Föreläsning 16: Repetition och kommentarer

Matematisk statistik

David Bolin
Chalmers University of Technology
Oktober 24, 2018

Faktorförsök

- Till exempel för ett 2^3-försök kan modellen då skrivas som

$$\begin{aligned} Y_{ijkl} =& \mu + Ax_i + Bx_j + Cx_k + ABx_ix_j \\ &+ ACx_ix_k + BCx_jx_k + ABCx_ix_jx_k + \varepsilon_{ijkl} \end{aligned}$$

  med $\varepsilon_{ijkl} \sim \mathsf{N}(0, \sigma^2)$.
- Här är A, B, C huvudeffekterna för de tre faktorerna.
- AB, AC, BC, ABC är samspelseffekterna.
- Dessa effekter skattas nu som för vanligt multipel regression.
- Hur parameterskattningarna ser ut kan sammanfattas i ett teckenschema.

Faktorförsök

- En vanlig situation är att man vill göra experiment för att undersöka vilka faktorer som påverkar en viss process.
- Man vill veta om faktorerna påverkar processen och om det finns några samspelseffekter mellan faktorerna.
- En enkel uppställning för att undersöka detta är att göra experiment där varje faktor testas dels på en låg nivå och dels på en hög nivå.
- Ett sådant försök med k faktorer kallas ett 2^k-faktorförsök. Försöket sägs ha n replikat om varje faktorkombination mäts n gånger.
- För att skatta effekterna formulerar vi modellen som en regressionsmodell: För varje faktor, inför en variabel som är 1 om faktorn har hög nivå och -1 annars.

Teckenschema för ett 2^3-försök

Försök	Medel	μ	A	B	C	AB	AC	BC	ABC
(1)	$\bar{y}_{111}$	+	-	-	-	+	+	+	-
a	$\bar{y}_{211}$	+	+	-	-	-	-	+	+
b	$\bar{y}_{121}$	+	-	+	-	-	+	-	+
ab	$\bar{y}_{221}$	+	+	+	-	+	-	-	-
c	$\bar{y}_{112}$	+	-	-	+	+	-	-	+
ac	$\bar{y}_{212}$	+	+	-	+	-	+	-	-
bc	$\bar{y}_{122}$	+	-	+	+	-	-	+	-
abc	$\bar{y}_{222}$	+	+	+	+	+	+	+	+

- För en given effekt θ har vi $\mathsf{d}(\hat{\theta}) = s/\sqrt{2^k n}$.
- $I_\theta = [\hat{\theta} \pm t_{\alpha/2}(2^k(n-1))\mathsf{d}(\hat{\theta})]$.

Yates schema

Vid räkning för hand på ett 2^k-försök kan beräkningarna underlättas avsevärt genom att använda Yates schema:

1. Skriv effekterna i standardordning i en kolumn (tex för $k = 3$: (1), a, b, ab, c, ac, bc, abc)
2. Skriv ner motsvarande medelvärden i en ny kolumn.
3. Bilda en ny kolumn (1) där första hälften av värdena är alla parsummor av värden i föregående kolumn, och där andra hälften av värdena är alla pardifferenser.
4. Bilda en ny kolumn (2) genom att upprepa Steg 3, fortsätt med detta tills kolumn (k) bildats.
5. Dela talen i kolumn (k) med $2^k n$ för att få skattningarna.

Yates schema (exempel)

Steg 1:

Försök	Medel	(1)	(2)	(3)	Effekt	Skattning
(1)	522	1068	-	-	-	-
a	546	-	-	-	-	-
b	557	-	-	-	-	-
ab	581	-	-	-	-	-
c	567	-	-	-	-	-
ac	579	-	-	-	-	-
bc	597	-	-	-	-	-
abc	609	-	-	-	-	-

Steg 2:

Försök	Medel	(1)	(2)	(3)	Effekt	Skattning
(1)	522	1068	-	-	-	-
a	546	1138	-	-	-	-
b	557	-	-	-	-	-
ab	581	-	-	-	-	-
c	567	-	-	-	-	-
ac	579	-	-	-	-	-
bc	597	-	-	-	-	-
abc	609	-	-	-	-	-

Steg 3:

Försök	Medel	(1)	(2)	(3)	Effekt	Skattning
(1)	522	1068	-	-	-	-
a	546	1138	-	-	-	-
b	557	1146	-	-	-	-
ab	581	-	-	-	-	-
c	567	-	-	-	-	-
ac	579	-	-	-	-	-
bc	597	-	-	-	-	-
abc	609	-	-	-	-	-

Steg 4:

Försök	Medel	(1)	(2)	(3)	Effekt	Skattning
(1)	522	1068	-	-	-	-
a	546	1138	-	-	-	-
b	557	1146	-	-	-	-
ab	581	1206	-	-	-	-
c	567	-	-	-	-	-
ac	579	-	-	-	-	-
bc	597	-	-	-	-	-
abc	609	-	-	-	-	-

Yates schema (exempel)

Steg 5:

Försök	Medel	(1)	(2)	(3)	Effekt	Skattning
(1)	522	1068	-	-	-	-
a	546	1138	-	-	-	-
b	557	1146	-	-	-	-
ab	581	1206	-	-	-	-
c	567	24	-	-	-	-
ac	579	-	-	-	-	-
bc	597	-	-	-	-	-
abc	609	-	-	-	-	-

Steg 6:

Försök	Medel	(1)	(2)	(3)	Effekt	Skattning
(1)	522	1068	-	-	-	-
a	546	1138	-	-	-	-
b	557	1146	-	-	-	-
ab	581	1206	-	-	-	-
c	567	24	-	-	-	-
ac	579	24	-	-	-	-
bc	597	-	-	-	-	-
abc	609	-	-	-	-	-

Steg 7:

