Avd. Matematisk statistik

KTH Matematik

TENTAMEN I SF1913 MATEMATISK STATISTIK FÖR IT OCH ME
ONSDAGEN DEN 12 JANUARI 2011 KL 14.00–19.00.

Examinator: Camilla Landén , tel. 790 8466.

Tillåtna hjälpmedel: Formel- och tabellsamling i Matematisk statistik. Räknare. Extrablad om icke-parametriska test finns sist i tentamen.

Införda beteckningar skall förklaras och definieras. Resonemang och uträkningar skall vara så utförliga och väl motiverade att de är lätta att följa. Numeriska svar skall anges med minst två siffrors noggrannhet. Tentamen består av 6 uppgifter. Varje korrekt lösning ger 10 poäng. Gränsen för godkänt är preliminärt 24 poäng. Möjlighet att komplettera ges för de tentander med 22–23 poäng. Det ankommer på dig själv att ta reda på om du har rätt att komplettera.

Uppgift 1

En maskin fyller en vätska på flaskor i ett bryggeri. Kontrollmätningar har visat att den påfyllda volymen kan betraktas som en normalfördelad stokastisk variabel med väntevärde m och standardavikelse $\sigma = 4$ ml, där m kan ställas in av maskinens operatör. På flaskornas etikett står det att innehållet är 330 ml.

(a) Hur bör m väljas för att sannolikheten att en flaska ska få ett innehåll mindre än 330 ml är 0.1? (5 p)

(b) Flaskorna ställs i backar med 20 st i varje. Om man väljer $m = 332$ ml vid påfyllningen, hur stor är då sannolikheten för att en back skall innehålla mindre än 6600 ml vätska? (Vätskemängderna i olika flaskor är oberoende av varandra.) (5 p)

Uppgift 2

En digitalt kommunikationssystem fungerar så att en sändare skickar en spänning som är antingen 0 volt (för en digital 0:a) eller 1.8 volt (för en digital 1:a). På vägen till mottagaren störs signalen av additivt normalfördelat brus med väntevärde 0 volt och standardavvikelse 0.45 volt. Den *mottagna* signalen X är alltså sådan att betingat att en 0:a har sänts är X normalfördelad med väntevärde 0 volt, och betingat att en 1:a har sänts är X normalfördelad med väntevärde 1.8 volt. Standardavvikelsen för inspänningen är 0.45 volt i båda fallen.

Beslutskretsen i mottagaren fungerar enligt följande. Om inspänningen är större än 0.9 volt fattas beslutet ”1:a sänd”, och om den är mindre fattas beslutet att ”0:a sänd”.

(a) Bestäm den betingade felsannolikheten för att beslutskretsen tar beslutet ”0:a sänd” givet att en 1:a i själva verket sändes. (3 p)

(b) Låt $p_0 = 0.4$ vara sannolikheten att en nolla är sänd och $p_1 = 0.6$ vara sannolikheten att en etta är sänd. Bestäm sannolikheten för att en etta faktiskt sändes givet att beslutskretsen tar beslutet ”1:a sänd”. (7 p)

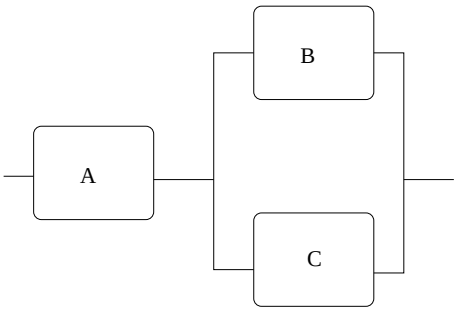
Uppgift 3

Ett system består av tre komponenter enligt figuren nedan. Systemet fungerar så länge komponent A och minst en av B och C fungerar.

