TENTAMENSSKRIVNING PÅ KURSERNA

GRUNDLÄGGANDE STATISTIK A4 (15 hp)
STATISTIK FÖR EKONOMER A8 (15 hp)

2014-03-26

UPPLYSNINGAR

A. Tillåtna hjälpmedel:
Kursspecifik formelsamling (utan anteckningar)
Språklexikon
Miniräknare

B. **Skrivtid: 8.00-13.00** Skrivningen omfattar 5 uppgifter, om sammanlagt 100 poäng.

C. För varje uppgift anges den maximala poäng som kan erhållas. Om en uppgift är uppdelad på deluppgifter anges den maximala poängen för varje deluppgift. Ibland kan inte deluppgifterna bedömas oberoende av varandra, vilket kan innebära att poäng inte utdelas på en senare uppgift om inte tidigare deluppgift lösts på ett i princip riktigt sätt. Dock gäller att utdelad poäng för varje deluppgift aldrig kan vara negativ.

D. Om Du känner Dig osäker på någonting (skrivningens genomförande, någon formulering i en uppgift, om något hjälpmedel är otillåtet), fråga då jourhavande skrivningsvakt eller den skrivningsansvariga läraren.

E. Efter skrivningens slut får Du behålla sidorna med frågeställningarna (de ska inte lämnas in!). Preliminära lösningar anslås på Pingpong.

UPPMANINGAR

A. Följ noga de anvisningar som finns på skrivningsförsättsbladet.

B. Redovisa Dina lösningar i en form som gör det lätt att följa Din tankegång! (Det dunkelt uttryckta förutsätts av rättaren vara dunkelt tänkt.) Motivera alla väsentliga steg i lösningen. Ange alla antaganden Du gör och alla förutsättningar Du utnyttjar.

C. Vid konfidensintervall måste Du dessutom ange vad intervallet avser att täcka samt teckna intervallet i symbolform innan de numeriska uppgifterna insätts. Verbal slutsats av det framräknade intervallet krävs för full poäng.

D. Vid hypotesprövning måste Du utöver vad som sägs i punkt B ovan ange vad hypotesprövningen avser att testa, hypoteserna i symbolform (då så är möjligt), signifikansnivå, testfunktion (inklusive antal frihetsgrader då detta är aktuellt) både i symbolform och med numeriska uppgifter, beslutsregel, resultat samt verbal slutsats.

E. Vid standardvägning ska metod anges och kalkylerna ska följas av en verbal slutsats för full poäng.

Uppgift 1

Facebook ger en mängd statistik på sin hemsida som i detalj beskriver tillväxten och popularitеten hos webbplatsen. Ett exempel på sådan statistik är att den genomsnittliga användaren har 130 vänner (på Facebook). På ett större universitet gjordes ett slumpmässigt urval om 30 Facebookanvändare där de utvalda fick ange hur många vänner de har på Facebook.

Resultatet av undersökningen finns i form av en Minitabkolumn på sista sidan i denna skrivning. Kontrollera att du har den! Vidare har vi följande sammanfattning i nedanstående Minitabutskrift.

Descriptive Statistics: Friends

```
Variable    Mean  SE Mean  StDev
Friends   119.07     5.40  29.57
```

(4) **A** Ange vad som är individ/element och vad som är variabel i den här situationen. Ange dessutom den aktuella variabelns datanivå samt huruvida den är diskret eller kontinuerlig. För full poäng måste svaren rörande datanivå och diskret/kontinuerlig motiveras.

(5) **B** Ge en ordentlig förklaring av innebörden av värdena i ovanstående sammanfattning.

(7) **C** Åskådliggör materialet grafiskt genom att konstruera ett lådagram/boxplot.

(8) **D** Använd resultatet i detta urval för att konstruera ett intervall som med 95% säkerhet täcker in medelvärdet för antal Facebookvänner gällande Facebookanvändare vid universitetet.

(4) **E** Intervallet i D-uppgiften ansågs som alldeles för brett. Hur stort stickprov behövs för att bredden av ett 95% konfidensintervall för populationsmedelvärdet ska bli högst 10 vänner brett. Använd information från detta stickprov för att göra beräkningen.

(12) **F** Anta att vi istället tolkar ”genomsnittligt” som att medianանvändaren har 130 vänner på Facebook. Undersök med ett hypotestest angående p om medianանvändaren (vad det gäller antal vänner på Facebook) vid detta universitet avviker från medianանvändaren i allmänhet. Använd en signifikansnivå på 5% och utför testet enligt klassisk metod.

Uppgift 2

På en stormarknad ville man undersöka hur försäljningen i 1000-kronor (y) av en vara påverkas av exponeringsytan i kvadratmeter (x). Man gjorde ett experiment under sju veckor då exponeringsytan varierade mellan 2 och 4 kvadratmeter. Resultatet sammanfattas via summorna

$$\sum x = 21, \sum y = 56.7, \sum x^2 = 67, \sum y^2 = 462.51, \sum xy = 172.5$$

(6) **A** Skatta den linjära regressionsmodellen som visar hur försäljningen beror på exponeringsytan. Uppskatta med hjälp av regressionsmodellen den försäljning vi förväntar oss vid en exponeringsyta på 3 kvadratmeter.

(4) **B** Tolka på ett begripligt sätt (dvs utan att använda statistiska facktermer) de båda regressionskoefficienterna i ord. Ange en orsak till varför tolkningen av interceptet bör tas med en nypa salt.

Uppgift 3

Elva anställda i ett företag fick träffa företagets sjuksköterska på grund av höga kolesterolvärden. Sköterskan informerade dem om farorna med höga kolesterolvärden och satte dem på en ny diet. I tabellen nedan finner du kolesterolvärden för de elva anställda i undersökningen både innan dieten påbörjades och även en månad efter att den påbörjats. Kan det i och med detta resultat, i statistisk mening, påstås att behandlingen uppnår sitt mål?

Anställd	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Före	255	230	290	242	300	250	215	230	225	219	236
Efter	197	225	215	215	240	235	190	240	200	203	223

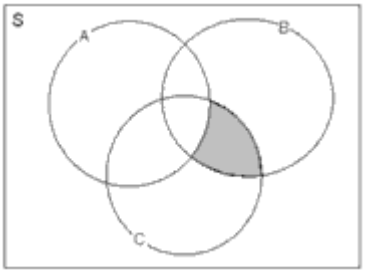
(3) **A** Valet av testfunktion i den här situationen avgörs av om den för inferensen så nödvändiga sannolikhetsbedömningen kan baseras på normalfördelningen eller inte. Vad är det som måste vara normalfördelat och varför är det nödvändigt i det här fallet? Förklara. Observera att du inte behöver göra någon normalfördelningskontroll.

(12) **B** Anta att normalfördelningsantagandet som diskuterades i A-uppgiften kan anses vara uppfyllt. Utför ett fullständigt hypotestest enligt *p*-värdesmetoden som utnyttjar detta antagande. Använd en signifikansnivå på 1%

(12) **C** Anta nu att normalfördelningsantagandet som diskuterades i A-uppgiften inte anses vara realistiskt. Utför ett fullständigt hypotestest på 1% signifikansnivå där du använder det test som utifrån förutsättningarna utnyttjar informationen på bästa sätt.

Uppgift 4

I ett visst slumpförsök studeras de tre händelserna A, B och C. Venn-diagrammet nedan beskriver utfallsrummet för detta slumpförsök.



(3) **A** Beskriv med mängdlärans symboler (dvs union, snitt och komplement) den händelse som är gråmarkerad i Venn-diagrammet.

(3) **B** Anta att de tre händelserna A, B och C är oberoende och att det dessutom gäller att Pr(A)=Pr(B)=Pr(C)=0.3. Visa att sannolikheten för den händelse som är gråmarkerad i Venn-diagrammet är 0.063.

(6) **C** Anta vidare att vi utför detta slumpförsök vid 25 tillfällen. Bestäm sannolikheten att den händelse som är gråmarkerad i Venn-diagrammet inträffar vid åtminstone två av dessa 25 tillfällen. Var noga med att ordentligt motivera dina beräkningar.

Uppgift 5

MyTVLab är en onlinetjänst som bl a tillåter användare att ladda upp och dela egna klipp. För att attrahera och behålla besökare på webbplatsen, måste man se till att användarna snabbt kan ladda ner önskade videoklipp. Tidigare data indikerar på att medelvärdet för nedladdningstiden av videoklipp är 7 sekunder med en standardavvikelse på 2 sekunder samt att ungefär två-tredjedelar av videoklippen laddas ner på mellan 5 till 9 sekunder.

(3) **A** Vi avser att i beräkningarna nedan använda normalfördelningen. Ange vad i de angivna förutsättningarna som ger stöd för detta val av sannolikhetsfördelning och ange även vad som mer måste till för att vi ordentligt ska kunna motivera dess användning.

(4) **B** Beräkna sannolikheten att nedladdningstiden för ett slumpmässigt valt videoklipp ligger mellan 8 och 9 sekunder.

(4) **C** MyTVLab arbetar kontinuerligt med att minska nedladdningstiden. Det närmsta målet är att 75% av alla videoklipp ska kunna laddas ner på under 7 sekunder. Vilket medelvärde för nedladdningstid måste man rikta in sig på för att målet ska nås (förutsatt att standardavvikelsen förblir oförändrad)?

Bilaga till Uppgift 1

↓	C1
	Friends
1	72
2	74
3	83
4	85
5	85
6	96
7	97
8	99
9	103
10	104
11	105
12	106
13	110
14	111
15	112
16	118
17	119
18	119
19	120
20	126
21	127
22	128
23	137
24	148
25	152
26	154
27	158
28	160
29	171
30	193

1. Statistik i samband med Facebookanvändande.

   (a) I den här situationen är det *facebookanvändare vid universitetet* som är individer/element och *antal facebookvänner* som är variabel. Eftersom denna variabel endast kan anta heltalsvärden gäller att den är *diskret*. Vidare gäller exempelvis att 4 vänner är dubbelt så många som 2 vänner varför det för denna variabels värden är meningsfullt att göra relativa jämförelser. Således mäts variabeln på *kvotskalan*.

   (b) Utifrån minitabutskriften finner vi att det för de undersökta i gruppen Facebookanvändare gäller att sammanfattningen fås att

$$\begin{aligned} \overline{x} &= 119.07 \\ s &= 29.57 \\ s_{\overline{x}} &= \frac{s}{\sqrt{n}} = 5.40 \end{aligned}$$

   För de Facebookanvändare som var med i undesökningen gäller alltså att det genomsnittliga antalet Facebookvänner var 119. Samtliga i gruppen hade dock inte lika många Facebookvänner utan detta antal avvek med i genomsnitt ca 30 vänner från det genomsnittliga antalet. Begreppet SE Mean betyder *Standard Error of the Mean* och översätts av oss till *medelfelet*. Stickprovsmedelvärdet används för att skatta populationsmedelvärdet och medelfelets uppgift är att ge oss en uppskattning av det genomsnittliga felet i denna skattning. Enligt denna uppskattning gäller således att stickprovsmedelvärdet (vid upprepade stickprov av denna storlek) i genomsnitt kommer att avvika från populationsmedelvärdet med 5.4 vänner.

(c) För att kunna konstruera ett lådagram behöver vi median och kvartiler. Utifrån informationen i appendix finner vi att

$$\begin{aligned} q_1 &= \left(\text{Värdet på observation } \frac{30+1}{4} = 7.75\right) = \\ &= 97 + 0.75 \cdot (99 - 97) = 98.5 \\ md &= \left(\text{Värdet på observation } \frac{30+1}{2} = 15.5\right) = \\ &= 112 + 0.5 \cdot (118 - 112) = 115 \\ q_3 &= \left(\text{Värdet på observation } \frac{3 \cdot (30+1)}{4} = 23.25\right) = \\ &= 137 + 0.25 \cdot (148 - 137) = 139.75 \end{aligned}$$

Ett och ett halvt kvartilavstånd ges av

$$1.5 \cdot (139.75 - 98.5) = 61.875$$

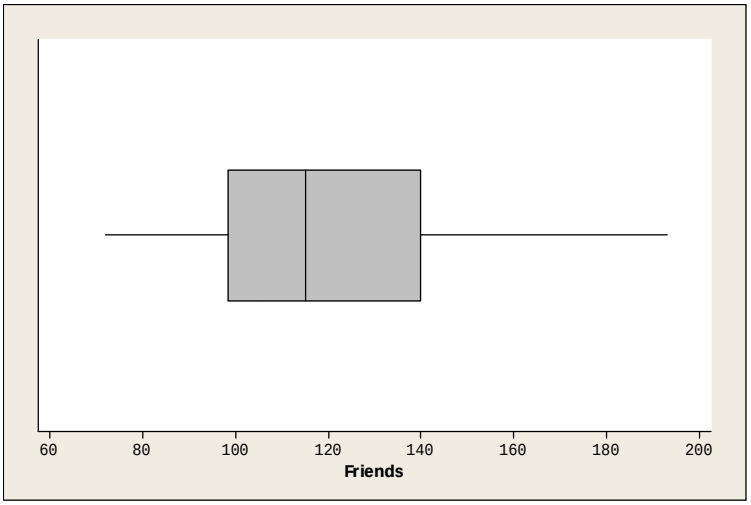
varför uteliggare är observationer under

$$98.5 - 61.875 = 36.625$$

och över

$$139.75 + 61.875 = 201.63$$

Det finns därmed inte några uteliggare i vårt material. Lådagrammet får följande utseende



(d) Vi ska konstruera ett 95% konfidensintervall för μ där

μ = Medelvärde för antal Facbookvänner för Facbookanvändare vid universitetet

I och med att $n = 30 \leq 30$ har vi ett någorlunda stort stickprov men inte tillräckligt stort för att utan vidare använda Centrala gränsvärdessatsen. Vi bör därför göra en inledande kontroll av vårt material och försäkra oss om att variabeln, dvs antral facebookvänner, är (någorlunda) symmetriskt fördelad. Vi kan exempelvis jämföra medelvärde och median och även studera lådagrammet från c-uppgiften. Vi konstaterar att det inte är ett helt symmetriskt material men samtidigt inte alarmerande asymmetriskt. Alltså borde vi kunna gå vidare och utgå från att stickprovsmedelvärdet approximativt kan betraktas som normalfördelat. Då detta är ett större universitet utgår vi från att antal Facebookanvändare vid universitetet är så pass många att ändlighetskorrektion kan bortses från. Vidare står i uppgiften att de 30 studenterna i urvalet är slumpmässigt valda vilket innebär att vi kan använda konfidensintervallet

$$\overline{x} \pm t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Här gäller att det aktuella tabellvärdet blir $t_{29,0.025} = 2.045$ vilket innebär att konfidensintervallet efter insättning av våra värden blir

$$119.07 \pm 2.045 \cdot \frac{29.57}{\sqrt{30}}$$

eller som ett intervall

$$\mathbf{108 \leq \mu \leq 130}$$

Med 95% säkerhet befinner sig μ, dvs genomsnittligt antal Facebookvänner hos Facebookanvändare vid universitetet, någonstans mellan 108 och 130.

