SF1901: Sannolikhetslära och statistik
Föreläsning 8.
Statistik: statistiska inferensproblem, maximum likelihood, minsta kvadrat

Jan Grandell & Timo Koski

10.02.2011

KTH Matematik

Statistisk inferens

En statistisk inferens är utsaga om en egenskap hos en hel population som baserar sig på ett studium av en (kanske relativt liten) del av populationen. Denna föreläsning handlar om två metoder att konstruera punktskattningar:

- maximum likelihood -metoden
- minsta-kvadrat - metoden

KTH Matematik

Skattning av väntevärdet μ

Sats

Stickprovsmedelvärdet $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ *som skattning av väntevärdet* μ *är*

1) *Väntevärdesriktig;*
2) *Konsistent;*
3) *Ej nödvändigtvis effektiv, dvs. den effektivaste möjliga skattningen.*

Bevis.

1) $E(\overline{X}) = \mu$.
2) $V(\overline{X}) = \sigma^2/n$ och stora talens lag gäller.
3) Det finns motexempel (den intresserade hänvisas till Blom et al.)

Punktskattning: ett vanligt inferensproblem

Exempel
På en laboration vill man bestämma den fysikaliska konstanten θ. Vi gör upprepade mätningar av θ t.ex.

$$\text{mätvärde} = \theta + \text{ slumpmässigt mätfel}$$

och erhåller följande mätvärden:

$$x_1, x_2, \ldots, x_n$$

Problem
Hur skall vi *skatta* θ så bra som möjligt. Här liksom i de föregående exemplen arbetar vi med *punktskattning* (till skillnad från *intervallskattning* som kommer att presenteras längre fram i kursen).

KTH VETENSKAP OCH KONST
KTH Matematik

Mätning av ljusets hastighet

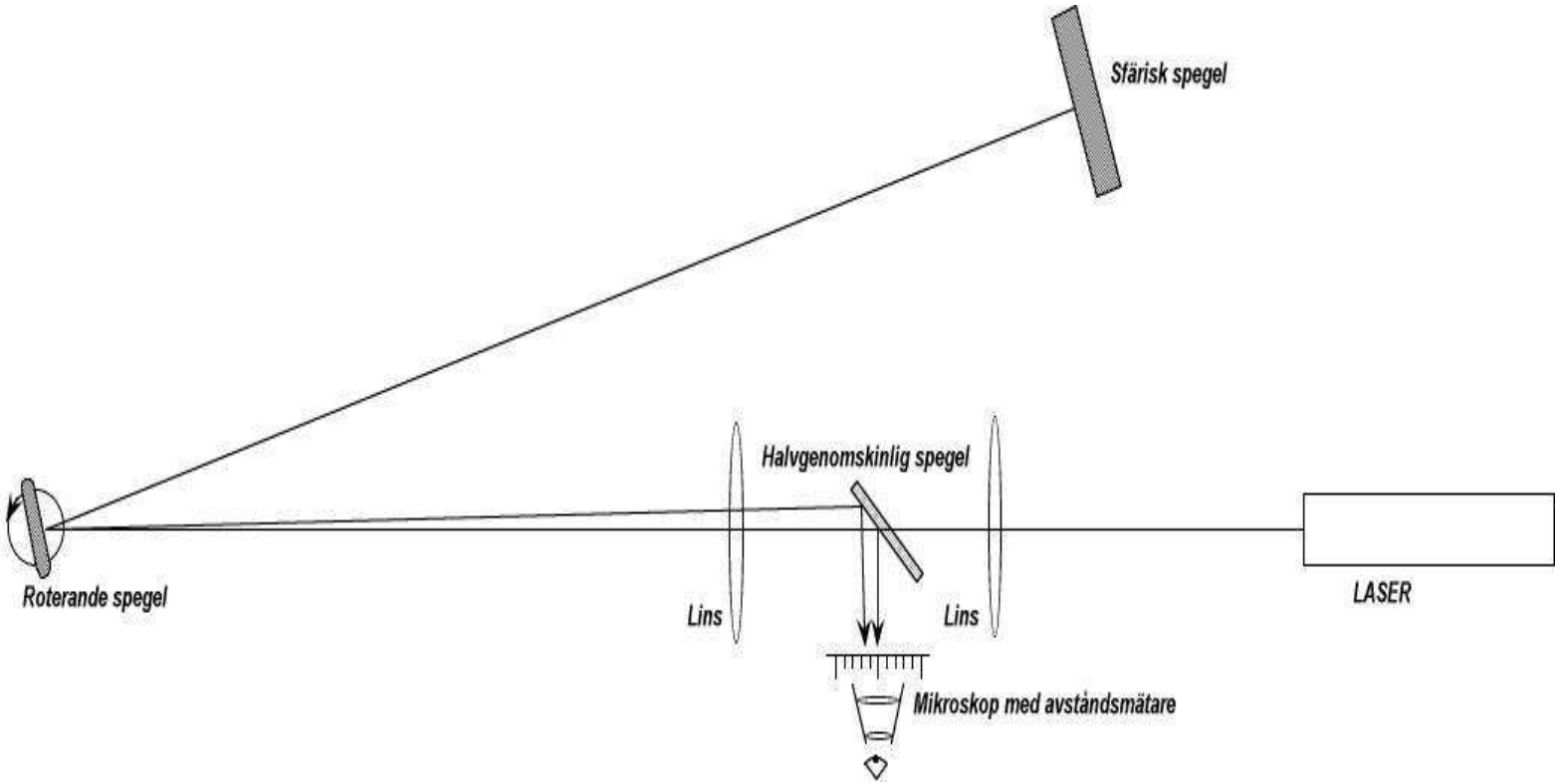
Exempel

För bestämning av ljusets hastighet (nu accepterad som $c = 299792458(\leftrightarrow \theta)$ m per sekund), utvecklade Simon Newcomb (1835 – 1909, astronom och matematiker vid U.S. Navy) och Albert Abraham Michelson (1852 1931, fysiker , året 1907s nobelpristagare i fysik) en optisk utrustning som bestod av en fast sfärisk spegel, placerad vid foten av Washington Monument (Washington D.C.) och en roterande spegel och en optisk lins vid nuvarande Fort Myer. Genom att reglera spegelns rotation, mäta avståndet mellan monumentet och Fort Myer, avståndet mellan en ljuskälla och den roterande spegeln, och förskjutningen mellan den skickade ljusstrålen och den reflekterade ljusstrålen, kan ljusets hastighet bestämmas (per en formel).

