2019-04-08
Föreläsning 5
Johan Swärd

Föreläsning 5

* Miniprojekt I inskickad 16.00 den 16/4. (Glöm inte att göra hemuppgifterna innan!)

Maila in pdf till:

masb11@matstat.lu.se

Ämnesraden i mailet skrivs som:

Miniprojekt1 av studid1 och studid2

där studid1 och studid2 är era SIL-id.

Dagens föreläsning

* Repetition.
* Samplingsfördelningar.
* Modell vs. verklighet.
* Centrala gränsvärdessatsen (CGS)
* Estimatorer.

Repetition

* Vi kan beräkna väntevärdet av X genom

1) $E[X] = \sum_{\text{alla } x} x \, f_X(x)$, om X diskret

2) $E[X] = \int_{-\infty}^{\infty} x \, f_X(x) \, dx$, om X kontinuerlig.

* Hur beräknar vi $E[g(X)]$, där $g(X)$ är en funktion av X?

1) $E[g(X)] = \sum_{\text{alla } x} g(x) \, f_X(x)$, om X diskret

2) $E[g(X)] = \int_{-\infty}^{\infty} g(x) \, f_X(x) \, dx$, om X kontinuerlig.

Exempel: $g(x) = x^2$ och $X \in U(0,10)$.

$$f_X(x) = \begin{cases} \frac{1}{10}, & 0 \le x \le 10 \\ 0, & \text{annars.} \end{cases}$$

$$E[g(X)] = \int_0^{10} x^2 \cdot \frac{1}{10} \, dx = \left[\frac{1}{3} x^3 \frac{1}{10} \right]_{x=0}^{10} = \frac{100}{3}$$

Repetition

* Linjära transformationer av slumpvariabler.
  - $E[X+b] = E[X]+b$
  - $E[aX+b] = a\,E[X]+b$
  - $Var[X+b] = Var[X]$
  - $Var[aX+b] = a^2 Var(X)$

* Summor av slumpvariabler

  Låt X och Y vara oberoende slumpvariabler. Då gäller:
  - $E[X+Y] = E[X]+E[Y]$ (alltid)
  - $V[X+Y] = V[X]+V[Y]$ (om ober.)
  - $E\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N} E[X_i]$ (alltid)
  - $V\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N} V[X_i]$ (om ober.)

Repetition

Om $X_1, \ldots, X_N$ alla är normalfördelade, då är också summan av dem normalfördelad.

Exempel:

$X_n \sim N(\mu, \sigma^2)$. X_n oberoende.

Vad har $\bar{X} = \frac{1}{N}\sum_{n=1}^{N} X_n$ för fördelning?

$$E[\bar{X}] = E\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right] = \frac{1}{N}\sum_{n=1}^{N} E[X_n]$$

$$= \frac{1}{N} \cdot N\mu = \mu$$

$$V[\bar{X}] = V\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right] = \frac{1}{N^2} V\left[\sum_{n=1}^{N} X_n\right]$$

Oberoende

$$= \frac{1}{N^2} N \cdot \sigma^2 = \frac{\sigma^2}{N}.$$

Uppgifter

1. Låt $X \in N(0,1)$ och $Y = 3X + 2$.
   Beräkna $E[Y]$ och $D[Y]$.

2. Låt X och Y vara oberoende och respektive $N(1,1)$ och $N(-1,2)$.
   Vilken fördelning har $X+Y$ och $X-Y$?

3. De oberoende stokastiska variablerna X_1 och X_2 tillhör båda $N(1,2)$.
   Ange fördelningen för $\bar{X} = (X_1 + X_2)/2$.

Samplingsfördelningar

* Hittills har vi tittat på slumpvariabler med kända parametrar. T.ex.

$X \sim N(3, 4)$

$Y \sim Bin(10, 0.2)$

$Z \sim Po(5)$

* I verkligheten vet vi oftast inte vilka värden dess parametrar har.

— Vi måste skatta dem från den insamlade datan.

Modellera data

* Antag att vi är intresserade av att mäta mängden av ett visst hormon hos människor.

