Matematisk statistik LKT325
Tentamen 2018-04-06 med lösningar

Tid: 8.30-12.30. **Tentamensplats:** Lindholmen
Hjälpmedel: Kursboken **Matematisk Statistik** av Ulla Dahlbom. Formelsamlingen **Tabell- och formelsamling i matematisk statistik, försöksplanering och kvalitetsstyrning** av Håkan Blomqvist. Boken och formelsamlingen får ej innehålla extra anteckningar, men understrykningar, sticks och markeringar är tillåtna. **Chalmersgodkänd räknare.**

Till varje uppgift skall fullständig lösning lämnas!
Betygsgränser: För betyg 3, 4 resp. 5 krävs minst 20, 30 resp. 40 poäng.

---

1. (2+4 poäng) Antag att mätvärdena 10.5, 11.8 och 9.6 kommer från en normalfördelning med väntevärde μ och standardavvikelse σ. Antag också att mätningarna är gjorda oberoende av varandra.

   (a) Beräkna ett 95% konfidensintervall för μ om $\sigma = 1$.

   (b) Beräkna ett 95% konfidensintervall för μ om σ okänt.

   Lösning:

   (a) $\bar{x} = 10.6333$. Intervallet blir

   $$10.6333 \pm 1.96 \times 1/\sqrt{3} = \boxed{10.6333 \pm 1.1316} = \boxed{[9.5017, 11.7649].}$$

   (b) $s = 1.10604$. t-tabellen, rad 2, kolumn 0.05, ger värdet 4.30. Intervallet blir

   $$10.6333 \pm 4.3 \times 1.10604/\sqrt{3} = \boxed{10.6333 \pm 2.7459} = \boxed{[7.8874, 13.3792]}$$

2. (3+3+4 poäng) Anna har fört statistik över hur långt hon går på en dag. Hon har kommit fram till att sträckan hon går på en dag kan betraktas som en kontinuerlig stokastisk variabel ξ med frekvensfunktion

   $$f(x) = \begin{cases} \frac{1}{2500}(30x^2 - 3x^3) & \text{för } 0 \leq x \leq 10 \\ 0 & \text{för övrigt} \end{cases}$$

   Enheten är kilometer.

   (a) Beräkna väntevärde och standardavvikelse för ξ.

   (b) Beräkna approximativt sannolikheten att hon går mindre än eller lika med 620 kilometer under en tidsperiod på 100 dagar. Vi antar att sträckorna hon går under olika dagar kan betraktas som oberoende.

   (c) Antag igen att vi tittar på 100 dagar. Låt η vara lika med antalet dagar av de 100 dagarna på vilka hon går längre än 5 kilometer. Beräkna approximativt $P(\eta > 65)$.

   Lösning:

(a)

$$E(\xi) = \int_0^{10} xf(x)dx = ... = \boxed{6}.$$

$$\sigma = \sqrt{\int_0^{10} x^2 f(x)dx - (E(\xi))^2} = ... = \sqrt{40 - 36} = \boxed{2}.$$

(b) Sätt ξ_i = sträckan hon går dag nummer i, $i = 1, ..., 100$. Totala sträckan hon går under 100 dagar betecknas med T. Då är $T = \sum_{i=1}^{100} \xi_i$. Enligt centrala gränsvärdessatsen gäller att T är approximativt $N(6 \times 100, 2 \times \sqrt{100}) = N(600, 20)$-fördelad, så $(T - 600)/20$ är approx $N(0, 1)$-fördelad. Så

$$P(T \leq 620) = P(\frac{T - 600}{20} \leq \frac{620 - 600}{20}) \approx \Phi(1) = \boxed{0.8413}.$$

(c)

$$P(\xi > 5) = \int_5^{10} f(x) = ... = \frac{11}{16}.$$

η är Bin($n = 100$, $p = 11/16$)-fördelad. Eftersom $np(1 - p) = 21.5 > 10$ är η approx $N(np, \sqrt{np(1-p)}) = N(68.75, 4.63512)$. Så

$$P(\eta > 65) = 1 - P(\eta \leq 65) = 1 - P(\frac{\eta - 68.75}{4.65312} \leq \frac{65 - 68.75}{4.63512})$$
$$\approx 1 - \Phi(-0.809) = 1 - (1 - \Phi(0.809)) = \Phi(0.809) = \boxed{0.791.}$$

3. (6 poäng) Antag att det i urna A finns 3 röda och 4 gröna bollar, och att det i urna B finns 2 röda och 5 gröna bollar. Först väljer man en boll slumpmässigt från urna A och lägger den i urna B. Sedan väljer man en boll slumpmässigt från urna B (som nu innehåller 8 bollar), och lägger den i urna A. I det sista och tredje steget drar vi slumpässigt en boll från urna A. Vad är sannolikheten att bollen vi drar i det sista steget är grön? Vi antar hela tiden att bollarna i urnorna är väl blandade.