Det finns många källor till mätfel i detta.

KTH Matematik

Mätning av ljusets hastighet: utrustningen (lite moderniserad)



Mätning av ljusets hastighet

Exempel

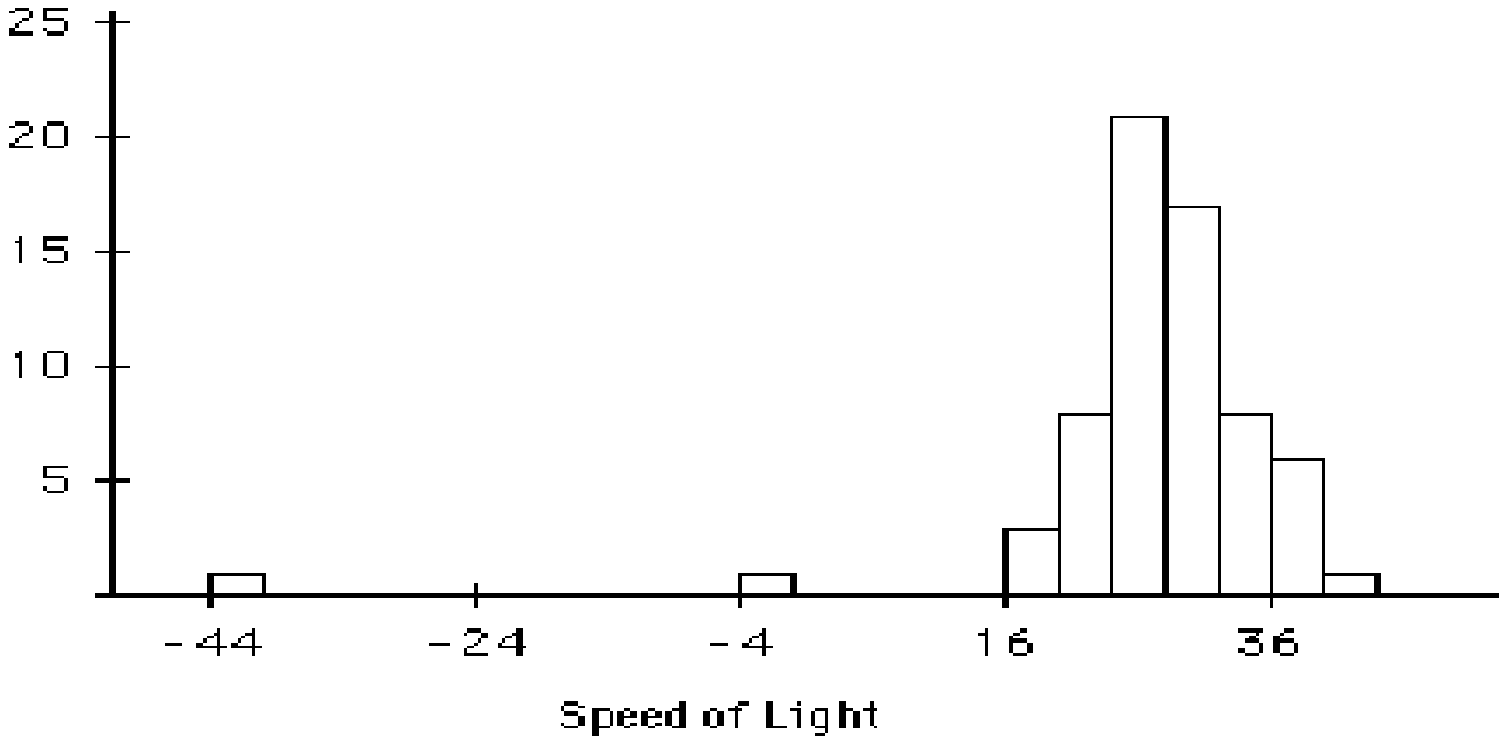
De exakta tekniska detaljerna för hur dessa mätningar gjordes redovisas i R.J. MacKay & R.W. Oldford : Statistical Method and the Speed of Light. *Statistical Science*, vol. 15, 2000, pp. 254−278.
W.E. Cater & M.S. Carter: The Newcomb-Michelson Velocity of Light Experiments. *Eos, Transactions, American Geophysical Union*, vol 83, 2002, pp. 405−410.

KTH VETENSKAP OCH KONST
KTH Matematik

Mätning av ljusets hastighet

Exempel

Newcomb-Michelsons data är differenserna från 24,800 ns. (Addera 299,800 för att få svaret i km/s.) Histogrammet för 66 differenser är enligt figuren:



Mätning av ljusets hastighet

Exempel

Newcomb-Michelsons svar blev 299860 km/s. Det intressanta för eftervärldens statistiker är inte svaret utan hur data hanterades och mätfel och systematiska fel eliminerades med statistiska metoder.

KTH Matematik

Punktskattning **Modell**

Vi uppfattar mätvärdena ($= \theta +$ mätfel), som utfall av n st. oberoende och lika fördelade s.v. $X_1, X_2, \ldots, X_n$ med $E(X_i) = \theta$ och $V(X_i) = \sigma^2$.
En *punktskattning* θ^*_{obs} av θ är en funktion av mätvärdena:

$$\theta^*_{\text{obs}} = \theta^*(x_1, \ldots, x_n).$$

KTH VETENSKAP OCH KONST
KTH Matematik

Punktskattning **Modell**

När vill vill analysera en skattning ersätter vi observationerna med de underliggande stokastiska variablerna. Vi säger då att

$$\theta^* = \theta^*(X_1, \ldots, X_n)$$

är en *stickprovsvariabel*.

Stickprovsvariabeln är själv en stokastisk variabel, vars fördelning beror av fördelningen för $X_1, X_2, \ldots, X_n$ och därmed av θ.

KTH Matematik

Punktskattning **Modell**

Om vi inte använder någon statistisk teori så väljer vi antagligen

$$\theta^*_{\text{obs}} = \overline{x} = \frac{1}{n}\sum x_i.$$

För motsvarande stickprovsvariabel $\theta^* = \overline{X}$ gäller (liksom tidigare konstaterats) att

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

och

$$V(\theta^*) = V(\overline{X}) = \sigma^2/n.$$

Maximum-likelihood-metoden: Likelihoodfunktion

Antag att X_i har täthetsfunktionen $f_X(x, \theta)$, θ okänd.
Vi ska nu studera en systematisk metod att hitta skattningar. Idén är att skatta θ så att utfallet blir så "troligt" som möjligt.