(a) Antag att komponenterna fungerar oberoende av varandra och med sannolikheter $p_A = 0.9$, $p_B = 0.7$ och $p_C = 0.8$. Beräkna sannolikheten att systemet fungerar. (3 p)

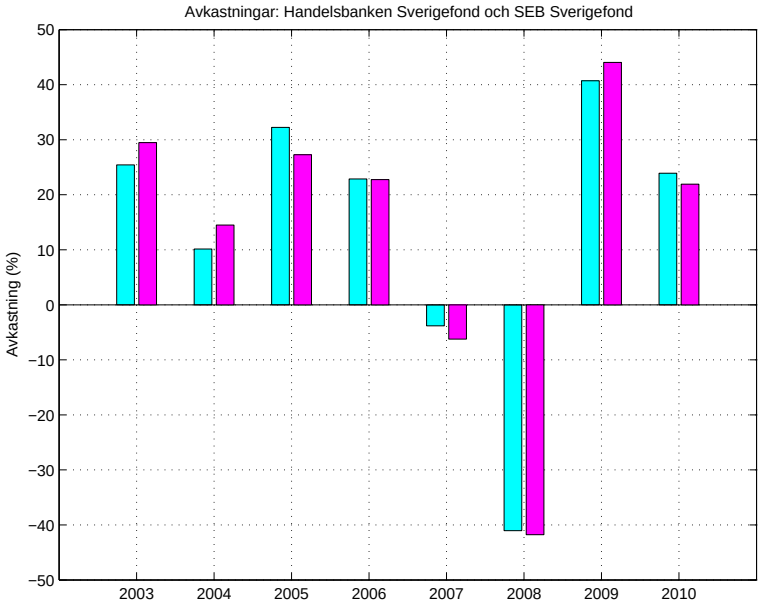
(b) Beräkna den betingade sannolikheten att komponent B fungerar, givet att systemet fungerar. (4 p)

(c) Antag komponenternas livslängder T_A, T_B respektive T_C är oberoende och alla exponentialfördelade med väntevärde $1/2$ år. Låt T vara systemets livslängd. Beräkna fördelningsfunktionen för T.
Ledning. Notera att om X är livslängd för en komponent eller ett system så är $P(X > t)$=sannolikheten att komponenten (systemet) fungerar vid tidpunkt t. (3 p)



Uppgift 4

Nedanstående figur visar avkastningarna, i procent, för aktiefonderna *Handelsbanken Sverigefond* och *SEB Sverigefond* för åren 2003–2010.



De numeriska värdena (i %, från fondsidorna på `dn.se`) är enligt följande tabell:

år	2003	2004	2005	2006	2007	2008	2009	2010
Handelsbanken	25.4	10.2	32.2	22.9	–3.8	–41.0	40.7	23.9
SEB	29.5	14.5	27.3	22.8	–6.2	–41.8	44.0	21.9

Det kan vara intressant att undersöka, speciellt eftersom båda fonderna har samma inriktning (sverigefonder), om det finns någon systematisk skillnad mellan deras förväntade avkastningar.

(a) Gör detta, utgående från de givna data, med ett test på nivån 0.05 i en modell baserad på normalfördelningsantaganden. Var noga med att ange vilken modell du arbetar med och vilka antaganden du gör. (5 p)

Antag att man är osäker på normalfördelningsantagandet. Genom att gå tillbaka till år 1995 fås data för 16 år. Det visar sig då att av differenserna mellan avkastningarna för SEB och Handelsbanken är 9 positiva (ingen differens är lika med noll) och rangsumman för SEB är 264.

b) Gör ett icke-parametriskt test, utgående från data för 16 år, av om det på nivå 5% finns någon systematisk skillnad mellan fondernas förväntade avkastningar. (5 p)

Uppgift 5

Låt oss anta att det i branschen för pälsschampoo för hundar finns två dominerande tillverkare, A och B, som tillsammans har drygt 50% av marknaden.
Tillverkare A gör en marknadsundersökning i vilken 1000 hundägare tillfrågas, och 184 av dessa säger sig föredra schampoo som A säljer. Tillverkare B, som är den största på marknaden och vill trycka ner konkurrensen, gör då en egen undersökning i vilken 196 av 500 tillfrågade hundägare säger sig föredra schampoot från B. Detta tar tillverkare B som intäkt för att i en stor kampanj påstå att "Vårt schampoo är mer än dubbelt så populärt som någon annan tillverkares schampoo". Vi skall undersöka, ur statistisk synvinkel, om detta håller.