(e) För att kunna bestämma hur stort stickprov som ska tas för att intervallet ska bli maximalt 10 vänner brett måste vi ha en uppfattning om standardavvikelsen i populationen. Stickprovet ovan gav oss en möjlighet att skatta denna via

$$\widehat{\sigma} = 29.57$$

Eftersom $z_{0.025} = 1.96$ och halva bredden $B = 5$ följer att den sökta stickprovsstorleken blir

$$n = \frac{1.96^2 \cdot 29.57^2}{5^2} = 134.36$$

För att uppfylla kraven krävs alltså ett stickprov om minst **135** Facebookanvändare vid universitetet.

(f) Vi börjar med att konstatera att frågan huruvida medianandvändaren av Facebook vid universitetet har 130 Facebookvänner kan ställas upp som en frågeställning angående p. Låter vi

$$p = \text{Andel Facebookanvändare vid universitetet som har färre än 130 Facebookvänner}$$

formuleras hypoteserna utifrån frågeställningen på följande sätt:

$$\begin{aligned} H_0 &: p = 0.5 \\ H_1 &: p \neq 0.5 \end{aligned}$$

Detta test ska utföras på 5% signifikansnivå. Stickprovet är ett slumpmässigt urval och precis som tidigare antar vi att antal Facebookanvändare vid universitetet är så pass många att ändlighetskorrektion kan bortses från. Eftersom

$$np_0\left(1-p_0\right) = 30 \cdot 0.5 \cdot 0.5 = 7.5 > 5$$

är stickprovet (med liten marginal) tillräckligt stort för att normalapproximation av binomialfördelningen ska vara tillåten. Därmed ska vi använda testfunktionen

$$z = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0 \cdot (1-p_0)}{n}}}$$

I och med att vi använder en signifikansnivå på 5% samtidigt som att mothypotsen är $H_1 : p \neq 0.5$ följer att nollhypotesen ska förkastas först om

$$z_{obs} > z_{0.025} = 1.96, \text{ eller } z_{obs} < -z_{0.025} = -1.96$$

I urvalet blev andelen Facebookanvändare med färre än 130 Facebookvänner

$$\widehat{p} = \frac{22}{30} = 0.7333$$

vilket alltså ger ett tillsynes starkt stöd åt att medianandvändaren av Facebook vid universitetet inte har 130 Facebookvänner. Frågan är hur övertygande resultatet är? Vi sätter in i testfunktionen

$$z = \frac{0.73 - 0.5}{\sqrt{\frac{0.5 \cdot 0.5}{30}}} = 2.56$$

och eftersom

$$z_{obs} = 2.56 > 1.96 = z_{0.025}$$

har vi hamnat i det kritiska området och därmed förkastas nollhypotesen. Det är alltså på 5% signifikansnivå statistiskt säkerställt att p, andel Facebookanvändare vid universitetet som har färre än 130 Facebookvänner, inte är 50%. Alltså är det på 5% signifikansnivå statistiskt säkerställt att medianandvändaren av Facebook vid universitetet inte har 130 Facebookvänner.

2. Utifrån den givna informationen börjar vi med att bestämma de tre nyckelsummorna

$$\begin{aligned}\sum (x-\overline{x})^2 &= 67 - \frac{21^2}{7} = \mathbf{4} \\ \sum (y-\overline{y})^2 &= 462.51 - \frac{56.7^2}{7} = \mathbf{3.24} \\ \sum (x-\overline{x})(y-\overline{y}) &= 172.5 - \frac{21 \cdot 56.7}{7} = \mathbf{2.4}\end{aligned}$$

(a) Utifrån nyckelsummorna ovan finner vi först att

$$b = \frac{2.4}{4} = 0.6$$

och sedan att

$$a = \frac{56.7}{7} - 0.6 \cdot \frac{21}{7} = 6.3$$

Regressionslinjens ekvation blir således

$$\widehat{\mathbf{y}} = \mathbf{6.3} + \mathbf{0.6} \cdot \mathbf{x}$$

Utifrån regressionsekvationen uppskattar vi den förväntade försäljningen vid en exponeringsyta på 3 kvadratmeter till

$$\widehat{y}_{x=3} = 6.3 + 0.6 \cdot 3 = 8.1$$

dvs **8 100** kronor.

(b) Vi tolkar b-koefficienten som att en extra kvadratmeters exponeringsyta i genomsnitt ökar försäljningen med 600 kronor. Interceptet, dvs a-koefficienten, anger att den genomsnittliga försäljningen då varan överhuvudtaget inte exponeras bör vara ungefär 6 300 kronor. Eftersom denna situation inte förekom i undersökningen blir detta en extrapolation och tolkningen av resultatet bör därför tas med en nypa salt.

3. Parvisa observationer.

   (a) Det är differenserna, dvs den effekt dieten har på kolesterolvärdet som måste vara approximativt *normalfördelad* i den bakomliggande populationen. Med population avser vi i det här fallet antingen personer med högt kolesterolvärde i företaget eller mer allmänt personer med höga kolesterolvärden. Om vi, rent hypotetiskt, kunde placera ut alla i populationen på en skala utifrån den effekt dieten har på just deras kolesterolvärde ska den resulterande kurvan vara mycket lik en normalfördelningskurva. Detta är ett nödvändigt antagande eftersom vi i b-uppgiften ska utföra ett hypotestest angående medeldifferensen i populationen och stickprovet är inte tillräckligt stort för att vi utan antaganden kan förutsätta att medeldifferensen i stickprovet är approximativt normalfördelad.

   (b) Vi har ett litet stickprov med endast 11 observationer (anställda) men eftersom *differenserna* antas vara normalfördelade löses många problem. Om vi vidare kan betrakta personerna i undersökningen som slumpmässigt utvalda ur den bakomliggande populationen (möjligtvis tveksamt) och inte på något sätt påverkar varandras resultat (också det något tveksamt) kan ett parametriskt t-test användas. Låter vi

   $$\mu_d = \text{Den genomsnittliga effekten av dieten}$$

   formuleras hypoteserna som

   $$\begin{aligned} H_0 &: \mu_d = 0 \\ H_1 &: \mu_d > 0 \end{aligned}$$

   där vi här mäter effekten som Före–Efter vilket betyder att ett positivt värde innebär att kolesterolvärdet gått ner. Det går givetvis lika bra att ta differenserna i omvänd ordning men då förstås med olikheten i mothypotesen vänd åt andra hållet. Vi ämnar utföra testet med en signifikansnivå på 1% vilket innebär att nollhypotesen ska förkastas först om testets p-värde understiger 1%. Eftersom detta handlar om parvisa observationer börjar vi med att bestämma differenserna, vilket vi som nämnts ovan gör genom att beräkna Före–Efter så att positiva värden är bra för påståendet att dieten har en önskad effekt.

Anställd	1	2	3	4	5	6	7	8	9	10	11
Före	255	230	290	242	300	250	215	230	225	219	236
Efter	197	225	215	215	240	235	190	240	200	203	223
Före−Efter	58	5	75	27	60	15	25	−10	25	16	13

Utifrån stickprovsdifferenserna beräknas medelvärde och standardavvikelse till

$$\begin{aligned} \overline{d} &= \frac{309}{11} = 28.1 \\ s_d &= \sqrt{\frac{15\,343 - \frac{309^2}{11}}{10}} = 25.8 \end{aligned}$$

Den aktuella testfunktionen rör populationens medelvärde och då den tänkta populationen bestående av personer med högt kolesterolvärde kan betraktas som stor kan ändlighetskorrektion bortses från. Med specifika beteckningar för parvisa observationer kan formeln skrivas som

$$t = \frac{\overline{d} - D_0}{s_d/\sqrt{n}}$$

varför vi efter insättning av värden får

$$t_{obs} = \frac{28.1 - 0}{25.8/\sqrt{11}} = 3.61$$

som skall jämföras med t_{10}-fördelningen. Eftersom

$$t_{10,0.005} 3.169 < t_{obs} = 3.61 < 4.144 = t_{10,0.001}$$

drar vi slutsatsen att

$$0.1\% < p\text{-värde} < 0.5\%$$

och då p-värdet understiger den uppsatta signifikansnivån på 1% förkastas nollhypotesen. Det är på 1% signifikansnivå statistiskt säkerställt att dieten, i genomsnitt (med avseende på medelvärde), sänker kolesterolvärdet för personer som före dieten hade en hög kolesterolnivå.

(c) Vi har ett litet stickprov med endast 11 observationer och eftersom differenserna *inte* kan antas vara normalfördelade måste vi använda ett icke-parametriskt test. Vi förutsätter som i föregående uppgift (det något tveksamma) att personerna i undersökningen är slumpmässigt utvalda och inte på något sätt påverkar varandras resultat. Detta tillsammans med det faktum att variabeln mäts på kvotskala innebär att ett *teckenrangtest* kan användas. I detta icke-parametriska test undersöks om fördelningen vad det gäller testresultat är samma i de "båda" populationerna (dvs före och efter genomgången diet), eller om medianen för differenserna är noll, dvs

H_0 : Fördelningen för kolesterolvärde är samma både före och efter dieten

H_1 : Fördelningen för kolesterolvärde efter dieten är förskjuten nedåt

Vi får

Anställd	1	2	3	4	5	6	7	8	9	10	11
Före	255	230	290	242	300	250	215	230	225	219	236
Efter	197	225	215	215	240	235	190	240	200	203	223
Före−Efter	58	5	75	27	60	15	25	−10	25	16	13
Rang	9	1	11	8	10	4	6.5	2	6.5	5	3
Tecken	+	+	+	+	+	+	+	−	+	+	+

Vi ser att det inte förekommer några ties, dvs observationer med samma resultat vid båda mätningarna. Vi förväntar oss låga rangtal på dom negativa differenserna vilket innebär att vi som testfunktion använder T_-. Det följer att

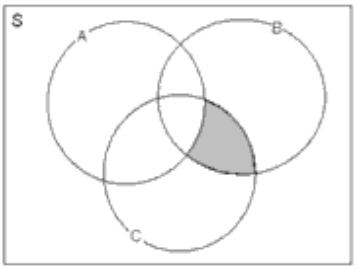
$$T_- = 2$$

och eftersom

$$T_- = 2 < 7 = T_{11,0.01}$$

har vi hamnat i det kritiska området och förkastar (nu även med den icke-parametriska testmetoden) nollhypotesen. Det är alltså på 1% signifikansnivå säkerställt att dieten, i genomsnitt (med avseende på median), sänker kolesterolvärdet för personer som före dieten hade en hög kolesterolnivå.

4. Det aktuella Venn-diagrammet har följande utseende.



(a) Det gråmarkerade området är $\overline{A} \cap B \cap C$.

(b) Enligt förutsättningarna gäller att $\Pr(A) = \Pr(B) = \Pr(C) = 0.3$. Eftersom händelserna dessutom är oberoende av varandra följer att

$$\begin{aligned} \Pr\left(\overline{A} \cap B \cap C\right) &= \Pr\left(\overline{A}\right) \cdot \Pr(B) \cdot \Pr(C) = [1 - \Pr(A)] \cdot \Pr(B) \cdot \Pr(C) = \\ &= 0.7 \cdot 0.3 \cdot 0.3 = \mathbf{0.063} \end{aligned}$$

(c) Vi börjar med att låta

$$X = \text{Antal gånger det gråmarkerade området inträffar}$$

Vi förutsätter att slumpförsöken utförs oberoende av varandra, dvs att resultatet i ett slumpförsök inte påverkar resultatet i något annat slumpförsök. Eftersom det är samma slumpförsök som utförs om och om igen bör vi hela tiden ha samma sannolikhet $p = 0.063$ att det gråmarkerade området inträffar. Eftersom X dessutom räknar hur många gånger det inträffar följer att X är $Bi(25, 0.063)$. Därmed följer att

$$\begin{aligned} \Pr(X \geq 2) &= 1 - \Pr(X \leq 1) = 1 - [\Pr(X = 0) + \Pr(X = 1)] = \\ &= 1 - \left[\tbinom{25}{0} \cdot 0.063^0 \cdot 0.937^{25} + \tbinom{25}{1} \cdot 0.063^1 \cdot 0.937^{24}\right] \\ &= 1 - (0.1966 + 0.3304) = \mathbf{0.473} \end{aligned}$$

5. Vi betraktar nu slumpvariabeln

$$X = \text{Nedladdningstid för ett slumpmässigt vald videoklipp}$$

som $N(7, 2)$ där enheten är sekunder.

(a) Det fakum att ungefär två-tredjedelar av nedladdningstiderna ligger inom en standardavvikelse från medelvärdet stämmer bra överens med vad som gäller för normalfördelningen. Dock säger det inget om att fördelningen är symmetrisk. Vi skulle alltså behövs se en graf över det historiska materialet för att kunna göra en symmetribedömning.

(b) Vi söker

$$\begin{aligned}\Pr(8 < X < 9) &= \Pr\left(\frac{8-7}{2} < Z < \frac{9-7}{2}\right) = \Pr(0.5 < Z < 1) = \\ &= \Pr(Z < 1) - \Pr(Z < 0.5) = 0.8413 - 0.6915 \approx \mathbf{0.15}\end{aligned}$$

(c) Enligt Tabell 5.2.B gäller att

$$z_{0.25} = 0.6745$$

vilket vi tolkar som att den angivna tiden på 7 sekunder ska befinna sig 0.6745 standardavvikelser över medelvärdet. Med bibehållen standardavvikelse på 2 sekunder betyder detta att vi får ekvationen

$$7 = \mu + 0.6745 \cdot 2$$

vilket innebär att man ska sikta in sig på medelvärdet

$$\mu = 7 - 0.6745 \cdot 2 = \mathbf{5.65}$$

TENTAMENSSKRIVNING PÅ KURSERNA

GRUNDLÄGGANDE STATISTIK A4 (15 hp)
STATISTIK FÖR EKONOMER A8 (15 hp)

2014-04-26

UPPLYSNINGAR

A. Tillåtna hjälpmedel:
Kursspecifik formelsamling (utan anteckningar)
Språklexikon
Miniräknare

B. **Skrivtid: 9.00-14.00** Skrivningen omfattar 5 uppgifter, om sammanlagt 100 poäng.

C. För varje uppgift anges den maximala poäng som kan erhållas. Om en uppgift är uppdelad på deluppgifter anges den maximala poängen för varje deluppgift. Ibland kan inte deluppgifterna bedömas oberoende av varandra, vilket kan innebära att poäng inte utdelas på en senare uppgift om inte tidigare deluppgift lösts på ett i princip riktigt sätt. Dock gäller att utdelad poäng för varje deluppgift aldrig kan vara negativ.

