2019-04-17
Föreläsning 6
Johan Swärd

Föreläsning 6

* Deadline för Mozquizto test 1
  klockan 23.59 den 8/5

* Miniprojekt 1 rättas/är rättad.

Dagens föreläsning

* Repetition
* Statistisk inferens (Kap 6.1 - 6.5 i boken)
    - Estimat av μ och σ^2
    - Vad utgör en bra estimator?
    - Konfidensintervall

Dagens föreläsning

* Repetition
* Statistisk inferens (kap 6.1 - 6.5 i boken)
  - Estimat av μ och σ^2
  - Vad utgör en bra estimator?
  - Konfidensintervall

Repetition

* Centrala gränsvärdessatsen (CGS)

  säger att om vi har oberoende och likafördelade slumpvariabler $X_1, X_2, \ldots, X_N$ med $E[X_i] = \mu$ och $V[X_i] = \sigma^2$

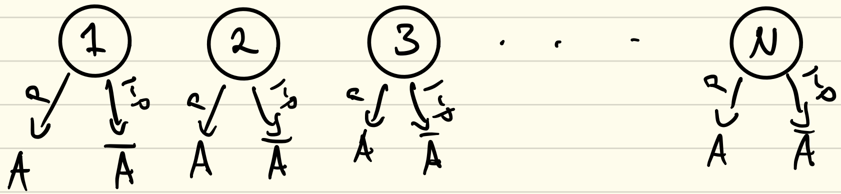
  och N är stort, då är

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

Approximativt

Repetition

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

Repetition

När kan vi normalapproximera?

* $\text{Bin}(n, p) \to N(np, np(1-p))$

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

* $\text{Po}(\lambda) \to N(\lambda, \lambda)$

  om $\lambda \geq 15$

* Annan fördelning. Normalapproximera när N är stort.

För att få full poäng så måste man motivera varför man får använda CGS.

<u>Exempel</u>: "Eftersom N är stort och alla X_k är oberoende och likafördelade så kan vi använda CGS."

Repetition

Exempel: En slumpmässigt vald flygplanspassagerares vikt har en fördelning med väntevärde 75 kg och en standardavvikelse på 12 kg. Ett flygplan kan ha max 3120 kg passagerare.

a) Vad är sannolikheten att vikten av 40 slumpm. valda passagerare överstiger maxkapaciteten?

b) Man vill sänka risken att överstiga maxkapaciteten med 40 passagerare ombord till ungefär 1/100. Hur stor ska maxkapaciteten vara?

Repetition

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 daghemsplatser ska planeras om sannolikheten att alla barn ska få daghemsplats ska vara 90%?

Repetition

Estimator: Ett sätt att använda data till att skatta en parameter. En bra estimator $\hat{\theta}$ av θ ska:

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

$V[\hat{\theta}]$ så liten som möjlig (effektiv).

Repetition

Estimator: Ett sätt att använda data till att skatta en parameter. En bra estimator $\hat{\theta}$ av θ ska:

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

$V[\hat{\theta}]$ så liten som möjlig (effektiv).

Exempel: Vilken skattare av μ är bäst? $\theta_1 = 0.7 X_1 + 0.3 X_2$ eller $\theta_2 = \frac{1}{2} X_1 + \frac{1}{2} X_2$ om X_1, X_2 oberoende och $E[X_1] = E[X_2] = \mu$ samt $V[X_1] = V[X_2] = \sigma^2$?

Konfidensintervall

* Eftersom vi får olika värden på vår skattning för olika stickprov så kan vi inte exakt lita på att vår skattning överestimmer med värdet för den verkliga parametern.

Exempel: Gränsen för radon i ett hus är $200 Bq/m^3$. Vi gör 10 mätningar. Medelvärdet blev $\bar{x} = 190 Bq/m^3$. Kan vi vara säkra på att radonnivån i huset är under gränsvärdet?

