Statistik

Språkligt och historiskt betyder statistik ungefär ”sifferkunskap om staten”

En Statistisk undersökning består av fyra delar:

- **Planering (kap 15)**
- **Datainsamling**
- **Bearbetning**
    - **Beskrivande statistik (kap 10)**
    - **Statistisk analys (kap 11-14)**
- **Presentation**

Statistiska undersökningar förekommer inom nästan alla vetenskaper.
Tex naturvetenskap, teknik och samhällsvetenskap.

Det finns tre sorters lögner: lögn, förbannad lögn och statistik

Fyra syften med statistik

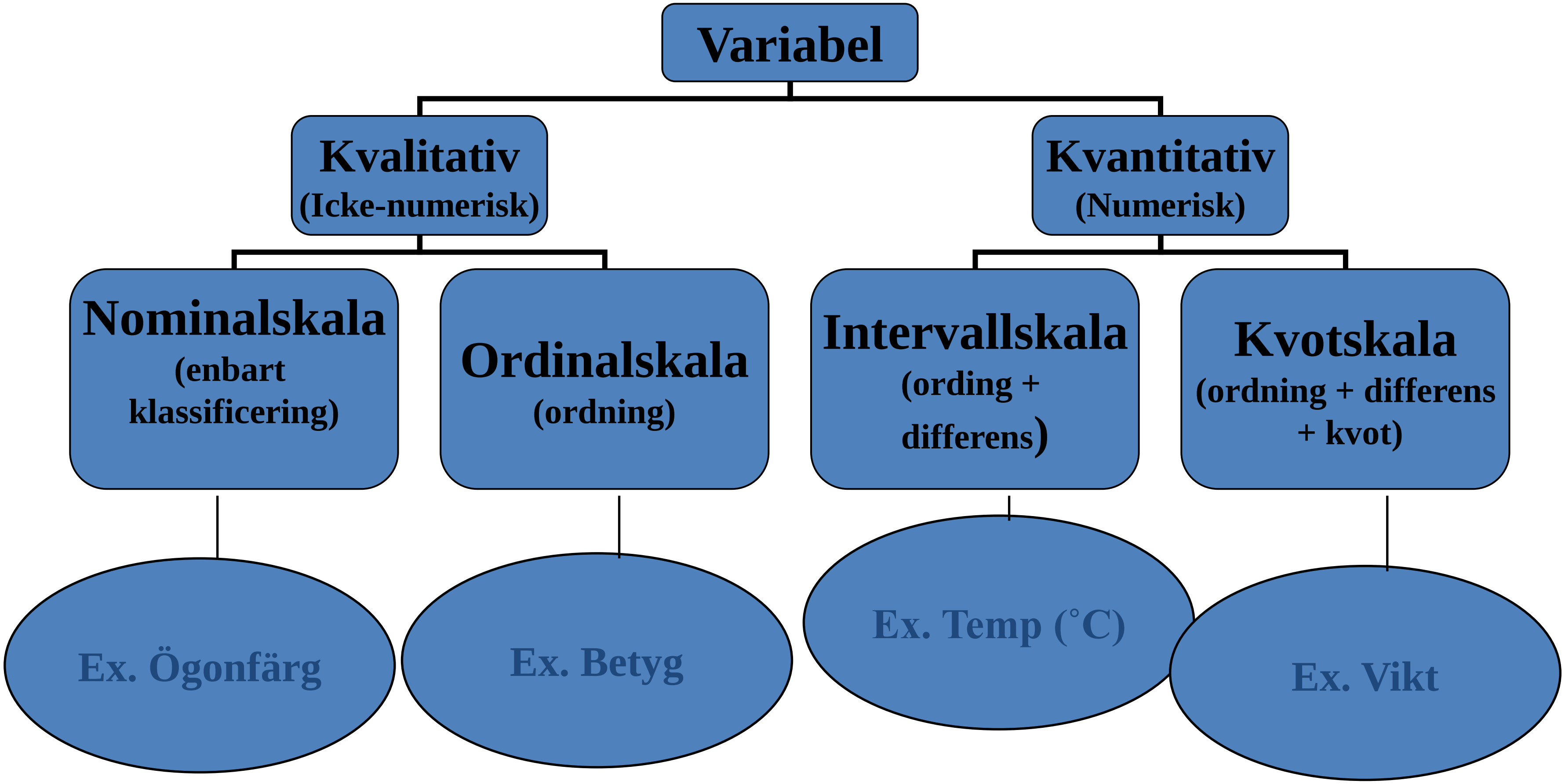
- **Deskriptiv**
  – informera, kartlägga
- **Hypotesprövande**
  – Verifiera eller förkasta ett antagande (hypotes)
- **Utredande**
  – kausala samband, orsakssammanhang
- **Prognosticerande**
  – vad händer i framtiden?, vad händer om vi gör så här?

”alltför många försöker spå om framtiden, utan att ens kunna historien”

Några vanliga begrepp

- **Element (individ) -** de som information söks om
  - Mängden av dessa element kallas ofta *population*.
  - Populationen kan vara ändlig eller oändlig.
- **Total undersökning –** hela populationen studeras
- **Stickprovsundersökning –** del av populationen studeras
- **Stickprov - en del av populationen**
- **Validitet -** mäter vi det vi avser att mäta?
- **Reliabilitet -** är de mätningar vi gör tillförlitliga?
- **Kategori variabel, (Kvalitativ, icke-numerisk variabel)**
  färg, ogift, god mat, attityd, servicegrad, kundnöjdhet (kan ges siffervärden)
- **Kvantitativ variabel (numerisk)**
  - *Kontinuerlig* - alla (oändligt antal) värden inom ett intervall
  - *Diskret* - vissa (ändligt antal) värden inom ett intervall

Något om mätskalor



Ett exempel på stickprovsundersökning

(icke-experimentell undersökning)

En firma tillverkar mätapparatur till vilken det behövs elektroniska kretskort. Det blir dyrt om man får in för många defekta kretskort i produktionen varför underleverantören lovar högst 0,5% defekta kretskort.

Kretskorten ligger i förpackningar med 10 000 i varje. Man undersöker 200 på måfå utvalda kort ur varje förpackning. I en sändning på 80 förpackningar fick man följande resultat.

(Detta är ett exempel på *diskret* variation)

Ett exempel på stickprovsundersökning
(icke-experimentell undersökning)

Antal defekta kretskort bland 200 utvalda i 80 förpackningar.