Definition

$$L(\theta) = f_{X_1}(x_1, \theta) \cdot \cdots \cdot f_{X_n}(x_n, \theta)$$

kallas Likelihood-funktionen.

Observera att likelihoodfunktionen betraktas som en funktion av θ, inte av $x_1, \ldots, x_n$.

KTH VETENSKAP OCH KONST
KTH Matematik

Maximum-likelihood-metoden

$$L(\theta) = f_{X_1}(x_1, \theta) \cdot \cdots \cdot f_{X_n}(x_n, \theta)$$

Definition

Det värde θ^*_{obs} för vilket $L(\theta)$ antar sitt största värde kallas ML-skattningen av θ.

För stora stickprov är denna skattning i allmänhet mycket bra.

KTH VETENSKAP OCH KONST
KTH Matematik

Exempel

X_i är $N(\theta, \sigma)$, dvs.

$$f(x,\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2}.$$

Vi kan t.ex. ha

$$\text{mätvärde} = \theta + \text{ slumpmässigt normalfördelat mätfel}$$

$$X_i = \theta + \sigma Z_i, \quad Z_i \in N(0,1), \quad i = 1, 2, \ldots, n$$

Vi observerar $x_1, \ldots, x_n$. Då fås

$$L(\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_1-\theta}{\sigma}\right)^2} \cdot \ldots \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_n-\theta}{\sigma}\right)^2} = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2}\sum_1^n \left(\frac{x_i-\theta}{\sigma}\right)^2}$$

KTH Matematik

Maximum-likelihood-metoden Exempel

Vi antar att σ är känt.

$$\ln L(\theta) = -\ln(\sigma^n (2\pi)^{n/2}) - \frac{1}{2\sigma^2} \sum_1^n (x_i - \theta)^2$$

$$\frac{d \ln L(\theta)}{d\theta} = \frac{1}{2\sigma^2} \sum_1^n 2(x_i - \theta).$$

$$\frac{d \ln L(\theta)}{d\theta} = 0$$

ger

$$\sum_1^n x_i = n\theta,$$

dvs.

$$\theta^*_{\text{obs}} = \overline{x}.$$

I detta fall är θ^*_{obs} *effektiv*!

Maximum-likelihood-metoden Kommentar

Varför kan vi använda den naturliga logaritmen av likelihoodfunktionen, $\ln L(\theta)$, för att hitta maximum ? Helt allmänt gäller: om $f(x) > 0$ för alla x i ett område där $f(x)$ är deriverbar,

$$\frac{d \ln f(x)}{dx} = \frac{\frac{d f(x)}{dx}}{f(x)}$$

dvs.

$$\frac{d \ln L(\theta)}{d\theta} = \frac{\frac{d L(\theta)}{d\theta}}{L(\theta)}$$

och således $\frac{d L(\theta)}{d\theta} = 0 \Leftrightarrow \frac{d \ln L(\theta)}{d\theta} = 0$. Vi arbetar givetvis med θ sådana att $L(\theta) > 0$. Dvs. $\ln L(\theta)$ maximeras av samma θ som maximerar $L(\theta)$ och omvänt.

Varför vill vi använda $\ln L(\theta)$? Ja, det är oftast enklast att derivera i summan

$$\ln L(\theta) = \ln f_{X_1}(x_1, \theta) + \cdots + \ln f_{X_n}(x_n, \theta)$$

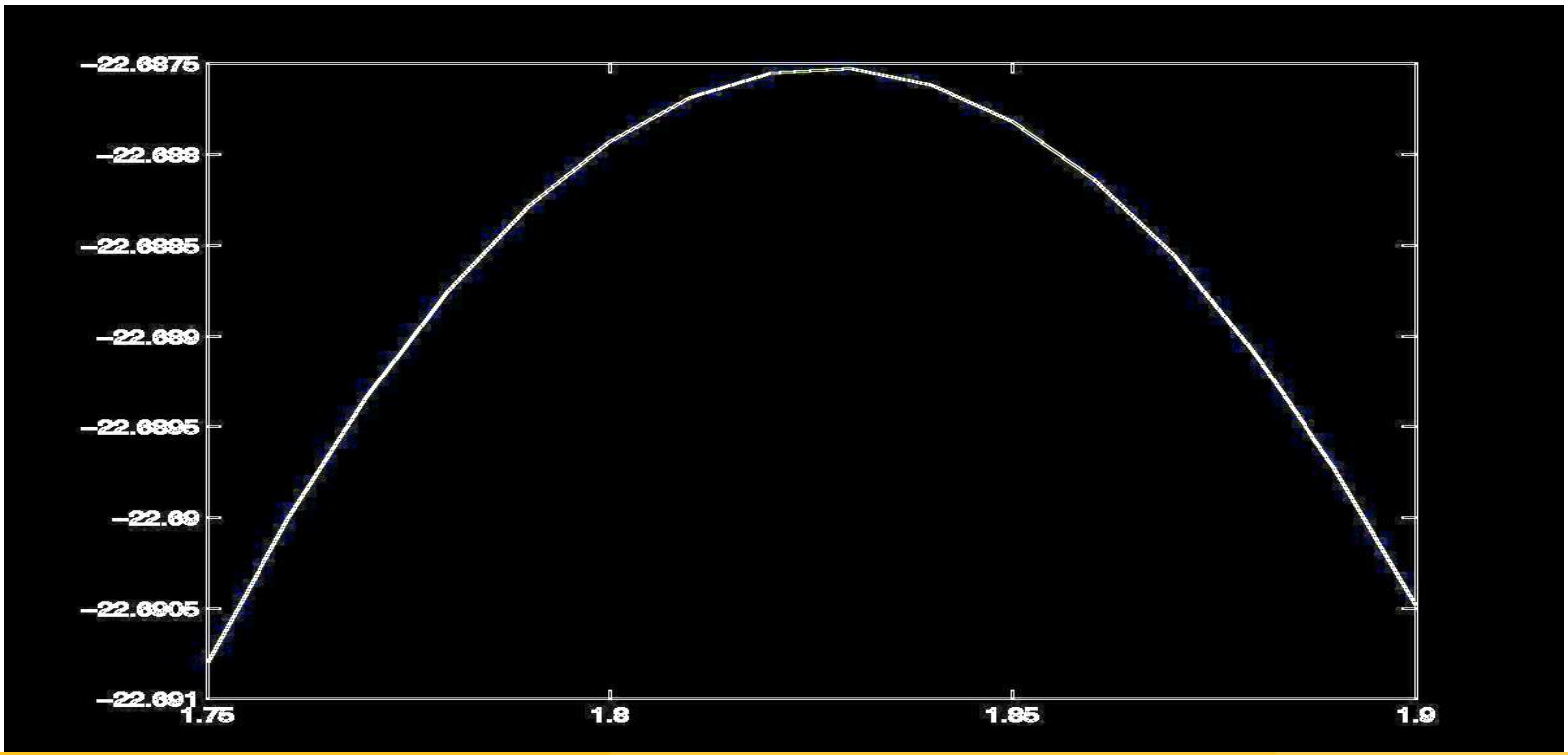
Maximum-likelihood-metoden Exempel: $\sigma = 3$

