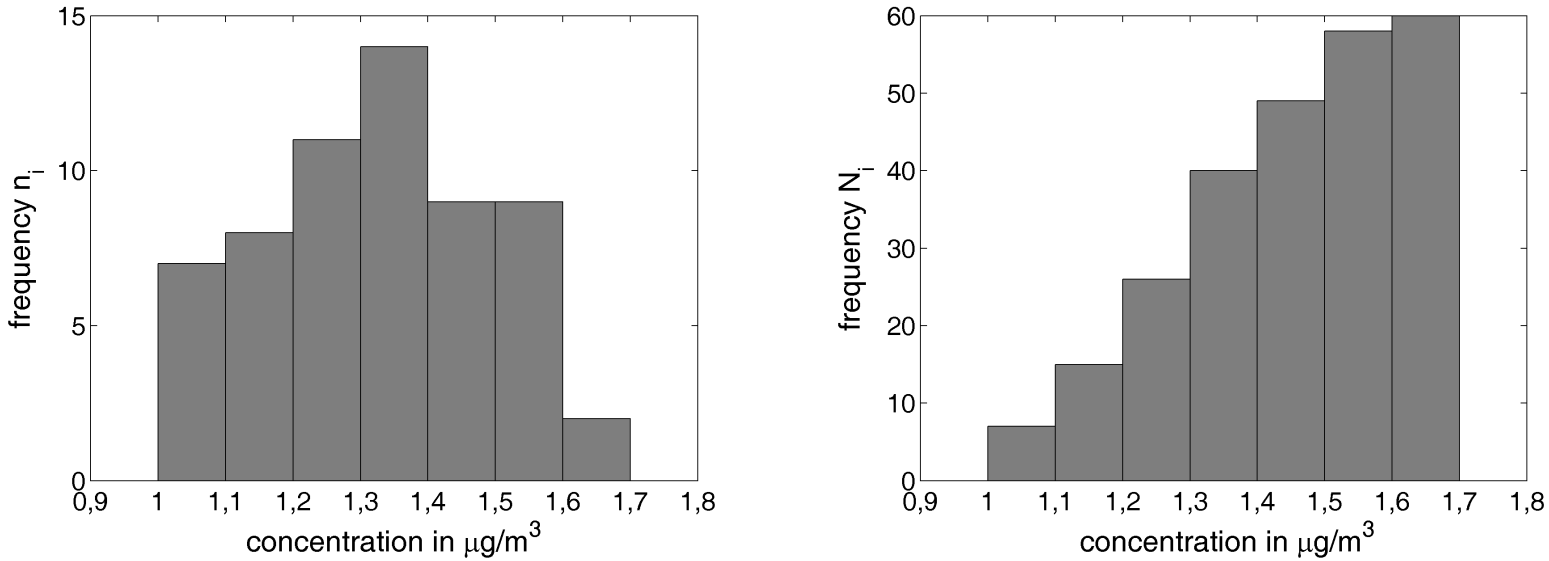

Figure: Frequency distribution – histograms

Organizing and Graphing Data – Quantitative Data

The frequency distribution is also possible to describe by an **empirical distribution function**, which is defined as

$$F_n(x) = \frac{\text{number of elements in the sample} \leq x}{n} = \frac{N(x_i \leq x)}{n}.$$

Organizing and Graphing Data – Quantitative Data

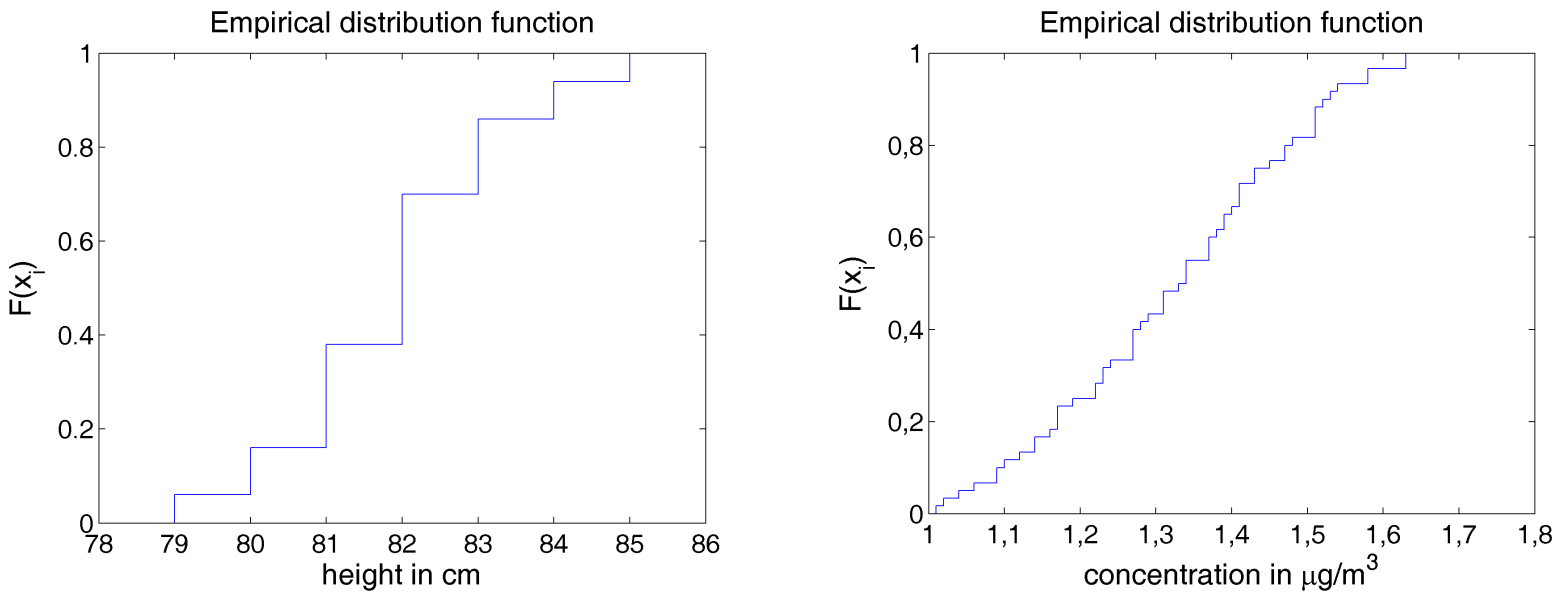

Figure: Empirical distribution function