SF1901: Sannolikhetslära och statistik
Föreläsning 9.
Statistik: statistiska inferensproblem, maximum likelihood, minsta kvadrat

Jan Grandell & Timo Koski

15.02.2016

KTH Matematik

Sammanfattning

- Vi inleder med några exempel på statistiska *inferensproblem* (inferens= slutledning):

KTH Matematik

Sammanfattning

- Vi inleder med några exempel på statistiska *inferensproblem* (inferens= slutledning):
  A) opiniondundersökning

KTH Matematik

Sammanfattning

- Vi inleder med några exempel på statistiska *inferensproblem* (inferens= slutledning):
  A) opinionsundersökning
  B) Hur får man ett svar på en fråga som man inte ställt?

KTH Matematik

Sammanfattning

- Vi inleder med några exempel på statistiska *inferensproblem* (inferens= slutledning):
  - A) opinionsundersökning
  - B) Hur får man ett svar på en fråga som man inte ställt?

Bonus Enigma

KTH VETENSKAP OCH KONST
KTH Matematik

Sammanfattning

- Vi inleder med några exempel på statistiska *inferensproblem* (inferens= slutledning):
  - A) opinionsundersökning
  - B) Hur får man ett svar på en fråga som man inte ställt?

Bonus Enigma

  - C) serienummerproblemet även känt som 'the German tank problem'.

KTH Matematik

Sammanfattning

- Vi inleder med några exempel på statistiska *inferensproblem* (inferens= slutledning):
  - A) opinionsundersökning
  - B) Hur får man ett svar på en fråga som man inte ställt?

Bonus Enigma

  - C) serienummerproblemet även känt som 'the German tank problem'.
  - D) mätvärden med brus (ex. ljusets hastghet; Newcomb-Michelson)

KTH Matematik

Sammanfattning

- Vi inleder med några exempel på statistiska *inferensproblem* (inferens= slutledning):
  - A) opinionsundersökning
  - B) Hur får man ett svar på en fråga som man inte ställt?

  Bonus Enigma
  - C) serienummerproblemet även känt som 'the German tank problem'.
  - D) mätvärden med brus (ex. ljusets hastghet; Newcomb-Michelson)
- Sedan definierar vi begrepp som punktskattning och stickprovsvariabel, och inför krav på deras prestanda.

KTH Matematik

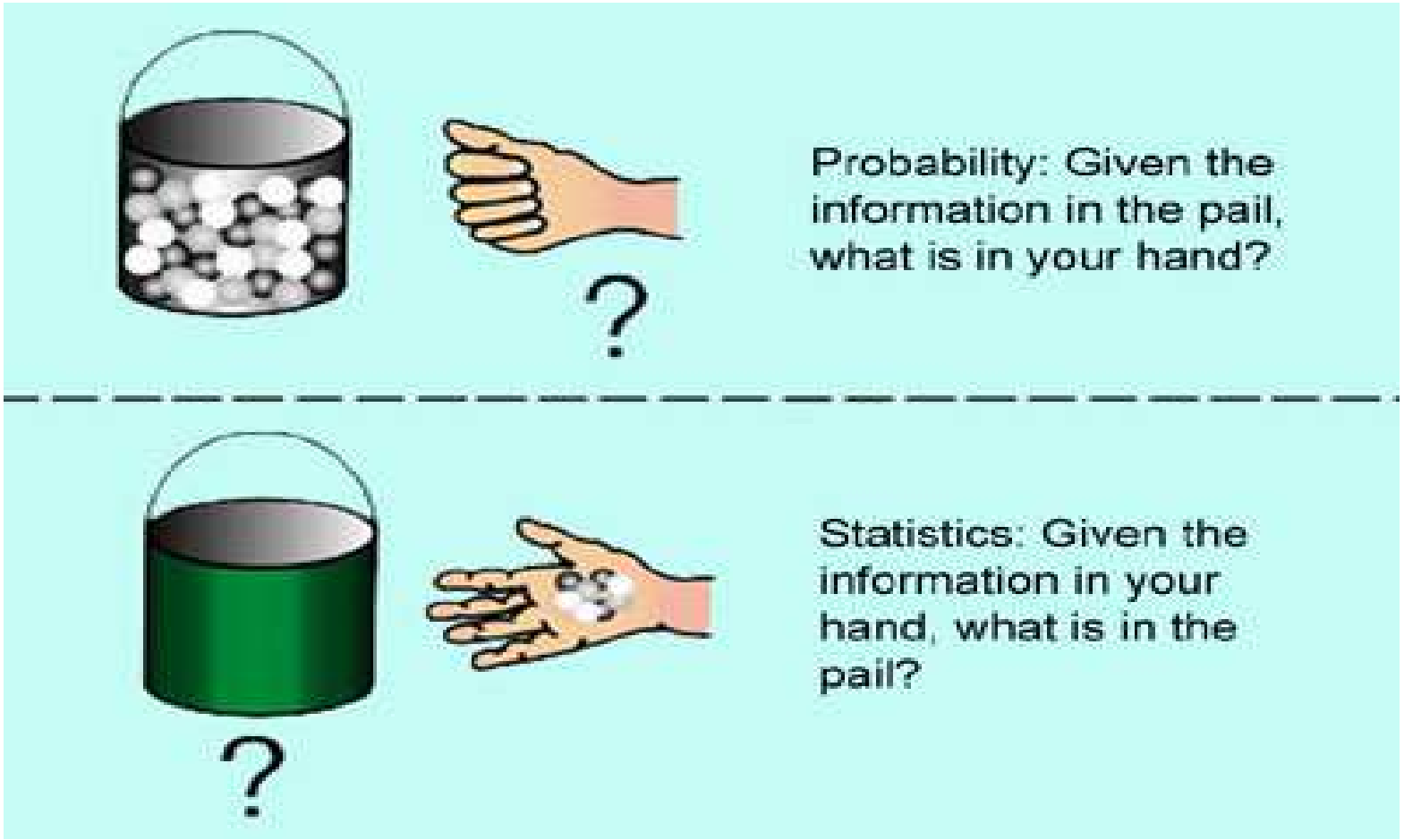
Sammanfattning

- Vi inleder med några exempel på statistiska *inferensproblem* (inferens= slutledning):
  - A) opinionsundersökning
  - B) Hur får man ett svar på en fråga som man inte ställt?

  Bonus Enigma
  - C) serienummerproblemet även känt som 'the German tank problem'.
  - D) mätvärden med brus (ex. ljusets hastghet; Newcomb-Michelson)
- Sedan definierar vi begrepp som punktskattning och stickprovsvariabel, och inför krav på deras prestanda.
- Vi inför två systematiska metoder för systematisk konstruktion av punktskattningar: *maximum-likelihood* och *minsta kvadrat* och tillämpar på exemplen A) -D).

KTH Matematik

Sann.stat.



KTH Matematik

Statistisk Inferens: Data $\rightarrow$ Kunskap. Ett citat:

Inference is the problem of turning data into knowledge, where knowledge often is expressed in terms of entities that are not present in the data per se but are present in models that one uses to interpret the data.

