Introduction to Machine Learning and Data Mining

Statistics: the basics

assoc. prof. Branko Kavšek

Outline

Basic definitions

Distributions

Probability

Patterns

About statistics …

Definition:

1. Statistics is a branch of mathematics dealing with data collection, organization, analysis, interpretation and presentation.
   (from: Wikipedia)
2. Statistics is a form of mathematical analysis that uses quantified models, representations and synopses for a given set of experimental data or real-life studies.
   (from: Investopedia)

“Statistical” statements – examples

- The most violent earthquake measured 9.2 on Richter scale.
- The probability for murderers of being men is 10 times higher then for women.
- Every eighth South-african is infected with the HIV virus.
- In the year 2022 there will be 15 people older than 64 for each newborn.

Thus, statistics …

- … uses mathematical calculations,
- … deals with numbers.

But, is also important …

- … how we choose those numbers,
- … how we interpret the results of calculations.

Let's take a look at some examples →

Example no. 1

“Statistical” finding/result:

Due to a new commercial campaign in May the sales of ice cream XYZ went up 30% in the next 3 months.

The sales of ice cream in the summer months (June, July, August) goes up regardless of the commercial.

“Historical effect” – interpreting the result depending on one variable when in reality it is dependent on another (variable) – in our case *time*.

Example no. 2

“Statistical” finding/result:

The highest the number of churches in a city, the highest the criminal rate. Hence: churches lead to criminal.

Both the increase in the number of churches and criminal rate can be bound to the increase in a city's population – bigger city, more churches, more criminal.

“Third variable effect” – we wrongly assume that there is a connection between two variables where in fact there is a third variable affecting both variables.

Example no. 3

“Statistical” finding/result:

This year there is 75% more interracial marriages than 25 years ago.

What if 25 years ago there were 1% interracial marriages, this year 1.75% (75% more). Does this really mean a so drastic increase? What about the fluctuations in the years in between?

Lack of data – we simply do not have enough data, to make sound conclusions.

Why is it important to know statistics?

- We hear “statistical” statements, similar to those on previous slides, every day
  - We can believe to some
  - But, most of them can be deceiving
- The knowing of statistics enables us to differentiate between truth and deception
- **Statistics is an introduction to Data Mining**

Basic terminology and definitions

- Descriptive statistics
- Inferential statistics
  - sampling
- Variables/attributes
- Percentiles
- Measuring
  - How to choose a measure?
  - Data collection basics
- (probabilistic) Distributions
- Linear transformations

Descriptive statistics

- Describe the data at hand
- Do not “make conclusions” based on this data

- **Descriptive statistic**:

  Interesting, Americans are paying more for people that take care of their teeth and feet than for those protecting and educating their children.

  (is Slovenia different?)

- **Example** – table representing the average annual income of people in the US by occupation for the year 1999:

$ 112,760	pediatritians
$ 106,130	dentists
$ 100,090	podiatritians
$ 76,140	fizicists
$ 53,410	architects
$ 49,720	psychologists
$ 47,910	hosteses
$ 39,560	elementary school teachers
$ 38,710	policemen
$ 18,980	florists

Inferential statistics

- From properties of a **sample** we try to draw conclusions about the whole **population**
  - How to choose a “good” / random sample?
  - What is a sample's bias?

How to choose a sample? sampling

Rule:

sample **bias**

The sample has to be **representative** = has to represent the properties of the polulation + beware of the sample **size**!

- Types of sampling:
  - (simple) random sampling
  - advanced samplings:
    - random assignment
    - stratified sampling

Sampling – examples (1)

- Random sampling:
  - each individual from the population has to have **the same probability** of being chosen (in the sample)
  - The selection of one individual must not affect the selection of the others = **independence**

Example:

Among the Slovenian population, aged 19 to 35 years we survey all those individuals whose last name begins with the letter “Z”, but just every hundredth such person.

What is the problem?

