Lösningsförslag till Matematisk statistik LKT325 Tentamen 20190115

Kursansvarig: Reimond Emanuelsson
Betygsgränser: för betyg 3 krävs minst 20 poäng, för betyg 4 krävs minst 30 poäng, för betyg 5 krävs minst 40 poäng.

1. (2+2+3 poäng) För händelserna A och B gäller att

$$P(A) = 0.5, \quad P(B) = 0.4, \quad P(A \cap B) = 0.3.$$

Beräkna...

(a)

$$P(A|B^c) = \frac{P(A) - P(A \cap B)}{1 - P(B)} = \frac{0.5 - 0.3}{1 - 0.4} = 0.33\ldots$$

(b)

$$P(A^c \cap B^c) = P((A \cup B)^c) = 1 - P(A \cup B) = 1 - [P(A) + P(B) - P(A \cap B)] = 0.4$$

(c)

$$\begin{aligned} P((A \cup B) \cap (A \cap B)^c) &= P((A \cup B) \setminus (A \cap B)) = \{\text{även}\} = \\ &= P(A \setminus B) + P(B \setminus A) = P(A) + P(B) - 2P(A \cap B) = 0.3. \end{aligned}$$

2. (1+3 poäng) I de lilla municipalsamhället Stuvesberg är telefonnumren femsiffriga, från 20 000 till 99 999.

(a) Antal telefonnummer, som är möjliga (finns) är 80 000.

(b) Man behåller alla telefonnummer som är möjliga i a) men tar bort de som börjar med siffran 7, d.v.s. $80\,000 - 10\,10\,000 = 70\,000$. Därefter inför man alla sexsiffriga telefonnummer, som börjar med 7, d.v.s. man får 100 000 nya telefonnummer.

Antal möjliga telefonnummer finns det totalt efter dessa två ändringar är alltså 170 000.

3. (2+2+3 poäng) Tiden för Börje att lägga ett golv i ett rum (på 60m^2), är normalfördelad med väntevärde 1.5 och standardavvikelse 0.40, dvs $\xi_1 \in \text{N}(1.5, 0.40)$ (enhet timmar). Efter att golvet är lagt, tapetseras rummet av Lasse och tar en tid, som är $\xi_2 \in \text{N}(2.5, 0.41)$ (enhet timmar).

(a) Sannolikheten golvläggningen tar mindre än 2.0 h är

$$\Phi\left(\frac{2.0 - 1.5}{0.40}\right) = 0.89.$$

(b) För sannolikheten att den totala tiden (golvläggning+tapetsering) har vi att

$$\xi_1 + \xi_2 \in \text{N}(1.5 + 2.5, \sqrt{0.40^2 + 0.41^2}) = \text{N}(1.5 + 2.5, 0.5728).$$

Sannolikheten att den totala tiden ligger mellan 3.5 och 4.5 h är

$$\Phi\left(\frac{4.5 - 4.0}{0.5728}\right) - \Phi\left(\frac{3.5 - 4.0}{0.5728}\right) = 2[\Phi\left(\frac{4.5 - 4.0}{0.5728}\right) - 1 = 0.62.$$

(c) Antag istället att de börjar jobba samtidigt, och att förutsättningarna i övrigt är som ovan. Sannolikheten att Börje blir klar först:

$$\xi_1 - \xi_2 \in \text{N}(1.0, 0.5728), \; P(\xi_1 \geq \xi_2) = 1 - P(\xi_1 - \xi_2 \leq 0) = \Phi\left(\frac{1.0}{0.5728}\right) = 1 - 0.04 = 0.96.$$

4. (2+3 poäng) Talen $(25.0, 26.0, 24.5, 22.0, 25)$ är ett observerat stickprov på en normalfördelning med väntevärde μ och standardavvikelse σ. Räknehjälp: Från stickprovet beräknas medelvärde $\overline{x} = 24.5$ och standardavvikelse $s = 1.5$.