För att bättre svara på frågan kan vi göra ett konfidensintervall för radonhalten i huset.

Konfidensintervall

Säg att vorje randonmätning är

X = "randonhalt i huset" $\in N(\mu, \sigma^2)$.

Vi vill uttala oss om värdet för μ.

Vi skattar μ genom medelvärdet:

$\bar{x} = 190$ Bq/m^3

Antag att vi vet att $\sigma^2 = 9$.

Inom vilka värden a och b

ligger μ med 95% sannolikhet?

Detta är ett konfidensintervall.

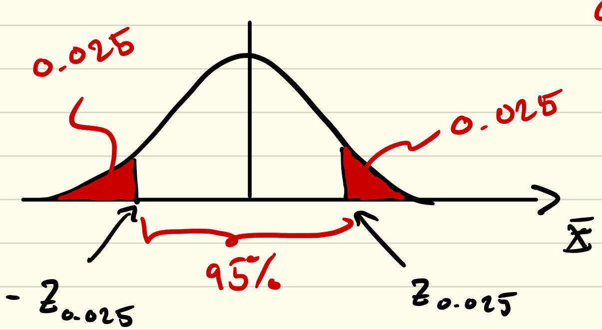
$P(a \leq \mu \leq b) = 0.95$

Konfidensintervall

Vi vet att $\bar{X} \in N\left(\mu, \frac{9}{10}\right)$

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{9}{10}}} \in N(0,1)$$

$0.025 + 0.025 = 0.05$

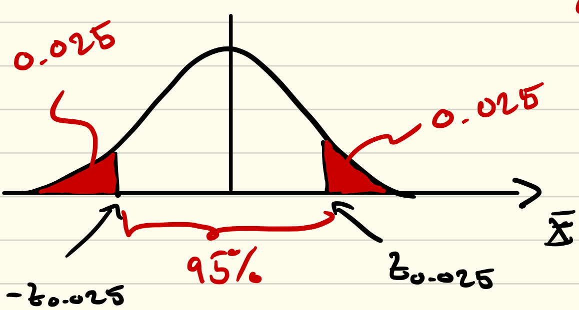


Konfidensintervall

Vi vet att $\bar{X} \in N\left(\mu, \frac{9}{10}\right)$

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{9}{10}}} \in N(0,1)$$

$0.025 + 0.025 = 0.05$



$$P\left(-z_{0.025} \leq \frac{\bar{X} - \mu}{\sqrt{\frac{9}{10}}} \leq z_{0.025}\right) = 0.95$$

$$\Rightarrow \quad P\left(\bar{X} - \sqrt{\frac{9}{10}} \cdot z_{0.025} \leq \mu \leq \bar{X} + \sqrt{\frac{9}{10}}\, z_{0.025}\right) = 0.95$$

Tabell ger $z_{0.025} = 1.96$

$$P\left(\bar{X} - \frac{3}{\sqrt{10}} 1.96 \leq \mu \leq \bar{X} + \frac{3}{\sqrt{10}} 1.96\right) = 0.95$$

Med $\bar{x} = 190$ får vi $P(188.1 \leq \mu \leq 192.0) = 0.95$

$$\boxed{\text{Skriv: } \quad I_\mu = (188.1\,,\,192.2)}$$

Konfidensintervall

Tolkningen av $I_\mu = (188.1, 192.0)$

I_μ täcker det verkliga värdet på μ med 95% sannolikhet.

Vi borde således vara ganska säkra i huset.

Konfidensintervall

Tolkningen av $I_\mu = (188.1, 192.0)$

I_μ täcker det verkliga värdet på μ med 95 % sannolikhet.

Vi borde således vara ganska säkra i huset.