Grunddata

1	2	1	0	3	3	4	2	4	7	4	1	1	0	0	1	1	0	0	4
1	2	2	2	2	2	2	5	2	2	3	5	1	2	2	4	0	1	4	1
5	1	3	3	1	1	3	2	1	4	2	1	3	2	1	1	4	3	1	3
5	2	2	4	1	3	3	0	0	1	2	4	3	2	0	3	1	1	1	1

Vad kan man säga om *p,* andel defekta kretskort i sändningen?
Frågan kan preciseras på 3 olika sätt:

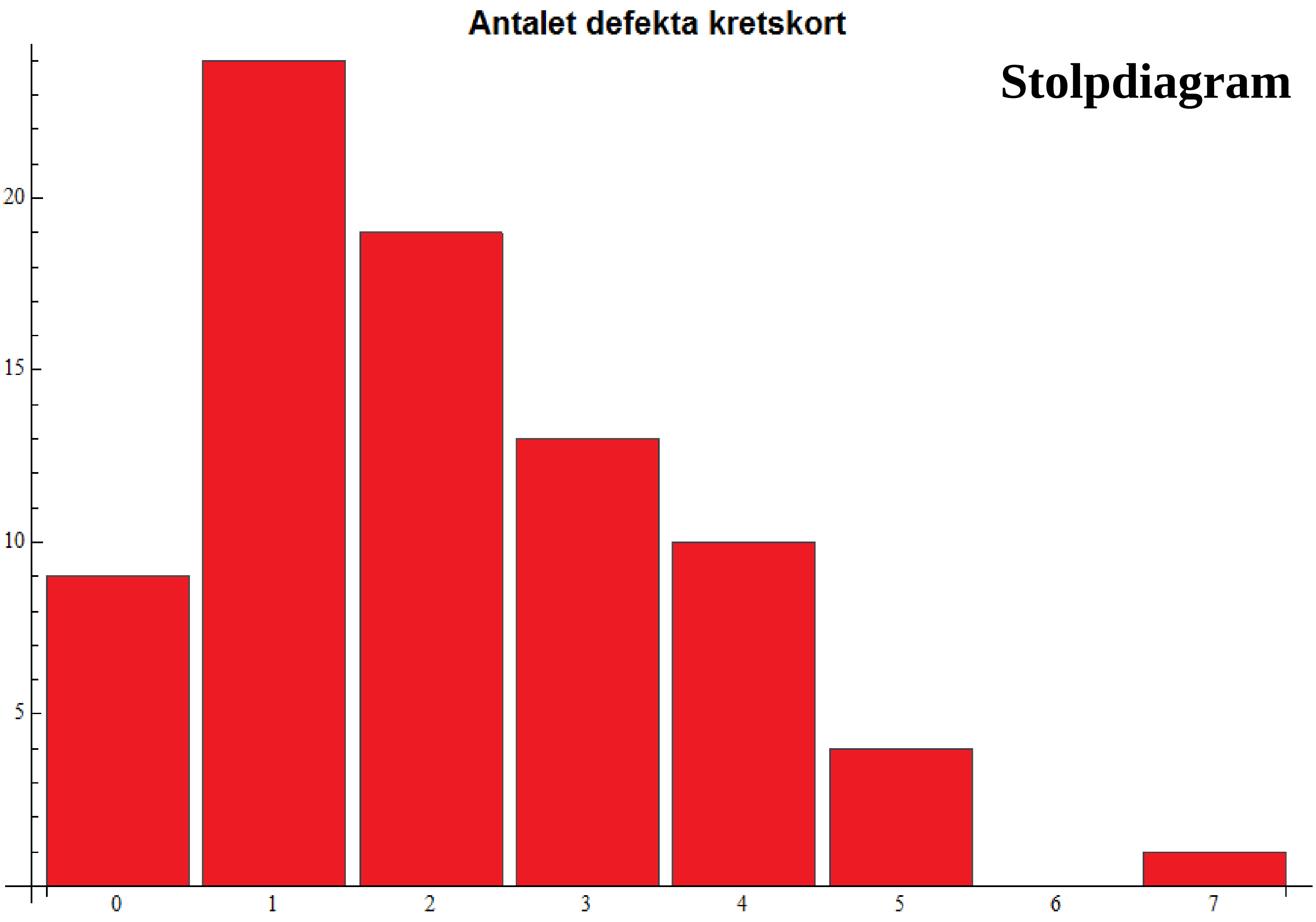
- ***Punktskattningsproblem* – hur skattar man *p*?**
- ***Intervallskattningsproblem* – hur anger man ett intervall som med given säkerhet innehåller *p?***
- ***Hypotesprövningsproblem* – hur prövar man hypoteser rörande *p?***

Ett exempel på stickprovsundersökning (icke-experimentell undersökning)

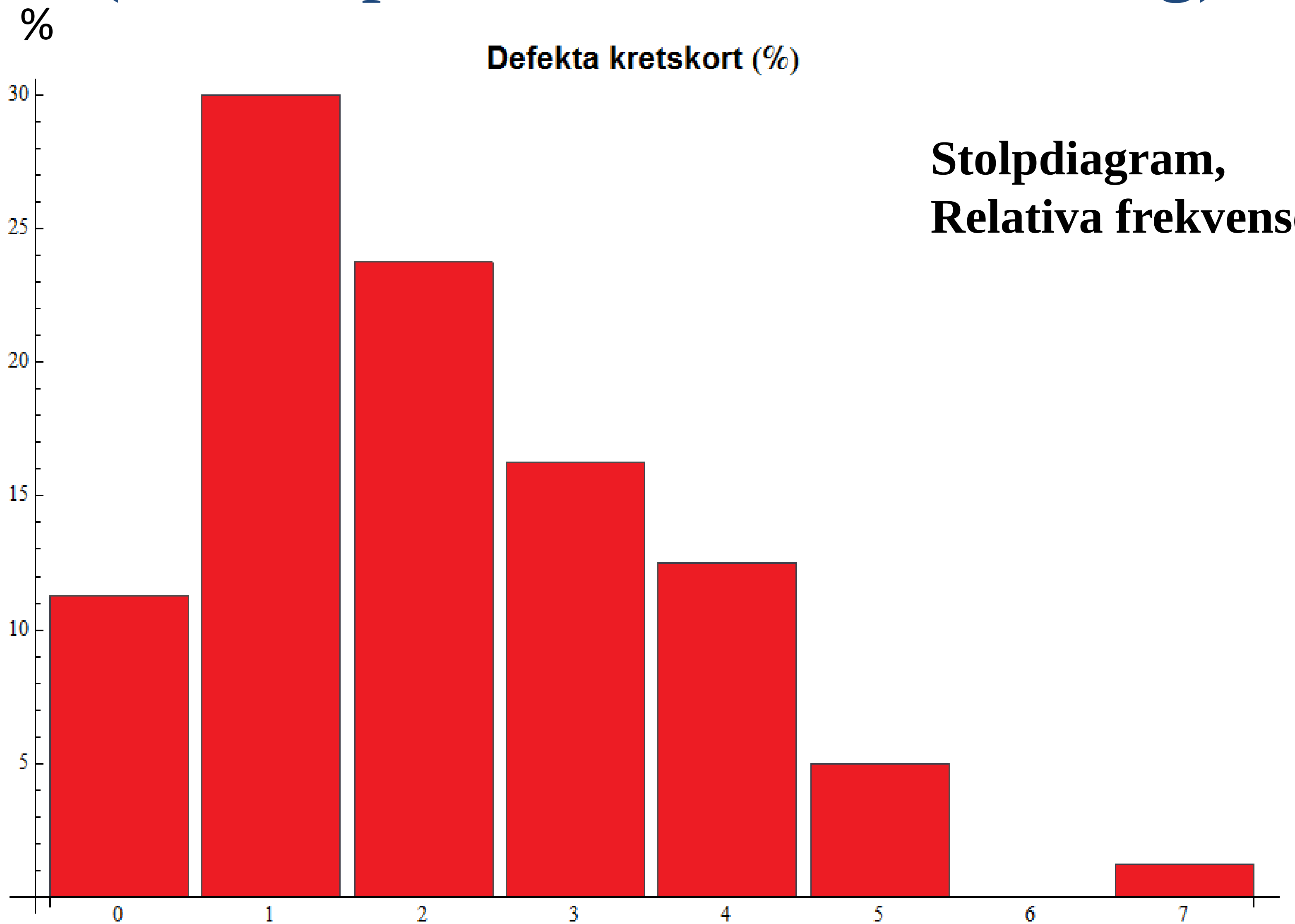
Frekvenstabell för antalet defekta kretskort

