TENTAMENSSKRIVNING PÅ KURSERNA

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

2013-03-22

UPPLYSNINGAR

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

B. **Skrivtid: 8.00-13.00** Skrivningen omfattar 4 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

I en studie där 859 slumpmässigt utvalda svenskars kostvanor undersöktes studerades bl.a. de båda variablerna Rökare (nej=0, ja=1) och Fiber (fiberintag mätt i gram/dag). Intressant information från denna undersökning presenteras i Minitabutskriften nedan.

Variable	Rökare	Total Count	Mean	SE Mean	StDev	Minimum	Q1	Median
Fiber	0	723	20,431	0,273	7,348	3,600	15,500	19,400
	1	136	17,089	0,607	7,081	2,200	12,500	16,300

Variable	Rökare	Q3	Maximum	Mode	N for Mode
Fiber	0	24,200	57,400	21,1	9
	1	20,275	47,300	15,5	4

(3) **A** Ange de aktuella variablernas datanivå. För full poäng måste svaret motiveras.

Låt oss fortsätta analysen med att i deluppgifterna B till D nedan mer ingående studera det dagliga fiberintaget för den grupp som röker.

(6) **B** Ge ordentliga tolkningar av medelvärde, standardavvikelse och median för den aktuella gruppen.

(3) **C** Visa hur värdet för SE Mean tagits fram för den aktuella gruppen och förklara även innebörden av detta värde.

(6) **D** Ge en grafisk beskrivning av det dagliga fiberintaget för den aktuella gruppen genom att skapa ett lådagram/boxplot. För att kunna göra detta behövs lite tilläggsinformation. De tre lägsta värdena i denna grupp var 2.2, 5.4 och 6.0, medan de nio största värdena var 28.3, 29.1, 31.6, 32.6, 33.2, 33.7, 33.9, 41.9, 47.3.

Vi går nu vidare i analysen och använder den givna informationen till att utföra statistisk inferens i samband med några intressanta frågeställningar.

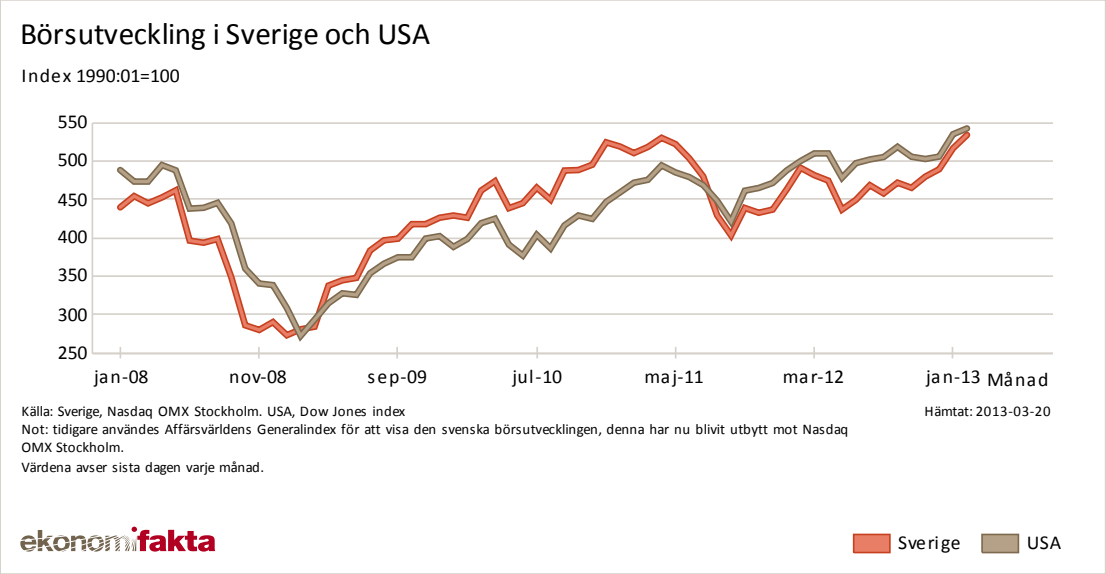
(12) **E** Är det i och med resultatet i denna undersökning statistiskt säkerställt att andelen rökare i Sverige understiger 20%. Utför en fullständig hypotesprövning enligt klassisk metod där du använder 5% signifikansnivå.

(5) **F** Beräkna *p*-värdet för testet i E-uppgiften. Ge en ordentlig tolkning av detta p-värde genom att börja med ”Om det är så att andelen rökare i Sverige är…”. Observera att tolkningen inte skall gälla huruvida nollhypotesen skall förkastas (detta är redan gjort i E-uppgiften).

(10) **G** Konstruera ett intervall som med 95% säkerhet täcker in skillnaden i genomsnittligt fiberintag mellan de båda grupperna, d.v.s. rökare vs icke-rökare i de bakomliggande populationerna. Här är det tillåtet att utan vidare kontroll anta att spridningen i de båda populationerna är samma. *Anmärkning. Vi betraktar här urvalet som två stickprov, ett taget från gruppen av rökare och ett taget från gruppen av icke-rökare*.

Uppgift 2

År 2008 skedde en kraftig nedgång på världens börser, så även i Sverige. Efter det har börsvärdet återhämtat sig vilket framgår av diagrammet nedan.



I tabellen nedan ses årliga relativa värdeförändringar på börsen i Sverige, d.v.s. den angivna värdeförändringen för exempelvis 2009 ska ses som den procentuella värdeförändringen från 31 december 2008 till 31 december 2009.

År	**2009**	**2010**	**2011**	**2012**
Värdeförändring (%)	46.7	23.0	-16.7	12.0

(5) **A** Skapa en indexserie för börsutvecklingen i Sverige där du använder 31 december 2008 som bastidpunkt.

(3) **B** Hur stor var den årliga relativa värdeförändringen i genomsnitt under perioden 31 december 2008 till 31 december 2012?

Uppgift 3

Ett slumpmässigt urval av 10 studenter ombads att i ett blindtest betygsätta smaken på två märken av glass, en glass med reducerat sockerinnehåll och en vanlig glass (d.v.s. utan reducerat sockerinnehåll). Betyg baserades på en skala från 1 (dåligt) till 10 (utmärkt). Den bifogade tabellen visar resultaten.

Student	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Reducerat socker	2	3	7	8	7	4	3	4	5	6
Vanlig	6	5	6	9	5	8	9	6	4	9

Är det i och med detta resultat statistiskt säkerställt att glasskonsumenter bland studenter i större utsträckning föredrar den vanliga glassen? För att undersöka denna frågeställning ska vi på 5% signifikansnivå utföra fullständig hypotesprövning enligt *två icke-parametriska* testmetoder.

(12) **A** *Icke-parametrisk metod 1*. Använd den icke-parametriska testmetod som bäst utnyttjar informationen (d.v.s. den av de båda metoderna som har den högsta styrkan). När du anger förutsättningar för testet ska du vara noga med att ange och kommentera vilken av förutsättningarna som inte nödvändigtvis är uppfylld.

(8) **B** *Icke-parametrisk metod 2*. Utför nu testet med den andra icke-parametriska testmetoden som kan användas för den aktuella situationen.

(7) **C** Vi betraktar nu situationen i B-uppgiften innan stickprovet var taget, d.v.s. någon analys är ännu inte utförd. Anta att det förhåller sig så att 65% av studenterna i vår population föredrar den vanliga glassen framför den med reducerat sockerinnehåll. Bestäm för denna situation risken för ett Typ2-fel i den kommande undersökningen (d.v.s. den som sedan utförs i B-uppgiften).
Ledning. Det första steget i beräkningen är att bestämma det kritiska området för det test som utförs i B-uppgiften.
Anmärkning: Vid beräkningen förutsätter vi att det i vår population inte förekommer några ties.

Uppgift 4

På en viss universitetskurs ingår sedan lång tid tillbaka två obligatoriska inlämningsuppgifter. För var och en av dessa båda uppgifter erhåller studenten 0, 1 eller 2 poäng inför en bedömning av slutbetyget på kursen. Resultaten från de båda inlämningsuppgifterna finns för detta mycket stora datamaterial sammanställda i en korstabell med relativa frekvenser (se nedan).

	Inlämning 2		
Inlämning 1	**y=0**	**y=1**	**y=2**
x=0	0.07	0.02	0.01
x=1	0.12	0.27	0.21
x=2	0.03	0.03	0.24

(3) **A** *(OBS! Denna uppgift ingår inte längre på kursen.)* Ur det stora datamaterialet väljer vi slumpmässigt en student. För denna student ser vi endast att den sammanlagda poängen för de båda inlämningsuppgifterna är två. Bestäm utifrån denna information sannolikheten att studenten erhöll åtminstone en poäng på den första inlämningsuppgiften.

(5) **B** Ur det stora datamaterialet väljer vi slumpmässigt tolv studenter. Bestäm sannolikheten att högst tio av dessa erhöll åtminstone en poäng på den första inlämningsuppgiften. För full poäng måste dina beräkningar ordentligt motiveras.

(5) **C** I en viss grupp om tolv studenter är det nio som erhöll full poäng (d.v.s. fyra poäng) på de båda inlämningsuppgifterna. Ur denna grupp väljs slumpmässigt tre studenter. Bestäm sannolikheten att åtminstone två av de valda var fullpoängare. För full poäng måste dina beräkningar ordentligt motiveras.

(7) **D** Låt oss sammanställa en tabell med relativa frekvenser för den sammanlagda poängen på de båda inlämningsuppgifterna.

Inlämning 1+2	**0**	**1**	**2**	**3**	**4**
Relativ frekvens	0.07	0.14	0.31	0.24	0.24

Ur det stora datamaterialet väljer vi slumpmässigt fyrtio studenter. Bestäm sannolikheten att dessa studenters sammanlagda poäng från de båda inlämningsuppgifterna överstiger 100.

1. Svenska folkets kostvanor.

   (a) Variabeln Rökare med värden/kategorier Rökare och Icke-rökare mäts på *nominalskala* eftersom de båda värdena inte på något objektivt sätt kan ställas i storleksordning. För variabeln Fiberintag gäller däremot att ett intag på exempelvis 10g är dubbelt så stort som ett intag på 5g vilket innebär att 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 rökare gäller att sammanfattningen fås att

   $$\begin{aligned} \overline{x} &= 17.089 \\ s &= 7.081 \\ md &= 16.3 \end{aligned}$$

   För de undersökta i gruppen rökare gäller alltså att det genomsnittliga fiberintaget (med avseende på medelvärdet) var 17.1 g. Samtliga i gruppen hade dock inte samma fiberintag utan detta avvek med i genomsnitt 7.1 g från det genomsnittliga intaget. Vidare ser vi att medianintaget var 16.3 g vilket innebär att hälften av de undersökta rökarna hade ett dagligt intag under 16.3 g och den andra halvan hade ett dagligt intag över detta värde.

   (c) Begreppet SE Mean betyder *Standard Error of the Mean* och översätts av oss till *medelfelet*. Det är alltså den skattade standardavvikelsen för medelvärdet och beräknas enligt formeln $s/\sqrt{n}$. Här beräknas den därmed som

   $$s_{\overline{x}} = \frac{s}{\sqrt{n}} = \frac{7.081}{\sqrt{136}} = 0.607$$

   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 0.61 g.

(d) för de undersökta i gruppen rökare har vi tillgång till median och kvartiler vilka ges av

$$\begin{aligned} q_1 &= 12.5 \\ md &= 16.3 \\ q_3 &= 20.275 \end{aligned}$$

Ett och ett halvt kvartilavstånd ges därmed av

$$1.5 \cdot (20.275 - 12.5) = 11.663$$

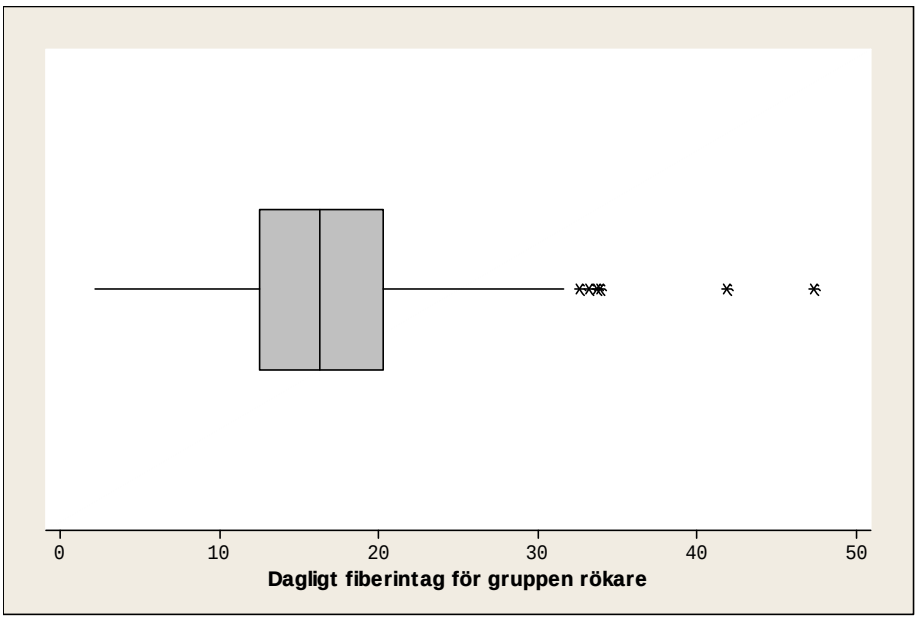
varför uteliggare är observationer under

$$12.5 - 11.663 = 0.837$$

och över

$$20.275 + 11.663 = 31.938$$

Utifrån tilläggsinformationen ser vi att vi inte har några uteliggare med små värden men att vi har inte mindre än sex uteliggare, nämligen de rökare med fiberintag på $32.6, 33.2, 33.7, 33.9, 41.9, 47.3$. Det undre morrhåret ska därmed dras till den minsta observationen, dvs 2.2, medan det övre morrhåret ska dras till värdet 31.6. Lådagrammet får således följande utseende



(e) Vi låter nu

$$p = \text{Andel rökare i den svenska befolkningen}$$

Utifrån frågeställningen formuleras hypoteserna på följande sätt

$$\begin{aligned} H_0 &: p = 0.2 \\ H_1 &: p < 0.2 \end{aligned}$$

vilka ska testas på 5% signifikansnivå. Vi förutsätter att de som ingår i undersökningen kan betraktas som ett slumpmässigt urval ur sveriges befolkning och eftersom

$$np_0\left(1-p_0\right) = 859 \cdot 0.2 \cdot 0.8 = 137.44 >> 5$$

är stickprovet med god marginal tillräckligt stort för att normalapproximation av binomialfördelningen ska vara tillåten. Eftersom sveriges befolkning kan betraktas som en mycket stor population följer att vi kan bortse från ändlighetskorrektion och 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 $H_1 : p < 0.2$ följer att nollhypotesen ska förkastas först om

$$z_{obs} < -z_{0.05} = -1.645$$

I urvalet blev andelen rökare

$$\widehat{p} = \frac{136}{859} = 0.158$$

vilket alltså ger ett tillsynes starkt stöd åt att andelen rökare i Sverige understiger 20%. Frågan är hur övertygande resultatet är? Vi sätter in i testfunktionen

$$z = \frac{0.158 - 0.2}{\sqrt{\frac{0.2 \cdot 0.8}{859}}} = -3.05$$

och eftersom

$$z_{obs} = -3.05 < -1.645 = -z_{0.05}$$

har vi hamnat i det kritiska området och därmed förkastas nollhypotesen. Det är alltså på 5% signifikansnivå statistiskt säkerställt att p, dvs andelen rökare i Sverige, understiger 20%.