Mer generellt, om σ^2 är känd, så blir konfidensintervallet för μ:

$$\boxed{I_\mu = \left(\bar{x} - z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}},\ \bar{x} + z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right)}$$

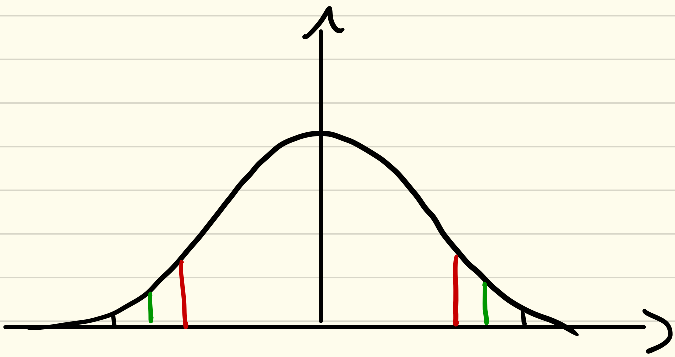
där α är konfidensgraden

($\alpha = 0.05$ i exemplet ovan)

som bestämmer kvantilen $z_{1-\alpha/2}$.

Konfidensintervall

Vi vill att konfidensintervallet är så smalt som möjligt samtidigt som α är så litet som möjligt



Svart motsvarar $\alpha = 0.001$

Grön motsvarar $\alpha = 0.01$

Röd motsvarar $\alpha = 0.05$

Ju lägre α desto mer säker är man.

Konfidensintervall

Vi vill att konfidensintervallet är så smalt som möjligt samtidigt som α är så litet som möjligt

$$I_\mu = \left(\bar{x} - z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}},\ \bar{x} + z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}} \right)$$

Smalare I_μ om:

- n är stort.
- σ^2 är litet.

Konfidensintervall

Hittills har vi antagit att vi vet σ^2, t.ex $N(\mu, 4)$. Då har våra konfidensintervall sett ut så här:

$$I_\mu = \left(\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

Om vi inte vet σ^2, hur gör vi då?

1. Vi skattar σ^2 med $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
2. Eftersom vi har skattat σ^2 med s^2 så har vi mer osäkerhet än om vi visste σ^2. Det betyder att vi måste ändra våra intervall så att vi tar hänsyn till denna osäkerhet.

Detta gör vi genom att byta ut $z_{1-\alpha/2}$ till $t^{(f)}_{1-\alpha/2}$

Konfidensintervall

Vad är då $t_{1-\alpha/2}(f)$?

- $t_{1-\alpha/2}(f)$ står för t-fördelningen. där f är frihetsgrader och α är konfidensgraden.

- Om $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ så är

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \in \underbrace{t_{1-\alpha/2}(n-1)}_{\text{t-kvantil}}$$

- Om vi måste skatta σ^2 så blir intervallet för μ

$$I_\mu = \left(\bar{x} - t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}},\ \bar{x} + t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right).$$

Sammanfattning

Estimator: Ett sätt att använda data till att skatta en parameter. En bra estimator $\hat{\theta}$ av θ ska:

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

$V[\hat{\theta}]$ så liten som möjlig (effektiv).

Konfidensintervall: Ett intervall som täcker det verkliga värdet av μ med sannolikhet $1-\alpha$ (t.ex. 95%, 99%, 99.9%).

Om σ^2 känd så är

$I_\mu = \left(\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$, $z_{1-\alpha/2}$ normal-kvantil

Om σ^2 okänd så är

$I_\mu = \left(\bar{x} - t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}},\ \bar{x} + t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right)$.

$t_{1-\alpha/2}(n-1)$ är t-kvantilen med $n-1$ frihetsgrader. (se s. 281).

Uppgifter

1. Man gör två oberoende mätningar av pH-värdet μ. Som skattning av μ använder man $\bar{x} = (x_1 + x_2)/2$.

Man antar att X_1 och X_2 är oberoende och $N(\mu, \sigma^2)$.

a) Vad är variansen för $\mu^* = \bar{X} = (X_1 + X_2)/2$?

b) Vad är fördelningen för μ^*?

c) Om σ^2 är okänd kan den skattas med s^2. Vad är medelfelet för skattningen μ^*? D.v.s. vad är $d(\mu^*)$?