Antalet defekta	Frekvens	Rel.frekvens	Kum.frekvens
0	9	11.25	11.25
1	24	30.00	41.25
2	19	23.75	65.00
3	13	16.25	81.25
4	10	12.50	93.75
5	4	5.000	98.75
6	0	0	98.75
7	1	1.250	100.0

Ett exempel på stickprovsundersökning (icke-experimentell undersökning)



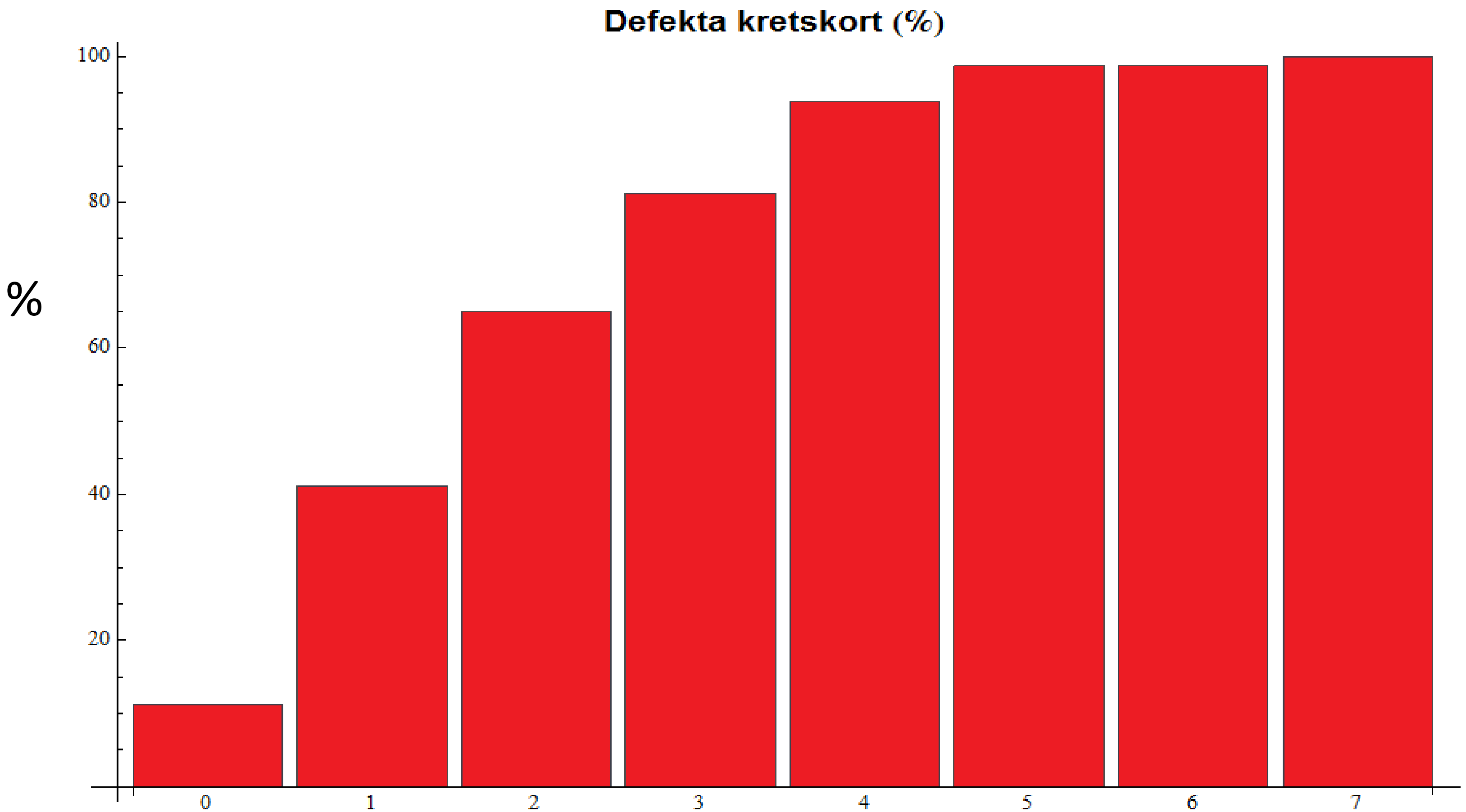
Ett exempel på stickprovsundersökning (icke-experimentell undersökning)