$$\ln L(\theta) = -\log(3^{10} \cdot (2\pi)^{10/2}) - \frac{1}{2 \cdot 3^2} \sum_{1}^{10} (x_i - \theta)^2$$

$$x_1 = 3.28 \quad x_2 = -0.21 \quad x_{31} = 3.69 \quad x_4 = -2.15 \quad x_5 = 3.38$$

$$x_6 = 3.89 \quad x_7 = 3.14 \quad x_8 = -1.04 \quad x_9 = 0.96 \quad x_{10} = 3.33$$

$$\overline{x} = 1.83$$



KTH Matematik

Skattning av σ^2

Maximum likelihood skattningen av σ^2 är

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

Men:

Sats

Stickprovsvariansen $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ *som skattning av* σ^2 *är*

1) *Väntevärdesriktig;*
2) *Konsistent;*
3) *Ej nödvändigtvis effektiv.*

1) används ofta som motivering för att man dividerar med $n-1$, men det är en dålig motivering, eftersom man oftast vill skatta σ. s som skattning av σ är dock ej väntevärdesriktig.

KTH Matematik

Maximum-likelihood-metoden för $\mathrm{Bin}(1000, p)$

Låt oss återkalla i minnet den inledande diskussionen om skattning av proportionen JA-sägare i en stor population. Vi har observerat $x = 350$ och tar detta som ett utfall av X med

$$X \in \mathrm{Bin}(1000, p)$$

Då är likelihoodfunktionen för p

$$L(p) = p_X(x) = \binom{1000}{x} p^x (1-p)^{1000-x}$$

Vi bildar *loglikelihoodfunktionen* $\ln L(p)$, dvs.

$$\ln L(p) = \ln \binom{1000}{x} + x \ln p + (1000 - x) \ln 1 - p$$

KTH Matematik

Maximum-likelihood-metoden för Bin$(1000, p)$

För att maximera $L(p)$ kan vi ekvivalent maximera

$$\ln L(p) = \ln \binom{1000}{x} + x \ln p + (1000 - x) \ln 1 - p$$

För detta deriverar vi $\ln L(p)$ m.a.p. p

$$\frac{d}{dp} \ln L(p) = x\frac{1}{p} - (1000 - x)\frac{1}{1-p}$$

och löser $\frac{d}{dp} \ln L(p) = 0$ m.a.p. p.

$$x\frac{1}{p} - (1000 - x)\frac{1}{1-p} = 0 \Leftrightarrow x\frac{1}{p} = (1000 - x)\frac{1}{1-p}$$

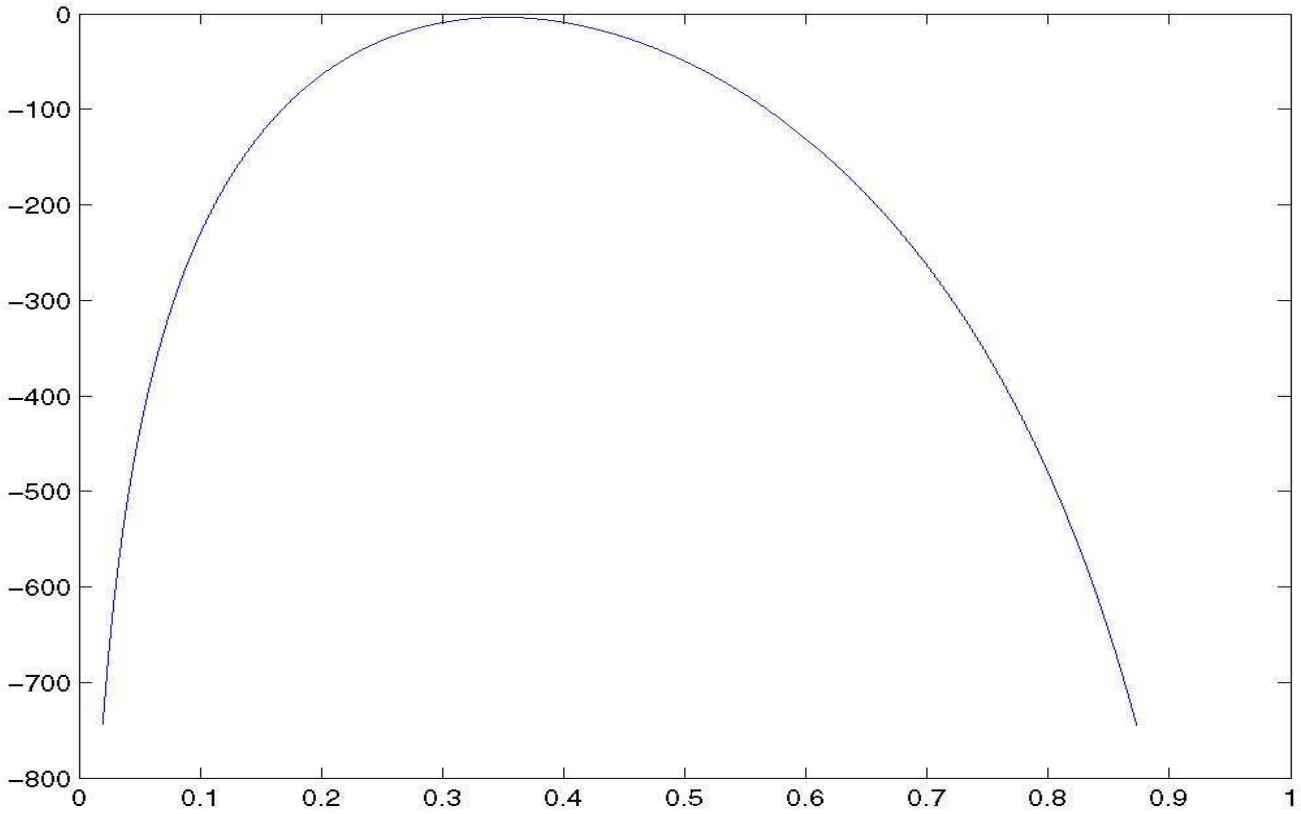
$$\Leftrightarrow (1-p)x = p(1000 - x) \Leftrightarrow x - px = p1000 - px$$