Uppgifter

2. Man har ett slumpmässigt stickprov

från $N(\mu, 2^2)$:

$44.3 \quad 45.1 \quad 46.1 \quad 45.3$

Ange ett 95%-konfidensintervall för μ.

Hur ser ett 99%-konfidensintervall ut?

3. Man mäter 10 gånger från $N(\mu, \sigma^2)$:

$7.3 \quad 7.2 \quad 7.8 \quad 7.1 \quad 8.0 \quad 6.9 \quad 7.5 \quad 8.1$

$7.7 \quad 7.5.$

Beräkna ett 95% konfidensintervall

för μ.

Sammanfattning

* Nu har vi gjort 60% av kursen.
* Här följer en kort sammanfattning av vad vi hittills har gjort.

Grundläggande begrepp.

Ett slumpmässigt försök: Ett "försök" vars utfall vi inte kan exakt förutsäga.

Utfall: Resultatet av ett slumpmässigt försök.

Utfallsrum: Alla utfall som kan ske.

Händelse: En eller flera utfall (delmängd av utfallsrummet)

Betecknas ofta med $A, B, C, \ldots$

* Sannolikheten för en händelse, A, måste vara

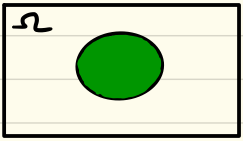
$$\boxed{0 \leq P(A) \leq 1}$$

* Summan av sannolikheterna för hela utfallsrummet (allt som kan ske i försöket) är 1. $(P(\Omega) = 1)$

Sammanfattning

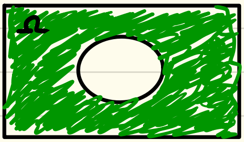
Utfallsrummet:

Händelsen A inträffar



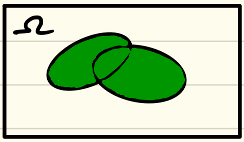
Komplementet till A

$\overline{A}$



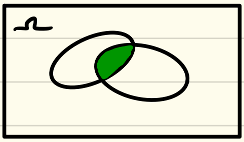
Union. $A \cup B$

A eller B eller båda inträffar.



Snitt. $A \cap B$.

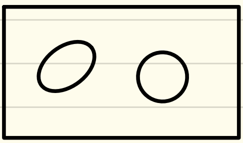
Både A och B inträffar.



A och B är oförenliga.

Kan ej ske samtidigt

$P(A \cap B) = 0$.



Sammanfattning

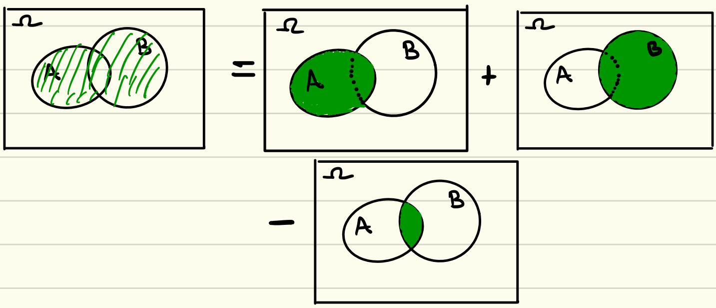
Oberoende: Om A och B är oberoende (ej påverkar varandra) så gäller att:

$$P(A\cap B)=P(A)\cdot P(B)$$

---

* Additionssatsen.

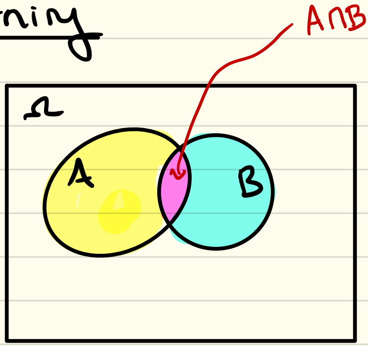
$$P(A\cup B)=P(A)+P(B)-P(A\cap B)$$



Sammanfattning

Betingad sannolikhet

$$P(A|B) = \frac{P(A\cap B)}{P(B)}$$



Bayes sats

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}.}$$

---

* Slumpvariabel X
  - Diskret
  - Kontinuerlig.