Stolpdiagram, Relativa frekvenser

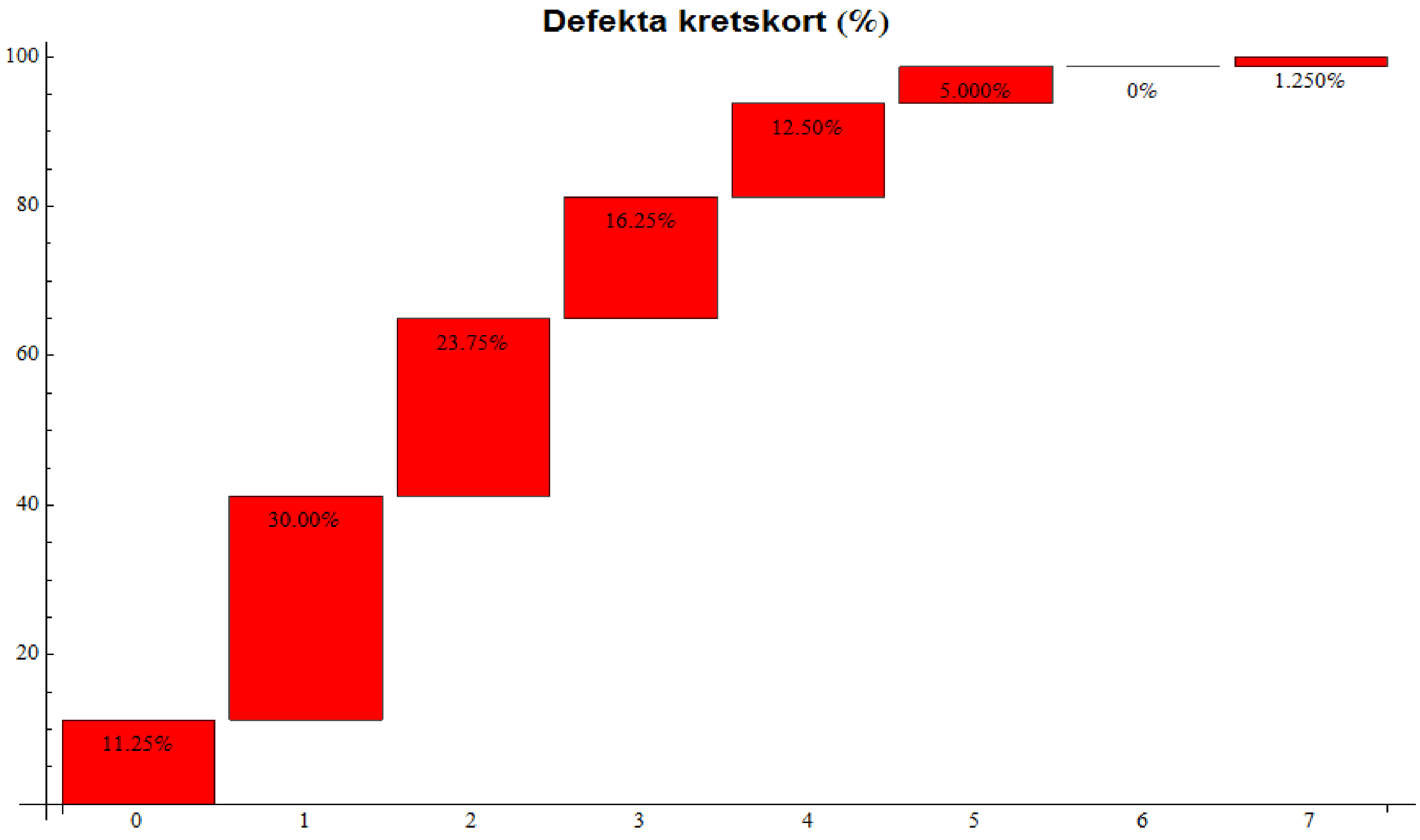
Ett exempel på stickprovsundersökning (icke-experimentell undersökning)

Trappstegskurva för antalet defekta kretskort Kumulativ relativ frekvens



Ett exempel på stickprovsundersökning (icke-experimentell undersökning)

Trappstegskurva för antalet defekta kretskort Kumulativ relativ frekvens



Ett exempel på stickprovsundersökning (icke-experimentell undersökning)

Totalt valdes

200*80 = 16000 kretskort ut för undersökning.

Stickprovstorlek är på 16000, n = 16000.

Stickprovet valdes ut bland totalt

80*10000 = 800000 kort.

Populationsstorleken är på 800000, N = 800000

Felkvoten i stickprovet var 168/16000 = 0.0105

dvs dubbelt så stor än den utlovade.

Vad kan man säga om felkvoten i sändningen?

Hur säkra uttalanden kan man göra om felkvoten?

Ett exempel till på stickprovsundersökning
(Experimentell undersökning)

I Grängesberg gjordes ett fullskaleförsök för att bl.a. studera hur lång tid det tar att fylla en 2 m^3 vagn med malm. Tiden noterades från det att lastmaskinen började köra in i bergshögen tills att lastaren kopplade loss vagnen.
Följande resultat erhölls.
(Detta är ett exempel på *kontinuerlig* variation)

Ett exempel till på stickprovsundersökning
(Experimentell undersökning)

Tidsåtgång vid lastning i sek.

Grunddata

**85,80,85,77,101,109,111,109,148,183,153,78,84,80,94,104,96,100
117,112,103,122,155,153,128,172,69,84,99,110,112,181,176,79,94
111,111,118,133,140,80,84,100,101,122,129,73,75,111,96,126,147
90,103,100,96,116,128,86,80,97,118,124,150,96,105,83,99,140,79
78,87,107,134,140,79,87,104,153,134,82,91,104,128,76,108,141
134,117,110,149,119,121,116,114,130,90,97,127,113,96,106,107,
108,128,110,109,85,95,116,118,110,91,126,97,121,107,104,129,
106,112,91,119,118,105**

Vad kan man säga om μ, den genomsnittliga tidsåtgången för att lasta en vagn?
Frågan kan preciseras på 3 olika sätt:

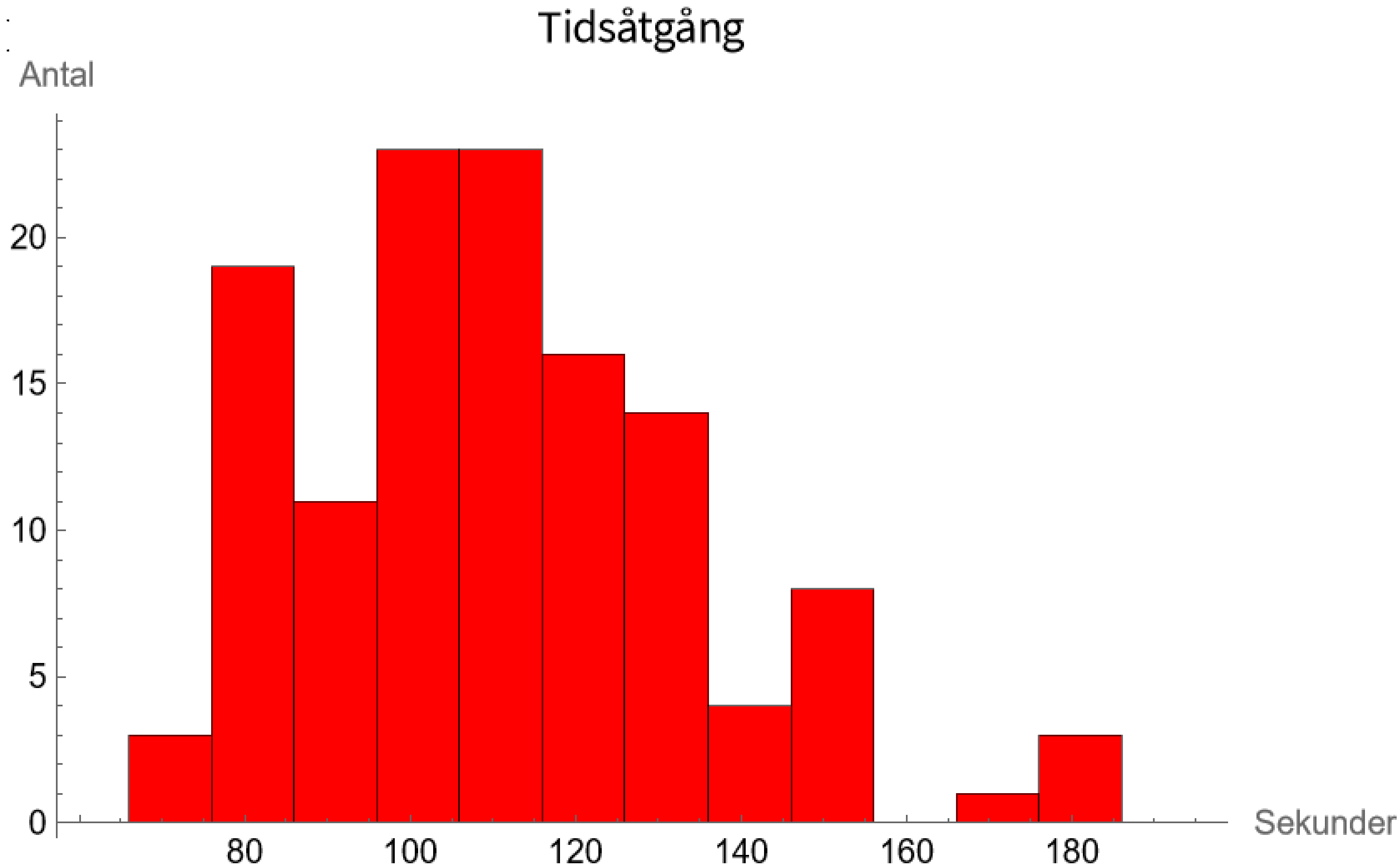
- ***Punktskattningsproblem* – hur skattar man μ?**
- ***Intervallskattningsproblem* – hur anger man ett intervall som med given säkerhet innehåller μ?**
- ***Hypotesprövningsproblem* – hur prövar man hypoteser rörande μ?**