Försök	Medel	(1)	(2)	(3)	Effekt	Skattning
(1)	522	1068	-	-	-	-
a	546	1138	-	-	-	-
b	557	1146	-	-	-	-
ab	581	1206	-	-	-	-
c	567	24	-	-	-	-
ac	579	24	-	-	-	-
bc	597	12	-	-	-	-
abc	609	-	-	-	-	-

Steg 8:

Försök	Medel	(1)	(2)	(3)	Effekt	Skattning
(1)	522	1068	-	-	-	-
a	546	1138	-	-	-	-
b	557	1146	-	-	-	-
ab	581	1206	-	-	-	-
c	567	24	-	-	-	-
ac	579	24	-	-	-	-
bc	597	12	-	-	-	-
abc	609	12	-	-	-	-

Yates schema (exempel)

Steg 9:

Försök	Medel	(1)	(2)	(3)	Effekt	Skattning
(1)	522	1068	2206	-	-	-
a	546	1138	-	-	-	-
b	557	1146	-	-	-	-
ab	581	1206	-	-	-	-
c	567	24	-	-	-	-
ac	579	24	-	-	-	-
bc	597	12	-	-	-	-
abc	609	12	-	-	-	-

Steg 10:

Försök	Medel	(1)	(2)	(3)	Effekt	Skattning
(1)	522	1068	2206	-	-	-
a	546	1138	2352	-	-	-
b	557	1146	-	-	-	-
ab	581	1206	-	-	-	-
c	567	24	-	-	-	-
ac	579	24	-	-	-	-
bc	597	12	-	-	-	-
abc	609	12	-	-	-	-

. . . Fortsätt på samma sätt . . .

Resultat:

Försök	Medel	(1)	(2)	(3)	Effekt	Skattning
(1)	522	1068	2206	4558	$\hat{\mu}$	569.75
a	546	1138	2352	72	$\hat{A}$	9
b	557	1146	48	130	$\hat{B}$	16.25
ab	581	1206	24	0	$\hat{AB}$	0
c	567	24	70	146	$\hat{C}$	18.25
ac	579	24	60	-24	$\hat{AC}$	-3
bc	597	12	0	-10	$\hat{BC}$	-1.25
abc	609	12	0	0	$\hat{ABC}$	0

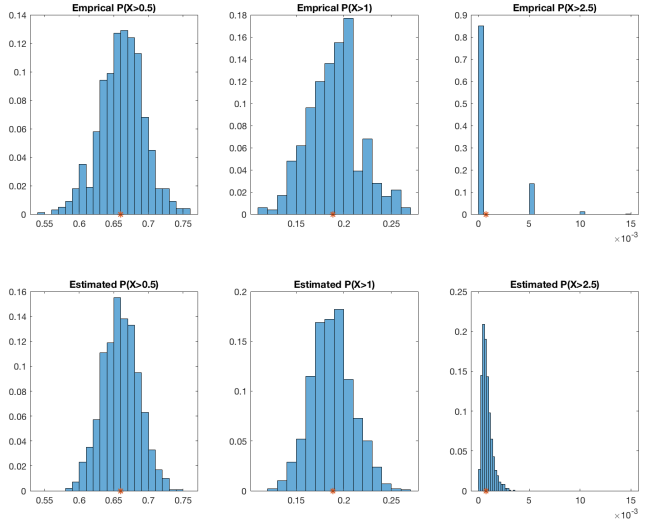
Empirisk skattning eller skattning från modell

- I projektet skattades empiriskt sannolikheterna att överskrida de två gränsvärdena.
- Ni beräknade också motsvarande sannolikheter baserat på en modell skattad från datan.
- Vad är för och nackdelar med de två metoderna?

Simuleringsstudie:

- Simulera 200 observationer från en $\Gamma(3.5, 0.2)$-fördelning.
- Baserat på datan, skatta de empiriska sannolikheterna.
- Baserat på datan, skatta en gammafördelning och beräkna sannolikheterna.
- Upprepa detta 1000 gånger och plotta histogram över de olika skattningarna.

UNIVERSITY OF GOTHENBURG CHALMERS



Projektet David Bolin

UNIVERSITY OF GOTHENBURG CHALMERS

Empirisk skattning eller skattning från modell

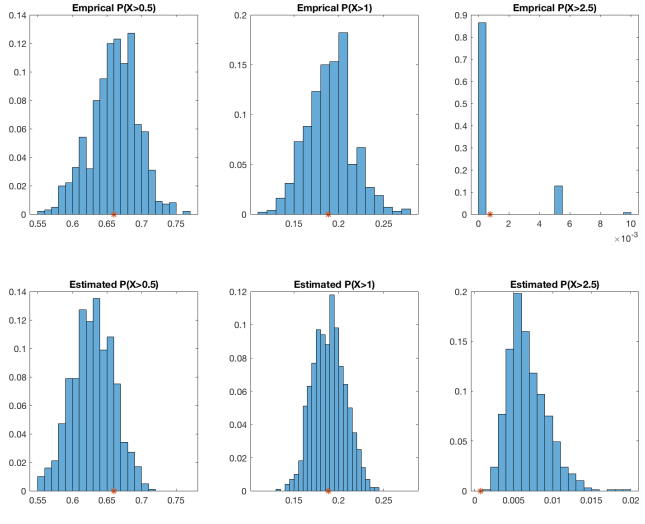
- Vi ser alltså att vi får högre precision genom att använda modellen.
- Vad händer om modellen är fel?

Simuleringsstudie 2:

- Simulera 200 observationer från en $\Gamma(3.5, 0.2)$-fördelning.
- Baserat på datan, skatta de empiriska sannolikheterna.
- Baserat på datan, skatta en log-normalfördelning och beräkna sannolikheterna.
- Upprepa detta 1000 gånger och plotta histogram över de olika skattningarna.