Lösning: Det finns 4 fall som gör att vi drar en grön i sista och tredje steget. Vi betecknar fallen med A_1, A_2, A_3 och A_4. Fallet A_1 är följande sekvens av händelser: Först flyttas en grön, sedan en grön, och en grön dras i tredje steget. A_2: grön - röd-grön. A_3: röd - grön - grön. A_4: röd - röd -grön. Sannolikheten för A_1 kan beräknas enligt följande:

$$P(A_1) = \frac{4}{7} \times \frac{6}{8} \times \frac{4}{7} = \frac{96}{392}.$$

Den första faktorn kommer sig av att vi har 4 gröna bollar av 7 bollar när vi väljer i första steget. Den andra faktorn kommer från att vi har 6 gröna av 8 i andra steget, ifall det flyttades en grön i första. Det tredje faktorn kommer från att vi har 4 gröna bollar av 7 i tredje steget ifall det flyttades gröna bollar i de två första stegen.

På liknande sätt fås att

$$P(A_2) = \frac{4}{7} \times \frac{2}{8} \times \frac{3}{7} = \frac{24}{392},$$

$$P(A_3) = \frac{3}{7} \times \frac{5}{8} \times \frac{5}{7} = \frac{75}{392},$$
$$P(A_4) = \frac{3}{7} \times \frac{3}{8} \times \frac{4}{7} = \frac{36}{392}.$$

Eftersom A_1, A_2, A_3, A_4 samtliga är disjunkta, fås

$$P(\text{grön dras i tredje steget}) = P(A_1)+P(A_2)+P(A_3)+P(A_4) = \frac{231}{392} = \boxed{\frac{33}{56}} \approx \boxed{0.5893.}$$

4. (4+1 poäng) Tre ingenjörsstudenter har tillsammans fått sommarjobb på en biltidning. Som en del i ett reportage får de i uppdrag att testa om tre olika biltillverkare producerar bilar med liknande bensinförbrukning. Det ska bara testa en viss typ av bilar så alla testade bilar har liknande prestanda (vikt, motorkapacitet osv.). Var för sig testar studenterna varsitt bilmärke, fyra bilar från varje märke. De kör med varje bil två timmar i stadstrafik under liknande förhållanden. Här är bensinförbrukningen (i liter per mil) de uppmäter.

Bilmärke 1	Bilmärke 2	Bilmärke 3
0.541	0.602	0.635
0.676	0.688	0.592
0.589	0.662	0.592
0.561	0.724	0.561

När de samlat in all data börjar de fylla i en ANOVA-tabell för testet.

Variationskälla	SS	df	MS	F
Mellan bilmärken	0.01527	...	...	...
Inom bilmärken	...	...	...	
Totalt	0.03659	...		

(a) Hjälp studenterna att komplettera ANOVA-tabellen. Utför ett hypotestest, tolka resultatet och dra slutsatser. Är det någon skillnad på den genomsnittliga bensinförbrukningen hos de olika bilmärkena?

(b) Under testandet kör de tre studenterna bilar från varsitt bilmärke. Dvs observationerna från bilmärke 1 kommer från en student, observationerna från bilmärke 2 från en annan student, osv. Är detta ett problem? Motivera!

Lösning:

(a) Den fullständiga ANOVA-tabellen bör se ut så här:

Variationskälla	SS	df	MS	F
Mellan bilmärken	0.01527	2	0.00764	3.22
Inom bilmärken	0.02132	9	0.00237	
Totalt	0.03659	11		

Vi använder de fem stegen.

Steg 1. Låt μ_1, μ_2 och μ_3 beteckna den genomsnittliga bilförbrukningen hos denna biltyp från de olika biltillverkarna. Våra hypoteser blir $H_0 : \mu_1 = \mu_2 = \mu_3$ och $H_1 :$ alla μ_i är inte lika.

Steg 2. Vi väljer signifikansnivå $\alpha = 0.05$. (Man är fri att välja signifikansnivå men det är bara 0.05 som finns i tabellen så om man väljer någon annan nivå får man problem.)

