

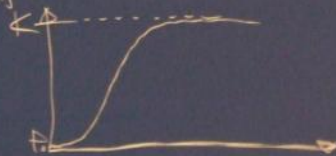
Modellera verkligheten med matematiska modeller.

Deterministiska modeller

Givet in-parametrar, talar dessa modeller om precis hur ett utfall blir.

Ex Logistisk populationstillväxt

$$P(t) = \frac{K P_0 e^t}{K + P_0 (e^t - 1)}$$



Dra slutsatser om den större population som data kommer ifrån

Verkligheten är sällan deterministisk!
"förutsägbar"

Statistiska (stokastisk/kannotikhetsteoretiska) modeller

Inkluderar osäkerhet/slump i det matematiska sammanhanget.

$$\text{Ex } P(t) = \frac{K P_0 e^t}{K + P_0 (e^t - 1)} + \epsilon$$

↖ osäkerhet/slump

Vi vill bedömma graden av osäkerhet/
beteendet hos osäkerheten.
Besvaras av sannolikheter!

Mha sannolikheter kan vi:

- Formulera modeller för den observerade verkligheten (= data)
- Dra slutsatser om den större population som data kommer ifrån



Population

- Tala om tillförlitligheten i våra slutsatser

Vad säga om ^{en}sannolikhet p

- $0 \leq p \leq 1$, ger oss chansen att ett (fysikaliskt) experiment lyckas

- $P \approx 1 \Rightarrow$ troligt att händelsen med slh P inträffar (ej garanterat)
sannolikhet
- $P \approx 0 \Rightarrow$ —||— ej inträffar
(ej garanterat)
- $P \approx \frac{1}{2} \Rightarrow$ lika troligt som otroligt
att händelsen inträffar
- slh kan uttryckas mha procent.
 $P = 0.7 \Leftrightarrow 70\%$ chans att inträffa

Kalla slh för en händelse A
för $P(A)$.

$\uparrow$
 $= \{ \text{Får krona med en slant} \}$

Hur tillskriva slh till händelsen
 A ?

Relativ frekvens:

Antag att vi kan repetera
experimentet n ggr.

$$\Rightarrow P(A) \approx \frac{\text{Antal ggr } A \text{ inträffar}}{n}$$

Ex: Singla slant $A = \{\text{får krona}\}$

Klassisk formel

Om man kan antaga att alla möjliga utfall i ett experiment har samma slk att inträffa.

$$P(A) = \frac{\text{Antal gynnsamma fall för } A}{\text{Antal möjliga utfall}}$$

Ex:

Både far och mor bär på gener för bruna och blåa ögon.

$A = \{\text{Deras barn får bruna ögon}\}$

Vad är $P(A)$?

Barnet får en gen från varje förälder.

Möjliga uppsättningar:

(brun, blå), (brun, brun), (blå, brun), (blå, blå)
↑ ↑
far mor

V varje förälder lika sannolik att föra vidare brun som blå, alltså är alla utfall lika sannolika att inträffa

Brun är dominant gen (dvs om vi stöter på ngn brun, får barnet bruna ögon).

$$P(A) = \frac{3}{4} = 0.75 = 75\%$$

utfall som inneh. åtminstone en brun
totala # utfall
antal

Def (utfallsrum)

Mängden, S , av alla möjliga utfall kallas utfallsrum.

Ett element (utfall) $x \in S$ kallas samplepunkt.

Ex: I tidigare exempel har vi

$$S = \{(\text{brun}, \text{brun}), (\text{brun}, \text{blå}), (\text{blå}, \text{brun}), (\text{blå}, \text{blå})\}$$

$\overset{||}{x_1} \quad \quad \overset{||}{x_2} \quad \quad \overset{||}{x_3} \quad \quad \overset{||}{x_4}$

(Kom ihåg: andra mängder $\mathbb{Z} = \{\text{alla heltal}\}$

$\mathbb{R} = \{\text{reella talen}\}$

$C = \{\text{Alla cyklar i Sverige}\}$

Def

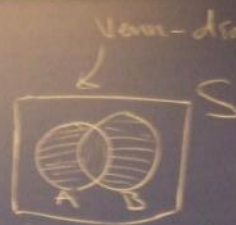
En delmängd $A \subseteq S$ kallas för
en händelse.

I exemplet: $A = \{\text{Barnet får bruna ögon}\}$
 $= \{x_1, x_2, x_3\}$

För händelser A och B:

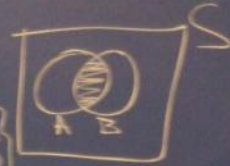
Unionen av A och B:

$$A \cup B = \{x \in S : x \in A \text{ och/eller } x \in B\}$$



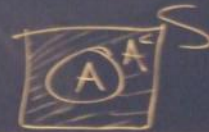
Snittet mellan A och B

$$A \cap B = \{x \in S : x \in A \text{ och samtidigt } x \in B\}$$



Komplementet till A:

$$A^c = \{x \in S : x \notin A\}$$



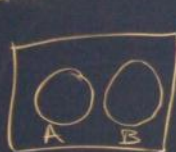
I boken används A' istället för A^c

I boken används A' istället för A^c

Def (Disjunkta ("mutually exclusive") mängder)

Händelserna A och B är disjunkta om

$$A \cap B = \emptyset$$



$\emptyset$ kallas tomma mängden, $\emptyset = \{\}$

Ex Två reglage väljs ut och man
man kollar om var och en uppfyller
tillverkningsspecifikationerna.
(ja = j, nej = n).

$$S = \{jn, nj, nn, jj\}$$

$$A = \{\text{åtminstone en fungerar}\} = \{jn, nj, jj\}$$

$$B = \{\text{ingen fungerar}\} = \{nn\}$$

$$A \cap B = \emptyset$$

$$A^c = \{nn\} = B$$

$$A \cup B = A \cup A^c = S$$



Def (permutation)

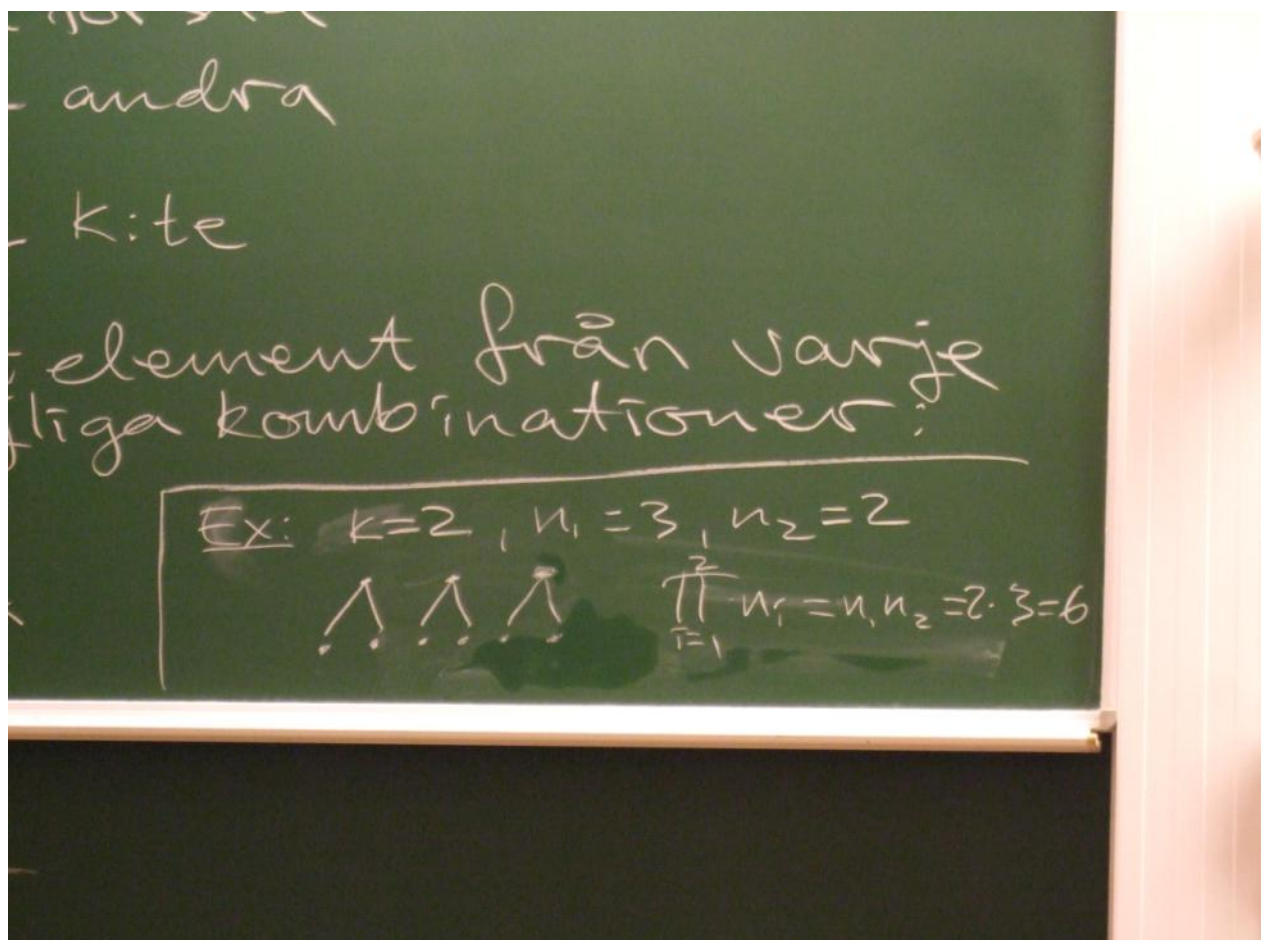
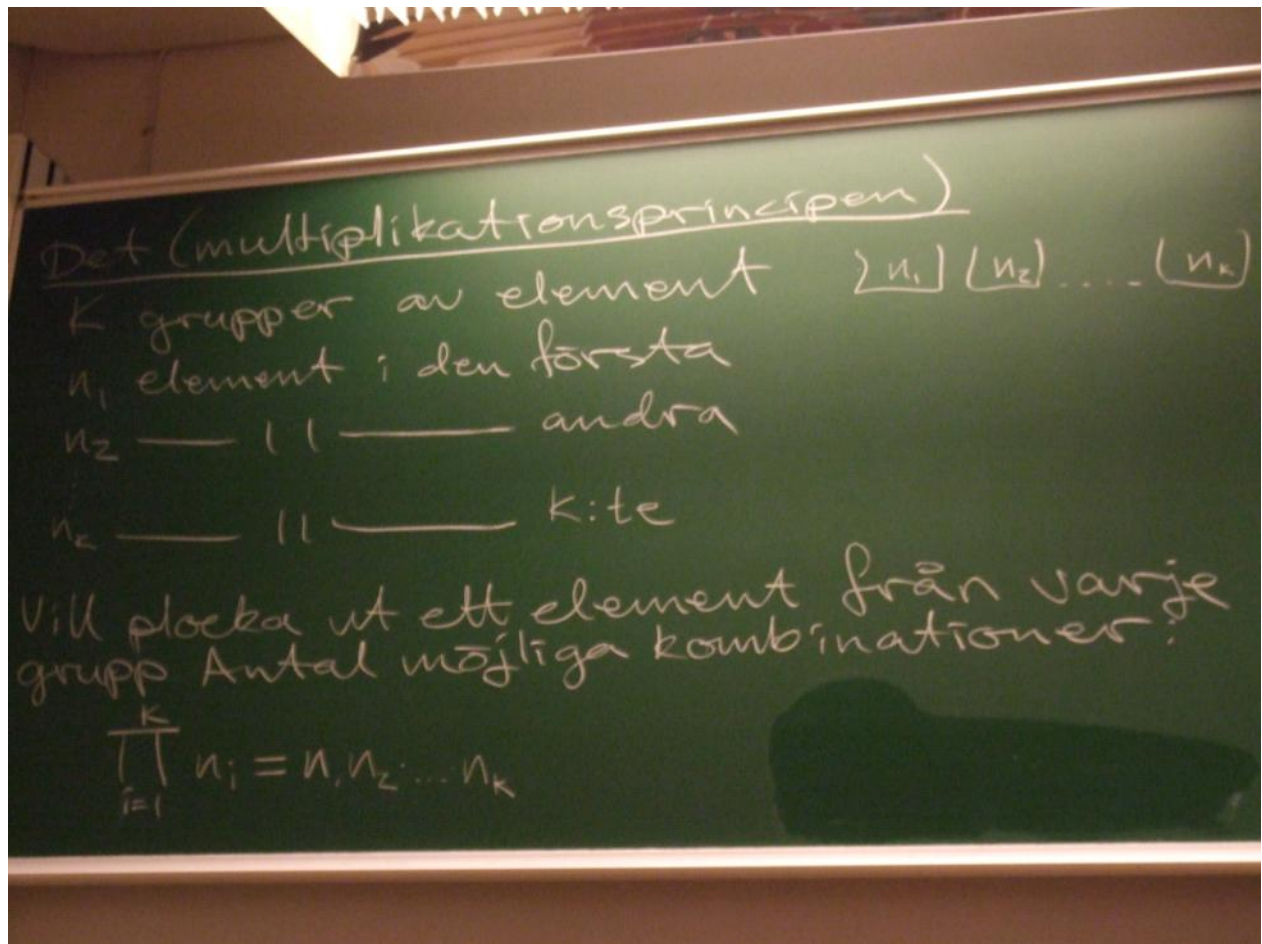
En permutation är ett ordnat
urval av elementen i en mängd

Def (Kombination)

En kombination är ett urval av element där man ej tar hänsyn till ordningen (letta formar en mängd).

Ex: 123 och 231 är ej samma permutation men de utgör samma kombination.

$$\subset \{1, 2, 3\}$$



grupp. Antal möjliga

$$\prod_{i=1}^k n_i = n_1 n_2 \dots n_k$$

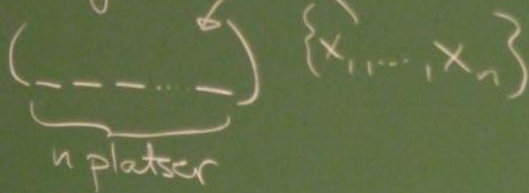
Ex: $k=2, n_1=3, n_2=2$
 $\prod_{i=1}^2 n_i = n_1 n_2 = 2 \cdot 3 = 6$

Def. (fakultet)

$$n! = n(n-1) \dots 2 \cdot 1 \quad \text{kallas 'n-fakultet'}$$

$$0! \equiv 1$$

$n!$ ger antal permutationer av n element.



Sats

Antal permutationer av längd r som kan väljas ut från n distinkta element ges av $nPr = \frac{n!}{(n-r)!}$

Bevis

Det första av de r elementen kan väljas på n sätt. Det finns nu $n-1$ tillgängliga element som ger oss element 2 osv.

Element nr.: 1 2 3 ... r

Möjliga val: n $n-1$ $n-2$... x

Vad är x ? Notera att # möjliga val ges av $n - (\text{objektnr.} - 1)$

$$\Rightarrow x = n - (r - 1) = n - r + 1$$

Enl. mult. principen får vi nu att

$$\begin{aligned} nPr &= n(n-1) \cdots (n-r+1) = \\ &= \frac{n(n-1) \cdots (n-r+1) \cdot (n-r)(n-r-1) \cdots 1}{(n-r)(n-r-1) \cdots 1} \\ &= \frac{n!}{(n-r)!} \end{aligned}$$



$$n = 4$$

$$n_1 = 2$$

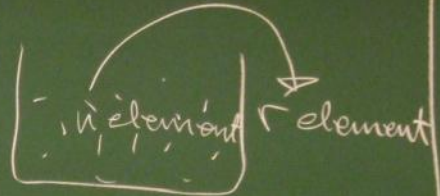
$$n_2 = 2$$

$$\frac{4!}{2!2!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 2 \cdot 1} = 6$$

Möjliga val: $n, n-1, n-2, \dots$

Sats (delmängder) (storlek)
kombinationer av längd r som kan väljas ut bland n element.

$$nCr = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$



Beweis

Vet att vi kan skapa nPr olika perm. av storl. r . De perm. som består av samma element representerar samma kombination.

Vi delar nPr med # perm. vi kan skapa av r element

$$\Rightarrow nCr = \frac{nPr}{r!} = \frac{n!}{r!(n-r)!}$$

Sats

Givet n_1 (^{element} objekt) av sort 1, n_2 av sort 2, $\dots$, n_k av sort k där $\sum_{i=1}^k n_i = n_1 + \dots + n_k = n$. # ^{särskilbara} permutationer av de n objekter ges av $\frac{n!}{n_1! \cdot \dots \cdot n_k!}$

Sth-axiomen (kom ihåg: S utfallsrum)

1) $P(S) = 1$, dvs "något händer" = 1

2) $P(A) \geq 0$ för varje händelse $A \subseteq S$

3) $A_1, A_2, \dots$ är parvis disjunkta ($A_i \cap A_j = \emptyset, i \neq j$)

$$\Rightarrow P\left(\bigcup_{i=1}^{\infty} A_i\right) = P(A_1 \cup A_2 \cup \dots) = \sum_{i=1}^{\infty} P(A_i) = P(A_1) + P(A_2) + \dots$$

Sats:

$$P(\emptyset) = 0$$

Beweis

$$S = S \cup \emptyset \quad \text{samtidigt} \quad S \cap \emptyset = \emptyset$$

$$\Rightarrow \stackrel{\text{ax 1)}}{=} P(S) = P(S \cup \emptyset) \stackrel{\text{ax 3)}}{=} P(S) + P(\emptyset)$$

$$\stackrel{\text{ax 1)}}{=} 1 + P(\emptyset) \quad \therefore P(\emptyset) = 0$$

Ex Blodgrupper $S = \{A, AB, O, B\}$

$$P(A) = 0.41$$

$$P(B) = 0.09$$

$$P(AB) = 0.04$$

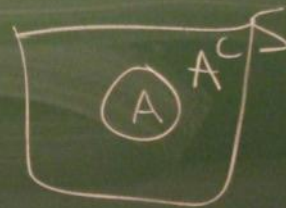
$$P(O) = 0.46$$

A, B, AB, O är ömsj. händelser
d.v.s. man kan ej ha 2
olika blodgr.

$$\begin{aligned} P(S) &= P(A) + P(B) + P(AB) + P(O) \\ &= 0.41 + 0.09 + 0.04 + 0.46 \\ &= 1 \end{aligned}$$

Sats

$$P(A^c) = 1 - P(A)$$



Beweis

$$S = A \cup A^c \quad \text{samnt} \quad A \cap A^c = \emptyset \quad (\text{d.v.s. disjunta})$$

$$1 \stackrel{\text{ax 1}}{=} P(S) \stackrel{\text{ax 3}}{=} P(A \cup A^c) \stackrel{\text{ax 3}}{=} P(A) + P(A^c)$$

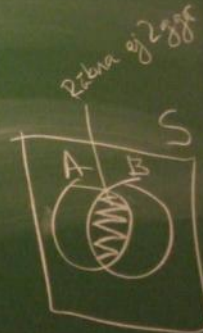
$$\Rightarrow P(A^c) = 1 - P(A) \quad \square$$

Sats

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

Beweis

Uppg. 12, kap 2



Ex Singling av 2 mynt.

H = krona ("head")

T = klave ("tail")

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

Antag alla utfall samma slh (dvs $\frac{1}{4}$)

$$A = \{\text{första krona}\} = \{(H, H), (H, T)\}$$

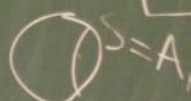
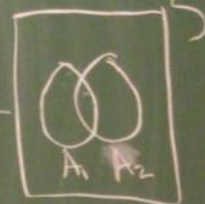
$$B = \{\text{andra krona}\} = \{(H, H), (T, H)\}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{2} + \frac{1}{2} - P(\overbrace{(H, H)}^{S, A}) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}$$

Def (Betingad slh)

A_1 och A_2 är händelser ^{s.a.} sådana att $P(A_1) \neq 0$. Den betingade sannolikheten för A_2 givet A_1 definieras av

$$P(A_2 | A_1) := \frac{P(A_2 \cap A_1)}{P(A_1)}$$



Ex Slantsingling (som innan)

$$A = \{\text{två kronor}\} = \{(H, H)\}$$

$$B = \{\text{första krona}\} = \{(H, H), (H, T)\}$$

Betingade slh att få två kronor givet krona i första singlingen?

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{(H, H)\})}{P(\{(H, H), (H, T)\})} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}$$

$$P(B) = \frac{P(\{(H, H), (H, T)\})}{P(\{(H, H), (H, T)\})} = \frac{\frac{1}{2}}{\frac{1}{2}} = 1$$

Multiplikationsregeln

Om $P(A_1) \neq 0$ får vi

$$P(A_1 \cap A_2) = P(A_2 | A_1) P(A_1)$$

För vissa händelser gäller $P(A|B) = P(A)$
dvs inträffandet av B påverkar ej
slh för A.

Det (oberoende)

Händelser A och B kallas oberoende
om $P(A \cap B) = P(A)P(B)$

Ex I en burk ligger en röd (R), en
gul (G) och en vit (V) boll.

Experiment: Plocka ut 2 bollar
(utan återläggning).

$S = \{(R, G), (R, V), (G, V), (V, R), (V, G), (G, R)\}$

Dessa utfall har samma slh.

$A = \{\text{första bollen gul}\}$

$B = \{\text{andra ———}\}$

$$P(HH|HT) = \frac{1}{2} = \frac{1}{2}$$

Intuition: A och B ej oberoende
B kan omöjligt inträffa
om A inträffat.

$$P(A \cap B) = P(\emptyset) = 0 \neq \underbrace{P(A)}_{\neq 0} \underbrace{P(B)}_{\neq 0}$$

I fortsättningen är frelagsförel. 8-10

Kom ihåg: A och B oberoende $\Leftrightarrow P(A \cap B) = P(A)P(B)$

Sats

Låt A och B vara händelser.

Om $P(A) \neq 0$ gäller att

A och B oberoende $\Leftrightarrow P(B|A) = P(B)$

Beweis

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \stackrel{\text{ob.}}{=} \frac{P(A)P(B)}{P(A)} = P(B)$$

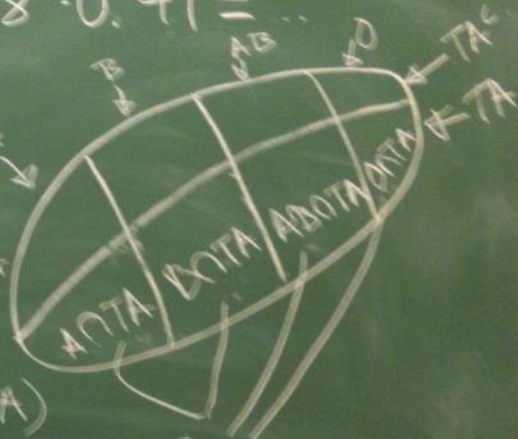
Men $P(A \cap TA) = P(TA|A)P(A)$

$= 0.88 \cdot 0.41 =$

$TA = (A \cap TA) \cup (B \cap TA) \cup (AB \cap TA) \cup (O \cap TA)$

Union av parvis
disjunkta händelser

$P(TA) \stackrel{\text{ax 3}}{=} P(A \cap TA) + P(B \cap TA) + P(AB \cap TA) + P(O \cap TA)$



Parvis disjunkt

$= P(A)P(TA|A) + P(B)P(TA|B) + P(AB)P(TA|AB) + P(O)P(TA|O)$

$P(A|TA) = \frac{P(A \cap TA)}{P(TA)}$ — kommer g till dessa!