Projektet David Bolin

UNIVERSITY OF GOTHENBURG CHALMERS



Projektet David Bolin

Polynomregression och kolinearitet

```
y = log(abborre.HG);
x = datenum(abborre.Datum);
b = regress(y,[ones(length(y),1) x]);
b_quadratic = regress(y,[ones(length(y),1) x x.^2]);
```

- I Del 4 testade vissa att göra polynomregression och fick ganska konstiga skattningar.
- Koden ovan ger varningen

  `Warning: X is rank deficient to within machine precision.`
- Problemen uppkommer eftersom x och x^2 nästan är kolinjära. Anledningen till detta är att datenum returnerar väldigt stora tal: min(x) = 720898, max(x) = 736160.

Polynomregression och kolinearitet

- Problemet kan lösas genom att helt enkelt använda x - min(x) som förklarande variabel. Detta ger bättre numerisk stabilitet.

  ```
  x = x - min(x);
  b_quadratic2 = regress(y,[ones(length(y),1) x x.^2]);
  ```
- `b' = [69.3783 -0.0001]`
- `b_quadratic' = 1.0e-04 *[0 0.9230 -0.0000]`
- `b_quadratic2' = [0.4189 -0.0004 0.0000]`

Tentan

- Tillåtna hjälpmedel: Formelsamling och "Chalmersgodkänd" miniräknare.
- Båda dessa kommer tillhandahållas så ni behöver inte ta med något.
- Enligt info från Chalmers: "De inköpta modellerna heter:
  1. Casio FX-82
  2. Sharp EL-W531
  3. Texas TI-30

  Det kommer att finnas några exemplar av de inköpta miniräknarna i Studentcentrum så att studenterna kan testa hur de fungerar innan tentamenstillfället. "
- Angående bevis:
  - Ni behöver inte kunna bevisa de satser som jag har visat under föreläsningarna.
  - Ni kan behöva definiera begrepp som vi har använt och härleda teoretiska egenskaper hos till exempel fördelningar och skattningar.

Utfall och utfallsrum

Slumpmässigt försök

Man brukar säga att ett slumpmässigt försök är ett försök som kan upprepas under väsentligen identiska förhållanden där utfallet av försöket inte exakt kan förutsägas i det enskilda fallet.

Utfall och utfallsrum

Resultatet av ett försök kallas för ett *utfall* ω, och mängden av alla möjliga utfall kallas för *utfallsrummet* Ω.

Händelser

Man kan också sätta samman olika utfall till *händelser*. En händelse A är en mängd av utfall, det vill säga en delmängd av utfallsrummet Ω.

Egenskaper hos sannolikhetsmåttet

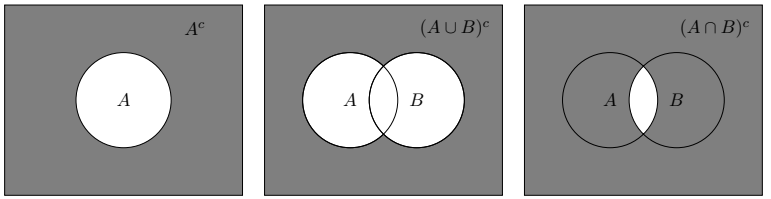
Ur axiomen kan vi härleda följande egenskaper.

Egenskaper

För ett sannolikhetsmått P gäller att

1. $\mathsf{P}(\emptyset) = 0$.
2. $\mathsf{P}(A^c) = 1 - \mathsf{P}(A)$.
3. $\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B)$.

Att dessa är uppfyllda inses enkelt genom att rita Venndiagram!



Kolmogorovs axiomsystem

Kolmogorovs axiomsystem

Låt Ω vara ett utfallsrum. En sannolikhet, eller sannolikhetsmått $\mathsf{P} : \Omega \to \mathbb{R}$ är en reell funktion som tar händelser i Ω som argument och returnerar ett reellt tal, och som uppfyller

1. $0 \leq \mathsf{P}(A) \leq 1$.
2. $\mathsf{P}(\Omega) = 1$.
3. Om $A_1, A_2, \ldots$ är än följd av disjunkta händelser så gäller att

$$\mathsf{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathsf{P}(A_i).$$

Speciellt så gäller att om händelserna A och B är disjunkta så är

$$\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B).$$

Betingade sannolikheter

Vi vill ibland beräkna sannolikheten för en händelse A givet att vi vet att en annan händelse B har inträffat. Vi vill då veta den så kallade betingade sannolikheten för A givet B.

Betingad sannolikhet

Antag att $\mathsf{P}(B) > 0$. Den betingade sannolikheten för A givet B definieras som

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

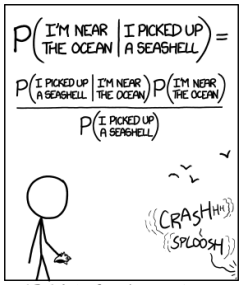
Multiplikationssatsen för sannolikheter

Om A och B är händelser så gäller att

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

Multiplikationssatsen är speciellt användbar för att beräkna sannolikheten för successiva händelser som påverkar varandra.

Bayes sats



Bayes sats

För två händelser A och B gäller att

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B)} = \frac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B|A)\mathrm{P}(A) + \mathrm{P}(B|A^c)\mathrm{P}(A^c)}$$

Slumpvariabler

En slumpvariabel beskriver utfallet av ett slumpmässigt försök med ett numeriskt värde:

Slumpvariabel

En stokastisk variabel (slumpvariabel) X är en funktion som tar element från Ω som argument och som returnerar ett reellt tal.

Vi betecknar ofta slumpvariabler med stora bokstäver, X, Y och Z. Utfall betecknas ofta med respektive små bokstäver, x, y och z.

Tvådimensionell slumpvariabel

En tvådimensionell slumpvariabel (X, Y) tillordnar två numeriska värden till varje utfall i Ω.

Diskreta fördelningar

Diskreta slumpvariabler

En diskret slumpvariabel kan endast anta ett ändligt eller uppräkneligt antal värden, i allmänhet någon delmängd av heltalen.

