

Torsdag 26.mai Repetisjon + Kahoot.

Torsdag 26.mai Hjelpetimar 12.15-16 i S5

Fredag 27.mai Eksamensoppgåvereknig 10.15-12 i EI5

- Eksamen Mai 2014, oppgave 2 (inkl normal fordeling, lin.reg. og deskriptiv statistikk)
- Eksamen August 2012, oppgave 2 a og b (inkl SME)

Fredag 27.mai Hjelpetimar i S6

Mandag 30.mai Eksamen

- 10 punkt.
- Alle punkta tel like mykje. VIKTIG å få til dei lette rett.
- Gult A5 ark. Lov å skrive på begge sider. Kan hentast i 7.etg i Sentralbygg 2 (i gangen utanfor instituttkontoret)
- Formalsamlinga. Her står det mykje, bli kjent med!

Never, never, never give up!

Kurset har tre delar:

- Deskriptiv statistikk (Kap 1 + litt i Kap 8)
- Sannsynsteori (Kap 1-7 + notat)
- Statistikk (Kap 8-11)

PS: Dette er ei oppsummering av dei viktigaste begrepa og metodane. Ikkje ein lovnad om kva som kjem og ikkje kjem på eksamen.

I kapittel 1, i anbefalte oppgaver 1 og i kapittel 8. Brukt i forbindelse med evaluering av lin.regresjon (kap 11)

- Histogram
- Boxplot
- Plotte variable mot kvarandre (sjå etter avhengighet)
- Normal-kvantil plott
- Gjennomsnitt
- Median
- Varians og standard avvik.

(Eksamen mai 2014, Oppg.2a og 2f)

- Sannsyn for hendingar.
- Union, snitt og komplimentær.
- Uavhengige hendingar: A og B er uavhengige dersom $P(A \cap B) = P(A)P(B)$
- Betinge sannsyn: $P(B|A) = P(A \cap B)/P(A)$

(Eksamen mai 2014, oppgave 2 b).

Kap 3 Stokastiske variable og sannsynsfordelingar

- Diskret stokastiske variable (X) og sannsynsfordelingar ($f(x) = P(X = x)$)
- Kontinuerleg stokastisk variable (X) og sannsynsfordelingar ($f(x)$)
- Kummulativ sannsynsfordeling, $F(x) = P(X \leq x)$

Stokastisk variabel

Ein *stokastisk variabel* er ein funksjon som assosierer eit reelt tal med kvart element i utfallsrommet. $X : S \rightarrow \mathbb{R}$

Notasjon:

- X, Y : Store bokstavar på stokastiske variable
- x, y : Små bokstavar på utfall (datapunkt)

Diskret stokastisk variabel:

Dersom den stokastiske variabelen gjev tellbart antall utfall.

Kontinuerleg stokastisk variabel

Dersom den kontinuerlege variabelen har utfall på kontinuerleg skala.

Diskret

Paret $(x, f(x))$ blir kalla sannsynsfordelinga til den diskret stok. var. X dersom

- $0 \leq f(x)$
- $\sum_{\forall x} f(x) = 1$ (summen over alle moglege x)
- $f(x) = P(X = x)$

Kontinuerleg

Funksjonen $f(x)$ definert for alle reelle tal $x \in \mathbb{R}$ blir kalla sannsynsfordelinga til den kontinuerlege stok. var. X dersom

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

Definisjon

Den *kummulative fordelingsfunksjonen* $F(x)$ for ein stok. var. X er:

$$F(x) = P(X \leq x)$$

- *Diskret:* $F(x) = \sum_{t \leq x} f(t)$
- *Kontinuerleg* $F(x) = \int_{-\infty}^x f(t) dt$

For diskret stokastiske variable er det viktig med $<$ eller $\leq$, for kontinuerlege stokastiske variable er det likegyldig.

