



Avd. Matematisk statistik

KTH Matematik

TENTAMEN I 5B1503 STATISTIK MED FÖRSÖKSPLANERING FÖR K OCH B
MÅNDAGEN DEN 25 AUGUSTI 2003 KL 14.00–19.00.

Examinator: Gunnar Englund, 790 7416. *Tillåtna hjälpmedel:* Formel- och tabellsamling i Statistik med försöksplanering. Kalkylator.

Varje korrekt löst uppgift ger 10 poäng så maximalpoängen blir 50 poäng. För godkänt krävs 20 poäng. Resultatet anslås senast måndagen den 9 september 2003 på matematisk statistik's anslagstavla i entréplanet, Lindstedtsvägen 25.

Införda beteckningar skall förklaras och definieras. Resonemang och räkningar skall vara så utförliga att de är lätta att följa. Numeriska svar skall anges med minst två siffrors noggrannhet.

Lycka till!

Uppgift 1

För en viss kassa kan betjäningstiden per kund betraktas som en exponentialfördelad stokastisk variabel med väntevärde m minuter. Olika kunders betjäningstider är oberoende och så snart en kund är betjänad så påbörjas betjäningen av nästa kund. Av erfarenhet vet man att kassan 1 gång på 10 kan betjäna 25 eller fler kunder inom 60 minuter.

Tolka denna uppgift som en sannolikhet och bestäm m . Lämpliga och välmotiverade approximationer är tillåtna. (10 p)

Uppgift 2

För att undersöka effekten av ett rostskyddsmedel gör en person 20 mätningar. Dessa mätningar uppfattas som utfall av oberoende $N(m, \sigma)$ -fördelade stokastiska variabler där m och σ är okända. Personen gör, efter konstens alla regler, ett exakt 95%-igt konfidensintervall för m och erhåller intervallet $I_m = (30.24, 31.96)$. Tyvärr har de ursprungliga mätresultaten kommit bort när personen får i uppdrag att även beräkna ett konfidensintervall för σ . Han vill inte erkänna sitt slarv, utan ber sin statistiskt kunnige vän teknologen (dvs Dig) om hjälp. Ge honom den hjälpen, dvs beräkna ett exakt 95% konfidensintervall för σ . (10 p)

Uppgift 3

Man har undersökt sambandet mellan halten av en viss fibertyp (x) och kvalitén av producerat papper (y). Följande 13 par av (lämpligt transformerade) observationer (x_i, y_i) (=halt fibrer, kvalité) har erhållits:

x_i	5	10	70	75	70	90	95	140	170	190	192	210	230
y_i	0	2	9	6	4	8	6	3	13	12	14	17	18

Vi antar att $(x_1, y_1), \dots, (x_{13}, y_{13})$ kan beskrivas av en enkel linjär regressionsmodell dvs. som oberoende utfall av $(x_1, Y_1), \dots, (x_{13}, Y_{13})$, där varje Y_i är normalfördelad $N(\alpha + \beta x_i, \sigma)$. Räknehjälp (grafisk lösning godtages ej):

$$\begin{aligned} \sum_{i=1}^{13} x_i &= 1547 & \sum_{i=1}^{13} y_i &= 112 \\ \sum_{i=1}^{13} y_i^2 &= 1368 & \sum_{i=1}^{13} x_i^2 &= 251139 \\ \sum_{i=1}^{13} x_i y_i &= 17978. \end{aligned}$$

Skatta förväntad kvalitet vid fiberhalten $x=100$ samt medelfelet för denna skattning. (10 p)

Uppgift 4

Vid ett s.k. ringtest skulle man jämföra fyra laboratorier (B_1, B_2, B_3 och B_4) med varandra. De fick mäta halten av ett visst enzym i tre stycken prover (A_1, A_2 , och A_3). Data (i lämplig enhet):

	B_1	B_2	B_3	B_4	$\bar{y}_i$
A_1	6.43	7.36	7.03	6.22	6.76
A_2	9.35	9.86	9.08	9.27	9.39
A_3	6.03	6.90	6.69	6.50	6.53
$\bar{y}_j$	7.27	8.04	7.60	7.33	$\bar{y}_{..} = 7.56$

$\sum_i \sum_j y_{ij}^2 = 707.6222$. Man antar att det kan finnas systematiska skillnader mellan laboratorierna, men att dessa skillnader i så fall är desamma oberoende av prov. Följande kvadratsummor beräknades

Mellan prov	20.1992
Mellan laboratorier	1.1070
Totalt	21.7790

a) Testa hypotesen att laboratorierna ej skiljer sig åt. 5 % signifikansnivå. (5 p)

b) Testa med konfidensmetoden på 5% nivå om det föreligger någon skillnad mellan laboratorium B_1 och laboratorium B_4 . (5 p)

Både i a) och b)-delarna skall slutsatserna klart framgå.