(a) Definiera storheten, eller parametern,

$$\begin{aligned}\Delta &= \text{andelen hundägare som föredrar pälsschampoo från B} \\ &- 2 \times \text{andelen hundägare som föredrar pälsschampoo från A.}\end{aligned}$$

Använd tillverkare A:s undersökning för att skatta andelen andelen hundägare som föredrar pälsschampoo från A, och tillverkare B:s undersökning för att skatta andelen andelen hundägare som föredrar pälsschampoo från B, för att konstruera en skattning Δ^* av Δ. Räkna också ut denna skattnings värde för de aktuella data. Det vill säga, i bokens terminologi, ange både stickprovsvariabel och skattning. (2 p)

(b) Beräkna variansen av Δ^* uttryckt i lämpliga parametrar. (3 p)

(c) Ange en uppskattning av standardavvikelsen för Δ^*, dvs dess medelfel, för de aktuella data. (3 p)

(d) Finns det fog för B:s påstående i kampanjen (jämför A och B)? Svara på frågan med hjälp av ett lämpligt konfidensintervall eller test och välj signifikansnivå själv. (2 p)

Uppgift 6

Enligt en modell har en tillverkad enhet fel av typ A med sannolikhet 0.03 och fel av typ B med sannolikhet 0.07 och A-fel och B-fel uppträder oberoende av varandra.
Man tillverkar $n = 3000$ enheter och delar in enheterna i kategorier baserat på deras fel:

	felfria	Enbart A	Enbart B	Både A och B	
Observationer, x_i:	2659	120	210	11	$n = 3000$

Testa på nivå $\alpha = 0.05$ ifall modellen är förenlig med mätningarna. (10 p)

Icke-Parametriska Test

- **Teckentestet.** Låt $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ vara ett stickprov i par. Bilda differenserna mellan x-observationerna och y-observationerna och låt t vara antalet gånger differensen är strikt positiv. Då är t en observation av T som är $Bin(n_z, 0.5)$, under förutsättning att x_i och y_i är observationer ur samma fördelning. Med n_z avses antalet differenser som inte är noll.

- **Wilcoxons Rangsummetest.** Låt $x_1, x_2, \ldots, x_{n_1}$ och $y_1, y_2, \ldots, y_{n_2}$ vara två oberoende stickprov. Låt r vara rangsumman för x-observationerna, då x-observationerna och y-observationerna storleksordnats. Då gäller att r är en observation av R för vilken

$$E(R) = n_1 \frac{n_1 + n_2 + 1}{2} \quad \text{och} \quad V(R) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12},$$

  under förutsättning att x-observationerna och y-observationerna kommer från samma fördelning. Förutom för små n_1 och n_2 är R approximativt normalfördelad.

LÖSNINGAR TILL
TENTAMEN I SF1913 MATEMATISK STATISTIK FÖR IT OCH ME
ONSDAGEN DEN 12 JANUARI 2011 KL 14.00–19.00.

Uppgift 1

(a) Sätt X = innehållet i en flaska, vilket innebär att X är $N(m, 4)$-fördelad. Vi vill ha

$$0.1 = P(X < 330) = P\left(\frac{X-m}{4} < \frac{330-m}{4}\right).$$

Av detta följer att $\frac{330-m}{4} = -\lambda_{0.10} = -1.2816$ eller $m = 330 + 4\lambda_{0.10} = 330 + 4 \cdot 1.2816 = \underline{335.13}$.

(b) Låt X_i vara innehållet flaska nr i och låt Y beteckna backens totala innehåll, dvs

$$Y = X_1 + \ldots + X_{20}.$$

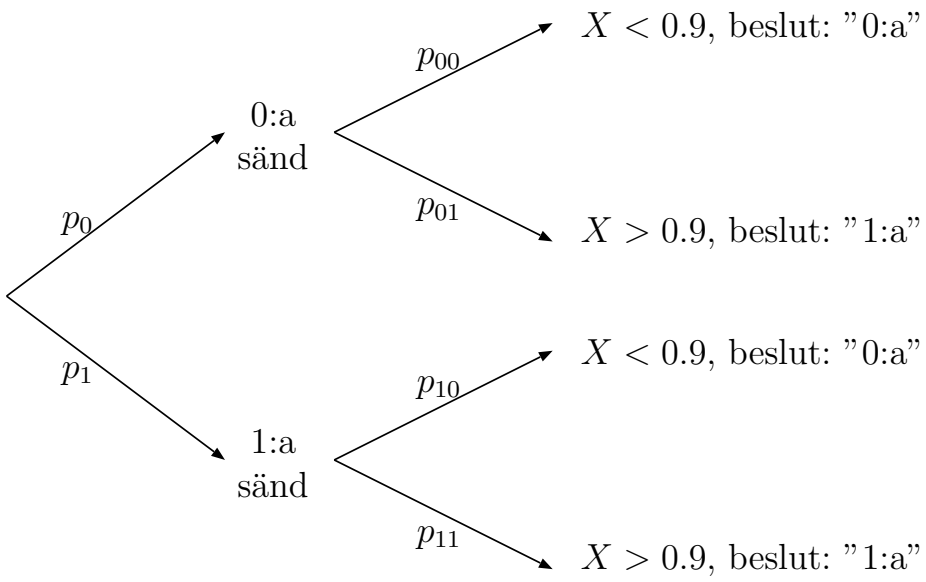
Eftersom summor av oberoende normalfördelade stokastiska variabler är normalfördelade så gäller det att Y är $N(20 \cdot 332, \sqrt{20} \cdot 4) = N(6640, 17.89)$-fördelad. Detta ger