Ett exempel på stickprovsundersökning
(experimentell undersökning)

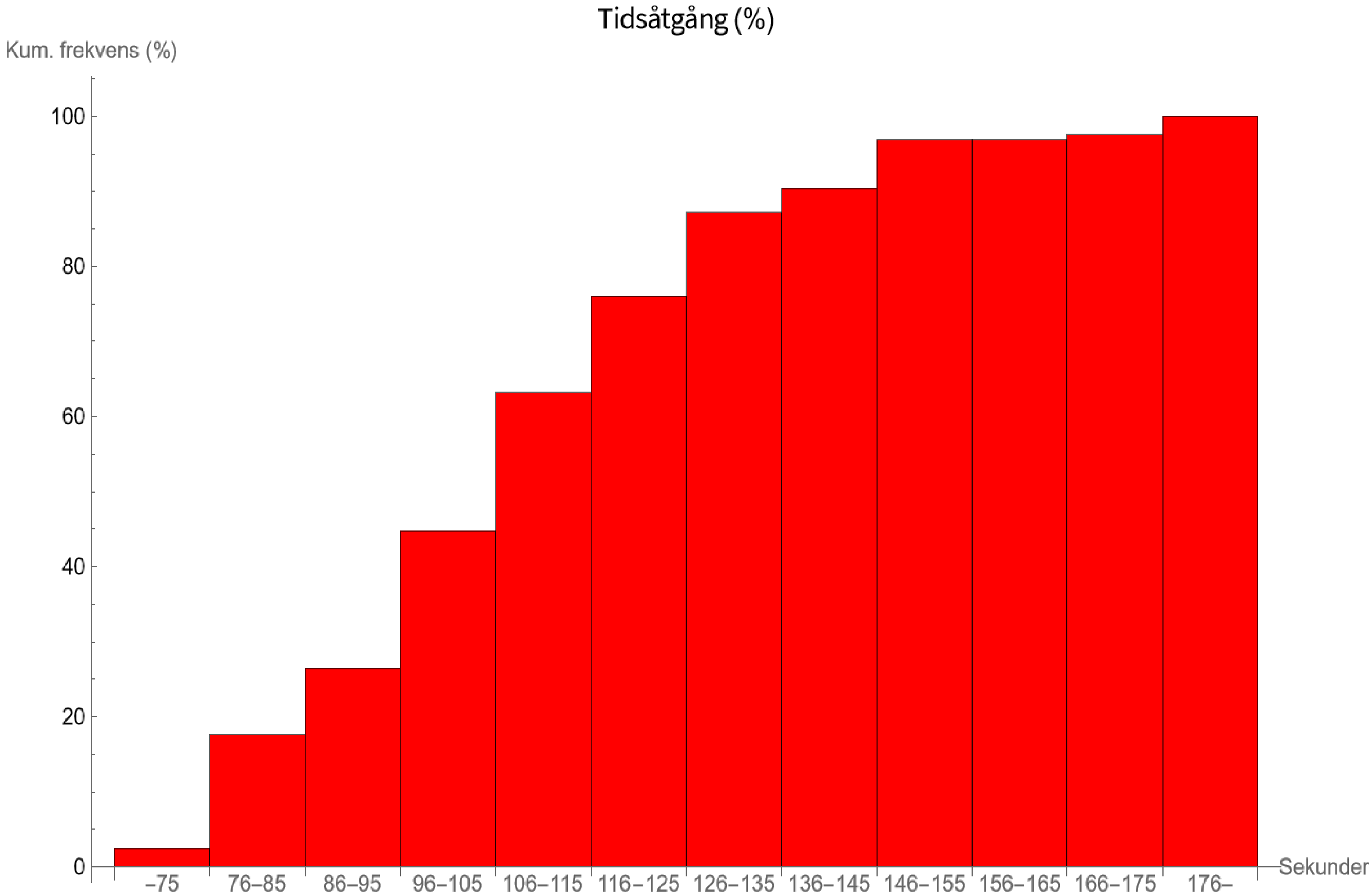
Frekvenstabell för tidsåtgång vid lastning, Klassindelat material

Tidsåtgång	Frekvens	Rel.frekvens (%)	Kum.frekvens (%)
<76	3	2.4	1.6
76-85	19	15.2	17.6
86-95	11	8.8	26.4
96-105	23	18.4	44.8
106-115	23	18.4	63.2
116-125	16	12.8	76.0
126-135	14	11.2	87.2
136-145	4	3.2	90.4
146-155	8	6.4	96.8
156-165	0	0	96.8
166-175	1	0.8	97.6
>175	3	2.4	100

