Oppsummering del 2

Torstein Fjeldstad
Institutt for matematiske fag, NTNU
23.11.2018

- Repetisjon del 2
- Eksamen mai 2009 oppgåve 3

DEL 2: Statistikk

Sentralgrenseteoremet

Viss $\bar{X}$ er gjennomsnittet av eit tilfeldig utval av storleik n tatt frå ein populasjon med forventningsverdi μ og varians $\sigma^2 < \infty$ vil

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to n(z; 0, 1) \qquad \text{når } n \to \infty.$$

Estimator

Situasjon: $X_1, X_2, \ldots, X_n$ tilfeldig utval med $X_i \sim f(x_i; \theta)$
Mål: ynskjer å anslå verdien til θ (og seie noko om tilhøyrande uvisse)

Observator: funksjon av stokastiske variablar

Estimator: ein observator $\widehat{\theta}$ som nyttes til å estimere verdien til den ukjende parameteren θ

Sannsynsmaksimering

Situasjon: $X_1, X_2, \ldots, X_n$ tilfeldig utval $X_i \sim f(x_i; \theta)$

1. Definer rimelighetsfunksjonen

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

2. Ofte er det enklare å sjå på log-rimelighetsfunksjonen

$$l(\theta) = \sum_{i=1}^{n} \ln f(x_i; \theta)$$

3. Maksimer $l(\theta)$ (ofte ved å derivere og setje lik null)

Eigenskapar til estimatorar

- Ynskjer ein forventningsrett (eng: unbiased) estimator

$$E(\widehat{\theta}) = \theta$$

- For to forventningsrette estimatorar $\widehat{\theta}_1, \widehat{\theta}_2$ vil me velge den med lågast varians:

$$Var(\widehat{\theta}_1) < Var(\widehat{\theta}) \Rightarrow \text{velg } \widehat{\theta}_1$$

χ^2-fordeling (kjikvadratfordeling)

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}\,, \quad x \geq 0\,, \quad \nu = 1, 2, \ldots.$$

$$\mathrm{E}(X) = \nu\,, \quad \mathrm{Var}(X) = 2\nu\,, \quad M_X(t) = \left(\frac{1}{1-2t}\right)^{\nu/2} \text{ for } t < \frac{1}{2}.$$

Kommentar: Dersom $X_1, \ldots, X_n$ er uavhengige og normalfordelte med forventning μ og varians σ^2 har vi at

$$\sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2} \quad \text{er } \chi^2\text{-fordelt med } n \text{ frihetsgrader,}$$

$$\sum_{i=1}^{n} \frac{(X_i-\bar{X})^2}{\sigma^2} \quad \text{er } \chi^2\text{-fordelt med } n-1 \text{ frihetsgrader,}$$

$$\sum_{i=1}^{n} (X_i-\bar{X})^2 \text{ og } \bar{X} \text{ er uavhengige.}$$

t-fordeling (Student t-fordeling)

$$f(x) = \frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)\sqrt{\pi\nu}} \left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}\,, \quad -\infty < x < \infty.$$

$$\mathrm{E}(X) = 0 \text{ hvis } \nu \geq 2\,, \quad \mathrm{Var}(X) = \frac{\nu}{\nu-2} \text{ hvis } \nu \geq 3\,, \quad M_X(t) \text{ eksisterer ikke.}$$

Spesialtilfeller: $\nu = 1$ gir Cauchyfordelingen.
$\nu = \infty$ gir Normalfordelingen.

Kommentar: Dersom Z er standard normalfordelt og V er χ^2-fordelt med ν frihetsgrader og Z og V er uavhengige, har vi at

$$\frac{Z}{\sqrt{V/\nu}} \quad \text{er } t\text{-fordelt med } \nu \text{ frihetsgrader.}$$

Spesielt gir dette at dersom $X_1, \ldots, X_n$ er uavhengige og normalfordelte med forventning μ og varians σ^2 har vi at

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \quad \text{er } t\text{-fordelt med } n-1 \text{ frihetsgrader,}$$

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

Fordeling til lineærkombinasjoner

La $X_1, \ldots, X_n$ være uavhengige variabler.