* $f(x)$ kallas sannolikhetsfunktion om X är diskret och täthetsfunktion om X är kontinuerlig.

Sammanfattning

Diskreta fallet:

* $f(x)$ beskriver vad sannolikheten är för utfallet x.

$$f(x) = P(X = x)$$

* $F(x)$ kallas fördelningsfunktion och svarar på:

$$F(x) = P(X \leq x) = \sum_{k \leq x} f(k)$$

* Väntevärdet skrivs:

$$\mu = E[X] = \sum_{\text{alla } x} x f(x)$$

* Varians skrivs:

$$V[X] = \sum_{\text{alla } x} (x - \mu)^2 f(x).$$

Sammanfattning

Kontinuerliga fallet:

* $f(x)$ beskriver hur sannolikheten är utsmetad över de oändligt många utfallen X kan anta.

* $F(x)$ kallas fördelningsfunktion och svarar på:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(z)\, dz$$

* Väntevärdet skrivs:

$$\mu = E[X] = \int_{-\infty}^{\infty} x f(x)\, dx$$

* Varians skrivs:

$$V[X] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$$

Sammanfattning

Diskreta fallet:

Standardfördelningar:

* Poisson. Svarar på hur många gånger något händer i ett visst intervall. Tre krav:

  1) λ får inte ändras

  2) I två separerade intervall är händelserna oberoende.

  3) Två händelser får ej ske samtidigt.

$X \in Po(\lambda)$

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, \ldots$$

$$F(x) = \sum_{k \leq x} \frac{\lambda^k}{k!} e^{-\lambda}$$

Sammanfattning

Diskreta fallet:

Standardfördelningar:

* Binomialfördelningen. Svarar på hur mångar något händer på n försök. Kravet är att försöken är oberoende av varandra. Händelsen sker med sannolikhet p.

$X \in Bin(n, p)$

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, \ldots, n$$

$$F(x) = \sum_{k \leq x} \binom{n}{k} p^k (1-p)^{n-k}$$

Sammanfattning

* Kontinuerliga slumpvariabler antar oändligt många värden. Det betyder att $P(X = x) = 0$

* För kontinuerliga slumpvariabler studerar man istället intervall: $P(a \leq X \leq b)$.

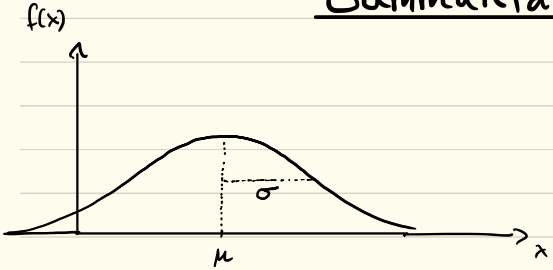
* Fördelningen av X beskrivs av en täthetsfunktion, $f(x) \geq 0$.

* Sannolikheter motsvaras av areor.

* Totala arean under täthetsfunktionen är alltid 1.

$$\boxed{F(a) = P(X \leq a) = \text{"arean till vänster om a"} = \int_{-\infty}^{a} f(x)\,dx.}$$

Sammanfattning

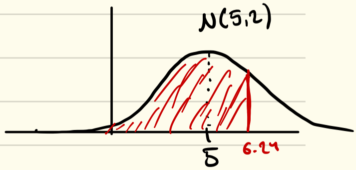


Normalfördelningen:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \text{ alla } x.$$

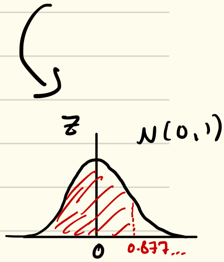
* Beräknas med hjälp av tabell eller dator/miniräknare.