   (a) Ett tvåsidigt 95%:s konfidensintervall för väntevärdet μ. Stickprovets storlek är $n = 5$, så vi behöver kvantilen $t_{4,0.025} = 2.78$. Intervallet skrivet som en olikhet och intervall är

   $$24.5 \pm \frac{t_{5,0.025} \cdot s}{\sqrt{5}} = 24.5 \pm 1.86 \text{ resp. } [22.6, 26.4]$$

   (b) Ett uppåt begränsat 95%:s konfidensintervall för standardavvikelsen σ: Vi behöver kvantilen $\chi^2_{0.95}(4) = 0.71$ så att intervallet för σ^2 och σ är

   $$\left[0, \frac{4s^2}{0.71}\right] = [0, 12.7] \text{ resp. } [0, 3.56].$$

5. (2+4 poäng) Antalet cyklister som passerar GB-bron i Stuvesbergs Centrum under en minut är Poissonfördelat med väntevärde 1.5. Antalet cyklister under olika minuter antas oberoende av varandra.

   (a) Sannolikheten att minst två cyklister passerar bron under en given minut är

   $$P(\xi \geq 2) = 1 - P(\xi = 0) - P(\xi = 1) = 1 - e^{-\lambda}\left(\frac{\lambda^0}{0!} + \frac{\lambda^1}{1!}\right) = 0.44$$

   (b) Låt $\zeta = \sum_{k=1}^{60} \xi_k$. Sannolikheten (approximativt) att minst 60 cyklister passerar bron under en timma är

   $$P(\zeta > 60)$$

   där ζ approximativt är $\mathrm{N}(60\lambda, \sqrt{60\lambda})$ enl. CGS och $60\lambda = 90$, $\sqrt{60\lambda} = 9.48683$.

   $$P(\zeta > 60) = 1 - P(\zeta \leq 60) \approx 1 - \Phi\left(\frac{60-90}{9.48683}\right) = \Phi\left(\frac{30}{9.48683}\right) = 0.999$$

6. (3 + 2 + 2 poäng)

   Familjen Yrsel åkte på gemensam cykelsemester förra sommaren. Semestern varade i 32 dagar och varje dag innebar en cykelsträcka på 3 mil (vi kan tänka oss att varje sträcka var likvärdig i svårighetsgrad).

   Vi kan tänka oss att tiden en sträcka cyklades är en responsvariabel (y). Responsvariabeln kan påverkas av tre olika faktorer (ålder, kön och om det regnar eller inte). Då familjen består av jämnåriga föräldrar och deras två barn som är tvillingsyskon (en flicka och en pojke) samt att det regnade 16 av dagarna som familjen cyklade så kan man undersöka responsvariabeln med hjälp av ett 2^3 försök där det existerar 4 replikat för varje provgrupp.

   (a) Beräkna huvud- och samspels-effekter.

   ### Lösning

   $$M = \frac{164.9 + 43 + 159.1 + 38.5 + 224.7 + 179 + 222.2 + 183}{8} = 151.8$$
   $$l_A = \frac{-164.9 + 43 - 159.1 + 38.5 - 224.7 + 179 - 222.2 + 183}{4} = -81.85$$
   $$l_B = \cdots = -2.2$$
   $$l_C = \cdots = 100.9$$
   $$l_{AB} = \cdots = 2.0$$
   $$l_{AC} = \cdots = 38.4$$
   $$l_{BC} = \cdots = 2.9$$
   $$l_{ABC} = \cdots = 1.3$$

   (b) Beräkna ett 95%-igt referensintervall och avgör vilka skattade effekter som är signifikanta. Till din hjälp får du att den skattade standardavvikelsen för effekterna har räknats ut till $s_{\text{effekt}} = 4.5$. Denna har man räknat ut med hjälp av alla s_i värdena i tabellen.

Faktorer	A: Ålder	B: Kön	C: Regn
Låg (-)	12	Man	Torrt
Hög (+)	42	Kvinna	Blött

Nr.	A	B	C	$\hat{y}_i$	s_i
1	-	-	-	164.9	4.3
2	+	-	-	43.0	3.9
3	-	+	-	159.1	4.8
4	+	+	-	38.5	5.1
5	-	-	+	224.7	7.0
6	+	-	+	179.0	4.9
7	-	+	+	222.2	7.7
8	+	+	+	183.0	10.5

Tabell 1: Faktorer samt försöksplan för tiden det tar att cykla en sträcka om 3 mil.

Nr.	A	B	C	AB	AC	BC	ABC	$\hat{y}_i$	s_i
1	-	-	-	+	+	+	-	164.9	4.3
2	+	-	-	-	-	+	+	43.0	3.9
3	-	+	-	-	+	-	+	159.1	4.8
4	+	+	-	+	-	-	-	38.5	5.1
5	-	-	+	+	-	-	+	224.7	7.0
6	+	-	+	-	+	-	-	179.0	4.9
7	-	+	+	-	-	+	-	222.2	7.7
8	+	+	+	+	+	+	+	183.0	10.5

Tabell 2: Provtagningsplan med samspelseffekterna tillagda.