$= 0.88 \cdot 0.41 + 0.04 \cdot 0.09 + 0.1 \cdot 0.04 + 0.04 \cdot 0.46$

$\therefore P(A|TA) = \frac{P(TA|A)P(A)}{P(TA|A)P(A) + P(TA|B)P(B) + P(TA|AB)P(AB) + P(TA|O)P(O)}$

$= \dots = 0.92$

Bayes sats i exempel: $ABABD$

$A_1, A_2, \dots, A_n$ parvis disjunkta ($A_i \cap A_j = \emptyset$ if $i \neq j$) händelser s.a. $\bigcup_{i=1}^n A_i = S$

(dvs $A_1, \dots, A_n$ utgör en partition av utfallsrummet)

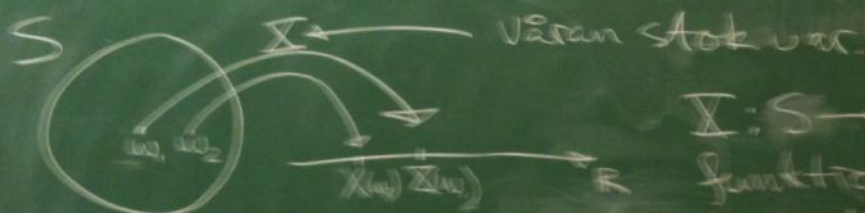
Låt B vara en händelse s.a. $P(B) \neq 0$

För en godtycklig händelse $A_j, j=1, \dots, n$, gäller nu

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^n P(B|A_i)P(A_i)}$$

Kap 3

En stokastisk variabel (slumpvariabel) beskriver ett slumpmässigt förlopp.



$X: S \rightarrow R$ är en funktion som går från utfallsrummet till R

Kap 3

En stokastisk variabel (slumpvariabel) beskriver ett slumpmässigt utfall.



$X: S \rightarrow R$ är en funktion som går från utfallsrummet till R .

"Slumpen" drar $w \in S$ och $X(w) \in Z$ skickar detta till R .
 $X(w) \in Z$ ← Mängden av möjliga värden på X

$$P(X=k) = P(\underbrace{\{w \in S : X(w)=k\}}_{\text{Är en händelse}})$$

Är en händelse

Ex Kasta 2 tärningar. Intresserad av händelserna

$A = \{\text{första tärningen } 1a\}$

$B = \{\text{summan är } 6\}$

I stället betraktar vi de stok. var.
 $X = \# \text{ prickar på första tärningen}$
 $Y = \text{summan av de två kasten}$

$$\Rightarrow P(A) = P(X=1)$$

$$P(B) = P(Y=6)$$

Def (Diskret stok. var)

X kallas diskret stok. var. om de möjliga värden den kan antaga är ändligt till antal ($S_X = \{x_1, \dots, x_n\}$) eller oändligt uppräkneligt ($S_X = \{x_1, x_2, \dots\}$). $\square$

Def (tätthetsfunktion)

X diskret stok. var.

Dess täthetsfunktion, f_X , ges av

$$f_X(k) = P(X=k), \quad k \in \mathbb{R}$$

(Notera att $f(k) = 0$ där $k \notin S_X \subseteq \mathbb{R}$)

Tätheten uppfyller.

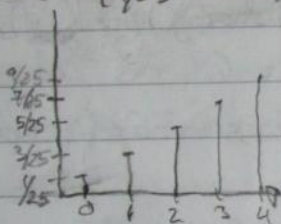
$$1) f(x) \geq 0 \text{ för alla } x \in \mathbb{R}$$

$$2) \sum_{k \in S_X} f(k) = 1$$

Ex: Låt $f(x) = \frac{2x+1}{25}$, $x=0,1,2,3,4$

Är $f(x)$ en täthet?

$$f(x) = \begin{cases} 1/25 & \text{om } x=0 \\ 3/25 & \text{om } x=1 \\ 5/25 & \text{om } x=2 \\ 7/25 & \text{om } x=3 \\ 9/25 & \text{om } x=4 \end{cases} \Rightarrow f(x) \geq 0 \text{ för alla } x=0,1,2,3,4$$



$$\sum_{x=0}^4 f(x) = \frac{1}{25} (1+3+5+7+9) = \frac{25}{25} = 1$$

Def: Låt X vara en diskret stok. var. med täthet $f_X(x)$.

Fördelningsfunktionen för X ges av

$$F(x) = P(X \leq x) = \sum_{k \in \{x \in S_X: x_i \leq x\}} f_X(k), \quad x \in \mathbb{R}$$

Ex: $f(x) = \frac{2x+1}{25}$, $x=0,1,2,3,4$

$$F(x) = \begin{cases} 0 & \text{om } x < 0 \\ 1/25 & \text{om } 0 \leq x < 1 \\ 4/25 & \text{om } 1 \leq x < 2 \\ 9/25 & \text{om } 2 \leq x < 3 \\ 16/25 & \text{om } 3 \leq x < 4 \\ 1 & \text{om } 4 \leq x < \infty \end{cases}$$

Täthetsfunktionen (och fördelningsfunktionen) säger allt om hur de tal X produceras fördelar sig. Kopplat till varje stok. var. finns pararna som talar om övergripande beteenden. Vi

Vi betraktar här "väntevärde", "varians" och "standardavvikelse" vilka alla kan hittas i en täthet

Kasta en rättvis tärning n gånger
och räkna ut medelvärdet av alla
 n kast: $\frac{1}{n} \sum_{i=1}^n X_i$

Vad är det teoretiska medelvärde
som vi får då $n \rightarrow \infty$, dvs då antalet kast
blir oändligt många?

Detta ges av väntevärdet:

Def Låt X vara en diskret stok. var. med
tätthet $f_X(x)$.

X har väntevärde

$$E[X] = \sum_{x \in S_X} x f_X(x)$$

Sist: ^{Diskret} Stok. var. X har täthet

$$f_X(x) = P(X=x), x \in \mathbb{R}$$

och fördelningsfunktion

$$F_X(x) = P(X \leq x) = \sum_{\substack{y \in S_X \\ y \leq x}} f_X(y)$$

och väntevärde

$$E[X] = \sum_{x \in S_X} x f_X(x) \quad (\text{teoretiskt medelvärde})$$



Definiera den nya stok. var. $Y = H(X)$
Vi får väntevärdet Lenfunkt.

$$E[Y] = E[H(X)] = \sum_{x \in S_X} H(x) f_X(x)$$

$S_X = \text{mögliga värden på } X$

Ibland betecknas $E[X]$ med μ_X eller med μ .

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{n \rightarrow \infty} E[X]$$

Sats Låt X och Y vara stok. var.
och låt $c \in \mathbb{R}$. Då gäller:

Sats Låt X och Y vara stok. var.
och låt $c \in \mathbb{R}$. Då gäller;

1) $E[c] = c$

2) $E[cX] = cE[X]$

3) $E[X+Y] = E[X] + E[Y]$

Ex $E[X] = 7$ $E[Y] = 5$

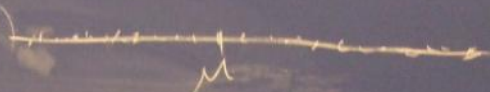
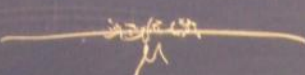
$$E[4X - 2Y + 6] \stackrel{3)}{=} E[4X] + E[-2Y] + E[6]$$

$$\stackrel{1,2)}{=} 4E[X] - 2E[Y] + 6 = 4 \cdot 7 - 2 \cdot 5 + 6 = 24$$

$\mu = E[X]$ talar om var X i medeltal
ligger, dvs μ är en lägesparameter.

Vi behöver ett mått på hur X
sprider sig!

Lite spridn.



Def (Varians)

X stok. var. med väntevärde μ_X
Variansen av X ges av

$$\sigma^2 = \sigma_X^2 = \text{Var}(X) = E[(X - \mu_X)^2]$$

Tre olika skrivsätt för

Kan tolka. Den är också ett
spridningsmått!

Sats (Beräkningsformel för $\text{Var}(X)$)

$$\sigma^2 = \sigma_X^2 = \text{Var}(X) = E[X^2] - \underbrace{E[X]^2}_{=\mu_X^2}$$

Beweis

$$\begin{aligned} \sigma^2 = \sigma_X^2 &= E[(X - \mu_X)^2] = E[X^2 - 2\overset{\text{konstant}}{\mu_X}X + \mu_X^2] \\ &= E[X^2] - 2\mu_X \underbrace{E[X]}_{=\mu_X} + \mu_X^2 \\ &= E[X^2] - \mu_X^2 \end{aligned}$$

$$\text{Var}(X) = E[X^2] - E[X]^2 = \frac{230}{25} - \left(\frac{14}{5}\right)^2$$

Ex X är en stok var. med
täthet $f_X(x) = \frac{2x+1}{25}$, $x \in S_X = \{0, 1, 2, 3, 4\}$

$$E[X] = \sum_{x=0}^4 x f_X(x) = 0 \cdot \frac{2 \cdot 0 + 1}{25} + 1 \cdot \frac{2 \cdot 1 + 1}{25} + 2 \cdot \frac{2 \cdot 2 + 1}{25} \\ + 3 \cdot \frac{2 \cdot 3 + 1}{25} + 4 \cdot \frac{2 \cdot 4 + 1}{25} = \frac{14}{5}$$

$$E[X^2] = \sum_{x=0}^4 x^2 f_X(x) = \sum_{x=0}^4 x^2 \frac{2 \cdot x + 1}{25} = \\ = 0^2 \cdot \frac{2 \cdot 0 + 1}{25} + 1^2 \cdot \frac{2 \cdot 1 + 1}{25} + \dots + 4^2 \cdot \frac{2 \cdot 4 + 1}{25} = \frac{230}{25}$$

Det (Standardavvikelse)
 X stok. var. med varians

$$\text{Var}(X) = E[X^2] - E[X]^2 = \frac{230}{25} - \left(\frac{14}{5}\right)^2 \\ = \frac{230 - 196}{25} = \frac{34}{25}$$

Skumt! Variansen beräknas
kvadrerad avvikelse!

V: tittar på standardavvikelsen!

Def (Standardavvikelse)

X stok. var. med varians
 $\sigma^2 = \text{Var}(X)$. X har
Standardavvikelse
 $\sigma = \sqrt{\text{Var}(X)}$

(σ har en skala man (oftast)
kan tolka. Den är också ett
spridningsmått!)

σ har samma skala som X .

Högt $\sigma \Rightarrow$ stor spridning, dvs ganska
osäkerhetsbar!

Sats (Räkneregler för varianser)

X och Y stok. var., $c \in \mathbb{R}$.

1) $\text{Var}(c) = 0$

2) $\text{Var}(cX) = c^2 \text{Var}(X)$

3) Om X och Y oberoende, $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$

(X och Y oberoende om X inte påverkar utfallet av Y och vice versa)

Ex: $\sigma_x^2 = 9$ $\sigma_y^2 = 3$, X och Y ober.
 $\text{Var}(4X - 2Y + 6) \stackrel{20}{=} \text{Var}(4X) + \text{Var}(-2Y) + \text{Var}(6)$
 $\stackrel{20}{=} 4^2 \text{Var}(X) + (-2)^2 \text{Var}(Y)$
 $= 16 \cdot 9 + 4 \cdot 3 = 156$

θ är en parameter-vektor som styr fördelningfam.

Olika diskreta fördelningar

Varje stok. var. X tillhör en fördelningsfamilj $F(\theta)$, skrivs $X \sim F(\theta)$.

Man säger " X är $F(\theta)$ -fördelad".

Fördelningsfamiljen/typen $F(\theta)$ talar om X 's sannolikheter (och därmed andra saker så som fördeln. funkt., väntevärde, varians).

θ är en parameter-vektor som styr fördelningfam.

Bernoulli-fördelningen, $X \sim \text{Ber}(p)$

Om $X \sim \text{Ber}(p)$ då gäller att

$$f_X(1) = P(X=1) = p \quad \text{och} \quad f_X(0) = P(X=0) = 1-p$$

$$0 < p < 1 \quad S_X = \{0, 1\}$$

Tänk slantsingling

1 $\Leftrightarrow$ lyckat försök (ex. krona)

0 $\Leftrightarrow$ misslyckat försök (ex. klave)

$$E[X] = \sum_{x \in S_X} x f_X(x) = \sum_{x=0}^1 x f_X(x) = 0 \cdot \overbrace{P(X=0)}^{=1-p} + 1 \cdot \overbrace{P(X=1)}^{=p}$$

$$= p$$

$$\text{Var}(X) = \underbrace{E[X^2]}_{=0^2 P(X=0) + 1^2 P(X=1)} - \underbrace{E[X]^2}_{=p^2} = p - p^2 = p(1-p)$$

Geometrisk fördelning, $X \sim \text{Geo}(p)$

$X = \# \text{ Bernoulliförsök}$

Geometrisk fördelning, $X \sim \text{Geo}(p)$

$X = \#$ Bernoulli-försök, med parameter p ,
som krävs för ett lyckat försök

Antal försök tills vi får krona för första gången.

$$S_X = \{1, 2, 3, \dots\} = \mathbb{Z}_+$$

$$f_X(k) = P(X=k) = p(1-p)^{k-1}, \quad k \in S_X, 0 < p < 1$$

Ex $P(\text{krona för första gånge efter } 3 \text{ försök då } p=0.5) = 0.5(1-0.5)^2 =$

Om $X \sim \text{Geo}(p)$, vad är då
 $E[X]$ och $\text{Var}(X)$?

Def (Momentgenererande funkt.)

X stok. var. med täthet $f_X(x)$

X har momentgenererande
funktion (mgf)

$$m_X(t) = E[e^{tX}]$$

(givet att $E[e^{tX}] < \infty$ för alla $t \in (-h, h)$, mgf $t > 0$)

Sats

Låt $m_X(t)$ vara mgf för X .

Då gäller att

$$\left. \frac{d^k m_X(t)}{dt^k} \right|_{t=0} = E[X^k]$$

mgf för $X \sim \text{Geo}(p)$:

Låt $q = 1 - p$.

$$m_X(t) = E[e^{tX}] = \sum_{k=1}^{\infty} e^{tk} f_X(k)$$

$$= \sum_{k=1}^{\infty} e^{tk} p (1-p)^{k-1} \frac{q}{p} = \frac{p}{q} \sum_{k=1}^{\infty} (e^t q)^k$$

$$= \frac{p}{q} \frac{q e^t}{1 - q e^t} = \frac{p e^t}{1 - (1-p)e^t}$$

geometrisk summa

$$E_{nl} = \alpha \lambda_s$$

$$E[X] = \left. \frac{d m_x(t)}{dt} \right|_{t=0} = \left. \frac{d}{dt} \left(\frac{e^t}{1 - qe^t} \right) \right|_{t=0} =$$

$$= \dots = \frac{1}{p}$$

$$E[X^2] = \left. \frac{d^2}{dt^2} \left(\frac{pe^t}{1 - qe^t} \right) \right|_{t=0} = \dots = \frac{1+q}{p^2}$$

mq

$$\Rightarrow \text{Var}(X) = E[X^2] - E[X]^2 = \frac{1-p}{p^2}$$

Sist: Diskreta fördelningarBer(p): $P(X=1)=p$, $P(X=0)=1-p$, $0 < p < 1$ Geo(p): $f_X(k) = (1-p)^{k-1}p$, $k=1,2,\dots$ mgt: $m_X(t) = E[e^{tX}]$, $E[X^k] = \left. \frac{d^k m_X(t)}{dt^k} \right|_{t=0}$ Om $X \sim \text{Geo}(p)$ så gäller att $F_X(x) = 1 - q^{(x)}$
(enl. uppg 26)Binomialfördelningen, $X \sim \text{Bin}(n, p)$ $Y_1, \dots, Y_n$ obero. och Ber(p)-fördelade, $0 < p < 1$,
dvs vi har n obero. slantsinglingar $X = \sum_{i=1}^n Y_i$ följer en Bin(n, p)-fördelning. $X = \#$ lyckade försök vid n försök.

$$f_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k=0,1,2,\dots,n$$

 $\overset{||}{P(X=k)}$

|||||

- $\frac{1}{n}$ stplöja ut
k st. som tar
värdet 1

Sats

Låt $X \sim \text{Bin}(n, p)$. Då gäller att

$$M_X(t) = ((1-p) + pe^t)^n$$

$$E[X] = np \quad \text{Var}(X) = np(1-p)$$

$$E[X(X-1)] = n(n-1)p^2$$

Fördelningsfunktionen ges av

$$F_X(t) = \sum_{k=0}^{\lfloor t \rfloor} f_X(k) = \sum_{k=0}^{\lfloor t \rfloor} \binom{n}{k} p^k (1-p)^{n-k}$$

$\lfloor t \rfloor$ = avrunda till närmaste mindre heltal

$$\lfloor 3.7 \rfloor = 3$$

Värden på $F_X(t)$ ges i tabell I i Appendix A.

Ex

X = # signaler som uppfattas ordentligt bland 10 sända signaler.

Antag att $p = 0.5$, dvs $X \sim \text{Bin}(n, p)$

då gäller

$$E[X] = np = 10 \cdot 0,5 = 5 \quad (\text{hälften för-} \\ \text{väntas mot-} \\ \text{tagas korrekt})$$

$$\text{Var}(X) = np(1-p) = 10 \cdot 0,5 \cdot 0,5 = 2,5$$

$$\sigma = \sqrt{\text{Var}(X)} = \sqrt{2,5} = 1,58$$

$$F(7) = \sum_{k=0}^7 \binom{10}{k} \underbrace{0,5^k 0,5^{n-k}}_{=0,5^n} \stackrel{\text{Tabell I}}{=} 0,9453$$

Hypergeometrisk fördeln. $X \sim \text{Hyp}(N, n, r)$



Urn med N st. bollar
 r st. är röda
 $N-r$ st. är blåa

Drag på måfå ut n st bollar
utan återläggning!

$X = \#$ röda bollar vi plockar ut.

sätt att välja k av de r st. röda: $\binom{r}{k}$

sätt kvar att välja $n-k$ blåa: $\binom{N-r}{n-k}$

$\binom{N}{n}$ sätt att välja de n bollarna

Enl. multiplikationsregeln:

$$P_X(k) = P(X=k) = \frac{\binom{r}{k} \binom{N-r}{n-k}}{\binom{N}{n}}$$

$$\max(0, n-(N-r)) \leq k \leq \min(n, r)$$

$$\begin{aligned} E[X] &= n \frac{r}{N} \\ \text{Var}(X) &= n \frac{r}{N} \frac{N-r}{N} \frac{N-n}{N-1} \end{aligned}$$

Ex Transport av motorblock sker i partier av 20 st. Man väljer slumpmässigt ut 3 block för test och godtar partiet om alla 3 är ok. Antag att ett givet parti faktiskt innehåller 5 felproducerade block. Låt X = # felaktiga block som påträffas.

$$\begin{aligned} N &= 20 \\ n &= 3 \\ r &= 5 \end{aligned} \quad P_X(k) = \frac{\binom{5}{k} \binom{20-5}{3-k}}{\binom{20}{3}} \quad 0 \leq k \leq 3$$

$$E[X] = n \frac{r}{N} = 3 \frac{5}{20} = 0.75$$

$$\text{Var}(X) = 3 \frac{5}{20} \frac{15}{20} \frac{17}{19} = \frac{153}{304} \approx 0.5$$

Om n är litet i förhållande till N kan denna fördeln. approximeras med Binomial-fördelningen.

Tumregel: Om $\frac{n}{N} \leq 0.05$ kan vi approximera $\text{Hyp}(N, n, r)$ med $\text{Bin}(n, \frac{r}{N})$

Poissonfördelningen, $X \sim \text{Poi}(\lambda)$

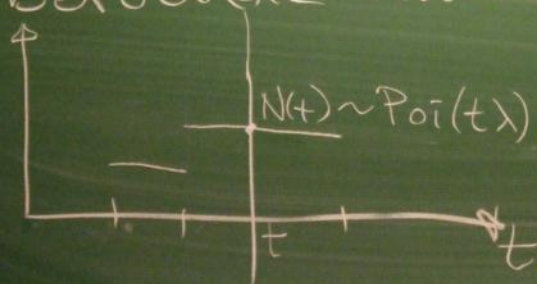
$X \sim \text{Poi}(\lambda), \lambda > 0$, har täthet

$$f_X(k) = \frac{e^{-\lambda} \lambda^k}{k!}, \quad k = 0, 1, 2, \dots$$

$$E[X] = \lambda \quad m_X(t) = e^{\lambda(e^t - 1)}$$

$$\text{Var}(X) = \lambda$$

X beskriver antal händelser som inträffar under en given tidsperiod då dessa händelser sker med en given genomsnittsfrekvens. Händelserna sker med oberoende tidsmellanrum.



Ex: Vid en viss vägkorsning kör det i genomsnitt förbi 5 bilar per halvtimme.

Låt $X = \#$ bilar som

kör förbi på en timme.



$\Rightarrow X \sim \text{Poi}(\lambda)$ där $\lambda = 2 \cdot 5 = 10$

$$P(X \geq 3) = 1 - P(X < 3) = 1 - P(X \leq 2)$$

$$= 1 - \sum_{k=0}^2 f_X(k) = 1 - \sum_{k=0}^2 \frac{e^{-10} 10^k}{k!} = 0.997$$

Kap 4: Kontinuerliga stok. var.

Def (kontinuerlig stok. var.)

X är kont. stok. var. om S_X består av ett eller flera reell-värda intervall.

t.ex. $S_X = (a, b] \cup [c, d]$, $a, b, c, d \in \mathbb{R}$

$$S_X = \mathbb{R}$$

$$S_X = \mathbb{R}_+$$

Det gäller att $P(X=a)=0$

Ex: X = tiden en lampa står påslagen en viss dag.

Den kontinuerliga tätheten, $f_X(x)$, ger ej slh att X tar värdet x (ty $P(X=x)=0$). f_X används f.a. beräkna $P(a \leq X \leq b)$ etc.

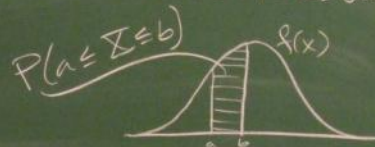
Def (kontinuerlig täthet)

Tätheten $f_X(x)$ för kont. stok. var. X uppfyller

$$1) f(x) \geq 0$$

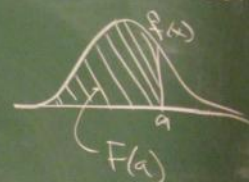
$$2) \int_{-\infty}^{\infty} f(x) dx = 1$$

$$3) P(a \leq X \leq b) = \int_a^b f(x) dx = \int_{-\infty}^b f(x) dx - \int_{-\infty}^a f(x) dx \\ = P(X \leq b) - P(X \leq a) = F(b) - F(a)$$

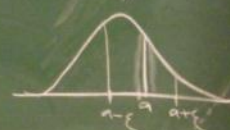
Def (Fördelningsfunktion)

Fördelningsfunktionen för den kont. stok. var. X ges av

$$F_X(a) = P(X \leq a) = \int_{-\infty}^a f(x) dx$$



Notera: $P(X=a) = \lim_{\epsilon \rightarrow 0} P(a-\epsilon \leq X \leq a+\epsilon) =$
 $= \lim_{\epsilon \rightarrow 0} \int_{a-\epsilon}^{a+\epsilon} f(x) dx = \int_a^a f(x) dx = 0$



Eftersom $P(X=a) = P(X=b) = 0$

$$P(a \leq X \leq b) = P(a < X \leq b) + \underbrace{P(X=a)}_{=0}$$

$$P(a \leq X \leq b) = P(a \leq X < b)$$

$$P(a \leq X \leq b) = P(a < X < b)$$

Sats (Täthet från förd. funkt.)