Committee on the Analysis of Massive Data: *Frontiers in Massive Data Analysis*.The National Academies Press, Washington D.C., 2013, sid.3.

KTH Matematik

A) Inledande om opinionsundersökning: punktskattning

Vi väljer 1000 personer ur en 'stor' population (t.ex. Sveriges befolkning). De ska svara 'JA' eller 'NEJ' till en fråga (t.ex. om medlemskap i militäralliansen NATO).
Ponera nu att

$$x = 350 \quad \text{svarat JA}$$

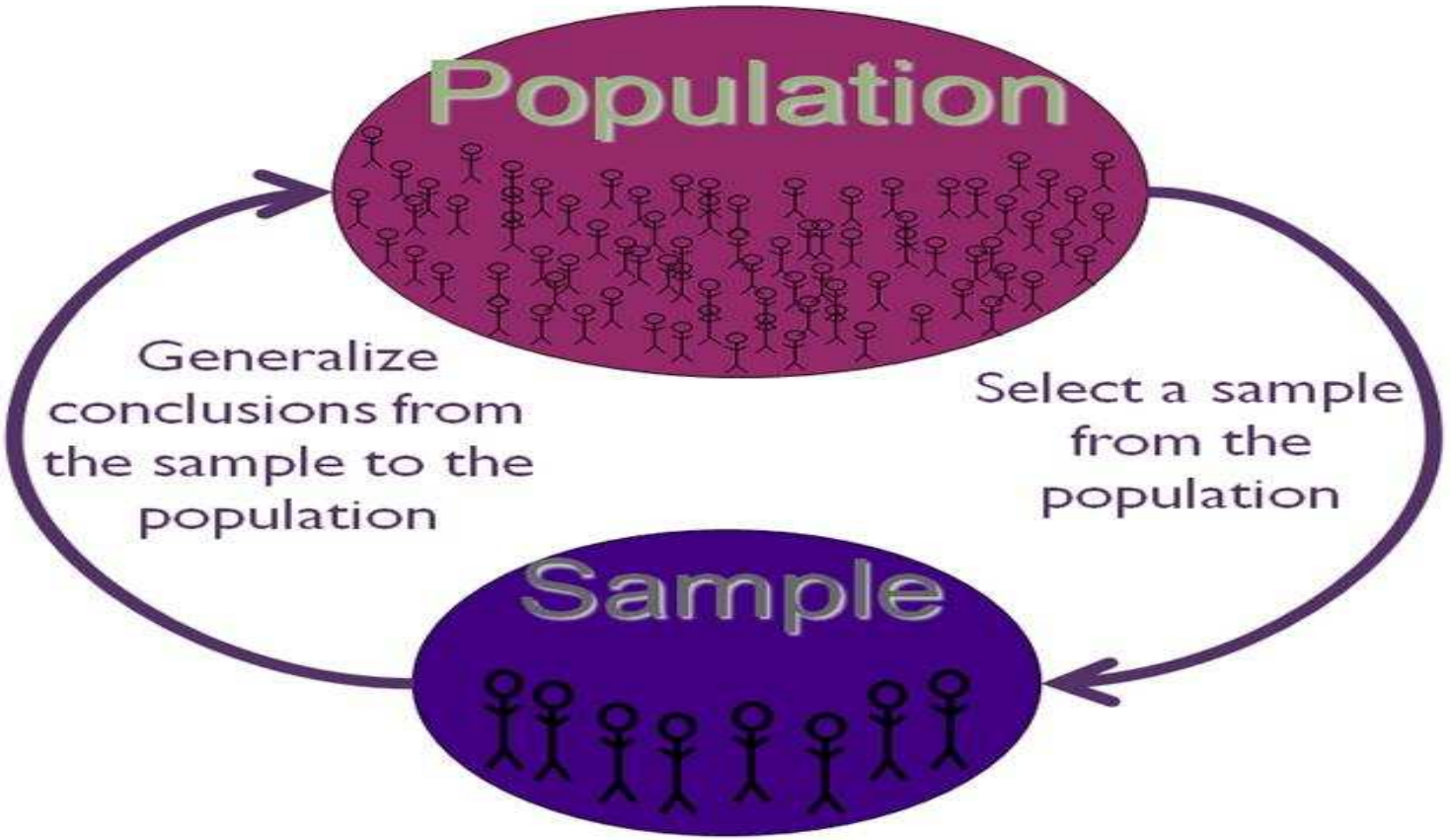
Vi vill veta proportionen JA-sägare i HELA populationen, dvs. punktskatta proportionen JA-sägare i HELA populationen.
En skattning av av p:

$$p^*_{obs} = \frac{x}{1000} \Rightarrow p^*_{obs} = 35\%$$

Hur stor är osäkerheten i denna skattning?

KTH Matematik

Sann.stat.



KTH Matematik

A) Inledande diskussion: modell för skattning

$p^*_{obs} = \frac{x}{1000} = 0.35$. Hur stor är osäkerheten i denna skattning? Vi använder oss av en sannolikhetsmodell.

URNMODELL: p = andelen JA-svar i populationen. X = antalet JA-svar när vi frågar 1000 individer. N = populationens storlek, t.ex. $N = 7{,}3$ miljoner svenskar över 15 år (enligt SCB var Sveriges befolkning 9 182 927 invånare på nyårsafton 2007)[1].

Vi har att

$$X \in \text{Hyp}(N, 1000, p)$$

[1] http://www.scb.se/templates/pressinfo___227515.asp

A) Inledande diskussion: modell för skattning

Vi återkallar i minnet

Sats

Om X är $\mathrm{Hyp}(N, n, p)$*-fördelad med* $n/N \leq 0.1$ *så är X approximativt* $\mathrm{Bin}(n, p)$*-fördelad.*

Då kan vi utan stor förlust ta

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

Vi uppfattar $x = 350$ som ett utfall av X. Vi kallar

$$p^* = \frac{X}{1000}$$

en *stickprovsvariabel*. $p^*_{obs} = \frac{x}{1000} = 0.35$ är ett utfall av p^*

KTH VETENSKAP OCH KONST
KTH Matematik

A) Inledande diskussion: modell för skattning

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

Vi vet att $E(X) = 1000 \cdot p$, $V(X) = 1000 \cdot p \cdot (1-p)$. Detta ger

$$E(p^*) = E\left(\frac{X}{1000}\right) = \frac{1}{1000} E(X) = \frac{1}{1000} \cdot 1000 \cdot p = p.$$

$$V(p^*) = V\left(\frac{X}{1000}\right) = \frac{1}{1000^2} V(X) = \frac{1}{1000^2} \cdot 1000 \cdot p(1-p) = \frac{p(1-p)}{1000}.$$

Svårigheten är att de ovanstående uttrycken beror på den okända (parameter) p som vi vill skatta!