D. Om Du känner Dig osäker på någonting (skrivningens genomförande, någon formulering i en uppgift, om något hjälpmedel är otillåtet), fråga då jourhavande skrivningsvakt eller den skrivningsansvariga läraren.

E. Efter skrivningens slut får Du behålla sidorna med frågeställningarna (de ska inte lämnas in!). Preliminära lösningar anslås på Pingpong.

UPPMANINGAR

A. Följ noga de anvisningar som finns på skrivningsförsättsbladet.

B. Redovisa Dina lösningar i en form som gör det lätt att följa Din tankegång! (Det dunkelt uttryckta förutsätts av rättaren vara dunkelt tänkt.) Motivera alla väsentliga steg i lösningen. Ange alla antaganden Du gör och alla förutsättningar Du utnyttjar.

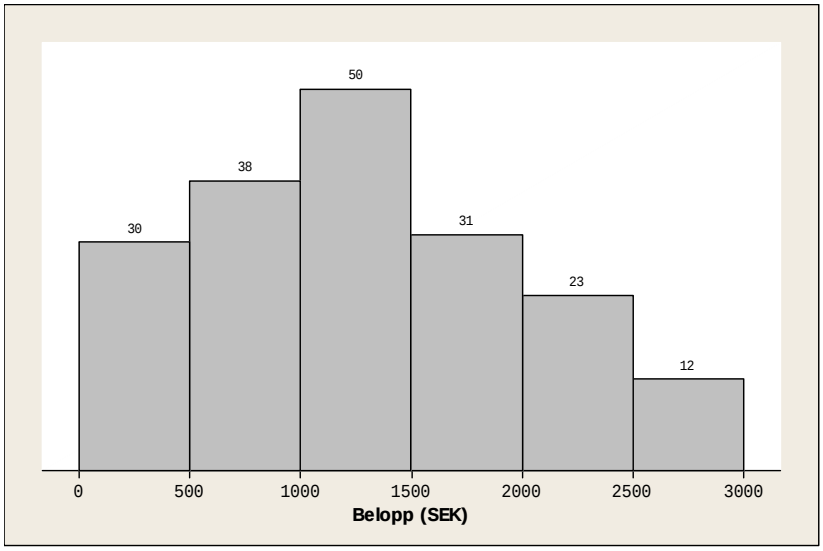
C. Vid konfidensintervall måste Du dessutom ange vad intervallet avser att täcka samt teckna intervallet i symbolform innan de numeriska uppgifterna insätts. Verbal slutsats av det framräknade intervallet krävs för full poäng.

D. Vid hypotesprövning måste Du utöver vad som sägs i punkt B ovan ange vad hypotesprövningen avser att testa, hypoteserna i symbolform (då så är möjligt), signifikansnivå, testfunktion (inklusive antal frihetsgrader då detta är aktuellt) både i symbolform och med numeriska uppgifter, beslutsregel, resultat samt verbal slutsats.

E. Vid standardvägning ska metod anges och kalkylerna ska följas av en verbal slutsats för full poäng.

Uppgift 1

Histogrammet nedan visar beloppen som ett slumpmässigt urval om 184 kunder spenderade i en viss affär.



För enkelhets skull används vid beräkningarna de angivna gränserna som de faktiska klassgränserna, dvs vi använder 0, 500, 1000 osv..

(2) **A** När vi utför beräkningarna nedan görs ett antagande. Vilket är detta antagande? *Anmärkning*: Antagandet i fråga är inte nödvändigt för att lösa deluppgift E.

(7) **B** Beräkna medelvärde och standardavvikelse för den aktuella variabeln. Ge en ordentlig förklaring av innebörden av dessa båda värden.

(4) **C** Beräkna den tredje kvartilen i stickprovet. Ge en ordentlig förklaring av innebörden av detta värde.

(7) **D** Åskådliggör den *kumulativa* frekvensfördelningen i materialet med ett lämpligt diagram. Använd diagrammet (en formell beräkning är redan gjord i C-uppgiften) till att göra en uppskattning av värdet på den första kvartilen för den aktuella variabeln.

(8) **E** Använd det aktuella urvalet för att konstruera ett 90 % konfidensintervall för andelen kunder som spenderar mindre än 1 000 kronor vid ett köptillfälle.

(12) **F** Är det i och med resultatet i vårt stickprov statistiskt säkerställt att det genomsnittliga köpbeloppet överstiger 1 200 kronor? Utför ett fullständigt hypotestest enligt *p*-värdesmetoden där du använder en signifikansnivå på 5%.

(8) **G** Vi fortsätter nu med situationen i F-uppgiften. Låt oss betrakta det aktuella hypotestestet *innan* resultatet av undersökningen sammanställdes, dvs vi har ännu inte några resultat från undersökningen. Anta att vi, baserat på tidigare undersökningar, har skattningen $\hat{\sigma}_x = 700$. Vad är med denna förutsättning testets styrka då μ=1 300? För full poäng måste situationen beskrivas grafiskt.

Uppgift 2

Vid tre företag A, B och C inom samma bransch ämnar man jämföra olycksfallsrisken under det senaste kalenderåret. Då olycksfallsrisken uppvisar stora skillnader mellan verkstadsarbetare, lagerarbetare och kontorsanställda och då de tre företagen har olika personalsammansättning, är det önskvärt att eliminera effekten av denna snedvridande faktor.

För respektive företag känner man antalet olyckor under året samt antalet anställda i varje yrkeskategori. Dessutom vet man att olycksfrekvensen (antal olyckor/ antal anställda) för branschen som helhet är 0.25, 0.14 och 0.03 för verkstadsarbetare, lagerarbetare respektive kontorsanställda.

Företag	Antal anställda			Antal olyckor
	Verkstads-arbetare	Lager-arbetare	Kontors-anställda	
A	140	20	40	42
B	150	30	100	42
C	100	20	20	28

(10) Gör en jämförelse mellan de tre företagen med avseende på olycksfallsfrekvensen på ett sådant sätt att resultatet blir oberoende av olikheter i fördelning över olika personalkategorier. (Av svaret ska framgå hur många procent under/över branschgenomsnittet företagen ligger).

Uppgift 3 (OBS! Denna uppgift ingår inte längre på kursen)

En tjänstemannaorganisation med totalt 3 000 medlemmar (av vilka 2 000 är i åldersgruppen under 45 år) önskade undersöka medlemmarnas inställning till flexibla arbetstider. För undersökningen utvaldes slumpmässigt två stickprov bestående av 300 medlemmar från åldersgruppen under 45 år och 150 medlemmar från åldersgruppen 45 år eller äldre.

De utvalda fick uppge sin inställning till flexibla arbetstider. Det visade sig att 225 personer i åldersgruppen under 45 år och 108 personer i åldersgruppen 45 år eller äldre var positiva till flexibla arbetstider.

(12) **A** Avgör med hypotesprövning enligt klassisk metod om andel medlemmar som är positivt inställda till flexibla arbetstider skiljer sig mellan de båda åldersgrupperna. Använd 5% signifikansnivå.

(5) **B** Beräkna *p*-värdet för testet i A-uppgiften. Ge en ordentlig tolkning av detta *p*-värde genom att börja med ”Om det är så att andelen medlemmar i tjänsteorganisationen som är positivt inställda till flexibla arbetstider är…”. Observera att tolkningen inte skall gälla huruvida nollhypotesen skall förkastas (detta är redan gjort i A-uppgiften).

Uppgift 4

En tulltjänsteman gör stickprovsundersökningar bland väskor.

(3**)** **A** Studera situationen i B-uppgiften nedan. För att finna den sökta sannolikheten används en slumpvariabel. Ange den aktuella slumpvariabeln och, med motivering, dess sannolikhetsfördelning.

(4) **B** Ett visst parti om 50 väskor innehåller 6 väskor med illegalt innehåll. Sin vana trogen väljer tulltjänstemannen slumpmässigt 5 väskor ur partiet. Beräkna sannolikheten att tulltjänstemannen hittar åtminstone en väska med illegalt innehåll.

Uppgift 5

Längden av en graviditet (hos människor) kan betraktas som approximativt normalfördelad med ett medelvärde/väntevärde på 266 dagar och en standardavvikelse på 16 dagar. *Anmärkning*. Den aktuella variabeln är kontinuerlig.

(4) **A** Bestäm sannolikheten att längden av graviditeten hos en slumpmässigt vald gravid kvinna understiger 240 dagar, dvs att den, grovt räknat, understiger 8 månader.

(3) **B** Bestäm värdet på den första kvartilen.

(6) **C** Anta att vi studerar 12 slumpmässigt valda gravida kvinnor. Bestäm sannolikheten att mer än hälften av dessa ”går över tiden”, dvs har en graviditet som varar längre än förväntat. För full poäng måste beräkningarna ordentligt motiveras.

(5) **D** Vi får reda på att en gravid kvinna har gått fyra dagar över tiden. Bestäm sannolikheten att det går ytterligare fyra dagar (eller mer) innan kvinnan föder sitt barn.

1. Vi börjar med att återge (och utöka) frekenstabellen där vi enligt anvisningarna (för enkelheten) använder de klassgränser som står givna. Klassbeteckningarna ser förbryllande ut men innebörden är, för att exemplifiera, att den *exakta* skiljelinjen mellan de båda första klasserna är 500 kronor.

Köpbelopp (SEK)	f_i	Mitt (x_i)	$f_i x_i$	$f_i x_i^2$	F_i
0–500	30	250	7 500	1 875 000	30
500–1 000	38	750	28 500	21 375 000	68
1 000–1 500	50	1 250	62 500	78 125 000	118
1 500–2 000	31	1 750	54 250	94 937 500	149
2 000–2 500	23	2 250	51 750	116 437 500	172
2 500–3 000	12	2 750	33 000	90 750 000	184
	184		**237 500**	**403 500 000**	

   (a) Eftersom vi endast har tillgång till klassindelat material måste vi förutsätta att observationerna är jämnt fördelade i klasserna för att kunna göra våra beräkningar.

   (b) Vi beräknar medelvärde och standardavvikelse till

$$\begin{aligned} \overline{x} &= \frac{237\,500}{184} = \mathbf{1\,290.8} \\ s &= \sqrt{\frac{403\,500\,000 - \frac{237\,500^2}{184}}{183}} = \mathbf{727.84} \end{aligned}$$

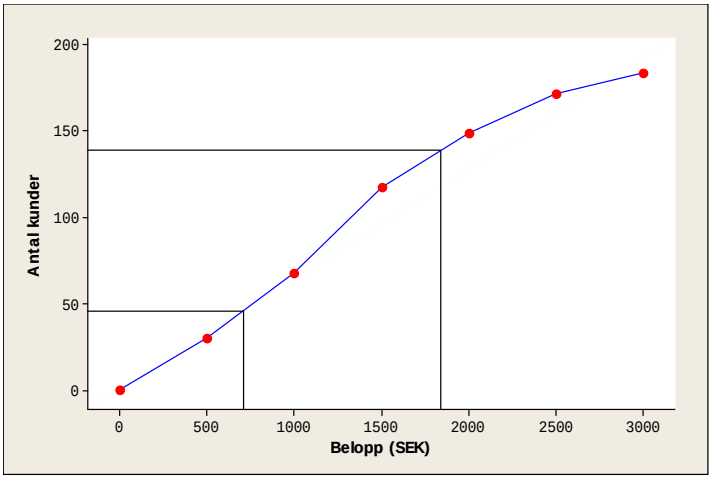
   De 184 kunderna i undersökningen handlade i genomsnitt för 1 291 kronor. Alla kunder handlade dock inte för lika mycket utan köpbeloppen avvek med i genomsnitt 728 kronor från medelvärdet.

   (c) Den första kvartilen är värdet på observation med observartionsnummer $3n/4 = 3 \cdot 184/4 = 138$ vilken befinner sig i klassen 1 500–2 000. Därmed får vi att

$$q_3 = 1\,500 + \frac{138 - 118}{31} \cdot 500 = \mathbf{1\,822.6}$$

   vilket innebär att det i undersökningen var så att en fjärdedel av kunderna handlade för mer än 1 823 kronor.

(d) En summapolygon används för att beskriva den kumulativa frekvensfördelningen. Vi använder därför dom kumulerade frekvenserna från vår frekvenstabell och får på så sätt följande diagram



Anmärkning. Här blev det ett litet tryckfel i uppgiften. Det står att man i diagrammet ska uppskatta den första kvartilen men det är förstås den tredje kvartilen som ska uppskattas. Båda varianterna accepteras. Genom att på y-axeln utgå från observation $n/4 = 184/4 = 46$ och dra en horisontell linje fram till summapolygonen och sedan därifrån dra en lodrät linje ner till x-axeln får vi en uppskattning av köpbeloppet för den första kvartilen. Beräknas den första kvartilen fås 710 kronor vilket verkar rimligt utifrån diagramuppskattningen. Den tredje kvartilen är värdet på observation $3n/4 = 3 \cdot 184/4 = 138$.Enligt resultatet i c-uppgiften ska denna bli 1 823 kronor vilket verkar rimligt.

(e) Vi ska konstruera ett 90% konfidensintervall för p där

$$p = \text{Andelen kunder som spenderar mindre än 1 000 kronor vid ett köptillfälle}$$

Vi förutsätter att kunderna i urvalet kan betraktas som ett slumpmässigt urval bland alla (potentiella) kunder och att populationen (av potentiella kunder) kan antas vara stor vilket betyder att vi kan bortse från ändlighetskorrektion. Det var 68 av de 184 kunderna i urvalet som handlade för under 1 000 kronor varför det följer att vår punktskattning ges av

$$\widehat{p} = \frac{68}{184} \approx 0.37$$

Eftersom

$$n\widehat{p}\,(1-\widehat{p}) = 184 \cdot \frac{68}{184} \cdot \frac{116}{184} = 42.87 > 5$$

är stickprovet med god marginal tillräckligt stort för att normalapproximation av binomialfördelningen ska vara tillåten. Vi använder därför konfidensintervallet

$$\widehat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}}$$

Eftersom $z_{0.05} = 1.6449$ följer efter insättning av våra stickprovsvärden att konfidensintervallet blir

$$0.37 \pm 1.6449 \cdot \sqrt{\frac{0.37 \cdot 0.63}{184}}$$

eller som ett intervall

$$\mathbf{0.311 \leq p \leq 0.428}$$

Med 90% säkerhet befinner sig p, dvs andelen kunder som spenderar mindre än 1 000 kronor vid ett köptillfälle, någonstans mellan 31% och 43%.