Kap 4. Matematisk forventning

- Forventningsverdi $E(X)$
- Varians $Var(X)$ og standardavvik $Std(X) = \sqrt{Var(X)}$
- Vit at kovarians og korrelesjon måler lineær avhengighet
- Forventning og varians av lineærkombinasjonar

Definisjon 4.1

La X vere ein stok.var. med sannsynsfordeling $f(x)$.
Forventningsverdien til X er då

$$\mu = E(X) = \sum_{\forall x} xf(x)$$

dersom X er diskret, og

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

dersom X er kontinuierleg.

Tolkning:

- Gjennomsnitt av uendeleg mange data trukke frå $f(x)$
- Massesenteret til fordelinga.

Definisjon 4.3

La X vere ein stok. var. med forventning μ . Variansen til X er då:
For diskret X :

$$\text{Var}(X) = E((X - \mu)^2) = \sum_{\forall x} (x - \mu)^2 f(x)$$

For kontinuerleg X :

$$\text{Var}(X) = E((X - \mu)^2) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

Tolkning Spredning, $\text{Std}(X) = \sqrt{\text{Var}(X)}$ same skala som X .

Eigenskap Må vere positiv (eller 0).

Teorem, VIKTIG

Dersom $a_0, a_1, \dots, a_n$ er konstanter og $X_1, X_2, \dots, X_n$ stok. var., så er

$$E(a_0 + \sum_{i=1}^n a_i X_i) = a_0 + \sum_{i=1}^n a_i E(X_i)$$

og

$$\text{Var}(a_0 + \sum_{i=1}^n a_i X_i) = \sum_{i=1}^n a_i^2 \text{Var}(X_i) + 2 \sum_{i=1}^n \sum_{j=1}^{i-1} a_i a_j \text{Cov}(X_i, X_j)$$

for *uavhengige* $X_1, \dots, X_n$ er

$$\text{Var}(a_0 + \sum_{i=1}^n a_i X_i) = \sum_{i=1}^n a_i^2 \text{Var}(X_i)$$

Kap 5 Diskret sannsynsfordelingar

- Binomisk fordeling og Bernoulli prosess
- Hypergeometrisk
- Geometrisk
- Poisson fordeling og poissonprosess

Bernoulli prosess

- 1 n uavhengige forsøk
- 2 Kvart forsøk resulterer i suksess, $I_i = 1$ eller ikkje-suksess $I_i = 0$.
- 3 Suksess-sannsynet $p = P(I_i = 1)$ er konstant.

Binomisk fordeling

Ser på antall suksess i ein Bernoulli prosess, $X = \sum_{i=1}^n I_i$. X er då binomisk fordelt;

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

- Normal fordeling
 - Kunne rekne i normal fordeling, dvs standardisere og slå opp i tabell.(Eksamen mai 2014, oppgåve 2)
 - Lineær kombinasjon av normal fordelte stok.var. er normal fordelt.
- Eksponensial fordeling
 - Tid til første hending i Possion prosess er eksponensial fordelt.
- Kji-kvadrat fordeling.

Kap 7. Funksjonar av stokastiske variable

- Ein funksjon av stok.var. er sjølv ein stok.var.
- Transformasjon av ein stok.var.
- Moment og momentgenererende funksjonar (for den spesielt interesserte/ambisiøse).
- Lineær kombinasjon av normalfordelte stok.var. er normalfordelt.
- Lineær kombinasjon av kji-kvadrat er kji-kvadrat.

Skal kunne finne sannsynsfordelinga til

- minimum (minste av n stokastiske variable)
- maximum (største av n stokastiske variable)

(Eksamen mai 2014, oppgave 2c)

- Kap.8: Utvalgsfordelingar
- Kap.9: Estimering og konfidensintervall (KI)
- Kap.10: Hypotesetesting
- Kap.11: Enkel lineær regresjon

- Tilfeldig utval
- Utvalsfordelingar.
- Utvalsfordelinga til gjennomsnittet $\bar{X}$
- Sentralgrenseteoremet (SGT)
- Utvalsfordeling for varians normalfordeling, χ^2_{n-1}
- Utvalsfordeling for gjennomsnitt normalfordeling og ukjent varians, T_{n-1}

