Avd. Matematisk statistik

KTH Matematik

TENTAMEN I SF1902 SANNOLIKHETSTEORI OCH STATISTIK
MÅNDAGEN DEN 14:E AUGUSTI 2017 KL 8.00–13.00.

Kursledare och examinator : Björn-Olof Skytt, tel 790 8649.

Tillåtna hjälpmedel: miniräknare, lathund till statistikfunktioner på Texas Instruments-räknare (TI-82 Stats och högre) utan egna tillägg, läroboken av Blom m.fl. utan egna tillägg, institutionens formelsamling utan egna tillägg, samt formelsamlingen BETA utan egna tillägg.

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.

Uppgift 1

En försäljare i den trendiga badbutiken Poolen och Plurret går in på lagret för att hämta en blå och en röd badbyxa. Inne på lagret är det kolmörkt eftersom strömmen gått, men hon hittar till korgen med badbyxorna. Hon vet att det endast är tre blå badbyxor och fyra röda badbyxor kvar i korgen och att de ligger huller om buller. Hon är stressad, så hon chansar och plockar två badbyxor ur korgen utan att se vad de har för färg. Vad är sannolikheten att hon fått med sig en badbyxa av varje färg?

Uppgift 2

En läkare har följande praxis gällande operation vid en viss sjukdom: Om hon är 80% säker på att en patient har sjukdomen rekommenderar hon operation. Om hon är mindre säker än så rekommenderar hon istället ytterligare test, vilka är dyra och kan vara smärtsamma för patienten. För en specifik patient är läkaren 60% säker på att patienten har sjukdomen - baserat på diverse test och tidigare liknande fall - och beställer därmed test A. Patienten i fråga är diabetiker och test A är sådant att det alltid ger ett positivt resultat om patienten har sjukdomen, men för diabetiker som *inte* har sjukdomen ger det ett positivt resultat i 30% av fallen. Antag nu att test A ger ett positivt resultat för patienten i fråga. Vad bör läkaren göra, ska hon rekommendera operation eller utföra ytterligare test? Motivera ditt svar.

Uppgift 3

Den trendiga badbutiken Poolen och Plurret vill i slutet av säsongen bli av med de 2 000 badbyxorna som är kvar av årets kollektion. Därför anordnar man en realisation på de kvarvarande badbyxorna. Av erfarenhet vet man att antalet badbyxor som en kund köper på sådana realisationer dels är oberoende av hur många badbyxor andra kunder köper, dels kan betraktas som en stokastisk variabel X med sannolikhetsfunktionen

$$p_X(k) = \begin{cases} 0.2 & \text{om} \quad k = 0, \\ 0.5 & \text{om} \quad k = 1, \\ 0.2 & \text{om} \quad k = 2, \\ 0.1 & \text{om} \quad k = 3, \\ 0 & \text{om} \quad k > 3. \end{cases}$$

Beräkna med lämplig och välmotiverad approximation sannolikheten att alla de 2000 badbyxorna blir sålda om det kommer 1700 kunder till butiken.

Uppgift 4

Ett flygbolag hävdar sig ha en av de bättre punktligheterna från en specifik större flygplats jämfört med konkurrerande flygbolag. “Punktlighet” mäts här i andelen flyg som avgår inom 20 minuter från utsatt tid. Bolaget påstår att det i genomsnitt endast är ett av 15 flyg är mer än 20 minuter försenat, vilket skulle bekräfta deras påstående om hur bolaget står sig gentemot sina konkurrenter. Flygbolaget har nyligen tecknat ett avtal med ett stort möbelföretag vilket innebär att flygbolaget får monopol på de flygresor som de anställda vid möbelföretaget gör i tjänsten i utbyte mot lägre biljettpriser. Ledningen för möbelföretaget tvivlar dock på flygbolagets påstående om sin punktlighet och ber sina anställda att under en månads tid notera om deras flyg från flygplatsen avgår i tid, det vill säga inom 20 minuter från utsatt tid, eller ej. Av de totalt 250 flygresor som möbelföretagets anställda gör under denna månad, avgår 20 av flygen senare än de tillåtna 20 minuterna.

Genomför ett statistiskt test som på (möjligen approximativa) nivån 5% prövar bolagets påstående om dess punktlighet. Var noga med att ange dina hypoteser och motivera dina slutsatser.

