TENTAMEN: Statistisk modellering för I3, TMS160, måndagen den 13 april 2004 kl 8:45 - 11:45 på V. **Jour:** Erik Brodin, ankn. 5077
Hjälpmedel: Utdelad formelsamling med tabeller, BETA, på kursen använd ordlista och typgodkänd räknedosa.
Poängberäkning: Uppgifterna är av flervalstyp, där endast ett alternativ är rätt. Korrekt besvarad uppgift ger 2 poäng, obesvarad uppgift (vet inte eller alternativ f) ger 0 poäng och felaktigt besvarad uppgift ger -0.5 poäng (flera ifyllda alternativ ger automatiskt -1/2 poäng). Inlämnade lösningar kommer ej tas hänsyn till vid rättningen. Fyll i och lämna in denna sida.
Svar: Läggs ut i studieportalen efter tentamens slut.

Uppgift	a	b	c	d	e	f (vet ej)	Poäng
1	☐	☐	☐	☐	☐	☐	
2	☐	☐	☐	☐	☐	☐	
3	☐	☐	☐	☐	☐	☐	
4	☐	☐	☐	☐	☐	☐	
5	☐	☐	☐	☐	☐	☐	
6	☐	☐	☐	☐	☐	☐	
7	☐	☐	☐	☐	☐	☐	
8	☐	☐	☐	☐	☐	☐	
9	☐	☐	☐	☐	☐	☐	
10	☐	☐	☐	☐	☐	☐	
11	☐	☐	☐	☐	☐	☐	
12	☐	☐	☐	☐	☐	☐	
13	☐	☐	☐	☐	☐	☐	
14	☐	☐	☐	☐	☐	☐	
15	☐	☐	☐	☐	☐	☐	

1 Nedanstående tabell visar antalet besökande (uppmätt som antalet övernattningar) för tio skidsportsorter under en viss period. För varje ort anges dessutom den samlade pistlängden (i miles) samt samt liftkapaciteten (i personer per timme).

Ort	Pist-längd	Lift - kapa-citet	Antal övernatt-ningar
1	10.5	2 200	19 929
2	2.5	1 000	5 839
3	13.1	3 250	23 696
4	4.0	1 475	9 881
5	14.7	3 800	30 011
6	3.6	1 200	7 241
7	7.1	1 900	11 634
8	22.5	5 575	45 684
9	17.0	4 200	36 476
10	6.4	1 850	12 068

Man vill använda linjär regressionsanalys till att beskriva hur antalet övernattningar hänger samman med pistlängd och liftkapacitet.

En eventuell multikollinearitet mellan pistlängd och liftkapacitet ger problem därför att

a ☐ Man får ett litet värde på R^2.

b ☐ Variansen av interceptttermen blir mycket stor.

c ☐ Man kan inte skilja effekten av pistlängden och liftkapaciteten åt.

d ☐ Residualkvadratsumman blir för stor.

e ☐ Inget av ovanstående.

f ☐ Vet ej.

2 Man har samlat in n observationer från en fördelning med okänd varians. Vilken av följande faktorer påverkar *inte* bredden på ett konfidensintervall för väntevärdet.

a ☐ Addition av en konstant till var och en av observationerna.

b ☐ Storleken på stickprovet.

c ☐ Storleken av skattningen av standardavvikelsen.

d ☐ Konfidensgraden.

e ☐ Alla ovanstående alternativ påverkar längden av konfidensintervallet.

f ☐ Vet ej.

3 Ett företag som tillverkar en bestämd produkt startar tillverkningen när den mottar en order. Storleken av ordern varierar från gång till gång. Man är intresserad av att beskriva hur produktionstiden beror av orderstorleken, och vill använda linjär regressionsanalys till detta.

Tabellen nedan visar orderstorlek (i antal tillverkade enheter) och produktionstid (i mantimmar) för ett antal order

Order nr	Order-storlek	Produktions-tid
1	30	73
2	20	50
3	60	128
4	80	170
5	40	87
6	50	108
7	60	135
8	30	69
9	70	148
10	60	132

Man kan kontrollera modellens normalfördelningsantagande genom att rita en Q-Q plott av

a ☐ Antalet mantimmar.

b ☐ Antalet mantimmars avvikelse från regressionslinjen.

c ☐ Orderstorleken.

d ☐ De standardiserade värdena av antalet mantimmar

e ☐ Inget av ovanstående.

f ☐ Vet ej.