Lösning

Först beräknar vi den gemensamma skattningen av variansen såsom

$$s_p^2 = \frac{3s_1^2 + 3s_2^2 + \ldots}{3 * 8} = 40.65.$$

Eftersom effekten har variansen $\sigma^2_{\text{effekt}} = 4\frac{\sigma^2}{8}$ så blir skattningen av effekternas varians

$$s^2_{\text{effekt}} = \frac{40.65}{2} = 20.3.$$

Alltså får effekten den skattade standardavvikelsen $s_{\text{effekt}} = 4.5$. Det tvåsidiga 95%-iga referensintervallet fås genom att multiplicera standardavvikelsen för effekten med 97.5% kvantilen för t-fördelningen med frihetsgraden $8(4-1) = 24$. Alltså blir referensintervallet

$$0 \pm t_{0.025}(df = 24) \cdot s_{\text{effekt}} = 0 \pm 2.06 \cdot 4.5 = 0 \pm 9.27.$$

Från referensintervallet kan vi slutleda oss till att effekterna A, C, och AC är signifikanta eftersom dessa alla har ett absolutvärde större än 9.27.

(c) Under familjen Yrsels cykelsemester så cyklade far och dotter på racingcyklar och mor och son på mountainbikes. Detta kan ha spelat in i hur snabbt de cyklade. Man kan alltså räkna in en fjärde faktor, D som motsvarar cykeltyp där låg (-) skulle motsvara racingcykel och (+) mountainbike.

Försöksplanen är alltså egentligen en reducerad försöksplan, 2^{4-1}-plan. Bestäm generator, ord och upplösning för denna reducerade försöksplan.

Lösning

Då en person är antingen gammal och kvinna eller ung och man så kommer de cykla på mountainbikes, annars inte. Det betyder alltså att D:s kolumn motsvarar tvåfaktorsamspelet AB då AB har positivt värde då A och B är båda låga eller A och B är båda höga.

Nr.	A	B	C	AB	AC	BC	ABC	D	$\hat{y}_i$	s_i
1	-	-	-	+	+	+	-	+	164.9	4.3
2	+	-	-	-	-	+	+	-	43.0	3.9
3	-	+	-	-	+	-	+	-	159.1	4.8
4	+	+	-	+	-	-	-	+	38.5	5.1
5	-	-	+	+	-	-	+	+	224.7	7.0
6	+	-	+	-	+	-	-	-	179.0	4.9
7	-	+	+	-	-	+	-	-	222.2	7.7
8	+	+	+	+	+	+	+	+	183.0	10.5

Tabell 3: Tabell med den exta faktorn D inkluderad.

Man har alltså generatorn $D = AB$. Detta ger den definierande relationen $I = ABD$. Ordet är alltså ABD som har upplösning III. Vi har alltså en 2^{4-1}_{III}-plan.

7. (6 poäng)
Alice, Bob och Carol är tre ölintresserade kemistudenter. Var för sig har de i sina lägenheter bryggt fyra på förhand utvalda ölsorter. De har använt samma recept och samma instruktioner, men de undrar om de olika omständigheterna i deras lägenheter och/eller deras utrustning har lett till en skillnad i alkoholhalten. I första tabellen nedan kan vi se den uppmätta alkoholhalten. I den andra tabellen nedan kan vi se den påbörjade ANOVA-tabellen.

	Ölsort 1	Ölsort 2	Ölsort 3	Ölsort 4
Alice	5.2	4.1	6.8	5.9
Bob	5.4	3.9	6.5	5.5
Carol	5.0	3.5	6.5	5.6

Variationskälla	SS	df	MS	F
Mellan ölbryggare	0.245	...	...	...
Mellan ölsorter	11.949	...	...	...
Okänd	...	...	...	
Totalt	12.363	...		

Fyll i resten av ANOVA-tabellen och utför ett hyptestest för att testa om alkoholhalten påverkas av vem som brygger ölet.

Lösning: Den fullständiga ANOVA-tabellen bör se ut så här:

Variationskälla	SS	df	MS	F
Mellan ölbryggare	0.245	2	0.123	4.366
Mellan ölsorter	11.949	3	3.983	141.97
Okänd	0.168	6	0.028	
Totalt	12.363	11		

Vi använder de fem stegen.