* Standardisering: X



Låt $X \sim N(5,2)$. Beräkna

a) $P(X \leq 6.24)$



Lösning:

Standardisera!

$$P(X \leq 6.24) = P\left(\frac{X-5}{\sqrt{2}} \leq \frac{6.24-5}{\sqrt{2}}\right)$$

$$= P\left(Z \leq \frac{1.24}{\sqrt{2}}\right) = \Phi\Big(\underbrace{\frac{1.24}{\sqrt{2}}}_{\approx 0.877}\Big) \approx 0.81$$

$Z \in N(0,1)$

Tabell.

Sammanfattning

Kvantiler:

$Z \sim N(0,1)$

$P(Z \leq z_{1-\alpha}) = 1-\alpha$, dvs $P(Z > z_{1-\alpha}) = \alpha$.

Linjära transformationer

- $E[aX+b] = aE[X]+b$
- $Var[aX+b] = a^2 Var(X)$
- Låt $X_1, \ldots, X_N$ vara slumpvariabler.

  Då gäller:

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

  och om de är oberoende

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

Sammanfattning

Om $X_1, X_2, \ldots, X_n$ alla är normalfördelade

med $X_i \in N(\mu, \sigma^2)$, $i = 1, \ldots, n$

då är

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

Sammanfattning

Om $X_1, X_2, \ldots, X_n$ alla är normalfördelade

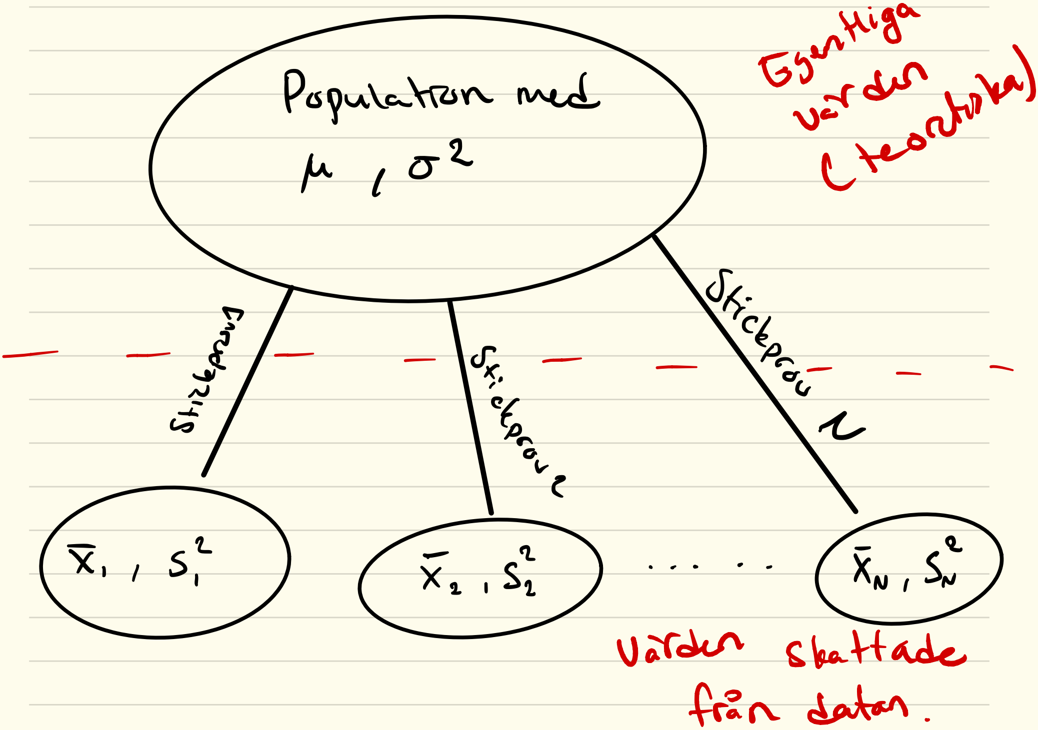
med $X_i \in N(\mu, \sigma^2)$, $i=1,\ldots,n$

då är

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

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

Sammanfattning



Vi vill med hjälp av insamlad data skatta μ, σ^2 så bra som möjligt.

