UPPSALA UNIVERSITET

Föreläsningsanteckningar

Inferensteori

Rami Abou Zahra

Inlämningsdatum
November 9, 2022

Contents

1. TODO

- Experiment in r (QQ-plot of exp vs n(0,1) data)
- Understand .dat files
- Add proof from book of theorem 4.9
- Problems 7.2.2 in the book
- Stora talens lag
- MLE better than methods of moments
- Derivatan av binomial

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: Mid-range

The mean between the biggest and smallest value in the sample

Definition/Sats 2.9: 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 2.10: 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 2.11: 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 2.12: 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

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

□

3. Statistical Inference

Anmärkning:
If $X \sim Hyp$, then for a large population $X \sim Bin$ (because the chance of picking the same one in a large population is so small)

Anmärkning:
If $X_1, \cdots, X_n$ are independent and equally distributed $N(\mu, \sigma^2)$ variables, then

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$$

In the same way, the mean of these random variables is normally distributed with

$$\overline{X}_n \sim N(\mu, \sigma^2/n)$$

Example (*):
In an opinion poll, 1000 randomly selected voters are asked about their politcal sympathies. Let X be the number of these voters who sympathise with the party P. $X \sim Hyp$, but because the number of voters is so large, we may assume that $X \sim Bin(1000, p)$
Suppose we know that 100 of the selcted voters sympathise with P, what can p be?
Well, it makes sense that $p = \dfrac{100}{1000}$, the probability of choosing 100 from our 1000 people.

Definition/Sats 3.1: Sample (stickprov)

$x_1, x_2, \cdots, x_n$ is a *sample* from the random variable X with distribution F_X

If $X = (X_1, \cdots, X_n)$ are independent, we have a *random sample* from X

We can purposely choose our random variable in such way that makes it easier for us to analyse. It also allows us to compare these observations.

Example:
Using the same environment as example (*), we can either have a random variable $X \sim Bin(1000, p)$, *or* we can have let the 1000 people all have an associated Bernoulli distributed random variable.
In the first case, our observed sample has size $n = 1$, so our $x_1 = 100$ is an observation of the random variable X_1
In the second case, our observed sample has size $n = 1000$, with each $X_1, \cdots, X_{1000} \sim Be(p)$ and

$$\sum_{i=1}^{1000} x_i = 100$$

The pros in splitting up our situation into smaller random variables, is as touched upon previously, easier to analyse since it is easier to find independent random variables.

4. Estimation

Suppose we already know what our distribution function F is. We know this F is associated to our random variable, that we have our measured observations on.
We also know that our distribution function normally takes on certain *parameters*, for example, the Poisson distribution

$$f_X(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

Is actually a function of both x and λ.
We call λ in our case, an *unknown parameter.* Given enough data, and knowing the distribution function, we should be able to estimate what the value of this parameter is.

We can write the following for our data:

$$x = (x_1, x_2, \cdots, x_n)$$
$$X = (X_1, X_2, \cdots, X_n)$$

Definition/Sats 4.1: Estimate (skattning)

An *estimate* $\theta^* = \theta^*(x)$ is a funciton of the sample x

The estimate is an observation of the estimator $\theta^*(X)$

This attempts to put the previous paragraphs into "functions". Given a sample from our random variable (given in $x = x_1, \cdots$), we want to find the unknown parameter θ. We can then construct a general formula for *any* data given that it comes from the random variable we have agreed upon beforehand.
This is the *estimate* function, as defined above, of a given sample x

Example:
Let $x = (x_1, x_2, x_3, x_4, x_5) = (200, 185, 210, 190, 190)$ be a random sample from $X \sim N(\mu, 100)$ (every $X_i \sim N(\mu, 100)$)