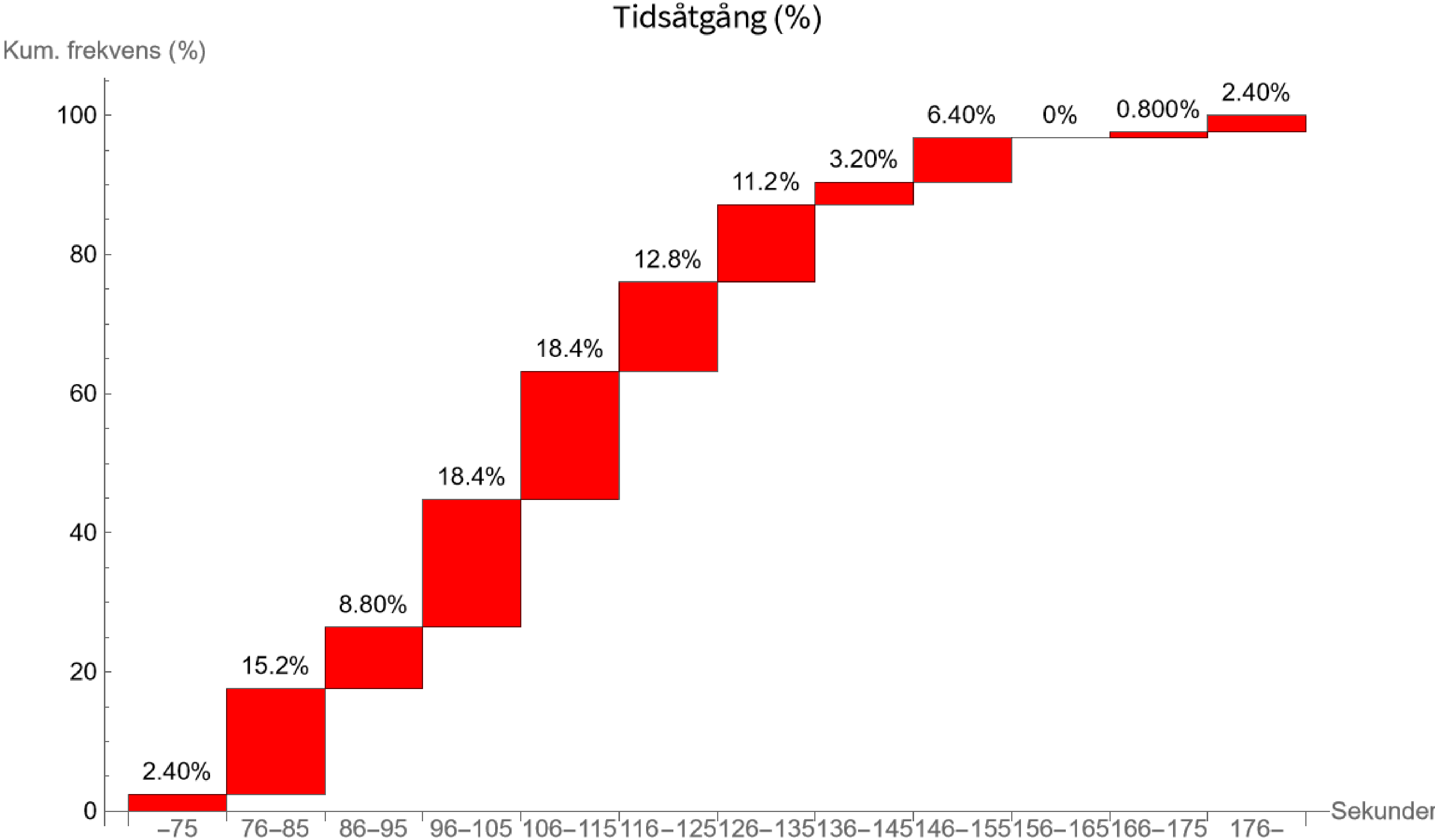
Ett exempel på stickprovsundersökning (experimentell undersökning)



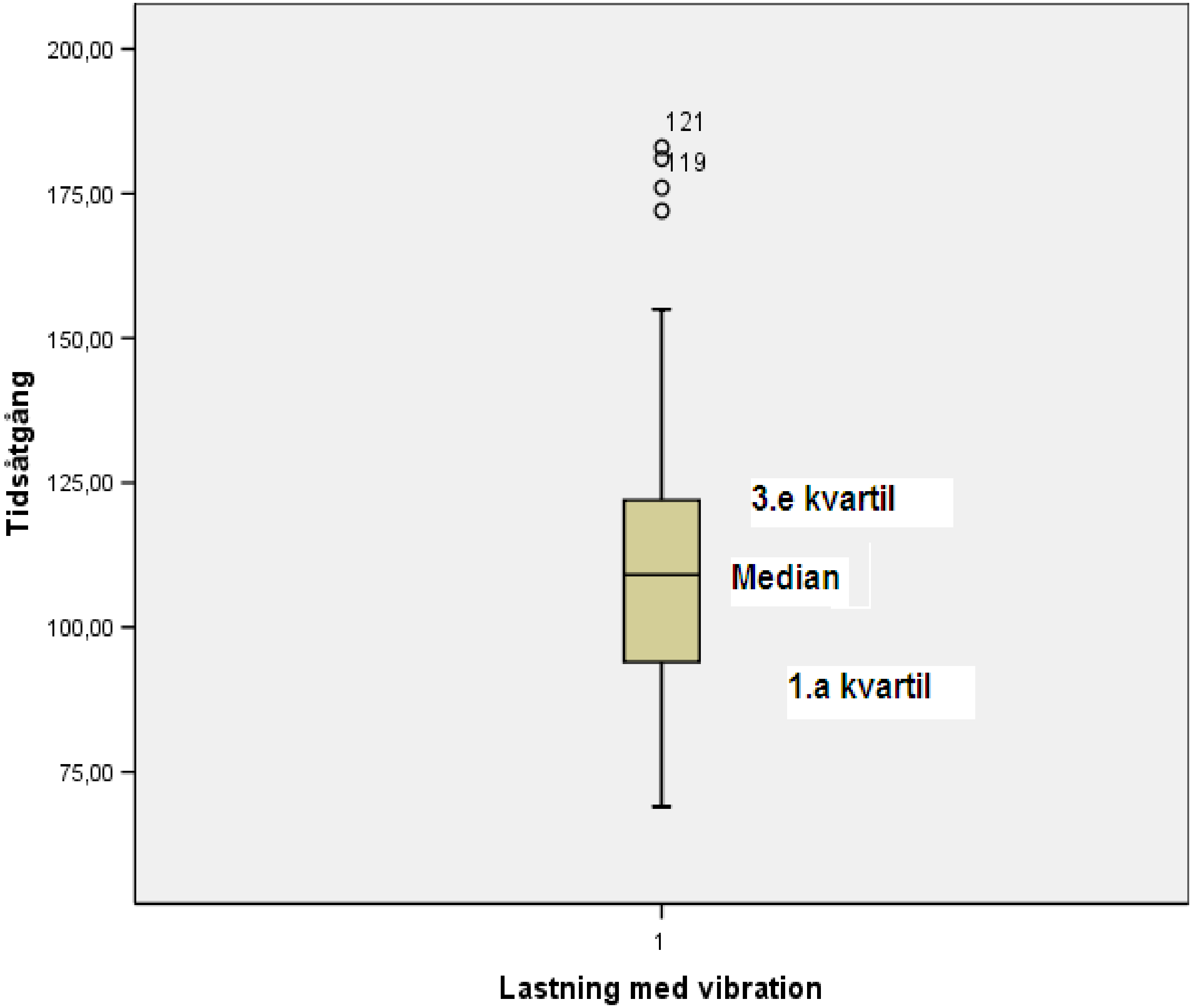
Ett exempel på stickprovsundersökning (experimentell undersökning)



Ett exempel på stickprovsundersökning (experimentell undersökning)