* Vilken fördelning har hormon-mängden i människor?

* Vi samlar data.

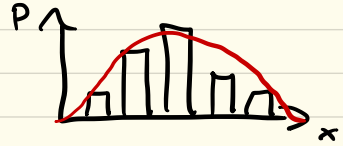
* Är det $Exp(\lambda)$? $N(\mu, \sigma^2)$? $R(a,b)$?

* Vi testar genom att skatta λ för $Exp(\lambda)$, μ och σ^2 för $N(\mu, \sigma^2)$ och a och b för $R(a,b)$

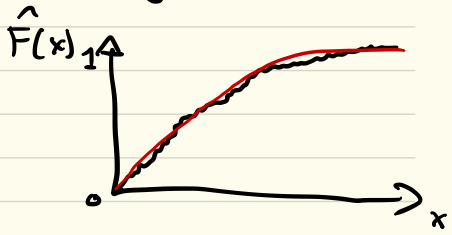
Modellera data

* Vi kan se vilka fördelningar som passar bäst genom att undersöka:

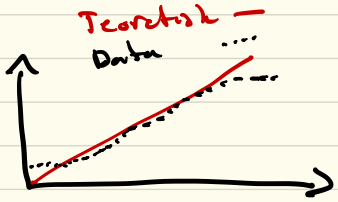
  - Histogram



  - Empirisk fördelningsfunktion



  - QQ-plot



Samplingsfördelningar

* Säg att vi vet att en myras vikt är normalfördelad men att vi inte vet vad väntevärdet, μ, är.

* Vi hittar en myrstack och väljer ut 10 myror.

Modell -> X_i = "Myran i:s vikt" $\in N(\mu, \sigma^2)$

Mätvärden -> x_i = "Den uppmätta vikten av en myra".

* Vi beräknar medelvärdet för att skatta μ:

$$\hat{\mu} = \frac{1}{10} \sum_{i=1}^{10} x_i$$

Samplingsfördelningar

* Säg att medelvärdet blev 5.2 mg. Om vi nu väljer ut 10 nya myror, kommer medelvärdet av dessa bli 5.2 mg?

Samplingsfördelningar

* Säg att medelvärdet blev 5.2 mg. Om vi nu väljer ut 10 nya myror, kommer medelvärdet av dessa bli 5.2 mg?

  — Troligtvis inte.

* Det betyder att $\hat{\mu}$ också är en slumpvariabel.

* Fördelning?

Samplingsfördelningar

* Säg att medelvärdet blev 5.2 mg. Om vi nu väljer ut 10 nya myror, kommer medelvärdet av dessa bli 5.2 mg?

  – Troligtvis inte.

* Det betyder att $\hat{\mu}$ också är en slumpvariabel.

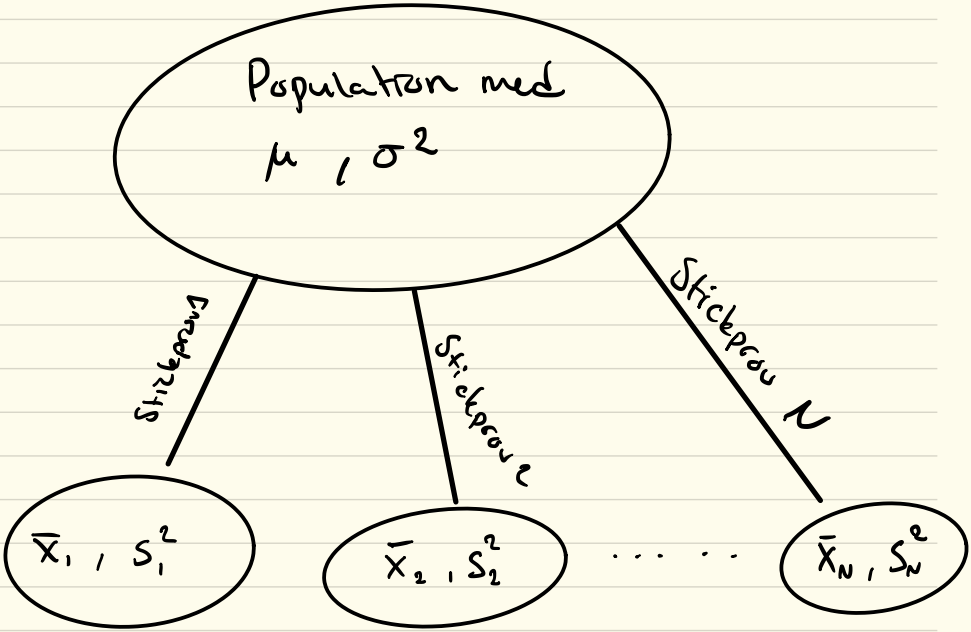
* Fördelning? Summa av normalfördelningar...