Uppgift 5

Den trendiga badbutikskedjan Poolen och Plurret säljer bl.a. badbyxor. Badbyxorna finns i fyra olika färger. För att kunna planera produktionen vill man undersöka om de fyra färgernas popularitet fördelar sig på ungefär samma sätt som året innan eller inte. Undersökningen går till på följande sätt. När innevarande års försäljning börjar så noterar man vilken färg var och en av de 2 000 först sålda badbyxorna har. Om de fyra färgerna fördelar sig på ungefär samma sätt i årets undersökning som de gjorde i förra årets så ställs produktionen inte om. Om däremot de fyra färgernas proportioner i föregående års undersökning skiljer sig tilläckligt mycket åt från hur deras proportioner är i innevarande års undersökning så ställs produktionen om. Tabellen nedan visar antalet sålda badbyxor av respektive färg i respektive års undersökning.

Färg	Blå	Svart	Grön	Röd
Föregående år	943	357	498	202
Innevarande år	860	347	476	317

Avgör utgående från ovanstående data om produktionen bör ställas om eller inte. Använd risknivån 5%. Var noga med att ange dina hypoteser och motivera dina slutsatser.

Uppgift 6

Antalet drunkningsolyckor i Sverige under sommarmånaderna juni-augusti 2015 var enligt Svenska Livräddningssällskapets statistik 75 stycken. Under motsvarande period 2016 inträffade 55 drunkningsolyckor enligt Svenska Livräddningssällskapets statistik.Kan man dra slutsatsen att minskningen mellan år 2015 och år 2016 vad gäller antalet drunkningsolyckor i Sverige under sommarmånaderna juni-augusti berodde på åtgärder man vidtagit, eller kan man inte utesluta att minskningen beror på slumpen? Undersök detta på signifikansnivån 5%. Ange tydligt vilka de uppställda hypoteserna är och vad slutsatsen är.

Uppgift 7

Man vill undersöka om kokta kräftor från företag A respektive företag B smakar lika gott, Man anlitade en panel bestående av 18 kräftätarexperter som fick smaka på kräftor från A respektive B. Varje kräftätarexpert fick sedan ange vilket företags kräftor som smakade bäst. Man fick följande 18 resultat: B B B A B A B B B B B A B B B A A B

Kan man på risknivån 10% avgöra om det finns en systematisk skilnad i smak mellan de två företagens kokta kräftor? Ange tydligt vilka de uppställda hypoteserna är och vad slutsatsen är.

Avd. Matematisk statistik

KTH Matematik

LÖSNINGSFÖRSLAG TENTAMEN I SF1902 MATEMATISK STATISTIK.
MÅNDAGEN DEN 14:E AUGUSI 2017 KL 8.00–13.00

Uppgift 1

Låt den stokastiska variabeln X vara antalet blå badbyxor försäljaren får med sig.
Då gäller att $X \in Hyp(N, n, p) = Hyp(7, 2, \frac{3}{7})$.
Vi söker $P(X = 1)$

$$P(X = 1) = \frac{\binom{3}{1}\binom{4}{1}}{\binom{7}{2}} = \frac{4}{7} = \underline{0.57}$$

Uppgift 2

Låt S beteckna händelsen att patienten har sjukdomen och A händelsen att testet visar positivt (det vill säga indikerar att patienten har sjukdomen). Det är givet att $P(S) = 0.6$ - läkarens uppfattning baserat på tidigare patienter och diverse test - samt att, på grund av patientens diabetes,

$$P(A|S) = 1, \quad P(A|S^*) = 0.3.$$

Vi söker sannolikheten att patienten har sjukdomen givet ett positivt testresultat, $P(S|A)$. Om denna sannolikhet överstiger 0.8 bör läkaren rekommendera operation, annars bör hon utföra ytterligare test. Med Bayes' sats fås

$$\begin{aligned} P(S|A) &= \frac{P(S \cap A)}{P(A)} \\ &= \frac{P(A|S)P(S)}{P(A|S)P(S) + P(A|S^*)P(S^*)} \\ &= \frac{1 \cdot 0.6}{1 \cdot 0.6 + 0.3 \cdot 0.4} \\ &= 0.833 \end{aligned}$$

Med ett positivt testresultat är läkaren alltså mer än 80% säker på att patienten har sjukdomen och bör därmed rekommendera operation.
Svar: Läkaren är nu mer än 80% säker och bör rekommendera operation.