(f) Vi finner via Tabell 5.2.A det sökta p-värdet till

$$p\text{-värde} = \Pr\left(Z < -3.05\right) = 0.0011$$

Om det är så att andelen rökare i Sverige är 20% är sannolikheten ca 0.1% att endast 136 (eller färre) av 859 slumpmässigt valda svenskar är rökare, dvs en så låg eller lägre andel får man i ungefär vart tusende stickprov. Det är alltså mycket ovanligt och det är förstås därför vi (på 5% signifikansnivå) förkastar nollhypotesen.

(g) Nu låter vi

$$\begin{aligned} \mu_{IR} &= \text{Genomsnittligt dagligt fiberintag för icke-rökare i Sverige} \\ \mu_R &= \text{Genomsnittligt dagligt fiberintag för rökare i Sverige} \end{aligned}$$

Vi tänker konstruera ett 95%-igt konfidensintervall för differensen $\mu_{IR} - \mu_R$. Vi förutsätter att båda stickproven är OSU dragna oberoende av varandra. Eftersom båda urvalen är stora behövs inte några vidare förutsättningar angående fördelningen för fiberintag i de båda populationerna. Vidare gäller att både rökare och icke-rökare i Sverige är är två mycket stora populationer varför ändlighetskorrektionen kan bortses från. Enligt anvisningarna kan vi utan vidare kontroll utgå från att antagandet $\sigma_{IR} = \sigma_R$ är rimligt varför vi använder konfidensintervallet

$$\overline{x}_1 - \overline{x}_2 \pm z_{\alpha/2} \cdot \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Minitabutskriften ger oss nödvändig information från stickproven.

$$\begin{aligned} \overline{x}_{IR} &= 20.431, \quad s_{IR} = 7.348, \quad n_{IR} = 723 \\ \overline{x}_R &= 17.089, \quad s_R = 7.081, \quad n_R = 136 \end{aligned}$$

Den sammanslagna (polade) variansen blir

$$s_p^2 = \frac{722 \cdot 7.348^2 + 135 \cdot 7.081^2}{857} = 53.386$$

Vi kan nu beräkna konfidensintervallet till

$$20.431 - 17.089 \pm 1.96 \cdot \sqrt{53.386 \left(\frac{1}{723} + \frac{1}{136}\right)}$$

eller mer kompakt

$$3.342 \pm 1.3385$$

Skrivet som ett intervall blir det

$$2.0 \leq \mu_{IR} - \mu_R \leq 4.7$$

Med 95% säkerhet befinner sig det genomsnittliga dagliga fiberintaget för icke-rökare i Sverige någonstans mellan 2 till 4.7 gram över motsvarande värde för rökare i Sverige.

2. Börsutveckling i Sverige.

   (a) Eftersom år 2008 ska vara basår får det året automatiskt värdet 100. Enligt ränta-på-ränta-principen följer övriga indexvärden på följande sätt.

2009	$100 \cdot 1.467 = 146.70$
2010	$146.70 \cdot 1.23 = 180.44$
2011	$180.44 \cdot 0.833 = 150.31$
2012	$150.31 \cdot 1.12 = 168.34$

   så vår indexserie blir nu

År	**2008**	**2009**	**2010**	**2011**	**2012**
Index	100.0	146.7	180.4	150.3	168.3

   (b) Den procentuella uppgången från 31 december 2008 till 31 december 2012 var 68.3% vilket innebär att den genomsnittliga årliga ökningen var

$$g = 1.6834^{1/4} = 1.139$$

   dvs 13.9%.

3. Icke-parametriska test i fallet med parvisa observationer. Här har vi två testmetoder att tillgå, dels Wilcoxons teckenrangtest och dels teckentest. Wilcoxons teckenrangtest är det test som bäst utnyttjar informationen så om materialet tillåter det är det detta test som bör användas.

   (a) *Wilcoxons teckenrangtest.* Om informationen tillåter det är Wilcoxons teckenrangtest det av de båda icke-parametriska testen som har den högsta styrkan vilket innebär att detta blir vårt förstahandsval. Vi förutsätter att studenterna är slumpmässigt utvalda och inte på något sätt påverkar varandras smakbetyg. Vidare måste vi förutsätta att den aktuella skalan för differenserna kan betraktas som en gemensam skala som åtminstone befinner sig på ordinalskalan. Det måste alltså gå att på ett objektivt sätt ställa olika studenters differenser i storleksordning. Att detta är fallet är inte helt klart i den här situationen eftersom smakbetyg är individuellt och subjektivt. Hypoteserna kan formuleras på lite olika sätt men här formulerar vi dem som

   H_0 : I studentpopulationen är fördelningen för smakbetygen samma för båda glasstyperna

   H_1 : I studentpopulationen är fördelningen för smakbetygen gällande den vanliga glassen förskjuten i positiv riktning

   vilka ska testas på 5% signifikansnivå. Vi förväntar oss låga rangtal på de negativa differenserna vilket innebär att vi som testfunktion använder T_-. I och med att det inte förekommer några ties ska vi enligt Tabell 5.7 förkasta nollhypotesen om

$$T_- \leq 10 = T_{10,0.05}$$

   Stickprovsinformationen ger oss att

Student	1	2	3	4	5	6	7	8	9	10
Reducerat socker	2	3	7	8	7	4	3	4	5	6
Vanlig	6	5	6	9	5	8	9	6	4	9
Vanlig−Reducerat	4	2	−1	1	−2	4	6	2	−1	3
Rang	8.5	5	2	2	5	8.5	10	5	2	7
Tecken	+	+	−	+	−	+	+	+	−	+

   och därmed följer att

$$T_- = 2 + 2 + 5 = 9$$

   och eftersom

$$T_- = 9 < 10 = T_{10,0.05}$$

   har vi hamnat i det kritiska området och förkastar nollhypotesen. Det är på 5% signifikansnivå säkerställt att det i vår studentpopulation gäller att fördelningen för smakbetygen gällande den vanliga glassen är förskjuten uppåt jämfört med motsvarande fördelning för glassen med reducerat sockerinnehåll.

(b) *Teckentest.* Om vi är tveksamma till att en jämförelse av differensernas storlek görs bör vi inte använda teckenrangtestet utan istället använda ett teckentest. Vi förutsätter som innan att studenterna är slumpmässigt utvalda och inte på något sätt påverkar varandras smakbetyg. Dock behöver vi som ytterligare förutsättning nu endast att vi för varje student kan avgöra vilken av glassarna som smakade bäst (vilket känns mycket rimligt). Hypoteserna kan formuleras på olika sätt men här formulerar vi (som ovan) hypoteserna som

H_0 : I studentpopulationen är fördelningen för smakbetygen samma för båda glasstyperna

H_1 : I studentpopulationen är fördelningen för smakbetygen gällande den vanliga glassen förskjuten i positiv riktning

vilka ska testas på 5% signifikansnivå. I och med att det inte förekommer några ties gäller att testfunktionen

$$X = \text{Antal minustecken}$$

är $Bi\,(10, 0.5)$ då nollhypotesen är sann. Vi har att

Student	1	2	3	4	5	6	7	8	9	10
Tecken	+	+	−	+	−	+	+	+	−	+

vilket innebär att

$$p\text{-värde} = \Pr\,(X \leq 3) = 0.1719$$

och då p-värdet överstiger den uppsatta signifikansnivån på 5% ska nollhypotesen accepteras. Det är på 5% signifikansnivå *inte* statistiskt säkerställt att det i vår studentpopulation gäller att fördelningen för smakbetygen gällande den vanliga glassen är förskjuten uppåt jämfört med motsvarande fördelning för glassen med reducerat sockerinnehåll. Det är alltså enligt denna testmetod inte säkerställt att den vanliga glassen är mer populär än den med reducerat sockerinnehåll.

(c) Studerar vi $Bi\,(10, 0.5)$ i Tabell 5.1 ser vi följande:

$$\begin{aligned} \Pr\,(X \leq 1) &= 0.0107 \\ \Pr\,(X \leq 2) &= 0.0547 \end{aligned}$$

Använder vi som kritiskt området $X \leq 2$ kommer den angivna signifikansnivån att överskridas och därmed gäller alltså att det kritiska området är $X \leq 1$ och att den faktiska signifikansnivån blir 1.07%. Om det nu är så att 65% av populationen föredrar den vanliga glassen gäller att testfunktionen

$$X = \text{Antal minustecken}$$

är $Bi\,(10, 0.35)$. Vi söker nu risken för att nollhypotesen (felaktigt) accepteras vilken ges av

$$\beta = \Pr\,(X \geq 2) = 1 - \Pr\,(X \leq 1) = 1 - 0.086 = 0.914$$

vilket alltså är en avsevärd risk trots att det är en tillsynes klart större andel som tycker att den vanliga glassen är godare.

4. Vi börjar med att komplettera den simultana fördelningen med marginalfördelningarna vilket ger

		y			
		0	1	2	$p\,(x)$
	0	0.07	0.02	0.01	0.10
x	1	0.12	0.27	0.21	0.60
	2	0.03	0.03	0.24	0.30
	$p\,(y)$	0.22	0.32	0.46	1.00

(a) Om vi uttrycker oss formellt gäller att det vi söker är $\Pr\,(X \geq 1 \mid X + Y = 2)$ och det följer därmed av räkneregler för betingade sannolikheter att

$$\begin{aligned} \Pr\,(X \geq 1 \mid X + Y = 2) &= \frac{\Pr\,(\{X \geq 1\} \cap \{X + Y = 2\})}{\Pr\,(X + Y = 2)} = \\ = \quad & \frac{\Pr\,(\{X = 1, Y = 1\} \cup \{X = 2, Y = 0\})}{\Pr\,(\{X = 0, Y = 2\} \cup \{X = 1, Y = 1\} \cup \{X = 2, Y = 0\})} = \\ = \quad & \frac{0.27 + 0.03}{0.01 + 0.27 + 0.03} = \frac{0.30}{0.31} = 0.968 \end{aligned}$$

där vi summerar sannolikheterna för de olika händelserna eftersom de är disjunkta. Om vi istället löser det mer informellt gäller att vi i korstabellen ser att det finns tre möjligheter för den sammanlagda poängen att bli 2. Vi vet alltså att vi befinner oss i antingen $x = 0, y = 2$ eller $x = 1, y = 1$ eller $x = 2, y = 0$. Eftersom kravet var att studenten skulle ha åtminstone en poäng på den första uppgiften söker vi (den betingade) sannolikheten att vi befinner oss i någon av de båda senare alternativen (givet att vi befinner oss i något av de tre alternativen). Utifrån de i korstabellen angivna sannolikheterna följer därmed att den sökta sannolikheten ges av

$$\frac{0.27 + 0.03}{0.01 + 0.27 + 0.03} = \frac{0.30}{0.31} = 0.968$$

som ovan.

(b) Sannolikheten att en slumpmässigt vald student fick åtminstone en poäng på den första inlämningsuppgiften, dvs $\Pr(X \geq 1)$, är 0.9. Nu väljer vi slumpmässigt 12 studenter och eftersom vi har ett mycket stort material av välja från följer att antal rätt på inlämningsuppgifterna för olika studenter är oberoende av varandra och om vi låter

$$U = \text{Antal studenter med åtminstone en poäng på inlämningsuppgift 1}$$

följer att U är binomialfördelad, $Bi(12, 0.9)$. Denna finns inte med i tabellen men om vi istället betraktar

$$V = \text{Antal studenter utan poäng på inlämningsuppgift 1}$$

gäller att V är binomialfördelad, $Bi(12, 0.1)$ och denna finns med i tabellen. Därmed följer att

$$\Pr(U \leq 10) = \Pr(V \geq 2) = 1 - \Pr(V \leq 1) = 1 - 0.659 = 0.341$$

(c) Eftersom urvalet nu sker från ett begränsat (och litet) material gäller att

$$T = \text{Antal studenter med full poäng på inlämningsuppgifterna}$$

är hypergeometriskt fördelad, $Hyp\left(3, \frac{9}{12}, 12\right)$. Stickprovet är för litet för att någon approximation ska vara tillåten så vi får lov att gå den hårda vägen.

$$\Pr(T \geq 2) = \frac{\binom{9}{2}\binom{3}{1}}{\binom{12}{3}} + \frac{\binom{9}{3}\binom{3}{0}}{\binom{12}{3}} = 0.8727$$

(d) Nu låter vi $W_1, W_2, \ldots, W_{40}$ representera totalt antal poäng på de båda inlämningsuppgifterna för var och en av studenterna i urvalet. Den totala antalet poäng för dessa studenter ges då av

$$S = W_1 + W_2 + \cdots + W_{40}$$

I och med att de 40 studenterna är slumpmässigt valda bland det stora antelet studenter i materialet följer att W-variablerna kan betraktas som o.l.f.s.v. där oberoendet skall tolkas som att rätt för olika studenter är oberoende av varandra och att de är likafördelade innebär att vi gör samma sannolikhetsbedömning angående antal rätt för alla studenter. I och med att vi har så pass många slumpvariabler ($n = 40 > 30$) ger Centrala gränsvärdessatsen gör att summan S blir approximativt normalfördelad. För att kunna använda normalfördelningen behöver vi ha väntevärde och standardavvikelse för S varför vi börjar med att bestämma detta för W.

$$\begin{aligned} E(W) &= 0 \cdot 0.07 + 1 \cdot 0.14 + 2 \cdot 0.31 + 3 \cdot 0.24 + 4 \cdot 0.24 = 2.44 \\ E\left(W^2\right) &= 0^2 \cdot 0.07 + 1^2 \cdot 0.14 + 2^2 \cdot 0.31 + 3^2 \cdot 0.24 + 4^2 \cdot 0.24 = 7.38 \end{aligned}$$

vilket leder till att

$$\sigma(W) = \sqrt{7.38 - 2.44^2} = 1.1943$$

För studenterna i vårt stora material gäller alltså att den genomsnittliga totalpoängen är 2.44. Dock gäller att studenters totalpoäng i genomsnitt avviker med ca 1.2 poäng från detta medelvärde. Nu är det emellertid den sammanlagda poängen för fyrtio studenter vi är ute efter vilket innebär att

$$\begin{aligned} E(S) &= 40 \cdot 2.44 = 97.6 \\ \sigma(S) &= \sqrt{40} \cdot 1.1943 = 7.5534 \end{aligned}$$

och alltså har vi att det totala antalet poäng för dessa studenter, S, approximativt är $N(97.6; 7.55)$. Vi finner nu den sökta sannolikheten (med kontinuitetskorrektion) via

$$\Pr(S > 100) \approx \Pr\left(Z > \frac{100.5 - 97.6}{7.55} \approx 0.38\right) \approx 0.352$$

dvs det är ungefär 35% chans att det totala antalet poäng för de fyrtio studenterna överstiger 100.