Uppgift 5

I ett försök för att utröna tillförlitligheten hos kullager bedömdes tre faktorer vara väsentliga; värmebehandling av innerringen (A), ytterrings s.k. oskulation (B) samt hållarkonstruktionen

(C). De tre faktorerna undersöktes vid standardutseende (–) och vid modifierad konstruktion (+). Resultaten från livslängdstestet (två observationer (y) för varje konstruktion):

Försök nr (i)	A	B	C	y_{i1}	y_{i2}	Medelvärde $\bar{y}_i$	Varians s_i^2
1	–	–	–	16	18	17	2
2	+	–	–	21	31	26	50
3	–	+	–	20	30	25	50
4	+	+	–	96	88	92	32
5	–	–	+	19	19	19	0
6	+	–	+	13	19	16	18
7	–	+	+	18	24	21	18
8	+	+	+	99	93	96	18

$$\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 = 16\,508$$

Som räknehjälp ges att med

$$\mathbf{M} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \end{pmatrix} \text{ så erhålles } \mathbf{M}\bar{\mathbf{y}} = \begin{pmatrix} 312 \\ 148 \\ 156 \\ -8 \end{pmatrix}$$

där $\bar{\mathbf{y}}$ är medelvärdeskolumnen ovan.

- a) Ställ upp en lämplig modell samt testa vilka effekter som är signifikanta (nivå 5%). (6 p)
- b) Reducera modellen för de icke-signifikanta effekterna i a). Beräkna i denna reducerade modell en ny σ -skattning samt ange hur många frihetsgrader den har. (4 p)



Avd. Matematisk statistik

KTH Matematik

LÖSNINGAR TILL TENTAMEN I 5B1503 STATISTIK MED FÖRSÖKSPLANERING FÖR K MÅNDAGEN DEN 25 AUGUSTI 2003 KL 14.00-19.00

Uppgift 1

Låt X_k vara betjäningstiden (i minuter) för kund k . Sätt
 $Y = X_1 + X_2 + \dots + X_{25}$ = betjäningstiden för 25 kunder.

Enligt uppgiften är $P(Y \leq 60) = 0.1$.

Enligt centrala gränsvärdessatsen gäller att Y är approximativt $N(25m, m\sqrt{25})$ -fördelad ty
 $E(Y) = m$ och $V(Y) = m^2$.

Detta ger

$$0.9 = P(Y > 60) = P\left(\frac{Y - 25m}{5m} > \frac{60 - 25m}{5m}\right) \approx 1 - \Phi\left(\frac{60 - 25m}{5m}\right) = \Phi\left(\frac{25m - 60}{5m}\right),$$

dvs.

$$\frac{25m - 60}{5m} \approx \lambda_{0.10} \approx 1.28 \quad \text{eller} \quad \underline{m \approx 3.23}.$$

Uppgift 2

Konfidsensintervallet för m var $\bar{x} \pm t_{0.025}(20-1) \frac{s}{\sqrt{20}}$ som ger (eftersom $t_{0.025}(19) \approx 2.09$ enligt tabell 3)

$$\bar{x} - 2.09 \frac{s}{\sqrt{20}} \approx 30.24 \text{ och } \bar{x} + 2.09 \frac{s}{\sqrt{20}} \approx 31.96$$

dvs $\bar{x} \approx \frac{1}{2}(30.24 + 31.96) = 31.10$ och

$s = \frac{\sqrt{20}}{2.09}(\bar{x} - 30.24) \approx 1.84$. Konfidsensintervallet för σ blir

$$\left(s \sqrt{\frac{(20-1)}{\chi_{0.025}^2(19)}}, s \sqrt{\frac{(20-1)}{\chi_{0.975}^2(19)}} \right) \approx \underline{(1.40, 2.69)}.$$

(Man kan även tänka sig ett enkelsidigt konfidsensintervall där man då har undre gränsen 0 och har $\chi_{0.95}^2(19)$ i stället för $\chi_{0.975}^2(19)$ i nämnaren i övre gränsen.)