Uppgift 3

Låt X_i vara den stokastiska variabeln antal köpta badbyxor för kund nr i. Vi får då

$$\mu = E[X_i] = \sum_{k=0}^{3} k \cdot p_{X_i}(k) = 0 \cdot 0.2 + 1 \cdot 0.5 + 2 \cdot 0.2 + 3 \cdot 0.1 = 1.2$$

och

$$\sigma = D(X_i) = \sqrt{V(X_i)}$$

där

$$V(X_i) = E(X_i^2) - E^2(X_i) = E(X_i^2) - 1.2^2$$

och

$$E(X_i^2) = \sum_{k=0}^{3} k^2 \cdot p_{X_i}(k) = 0^2 \cdot 0.2 + 1^2 \cdot 0.5 + 2^2 \cdot 0.2 + 3^2 \cdot 0.1 = 2.2$$

Detta medför att

$$V(X_i) = 2.2 - 1.2^2 = 0.76$$

Vilket ger att

$$\sigma = D(X_i) = \sqrt{0.76}$$

Låt nu

$$Y = \sum_{i=1}^{1700} X_i$$

där Y är det totala antalet sålda badbyxor.
Vi antar nu att X_i:na är många och oberoende. Eftersom de dessutom är likafördelade gäller då enligt Centrala Gränsvärdessatsen C.G.S. att

$$Y \sim N(1700\mu, \sigma\sqrt{1700}) = N(1700 \cdot 1.2, \sqrt{0.76} \cdot \sqrt{1700}) = N(2040, \sqrt{0.76} \cdot \sqrt{1700})$$

$$P(Y \geq 2000) = P(\frac{Y - 2040}{\sqrt{0.76 \cdot 1700}} \geq \frac{2000 - 2040}{\sqrt{0.76 \cdot 1700}}) = [Z = \frac{Y - 2040}{\sqrt{0.76 \cdot 1700}} \sim N(0,1)] =$$

$$= P(Z \geq \frac{2000 - 2040}{\sqrt{0.76 \cdot 1700}}) =$$

$$= 1 - P(Z \leq \frac{2000 - 2040}{\sqrt{0.76 \cdot 1700}}) = 1 - \Phi(\frac{-40}{\sqrt{0.76 \cdot 1700}}) = 1 - [1 - \Phi(\frac{40}{\sqrt{0.76 \cdot 1700}})] \approx \Phi(1.11) \approx 0.8665$$

Svar: Sannolikheten att alla badbyxor blir sålda är ungefär 86.65%.

Uppgift 4

Vi antar att varje flygresa har samma sannolikhet p att vara försenad. Då gäller att antalet undersökta flygresor X som blir försenade $\in Bin(n,p) = Bin(250,p)$ i vårt fall. Vi ska nu testa nollhypotesen $H_0 : p = \frac{1}{15}$ mot mothypotesen $H_1 : p \neq \frac{1}{15}$ på (aproximativa) risknivån 5%.
En punktskattning av sannolikheten p ges nu av

$$p^*_{obs} = \frac{x}{n} = \frac{20}{250} = 0.08.$$

Vi vill nu ta fram ett approximativt konfidensintervall för p, men då krävs att $n \cdot p^*_{obs} \cdot (1 - p^*_{obs}) \geq 10$. Då $p^*_{obs} = 0.08$ och $n = 250$ gäller att

$$np^*_{obs}(1 - p^*_{obs}) = 250 \cdot \frac{20}{250} \cdot \frac{230}{250} = 18.4 > 10,$$

varför normalapproximationen till binomialfördelningen är tillåten.

Vi bildar nu ett konfidensintervall för p enligt läroboken:

$$I_p = p^*_{obs} \pm \sqrt{\frac{p^*_{obs}(1 - p^*_{obs})}{n}} \cdot \lambda_{\frac{\alpha}{2}}$$

$$I_p = 0.08 \pm \sqrt{\frac{0.08 \cdot 0.92}{250}} \cdot \lambda_{0.025}$$

$$I_p = 0.08 \pm \sqrt{\frac{0.08 \cdot 0.92}{250}} \cdot 1.96$$

$$I_p = 0.08 \pm 0.0336$$

$$I_p = (0.0464, 0.1136)$$

Vi har följande nollhypotes: $H_0 : p = \frac{1}{15} = 0.067$

$$0.067 \in I_p = (0.0464, 0.1136)$$

Eftersom intervallet innehåller $p = 0.067$ förkastar vi ej H_0 på den approximativa nivån 5%.

Svar: <u>Vi kan **ej** förkasta flygbolagets påstående på nivån 5%.</u>

Uppgift 5

Vi gör här ett homogenitetstest (avsnitt 14.3 i formelsamlingen) eftersom vi ska undersöka om sannolikheterna för färgerna är desamma i de båda försöksserierna.Nollhypotesen H_0 är då att fördelningen av färgerna är oförändrad mellan de båda undersökningstillfällena. Mothypotesen H_1 är då att det skett en sådan förändring.