TENTAMENSSKRIVNING PÅ KURSERNA

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

2013-04-27

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

En läkare har telefonmottagning för sina patienter, vilket förväntas ta 60 minuter per dag. Dock misstänker läkaren att det tar mer tid än så, och kontrollmäter därför hur lång tid telefonmottagningen tar under 10 slumpmässigt valda dagar. Hon fick följande värden:

79, 84, 75, 52, 71, 62, 68, 68, 76, 62

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

(8) **B** Konstruera en boxplot för materialet.

(6) **C** I D- och E-uppgiften nedan görs ett normalfördelningsantagande. Vad är det som måste vara normalfördelat? Varför är antagandet nödvändigt i D-uppgiften? Varför är antagandet nödvändigt i E-uppgiften? Förklara kortfattat hur man utifrån olika mått på genomsnitt kan göra en första kontroll över huruvida frekvensfördelningen är symmetrisk. Vad ser vi utifrån genomsnittsmåtten tecken på (med avseende på en symmetrisk fördelning) i det här materialet?

(12) **D** Går det att statistiskt säkerställa att *medeltiden* för telefonmottagningen överstiger sextio minuter? Utför ett fullständigt hypotestest enligt klassisk metod där du använder en signifikansnivå på 5%. Glöm inte att ange alla antaganden du gör.

(8) **E** Vad kan utifrån det insamlade materialet sägas om *spridningen* i mottagningstid för denna läkare? Besvara frågan genom att konstruera ett konfidensintervall där du använder en konfidensgrad på 90%.

(10) **F** Anta att vi kom fram till att det för testet i D-uppgiften nödvändiga normalfördelningsantagandet inte var rimligt. Vi beslutar oss istället för att utföra ett hypotestest angående *medianen*. Går det att statistiskt säkerställa att *mediantiden* för telefonmottagningen överstiger sextio minuter? Utför ett fullständigt hypotestest med 5% signifikansnivå. Ange alla antaganden du gör.

(8) **Uppgift 2**

På en viss universitetskurs består examinationen dels av en skriftlig tentamen i slutet av kursen och dels av en större inlämningsuppgift. Nedan presenteras resultaten från både den skriftliga tentamen och inlämningsuppgiften för ett slumpmässigt urval av 10 studenter på kursen.

Student	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Skriftlig tentamen	81	62	74	78	93	69	72	83	90	84
Inlämningsuppgift	78	71	69	76	87	62	80	75	92	79

Hur starkt samband har vi mellan de båda variablerna? Beräkna rangkorrelationskoefficienten och tolka även denna genom att förklara vad det framräknade värdet är ett tecken på.

Uppgift 3

Hur får konsumenter upp ögonen för en ny produkt? För att få svar på denna fråga gjordes ett slumpmässigt urval av 200 användare av en ny produkt. I undersökningen tog man bland annat även reda på konsumenternas ålder.

Utav de undersökta var 50 personer under 21 år samt 90 personer mellan 21 och 35 år. Övriga personer i undersökningen var över 35 år.

Utav de tillfrågade under 21 år hade 60% hört talas om produkten från en vän och de resterande hade läst om det i en annons i lokaltidningen. Utav de tillfrågade mellan 21 och 35 år hade två tredjedelar hört talas om produkten från en vän och de resterande hade läst om det i en annons i lokaltidningen. Utav de tillfrågade över 35 år hade 30% hört talas om produkten från en vän och de resterande hade läst om det i en annons i lokaltidningen.

Den fråga vi ska besvara i deluppgifterna nedan är följande: Föreligger det ett samband mellan konsumenternas ålder och det sätt på vilket konsumenter fick upp ögonen för den nya produkten?

(6) **A** Åskådliggör situationen grafiskt genom att skapa ett diagram med relativa frekvenser som på bästa sätt visar på eventuella skillnader mellan de tre ålderskategoriseringarna vad det gäller sättet på vilket man fick upp ögonen för den nya produkten.

(12) **B** Låt oss nu besvara frågan genom att utföra en ordentlig statistisk analys. Ställ upp hypoteser och utför ett <u>fullständigt</u> hypotestest med 5% signifikansnivå enligt *p*-värdesmetoden.

Uppgift 4

Ett bokförlag kan använda alla, några eller inga av tre möjliga strategier för att förbättra försäljningen av en bok:

A. Extra påkostad marknadsföring.
B. Ett påkostat omslag.
C. En bonus för säljare som uppfyller förutbestämda försäljningsnivåer.

Fram till nu har dessa tre strategier använts samtidigt vid endast två procent av förlagets böcker. Var femte bok har fått påkostade omslag och av dessa har åttio procent även haft extra påkostad marknadsföring.

(7) **A** Vid en viss tidpunkt ska förlaget ge ut 30 nya böcker. Vi är intresserade av att göra en sannolikhetsbedömning för hur många av dessa böcker som kommer att få tillgång till samtliga tre strategier.

(i) För att göra denna sannolikhetsbedömning beslutar vi oss för att utgå från binomialfördelningen. Förklara resonemanget bakom detta beslut. Åtminstone ett av de för beräkningen nödvändiga antagandena är tvivelaktigt. Förklara.

(ii) Bestäm sannolikheten att de tre strategierna används samtidigt för högst en av dessa 30 böcker.

(5) **B** Ett konkurrerande förlag får reda på att en ny bok från det aktuella förlaget ska få både extra påkostad marknadsföring och ett påkostat omslag. Hur troligt är det att man vid marknadsföringen av boken även ger en bonus för säljare som uppfyller förutbestämda försäljningsnivåer?
Anmärkning. En välmotiverad lösning med hjälp av ett Venn-diagram är tillräckligt för full poäng.

Uppgift 5

En möbeltillverkare har konstaterat att den tid det tar för en arbetare att montera ett visst bord kan ses som normalfördelad med ett medelvärde på 150 minuter och en standardavvikelse på 20 minuter.

(3) **A** Bestäm den första kvartilen vad det gäller monteringstid för denna typ av bord.

(7) **B** Anta att vi en viss dag har en beställning på tre sådana bord och endast en arbetare tillgänglig för monteringen av dessa. Bestäm sannolikheten att beställningen blir färdig under denna arbetsdag (åtta timmar). Kritiska steg i beräkningen måste motiveras.

1. Vi studerar här den kvantitativa variabeln

$$x = \text{Antal minuter per dag läkaren sitter i telefonmottagning}$$

(a) Summerar vi värdena och även deras kvadrater fås summorna

$$\begin{aligned} \sum x &= 697 \\ \sum x^2 &= 49\,379 \end{aligned}$$

vilket leder fram till att

$$\overline{x} = \frac{697}{10} = \mathbf{69.7}$$

och att

$$s = \sqrt{\frac{49379 - \frac{697^2}{10}}{10-1}} = \mathbf{9.42}$$

Under de 10 dagarna som var med i urvalet satt läkaren i genomsnitt 69.7 minuter i telefonmottagning. Mottagningen tog dock inte lika lång tid varje dag utan i genomsnitt avvek tiden som avsattes med 9.42 minuter från medelvärdet.

(b) Vi ställer observationerna i storleksordning vilket ger

$$52, 62, 62, 68, 68, 71, 75, 76, 79, 84$$

varpå kvartilerna följer som

$$\begin{aligned} q_1 &= \left(\text{Värdet på obs } \frac{10+1}{4} = 2.75\right) = 62 + 0.75 \cdot (62 - 62) = 62 \\ md &= \left(\text{Värdet på obs } \frac{10+1}{2} = 5.5\right) = 68 + 0.5 \cdot (71 - 68) = 69.5 \\ q_3 &= \left(\text{Värdet på obs } \frac{3 \cdot (10+1)}{4} = 8.25\right) = 76 + 0.25 \cdot (79 - 76) = 76.75 \end{aligned}$$

Ett och ett halvt kvartilavstånd ges av

$$1.5 \cdot (76.75 - 62) = 22.125$$

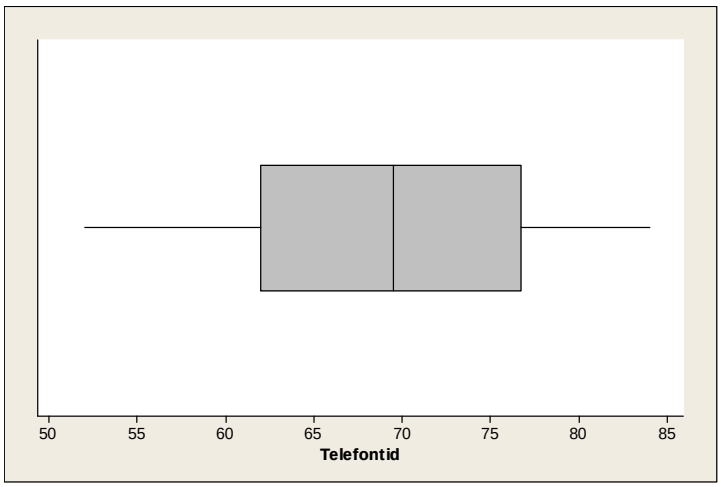
varför uteliggare är observationer under

$$62 - 22.125 = 39.875$$

och över

$$76.75 + 22.125 = 98.875$$

Vi har således inte några uteliggare i vårt material och morrhåren dras till minsta respektive största värde. Boxploten får därmed följande utseende.



(c) I den här situationen måste vi förutsätta att populationen är normalfördelad med avseende på den aktuella variabeln, dvs den tid läkaren sitter i telefonmottagning måste variera i likhet med en normalfördelning. Anta att vi för den aktuella läkaren mäter mottagningstiden under ett mycket stort antal dagar och sedan beskriver detta grafiskt (exempelvis) i form av ett histogram. För att antagandet ska stämma ska detta histogram uppvisa stora likheter med en normalfördelningskurva. Varför behöver antagandet göras? Våra slutsatser i d-uppgiften bygger på att stickprovsmedelvärdet är (approximativt) normalfördelat och eftersom stickprovet är litet kan vi inte förlita oss på att Centrala gränsvärdessatsen hunnit göra stickprovsmedelvärdet (approximativt) normalfördelat. Därför är det tidigare angivna normalfördelningsantagandet nödvändigt. Våra slutsatser i e-uppgiften bygger på stickprovsvariansen och för att dessa slutsatser ska kunna göras via χ^2-fördelningen krävs (obeoende av stickprovsstorlek) att populationen är normalfördelad. I en helt symmetrisk fördelning gäller att de båda genomsnittsmåtten medelvärde och median är samma vilket innebär att vi kan göra en grov kontroll av symmetri (i stickprovet) genom att jämföra dessa i vårt stickprov. Observera dock att ett symmetriskt/asymmetriskt stickprov inte är en garanti för att populationen är symmetrisk/asymmetrisk. Dock gäller att det ger oss en första indikation om hur populationen ser ut. I vårt stickprov är medelvärde och median nästa samma vilket åtminstone inte talar emot vårt normalfördelningsantagande.

(d) Låter vi först

$$\mu = \text{Medeltiden för läkarens telefonmottagning}$$

följer av frågeställningen i uppgiften att hypoteserna ska formuleras som

$$\begin{aligned} H_0 &: \mu = 60 \\ H_1 &: \mu > 60 \end{aligned}$$

Detta ska nu undersökas med ett test på 5% signifikansnivå. Om vi förutsätter att de dagar som blivit bedömda är slumpmässigt utvalda ur en mycket stor (oändlig) population av (potentiella) dagar samt att den tid läkaren sitter i telefonmottagning kan ses som approximativt normalfördelad (se c-uppgiften för utförligare förklaring) följer att testfunktionen

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

är t-fördelad med $n-1$ frihetsgrader då nollhypotesen är sann. Eftersom hypotesen är ensidig och $n = 10$ följer att nollhypotesen ska förkastas endast då

$$t_{obs} > 1.833 = t_{9,0.05}$$

Vi får följande värde på testfunktionen

$$t = \frac{69.7 - 60}{9.42/\sqrt{10}} = 3.26$$

och eftersom

$$t_{obs} = 3.26 > 1.833$$

har vi hamnat i det kritiska området och nollhypotesen förkastas. Vi har på 5% signifikansnivå statistiskt säkerställt att medeltiden läkaren sitter i telefonmottagning överstiger sextio minuter.

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

$$\sigma = \text{Standardavvikelsen för läkarens telefonmottagningstid}$$

Om vi förutsätter att de dagar som blivit bedömda är slumpmässigt utvalda ur en mycket stor (oändlig) population av (potentiella) dagar samt att den tid läkaren sitter i telefonmottagning kan ses som approximativt normalfördelad (se c-uppgiften för utförligare förklaring) följer att vi kan använda intervallet

$$\sqrt{\frac{(n-1)\,s^2}{\chi^2_{n-1,\alpha/2}}} < \sigma < \sqrt{\frac{(n-1)\,s^2}{\chi^2_{n-1,1-\alpha/2}}}$$

Eftersom $\chi^2_{9,0.95} = 3.325$ och $\chi^2_{9,0.05} = 16.919$ följer efter insättning av våra stickprovsvärden att konfidensintervallet blir

$$\sqrt{\frac{9 \cdot 9.42^2}{16.919}} < \sigma < \sqrt{\frac{9 \cdot 9.42^2}{3.325}}$$

eller mer kortfattat

$$\mathbf{6.9} \leq \boldsymbol{\sigma} \leq \mathbf{15.5}$$

Med 90% säkerhet är standardavvikelsen σ för den tid läkaren sitter i telefonmottagning, dvs den genomsnittliga avvikelsen från medelmottagningstiden, någonstans mellan 6.9 minuter och 15.5 minuter.

(f) Låter vi först

$$md = \text{Mediantiden för läkarens telefonmottagning}$$

följer av frågeställningen i uppgiften att hypoteserna ska formuleras som

$$\begin{aligned} H_0 &: md = 60 \\ H_1 &: md > 60 \end{aligned}$$

Detta ska nu undersökas med ett test på 5% signifikansnivå. Om vi förutsätter att de dagar som blivit bedömda är slumpmässigt utvalda ur en mycket stor (oändlig) population av (potentiella) dagar följer att testfunktionen

$X =$ Antal dagar i urvalet läkaren sitter längre än sextio minuter i telefonmottagning

är $Bi\,(10, 0.5)$ då nollhypotesen är sann. I urvalet satt läkaren längre än sextio minuter i telefonmottagning i inte mindre än nio av de tio dagarna vilket är ett tillsynes mycket starkt tecken på att mediantiden överstiger sextio minuter. Testets p-värde blir

$$p\text{-värde} = \Pr\,(X \geq 9) = 1 - \Pr\,(X \leq 8) = 1 - 0.9893 = 0.0107$$

Eftersom

$$p\text{-värde} < 5\%$$

förkastas nollhypotesen. Det är därmed på 5% signifikansnivå statistiskt säkerställt att mediantiden läkaren sitter i telefonmottagning överstiger sextio minuter.