$$E[\hat{\mu}] = \frac{1}{N} \cdot N \cdot \mu = \mu$$

$$V[\hat{\mu}] = \frac{1}{N^2} \cdot N\sigma^2 = \frac{\sigma^2}{N}$$

$$\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{N}\right)$$

Samplingsfördelningar



Samplingsfördelningar

Vad händer om vi inte vet vilken fördelning myrorna har?

Vilken fördelning kommer $\hat{\mu}$ ha då?

Samplingsfördelningar

Vad händer om vi inte vet vilken fördelning myrorna har?

Vilken fördelning kommer $\hat{\mu}$ ha då?

För att svara på detta vill jag först göra ett experiment.

Låt X vara utfallet vid ett tärningskast.

$$f_X(x) = \begin{cases} 1/6, & x = 1, 2, 3, 4, 5, 6 \\ 0, & \text{annars.} \end{cases}$$

Om $T_n = \sum_{i=1}^{n} X_i$. (Dvs. summen av n tärningskast)

Vad är fördelningen för T_n?

CGS

* Centrala gränsvärdessatsen (CGS) säger att om vi har oberoende och likafördelade slumpvariabler

$X_1, X_2, \dots, X_N$ med $E[X_i] = \mu$
och $V[X_i] = \sigma^2 < \infty$

och N är stort, då är

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

CGS

* Centrala gränsvärdessatsen (CGS) säger att om vi har oberoende och likafördelade slumpvariabler

$X_1, X_2, \dots, X_N$ med $E[X_i] = \mu$ och $V[X_i] = \sigma^2$

och N är stort, då är

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

* Det betyder att medelvärdet av oberoende och likafördelade slumpvariabler är

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \sim N\left(\mu, \frac{\sigma^2}{N}\right)$$

CGS

Exempel: Vikten (i gram) av en slumpmässigt vald magnecyltablett är en S.V. med väntevärde 0.65 och varians 0.0004.

a) Vad är väntevärde och varians för den totala vikten av 100 tabletter (vars vikt är oberoende av varandra)?

b) Vad är sannolikheten att 100 magnecyltabletter tillsammans väger högst 65.3g?

CGS

Exempel: Vikten (i gram) av en slumpmässigt vald magnecyltablett är en S.V. med väntevärde 0.65 och varians 0.0004.

a) Vad är väntevärde och varians för den totala vikten av 100 tabletter (vars vikt är oberoende av varandra)?

b) Vad är sannolikheten att 100 magnecyltabletter tillsammans väger högst 65.3g?

Lösning: X = "Vikt hos en tablett"

a) $E\left[\sum_{k=1}^{100} X_k\right] = 100 \cdot 0.65 = 65$

$V\left[\sum_{k=1}^{100} X_k\right] \overset{\text{obes.}}{=} 100 \cdot 0.0004 = 0.04$

b) $P\left(\sum_{k=1}^{100} X_k \leq 65.3\right)$? $\quad Y = \sum_{k=1}^{100} X_k$

