Introductory Statistics

2024 Lectures
Part 2 - Data and Statistics

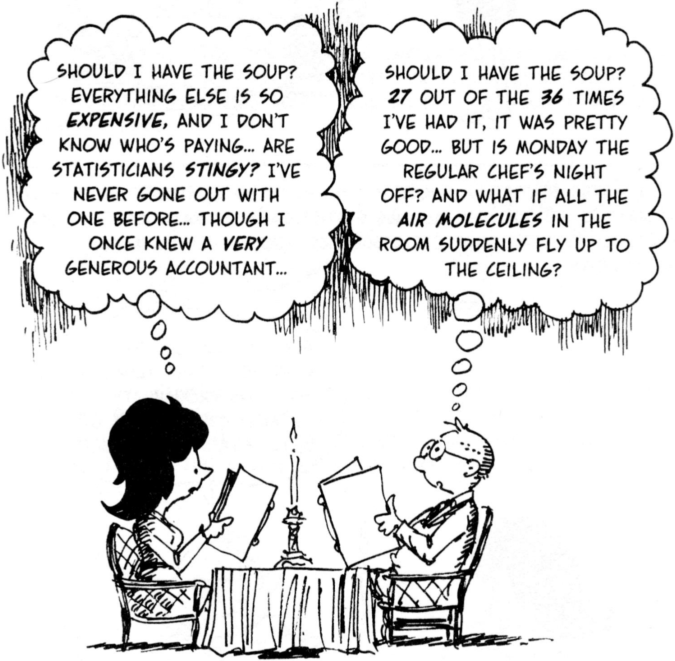
Institute of Economic Studies
Faculty of Social Sciences
Charles University in Prague

Statistics around us

Examples of reports in media:

- With respect to last week, the average price of the Natural 95 petrol increased by 38 hellers to CZK 37.15 a liter in the Czech Republic. The average price is highest since December 2023. (ČTK České noviny, February 8, 2024).
- The employment rate, seasonally adjusted, reached 75.0% in December 2023 and it decreased by 0.4 percentage point compared to that in December 2022. (ČSÚ, January 31, 2024).
- In Q3 2023 the median wage (CZK 37 492) increased by 7.1% compared to the same period of the previous year. (ČSÚ, December 4, 2024).
- The Dow Jones Industrial Average closed at 38,773.12 (The Wall Street Journal, February 15, 2024).

Descriptive statistics refers to numerical facts such as averages, medians, percents, and index numbers that help us understand the reality via available data.


SHOULD I HAVE THE SOUP? EVERYTHING ELSE IS SO EXPENSIVE, AND I DON'T KNOW WHO'S PAYING... ARE STATISTICIANS STINGY? I'VE NEVER GONE OUT WITH ONE BEFORE... THOUGH I ONCE KNEW A VERY GENEROUS ACCOUNTANT...
SHOULD I HAVE THE SOUP? 27 OUT OF THE 36 TIMES I'VE HAD IT, IT WAS PRETTY GOOD... BUT IS MONDAY THE REGULAR CHEF'S NIGHT OFF? AND WHAT IF ALL THE AIR MOLECULES IN THE ROOM SUDDENLY FLY UP TO THE CEILING?

Subject of Statistics

- in most scientific/industrial studies decisions are made based on data - past experiences/observations, results of some controlled process/experiments
- producing data enables analysis and drawing useful conclusions from them
- an inherent part of decision making is the knowledge of relevant information/data to the decision (possibly unknown or subject to future events/uncertainty) - to reduce guesswork
- statistics as a science: to analyze/approximate a part of reality based on limited information – science of data
- statistics involve collecting, classifying, summarizing, organizing, analyzing, presenting and interpreting data and also model building
- application in agriculture, astronomy, biology, business, economics, education, electronics, geology, medicine, engineering, weather forecast etc.