Steg 1. Låt α_1, α_2 och α_3 beteckna hur stor effekt respektive ölbryggare har på alkoholhalten. Våra hypoteser blir då $H_0 : \alpha_1 = \alpha_2 = \alpha_3$ 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

$$F_r = \frac{MS_r}{MS_o} = \frac{SS_r/(r-1)}{SS_o/(r-1)(k-1)}.$$

I vårt fall är $r - 1 = 3 - 1 = 2$ och $(r-1)(k-1) = (3-1)(4-1) = 6$ så vår testvariabel är F-fördelad med 2 och 6 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 $SS_t = SS_r + SS_k + SS_o$ så vi vet att $SS_o = 12.363 - 0.245 - 11.949 = 0.168$. Frihetsgraderna är $r - 1 = 2$, $k - 1 = 3$, $(r-1)(k-1) = 6$ och $n - 1 = 11$. Vi vet också att $MS = SS/df$ vilket ger att $MS_r = 0.245/2 = 0.123$, $MS_k = 11.949/3 = 3.983$ och att $MS_o = 0.168/6 = 0.028$. Tillslut får vi att $F_r = MS_r/MS_o = 0.123/0.028 = 4.366$ och att $F_k = MS_k/MS_o = 3.983/0.028 = 141.97$. Eftersom det är variationen mellan raderna (ölbryggarna) vi är intresserade av är det F_r vi vill jämföra med tabellen.

Steg 5. Vi tittar i F-fördelningstabellen med 2 och 6 frihetsgrader och ser att det kritiska värdet är 5.14. Eftersom vårt observerade värde $F_r = 4.366$ är mindre än det kritiska värdet så kan vi inte förkasta H_0 på signifikansnivå $\alpha = 0.05$.

Slutsats: Vi kan inte dra slutsatsen att alkoholhalten påverkas av vem utav studenterna som brygger ölet.

8. (4 + 2 + 2 poäng)

En forskare är intresserad av ett specifikt väntevärde. Hen misstänker att väntevärdet är större än 20. För att testa detta har hen samlat in ett stickprov av storlek 64 från den relevanta populationen och resultaten blev $\bar{x} = 21.5$ och $s^2 = 38.4$.

(a) Utför ett hypotestest för att bedöma om forskarens misstankar stämmer. Använd signifikansnivå $\alpha = 0.05$.

(b) Hur stort är p-värdet för testet?

(c) Givet att stickprovsstorleken och stickprovsvariansen förblir samma, vilket är det minsta värde på stickprovsmedelvärdet som leder till att nollhypotesen förkastas på signifikansnivå $\alpha = 0.05$?

Lösning:

(a) Vi använder de fem stegen. Låt μ vara väntevärdet av intresse.

Steg 1 Testet bör vara ensidigt eftersom forskaren vill påvisa att väntevärdet är större än 20.

$$H_0 : \mu \leq 20 \qquad \text{mot} \qquad H_1 : \mu > 20$$

Steg 2 Signifikansnivån var given: $\alpha = 0.05$.

Steg 3 Testvariabeln är

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

som är normalfördelad eftersom $n = 64 > 30$.

Steg 4 Den observerade testvariabeln blir

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{21.5 - 20}{\sqrt{38.4}/\sqrt{64}} = \frac{1.5}{6.2/8} = 1.935.$$

Steg 5 Vi tittar i $N(0,1)$-tabellen och får fram det kritiska värdet 1.645. Eftersom den observerade testvariabeln är större (längre ifrån 0) än det kritiska värdet så bör nollhypotesen förkastas.

(b) Vi utgår från vår observerade testvariabel 1.935, tittar i $N(0,1)$-tabellen och får fram värdet 0.026 som är vårt p-värde.

(c) Vi söker alltså det väre på $\bar{x}$ som gör att p-värdet blir just under 0.05. Tabellvärdet för 0.05 är 1.645 så vår observerade testvariabel borde vara just under 1.645, eller med andra ord vill vi att

$$\frac{\bar{x} - 20}{6.2/8} < 1.645.$$

Vi löser olikheten för $\bar{x}$ och får att

$$\bar{x} < 1.645 * 6.2/8 + 20 = 21.27.$$

Alltså är 21.27 det minsta värdet på $\bar{x}$ som leder till att nollhypotesen förkastas.