dvs. maximum likelihood - skattningen är

$$p^*_{obs} = \frac{x}{1000} = \frac{35}{1000} = 0.35.$$

KTH Matematik

Plot av loglikelihoodfunktionen $\ln L(p)$ för Bin$(1000, p)$ med $x = 350$



$$\ln L(p) = \ln \binom{1000}{350} + 350 \ln p + (1000 - 350) \ln 1 - p, \quad 0 < p < 1$$

Minsta-kvadrat-metoden

Låt $x_1, \ldots, x_n$ vara ett stickprov från en fördelning med $E(X) = \mu(\theta)$ där $\mu(\theta)$ är en *känd* funktion av en *okänd* parameter θ.
Sätt $Q(\theta) = \sum_{i=1}^{n}(x_i - \mu(\theta))^2$ och minimera $Q(\theta)$ map. θ. Lösningen θ^*_{obs} till detta problem kallas *MK-skattningen* av θ.
Ifall den inversa funktionen $\mu^{-1}(\theta)$ existerar ges MK-skattningen θ^*_{obs} av

$$\theta^*_{obs} = \mu^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \mu^{-1}\left(\overline{x}\right)$$

KTH Matematik

Minsta-kvadrat-metoden

$$Q(\theta) = \sum_{i=1}^{n}(x_i - \mu(\theta))^2$$

$$\frac{d}{d\theta}Q(\theta) = 2\sum_{i=1}^{n}(x_i - \mu(\theta)) \cdot \left(-\frac{d}{d\theta}\mu(\theta)\right)$$

$$\frac{d}{d\theta}Q(\theta) = 0 \Leftrightarrow -2\frac{d}{d\theta}\mu(\theta)\sum_{i=1}^{n}(x_i - \mu(\theta)) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n}(x_i - \mu(\theta)) = 0 \Leftrightarrow \sum_{i=1}^{n}x_i - n\mu(\theta) = 0$$

$$\mu(\theta) = \frac{1}{n}\sum_{i=1}^{n}x_i$$

Ifall den inversa funktionen $\mu^{-1}(\theta)$ existerar ges MK-skattningen θ^*_{obs} av

$$\theta^*_{obs} = \mu^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) = \mu^{-1}(\overline{x})$$

KTH Matematik

Minsta-kvadrat-metoden

$x_1, x_2, \ldots, x_n$ utfall av s.v. $X_1, X_2, \ldots, X_n$, $X_i \in U(0, \theta)$. Kursens formelsamling eller en enkel härledning ger

$$E(X_i) = \frac{\theta}{2}, \quad V(X_i) = \frac{\theta^2}{12}$$

dvs. vi tar $\mu(\theta) = \frac{\theta}{2}$ i $Q(\theta) = \sum_{i=1}^{n}(x_i - \mu(\theta))^2$, dvs.

$$Q(\theta) = \sum_{i=1}^{n}(x_i - \frac{\theta}{2})^2$$

MK-skattningen θ^* ges enligt ovan av

$$\theta^*_{obs} = \mu^{-1}(\overline{x}) = 2\overline{x}$$

KTH Matematik

$$\theta^*_{obs} = \mu^{-1}(\overline{x}) = 2\overline{x}$$

Denna punktskattning är väntevärdesriktig:

$$E(\theta^*) = 2E(\overline{X}) = 2\frac{\theta}{2} = \theta$$

och konsistent (Tjebysjovs olikhet)

$$V(\theta^*) = V(2\overline{X}) = 4\frac{1}{n^2} \cdot n\frac{\theta^2}{12}$$

$$= \frac{\theta^2}{3n}.$$

KTH Matematik

Maximum likelihood : samma exempel

$x_1, x_2, \ldots, x_n$ utfall av s.v. $X_1, X_2, \ldots, X_n$, $X_i \in U(0, \theta)$. Vad är maximum likelihood skattningen av θ?

$$f_X(x) = \begin{cases} \frac{1}{\theta} & \text{för } 0 \leq x \leq \theta \\ 0 & \text{annars.} \end{cases}$$

$$L(\theta) = f_{X_1}(x_1, \theta) \cdot \cdots \cdot f_{X_n}(x_n, \theta) =$$

och eftersom $x \geq \theta \Rightarrow f_X(x) = 0$

$$= \begin{cases} \frac{1}{\theta^n} & \text{för } \theta \geq \max(x_1, x_2, \ldots, x_n) \\ 0 & \text{annars.} \end{cases}$$

$L(\theta)$ är en avtagande funktion av θ och detta ger (utan derivering)

$$\theta_{obs}^{ML} = \max(x_1, x_2, \ldots, x_n).$$

KTH VETENSKAP OCH KONST
KTH Matematik

Korrigerad maximum likelihood i $U(0,\theta)$

$X_1, X_2, \ldots, X_n$, $X_i \in U(0,\theta)$. Maximum likelihood skattningen av θ är

$$\theta_{obs}^{ML} = \max(x_1, x_2, \ldots, x_n).$$

Man kan checka (Blom et.al. sid. 258) att den korrigerade skattningen

$$\frac{n+1}{n}\theta_{obs}^{ML}$$

är väntevärdesriktig, konsistent samt effektivare än minst-kvadrat skattningen.

German tank problem (1)

During World War II, the Germans tried to make a complete tabulation of how much they were producing, but reports from individual factories were often late and not always reliable. British and U.S. statisticians working for military intelligence were keenly interested in estimating German war production (especially the production of Mark V German tanks (Panther)), too, but they could hardly ask the German factories to send them reports. Instead, they based their estimates on the manufacturing serial numbers of captured equipment (specifically the tank gearboxes). These numbers were consecutive and didn't vary - because that was a rational system in terms of maintenance and spare parts. These serial numbers provided a sample that was very small, but reliable.

