

TENTAMEN: Dataanalys och statistik för I2, TMS136

Onsdagen den 25 oktober kl. 8.30-13.30 på M.

Jour: Jenny Andersson, ankn 5317

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) ger 0 poäng och felaktigt besvarad uppgift ger -0.5 poäng (flera ifyllda alternativ ger automatiskt -0.5 poäng). Inlämnade lösningar kommer ej tas hänsyn till vid rättningen. Fyll i och lämna in denna sida.

Svar: Lägg ut i studieportalen efter tentamens slut.

Uppgift	a	b	c	d	e	Vet inte	Poäng
1	☐	☐	☐	☐	☐	☐	
2	☐	☐	☐	☐	☐	☐	
3	☐	☐	☐	☐	☐	☐	
4	☐	☐	☐	☐	☐	☐	
5	☐	☐	☐	☐	☐	☐	
6	☐	☐	☐	☐	☐	☐	
7	☐	☐	☐	☐	☐	☐	
8	☐	☐	☐	☐	☐	☐	
9	☐	☐	☐	☐	☐	☐	
10	☐	☐	☐	☐	☐	☐	
11	☐	☐	☐	☐	☐	☐	
12	☐	☐	☐	☐	☐	☐	
13	☐	☐	☐	☐	☐	☐	
14	☐	☐	☐	☐	☐	☐	
15	☐	☐	☐	☐	☐	☐	

- 1 Det är känt att hushållens inkomster i ett visst samhälle följer en normalfördelning med väntevärde $\mu = \$42000$ och standardavvikelse $\sigma = \$5000$. Medianen av hushållens inkomster i detta samhälle är
- a ☐ $= 47000$ \$.
 - b ☐ < 42000 \$.
 - c ☐ > 42000 \$.
 - d ☐ $= 42000$ \$.
 - e ☐ Det går inte att avgöra vad medianen är.
- 2 För drygt 10 år sedan infördes kunskapsrelaterade betyg i skolorna istället för relativa betyg. När de relativa betygen sattes skulle de bästa 7% av eleverna i sin årskurs få betyg 5. I gymnasiet gavs centralprov för alla elever i en årskurs. Antag att resultaten på ett sådant centralprov var normalfördelade med $\mu = 75$ och $\sigma = 5$. Vilket var det lägsta poängantalet, avrundat till närmaste heltal, som krävdes för att få betyg 5 på provet?
- a ☐ 66.
 - b ☐ 84.
 - c ☐ 68.
 - d ☐ 82.
 - e ☐ 87.
- 3 Du vill uppskatta andelen göteborgare som är för kraftiga böter för miljöbrott. Hur stort stickprov behövs för att vara 90% säker på att stickprovsandelen inte kommer skilja sig från den sanna andelen i populationen med mer än 0.01?
- a ☐ ca 6800.
 - b ☐ ca 9600.
 - c ☐ ca 40.
 - d ☐ ca 80.
 - e ☐ ca 100.

4 I en undersökning motsvarande den i uppgift 3 utförd i Kalifornien, svarade 80% av 1000 personer ja på frågan angående kraftiga böter för miljöbrott. Vilket av följande är ett 90% konfidensintervall för den sanna andelen som vill ha kraftiga böter? (Undre gränser avrundas neråt och övre gränser uppåt.)

- a ☐ ca (0.77, 0.83).
- b ☐ ca (0.74, 0.86).
- c ☐ ca (0.70, 0.85).
- d ☐ ca (0.69, 0.89).
- e ☐ ca (0.79, 0.81).

5 Vilket av följande påståenden är sant om en fördelning som är vänsterskev?

- a ☐ Medianen är mindre än medelvärdet.
- b ☐ Medianen och medelvärdet är lika stora.
- c ☐ Medianen är större än medelvärdet.
- d ☐ Standardavvikelsen är liten
- e ☐ Standardavvikelsen är stor.

6 En dag i december är sannolikheten 0.4 att det kommer snöa i Boston och sannolikheten är 0.7 att det kommer snöa i Moskva. Antag att vädret i Boston och Moskva är oberoende. Vad är sannolikheten att det inte snöar i någon av de bägge städerna en viss dag i december?

- a ☐ 0.9
- b ☐ 0.28
- c ☐ 0.18
- d ☐ 0.03
- e ☐ 0

7 Föraren till en lastbil vars last består av 900 lådor kommer bötfällas om totala vikten på lådorna överstiger 36450 pounds. Vikten hos en låda har väntevärde 40 pounds och standardavvikelse 6 pounds. Vad är sannolikheten att föraren bötfälls? Tips: Använd centrala gränsvärdessatsen.

