TENTAMENSSKRIVNING PÅ KURSERNA

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

2015-03-26

UPPLYSNINGAR

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

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

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

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

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

UPPMANINGAR

A. Följ noga de anvisningar som finns på skrivningsförsättsbladet.

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

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

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

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

Uppgift 1

Åldersfördelningen för 15 slumpmässigt utvalda racerförare återges i stam-blad-diagrammet nedan.

```
Stem-and-leaf of Ålder  N  = 15
Leaf Unit = 1.0

 1    2  3
 5    2  5799
(6)   3  222333
 4    3  8
 3    4  1
 2    4  5
 1    5  2
```

Som stöd för fortsatta beräkningar har vi även att $\Sigma x = 504$ och $\Sigma x^2 = 17\,758$.

(5) **A** Använd den givna informationen till att beräkna medelvärde och standardavvikelse för den aktuella variabeln. Ge en ordentlig förklaring av innebörden av de värden du beräknat.

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

(12) **C** Undersök med ett hypotestest angående p om medianåldern för racerförare överstiger 30 år. Använd en signifikansnivå på 5% och utför testet enligt p-värdesmetoden. *Observera att normalfördelningen inte får användas för detta test.*

(4) **D** Anta att vi i C-uppgiften istället utför hypotestestet enligt klassisk metod. Ange det kritiska området för testet uttryckt i termer av den testfunktion du använde i C-uppgiften.

(4) **E** I F-uppgiften nedan görs ett normalfördelningsantagande. Vad är det som måste vara normalfördelat? Varför är antagandet nödvändigt? Förklara kortfattat hur man utifrån stam-blad-diagrammet ovan kan göra en rimlighetsbedömning av detta antagande.

(8) **F** Trots eventuell tveksamhet (baserat på det vi såg E-uppgiften ovan) anser vi att normalfördelningsantagandet är rimligt. Använd resultatet i detta urval för att konstruera ett intervall som med 90% säkerhet täcker in medelåldern för racerförare.

Uppgift 2 Källa: *www.SCB.se*

Tabellen nedan visar Konsumentprisindex (1980=100), årsmedeltal, under perioden 2008-2013.

År	2008	2009	2010	2011	2012	2013
KPI	300,61	299,66	303,46	311,43	314,20	314,06

(3) **A** Med hur många procent har prisnivån (KPI) förändrats från mitten av 2008 till mitten av 2013? Beräkna också den genomsnittliga årliga procentuella förändringen av KPI under denna tidsperiod. (Svara med två decimalers noggrannhet).

(2) **B** Hur har kronans köpkraft förändrats under denna tidsperiod? Besvara frågan genom att ange kronans köpkraft 2013 i 2008 års prisnivå. (Svara med två decimalers noggrannhet).

(6) **C** Den genomsnittliga månadslönen för anställda inom statlig sektor var år 2008 29100 kr och för år 2013 33300 kr. Med hur många procent förändrades genonsnittslönen under denna period (2008-2013)? Beräkna också den genomsnittliga årliga procentuella förändringen av månadslönen under denna tidsperiod. (Svara med två decimalers noggrannhet).

Beräkna den reala genomsnittslönen (*fast pris*) för år 2013 i 2008 års prisnivå.
Med hur många procent förändrades den reala genomsnittslönen under denna period (2008-2013)? Beräkna också den genomsnittliga årliga procentuella förändringen av den reala månadslönen under denna tidsperiod. (Svara med två decimalers noggrannhet).

Uppgift 3

Pratar män mindre än kvinnor? Detta var frågeställningen i artikeln *"Are Women Really More Talkative Than Men?"* av *Mehl, et al., Science, Vol 317, No. 5834)*. I undersökningen studerades ett antal män och kvinnor där det mättes hur många ord som sades under en dag. Minitabutskriften nedan ger en sammanfattning av resultatet av undersökningen.