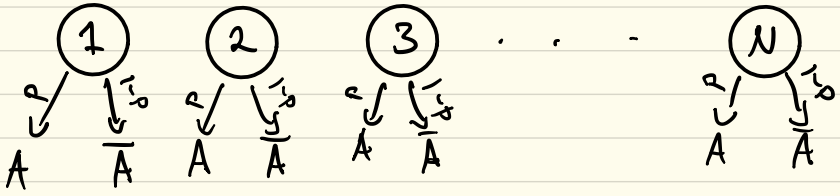
Eftersom oberoende och likafördelade X_k

CGS ger: $Y \underset{\sim}{\in} N(65, 0.04)$, $P(Y \geq 65.3) = \Phi\left(\frac{65.3-65}{\sqrt{0.04}}\right)$

$= \Phi\left(\frac{0.3}{0.2}\right) = \Phi(1.5) \overset{\text{tabell}}{=} 0.9332$

CGS och Binomialfördelningen

* Vi har N oberoende försök och $P(A) = p$:



Om $X_i = \begin{cases} 1 & , \text{ om } A \text{ sker i försök } i \\ 0 & , \text{ om } A \text{ ej sker i försök } i \end{cases}$

$X =$ "Antal gånger A sker i N försök"

$= \sum_{i=1}^{N} X_i \in Bin(N, p)$

CGS säger att X kan approximeras med

$X \underset{\sim}{\in} N(Np, Np(1-p))$

om $Np(1-p) \geq 10$

CGS och Binomialfördelningen

Exempel: 100 slumpmässigt valda svenskar frågas huruvida de röker. Vad är sannolikheten att fler än 40 svarade att de röker om ca 30 % av alla svenskar röker?

CGS och Binomialfördelningen

Exempel: 100 slumpmässigt valda svenskar frågas huruvida de röker. Vad är sannolikheten att fler än 40 svarade att de röker om ca 30% av alla svenskar röker?

Lösning: X = "Antal rökare" $\in Bin(100, 0.3)$

$P(X \geq 40)$? Tabell räcker ej!

Men $100 \cdot 0.3(1-0.3) = 21 > 10$

Vi kan (cas) använda CGS:

$$X \overset{\text{cas}}{\sim} N(np, np(1-p)) = N(30, 21)$$

$$P(X \geq 40) = 1 - P(X \leq 39) \approx 1 - P\left(\frac{X-30}{\sqrt{21}} \leq \frac{39-30}{\sqrt{21}}\right)$$

Normalapprox här!!!

$$= 1 - \Phi(1.96) = 1 - 0.975 = 0.025$$

Uppgifter

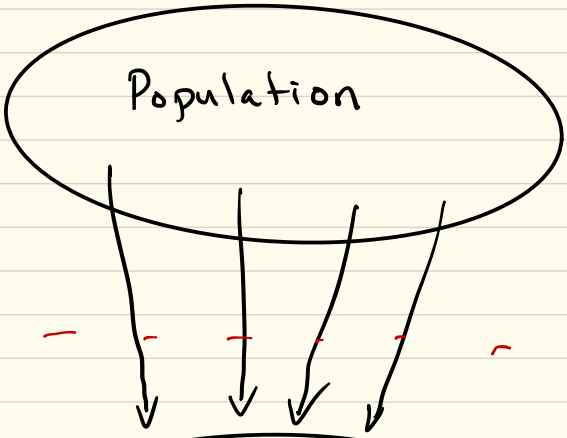
Ett bostadsområde med 1000 familjer. Sannolikhetsfunktionen för antal barn i förskoleålder i en slumpvald familj är

$$f(x) = \begin{cases} 0.4 & x = 0 \\ 0.2 & x = 1 \\ 0.3 & x = 2 \\ 0.1 & x = 3 \\ 0 & \text{annars} \end{cases}$$

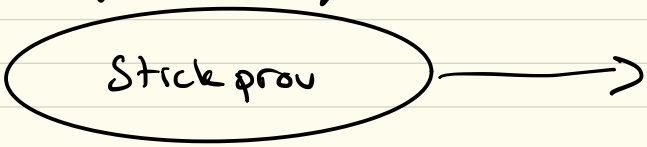
Antal barn mellan olika familjer är oberoende. Hur många dagbemsplatser ska planeras om sannolikheten att alla barn ska få dagbemsplats ska vara 90%?