2. Vi låter här

$$\begin{aligned} x &= \text{Poäng på skriftlig tentamen} \\ y &= \text{Poäng på inlämningsuppgift} \end{aligned}$$

För att mäta sambandets styrka ska vi använda Spearmans rangkorrelationskoefficient vilket innebär att vi beräknar den "vanliga" korrelationskoefficienten för de till våra värden associerade rangtalen. I och med att det inte förekommer "ties" kan dock den förenklade formeln användas. Vi sammanställer därmed den givna informationen i en tabell där vi utökar med rangtalen och de associerade differenserna och deras kvadrater.

x	y	R_x	R_y	d_i	d_i^2
81	78	6	6	0	0
62	71	1	3	−2	4
74	69	4	2	2	4
78	76	5	5	0	0
93	87	10	9	1	1
69	62	2	1	1	1
72	80	3	8	−5	25
83	75	7	4	3	9
90	92	9	10	−1	1
84	79	8	7	1	1
		55	55		46

I och med denna information kan vi nu via den förenklade formeln beräkna rangkorrelationskoefficienten till

$$r_s = 1 - \frac{6 \cdot 46}{10 \cdot (10^2 - 1)} = \mathbf{0.721}$$

Den beräknade korrelationskoefficienten är ett förhållandevis stort positivt värde vilket innebär att vi ser tydliga tecken på att de som preseterar bra (dåligt) på den ena examinationsformen även tenderar att prestera bra (dåligt) på den andra. Om vi istället

går den "hårda" vägen behöver vi följande information.

x	y	R_x	R_y	R_x^2	R_y^2	R_xR_y
81	78	6	6	36	36	36
62	71	1	3	1	9	3
74	69	4	2	16	4	8
78	76	5	5	25	25	25
93	87	10	9	100	81	90
69	62	2	1	4	1	2
72	80	3	8	9	64	24
83	75	7	4	49	16	28
90	92	9	10	81	100	90
84	79	8	7	64	49	56
		55	55	385	385	362

varefter vi först beräknar de tre nyckelsummorna

$$\begin{aligned}
\sum \left(R_x - \overline{R}_x\right)^2 &= 385 - \frac{55^2}{10} = 82.5 \\
\sum \left(R_y - \overline{R}_y\right)^2 &= 385 - \frac{55^2}{10} = 82.5 \\
\sum \left(R_x - \overline{R}_x\right)\left(R_y - \overline{R}_y\right) &= 362 - \frac{55 \cdot 55}{10} = 59.5
\end{aligned}$$

vilket (givetvis som vid beräkningen med den förenklade formeln) ger oss att

$$r_s = \frac{59.5}{\sqrt{82.5 \cdot 82.5}} = \mathbf{0.721}$$

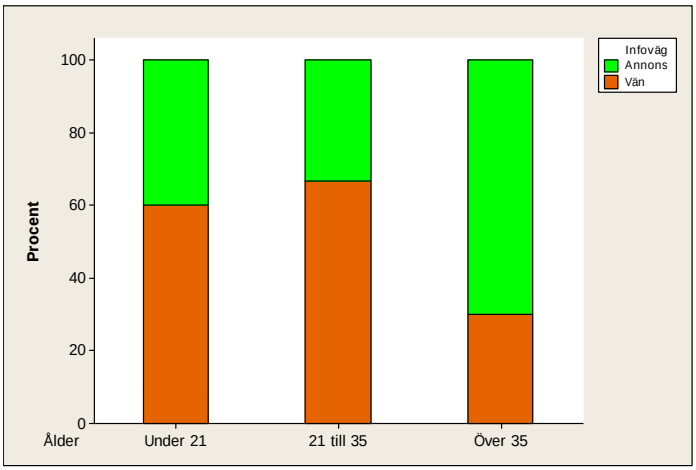
3. Föreligger det ett samband mellan konsumenternas ålder och det sätt på vilket konsumenter fick upp ögonen för den nya produkten?

   (a) I och med att vi ska skapa ett diagram med relativa frekvenser som på bästa sätt visar på eventuella skillnader mellan de tre ålderskategoriseringarna vad det gäller sättet på vilket man fick upp ögonen för den nya produkten är det kolumnerna (ålderskategorierna) som ska summera till 100 procent. Vi får därmed följande tabell

Rows: Infoväg Columns: Ålder

	Under 21	21 till 35	Över 35	All
Vän	30	60	18	108
	60,00	66,67	30,00	54,00
Annons	20	30	42	92
	40,00	33,33	70,00	46,00
All	50	90	60	200
	100,00	100,00	100,00	100,00

vilket i sin tur ger oss det uppdelade stapeldiagrammet nedan.



Vi ser tillsynes endast små skillnader mellan de två första åldersgrupperna medan de över 35 skiljer sig från de båda första. För att avgöra hur stora skillnaderna är behövs ett formellt hypotestest. Så låt oss gå vidare till *b*-uppgiften.

(b) I och med att det inte är binära variabler (det är tre ålderskategorier) måste det bli ett χ^2-test. För att undersöka detta ställer vi upp hypoteserna

H_0 : Inget samband mellan ålder och det sätt på vilket konsumenter fick upp ögonen för den nya produkten

H_1 : Samband mellan ålder och det sätt på vilket konsumenter fick upp ögonen för den nya produkten

och testar detta på 5% signifikansnivå. Eftersom konsumenterna i urvalet är slumpmässigt utvalda (ur en stor population) samt att inga av de förväntade frekvenserna understiger 5 följer att testfunktionen

$$\chi^2 = \sum\sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

är χ^2-fördelad med $(2-1)\cdot(3-1) = 2$ frihetsgrader då nollhypotesen är sann. Nollhypotesen förkastas först då p-värdet understiger 5%. För att kunna beräkna testfunktionens värde behöver vi dom förväntade frekvenserna och en korstabell där både observerade och förväntade frekvenser är angivna får följande utseende

Rows: Infoväg Columns: Ålder

	Under 21	21 till 35	Över 35	All
Vän	30	60	18	108
	27,00	48,60	32,40	108,00
Annons	20	30	42	92
	23,00	41,40	27,60	92,00
All	50	90	60	200
	50,00	90,00	60,00	200,00

Hur har vi då funnit de förväntade frekvenserna? Eftersom nollhypotesen förutsätts vara sann ska det samma proportioner mellan Vän och Annons för samtliga ålderskategorier varför vi exempelvis för personer under 21 finner förväntat antal som fått informationen via en vän via

$$E_{Vän,<21} = \frac{108 \cdot 50}{200} = 27$$

Ingen av de förväntade frekvenserna understiger 5 vilket innebär att vi kan gå vidare och jämföra observerade och förväntade frekvenser i testfunktionen

$$\chi^2 = \frac{(30-27)^2}{27} + \frac{(60-48.6)^2}{48.6} + \cdots + \frac{(42-27.6)^2}{27.6} = 20.45$$

Eftersom

$$\chi^2_{\text{obs}} = 20.45 > 10.597 = \chi^2_{2,0.005}$$

följer att

$$p\text{-värde} < 0.5\%$$

Vi har hamnat i det kritiska området och förkastar därmed nollhypotesen. Det finns således på 5%-nivån ett statistiskt säkerställt samband mellan konsumenternas ålder och det sätt på vilket konsumenter fick upp ögonen för den nya produkten.

4. Vi börjar med att för den nya boken definiera följande händelser.

$$\begin{aligned} A &= \text{Boken får extra påkostad marknadsföring} \\ B &= \text{Boken får ett påkostat omslag} \\ C &= \text{Boken associeras med en bonus för säljare som uppfyller förutbestämda försäljningsnivåer} \end{aligned}$$

Om vi förutsätter att förlaget kommer att fortsätta som tidigare finner vi utifrån den givna informationen att

$$\begin{aligned} \Pr(A \cap B \cap C) &= 0.02 \\ \Pr(B) &= 0.2 \\ \Pr(A \mid B) &= 0.8 \end{aligned}$$

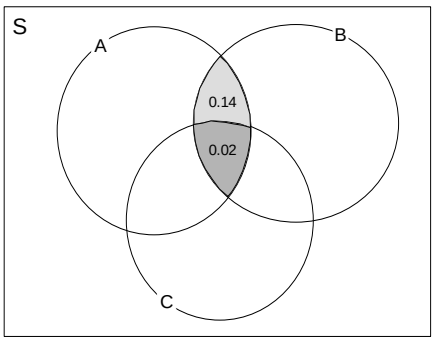
(a) Om vi förutsätter att förlaget kommer att fortsätta med sina försäljningsförbättrande strategier som tidigare samtidigt som att de åtgärder som görs för att öka försäljningen av en bok inte på något sätt påverkar eller påvkeras av vilka åtgärder som görs för att öka försäljningen av en annan bok gäller att

$$X = \text{Antal böcker där de tre strategierna används samtidigt}$$

är $Bi(30, 0.02)$. Det är förstås det andra antagandet (det rörande oberoende) som är tveksamt. Vi finner nu att

$$\Pr(X \leq 1) = \Pr(X = 0) + \Pr(X = 1) = 0.98^{30} + \binom{30}{1} \cdot 0.02 \cdot 0.98^{29} = \mathbf{0.88}$$

(b) Om vi löser uppgiften grafiskt får det associerade Venn-diagrammet följande utseende:



Observera dock att cirklarnas/händelsernas storlek inte nödvändigtvis reflekterar de sannolikheter som förknippas med dem. Händelsen $A \cap B \cap C$ är det mörkgrå området och enligt den givna informationen ska $\Pr(A \cap B \cap C) = 0.02$. Hela det gråmarkerade området är $A \cap B$ och sannolikheten för denna händelse finner vi via sannolikhetslärans multiplikationssats till

$$\Pr(A \cap B) = \Pr(B)\Pr(A \mid B) = 0.2 \cdot 0.8 = 0.16$$

(vilket innebär att det område som är markerat i ljusgrått har sannolikheten $0.1\dot{4}$). Det vi söker är $\Pr(C \mid A \cap B)$ vilket i Venn-diagrammet kan uttryckas som att vi söker hur stor del, rent sannolikhetsmässigt, av det gråmarkerade området som utgörs av det mörkgrå området. Alltså följer att

$$\Pr(C \mid A \cap B) = \frac{0.02}{0.16} = \frac{\mathbf{1}}{\mathbf{8}} = \mathbf{0.125}$$

Den formella beräkningen blir

$$\Pr(C \mid A \cap B) = \frac{\Pr(A \cap B \cap C)}{\Pr(A \cap B)} = \frac{\Pr(A \cap B \cap C)}{\Pr(B)\Pr(A \mid B)} = \frac{0.02}{0.2 \cdot 0.8} = \mathbf{0.125}$$

5. Om vi nu låter

$$X = \text{Monteringstid i minuter för denna typ av bord}$$

gäller enligt den givna informationen att X är $N(150, 20)$.

(a) Vi söker den första kvartilen och enligt Tabell 5.2.B gäller att $z_{0.75} = -z_{0.25} = -0.6745$ vilket alltså betyder att den första kvartilen befinner sig 0.6745 standardavvikelser under medelvärdet, dvs

$$q_1 = 150 - 0.6745 \cdot 20 = \mathbf{136.5}$$

I långa loppet kommer en fjärdedel av dessa bord att ta kortare tid än 136.5 minuter att montera.

(b) Om vi nu låter X_1, X_2 och X_3 representera tiden det tar att montera vart och ett av de tre borden och samtidigt lägger till antagandet att monteringstiden för olika bord är oberoende av varandra följer att den sammanlagda monteringstiden för de tre borden är

$$Y = X_1 + X_2 + X_3$$

dvs en summa av tre oberoende och normalfördelade slumpvariabler och därmed själv normalfördelad. Det återstår att finna väntevärde och standardavvikelse för Y vilka ges av

$$\begin{aligned} E(Y) &= 3 \cdot E(X) = 3 \cdot 150 = 450 \\ \sigma(Y) &= \sqrt{3} \cdot \sigma(X) = \sqrt{3} \cdot 20 = 34.64 \end{aligned}$$

dvs Y är $N(450, 34.64)$. Eftersom åtta timmar är samma som 480 minuter söker vi

$$\Pr(Y \leq 480) = \Pr\left(Z \leq \frac{480 - 450}{34.64} = 0.87\right) = \mathbf{0.81}$$

dvs det är ungefär 81% chans att borden blir färdiga under denna arbetsdag.

TENTAMENSSKRIVNING PÅ KURSERNA

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

2013-10-31

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

Åldersfördelningen för samtliga vinnare av en Oscar (eller Academy Award) i kategorin bästa kvinnliga huvudroll sammanfattas i frekvenstabellen nedan.

Ålder	Frekvens
21-30	32
31-40	32
41-50	13
51-60	2
61-70	4
71-80	2

Anmärkning. Samtliga beräkningar för deluppgifterna A-C nedan ska göras utifrån ovanstående frekvenstabell.

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

(4) **B** Beräkna den första kvartilen. Ge en ordentlig förklaring av innebörden av det framräknade värdet.

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

(7) **D** Åldersfördelningen för samtliga vinnare av en Oscar (eller Academy Award) i kategorin bästa kvinnliga huvudroll återges i stam-blad-diagrammet nedan.

```
 29   2  12244555566667777888899999999
(34)  3  000111222333333444555555677888899
 22   4  0111112235569
 9    5  04
 7    6  01123
 2    7  4
 1    8  0
```

Använd informationen i stam-blad-diagrammet ovan för att återge åldersfördelningen i ett lådagram.

(8) **Uppgift 2**

Går det att finna något samband mellan den budget en film har och de biljettintäkter filmen inbringar? För ett litet urval av sju filmer har vi följande information (där alla belopp är i miljoner dollar):

Budget	62	90	50	35	200	100	90
Biljettintäkter	65	64	48	57	601	146	47

Anpassa en linjär regressionsmodell som beskriver hur biljettintäkterna beror på budgeten. Ge en ordentlig tolkning av de båda framräknade regressions-koefficienterna a och b i termer av de ingående variablerna (budget och biljettintäkter). Förklara även varför en tolkning av a-koefficienten i det här fallet bör tas med en rejäl nypa salt?

Följande information från Minitab kan kanske vara till viss hjälp vid beräkningarna.

Kolumn C1 innehåller variabeln ”Budget”.
Kolumn C2 innehåller variabeln ”Biljettintäkter”.

```
MTB > let c3=c1*c2
MTB > let c4=c1**2
MTB > let c5=c2**2

Variable     Sum
C1         627,0
C2        1028,0
C3        153215
C4         73769
C5        398600
```

Uppgift 3

PTC är en förening som har en stark bitter smak för vissa människor och är smaklös för andra. Förmågan att känna den bittra smaken av PTC är ett ärftligt anlag. Många studier har för olika populationer bedömt andelen människor som kan känna smaken av PTC. Följande tabell ger resultat för stickprov från några länder:

	Land			
Smak	**Irland**	**Portugal**	**Norge**	**Italien**
Bitter	558	345	185	402
Smaklös	225	109	81	134

(12) **A** Undersök med ett hypotestest på 5% signifikansnivå om det går att statistiskt säkerställa att andelen individer som kan känna smaken av PTC skiljer sig mellan de olika länderna.

