Avd. Matematisk statistik

KTH Matematik

TENTAMEN I SF1913 MATEMATISK STATISTIK FÖR IT OCH ME
LÖRDAGEN DEN 11 FEBRUARI 2012 KL 14.00–19.00.

Examinator: Gunnar Englund, tel. 073 3213745

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

Ett elektronikföretag köper IC-kretsar från tre olika underleverantörer, A, B och C. Man köper dubbelt så många kretsar från B som från A, och tre gånger så många kretsar från C som från A. Man vet att i snitt är 1% av kretsarna som levereras av A defekta på något sätt; för leverantörerna B och C är motsvarande andelar 0.5% och 0.8%.

Alla kretsar läggs i ett enda förråd. Om en slumpmässigt vald krets i förrådet visar sig vara defekt, vad är sannolikheten att den kommer från leverantör A? (10 p)

Uppgift 2

Två defekta enheter har av misstag hamnat tillsammans med tre felfria enheter. För att finna de felfria testar man i tur och ordning en enhet i taget tills man antingen har funnit de båda defekta eller de tre felfria.

(a) Bestäm sannolikheten att båda de defekta enheterna behöver testas. (5 p)

(b) Bestäm det förväntade antalet enheter som behöver testas. (5 p)

Uppgift 3

Ett större företag vill undersöka om det finns intresse bland sina tjänstemän för att gå över till flextid. Ett slumpmässigt urval på 200 tjänstemän tillfrågas. Av de som valdes ut i stickprovet var 120 kvinnor, och av dessa var 90 positiva till flextid; bland de 80 männen i stickprovet var 50 positiva. Kan man hävda att inställningen till flextid skiljer sig åt mellan könen?

Svara på frågan med hjälp av ett lämpligt statistiskt test på nivån 1%. (10 p)

Uppgift 4

Lisa funderar på att installera solceller på taket på sitt hus. Hon kan välja mellan två typer, A och B. Solceller av den enkla typen A kostar 750 kr/styck och levererar under vissa förhållanden en effekt som beskrivs av en stokastisk variabel med väntevärde 150 W och standardavvikelse 60 W.

Den mer avancerade solcellen B kostar 2500 kr/styck men levererar en effekt som beskrivs av en stokastisk variabel med värde 550 W och standardavvikelse 210 W.

Om levererad effekt av skilda solceller beskrivs av oberoende stokastiska variabler, bestäm approximativt sannolikheten att 49 solceller av typ B ger en större total effekt *per krona* än 100 solceller av typ A. (10 p)

Uppgift 5

Hösten 2005 hade Lomma kommun problem med förhöjda halter av legionellabakterier i en del av sina lokaler (Pilängsbadet, Smultronställets förskola). Ett gränsvärde för accepterad halt av dessa bakterier är i genomsnitt 100 bakteriekolonier per 100 ml vatten. Detta är uppenbarligen detsamma som i genomsmitt 1 koloni per ml vatten.
Antag att vid provtagning en volym V (enhet ml) av vatten samlas in. En enkel statistisk modell är att antalet bakteriekolonier i denna är Poissonfördelat med väntevärdet γV, där γ är genomsnittlig koncentrationen av kolonier (i enheten kolonier per ml).

(a) Det första provet från Smultronställets förskola innehöll 620 kolonier per 100 ml vatten. Provets totala storlek var 400 ml. Undersök med lämpligt test eller konfidensintervall om legionellakoncentrationen i Smultronställets vatten under- eller översteg gränsvärdet ovan. Välj felrisk själv. (4 p)

(b) Antag att legionellakoncentrationen i ett vattenledningssystem är 1.1 kolonier per ml vatten. En volym V analyseras, och man utför ett test av nollhypotesen γ =1 koloni/ml mot alternativet γ >1 koloni/ml på nivån 0.001. Hur stor volym måste provtas för att sannolikheten att nollhypotesen förkastas (rätt beslut alltså i den aktuella situationen) skall vara minst 0.999? (6 p)

Uppgift 6

Tolv fyraåriga pojkar och tolv fyraåriga flickor observerades under två 15 minutersperioder och varje barns aggressionsnivå bedömdes enligt en poängskala:

Pojkar:	25	26	41	50	65	69	72	86	104	113	118	141
Flickor:	7	9	15	20	22	27	36	40	49	55	58	75

Data kan sammanfattas med (x för pojkar och y för flickor):

$$\sum_1^{12} x_i = 910,\ \sum_1^{12} y_i = 413,\ \ \sum_1^{12} x_i^2 = 84438,\ \sum_1^{12} y_i^2 = 19279.$$

a) Man antog att ovanstående data kom från två normalfördelningar med samma spridning. Analysera under dessa antaganden om det finns någon skillnad i aggressionsnivå mellan pojkar och flickor (nivå 5%). (5 p)

b) En konsulterad statistiker ansåg att antagandena i a-delen var alltför äventyrliga och föreslog att data skulle analyseras med ett icke-parametriskt test. Genomför denna analys. (5 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.