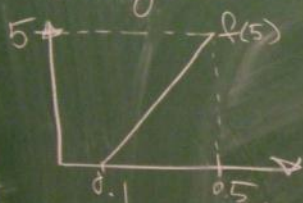
Låt X ha förd. funkt. $F(x)$

$$F'(x) = \frac{d}{dx} \int_{-\infty}^x f(t) dt = f(x)$$

Ex X = blykoncentrationen i bensin

$$f_X(x) = \begin{cases} 12.5x - 1.25 & \text{om } 0.1 \leq x \leq 0.5 \\ 0 & \text{annars} \end{cases}$$

(blykoncentr. ligger alltså mellan 0.1 och 0.5)



Är detta en täthet?

1) $f(x) \geq 0$ ja!

$$2) \int_{-\infty}^{\infty} f(x) dx = \int_{0.1}^{0.5} 12.5x - 1.25 dx =$$

$$= \left[\frac{12.5}{2} x^2 - 1.25x \right]_{0.1}^{0.5} = \dots = 1$$

$$P(0.2 \leq X \leq 0.3) = \int_{0.2}^{0.3} 12.5x - 1.25 \, dx =$$

$$= \left[\frac{12.5}{2} x^2 - 1.25x \right]_{0.2}^{0.3} = 0.1875$$

$$F_X(x) = \int_{-\infty}^x f(t) \, dt = \int_{0.1}^x 12.5t - 1.25 \, dt$$

$$= \left[\frac{12.5}{2} t^2 - 1.25t \right]_{0.1}^x = 6.25x^2 - 1.25x - (0.65 - 0.125)$$

Notera igen att $P(a \leq X \leq b) = F(b) - F(a)$

$$F'(x) = 12.5x - 1.25 = f(x)$$

Def (väntevärde)

$$E[X] = \sum_{x \in S_X} x \cancel{P(X=x)} = \int_{-\infty}^{\infty} x f(x) \, dx$$

funktion

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

Detta ger bl.a. att

$$\text{Var}(X) = E[X^2] - E[X]^2 = \int_{-\infty}^{\infty} x^2 f(x) \, dx - \left(\int_{-\infty}^{\infty} x f(x) \, dx \right)^2$$

Ex. $f(x) = 12.5x - 1.25$, $x \in [0.1, 0.5]$

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx = \int_{0.1}^{0.5} x(12.5x - 1.25) dx$$

$$= \left[\frac{12.5}{3} x^3 - \frac{1.25}{2} x^2 \right]_{0.1}^{0.5} = 0.3667$$

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_{0.1}^{0.5} x^2(12.5x - 1.25) dx$$

$$= \left[\frac{12.5}{4} x^4 - \frac{1.25}{3} x^3 \right]_{0.1}^{0.5} = 0.1433$$

$$\Rightarrow \text{Var}(X) = E[X^2] - (E[X])^2 = 0.1433 - 0.3667^2 = 0.09396$$

Momentgenererande funktioner

X kont. stok. var.

$$m_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

$$E[X^k] = \left. \frac{d^k m_X(t)}{dt^k} \right|_{t=0}$$

Om $E[e^{tX}] < \infty$ för $t \in (-h, h)$, ng $h > 0$

Ex $f_X(x) = e^{-x}, x > 0$ (exponentialfördeln. med $\beta=1$)

$$m_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x) dx = \int_0^{\infty} e^{tx} e^{-x} dx = \int_0^{\infty} e^{x(t-1)} dx$$

om $|t| < 1$ blir integralen ändlig

$$= \left[\frac{1}{t-1} e^{x(t-1)} \right]_0^{\infty} \stackrel{|t|<1}{=} \frac{1}{t-1} \left(\underbrace{\lim_{x \rightarrow \infty} e^{x(t-1)}}_{=0} - 1 \right) = \frac{1}{1-t}$$

$$E[X] = \frac{d}{dt} m_X(t) \Big|_{t=0} = \frac{d}{dt} \left(\frac{1}{1-t} \right) \Big|_{t=0} = 1$$

$$E[X^2] = \frac{d^2}{dt^2} m_X(t) \Big|_{t=0} = \frac{d^2}{dt^2} \left(\frac{1}{1-t} \right) \Big|_{t=0} = 2 \quad \therefore \text{Var}(X) = 2 - 1^2 = 1$$

Exponentialfördelningen, $X \sim \text{Exp}(\beta)$

Används typiskt för att beskriva livslängder.

$$f_X(x) = \frac{1}{\beta} e^{-x/\beta}, x > 0, \beta > 0$$

$$E[X] = \beta$$

$$\text{Var}(X) = \beta^2$$

$$F_X(x) = \int_0^x \frac{1}{\beta} e^{-t/\beta} dt = \left[-e^{-t/\beta} \right]_0^x = 1 - e^{-x/\beta}$$

Alternativt parametriseras $\text{Exp}(\beta)$ enl.

$\text{Exp}(\alpha)$: Notera: $\alpha = \frac{1}{\beta}$

$$f_X(x) = \alpha e^{-x\alpha}, \quad x > 0, x > 0$$

$$E[X] = \frac{1}{\alpha} \quad \text{Var}(X) = \left(\frac{1}{\alpha}\right)^2$$

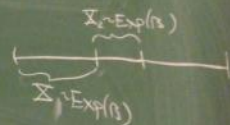
$$F_X(x) = 1 - e^{-x\alpha}$$

Ex. Vissa läkemedel producerar
"mördarpartiklar" som kan vara dödliga.
Förväntad tid mellan utsöndrade
partiklar är 5 timmar.
Antag $X \sim \text{Exp}(\beta)$.

Här är $\beta = 5$. X = tid till första partikel
 $f_X(x) = \frac{1}{5} e^{-x/5}$ $\beta = E[X] = 5$

$P(\text{måste vänta som mest 4 timmar innan första part.})$

$$= P(X \leq 4) = F(4) = 1 - e^{-4/5} = 0.5507$$



Normalfördelningen, $X \sim N(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \begin{array}{l} \mu \in \mathbb{R}, \sigma > 0 \\ -\infty < x < \infty \end{array}$$



μ är lägesparameter $X \sim N(\mu, \sigma^2)$

σ^2 är spridningsparameter $X_2 \sim N(\mu, \sigma_2^2)$

$$\sigma_2^2 > \sigma_1^2$$

$N(\mu, \sigma^2)$ har μ som väntevärde, dvs

$$E[X] = \mu$$

och σ^2 som varians, dvs

$$\text{Var}(X) = \sigma^2$$

Ex X = gram kolväten som en bil släpper ut per mil

Antag $X \sim N(\mu, \sigma^2)$ med $\mu = 1$, $\sigma = 0.25$



konstrukt att anv. en fördeln. som tar både positiva och negativa värden. Vi accepterar $N(\mu, \sigma^2)$ ändå.

$F(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$ har inget slutet uttryck.

I tabell V, Appendix A, finner vi värden på $\Phi(x) = F_z(x)$ då $Z \sim N(0,1)$.

Sats

Om $X \sim N(\mu, \sigma^2)$ då gäller att

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1) \Leftrightarrow X = \mu + \sigma Z \sim N(\mu, \sigma^2)$$

Beweis

$$E[Z] = E\left[\frac{X - \mu}{\sigma}\right] = \frac{1}{\sigma} E[X - \mu] = 0$$

$= E[X] - \mu = \mu - \mu = 0$

$$\text{Var}(Z) = \text{Var}\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma^2} \underbrace{\text{Var}(X - \mu)}_{= \text{Var}(X) = \sigma^2} = 1$$

$P(Z)$

$= f_Z(z)$

Sist: Kontinuerliga stok. var.

X har täthet $f_X(x)$ och fördeln. funkt.

$$F_X(x) = \int_{-\infty}^x f_X(t) dt \quad (P(X=a)=0)$$

$$P(a \leq X \leq b) = \int_a^b f_X(t) dt$$

$$E[H(X)] = \int_{-\infty}^{\infty} H(t) f_X(t) dt$$

anv. t.ex. för $\text{Var}(X) = E[X^2] - E[X]^2 = \int_{-\infty}^{\infty} x^2 f_X(x) dx - \left(\int_{-\infty}^{\infty} x f_X(x) dx \right)^2$

Normalfördeln. $X \sim N(\mu, \sigma^2)$, $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Exempel på kont. fördelningsfamiljer.

Normalfördeln. $X \sim N(\mu, \sigma^2)$, $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
 $E[X] = \mu$ $\text{Var}(X) = \sigma^2$

SåAs

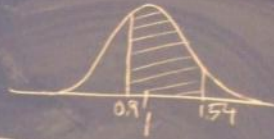
Om $X \sim N(\mu, \sigma^2)$ då gäller

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

$\Phi(x) = P(Z \leq x) = F_Z(x)$ finns i tabell V, Appendix

Idag:

Ex: $X \sim N(1, 0.25^2)$, $\mu=1$, $\sigma=0.25$



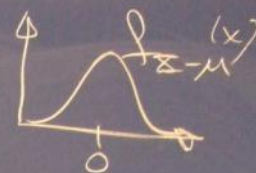
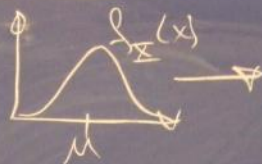
$$P(0.9 \leq X \leq 1.54) = P(X \leq 1.54) - P(X \leq 0.9)$$

$$= P\left(\underbrace{\frac{X-1}{0.25}}_{\substack{\text{vars} \\ \sim N(0,1)}} \leq \underbrace{\frac{1.54-1}{0.25}}_{=2.16}\right) - P\left(\underbrace{\frac{X-1}{0.25}}_{\sim N(0,1)} \leq \underbrace{\frac{0.9-1}{0.25}}_{=-0.4}\right) = \Phi(2.16) - \Phi(-0.4)$$

$$= 0.9846 - 0.3446 \approx 0.64$$

↑
Tabell

$$X \sim N(\mu, \sigma^2)$$



$$P(-\sigma \leq X - \mu \leq \sigma) =$$

$$= P\left(-1 \leq \underbrace{\frac{X - \mu}{\sigma}}_{\sim N(0,1)} \leq 1\right) =$$

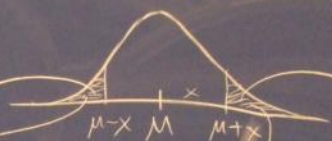


$$= \Phi(1) - \Phi(-1) \stackrel{\text{Tabell}}{=} 0.8413 - 0.1587 = 0.68$$

$$P(-2\sigma \leq X - \mu \leq 2\sigma) = 0.95$$

$$P(-3\sigma \leq X - \mu \leq 3\sigma) = 0.997$$

Om $X \sim N(\mu, \sigma^2)$ då är $f_X(x)$ symmetrisk
 $f_X(\mu-x) = f_X(\mu+x)$



$$\Rightarrow F_X(\mu-x) = P(X \leq \mu-x) = P(X \geq \mu+x) \\ = 1 - P(X \leq \mu+x) = 1 - F_X(\mu+x)$$

ex. $\Phi(x) = 1 - \Phi(-x)$

χ^2 -fördelningen, $X \sim \chi^2(\gamma)$

Viktig i statistikdelen av kursen!

Parametern $\gamma \in \mathbb{Z}_+ = \{1, 2, \dots\}$

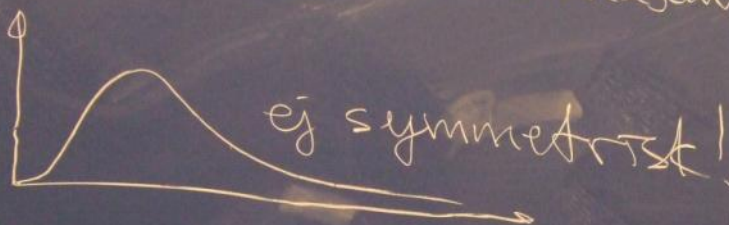
kallas frihetsgradsantal

$$E[X] = \gamma \quad \text{Var}(X) = 2\gamma$$

Värden på $F_X(t)$ återfinnes i
tabell IV, Appendix A.

$$f_X(x) = \frac{1}{\Gamma(\frac{x}{2}) 2^{x/2}} x^{x/2-1} e^{-x/2}$$

Gammafunktionen



Binomial-täthet och Normal-täthet:



Binomialtättheten
liknar en normal-
täthet!

Såts (Normalapproximation av Bin(n,p))

Låt $X \sim \text{Bin}(n,p)$. Då $n \rightarrow \infty$ kommer
fördelningen för X (dvs Bin(n,p)) att
gå mot en normalfördelning.

Annorlunda uttryckt, då n är stort

är X approximativt normal-
fördelad, med väntevärde np
och varians $np(1-p)$, $N(np, np(1-p))$

Bra sätt finna $F_X(x) = P(X \leq x)$
då $X \sim \text{Bin}(n, p)$ och n är stort!
(tabellen i boken går bara fram till
 $n=20$).

Vad betyder "n stort"?

Thmregel:

Vi kan använda satsen då
antingen

1) $p \leq \frac{1}{2}$ och $np > 5$

eller

2) $p > \frac{1}{2}$ och $n(1-p) > 5$

(Gör att binomialtätheten blir
tillräckligt symmetrisk!) Läs kap 4.6!

CHALMERSSPEYET
BOB

BILSETTER I KÅR-
HUSET PÅ LUNCH!

Läs kap. 4.6 på egen hand!

Ex

Antal bits som mottages felaktigt över en kanal modelleras med en binomialfördeln. där

$$p = P(\text{bit mottages felaktigt}) = 10^{-5}$$

$n = 16$ miljoner bit skickas

$X = \#$ felaktigt mottagna bit.

$$X \sim \text{Bin}(n, p) = \text{Bin}(16 \cdot 10^6, 10^{-5})$$

Finn $P(\text{inträffar fler än 150 fel})$!

$$P(X > 150) = 1 - P(X \leq 150) = 1 - \sum_{i=0}^{150} \binom{16 \cdot 10^6}{i} (10^{-5})^i (1 - 10^{-5})^{16 \cdot 10^6 - i}$$

? jobbigt räkna ut!

Enl. normalapprox. kan $\text{Bin}(16 \cdot 10^6, 10^{-5})$ approximeras med

$$N(\underbrace{16 \cdot 10^6 \cdot 10^{-5}}_{=n}, \underbrace{16 \cdot 10^6 \cdot 10^{-5} (1 - 10^{-5})}_{=n \cdot \underbrace{10^{-5} (1 - 10^{-5})}_{=p \cdot (1-p)}}) = N(160, 160(1 - 10^{-5}))$$

Enl. tumregeln:

$$p = 10^{-5} \leq \frac{1}{2} \quad \text{och} \quad np = 160 > 5$$

$$\text{Låt } Y \sim N(\underset{=\mu}{160}, \underset{=\sigma^2}{160(1-10^{-5})})$$

Normalapprox

$$\begin{aligned} P(X > 150) &= 1 - P(X \leq 150) \approx 1 - P(Y \leq 150) = \\ &= 1 - P\left(\underbrace{\frac{Y-\mu}{\sigma}}_{=Z \sim N(0,1)} \leq \underbrace{\frac{150-\mu}{\sigma}}_{\substack{= \frac{150-160}{\sqrt{160(1-10^{-5})}} \\ = -0.79}}\right) = \\ &= 1 - \Phi(-0.79) = 1 - (1 - \Phi(0.79)) = \Phi(0.79) \stackrel{\text{Tabell}}{=} 0.785 \end{aligned}$$

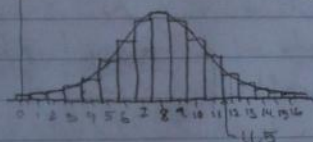
Kontinuitets-korrektion

$$X \sim \text{Bin}(n, p) \quad (\text{säg } n=20, p=0.4)$$

X är approximativt $N(\mu, \sigma^2)$ -fördelat.

$$\Rightarrow \frac{X - \mu}{\sigma} \approx N(0, 1)$$

$$\text{Låt } Y \sim N(\mu, \sigma^2)$$



Vi är intresserade av $P(X \geq 12)$ där $X \sim \text{Bin}(20, 0.4)$

$$\Rightarrow P(X \geq 12) = 1 - P(X < 12) = 1 - P(X \leq 11) \approx 1 - P(Y \leq 11)$$

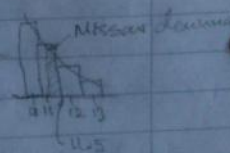
Vi drar bort för mycket i vår normalförd.

I stället betraktar vi

$$P(X \geq 12) \approx 1 - P(Y \leq 11 + 0.5)$$

vilket ligger närmare.

Vi drar bort halva stapeln



$$1 - \Phi(0.79) = 1 - \Phi(0.79) = \Phi(0.79) \approx 0.785$$

Läs kap 4.6 om kontinuitetskorrektur och storleksmed frågor! (Läxa!)

Kap 5 (Två-dimensionella stok. var.)

Def (Diskret 2-dim. stok. var.)

X och Y diskreta stok. var.

Det ordnade paret $T = (X, Y)$ kallas en 2-dim. stok. var.

Simultan täthet

Def (Simultan täthet)

En funktion $f_{x,y}(x,y)$ sådan att

$$f_{x,y}(x,y) = P(X=x, Y=y) = P(\{X=x\} \cap \{Y=y\})$$

kallas den simultana tätheten för den diskreta stok. var. $T = (X, Y)$.
(2-dim.)

SistTäthet (diskret) 2-dim stok. var. (X, Y)

$$f_{X,Y}(x,y) = P(X=x, Y=y) = P(\{X=x\} \cap \{Y=y\})$$

"den simultana tätheten"

Egenskaper hos tätheten (diskret):

$$1) f_{X,Y}(x,y) \geq 0 \quad \forall x \in S_X, y \in S_Y$$

$$2) \sum_{x \in S_X} \sum_{y \in S_Y} f_{X,Y}(x,y) = 1$$

möjliga värden på X Ex

En robot i en bilfabrik utför två sysslor:

1) Fastsvetsning av 2 ledaxlar

2) Fastskruvande av 3 bultar

 $X = \#$ felaktiga svetsningar per bil $Y = \#$ felaktigt fastskruvade bultar per bil (X, Y) är en 2-dim stok. var. med
täthet $f_{X,Y}(x,y)$, $S_X = \{0, 1, 2\}$, $S_Y = \{0, 1, 2, 3\}$

$f_{XY}(x,y)$ ges av

$x \backslash y$	0	1	2	3
0	0.84	0.03	0.02	0.01
1	0.06	0.01	0.008	0.002
2	0.01	0.005	0.004	0.001

$$\sum_{x \in S_X} \sum_{y \in S_Y} f_{XY}(x,y) =$$
$$= 0.84 + 0.03 + \dots + 0.004 + 0.001$$
$$= 1$$

$$P(\text{inga fel}) = f_{XY}(0,0) = 0.84$$

$$P(\text{exakt ett fel}) = f_{XY}(1,0) + f_{XY}(0,1)$$
$$= 0.06 + 0.03 = 0.09$$

$$P(\text{inga felaktigt åtskruvade bultar}) =$$
$$= P(Y=0) = \sum_{x \in S_X} f_{XY}(x,0) = \sum_{x=0}^2 f_{XY}(x,0)$$
$$= 0.84 + 0.06 + 0.01 = 0.91$$

Notera: $P(Y=0)$ ges av summan
av alla slh s.a. $y=0$.

Def (marginaltätthet)

(X, Y) diskret 2-dim. stok. var.
med täthet $f_{XY}(x, y)$.

Marginaltättheterna ges av:

$$f_X(x) = P(X=x) = \sum_{y \in S_Y} f_{XY}(x, y)$$

$$f_Y(y) = P(Y=y) = \sum_{x \in S_X} f_{XY}(x, y)$$

Ex I det tidigare exemplet:

$x \backslash y$	0	1	2	3	$f_X(x)$
0					0.9 = $\sum_{y=0}^3 f_{XY}(0, y)$
1					0.08
2					0.02
$f_Y(y)$	0.91	0.045	0.032	0.013	1

Def (kontinuerlig 2-dim stok. var.)

X och Y kont. stok. var.

Det ordnade paret $T = (X, Y)$
kallas en 2-dim. kont. stok. var.

En funk. $f_{X,Y}(x,y)$ kallas (X,Y) 's
simultana täthet om:

$$1) f_{X,Y}(x,y) \geq 0, \quad -\infty < x < \infty, -\infty < y < \infty$$

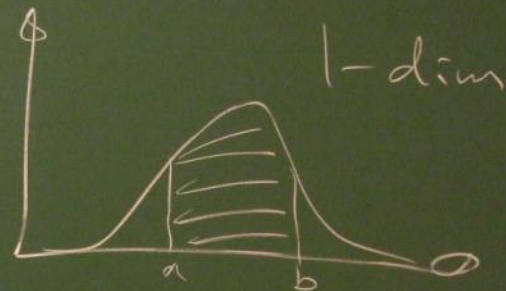
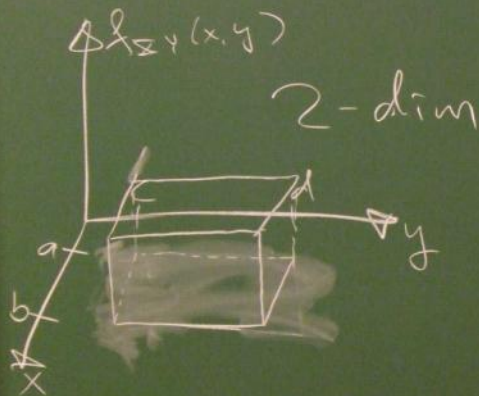
$$2) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy = 1$$

3) För $a, b, c, d \in \mathbb{R}$ gäller

$$P(a \leq X \leq b, c \leq Y \leq d) =$$

$$= \int_a^b \int_c^d f_{X,Y}(x,y) dy dx$$

Sannolikheter ges av volymer!



Ex Hos en försk 20-29 åring
 ligger kalciumnivån i blodet, X ,
 mellan 8.5 och 10.5 mg/dl och
 kolesterollnivån, Y , mellan 120 och 240
 mg/dl.
 Antag att (X,Y) är likformigt fördelad
 på rektangeln $\underbrace{(8.5, 10.5)}_{=S_X} \times \underbrace{(120, 240)}_{=S_Y} = S_{XY} \subseteq \mathbb{R}^2$
 dvs $f_{X,Y}(x,y) = C$ $(x,y) \in S_{XY}$ (se lädan ovan!)

Hitta det c som gör f_{xy} till en
tätthet!

$$D(c) \geq 0$$

$$2) 1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{xy}(x, y) dx dy = \int_{120}^{240} \int_{8.5}^{10.5} c dx dy$$

$$= c \int_{120}^{240} [x]_{8.5}^{10.5} dy = c(10.5 - 8.5) [y]_{120}^{240} = c \cdot 2 \cdot \frac{(240 - 120)}{= 120}$$

$$= 240c \quad \therefore c = \frac{1}{240}$$

$P(\text{persons kalciumnivå ligger mellan 9 och 10 mg/dl samt dess kolesterolnivå ligger mellan 125 och 140 mg/dl}) =$

$$= P(9 \leq X \leq 10, 125 \leq Y \leq 140) =$$

$$= \int_{125}^{140} \int_9^{10} f_{xy}(x, y) dx dy = \frac{1}{240} \int_{125}^{140} \int_9^{10} dx dy = \frac{140 - 125}{240} \cdot \frac{10 - 9}{= 1} = \frac{15}{240}$$

Def (Kontinuerlig marginaltätthet)

(X, Y) 2-dim. kont. stok. var. med
tätthet $f_{XY}(x, y)$. Marginaltättheterna
ges av:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y) dx$$

Ex | exempel 1 ovan:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy = \frac{1}{240} \int_{120}^{240} dy = \frac{140-120}{240} = \frac{1}{2}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y) dx = \frac{1}{240} \int_{8.5}^{10.5} dx = \frac{10.5-8.5}{240} = \frac{1}{120}$$

$$P(150 \leq Y \leq 200) = \int_{150}^{200} f_Y(y) dy = \frac{1}{120} \int_{150}^{200} dy = \frac{50}{120} = \frac{5}{12}$$

Notera att både X och Y likformigt fördelade då (X, Y) är det!