The challenge is to choose a good estimator for the total number of tanks.

KTH Matematik

German tank problem (2)

Detta är även känt som 'the serial number problem'. Vi kan renodla det enligt följande:

Någon har levererat oss heltalen $x_1, x_2, \ldots, x_n$, som är slumpmässiga stickprov från mängden av heltal

$$\{1, 2, 3, \ldots, N\}$$

där *N är okänt*.

Uppgiften (uppgiften för statistisk inferens) är att skatta N på basis av $x_1, x_2, \ldots, x_n$. (forts.)

KTH VETENSKAP OCH KONST
KTH Matematik

German tank problem (3)

$\theta =$ total number of tanks (true population value)

$n =$ number of tanks captured

$m =$ largest serial number of the captured tanks

The following *estimator* (=*stickprovsvariabel* i vår terminologi) was invented and used by statisticians in the the WW II allied military intelligence[1] :

$$\theta^* = [(n+1)/n]m \leftrightarrow \theta^* = m + (m/n)$$

which is interpreted as adding the average size of the gap to the highest serial number.

[1] Economic Warfare Division of the American Embassy in London

German tank problem (4)

Betrakta ett fiktivt exempel[2]. Antag att vi har $n = 15$ och att serienumren är

$$5 \quad 10 \quad 15 \quad 20 \quad 25 \quad 30 \quad 35 \quad 40 \quad 45 \quad 50 \quad 55 \quad 60 \quad 65 \quad 70 \quad 75$$

så att $m = 75$. Då ger formeln ovan *punktskattningen*

$$\theta^*_{\text{obs}} = 75 + (75/15) = 80$$

Vi kommer nedan att tolka stickprovsvariabeln $\theta^* = m + (m/n)$ som en för väntevärdesriktighet korrigerad maximum-likelihood-skattning av θ i $U(0, \theta)$ med n observationer !

[2] `http://en.wikipedia.org/wiki/German_tank_problem`

Korrigerad maximum likelihood i $U(0,\theta)$ & the German tank problem (5)

Den korrigerade skattningen

$$\frac{n+1}{n}\theta_{obs}^{ML} = \max(x_1, x_2, \ldots, x_n) + \frac{\max(x_1, x_2, \ldots, x_n)}{n}$$

är ingenting annat än den ovan framlagda skattningen i 'the German tank problem'. MEN: vi har då gett oss in på en kontinuerlig approximation av det ursprungliga problemet, dvs. vi har approximerat en likformig sannolikhetsfunktion på heltalen $\{1, 2, \ldots, \theta\}$ med en likformig täthet på $(0, \theta)$. Detta låter sig göras om θ är stort.
Exakt räknat kan man visa för det diskreta problemet att

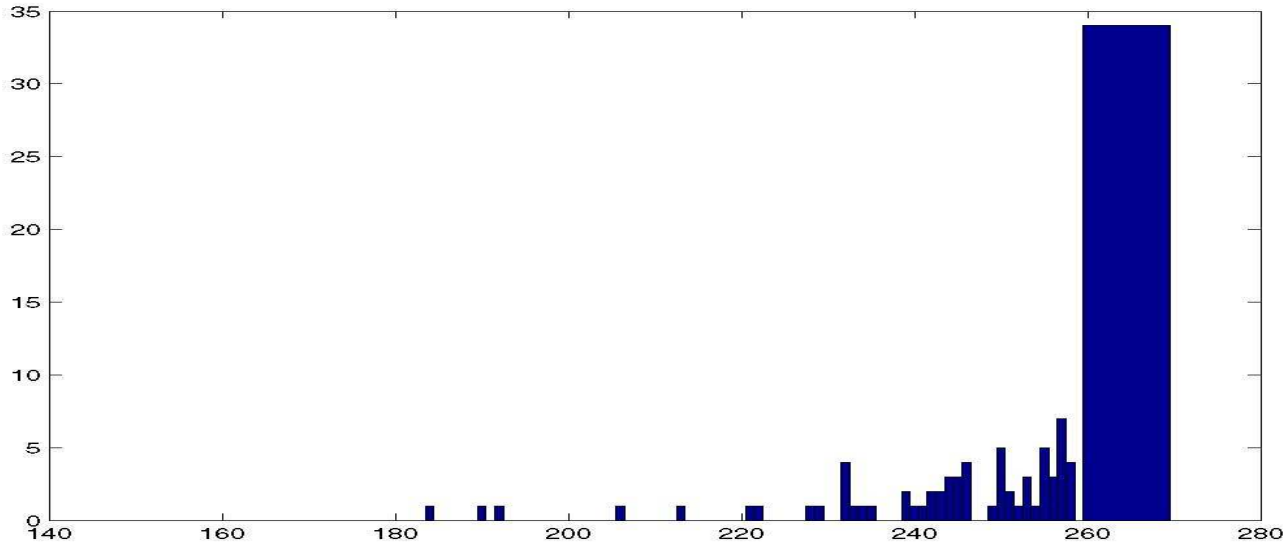
$$\theta_{obs}^{*} = \max(x_1, x_2, \ldots, x_n) + \frac{\max(x_1, x_2, \ldots, x_n)}{n} - 1$$

är en väntevärdesriktig och effektiv skattning.

KTH Matematik

Simulerad korrigerad maximum likelihood & the German tank problem (6)