Avd. Matematisk statistik

KTH Matematik

LÖSNINGAR TILL
TENTAMEN I SF1913 MATEMATISK STATISTIK FÖR IT OCH ME
LÖRDAGEN DEN 11 FEBRUARI 2012 KL 14.00–19.00.

Inför beteckningarna A, B och C för händelserna att den slumpvisvalda kretsen kommer från leverantör A, B respektive C, och låt D beteckna händelsen att kretsen är defekt. Vi söker $P(A|D)$.

Vi har givet att $P(B) = 2P(A)$ och $P(C) = 3P(A)$, och då $P(A) + P(B) + P(C) = 1$ måste det gälla $P(A) = 1/6$, $P(B) = 1/3$ och $P(C) = 1/2$.

Vidare har vi givet att $P(D|A) = 0.01$, $P(D|B) = 0.005$ och $P(D|C) = 0.008$. Bayes sats (alternativt, definitionen av betingad sannolikhet och satsen om total sannolikhet) ger nu

$$\begin{aligned} P(A|D) &= \frac{P(D \cap A)}{P(D)} \\ &= \frac{P(D|A)P(A)}{P(D|A)P(A) + P(D|B)P(B) + P(D|C)P(C)} \\ &= \frac{0.01/6}{0.01/6 + 0.005/3 + 0.008/2} = \frac{10}{23} \approx 0.435. \end{aligned}$$

Uppgift 2

Uppgiften kan lösas på olika sätt; nedan redovisas en variant.

(a) Betrakta en urna med 2 kulor märkta D (defekta) och 3 kulor märkta F felfria. Händelsen att båda de defekta enheterna måste testas, kan ses som händelsen att om vi drar fyra kulor ur urnan så får vi två F bland dessa. I praktiken testar vi förstås inte fyra enheter om t ex de två första är defekta, men vi kan ändå tänka oss att vi alltid drar fyra kulor.

Sannolikheten för händelsen ovan är (hypergeometrisk fördelning)

$$\frac{\binom{2}{2}\binom{3}{2}}{\binom{5}{2+2}} = \frac{1 \cdot 3}{5} = 3/5.$$

Ett annat sätt att se på saken är att båda de defekta enheterna måste testas om och endast om den sista enheten som (potentiellt) väljs ut är felfri. Eftersom det finns tre felfria enheter av totalt fem, är sannolikheten för denna händelse 3/5.

(b) Låt X vara antalet enheter som behöver testas totalt. Vi har att X tar något av värdena 2, 3 eller 4. Händelsen $X = 2$ inträffar om de två första testade enheterna är defekta. I urnmodellen ovan kan vi identifiera det med att vi drar två kulor, och får två stycken F. Sannolikheten för detta är

$$\frac{\binom{2}{2}\binom{3}{0}}{\binom{5}{2+0}} = \frac{1 \cdot 1}{10} = 1/10.$$

Händelsen $X = 3$ inträffar om antingen de tre första testade enheterna är feldria, eller om de två första enheterna är en defekt och en felfri (i någon ordning) och den tredje enheten är defekt. Sannolikheten för detta är

$$\frac{\binom{2}{0}\binom{3}{3}}{\binom{5}{0+3}} + \frac{\binom{2}{1}\binom{3}{1}}{\binom{5}{1+1}} \cdot \frac{1}{3} = \frac{1 \cdot 1}{10}\frac{2 \cdot 3}{10} \cdot \frac{1}{3} = 3/10.$$

Här är det andra kombinatoriska uttrycket sannolikheten att få en defekt och en felfri bland de två första valda enheterna, och 1/3 är den betingade sannolikheten att den tredje testade enheten är defekt, givet att de två första var en av varje sort (en defekt kvar av totalt tre).

Slutligen inträffar händelsen $X = 4$ om bland de tre första testade enheterna en är defekt och två är felfria (i någon ordning; den fjärde enheten är då antingen defekt eller felfri, och vi är klara). Sannolikheten för detta är (urnmodellen igen)

$$\frac{\binom{2}{1}\binom{3}{2}}{\binom{5}{1+2}} = \frac{2 \cdot 3}{10} = 3/5.$$

Vi kan också använda $P(X = 2) + P(X = 3) + P(X = 4) = 1$ för att bestämma en sannolikhet när vi har två valfria andra.