- a ☐ 0.0319.
- b ☐ 0.4938.
- c ☐ 0.0062.
- d ☐ 0.9938.
- e ☐ < 0.0001 .

8 Tjugo män har deltagit i ett viktnedskningsprogram. Deras vikter mättes före och efter programmets genomförande. Vikten före minus vikten efter beräknades för alla män. Medelvärdet av de beräknade viktskillnaderna var 2.5 kg och standardavvikelsen 0.5 kg. Skillnaderna är approximativt normalfördelade. Man vill testa om männen har minskat i vikt. Vad blir p-värdet för ett sådant test?

- a ☐ $p\text{-värde} < 0.01$.
- b ☐ $0.01 \leq p\text{-värde} < 0.05$.
- c ☐ $0.05 \leq p\text{-värde} < 0.1$.
- d ☐ $0.1 \leq p\text{-värde} < 0.2$.
- e ☐ $p\text{-värde} \geq 0.2$.

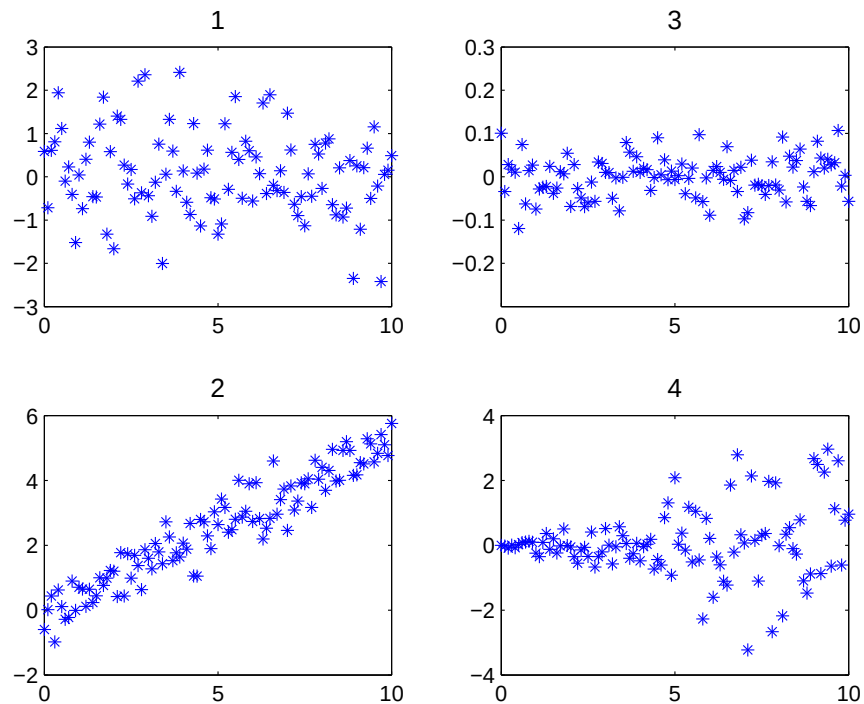
- 9 En plot av segertiderna, Y , i minuter i Boston maraton mellan 1927 och 1997 mot årtal, T , antyder ett linjärt samband. En regression där segertiden är responsvariabel och årtalet är regressorvariabel utförs. Residualplottarna visar inte på något mönster. Följande beräknades vid regressionen:

$$\hat{Y} = 1029 - 0.453 \cdot T$$
$$s = 3.56775 \quad R^2 = 0.874$$

Både vid test av om skärningen är noll och om lutningen är noll gavs p-värdet till < 0.0005 . Skatta vad segertiden bör ha varit 1980 i timmar:min:sek.

- a ☐ 1:32:06
- b ☐ 2:12:06
- c ☐ 2:12:04
- d ☐ 2:03:54
- e ☐ Det går inte att skatta från data i uppgiften.

10 Givet är 4 plottar med data mot tid. Avgör vilka som tyder på stationära data.



- a ☐ 1 stationär 2, 3, 4 ickestationära.
- b ☐ 1, 3 stationära 2, 4 ickestationära.
- c ☐ 1, 4 stationära 2, 3 ickestationära.
- d ☐ 3 stationär 1, 2, 4 ickestationära.
- e ☐ 4 stationär 1, 2, 3 ickestationära.

- 11 Ett 95% konfidensintervall för parametern μ , baserat på ett stickprov på $n = 40$ mätningar, ges av

(0.19, 0.35).

Vilket av följande påståenden är sant?