Boxplot tidsåtgång vid lastning



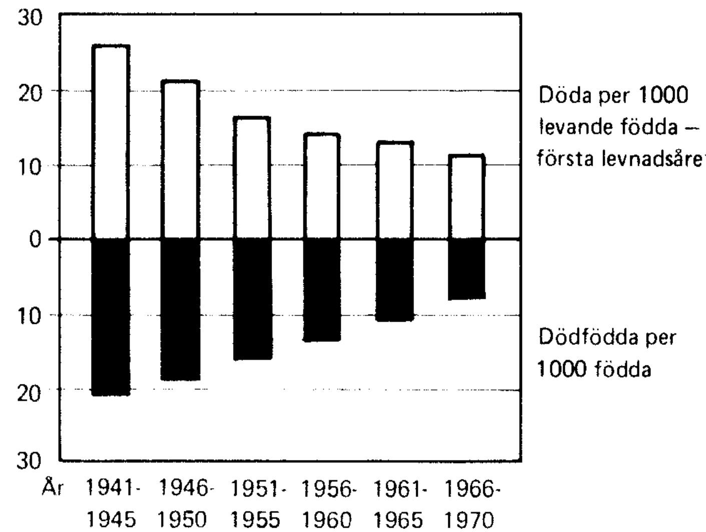
Ett exempel på stickprovsundersökning (experimentell undersökning)

1) Vad är den genomsnittliga tidsåtgången?
   Medelvärdet i stickprovet är $\overline{x} = 110.2\ s$.
2) Hur mycket varierar det?
   Standardavvikelsen i stickprovet är $s = 23.7\ s$.
3) Hur stor andel av lastningen av vagnarna överstiger 2 min?
   Andelen som överstiger 2 min är 28%.

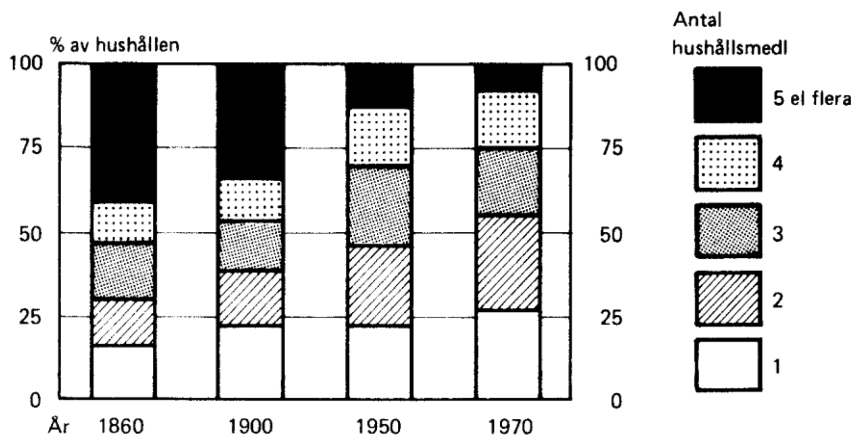
Hur säkra är dessa uttalanden?

Kvalitativa data - exempel

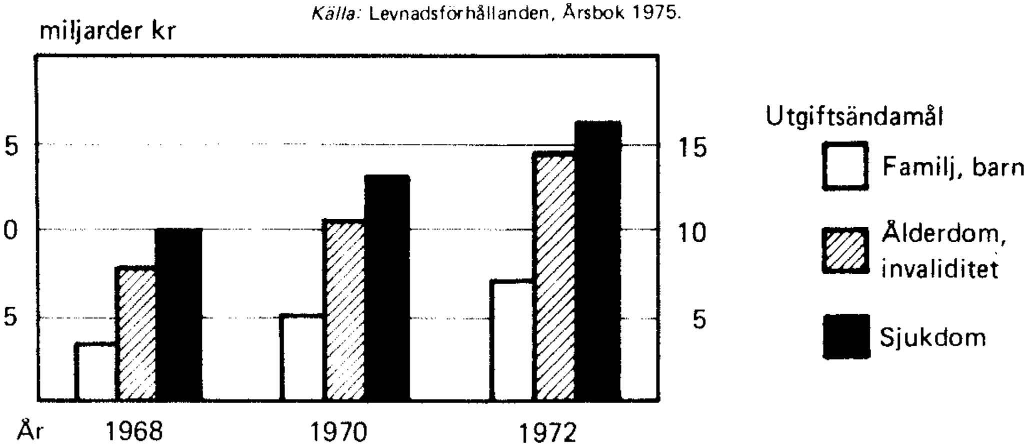
Val-deltagande	Gifta				Ej gifta				Samtliga	
	Män		Kvinnor		Män		Kvinnor			
Ej röstat	25	(3,0)	29	(4,0)	41	(10,9)	44	(13,1)	139	(6,1)
har röstat	806	(97,0)	690	(96,0)	335	(89,1)	293	(86,9)	2124	(93,9)
Summa	831	(100,0)	719	(100,0)	375	(100,0)	337	(100,0)	2263	(100,0)



Motställda staplar: Spädbarnsdödlighet bland flickor 1941–70.
Källa: Levnadsförhållanden, Årsbok 1975.



Uppdelade staplar: Hushåll efter storlek. Procent

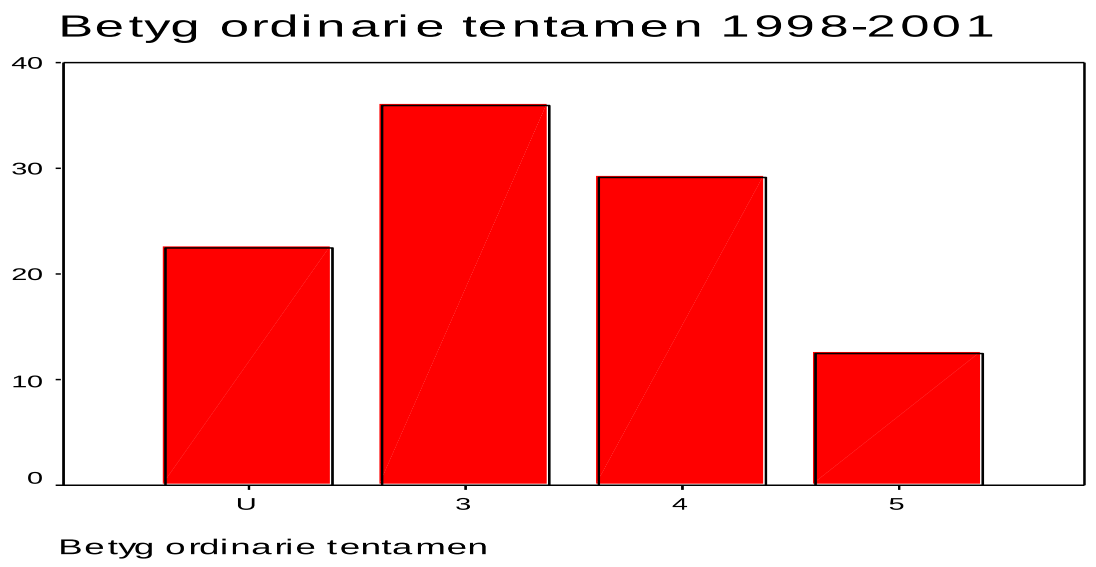


Grupperade staplar: Socialutgifter 1968–72. Miljarder kr.