Sannolikhetsfunktion

Till en diskret stokastisk variabel X definierar vi sannolikhetsfunktionen $f(k)$ genom $f(k) = \mathrm{P}(X = k)$, eller $f_{X,Y}(i, j) = \mathrm{P}(X = i, Y = j)$ för en tvådimensionell variabel.

Eftersom $f(k)$ beskriver sannolikheter har vi

- $f(k) \geq 0$ för alla k.
- $\sum_{k=-\infty}^{\infty} f(k) = 1$.
- $\mathrm{P}(m \leq X \leq n) = \sum_{k=m}^{n} f(k)$ om m och n är heltal.

Kontinuerliga fördelningar

I praktiken stöter vi ofta på situationer då de möjliga värdena är alla värden i ett intervall, tex $[0, 1]$, eller är alla positiva reella tal.

Kontinuerlig slumpvariabel

En kontinuerlig slumpvariabel kan anta alla värden i något (eller några) intervall av reella tal och sannolikheten för att den antar varje specifikt värde är noll.

Täthetsfunktion

Låt X vara en kontinuerlig slumpvariabel, en funktion $f(x)$ så att

$$f(x) \geq 0, \quad \int_{-\infty}^{\infty} f(x)dx = 1, \quad \mathrm{P}(a \leq X \leq b) = \int_a^b f(x)dx,$$

kallas en täthetsfunktion.

Fördelningsfunktion och marginalfördelningar

Ett annat vanligt sätt att beskriva en stokastisk variabel är att använda den så kallade fördelningsfunktionen, $F(x) = \mathrm{P}(X \leq x)$. Vi har

$$F(x) = \begin{cases} \sum_{k \leq x} f(k) & X \text{ diskret} \\ \int_{-\infty}^{x} f(y)dy & X \text{ kontinuerlig} \end{cases}$$

Givet en diskret tvådimensionell slumpvariabel (X, Y) definieras marginalfördelningarna för X och Y som

$$f_X(i) = \sum_{j=-\infty}^{\infty} f_{X,Y}(i,j), \qquad f_Y(j) = \sum_{i=-\infty}^{\infty} f_{X,Y}(i,j)$$

I det kontinuerliga fallet har vi på motsvarande sätt

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy, \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

Väntevärden

Väntevärde

Väntevärdet för en slumpvariabel definieras som

$$\mathrm{E}(X) = \begin{cases} \sum_{k=-\infty}^{\infty} k f(k) & \text{om } X \text{ är diskret,} \\ \int_{-\infty}^{\infty} x f(x)dx & \text{om } X \text{ är kontinuerlig.} \end{cases}$$

Väntevärdet är "tyngdpunkten" i fördelningen.

Sats

Vi har att

$$\mathrm{E}(g(X)) = \begin{cases} \sum_{k=-\infty}^{\infty} g(k) f(k), & \text{om } X \text{ är diskret,} \\ \int_{-\infty}^{\infty} g(x) f(x)dx, & \text{om } X \text{ är kontinuerlig.} \end{cases}$$

Beroende

Betingad fördelning

Den betingade fördelningen för X givet $Y = y$ definieras som

$$f_{X|y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)},$$

givet att $f_Y(y) > 0$.

Oberoende händelser

Två händelser A och B är oberoende om $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$. Detta innebär också att $\mathrm{P}(A|B) = \mathrm{P}(A)$.

Oberoende slumpvariabler

Två slumpvariabler X och Y är oberoende om $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. Detta innebär att $f_{X|y}(x) = f_X(x)$.

Varians och standardavvikelse

Varians

Variansen av en stokastisk variabel definieras som $\mathrm{V}(X) = \mathrm{E}[(X-\mu)^2]$, där μ är väntevärdet av X.

Vi ser alltså att variansen definieras som väntevärdet av den kvadratavvikelsen mellan X och dess väntevärde. Vi beräknar variansen som

$$\mathrm{V}(X) = \begin{cases} \sum_{k=-\infty}^{\infty} (k-\mu)^2 f(k), & \text{om } X \text{ är diskret} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx, & \text{om } X \text{ är kontinuerlig.} \end{cases}$$

Ett enklare sätt är ofta att beräkna variansen som

$$\mathrm{V}(X) = \mathrm{E}(X^2) - \mu^2$$

Standardavvikelsen av en stokastisk variabel ges av $\sigma = \sqrt{\mathrm{V}(X)}$.

Kovarians och korrelation

Kovarians

Kovariansen mellan X och Y definieras som
$\mathsf{C}(X,Y) = \mathsf{E}[(X-\mu_X)(Y-\mu_Y)]$ där $\mu_X = \mathsf{E}(X)$ och $\mu_Y = \mathsf{E}(Y)$.

- Enligt definitionen kan vi skriva kovariansen som

$$\mathsf{C}(X,Y) = \begin{cases} \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty}(i-\mu_X)(j-\mu_Y)f_{X,Y}(i,i), & \text{diskret} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)f_{X,Y}(x,y), & \text{kont.} \end{cases}$$

- Notera att vi har att $\mathsf{C}(X,X) = \mathsf{V}(X)$.
- $\mathsf{C}(X,Y)$ kan beräknas som $\mathsf{C}(X,Y) = \mathsf{E}(XY) - \mathsf{E}(X)\mathsf{E}(Y)$.

Normalfördelningen

Normalfördelningen

En kontinuerlig slumpvariabel X är normalfördelad, $\mathsf{N}(\mu,\sigma^2)$, med parametrar $\mu \in \mathbb{R}$ och $\sigma > 0$, om den har täthetsfunktionen

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

Fördelningsfunktionen ges av

$$F(x) = \int_{-\infty}^{x}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)dy$$

Parametrar

Om $X \sim \mathsf{N}(\mu,\sigma^2)$ har vi att $\mathsf{E}(X) = \mu$ och $\mathsf{V}(X) = \sigma^2$.