$$P(Y < 6600) = P\left(\frac{Y-6640}{17.89} < \frac{6600-6640}{17.89}\right) = \Phi\left(\frac{6600-6640}{17.89}\right)$$

$$= \Phi(-2.24) = 1 - 0.98745 = \underline{0.0126}.$$

Uppgift 2

Med X som uppmätt spänning kan vi beskriva beslutskretsen med följande träddiagram:



Om en 1:a sändes är X N(1.8, 0.45) och

$$\begin{aligned} p_{10} &= P\left(X < 0.9\right) = P\left(\frac{X-1.8}{0.45} < \frac{0.9-1.8}{0.45}\right) = \Phi\left(\frac{0.9-1.8}{0.45}\right) = \Phi(-2) \\ &= 1 - \Phi(2) = 0.0228. \end{aligned}$$

Detta är den sökta sannolikheten i a).
Vidare, $p_{11} = 1 - p_{10} = 0.9772$. Givet att en 0:a sändes är X N(0, 0.45) och

$$p_{00} = P\left(X < 0.9\right) = P\left(\frac{X-0}{0.45} < \frac{0.9-0}{0.45}\right) = \Phi\left(\frac{0.9-1.8}{0.45}\right) = \Phi(2) = 0.9772$$

och $p_{01} = 1 - p_{00} = 0.0228$. (Vi har samma felsannolikhet oavsett utsänt bit.) Givet är att

$$p_0 = P\left(\text{0:a sänd}\right) = 0.4 \qquad p_1 = P\left(\text{1:a sänd}\right) = 0.6.$$

Alltså,

$$\begin{aligned} P\left(\text{1:a sänd}|\text{beslut: "1:a sänd"}\right) &= \frac{P\left(\text{1:a sänd och beslut: "1:a"}\right)}{P\left(\text{beslut: "1:a"}\right)} \\ &= \frac{p_1 \cdot p_{11}}{p_0 \cdot p_{01} + p_1 \cdot p_{11}} \\ &= \frac{0.60 \cdot 0.9772}{0.40 \cdot 0.0228 + 0.60 \cdot 0.9772} = 0.9847. \end{aligned}$$

Uppgift 3

a) Låt A, B och C stå för händelserna att komponent A, B respektive C fungerar och S för att systemet fungerar. Vi har då att $S = A \cap (B \cup C)$ och alltså

$$\begin{aligned} P(S) = P\big(A \cap (B \cup C)\big) = (\text{oberoendet}) = P(A)P(B \cup C) = \\ P(A)\big(P(B) + P(C) - P(B)P(C)\big) = 0.9 \cdot (0.7 + 0.8 - 0.7 \cdot 0.8) = \underline{0.846}. \end{aligned}$$

b) Vi söker $P(B \mid S) = \dfrac{P(B \cap S)}{P(S)}$. Nämnaren är beräknad i a). Vidare har vi att $B \cap S = B \cap A \cap (B \cup C) = B \cap A$ och således är $P(B \cap S) = P(B \cap A) = P(B)P(A) = 0.9 \cdot 0.7 = 0.63$. Det ger oss $P(B \mid S) = \frac{0.63}{0.846} = \underline{0.7447}$.
c) Med beteckningar som a) ser vi att $P(T > t) = P(S) = P(A)\big(P(B) + P(C) - P(B)P(C)\big) = P(T_{\mathrm{A}} > t)\big(P(T_{\mathrm{B}} > t) + P(T_{\mathrm{C}} > t) - P(T_{\mathrm{B}} > t)P(T_{\mathrm{C}} > t)\big)$. Eftersom T_A är exponentialfördelad är $P(T_{\mathrm{A}} > t) = \int_t^\infty 2e^{-2x}dx = e^{-2t}$ och likadant för de övriga komponenternas livslängder. Vi erhåller därför till slut $P(T_S > t) = 2e^{-4t} - e^{-6t}$ och fördelningsfunktionen ges av