Sampling – examples (2)

- The size of a sample:
  - Small samples are often **non-representative** = they do not represent the properties of the entire population

Example:

We infer the probabilities of a fair coin toss "coming out" head or tails form tossing such a coin 10 times.

What is the problem?

Sampling – examples (3)

- Random assignment:
  - there is no actual population; we deal with a **hypothetical population**
  - the sample from this hypothetical population is randomly split in 2 or more groups = the individuals from the sample get **randomly assigned** to groups

Example:

When testing the effect of a drug, we split a sample of people into 2 groups. To one group (the controls) we give the *placebo*, to the other the actual drug. We then observe whether there are differences between the two groups.

What could be the problem?

Sampling – examples (4)

- Stratified sampling:
  - We sample in layers (***stratus* = layer**) based on some property of the population

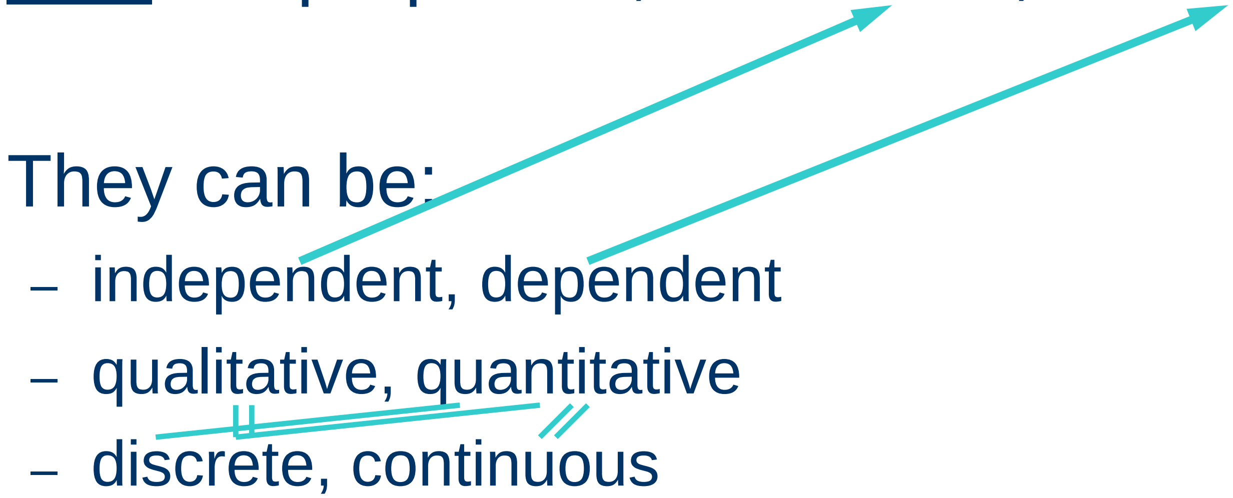
Example:

There are 1000 balls in the basket (population),
70% are **red**, **20%** are **green** and **10%** are **blue**. The property used for stratification is thus the ***color*** of the balls.

How to sample this population to get a representative sample?

Variables / attributes

- Also: properties, **attributes**, **classes**, …



- They can be:
  - independent, dependent
  - qualitative, quantitative
  - discrete, continuous

- More – a bit later in "**measuring things**"

Percentiles

- What is a **percentile**? – example:

Say, you did a test of motoric abilities and you scored 35 points out of a total of 50 points. What does this tell you about your motoric abilities? What are your motoric abilities compared to other participants on the testing?

A more informative indicator would be: “what percentage of people is (motorically) less capable than me?” → this percentage is called a **percentile**.

If your score is in the **65th percentile**, this means that **65%** of all people taking the test scored **worse** than you. In your case the ***65th percentile = 35***.

3 definitions of a percentile

Definition 1:

The *N^{th} percentile* is the lowest value that is *strictly greater than N%* of all values.

Definition 2:

The *N^{th} percentile* is the lowest value that is *greater than or equal to N%* of all values.