Dersom X_i er normalfordelt med forventning μ_i og varians σ_i^2 vil $Y = \sum_{i=1}^n a_i X_i$ være normalfordelt med forventning $\sum_{i=1}^n a_i \mu_i$ og varians $\sum_{i=1}^n a_i^2 \sigma_i^2$.

Dersom X_i er binomisk fordelt med parametre m_i og p vil $Y = \sum_{i=1}^n X_i$ være binomisk fordelt med parametre $\sum_{i=1}^n m_i$ og p.

Dersom X_i er Poissonfordelt med parameter μ_i vil $Y = \sum_{i=1}^n X_i$ være Poissonfordelt med parameter $\sum_{i=1}^n \mu_i$.

Dersom X_i er χ^2-fordelt med ν_i frihetsgrader vil $Y = \sum_{i=1}^n X_i$ være χ^2-fordelt med $\sum_{i=1}^n \nu_i$ frihetsgrader.

Konfidensintervall

Situasjon: $X_1, X_2, \ldots, X_n$ tilfeldig utval $X_i \sim f(x_i; \theta)$
Mål: ynskjer konfidensintervall $[\widehat{\theta}_L(X_1, X_2, \ldots, X_n, \widehat{\theta}_U(X_1, X_2, \ldots, X_n)]$
slik at

$$P(\widehat{\theta}_L(X_1, X_2, \ldots, X_n \leq \theta \leq \widehat{\theta}_U(X_1, X_2, \ldots, X_n)]) = 1 - \alpha$$

Generell framgangsmåte for konfidensintervall

1. Estimator for θ, $\hat{\theta}$ (t.d. SME)
2. La $Z = h(\hat{\theta}, \theta)$ der $h(\cdot, \cdot)$ er ein funksjon s.a. Z har ei kjend fordeling.
3. Har då
$$P(z_{1-\alpha/2} \leq h(\hat{\theta}, \theta) \leq z_{\alpha/2}) = 1 - \alpha$$
4. Løys ulikskapane (mhp. θ) kvar for seg og finn eit uttrykk med θ i midten
$$P(\hat{\theta}_L(X_1, X_2, \ldots, X_n) \leq \theta \leq \hat{\theta}_U(X_1, X_2, \ldots, X_n)) = 1 - \alpha$$
5. Et $(1 - \alpha) \cdot 100$ % konfidensintervall for θ er
$$\left[\hat{\theta}_L(X_1, X_2, \ldots, X_n), \hat{\theta}_U(X_1, X_2, \ldots, X_n)\right]$$

Hypotesetesting

	H_0 riktig	H_1 riktig
Forkast H_0	Type I-feil	Ok
Ikkje forkast H_0	Ok	Type II-feil

Ide: vi må vere "sikre" før me påstår at H_1 er rett. Me velg signinifikansnivået α liten og krev

$$P(\text{Type I-feil}) = P(\text{Forkast H}_0 \text{ når H}_0 \text{ er riktig}) \leq \alpha$$

$$\beta = P(\text{Type II-feil}) = P(\text{Ikkje forkast H}_0 \text{ når H}_1 \text{ er riktig})$$

Generell framgangsmåte

Situasjon: $X_1, X_2, \ldots, X_n$ tilfeldig utval med $X_i \sim f(x_i; \theta)$.