4 En studie av 1436 kvinnor som varit gifta minst en gång gav följande resultat.

Utbildning	Gifta en gång	Gifta flera gånger	Totalt
Högskola	550	61	611
Ingen Högskola	681	144	825
Totalt	1231	205	1436

Vad kan man som starkast säga om hypotesen "antalet gånger kvinnorna varit gifta är oberoende av deras utbildning"?

a ☐ Vi kan förkasta på signifikansnivå 10%

b ☐ Vi kan förkasta på signifikansnivå 5%

c ☐ Vi kan förkasta på signifikansnivå 2.5%

d ☐ Vi kan förkasta på signifikansnivå 1%

e ☐ Inget av ovanstående.

f ☐ Vet ej.

5 Vilka av följande datamängder kan man direkt, utan att transformera, med fördel applicera linjär regression på?

Figur 1: Datamängd A visas i figur 14.10, datamängd B visas i figur 14.11, datamängd C visas i figur 14.12 och datamängd D visas i figur 14.13

a ☐ A och C.

b ☐ B.

c ☐ Alla.

d ☐ A och D.

e ☐ Inget av ovanstående.

f ☐ Vet ej.

6 Ett test har utvecklas för att hitta en sjukdom bland individer som är äldre än 50 år. Vi vet att approximativt 10% av den åldersgruppen är smittade. Vi vet även att vid en undersökning av smittade personer gav testet ett korrekt resultat i 85% av fallen samt att vid en undersökning av friska människor rapporterade testet att 4% var smittade.

Vad är den avrundade sannolikheten att en individ har sjukdomen givet att testet indikerar detta?

a ☐ 0.66

b ☐ 0.68

c ☐ 0.70

d ☐ 0.72

e ☐ Inget av ovanstående.

f ☐ Vet ej.

7 För att mäta något man är intresserad av är det vanligt att man först genomför en referensmätning. Vi antar att man gör oberoende mätningar. Först gör man 25 mätningar på referensämnet och beräknar standardavvikelsen till 1.08 g/ml.

Därefter gör man 12 mätningar av halten i det prov man verkligen är intresserad av. Medelvärdet av dessa mätningar blev 91.12 och standardavvikelsen blev 0.73 g/ml

Man vet att standardavvikelsen är den samma vid referensmätningar och riktiga mätningar.

Det mest korrekta 95%- iga konfidensintervallet för halten i provet ges då av:

a ☐ $91.12 \pm t_{35,0.975} \cdot \sqrt{\frac{1}{12} \cdot \frac{24 \cdot 1.08^2 + 11 \cdot 0.73^2}{43}}$ [g/ml]
(Rätt)

b ☐ $91.12 \pm t_{35,0.95} \cdot \sqrt{\frac{1}{12} \cdot \frac{24 \cdot 1.08^2 + 11 \cdot 0.73^2}{43}}$ [g/ml]

c ☐ $91.12 \pm t_{11,0.975} \cdot \sqrt{\frac{0.73^2}{15}}$ [g/ml]

d ☐ $91.12 \pm t_{11,0.95} \cdot \sqrt{\frac{0.73^2}{15}}$ [g/ml]

e ☐ $91.12 \pm t_{11,0.975} \cdot \sqrt{\frac{1}{11} \cdot \frac{24 \cdot 1.08^2 + 11 \cdot 0.73^2}{43}}$ [g/ml]

f ☐ vet ej

8 Vad undersöks med randomized complete block design (RBCD)?

a ☐ Medelvärdet av k stycken populationer som också påverkas av en extern störvariabel.

b ☐ Om en mängd observationer är dragna av en speciell fördelning.

c ☐ Linjärt beroende mellan variabler.

d ☐ Effekten av k stycken behandlingsformer på en homogen mängd.

e ☐ Inget av ovanstående.

f ☐ Vet ej.

9 Vi kategoriserar frukostflingor efter sockerhalt. Vi låter kategori 0 stå för flingor av müslityp, 1 för majsflingor och 2 för flingor av typen honungsrostadechokladdoppadesockerpuffar. 90 butiker valdes ut slumpmässigt. I 30 butiker så undersöktes flingorna närmast kassorna, i 30 butiker flingorna i mitten och i 30 butiker flingorna längst bort från kassorna.