Sammanfattning

* Centrala gränsvärdessatsen (CGS) säger att om vi har oberoende och likafördelade slumpvariabler $X_1, X_2, \ldots, X_N$ med $E[X_i] = \mu$ och $V[X_i] = \sigma^2$

och N är stort, då är

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

Approximativt

* Det betyder att medelvärdet av oberoende och likafördelade slumpvariabler är

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

Sammanfattning

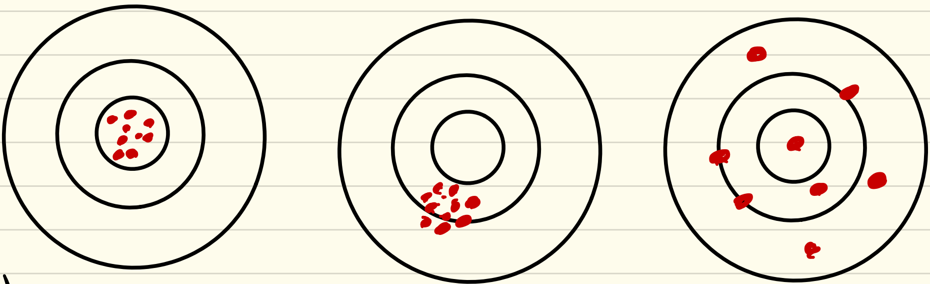
* En estimator är en skattning av en parameter.

  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.



Liten varians (effektiv)
Väntevärdesriktig.

Liten varians (effektiv)
Ej vänte-värdesriktig

Stor varians (ej effektiv)
Väntevärdesriktig.

Sammanfattning

* Skattning av väntevärdet:

$$\bar{X} = \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i \quad , \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Slumpvariabel

Ett värde, data som input.

* Skattning av varians:

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \hat{\mu})^2 , \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Slumpvariabel

Ett värde, data som input.

* Medelfelet:

$$D[\bar{X}] = \sqrt{V[\bar{X}]} = \sqrt{V\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right]}$$

Slumpvariabel

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

$$d[\bar{X}] = \frac{s}{\sqrt{n}}$$

Ett värde, data som input.

Sammanfattning

Konfidensintervall: Ett intervall som täcker det verkliga värdet av μ med sannolikhet $1-\alpha$ (t.ex. 95%, 99%, 99.9%).

Om σ^2 känd så är

$$I_\mu = \left(\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right),$$ $z_{1-\alpha/2}$ normalkvantil

Om σ^2 okänd så är

$$I_\mu = \left(\bar{x} - t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}},\ \bar{x} + t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right),$$

$t_{1-\alpha/2}(n-1)$ är t-kvantilen med $n-1$ frihetsgr. (se s. 281).

Sammanfattning

Konfidensintervall: Ett intervall som täcker det verkliga värdet av μ med sannolikhet $1-\alpha$ (t.ex. 95%, 99%, 99.9%).

Om σ^2 känd så är

$$I_\mu = \left(\bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} , \bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right),$$

$z_{1-\alpha/2}$ normal-kvantil

Om σ^2 okänd så är

$$I_\mu = \left(\bar{x} - t_{1-\alpha/2}(n-1) \frac{s}{\sqrt{n}} , \bar{x} + t_{1-\alpha/2}(n-1) \frac{s}{\sqrt{n}} \right),$$

$t_{1-\alpha/2}(n-1)$ är t-kvantilen med $n-1$ frihetsgr. (se s. 281).

GLAD PÅSK