1. Ynskjer å teste:
   a) $H_0 : \theta = \theta_0$ mot $H_1 : \theta > \theta_0$
   b) $H_0 : \theta = \theta_0$ mot $H_1 : \theta < \theta_0$
   c) $H_0 : \theta = \theta_0$ mot $H_1 : \theta \neq \theta_0$
2. Estimator for θ; $\widehat{\theta}$
3. La $Z = h(\widehat{\theta}, \theta_0)$, der $h(\cdot, \cdot)$ er ein funksjon s.a. Z har ei kjend fordeling under H_0
4. Bestem eit forkastningskriterium (antar Z stor når $\widehat{\theta}$ stor)
   a) Forkast H_0 dersom $Z > k$
   b) Forkast H_0 dersom $Z < k$
   c) Forkast H_0 dersom $Z < k_l$ eller $Z > k_u$

   der k bestemmes frå kravet

$$P(\text{Forkast H}_0 \text{ når H}_0 \text{ er riktig}) \leq \alpha$$

5. Sett inn tal og konkluder

p-verdi

Ein *p*-verdi er det lågaste signifikansnivået α slik at observert verdi for observatoren gjev at me skal forkaste H_0. Det vil seie, forkast H_0 dersom *p*-verdien er *mindre* enn α.

Teststyrke

Styrken til ein test er sannsynet for å forkaste H_0 gitt at ein spesifikk alternativ hypotese er sann.

H_0	Value of Test Statistic	H_1	Critical Region
$\mu = \mu_0$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$; σ known	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
$\mu = \mu_0$	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$; $v = n - 1$, σ unknown	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$z = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$; σ_1 and σ_2 known	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}}$; $v = n_1 + n_2 - 2$, $\sigma_1 = \sigma_2$ but unknown, $s_p^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t' = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$; $v = \dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$, $\sigma_1 \neq \sigma_2$ and unknown	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t' < -t_\alpha$ $t' > t_\alpha$ $t' < -t_{\alpha/2}$ or $t' > t_{\alpha/2}$
$\mu_D = d_0$ paired observations	$t = \dfrac{\bar{d} - d_0}{s_d/\sqrt{n}}$; $v = n - 1$	$\mu_D < d_0$ $\mu_D > d_0$ $\mu_D \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$

Lineær regresjon

Situasjon: observert $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
Modell:

$$Y_i | x_i \sim n(y_i; \alpha + \beta x_i, \sigma)$$

Kan tilpasse α, β, σ

- Minste kvadraters metode (kun α, β)
- Sannsynsmaksimering

La $Y_1, \ldots, Y_n$ være uavhengige variabler med samme varians σ^2 og forventningsverdier

$$\mathrm{E}(Y_i) = \alpha + \beta x_i\,, \quad i = 1, \ldots, n.$$

Minste kvadratsumsestimatorene for α og β er da

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\,, \qquad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x},$$

og en forventningsrett estimator for σ^2 er gitt ved

$$S^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2.$$

Dersom i tillegg $Y_1, \ldots, Y_n$ er normalfordelte vil

$$\frac{(n-2)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

være χ^2-fordelt med $n-2$ frihetsgrader. Det kan også vises at $(n-2)S^2/\sigma^2$ er uavhengig av $\hat{\alpha}$ og $\hat{\beta}$.

Prediksjon lineær regresjon

Merk at det er skilnad på følgande:

- Forventa respons $\mu_{Y|x_0}$
- Ein ny observasjon Y_0 (gitt x_0)

Eksamen mai 2009 oppgåve 3

Eksamen mai 2009 oppgåve 3

Oppgåve 3 Kontrastmiddel

Effekten av ulike typer kontrastmiddel brukt ved røntgenundersøkelsar av hender skal studerast. Kontrastmiddelet blir injisert i handflata før røntgenbiletet blir teke. Ein ynskjer å minske strålingsfaren ved å ta få bileter - helst berre eit av kvar hand.

For å måle effekten har ein utvikla eit kontrastmål for eit bilete av ei hand. Utan kontrastmiddel kallast målet K_0 og det varierer frå person til person, men kan sjåast på som identisk for begge hendene på ein person. Tidligare erfaring tilseier at K_0 er normalfordelt med forventningsverdi μ_0 og standardavvik σ_0. Det vil sei at K_0 er $n(k_0; \mu_0, \sigma_0)$.