(f) Låter vi först

$$\mu = \text{Medelköpbeloppet för kunderna i populationen}$$

följer av frågeställningen i uppgiften att hypoteserna ska formuleras som

$$\begin{aligned} H_0 &: \mu = 1\,200 \\ H_1 &: \mu > 1\,200 \end{aligned}$$

Detta ska nu undersökas med ett test på 5% signifikansnivå vilket innebär att vi ska förkasta nollhypotesen först om p-värdet understiger 5%. Vi förutsätter som ovan att kunderna i urvalet kan betraktas som ett slumpmässigt urval bland alla (potentiella) kunder och att populationen (av potentiella kunder) kan antas vara stor vilket betyder att vi kan bortse från ändlighetskorrektion. Eftersom vi dessutom har att $n = 184 > 30$ följer att vi kan använda testfunktionen

$$z = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

I b-uppgiften fann vi att $\overline{x} = 1\,290.8$ och $s = 727.84$. Insättning av dessa värden ger oss följande värde på testfunktionen

$$z = \frac{1\,290.8 - 1\,200}{727.84/\sqrt{184}} = 1.69$$

vilket utifrån utseendet på mothypotesen innebär att

$$p\text{-värde} = \Pr\left(Z > 1.69\right) = 0.045$$

Eftersom p-värdet understiger den uppsatta signifikansnivån på 5% förkastas nollhypotesen. Det är således på 5% signifikansnivå statistiskt säkerställt att medelköpbeloppet för kunderna i populationen överstiger 1 200 kronor.

(g) Detta är en uppgift som måste lösas i två steg. Först måste vi under nollhypotesantagendet, dvs att $\mu = 1\,200$, ta reda på för vilka värden på stickprovsmedelvärdet nollhypotesen kommer att förkastas och sedan måste vi under den nya förutsättningen, dvs att $\mu = 1\,300$, ta reda på sannolikheten att detta kommer att inträffa (vilket är testets styrka).

i. För vilka värden på $\overline{x}$ kommer nollhypotesen att förkastas? Nollhypotesen förkastas om

$$\frac{\overline{x} - 1\,200}{700/\sqrt{184}} > 1.6449$$

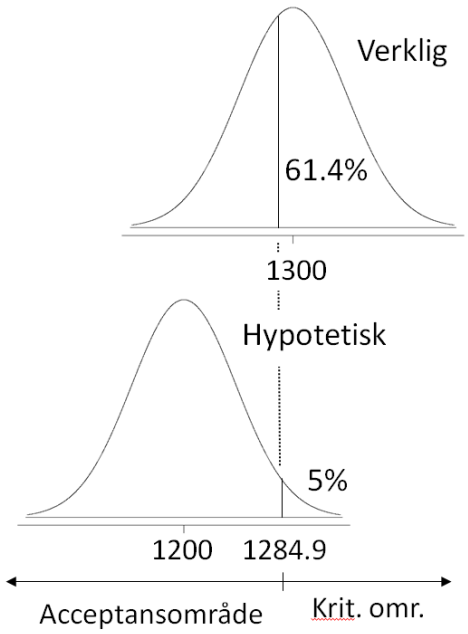
vilket vi översätter till

$$\overline{x} > 1\,200 + 1.6449 \cdot \frac{700}{\sqrt{184}} = 1\,284.9$$

ii. Vad blir $\Pr\left(\overline{X} > 1\,284.9\right)$ under den nya förutsättningen att $\mu = 1\,300$, dvs att $\overline{X}$ är $N\left(1\,300, 700/\sqrt{184}\right)$. På vanligt normalfördelningsmanér uttrycker vi detta i standardavvikelser, dvs

$$\Pr\left(\overline{X} > 1\,284.9\right) = \Pr\left(Z > \frac{1\,284.9 - 1\,300}{700/\sqrt{184}} = -0.29\right) = \mathbf{0.614}$$

Testets styrka, dvs sannolikheten att förkasta en felaktig nollhypotes, blir i den här situationen ca 0.61. Chansen att vi under dessa omständigheter kommer att få tillräckligt övertygande bevis om att medelköpbeloppet bland kunderna (i populationen) överstiger 1 200 kronor är alltså lite drygt 61%. Hela situationen beskrivs väl med följande graf



2. Standardvägning enligt kapacitetsmetoden. För att jämförelsen av olycksfallsfrekvensen på de tre företagen ska bli rättvis använder vi *kapacitetsmetoden* som är en indirekt standardvägningsmetod. På så sätt kan vi eliminera den (eventuella) snedvridande effekt som följer av skillnader i företagens fördelning över olika personalkategorier. Vi finner först de faktiska olycksfallsfrekvenserna för de tre företagen till

$$\overline{x}^A = \frac{42}{200} = 0.21, \quad \overline{x}^B = \frac{42}{280} = 0.15, \quad \overline{x}^C = \frac{28}{140} = 0.2$$

Man bör dock vara försiktig med att jämföra dessa värden eftersom de eventuellt är missvisande. För en bättre jämförelse beräknar vi först *kapacitetstal*, dvs hypotetiska olycksfallsfrekvenser, vilket ger oss de olycksfallsfrekvenser kontoren skulle haft om de följt branschgenomsnitt.

$$\begin{aligned}
\overline{x}^A_{hyp} &= \frac{140}{200} \cdot 0.25 + \frac{20}{200} \cdot 0.14 + \frac{40}{200} \cdot 0.03 = 0.195 \\
\overline{x}^B_{hyp} &= \frac{150}{280} \cdot 0.25 + \frac{30}{280} \cdot 0.14 + \frac{100}{280} \cdot 0.03 = 0.15964 \\
\overline{x}^C_{hyp} &= \frac{100}{140} \cdot 0.25 + \frac{20}{140} \cdot 0.14 + \frac{20}{140} \cdot 0.03 = 0.20286
\end{aligned}$$

varför det följer att respektive *kapacitetsindex* ges av

$$\begin{aligned}
I^A_{kap} &= \frac{0.21}{0.195} \cdot 100 = 107.7 \\
I^B_{kap} &= \frac{0.15}{0.15964} \cdot 100 = 94.0 \\
I^C_{kap} &= \frac{0.2}{0.20286} \cdot 100 = 98.6
\end{aligned}$$

Vi får alltså att

$$I^A_{kap} > I^{\text{branchen}} = 100 > I^C_{kap} > I^B_{kap}$$

Efter att vi tagit hänsyn till skillnader i fördelning över personalkategori finner vi att olycksfallsrisken är störst i företag A (ca 7.7% över branschgenomsnittet) och minst i företag B (ca 6.2% under branschgenomsnittet). Företag C ligger ca 1.5% under branschgenomsnittet.

3. Skiljer sig andelen medlemmar som är positivt inställda till flexibla arbetstider i de båda åldersgrupperna?

   (a) Låter vi p_{-44} och p_{45-} representera den andel av respektive åldersgrupp som är positivt inställda till flexibla arbetstider ska våra hypoteser utifrån frågeställningen formuleras på följande sätt:

$$\begin{aligned} H_0 &: p_{-44} = p_{45-} \\ H_1 &: p_{-44} \neq p_{45-} \end{aligned}$$

   vilka vi tänker undersöka med ett hypotestest på 5% signifikansnivå. Vi har enligt uppgiften två slumpmässiga stickprov, OSU, och förutsätter att dom är dragna oberoende av varandra. Eftersom

$$\begin{aligned} n_{-44} \cdot \widehat{p}_{-44} \cdot (1 - \widehat{p}_{-44}) &= 300 \cdot \frac{225}{300} \cdot \frac{75}{300} = 56.25 > 5 \\ n_{45-} \cdot \widehat{p}_{45-} \cdot (1 - \widehat{p}_{45-}) &= 150 \cdot \frac{108}{150} \cdot \frac{42}{150} = 30.24 > 5 \end{aligned}$$

   är stickproven tillräckligt stora för att normalapproximation ska kunna användas. Eftersom de båda urvalen utgör 15% av respektive delpopulation är det enligt tumregeln inte tillåtet att bortse från ändlighetskorrektion. Vi använder därmed testfunktionen

$$Z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}\,(1-\widehat{p})\left(\frac{1}{n_1} \cdot \frac{N_1 - n_1}{N_1} + \frac{1}{n_2} \cdot \frac{N_2 - n_2}{N_2}\right)}}$$

   som approximativt är $N(0,1)$ då nollhypotesen är sann. Eftersom vi här använder den klassiska metoden på 5% signifikansnivå och mothypotesen är tvåsidig blir vår beslutsregel att nollhypotesen ska förskastas först om

$$|Z_{\text{obs}}| > 1.96$$

   dvs om $Z_{\text{obs}} > 1.96$ eller $Z_{\text{obs}} < -1.96$. Den polade stickprovsandelen ges av

$$\widehat{p} = \frac{300 \cdot 0.75 + 150 \cdot 0.72}{300 + 150} = 0.74$$

   vilket innebär att testfunktionen får värdet

$$z_{\text{obs}} = \frac{0.75 - 0.72}{\sqrt{0.74 \cdot 0.26 \cdot \left(\frac{1}{300} \cdot \frac{2000-300}{2000} + \frac{1}{150} \cdot \frac{1000-150}{1000}\right)}} = 0.74$$

   Eftersom

$$-1.96 < z_{\text{obs}} = 0.74 < 1.96$$

   har vi hamnat i acceptansområdet och nollhypotesen accepteras. Vi har på 5% signifikansnivå *inte* kunnat statistiskt säkerställa att det är någon skillnad i de båda åldersgrupperna vad det gäller andelen medlemmar som är positivt inställda till flexibla arbetstider.

(b) Vi finner testets p-värde till

$$p\text{-värde} = 2 \cdot \Pr(Z > 0.74) \approx 2 \cdot 0.23 = \mathbf{0.46}$$

Om det är så att andelen medlemmar i tjänsteorganisationen som är positivt inställda till flexibla arbetstider är densamma i de båda åldersgrupperna är sannolikheten att få ett så pass här avvikande (eller ännu mer avvikande) stickprovsresultat ca 46%. Det är alltså inte på något sätt ovanligt att få ett sådant här resultat och det är förstås därför vi (på 5% signifikansnivå) accepterar nollhypotesen.

4. Hypergeometrisk fördelning.

(a) Den aktuella slumpvariabeln är

$$X = \text{Antal väskor med illegalt innehåll bland de utvalda}$$

En väska har antingen illegalt innehåll eller inte. Antalet väskor i partiet är ändligt och tulltjänstemannen väljer förstås väskor utan återläggning. Vidare gäller att vår slumpvariabel räknar antalet väskor med illegalt innehåll bland de utvalda. Detta gör att X är $Hyp\left(5, \frac{6}{50}, 50\right)$.

(b) Sannolikheten att tulltjänstemannen hittar åtminstone en väska med illegalt innehåll ges av

$$\Pr(X \geq 1) = 1 - \Pr(X = 0) = 1 - \frac{\binom{6}{0}\binom{44}{5}}{\binom{50}{5}} = 1 - 0.51 = \mathbf{0.49}$$

Det är alltså ungefär 50/50 att tulltjänstemannen hittar (åtminstone) en väska med illegalt innehåll.

5. Vi betraktar nu slumpvariabeln

$$X = \text{Längden av en graviditet}$$

som enligt den givna informationen kan betraktas som approximativt $N(266, 16)$ där enheten är dagar.

(a) Vi söker nu

$$\Pr(X < 240) = \Pr\left(Z < \frac{240 - 266}{16} = -1.625\right) = 1 - \Pr(Z < 1.625) \approx \mathbf{0.052}$$

Vi tolkar detta som att ungefär var tjugonde graviditet är kortare än 240 dagar.

(b) Enligt Tabell 5.2.B gäller att

$$z_{0.75} = -0.6745$$

vilket innebär att den första kvartilen ges av

$$q_1 = 266 - 0.6745 \cdot 16 = \mathbf{255.2}$$

Vi tolkar detta som att var fjärde graviditet är kortare än 255 dagar.

(c) Låt

$$Y = \text{Antal gravida kvinnor i urvalet som går över tiden}$$

Sannolikheten att en slumpmässigt vald gravid kvinna går över tiden är 0.5. En gravid kvinna går antingen över tiden eller gör det inte. Vi utgår från att gravida kvinnor går över tiden (eller gör det inte) oberoende av varandra. Vidare gäller att vår slumpvariabel räknar antalet gravida kvinnor i urvalet som går över tiden vilket innebär att Y är $Bi(12, 0.5)$. Vi söker

$$\Pr(Y \geq 7) = 1 - \Pr(Y \leq 6) = 1 - 0.6128 = \mathbf{0.3872}$$

(d) Här söker vi en betingad sannolikhet. Eftersom händelsen $\{X > 274\}$ ingår i händelsen $\{X > 270\}$ följer att

$$\Pr(\{X > 274\} \cap \{X > 270\}) = \Pr(X > 274)$$

Formeln för betingad sannolikhet ger därmed att

$$\begin{aligned} \Pr(X > 274 \mid X > 270) &= \frac{\Pr(X > 274)}{\Pr(X > 270)} = \frac{\Pr\left(Z > \frac{274-266}{16}\right)}{\Pr\left(Z > \frac{270-266}{16}\right)} = \\ &= \frac{\Pr(Z > 0.5)}{\Pr(Z > 0.25)} = \frac{1 - 0.6915}{1 - 0.5987} \approx \mathbf{0.77} \end{aligned}$$

Uppsala universitet
Statistiska institutionen

A4 & A8

2014-10-30

TENTAMENSSKRIVNING PÅ KURSERNA

Grundläggande statistik A4, 15 hp
Statistik för ekonomer A8, 15 hp

UPPLYSNINGAR

A. Tillåtna hjälpmedel:
Kursspecifik formelsamling (utan anteckningar)
Språklexikon
Miniräknare

B. **Skrivtid: 14.00-19.00** Skrivningen omfattar **6** uppgifter, om sammanlagt 100 poäng.

C. För varje uppgift anges den maximala poäng som kan erhållas. Om en uppgift är uppdelad på deluppgifter anges den maximala poängen för varje deluppgift. Ibland kan inte deluppgifterna bedömas oberoende av varandra, vilket kan innebära att poäng inte utdelas på en senare uppgift om inte tidigare deluppgift lösts på ett i princip riktigt sätt. Dock gäller att utdelad poäng för varje deluppgift aldrig kan vara negativ.

D. Om Du känner Dig osäker på någonting (skrivningens genomförande, någon formulering i en uppgift, om något hjälpmedel är otillåtet), fråga då jourhavande skrivningsvakt eller den skrivningsansvariga läraren.

E. Efter skrivningens slut får Du behålla sidorna med frågeställningarna (de ska ej lämnas in!). Preliminära lösningar anslås på Pingpong.

UPPMANINGAR

A. Följ noga de anvisningar som finns på skrivningsförsättsbladet.

B. Redovisa Dina lösningar i en form som gör det lätt att följa Din tankegång! (Det dunkelt uttryckta förutsätts av rättaren vara dunkelt tänkt). Motivera alla väsentliga steg i lösningen. Ange alla antaganden Du gör och alla förutsättningar Du utnyttjar.

C. Vid konfidensintervall måste Du dessutom ange vad intervallet avser att täcka samt teckna intervallet i symbolform innan de numeriska uppgifterna insätts. Verbal slutsats av det framräknade intervallet krävs för full poäng.