(6) **B** Ge en ordentlig förklaring till de båda begreppen *signifikansnivå* och *Typ2-fel* genom att *utgå från situationen i den här uppgiften*. Någon allmän förklaring av begreppen ger inte några poäng.

Uppgift 4

I en studie av förskoleverksamheten fick ett antal förskoleelever återberätta en saga som hade blivit uppläst för dem tidigare i veckan. Av de tio barn som ingick i studien klassificerades fem som långt framskridna vad det gäller språkförståelse medan de övriga fem barnen klassificerades som normalt framskridna beträffande språkförståelse. En expert lyssnade på inspelningar av barnens återberättelser och tilldelade därefter varje barn en poäng för att belysa nivån på deras språkbruk. Bedömningen gjordes på en heltalsskala från 1 till 100 där 1 betyder en mycket låg nivå på språkbruket medan 100 betyder en mycket hög nivå på språkbruket. I tabellen nedan återges denna bedömning för de tio barnen.

Språkförståelse	**Expertbedömning**				
Normal	77	49	66	28	38
Hög	80	82	54	79	89

Aktuell frågeställning. Är det utifrån resultatet av denna studie statistiskt säkerställt att den genomsnittliga språknivån i en sådan återberättelse är högre för barn som klassificeras som långt framskridna vad det gäller språkförståelse jämfört med barn klassificerade som normalt framskridna beträffande språkförståelse?

(4) **A** Vad är variabel i den här situationen och vad kan sägas om dess datanivå? Resonera kortfattat kring hur den aktuella variabelns datanivå påverkar val av hypotestest i den här situationen.

(4) **B** I C-uppgiften nedan görs ett normalfördelningsantagande. Vad är det som måste vara normalfördelat och varför är det nödvändigt i det här fallet? Förklara.

(10) **C** Besvara den aktuella frågeställningen via ett *parametriskt* test enligt klassisk metod på 5% signifikansnivå. Till vår hjälp har vi följande Minitabutskrift:

Variable	Språkklass	N	Mean	SE Mean	StDev	Q1	Median	Q3
Expertpoäng	Normal	5	51,60	8,95	20,01	33,00	49,00	71,50
	Hög	5	76,80	5,96	13,33	66,50	80,00	85,50

Anmärkning. Eftersom urvalen är små är det här tillåtet att använda tumregeln för att bestämma lämplig testfunktion.

(2) **D** Uppskatta *p*-värdet för det test som utfördes i C-uppgiften ovan.

(10) **E** Besvara den aktuella frågeställningen via ett *icke-parametriskt* test på 5% signifikansnivå.

Uppgift 5

Anmärkning. För full poäng på uppgifterna nedan måste samtliga kritiska steg i beräkningarna ordentligt motiveras.

Aron är vaktmästare på en skola.

(5) **A** Han ska byta fyra glödlampor i ett av klassrummen. I verkstaden har han en låda med 20 glödlampor. Dumt nog har han blandat nya och gamla glödlampor i lådan, så bland de 20 lamporna finns det 15 nya och 5 gamla, trasiga, lampor. Aron vill gärna ta med sig tillräckligt många glödlampor så att han kan byta samtliga fyra trasiga lampor i klassrummet på en gång. Eftersom han är medveten om att några lampor i hans låda är trasiga tar han med sig fem glödlampor som han väljer slumpmässigt ur sin låda. Hur stor är sannolikheten att han måste gå tillbaka till verkstaden en gång till för att hämta fler lampor?

(7) **B** Han vet dessutom att några stolar i klassrummet behöver lagas. För detta behöver han sex skruvar av typen CT-4A. I en låda har han ett mycket stort antal skruvar varav 30 procent är av typ CA-35T, 20 procent av typ CT-4A och 50 procent av typ ST04-X. Han tar med sig 40 slumpmässigt utvalda skruvar ur lådan. Hur stor är sannolikheten att Aron kan laga stolarna med de skruvar han tagit med sig?

(6) **C** Arons skruvar, beskrivna i uppgift B, är inköpta vid två olika tillfällen. De nyare skruvarna har röda huvuden, de gamla vita. I hans lager har 15 procent av skruvarna av typen CT-4A röda huvuden, bland de båda typerna CA-35T och ST04-X är det 30 procent av skruvarna som har röda huvuden. Nedanför skruvlådan hittar Aron en skruv med ett rött huvud som har ramlat ur lådan. Hur stor är sannolikheten att skruven är av typen CA-35T?

(2) **D** Utgående från uppgift C – är händelserna *skruvtyp* och *rött huvud* oberoende av varandra? Ditt svar måste motiveras.

1. *Anmärkning.* Denna situation kan uppfattas på två sätt; dels som ett urval från en bakomliggande population av skådespelerskor och dels som en totalundersökning. Jag har här i lösningarna valt att använda den första tolkningen av situationen men det är inte fel att betrakta det som en totalundersökning.
Vi börjar med att återge (och utöka) frekenstabellen

Ålder	f_i	Mitt (x_i)	$f_i x_i$	$f_i x_i^2$	Kum
21–30	32	26	832	21 632	32
31–40	32	36	1 152	41 472	64
41–50	13	46	598	27 508	77
51–60	2	56	112	6 272	79
61–70	4	66	264	17 424	83
71–80	2	76	152	11 552	85
	85		**3 110**	**125 860**	

(a) Vi beräknar medelvärde och standardavvikelse till

$$\begin{aligned} \overline{x} &= \frac{3110}{85} = \mathbf{36.6} \\ s &= \sqrt{\frac{125860 - \frac{3110^2}{85}}{84}} = \mathbf{12.0} \end{aligned}$$

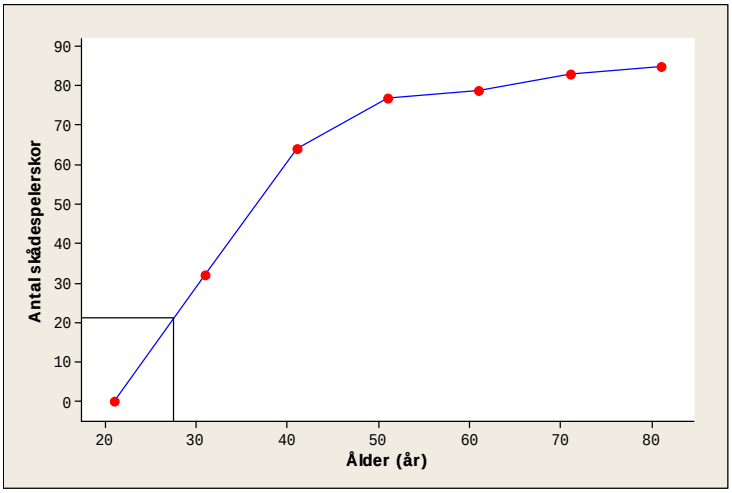
Medelåldern (vid tidpunkten för utnämningen) för de 85 skådespelerskor som tilldelats en Oscar för bästa kvinnliga huvudroll är 36.6 år. Dock gäller att inte alla dessa skådespelerskor var lika gamla då de tilldelades priset. Standardavvikelsen anger att de avvek med i genomsnitt 12 år från medelåldern.

(b) Vi ser att den observation vars värde är den första kvartilen, dvs observationen med ordningsnummer 21.25, befinner sig i klassen 21–30. Den första kvartilen beräknas därmed till

$$q_1 = 21 + \frac{21.25 - 0}{32} \cdot 10 = \mathbf{27.6}$$

dvs den första kvartilen är **27.6** år. Om vårt antagande om att skådespelerskorna är jämnt utspridda i åldersklasserna gäller att en fjärdedel av dem är under 27.6 år. (I d-uppgiften nedan ser vi att det korrekta värdet på q_1 är 28.5 år).

(c) I och med att det är klassindelat material används en summapolygon för att beskriva den kumulativa frekvensfördelningen. Vi använder därför de kumulerade frekvenserna från vår frekvenstabell och får på så sätt följande diagram



Genom att på y-axeln utgå från observation $n/4 = 21.25$, dvs observationen vars värde är q_1, och dra en horisontell linje fram till summapolygonen och sedan därifrån dra en lodrät linje ner till x-axeln får vi en uppskattning av värdet på q_1. Enligt resultatet i b-uppgiften ska denna bli 27.6 vilket verkar rimligt.

(d) Eftersom observationerna i ett stam-blad-diagram är uppställda i storleksordning finner vi

$$\begin{aligned} q_1 &= \left(\text{Värdet på observation } \frac{85+1}{4} = 21.5\right) = 28.5 \\ md &= \left(\text{Värdet på observation } \frac{85+1}{2} = 43\right) = 33 \\ q_3 &= \left(\text{Värdet på observation } \frac{3\cdot(85+1)}{4} = 64.5\right) = 40.5 \end{aligned}$$

Ett och ett halvt kvartilavstånd ges av

$$1.5 \cdot (40.5 - 28.5) = 18$$

varför uteliggare är observationer under

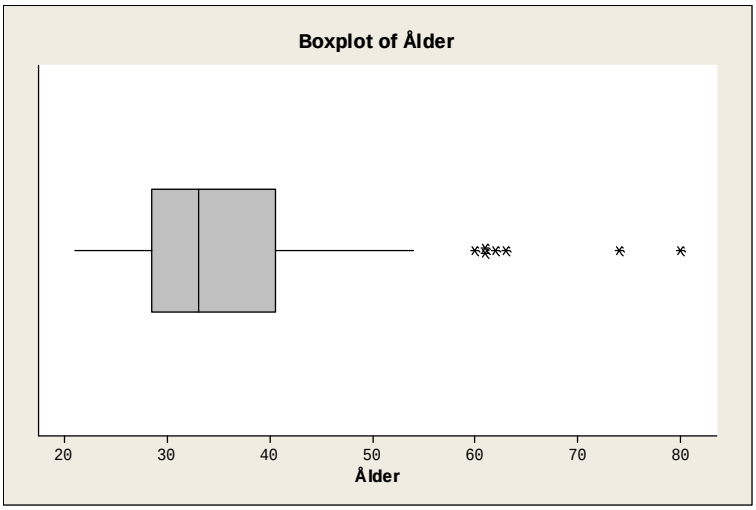
$$28.5 - 18 = 10.5$$

och över

$$40.5 + 18 = 58.5$$

Vi har ett antal uteliggare vilka är samtliga skådespelerskor som vid tilldelningen av priset var 60 år eller äldre. I och med att den äldsta skådespelerskan som

samtidigt inte är en uteliggare var 54 år ska det övre/högra morrhåret dras till denna ålder. Boxploten får därmed följande utseende:



2. I och med att vi ska studera det linjära regressionssamband som beskriver hur biljettintäkterna beror på budgeten gäller att $y = \textit{Biljettintäkter}$ och $x = \textit{Budget}$. Vi börjar med att strukturera de givna summorna.

$$\begin{aligned}
\sum (x-\overline{x})^2 &= 73\,769 - \frac{627^2}{7} = 17\,608 \\
\sum (y-\overline{y})^2 &= 398\,600 - \frac{1\,028^2}{7} = 247\,631 \\
\sum (x-\overline{x})(y-\overline{y}) &= 153\,215 - \frac{627 \cdot 1\,028}{7} = 61\,136
\end{aligned}$$

Utifrån ovanstående summor beräknas regressionskoefficienterna till

$$\begin{aligned}
b &= \frac{61\,136}{17\,608} = \mathbf{3.4721} \\
a &= \frac{1\,028}{7} - 3.4721 \cdot \frac{627}{7} = \mathbf{-164.14}
\end{aligned}$$

vilket innebär att modellen blir

$$\widehat{\mu}_{y|x} = -164 + 3\,.5 \cdot x$$

Eftersom a-koefficienten är en skattning av genomsnittligt y-värde då $x = 0$ ska den här tolkas som att filmer utan budget i genomsnitt ger biljettintäkter på -164 miljoner dollar. Detta är förstås orimligt på flera sätt. En anledning är att det knappast går att göra en film utan någon budget men att vi här får ett så orimligt resultat är att värdet $x = 0$ befinner sig långt utanför undersökningsområdet, dvs det rör sig om en rejäl extrapolation. b-koefficienten är en skattning som ska tolkas som att en extra miljon i budget i genomsnitt ger en ökning av biljettintäkterna med 3.5 miljoner.

3. Går det att statistiskt säkerställa att andelen individer som kan känna smaken av PTC skiljer sig mellan de olika länderna?

   (a) I och med att det är kvalitativa variabler samtidigt som att det är fler än två länder som ingår i jämförelsen måste det bli ett χ^2-test. För att undersöka detta ställer vi upp hypoteserna

   H_0 : Ingen skillnad mellan länderna vad det gäller andelen som kan känna smaken av PTC

   H_1 : Det finns skillnader mellan länderna vad det gäller andelen som kan känna smaken av PTC

   och testar detta på 5% signifikansnivå. Om inga av de förväntade frekvenserna understiger 5 följer att testfunktionen

   $$\chi^2 = \sum\sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

   är χ^2-fördelad med $(4-1)\cdot(2-1) = 3$ frihetsgrader då nollhypotesen är sann. Enligt Tabell 5.4 får vi därmed beslutsregeln att förkasta nollhypotesen först om

   $$\chi^2_{\text{obs}} > \chi^2_{3,0.05} = 7.815$$

   För att kunna beräkna testfunktionens värde behöver vi dom förväntade frekvenserna och en korstabell där både observerade och förväntade frekvenser är angivna får följande utseende

Rows: Smak Columns: Land

	Irland	Portugal	Norge	Italien	All
Bitter	558	345	185	402	1490
	572,2	331,8	194,4	391,7	1490,0
Smaklös	225	109	81	134	549
	210,8	122,2	71,6	144,3	549,0
All	783	454	266	536	2039
	783,0	454,0	266,0	536,0	2039,0

   Hur har vi då funnit de förväntade frekvenserna? Eftersom nollhypotesen förutsätts vara sann ska det inte vara någon skillnad i smakfördelning Bitter/Smaklös

mellan de olika länderna varför vi exempelvis finner förväntat antal som känner den bittra smaken i Irland via

$$E_{L=Irl,S=B} = \frac{783 \cdot 1\,490}{2\,039} = 572.18$$

Ingen av de förväntade frekvenserna understiger 5 vilket innebär att vi kan gå vidare och jämföra observerade och förväntade frekvenser i testfunktionen

$$\chi^2_{\text{obs}} = \frac{(558 - 572.2)^2}{572.2} + \frac{(345 - 331.8)^2}{331.8} + \cdots + \frac{(134 - 144.3)^2}{144.3} = 5.96$$

Eftersom

$$\chi^2_{\text{obs}} = 5.96 < 7.815 = \chi^2_{3,0.05}$$

har vi hamnat i accpetansområdet och accepterar därmed nollhypotesen. Det är alltså på 5% signifikansnivå *inte* statistiskt säkerställt att andelen individer som kan känna smaken av PTC skiljer sig mellan de olika länderna. Om vi istället använder p-värdesmetoden ser vi först att

$$\chi^2_{\text{obs}} = 5.96 < 6.251 = \chi^2_{3,0.1}$$

varför vi drar slutsatsen att

$$p\text{-värde} > 10\%$$

och då p-värdet överstiger den uppsatta signifikansnivån ska nollhypotesen accepteras med samma tolkning som ovan. Exakt p-värde blir 11.4% enligt Minitab.