In order to estimate μ by the sample mean, we induce the function μ^* on our sample set and calculate the mean:

$$\mu^*(x) = \overline{x} = 195$$

Since the estimate is an observation of the *estimator*, lets look at the estimator:

$$\mu^*(X) = \overline{X} = \frac{1}{5}\sum_{i=1}^{5} X_i \sim N(\mu, 100) \Rightarrow \overline{X} \sim N(\mu, 100/5) = N(\mu, 20)$$

Example:
Using the poll example from above, what we really did when we said that p reasonably has to be $\frac{100}{1000}$ is determine an estimate p^*

The greater the stickprov the better the estimate (because less and less variance)

Anmärkning:
The estimator is not distributed with the same distribution, since the estimator is not always an integer. We saw this in the above example, where the estimator was $N(\mu, 20)$ distributed but our estimate was $N(\mu, 100)$ distributed.

If we have different estimates, we need to make a reasonable choice such that our error is as little as possible (this is why we introduce estimators)

4.1. Properties of estimates.

The purpose of our estimates is to estimate θ. When we calculate using our function θ^* on our sample data we get a value that may or may not deviate from the actual value θ

We can take $\theta^* - \theta = E(\theta^*(X)) - \theta + (\theta^* - E(\theta^*(X)))$
It turns out, this is equal to the systematic error + random error

Definition/Sats 4.2: Unbiased (väntevärdesriktigt)

An estimate θ^* is said to be *unbiased* if it satisfies $E(\theta^*(X)) - \theta = 0$

This is the same as saying it has no systematic error (therefore, we only have the random error left)

Example:
We show this by:

$$E(\mu^*(X)) = E(\overline{X}) = \mu$$

Let

$$p^*(X) = \frac{X}{1000} \qquad X \sim Bin(1000, p)$$

Is $p^*(x)$ an unbiased estimate of p?
Take the expected value function on both sides:

$$E\left(\frac{X}{1000}\right) = E(p^*(X)) \qquad X \sim Bin(1000, p)$$
$$\frac{1}{1000}E(X) = \frac{1}{1000} \cdot 1000 \cdot p = p$$
$$\Rightarrow E(p^*(X)) = p \Rightarrow p - p = 0$$

If we have more than one unbiased estimate, which is the best one? Well, in that case we need to start looking at the random error. We can study this by looking at the variance

Definition/Sats 4.3: Efficiency comparison of estimates

If θ_1^* and θ_2^* are unbiased estimates of θ and

$$V(\theta_1^*(X)) \leq V(\theta_2^*(X))$$

For all θ with strict inequality for some. We say that θ_1^* is more *efficient* than θ_2^* (less random error)

Example:
See slide 6 & 7

Example: (stratification)
We are interested in the proportion p of Swedish citizens that last year have traveled by plane in connection with work. Take a sample consisting of $n = 1000$ people
It is safe to assume that the number of men that take the plane will be greater than the number of woman, we can denote this using $p \pm a$. Since we are looking at the combinations of ways we can choose men and women from our set, a reasonable distribution would be a binomial distribution for men $Bin(500, p + a)$ while for women $Bin(500, p - a)$. One can also look at Bernoulli distributions and get the same results.

$$p_1^*(X) = \frac{1}{1000}X \Rightarrow E(p_1^*) = p$$
$$p_2^*(y, z) = \frac{1}{1000}y + \frac{1}{1000}z$$
$$\Rightarrow \frac{1}{1000}E(y) + \frac{1}{1000}E(z)$$
$$\frac{1}{1000}500(p + a) + \frac{1}{1000}500(p - a) = p$$
$$V(p_1^*) = \frac{p(1-p)}{1000} \geq V(p_2^*) = \frac{p(1-p)}{1000} - \frac{a^2}{1000}$$

Definition/Sats 4.4: Standard error (medelfelet)

We want to assign a numerical value to the dispersion of an estimate.
Therefore, we define the *standard error* of the estimate θ^* is an estimate of the standard deviation $D(\theta^*)$

Denoted by $d(\theta^*(x)) = d(\theta^*)$

Recall that the standard deviation is given by $\sqrt{Var}$

Example:
Given $x = 100$ (one observation) of the random variable $X \sim Bin(1000, p)$, estimate $p^* = \dfrac{x}{1000} = 0.1$ and calculate the standard error of this estimate.