D. Vid signifikansanalys måste Du utöver vad som sägs i punkt B ovan ange vad hypotesprövningen avser att testa, hypoteserna i symbolform (då så är möjligt), signifikansnivå, testfunktion (inklusive antal frihetsgrader då detta är aktuellt) både i symbolform och med numeriska uppgifter, beslutsregel, resultat samt verbal slutsats.

E. Vid standardvägning ska metod anges och beräkningarna ska följas av en verbal slutsats för full poäng.

Lycka till!

Uppgift 1

Följande datamaterial är från USA och beskriver BMI (*Body Mass Index*) för 40 slumpmässigt utvalda män från en viss population.

Tabell 1.1 BMI, män

23,8	23,2	24,6	26,2	23,5	24,5	21,5	31,4	26,4	22,7	27,8
28,1	25,2	23,3	31,9	33,1	33,2	26,7	26,6	19,9	27,1	23,4
27,0	21,6	30,9	28,3	25,5	24,6	23,8	27,4	28,7	26,2	26,4
32,1	19,6	20,7	26,3	26,9	25,6	24,2				

Tabell 1.2 BMI, män (sorterade data)

19,6	19,9	20,7	21,5	21,6	22,7	23,2	23,3	23,4	23,5	23,8
23,8	24,2	24,5	24,6	24,6	25,2	25,5	25,6	26,2	26,2	26,3
26,4	26,4	26,6	26,7	26,9	27,0	27,1	27,4	27,8	28,1	28,3
28,7	30,9	31,4	31,9	32,1	33,1	33,2				

X: BMI Beräkningshjälp $\sum_{i=1}^{40} x_i = 1039{,}9$ $\sum_{i=1}^{40} x_i^2 = 27493{,}8$

(6) **A** Beräkna standardavvikelse och medelfel (*standard error*) för BMI. Vilket av dessa två mått är lämpligast att använda som ett deskriptivt mått för spridningen? Motivera ditt svar kortfattat.

(10) **B** Illustrera fördelningen med ett lådagram (*box-plot*).

(8) **C** Gränsen för övervikt sägs vara vid ett BMI på 25 eller högre. Beräkna ett 90% konfidensintervall för andelen överviktiga män i den aktuella populationen.

(6) **D** I en rapport baserad på denna studie angavs den statistiska felmarginalen för andelen överviktiga män till 0,1802. Vilken konfidensnivå motsvarar det?

Uppgift 2

En viss typ av förkylning för med sig vissa symptom. Det har visat sig att den som råkar ut för denna förkylning får feber med 35% sannolikhet och halsont med 80% sannolikhet. Sannolikheten att den som fått förkylningen råkar ut för båda symptomen är 18%.

(4) **A** Beräkna sannolikheten att en person som råkar ut för denna förkylning varken får feber eller halsont.

(4) **B** En person som fått förkylningen har halsont. Beräkna sannolikheten att personen dessutom har feber.

Uppgift 3

De årliga intäkterna från en viss del av statsbudgeten kan beskrivas av en normalfördelning med genomsnittet 140 miljoner kronor och standardavvikelsen 20 miljoner kronor.

(3) **A** Vad är sannolikheten att intäkterna ett år överstiger 160 miljoner kronor?

(4) **B** Vad är sannolikheten att intäkterna överstiger 120 miljoner kronor fyra år i rad? Intäkterna under dessa år kan anses vara oberoende av varandra.

(3) **C** Bilda ett symmetriskt intervall kring genomsnittet som med 99% sannolikhet kommer att innehålla årets intäkter.

Uppgift 4

Inom ett pedagogiskt projekt på universitet ville man analysera ett eventuellt samband mellan hur studenterna uppfattar kursen och studieresultatet för studenten. I en pilotstudie på Statistiska institutionen valdes 10 studenter slumpmässigt. En av institutionens doktorander samlade i slutet av kursen in dessa studenters helhetsomdömen om kursen med försäkran om att ej avslöja resultatet före det att kursbetygen var inrapporterade. Omdömena gavs som värden på en 10-gradig skala (Från 1=Mycket dålig till 10=Mycket bra). Tabellen nedan visar studenternas bedömning av kursen samt dessa studenters tentamensresultat (antal poäng 1-20).

Student	Omdöme	Tentamens resultat
1	9	20
2	5	3
3	2	7
4	8	5
5	8	18
6	4	15
7	5	11
8	7	20
9	8	19
10	3	8

(2) **A** Ange datanivå (skalnivå) för de två variablerna ”omdöme” och ”tentamensresultat”.

(10) **B** Beräkna och tolka korrelationen mellan de två variablerna.

Uppgift 5

Skolverket ville före midsommar 2014 ha din hjälp med att snabbt utröna om kommunala grundskolor har förbättrad lärartäthet 2013 jämfört med 2012. Eftersom statistik för lärartäthet ännu inte hade inrapporterats till Skolverket ringde du upp 13 kommuner vilka valts ut med ett obundet slumpmässigt urval från Sveriges 290 kommuner. Du bad att få ta del av tillgängliga uppgifter och kunde sedan för dessa kommuner beräkna lärartätheten (som definieras som antal lärare per 100 elever). Med hjälp av inrapporterad data från föregående år kunde du sedan konstruera Tabell 1.

Tabell 1. Lärartäthet (antal lärare/100 elever) år 2012 och 2013.

Huvudman: Kommun	**Lärartäthet 2012**	**Lärartäthet 2013**
Hagfors	12,3	12,7
Sollentuna	12,4	12,3
Gnosjö	9,5	10,0
Knivsta	14,1	14,2
Gnesta	10,3	11,7
Kungsbacka	12,2	12,6
Vadstena	10,8	11,3
Kinda	12,1	12,3
Hedemora	11,5	11,5
Lilla Edet	13,9	14,5
Grums	11,1	10,8
Borlänge	12,5	11,4
Ljusnarsberg	11,8	11,3

Källa: Skolverket

Vi vet från tidigare år, baserat på statistik för samtliga 290 kommuner, att variabeln lärartäthet är approximativt normalfördelad (se Figur 1 på nästa sida). Vi antar därför att lärartäthet är approximativt normalfördelad även år 2013.

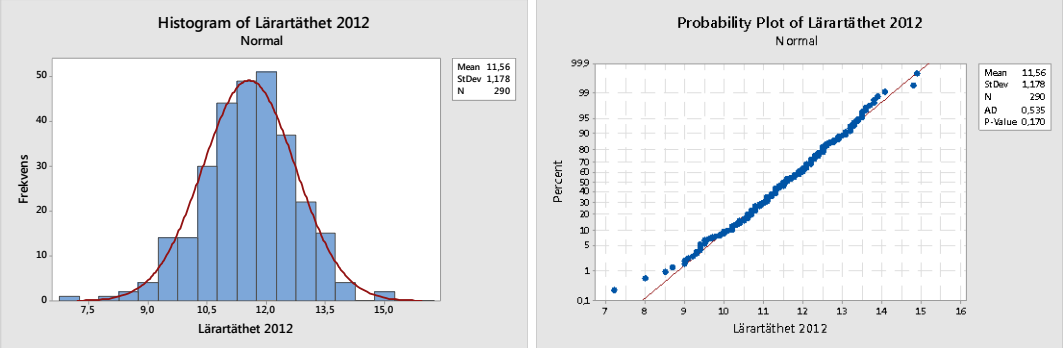
(14) **A** Avgör med hjälp av klassisk hypotesprövning om den genomsnittliga lärartätheten i Sveriges kommuner har ökat från 2012 till 2013. Använd signifikansnivån 10 %. Var noga med att redogöra för testförfarandets alla steg.

(2) **B** I oktober 2014, när statistik från alla kommuner var inrapporterad, visade det sig att den genomsnittliga ökningen i lärartäthet var 0,091 tjänster per 100 elever. Har vi i testet ovan gjort ett felbeslut? Om så är fallet, vad kallas detta fel?

(7) **C** Beräkna testets styrka givet att den genomsnittliga ökningen i lärartäthet i populationen faktiskt var 0,091. Tolka ditt framräknade resultat med ord och kommentera värdet på den framräknade styrkan. Vilken slutsats kan du i efterhand göra gällande lämpligheten av din undersökning?

(Utan ett statistikprogram kan du inte att beräkna styrkan exakt. För att möjliggöra en approximation får du därför lov att i denna deluppgift anta att populations-standardavvikelsen är känd och lika med standardavvikelsen i stickprovet, dvs $s = \sigma$. *Ledning: Tänk på att du nu har en ny testfunktion med en ny beslutsregel.)*

(3) **D** Utgå från situationen i uppgift C. Beräkna testets styrka om stickprovsstorleken utökas med ytterligare 14 observationer.



Figur 1. Histogram samt *probability plot* för lärartäthet år 2012 i Sveriges 290 kommuner. *Källa: Skolverket*

Uppgift 6

(14) Rosén et al. (2014) undersöker i sin artikel ”*Priority setting in Swedish health care: Are the politicians ready?”* om det finns något samband mellan olika aktörer inom sjukvården och uppfattning om resurstilldelning. För att få svar på sin frågeställning skickade artikelförfattarna ut en enkät till ett slumpmässigt urval av politiker, administratörer och läkare i Region Skåne. Enkäten innehöll flera frågor om resurser och kvalitet. En utav frågorna löd:
”Do you think today’s health care resources are sufficient to meet all the health care needs?”

De insamlade svaren för denna fråga (exklusive respondenter som svarat ”Vet ej”) redovisas i Tabell 1.

Använd *p*-värdesmetoden för att undersöka om det finns ett samband mellan aktör och uppfattning om dagens resurser är tillräckliga. Använd en signifikansnivå på 5%. Var noga med att redogöra för testförfarandets alla steg.

Tabell 1.
”Do you think today’s health care resources are sufficient to meet all the health care needs?”
(Siffror inom parentes indikerar antal)

	Politiker	Adminstratörer	Läkare
Ja	23 % (36)	13 % (26)	10 % (119)
Nej	77 % (119)	87 % (174)	90 % (1039)
	100 % (155)	100 % (200)	100% (1158)

Källa: Rosén et al. (2014)

Preliminära lösningar till tentamensskrivning på kurserna
Grundläggande statistik, A4 och Statistik för ekonomer, A8
2014-10-30 rev 2015-09-25/LH, RP

Uppgift 1

Variable	N	Mean	SE Mean	StDev	Min	Q1	Median	Q3	Max
BMI_M	40	25,998	0,542	3,431	19,600	23,575	26,2	27,700	33,200

A Standardavvikelse: $s = \sqrt{\dfrac{\sum x_i^2 - \left(\sum x_i\right)^2 / n}{n-1}} = \sqrt{\dfrac{27493{,}8 - 1039{,}9^2 / 40}{39}} \approx 3{,}431$

Medelfel (*Standard Error*): $SE = \hat{\sigma}_{\bar{x}} = \dfrac{s}{\sqrt{n}} = \dfrac{3{,}431}{\sqrt{40}} \approx 0{,}542$

För att beskriva spridningen i ett datamaterial används standardavvikelsen då värdet på medelfelet i stor grad speglar stickprovsstorleken.

B För att konstruera ett lådagram behövs median samt kvartiler.
Medianen är medelvärdet av de två mittersta observationerna då observationerna är ordnade i storleksordning: dvs medelvärdet av observation nr 20 ($x_{20} = 26{,}2$) och observation nr 21 ($x_{21} = 26{,}2$) $\Rightarrow$ $Md = \dfrac{x_{20} + x_{21}}{2} = \dfrac{26{,}2 + 26{,}2}{2} = 26{,}2$

Första kvartilen: q_1: Observation nr. $\dfrac{1}{4}(n+1) = \dfrac{1}{4}(41) = 10{,}25$

$$\left.\begin{matrix} x_{10} = 23{,}5 \\ x_{11} = 23{,}8 \end{matrix}\right\rangle \qquad q_1 = 23{,}5 + 0{,}25(23{,}8 - 23{,}5) = 23{,}575 \approx 23{,}6$$

Tredje kvartilen q_3: Observation nr. $\dfrac{3}{4}(n+1) = \dfrac{3}{4}(41) = 30{,}75$

$$\left.\begin{matrix} x_{30} = 27{,}4 \\ x_{31} = 27{,}8 \end{matrix}\right\rangle \qquad q_3 = 27{,}4 + 0{,}75(27{,}8 - 27{,}4) = 27{,}7$$

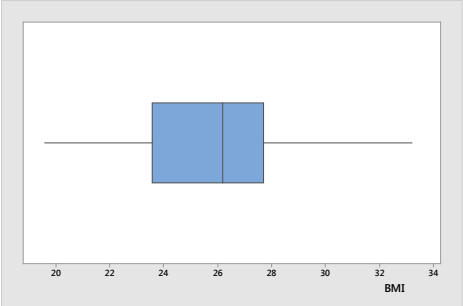
Eventuella extremvärden: Kvartilavståndet fås som: $q_3 - q_1 = 27{,}7 - 23{,}6 = 4{,}1$

Gränser för extremvärden: Nedre gräns: $q_1 - 1{,}5(q_3 - q_1) = 23{,}6 - 1{,}5 \cdot 4{,}1 = 17{,}45$
Övre gräns: $q_3 + 1{,}5(q_3 - q_1) = 27{,}7 + 1{,}5 \cdot 4{,}1 = 33{,}85$

Värden lägre än den nedre gränsen eller högre än den övre gränsen klassas som extremvärden (*uteliggare*). Då vi inte har något BMI under 17,45 eller över 33,85 har vi har således inga extremvärden.

Lådagrammet kan då ritas, där lådan begränsas av första och tredje kvartilen och medianen markeras i lådan. Morrhåren dras då ut till minsta värdet x_{min}= 19,6 respektive högsta värde x_{max}=33,2

Figur 1: *BMI för 40 slumpmässigt utvalda män*



C p = andelen överviktiga män (med BMI ≥ 25) i en viss population

Av de 40 männen i stickprovet hade 24 stycken ett BMI på 25 eller högre.

$n = 40$ (stickprovsstorlek) Stickprovsandelen: $\hat{p} = \frac{24}{40} = 0{,}60$

90% konfidensintervall för andelen överviktiga män dvs för p

$\alpha = 0{,}10$ $z_{\alpha/2} = z_{0{,}05} = 1{,}6449$

$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}$ ÄK ≈ 1 antages (mycket stor population)

Förutsättningar: Slumpmässigt urval av de 40 männen ur populationen. Stort stickprov. $n \cdot \hat{p}(1-\hat{p}) > 5$ Kontroll: $n \cdot \hat{p}(1-\hat{p}) = 40 \cdot 0{,}60 \cdot 0{,}40 = 9{,}6 > 5$ OK!