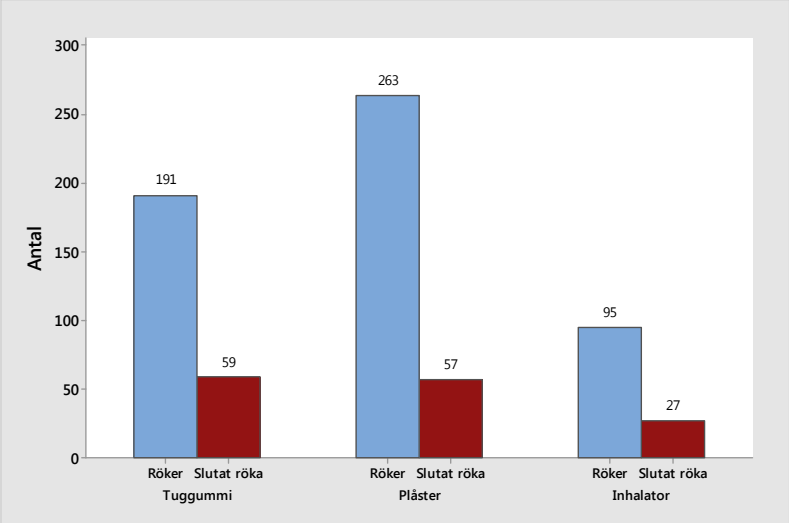
Variable	N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Män	186	15669	633	8633	695	9997	14290	20607	47016
Kvinnor	210	16215	504	7301	1674	10964	15917	20587	40055

(12) **A** Undersök med ett fullständigt hypotestest enligt klassisk metod på 10% signifikansnivå om det går att påvisa någon skillnad i spridning vad det gäller hur många ord som sägs under en dag, dvs skillnad i standardavvikelse/varians, mellan män och kvinnor.

(12) **B** *Pratar män mindre än kvinnor?* Besvara frågan genom att utföra ett fullständigt hypotestest enligt p-värdesmetoden på 5% signifikansnivå där du utnyttjar det du kommit fram till i A-uppgiften.

(12) **Uppgift 4**

I en undersökning av personer som röker men som vill sluta röka var man intresserad av huruvida det finns någon skillnad i effektivititet hos vanligt förekommande behandlingar. De behandlingar som studerades var nikotintuggummi, nikotinplåster samt nikotininhalator. Stapeldiagrammet nedan åskådliggör hur de undersökta personerna, uppdelade på de olika behandlingsmetoderna, lyckats respektive misslyckats med att sluta röka efter fem månaders behandling.



Undersök med ett hypotestest på 5% signifikansnivå om det går att statistiskt säkerställa att effektiviteten hos de tre behandlingsmetoderna skiljer sig åt.

Uppgift 5

Födelsevikten hos barn (födda i USA) kan betraktas som approximativt normalfördelad med medelvärde *3 369 g* och standardavvikelse *567 g* (enligt *”Comparison of Birth Weight Distributions between Chinese and Caucasian Infants”, av Wen, Kramer, Usher, American Journal of Epidemiology, Vol. 172, No 10*).

(4) **A** En definition av för tidigt född är att födelsevikten understiger *2 500 g*. Bestäm sannolikheten att ett slumpmässigt valt nyfött barn klassificeras som för tidigt född enligt denna definition.

(3) **B** En annan definition av för tidigt född är att födelsevikten är bland de 10% som väger minst. Bestäm den vikt som enligt denna definition anger gränsen för att klassificeras om för tidigt född.

(6) **C** Bestäm, genom att göra en lämplig approximation, sannolikheten att åtminstone 10 av 80 nyfödda barn klassificeras som för tidigt födda enligt definitionen i B-uppgiften ovan.

1. Statistik i samband med åldersfördelning av racerförare.

   (a) Utifrån den givna informationen finner vi att det för de undersökta i gruppen av racerförare gäller att

   $$\overline{x} = \frac{\sum x}{n} = \frac{504}{15} = 33.6$$

   och

   $$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}} = \sqrt{\frac{17\,758 - \frac{504^2}{15}}{14}} = 7.67$$

   För de racerförare som var med i undesökningen gäller alltså att medelåldern var 33.6 år. Samtliga i gruppen var dock inte lika gamla utan avvek med i genomsnitt ca 7.7 år från denna medelålder.

   (b) För att kunna konstruera ett lådagram behöver vi median och kvartiler. Utifrån informationen i stam-blad-diagrammet finner vi att

   $$\begin{aligned} q_1 &= \left(\text{Värdet på observation } \frac{15+1}{4} = 4\right) = 29 \\ md &= \left(\text{Värdet på observation } \frac{15+1}{2} = 8\right) = 32 \\ q_3 &= \left(\text{Värdet på observation } \frac{3\cdot(15+1)}{4} = 12\right) = 38 \end{aligned}$$

