

UPPSALA UNIVERSITET

FÖRELÄSNINGSANTECKNINGAR

Inferensteori

Rami Abou Zahra

Inlämningsdatum
November 1, 2022

CONTENTS

1. TODO	2
2. Data Analysis (K6)	3
2.1. Location Measures	3
2.2. Dispersion measures	3
2.3. Graphical illustration	5
2.4. Data materials in several dimensions	6

1. TODO

- Add slide data and calculation
- Experiment in r
- Understand .dat files
- Look at def. of QQ-plot and do some plotting problems
- Understand proof 2.1

2. DATA ANALYSIS (K6)

Vi kommer undersöka statistisk säkerställd skillnad (Opinion polls example), hypotestestning (räknar sannolikheten att hypotesen är sann).

Anmärkning:

Vanligtvis antar vi att datan är normalfördelad, men inte i alla fall (såsom stickprov av lön)

2.1. Location Measures.

A data set is given by x_1, \dots, x_n

Definition/Sats 2.1: Sample mean

Sample mean is given by:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Definition/Sats 2.2: Median

The "middle value" of the sorted data. Different from the mean.

If n is even, the median is defined as the mean of the two middle values

Definition/Sats 2.3: Mode

This doesn't work if it's continuous data but it can be made discrete (such as age/time)

Mode is the most common data value

Example:

Let our data points be:

32 34 41 44 45 50 50 54 55 57 58 60 63

Find mean, median mode:

Mean: 13 data sets $\Rightarrow n = 13$:

$$\frac{1}{13}(32 + 34 + 41 + 44 + 45 + 50 + 50 + 54 + 55 + 57 + 58 + 60 + 63) \approx 49.46$$

Median: The middle value is 50

Mode: 50 is the only data value appearing more than once.

Anmärkning:

In this example, the median = mode. This is not always the case!

2.2. Dispersion measures.

Describes the "spread" of the data, such as the variance. We have the following:

Definition/Sats 2.4: Sample variance

The sample variance is given by:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Definition/Sats 2.5: Sample standard variance

Is given by:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Definition/Sats 2.6: Range

Variationsbredden. The difference between the largest and the smallest values of the data

Definition/Sats 2.7: Inter quartile range

Kvartilavståndet is the difference between the upper and lower quartiles.
If we have an odd amount of data it is including the median!

Example:

See slide 2

Definition/Sats 2.8: Lower/Upper quartile

The *lower quartile* is the median of the lower half of the data material including the median if n is odd

The *upper quartile* is the median of the upper half of the data material including the median if n is odd

Example:

Slide 3

2.3. Graphical illustration.**Stem and leafplots:**

```
u = c(32,34,...)
stem(u)
```

Boxplots:

Uses quartiles, max min, and median. Useful if you want a quick look at the dispersion of data.

Bar chart:

Good for illustrating the frequency of each data point, but for large data points the data is hard to read

Histogram:

Attempts to fix the readability issues with the bar chart and is easier to compare with probability density functions.

Easier to manipulate data for readability (use bigger/smaller intervals) (one can ask what the optimal width for a histogram would be)

Very often you can ask if the data follow a normal distribution, which can be hard by just looking at the histogram (because the width varies)

Thoughts:

Dynamically widths on histograms, the more sparse data the greater the width and the more dense, the less the width

QQ-plot:

Is the data normally distributed? You order your data and construct a table with your data and compare it with if it was normally distributed:

$$\Phi(z) = \frac{i - 0.5}{n}$$

If data was perfectly normal, x_i would be a linear function of z .
We plot z on the x -axis and x_i on the y -axis

2.4. Data materials in several dimensions.

We can calculate correlation through sample covariance:

Definition/Sats 2.9: Sample covariance

$$\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Not scale invariant (if you measure x in meters and go to cm then it is not the same). Therefore we need to norm it with something, which is where the correlation comes in:

Definition/Sats 2.10: Sample correlation coefficient

$$r_{xy} = \frac{c_{xy}}{s_x s_y}$$

Where s_x and s_y are the sample standard deviations for x and y

Definition/Sats 2.11: Sample correlation satisfies

The sample correlation coefficient satisfies

$$-1 \leq r_{xy} \leq 1$$

If it is 1, then there is a strong positive correlation (the linear regression has a line with positive derivative), similarly for negative.

When it is 0 there is no *linear* relation. There might be other, for example quadratic relation.

Bevis 2.1: Sample correlation satisfaction

$$\begin{aligned} 0 &\leq \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} - \frac{y_i - \bar{y}}{s_y} \right)^2 \\ &= \underbrace{\frac{1}{s_x^2} \frac{1}{n-1} \sum_i (x_i - \bar{x})^2}_{s_x^2} + \underbrace{\frac{1}{s_y^2} \frac{1}{n-1} \sum_i (y_i - \bar{y})^2}_{s_y^2} - 2 \underbrace{\frac{1}{s_x s_y} \frac{1}{n-1} \sum_i (x_i - \bar{x})(y_i - \bar{y})}_{c_{xy}} \\ &= 2 - 2r_{xy} \Rightarrow r_{xy} \leq 1 \\ 0 &\leq \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} + \frac{y_i - \bar{y}}{s_y} \right)^2 = 2 + 2r_{xy} \\ &\Rightarrow -1 \leq r_{xy} \end{aligned}$$

□