Vi gör här en tabell med observerade antal enligt

Observerade antal	Blå	Svart	Grön	Röd	Totalt
Föregående år	943	357	498	202	2000
Innevarande år	860	347	476	317	2000
Totalt	1803	704	974	519	4000

Teststorheten blir

$$Q = \frac{(943 - \frac{2000\cdot 1803}{4000})^2}{\frac{2000\cdot 1803}{4000}} + \frac{(357 - \frac{2000\cdot 704}{4000})^2}{\frac{2000\cdot 704}{4000}} + \frac{(498 - \frac{2000\cdot 974}{4000})^2}{\frac{2000\cdot 974}{4000}} +$$

$$+ \frac{(202 - \frac{2000\cdot 519}{4000})^2}{\frac{2000\cdot 519}{4000}} + \frac{(860 - \frac{2000\cdot 1803}{4000})^2}{\frac{2000\cdot 1803}{4000}} + \frac{(347 - \frac{2000\cdot 704}{4000})^2}{\frac{2000\cdot 704}{4000}} +$$

$$+ \frac{(476 - \frac{2000\cdot 974}{4000})^2}{\frac{2000\cdot 974}{4000}} + \frac{(317 - \frac{2000\cdot 519}{4000})^2}{\frac{2000\cdot 519}{4000}} = 29.94$$

Om H_0 är sann så är 29.94 ett utfall från en stokastisk variabel som approximativt har en χ^2-fördelning med $(4-1)(2-1) = 3$ frihetsgrader. Eftersom $\chi^2_{0.05}(3) = 7.81 < 29.94$ så kan H_0 förkastas på nivån 5%. Alternativt kan vi beräkna sannolikheten att en $\chi^2(2)$-variabel är större än eller lika med 29.94 (`X2cdf` på en TI-räknare). Denna sannolikhet, dvs p-värdet för testet, är$1.4 \cdot 10^{-6}$. Detta p-värde är så lågt att vi förkastar H_0 på risknivån 5%. Både teststorheten och p-värdet fås direkt med funktionen `X2-Test` på en TI-räknare.

Svar: Vi drar slutsatsen att förändringen i kundernas färgval är statistiskt säkerställd på risknivån 5% och att produktionen därför skall ställas om.

Uppgift 6

Här antar vi att antalet drunkningsolyckor under sommarmånaderna är Poissonfördelat. Vi antar att X_1 är antalet drunkningsolyckor maj-aug 2015 och att X_2 är antalet drunkningsolyckor maj-aug 2016 , och att $X_1 \in Po(\mu_1)$ och att $X_2 \in Po(\mu_2)$. Nollhypotesen blir i detta fall att det inte är någon signifikant skillnad i antalet drunkningsolyckor mellan de båda årens sommarmånader.D.v.s. att skillnaden beror av slumpen.
Således har vi $H_0 : \mu_1 = \mu_2$. $H_1 : \mu_1 \neq \mu_2$.
Här använder vi χ^2-test:

$$Q = \sum_{i=1}^{2} \frac{(x_i - \mu^*_{obs})^2}{\mu^*_{obs}}$$

Eftersom $E(X_i) = \mu$ när $X_i \in Po(\mu_i)$ så skattar vi μ med

$$\bar{x} = \frac{x_1 + x_2}{2} = \frac{75 + 55}{2} = 65$$

Detta ger

$$Q = \frac{(75 - 65)^2}{65} + \frac{(55 - 65)^2}{65} \approx 3.08$$

$$\chi^2_\alpha(r - 1) = \chi^2_{0.05}(2 - 1) = 3.84$$

Eftersom $Q < \chi^2_{0.05}(1)$ så förkastar vi inte H_0 på risknivån 5%.
Svar:Vi drar slutsatsen att det på risknivån 5% inte skett någon signifikant förändring i antalet drunkningsolyckor under sommarmånaderna mellan 2015 och 2016 och att vi därför inte kan utesluta att skillnaden beror på slumpen.

Uppgift 7

Här har vi följande hypoteser:
H_0 : Kräftorna från de båda företagen smakar lika bra.
H_1 : Kräftorna från de båda företagen smakar inte lika bra.

Vi använder oss av teckentestet. Vi inför den stokastiska variabeln Z som är antal kräftexperter som tycker att kräftorna från företag A smakar bäst. S. Om H_0 gäller ska $Z \in Bin(n, 0.5) = Bin(18, 0.5)$. Vi får alltså att

p-värdet $= 2 \cdot P(Z \geq 13) = 2 \cdot P(Z \leq 5) = [\text{se tab } 6; x = 5, n = 18, p = 0.5] = 2 \cdot 0.04813 = 0.09626$

D.v.s. p-värdet är 9.6%. Detta innebär att vi kan förkasta H_0 på risknivån 10%.

Svar:Vi drar slutsatsen att kräftorna från de båda företagen inte smakar lika bra (och att alltså kräftorna från företag B är de som smakar bäst).