KTH Matematik

A) Inledande diskussion: medelfelet för skattning

$$E\left(p^*\right) = p, \quad V\left(p^*\right) = \frac{p(1-p)}{1000}.$$

$$D\left(p^*\right) = \sqrt{V\left(p^*\right)} = \sqrt{\frac{p(1-p)}{1000}}$$

Vår skattning av osäkerhet i skattningen är då *medelfelet*

$$d\left(p^*\right) = \sqrt{\frac{p^*_{obs}(1-p^*_{obs})}{1000}} = \sqrt{\frac{0.35(1-0.35)}{1000}} \approx 0.015 = 1.5\%$$

KTH Matematik

B) Hur får man ett svar på en fråga som man inte ställt?

Det föregående bygger på antagandet att intervjuobjekten svarar ärligt. Ibland frågan är s.a.s. 'känslig', så att svaret 'Ja' kan vara stigmatiserande. Således kan man inte förvänta sig enbart ärliga svar.

KTH Matematik

B) Hur får man ett svar på en fråga som man inte ställt?

Man vill undersöka andelen p av personer i en stor population med en viss egenskap som ger svaret 'Ja' till en känslig frga. (Exempel: "Har Du under det senaste året använt narkotika?").

En randomiserad undersökning:

För att få ökad personlig sekretess (och mer korrekt undersökningsresultat) lät man de tillfrågade dra ett kort. Med sannolikheten 2/3 skulle de dra ett kort av typ I som säger de skall svara (ärligt: Ja/Nej) på den känsliga frågan och med sannolikheten 1/3 skulle de dra kort av typ II som säger att de skall svara (ärligt: Ja/Nej) på en irrelevant fråga (T.ex. är den sista siffran i ditt personnummer ett jämnt tal?). Denna lottning fick den utfrågade själv utföra utan att visa resultatet till andra.

Antag att du genomförde denna undersökning och fick 40% Ja-svar. Vad är p =andelen individer som t.ex. under det senaste året använt narkotika ?

KTH Matematik

B) Hur får man ett svar på en fråga som man inte ställt?

Låt $\lambda =$ sannolikheten för svaret 'Ja'. $p = P\left(\text{svar:Ja} \mid \text{kort typ I}\right)$. Vi antar att det är lika sannolikt med Ja och Nej till den irrelevanta frågan. **Lagen om total sannnolikhet** ger:

$$\lambda = P\left(\text{svar:Ja} \mid \text{kort typ I}\right) P\left(\text{kort I}\right) + P\left(\text{svar:Ja} \mid \text{kort typ II}\right) P\left(\text{kort II}\right)$$

$$= p \cdot \frac{2}{3} + \frac{1}{2}\frac{1}{3}$$

D.v.s

$$p = \frac{3\lambda}{2} - \frac{3}{2} \cdot \frac{1}{6} = \frac{3\lambda}{2} - \frac{1}{4}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

B) Hur får man ett svar på en fråga som man inte ställt?

$$p^* = \frac{3\lambda^*}{2} - \frac{1}{4}.$$

Med $\lambda^*_{\text{obs}} = 0.40$ fås

$$p^*_{\text{obs}} = 1.5 \cdot 0.4 - 0.25 = 0.35.$$

KTH Matematik

B) λ^*

$\lambda^* = \frac{X}{n}$, där X är antalet Ja-svar i den randomiserade undersökningen och n är antalet deltagare i samma undersökning. $X \in \text{Bin}(n, \lambda)$. Då fås

$$V(p^*) = V(1.5 \cdot \lambda^* - 0.25) =$$

$$= 1.5^2 V(\lambda^*) = 2.25 \cdot \frac{1}{n^2} V(X) =$$

$$= 2.25 \cdot \frac{1}{n} \lambda \cdot (1 - \lambda).$$

$$= 2.25 \cdot \frac{1}{n} \left(\frac{2p}{3} + \frac{1}{6} \right) \cdot \left(\frac{5}{6} - \frac{2p}{3} \right).$$

KTH Matematik

B) En svårighet med detta:

$$p = \frac{3\lambda}{2} - \frac{1}{4}.$$

Med $\lambda^*_{\text{obs}} = 0.15$ fås

$$p^*_{\text{obs}} = 1.5 \cdot 0.15 - 0.25 = -0.025$$

Ett uppenbarligen orimligt resultat !

KTH VETENSKAP OCH KONST
KTH Matematik

C) 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

C) 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

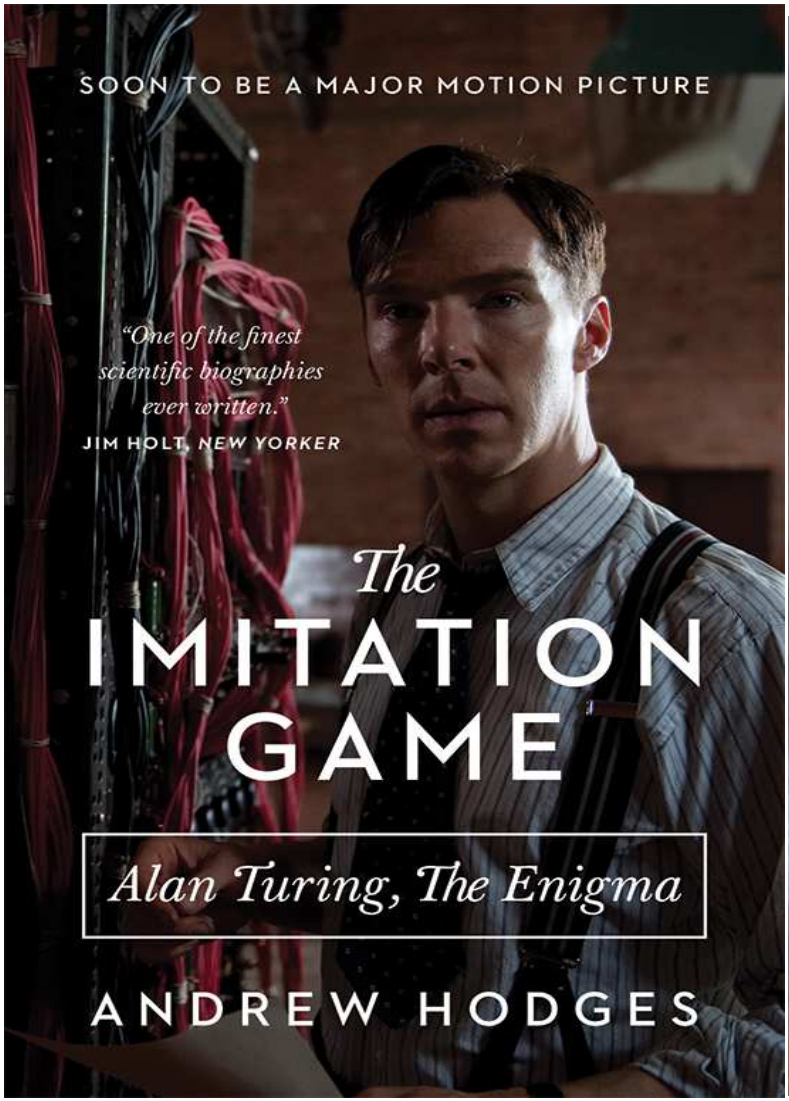
$$\{1, 2, 3, \ldots, N\}$$

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

Enigma


SOON TO BE A MAJOR MOTION PICTURE
"One of the finest scientific biographies ever written."
JIM HOLT, NEW YORKER
The
IMITATION
GAME
Alan Turing, The Enigma
ANDREW HODGES