Man vill undersöka om sockerinnehållet har ett samband med placeringarna relativt kassorna. Hur gör man det bäst?

a ☐ χ^2-test för oberoende i en tvåsidig tabell för kategoriska data.

b ☐ Ensidig variansanalys utan blockindelning (one-way model).

c ☐ Ensidig variansanalys med blockindelning (RCBD)

d ☐ Regressionsanalys.

e ☐ Inget av ovanstående.

f ☐ vet inte

10 Den stokastiska variabeln X har väntevärde 2 och standardavvikelse 3 och variablen Y har väntevärde 3 och standardavvikelse 4. Vidare är korrelationen mellan X och Y lika med 0.25. Sätt $Z = X - Y + 5$.

Då är variansen för Z lika med

a ☐ 13

b ☐ 15

c ☐ 16

d ☐ 17

e ☐ Inget av ovanstående.

f ☐ vet ej

11 Vi låter 400 fotbollspelare, proffs och amatörer, välja mellan Nike och Adidasskor.

	Nike	Adidas	Totalt
Proffs	131	84	215
Amatörer	99	86	185
Totalt	230	170	400

Vid ett χ^2-test av radandelar fås följande datorutskrift.

```
Expected counts are printed below observed counts

         Nike      Adidas    Total
    1      131        84      215
         123,63    91,38

    2       99        86      185
         106,38    78,63

Total      230       170      400

Chi-Sq =  0,440 +  0,595 +
          0,511 +  0,692 = 2,238
DF = 1, P-Value = 0,135
```

Vilken är den starkaste slutsatsen som kan dras från denna utskrift?

a ☐ vi kan på signifikansnivå 0.05 säga att proffs föredrar Nike mer än Adidas

b ☐ vi kan på signifikansnivå 0.10 säga att amatörer föredrar Adidas.

c ☐ vi kan på signifikansnivå 0.05 säga: "det är inte någon preferensskillnad mellan proffs och amatörer"

d ☐ vi kan på signifikansnivå 0.15 säga att det är preferensskillnad mellan proffs och amatörer.

e ☐ inget av ovanstående.

f ☐ vet inte.

12 Betrakta följande frågor:

1 Hur kan man upptäcka interaktion i en interaktionsgraf?

2 Var kan man förvänta sig interaktion?

Dessa frågor besvaras bäst med:

a ☐ 1: Korsande linjer. 2: Kvinnors längd relativt storlek på klack.

b ☐ 1: Korsande linjer. 2: Kvinnors längd relativt färg på byxor.

c ☐ 1: Parallella linjer. 2: Kvinnors längd relativt storlek på klack.

d ☐ 1: Parallella linjer. 2: Kvinnors längd relativt färg på byxor.

e ☐ Inget av ovanstående

f ☐ vet inte

13 Vad är sant om faktor design relativt OFAT (one factor at a time).

a ☐ OFAT kräver ett mindre antal observationer än ett faktorförsök för att få samma information om faktoreffekter.

b ☐ OFAT kan ej modelera interaktion.

c ☐ Bruset i faktor design behöver ej vara oberoende.

d ☐ Bruset i faktor design behöver ej vara stationärt.

e ☐ Inget av ovanstående.

f ☐ vet ej.

14 Vi har följande ofullständiga ANOVA tabell:

Analysis of variance					
Source	DF	Sum of square	Mean sauare	F Stat	Prob> F
Model	*	12.12	4.04	*	*
Error	21	18.90	*		
C Total	*	*			

Hur många observationer har gjorts?

a ☐ 22

b ☐ 23

c ☐ 24

d ☐ 25

e ☐ Inget av ovanstående.

f ☐ vet inte

15 Vid en ensidig variansanalys gör man vissa antagande. Man bör då kontrollera om antaganden stämmer för en datamängd man vill analysera. Vilket av följanden påstående stämmer bäst?

a ☐ Man antar bland annat stationaritet och detta kan kontrolleras med en QQ-plot.

b ☐ Man antar bland annat oberoende och detta kan kontrolleras med en QQ-plot.

c ☐ Man antar bland annat att data är t-fördelade och detta kan kontrolleras med en QQ-plot.

d ☐ Man antar bland annat stationaritet och detta kan kontrolleras med ett histogram.

e ☐ Inget av ovanstående.

f ☐ vet ej.