Korrelation och oberoende

Korrelation

Den så kallade korrelationskoefficienten definieras som

$$\rho(X,Y) = \frac{\mathsf{C}(X,Y)}{\sqrt{\mathsf{V}(X)\mathsf{V}(Y)}}.$$

- Detta mått beskriver linjär samvariation mellan X och Y.
- Vi har att $-1 \leq \rho \leq 1$.
- Vi säger att X och Y är okorrelerade om $\rho(X,Y) = 0$.

Vi har följande samband mellan beroende och korrelation:

- Om X och Y är oberoende så är de okorrelerade.
- Om X och Y är okorrelerade så behöver de inte vara oberoende.

Dessa samband är naturliga eftersom två slumpvariabler är oberoende om det inte finns någon samvariation alls mellan dem, medan de är okorrelerade om det saknas *linjär* samvariation.

Centrala gränsvärdessatsen

En av de viktigaste satserna inom statistikteorin, och en av anledningarna till varför normalfördelningen är så viktig.

CGS

Låt $X_1,\ldots,X_n$ vara oberoende och likafördelade slumpvariabler med väntevärde μ och varians $\sigma^2 < \infty$. Då gäller att

$$\mathsf{P}\left(\frac{\sum_{i=1}^{n}X_i - n\mu}{\sigma\sqrt{n}} \leq x\right) \to \Phi(x), \quad \text{då } n \to \infty.$$

Om n är stort har vi enligt satsen att

- $\sum_{i=1}^{n}X_i$ är approximativt $\mathsf{N}(n\mu, n\sigma^2)$-fördelad.
- $\bar{X} = n^{-1}\sum_{i=1}^{n}X_i$ är approximativt $\mathsf{N}(\mu,\sigma^2/n)$-fördelad.

Hur stor n måste vara beror på fördelningen av X_i.

Punktskattningar

Stickprov

Ett stickprov är n oberoende observationer av en slumpvariabel X. Vi kan alltså skriva stickprovet som observationer av n slumpvariabler $X_1, \ldots, X_n$ där alla X_i är oberoende och likafördelade.

Skattning

En skattning av en parameter θ är en funktion $\hat{\theta}(X_1, \ldots, X_n)$ av observationerna.

Två egenskaper vi vill att vår skattare ska ha är att

- Den ska vara *väntevärdesriktig* (unbiased på engelska), vilket betyder att vi vill att $\mathsf{E}(\hat{\theta}(X_1, \ldots, X_n)) = \theta$.
- Den ska ha låg varians om n är stort, helst vill vi att $\mathsf{V}(\hat{\theta}(X_1, \ldots, X_n)) \to 0$ då $n \to \infty$.

Maximum likelihood-metoden

Tanken bakom ML-metoden är att vi vill hitta det parametervärde som gör att det stickprov vi har blir så troligt som möjligt.

Metoden är baserad på den så kallade likelihood-funktionen, som för ett stickprov $x_1 \ldots, x_n$ är

$$L(\theta) = \prod_{i=1}^{n} f(x_i)$$

där $f(x)$ är täthetsfunktionen eller sannolikhetsfunktionen för fördelningen och θ är parametrarna vi vill skatta.

ML skattare

Maximum likelihood-skattaren av en parameter θ ges av

$$\hat{\theta}_{ML} = \arg\max_{\theta} L(\theta)$$

Skattning av väntevärde och varians

Låt $x_1, \ldots, x_n$ vara observationer av oberoende och likafördelade s.v. $X_1, \ldots, X_n$ med väntevärde μ och standardavvikelse σ. Väntevärdesriktiga skattningar av μ och σ^2 är då

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

Om $X_i \in \mathsf{N}(\mu, \sigma^2)$ så har vi $\mu^* \sim \mathsf{N}(\mu, \frac{\sigma^2}{n})$ och $\frac{Q}{\sigma^2} \in \chi^2(n-1)$

Låt $x_{i1}, \ldots, x_{in_i}$ vara ober. obs. från $\mathsf{N}(\mu_i, \sigma)$ då $i = 1, \ldots, k$. Då är en poolad variansskattning

- $(\sigma^2)^* = S_p^2 = \frac{Q}{f} = \frac{(n_1-1)\,S_1^2 + \cdots + (n_k-1)\,S_k^2}{(n_1-1)+\cdots+(n_k-1)}$

Eftersom $X_{ij} \sim \mathsf{N}(\mu_i, \sigma^2)$ så har vi $\frac{Q}{\sigma^2} \sim \chi^2(f)$

Konfidensintervall

Konfidensintervall

Låt $X_1, \ldots, X_n$ vara slumpvariabler med en fördelning som har θ som en parameter med θ_0 som sant okänt värde. Ett $100(1-\alpha)\%$ konfidensintervall för θ med konfidensgraden $1-\alpha$ är ett intervall $[a(X_1, \ldots, X_n), b(X_1, \ldots, X_n)]$ sådant att

$$\mathsf{P}(a \leq \theta_0 \leq b) = 1 - \alpha.$$

- (a, b) är ett slumpmässigt intervall, eftersom a och b är slumpvariabler som beror av $X_1, \ldots, X_n$.
- Konfidensintervallet ska alltså tolkas som att om vi gör upprepade mätningar av $X_1, \ldots, X_n$ och bildar konfidensintervallet för alla dessa mätningar så kommer $100(1-\alpha)\%$ av dessa intervall täcka det sanna värdet θ_0.