   Ett och ett halvt kvartilavstånd ges av

   $$1.5 \cdot (38 - 29) = 13.5$$

   varför uteliggare är observationer under

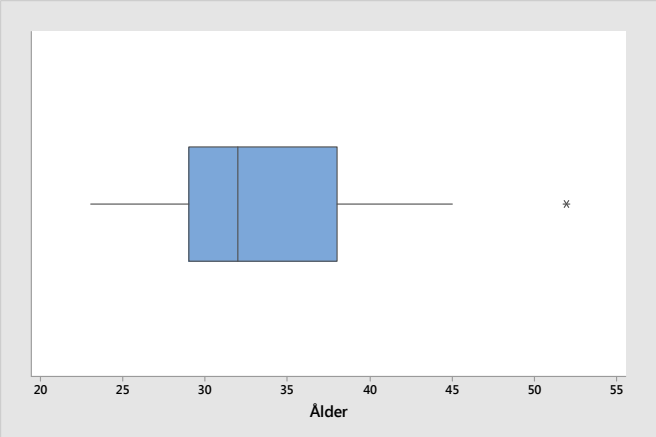
   $$29 - 13.5 = 15.5$$

   och över

   $$38 + 13.5 = 51.5$$

   Vi konstaterar att den racerförare som är 52 år gammal är en uteliggare i vårt

material. Lådagrammet får därmed följande utseende



(c) Vi börjar med att konstatera att frågan huruvida medianåldern för racerförare överstiger 30 år kan ställas upp som en frågeställning angående p. Låter vi

$$p = \text{Andel racerförare som är äldre än 30 år}$$

formuleras hypoteserna utifrån frågeställningen på följande sätt:

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

Detta test ska utföras på 5% signifikansnivå. Stickprovet är ett slumpmässigt urval men eftersom

$$np_0\left(1-p_0\right) = 15 \cdot 0.5 \cdot 0.5 = 3.75 < 5$$

är stickprovet inte tillräckligt stort för att normalapproximation av binomialfördelningen ska vara tillåten. Därmed använder vi testfunktionen

$$X = \text{Antal racerförare i urvalet som överstiger 30 år}$$

Eftersom populationen av racerförare kan betraktas som mycket stor gäller att X är $Bi\left(15, 0.5\right)$ då nollhypotesen är sann. I och med att vi utför testet med p-värdesmetoden och använder en signifikansnivå på 5% ska nollhypotesen förkastas om p-värdet understiger 5%. Eftersom det i urvalet var 10 racerförare som var äldre än 30 år följer av tabellen över binomialfördelningen att p-värdet blir

$$p\text{-värde} = \Pr\left(X \geq 10\right) = 1 - \Pr\left(X \leq 9\right) = 1 - 0.8491 = \mathbf{0.1509}$$

Eftersom

$$p\text{-värde} = 0.1509 > 0.05 = \alpha$$

ska nollhypotesen accepteras. Det är alltså på 5% signifikansnivå inte statistiskt säkerställt att medianåldern för racerförare överstiger 30 år.

(d) Hur många av racerförarna i urvalet hade behövt vara över 30 år för att vi på 5% signifikansnivå skulle ha förkastat nollhypotesen? Vi får testa oss fram. Uppenbarligen var inte 10 tillräckligt. Om vi låter $X = 11$ följer att

$$\alpha = \Pr(X \geq 11) = 1 - \Pr(X \leq 10) = 1 - 0.9408 = 0.0592$$

vilket är en signifikansnivå över 5%. Vi får därmed söka vidare och om vi låter $X = 12$ följer att

$$\alpha = \Pr(X \geq 12) = 1 - \Pr(X \leq 11) = 1 - 0.9824 = 0.0176$$

och detta kan vi använda oss av. Det följer således att det kritiska området för testet ges av $\mathbf{X \geq 12}$.

(e) I den här situationen måste vi förutsätta att populationen är approximativt normalfördelad med avseende på den aktuella variabeln, dvs åldersfördelningen för racerförare måste uppvisa en likhet med en normalfördelning. Anta att vi, rent hypotetiskt, har tillgång till den faktiska åldersfördelningen för samtliga racerförare och sedan beskriver denna 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 f-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. Hur gör vi då en rimlighetsbedömning? Vi har endast tillgång till åldersfördelningen för de racerförare i urvalet och det är utifrån denna fördelning vi får göra vår rimlighetsbedömning. Ett stam-blad-diagram är ett liggande histogram och vi gör en bedömning om detta är symmetriskt och normalfördelningslikt. 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. Vårt stam-blad-diagram uppvisar en viss asymmetri vilket gör oss tveksamma till om åldersfördelningen för racerförare kan antas vara normalfördelad. Dock gäller att vi har tillgång till 15 observationer vilket innebär att populationen inte måste vara perfekt symmetriskt och normalfördelningslikt för att metoden ska vara tillförlitlig.