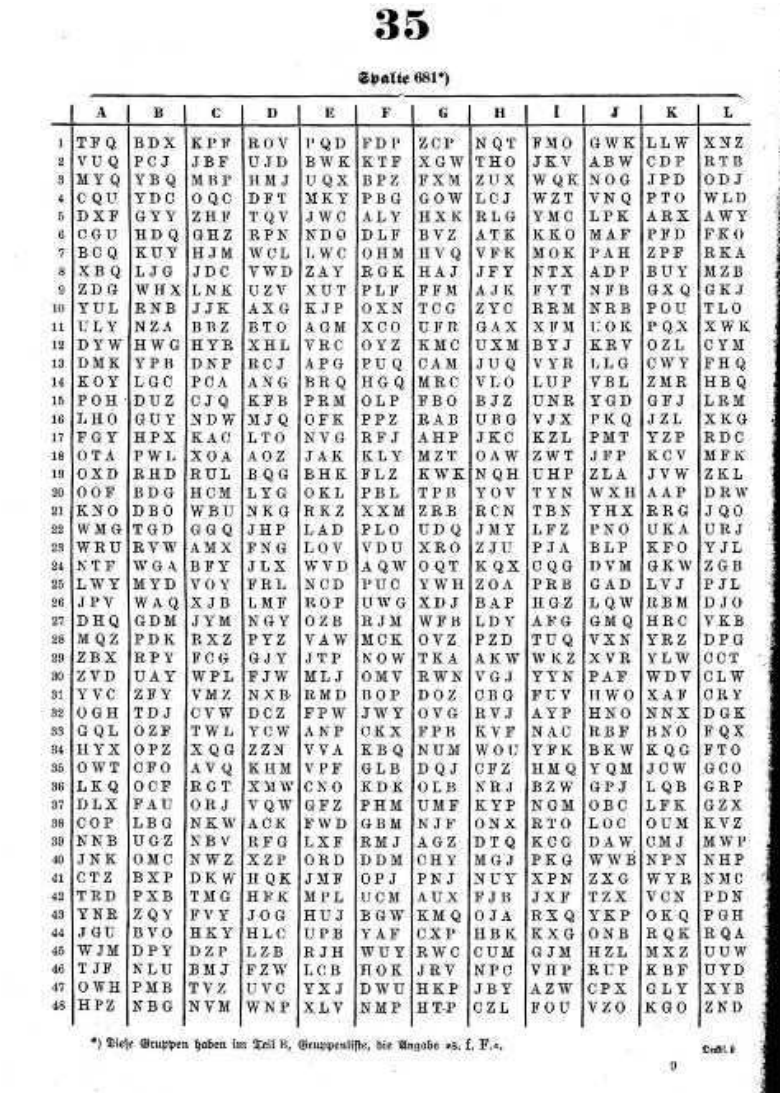
Enigma & Good-Turings punktskattning

Enigma var en portabel krypteringsmaskin som användes av Nazitysklands väpnade styrkor både före och under andra världskriget. Britterna hade lyckats i att kapa åt sig Kenngruppenbuch (KG), en sekretessbelagd bok som innehöll i en slumpmässig ordning varje ord med tre bokstäver. Ett sådant ord, dagligen plockat ur en sida i KG enligt vissa instruktioner, ingick i krypteringsnyckeln för krigsmarinens meddelanden per enigma.

KTH Matematik

Enigma

De omständliga detaljerna i denna procedur genomgås i
`http://users.telenet.be/d.rijmenants/en/enigmaproc.htm`



35

Spalte 681*)

	A	B	C	D	E	F	G	H	I	J	K	L
1	TFQ	BDX	KPF	ROV	PQD	FDP	ZCP	NQT	FMO	GWK	LLW	XNZ
2	VUQ	PCJ	JBF	UJD	BWK	KTF	XGW	THO	JKV	ABW	CDP	RTB
3	MYQ	YBQ	MBP	HMJ	UQX	BPZ	FXM	ZUX	WQK	NOG	JPD	ODJ
4	CQU	YDC	OQC	DFT	MKY	PBG	GOW	LCJ	WZT	VNQ	PTO	WLD
5	DXF	GYY	ZHF	TQV	JWC	ALY	HXK	RLG	YMC	LPK	ARX	AWY
6	CGU	HDQ	GHZ	RPN	NDO	DLF	BVZ	ATK	KKO	MAF	PFD	FKO
7	BCQ	KUY	HJM	WCL	LWC	OHM	HVQ	VFK	MOK	PAH	ZPF	RKA
8	XBQ	LJG	JDC	VWD	ZAY	RGK	HAJ	JFY	NTX	ADP	BUY	MZB
9	ZDG	WHX	LNK	UZV	XUT	PLF	FFM	AJK	FYT	NFB	GXQ	GKJ
10	YUL	RNB	JJK	AXG	KJP	OXN	TCG	ZYC	RRM	NRB	POU	TLO
11	ULY	NZA	BBZ	BTO	AGM	XCO	UFR	GAX	XFM	UOK	PQX	XWK
12	DYW	HWG	HYR	XHL	VRC	OYZ	KMC	UXM	BYJ	KRV	OZL	CYM
13	DMK	YPB	DNP	RCJ	APG	PUQ	CAM	JUQ	VYR	LLG	CWY	FHQ
14	KOY	LGC	PCA	ANG	BRQ	HGQ	MRC	VLO	LUP	VBL	ZMR	HBQ
15	POH	DUZ	CJQ	KFB	PRM	OLP	FBO	BJZ	UNR	YGD	GFJ	LRM
16	LHO	GUY	NDW	MJQ	OFK	PPZ	RAB	UBG	VJX	PKQ	JZL	XKG
17	FGY	HPX	KAC	LTO	NVG	RFJ	AHP	JKC	KZL	PMT	YZP	RDC
18	OTA	PWL	XOA	AOZ	JAK	KLY	MZT	OAW	ZWT	JFP	KCV	MFK
19	OXD	RHD	RUL	BQG	BHK	FLZ	KWK	NQH	UHP	ZLA	JVW	ZKL
20	OOF	BDG	HCM	LYG	OKL	PBL	TPB	YOV	TYN	WXH	AAP	DRW
21	KNO	DBO	WBU	NKG	RKZ	XXM	ZRB	RCN	TBN	YHX	RRG	JQO
22	WMG	TGD	GGQ	JHP	LAD	PLO	UDQ	JMY	LFZ	PNO	UKA	URJ
23	WRU	RVW	AMX	FNG	LOV	VDU	XRO	ZJU	PJA	BLP	KFO	YJL
24	NTF	WGA	BFY	JLX	WVD	AQW	OQT	KQX	CQG	DVM	GKW	ZGB
25	LWY	MYD	VOY	FRL	NCD	PUC	YWH	ZOA	PRB	GAD	LVJ	PJL
26	JPV	WAQ	XJB	LMF	ROP	UWG	XDJ	BAP	HGZ	LQW	RBM	DJO
27	DHQ	GDM	JYM	NGY	OZB	RJM	WFB	LDY	AFG	GMQ	HRC	VKB
28	MQZ	PDK	RXZ	PYZ	VAW	MCK	OVZ	PZD	TUQ	VXN	YRZ	DPG
29	ZBX	RPY	FCG	GJY	JTP	NOW	TKA	AKW	WKZ	XVR	YLW	CCT
30	ZVD	UAY	WPL	FJW	MLJ	OMV	RWN	VGJ	YYN	PAF	WDV	CLW
31	YVC	ZFY	VMZ	NXB	RMD	BQP	DOZ	CBQ	FUV	HWO	XAF	CRY
32	OGH	TDJ	CVW	DCZ	FPW	JWY	OVG	RVJ	AYP	HNO	NNX	DGK
33	GQL	OZF	TWL	YCW	ANP	CKX	FPB	KVF	NAC	RBF	BNO	FQX
34	HYX	OPZ	XQG	ZZN	VVA	KBQ	NUM	WOU	YFK	BKW	KQG	FTO
35	OWT	CFO	AVQ	KHM	VPF	GLB	DQJ	CFZ	HMQ	YQM	JCW	GCO
36	LKQ	OCF	RGT	XMW	CNO	KDK	OLB	NRJ	BZW	GPJ	LQB	GRP
37	DLX	FAU	ORJ	VQW	GFZ	PHM	UMF	KYP	NGM	OBC	LFK	GZX
38	COP	LBG	NKW	ACK	FWD	GBM	NJF	ONX	RTO	LOC	OUM	KVZ
39	NNB	UGZ	NBV	RFG	LXF	RMJ	AGZ	DTQ	KCG	DAW	CMJ	MWP
40	JNK	OMC	NWZ	XZP	ORD	DDM	CHY	MGJ	PKG	WWB	NPN	NHP
41	CTZ	BXP	DKW	HQK	JMF	QPJ	PNJ	NUY	XPN	ZXG	WYR	NMC
42	TRD	PXB	TMG	HFK	MPL	UCM	AUX	FJB	JXF	TZX	VCN	PDN
43	YNR	ZQY	FVY	JOG	HUJ	BGW	KMQ	OJA	RXQ	YKP	OKQ	PGH
44	JGU	BVO	HKY	HLC	UPB	YAF	CXP	HBK	KXG	ONB	RQK	RQA
45	WJM	DPY	DZP	LZB	RJH	WUY	RWC	CUM	GJM	HZL	MXZ	UUW
46	TJF	NLU	BMJ	FZW	LCB	HOK	JRV	NPC	VHP	RUP	KBF	UYD
47	OWH	PMB	TVZ	UVC	YXJ	DWU	HKP	JBY	AZW	CPX	GLY	XYB
48	HPZ	NBG	NVM	WNP	XLV	NMP	HTP	CZL	FOU	VZO	KGO	ZND