KI blir: $0{,}60 \pm 1{,}6449\sqrt{\frac{0{,}60 \cdot 0{,}40}{40}}$

$0{,}60 \pm 0{,}1274$ ung. $0{,}60 \pm 0{,}13$

Med 90% säkerhet innefattar intervallet från 47% till 73% andelen överviktiga män (med BMI ≥ 25) i denna population

D Statistiska felmarginalen (vid KI för p): $z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0{,}1802$ där $z = z_{\alpha/2}$

$$\rightarrow z = \frac{0{,}1802}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} = \frac{0{,}1802}{\sqrt{\frac{0{,}6 \cdot 0{,}4}{40}}} \approx 2{,}33$$

Tabell 5.2.A ger $P(z < 2{,}33) = 0{,}9901 \approx 0{,}99$ (alternativt använd tabell 5.2.B)

så $\alpha/2 = P(z > 2{,}33) = 1 - P(z < 2{,}33) = 1 - 0{,}99 = 0{,}01$ $\alpha = 2 \cdot 0{,}01 = 0{,}02$

Konfidensnivå: $(1-\alpha) \cdot 100\% = (1 - 0{,}02) \cdot 100\% = 98\%$

Konfidensnivå är ca 98%

Uppgift 2

Låt F = Feber H = Halsont

$\Pr(F) = 0{,}35$ $\Pr(H) = 0{,}8$ $\Pr(F \cap H) = 0{,}18$

Additionssatsen ger att sannolikheten för minst ett av symptomen blir:

$\Pr(F \cup H) = \Pr(F) + \Pr(H) - \Pr(F \cap H) = 0{,}35 + 0{,}8 - 0{,}18 = 0{,}97$

Den sökta sannolikheten är komplementet till detta dvs sannolikheten att en förkyld varken får feber eller halsont blir:

$\Pr(sökt) = \Pr(\overline{F \cup H}) = 1 - 0{,}97$ = **0,03**

Alternativ lösning

Fyll i de givna sannolikheterna i en fyrfältstabell:

	F	$\overline{F}$	
H	0,18		0,80
$\overline{H}$			
	0,35		1,00

Komplettera tabellen med resterande sannolikheter så att marginalsannolikheterna stämmer.

	F	$\overline{F}$	
H	0,18	0,62	0,80
$\overline{H}$	0,17	**0,03**	0,20
	0,35	0,65	1,00

Den sökta sannolikheten fås direkt ur tabellen: $\Pr(sökt) = \Pr(\overline{F} \cap \overline{H})$ = **0,03**

B $\Pr(sökt) = \Pr(F|H) = \dfrac{\Pr(F \cap H)}{\Pr(H)} = \dfrac{0{,}18}{0{,}8}$ = **0,225**

Uppgift 3

X: Intäkterna från en viss del av statsbudgeten. X är $N(\mu = 140;\ \sigma = 20)$

A $\Pr(X > 160) = \Pr\left(z > \dfrac{160-140}{20}\right) = \Pr(z > 1) = 1 - \Pr(z < 1) = 1 - 0{,}8413 =$ **0,1587**

B $\Pr(X > 120) = \Pr\left(z > \dfrac{120-140}{20}\right) = \Pr(z > -1) = \Pr(z < 1) = 0{,}8413$ (tabell)

Y: Antal år (av fyra) då intäkterna överstiger 120 miljoner kronor.

Y är $Bi\ (n = 4;\ p = 0{,}8413)$ $\Pr(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$

$$\Pr(Y = 4) = \binom{4}{4} \cdot 0{,}8413^4 \cdot (1 - 0{,}8413)^0 = 1 \cdot 0{,}8413^4 \cdot 1 = 0{,}8413^4 \approx 0{,}50$$

Sannolikheten att intäkterna överstiger 120 miljoner fyra år i rad är ca 50%

C Med 99% sannolikhet kommer årets intäkter att hamna i intervallet: $\mu \pm z_{\alpha/2} \cdot \sigma$

där α=1%=0,01 och $\alpha/2$=0,005 $z_{0{,}005} = 2{,}5758$ enligt tabell 5.2.B

dvs $140 \pm 2{,}5758 \cdot 20$ $140 \pm 51{,}5$

Med 99% sannolikhet kommer årets intäkter att hamna i intervallet 88,5 till 191,5

Uppgift 4

A Omdöme: ordinalskala Tentamensresultat: kvotskala

B Då variabeln ”Omdöme” är på ordinalnivå så ska Spearmans rangkorrelations- koefficient (r_S) beräknas. Beräkningen baseras på rangerna (ordningstalen). Medelranger vid ”ties”.

	x	y	R_x	R_y	R_xR_y	R_x^2	R_y^2
1	9	20	10	9,5	95,0	100	90,25
2	5	3	4,5	1	4,5	20,25	1
3	2	7	1	3	3,0	1	9
4	8	5	8	2	16,0	64	4
5	8	18	8	7	56,0	64	49
6	4	15	3	6	18,0	9	36
7	5	11	4,5	5	22,5	20,25	25
8	7	20	6	9,5	57,0	36	90,25
9	8	19	8	8	64,0	64	64
10	3	8	2	4	8,0	4	16
Summa			**55**	**55**	**344**	**382,5**	**384,5**

$$r_S = \frac{n\sum R_xR_y - \sum R_x \sum R_y}{\sqrt{(n\sum R_x^2 - (\sum R_x)^2)(n\sum R_y^2 - (\sum R_y)^2)}} = \frac{10 \cdot 344 - 55 \cdot 55}{\sqrt{(10 \cdot 382{,}5 - 55^2)(10 \cdot 384{,}5 - 55^2)}} \approx \mathbf{0{,}512}$$

Tolkning: Korrelationskoefficienten visar på ett positivt samband mellan de två variablerna. Vi har dock ett mycket litet stickprov här. Studenter som ger höga värden på variabeln ”omdöme” har höga tentamensresultat. Låga omdömen - låga tentamenspoäng.

Notera: I detta exempel har vi ”*ties*” så koefficienten kan ej beräknas enligt formel

$$r_S = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad \text{där } d_i = R_x - R_y$$

Uppgift 5

A.

Frågeställning

Vi vill veta om kommunala grundskolor har förbättrad lärartäthet 2013 jämfört med 2012.

Hypoteser

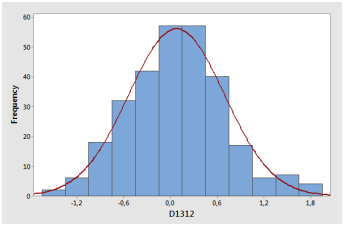
Frågeställningen leder fram till hypoteserna

$$H_0 : \mu_{2013} - \mu_{2012} = \mu_0 = 0,$$
$$H_1 : \mu_{2013} - \mu_{2012} = \mu_0 > 0,$$

där μ_{2013} anger genomsnittlig lärartäthet i alla kommuner 2013 och μ_{2012} anger genomsnittlig lärartäthet i alla kommuner 2012.

Förutsättningar

Vi drar ett *obundet slumpmässigt urval (OSU)* bestående av $n = 13$ kommuner och låter $X_{2013,i}$ och $X_{2012,i}$ beteckna lärartätheten respektive år i en slumpmässig vald kommun, i. *Parvisa observationer* innebär att variabeln av intresse är $D_i = X_{2013,i} - X_{2012,i}$, som enligt nollhypotesen har väntevärde μ_d och standardavvikelse σ_d. *Standardavvikelsen i populationen* är dock *okänd* och vi måste således använda s_d för att skatta σ_d. Eftersom *stickprovsstorleken är liten* måste vi *anta att D_i är normalfördelad.* Visserligen har data från tidigare år visat att $X_{2012,i}$ är approximativt normalfördelad (vilket gör att vi antar att även $X_{2013,i}$ är approximativt normalfördelad), men eftersom $X_{2012,i}$ och $X_{2013,i}$ inte är oberoende innebär det inte nödvändigtvis att även D_i är approximativt normalfördelad. En lösning för att ge stöd för vårt antagande om D_i:s normalitet är naturligtvis att titta på differenser från tidigare år. Dock hade uppgiften i så fall i för stor utsträckning hjälpt er med en viktig del i uppgiften, nämligen att identifiera att analysen måste baseras på differenserna. Bara för att övertyga er om en approximativ normalitet redovisar jag en graf över differenserna för alla kommuner baserat på data från 2013 och 2012.



Testfunktion

Förutsättningarna innebär att vi som testfunktion använder oss av

$$t = \frac{\bar{D} - \mu_0}{s/\sqrt{n}},$$

där $\bar{D} = (1/n)\sum_{i=1}^{n} D_i$. Enligt nollhypotesen är $\mu_0 = 0$ och om nollhypotesen är sann så är testfunktionens fördelning en t-fördelning med $n-1$ frihetsgrader.

Beslutsregel

Vi bestämmer oss för signifikansnivån 10%, dvs vi sätter $\alpha = 0,1$. Testet är ensidigt och enligt mothypotesen finns ett kritiskt område enbart i höger svans. Tabell 5.3 ges oss att den kritiska punkten är $t_{krit} = 1,356$. Beslutsregeln blir att vi förkastar H_0 om $t_{obs} > 1,356$.

Datainsamling

Data samlades in med OSU och presenteras i Tabell 1.

Beräkning

Vi inleder med att beräkna differenserna för alla observerade parvisa lärartätheter, d_i, där skillnaden mellan 2013 och 2012 för exempelvis Hagfors är $\mu_1 = x_{2013,1} - x_{2012,1} = 12,7 - 12,4 = 0,4$. På samma vis erhålls övriga differenser: -0,1; 0,5; 0,1; 1,4; 0,4; 0,5; 0,2; 0,0; 0,6; -0,3; -1,1; -0,5. Vi beräknar de aktuella nyckelsummorna $\sum_{i=1}^{13} d_i = 4,41$ och $\sum_{i=1}^{13} d_i^2 = 4,75$.

Medeldifferensen samt standardavvikelsen för differenserna, d, är

$$\bar{d} = \frac{\sum d_i}{n} = \frac{4,41}{13} = 0,162,$$

$$s = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{\sum d_i^2 - (\sum d_i)^2/n}{n-1}} = \sqrt{\frac{4,75 - \frac{4,41^2}{13}}{13-1}} = 0,6063.$$

Insättning av värden i testfunktionen ger att

$$t_{obs} = \frac{0,162}{0,6063/\sqrt{13}} = 0,963 < 1,356 = t_{krit}$$

vilket enligt beslutsregeln innebär att vi inte förkastar H_0.

Slutsats

Vi kan inte förkasta nollhypotesen på 10% signifikansnivå. Undersökningen ger alltså inte stöd för att lärartätheten i genomsnitt har förbättrats från 2012 till 2013.

B.

Vi har gjort ett felbeslut. I det här fallet har vi begått ett Typ-II-fel, dvs vi har förkastat en mothypotes som sann. Lärartätheten har ju faktiskt ökat!

C.

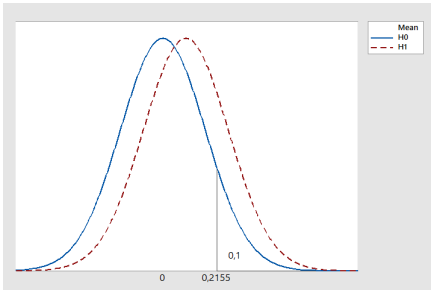
Eftersom $\sigma_d = s_d = 0,6063$ nu är känd använder vi nu testfunktionen

$$Z = \frac{\bar{D} - \mu_0}{\sigma_d/\sqrt{n}}$$

som är $Z \sim N(0,1)$ vilket ger den nya kritiska punkten $z_{krit} = 1,282$. Givet att nollhypotesen är sann så är $\mu_0 = 0$. Dessutom vet vi att $n = 13$. För att beräkna styrkan börjar vi med att lösa ut $\bar{D}$ ur den nya testfunktionen, efter insättning av värden får vi att

$$\bar{d}_{krit} = 0 + 1,282\frac{0,6063}{\sqrt{13}} = 0,216.$$

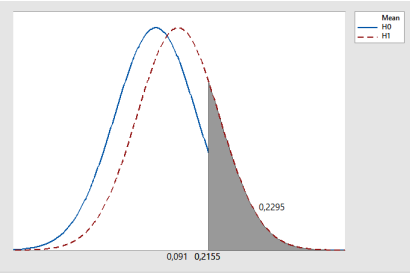
Vi har genom denna transformation omformulerat beslutsregeln till att vi förkastar nollhypotesen om $\bar{d} > 0,216 = \bar{d}_{krit}$. Denna kritiska punkt redovisas i den blå grafen nedan.



Nu visade det sig att den genomsnittliga skillnaden i lärartäthet faktiskt var $\mu = 0,091$. Detta innebär att verkligheten är den röda grafen, inte den blåa. Det innebär således att att $\bar{D}$ är $N\left(0,091, \frac{0,6063}{\sqrt{13}}\right)$ och inte att $\bar{D}$ är $N\left(0, \frac{0,6063}{\sqrt{13}}\right)$. För att beräkna styrkan beräknar vi sannolikheten för att en observation från den röda fördelningen hamnar i det kritiska området, dvs i det här fallet till höger om 0,216.

$$
\begin{aligned}
1-\beta = \Pr(Förkasta\, H_0 | H_1 sann) &= \Pr(\bar{D} > 0,216 | H_1 \, sann) \\
&= \Pr(\frac{\bar{D} - 0,091}{0,6063/\sqrt{13}} > \frac{0,216 - 0,091}{0,6063/\sqrt{13}}) \\
&= \Pr(Z > \frac{0,216 - 0,091}{0,6063/\sqrt{13}}) \\
&= \Pr(Z > 0,743) = 1 - \Pr(Z \leq 0,743) \\
&\approx 1 - 0,7704 = 0,23.
\end{aligned}
$$

Styrkan illustreras av det gråa fältet i grafen nedan.



Slutsats

Sannolikheten att ha upptäckt att det finns en skillnad när vi antog att nollhypotesen var sann var 0,23, dvs styrkan är 23%. Det innebär att om vi tänker oss att vi upprepar undersökningen utifrån samma förutsättningar (men med hypotetiska nya stickprov) så kommer vi endast i 23% av fallen förkasta den falska nollhypotesen. Således är 23% att betrakta som en liten styrka och i studier är det i regel önskvärt att styrkan är 80%. Undersökningen är därför inte lämplig att genomföra (eftersom det eventuellt är ett slöseri med resurser) då vi förmodligen inte kommer upptäcka att det faktiskt skett en ökning i lärartätheten. Stickprovet är helt enkelt för litet!

D.