(b) Tolkningar.

- *Testets signifikansnivå* är risken att göra ett Typ1-fel, dvs att förkasta en korrekt nollhypotes. Denna risk är här bestämd till 5%. Alltså; om det är så att det inte är någon skillnad mellan länderna vad det gäller andelen som kan känna smaken av PTC finns ändå en risk för att vi på grund av slumpmässig variation i vår undersökning får indikationer om att sådana skillnader finns. Risken för att få så starka indikationer att vi blir övertygade om att det verkligen finns skillnader mellan länderna (trots att de egentligen inte är några) är här begränsad till 5%.
- Ett *Typ2-fel* innebär att felaktigt acceptera nollhypotesen. Alltså; om det är så att det finns skillnader mellan länderna vad det gäller andelen som kan känna smaken av PTC men att de indikationer vi i undersökningen får angående detta inte är tillräckligt starka för att förkasta nollhypotesen gör vi således ett Typ2-fel.

4. Statistisk inferens vid jämförelser.

   (a) Den aktuella variabeln är "Expertbedömning av nivå på språkbruk" och denna variabels datanivå är beroende på expertens kunnighet och objektivitet. Eftersom det rör sig om en expert bör vi åtminstone kunna förutsätta att variabeln mäts på en ordinalskala och eventuellt kan vi åtminstone approximativt även betrakta detta som en kvotskala. För att få använda det parametriska t-testet i c-uppgiften krävs (bland annat) att variabeln mäts på kvotskala (så att medelvärden kan beräknas) varför en sådan bedömning av variabelns datanivå blir avgörande. Det icke-pararmetriska testet kräver endast ordinalskala varför detta test bör kunna användas utan några större problem vad det gäller nödvändiga förutsättningar.

   (b) Vi ska i c-uppgiften använda ett parametriskt hypotestest för att jämföra genomsnittlig expertpoäng (vad det gäller språknivå vid återberättelse av saga) för två populationer av barn. Ett sådant parametriskt test rör skillnader i *medelpoäng* i de båda populationerna och baseras på en jämförelse/differens av medelpoäng i stickprov från respektive population. För att vi ska kunna dra slutsatser utifrån differensen $\overline{x}_1 - \overline{x}_2$ förutsätts bl a att de båda stickprovsmedelvärdena är approximativt normalfördelade. Enligt Centrala gränsvärdessatsen uppfylls detta approximativt om de båda stickproven är stora. Här gäller emellertid att de båda stickproven är små varför det måste gälla att variabeln "Expertpoäng" själv är approximativt *normalfördelad* i båda populationerna. Alltså; då samtliga barn i en barnpopulation placeras ut på en skala utifrån deras expertpoäng ska den resulterande kurvan (eller högen av barn) vara mycket lik en normalfördelningskurva. Detta ska gälla för båda barnpopulationerna.

   (c) Nu låter vi

   $$\begin{aligned} \mu_H &= \text{Medelpoäng för barn med hög språkförståelse} \\ \mu_N &= \text{Medelpoäng för barn med normal språkförståelse} \end{aligned}$$

   Utifrån frågeställningen ställer vi upp följande hypoteser

   $$\begin{aligned} H_0 &: \mu_H = \mu_N \\ H_1 &: \mu_H > \mu_N \end{aligned}$$

   vilka vi tänker undersöka med ett hypotestest på 5% signifikansnivå. Vi förutsätter att de båda urvalen är OSU och att de båda urvalen dragits oberoende av varandra. Dessutom förutsätts att den aktuella variabeln mäts på kvotskalan (som diskuterades i b-uppgiften ovan) samt att variabeln även kan betraktas som approximativt normalfördelad (som diskuterades i c-uppgiften ovan). Vidare förutsätts att det är tillräckligt med barn i de båda populationerna för att ändlighetskorrektionen ska kunna bortses från. För att slutligen avgöra vilken testfunktion som ska användas jämförs stickprovsstandardavvikelserna via kvoten

   $$\frac{s_N}{s_H} = \frac{20.01}{13.33} = 1.50 < 2$$

varför vi enligt tumregeln (som är tillförlitlig då stickproven är små) anser att antagandet $\sigma_H = \sigma_N$ inte är orimligt. Vi använder därmed testfunktionen

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

som är t-fördelad med $n_1 + n_2 - 2 = 5 + 5 - 2 = 8$ frihetsgrader då nollhypotesen är sann. Detta innebär att nollhypotesen ska förkastas först om

$$t_{obs} > 1.86 = t_{8,0.05}$$

Utifrån den givna informationen beräknas först den sammanslagna (polade) variansen till

$$s_p^2 = \frac{4 \cdot 13.33^2 + 4 \cdot 20.01^2}{8} = \frac{13.33^2 + 20.01^2}{2} = 289.04$$

varpå testfunktionen får värdet

$$t_{obs} = \frac{76.8 - 51.6}{\sqrt{289.04\left(\frac{1}{5} + \frac{1}{5}\right)}} = 2.34$$

Eftersom

$$t_{obs} = 2.34 > 1.86 = t_{8,0.05}$$

har vi hamnat i det kritiska området och nollhypotesen förkastas. Det är därmed på 5% signifikansnivå statistiskt säkerställt att den genomsnittliga språknivån (med avseende på medelvärde) i en sådan återberättelse är högre för barn som klassificeras som långt framskridna vad det gäller språkförståelse jämfört med barn klassificerade som normalt framskridna beträffande språkförståelse.

(d) Eftersom

$$t_{8,0.025} = 2.306 < t_{obs} = 2.34 < 2.896 = t_{8,0.01}$$

följer att testets p-värde uppskattas till

$$1\% < p\text{-värde} < 2.5\%$$

(Exakt p-värde är 2.4% enligt Minitab.)

(e) För det parametriska testet i c-uppgiften förutsattes först att den aktuella variabeln mäts på kvotskala och dessutom att den kan betraktas som normalfördelad i de båda populationerna, dvs det ställs höga krav på den aktuella variabeln. Om dessa förutsättningar inte verkar vara uppfyllda bör det parametriska t-testet inte användas. Vi kan då istället använda oss av det icke-parametriska test som kallas *Mann-Whitney* (eller *Wilcoxons rangsummatest*). Vi låter nu

$$\begin{aligned} M_H &= \text{Medianpoäng för barn med hög språkförståelse} \\ M_N &= \text{Medianpoäng för barn med normal språkförståelse} \end{aligned}$$

Utifrån frågeställningen ställer vi upp följande hypoteser

$$\begin{aligned} H_0 &: M_H = M_N \\ H_1 &: M_H > M_N \end{aligned}$$

vilka vi tänker undersöka med ett hypotestest på 5% signifikansnivå. Vi förutsätter att de båda urvalen är OSU och att de båda urvalen dragits oberoende av varandra. Dessutom förutsätts att den aktuella variabeln mäts på ordinalskala eller högre (som diskuterades i b-uppgiften ovan) vilket verkar rimligt. Vi har här att $n_H = n_N$ vilket innebär att vi i analysen fritt kan välja testpopulation. Eftersom vi enligt mothypotesen förväntar oss låga rangtal från barnen med normal språkförståelse blir det naturligt att använda den populationen som testpopulation. Nollhypotesen ska förkastas först om

$$R_{obs} < 19 = R_{5,5,0.05}$$

Resultatet av rangordningen blir

Normal språkförståelse	Expertbedömning	77	49	66	28	38
	Rangtal	6	3	5	1	2
Hög språkförståelse	Expertbedömning	80	82	54	79	89
	Rangtal	8	9	4	7	10

vilket innebär att $R_1 = 6 + 3 + 5 + 1 + 2 = 17$ och eftersom

$$R_{obs} = 17 < 19 = R_{5,5,0.05}$$

har vi hamnat i det kritiska området och nollhypotesen förkastas. Det är därmed på 5% signifikansnivå statistiskt säkerställt att den genomsnittliga språknivån (med avseende på medianen) i en sådan återberättelse är högre för barn som klassificeras som långt framskridna vad det gäller språkförståelse jämfört med barn klassificerade som normalt framskridna beträffande språkförståelse. (Minitab ger p-värdet till 1.84%.)

5. Arons sannolikhetsbestyr.

   (a) Eftersom antalet lampor i lådan är begränsat samtidigt som att lampor (förstås) väljs utan återläggning följer att

   $$X = \text{Antal nya lampor bland de utvalda}$$

   är hypergeometriskt fördelad, $Hyp\left(5, \frac{15}{20}, 20\right)$. För att han ska behöva gå tillbaka för att hämta nya lampor får högst tre av de valda vara nya lampor. Sannolikheten för detta ges av

   $$\Pr(X \leq 3) = 1 - \Pr(X \geq 4) = 1 - \frac{\binom{15}{4}\binom{5}{1} + \binom{15}{5}\binom{5}{0}}{\binom{20}{5}} = 1 - 0.634 = \mathbf{0.366}$$

   (b) Eftersom antalet skruvar i lådan är mycket stort följer, trots att dragning sker utan återläggning, att

   $$X = \text{Antal skruvar av typen CT-4A bland de utvalda}$$

   approximativt är binomialfördelad, $Bi(40, 0.2)$. I och med att

   $$np(1-p) = 40 \cdot 0.2 \cdot 0.8 = 6.4 > 5$$

   samtidigt som att

   $$\begin{aligned} E(X) &= 40 \cdot 0.2 = 8 \\ \sigma(X) &= \sqrt{40 \cdot 0.2 \cdot 0.8} = \sqrt{6.4} = 2.53 \end{aligned}$$

   gäller att X approximativt är $N(8, 2.53)$. För att Aron ska kunna laga stolarna måste det bland de skruvar han har med sig finnas åtminstone sex av typen CT-4A. Med kontinuitetskorrektion finner vi den sökta sannolikheten till

   $$\Pr(X \geq 6) \approx \Pr\left(Z \geq \frac{5.5 - 8}{2.53} = -0.99\right) = \mathbf{0.839}$$

   (Den exakta binomiala sannolikheten är 0.8387 vilket innebär att vår approximation är alldeles utmärkt.)

(c) Om vi nu först definierar händelserna

$$\begin{aligned} A_1 &= \text{Den upphittade skruven är av typen CT-4A} \\ A_2 &= \text{Den upphittade skruven är av typen CA-35T} \\ A_3 &= \text{Den upphittade skruven är av typen ST04-X} \end{aligned}$$

och vidare låter

$$B = \text{Den upphittade skruven har rött huvud}$$

följer utifrån den givna informationen i *b*-uppgiften att

$$\begin{aligned} \Pr(A_1) &= 0.2 \\ \Pr(A_2) &= 0.3 \\ \Pr(A_3) &= 0.5 \end{aligned}$$

samt utifrån informationen i denna uppgift att

$$\begin{aligned} \Pr(B \mid A_1) &= 0.15 \\ \Pr(B \mid A_2) &= 0.3 \\ \Pr(B \mid A_3) &= 0.3 \end{aligned}$$

Vi söker nu $\Pr(A_2 \mid B)$ och använder därför Bayes' sats, dvs

$$\Pr(A_2 \mid B) = \frac{\Pr(B \mid A_2)\Pr(A_2)}{\Pr(B)}$$

För att kunna utföra beräkningen behövs $\Pr(B)$, dvs sannolikheten att en upphittad skruv har rött huvud, vilken vi via satsen om total sannolikhet finner till

$$\begin{aligned} \Pr(B) &= \Pr(B \mid A_1)\Pr(A_1) + \Pr(B \mid A_2)\Pr(A_2) + \Pr(B \mid A_3)\Pr(A_3) = \\ &= 0.15 \cdot 0.2 + 0.3 \cdot 0.3 + 0.3 \cdot 0.5 = 0.27 \end{aligned}$$

och därmed följer att

$$\Pr(A_2 \mid B) = \frac{\Pr(B \mid A_2)\Pr(A_2)}{\Pr(B)} = \frac{0.3 \cdot 0.3}{0.27} = \frac{\mathbf{1}}{\mathbf{3}} \approx \mathbf{0.33}$$

(d) För att Skruvtyp och Rött huvud ska vara oberoende av varandra måste det gälla att vetskap om vilket skruvtyp det rör sig om inte påverkar sannolikheten för att skruven har rött huvud. Här gäller dock exempelvis att

$$\Pr(B \mid A_1) = 0.15 \neq 0.3 = \Pr(B \mid A_2)$$

och det står därmed klart att Skruvtyp och Rött huvud *inte* är oberoende av varandra.

TENTAMENSSKRIVNING PÅ KURSERNA

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

2013-11-30

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.

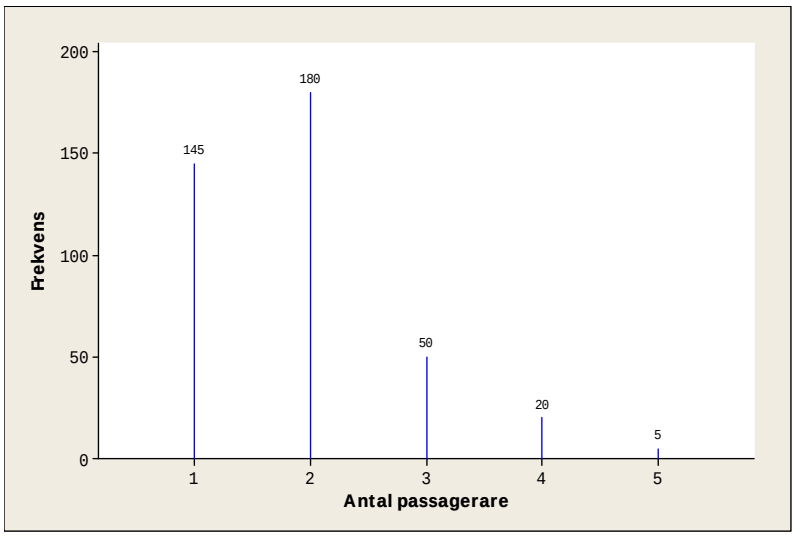
D. Vid hypotesprövning måste Du utöver vad som sägs i punkt B ovan ange vad hypotesprövningen avser att testa, hypoteserna i symbolform (då så är möjligt), signifikansnivå, testfunktion (inklusive antal frihetsgrader då detta är aktuellt) både i symbolform och med numeriska uppgifter, beslutsregel, resultat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

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

Uppgift 1

I samband med en utredning om behovet av taxibilar (ej storbilar) i en mellansvensk stad, registrerades 400 taxifärder med avseende på bland annat antalet passagerare.

Fördelningen över antalet passagerare per färd framgår av följande stolpdiagram:



(4) **A** Ange vad det är 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.