Application in Economics and Finance

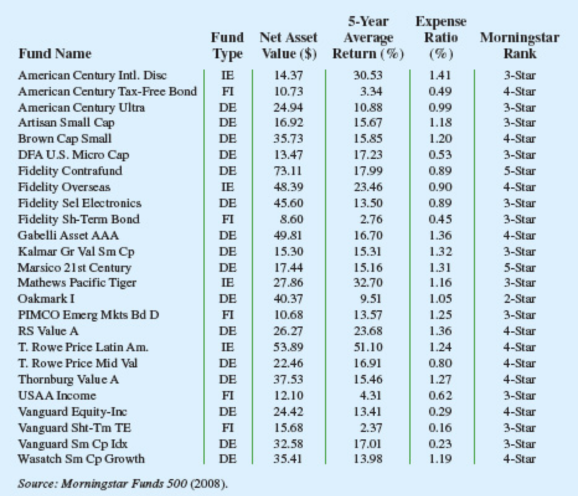
- accounting - sampling procedures when conducting audits - reviewing and validating every account can be too time-consuming and expensive - select and review a subsets of accounts and draw a conclusion about all accounts
- finance - statistical information as a guide to investment recommendations - e.g. in case of stocks review financial data including price/earnings ratios and dividend yields to draw a conclusion whether a stock is over- or underpriced
- marketing - electronic scanners and retail checkout counters collect data for a variety of marketing research applications - brand managers can review the scanner statistics to establish future promotional activities
- economics - statistical data for variety of forecasts about the future of aspects of economy - e.g. forecasting inflation rates, economists enter various indicators (Producer Price Index, unemployment rate, etc.) into a forecasting model

Data

- data are facts and figures collected, analysed and summarized for presentation and interpretation
- data set is a collection of all data in a particular study
- elements are the entities on which data are collected
- variable is a characteristic of interest for the elements
- measurements on each variable for every element in the study provide data set
- observation is the set of measurements for a particular element

Example Morningstar

Example 1: Consider the following data set containing information for 25 mutual funds that are part of the Morningstar Funds 500 for 2008.



Fund Name	Fund Type	Net Asset Value ($)	5-Year Average Return (%)	Expense Ratio (%)	Morningstar Rank
American Century Intl. Disc	IE	14.37	30.53	1.41	3-Star
American Century Tax-Free Bond	FI	10.73	3.34	0.49	4-Star
American Century Ultra	DE	24.94	10.88	0.99	3-Star
Artisan Small Cap	DE	16.92	15.67	1.18	3-Star
Brown Cap Small	DE	35.73	15.85	1.20	4-Star
DFA U.S. Micro Cap	DE	13.47	17.23	0.53	3-Star
Fidelity Contrafund	DE	73.11	17.99	0.89	5-Star
Fidelity Overseas	IE	48.39	23.46	0.90	4-Star
Fidelity Sel Electronics	DE	45.60	13.50	0.89	3-Star
Fidelity Sh-Term Bond	FI	8.60	2.76	0.45	3-Star
Gabelli Asset AAA	DE	49.81	16.70	1.36	4-Star
Kalmar Gr Val Sm Cp	DE	15.30	15.31	1.32	3-Star
Marsico 21st Century	DE	17.44	15.16	1.31	5-Star
Mathews Pacific Tiger	IE	27.86	32.70	1.16	3-Star
Oakmark I	DE	40.37	9.51	1.05	2-Star
PIMCO Emerg Mkts Bd D	FI	10.68	13.57	1.25	3-Star
RS Value A	DE	26.27	23.68	1.36	4-Star
T. Rowe Price Latin Am.	IE	53.89	51.10	1.24	4-Star
T. Rowe Price Mid Val	DE	22.46	16.91	0.80	4-Star
Thornburg Value A	DE	37.53	15.46	1.27	4-Star
USAA Income	FI	12.10	4.31	0.62	3-Star
Vanguard Equity-Inc	DE	24.42	13.41	0.29	4-Star
Vanguard Sht-Tm TE	FI	15.68	2.37	0.16	3-Star
Vanguard Sm Cp Idx	DE	32.58	17.01	0.23	3-Star
Wasatch Sm Cp Growth	DE	35.41	13.98	1.19	4-Star

Source: Morningstar Funds 500 (2008).