Om vi ökar stickprovsstorleken till $n = 13 + 14 = 27$ individer så får vi att det kritiska området i termer av $\bar{D}$ nu ges av

$$
\bar{d}_{krit} = 0 + 1,282\frac{0,6063}{\sqrt{27}} = 0,150.
$$

Vi har genom denna transformation omformulerat beslutsregeln till att vi förkastar nollhypotesen om $\bar{d} > 0,150 = \bar{d}_{krit}$. Återigen så får vi att den sanna genomsnittliga differensen var $\mu = 0,091$. Det innebär att $\bar{D} \sim N\left(0,091, \frac{0,6063}{\sqrt{27}}\right)$. Vi utför samma beräkning som förut

$$
\begin{aligned}
1-\beta = \Pr(\textit{Förkasta } H_0 | H_1 \textit{sann}) &= \Pr(\bar{D} > 0,150 | H_1 \textit{ sann}) \\
&= \Pr(\frac{\bar{D} - 0,091}{0,6063/\sqrt{27}} > \frac{0,150 - 0,091}{0,6063/\sqrt{27}}) \\
&= \Pr(Z > \frac{0,150 - 0,091}{0,6063/\sqrt{27}}) \\
&= \Pr(Z > 0,506) = 1 - \Pr(Z \leq 0,506) \\
&\approx 1 - 0,695 = 0,31.
\end{aligned}
$$

Slutsats

Styrkan har nu ökat till 31%, men är fortfarande att betrakta som alldeles för liten. Vi behöver ett större stickprov! (En annan tanke är dock att den genomsnittliga ökningen i lärartäthet $\mu = 0,091$ kanske är att betrakta som så liten att den inte är av praktisk betydelse. Den blir därmed svår att upptäcka.).

Uppgift 6

Frågeställning

Vi vill veta om det finns ett samband mellan aktörer och uppfattning vad gäller om resurserna i hälso- och sjukvården är tillräckliga för att möta behovet i Region Skåne.

Hypoteser

Frågeställningen leder fram till hypoteserna

H_0 :Det finns inget samband mellan aktör och uppfattning,
H_1 :Det finns ett samband mellan aktör och uppfattning.

Förutsättningar

Vi drar ett *obundet slumpmässigt urval (OSU)* bestående av $n = 1513$ individer. Bägge variablerna är på nominal datanivå, en med 2 kategegorier och en med 3 kategorier, och kan presenteras i en korstabell. Vi bestämmer oss för att göra ett χ^2-test, men måste då vara uppmärksam på att *inga förväntade frekvenser får understiga 5.*

Testfunktion

Låt O beteckna observerade frekvens. Förutsättningarna innebär att vi som testfunktion använder oss av

$$\chi^2 = \sum \frac{(O-E)^2}{E}.$$

där

$$E = \frac{\text{radtotal}\times\text{kolumntotal}}{\text{radtotal}+\text{kolumntotal}}$$

är förväntade frekvens om nollhypotesen är sann. Testfunktion X^2 följer om nollhypotesen är sann en χ^2-fördelning med $(r-1)\times(k-1)$, där r = antal raderoch k = antal kolumner i korstabellen.

Beslutsregel

Vi bestämmer oss för signifikansnivån 5%, dvs vi förkastar H_0 om p-värde$<$0,05$=\alpha$.

Datainsamling

Data samlades in med OSU och presenteras i Tabell 1.

Beräkning

Vi inleder med att beräkna de förväntade frekvenser i respektive cell. I första cellen får vi att förväntad frekvens är $\frac{181\times 155}{1513} = 18,5$. Vi beräknar på motsvarande sätt förväntade frekvenser för alla celler, noterar att inga förväntade frekvenser understiger 5, och stoppar in värdena i testfunktionen

$$\chi^2_{obs} = \frac{(19-18,5)^2}{18,5} + \frac{(26-23,9)^2}{23,9} + \cdots + \frac{(1039-1019,5)^2}{1019,5} \approx 22.$$

För att räkna ut p-värdet så börjar vi med att konstatera att under nollhypotesen så vet vi att χ^2 ärχ^2_2-fördelad, eftersom $r = 2$ och $k = 3$. Då χ^2-testet kan betraktas som ensidigt i bemärkelsen att vi bara studerar högra svansen får vi att

$$p - värde = \Pr(X^2 \geq 22 | H_0 \text{ sann}) < 0,005.$$

p-värdet kan endast approximeras utifrån det faktum att $\chi^2_{obs} = 22 > 10,59$, där $10,59$ är den kritiska punkten i χ^2_2-fördelning om vi hade valt $\alpha = 0,005$. Notera att den kritiska punkten bara används här för att få en uppfattning om p-värdet. Själva beslutet att förkasta avgörs om p-värdet understiger signifikansnivån.

Slutsats

Vi förkastar på 5% signifikansnivå nollhypotesen att det inte finns ett samband mellan aktör och uppfattning. Det finns ett samband mellan beslutsfattande roll och uppfattning och vi drar slutsatsen att politiker, administratörerna och läkare har olika uppfattning om resurserna i hälso- och sjukvården i Region Skåne är tillräckliga för att möta de behov som finns.

Uppsala universitet
Statistiska institutionen

A4 & A8

2014-11-29

TENTAMENSSKRIVNING PÅ KURSERNA

Grundläggande statistik A4, 15 hp
Statistik för ekonomer A8, 15 hp

UPPLYSNINGAR

A. Tillåtna hjälpmedel:
Kursspecifik formelsamling (utan anteckningar)
Språklexikon
Miniräknare

B. **Skrivtid: 9.00-14.00** Skrivningen omfattar 6 uppgifter, om sammanlagt 100 poäng.

C. För varje uppgift anges den maximala poäng som kan erhållas. Om en uppgift är uppdelad på deluppgifter anges den maximala poängen för varje deluppgift. Ibland kan inte deluppgifterna bedömas oberoende av varandra, vilket kan innebära att poäng inte utdelas på en senare uppgift om inte tidigare deluppgift lösts på ett i princip riktigt sätt. Dock gäller att utdelad poäng för varje deluppgift aldrig kan vara negativ.

D. Om Du känner Dig osäker på någonting (skrivningens genomförande, någon formulering i en uppgift, om något hjälpmedel är otillåtet), fråga då jourhavande skrivningsvakt eller den skrivningsansvariga läraren.

E. Efter skrivningens slut får Du behålla sidorna med frågeställningarna (de ska ej lämnas in!). Preliminära lösningar anslås på Pingpong.

UPPMANINGAR

A. Följ noga de anvisningar som finns på skrivningsförsättsbladet.

B. Redovisa Dina lösningar i en form som gör det lätt att följa Din tankegång! (Det dunkelt uttryckta förutsätts av rättaren vara dunkelt tänkt). Motivera alla väsentliga steg i lösningen. Ange alla antaganden Du gör och alla förutsättningar Du utnyttjar.

C. Vid konfidensintervall måste Du dessutom ange vad intervallet avser att täcka samt teckna intervallet i symbolform innan de numeriska uppgifterna insätts. Verbal slutsats av det framräknade intervallet krävs för full poäng.

D. Vid signifikansanalys måste Du utöver vad som sägs i punkt B ovan ange vad hypotesprövningen avser att testa, hypoteserna i symbolform (då så är möjligt), signifikansnivå, testfunktion (inklusive antal frihetsgrader då detta är aktuellt) både i symbolform och med numeriska uppgifter, beslutsregel, resultat samt verbal slutsats.

E. Vid standardvägning ska metod anges och beräkningarna ska följas av en verbal slutsats för full poäng.

Lycka till!

Uppgift 1

Tabellen nedan visar BMI (*Body Mass Index*) för 160 slumpmässigt utvalda kvinnor från en mycket stor population.

BMI	Antal kvinnor
16-20*	28
20-25	56
25-30	48
30-35	15
35-50	13
Summa	**160**

* eg. $(16{,}0 \le x < 20{,}0)$

(8) **A** Beskriv frekvensfördelningen ovan med lämpligt diagram.

(12) **B** Beräkna ett 95% konfidensintervall för genomsnittligt BMI för kvinnor i denna population.

(14) **C** Gränsen för övervikt sägs vara vid ett BMI på 25 eller högre. Tio år tidigare var andelen överviktiga i denna population kvinnor 45%. Nu vill man med hjälp av stickprovsresultaten ovan avgöra huruvida andelen överviktiga förändrats.

Som beslutsregel används följande:
H_0 förkastas om 61 kvinnor eller färre, alternativt 83 kvinnor eller fler, av de tillfrågade 160 kvinnorna är överviktiga.

Ange hypoteser (nollhypotes och mothypotes) samt vilken signifikansnivå dessa förkastelsegränser motsvarar. Vilken slutsats kan man dra utifrån stickprovsresultatet ovan och med denna framräknade signifikansnivån; har andelen överviktiga i denna population kvinnor signifikant förändrats under 10-års perioden?

Beräkna *p*-värdet för testet.

(8) **D** Utgå från testet i C-uppgiften. Beräkna β (sannolikheten för fel av typ II) för detta test om den sanna andelen överviktiga i denna population kvinnor är dels 48% dels 52%.

Uppgift 2 **Mjölkkonsumtionen i Sverige**

Tabellen nedan visar hur mycket mjölk svenskarna dricker per person och år. Mätningarna startade 1939. *Källa: SCB*

År	1939	1960	1980	2000	2010	2013
Mjölk-konsumtion*	190	177	162	113	94	89

* Mjölkkonsumtion i liter per år och per person

(6) **A** Beräkna en indexserie med basår 1939 och en indexserie med basår 2000 som beskriver mjölkkonsumtionen i Sverige. Redovisa indextalen med en decimal.

(6) **B** Hur stor är den årliga genomsnittliga procentuella förändringen av mjölkkonsumtionen? Redovisa procenttalen med två decimaler.

1. Under perioden 1939-2013
2. Under perioden 2000-2013

Uppgift 3

(14) En mäklare påstår att husen i område A är mer värda än husen i område B. Ett slumpmässigt urval av taxeringsvärden för hus i område A gav följande resultat (i 1000-tals $): 85, 70, 74, 69, 88, samt 89. Ett slumpmässigt urval från område B gav följande värden (i 1000-tals $): 71, 64, 68, 73, 81, 89 samt 72.

Avgör med lämpligt hypotestest om taxeringvärdet i område A är högre än taxeringvärdet i område B. Taxeringsvärdet kan **ej** anses vara en normalfördelad variabel. Använd signifikansnivå 5%.

Uppgift 4

Om en kommun har ett konstant lägenhetsbestånd under en längre period kommer befolkningen att minska år efter år på grund av den s.k. utglesningen. Utglesningen innebär att det bor färre personer/lägenhet d.v.s. man ökar utrymmesstandarden. För att förhindra att befolkningen minskar måste kommunen därför ha ett visst nettotillskott av lägenheter. Följande siffror har hämtats från Solna kommun som har drygt 50 000 invånare.

År	Nettoproduktion av lägenheter	Befolknings förändring
1970	132	-1029
1971	1080	341
1972	928	-489
1973	1128	536
1974	2	-1025
1975	27	-1063
1976	-22	-1037

Källa: Solna i siffror 1977. Solna PLU.

$\sum x_i = 3275$ $\sum y_i = -3766$ $\Sigma x_i^2 = 3318609$

$\sum y_i^2 = 4957502$ $\sum x_i y_i = 375331$

(8) **A** Anpassa en enkel linjär regressionsmodell som beskriver befolkningsförändring (Y) som en funktion av nettoproduktion (X). Tolka de framräknade koefficienterna i termer av de ingående variablerna befolkningsförändring och nettoproduktion.

(4) **B** Beräkna vilken nettoproduktion (lägenheter/år) som i genomsnitt erfordrats för att folkmängden skulle varit oförändrad (förändring=0) under denna period och enligt den anpassade regressionsmodellen.

Uppgift 5

Högskoleprovet kan kategoriseras i två stycken block: 1) ett kvantitativt block och 2) ett verbalt block. Bland alla skrivande under ett år var medelpoängen och standardavvikelsen i det kvantitativa blocket 22,4 respektive 6,9 poäng. I det verbala blocket var medelpoängen 45,9 och standardavvikelsen 12,8.

(4) **A** Beräkna den förväntade totala poängen på högskoleprovet för en slumpvis vald skrivande från den stora populationen som skrev provet.

(8) **B** Standardavvikelsen för den totala poängen var 17. Hur stor var korrelationen mellan antal poäng i det kvantitativa blocket och det verbala blocket?

Uppgift 6

På Nya Zeeland är det sedan 1996 möjligt (givet visa kriterier) att i passet ersätta kvinna (K) eller man (M) med ett X. År 2008 fanns det 2 450 000 Nya Zeeländare över 15 år som hade pass. Av dessa hade 27 individer ändrat K till X, 163 hade ändrat M till X och 119 hade valt X direkt. Vi utgår från att dessa uppgifter fortfarande är aktuella.

(2) **A** Anta att du arbetar som passkontrollant på Arlanda och ser ett X i ett Nya Zeeländskt pass, vad är sannolikheten att den personen tidigare kategoriserades som man (M) i passet?

(6) **B** Anta att du drar ett obundet slumpmässigt urval från populationen Nya Zeeländare med pass. Hur stort måste urvalet vara om du vill att sannolikheten, för att minst en individ i urvalet har ett X i passet, ska vara 20%? (Om du i sista steget inte klarar av att lösa uppgiften analytiskt så får du lov att pröva dig fram numeriskt till det rätta svaret.)

Hämtat från Veale J. F (2008). The Prevalence of Transsexualism Among New Zealand Passport Holders. The Australian and New Zealand Journal of Psychiatry.

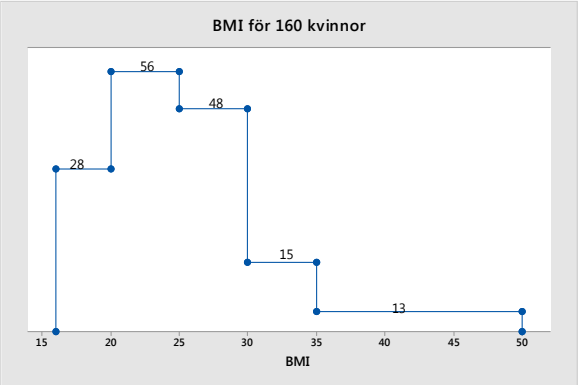
Preliminära lösningar till tentamensskrivning på kurserna
Grundläggande statistik, A4 och Statistik för ekonomer, A8
2014-11-29/LH, RP

Uppgift 1

Vi har här en klassindelad variabel med varierande klassbredder. →
Histogram för att illustrera frekvensfördelningen, med ytor proportionella mot frekvenser.

BMI	f_i	w_i (klassbredd)	Hjälpaxel* $\frac{f_i}{w_i} \cdot 5$
16-20	28	4	35
20-25	56	5	56
25-30	48	5	48
30-35	15	5	15
35-50	13	15	4,33
Summa	**160**		

* Hjälpaxel ger höjdförhållandet mellan de olika rektanglarna i histogrammet: Kvoten f_i / w_i kan multipliceras med valfri konstant (här har kvoten multiplicerats med siffran 5).