- a ☐ Av alla fortsatta mätningar så kommer 95% ligga i konfidensintervallet.
 - b ☐ Nästa mätning kommer med 95% sannolikhet ligga i intervallet.
 - c ☐ Om många stickprov av storlek $n = 40$ tas och ett nytt konfidensintervall beräknas för varje stickprov så kommer i genomsnitt 95% av intervallen innehålla det sanna värdet på μ .
 - d ☐ Det är helt säkert att det sanna värdet på μ ligger i intervallet.
 - e ☐ Ett 90% konfidensintervall kommer vara längre än konfidensintervallet ovan.
- 12 I vilket av följande fall är det lämpligt att analysera data med ett parat t-test, d.v.s. de två populationerna som studeras är inte oberoende?
- a ☐ Man vill testa om en lyxig hudkräm har bättre effekt än en billigare. Två grupper om vardera 15 personer har slumpvis valts ur en större grupp på 1000 personer. Den ena gruppen får använda den lyxiga hudkrämen i en månad och den andra gruppen får den billiga hudkrämen (i samma lyxiga förpackning) i en månad.
 - b ☐ PCB-halten i pilgrimsfalksbon mäts i 15 slumpvis utvalda bon 1985 och i 15 slumpvis utvalda bon 2005.
 - c ☐ En forskare vill undersöka om uthålligheten under ett träningspass är bättre när senaste måltiden före passet hade lågt GI jämfört med om den hade högt GI. Forskaren har hittat 10 cyklister som ställer upp som försökspersoner. En dag får de äta den ena typen av måltid och sedan cykla testcykel. Nästa dag får de äta den andra typen av måltid och cykla testcykel. Forskaren singlar slant om vilken måltid som ska ätas första dagen.
 - d ☐ En affär har frågat 80 (31 män och 49 kvinnor) av sina kunder om de kan tänka sig att komma tillbaka till affären.
 - e ☐ Vikten hos 100 svenska förskolebarn jämförs med vikten hos 100 amerikanska förskolebarn.

13 Ett test av

$$H_0 : \mu_1 = \mu_2$$

mot

$$H_a : \mu_1 \neq \mu_2$$

förkastar nollhypotesen på signifikansnivå 0.05. Vilket påstående är alltid sant?

- a ☐ Ett 90% konfidensintervall för $\mu_1 - \mu_2$ kommer innehålla 0.
- b ☐ Ett 99% konfidensintervall för $\mu_1 - \mu_2$ kommer inte innehålla 0.
- c ☐ Ett 95% konfidensintervall för $\mu_1 - \mu_2$ kommer innehålla 0.
- d ☐ Ett 95% konfidensintervall för $\mu_1 - \mu_2$ kommer inte innehålla 0.
- e ☐ Inget av påståendena a-d är sant.

14 När det gäller regression vill man avgöra om den modell som använts passar till data. Vilket av följande är det viktigaste verktyget för att avgöra detta?

- a ☐ R^2
- b ☐ r
- c ☐ Ett 95% konfidensintervall för responsvariabeln.
- d ☐ Ett test av om lutningen är noll.
- e ☐ Residualplottar.

- 15 Föräldrar i Sverige har fått svara på vad de helst väljer för färg på nappen till sina barn och om barnet är en pojke eller en flicka. Resultatet blev följande.

	Grön	Röd	Gul	Turkos	Rosa	Blå	Tot
Flickor	110	106	91	48	87	38	480
Pojkar	83	114	30	135	2	156	520
Tot	193	220	121	183	89	194	1000

Pearson residualerna anges i tabellen nedan:

	Grön	Röd	Gul	Turkos	Rosa	Blå
Flickor	1.8	0.04	4.3	-4.3	6.8	-5.7
Pojkar	-1.7	-0.04	-4.2	4.1	-6.5	5.5

Det vanliga χ^2 -testet gav ett p-värde på < 0.005 . Låt p_{fi} vara den sanna andelen av föräldrar till flickor som väljer färgen i kolumn i på nappen och låt p_{pi} vara motsvarande andel om barnet är en pojke. Vilket av följande påståenden är **inte** sant?

- a ☐ Ett test av

H_0 : kön och nappfärg oberoende

H_a : kön och nappfärg beroende

förkastar nollhypotesen.

- b ☐ Ett test av

H_0 : $p_{fi} = p_{pi}$ för alla $i = 1, 2, \dots, 6$

H_a : H_0 ej sann

förkastar nollhypotesen.

- c ☐ Ett test av

H_0 : $p_{fi} = p_{pi}$ för alla $i = 1, 2, \dots, 6$

H_a : H_0 ej sann

kan inte förkasta nollhypotesen.

- d ☐ Pearson residualen för rosa napp och flicka är stor och positiv, vilket betyder att rosa nappar är överrepresenterade bland flickor jämfört med om nappfärg och kön är oberoende.

- e ☐ Pearson residualen för röd napp och pojke är liten, vilket betyder att röda nappar inte verkar avvika nämvärt från oberoende antagandet.