Statistisk inferens

"Idealvärld"



Populationens parametrar
μ, σ^2 etc.



Sammanfattande mått för stickprovet
$\bar{x}, s^2$ etc.

"Verklighet"

Beteckningar:	Population	Stickprov
Antal enheter	N	n
Medelvärde	μ	$\bar{x}$
Varians	σ^2	s^2
Standard-avvikelse	σ	s
proportion	p	$\hat{p}$

BIOSTATISTISK GRUNDKURS, MASB11
OH-BILDER 5

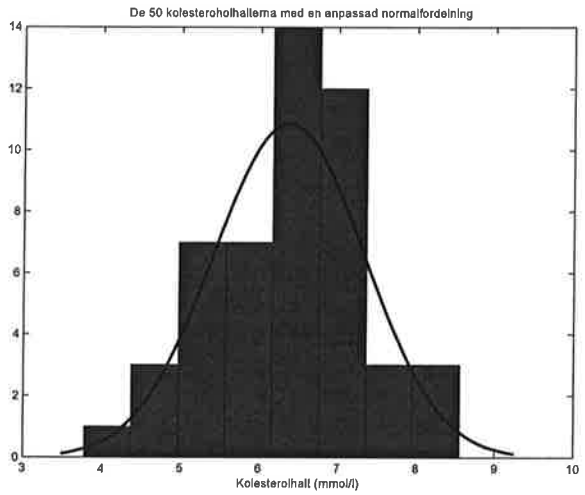
EXEMPEL PÅ FRÅGESTÄLLNINGAR INOM STATISTISK INFERENS:

VAD ÄR FÖRVÄNTAT VÄRDE?

I en hälsoundersökning valde man slumpmässigt ut 50 manliga bussförare och noterade bl.a. deras kolesterolhalt. De uppmätta halterna varierar och man antar att de kommer från en $N(\mu, \sigma^2)$. Några av halterna (mmol/l):

6.9 6.6 5.6 3.8 7.2 ... 5.3 6.3 5.9 6.2

Histogram över samtliga 50 halter med en anpassad normalfördelning:



- **Vad kan man säga om μ — förväntad kolesterolhalt hos manliga bussförare?**
- **Hur kan man bäst estimera (skatta) μ och σ?**
- **Intressant är att få så mycket information som möjligt om μ. Kan man göra mer än en estimation (skattning)?**
- **Antag att för den manliga "normalbefolkningen" gäller att populationsmedelvärdet beträffande kolesterolhalt är 6.0 mmol/l. Är det troligt att μ — förväntad kolesterolhalt hos bussförarna — också är 6.0 eller skiljer det sig från "normalbefolkningens"?**

FRÅGESTÄLLNINGAR (FORTS):

FINNS DET EN SIGNIFIKANT SKILLNAD MELLAN BEHANDLINGAR?
EXEMPEL: För att undersöka om en viss medicin har som primär biverkan att höja ett visst levervärde mättes detta, på 50 personer som ej behandlats med medicinen samt på 25 personer som behandlats med medicinen. Resultat:

Behandling	Medelvärde	Standardavvikelse	n-antal
Utan medicin	148.2	10.0	50
Med medicin	151.7	8.0	25

Kan man dra några slutsatser om att medicinen höjer levervärdet?

EXEMPEL: Man ville göra en jämförelse mellan två olika läkemedels botande förmåga. Sammanlagt 110 patienter med urinvägsinfektion förorsakad av en viss bakterie ingick i försöket. Antibiotikum A gavs till 60 kvinnor varav 80 % blev friska. Antibiotikum B gavs till 50 kvinnor varav 60 % blev botade. Kan vi påstå att det finns någon skillnad mellan andelen botade med de respektive läkemedlen?

Vad är skillnaden mellan SANNOLIKHETSTEORI och STATISTISK INFERENS?