*) Diese Gruppen haben im Teil B, Gruppenliste, die Angabe »s. f. F.«.

Enigma

Alan Turing & gruppen hade stickprov av ord med tre bokstäver, som använts av ubåtskaptenerna i meddelanden som tidigare forcerats. Turing antog att dessa fartygschefer ofta plockade samma ord som tidigare.

Bundesarchiv, Bild 101II-MW-4222-02A
Foto: Dietrich | März 1941

http://commons.wikimedia.org/wiki/File:Bundesarch

KTH VETENSKAP OCH KONST
KTH Matematik

Enigma: Good-Turings punktskattning

Turing ville skatta sannolikheten att den nya nyckelns ord är ett av dem, som redan använts, och fördelningen för alla observerade ord.

Relativ frekvens visade sig inte fungera, Turing byggde sin skattning på storleksfördelningen *(size distribution)* $\{N_r\}$, *där* N_r *är antalet ord som förekommer* r *gånger bland de forcerade orden (med tre bokstäver). Då är totalantalet ord* $N = \sum_{r=1} rN_r$.

Enigma: Good-Turings punktskattning

p_r(= sannolikheten för ett ord som setts r gånger) skattas av

$$\widehat{p}_r = \frac{(r+1) \cdot N_{r+1}}{N \cdot N_r}.$$

I.J. Good förfinade detaljerna, d.v.s algoritmen genom en funktion $S(N_r)$. Denna skattning få sägas vara rätt så icke-intuitiv. Det skulle ta oss för långt att gå vidare med detta.
Med dessa skattningar kunde Turing designa *statistiska test* för att reducera antalet initiella inställningar som checkades av den av honom konstruerade maskinen (datorn?, kallad bombe) för att hitta Enigmas dagliga inställning.