Definition 3:

A *weighted average* of the percentiles from the first two definitions (the most accurate definition that we are going to use)

Percentile definitions – example

Value	Rank
3	1
5	2
7	3
8	4
9	5
11	6
13	7
15	8

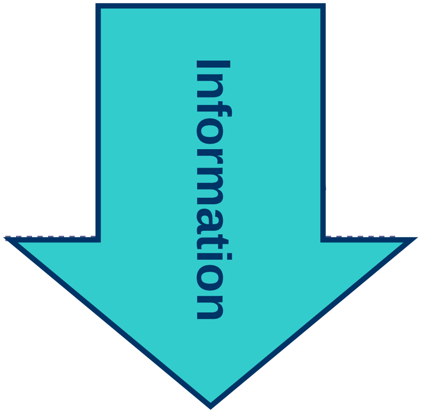
25th percentile = 5.5

Definition 3

How do we measure things?

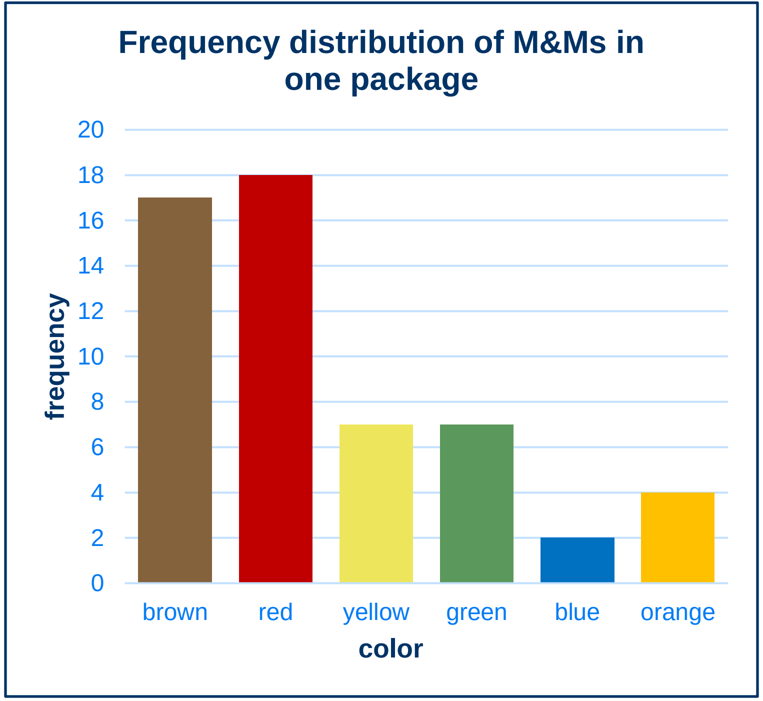
- In science data often come from measurings
- How can we measure?
  - Nominal (descriptive) values
  - Ordinal (ordered) values
  - Interval values
  - Ratio values
- Transformations between different types