Oberoende stok. var.

Låt (X, Y) vara en 2-dim. stok. var.

Antag att (X, Y) är diskret.

Erinran: Händelser A och B är oberoende om $P(A \cap B) = P(A)P(B)$

Låt $A_x = \{X=x\}$ $B_y = \{Y=y\}$

som är oberoende om

$$P(A_x \cap B_y) = P(A_x)P(B_y)$$

dvs

$$\begin{aligned} f_{X,Y}(x,y) &= P(X=x, Y=y) = P(A_x \cap B_y) = \\ &= P(A_x)P(B_y) = P(X=x)P(Y=y) = f_X(x)f_Y(y) \end{aligned}$$

Def (oberoende stok. var.)

Låt (X, Y) vara en 2-dim. stok. var.
(antingen diskret eller kontinuerlig)
med täthet $f_{XY}(x, y)$.

X och Y är oberoende om

$$f_{XY}(x, y) = f_X(x) f_Y(y) \quad \text{för alla } (x, y) \in S_{XY}$$

Ex

a) I svets-/skruv-exemplet är
 X och Y ej oberoende ty

$$f_{XY}(0, 0) = 0.84 \neq \underbrace{0.9 \cdot 0.91}_{=0.819} = f_X(0) f_Y(0)$$

b) Kalcium-/blod-exemplet:

$$f_{XY}(x, y) = \frac{1}{240}$$

$\therefore X$ och Y ober.

$$f_X(x) f_Y(y) = \frac{1}{2} \cdot \frac{1}{120} = \frac{1}{240}$$

Def (väntevärde)

(X, Y) diskret: $E[H(X, Y)] = \sum_{x \in S_X} \sum_{y \in S_Y} H(x, y) f_{X,Y}(x, y)$

(X, Y) kontinuerlig: $E[H(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(x, y) f_{X,Y}(x, y) dx dy$

$H(\cdot, \cdot)$ är en funktion.

I synnerhet:

$E[X^k] = \sum_{x \in S_X} \sum_{y \in S_Y} x^k f_{X,Y}(x, y)$ ← disk. (X, Y)

$E[XY] = \sum_{x \in S_X} \sum_{y \in S_Y} xy f_{X,Y}(x, y)$ ← disk. (X, Y)

Sista:

$$E[H(X,Y)] = \begin{cases} \sum_{x \in S_X} \sum_{y \in S_Y} H(x,y) f_{XY}(x,y) & \text{om } (X,Y) \text{ diskret} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(x,y) f_{XY}(x,y) dx dy & \text{kont.} \end{cases}$$

H funktion

1-dim: $E[H(X)] = \sum_{x \in S_X} H(x) f_X(x)$

$$E[X^k] = \begin{cases} \sum_{x \in S_X} \sum_{y \in S_Y} x^k f_{XY}(x,y) & \text{om } (X,Y) \text{ diskret} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^k f_{XY}(x,y) dx dy & \text{kont.} \end{cases}$$

$$E[XY] = \begin{cases} \sum_{x \in S_X} \sum_{y \in S_Y} xy f_{XY}(x, y) \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{XY}(x, y) dx dy \end{cases}$$

Idag:

Def (Kovarians)

Kovariansen mellan X och Y ges av

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

$\text{Cov}(X, Y)$ beskriver ett linjärt beroende mellan X och Y .

Om $X > E[X]$ och $Y > E[Y]$ medför varandra
samt om $X < E[X]$ och $Y < E[Y]$ — || —
får vi att $\text{Cov}(X, Y) > 0$.

Om $X > E[X]$ och $Y < E[Y]$ medför varandra
eller om $X < E[X]$ och $Y > E[Y]$ — || —
får vi $\text{Cov}(X, Y) < 0$.

SaAs

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$$

Bewis

$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] = \\ &= E[XY - \overbrace{X\mu_Y - Y\mu_X}^{\text{icke-stokastiska (konst.)}} + \mu_X\mu_Y] = \\ &= E[XY] - \underbrace{\mu_Y E[X]}_{=\mu_X} - \underbrace{\mu_X E[Y]}_{=\mu_Y} + \mu_X\mu_Y \\ &= E[XY] - \mu_X\mu_Y\end{aligned}$$

Ex

1 snets/skruurobot - exempel

för vi:

$$E[XY] = \sum_{x=0}^2 \sum_{y=0}^3 xy f_{XY}(x, y) =$$



$$= (0 \cdot 0) \cdot 0.84 + (0 \cdot 1) \cdot 0.03 + \dots + (2 \cdot 3) \cdot 0.00 = 0.064$$

$$\begin{aligned}E[X] &= \sum_{x=0}^2 \sum_{y=0}^3 x f_{XY}(x, y) = 0 \cdot 0.84 + 0 \cdot 0.03 + \dots + 2 \cdot 0.00 \\ &= 0.12\end{aligned}$$

P.S.S.

$$E[Y] = \sum_{x=0}^2 \sum_{y=0}^3 y f_{X,Y}(x,y) = 0.148$$

$$\Rightarrow \text{Cov}(X,Y) = 0.064 - 0.12 \cdot 0.148 \\ = 0.04624$$

Sats

Om X och Y ober. $\Rightarrow \text{Cov}(X,Y) = 0$
(omvändningen gäller ej nödvändigtvis)



Sats

Om X och Y är normalfördelade
då gäller att

$$X \text{ och } Y \text{ ober.} \Leftrightarrow \text{Cov}(X,Y) = 0$$



Bara för att $|\text{Cov}(X,Y)|$ är
stor innebär det ej att X och Y
har ett starkt beroende!
(De kan ha stora varianser)

Def (Korrelation)
Korrelationen mellan X och Y ges av

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \quad -1 \leq \rho_{XY} \leq 1$$

Notera:

$$\text{Var}(X) = \text{Cov}(X, X) = E[(X - \mu_X)^2]$$

Sats

För de stok. var. X och Y gäller

$$|\rho_{XY}| = 1 \iff Y = \beta_0 + \beta_1 X$$

för några (deterministiska) konstanter

$$\beta_0, \beta_1 \in \mathbb{R}, \beta_1 \neq 0$$

Om $\rho_{ZY} \approx 1$ då är Z och Y starkt linjärt beroende.

" Z och Y är starkt positivt korrelerade"
om $\rho_{ZY} \approx 1$.

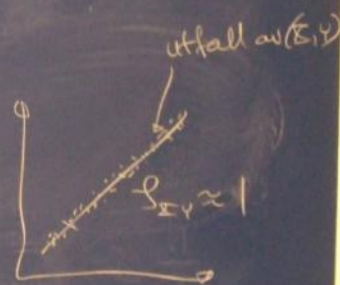
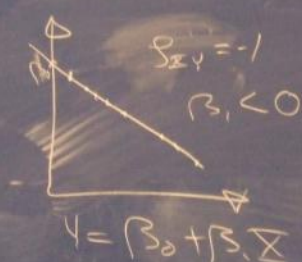
$\rho_{ZY} \approx -1$ (starkt negativt korrelerade)

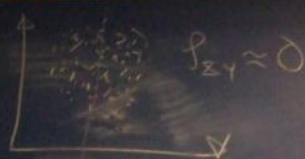
Z och Y hämmar varandra starkt (i modellen)

" Z och Y starkt negativt korrelerade"

$\rho_{ZY} = 0$ (Inget linjärt beroende)

" Z och Y är okorrelerade"





Linjärt beroende
säger ej allt!
(såvida inte X och
 Y är normalfördelade)

För (X, Y) diskret, definiera händelserna
 $A_x = \{X=x\}$, $B_y = \{Y=y\}$, $x \in S_X$, $y \in S_Y$.

$$f_{X|Y=y}(x) = P(A_x | B_y) = \frac{P(A_x \cap B_y)}{P(B_y)} = \frac{P(X=x, Y=y)}{P(Y=y)} = \frac{f_{XY}(x, y)}{f_Y(y)}$$

$f_Y(y) > 0$

Def (Betingad täthet)

(X, Y) 2-dim. stok. var. (antingen diskret eller kont.) med simultan täthet $f_{XY}(x, y)$ och Y har marginaltäthet $f_Y(y)$. Den betingade tätheten för X givet $Y=y$

ges av

$$f_{X|Y=y}(x) = \frac{f_{XY}(x, y)}{f_Y(y)} \quad \text{om } f_Y(y) > 0$$

P.S.S. får vi att den betingade tätheten för Y givet $X=x$ ges av

$$f_{Y|X=x}(y) = \frac{f_{XY}(x,y)}{f_X(x)} \quad \text{om } f_X(x) > 0.$$

Stokastiska processer

Def

En samling stokastiska variabler $\{X_t\}_{t \in T}$ där index-mängden T antingen ges av $T = \mathbb{N} = \{0, 1, 2, \dots\}$ eller $T = \mathbb{R}^+ = [0, \infty)$ som (möjligen) beror på varandra kallas för en stokastisk process.

Ex (slumpvandring)

$$T = \{0, 1, 2, \dots\} \text{ dvs } \{X_t\}_{t \in \{0, 1, 2, \dots\}}$$

$$X_0 = 0$$

Låt $Y_1, Y_2, Y_3, \dots$ vara oberoende stok. var. s.a.

$$Y_i = \begin{cases} 1 & \text{med slh } p \\ -1 & \text{med slh } 1-p \end{cases}$$

$$X_t = \sum_{i=1}^t Y_i = Y_1 + \dots + Y_t, \quad t \geq 1$$



Slugga slant.
Varje gång vinner vi en kr med slh p och förlorar en med slh $1-p$.

X_t = totala vinsten efter t spelomgångar.

$$\begin{aligned} \text{Cov}(X_1, X_2) &= \text{Cov}(Y_1, Y_1 + Y_2) = \\ &= E[(Y_1 - E[Y_1])(Y_1 + Y_2 - E[Y_1 + Y_2])] \\ &= E[(Y_1 - E[Y_1])(Y_1 - E[Y_1] + Y_2 - E[Y_2])] \\ &= E[(Y_1 - E[Y_1])(Y_1 - E[Y_1]) + (Y_1 - E[Y_1])(Y_2 - E[Y_2])] \\ &= E[(Y_1 - E[Y_1])^2] + E[(Y_1 - E[Y_1])(Y_2 - E[Y_2])] \end{aligned}$$

$$= \text{Var}(Y_1) + \underbrace{\text{Cov}(Y_1, Y_2)}_{=0 \text{ ty } Y_1 \text{ och } Y_2 \text{ oberoende}} = 1 - (2p-1)^2$$

$$E[Y_1] = \sum_{y \in S_{Y_1}} y P(Y_1 = y) = 1 \cdot p + (-1)(1-p) = 2p-1$$

$$E[Y_1^2] = 1^2 p + (-1)^2 (1-p) = p + 1-p = 1$$

$$\text{Var}(Y_1) = E[Y_1^2] - E[Y_1]^2 = 1 - (2p-1)^2$$

Slust:

$$\{X_t\}_{t \in \{0,1,2,\dots,3\}}$$

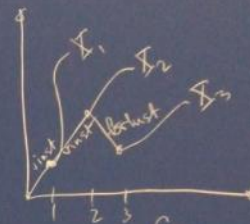
Slumpvandring

$$X_0 = 0$$

$$X_t = \sum_{i=1}^t Y_i, t \geq 1, \text{ där } Y_1, Y_2, \dots \text{ är oberoende, } Y_i = \begin{cases} 1 & \text{med slh } p \\ -1 & \text{med slh } 1-p \end{cases}$$

Vi singlar slant där vi vinner 1 kr om vi myntet visar krona (slh p) och förlorar 1 kr om vi får klave. X_t = total vinst efter t spelgångar.

Vi räknade ut $\text{cov}(X_1, X_2)$



Väntevärdessfunktionen:

$$\mu_t = E[X_t] = E\left[\sum_{i=0}^t Y_i\right] = \sum_{i=1}^t \overbrace{E[Y_i]}^{=2p-1} = \sum_{i=1}^t (2p-1) = t(2p-1)$$

notera att då $p = \frac{1}{2}$, $\mu_t = 0$

$$\gamma(t) = \text{Var}(X_t) = \text{Var}\left(\sum_{i=1}^t Y_i\right) \stackrel{\text{ober.}}{=} \sum_{i=1}^t \overbrace{\text{Var}(Y_i)}^{=1-(2p-1)^2} = t(1-(2p-1)^2)$$

Om $p = \frac{1}{2} \Rightarrow \gamma(t) = \text{Var}(X_t) = t$

Observation av X_t

Ex (AR(1)-processen)

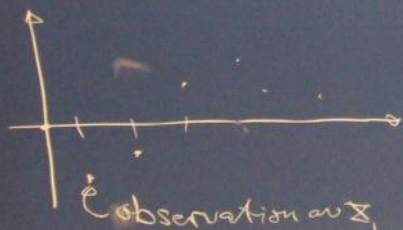
$\{X_t\}_{t \in \{0,1,2,\dots\}}$

$$X_0 = 0, X_t = \alpha X_{t-1} + e_t, \alpha \in \mathbb{R}$$

$e_1, e_2, \dots$ ober. all med fördeln. $e_i \sim N(0, \sigma^2)$

$$X_1 = \alpha \overset{=0}{X_0} + e_1 = e_1$$

$$X_2 = \alpha X_1 + e_2$$



Observation av X_t



Detta är en så kallad (svagt) stationär process, dvs

$$\mu_t = E[X_t] = m \in \mathbb{R} \text{ (konstant)}$$

Autokovarians

$$\text{Cov}(X_t, X_{t+h}) = \text{Cov}(X_1, X_{1+h}) = R(h), h \in \mathbb{N}$$

Beroendet i processen beror bara på avståndet mellan två tidpunkter.

$$\text{Notera att } \text{Var}(X_t) = \text{Cov}(X_t, X_t) = R(0)$$

Def (Auto-korrelation)

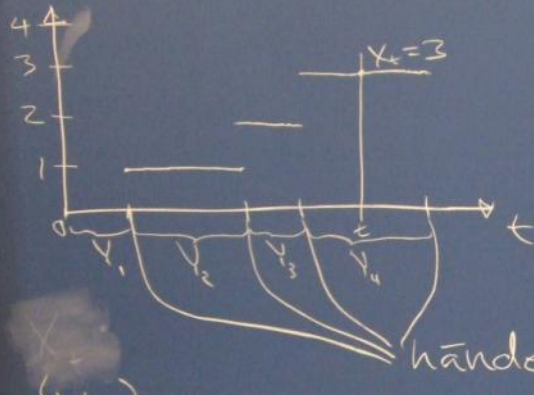
$$r(h) = \frac{\text{Cov}(X_t, X_{t+h})}{\sqrt{\text{Var}(X_t) \text{Var}(X_{t+h})}} = \frac{R(h)}{R(0)}$$

$\underbrace{\text{Var}(X_t)}_{= R(0)} \quad \underbrace{\text{Var}(X_{t+h})}_{= R(0)}$

$$-1 \leq r(h) \leq 1$$

Om beroendet avtar med tiden då kommer också $r(h)$ att avta med ökande h .

Ex (Poisson-processen)



Kontinuerlig tid!
dvs $T = [0, \infty)$

$Y_1, Y_2, \dots$ är oberoende och

$Y_i \sim \text{Exp}(\beta)$, $f_{Y_i}(t) = \frac{1}{\beta} e^{-\frac{1}{\beta}t}$

$\{X_t\}_{t \geq 0}$, där X_t = antal händelser som har inträffat vid tiden t

Poisson-processen (parameter $\beta > 0$)

$$X_t = \begin{cases} 0 & \text{då } 0 \leq t < Y_1 \\ 1 & \text{då } Y_1 \leq t < Y_1 + Y_2 \\ 2 & \text{då } Y_1 + Y_2 \leq t < Y_1 + Y_2 + Y_3 \\ \vdots & \end{cases}$$

Sats i boken knyter ihop de $\text{Exp}(\beta)$ -fördelade tids mellanrummen med X_t 's $\text{Poi}(\beta t)$ -fördeln.

Sats

$$X_t \sim \text{Poi}\left(\frac{1}{\beta}t\right), \text{ dvs } P(X_t = k) = \frac{e^{-\frac{1}{\beta}t} \left(\frac{1}{\beta}t\right)^k}{k!}$$

Ex (Poi, exp.)

Medeltalet bilar som passerar ett visst övergångsställe är 1 var 5:te timma. Ober. trider mellan ankomster (dessa är även exponentialford.).

Vi har en Poissonprocess med parameter $\lambda = \frac{1}{5}$.

$P(\text{Vänta mindre än 4 timmar mellan bil 1 och bil 2}) =$

$$\begin{aligned} &= P(Y_2 \leq 4) = \int_0^4 \underbrace{\frac{1}{5} e^{-\frac{1}{5}x}}_{\text{tätheten för Exp}(\lambda)\text{-ford.}} dx = \\ &= \left[-e^{-\frac{1}{5}x} \right]_0^4 \stackrel{\lambda=\frac{1}{5}}{=} -e^{-\frac{1}{5} \cdot 4} + e^0 = 1 - e^{-\frac{4}{5}} = 0.5507 \end{aligned}$$

$$\begin{aligned} &P(\text{vi har sett endast en bil efter 10 timmar}) = \\ &= P(X_{10} = 1) = \frac{e^{-\frac{10}{5}} \left(\frac{10}{5}\right)^1}{1!} = 2e^{-2} \approx 0.27 \end{aligned}$$

Matematisk statistik (kap 6)

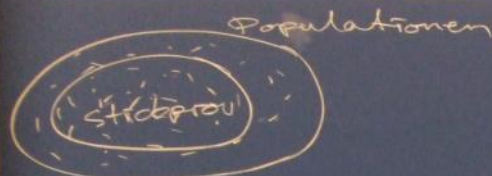
Vi vill säga ngt om en population



t.ex. alla längder på alla män
i åldrarna 25-35

Vi kan (av olika anledningar) ej mäta upp
hela populationen.

Vi väljer ut ett "stickprov" från populationen
dvs vi väljer ut en delmängd av populationen
som vi studerar vidare.



Vi antar att uppmätningarna ej har ngt att göra
med varandra.

Def (Stickprov)

Ett stickprov $X_1, \dots, X_n$ är oberoende stot. var.
som har samma fördelning.

Före
uppmättn

Efter
uppmättn

När vi har observerat stickprovet, skrivs det som
 $x_1, \dots, x_n$ dvs $X_1 = x_1, \dots, X_n = x_n$

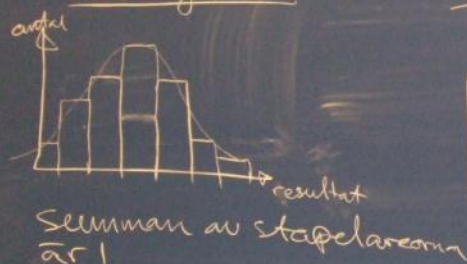
Baserat på de slutsatser vi gör angående
vårt stickprov så önskar vi kunna dra
slutsatser kring populationen.

Vi bryr oss ej om data-insamlingen!

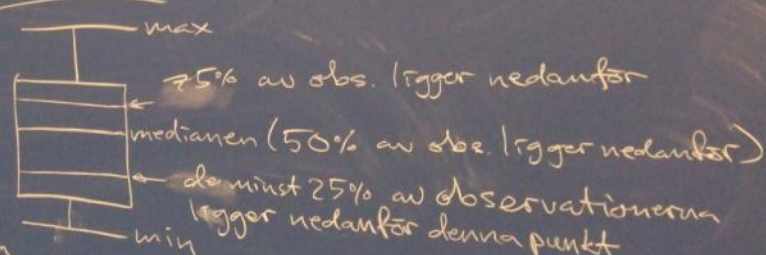
Vi vill kunna avgöra vilken fördelning
stickprovet följer.

Grafisk illustration av fördelningen:

Histogram



Box-plot



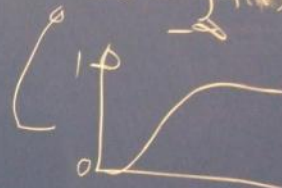
Empirisk fördelningsfunktion

$$F_n(x) = \frac{\# \text{ observationer som är } \leq x}{n}$$

$n \leftarrow$ totala antalet



$\hookrightarrow$ Efterliknar $F(x) = \int_{-\infty}^x f(t) dt$



* Jämför med $F(x)$
genom t.ex.

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \leftarrow \text{Kolmogorov-Smirnov}$$

Def (stickprovsmedelvärde, "sample mean")

För stickprovet $X_1, \dots, X_n$ ges stickpr. medelv. av

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (\text{observeras som } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i)$$

$\hookrightarrow$ Används f.a. uppskatta $E[X_i]$

Def (stickprovsvarians, "sample variance")

$$\text{Var}(X_i) = E[(X_i - E[X_i])^2] \quad \checkmark \text{ Anv. f.a. uppskattar } \text{Var}(X_i)$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (\text{obs. som } s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2)$$

Def (statistika)

En funktion $H(X_1, \dots, X_n)$ kallas en statistika.

Def (skattare)

En statistika $H(X_1, \dots, X_n)$ som används för att uppskatta (approximera) en fördelnings-parameter kallas för en (punkt-)skattare. t.ex. $\bar{X}, S^2$

STSTDef (Punktskattare)

En statistiska (funktion av stickprov) som används för att uppskatta ngn parameter θ kallas en punktskattare

$$\hat{\theta} = \hat{\theta}(X_1, \dots, X_n) \leftarrow \text{stok. var.}$$

$\hat{\theta}(x_1, \dots, x_n)$ kallas punktskattning

↑
observerade stickprovet
ej stokastiskt

Ex

Vill uppskatta $E[X_i] = \mu$

Vi använder $\hat{\theta}(X_1, \dots, X_n) = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

Efter vi mätt upp $x_1, \dots, x_n$ får vi $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n) = \bar{x}$

Hur bra är $\hat{\theta}$ på att skatta θ ?

Def (Väntevärdesriktighet (V.V.R.) - "Unbiased estimator")

Skattaren $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ kallas väntevärdesriktig för θ om $E[\hat{\theta}] = \theta$.

skattningar från 5 olika st. pr.
 $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4, \hat{\theta}_5$

$\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$ är en stok. var.
som anv. f.a. uppskatta θ .

Vi mäter upp, säg, 10 stickprov.

$X_1 = (x_1, \dots, x_n), \dots, X_{10} = (x_1, \dots, x_n)$ (gör om experimentet)

Räkna ut medelvärdet 10 ggr

är $\hat{\theta}(x_1), \dots, \hat{\theta}(x_{10})$ ← 10 skattningar

Vi vill att detta medelv. skall ligga nära
den riktiga parametern, dvs att
vår skattare producerar skattningar
som i snitt hamnar i θ .