Scales of measurements

- nominal scale - data for a variable consist of labels or names used to identify an attribute of the element. For numerical purposes may be replaced by a numerical code 1, 2, ... E.g. Fund type, sex, nationality.
- ordinal scale - data exhibit the properties of nominal data and the order or rank of the data is meaningful. Ordinal data can also be provided using a numerical code. E.g. Morningstar Rank, highest achieved education
- interval scale - data have all the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure. Interval data are always numerical. E.g. year of birth, temperature in degrees Celsius.
- ratio scale - data have all the properties of interval data and the ratio of two values is meaningful. This scale requires that zero value be included. The scale of most of the variables we measure is a ratio scale. E.g. cost of a car, time, weight, distance.

Another types of classification

Categorical vs Quantitative:

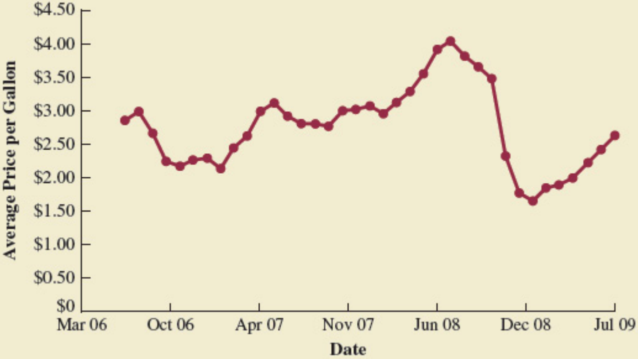
- categorical data - data grouped by specific categories; of either nominal or ordinal scale of measurement. Sometimes referred to as qualitative data. Statistical analysis of categorical variables is limited, we can summarize by counting number or proportion of observations in categories. It may not have sense to consider characteristics such as sum or average.
- quantitative data - data with numerical values indicating “how much” or “how many”. Arithmetic operations provide meaningful results for quantitative variables.

Cross-sectional vs Time series

- cross-sectional data - collected at the same or approximately the same point of time. E.g. Morningstar data set
- time series data - collected over several time periods. E.g. average price per gallon of gasoline between 2006 to 2009.

Example Average price per gallon of gasoline

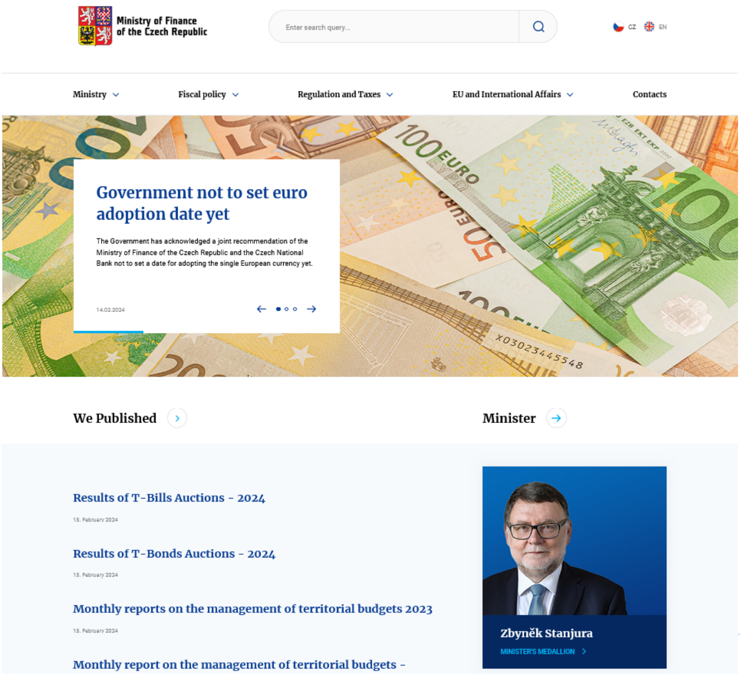
Example 2:



Source: Energy Information Administration, U.S. Department of Energy, July 2009.

