UPPSALA UNIVERSITET

Föreläsningsanteckningar

Inferensteori

Rami Abou Zahra

Inlämningsdatum
November 1, 2022

Contents

1. TODO

- Experiment in r (QQ-plot of exp vs n(0,1) data)
- 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 $x_1, \cdots, x_n$

Definition/Sats 2.1: Sample mean

Sample mean is given by:

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

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 doesnt work if its continous 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 datavalue 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:

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

Definition/Sats 2.5: Sample standard variance

Is given by:

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

Definition/Sats 2.6: Range

Variationsbredden. The differnece 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!

Definition/Sats 2.8: Lower/Upper quartile

The *lower quartile* is the median of th elower 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:

$$0\ 0\ 1\ 1\ 2\ 2$$

Here, the mean is given by $\dfrac{(1+1+2+2)}{6} = 1$.
Therefore, the sample variance is given by $\dfrac{4}{5}$ and the sample standard deviation $\sqrt{\dfrac{4}{5}}$
We can find the inter quartile range by looking at the half, like this:

$$[0\ \underbrace{0}_{\Delta}\ 1]\ [1\ \underbrace{2}_{\Delta}\ 2]$$

Therefore, the inter quartile range here is $2 - 0 = 2$

2.3. **Graphical illustration.**

Stem av 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:
Attemps 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 on both axis, x_i would be a linear funcfion of z, ie. normally distributed $N(0, 1)$
We plot z on the x-axis and x_i on the y-axis

The name comes from quantile-quantile-plot (QQ-plot). It is a graphical way of comparing two probability distributions (sannolikhetsfördelning)

2.4. **Data materials in several dimensions.**

We can calculate correlation through sample covariance:

Definition/Sats 2.9: Sample covariance

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

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

$$0 \leq \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{s_x}-\frac{y_i-\overline{y}}{s_y}\right)^2$$

$$= \frac{1}{s_x^2}\underbrace{\frac{1}{n-1}\sum_i (x_i-\overline{x})^2}_{s_x^2} + \frac{1}{s_y^2}\underbrace{\frac{1}{n-1}\sum_i (y_i-\overline{y})^2}_{s_y^2} - 2\frac{1}{s_x s_y}\underbrace{\frac{1}{n-1}\sum_i (x_i-\overline{x})(y_i)(\overline{y})}_{c_{xy}}$$

$$= 2-2r_{xy} \Rightarrow r_{xy} \leq 1$$

$$0 \leq \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{s_x}+\frac{y_i-\overline{y}}{s_y}\right)^2 = 2+2r_{xy}$$

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

□