Konfidensintervall

Om $\theta^*(X_1, \ldots, X_n) \in \mathsf{N}(\theta, \mathsf{V}(\theta^*))$ så har vi

- $I_\theta = (\theta^* \pm \lambda_{\alpha/2} \cdot \sqrt{\mathsf{V}(\theta^*)})$ om $\mathsf{V}(\theta^*)$ är känd
- $I_\theta = (\theta^* \pm t_{\alpha/2}(f) \cdot s \cdot c)$ om $\mathsf{V}(\theta^*) = c^2 \cdot \sigma^2$ där $\sigma^2 = \mathsf{V}(X_i)$, c är en konstant och $(\sigma^2)^* = S^2 = \frac{Q}{f}$ med $\frac{Q}{\sigma^2} \sim \chi^2(f)$

Om $\theta^*(X_1, \ldots, X_n)$ är approximativt $\mathsf{N}(\theta, \mathsf{V}(\theta^*))$-fördelad enligt CGS (el. dyl.) så kan vi använda normalbaserade konfidensintervall approximativt.

Om $X_1, \ldots, X_n \sim \mathsf{N}(\mu, \sigma^2)$ med $(\sigma^2)^* = S^2 = \frac{Q}{f}$ och $\frac{Q}{\sigma^2} \in \chi^2(f)$ så ges ett konfidensintervall för σ^2 av

$$I_{\sigma^2} = \left(\frac{f \cdot s^2}{\chi^2_{\alpha/2}(f)}, \frac{f \cdot s^2}{\chi^2_{1-\alpha/2}(f)} \right).$$

Konfidensintervallmetoden

Om vi bildar ett 95% konfidensintervall I_θ för θ kan vi direkt använda detta för att göra hypotestestet.

- Vi förkastar H_0 om intervallet inte innehåller θ_0.
- Om intervallet däremot innehåller θ_0 kan vi inte förkasta H_0.

Konfidensgraden vi använder när vi beräknar konfidensintervallet är också konfidensgraden för hypotestestet vi gör.

Definition

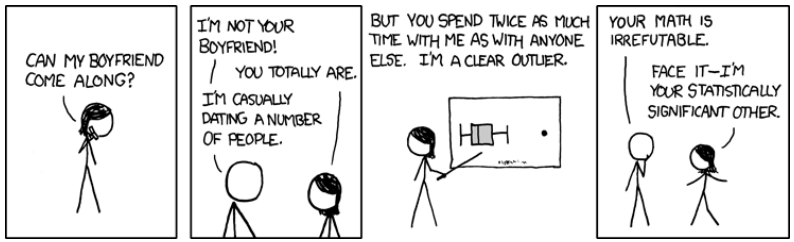
Felrisken eller signifikansnivån definieras som

$$\alpha = \mathsf{P}(H_0 \text{ förkastas } | H_0 \text{ sann}).$$

Översatt till konfidensintervall är signifikansnivån

$$\alpha = \mathsf{P}(\theta \notin I_\theta \text{ om } \theta = \theta_0) = \mathsf{P}(\theta_0 \notin I_\theta).$$

Hypotesprövning



Vi har ett stickprov med någon fördelning med okänd parameter θ. Vi vill testa *nollhypotesen*

$$H_0 : \theta = \theta_0$$

mot *mothypotesen*

$$H_1 : \theta \neq \theta_0.$$

Om testet bekräftar att $\theta \neq \theta_0$ säger vi att H_0 förkastas till förmån för H_1. Man brukar då också säga att θ är signifikant skild från θ_0.

Teststorheter och p-värden

Teststorhet

En teststorhet $T = T(X_1, \ldots, X_n)$ är en funktion av observationerna, och alltså en slumpvariabel. $T_{obs} = T(x_1, \ldots, x_n)$ är ett observerat värde av teststorheten för givna observationer.

p-värde

p-värdet eller signifikanssannolikheten definieras som sannolikheten under nollhypotesen att vi får ett värde på T som är åtminstone lika "extremt" som det observerade värdet T_{obs}.

Kritiskt område

Givet en signifikansnivå α definierar vi det kritiska området C_α som de värden på teststorheten T som leder till att man förkastar H_0 på nivån α.

Exempel: Binomialfördelningen

Låt $X \sim \text{Bin}(n, p)$. $p^* = x/n$ är en väntevärdesriktig skattning av p med varians $\text{V}(p^*) = p(1-p)/n$.

Enligt CGS är X approximativt $\text{N}(np, np(1-p))$-fördelad om n är stor ($np(1-p) > 10$). Vi har då att

$$T = \frac{p^* - p}{\sqrt{p^*(1-p^*)/n}}$$

är approximativt $\text{N}(0,1)$-fördelad.

Om vi vill testa

$$H_0 : p = p_0$$
$$H_1 : p \neq p_0$$

eller en ensidig hypotes $H_1 : p > p_0$ eller $H_1 : p < p_0$ så är $X \sim \text{Bin}(n, p_0)$ under H_0, och vi kan direkt beräkna p-värdet

Olika typer av jämförelser

Vi har tittat på olika typer av jämförelser mellan populationer:

- Stickprov i par.
- Jämförelser av oberoende populationer.
  - t-test: Två populationer som inte nödvändigtvis har samma varians.
  - Ensidig variansanalys: k populationer som skiljer sig åt på grund av att de har olika nivåer på en förklarande faktor A.
  - Tvåsidig variansanalys: Vi har två förklarande faktorer A och B som har a och b olika nivåer. Detta ger upphov till ab olika populationer.
  - 2^k faktorförsök: Vi har k förklarande faktorer som har två nivåer vardera. Detta ger upphov till 2^k olika populationer.
- Regression och sambandsanalys.
  - Regression används för att studera linjära (eller icke-linjära) samband mellan en responsvariabel Y och en förklarande variabel x som inte antas vara stokastisk.
  - Korrelation är ett mått på linjärt samband mellan två stokatiska variabler Y och X.

Stickprov i par

En vanlig situation är att mätningar uppkommer i par, till exempel:

- Man vill studera hur mycket rökare går upp i vikt när de slutar röka. Man mäter då vikten före och efter för varje person före och efter den slutar röka och jämförelsen sker för varje person.
- Man vill studera systematiska skillnader mellan två mätmetoder och använder varje metod på var och ett av ett antal prover och jämför de två metoderna för varje prov.