Existing data sources

- data can be obtained from existing sources or from surveys and experimental studies designed to collect new data
- there are organizations that specialize in collecting and maintaining data, e.g. Bloomberg, Dow Jones & Company, Median, SCIO, Ministry of Finance, Czech Statistical office, Google
- examples of data from internal company records:
  - employee records: name, address, salary, number of vacation days, number of sick days, bonus
  - production records: part or product number, quantity produced, direct labor cost, materials cost
  - sales records: product number, sales volume, sales volume by region or customer type
  - credit records: customer name, address, phone number, credit limit, accounts receivable balance



Ministry of Finance of the Czech Republic

Enter search query...

CZ EN

Ministry | Fiscal policy | Regulation and Taxes | EU and International Affairs | Contacts

Government not to set euro adoption date yet

The Government has acknowledged a joint recommendation of the Ministry of Finance of the Czech Republic and the Czech National Bank not to set a date for adopting the single European currency yet.

14.02.2024

We Published

Results of T-Bills Auctions - 2024

15. February 2024

Results of T-Bonds Auctions - 2024

15. February 2024

Monthly reports on the management of territorial budgets 2023

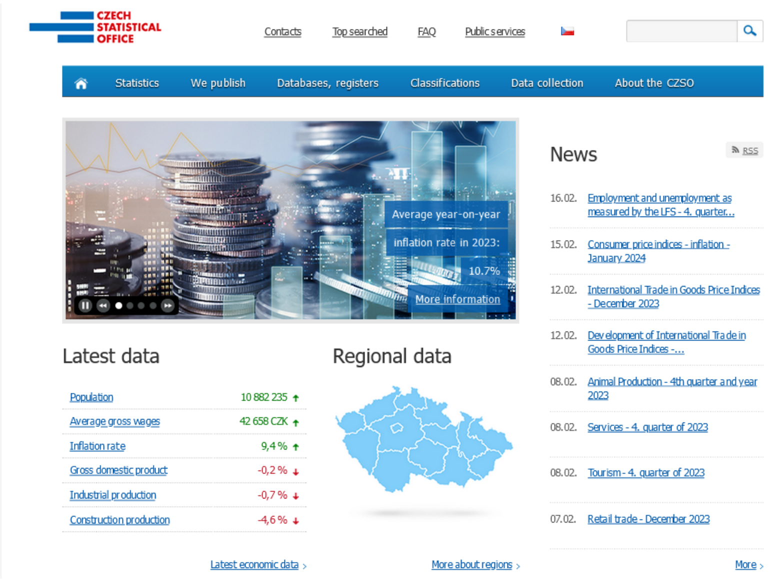
15. February 2024

Monthly report on the management of territorial budgets -

Minister

Zbyněk Stanjura

MINISTER'S MEDALLION



CZECH STATISTICAL OFFICE

Contacts · Top searched · FAQ · Public services

Statistics · We publish · Databases, registers · Classifications · Data collection · About the CZSO

Average year-on-year inflation rate in 2023: 10.7%

More information

News

RSS

- 16.02. Employment and unemployment as measured by the LFS - 4. quarter...
- 15.02. Consumer price indices - inflation - January 2024
- 12.02. International Trade in Goods Price Indices - December 2023
- 12.02. Development of International Trade in Goods Price Indices -...
- 08.02. Animal Production - 4th quarter and year 2023
- 08.02. Services - 4. quarter of 2023
- 08.02. Tourism - 4. quarter of 2023
- 07.02. Retail trade - December 2023

More ›

Latest data

Population	10 882 235 ↑
Average gross wages	42 658 CZK ↑
Inflation rate	9,4 % ↑
Gross domestic product	-0,2 % ↓
Industrial production	-0,7 % ↓
Construction production	-4,6 % ↓