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

(4) **C** Använd variationsvidden till att göra en snabbuppskattning av standardavvikelsen. Då vi jämför denna snabbuppskattning med den faktiska standardavvikelsen som beräknades i B-uppgiften ser vi att de skiljer sig något. En anledning till detta är att snabb-uppskattningen inte är helt lämplig för vårt material. Förklara kortfattat varför det är så, dvs för vilken typ av fördelning fungerar snabbuppskattningen bäst?

(6) **D** Åskådliggör materialet grafiskt genom att konstruera ett lådagram/boxplot.

(8) **E** Använd resultatet i detta urval för att konstruera ett 90% konfidensintervall för andel taxifärder med ensamåkande resenärer.

(12) **F** Två år tidigare var det genomsnittliga antalet passagerare per taxifärd i denna stad 2 (dvs $\mu = 2$). Nu vill man med hjälp av hypotestest undersöka huruvida det genomsnittliga antalet passagerare per taxifärd förändrats under denna tvåårsperiod. Utför ett fullständigt hypotestest med 5% signifikansnivå enligt *p*-värdesmetoden och besvara frågan.

(8) **G** Vi fortsätter nu med situationen i F-uppgiften. Låt oss nu betrakta det aktuella hypotestestet *innan* resultatet av undersökningen sammanställdes, dvs vi har ännu inte några resultat från undersökningen. För att ändå kunna göra beräkningar används $\hat{\sigma}_x$=0.9 (som är en skattning vi fått från den undersökning som gjordes två år tidigare). Anta att det under detta år gjordes totalt 100 000 taxifärder och att det totala antalet passagerare var 195 000. Vad är med denna förutsättning risken för ett Typ2-fel? För full poäng måste situationen beskrivas grafiskt.

(4) **H** Hur stort stickprov hade vi i denna undersökning behövt ta för att vi på 5% signifikansnivå ska få en styrka på 70% om det faktiska genomsnittliga antalet passagerare per taxifärd är som i G-uppgiften ovan?

Uppgift 2

På sista sidan i denna skrivning finns två tabeller från Statistisk årsbok för Sverige 2013. Använd dessa tabeller för att lösa denna uppgift. Vi är intresserade av att studera gruppen kvinnor 20-29 år vad det gäller tidpunkterna 2006, 2008 och 2010.

(7) **A** Konstruera en indexserie som i fasta priser (2010 års priser) visar hur sammanräknad förvärvsinkomst, medianvärden har utvecklats för denna grupp under de angivna tidpunkterna. Använd 2010 som basår och redovisa indextalen med en decimal.

(3) **B** Hur stor (beräknat i fasta priser) har den årliga förändringstakten vad det gäller sammanräknad förvärvsinkomst, medianvärden för den aktuella gruppen i genomsnitt varit under perioden mellan 2006 och 2010?

(12) **Uppgift 3**

Chokladtillverkaren Hershey Company vet från historiska data att 30% av deras kunder föredrar Mr. Goodbar (MrG), 50% föredrar Hershey's Milk Chocolate (HMC), 15% föredrar Hershey's Special Dark Chocolate (HSD) och resterande föredrar Krackel (Kr). För att undersöka om denna uppdelning i preferenser ändrats under senare tid gjordes en undersökning där 200 slumpmässigt valda kunder tillfrågades om vilken chokladbit de tyckte bäst om i sortimentet (bland de ovan nämnda). Resultatet blev följande:

Chokladbit	**MrG**	**HMC**	**HSD**	**Kr**
Antal	50	95	39	16

Är det i och med detta resultat statistiskt säkerställt att uppdelning i preferenser ändrats under senare tid? Utför en fullständig hypotesprövning enligt klassisk metod där du använder 5% signifikansnivå.

Uppgift 4

(5) **A** Betrakta de båda händelserna A och B. Vi har information om att händelsen A inträffar dubbelt så ofta som händelsen B. Vidare ges informationen att *både* A och B inträffar vid 10% av alla försök medan *åtminstone* en av händelserna inträffar vid 80% av alla försök. Ett försök ska precis till att utföras. Bestäm sannolikheten att *endast* B inträffar, dvs att B inträffar men inte A.

(5) **B** En viss elektronisk komponent säljs i partier om 20 komponenter. Före leveransen kontrolleras partierna. Tillverkaren kan dock av kostnadsskäl inte kontrollera samtliga komponenter utan får nöja sig med stickprov om fem komponenter ur varje parti. Ett visst parti innehåller tre defekta komponenter. Vad är sannolikheten att tillverkaren får åtminstone två defekta komponenter i stickprovet?

Uppgift 5

En enkrona ska enligt Riksbanken väga 7 gram men anta att denna vikt i själva verket är en slumpvariabel som kan betraktas som normalfördelad med väntevärde 7 gram och standardavvikelse 0.07 gram.

(4) **A** Förklara innebörden av uttalandet att en kronas vikt är normalfördelad med väntevärde 7 gram och standardavvikelse 0.07 gram.

(3) **B** Beräkna sannolikheten att en slumpmässigt utvald enkrona väger mer än 6.9 gram.

(4) **C** En viss varuautomat accepterar enbart enkronor vars vikt inte avviker för mycket från idealvikten på 7 gram. Den är inställd så att 98% av alla enkronor accepteras. Ange (i gram) hur mycket en enkrona maximalt får avvika från 7 gram för att accepteras av automaten.

(5) **D** Anta att du plötsligt får en oemotståndlig längtan efter choklad och att det enda sättet att stilla detta sug är att köpa en chokladbit ur varuautomaten som beskrevs i C-uppgiften ovan. Du har nio enkronor i fickan och konstaterar att du bara kan köpa den billigaste chokladbiten som kostar 8 kronor. Beräkna sannolikheten att du kan stilla ditt chokladsug. För full poäng måste dina beräkningar motiveras.

13.5 Sammanräknad förvärvsinkomst, medianvärden i löpande priser efter kön och ålder, tkr

Total income from employment and business, median values in current prices by sex and age, SEK thousands

Kön/ålder *Sex/age*	2004[1]	2005[1]	2006[1]	2007[1]	2008[1]	2009[1]	2010[1]
Män och kvinnor *Both sexes*	189,4	192,9	198,8	206,2	215,1	218,7	219,7
16–19 år *years*	4,9	4,5	5,3	6,1	6,3	3,8	4,9
20–29 år	136,0	135,4	142,6	150,8	154,9	138,7	134,6
30–49 år	232,3	238,3	247,3	258,4	270,8	276,4	283,5
50–64 år	236,0	241,7	249,5	259,0	270,3	278,1	284,6
65– år	144,1	148,9	153,5	159,2	167,1	176,6	175,5
20–64 år	219,3	224,2	232,0	241,6	252,6	256,5	261,2
25–64 år	228,8	234,3	242,5	252,8	264,4	270,0	276,1
Män *Men*	221,6	226,2	232,9	242,7	252,6	254,6	257,2
16–19 år *years*	4,3	4,0	4,7	5,5	5,5	2,8	4,2
20–29 år	162,2	161,9	169,7	182,6	189,5	166,2	159,7
30–49 år	265,6	273,1	283,6	297,4	309,6	312,2	321,0
50–64 år	264,2	271,0	279,7	291,4	302,3	307,8	314,7
65– år	174,7	180,2	185,2	191,4	200,6	210,6	209,1
20–64 år	248,5	254,7	263,7	275,6	286,5	288,3	294,2
25–64 år	258,9	265,6	275,2	287,9	299,6	302,3	309,6
Kvinnor *Women*	164,0	167,4	172,6	178,0	185,1	189,6	190,1
16–19 år *years*	5,6	5,0	5,9	6,8	7,3	4,8	5,5
20–29 år	118,4	117,0	123,1	128,2	131,3	120,1	116,6
30–49 år	204,2	209,1	217,0	225,9	238,2	245,2	251,2
50–64 år	213,2	218,5	225,5	233,9	245,6	254,5	260,9
65– år	119,0	122,7	126,9	132,4	139,5	148,3	147,9
20–64 år	195,3	199,4	206,4	214,2	224,7	230,0	233,8
25–64 år	203,6	208,3	215,7	224,2	235,8	242,6	247,9

1) Sammanräknad förvärvsinkomst består av inkomst av tjänst och inkomst av näringsverksamhet. Medianvärden för befolkningen (16 år och äldre) folkbokförd i Sverige både den 1 januari och den 31 december. *Median values for the population (16 years and older) entered in the population register both January 1 and December 31.*

Se Tabellanmärkningar.

Källa: SCB Inkomst- och taxeringsregistret, IoT (bearbetning), Inkomster och skatter (www.scb.se/HE0110).

14.7 Konsumentprisindex (totalindex), månadstal (1980=100), fastställda tal och beräknade årsmedeltal

Consumer price index (total), monthly, fixed index and calculated annual averages

År *Year*	Jan.	Feb.	Mars *March*	April	Maj *May*	Juni *June*	Juli *July*	Aug.	Sept.	Okt. *Oct.*	Nov.	Dec.	Årsmedeltal[1] *Annual average*
1990	199,0	199,9	205,4	205,2	206,4	206,2	208,2	209,6	212,0	213,4	214,1	213,9	207,8
1991	218,9	225,0	225,8	227,1	227,3	227,0	227,1	226,7	229,2	230,1	231,1	230,8	227,2
1992	230,2	230,3	231,3	231,9	232,0	231,5	231,2	231,3	234,6	235,1	234,0	234,9	232,4
1993	241,0	241,6	242,7	243,7	243,1	242,3	241,9	242,3	244,5	245,2	245,3	244,3	243,2
1994	245,1	245,9	246,8	247,8	248,3	248,4	248,4	248,5	250,7	251,0	250,8	250,4	248,5
1995	251,3	252,3	253,3	255,0	255,3	255,1	254,8	254,5	256,2	256,9	256,8	256,0	254,8
1996	255,6	255,8	257,0	257,6	257,3	256,3	255,7	254,5	256,0	255,9	255,3	254,9	256,0
1997	254,6	254,2	255,2	257,0	257,0	257,4	257,3	257,4	259,8	259,6	259,2	259,1	257,3
1998	256,9	256,6	257,0	257,7	258,1	257,6	257,0	255,7	256,8	257,3	256,7	256,2	257,0
1999	256,2	256,3	257,3	257,9	258,3	258,7	257,6	257,6	259,4	259,7	259,0	259,6	258,1
2000	257,5	258,7	259,9	260,0	261,3	261,2	260,0	260,2	262,0	262,6	262,7	262,5	260,7
2001	261,7	262,6	264,6	266,9	268,7	268,3	266,9	267,6	269,9	269,1	269,2	269,5	267,1
2002	268,8	269,4	271,8	272,9	273,6	273,2	272,3	272,4	274,5	275,4	274,7	275,1	272,8
2003	276,0	278,4	279,8	278,8	278,5	277,7	276,8	276,7	278,7	278,9	278,3	278,6	278,1
2004	278,0	277,3	279,4	279,4	280,1	278,9	278,5	278,2	280,2	281,0	279,4	279,4	279,2
2005	277,9	279,2	279,8	280,2	280,3	280,4	279,4	279,9	281,9	282,4	281,7	281,8	280,4
2006	279,59	280,90	282,89	284,32	284,76	284,68	284,19	284,38	286,04	286,07	286,43	286,43	284,22
2007	285,01	286,45	288,33	289,79	289,48	289,95	289,49	289,41	292,30	293,85	295,75	296,32	290,51
2008	294,09	295,28	298,08	299,67	300,99	302,45	302,11	301,98	305,08	305,56	303,06	298,99	300,61
2009	297,88	297,95	298,80	299,26	299,45	300,17	298,80	299,42	300,35	301,11	301,03	301,69	299,66
2010	299,79	301,59	302,32	302,36	302,92	302,97	302,04	302,06	304,60	305,57	306,58	308,73	303,46
2011	306,15	308,02	310,11	311,44	312,02	311,28	311,13	311,23	313,41	313,42	314,16	314,78	311,43
2012	311,85	313,92	314,80	315,49	315,23	314,45	313,23	313,55	314,81	314,59			

1) Årsmedeltalen kan skilja sig från årsmedeltalen i tabell 14.8, se not 1 för tabell 14.8.

Se Tabellanmärkningar.

Källa: SCB Statistiska meddelanden serie P 14 (–1999), PR 14 (2000–), Konsumentprisindex (www.scb.se/PR0101); Statistikdatabasen: Priser och konsumtion.

1. Låt oss börja med utifrån stolpdiagrammet skapa en frekvenstabell.

Antal passagerare (x)	f	fx	fx^2	F
1	145	145	145	145
2	180	360	720	325
3	50	150	450	375
4	20	80	320	395
5	5	25	125	400
	400	**760**	**1 760**	

(a) I den här situationen är det *taxifärder (i den aktuella staden)* som är individer/element och *antal resenärer (per taxifärd)* som är variabel. Eftersom denna variabel endast kan anta heltalsvärden gäller att den är *diskret*. Vidare gäller exempelvis att 4 resenärer är dubbelt så många som 2 resenärer varför det för denna variabels värden är meningsfullt att göra relativa jämförelser. Således mäts variabeln på *kvotskalan*.

(b) Utifrån vår frekvenstabell ovan fås att

$$\overline{x} = \frac{760}{400} = \mathbf{1.9}$$

$$s = \sqrt{\frac{1760 - \frac{760^2}{400}}{399}} = \mathbf{0.89}$$

För taxifärderna i urvalet gällde att det i genomsnitt var 1.9 passagerare. Alla taxifärder hade dock inte samma antal passagerare utan avvek med i genomsnitt 0.9 passagerare från det genomsnittliga antalet passagerare.

(c) En snabbuppskattning av standardavvikelsen fås genom att dela variationsvidden med 4, dvs här får vi

$$s \approx \frac{\max - \min}{4} = \frac{5-1}{4} = 1$$

Ett resonemang kring denna snabbuppskattning finner du i avsnitt 4.4.3 i boken Praktisk Statistik (sid. 106) där vi läser "För inte alltför stora symmetriskt fördelade material kan det därför vara rimligt att räkna med att differensen mellan det största och det minsta värdet – *variationsvidden* – är cirka fyra standardavvikelser.". Problemet här är dels att materialet inte är symmetriskt och dels att materialet inte är litet. Vidare gäller att de värden vår variabel kan anta är mycket begränsat.