Tidigare SANNOLIKHETSTEORI:

- **EX:** X **= kolesterolhalten** $\sim N(6.0,\ 1.1)$**. Vad är sannolikheten att halten överstiger 6.5, d.v.s.** $P(X > 6.5)$ **söks.**
- **EX:** P**(en person röker)** $= 0.2$**. Välj ut 5 personer.** X **= antal rökare bland de 5. Vad är sannolikheten att det finns minst 2 personer av de fem som röker, d.v.s. beräkna** $P(X \geq 2)$**.**

FÖRDELNINGARNA ÄR HELT KÄNDA (Vi känner alla populationsparametrar).

Nu STATISTISK INFERENS:

- **EX:** X **= kolesterolhalten** $\sim N(\mu, \sigma^2)$ **där** μ **och** σ **är okända. Vi mäter halten på** n **personer och får värdena** $x_1, \ldots, x_n$**. Kan vi nu säga något** μ **och** σ**?**
- **EX:** P**(en person röker)** $= p$ **(okänd). Av 50 utvalda personer var det 17 som rökte. Vad kan vi nu säga om** p**?**

FÖRDELNINGARNA INNEHÅLLER OKÄNDA PARAMETRAR
Vi använder data för att dra slutsatser om parametrarna.

VI ARBETAR MED DESSA METODER I DEN STATISTISKA INFERENSEN:

SKATTNINGAR:

- **Hur ska vi skatta μ och σ i kolesterolhaltexemplet?**
- **Hur nära ligger våra estimatorer (skattningar) de sanna (okända) värdena på μ och σ?**

KONFIDENSINTERVALL:

- **Hur skaffa ett intervall $I_\mu = (a, b)$ sådant att vi med en viss säkerhet (t.ex. 95 %) kan säga att det täcker över μ — förväntad kolesterolhalt hos en bussförare?**

HYPOTESTEST:

- **Skiljer sig μ — förväntad kolesterolhalt hos en bussförare — från 6.0 ("normalhalt")? Ställ upp hypoteser:**

  $H_0 : \mu = 6.0$
  $H_1 : \mu \neq 6.0$

  Undersök med ett test om H_0 kan förkastas till förmån för H_1.

Både konfidensintervall och hypotestest baseras på estimatorer och sannolikhetsberäkningar kring estimatorer.

ESTIMATORER (Skattningar):

Exempel: $X =$ **kolesterolhalten hos bussförare;** $X \sim N(\mu, \sigma^2)$
Vi har mätningar (halter) $x_1, \ldots, x_{50}$

$\hat{\mu}$ **betecknar en estimator av** μ
$\hat{\sigma}$ **betecknar en estimator av** σ

Vi väljer:

$$\hat{\mu} = \bar{x} = \frac{1}{50}\sum_{i=1}^{50} x_i$$

$$\hat{\sigma} = s = \sqrt{\frac{1}{50-1}\sum_{i=1}^{50}(x_i - \bar{x})^2}$$

I det aktuella stickprovet visade sig $\bar{x} = 6.369$ **och** $s = 0.96$.

Hade vi tagit ett nytt stickprov om 50 andra män hade $\bar{x}$ **(och** s**) förmodligen fått andra värden. Hur mycket kan** $\bar{x}$ **variera?**

Tnligt tidigare:

$$\bar{x} \sim N(\mu, \frac{\sigma^2}{n}) = N(\mu, \frac{\sigma^2}{50})$$

Storheten $\frac{\sigma}{\sqrt{n}}$ **kallas ofta "standard error of the mean" (SEM)**

MAN VILL ATT ESTIMATORER SKA

- **vara "unbiased" (väntevärdesriktiga) — d.v.s. ska i genomsnitt verkligen skatta rätt värde.**

  EX: Det förväntade värdet för $\hat{\mu} = \bar{x}$ **är** μ**. BRA!**

- **ha en så liten spridning som möjligt — d.v.s. estimatorns varians ska vara låg.**

  EX: $Var(\bar{x}) = \frac{\sigma^2}{n}$**. Ju större** n **desto mindre varians (desto effektivare är estimatorn).**

Estimator

* En estimator är en skattning av en parameter.