Latest economic data ›

Regional data

More about regions ›

Statistical studies

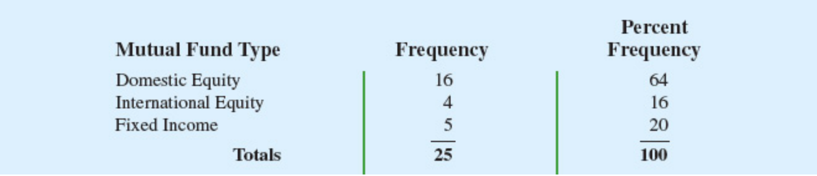
- sometimes data are not available - can be obtained by conducting a statistical study
- in an experimental study a variable of interest is first identified. Then one or more other variables are identified and controlled so that data can be obtained about how they influence the variable of interest. E.g. effects of a new drug on blood pressure.
- observational or non-experimental studies make no attempt to control the variables of interest. E.g. survey, when research questions are first identified and a questionnaire is designed and administered to a sample of individuals
- the type of source used depends on time and cost required to obtain the data

Descriptive statistics

- most of information in newspapers, company reports, etc. consists of data summarized in a form that is easy for the reader to understand (tables, graphs, numerical values) - descriptive statistics. They refer to a specific data set, results cannot be generalized to full population.

Example 1 (cont.):

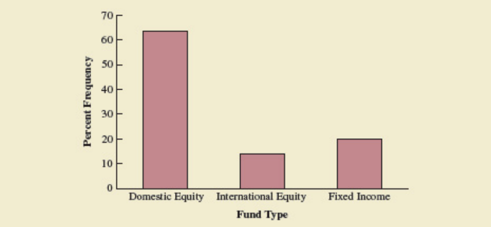
Tabular summary:

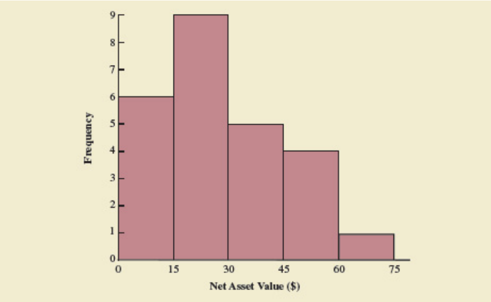


Mutual Fund Type	Frequency	Percent Frequency
Domestic Equity	16	64
International Equity	4	16
Fixed Income	5	20
Totals	**25**	**100**

Descriptive statistics

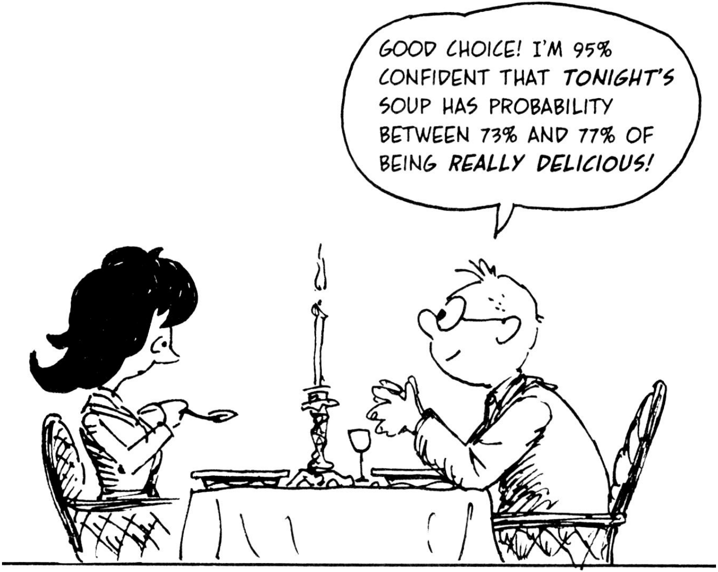
Graphical summary:





Statistical inference

- we often seek information about large group of elements - population - set of all elements of interest in a particular study (e.g. voters, companies, products etc.) but because of some reason (time, cost, etc.) data are collected from only a small portion of a group - sample - a subset of population
- conducting a survey to collect data for the entire population is called a census; collecting data for a sample is called sample survey
- statistics uses data from a sample to make estimates or test hypotheses about the characteristics of a population through a statistical inference


GOOD CHOICE! I'M 95% CONFIDENT THAT TONIGHT'S SOUP HAS PROBABILITY BETWEEN 73% AND 77% OF BEING REALLY DELICIOUS!