KTH Matematik

D) Mätning av ljusets hastighet

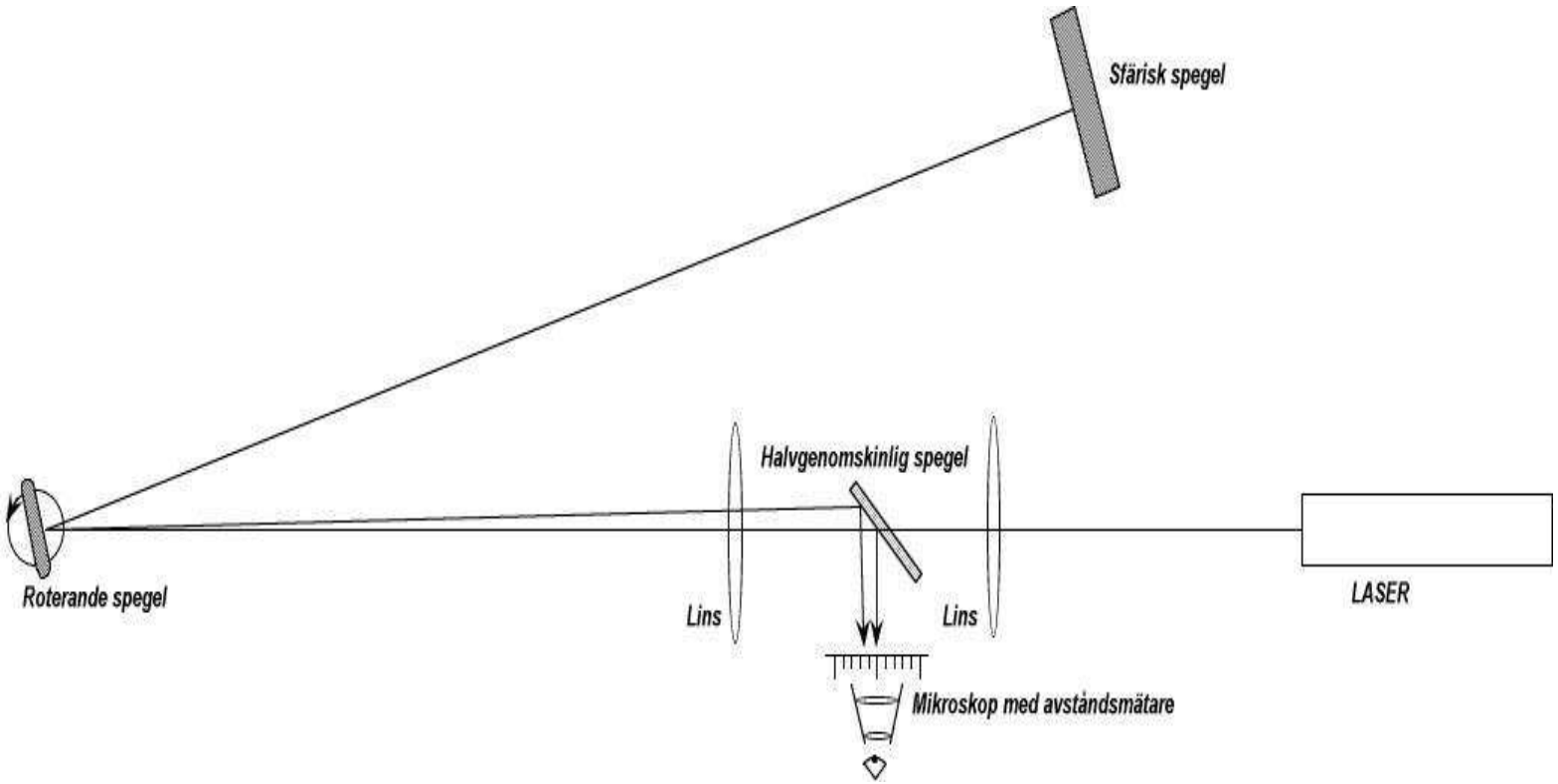
Exempel

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

D) Mätning av ljusets hastighet: utrustningen (lite moderniserad)



D) Mätning av ljusets hastighet

Exempel

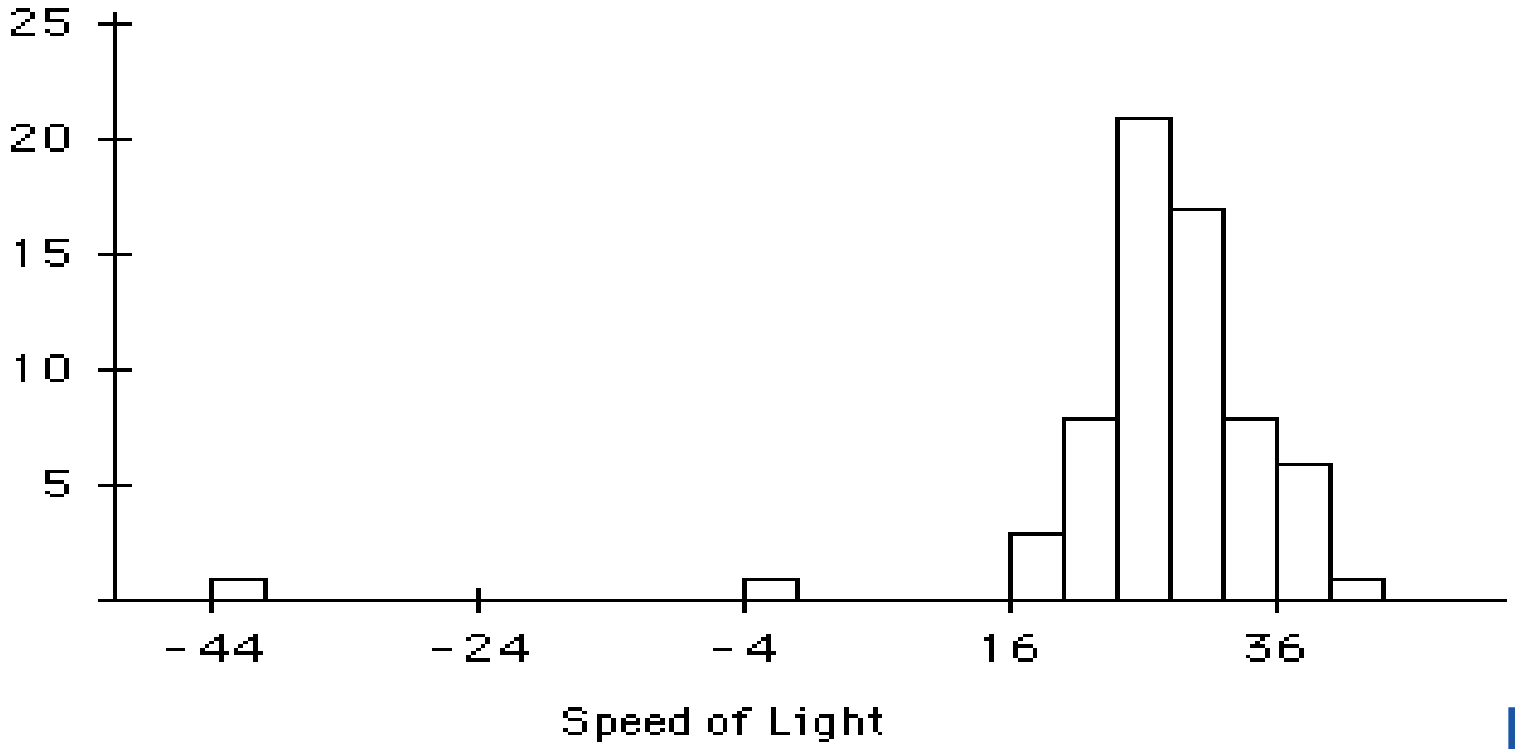
De exakta tekniska detaljerna för hur dessa mätningar gjordes redovisas i R.J. MacKay & R.W. Oldford : Statistical Method and the Speed of Light. *Statistical Science*, vol. 15, 2000, pp. 254−278.
W.E. Cater & M.S. Carter: The Newcomb-Michelson Velocity of Light Experiments. *Eos, Transactions, American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

KTH Matematik

D) 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:



D) 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

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

KTH VETENSKAP OCH KONST
KTH Matematik

Allmänt

Vi har en uppsättning data

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

som ses som utfall av s.v.

$$X_1, X_2, \ldots, X_n.$$

Dessa variabler antages vara oberoende och likafördelade och deras gemensamma fördelning beror av en okänd parameter θ, t.ex. $N(\theta, \sigma)$, $\mathrm{Po}(\theta)$, $N(\theta_1, \theta_2)$, osv.

KTH VETENSKAP OCH KONST
KTH Matematik

Allmänt

- En punktskattning θ^*_{obs} av θ är en funktion $\theta^*(x_1, \ldots, x_n)$

KTH Matematik

Allmänt

- En punktskattning θ^*_{obs} av θ är en funktion $\theta^*(x_1, \ldots, x_n)$
- och motsvarande stickprovsvariabel θ^* är $\theta^*(X_1, \ldots, X_n)$

KTH Matematik

Allmänt

- En punktskattning θ^*_{obs} av θ är en funktion $\theta^*(x_1, \ldots, x_n)$
- och motsvarande stickprovsvariabel θ^* är $\theta^*(X_1, \ldots, X_n)$
- Exemplet A) ovan:

$$p^*_{obs} = \frac{350}{1000} = 0.35$$

är en punktskattning.

$$p^* = \frac{X}{1000}, \quad X \in \text{Bin}(1000, p)$$

är motsvarande stickprovsvariabel.