(d) För att kunna konstruera ett lådagram behöver vi median och kvartiler.

$$\begin{aligned} q_1 &= \left(\text{Värdet på observation } \frac{400+1}{4} = 100.25\right) = 1 \\ md &= \left(\text{Värdet på observation } \frac{400+1}{2} = 200.5\right) = 2 \\ q_3 &= \left(\text{Värdet på observation } \frac{3\cdot(400+1)}{4} = 300.75\right) = 2 \end{aligned}$$

Ett och ett halvt kvartilavstånd ges av

$$1.5 \cdot (2-1) = 1.5$$

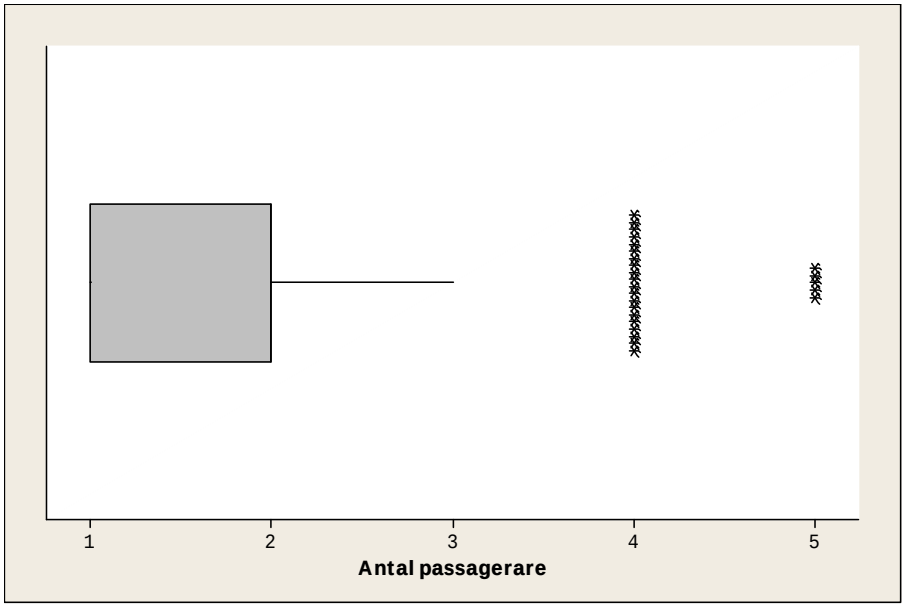
varför uteliggare är observationer under

$$1 - 1.5 = -0.5$$

och över

$$2 + 1.5 = 3.5$$

Vi har därmed inte mindre än 25 uteliggare (20 med värdet 4 och 5 med värdet 5). Lådagrammet får följande utseende



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

$$p = \text{Andel taxifärder med ensamåkande resenärer}$$

Vi har här att

$$\widehat{p} = \text{Andel taxifärder med ensamåkande resenärer i urvalet} = \frac{145}{400} = 0.3625$$

som är vår punktskattning. Vi förutsätter att taxifärderna i urvalet kan betraktas som ett slumpmässigt urval bland alla taxifärder i den aktuella staden och eftersom

$$n\widehat{p}(1-\widehat{p}) = 400 \cdot 0.3625 \cdot 0.6375 = 92.4 >> 5$$

är stickprovet (med god marginal) tillräckligt stort för att normalapproximation av binomialfördelningen ska vara tillåten. Vidare gäller att populationen (av taxifärder) kan antas vara stor vilket betyder att vi kan bortse från ändlighetskorrektion och använda 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

$$0.3625 \pm 1.6449 \cdot \sqrt{\frac{0.3625 \cdot 0.6375}{400}}$$

eller som ett intervall

$$\mathbf{0.323 \leq p \leq 0.402}$$

Med 90% säkerhet befinner sig p, dvs andel taxifärder med ensamåkande resenärer, någonstans mellan 32.3% och 40.2%.

(f) Låter vi först

$\mu =$ Medelvärde för antal passagerare under taxifärder i den aktuella staden

följer av frågeställningen i uppgiften att hypoteserna ska formuleras som

$$\begin{aligned} H_0 &: \mu = 2 \\ H_1 &: \mu \neq 2 \end{aligned}$$

Detta ska nu undersökas med ett test på 5% signifikansnivå. Vi förutsätter som ovan att taxifärderna i urvalet kan betraktas som ett slumpmässigt urval bland alla taxifärder i den aktuella staden samt att populationen (av taxifärder) kan antas vara stor vilket betyder att vi kan bortse från ändlighetskorrektion. Eftersom vi dessutom har att $n = 400 >> 30$ följer att vi kan använda testfunktionen

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

Insättning av våra stickprovsvärden från a-uppgiften ger oss följande värde på testfunktionen

$$z = \frac{1.9 - 2}{0.89/\sqrt{400}} = -2.25$$

vilket utifrån utseendet på mothypotesen innebär att

$$p\text{-värde} = 2 \cdot \Pr\left(Z < -2.25\right) = 2 \cdot 0.0122 = 0.024$$

Eftersom

$$p\text{-värde} = 0.024 < 0.05 = \alpha$$

förkastas nollhypotesen. Det är således på 5% signifikansnivå statistiskt säkerställt att medelvärdet för antal passagerare under taxifärder i den aktuella staden har förändrats sedan den tidigare mätningen.

(g) Detta är en uppgift som måste lösas i två steg. Först måste vi under nollhypotesantagendet, dvs att $\mu = 2$, ta reda på för vilka värden på stickprovsmedelvärdet nollhypotesen kommer att accepteras och sedan måste vi under den nya förutsättningen, dvs att $\mu = 1.95$, ta reda på sannolikheten att detta kommer att inträffa (vilket är risken för ett Typ2-fel dvs β).

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

$$-1.96 < \frac{\overline{x} - 2}{0.9/\sqrt{400}} < 1.96$$

vilket leder till att

$$\overline{x} > 2 - 1.96 \cdot \frac{0.9}{\sqrt{400}} = 1.91$$

samt

$$\overline{x} < 2 + 1.96 \cdot \frac{0.9}{\sqrt{400}} = 2.09$$

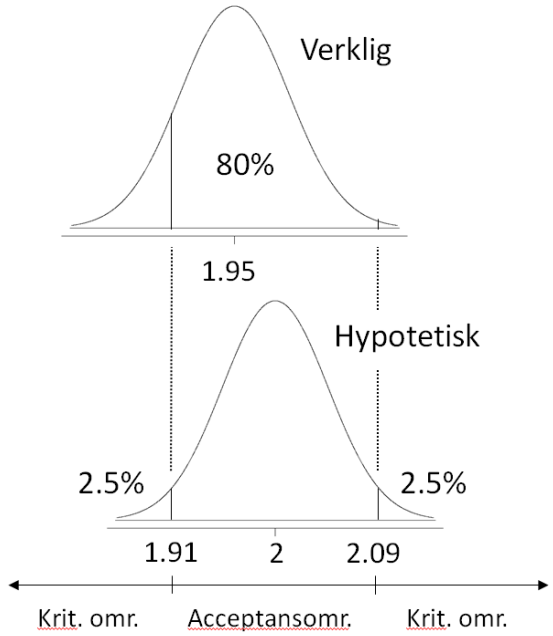
dvs

$$1.91 < \overline{x} < 2.09$$

ii. Vad blir $\Pr\left(1.91 < \overline{X} < 2.09\right)$ under den nya förutsättningen att $\mu = 1.95$, dvs att $\overline{X}$ är $N\left(1.95, 0.9/\sqrt{400}\right)$? På vanligt normalfördelningsmanér uttrycker vi detta i standardavvikelser, dvs

$$\begin{aligned}\Pr\left(1.91 < \overline{X} < 2.09\right) &= \Pr\left(\frac{1.91 - 1.95}{0.9/\sqrt{400}} < Z < \frac{2.09 - 1.95}{0.9/\sqrt{400}}\right) = \\ &= \Pr\left(-0.85 < Z < 3.07\right) = \\ &= 0.9989 - (1 - 0.8023) = \mathbf{0.80}\end{aligned}$$

Risken för ett Typ2-fel, dvs sannolikheten att acceptera en felaktig nollhypotes, blir i den här situationen ca 0.8. Risken för att vi under dessa omständigheter inte kommer att få tillräckligt övertygande bevis om att medelvärdet för antal resenärer under taxifärderna förändrats sedan senaste mätningen för två år sedan är alltså hela 80%. Hela situationen beskrivs väl med följande graf



(h) I G-uppgiften ovan såg vi att risken för ett Typ2-fel (då $\mu = 1.95$) är överhängande då vi använder en stickprovsstorlek på $n = 400$ taxifärder. För att få ner denna risk (dvs få upp testets styrka) behövs ett större urval. Hur stort måste urvalet vara för att få ner β, dvs risken för ett Typ2-fel till 30%? De uppställda kraven innebär att

$$\begin{aligned} z_\alpha &= z_{0.025} = 1.96 \\ z_\beta &= z_{0.3} = 0.5244 \end{aligned}$$

och då vidare

$$\begin{aligned} \mu_0 &= 2 \\ \mu_1 &= 1.95 \\ \widehat{\sigma} &= 0.9 \end{aligned}$$

följer från formelsamlingen att den sökta stickprovsstorleken ges av

$$n = \left(\frac{(1.96 + 0.5244) \cdot 0.9}{1.95 - 2} \right)^2 = \mathbf{1999.8}$$

dvs vi behöver ett stickprov om åtminstone 2 000 taxifärder för att uppnå de uppställda kraven.

2. Index.

(a) Den resulterande sammanställningen för gruppen kvinnor 20–29 år gällande den aktuella variabeln blir

År	2006	2008	2010
Sammanräknad förvärvsinkomst, tkr (Löpande)	123.1	131.3	116.6
KPI	284.22	300.61	303.46
Sammanräknad förvärvsinkomst, tkr (2010 års nivå)	131.4	132.5	116.6
Index	112.7	113.7	100.0

där vi med hjälp av KPI räknat om värdena till 2010 års penningvärde via

$$\begin{array}{ll} 2006 & 123.1 \cdot \frac{303.46}{284.22} = 131.43 \\ 2008 & 131.3 \cdot \frac{303.46}{300.61} = 132.54 \\ 2010 & 116.6 \end{array}$$

som sedan ger oss vår indexserie.

(b) Den procentuella förändringen under perioden 2006 till 2010 var

$$\frac{116.6}{131.4} = 0.887$$

dvs en minskning med totalt 11.3%. Detta innebär att den genomsnittliga årliga förändringen var

$$g = 0.887^{1/4} = 0.971$$

dvs en minskning på ca 2.9%.

3. Här har vi en situation med en kvalitativ variabel med mer än två variabelvärden/kategorier vilket innebär att vi måste använda χ^2-metoden (Goodness of Fit) för vårt hypotestest. Utifrån företagets historiska data ställer vi upp följande hypoteser:

$$\begin{aligned} H_0 &: p_{MrG} = 0.3,\ p_{HMC} = 0.5,\ p_{HSD} = 0.15,\ p_{Kr} = 0.05 \\ H_1 &: \text{Någon annan fördelning} \end{aligned}$$

Enligt informationen kan de 200 kunderna betraktas som ett slumpmässigt urval bland alla företagets kunder. Under förutstättning att preferensfördelningen vad det gäller de olika chokladbitarna är som tidigare förväntar vi oss denna fördelning bland de utvalda kunderna. Slumpen kommer dock som vanligt att ställa till det något så att stickprovsfördelningen inte blir helt perfekt. Frågan är vad sannolikheten är att slumpen ställer till det som i det här fallet. Vi får följande värden på observerade och förväntade frekvenser:

	MrG	HMC	HSD	Kr
Observerade	50	95	39	16
Förväntade	60	100	30	10

och eftersom ingen av de förväntade frekvenserna understiger 5 kan vi utföra testet genom att använda testfunktionen

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$$

som är χ^2-fördelad med 3 frihetsgrader då nollhypotesen är sann. Vi beräknar testfunktionen till

$$\chi^2 = \frac{(50-60)^2}{60} + \frac{(95-100)^2}{100} + \frac{(39-30)^2}{30} + \frac{(16-10)^2}{10} = 8.22$$

Eftersom

$$\chi^2_{\text{obs}} = 8.82 > 7.815 = \chi^2_{3,0.05}$$

har vi hamnat i det kritiska området och därmed förkastas nollhypotesen. Det är således på 5% signifikansnivå statistiskt säkerställt att kundernas preferensfördelning vad det gäller de olika chokladbitarna har förändrats under senare tid.

4. Blandad sannolikhetslära

(a) Uttrycker vi den givna informationen med sannolikhetsterminologi har vi att

$$\begin{aligned} \Pr(A) &= 2 \cdot \Pr(B) \\ \Pr(A \cap B) &= 0.1 \\ \Pr(A \cup B) &= 0.8 \end{aligned}$$

Additionssatsen tillsammans med informationen ovan ger att

$$0.8 = \Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) = 2 \cdot \Pr(B) + \Pr(B) - 0.1$$

vilket alltså innebär att $3 \cdot \Pr(B) = 0.9$ eller ekvivalent att $\Pr(B) = 0.3$. Att endast händelsen B inträffar är (rita Venn-diagram)

$$\Pr(\text{Endast } B) = \Pr\left(B \cap \overline{A}\right) = \Pr(B) - \Pr(A \cap B) = 0.3 - 0.1 = \mathbf{0.2}$$

(b) Låter vi

$$X = \text{Antal defekta bland de utvalda}$$

följer enligt förutsättningarna att X är $Hyp\left(5, \frac{3}{20}, 20\right)$. Vi söker

$$\Pr(X \geq 2) = \Pr(X = 2) + \Pr(X = 3) = \frac{\binom{3}{2}\binom{17}{3}}{\binom{20}{5}} + \frac{\binom{3}{3}\binom{17}{2}}{\binom{20}{5}} = \mathbf{0.14}$$

5. Vi betraktar nu slumpvariabeln

$$X = \text{Vikten av en slumpmässigt vald enkrona}$$

vilken enligt förutsättningarna är $N(7, 0.07)$ där enheten är gram.

(a) Vad är innebörden av att X är $N(7, 0.07)$? Om alla enkronor som finns skulle läggas i en hög där myntens placering från vänster till höger beror på dess vikt (och mynt som väger ungefär lika mycket placeras ovanpå varandra) skulle denna hög få formen av en normalfördelningskurva. Myntens genomsnittliga vikt är 7 gram (mynthögens mitt) och myntens genomsnittliga avvikelse till idealvikten är 0.07 gram (mynthögens utbredning).

(b) Vi söker

$$\Pr(X > 6.9) = \Pr\left(Z > \frac{6.9 - 7}{0.07} = -1.43\right) \approx \mathbf{0.924}$$

(c) Automaten är inställd så att den ska accepterar alla enkronor förutom de 1% tyngsta mynten och de 1% lättaste mynten. Enligt Tabell 5.2.B gäller att

$$z_{0.01} = 2.3263$$

vilket betyder att automaten accepterar alla mynt som inte avviker med mer än 2.3263 standardavvikelser från idealvikten. Detta innebär att vikten för en enkrona maximalt får avvika med $2.3263 \cdot 0.07 = \mathbf{0.16}$ gram för att accepteras av automaten.

(d) Låt

$$Y = \text{Antal av dina enkronor som accepteras av automaten}$$

Förutsätter vi att vikten hos de enkronor du har kan betraktas som oberoende av varandra följer av förutsättningarna från c-uppgiften att Y är $Bi(9, 0.98)$. Vi söker

$$\begin{aligned} \Pr(Y \geq 8) &= \Pr(Y = 8) + \Pr(Y = 9) = \\ &= \binom{9}{8} \cdot 0.98^8 \cdot 0.02^1 + \binom{9}{9} \cdot 0.98^9 \cdot 0.02^0 = \mathbf{0.987} \end{aligned}$$

dvs det är goda chanser att ditt chokladsug stillas.