  T. ex. $\hat{\theta}$ är en skattning av θ

  $\hat{\sigma}$ är en skattning av σ

  $\widehat{Var}(\hat{\theta})$ är en skattning av $Var(\hat{\theta})$.

* Estimatorn beror på stickprovet och är själv en slumpvariabel.

Estimator

* Vad är en bra estimator?

  - Väntevärdesriktig:

    $$E[\hat{\theta}] = \theta$$

  - Effektiv

    $V[\hat{\theta}]$ är så liten som möjligt.

Det vill säga att om vi vill skatta θ så vill vi att vår estimator ska ge värdet θ i snitt (väntevärdesriktig) och med så liten varians som möjligt.

Estimator

Exempel: Vi har X_1, X_2 och X_3 (oberoende)

$E[X_k] = \mu$, $V[X_k] = \sigma^2$ för $k = 1, 2, 3$.

Två estimatorer: $\hat{\theta}_1 = \frac{1}{3}(X_1 + X_2 + X_3)$

$\hat{\theta}_2 = X_2$

Vilken är bäst för att skatta μ?

Estimator

Exempel: Vi har X_1, X_2 och X_3 (oberoende)

$E[X_k] = \mu$, $V[X_k] = \sigma^2$ för $k = 1, 2, 3$.

Två estimatorer: $\hat{\theta}_1 = \frac{1}{3}(X_1 + X_2 + X_3)$

$\hat{\theta}_2 = X_2$

Vilken är bäst för att skatta μ?

Väntevärdesriktighet:

$\hat{\theta}_1$: $E[\hat{\theta}_1] = \frac{1}{3}(E[X_1] + E[X_2] + E[X_3]) = \mu$

$\hat{\theta}_2$: $E[\hat{\theta}_2] = E[X_2] = \mu$

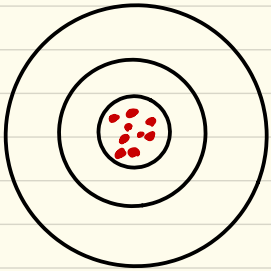
Båda väntevärdesriktiga. Vilken är mest effektiv?

$\hat{\theta}_1$: $V[\hat{\theta}_1] = V\left[\frac{1}{3}(X_1 + X_2 + X_3)\right] \overset{\text{oberoende}}{=} \frac{1}{9} \cdot 3\sigma^2 = \frac{\sigma^2}{3}$

$\hat{\theta}_2$: $V[\hat{\theta}_2] = V[X_2] = \sigma^2$

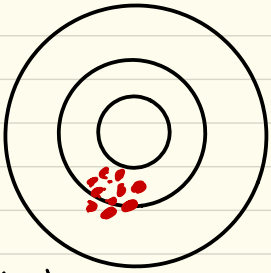
Eftersom $\frac{\sigma^2}{3} < \sigma^2$ så är $\hat{\theta}_1$ bättre än $\hat{\theta}_2$.

Estimator



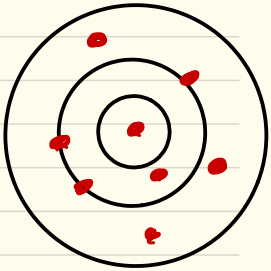
Liten varians (effektiv)

Väntevärdesriktig.



Liten varians (effektiv)

Ej väntevärdesriktig



Stor varians (ej effektiv)

Väntevärdesriktig.

För Normalfördelningen använder vi följande estimatorer:

$$\hat{\mu} = \frac{1}{N}\sum_{k=1}^{N} X_k = \bar{X}$$

Med mätningar $x_1, x_2, \dots, x_n$

$$\hat{\mu}_{obs} = \bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{k=1}^{N} (X_k - \mu)^2$$

Med mätningar $x_1, x_2, \dots, x_n$

$$\sigma^2_{obs} = s^2 = \frac{1}{n-1}\sum_{k=1}^{n} (x_i - \bar{x})^2$$