Modellen vi nu ansätter är att vi har n observationer i två stickprov

$$X_1, X_2, \ldots, X_n \qquad Y_1, Y_2, \ldots, Y_n$$

För varje mätning bildar vi differensen som antas vara normalfördelad:

$$D_i = X_i - Y_i \sim \text{N}(\Delta, \sigma^2)$$

Vi vill nu testa om $\Delta = 0$, vilket görs som vanligt för normalfördelade mätningar.

Oberoende stickprov

Vi antar att vi har två oberoende stickprov

- n_1 observationer $X_{11}, X_{12}, \ldots, X_{1n_1}$ från $\mathsf{N}(\mu_1, \sigma_1^2)$.
- n_2 observationer $X_{21}, X_{22}, \ldots, X_{2n_2}$ från $\mathsf{N}(\mu_2, \sigma_2^2)$.

Den intressanta parametern är nu $\theta = \mu_1 - \mu_2$ som vi skattar med $\theta^* = \bar{X}_1 - \bar{X}_2$ och vi vill nu bilda ett konfidensintervall för θ samt testa hypotesen

$$H_0 : \theta = 0,$$

vilket är samma sak som att testa $\mu_1 = \mu_2$, mot

$$H_1 : \theta \neq 0$$

eller någon ensidig hypotes, $\theta > 0$ (som motsvarar $\mu_1 > \mu_2$) eller $\theta < 0$ (som motsvarar $\mu_1 < \mu_2$). Vi skiljer på tre fall:

1. σ_1 och σ_2 är kända.
2. $\sigma_1 = \sigma_2 = \sigma$ där σ är okänd.
3. σ_2 och σ_2 är okända och ej säkert lika.

Testet i de olika fallen

Vi bildar teststorheten

$$T = \frac{\theta^*}{\sqrt{\mathsf{V}(\theta^*)}}$$

- Om σ_1 och σ_2 är kända gäller att $\mathsf{V}(\theta^*) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ och T är $\mathsf{N}(0,1)$-fördelad.
- Om $\sigma_1 = \sigma_2 = \sigma$ där σ är okänd använder vi en poolad variansskattning och skattar $\mathsf{V}(\theta^*)$ med $s_p^2(1/n_1 + 1/n_2)$. T är då $t(n_1 + n_2 - 2)$-fördelad.
- Om σ_1 och σ_2 är okända och $\sigma_1 \neq \sigma_2$ skattar vi $\mathsf{V}(\theta^*)$ med $s_1^2/n_1 + s_2^2/n_2$. T är då approximativt $t(f)$-fördelad med

$$f = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Ensidig variansanalys

- För jämförelser av fler än två populationer kan vi använda variansanalys.
- Vi vill studera om en faktor A påverkar en responsvariabel.
- Vi gör totalt $N = \sum_{i=1}^{a} n_i$ mätningar vid a olika faktornivåer.
- Ansätt modell $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, där $\varepsilon_{ij} \sim \mathsf{N}(0, \sigma^2)$
- Vi vill nu testa

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$$
$$H_1 : \alpha_i \neq \alpha_j \text{ för några } i, j$$

- Beräkningarna som behövs göras sammanfattas i en ANOVA-tabell:

Variation	Kvadratsumma	Frihetsgrader	Medelkvadratsumma	Teststorhet
Faktor A	$SS_A = \sum_{ij}(\bar{y}_i - \bar{y})^2$	$f_A = a - 1$	$MS_A = SS_A/f_A$	MS_A/MS_E
Residual	$SS_E = \sum_{ij}(y_{ij} - \bar{y}_i)^2$	$f_E = \sum_i n_i - a$	$MS_E = SS_E/f_E$	
Total	$SS_{Tot} = SS_E + SS_A$	$f = \sum_i n_i - 1$		

Tvåsidig variansanalys

- Vi kan analysera ett försök med två faktorer, A och B, med ANOVA.
- Vi gör totalt abn mätningar vid a olika faktornivåer för A och b olika nivåer för B.
- Ansätt modell $y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij}$, där $\varepsilon_{ij} \sim \mathsf{N}(0, \sigma^2)$ och $(\alpha\beta)_{ij}$ modellerar en möjlig samspelseffekt mellan A och B.
- Vi vill testa tre saker
  1. Finns det någon samspelseffekt: $H_0 : (\alpha\beta)_{ij} = 0$ för alla ij?
  2. Huvudeffekt för A: $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$?
  3. Huvudeffekt för B: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_b = 0$?

Variation	Kvadratsumma	Frihetsgrader	Medelkvadrat	Teststorhet
Faktor A	SS_A	$f_A = a - 1$	$MS_A = SS_A/f_A$	MS_A/MS_E
Faktor B	SS_B	$f_B = b - 1$	$MS_B = SS_B/f_B$	MS_B/MS_E
Faktor AB	SS_{AB}	$f_{AB} = f_A f_B$	$MS_{AB} = SS_{AB}/f_{AB}$	MS_{AB}/MS_E
Residual	SS_E	$f_E = ab(n-1)$	$MS_E = SS_E/f_E$	
Total	SS_{Tot}	$abn - 1$		

Enkel linjär regression

- ANOVA används för att undersöka om förklarande variabler (faktorer) påverkar en responsvariabeln. Vi får inget svar på hur påverkan ser ut.
- Regression används för att undersöka formen på sambandet.
- För enkel linjär regression har vi talpar (x_i, Y_i) där x_i är ett fixt värde på den förklarande variabeln och Y_i är en slumpvariabel. Vi antar att

$$\mathsf{E}(Y_i) = \beta_0 + \beta_1 x_i$$

- För polynomregression antar vi istället

$$\mathsf{E}(Y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p$$

- För multipel regression har vi talpar $(x_{1i}, \ldots, x_{ki}, Y_i)$ där x_{ki} är värdet på förklarande variabel k. Vi antar

$$\mathsf{E}(Y_i) = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_p x_{p,i}$$