Gå no ut i frå at μ_0 og σ_0 er ukjende. Ei studie på 10 forsøkspersonar blir brukt til å kartleggje kontrastmålet. Eit røntgenbilde utan bruk av kontrastmiddel blir teke av ei av hendene til kvar av dei 10 forsøkspersonane. Det resulterer i 10 uavhengige observasjonar av kontrastmålet K_0, sjå tabell 1.

Forsøksnr. i	1	2	3	4	5	6	7	8	9	10
$k_0(i)$	21	28	19	23	31	32	28	23	28	27

Tabell 1: Målt kontrast uten bruk av kontrastmiddel. Her blir $\bar{k}_0 = 1/10 \sum_{i=1}^{10} k_0(i) = 26$ og $\sum_{i=1}^{10}(k_0(i) - \bar{k}_0)^2 = 166$.

b) Utlei eit 90% konfidensintervall for forventa kontrastmål μ_0, og finn talsvar.

Eksamen mai 2009 oppgåve 3

Ved bruk av kontrastmiddel blir kontrasten i røntgenbileta endra slik at kontrastmålet blir:

$$K = K_0 + R$$

der R er effekten av kontrastmiddelet.

Gå ut i frå at R er normalfordelt med forventning μ_R og standardavvik σ_R, dvs $n(r; \mu_R, \sigma_R)$. Vidare går vi ut i frå at K_0 og R har ein korrelasjon på ρ_{0R}, og at K er normalfordelt $n(k; \mu_K, \sigma_K)$.

c) Utlei uttrykk for forventninga μ_K og standardavviket σ_K til kontrastmålet ved bruk av kontrastmiddel.

Eksamen mai 2009 oppgåve 3

Vi ønskjer no å samanlikne kontrastmåla ved bruk av to ulike kontrastmiddel, type A og type B. La effekten av kvar av desse vere R_A og R_B, og tilsvarande blir kontrastmåla:

$$K_A = K_0 + R_A$$

$$K_B = K_0 + R_B$$

Vi går ut i frå at alle variablane er normalfordelte, og at R_A og R_B er uavhengige. For å undersøkje kontrastmåla for dei to ulike kontrastmiddela gjennomfører vi et forsøksopplegg: For kvar type blir det gjort 10 forsøk. For dei 20 forsøkspersonane blir kontrastmiddelet injisert i ei av hendene, eit røntgenbilde blir teke, og kontrastmålet blir registrert. Dette gjev eit sett av uavhengige observasjonar av K_A og K_B, sjå tabell 2 og 3.

Forsøk nr (i)	1	2	3	4	5	6	7	8	9	10
$k_A(i)$	29	38	26	32	40	43	37	31	38	36

Tabell 2: Målt kontrast ved bruk av kontrastmiddel A. Her blir $\bar{k_A} = 1/10 \sum_{i=1}^{10} k_A(i) = 35$.

Forsøk nr (i)	1	2	3	4	5	6	7	8	9	10
$k_B(i)$	44	37	46	40	33	29	36	42	35	38

Tabell 3: Målt kontrast ved bruk av kontrastmiddel B. Her blir $\bar{k_B} = 1/10 \sum_{i=1}^{10} k_B(i) = 38$.

Gå i punkt d) og e) ut i frå at standardavvika til K_0 og R er kjende, $\sigma_0 = 4$ og $\sigma_R = 2$, at korrelasjonen mellom K_0 og R er kjend, $\rho_{0R} = 5/16$, og at standardavviket σ_R og korrelasjonen ρ_{0R} er lik for dei to kontrastmiddela, det vil sei $\text{Var}(R_A) = \text{Var}(R_B) = 2^2$ og $Corr(K_0, R_A) = Corr(K_0, R_B) = 5/16$.

Følgjande hypotese blir framsett: Forventa kontrastmål ved bruk av kontrastmiddel type A og type B er identiske. Denne hypotesen skal testast mot alternativet at dei to forventningane er ulike.

d) Test hypotesen over på signifikansnivå 0.1 ved å bruke dataane i tabell 2 og 3.

Utlei styrken for denne testen for forskjell i forventa kontrastmål lik 2.