Avd. Matematisk statistik

KTH Matematik

TENTAMEN I SF1901,SF1905,SF1907 OCH SF1908 SANNOLIKHETSTEORI OCH STATISTIK, ONSDAGEN DEN 12:E JANUARI 2011 KL 14.00–19.00.

Kursledare: Gunnar Englund för D och I, tel. 790 7416.
Kursledare: Tobias Rydén för T och M, tel. 790 8469.

Tillåtna hjälpmedel: Formel- och tabellsamling i Matematisk statistik, Mathematics Handbook (Beta), räknare.

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 tentander med 22–23 poäng. Tid och plats för komplettering kommer att anges på kursens hemsida. Det ankommer på dig själv att ta reda på om du har rätt att komplettera.

Ingen poäng från kontrollskrivningar tillgodoräknas.

Tentamen kommer att vara rättad inom tre arbetsveckor från skrivningstillfället och kommer att finnas tillgänglig på studentexpeditionen minst sju veckor efter skrivningstillfället.

Uppgift 1

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 2

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)

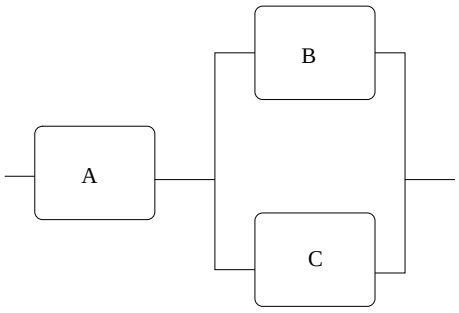
Uppgift 3

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)

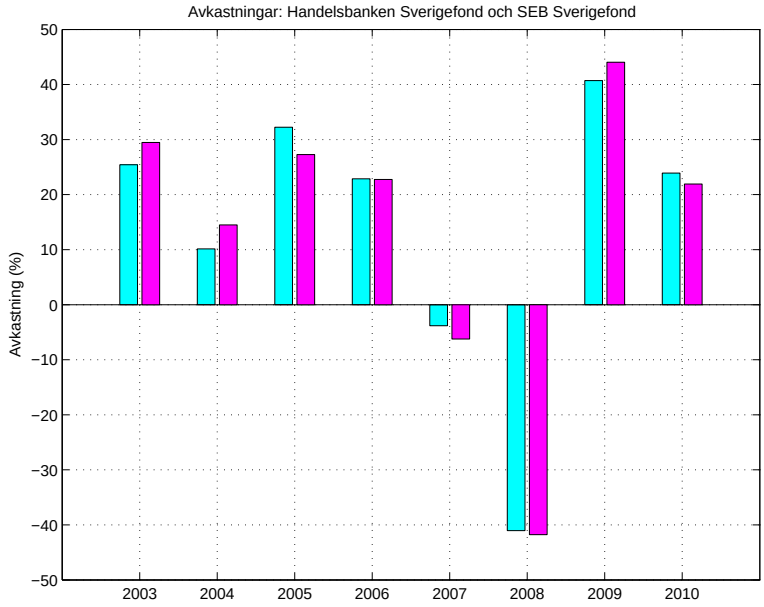
(b) Beräkna den betingade sannolikheten att komponent B fungerar, givet att systemet fungerar. (4 p)

(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. 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. (10 p)

Uppgift 5

I tidskriften *The American Journal of Public Health* kan vi läsa om en undersökning av antalet årligen rökta cigaretter per vuxen (x_i) och antalet dödsfall (y_i) per 100 000 invånare och år, i åldern 35–64 år, förorsakade av hjärtinfarkt. Tidskriften ger 21 par (x_i, y_i) från 21 olika länder i den industrialiserade västvärlden. (Sverige representeras av paret $(x_{19}, y_{19}) = (1270, 127)$.)

x_i	3900	3350	3220	3220	2790	2780	2770	2290	2160	1890	1810
y_i	265	211	238	212	194	160	187	208	208	150	125
x_i	1800	1770	1700	1680	1510	1500	1410	1270	1200	1090	
y_i	41	182	118	32	114	120	60	127	44	90	

Summor som kan vara användbara är

$$\sum_{i=1}^{21} x_i = 45110, \quad \sum_{i=1}^{21} y_i = 3086, \quad \sum_{i=1}^{21} (x_i - \overline{x})(y_i - \overline{y}) = 8.9896 \cdot 10^5$$

$$\sum_{i=1}^{21}(x_i-\overline{x})^2 = 1.3057\cdot 10^7, \quad \sum_{i=1}^{21}(y_i-\overline{y})^2 = 9.2611\cdot 10^4$$