Uppgift 3

Man har att

$$\bar{x} = \frac{1}{13} \sum_{i=1}^{13} x_i = 119, \quad \bar{y} = \frac{1}{13} \sum_{i=1}^{13} y_i = 8.61538.$$

Enligt formelsamlingen fås

$$\beta^* = \frac{\sum_{i=1}^{13} x_i y_i - 13 \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^{13} x_i^2 - 13 \cdot \bar{x}^2} = \frac{17978 - 13 \cdot 119 \cdot 8.61538}{251139 - 13 \cdot 119^2} = 0.069364.$$

Minstakvadratskattningarna är:

$$\alpha^* = \bar{y} - \beta^* \bar{x} = 8.61538 - 0.069364 \cdot 119 = 0.361064, \text{ och } \beta^* = 0.0694.$$

Den skattade regressionslinjen är med andra ord

$$\hat{y}(x) = \alpha^* + \beta^* x = 0.361064 + 0.0694 \cdot x$$

Med $x_0 = 100$ ser vi att vi skattar förväntad papperskvalité med $\alpha^* + \beta^* \cdot 100 \approx 7.297 \approx \underline{7.3}$.

För att skatta s beräknas residualkvadratsumman Q_0 . Enligt formelsamlingen gäller att

$$\begin{aligned} Q_0 &= \sum_{i=1}^{13} y_i^2 - 13 \bar{y}^2 - (\beta^*)^2 \left(\sum_{i=1}^{13} x_i^2 - 13 \bar{x}^2 \right) = \\ &= 1368 - 13 \cdot 8.61538^2 - 0.0694^2 (251139 - 13 \cdot 119^2) = 80.1603. \end{aligned}$$

Därmed skattas σ^2 av

$$s^2 = \frac{Q_0}{13 - 2} = \frac{80.1603}{11} = 7.2873.$$

$\alpha^* + \beta^* \cdot 100 = \bar{Y} + \beta^*(100 - \bar{x})$ ser vi har variansen ($\bar{Y}$ och β^* är ju oberoende)

$$\begin{aligned} V(\bar{Y}) + (100 - \bar{x})^2 V(\beta^*) &= \frac{\sigma^2}{13} + (100 - 119)^2 \frac{\sigma^2}{\sum_{i=1}^{13} (x_i - \bar{x})^2} = \sigma^2 \cdot \left(\frac{1}{13} + \frac{(100 - 119)^2}{\sum_{i=1}^{13} (x_i - \bar{x})^2} \right) = \\ &= \sigma^2 \cdot \left(\frac{1}{13} + \frac{19^2}{251139 - 13 \cdot 119^2} \right) \approx \sigma^2 \cdot 0.0823 \end{aligned}$$

och vi får alltså att medelfelet är $s\sqrt{0.0823} = \sqrt{7.2873 \cdot 0.0823} \approx \underline{0.774}$.

Svar: Skattningen är 7.3 med medelfelet 0.78.

Uppgift 4

Vi har här en två-sidig variansanalys med en replikation/cell och vi har en additiv modell dvs $Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$ där Y_{ij} =resultatet för prov nr i vid laboratorium j , $i = 1, 2, 3$, $j = 1, 2, 3, 4$. Vi får enligt formelsamlingen avsnitt 14.2 följande variansanalystabell där $p = 3$, $q = 4$ och där kvs för residualen fås med subtraktion till $21.7790 - 20.1992 - 1.1070 = 0.4728$.