Slutligen har vi

$$E(X) = \sum_k k \cdot P(X = k) = 2 \cdot \frac{1}{10} + 3 \cdot \frac{3}{10} + 4 \cdot \frac{3}{5} = \frac{7}{2}.$$

Uppgift 3

Vi använder ett χ^2-oberoendetest för att pröva nollhypotesen att inställningen till flextid är oberoende av kön. Eftersom vi måste ha svarsklasser som täcker alla möjligheter så inför vi klasserna $j = 1$ för "positiv till flextid" och $j = 2$ för "annat svar". Det senare kan betyda t ex negativt eller neutralt svar, men det spelar ingen roll för den här uppgiften.

Då är p_{ij} andelen av hela populationen av tjänstemän på företaget, av kön i ($i = 1$ för kvinna, $i = 2$ för man) som har inställning j ($j = 1$ eller 2 enligt ovan). Nollhypotesen H_0 är nu $p_{ij} = p_{i\cdot}p_{\cdot j}$ där $p_{i\cdot}$ är andelen av hela populationen av tjänstemän på företaget som har kön i, och $p_{\cdot j}$ är andelen av samma hela population som har inställning j enligt ovan. Mothypotesen H_1 är $p_{ij} \neq p_{i\cdot}p_{\cdot j}$ för något i och j.
Vi får tabellen

	$j = 1$	$j = 2$	n_i
$i = 1$	90	30	120
$i = 2$	50	30	80
S:a	140	60	200

Teststorheten blir

$$\begin{aligned} Q &= \frac{(90-120\cdot 140/200)^2}{120\cdot 140/200}+\frac{(30-120\cdot 60/200)^2}{120\cdot 60/120} \\ &+ \frac{(50-80\cdot 140/200)^2}{80\cdot 140/200}+\frac{(30-80\cdot 60/200)^2}{80\cdot 60/200} = 3.57 \end{aligned}$$

Under H_0 är detta en observation från en χ^2-fördelning med $(2-1)(2-1)=1$ frihetsgrad, och vi skall förkasta H_0 för stora värden. Då $\chi^2_{0.01}(1) = 6.63$ och $3.57 < 6.63$ finns det inte stöd (på signifikansnivån 1%) för slutsatsen att det finns ett beroende mellan kön och inställning till flextid.

Uppgift 4

(a) Vi ser att körtidena skiljer sig kraftigt åt mellan olika bilister; t ex har bilist 7 en lång körtid och bilist 10 en kort. Detta är naturligt då bilisterna kör från olika platser (hem) till olika arbetsplatser. Vi kan alltså inte anta att alla tider före omläggningen är observationer från en och samma fördelning, och motsvarande gäller efter omläggningen. Däremot kan vi anta att deras respektive tidsvinster kommer från en gemensam fördelning, eftersom tidsvinsterna bara beror på passagen av det område där omläggningen skett och de alla mätt tiderna samma två dagar. Vi har alltså situationen *stickprov i par*, eller *parvisa observationer*, och bildar tidsvinsterna $z_i = x_i - y_i$. Modellen är att z_i är oberoende observationer från en normalfördelning $N(\Delta, \sigma)$, där μ är den förväntade tidsvinsten. I läroboken antas också att x_i och y_i är normalfördelade, men det är helt överflödigt.

Vi har nu skattningarna $\Delta^*_{\text{obs}} = \overline{z} = 0.67$ och och $\sigma^*_{\text{obs}} = s_z = 2.29$ (minuter). Här är $\overline{z}$ ett utfall av $\overline{Z}$, som har fördelningen $N(\Delta, \sigma/\sqrt{n})$ med $n = 12$. Vidare har $T = (\overline{Z}-\mu)/(S/\sqrt{n})$ en t-fördelning med $n-1 = 11$ frihetsgrader, och genom att lösa ut μ i mitten av olikheten $-t_{0.025}(n-1) \le T \le t_{0.025}(n-1)$ (en händelse som har sannolikheten 0.95) så får vi ett konfidensintervall för μ som

$$\Delta \in \overline{z} \pm t_{0.025}(n-1)\frac{s_z}{\sqrt{n}} = 0.67 \pm 2.20\frac{2.29}{\sqrt{12}} = 0.67 \pm 1.45 = (-0.78,\, 2.12).$$

(b) Eftersom $\Delta = 3$ inte ingår i konfidensintervallet ovan kan vi *inte* förkasta hypotesen $\Delta = 3$ mot hypotesen $\Delta \neq 3$ på nivån 5%. Det finns alltså inget stöd för att hävda att målsättningen uppnåtts.