Ansätt en enkel linjär regressionmodell där y beror av x och avgör om regressionslinjens lutning är signifikant skild från noll (på signifikansnivån 5%). Detta innebär att du undersöker om det finns ett statistiskt påvisat samband mellan rökning och hjärtinfarkt. (10 p)

Uppgift 6

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)

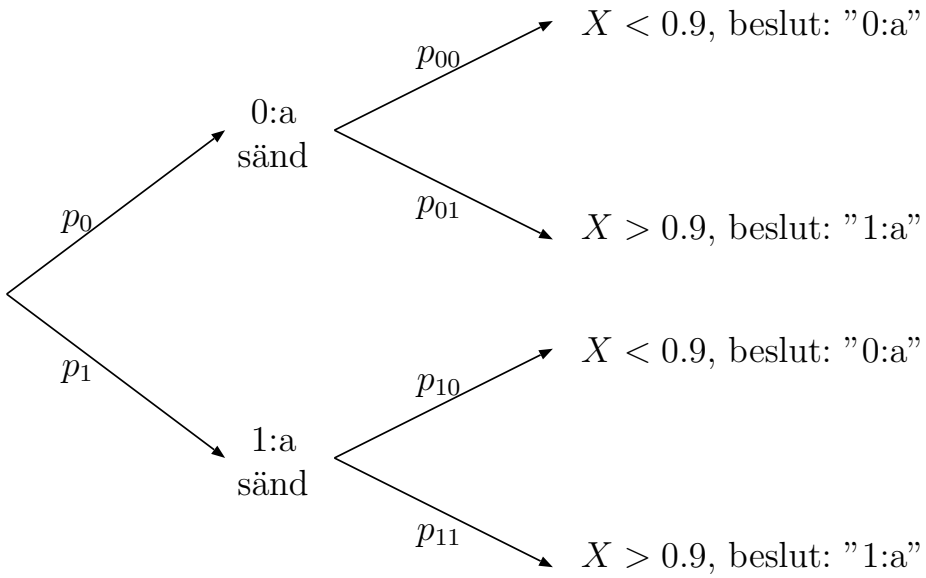
Avd. Matematisk statistik

KTH Matematik

LöSNINGAR TILL
TENTAMEN I SF1901, SF1905, SF1907 OCH SF1908 SANNOLIKHETSLÄRA OCH STATISTIK
ONSDAGEN DEN 12 JANUARI 2011 KL 14.00–19.00.

Uppgift 1

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

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

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

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

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

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

Alltså,

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

Uppgift 2

(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 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

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

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

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 samvaration (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.

Uppgift 5

Med $n = 21$ observationspar $(x_1, y_1), \ldots, (x_n, y_n)$ har vi

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 2148.1, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 146.95.$$

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2 = 109957100 - 21 \cdot 2148.1^2 = 13056524 \\ S_{xy} &= \sum_{i=1}^{n} x_i y_i - n \cdot \overline{x} \cdot \overline{y} = 7527980 - 21 \cdot 2148.1 \cdot 146.95 = 898958 \\ S_{yy} &= \sum_{i=1}^{n} y_i^2 - n \cdot \overline{y}^2 = 546106 - 21 \cdot 146.95^2 = 92610.9 \end{aligned}$$

Formelsamlingen ger att β skattas med

$$\beta^*_{\text{obs}} = \frac{S_{xy}}{S_{xx}} \approx 0.0689.$$

Det sökta konfidensintervallet för β med konfidensgrad 0.95 är

$$I_\beta : \beta^*_{\text{obs}} \pm t_{0.025}(n-2)\frac{s}{\sqrt{S_{xx}}}.$$

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

$$s^2 = \frac{1}{n-2}\left(S_{yy} - \beta^*_{\text{obs}} S_{xy}\right) = \frac{30716.5}{21-2} = 1616.67,$$

dvs $s = 40.21$. Här fås med $t_{0.025}(19) = 2.09$ att

$$I_\beta : 0.0689 \pm 2.09\frac{40.21}{\sqrt{13056524}}$$

dvs

$$I_\beta : 0.069 \pm 0.023 \quad \text{eller} \quad I_\beta = [0.046, 0.092].$$

Vi ser att $\beta = 0$ inte ligger i intervallet I_β. Således förkastas nollhypotesen $\beta = 0$ mot alternativet $\beta \neq 0$ på nivån 5%, dvs hypotesen om ett samband mellan rökning och hjärtinfarkt anses statistiskt påvisad på signifikansnivån 5%

Uppgift 6

(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.