Sats

$X_1, \dots, X_n$ stickpr. där $E[X_i] = \mu$

$\bar{X}$ är v.v.r. för μ , dvs $E[\bar{X}] = \mu$

Beweis

$$E[\bar{X}] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n \overbrace{E[X_i]}^{=\mu} = \frac{1}{n} n\mu = \mu \quad \square$$

Sats

$X_1, \dots, X_n$ stickpr. med $E[X_i] = \mu$ och $\text{Var}(X_i) = \sigma^2$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \text{ är v.v.r. för } \sigma^2$$

Det (standardfel)

$\hat{\theta}$'s standardfel ges av

$$SE(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})} \quad \leftarrow \text{skattarens standardavvikelse}$$

Sats

$X_1, \dots, X_n$ stickor med $E[X_i] = \mu$, $\text{Var}(X_i) = \sigma^2$

Det gäller att $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$, dvs

$$SE(\bar{X}) = \sqrt{\text{Var}(\bar{X})} = \frac{\sigma}{\sqrt{n}}$$

Vad händer då vi får fler och fler observationer, dvs då $n \rightarrow \infty$?

$$\lim_{n \rightarrow \infty} SE(\bar{X}) = \lim_{n \rightarrow \infty} \frac{\sigma}{\sqrt{n}} = 0$$

Med ett stort antal observationer gör $\bar{X}$ ett bra jobb vad gäller att uppskatta μ .

Ex

X = utfall i ett tärningskast.

$S_X = \{1, \dots, 6\}$, $P(X=k) = \frac{1}{6}$, $k \in S_X$

$$E[X] = 3.5$$

Vi kastar tärning 30 ggr.

(Vi räknar ut $\bar{X}$)

Gör om 56 ggr. $\Rightarrow \bar{X}_1, \dots, \bar{X}_{56}$ (olika värden)

Dessa är observationer av $\bar{X}$.

Vi finner att deras medelvärde ges av

$$\frac{1}{56} \sum_{i=1}^{56} \bar{X}_i = 3.548 > 3.5 = \mu$$



Vi vill ha 2 huvudsakliga egenskaper hos de skattare vi konstruerar.

1) v.v.r., dvs $E[\hat{\theta}(x_1, \dots, x_n)] = \theta$

2) Lågt standardfel, dvs $SE[\hat{\theta}(x_1, \dots, x_n)] = \sqrt{\text{Var}(\hat{\theta})}$ litet.

Metoder för att finna skattare

Säg att $X_1, \dots, X_n$ följer en given fördelningsfamilj (t.ex. $N(\mu, \sigma^2)$, $\text{Exp}(\lambda)$, $\text{Bin}(n, p)$).

Hur finner vi skattare för parametrarna i denna fördeln.?

Momentmetoden

$X_1, \dots, X_n$ stickpr. som kommer från en fördeln. med parametrar $\theta_1, \dots, \theta_s$ (t.ex. för $N(\mu, \sigma^2)$, $\theta_1 = \mu$, $\theta_2 = \sigma^2$).

Varje moment $E[X_i^k]$ skattas med
 $M_k = \frac{1}{n} \sum_{i=1}^n X_i^k$ (t.ex. $M_1 = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, $M_2 = \frac{1}{n} \sum_{i=1}^n X_i^2$)

Notera att $E[X_i^k]$ är en funktion av $\theta_1, \dots, \theta_s$ (t.ex. då $X_i \sim N(\mu, \sigma^2)$, $E[X_i] = \mu$, $E[X_i^2] = \sigma^2 + \mu^2$).

Vi finner skattarna $\hat{\theta}_1, \dots, \hat{\theta}_s$ genom att lösa ekvationssystemet

$$\begin{cases} M_1 = E[X_i] \\ M_2 = E[X_i^2] \\ \vdots \\ M_s = E[X_i^s] \end{cases}$$

Ex

$X_1, \dots, X_n$ st. pr. från $N(\mu, \sigma^2)$

$$\left\{ \frac{1}{n} \sum_{i=1}^n X_i = E[X_i] = \mu \right.$$

$$\left\{ \frac{1}{n} \sum_{i=1}^n X_i^2 = E[X_i^2] = \text{Var}(X_i) + E[X_i]^2 = \sigma^2 + \mu^2 \right.$$

$$\Rightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X} \quad (\text{dvs momentmetoden ger oss att väntevärdet skattas med medelvärdet})$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \hat{\mu}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

Notera $\hat{\sigma}^2 = \frac{n}{n-1} S^2 = \frac{n}{n-1} \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

Eftersom S^2 är vvr för $\text{Var}(X_i) = \sigma^2$
så är ej $\hat{\sigma}^2$ vvr för σ^2

Däremot är $\hat{\mu} = \bar{X}$ vvr för μ .

Maximum-likelihood-skattare
(trolighetsskattare)

Den simultana tätheten för st. pr.
 $X_1, \dots, X_n$ är $f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i)$

Vilka parametrar har gett upphov
till $\underline{X}_1 = x_1, \dots, \underline{X}_n = x_n$? (t.ex. vilket (μ, σ^2) är $\underline{X}_1, \dots, \underline{X}_n$)
ett givet utfall

Ide: de parametrar $\theta_1, \dots, \theta_s$ som maximerar

$$f_{\underline{X}_1, \dots, \underline{X}_n}(x_1, \dots, x_n) = f_{\underline{X}_1, \dots, \underline{X}_n}(x_1, \dots, x_n; \theta_1, \dots, \theta_s)$$

↪ då $\underline{X}_1, \dots, \underline{X}_n$ diskreta är detta s.k. att

$\underline{X}_1 = x_1, \underline{X}_2 = x_2, \dots, \underline{X}_n = x_n$ inträffar samtidigt

Vi vill maximera (m.a.p. $\theta_1, \dots, \theta_s$) s.k. att $\underline{X}_1 = x_1, \dots, \underline{X}_n = x_n$

Likelihood-funktionen ges av

$$L(\theta_1, \dots, \theta_s) = f_{\underline{X}_1, \dots, \underline{X}_n}(x_1, \dots, x_n; \theta_1, \dots, \theta_s) = \prod_{i=1}^n f_{\underline{X}_i}(x_i; \theta_1, \dots, \theta_s)$$

↑
tättheten för $\underline{X}_i$

Vill finna $(\theta_1, \dots, \theta_s)$ som maximerar
 $L(\theta_1, \dots, \theta_s)$.

Vi skall derivera $L(\theta_1, \dots, \theta_s)$.

Svårt att derivera en produkt!

De $\theta_1, \dots, \theta_s$ som maximerar $L(\theta_1, \dots, \theta_s)$
 maximerar även $l(\theta_1, \dots, \theta_s) = \ln(L(\theta_1, \dots, \theta_s)) =$
 $= \ln\left(\prod_{i=1}^n f_{\mathbf{x}_i}(x_i; \theta_1, \dots, \theta_s)\right) = \sum_{i=1}^n \ln(f_{\mathbf{x}_i}(x_i; \theta_1, \dots, \theta_s))$

Vi vill lösa

$$\begin{cases} \frac{\partial}{\partial \theta_1} l(\theta_1, \dots, \theta_s) = 0 \\ \frac{\partial}{\partial \theta_2} l(\theta_1, \dots, \theta_s) = 0 \\ \vdots \\ \frac{\partial}{\partial \theta_s} l(\theta_1, \dots, \theta_s) = 0 \end{cases}$$

Ex $X_1, \dots, X_n$ ströckpr. från $N(\mu, \sigma^2)$ -fördeln.

$$f_{\mathbf{x}_i}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$L(\mu, \sigma^2) = \prod_{i=1}^n f_{\mathbf{x}_i}(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$l(\mu, \sigma^2) = \sum_{i=1}^n \ln(f_{\mathbf{x}_i}(x_i)) = \sum_{i=1}^n \left(\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(x_i-\mu)^2}{2\sigma^2} \right)$$

$$0 = \frac{\partial}{\partial \mu} l(\mu, \sigma^2) = \sum_{i=1}^n \frac{\partial}{\partial \mu} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \sum_{i=1}^n \frac{\partial}{\partial \mu} \frac{(x_i-\mu)^2}{2\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n -2(x_i-\mu)$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^n x_i - \mu = \frac{1}{\sigma^2} \sum_{i=1}^n x_i - \frac{n}{\sigma^2} \mu \Leftrightarrow \mu = \frac{1}{n} \sum_{i=1}^n x_i = \bar{X}$$

$$\begin{aligned}\frac{\partial}{\partial \sigma} \log(h(\mu, \sigma^2)) &= \sum_{i=1}^n -\frac{\partial}{\partial \sigma} (\log(\sigma)) - \frac{\partial}{\partial \sigma} \left(\frac{(x_i - \mu)^2}{\sqrt{2} \sigma} \right) = \\ &= -\sum_{i=1}^n \frac{1}{\sigma} - \sum_{i=1}^n 2 \left(-\frac{x_i - \mu}{\sqrt{2} \sigma^2} \right) \left(\frac{x_i - \mu}{\sqrt{2} \sigma} \right) = \\ &= -\frac{n}{\sigma} + 2 \sum_{i=1}^n \frac{(x_i - \mu)^2}{2 \sigma^2} = -\frac{n}{\sigma} + \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 = 0\end{aligned}$$

$$n = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \Leftrightarrow \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

• Vi får ekvationssystemet

$$\begin{cases} \mu = \bar{x} \\ \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \end{cases} \Leftrightarrow \begin{cases} \mu = \bar{x} \\ \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \end{cases}$$

• Uttrycker vi detta i stickprovet $X_1, \dots, X_n$ istället för i det observerade för vi skattarna.

$$\hat{\mu} = \bar{X} \quad \text{och} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n}{n-1} S^2$$

• (Räknar vara samma som moment-skattarna)

Sist:

Punktskattare

Momentmetoden

Maximum-likelihood

Vi visade att då $X_1, \dots, X_n$ är ett $N(\mu, \sigma^2)$ -fördelat stickprov, då gäller att vi skattar μ och σ^2 med

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$\hat{\sigma}^2 = \frac{n-1}{n} S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

 $\hat{\mu} \in \mathcal{E}_\mu$ v.u.r.

Skattare (dvs. stok. var.)

Ex

Vi har mätt upp $X_1 = -1.1465$ $X_2 = 1.1909$ $X_3 = 1.1892$

$X_4 = -0.0376$ $X_5 = 0.3273$

Dessa antas komma från ett normalfördelat stickprov med okänt μ och σ^2 . $\hat{\mu} = \bar{X} = \frac{1}{5}(-1.1465 + 1.1909 + 1.1892 - 0.0376 + 0.3273) \approx 0.3047$

$$P(\theta \in [L_1, L_2]) = 1 - \alpha, \quad 0 \leq \alpha \leq 1$$

$$\text{där } L_1 = L_1(\bar{X}, \bar{X}_0, \alpha) \quad L_2 = L_2(\bar{X}, \bar{X}_0, \alpha)$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{5} \left((-1.1465 - 0.3047)^2 + \dots + (0.3275 - 0.3047)^2 \right) = 0.7583$$
 skattning (ej/stok. var. utan uppmätta värden)
 Ligger $\hat{\mu} = 0.3047$ nära det riktiga värdet på μ (om $x_1, \dots, x_5$ faktiskt är observationer av en $N(\mu, \sigma^2)$ -fördelning)?
 (I själva verket är $x_1, \dots, x_5$ värden som har simulerats från $N(0, 1)$)!

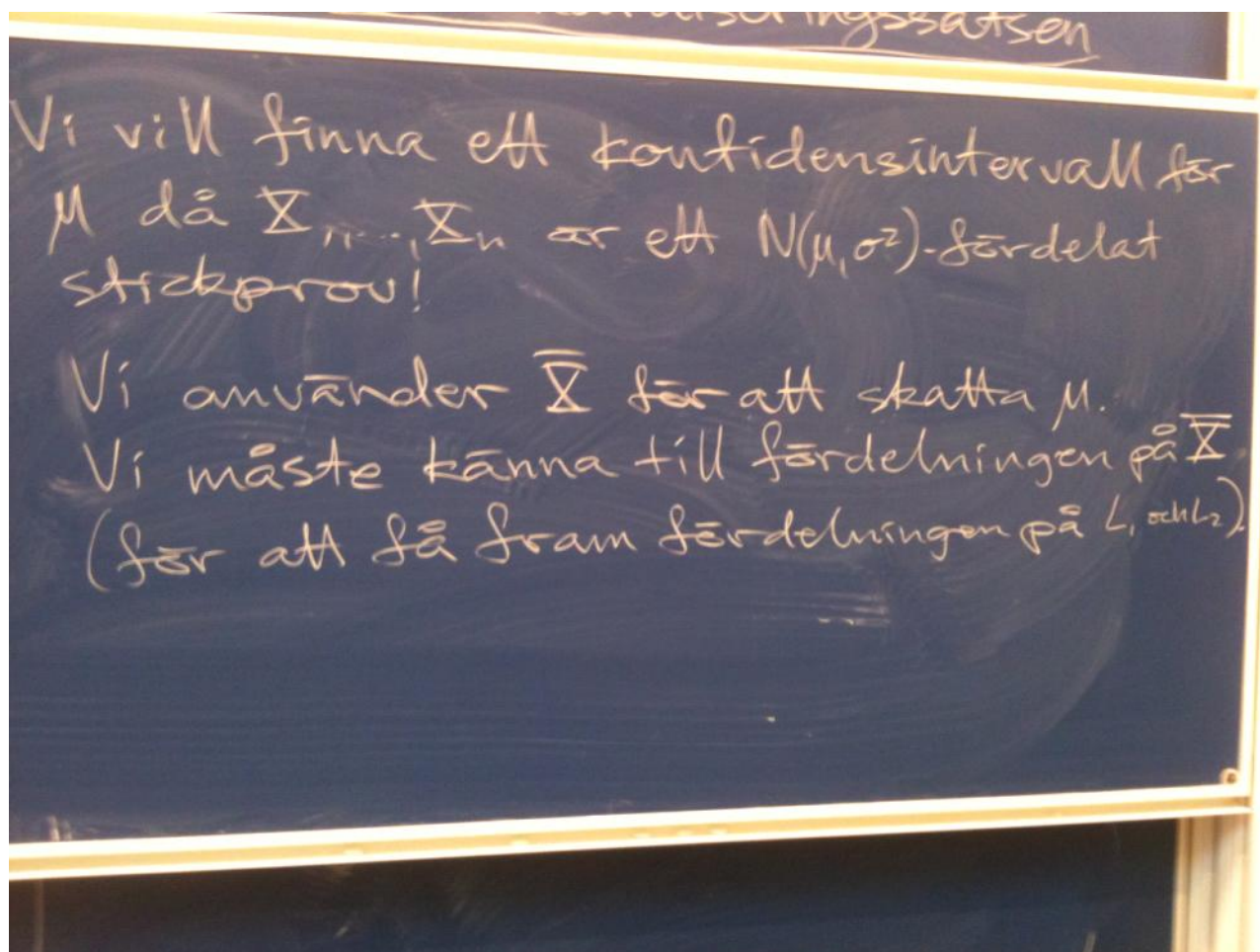
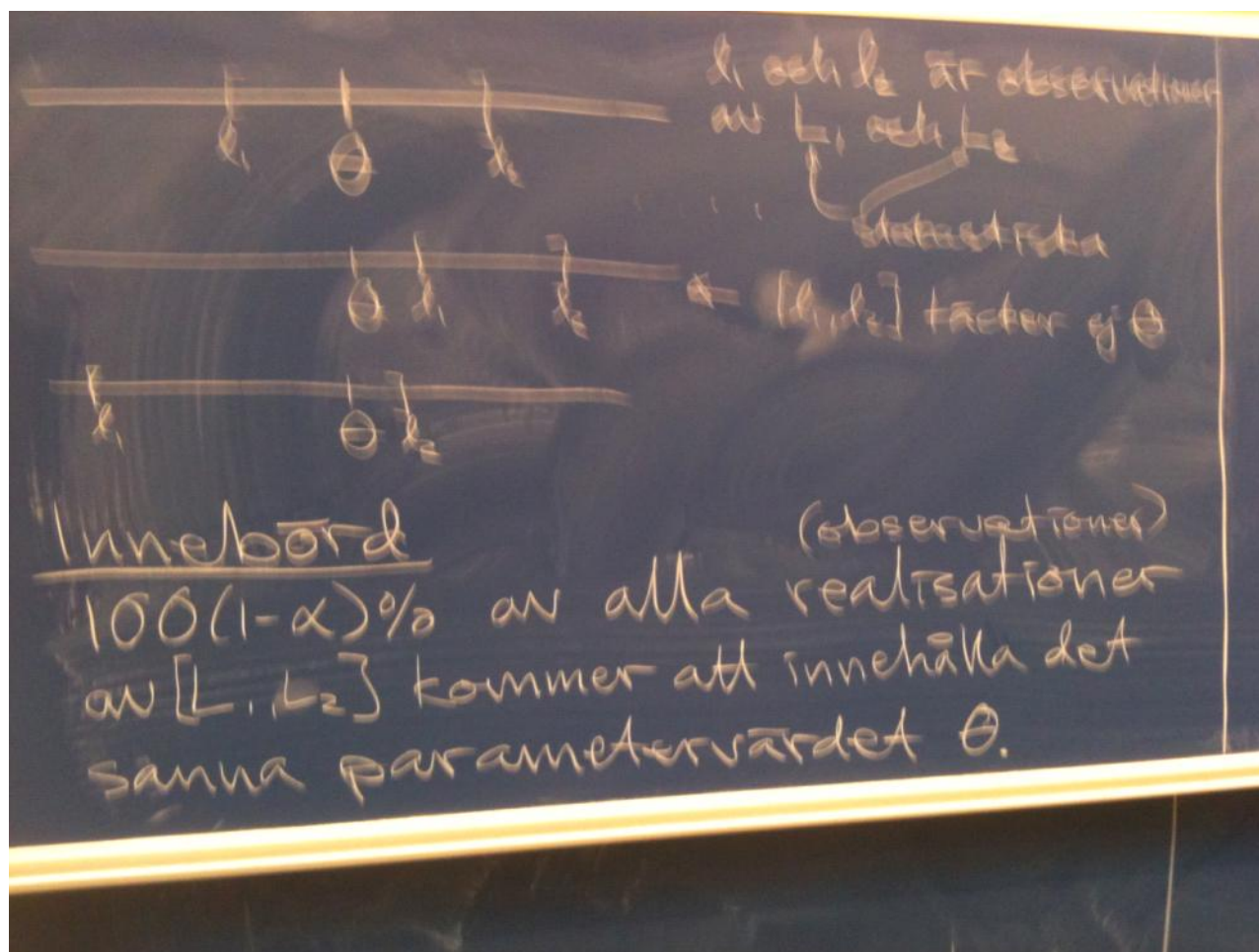
Vi kan ej vara försäkrade om att en viss uppmätt skattning, $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$, ligger nära den bakomliggande parametern θ .
det st. pr.

Def (Konfidenstervall)

Ett $100(1-\alpha)\%$ konfidenstervall, $0 \leq \alpha \leq 1$, för en parameter θ är ett stokastiskt intervall $[L_1, L_2]$ baserat på data $\Sigma_n, \bar{X}_n$ s.a.

$$P(\theta \in [L_1, L_2]) = 1 - \alpha, \quad 0 \leq \alpha \leq 1,$$

där $L_1 = L_1(\Sigma_n, \bar{X}_n, \alpha)$ $L_2 = L_2(\Sigma_n, \bar{X}_n, \alpha)$



Vi vet sedan tidigare att

$$E[\bar{X}] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \mu$$

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

Sats

X och Y stok. var. med mgt $m_X(t)$

$$\text{och } m_Y(t) = E[e^{tY}]$$

X och Y har samma fördelning

$$m_X(t) = m_Y(t) \quad (\Rightarrow) \quad \text{för alla } t \in (-\varepsilon, \varepsilon), \text{ något } \varepsilon > 0$$

Exempel: Standardiseringsatsen

Om $\bar{X}$ är $N(\mu, \sigma^2)$ -fördelad gäller

$$Z = \frac{\bar{X} - \mu}{\sigma} \sim N(0, 1) \quad \left(\begin{array}{l} \text{anv. t. ex. för att räkna ut} \\ P(\bar{X} \leq a) = P\left(\frac{\bar{X} - \mu}{\sigma} \leq \frac{a - \mu}{\sigma}\right) = \\ \sim N(0, 1) \end{array} \right)$$

Bevis

$$Z \sim N(0, 1) \text{ har mgt } m_Z(t) = e^{t^2/2}$$

$$\bar{X} \sim N(\mu, \sigma^2) \text{ har mgt } m_{\bar{X}}(t) = e^{\mu t + \sigma^2 t^2/2} \quad \leftarrow \text{se tabell}$$

Vi skall visa att $\bar{X} - \mu$ har samma mgt som Z

$$m_{\bar{X} - \mu}(t) = E\left[e^{t(\bar{X} - \mu)}\right] = E\left[e^{-\mu t/\sigma} e^{\frac{t}{\sigma} Z}\right]$$

$$= e^{-\mu t/\sigma} E\left[e^{\frac{t}{\sigma} Z}\right] = e^{-\mu t/\sigma} m_Z\left(\frac{t}{\sigma}\right) = e^{-\mu t/\sigma} e^{\frac{1}{\sigma^2} \left(\frac{t}{\sigma}\right)^2/2} = e^{-\mu t/\sigma + t^2/2}$$

$$= e^{i/\sigma} = m_Z(t)$$

$\therefore \frac{Z-\mu}{\sigma}$ har samma fördeln. som $Z \sim N(0,1)$

Sats

Om X och Y är oberoende då gäller

$$m_{X+Y}(t) = m_X(t) m_Y(t)$$

Beris

$$\begin{aligned} m_{X+Y}(t) &= E[e^{t(X+Y)}] = E[e^{tX} e^{tY}] = E[e^{tX}] E[e^{tY}] \\ &= m_X(t) m_Y(t) \end{aligned}$$

Sats (summer av oberoende normalfördelade)

Låt $X_1, \dots, X_n$ vara oberoende normalfördelade med väntevärden $\mu_1, \dots, \mu_n$ och varianser $\sigma_1^2, \dots, \sigma_n^2$.

Summan $Y = \sum_{i=1}^n X_i$ är $N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$ -fördelad.

Beris

$$\begin{aligned} m_Y(t) &= m_{X_1 + \dots + X_n}(t) \stackrel{\text{sats}}{=} \prod_{i=1}^n m_{X_i}(t) = \prod_{i=1}^n e^{\mu_i t + \sigma_i^2 t^2 / 2} \\ &= e^{t \left(\sum_{i=1}^n \mu_i \right) + \frac{t^2}{2} \left(\sum_{i=1}^n \sigma_i^2 \right)} = m_{N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)}(t) \end{aligned}$$

Jmför med $m_{N(\mu, \sigma^2)}(t) = e^{\mu t + \sigma^2 t^2 / 2}$

Enl. tidigare sats ($\sum_{i=1}^n Y_i$ likad. $\Leftrightarrow \mu_{\sum_{i=1}^n Y_i} = \mu_1 + \dots + \mu_n$)
 så gäller att $Y = \sum_{i=1}^n Z_i \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$

Sats

Låt $Z_1, \dots, Z_n$ vara ett stickprov (avs. obs., i. f. stickprov)
 från $N(\mu, \sigma^2)$. Då gäller att

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

Beweis:

$$\bar{X} = \sum_{i=1}^n \frac{1}{n} Z_i \quad \text{Varje } Y_i \text{ är normalford.}$$

$$E[Y_i] = E\left[\frac{1}{n} Z_i\right] = \frac{1}{n} \mu$$

$$\text{Var}(Y_i) = \text{Var}\left(\frac{1}{n} Z_i\right) = \frac{1}{n^2} \overbrace{\text{Var}(Z_i)}^{\sigma^2} = \frac{\sigma^2}{n^2}$$

$$\text{Från förra satsen gäller att } \bar{X} = \sum_{i=1}^n Y_i \sim N\left(\sum_{i=1}^n \frac{1}{n} \mu, \sum_{i=1}^n \frac{\sigma^2}{n^2}\right)$$

$$= N\left(\mu, \frac{\sigma^2}{n}\right)$$

Nu härleder vi vårt konfidenzintervall för μ då $X_1, \dots, X_n$ är ett stickpr. från $N(\mu, \sigma^2)$.