Källa	fr.gr.	kvs	mkvs
Mellan prov	3-1=2	20.1992	10.0996
Mellan laboratorier	4-1=3	1.1070	0.3690
Residual	(3-1)(4-1)=6	0.4728	$0.0788=\hat{\sigma}^2$
Totalt	12-1=11	21.7790	

a) Test av om laboratorier är lika. $H_0 : \beta_i = 0, i = 1, 2, 3, 4$. Testkvot för test av H_0 är $\frac{0.3690}{0.0788} \approx 4.6827 < F_{0.05}(3, 6) \approx 4.76$ och således kan ej H_0 förkastas på nivån 5%.

b) $\beta_1 - \beta_4$ skattas med $\bar{y}_{.1} - \bar{y}_{.4} = 7.27 - 7.33 = -0.06$. Vidare är $V(\hat{\beta}_1 - \hat{\beta}_4) = V(\bar{Y}_{.1} - \bar{Y}_{.4}) =$ (oberoende) $= V(\bar{Y}_{.1} + \bar{Y}_{.4}) = \frac{\sigma^2}{3} + \frac{\sigma^2}{3} = \frac{2\sigma^2}{3}$ och alltså är $D(\hat{\beta}_1 - \hat{\beta}_4) = \sigma\sqrt{\frac{2}{3}}$ och vi får ett 95%-igt konfidensintervall för $\beta_1 - \beta_4$ till

$$-0.06 \pm t_{0.025}(6)\hat{\sigma}\sqrt{\frac{2}{3}} \approx -0.06 \pm 2.45\sqrt{0.0788}\sqrt{\frac{2}{3}} \approx -0.06 \pm 0.56$$

Eftersom 0 tillhör konfidensintervallet kan $H_0 : \beta_1 = \beta_4$ ej förkastas på nivån 5%.

Uppgift 5

Vi har här en 2^3 -modell med två observationer per försökspunkt. Vi får med hjälp av ledningen att $\hat{\mu} = \frac{312}{8} = 39$, $\hat{A} = \frac{148}{8} = 18.5$, $\hat{B} = \frac{156}{8} = 19.5$, $\hat{C} = \frac{-8}{8} = -1$, Vi får vidare med hjälp av teckenschemat från formelsamlingen avsnitt 15.2 att

$$\widehat{AB} = \frac{1}{8}(17 - 26 - 25 + 92 + 19 - 16 - 21 + 96) = 17$$

$$\widehat{AC} = \frac{1}{8}(17 - 26 + 25 - 92 - 19 + 16 - 21 + 96) = -0.5$$

$$\widehat{BC} = \frac{1}{8}(17 + 26 - 25 - 92 - 19 - 16 + 21 + 96) = 1$$

$$\widehat{ABC} = \frac{1}{8}(-17 + 26 + 25 - 92 + 19 - 16 - 21 + 96) = 2.5$$

Vi får vidare

$$\hat{\sigma}^2 = \frac{1}{8}(s_1^2 + s_2^2 + s_3^2 + s_4^2 + s_5^2 + s_6^2 + s_7^2 + s_8^2) = \frac{188}{8} = 23.50$$

som är baserad på 8 frihetsgrader ($\hat{\sigma} = \sqrt{23.50} \approx 4.85$). De enskilda effektskattningarna har alla variansen

$$\frac{\sigma^2}{2^3 \cdot 2} = \frac{\sigma^2}{16}$$

vilket gör att vi får konfidensintervallen effektskattning $\pm t_{0.025}(8)\hat{\sigma}\frac{1}{\sqrt{16}} = \text{skattn.} \pm 2.80$.

Alltså är de effektskattningar som är till beloppet större än 2.80 signifikanta, vilket innebär A, B, AB förutom μ .

b) Vi bildar en ny kvadratsumma genom att till residualkvadratsumman (Inom celler) lägga

$$2^3 \cdot 2(\widehat{C}^2 + \widehat{AC}^2 + \widehat{BC}^2 + \widehat{ABC}^2)$$

Alltså får vi den nya σ -skattningen till

$$\frac{1}{12} (8 \cdot \hat{\sigma}^2 + 16 ((-1)^2 + (-0.5)^2 + 1^2 + (2.5)^2)) = \frac{1}{12}(188 + 16 \cdot 8.5) = 27$$

dvs den nya σ -skattningen blir $\sqrt{27} \approx 5.20$ och denna skattning är baserad på $8 + 4 = 12$ frihetsgrader.

För övrigt kan man inse denna nya reducerade modell faktiskt innebär ett fullständigt 2^2 -försök med 4 observationer per försökspunkt, där man alltså bortser från faktorn C .