4.2. **Asymptotic properties.**

The accuracy of an estimate should improve as the sample size increases, seems reasonable

Definition/Sats 4.5: Bias

The *bias* (väntevärdesfelet) for the estimate θ^* is defined as

$$B(\theta^*) = E(\theta^*) - \theta$$

Anmärkning:
An unbiased estimate has bias 0

Definition/Sats 4.6: Asymptotically unbiased

If the bias $B(\theta_n^*)$ tends to zero as $n \to \infty$ for all θ, the estimate θ_n^* is said to be *asymptotically unbiased*

Think of it like, the more data you gather, the less the bias.

Example:
Let $x_1, \cdots, x_n$ be a random sample from $N(\mu, \sigma^2)$ where μ is unknown. We want to estimate σ^2

The estimate $\sigma_n^{2*} = \dfrac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$ is biased, but it is asymptotically unbiased

The estimate s_n^2 is unbiased for σ^2, why?

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

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

$$E(s^2) = \frac{1}{n-1}\left(\sum_{i=1}^{n} E(x_i^2) - nE(\overline{X}^2)\right)$$

$$E(x_i^2) = V(x_i) + (E(x_i))^2 = \sigma^2 + \mu^2$$

$$E(\overline{x}^2) = V(\overline{x}) + (E(\overline{x}))^2 = \frac{\sigma^2}{n} + \mu^2$$

$$E(s^2) = \frac{1}{n-1}\left(\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right)$$

$$= \frac{1}{n-1}\left(n(\sigma^2 + \mu^2) - \sigma^2 - n\mu^2\right) = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2$$

Definition/Sats 4.7: Consistent estimate

The estimate θ_n^* is said to be *consistent* for θ if the corresponding estimator converges to θ in probability for all θ

Definition/Sats 4.8: Convergence in probability

The estimator θ_n^* converges to θ in probability if $\forall \varepsilon > 0$:

$$\lim_{n\to\infty} P(|\theta_n^* - \theta| > \varepsilon) = 0$$

Definition/Sats 4.9

If the estimate θ_n^* is asymptotically unbiased and

$$\lim_{n\to\infty} V(\theta_n^*) = 0$$

Then it is consistent

Example:
Let $x_1, \cdots, x_n$ be a random sample from $N(\mu, \sigma^2)$ where σ^2 is known.
Estimate μ by

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

Show that the estimate is unbiased: $\mu - \mu = 0$

Calculate the variance for the corresponding estimator: $\dfrac{\sigma^2}{n}$

Show that the estimate is consistent: $\lim_{n\to\infty} \dfrac{\sigma^2}{n} = 0$

If an estimate is not consistent, then it does not matter if the sample size is increased, it wont yield better results.

5. Methods of estimation

5.1. Methods of moments.

Often, this method works quite well but it also tends to fail (**CHECK**)

Definition/Sats 5.1: Method of moments

Let $x_1, \cdots, x_n$ be a random sample from the random variable X with $E(X) = m(\theta)$ where θ is the parameter in teh distribution for X

If θ is one dimensional, the moment estimate $\theta = \theta^*$ solves the equation $m(\theta) = \overline{x}$

Example:
Slide 1

$m(\beta) = E(X) = \dfrac{1}{\beta}$
Solve $\overline{x} = m(\beta) = \dfrac{1}{\beta} \Rightarrow \beta = \dfrac{1}{\overline{x}}$

Therefore, moment estimate is $\beta^* = \dfrac{1}{\overline{x}}$

Example:
Slide 2

$m(p) = E(X) = np$
Solve $x = \overline{x} = m(p) = np \Rightarrow p = \dfrac{x}{n}$ (recall political party example)