Steg 3. Testvariabeln vi vill använda är

$$F = \frac{MSB}{MSE} = \frac{(SSB)/(k-1)}{(SSE)/(n-k)}.$$

I vårt fall är $k - 1 = 3 - 1 = 2$ och $n - k = 12 - 3 = 9$ så vår testvariabel är F-fördelad med 2 och 9 frihetsgrader.

Steg 4. Vi vill ta fram vårt observerade värde på F. För att göra det så fyller vi i ANOVA-tabellen succesivt. Vi vet att $SSE = SST - SSB$ så vi vet att $SSE = 0.03659 - 0.01527 = 0.02132$. Eftersom det är tre grupper och fyra observationer i varje grupp blir frihetsgraderna 2 och 9. Vi vet också att $MSB = SSB/df = 0.00764$ och att $MSE = SSE/df = 0.00237$. Tillslut får vi att $F = MSB/MSE = 3.22$.

Steg 5. Vi tittar i F-fördelningstabellen med 2 och 9 frihetsgrader och ser att det kritiska värdet är 4.26. Eftersom vårt observerade värde är mindre än det kritiska värdet så förkastar vi inte H_0 på signifikansnivå $\alpha = 0.05$.

Slutsats: Vi kan inte utestluta att de olika biltillverkarn producerar bilar av denna typ med samma bensinförbrukning.

(b) Det är ett potentiellt problem att de testar bilar från varsin biltillverkare. Det skulle kunna vara så att de olika studenternas körsätt påverkar bensinförbrukningen. Om det är stor skillnad i hur bensinsnål körteknik de har så påverkas oberoendet i observationerna och därmed påverkas även den statistiska analysen. I värsta fall är det större variation mellan studenternas körteknik än mellan bilarnas bensinförbrukning. I så fall säger vårt test mer om skillnaden mellan studenternas körteknik än om skillnaden i bensinförbrukningen.

5. (4+1 poäng) Ett renhållningsbolag hämtar sopor i ett lägenhetskomplex en gång i veckan. Bostadsföreningen och renhållningbolaget har en överenskommelse om att det ska i genomsnitt hämtas max 1100 kg sopor varje vecka. Renhållningsarbetarna misstänker dock att de faktiskt hämtar mer. Om så är fallet vill renhållningsbolaget ta mer betalt av bostadsföreningen. För att undersöka om det hämtas för mycket sopor så väger renhållningsarbetarna soporna som hämtas varje vecka under ett helt år (52 veckor på ett år). Stickprovsmedelvärdet blev $\bar{x} = 1146$ och stickprovsstandardavvikelsen blev $s = 106$.

(a) Hjälp renhållningsarbetarna att utföra ett hypotestest. Kan vi dra slutsatsen att de hämtar för mycket sopor i genomsnitt?

(b) Beräkna p-värdet för testet.

Lösning:

(a) Låt μ beteckna den genomsnittliga mängd sopor som hämtas från lägenhetskomplexet varje vecka.

Steg 1 Eftersom vi bara är intresserade av huruvida mängden sopor överstigs eller inte så gör vi ett ensidigt test. Hypoteserna blir alltså: $H_0 : \mu \leq 1100$ och $H_1 : \mu > 1100$.

Steg 2 Vi bestämmer signifikansnivå: $\alpha = 0.05$. Man är fri att välja signifikansnivå själv här.

Steg 3 Stickprovet är stort nog för att vi kan använda normalapproximationen, vi använder oss därför av testvariabeln

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

som är approximativt $N(0, 1)$-fördelad.

Steg 4 I vårt fall är $\bar{x} = 1146$, $\mu_0 = 1100$, $s = 106$ och $n = 52$. Vår observerade teststatistika blir då 3.13.

Steg 5 Det kritiska värdet tas fram genom att titta i normalfördelningstabellen: 1.645. Eftersom vår observerade teststatistika är större än det kritiska värdet så förkastar vi H_0 på signifkansnivå $\alpha = 0.05$.

Slutsats: Vi kan dra slutsatsen att det i genomsnitt hämtas mer än 1100 kg sopor från lägenhetskomplexet. Renhållningsbolaget borde ta mer betalt från bostadsföreningen!

(b) Den observerade teststatistikan var 3.13. Från normalfördelningstabellen får vi då fram p-värdet: 1-0.9991 = 0.0009.

6. (2+2+2 poäng) Antag att ξ är en diskret stokastisk variabel. Det gäller att $P(\xi = -1) = 0.3$, $P(\xi = 1) = 0.2$, $P(\xi = 0) = 0.5$ och $P(\xi = x) = 0$ för övriga x. Vi definierar den diskreta stokastiska variabeln η genom sambandet $\eta = 1 + \xi^2$.