Enkel linjär regression

Vi har talpar (x_i, Y_i) där x_i är ett fixt värde och Y_i är en slumpvariabel. Modellen vi ansätter för Y_i är

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

där ε_i är oberoende $\mathsf{N}(0, \sigma^2)$ slumpvariabeler som beskriver mätfelet och β_0 och β_1 är okända parametrar som beskriver det linjära sambandet.
Vi har alltså att väntevärdet av Y ges av funktionen

$$\mathsf{E}(Y_i) = f(\beta_0, \beta_1, x_i)$$

Vi skattar parametrarna enligt minsta-kvadratmetoden genom att minimera kvadratfelet för den datan vi har

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - f(\beta_0, \beta_1, x_i))^2$$

Multipel linjär regression

Vi har gjort mätningar av en responsvariabel Y för fixerade värden på förklarande variabler $x_1, \ldots, x_p$, som kan väljas utan fel.
Vi ansätter en linjär modell för $(Y_i, x_{1,i}, \ldots, x_{p,i}), i = 1, \ldots, n$:

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_p x_{p,i} + \varepsilon_i \qquad (1)$$

där ε_i är oberoende $\mathsf{N}(0, \sigma^2)$ slumpvariabeler som beskriver mätfelet och β_i är parametrar som beskriver beroendet.
Vi kan formulera modellen på matrisform som

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$$

där

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{p,1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & \cdots & x_{p,n} \end{pmatrix} \quad \mathbf{E} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$$

Parameterskattning

Sats

Givet att $\mathbf{X}^T\mathbf{X}$ är inverterbar fås minstakvadrat-skattningen av $\boldsymbol{\beta}$ som $\boldsymbol{\beta}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. Skattningarna är väntevärdesriktiga och vi har

$$V(\beta_i^*) = \sigma^2(\text{diagonalelement nummer i+1 i } (\mathbf{X}^T\mathbf{X})^{-1}).$$

Vidare är $s^2 = Q_0/(n-p-1)$ där

$$Q_0 = \sum_{i=1}^{n} (y_i - \beta_0^* - \beta_1^* x_{1,i} - \ldots - \beta_p^* x_{p,i})^2 = \mathbf{Y}^T\mathbf{Y} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y}.$$

Sats

Med $\mu_Y^*(\mathbf{x_0}) = \beta_0^* + \beta_1^* x_{0,1} + \cdots + \beta_p x_{0,p}$ är
$\mathsf{V}(\mu_Y^*(\mathbf{x_0})) = \sigma^2 \mathbf{x_0}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0$

Egenskaper hos skattningarna

- Om ε_i är normalfördelade är också β_i^* normalfördelade.
- $(n-p-1)s^2/\sigma^2 \sim \chi^2(n-p-1)$.
- med $\theta = \beta_i$ eller $\mu_Y(\mathbf{x}_0)$ gäller $(\theta^* - \theta)/d(\theta^*) \sim t(n-p-1)$.
- Ett konfidensintervall för θ är

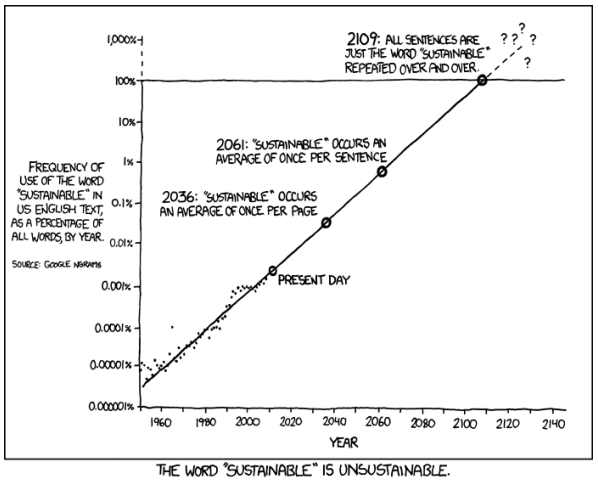
$$I_\theta = [\theta^* \pm t_{\alpha/2}(n-p-1)d(\theta^*)]$$

- För att testa $H_0 : \theta = \theta_0$ mot $H_1 : \theta \neq \theta_0$ används teststorheten $T = (\theta^* - \theta_0)/d(\theta^*)$ som har kritiskt område $C_\alpha = \{T : |T| > t_{\alpha/2}(n-p-1)\}$.
- Ett prediktionsintervall för en framtida observation $Y(\mathbf{x}_0)$ ges av

$$I_{Y(\mathbf{x}_0)} = \left[\mu_Y^*(\mathbf{x}_0) \pm t_{\alpha/2}(n-p-1) \cdot s\sqrt{1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}\right]$$

Ickeparametriska metoder

- Då vi använder ANOVA, regression och t-test antar vi att datan är normalfördelad.
- Dessa metoder kan generaliseras till andra typer av parametriska fördelningar.
- Om vi inte vill anta någon parametrisk fördelning kan vi istället använda en icke-parametrisk metod.
- I kursen har vi använt följande metoder:
  - Teckentest: Används för att undersöka om vi kan förkasta nollhypotesen att fördelningen som generade stickprovet har en viss median.
  - Wilcoxons ranktest: Test för medianen med högre styrka som kräver att fördelningen är symmetrisk.
  - Wilcoxsons ranksummetest: Test för att jämföra medianen hos två fördelningar.



Det viktigt att validera att modellantagandena är uppfyllda genom att studera residualerna. Det är också viktigt att inse att vi inte kan vara säkra på att modellen stämmer för x utanför mätområdet.

The end

Lycka till på tentan!