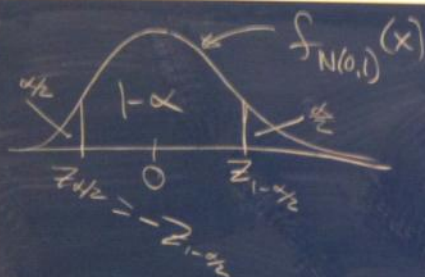
Vi vet att $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ dvs

$$\frac{\bar{X} - E[\bar{X}]}{\sqrt{\text{Var}(\bar{X})}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Således

Konf. int. för μ ges av

$$[\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}] = [L_1, L_2]$$



$$P(Z \leq z_{1-\alpha/2}) = 1 - \frac{\alpha}{2}$$

Kvantil (finns i tabell)

Bevis

$$\begin{aligned} 1-\alpha &= P(\underbrace{-z_{\alpha/2}}_{\sim N(0,1)} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}) = \\ &= P(-z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}) = \end{aligned}$$

$$= P\left(\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right)$$

$$= P\left(\mu \in \underbrace{\left[\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right]}_{=L_1}\right)$$



$100(1-\alpha)\%$ kallas för intervallets konfidensgrad.

Längden på intervallet beror på olika faktorer:

För det tidigare intervallet gäller att längden är

$$l(\alpha, \sigma, n) = \underbrace{\left|\left(\bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right) - \left(\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right)\right|}_{=L_2}$$

$$= 2 z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Och intervalllet är centrerat i $\bar{x}$.

$$\left[\bar{x} - 2z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + 2z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

Om α minskar då ökar $z_{1-\frac{\alpha}{2}}$ och $l(x, \sigma, n)$ växer

Om σ ökar då ökar $l(x, \sigma, n)$

Om n ökar då krymper $l(x, \sigma, n)$

Vi vill ha små intervall!

Sist $Y = \sum_{i=1}^n X_i$ är normalfördelad med

$$E(Y) = \sum_{i=1}^n \mu_i, \quad \text{Var}(Y) = \sum_{i=1}^n \sigma_i^2$$

$\frac{1}{n} \sum_{i=1}^n X_i$ där $X_1, \dots, X_n$ är oberoende. $X_i \sim N(\mu_i, \sigma_i^2)$

$\Rightarrow \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ då $X_1, \dots, X_n$ är ett normalford. stickprov med $E(X_i) = \mu$, $\text{Var}(X_i) = \sigma^2$

$\Rightarrow$ Ett $100(1-\alpha)\%$ konfidensintervall för μ (dvs μ är okänd, här är σ^2 känt) då $X_1, \dots, X_n$ $N(\mu, \sigma^2)$ -stickpr. ges av

$$\left[\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right] \text{ skrivs även } \bar{X} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Vad detta egentligen säger är att

$$1-\alpha = P\left(\underbrace{\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}}_{=L_1} \leq \mu \leq \underbrace{\bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}}_{=L_2} \right)$$

Idag:

Ex

V : observatorer

8.0	13.6	13.2	13.6
12.5	14.2	14.9	14.5
13.4	8.6	11.5	16.0
14.2	19.0	17.9	17.0

som kommer från en normalfördelning med varians $\sigma^2 = 9$. Finn ett 95% konf. int. för μ .

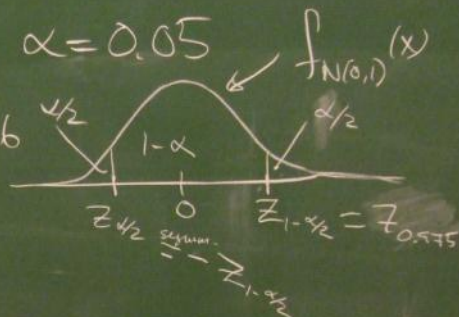
Vi får

$$\hat{\mu} = \bar{x} = 13.88$$

$$95\% = 100(1-\alpha)\%, \text{ dvs } \alpha = 0.05$$

$$z_{1-\alpha/2} = z_{1-\frac{0.05}{2}} = z_{0.975} = 1.96$$

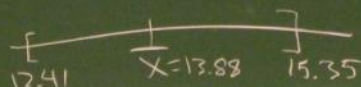
Från tabell



$$n=16, \sigma=3$$

Konfidensintervallet ges av

$$\bar{x} \pm z_{0.975} \frac{\sigma}{\sqrt{n}} = 13.88 \pm 1.96 \frac{3}{4} = 13.88 \pm 1.47$$



Vi har ställt kravet att vårt stickprov $X_1, \dots, X_n$ skall vara normalfördelat. Vad händer om det ej är normalförd. (dvs $X_1, \dots, X_n$ ober. med samma fördeln., ej $N(\mu, \sigma^2)$)?

GS (Centrala gränsvärdessatsen)

$X_1, \dots, X_n$ stickprov från en godtycklig fördelning med $E[X_i] = \mu$, $\text{Var}(X_i) = \sigma^2$

Då gäller att

$$\lim_{n \rightarrow \infty} \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \lim_{n \rightarrow \infty} \frac{\frac{1}{n} \sum_{i=1}^n X_i - E[\bar{X}]}{\sqrt{\text{Var}(\bar{X})}} \stackrel{\text{förd.}}{=} Y \sim N(0,1)$$

dvs

$$\lim_{n \rightarrow \infty} P\left(\frac{\frac{1}{n} \sum_{i=1}^n X_i - E[\bar{X}]}{\sqrt{\text{Var}(\bar{X})}} \leq a\right) = P(Y \leq a)$$

där $Y \sim N(0,1)$, $a \in \mathbb{R}$

∴ summor av oberoende likafördelade stok. var.
konvergerar (i fördelning) mot
en $N(0,1)$ -fördelning

Vi säger även att då n är stort

då är $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ approximativt $N(0,1)$ -fördelat

Från CGS får vi att då n stort

$$1 - \alpha \approx P\left(-z_{1-\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right) \quad \left\{ \begin{array}{l} \text{och } X_1, \dots, X_n \text{ ej} \\ \text{normalförd.} \end{array} \right.$$

$$1 - \alpha \approx P\left(\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right)$$

Vi har ett approx. kont. int.

$$\bar{X} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

då $X_1, \dots, X_n$ stickprov från annan fördeln. än
normalförd.

Vad är ett stort n ?

Tunnregel: $n \geq 25$ då kan vi använda intervallet trots att $X_1, \dots, X_n$ ej normalförd.

Normal approx. av $\text{Bin}(n, p)$:

Låt $X_1, \dots, X_n$ vara oberoende, $X_i \sim \text{Ber}(p) \Rightarrow Y = \sum_{i=1}^n X_i \sim \text{Bin}(n, p)$

Eml. CGS

$$\frac{Y - E(Y)}{\sqrt{\text{Var}(Y)}} = \frac{\frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} E\left[\sum_{i=1}^n X_i\right]}{\frac{1}{n} \sqrt{\text{Var}\left(\sum_{i=1}^n X_i\right)}} \xrightarrow{n \rightarrow \infty} N(0, 1)$$

Kont. int. för μ (då σ okänd)

$X_1, \dots, X_n$ stickpr. från $N(\mu, \sigma^2)$ där μ, σ^2 är okända.

Det gäller ju att (sett tidigare) $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$

Hur gör vi då σ ej känt?

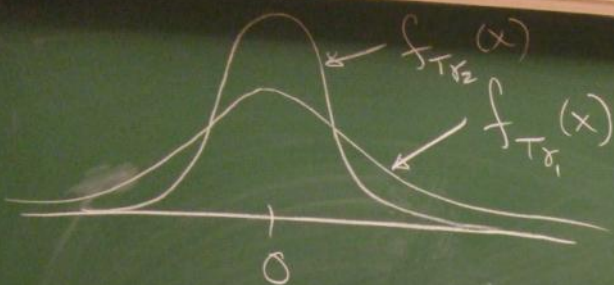
Förslag! Byt ut σ mot dess skattare, dvs betrakta $\frac{\bar{X} - \mu}{S/\sqrt{n}}$

Vad har $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ för fördelning?

Den är ej normalfördelad utan följer
en så kallad T-fördelning!

(Student) T-fördelningen, $Z \sim T(\delta)$

- 1) T-fördelningen styrs av parametern δ ,
ett positivt heltal.
 δ kallas fördelningens frihetsgradsantal.
- 2) En kontinuerlig fördelning,
Tätheten är en (symmetrisk) klock-kurva centrerad
i $E[Z] = 0$ (precis som $N(0,1)$ fördeln.)

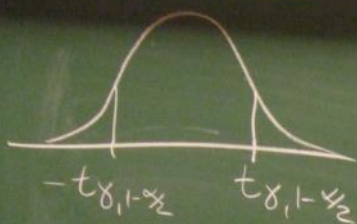


$$\delta_2 > \delta_1$$

Med ökande δ minskar
variansen

En skillnad mot $N(0,1)$ är att
 T_δ -fördelningen har "tjockare svansar"
än $N(0,1)$ -förd, dvs $P(T_\delta > u) > P(N(0,1) > u)$
då u stort.

3) Då $\delta \rightarrow \infty$ kommer T_δ -fördelningen
att gå mot en $N(0,1)$ -fördeln.



$P(\bar{X} \leq t_{\gamma, 1-\alpha/2}) = 1 - \frac{\alpha}{2}$ finns i tabell
för olika γ, α .

Sats

$X_1, \dots, X_n$ st. pr. från $N(\mu, \sigma^2)$.
Då gäller att

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim T_{n-1}$$

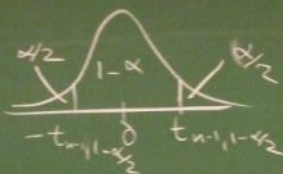
"är T-fördelad med $n-1$ frihetsgrader"

Sats (konf. int. för μ då μ och σ^2 skända)

$X_1, \dots, X_n$ st. pr. från $N(\mu, \sigma^2)$

En $100(1-\alpha)\%$ konfidensintervall för μ
ges av $\bar{X} \pm t_{n-1, 1-\alpha/2} \frac{S}{\sqrt{n}}$

Beweis



$$1 - \alpha = P\left(-t_{n-1, 1-\alpha/2} \leq \underbrace{\frac{\bar{X} - \mu}{S/\sqrt{n}}}_{\sim T_{n-1}} \leq t_{n-1, 1-\alpha/2}\right)$$

$$\Rightarrow 1 - \alpha = P\left(-t_{n-1, 1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \leq \bar{X} - \mu \leq t_{n-1, 1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right)$$

$$= P\left(\bar{X} - t_{n-1, 1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1, 1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right)$$

Ex

Vi har fått $n=24$ mätningar där man kunnat antaga att värdena kommer från en normalford (stånd μ och σ^2).

Vi vet $\bar{X} = 53.92$, $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = 10.07$

$n=24 \Rightarrow$ betrakta en $T_{24-1} = T_{23}$ fördeln

$$t_{n-1, 1-\frac{\alpha}{2}} = t_{23, 0.975} \stackrel{\text{tabell}}{=} 2.069$$

$\alpha=0.05$ dvs ett 95% konf. int.

Ett 95% konf.-int. för μ ges nu av

$$\bar{X} \pm t_{n-1, 1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} = 53.92 \pm 2.069 \frac{10.07}{\sqrt{24}}$$

dvs $[49.67, 58.17]$

Hypotestest

Vi har en uppfattning/påstående (hypotes)
angående ngt och vill se om hypotesen stämmer.

Ex

- Ett nytt läkemedel förlänger leukemipatienters liv.
- Efter 4 år är medelnivån av alkohol i blodet över den i trafiken tillåtna.

Våra hypoteser berör en hel population.
Vi har endast tillgång till ett stickprov, $X_1, \dots, X_n$

Ex

Hypotes: mer än 50% av alla bilar har feljusterade framlykter

Intressant eftersom om hypotesen stämmer kommer ett nytt kontrollsystem att införas.

Låt p = faktisk andel feljusterade ljus i populationen.

Vill testa om majoriteten av alla lykter är feljusterade.

De två hypoteserna:

1) Modhypotesen, H_1 (alternativhypotesen)

Den hypotes som forskaren tror på.

2) Nollhypotesen, H_0

Motsatsen till forskarens hypotes

Exemplet:

Vi testar $H_0: p \leq 0.5$ mot $H_1: p > 0.5$

Man antar att H_0 är sann tills vi lyckats hitta bevis för att förbasta den.

Def (Teststatistika)

En statistika (funktion av stickprovet) som används för att besluta om vi kan förkasta H_0 .

Notera att H_0 ger/kan ge oss en bestämd fördeln. på teststatistikan.

Exemplet:

Vi använder teststatistikan

$\underline{X} = \#$ bilar med felriktade lyktor

(bland $X_1, \dots, X_n$)

Antag att $H_0: p \leq 0.5$ är sann.

Jämför sedan $\frac{\underline{X}}{n}$ med $p_0 = 0.5$
↗ H_0 -värde närmast H_0

Om vi kan förkasta hypotesen att $p=0.5$ (H_0 -värde närmast H_1), då kan vi även förkasta $H_0: p \leq 0.5$!

När förkastar vi H_0 ?

Vi förkastar $H_0: p \leq 0.5$ då vi får ett stort värde på teststatistikan, $\frac{\sum}{n}$

Antar vi nu att H_0 faktiskt gäller (dvs $p=0.5$)
då blir $X \sim \text{Bin}(n, p)$

Vi undersöker $n=20$ bilar ($X_1, \dots, X_{20}$)

$\Rightarrow X \sim \text{Bin}(20, 0.5)$

Om X (observerade teststatistikan) avviker mycket från de värden som $\text{Bin}(n, p)$ -fördelningen producerar med stor slh, då förkastar vi $H_0: p \leq 0.5$

$f_{\text{Bin}(20, 0.5)}(x)$

x är en funktion av $x_1, \dots, x_n$
 $\uparrow$ obs. stickpr.

Antag nu att x (obs.-teststat.) får ett stort värde och vi därmed väljer att förkasta H_0 .

Två möjligheter:

1) Vi har rätt: H_0 är sann och vi förkastar H_0 .

2) Vi gör ett Typ I-fel:

H_0 var i själva verket sann men det obs. stickpr. vi fick var sållsynt/extremt och därför förkastade vi H_0 .

Vanligen tror man att 1) gäller men det finns en slh att 2) gäller.

Slh för ett Typ I-fel,

$$P(\text{Typ I-fel}) = P(\text{förkasta } H_0 \mid H_0 \text{ sann}) = \alpha$$

kallas signifikansnivå.

Exemplet:

Vi testar $H_0: p \leq 0.5$ mot $H_1: p > 0.5$ ($p=0.5$)

Teststat: $X = \#$ bilar med felriktade lyktor

Under $H_0: p=0.5$ är $X \sim \text{Bin}(n, p_0) = \text{Bin}(20, 0.5)$

Eftersom $E[X] = np_0 = 20 \cdot 0.5 = 10$ väljer vi en förkastningsgräns α mindre än 10, såg att vi förkastar H_0 då $X \in \{14, 15, \dots, 20\}$
 $\uparrow$ förk. gräns

Signifikansnivån ges nu av

$$\alpha = P(\underbrace{X \in \{14, \dots, 20\}}_{X \geq 14} \mid \underbrace{X \sim \text{Bin}(n, p_0) = \text{Bin}(20, 0.5)}_{H_0 \text{ sann}}) \approx 0.05$$

Det (kritiskt område/förkastningsregion)

Förkastningsregionen, C , är den mängd av värden som teststatistikan kan antaga sådana att vi förkastar H_0 .

Exemplet: $C = \{14, \dots, 20\}$

Säg att vi ej kan förkasta H_0 .

Two möjligheter:

1) Vi har rätt: H_0 är sann och vi kan inte förkasta H_0

2) Typ II-fel:

Vi accepterar H_0 fast H_1 är sann.

Låt

$$\beta = P(\text{Typ II-fel}) = P(\text{Acceptera } H_0 \mid H_1, \text{ sann})$$

Exemplet:

$$\beta(p) = P(\text{Acceptera } H_0 \mid X \sim \text{Bin}(20, p)) = P(X < 14 \mid X \sim \text{Bin}(20, p))$$

$p > 0.5$

$H_1: p > 0.5$

Det (styrka på ett test)

styrkan ges av $P(\text{Förkasta } H_0 \mid H_1, \text{ sann}) =$

$$= 1 - P(\text{Acceptera } H_0 / H_1 \text{ sann}) = 1 - \beta$$

Exemplet: styrkan ges av $1 - \beta(p)$, $p > 0.5$
 för $p = 0.7$ får vi en funktion

$$\beta(0.7) = P(\text{Acceptera } H_0 / X \sim \text{Bin}(20, 0.7)) =$$

$$= P(X \leq 14 / X \sim \text{Bin}(20, 0.7)) = 0.392$$

$\Rightarrow$ Styrkan på nivå $p = 0.7$ är $1 - \beta(0.7) = 0.608$
 $X \leq 13$ tabell

Vi vill designa test med liten signifikansnivå
 och hög styrka!

Hittills har vi valt förkastningsregion och
 sedan fått vår signifikansnivå, α .

Hädanetter vänder vi på det, dvs vi fixerar
 en signifikansnivå och får då en förkastningsregion.

Procedur:

- 1) Definiera H_0 och H_1
- 2) Fixera signifikansnivån, α och definiera teststatistikan $T = T(\underbrace{x_1, \dots, x_n}_{\text{stok. var.}})$
- 3) Mha $\alpha = P(T \in C \mid H_0 \text{ sann})$ finner vi det kritiska området C .
- 4) Beräkna den observerade teststat., $T_{\text{obs}} = T(\underbrace{x_1, \dots, x_n}_{\text{ej stokastisk}})$ och förk. H_0 om $T_{\text{obs}} \in C$

Om $T_{\text{obs}} \in C$ säger vi att
"vi förkastar H_0 på signifikansnivån α "

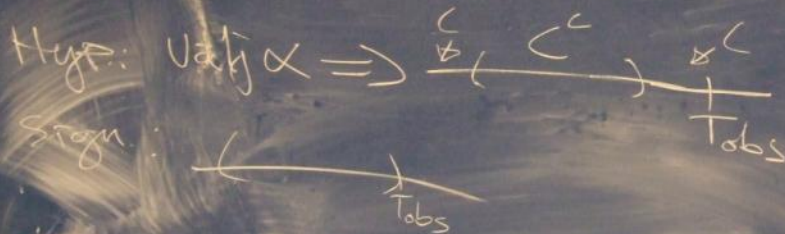
Signifikanstest

Vi sätter upp våra hypoteser, H_0, H_1 ,
Definierar T , räknar ut T_{obs} .

Vi bestämmer nu slh att T antar
ett värde åtminstone lika extremt
som T_{obs} (Givet H_0 -sann, dvs under H_0 -fördeln.)

Denna sth kallas för p-värdet.

p-värdet är minsta möjliga signifikans-nivå, dvs minsta α som kan väljas så att H_0 fortfarande kan förkastas.



Sist

- Gick igenom hypotestest & signifikanstest

Ex

I hopp om att bland annat minska bränsleförbrukningen hos bilar används aluminium som material.

En viss modell kan i genomsnitt köra 26 miles/gallon (mpg) med en standardavvikelse på $\sigma = 5$ mpg.

Vi tror att en ny, aluminiumrikare, design kommer att förlänga körsträckan.

Vi testar $H_0: \mu \leq 26$ mot $H_1: \mu > 26$
(μ anger förväntad körsträcka)

Vi väljer teststatistikan $T = \bar{X}$ (ty den uppskattar μ)

Vi gör 36 mätningar, $x_1, \dots, x_{36}$, och av dessa får vi $\bar{x} = 28.04$
dvs $T_{\text{obs}} = 28.04$.

Under H_0 gäller att (kom ihåg att vi byter ut $H_0: \mu \leq 26$
mot $H_0: \mu = 26$) $E[\bar{X}] = E\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{1}{n} \sum_{i=1}^n E[x_i] = \mu = 26$

$$\text{Var}(\bar{X}) \stackrel{H_0}{=} 25$$

Enl. CGS gäller det att $\bar{X} \stackrel{\text{approx}}{\sim} N(26, 25)$
(ty $n \geq 25$).

$$\text{p-värde} = P(T > T_{\text{obs}} | H_0) =$$

$$\approx P(\bar{X} > \bar{x} | \bar{X} \sim N(26, 25)) =$$

$$= P\left(\underbrace{\frac{\bar{X} - 26}{\sigma/\sqrt{n}}}_{\sim N(0,1)} > \frac{28.04 - 26}{\sqrt{25}/\sqrt{36}}\right) = P(Z > 2.45)$$

$$= 1 - \underbrace{\Phi(2.45)}_{\text{t tabell}} = 0.0071$$

Vi kan alltså förkasta $H_0: \mu = 26$ ända ner till signifikansnivån 0.0071

∴ Den nya processen ger längre körsträcka.

Test för väntevärdet (σ okänd)

T-test $X_1, \dots, X_n$ från $N(\mu, \sigma^2)$

Vill göra utsaga om μ (både μ och σ okända).

Tre möjliga test:

$$H_0: \mu \leq \mu_0$$

$$H_0: \mu \geq \mu_0$$

$$H_0: \mu = \mu_0$$

$$H_1: \mu > \mu_0$$

$$H_1: \mu < \mu_0$$

$$H_1: \mu \neq \mu_0$$

Här använder vi egentligen $H_0: \mu = \mu_0$

Teststatistikan ges av

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$$\left(S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

som är T_{n-1} -fördelad.

2 vägar. Antingen fixera signifikansnivån eller så räknar vi ut p-värdet.