Uppgift 5

Låt $X_1, \ldots, X_n$ vara stokastiska variabler som beskriver effekterna levererade av n solceller av typ A. Förutsättningarna ger att dessa är oberoende och likafördelade. För den totala effekten $S_n = X_1 + \cdots + X_n$ gäller då

$$\begin{aligned} E(S_n) &= E\left(\sum_{k=1}^n X_k\right) = \sum_{k=1}^n E(X_k) = n\mu_{\text{A}}, \\ V(X_n) &= V\left(\sum_{k=1}^n X_k\right) = \sum_{k=1}^n V(X_k) = n\sigma^2_{\text{A}}, \\ D(S_n) &= \sqrt{V(S_n)} = \sqrt{n}\sigma_{\text{A}} \end{aligned}$$

(enhet: Watt). Enligt centrala gränsvärdessatsen gäller också att S_n är approximativt normalfördelad, $S_n \in N(n\mu_{\text{A}}, \sqrt{n}\sigma_{\text{A}})$ (approximativt). Effekten per krona, R_{A} säg, är $R_{\text{A}} = S_n/(750n)$,

och eftersom detta är en approximativt normalfördelad variabel delat med en konstant, är även $R_{\rm A}$ approximativt normalfördelad, dvs $R_{\rm A} \in N(\mu_A/750, \sigma_{\rm A}/(750\sqrt{n}))$ (approximativt). Med insatta värden får vi $N(0.2, 0.008)$ (enhet: W/kr).

På samma sätt beskrivs totala effekten/krona i W/kr för 49 solceller av typ B av en approximativt $N(0.22, 0.012)$-fördelad stokastisk variabel R_B.

Eftersom effekterna från olika solceller antas oberoende är $R_{\rm A}$ och $R_{\rm B}$ oberoende, och skillnaden $D = R_{\rm B} - R_{\rm A}$ är därför approximativt normalfördelad (eftersom en differens av två oberoende normalfördelade variabler är normalfördelad). Variabeln D har värdevärde $0.22 - 0.2 = 0.02$ och varians $0.08^2 + (-1)^2 \cdot 0.012 = 0.006544$. Vi får

$$\begin{aligned} P\left(W_B > W_A\right) &= P\left(D > 0\right) = P\left(\frac{D - 0.02)}{\sqrt{0.006544}} > \frac{0 - 0.02}{\sqrt{0.006544}}\right) \\ &\approx 1 - \Phi(-0.2472) = \Phi(0.2474) = 0.598. \end{aligned}$$

Uppgift 6

a) Två oberoende stickprov. Vi erhåller $\bar{x} = 9910/14 = 75.83$ och $\bar{y} = 413/12 = 34.42$. Vidare får vi

$$s_x = \sqrt{\frac{1}{12-1}(sum_1^{12} x_i^2 - 12(\bar{x})^2)} = 37.45$$

$$s_y = \sqrt{\frac{1}{12-1}(sum_1^{12} y_i^2 - 12(\bar{y})^2)} = 21.46$$

som ger

$$s^2 = \frac{(12-1)s_x^2 + (12-1)s_y^2}{12+12-2} = 933.66 \text{ dvs } s = 30.5.$$

Vi får med t-metoden ett 95%-igt konfidensintervall för skillnaden i väntevärden

$$\bar{x} - \bar{y} \pm t_{0.025}(22) s \sqrt{\frac{1}{12} + \frac{1}{12}} = 75.8 - 34.2 \pm 2.07 \cdot 30.5\sqrt{1/6} = 41.6 \pm 25.7.$$

Eftersom 0 inte ingår i konfidensintervallet kan vi på nivån 5% förkasta hypotesen att väntevärdena är lika.

b) Använd Wilcoxons tvåsampeltest.

Pojkar:	25	26	41	50	65	69	72	86	104	113	118	141
Ranger:	6	7	11	13	16	17	18	20	21	22	23	24
Flickor:	7	9	15	20	22	27	36	40	49	55	58	75
Ranger:	1	2	3	4	5	8	9	10	12	14	15	19

Vi vill testa H_0 : samma aggressionsnivå hos flickor och pojkar med ett dubbelsidigt test. Rangsumman för flickor blir $T_f = 1 + 2 + \cdots + 19 = 102$. Under H_0 att
$E(T_f) = n_1(n_1 + n_2 + 1)/2 = 150$
och $V(T_f) = n_1 n_2(n_1 + n_2 + 1)/12 = 12 \cdot 12 \cdot 25/12 = 300$.
Vi får signifikansnivån (p-värdet)

$$2P(T_f \le 102) = P(\frac{T_f - 150}{\sqrt{300}} \le \frac{102 - 150}{\sqrt{300}}) \approx 2\Phi(\frac{102 - 150}{\sqrt{300}}) \approx 2\Phi(-2.77) \approx 0.0056$$

och H_0 förkastas alltså på nivån 5%.