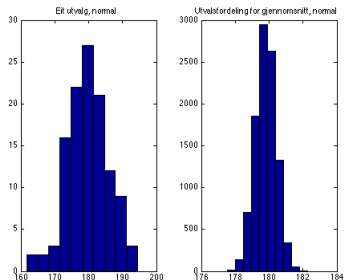
Utvalgsfordelinga til $\bar{X}$

Algoritme

For $m = 1 : M$

- Trekk $n=117$ datapunkt frå $N(179.8, 6.5^2)$ $m = 1, \dots, M$ gongar $\Rightarrow x_{m1}, x_{m2}, \dots, x_{mn}$.
- Finn gjennomsnittet $\bar{x}_m = 1/n \sum_{i=1}^n x_{mi}$

Plott histogram for $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_M$



Dersom X_i er normalfordelt, og kjent varians

$X_i \sim N(\mu, \sigma^2)$, for $i = 1, 2, \dots, n$. Dvs eit utvalg på n .
Då er

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

Sentralgrenseteoremet. Fordelinga til gjennomsnittet, $\bar{X}$.

Sentralgrenseteoremet, teorem 8.2

La X_i , $i = 1, 2, \dots, n$ vere uavhengige identisk fordelte (u.i.f.) stokastiske variable med $E(X_i) = \mu$ og $\text{Var}(X_i) = \sigma^2 < \infty$ (endeleg varians). La $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ og $Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}$. Då

$$Z \rightarrow N(0, 1)$$

når $n \rightarrow \infty$.

Dette tilsvarer

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

når $n \rightarrow \infty$.

PS: Gjeld uansett fordeling for X_i .

PSS: $Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}$ er og ein observator

Dersom $X_i \stackrel{u.i.f.}{\sim} N(\mu, \sigma^2)$, kjent σ^2 eller pga SGT

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

Dersom $X_i \stackrel{u.i.f.}{\sim} N(\mu, \sigma^2)$ og $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

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

Student-t fordelt med $\nu = n - 1$ fridomsgrader.

Dersom $X_i \stackrel{u.i.f.}{\sim} N(\mu, \sigma^2)$ og $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

Kji-kvadrat fordelt med $\nu = n - 1$ fridomsgrader.

- Antar fordeling frå kjennskap til fenomenet; $X_i \sim f(x; \theta)$.
- Har data $x_1, x_2, \dots, x_n$
- *Estimerer/ berekner θ frå data v.h.a. ein estimator $\hat{\theta}$.*

- Kva er ein god estimator (forventningsrett og minst mogeleg varians).
- Korleis finne ein estimator (SME eller for lineær regresjon MKM)
- Korleis kvantifisere usikkerheita i estimat (Konfidensintervall)

Konfidensintervall: Dersom forsøket blir repetert, vil sann parameter vere i KI i $1 - \alpha$ del av forsøk.

Finn den verdien for parameteren θ (påske-eksempel $\theta = p$) som gjev høgast sannsyn for å observere dei dataene vi har observert.

Korleis

- 1 Finn likelihoodfunksjonen

$$L(\theta; x_1, x_2, \dots, x_n) = f(x_1, x_2, \dots, x_n; \theta)$$

Sannsynet /sanns.tettheten for våre data.

- 2 Finn toppunkt av likelihoodfunksjonen: $\underset{\theta}{\operatorname{argmax}} L(\theta; x_1, \dots, x_n)$

August 2012 oppgave 3b

Finn den verdien for parameteren θ (påske-eksempel $\theta = p$) som gjev høgast sannsyn for å observere dei dataene vi har observert.

OPPSKRIFT

- 1 Finn likelihoodfunksjonen

$$L(\theta; x_1, x_2, \dots, x_n) = f(x_1, x_2, \dots, x_n; \theta) \stackrel{\text{uavh}}{=} \prod_{i=1}^n f(x_i; \theta)$$

- 2 Finn toppunkt av likelihoodfunksjonen:

- Tar ln av L ; $l(\theta; x_1, x_2, \dots, x_n) = \ln(L(\theta; x_1, x_2, \dots, x_n))$.
Reknetriks som nesten alltid blir brukt. L og l har same toppunkt.