Moment estimation is $p^* = \dfrac{x}{n}$

Example:
Slide 3 (a third of those 15 birds have rings, so we can get 30 from there)

The number of birds captured with ring $= X \sim Hyp(N, n, m)$, where $N =$ number of birds, $n =$ how many were captured the second day (15) and $m =$ how many with a ring in total the second day $= 10$:

$$p_X(x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}$$

We want to estimate N using the method of moments:

$$E(X) = n\frac{m}{N} = 15\frac{10}{N} = \frac{150}{N}$$

Solve for $f = x = \overline{x} = \dfrac{150}{N} \Rightarrow N = \dfrac{150}{5} = 30$. Moment estimation is $N^* = 30$

Definition/Sats 5.2: Method of moments with multiple parameters

If the parameter $\theta = (\theta_1, \theta_2)$, then the moment estimates solves the system:

$$E(X) = m_1(\theta_1, \theta_2) = \overline{x}$$

$$E(X^2) = m_2(\theta_1, \theta_2) = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

Example:
Slide 4

$$m_1(\mu, \sigma^2) = E(X) = \mu$$
$$m_2(\mu, \sigma^2) = E(X^2) = V(X) + (E(X))^2 = \sigma^2 + \mu^2$$

Solve

$$\begin{cases} \mu = \overline{x} \\ \sigma^2 + \mu^2 = \dfrac{1}{n} \sum_{i=1}^n x_i^2 \end{cases}$$

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

Almost looks like s^2, but here in the denominator we have n instead of $n - 1$

Moment estimates are:

$$\begin{cases} \mu^* = \overline{x} \\ \sigma^{2^*} = \dfrac{1}{n} \sum_{i=1}^n (x_i - \overline{x})^2 \end{cases}$$

Definition/Sats 5.3

Let $x_1, \cdots, x_n$ be a random sample from the random variable X where $V(X) = \sigma^2$

Then the sample variance is given by:

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

is an *unbiased estimate of* σ^2

5.2. **Maximum likelihood.**

Example:
Slide 5

Definition/Sats 5.4: Maximum likelihood

Let $x_1, \cdots, x_n$ be a random sample from X which has distribution $F(X; \theta)$

The likelihood function $L(\theta)$ is defined by:

$$L(\theta) = \begin{cases} \prod_{i=1}^n p(x_i; \theta) \quad X \text{ discrete} \\ \prod_{i=1} n f(x_i; \theta) \quad X \text{ continous} \end{cases}$$

The *maximum likelihood* estimate (MLE, ML-skattning) of θ is the θ that maximizes the likelihood function

Example:
Slide 6

Example:
Slide 7

Example:
Slide 8

Example:
Let $x_1, \cdots, x_n$ be a random sample from $X \sim N(\mu, \sigma^2)$ where μ and σ^2 are both unknown.

Estimate μ and σ^2 by MLE:

$$
\begin{aligned}
f_X(x;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \\
L(\mu,\sigma^2) &= \prod_{i=1}^{n} f_X(x_i;\mu,\sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \\
&= (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \\
l(\mu,\sigma^2) &= ln\left((2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}\right) \\
&= -\frac{n}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \\
\frac{\partial l}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = \frac{1}{v}\left(\sum_{i=1}^{n} x_i - n\mu\right) = \frac{n}{v}\left(\overline{x}-\mu\right) \\
\frac{\partial l}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 \\
\frac{\partial l}{\partial \mu} &= 0 \Rightarrow \mu = \overline{x} \\
\frac{\partial l}{\partial \sigma^2} &= 0 \Rightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 \qquad (\mu = \overline{x})
\end{aligned}
$$

MLE is therefore:

$$
\begin{aligned}
\mu^* &= \overline{x} \\
\sigma^{2^*} &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2
\end{aligned}
$$

Example:

Slide 9 (not supposed to follow, research level) (ibland för att få fram skattning måste man ta till med numeriska metoder)