KTH VETENSKAP OCH KONST
KTH Matematik

Vad menas med en bra skattning?

Vi ger tre kriterier:

Definition

1) *En punktskattning* θ^*_{obs} *av* θ *är väntevärdesriktig om* $E(\theta^*(X_1, \ldots, X_n)) = \theta$.

Vad menas med en bra skattning?

Vi ger tre kriterier:

Definition

1) *En punktskattning* θ^*_{obs} *av* θ *är väntevärdesriktig om* $E(\theta^*(X_1, \ldots, X_n)) = \theta$.
2) *En punktskattning* θ^*_{obs} *av* θ *är konsistent om* $P(|\theta^*(X_1, \ldots, X_n) - \theta| > \varepsilon) \to 0$ *då* $n \to \infty$.

KTH VETENSKAP OCH KONST
KTH Matematik

Vad menas med en bra skattning?

Vi ger tre kriterier:

Definition

1) *En punktskattning* θ^*_{obs} *av* θ *är väntevärdesriktig om* $E(\theta^*(X_1, \ldots, X_n)) = \theta$.
2) *En punktskattning* θ^*_{obs} *av* θ *är konsistent om* $P(|\theta^*(X_1, \ldots, X_n) - \theta| > \varepsilon) \to 0$ *då* $n \to \infty$.
3) *Om* θ^*_{obs} *och* θ^{**}_{obs} *är väntevärdesriktiga skattningar av* θ *så säger man att* θ^*_{obs} *är effektivare än* θ^{**}_{obs} *om* $V(\theta^*(X_1, \ldots, X_n)) < V(\theta^{**}(X_1, \ldots, X_n))$.

KTH VETENSKAP OCH KONST
KTH Matematik

A) *P*roportionen JA-sägare & Good-Turing

- $$X \in \text{Bin}(1000, p)$$
$$p^* = \frac{X}{1000}, \quad E\left(p^*\right) = p.$$
Således är p^*_{obs} väntevärdesriktig.
- Good-Turings punktskattning är inte väntevärdesriktig, men felet, även kallat bias, är litet.

KTH VETENSKAP OCH KONST
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.)

□

KTH Matematik

Skattning av σ^2

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: 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

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

Låt oss återkalla i minnet den inledande diskussionen A) 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.

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

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.

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

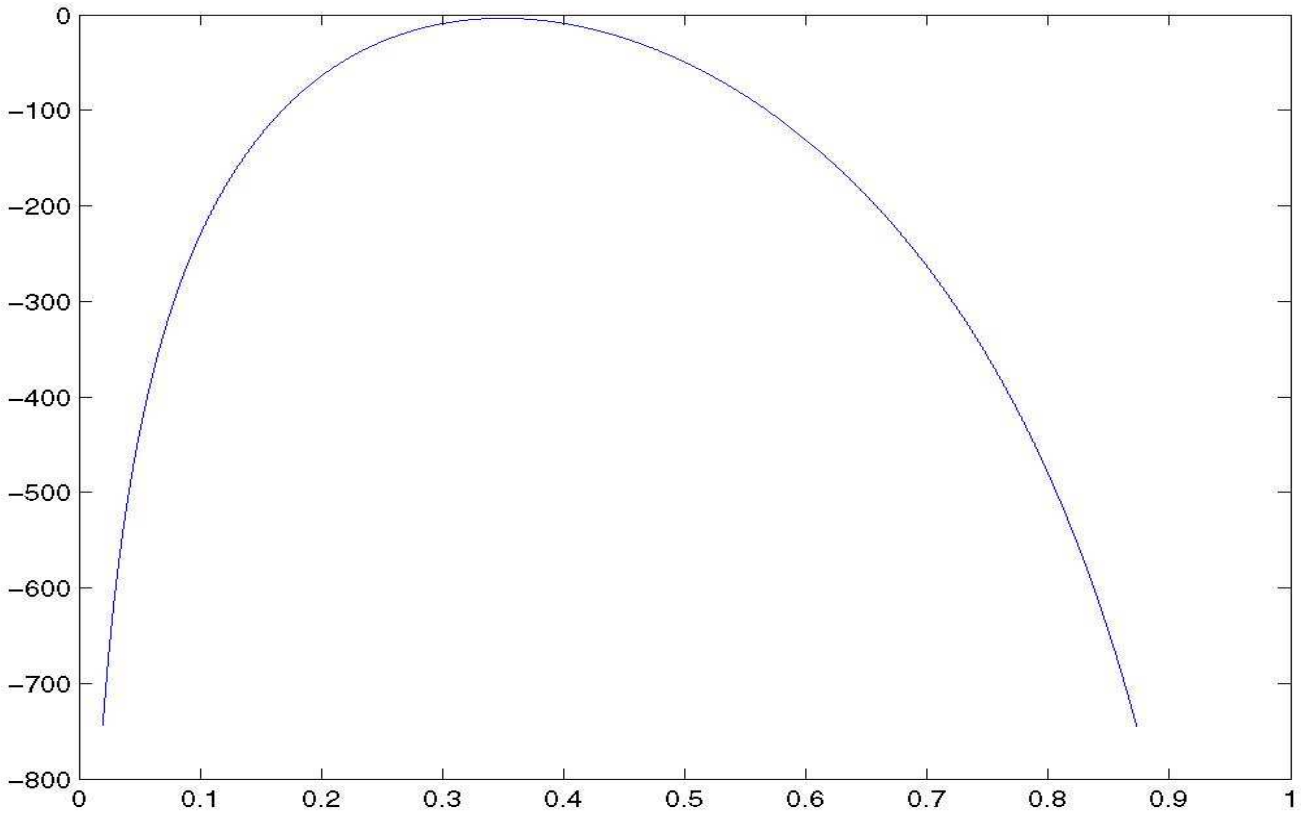
$$\Leftrightarrow (1-p)x = p(1000 - x) \Leftrightarrow x - px = p1000 - px$$

dvs. maximum likelihood - skattningen är

$$p^*_{obs} = \frac{x}{1000} = \frac{350}{1000} = 0.35.$$

KTH VETENSKAP OCH KONST
KTH Matematik

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



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

B) Ny metod: Hur får man ett svar på en fråga som man inte ställt?

Man vill undersöka andelen p av personer som svarar ja p en känslig frga. (Exempel: "Har Du under det senaste året använt narkotika?").
En ny randomiserad undersökning:
Vi har en kortlek med n=jämnt tal kort. De tillfrågade drar var sitt kort **utan återläggning**. $n/2$ kort ger instruktionen att de skall svara (ärligt: Ja/Nej) på den känsliga frågan och $n/2$ kort säger att de skall blankt svara Ja. Alla drar sitt kort och visar inte resultatet till andra.
$X =$ antalet Ja-svar. Vi vet att det finns minst $n/2$ Ja-svar. Då är $Y = X - \frac{n}{2} \in \mathrm{Bin}(\frac{n}{2}, p)$, med den okända proportionen p som förut. Då är

$$p^* = \frac{Y}{\frac{n}{2}}$$

enligt ovan maximum likelihood skattning av p, väntevärdesriktig och effektiv. Dessutom gäller $0 \leq p^* \leq 1$!