H_1

Förkasta H_0 om

$\mu > \mu_0$

$$T_{\text{obs}} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} > t_{1-\alpha, n-1}$$

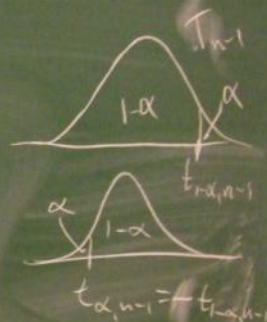
$\mu < \mu_0$

$$T_{\text{obs}} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} < t_{\alpha, n-1}$$

$\mu \neq \mu_0$

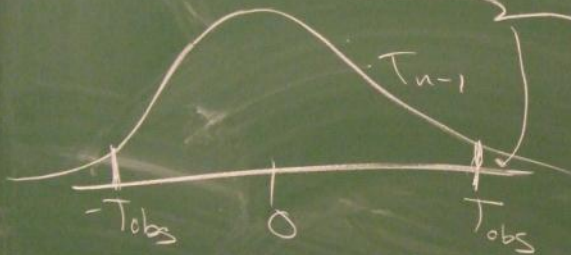
$$|T_{\text{obs}}| = \left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right| > t_{1-\frac{\alpha}{2}, n-1}$$

← dubbelsidigt



p-värdet för det dubbelsidiga testet:

$$P = 2P(T > |T_{\text{obs}}|) = 2P(T > \left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right|)$$

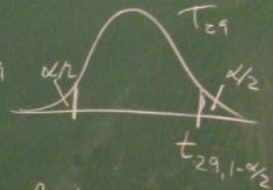


Ex Vi testar $H_0: \mu = 15$ mot $H_1: \mu \neq 15$

$X_1, \dots, X_n$, där $n=30$, ger oss $\bar{x}=14$, $s=3$
mätningar

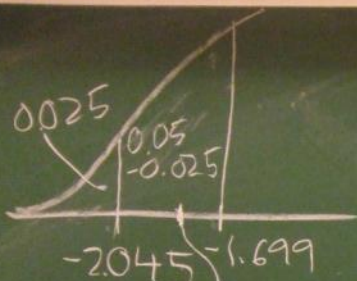
$$\text{Teststat.: } T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{\bar{X} - 15}{S/\sqrt{n}} \sim T_{30-1} = T_{29}$$

$$\text{Obs. test. stat.: } T_{\text{obs}} = \frac{14 - 15}{3/\sqrt{30}} = -1.83$$



För att finna p-värdet observerar vi följande.

$$P(T_{29} \leq -1.699) = 0.05 \quad P(T_{29} \leq -2.045) = 0.025$$



$$\Rightarrow 0.025 < P(T_{29} \leq -1.83) < 0.05$$

p-värdet ges av

$$\begin{aligned} p &= 2P(T > |T_{\text{obs}}|) = \\ &= 2P(T_{29} > |-1.83|) \\ &= 2P(T_{29} < -1.83) \end{aligned}$$

$$\Rightarrow 0.05 = 2 \cdot 0.025 < 2P(T_{29} < -1.83) < 2 \cdot 0.05 = 0.1$$

Anledningen till att vi gör på detta sätt är att $P(T_{29} \leq -1.83)$ ej finns i tabell

men det gör däremot $P(T_{29} \leq -1.699)$ och $P(T_{29} \leq -2.045)$!

Notera att vi antar att

$X_1, \dots, X_n$ kommer från en normalford.
Testen fungerar även då så ej är fallet
(enl. GTS) under förutsättn. att $n \geq 25$
(vi får ett approximativt test)

Konf. int. och test för σ

Erinran:
$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Sats

$X_1, \dots, X_n$ st. pr. från $N(\mu, \sigma^2)$

Den stok. var.

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n (X_i - \bar{X})^2 / \sigma^2$$

har en χ^2 -fördelning med $n-1$ frihetsgrader.

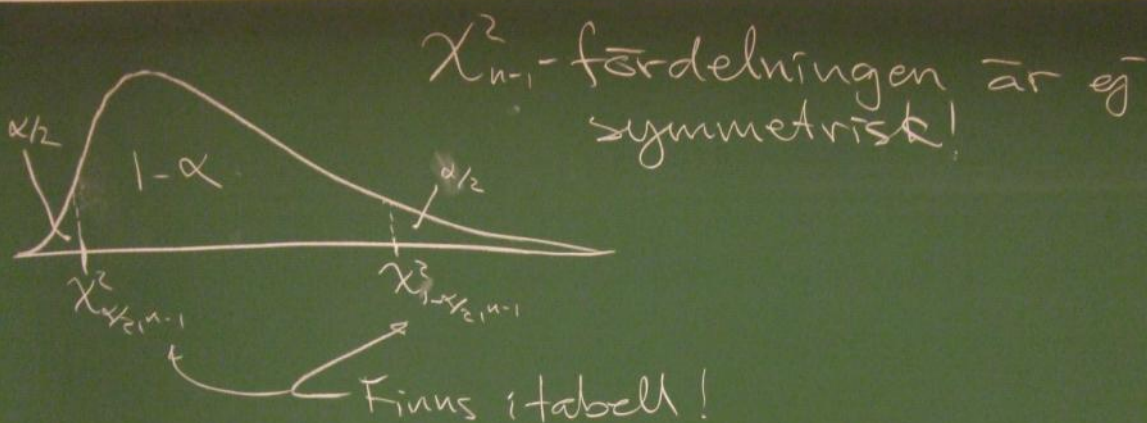
Detta används f.a. konstruera
konf. int. och test för variansen.

Sats

$X_1, \dots, X_n$ st.p.r. från $N(\mu, \sigma^2)$.

En $100(1-\alpha)\%$ konf. int. för σ^2
ges av $[L_1, L_2]$ där

$$L_1 = \frac{(n-1)S^2}{\chi^2_{1-\alpha/2, n-1}} \quad L_2 = \frac{(n-1)S^2}{\chi^2_{\alpha/2, n-1}}$$



Beweis

$$P\left(\chi^2_{\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{1-\alpha/2}\right) = 1-\alpha \text{ enl. sats}$$

$$P\left(\frac{1}{\chi^2_{1-\alpha/2}} \leq \frac{\sigma^2}{(n-1)S^2} \leq \frac{1}{\chi^2_{\alpha/2}}\right) = P\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\alpha/2}}\right)$$

Ex

$n=25$ observationer har gjorts

Vi fann att variansen skattas till

$$s^2 = 1.408$$

$\Rightarrow$ Vi betraktar en $\chi^2_{n-1} = \chi^2_{24}$ -fördeln.

Från tabell: $\chi^2_{0.025, 24} = 12.4$ $\chi^2_{0.975, 24} = 39.4$

Ett 95% konf. int. för σ^2 ges av $[L_1, L_2]$

$$\text{där } L_1 = \frac{(n-1)s^2}{\chi^2_{0.975, 24}} = \frac{24 \cdot 1.408}{39.4} = 0.858$$

$$L_2 = \frac{(n-1)s^2}{\chi^2_{0.025, 24}} = \frac{24 \cdot 1.408}{12.4} = 2.725$$

Test för variansen:

$X_1, \dots, X_n$ från $N(\mu, \sigma^2)$

$$H_0: \sigma^2 = \sigma_0^2$$

$$\text{Teststatistika: } T = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

H_1

$$\sigma^2 > \sigma_0^2$$

$$\sigma^2 < \sigma_0^2$$

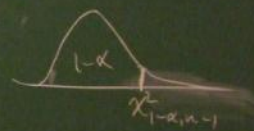
$$\sigma^2 \neq \sigma_0^2$$

Förk. H_0 om

$$\frac{(n-1)S^2}{\sigma_0^2} > \chi^2_{1-\alpha, n-1}$$

$$\frac{(n-1)S^2}{\sigma_0^2} < \chi^2_{\alpha, n-1}$$

$$\frac{(n-1)S^2}{\sigma_0^2} < \chi^2_{\alpha/2, n-1} \text{ eller } \frac{(n-1)S^2}{\sigma_0^2} > \chi^2_{1-\alpha/2, n-1}$$



LUNCHFÖRELÄSNING
MED ÅFS HÅLLBARHETS CHEF
ALICE BÄH

11/5, HÅ3

LÅT STA

Kap 9.1 - 9.2 läses på egen hand.

Jämförelse av 2 proportioner

Grupp 1: $X_1, \dots, X_{n_1}$ st. pr. $X_i \sim \text{Ber}(p_1)$

Grupp 2: $Y_1, \dots, Y_{n_2}$ st. pr. $Y_i \sim \text{Ber}(p_2)$ (möjligen $n_1 \neq n_2$)

De två stickpr. är oberoende.

Är intresserad av differensen $p_1 - p_2$ ← okända
Skatta med

$$\hat{p}_1 - \hat{p}_2 = \frac{\bar{X}}{n_1} - \frac{\bar{Y}}{n_2} = \frac{\sum_{i=1}^{n_1} X_i}{n_1} - \frac{\sum_{i=1}^{n_2} Y_i}{n_2}$$

Sats

Då n_1 och n_2 stora gäller att

$$\hat{p}_1 \overset{\text{app.}}{\sim} N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right), \quad \hat{p}_2 \overset{\text{app.}}{\sim} N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right) \text{ samt,}$$
$$\hat{p}_1 - \hat{p}_2 \overset{\text{app.}}{\sim} N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

□

Bevis

$\hat{p}_1$ och $\hat{p}_2$ är summor av oberoende likafördelade stok. var.

⇒ enl. CGS är de approx. normalford.

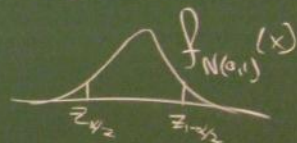
Linjärkombinationer av oberoende normalford. stok. var.
är normalford.

Notera att $E[\hat{p}_1 - \hat{p}_2] = p_1 - p_2$

dvs $\hat{p}_1 - \hat{p}_2$ är vur. för $p_1 - p_2$

Mha standardiseringssatsen

$$T = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \overset{\text{app.}}{\sim} N(0,1)$$



Sätter vi $P(z_{\alpha/2} \leq T \leq z_{1-\alpha/2}) = 1 - \alpha$

och löser ut så att vi får

$$P(L_1 \leq p_1 - p_2 \leq L_2) = 1 - \alpha$$

då har vi skapat det $100(1-\alpha)\%$ konf. int

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

(Notera att vi har bytt ut p_1 och p_2 mot $\hat{p}_1$ och $\hat{p}_2$)

En approx. intervall

Thumregel: stora strökor för att CGS skall gälla!

Ex: 2 patientgrupper

n_1 st. får en viss medicin

n_2 st. får placebo (sockerpiller)

Efter en tid kollas om förbättring skett hos patienterna.

två st. pr.

$$X_1, \dots, X_{n_1}, X_i \sim \text{Ber}(p_1) \quad (n_1 = 100)$$

$$Y_1, \dots, Y_{n_2}, Y_i \sim \text{Ber}(p_2) \quad (n_2 = 100)$$

($X_i = 1$ och $Y_i = 1$ innebär förbättring)

Resultaten blev

$$\hat{p}_1 = \frac{\sum_{i=1}^{n_1} x_i}{n_1} = \frac{27}{100}$$

$$\hat{p}_2 = \frac{\sum_{i=1}^{n_2} y_i}{n_2} = \frac{19}{100}$$

En 95% konf. int. för $p_1 - p_2$ ges av

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = 0.08 \pm 0.116$$

$z_{1-\frac{\alpha}{2}} = 1.96$

eller $[-0.036, 0.196]$

Viktigt: Då ett konf. int. innehåller

0 kan vi ej dra ngn
slutsats angående en skillnad

Hypotestest för $p_1 - p_2$

Vill testa $H_0: p_1 = p_2 = p$

dvs $H_0: p_1 - p_2 = 0$

Under H_0 gäller $\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}} \stackrel{H_0: p_1 = p_2}{=} \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}} \sim N(0,1)$

Vi skattar p med

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{\sum_{i=1}^{n_1} X_i + \sum_{i=1}^{n_2} Y_i}{n_1 + n_2}$$

Testet:

$$H_0: p_1 - p_2 = 0$$

$$\text{Test-stat.: } T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \stackrel{\text{app.}}{\sim} N(0,1)$$

$$\text{där } \hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{\sum_{i=1}^{n_1} X_i + \sum_{i=1}^{n_2} Y_i}{n_1 + n_2}$$

H_1

$$p_1 \neq p_2$$

$$p_1 > p_2$$

$$p_1 < p_2$$

Förk. H_0 ^{på sign. nivå α} (om

$$T_{\text{obs}} < -z_{\alpha/2} \text{ eller } T_{\text{obs}} > z_{1-\alpha/2}$$

$$|T_{\text{obs}}| > z_{1-\alpha/2}$$

$$T_{\text{obs}} > z_{1-\alpha}$$

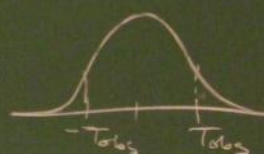
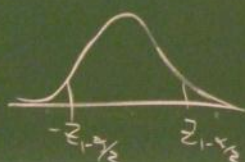
$$T_{\text{obs}} < -z_{1-\alpha}$$

p-värde

$$2(1 - \Phi(|T_{\text{obs}}|))$$

$$1 - \Phi(|T_{\text{obs}}|)$$

$$1 - \Phi(|T_{\text{obs}}|)$$



Ex Patientgrupperna från tidigare exempel.

Vi testar $H_0: p_1 = p_2$ mot $H_1: p_1 \neq p_2$
på $\alpha = 0.05$ nivå.

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \stackrel{\text{app}}{\sim} N(0,1)$$

Från tidigare: $n_1 = n_2 = 100$, $\hat{p}_1 = 0.27$, $\hat{p}_2 = 0.19$

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{27 + 19}{200} = 0.23$$

$$\text{Obs. teststat: } T_{\text{obs}} = \frac{0.27 - 0.19}{\sqrt{0.23(1-0.23)\left(\frac{1}{100} + \frac{1}{100}\right)}} = 1.35$$

Vi skall förk. H_0 om $|T_{\text{obs}}| > Z_{1-\alpha/2} = Z_{0.975} = 1.96$

$\therefore$ Vi kan ej förk. H_0 på $\alpha = 0.05$ nivå

$$\text{p-värdet} = 2(1 - \Phi(|T_{\text{obs}}|)) = 2(1 - \Phi(1.35)) = 2(1 - 0.9115) = 0.177$$

$\underbrace{\hspace{1.5cm}}_{\substack{P(Z \leq |T_{\text{obs}}|) \\ \sim N(0,1)}} \quad \quad \quad \underbrace{\hspace{1.5cm}}_{\text{tabell}}$

Dvs, minsta möjliga signifikansnivå (för förkastande)
är $\alpha = 0.177$

Kap 10

$$X_1^{(1)}, \dots, X_{n_1}^{(1)}$$

$$X_1^{(2)}, \dots, X_{n_2}^{(2)}$$

2 obero. st. pr.

$$E[X_i^{(1)}] = \mu_1$$

$$E[X_i^{(2)}] = \mu_2$$

Vi skattar $\mu_1 - \mu_2$ genom $\hat{\mu}_1 - \hat{\mu}_2 = \bar{X}_1 - \bar{X}_2$

$$\text{där } \bar{X}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i^{(1)} \quad \bar{X}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} X_i^{(2)}$$

Sats

Om $X_1^{(1)}, \dots, X_{n_1}^{(1)}$ st. pr. från $N(\mu_1, \sigma_1^2)$ och

$X_1^{(2)}, \dots, X_{n_2}^{(2)}$ st. pr. från $N(\mu_2, \sigma_2^2)$ då gäller

$$\text{att } \bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Beweis

$$\bar{X}_1 - \bar{X}_2 = \sum_{i=1}^{n_1} \frac{1}{n_1} X_i^{(1)} - \sum_{i=1}^{n_2} \frac{1}{n_2} X_i^{(2)}$$

är en linjärkombination av obero. normalfördelade stok. var.

$$\Rightarrow \bar{X}_1 - \bar{X}_2 \sim N\left(\underbrace{E[\bar{X}_1 - \bar{X}_2]}_{=\mu_1 - \mu_2}, \underbrace{\text{Var}(\bar{X}_1 - \bar{X}_2)}_{=\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

$$\text{Var}(\bar{X}_1) = \text{Var}\left(\frac{1}{n_1} \sum_{i=1}^{n_1} X_i^{(1)}\right) \stackrel{\text{obero.}}{=} \frac{1}{n_1^2} \sum_{i=1}^{n_1} \underbrace{\text{Var}(X_i^{(1)})}_{=\sigma_1^2} = \frac{1}{n_1^2} n_1 \sigma_1^2 = \frac{\sigma_1^2}{n_1}$$

Om st.proven ej är normalford.
så går resultatet att använda
approximativt, givet att $n_1, n_2 \geq 25$

$$\bar{X}_1 - \bar{X}_2 \approx N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

Om vi nu kan antaga att $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$N(0,1) \sim \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Skatta σ_1^2 och σ_2^2 med S_1^2 resp. S_2^2

Under antagandet $\sigma_1^2 = \sigma_2^2 = \sigma^2$
slås skattningarna samman enligt

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

(notera: $(n_1 - 1)S_1^2 = \sum_{i=1}^{n_1} (x_i^{(1)} - \bar{x})^2$)

Det gäller att

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim T_{n_1 + n_2 - 2}$$

$\Rightarrow$

En $100(1-\alpha)\%$ konf. int. för $\mu_1 - \mu_2$ då $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$(\bar{X}_1 - \bar{X}_2) \pm t_{1-\alpha/2, n_1+n_2-2} \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

Test för $\mu_1 - \mu_2$ då $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Testa $H_0: \mu_1 = \mu_2$ mot H_1 (se nedan)

$$\text{Teststat: } T = \frac{(\bar{X}_1 - \bar{X}_2) - \overbrace{(\mu_1 - \mu_2)}^{=0 \text{ under } H_0}}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim T_{n_1 + n_2 - 2}$$

$$\begin{array}{l} H_1 \\ \mu_1 - \mu_2 > 0 \end{array}$$

$$\begin{array}{l} \text{Förk. } H_0 \text{ om} \\ T_{\text{obs}} > t_{1-\alpha, n_1+n_2-2} \end{array}$$

$$\begin{array}{l} \text{p-värde} \\ P(T_{n_1+n_2-2} > T_{\text{obs}}) \end{array}$$

$$\mu_1 - \mu_2 < 0$$

$$T_{\text{obs}} < -t_{1-\alpha, n_1+n_2-2}$$

$$P(T_{n_1+n_2-2} < -T_{\text{obs}})$$

$$\mu_1 \neq \mu_2$$

$$|T_{\text{obs}}| > t_{1-\alpha/2, n_1+n_2-2}$$

$$2P(T_{n_1+n_2-2} > |T_{\text{obs}}|)$$

$$\mu_1 - \mu_2 \leq 0$$

$$T_{obs} \leftarrow t_{obs}$$

Sist

2 oberoende stickpr.

$$X_1^{(1)}, \dots, X_{n_1}^{(1)}$$

$$E[X_i^{(1)}] = \mu_1$$

$$X_1^{(2)}, \dots, X_{n_2}^{(2)}$$

$$E[X_i^{(2)}] = \mu_2$$

$$\mu_1 - \mu_2 \text{ skattas med } \widehat{\mu_1 - \mu_2} = \bar{X}_1 - \bar{X}_2$$

Vi förut sätter kont. int

$$(\bar{X}_1 - \bar{X}_2) \pm t_{1-\alpha/2, n_1+n_2-2} \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$\text{där } S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

$$\frac{1}{n_1} \sum_{i=1}^{n_1} X_i^{(1)} \quad \frac{1}{n_2} \sum_{i=1}^{n_2} X_i^{(2)}$$

Test:

$H_0: \mu_1 = \mu_2$ mot H_1

$$\text{Teststat.: } T = \frac{\bar{X}_1 - \bar{X}_2 - \overbrace{(\mu_1 - \mu_2)}^{=0 \text{ under } H_0}}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim T_{n_1+n_2-2}$$

H_1

$\mu_1 - \mu_2 > 0$

$\mu_1 - \mu_2 \leq 0$

$\mu_1 \neq \mu_2$

Förk. H_0 om

$$T_{\text{obs}} > t_{1-\alpha, n_1+n_2-2}$$

$$T_{\text{obs}} < -t_{1-\alpha, n_1+n_2-2}$$

$$|T_{\text{obs}}| > t_{1-\frac{\alpha}{2}, n_1+n_2-2}$$

Ovanstående gäller under antagandet $\sigma_1^2 = \sigma_2^2 = \sigma^2$

och därtill får vi

(kom ihåg n_1)

Om vi ej kan antaga att $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Vi vet att

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0, 1)$$

(dä $X_{1,1}, \dots, X_{1,n_1}$ från $N(\mu_1, \sigma_1^2)$
 $X_{2,1}, \dots, X_{2,n_2}$ från $N(\mu_2, \sigma_2^2)$)

och därtill får vi

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim T_8$$

(kom ihåg

$$S_1^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (X_{1,i} - \bar{X}_1)^2$$
$$S_2^2 = \frac{1}{n_2-1} \sum_{i=1}^{n_2} (X_{2,i} - \bar{X}_2)^2$$

Vi skattar γ från data

$$\gamma = \left[\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}} \right]$$

$\lfloor x \rfloor$ = största heltal $\leq x$ t.ex. $\lfloor 3.5 \rfloor = 3$

Vi får samma konf. int. och test
som då $\sigma_1^2 = \sigma_2^2 = \sigma^2$ med den skillnaden
att $n_1 + n_2 - 2$ byts ut mot γ och $\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$
byts ut mot $\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$.

Det verkar ha skett en förändring
(indikationer på att $\mu_2 < \mu_1$)

När antar vi att $\sigma_1^2 = \sigma_2^2 = \sigma^2$?

Om experiment upplägget är detsamma
i båda stickproven och S_1^2 och S_2^2
ej avviker mycket från varandra, då
kan vi antaga att $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

I övriga fall antar vi att $\sigma_1^2 \neq \sigma_2^2$.

Ex

Ex

Exponering för radioaktivitet i gruva.

1973

$$n_1 = 16$$

$$\bar{X}_1 = 0.94 \text{ rem}$$

$$S_1^2 = 0.04$$

1979

$$n_2 = 16$$

$$\bar{X}_2 = 0.62 \text{ rem}$$

$$S_2^2 = 0.028$$

Liknande experimentförhållanden
och ej för stor skillnad i S_1^2 och S_2^2

$$\Rightarrow \text{antar } \sigma_1^2 = \sigma_2^2 = \sigma^2$$

Vill finna ett 95% konf. int. för $\mu_1 - \mu_2$.

$$\underbrace{\bar{X}_1 - \bar{X}_2}_{= 0.94 - 0.62 = 0.32} \pm \underbrace{t_{1-\frac{\alpha}{2}, n_1+n_2-2}}_{= t_{0.975, 30} = 2.042 \text{ från tabell}} \sqrt{\underbrace{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}_{= 2/16 = \frac{1}{8}}} = 0.32 \pm 0.13$$

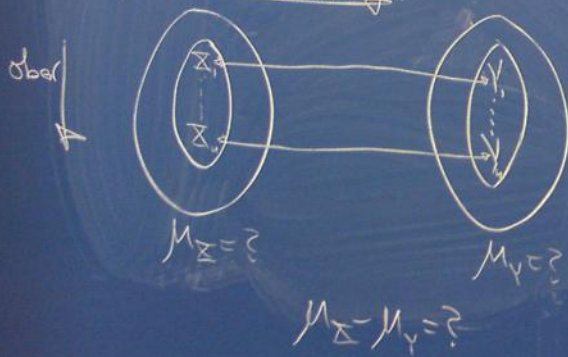
$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{15 \cdot 0.04 + 15 \cdot 0.028}{16+16-2} = 0.034$$

eller $[0.19, 0.45]$

Det verkar ha skett en förändring
(indikationer på att $\mu_2 < \mu_1$)

Notera att $0 \notin [0.19, 0.45]$

Jämförelse av väntevärden:
parade observationer (parat t-test)



Parvisa observationer
 $(X_1, Y_1), \dots, (X_n, Y_n)$
där X_i och Y_i ej ober.

Upplägg:

Vi ser till differenserna

$$D_i = X_i - Y_i, \quad i = 1, \dots, n$$

↑ Dessa antas vara ett st. pr. från $N(\mu_D, \sigma_D^2)$

$$\mu_D = \mu_X - \mu_Y = E[D_i] = E[X_i] - E[Y_i]$$

Konf. int. för μ_D (okänd σ_D^2), $100(1-\alpha)\%$:

$$\frac{\bar{D} - \mu_D}{S_D / \sqrt{n}} \sim T_{n-1}$$

$$\bar{D} = \frac{1}{n} \sum_{i=1}^n D_i, \quad S_D^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2$$

$\Rightarrow$ Konf. int.:

$$\bar{D} \pm t_{1-\frac{\alpha}{2}, n-1} \frac{S_D}{\sqrt{n}}$$

(Var. konf. för μ för ett normalford. st. pr.)