$$F_{T_S}(t) = P(T_S \le t) = 1 - P(T_S > t) = \underline{1 - 2e^{-4t} + e^{-6t}}, \; t \ge 0.$$

Uppgift 4

(a) Låt $x_1, \ldots, x_8$ och $y_1, \ldots, y_8$ beteckna avkastningarna för Handelsbanken respektive SEB under de åtta åren. Eftersom det finns en kraftig samvariation (som är börsens generella utveckling under de olika åren) så är det lämpligt att arbeta med modellen *stickprov i par*. Med normalfördelningsantaganden blir alltså modellen att med $z_i = y_i - x_i$ är $z_1, \ldots, z_8$ oberoende observationer från $N(\mu, \sigma)$. Uppgiften är att undersöka om en systematisk skillnad finns, dvs om $\mu \neq 0$.

De $n = 8$ z-värdena är 4.1, 4.3, –4.9, –0.1, –2.4, –0.8, 3.3, –2.0, vilket ger $\sum z_i = 1.5$, $\sum z_i^2 = 80.61$, $\overline{z} = 0.1875$, $Q_z = \sum z_i^2 - n^{-1}(\sum z_i)^2 = 80.3288$, $s^2 = Q_z/(n-1) = 11.4755$, $s = 3.3876$.
Vi vill nu testa $H_0 : \mu = 0$ mot $H_1 : \mu \neq 0$. Under våra antaganden är $(\overline{z} - \mu)/(s/\sqrt{n})$ en observation från en t-fördelning med $n - 1 = 7$ frihetsgrader. Sätter vi här $\mu = 0$ får vi den observerade teststorheten 0.1566. Vi skall förkasta H_0 om detta tal är långt ute i svansarna på t-fördelningen, närmare bestämt om $|0.1566| > t_{0.025}(7) = 2.36$. Detta är uppenbarligen inte fallet, så H_0 kan inte förkastas på nivån 5%; vi kan inte hitta någon systematisk skillnad i förväntad avkastning mellan fonderna.

(b) Teckentest. Om de förväntade avkastningarna var lika stora borde T=antalet positiva differenser vara Bin$(16, 1/2)$-fördelad. Vi har fått $t = 9$ som utfall och beräknar sannolikheten att få lika extremt (eller extremare) utfall och gör detta "dubbelsidigt".

$$P = P(T \leq 7 \text{ eller } T \geq 9) = 2P(T \leq 7) = 2 \cdot 0.40181 \approx 0.80.$$

Eftersom denna sannolikhet $\geq 5\%$ kan
H_0 : "de förväntade avkastningarna för fonderna är lika stora" inte förkastas på nivån 5%.

Uppgift 5

(a) Låt p_A och p_B beteckna andelen hundägare som föredrar schampoo från tillverkare A respektive B, låt n_1 och n_2 beteckna antalet tillfrågade i de båda undersökningarna (vi har $n_1 = 1000$ och $n_2 = 500$) och låt x_A och x_B beteckna antalet tillfrågade som i undersökning 1 föredrog A respektive i undersökning 2 föredrog B (vi har $x_A = 184$ och $x_B = 196$).

Vi kan skatta p_A och p_B med $p_A^* = X_A/n_1$ respektive $p_B^* = X_B/n_2$, där X_A och X_B är de stokastiska variabler som x_A respektive x_B är observationer av. Som skattning av Δ kan vi sedan ta $\Delta^* = p_B^* - 2p_A^*$. Med de aktuella data får vi skattningen $196/500 - 2 \times 184/1000 = 0.024$.