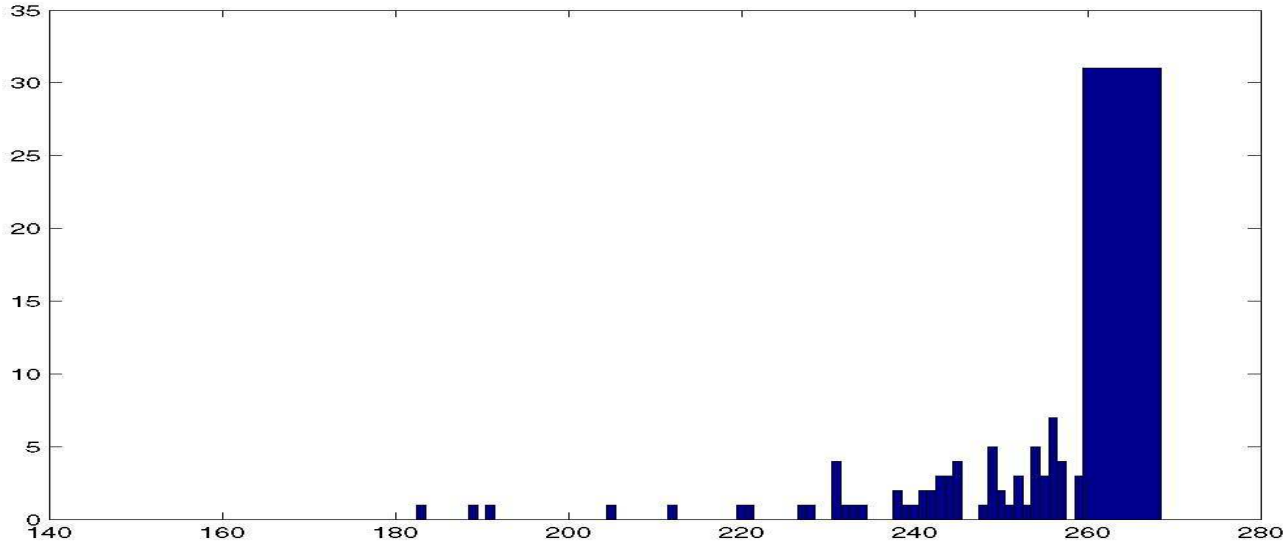
Vi drar hundra gånger tio värden $x_1, x_2, \ldots, x_{10}$ ur $1, 2, \ldots, \theta$ med $\theta = 245$ med hjälp av slumptalsgeneratorn *unidrnd* i MATLAB Statistics toolbox
Sedan tar vi maximum $y_i = \max(x_1, x_2, \ldots, x_{10})$, $i = 1, \ldots, 100$ för vart och ett av dessa hundra sampel och bildar $\theta^*_{obs\ \text{nr } i} = y_i + y_i/10$.
Histogrammet för de hundra θ^*_{obs} syns i bilden. $\overline{\theta^*_{obs}} = 250.64$.



KTH Matematik

Simulerad korrigerad maximum likelihood & the German tank problem (7)

Samma som ovan men med $\theta^*_{\underline{obs}\ \mathrm{nr}\ i} = y_i + y_i/10 - 1$. Histogrammet för de hundra θ^*_{obs} syns i bilden. $\overline{\theta^*_{obs}} = 249.65$.



KTH Matematik

German tank problem (8)

By using the formula $\theta^* = \theta^{ML} + (\theta^{ML}/n)$, statisticians reportedly estimated that the Germans produced 246 tanks per month between June 1940 and September 1942. At that time, standard intelligence estimates had believed the number was at around 1,400. After the war, the allies captured German production records of the Ministry, which was in charge of Germany's war production, showing that the true number of tanks produced in those three years was 245 per month, almost exactly what the statisticians had calculated, and less than one fifth of what standard intelligence had thought likely, and were more accurate and timely than Germany's own estimates.
Emboldened, the allies attacked the western front in 1944 and overcame the Panzers on their way to Berlin. And so it was that statisticians won the war - in their own estimation, at any rate.

KTH Matematik

German tank problem (9): lätt tillgängliga referenser

- G. Davies: How a statistical formula won the war.
  The Guardian, Thursday July 20 2006

  `http://www.guardian.co.uk/world/2006/jul/20/secondworldwar.tvandradio`

- Robert Matthews: Hidden truths.
  New Scientist 23 May 1998

  `http://www.newscientist.com/article/mg15821355.000-hidden-truths.html`

KTH Matematik

Exempel

Vi vill undersöka hur en termometer mäter temperatur. Vi prövar därför termometern i vätskor med olika temperaturer $x_1, \ldots, x_n$. Dessa temperaturer anser vi helt kända. Motsvarande mätvärden $y_1, \ldots, y_n$ antar vi är ungefär en linjär funktion av den verkliga temperaturen: $y_k \approx \alpha + \beta x_k$. Som vanligt uppfattas mätvärdena $y_1, \ldots, y_n$ som utfall av s.v. $Y_1, \ldots, Y_n$.