Test för $\mu_D = \mu_x - \mu_y$ (Var. en-st.pr. t-test)

$H_0: \mu_D = 0$ mot H_1

$$\text{Teststat.: } T = \frac{\bar{D} - \tilde{\mu}_D^0}{S_D/\sqrt{n}} \sim T_{n-1}$$

H_1

$\mu_D \neq 0$

$\mu_D > 0$

$\mu_D < 0$

Förk. H_0 om

$$|T_{\text{obs}}| > t_{1-\frac{\alpha}{2}, n-1}$$

$$|T_{\text{obs}}| > t_{1-\alpha, n-1}$$

— 1 —

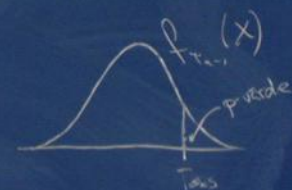


p-värde

$$2(1 - P(T_{n-1} \leq |T_{\text{obs}}|))$$

$$1 - P(T_{n-1} \leq |T_{\text{obs}}|)$$

$$1 - P(T_{n-1} \leq |T_{\text{obs}}|)$$

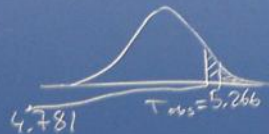


Ex Tid det tar att skicka info mellan 2 datorer med 2 olika kablar.

Datorpar	Tid kabel A (x_i)	Tid kabel B (y_i)	Diff. (d_i)
1	8.05	0.71	7.34
2	24.74	0.74	24.00
3	:	:	27.59
4	:	:	7.68
5	:	:	8.39
6	:	:	24.37
7	:	:	13.23
8	:	:	19.56
9	:	:	4.11
10	8.54	0.72	7.82

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i = 14.409$$

$$s_D = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2} = 8.653$$



Vi testar $H_0: \mu_D = \mu_A - \mu_B = 0$ mot $H_1: \mu_D = \mu_A - \mu_B > 0$

$$\text{Vi får } T_{\text{obs}} = \frac{\bar{d} - \mu_0}{s_D / \sqrt{n}} = \frac{14.409}{8.653 / \sqrt{10}} = 5.266$$

$$p\text{-värdet} = 1 - P(T_{10-1} \leq |T_{\text{obs}}|) = 1 - P(T_9 \leq 5.266) < \text{småst av i tabellen}$$

$$= 1 - P(T_9 \leq 4.781) = 1 - 0.9995 = 0.0005$$

minsta värdet i tabellen

Mycket litet p-värde $\Rightarrow$ vi förkastar H_0 !

Linjär regression

Vi vill undersöka beroendet
för ett par (X, Y)

Vi vill veta var Y förväntas ligga
lä vi vet att $X=x$.

Kalla detta väntevärde $\mu_{Y|x} = E[Y|X=x]$

Ex X = havsdjup Y = temperatur (i vattnet)
 X och Y har ett beroende

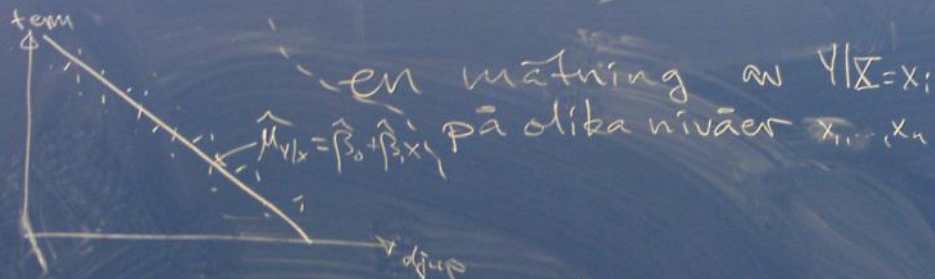
När vi gör mätningar vet vi vilket djup
vi mäter på, dvs vi fixerar $X = x_i, i = 1, \dots, n$

Grafen för $\mu_{Y|x}$ (betraktad som en funktion av x)
kallas regressionskurvan

Y kallas responsvariabel/beroende variabel

X kallas förklarande variabel/prediktorvariabel

(ibland kallas (fixerad) X för
oberoende variabel).



Vi vill skatta $\mu_{Y|X}$ från observationerna
 $(x_1, Y|X_1), \dots, (x_n, Y|X_n)$

Om sambandet $\mu_{Y|X}$ är linjärt får vi

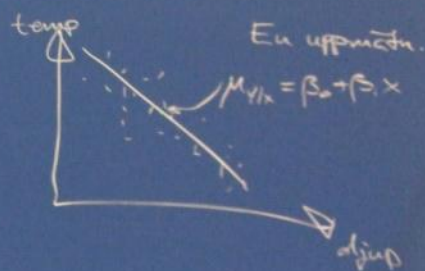
$$\mu_{Y|X} = \beta_0 + \beta_1 x, \quad \beta_0, \beta_1 \in \mathbb{R}$$

Detta kallas enkel linjär regression.

Sist Simpel linjär regression (X, Y) uppmätt för olika värden $x_1, \dots, x_n$ på X $\Rightarrow Y|X=x_i$ Vill beskriva $\mu_{Y|x_i} = E[Y|X=x_i]$

Simpel linjär regression:

$$\mu_{Y|x} = \beta_0 + \beta_1 x, \quad \beta_0, \beta_1 \in \mathbb{R}$$

dvs sambandet kan beskrivas
linjärt

β_0 : den punkt där $\mu_{Y|X}$ skär y-axeln

β_1 : lutningen på linjen $\mu_{Y|X}$

Vill skatta $\mu_{Y|X}$ från observationer, dvs skatta β_0 och β_1 .

Låt $E_i = \overbrace{Y|X=x_i}^{\text{stok.}} - \mu_{Y|X}$ vara (den stokastiska)

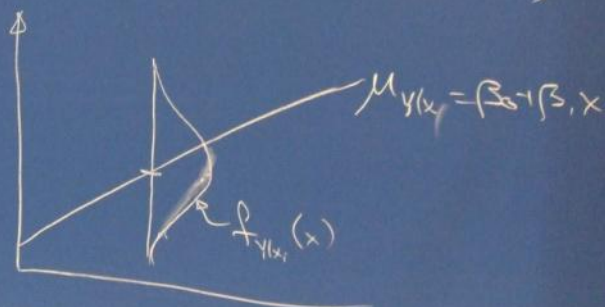
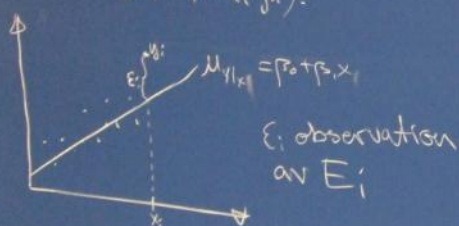
differensen mellan (den stokastiska) observationen i x_i och dess väntevärde

Omflyttning:

$$Y_i \equiv Y|X=x_i = \mu_{Y|X} + E_i = \beta_0 + \beta_1 x_i + E_i, \quad E[E_i] = 0$$

$$E[E_i] = \underbrace{E[Y|X=x_i]}_{=\mu_{Y|X}} - \underbrace{E[\mu_{Y|X}]}_{=\mu_{Y|X}} = 0$$

Vi kallar observationerna av (x_i, Y_i) , ..., (x_n, Y_n) för $(x_1, y_1), \dots, (x_n, y_n)$.



Hur skattar man β_0 och β_1 från data?

Kalla dessa skattare/skattningar
för b_0 och b_1 , så att $\hat{\mu}_{y|x} = b_0 + b_1 x$

Minsta kvadrat-metoden!

Differenserna

$$e_i = y_i - \hat{\mu}_{y|x_i} = y_i - b_0 - b_1 x_i$$

$\uparrow$
observation

kallas residualer



Minsta kvadr. skattarna finnes enl.

Finn b_0 och b_1 s.a.

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

minimeras. (vi minimerar över alla rätta linjer $b_0 + b_1 x$)

Detta görs genom lösning av

$$\begin{cases} 0 = \frac{\partial SSE}{\partial b_0} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) \\ 0 = \frac{\partial SSE}{\partial b_1} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) x_i \end{cases}$$

Delta ger oss

$$b_1 = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$\uparrow$ $\frac{1}{n} \sum_{i=1}^n y_i$ $\uparrow$ $\frac{1}{n} \sum_{i=1}^n x_i$

ofta skrivs $\hat{y}_i = b_0 + b_1 x_i$

Ex $(x_1, y_1), \dots, (x_n, y_n)$ är observationer

$$n = 25$$

$$\sum x_i = 1314.90$$

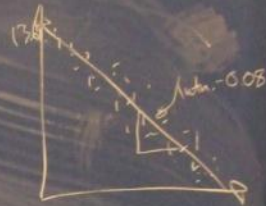
$$\sum y_i = 235.70$$

$$\sum x_i^2 = 76308.53$$

$$\sum y_i^2 = 2286.07$$

$$\sum x_i y_i = 11824.44$$

$$\hat{\beta}_1 = b_1 = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2} \stackrel{\text{insätt}}{=} -0.08$$



$$\hat{\beta}_0 \pm b_0 = \bar{y} - b_1 \bar{x} = \frac{1}{25} 235.70 - (-0.08) \frac{1}{25} 134.90 = 13.64$$

$$\Rightarrow \hat{M}_{y|x} = \hat{y} = 13.64 - 0.08x$$

Vill vi skatta var y i medel ^{ligger} då $\bar{x} = 50$ får vi

$$\hat{M}_{y|50} = 13.64 - 0.08 \cdot 50 = 9.64$$

En störsam $(x_1, y_1), \dots, (x_n, y_n)$ är observationer av $(x_1, Y_1), \dots, (x_n, Y_n)$ är b_0 och b_1 (skattningar) observationer av

$$B_1 = \hat{\beta}_1 = \frac{n \sum_{i=1}^n x_i Y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n Y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

$$B_0 = \hat{\beta}_0 = \bar{Y} - B_1 \bar{x}$$

Skallare (dvs stok. var.) som har ngn fördeln.

Vad har β_0 och β_1 för fördelning?

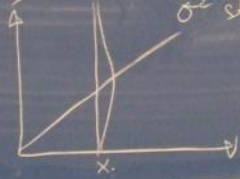
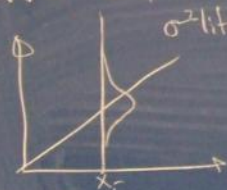
Mha dessa kan skapa intervall och test för β_0 och β_1 .

Måste säga ngt om $E_1, \dots, E_n$ ($Y_i = \beta_0 + \beta_1 x_i + E_i$)

Vet att $E[E_i] = 0$.

Antag att $E_1, \dots, E_n$ är ett stickpr. från $N(0, \sigma^2)$ -förd,

dvs $E_i \sim N(0, \sigma^2)$ är alla ober., $i=1, \dots, n$.



stor σ^2 innebär mycket fluktuation kring μ_{ex}

Sats

$$\beta_1 \sim N\left(\beta_1, \sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2\right)$$

Notera: $E[\beta_1] = \beta_1$, dvs β_1 är vur.

Sats

$$\beta_0 \sim N\left(\beta_0, \frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} \sigma^2\right)$$

Notera: β_0 är vur för β_0 .

Vi känner ej till σ^2 så vi skattar den med

$$\hat{\sigma}^2 = S^2 = \frac{SSE}{n-2} = \frac{1}{n-2} \sum_{i=1}^n (Y_i - B_0 - B_1 x_i)^2$$

som är vår för σ^2 (Ans $E[S^2] = \sigma^2$)

Förenklad notation:

$$S_{xx} = \sum (x_i - \bar{x})^2 = (n \sum x_i^2 - (\sum x_i)^2) / n$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = (n \sum y_i^2 - (\sum y_i)^2) / n$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = (n \sum x_i y_i - (\sum x_i)(\sum y_i)) / n$$

$$SSE = \sum_{i=1}^n (Y_i - B_0 - B_1 x_i)^2 = \sum E_i^2$$

$$\Rightarrow B_1 = \frac{S_{xy}}{S_{xx}}$$

$$B_0 = \bar{y} - B_1 \bar{x} = \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x}$$

$$SSE = S_{yy} - B_1 S_{xy}$$

Finns tillräckliga bevis för att $\beta_1 \neq 0$?

Om $\beta_1 = 0 \Rightarrow Y_i = \beta_0 + E_i$ (slumpmässig fluktuation)
 kring linjen $y = \beta_0$
 $\sim N(\beta_0, \sigma^2)$

Från sats får vi

$$\frac{B_1 - E[B_1]}{\sqrt{\text{Var}(B_1)}} = \frac{B_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0,1)$$

Byter vi ut σ mot $S = \sqrt{\frac{SSE}{n-2}}$ får vi

$$\frac{B_1 - \beta_1}{S/\sqrt{S_{xx}}} \sim T_{n-2}$$

Test för $\beta_1 = 0$

$H_0: \beta_1 = \beta_1^0$ (t.ex. $\beta_1^0 = 0$, dvs X och Y påverkar varandra ej)

Teststat.: $T = \frac{B_1 - \beta_1^0}{S/\sqrt{S_{xx}}} \sim T_{n-2}$

observ. testst.: $T_{\text{obs}} = \frac{b_1 - \beta_1^0}{s/\sqrt{S_{xx}}}$

H_1

$\beta_1 \neq \beta_1^0$

$\beta_1 > \beta_1^0$

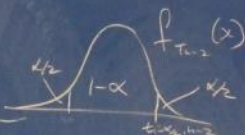
$\beta_1 < \beta_1^0$

Förb. H_0 om

$|T_{\text{obs}}| > t_{1-\alpha/2, n-2}$

$|T_{\text{obs}}| > t_{1-\alpha, n-2}$

$|T_{\text{obs}}| > t_{1-\alpha, n-2}$



P-värden

$$2P(T_{n-2} > |T_{\text{obs}}|) = P(1 - P(T_{n-2} \leq |T_{\text{obs}}|))$$

$$1 - P(T_{n-2} \leq |T_{\text{obs}}|)$$

$$1 - P(T_{n-2} \leq |T_{\text{obs}}|)$$

Ex 100(1- α)% konf. int. för β_1

$$\hat{\beta}_1 \pm t_{1-\alpha/2, n-2} \frac{S}{\sqrt{S_{xx}}}$$

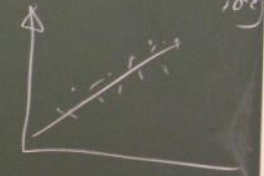
SIST Test för $H_0: \beta_1 = \beta_1^0$ (typiskt $\beta_1^0 = 0$)

$$\hat{M}_{y|x} = \hat{\beta}_0 + \hat{\beta}_1 x$$

skattaren för β_1 $y_i = \beta_0 + \beta_1 x_i + E_i$
 $E_i \sim N(0, \sigma^2)$

Teststat.: $T = \frac{\hat{\beta}_1 - \beta_1^0}{S / \sqrt{S_{xx}}} \sim T_{n-2}$

$$\hat{\sigma}^2 = S^2 = \frac{SSE}{n-2} = \frac{1}{n-2} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$



Mothypoteser, $H_1: \beta_1 \neq \beta_1^0$ $H_1: \beta_1 > \beta_1^0$ $H_1: \beta_1 < \beta_1^0$

Ex

$$n=25 \quad \sum x_i = 1314.9 \quad \sum y_i = 235.7$$

$$\Rightarrow \sum x_i^2 = 76308.53 \quad \sum y_i^2 = 2286 \quad \sum x_i y_i = 11824.44$$

$$\hat{M}_{y|x} = \underbrace{13.64}_{\hat{\beta}_0} - \underbrace{0.08x}_{\hat{\beta}_1}$$

Testa $H_0: \beta_1 = 0$ mot $H_1: \beta_1 \neq 0$

För att finna T_{obs} måste vi finna SSE

$$S_{xx} = \sum (x_i - \bar{x})^2 = (n \sum x_i^2 - (\sum x_i)^2) / n = 7150.05$$

$$S_{yy} = (n \sum y_i^2 - (\sum y_i)^2) / n = 63.89$$

$$S_{xy} = (n \sum x_i y_i - (\sum x_i)(\sum y_i)) / n = -572.44$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n \bar{x} \bar{y} = \sum x_i y_i - \frac{(\sum x_i)^2}{n} - \frac{(\sum y_i)^2}{n} + n \bar{x} \bar{y}$$

$$\Rightarrow SSE = S_{yy} - b \cdot S_{xy} = 63.89 - (-0.08)(-572.44) = 18.09$$

$$S^2 = \frac{SSE}{n-2} = \frac{18.09}{23} = 0.79$$

$$\Rightarrow T_{obs} = \frac{b_1 - \beta_0}{S / \sqrt{S_{xx}}} = \frac{-0.08}{\sqrt{0.79} / \sqrt{7150.05}} = -7.62$$

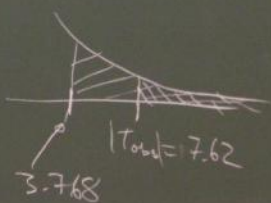
Teststat. $T = \frac{B_1 - \beta_0}{S / \sqrt{S_{xx}}} \sim T_{n-2} = T_{23}$

p-värdet $= 2P(T_{23} > |T_{obs}|) =$ Finns ej i T-tabell

$$= 2(1 - P(T_{23} \leq |-7.62|)) <$$

$$< 2(1 - \underbrace{P(T_{23} \leq 3.768)}_{= 0.9995}) = 0.001$$

$\therefore$ Litet p-värde!
Vi kan fork. H_0 (åtminstone)
på signifikansnivån 0.001



Från förra föreläsning. vet vi att

$$B_0 \sim N\left(\beta_0, \frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} \sigma^2\right)$$

$\uparrow$ β_0 's skattare

$$\Rightarrow \frac{B_0 - E[B_0]}{\sqrt{\text{Var}(B_0)}} = \frac{B_0 - \beta_0}{\sigma \sqrt{\sum x_i^2} / \sqrt{n S_{xx}}} \sim N(0,1)$$

$$\Rightarrow \frac{B_0 - \beta_0}{S \sqrt{\sum x_i^2} / \sqrt{n S_{xx}}} \sim T_{n-2}$$

Test $H_0: \beta_0 = \beta_0^0$ (typiskt $\beta_0^0 = 0$)

$$\text{Teststat.} : T = \frac{B_0 - \beta_0^0}{S \sqrt{\sum x_i^2} / \sqrt{n S_{xx}}} \sim T_{n-2}$$

Förkastn. regioner och p-värden för de olika mothypoteserna ($H_1: \beta_0 \neq \beta_0^0$, $H_1: \beta_0 > \beta_0^0$, $H_1: \beta_0 < \beta_0^0$) finnes på samma sätt som övriga t-fördeln. fall ($n-2$ frihetsgr.)

100(1- α)% konf. int. för β_0

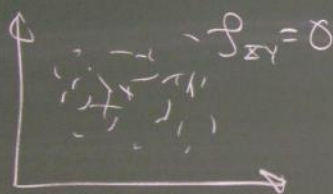
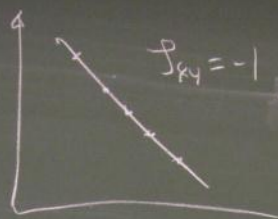
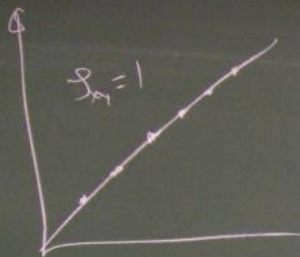
$$B_0 \pm t_{1-\alpha/2, n-2} \frac{S \sqrt{\sum x_i^2}}{\sqrt{n S_{xx}}}$$

Korrelation

Antag nu att (X, Y) är en 2-dim
stok-var.

Vi är intresserade av styrkan
på det linjära beroendet/sambandet
mellan X och Y (snarare än att beskriva
det som $\mu_{Y|X=x} = \beta_0 + \beta_1 x$ gör). Detta görs med

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}} \quad -1 \leq \rho_{XY} \leq 1$$



Då vi har observationer $(x_1, y_1), \dots, (x_n, y_n)$ av $(X, Y), \dots, (X_n, Y_n)$ (ett stickpr. från (X, Y) 's fördeln.) skattar vi ρ_{XY} .

$$\widehat{\text{Var}}(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{S_{XX}}{n}$$

$$\widehat{\text{Var}}(Y) = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{S_{YY}}{n}$$

$$\widehat{\text{Cov}}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{S_{XY}}{n}$$

$$\Rightarrow R = \hat{\rho}_{XY} = \frac{\widehat{\text{Cov}}(X, Y)}{\sqrt{\widehat{\text{Var}}(X)} \sqrt{\widehat{\text{Var}}(Y)}} = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$$

↑ Pearson's korrelationskoefficient

Ex

$$n=10 \quad \sum x_i^2 = 900775 \quad \sum y_i^2 = 919489$$

$$\sum x_i = 2405 \quad \sum y_i = 2503 \quad \sum x_i y_i = 902475$$

$$\Rightarrow S_{xx} = 322372.5$$

$$S_{yy} = 292988$$

$$S_{xy} = 300503.5$$

$$\Rightarrow \hat{\rho} = r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = 0.978$$

$\therefore$ starkt linjärt
beroende!

Determinationskoefficienten

$$\text{ges av } R^2 = \frac{S_{yy} - SSE}{S_{yy}}$$

↳ Korrelationen kvadrerad
Kan utläsas

S_{yy} mäter totala variationen i Y

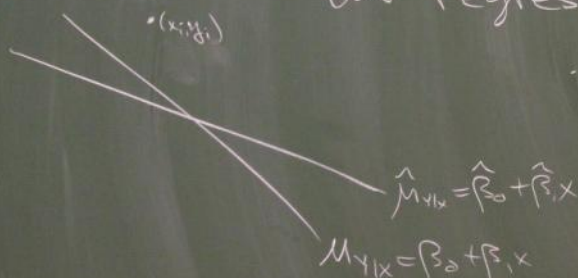
$$(S_{yy} = \sum (y_i - \bar{y})^2)$$

SSE mäter den slumpmässiga variationen
(i Y kring den skattade regressionslinjen)

$$\sum E_i^2 = \sum (y_i - \hat{y}_i)^2$$

$\Rightarrow S_{yy} - SSE$ mäter variationen i Y som kan förklaras av regressionsmodellen

$\Rightarrow R^2 = \frac{S_{yy} - SSE}{S_{yy}}$ mäter andelen av totala variationen i Y som förklaras av regressionsmodellen



Stort R^2 innebär att modellen passar väl

Residualanalys

Residualerna

$$e_i = y_i - (b_0 + b_1 x_i) = y_i - \hat{y}_i$$

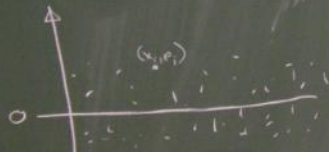
kan användas för att bedömma om regr. modellen är lämplig.

Plottar man paren $(x_1, e_1), \dots, (x_n, e_n)$ får man en residualplott.

I denna ser man:

1) om $\mathbb{E}[E_i] = 0$ (feltermerna $(x_1, e_1), \dots, (x_n, e_n)$ bör sprida sig runt 0)

2) om $\text{Var}(E_i) = \sigma^2$ konstant (dvs ej beror på x_i)
(punkterna sprider sig lika mycket genom plotten)



$E_1, \dots, E_n$ st.pr. från $N(0, \sigma^2)$



↑ modellen ej lämplig!