Kvalitativa data - exempel

Betyg ordinarie tentamen 1998-2001

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	U	63	22.4	22.4	22.4
	3	101	35.9	35.9	58.4
	4	82	29.2	29.2	87.5
	5	35	12.5	12.5	100.0
	Total	281	100.0	100.0	



Beskrivande statistik

Numerisk beskrivning av ett kvantitativt material

- Lägesmått
  - Medelvärde, $\bar{x}$
  - Median (andra kvartil), md, (Q_2)
  - Typvärde, T
- Spridningsmått
  - » Standardavvikelse, s (varians, $V = s^2$)
  - » Kvartilavstånd, Q $(= Q_3 - Q_1)$
  - » Variationsvidd (-bredd), R
- Beroende mått (Korrelation)
  - » Kovarians, c_{xy}
  - » Korrelationskoefficient, r

Lägesmått

✓ **<u>Medelvärde:</u>** $\bar{x} = \frac{1}{n}\sum_1^n x_i$

”Summan av alla värden delat med antalet värden”

✓ **<u>Median</u>:** $m = Q_2$

En *storleksordnad* datamängd kan delas in i 4 kvartiler, Q_i

25% av materialet är $\leq Q_1, 50\%$ är $\leq Q_2$ och

75% är $\leq Q_3$ eller 25% är $\geq Q_3$

✓ **<u>Typvärde</u>, *T***

Det värde som förekommer flest gånger.

Spridningsmått

✓**Standardavvikelse:** $s = \sqrt{\frac{1}{n-1}\sum_1^n (x_i - \overline{x})^2}$

"Genomsnittliga kvadratiska skillnaden mot medelvärdet"

Varians: $V = s^2$

✓**Kvartilavstånd:** $m = Q_3 - Q_1$

50% av materialet ligger mellan Q_1 och Q_3

✓**Variationsbred:** $R = Max - Min$

Beroendemått

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

Kallas kovariansen mellan *x* och *y*.

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

s_x, s_y är standardavvikelser för *x* resp. *y*

Kallas korrelationskoefficienten för *x* och *y*.

Huvudproblem inom statistikteorin

Verklighet	Modell
1. Formulera praktiskt problem	
	2. Gör slumpmodell
3. Insamla data	
	4. Gör statistisk analys
5. Drag praktiska slutsatser	

Vi kommer att syssla mest med teorin kring punkt 2, 4 och 5
(Projektuppgiften täcker alla 5 stegen)

Punktskattning

Definition

Ett *slumpmässigt stickprov* $x_1, x_2, \ldots x_n$ från någon fördelning F utgörs av observationer av oberoende stokastiska variabler $X_1, X_2, \ldots X_n$ var och en med fördelningen F.
Ett utfall $x_1, \ldots, x_n$ av stokastiska variabler $X_1, \ldots, X_n$ kallas för ett observerat stickprov av storleken n

Fördelningen F beror av en (eller flera) okänd parameter θ som vi är intresserade av att få information om. Parametern kan ta värden i ett parameterrum Ω_θ .

Ex. $\Omega_\theta = (-\infty < \theta < \infty)$ eller $\Omega_\theta = (0 < \theta < 1)$

Punktskattningar - även dessa beror av slumpen

Vi är intresserade av att skatta den okända parametern baserat på våra mätdata, $x_1, x_2,.... x_n$ med någon lämplig funktion.

Definition

En *punktskattning* $\theta^*_{obs} = \theta(x_1, x_2, \ldots, x_n)$ (tal) av en okänd parameter θ är en funktion av stickprovet, $x_1, x_2, \ldots, x_n$.
Detta stickprov ska se som utfall av stokastiska variabler,
$X_1, X_2, \ldots, X_n$, med fördelningar som alla beror på θ.
Punktskattning θ^*_{obs} är ett utfall av *stickprovsvariabeln*
$\theta^* = \theta(X_1, X_2, \ldots, X_n)$, (stokastisk variabel)

Önskvärda egenskaper på en punktskattning

En punktskattning θ^* sägs vara:

- ***Väntevärdesriktig,*** om skattningens, θ^*, väntevärde är lika med θ,

dvs **E[** θ^***] =** θ (i genomsnitt hamnar man ”rätt”)

- ***Konsistent,*** om för varje fixt $\theta \in W_Q$ och för givet $\varepsilon > 0$

gäller att ***P(|*** θ_n^* **-** θ **|<** $\boldsymbol{\varepsilon}$**)**$\rightarrow 1$, stickprovsstorleken $n \rightarrow \infty$ (**S**tora talens lag)

- ***Effektiv,*** om θ_1^* och θ_2^* är två väntevärdesriktiga skattningar av θ.

Om **V[**θ_1^***] < V[**θ_2^* **]** sägs θ_1^* vara en effektivare, sannolikt bättre, skattning av θ än θ_2^*.

- Ha ett litet eller inget *systematiskt fel, bias,* **E[**θ_2^***] -** $\theta \approx$ **0.**

Om θ^* är VVR är **E[** θ^***] -** θ **= 0**

Maximum-Likelihood-metoden

Definition

Låt $x_1, x_2, .., x_n$ vara ett stickprov.

Funktionen

$$L(\theta) = \begin{cases} P(X_1 = x_1, X_2 = x_2, .., X_n x_n\,; \theta) & \text{(diskreta variabler)} \\ f_{X_1, X_2, ..., X_n}(x_1, x_2, .., x_n\,; \theta) & \text{(kontinuerliga variabler} \end{cases}$$

kallas $likelihood - funktionen$ eller $L - funktionen$

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

Minsta-kvadrat-metoden

Definition

Låt $x_1, x_2, \ldots, x_n$ vara ett stickprov på $X_1, X_2, \ldots. X_n$ vars väntevärde är kända men beror av en okänd parameter θ, $E(X_i)=\mu_i(\theta)$.

Det värde θ^*_{obs}, för vilket funktionen

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

antar sitt minsta värde kallas *MK-skattningen av* θ.

Allmänna väntevärdesriktiga punktskattningar

- Låt $X_1, X_2, \dots X_n$, där X_i är oberoende och likafördelade stokastiska variabler.
- Låt $x_1, x_2, \dots, x_n$ vara ett stickprov på X

"Bästa" sättet att skatta ett okänt väntevärde, μ, är

$\mu^* = \overline{X}$ och $\mu^*_{obs} = \bar{x}$ eftersom denna är VVR och konsistent.

"Bästa" sättet att skatta en okänd varians, σ^2, är

$$\left(\sigma^2\right)^* = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \text{ och } \left(\sigma^2\right)^*_{obs} = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

Eftersom denna är VVR.