- Deriverer og set lik 0; $\frac{\partial}{\partial \theta} l(\theta; x_1, x_2, \dots, x_n) = 0$

- Løyer ut for $\theta = h(x_1, x_2, \dots, x_n)$.

- 3 **Estimator:** $\hat{\theta} = h(X_1, X_2, \dots, X_n)$ (stok.var.)

Estimat: $\theta^* = h(x_1, x_2, \dots, x_n)$ (talverdi)

Har eit nivå av trygghet for at sann parameter ligg i intervallet.

- ❶ Finn estimator for parameteren vi ønsker å finne KI for.
- ❷ Finn observator der parameter av interesse og estimator inngår:
 - $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$,
når SGT eller $X_i \overset{u.i.f.}{\sim} N(\mu, \sigma^2)$
 - $T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$,
når $X_i \overset{u.i.f.}{\sim} N(\mu, \sigma^2)$ og $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$
 - $\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$,
når $X_i \overset{u.i.f.}{\sim} N(\mu, \sigma^2)$ og $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$
- ❸ Har at (f.eks) $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$.
- ❹ Løyer ut for parameter, (f.eks. μ); $P(\hat{\mu}_L < \mu < \hat{\mu}_U) = 1 - \alpha$.
- ❺ Konfidensintervall; sett inn for data: $[\mu_L^*, \mu_U^*]$

- H_0 : Null hypotese. Konservativ. (f.eks. $\mu = \mu_1$)
- H_1 : Alternativ hypotese. Endring. (f.eks. $\mu < \mu_1$)

Moglege beslutningar

- Forkastar H_0 , og aksepterer H_1 .
Påstand H_1 'bevist' ved data.
- Forkaster ikkje H_0 .
Data underbygger ikkje påstand H_1 .

Beslutningsfeil

- *Type-I-feil*: Forkastar H_0 når H_0 er sann.
- *Testnivå*: $P(\text{Type-I-feil}) = \alpha$.
- *Type-II-feil*: Forkastar ikkje H_0 når H_1 er sann.
- *Teststyrke*: $1 - P(\text{Type-II-feil} | \mu = \mu_1) = 1 - \beta(\mu_1)$

Metode p-verdi

- 1 Antar H_0 er sann.
- 2 Finn p -verdi: $P(\text{vårt estimat eller meir ekstremt} \mid H_0 \text{ er sann})$
- 3 Forkastar H_0 dersom liten p -verdi ($< \alpha$).

Metode forkastningsområde

- 1 Antar H_0 er sann.
- 2 Finn testobservator og område for 'testobservasjon' (evt. estimat) som fører til forkasting.
- 3 Forkastar H_0 dersom 'testobservasjon'/estimat i forkastningsområdet.

Enkel lineær regresjon

$$Y_i = \alpha + \beta x_i + \epsilon_i \text{ med } \epsilon_i \overset{u.i.f.}{\sim} N(0, \sigma_\epsilon^2)$$

- Finne estimatorer / estimere α , β og σ_ϵ^2 (for α og β SME eller MKM, for σ^2 SME)
- Hypotesetest / KI for α , β eller σ_ϵ^2
- Hypotesetest / KI for $\mu_{Y_0|x_0} = E(Y|x = x_0)$
- Prediksjonsintervall for $Y_0|x_0$

- $A = \bar{Y} - B\bar{x}$ og $A \sim N(\alpha, \sigma_A^2)$
 - der $\sigma_A^2 = \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \sigma_\epsilon^2$
- $B \sim N(\beta, \sigma_B^2)$
 - der $\sigma_B^2 = \text{Var}(B) = \frac{\sigma_\epsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma_\epsilon^2}{S_{xx}}$
- $S_\epsilon^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - 2}$
 - der $\hat{Y}_i = A + Bx_i$ og $\frac{(n-2)S^2}{\sigma_\epsilon^2} \sim \chi_{n-2}^2$