(b) Det är rimligt att anta att de olika hundägare som ingick i undesökningarna har åsikter som är oberoende av varandra. Vi får då $X_A \in \text{Bin}(n_1, p_A)$ och $X_B \in \text{Bin}(n_2, p_B)$. Eftersom resultaten kommer från olika undersökningar är X_A och X_B oberoende (det hade inte varit fallet om de kom från en och samma undersökning), och därför gäller

$$V(\Delta^*) = V(X_B/n_2 - 2X_A/n_1) = \frac{V(X_B)}{n_2^2} + (-2)^2\frac{V(X_A)}{n_A^2} = \frac{p_B(1-p_B)}{n_2} + 4\frac{p_A(1-p_A)}{n_1}.$$

(c) Vi kan få en skattning av variansen för Δ^* genom att ersätta p_A och p_B i ovanstående uttryck med motsvarande skattningar $196/500 = 0.392$ och $184/1000 = 0.184$. Detta ger variansskattningen 0.00108. Roten ur detta, 0.0328, är en skattning av standardavvikelsen för Δ^*, dvs det är medelfelet för denna skattning.

(d) Låt $d(\Delta^*)$ beteckna medelfelet för skattningen Δ^*. Under våra förutsättningar på n_1, n_2, p_A och p_B gäller att fördelningarna för både X_A och X_B kan approximeras med normalfördelningar. Eftersom Δ^* är en linjärkombination av dessa två oberoende variabler kan även Δ^* anses vara approximativt normalfördelad. Denna variabel har väntevärde $\Delta = p_B - 2p_A$, ty $E(p_A^*) = p_A$ och $E(p_B^*) = p_B$ (båda skattningarna är väntevärdesriktiga). Därför gäller att $(\Delta^* - \Delta)/d(\Delta^*)$ ungefär är fördelad som en $N(0, 1)$-variabel.

Påståendet i kampanjen är $p_B > 2p_A$, dvs $\Delta > 0$. Vi kontrollerar om det är rimligt att påstå detta genom att testa $H_0 : \Delta = 0$ mot $H_1 : \Delta > 0$. Om H_0 är sann gäller således att $(\Delta^* - 0)/d(\Delta^*) = 0.024/0.0328 = 0.73$ är en observation från $N(0, 1)$. Vi skall förkasta H_0 till förmån för H_1 om

detta värde ligger långt ut i högra svansen på fö@rdelningen $N(0,1)$, mer precist om det är större än t ex 5%-kvantilen $\lambda_{0.05} = 1.64$. Detta är inte fallet, så det finns inget statistiskt underlag för vad som hävdas i kampanjen.

Alternativt kan vi göra ett nedåt begränsat approximativt 95%-igt konfidensintervall för Δ: $[\Delta^* - \lambda_{0.05} d(\Delta^*), \infty) = [-0.030, \infty)$. Intervallet innehåller talet 0, och därför kan inte H_0 förkastas mot H_1.

Uppgift 6

Låt A och B vara händelserna att en tillverkad enhet har A-fel resp. B-fel. Enligt modellen är

$$\begin{aligned} P(A^* \cap B^*) &= P(A^*)\,P(B^*) = 0.9021 \\ P(A \cap B^*) &= P(A)\,P(B^*) = 0.0279 \\ P(A^* \cap B) &= P(A^*)\,P(B) = 0.0679 \\ P(A \cap B) &= P(A)\,P(B) = 0.0021 \end{aligned}$$

så

	felfria	Enbart A	Enbart B	Både A och B	
Observationer, x_i:	2659	120	210	11	3000
Hypotes, p_i:	.9021	.0279	.0679	.0021	1
Förväntat, np_i:	2706.3	83.7	203.7	6.3	3000

Notera att $np_i \geq 5$, $i = 1, \ldots, 4$. De observerade antalen jämförs mot de förväntade med

$$q = \sum_{i=1}^{4} \frac{(x_i - np_i)^2}{np_i}$$

som om modellen stämmer är ett utfall från en approx. $\chi^2(4-1) = \chi^2(3)$-fördelning. Ur tabell fås att $\chi^2_{0.05}(3) = 7.81$. Vi observerar utfallet

$$q = \sum_{i=1}^{4} \frac{(x_i - np_i)^2}{np_i} = 0.83 + 15.74 + 0.195 + 3.51 = 20.3 > 7.81$$

så *vi förkastar hypotesen* om att A- och B-fel uppträder oberoende med de givna sannolikheterna.