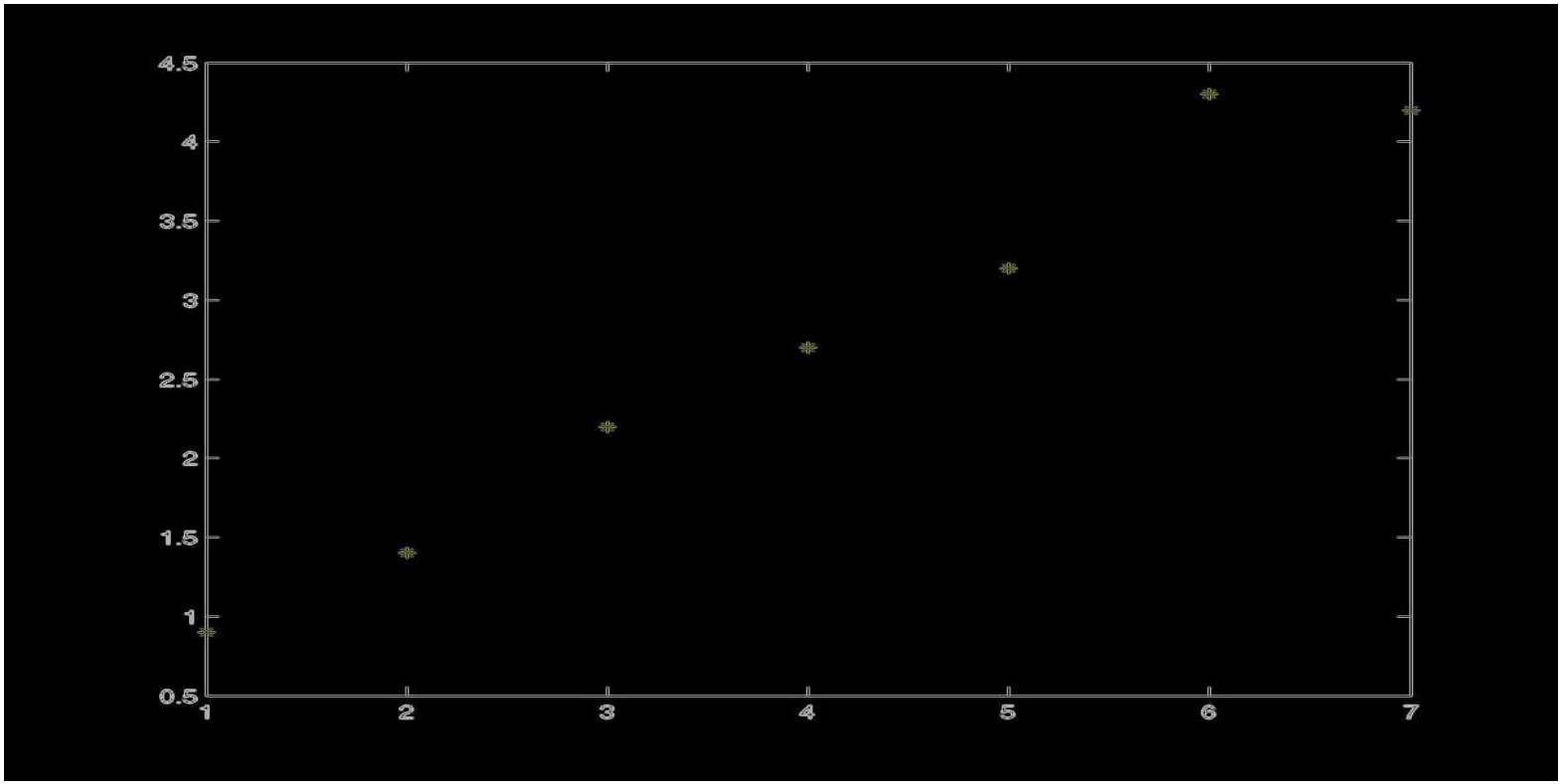
KTH Matematik

Minsta-kvadrat-metoden: Regressionsanalys

Ett fiktivt exempel

$$x = 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7$$

$$y = 0.9 \quad 1.4 \quad 2.2 \quad 2.7 \quad 3.2 \quad 4.3 \quad 4.2$$



KTH Matematik

Minsta-kvadrat-metoden: Enkel linjär regressionsanalys

Exemplen handlar om *enkel linjär regression*. En ofta använd modell för detta är följande: Det föreligger n par av värden

$$(x_1, y_1), \ldots, (x_n, y_n)$$

där $x_1, \ldots, x_n$ är givna storheter och $y_1, \ldots, y_n$ är observationer av oberoende s.v. $Y_1, \ldots, Y_n$, där $Y_i \in N(\mu_i, \sigma)$. Observera att σ förutsätts att *ej* bero av x, vilket ofta är det kritiska antagandet. Varje väntevärde μ_i är linjärt beroende av x_i, d.v.s.

$$\mu_i = \alpha + \beta x_i, \quad i = 1, \ldots, n.$$

Linjen

$$y = \alpha + \beta x \tag{1}$$

kallas den *teoretiska regressionslinjen*.

Minsta-kvadrat-metoden: Enkel linjär regressionsanalys

Koefficienten β är betydelsefull, ty den anger hur mycket väntevärdet ökar, då x ökas med en enhet. Om speciellt β skulle vara noll, är väntevärdet konstant, d.v.s beror inte av x. Med hjälp av regressionslinjen kan man för varje givet x bestämma tillhörande väntevärde.
Ju mindre σ är, desto mindre är på det hela taget dessa avstånd, d.v.s. desto bättre ansluter sig punkterna till linjen.

KTH Matematik

Minsta-kvadrat-metoden: Regressionsanalys

Vi skattar parametrarna α och β med Minsta-Kvadratmetoden, dvs. vi minimerar

$$Q(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

m.a.p. α och β. De värden α^*_{obs} och β^*_{obs} som ger minimum kallas *MK-skattningarna* av α och β.

KTH Matematik

Vi får nu:

$$\frac{\partial Q}{\partial \alpha} = -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = -2n(\overline{y} - \alpha - \beta\overline{x})$$

$$\frac{\partial Q}{\partial \beta} = -2\sum_{i=1}^{n} x_i(y_i - \alpha - \beta x_i).$$

Sätter vi derivatorna $= 0$, så fås av första ekvationen

$$\alpha = \overline{y} - \beta\overline{x},$$

vilket insatt i andra ekv. ger

$$0 = \sum_{i=1}^{n} x_i(y_i - \overline{y} - \beta(x_i - \overline{x})) = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y} - \beta(x_i - \overline{x})).$$

KTH VETENSKAP OCH KONST
KTH Matematik

Minsta-kvadrat-metoden: Regressionsanalys

Sätter vi ihop detta så får vi

$$\alpha^*_{\text{obs}} = \overline{y} - \beta^*_{\text{obs}}\overline{x} \quad \text{och} \quad \beta^*_{\text{obs}} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

KTH Matematik

Bilaga: räkneoperationer med summor

Definition

(1) $\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$.

KTH Matematik

Bilaga: räkneoperationer med summor

Sats

(2) $\underline{\sum_{i=1}^{n} a \cdot x_i = a \sum_{i=1}^{n} x_i.}$

Bevis. Definitionen (1) ger $\sum_{i=1}^{n} a \cdot x_i = ax_1 + ax_2 + \ldots + ax_n$
$= a\left(x_1 + x_2 + \ldots + x_n\right) = a \sum_{i=1}^{n} x_i.$
Exempel: $x_i = 1$, $i = 1. \ldots, n$
$\sum_{i=1}^{n} ax_i = a \sum_{i=1}^{n} 1 = a\left(1 + 1 + \ldots + 1\right) = a \cdot n.$

KTH VETENSKAP OCH KONST
KTH Matematik

Bilaga: räkneoperationer med summor

Sats

(3) $\underline{\sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.}$

Bevis. Definition (1) ger

$\sum_{i=1}^{n}(x_i + y_i) = (x_1 + y_1) + (x_2 + y_2) + \ldots + (x_n + y_n)$

$= x_1 + x_2 + \ldots + x_n + y_1 + y_2 + \ldots + y_n = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.$

KTH Matematik

Bilaga: räkneoperationer med summor

Sats

(4) $\underline{\sum_{i=1}^{n}(ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i}$

Bevis. Detta fås av (3) och (2).

Sats

(5) $\underline{\sum_{i=1}^{n}(x_i + y_i)^2 = \sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2.}$

Bevis. Använd $(x_i + y_i)^2 = x_i^2 + 2x_i y_i + y_i^2$ och (4) samt (2) med $a = 2$.

KTH Matematik