(a) Beräkna väntevärdet för η.

(b) Beräkna $P(\eta = x)$ för alla heltal x. (Det vill säga, bestäm sannolikhetsfunktionen för η.)

(c) Beräkna den betingade sannolikheten $P(\xi = 1 \,|\, \eta = 2)$.

Lösningar:

(a)

$$E(\xi^2) = (-1)^2 \times 0.3 + 0^2 \times 0.5 + 1^2 \times 0.2 = 0.5.$$

$$E(\eta) = 1 + E(\xi^2) = \boxed{1.5}.$$

(b) η kan anta värdena 1 och 2. Det gäller att $P(\eta = 1) = P(\xi = 0) = \boxed{0.5}$ och $P(\eta = 2) = P(\xi = 1) + P(\xi = -1) = \boxed{0.5}$ samt $P(\eta = x) = \boxed{0}$ för alla övriga x.

(c)

$$P(\xi = 1 \,|\, \eta = 2) = \frac{P(\{\xi = 1\} \cap \{\eta = 2\})}{P(\eta = 2)} = \frac{P(\xi = 1)}{P(\eta = 2)} = \frac{0.2}{0.5} = \boxed{\frac{2}{5}}.$$

7. (6 poäng) Följande gäller för en viss typ av datachips. De flesta datachipsen genomgår en kvalitetskontroll innan de går till försäljning, och repareras om de är defekta. Ibland misslyckas reparationen. Dessutom finns det datachips som inte kontrolleras alls.

   Antag att sannolikheten att ett datachips genomgår kvalitetskontroll är 0.99. Sannolikheten att ett datachips som genomgått kvalitetskontroll är defekt är 0.02. Sannolikheten att ett datachips som inte genomgått kvalitetskontroll är defekt är 0.12.

   Antag att du köper ett datachips, och ser att det är defekt. Beräkna den betingade sannolikheten att datachipset genomgått kvalitetskontroll givet detta.

   Lösning: Sätt $D = \{\text{chips defekt}\}$ och $K = \{\text{chips kontrollerat}\}$. Vi vet att $P(K) = 0.99$, $P(D|K) = 0.02$ och $P(D|K^c) = 0.12$. Vi får att

   $$P(D) = P(D|K)P(K)+P(D|K^c)P(K^c) = 0.02\times 0.99+0.12\times 0.01 = 0.021.$$

   Enligt Bayes sats blir

   $$P(K|D) = \frac{P(D|K)P(K)}{P(D)} = \frac{0.02 \times 0.99}{0.021} = \boxed{0.9429}.$$

8. (2+4 poäng) Man genomförde ett fullständigt faktorförsök för att undersöka hur de 3 faktorerna A, B och C påverkade en speciell situation. Man fick följande resultat från de åtta försöken.:

Nr.	A	B	C	Resultat y
1	-	-	-	41
2	+	-	-	51
3	-	+	-	42
4	+	+	-	51
5	-	-	+	38
6	+	-	+	55
7	-	+	+	40
8	+	+	+	51

   (a) Beräkna huvudeffekten l_A, två-faktorsamspelet l_{AB}.

   (b) Antag att man också var intresserad av faktorerna D och E. Man har bara råd att göra 8 försök, så man får göra ett reducerat faktorförsök. Antag att man väljer teckenkolumner för A, B och C precis som ovan. Det finns sedan olika sätt att välja teckenkolumner för D och E. Ingenjörerna Benedikt och Beata har fått i uppdrag att lägga upp det reducerade faktorförsöket med kravet att A inte sammanblandas med BC. Benedikt väljer generatorerna $D = AB$ och $E = AC$, medan Beata väljer generatorerna $D = ABC$ och $E = AB$. Kommer Benedikts reducerade faktorförsök att uppfylla villkoret? Kommer Beatas reducerade faktorförsök att uppfylla villkoret? Beräkna också upplösningarna för de bägge planerna.

   Lösningar:

   (a) $\boxed{l_A = 11.75}$, $\boxed{l_{AB} = -1.75}$.

(b) Benedikt: $I_1 = ABD$, $I_2 = ACE$, $I_3 = ABDACE = BCDE$. Alias till A blir BD, CE och $ABCDE$. Uppfyller således kravet, och upplösningen blir III.

Beata: $I_1 = ABCD$, $I_2 = ABE$, $I_3 = ABCDABE = CDE$. Alias till A blir BCD, BE och $ACDE$. Uppfyller således kravet, och upplösningen blir III.

Lycka till!