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 D))

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*!

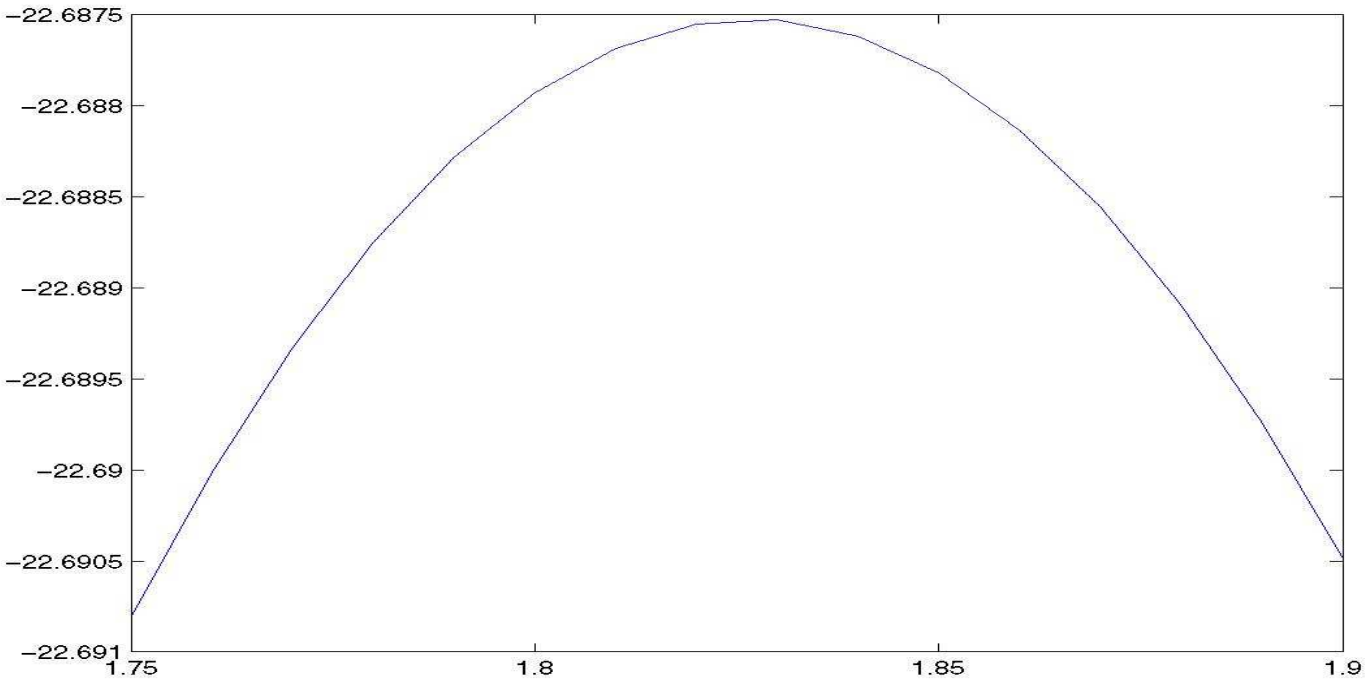
Maximum-likelihood-metoden Exempel: $\sigma = 3$

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

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

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

$$\overline{x} = 1.83$$



KTH Matematik

Minsta-kvadrat-metoden

Om vi inte känner fördelningen helt kan inte ML-metoden användas. Ibland ger den även upphov till svåra beräkningsproblem. Man kan då gå tillväga på följande sätt:
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.

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

Minsta-kvadrat-metoden

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

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

C) 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[2] :

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

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

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

C) German tank problem (4)

Betrakta ett fiktivt exempel[3]. 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 !

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

KTH Matematik

Maximum likelihood för θ i $U(0,\theta)$

$x_1, x_2, \ldots, x_n$ 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.

C) 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 C) '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

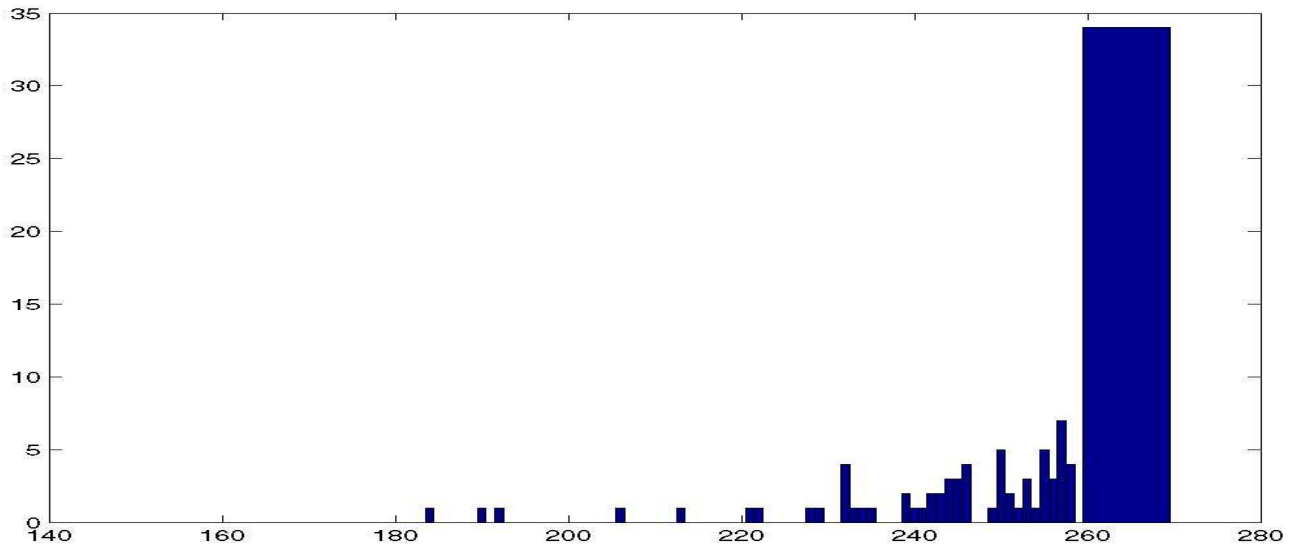
$$\theta_{obs}^{*} = \max(x_1, x_2, \ldots, x_n) + \frac{\max(x_1, x_2, \ldots, x_n)}{n} - 1$$

är en väntevärdesriktig och effektiv skattning.

KTH VETENSKAP OCH KONST
KTH Matematik

C) 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 $\theta^*_{obs\ \underline{\text{nr } i}} = y_i + y_i/10$.
Histogrammet för de hundra θ^*_{obs} syns i bilden. $\overline{\theta^*_{obs}} = 250.64$.



KTH Matematik

C) German tank problem (7)

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. Even Alan Turing is acknowledged as one made victory possible.

KTH Matematik

C) German tank problem (8): 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

Bilaga: räkneoperationer med summor

Definition

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

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

Bevis. Detta fås av (3) och (2).

Sats

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

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 VETENSKAP OCH KONST
KTH Matematik