BMI	Klassmitt x_i	f_i	$f_i x_i$	$f_i x_i^2$
16-20	18	28	504,0	9072,0
20-25	22,5	56	1260,0	28350,0
25-30	27,5	48	1320,0	36300,0
30-35	32,5	15	487,5	15843,8
35-50	42,5	13	552,5	23481,3
		160	**4124,0**	**113047,1**

B Medelvärde: $\bar{x} = \frac{\sum f_i x_i}{n} = \frac{4124}{160} = 25{,}775$

Standardavvikelse:

$$s = \sqrt{\frac{\sum f_i x_i^2 - \left(\sum f_i x_i\right)^2 / n}{n-1}} = \sqrt{\frac{113047 - 4124^2/160}{159}} \approx \sqrt{42{,}45849} \approx 6{,}5160$$

95% konfidensintervall för μ (genomsnittligt BMI i hela populationen) α=0,05

$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$ $\quad z_{\alpha/2} = z_{0{,}025} = 1{,}96$ ÄK≈1 antages ty mkt stor population

Förutsättningar: Slumpmässigt urval av de 160 kvinnorna. Stickprovet är stort. Tumregel: n>30; vilket är uppfyllt.

Så KI blir: $25{,}78 \pm 1{,}96\frac{6{,}516}{\sqrt{160}}$ $\quad 25{,}78 \pm 1{,}01$

Lägre gräns: 24,77 övre gräns: 26,79 $24{,}8 < \mu < 26{,}8$

Med 95% konfidens innefattar intervallet 24,8 till 26,8 genomsnittligt BMI för kvinnor i den bakomliggande populationen.

C p : Andel överviktiga dvs andel kvinnor med ett BMI på 25 eller högre

n=160 Signifikansnivå: α Dubbelsidigt test

$H_0 : p = 0{,}45$

$H_1 : p \neq 0{,}45$

Som beslutsregel används följande:

H_0 förkastas om $\hat{p} < \frac{61}{160} = 0{,}38125$ eller om $\hat{p} > \frac{83}{160} = 0{,}51875$

Testfunktion: $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ där $\sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0{,}45 \cdot 0{,}55}{160}} \approx 0{,}03933$

Nollhypotesen förkastas då

$z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} < -z_{\alpha/2}$ eller då $z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > z_{\alpha/2}$

Används den högra förkastelsegränsen fås följande:

$$z_{\alpha/2} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0{,}51875 - 0{,}45}{0{,}03933} \approx 1{,}75$$

$\alpha/2 = 1 - P(z < 1{,}75) = 1 - 0{,}9599 = 0{,}0401$ (från tabell)

$\alpha = 2 \cdot 0{,}0401 = 0{,}0802$

Dessa förkastelsegränser motsvarar således ett test med signifikansnivå ca 8%

Förutsättningar: Slumpmässigt urval av de 160 kvinnorna. Stickprovet är stort.
Tumregel: $np_0(1-p_0)>5$. Kontroll: $160\cdot 0{,}45(0{,}55_0)=39{,}6>5$ OK!

I vårt stickprov hade **76** kvinnor ett BMI på 25+. Vi förkastar nollhypotesen om antalet kvinnor med BMI på 25 eller högre är ≤61 alt. ≥83 .
Resultat: 61<76< 83 dvs ett icke signifikant resultat; H_0 förkastas ej.

Resultatet ger ej belägg för att andelen överviktiga i denna population har förändrats under 10-årsperioden.

Beräkning av *p*-värde:

$$z=\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}=\frac{0{,}475-0{,}45}{\sqrt{\frac{0{,}45(1-0{,}45)}{160}}}=0{,}6356\approx 0{,}64$$

$$p-värde=2\cdot\Pr(z>0{,}64)=2\cdot\{1-\Pr(z<0{,}64)\}=2\cdot\{1-0{,}7389\}=2\cdot 0{,}2611=0{,}5222$$

***p-värdet* ca 52%**

D Beräkna $\beta=\Pr(\textit{fel typ II})$ Z-transformering: $z=\dfrac{\hat{p}-p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}}$

$$\beta=\Pr(0{,}38125<\hat{p}<0{,}51875\,|\,p_1=0{,}48)=$$

$$=\Pr\left(\frac{0{,}38125-0{,}48}{\sqrt{\frac{0{,}48\cdot 0{,}52}{160}}}<z<\frac{0{,}51875-0{,}48}{\sqrt{\frac{0{,}48\cdot 0{,}52}{160}}}\right)=\Pr(-2{,}50<z<0{,}98)=$$

$$=P(z<0{,}98)-[1-P(z<2{,}50)]=0{,}8365-[1-0{,}9938]=0{,}8365-0{,}0062=0{,}8303\ (\textit{Tabell})$$

Svar: $\beta=$ ***0,8303*** (ca 83%)

$$\beta=\Pr(0{,}38125<\hat{p}<0{,}51875\,|\,p_1=0{,}52)=$$

$$=\Pr\left(\frac{0{,}38125-0{,}52}{\sqrt{\frac{0{,}52\cdot 0{,}48}{160}}}<z<\frac{0{,}51875-0{,}52}{\sqrt{\frac{0{,}52\cdot 0{,}48}{160}}}\right)=\Pr(-3{,}51<z<-0{,}03)=$$

$$=\Pr(z<3{,}51)-\Pr(z<0{,}03)=0{,}9998-0{,}5120=0{,}4878\qquad(\textit{Tabell})$$

Svar: $\beta=$ ***0,4878*** (48,78%)

Uppgift 2

År (t)	*1939*	*1960*	*1980*	*2000*	*2010*	*2013*
Mjölk-konsumtion*	190	177	162	113	94	89
Index (basår 1939)	100,0	93,2	85,3	59,5	49,5	46,8
Index (basår 2000)	168,1	156,6	143,4	100,0	83,2	78,8

A $$Index\ basår\ 1939 = \frac{Mjölkkonsumtion\ år\ t}{Mjölkkonsumtion\ 1939} \cdot 100$$

$$Index\ basår\ 2000 = \frac{Mjölkkonsumtion\ år\ t}{Mjölkkonsumtion\ 2000} \cdot 100$$

B Mjölkonsumtionen sjönk från 190 till 89 mellan år 1939-2013 (74 år)

Genomsnittlig förändring per år: $\sqrt[74]{\frac{89}{190}} = (0{,}4684)^{1/74} = 0{,}9898$

(0,9898-1)*100% =-1,02% d**vs en minskning med i genomsnitt 1,02% per år**

Mjölkonsumtionen sjönk från 113 till 89 mellan år 2000-2013 (13 år)

Genomsnittlig förändring per år: $\sqrt[13]{\frac{89}{113}} = (0{,}7876)^{1/13} = 0{,}9818$

(0,9818-1)*100% =-1,82% d**vs en minskning med i genomsnitt 1,82% per år. Minskningen har således gått lite snabbare de senare åren.**

Uppgift 3

Två oberoende stickprov, taxeringsvärdet kan ej anses var en normalfördelad variabel $\Rightarrow$ Mann-Whitneys test, Wilcoxons rangsummatest *(Wilcoxon rank sum test for independent sample)*

Låt η beteckna populationsmedianen för taxeringsvärde.

H_0: $\eta_A \leq \eta_B$ (Taxeringsvärdet är lägre i område A jämfört med område B eller taxeringsvärdena är lika i de två områdena)

H_1: $\eta_A > \eta_B$ (Taxeringsvärdet är högre i område A jämfört med område B)

Signifikansnivå $\alpha = 0.05$, enkelsidigt test

Testvariabel: R

Förutsättningar: Stickproven är dragna oberoende av varandra med OSU ur symmetriska populationer.
Taxeringsvärdet mäts på minst ordinalskalenivå.

Två oberoende stickprov om n_1 respektive n_2 observationer så att $n_1 \leq n_2$.
dvs $n_1 = n_A = 6,\ \ n_2 = n_B = 7$)

Förkastelseområde ($\alpha = 0.05$, enkelsidigt test, $n_1 = n_A = 6,\ \ n_2 = n_B = 7$):

H_0 förkastas om $R \leq R_{kritisk} = 29$ Tabell 5.6

Resultat:

	A	rang A	B	rang B
	85	10	71	5
	70	4	64	1
	74	8	68	2
	69	3	73	7
	88	11	81	9
	89	12,5	89	12,5
			72	6
Summa		48,5		42,5

Observationerna rangordnas från lägst till högst; medelvärde vid *ties*. R_1 är rangsumman i det mindre stickprovet dvs R_1=R_A=48,5

Symmetrivärdet $R_S = n_1(n_1 + n_2 + 1) - R_1 = 6(6 + 7 + 1) - 48{,}5 = 35{,}5$
Testvariabel: $R = \min(R_1, R_S) = \min(48{,}5\ ;\ 35{,}5) = 35{,}5$

$R = 35{,}5 > R_{kritisk} = 29$ Icke signifikant resultat. H_0 förkastas ej på 5%-nivån.

Testresultatet ger ej belägg för, på fem procents-nivån, att taxeringsvärdena i område A är högre än område B.

Jämför med Minitab:
Mann-Whitney Test and CI: A; B

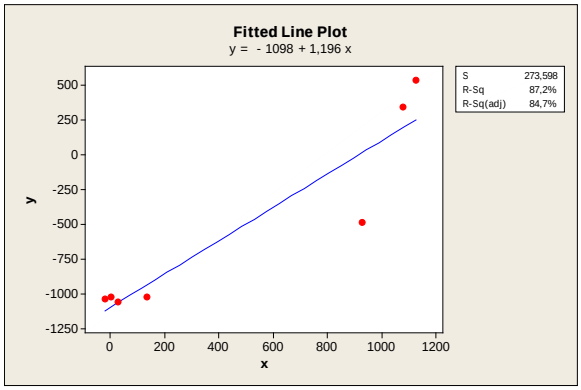
```
   N  Median
A  6   79,50
B  7   72,00

Point estimate for η1 - η2 is 4,50
96,2 Percent CI for η1 - η2 is (-4,00;17,00)
W = 48,5
Test of η1 = η2 vs η1 ≠ η2 is significant at 0,3914
```

Uppgift 4

A Enkel linjär regression $\hat{y} = a + bx$

där *Y* är befolkningsförändring (den beroende variabeln) och *X* är nettoproduktion (den förklarande variabeln).



x_i	y_i	x_i^2	y_i^2	$x_i y_i$
132	-1029	17424	1058841	-135828
1080	341	1166400	116281	368280
928	-489	861184	239121	-453792
1128	536	1272384	287296	604608
2	-1025	4	1050625	-2050
27	-1063	729	1129969	-28701
-22	-1037	484	1075369	22814
3275	**-3766**	**3318609**	**4957502**	**375331**

$\sum x_i = 3275 \quad \sum y_i = -3766 \quad \Sigma x_i^2 = 3318609 \qquad \sum y_i^2 = 4957502$

$\sum x_i y_i = 375331 \qquad n=7$

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \sum x_i \sum y_i / n}{\sum x_i^2 - (\sum x_i)^2 / n} = \frac{375331 - (3275 \cdot -3766)/7}{3318609 - 3275^2/7} \approx 1{,}20$$

$$a = \frac{\sum y_i}{n} - b\ \frac{\sum x_i}{n} = \frac{-3766}{7} - (1{,}20)\frac{3275}{7} \approx -1098$$

<u>Tolkningar (gällande i undersökningsområdet)</u>

b : 1,20
För varje ytterligare lägenhet som produceras förväntas befolkningsändringen i genomsnitt öka med 1,20 personer.
a =-1098
Ett förväntat värde på genomsnittligbefolkningsförändring då nettoproduktionen är 0 lägenheter.

B Teckna ekvationen $0 = -1098 + 1{,}20 \cdot x$ och lös ut x.

$1{,}20 \cdot x = 1098 \qquad x = \frac{1098}{1{,}20} = 915$ Svar: **En nettoproduktion på 915 lägenheter.**

Uppgift 5

a) Låt $\mu_X = 22,4$ vara medelpoängen i det kvantitativa blocket för alla skrivande under ett år och låt X beteckna antalet poäng en slumpmäsigt vald skrivande har på det kvantititativa blocket. Låt$\mu_Y = 45,9$ vara medelpoängen i det kvalitativa blocket för alla skrivande under ett år och låt Y beteckna antalet poäng en slumpmäsigt vald skrivande har på det kvalitativa blocket. Låt $Z = X + Y$ beteckna det totala antalet poäng en slumpvis vald skrivande har på högskoleprovet. Eftersom $E(Z) = E(X + Y) = E(X) + E(Y) = \mu_X + \mu_Y = 22,4 + 46,9 = 68,3$ får vi att den förväntade totala poängen på högskoleprovet för en slumpvis vald skrivande är 68,3 poäng.
b) Standardavvikelsen för den totala poängen var $\sigma_Z = 17$. Det innebär att variansen för totalpoängen kan skrivas $\sigma_Z^2 = Var(Z) = Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y) = 6,9^2 + 12,8^2 + 2Cov(X, Y) = 17^2$. Vi kan skriva om detta som $Cov(X, Y) = (17^2 - 6,9^2 - 12,8^2)/2 = 38,775$. Eftersom korrelationen är den standardiserade kovariansen får vi att

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \frac{38,775}{6,9 \cdot 12,8} = 0,439.$$

Vi konstaterar att korrelationen mellan antalet poäng i det kvantativa blocket och det kvalitativa blocket är 0,439.

Uppgift 6

a) Låt $X = 309$ beteckna antalet personer med X i passet och låt $M = 163$ beteckna antalet som tidigare haft ett M i passet. Sannolikheten vi söker ges av andelen utav X som tidigare haft M, vilket direkt erhålls genom

$$\Pr(M|X) = \frac{163}{27 + 163 + 119} = 0,527.$$

Sannolikheten att personen med X i passet tidigare kategoriserades som man är alltså 0,527.
b) Eftersom population är stor antar vi att dragningarna sker oberoende. Låt Y beteckna antalet individer i ett OSU som ett X i passet. Vi noterar att Y är binomialfördelad och för att förenkla beräkningar använder vi komplementet, dvs vi söker stickprovsstorleken givet att sannolikheten att ingen individ i stickprovet har ett X i passet ska vara 80%. Det innebär att vi ska lösa ut n ur binomafördelningen

$$\begin{aligned}
\Pr(Y = 0) &= \binom{n}{y} \pi^y (1 - \pi)^{n-y} \\
&= \binom{n}{0} \left(\frac{27 + 163 + 119}{2450000}\right)^0 \left(1 - \frac{27 + 163 + 119}{2450000}\right)^n \\
&= 1 \cdot 1 \cdot \left(1 - \frac{309}{2450000}\right)^n = 0,8.
\end{aligned}$$

Genom att pröva olika värden på n så finner vi att $n = 1770$ är en approximativ lösning för ekvationen. Det krävs således en stickprovsstorlek på ungefär 1770 individer för att sannolikheten för att minst en individ ska ha ett X passet ska vara 20%.