  = basis of data collection / **errors**

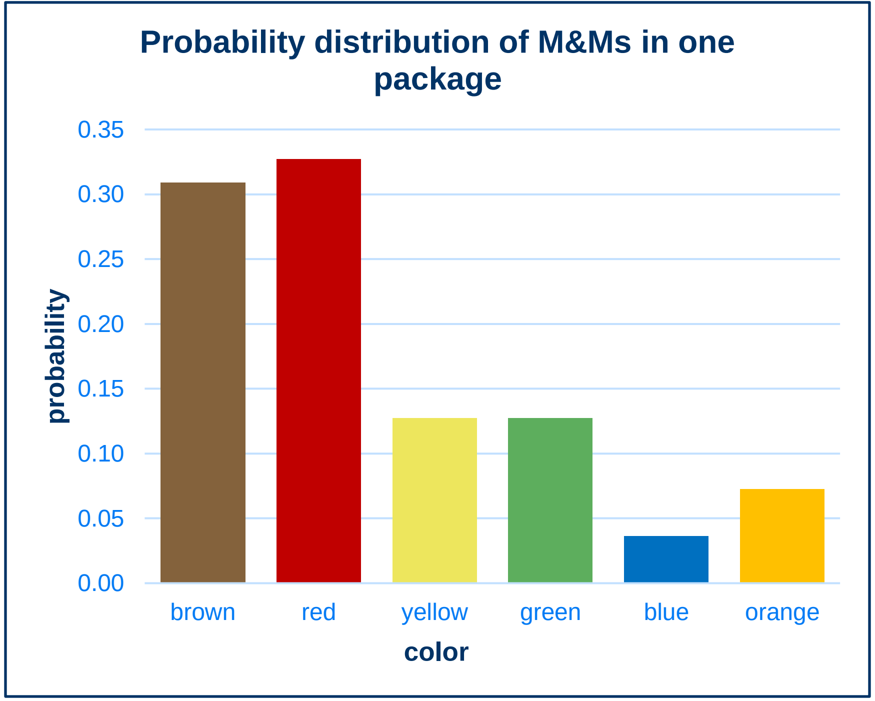


Distributions of discrete variables

Frequency distribution:

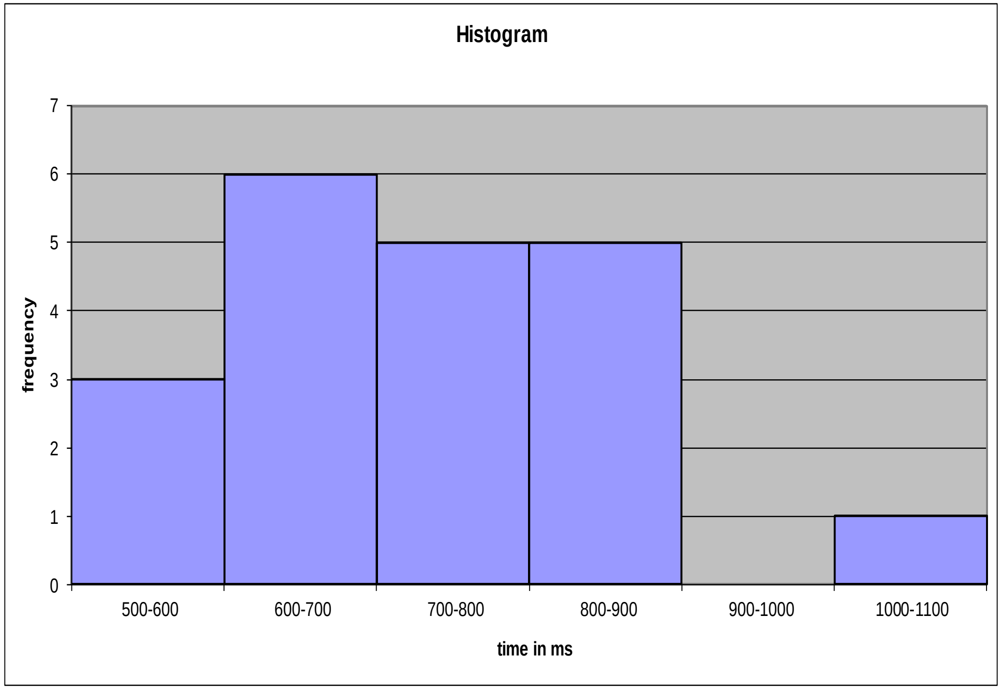


Probability distribution:



Distributions of continuous variables

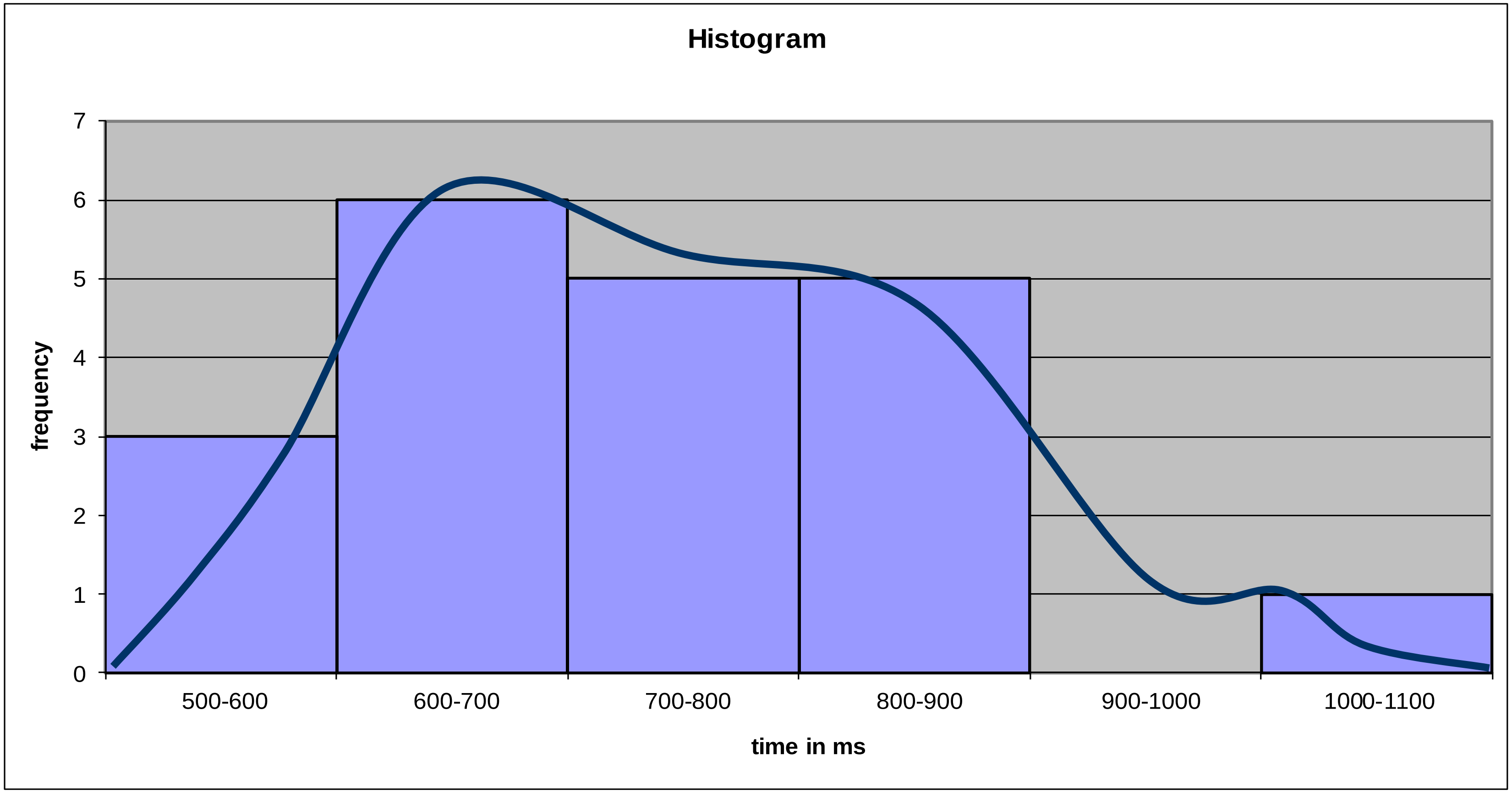
- Grouped frequency distribution
  - graphic → histogram



Interval	Frequency
500-600	3
600-700	6
700-800	5
800-900	5
900-1000	0
1000-1100	1

Time in *ms*
568
577
581
640
641
645
657
673
696
703
720
728
729
777
808
824
825
865
875
1007

Probability density



Linear transformations

- Transformation = to change/transform
- Linear = using only multiplication /w constant and/or adding a constant
  - if “original” and transformed values are depicted as a scatter plot, we “observe” a linear function.
- **Examples:**
  - Transformation of inches into centimeters (`x 2.54`)
  - Transformation from ºF into ºC (`x 9/5 + 32`)