(f) Vi ska konstruera ett 90% konfidensintervall för μ där

$$\mu = \text{Medelålder för racerförare}$$

I och med att $n = 15 < 30$ har vi ett för litet stickprov för att utan vidare använda Centrala gränsvärdessatsen. Vi bör därför göra en inledande kontroll av vårt material och försäkra oss om att variabeln, dvs åldersfördelningen för racerförare är (någorlunda) symmetriskt fördelad. Det var detta vi gjorde i *e*-uppgiften ovan och trots en viss tveksamhet beslutar vi oss för att gå vidare. Det verkar rimligt att utgå från att populationen av racerförare är stor vilket innebär att ändlighetskorrektion kan bortses från. Vidare står i uppgiften att de 15 racerförarna i urvalet är slumpmässigt valda vilket innebär att vi kan använda konfidensintervallet

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

Här gäller att det aktuella tabellvärdet blir $t_{14,0.05} = 1.761$ vilket innebär att konfidensintervallet efter insättning av våra värden från *a*-uppgiften blir

$$33.6 \pm 1.761 \cdot \frac{7.67}{\sqrt{15}}$$

eller som ett intervall

$$\mathbf{30.1} \leq \boldsymbol{\mu} \leq \mathbf{37.1}$$

Med 90% säkerhet befinner sig μ, dvs medelåldern för racerförare, någonstans mellan 30.1 år och 37.1 år.

Uppgift 2

A $KPI_{2008} = 300{,}61$ $KPI_{2013} = 314{,}06$

Förändring: $\frac{KPI_{2013}}{KPI_{2008}} = \frac{314{,}06}{300{,}61} = 1{,}04474$

Procentuell förändring: (1,04474-1)*100% **= 4,47% (förändring under 5 år)**

Genomsnittlig förändring per år: $\left(\frac{KPI_{2013}}{KPI_{2008}}\right)^{1/5} = \left(\frac{314{,}06}{300{,}61}\right)^{0{,}2} = 1{,}0088$

Genomsnittlig procentuell förändring per år: (1,0088-1)*100% **= 0,88 % (per år)**

B Kronans köpkraft: $\frac{KPI_{2008}}{KPI_{2013}} = \frac{300{,}61}{314{,}06} \approx 0{,}96$ 1 kr 2008 är värd **0,96 kr** år 2013.

C Förändring: $\frac{Lön_{2013}}{Lön_{2008}} = \frac{33300kr}{29100kr} = 1{,}1443$

Procentuell förändring: (1,1443-1)*100% **= 14,43% (förändring under 5 år)**

Genomsnittlig förändring per år: $\left(\frac{Lön_{2013}}{Lön_{2008}}\right)^{1/5} = \left(\frac{33300kr}{29100kr}\right)^{0{,}2} = 1{,}0273$

Genomsnittlig procentuell förändring per år: (1,0273-1)*100% **= 2,73% (per år)**

Fast pris år *t*, prisnivå år t_0 = löpande pris år $t \cdot \frac{KPI_{t_0}}{KPI_t}$

Fast pris (reallön) år *2013*, prisnivå år 2008 = $33300 \cdot \frac{300{,}61}{314{,}06} = 31873{,}89$ dvs ca **31 874 kr**

Förändring: $\left(\frac{\text{Reallön}_{2013}}{\text{Lön}_{2008}}\right) = \frac{31873{,}89\ kr}{29100\ kr} = 1{,}0953$

Procentuell förändring: (1,0953-1)*100% **= 9,53% (förändring under 5 år)**

Genomsnittlig förändring per år: $\left(\frac{\text{Reallön}_{2013}}{\text{Lön}_{2008}}\right)^{1/5} = \left(\frac{31873{,}89kr}{29100kr}\right)^{0{,}2} = 1{,}0184$

Genomsnittlig procentuell förändring per år: (1,0184-1)*100% **= 1,84% (per år)**

I genomsnitt ökade KPI med 0,88% per år, den genomsnittliga månadslönen inom statlig sektor 2,73% per år och motsvarande reallön 1,84% per år (2008-2013).

3. Statistisk inferens i samband med två oberoende stickprov.

   (a) Vi låter

   $$\begin{aligned} \sigma_m &= \text{Standardavvikelse för antal ord per dag för män} \\ \sigma_k &= \text{Standardavvikelse för antal ord per dag för kvinnor} \end{aligned}$$

   Vi är nu intresserade av att testa hypoteserna

   $$\begin{aligned} H_0 &: \sigma_m = \sigma_k \\ H_1 &: \sigma_m \neq \sigma_k \end{aligned}$$

   Vi förutsätter att båda stickproven är OSU dragna oberoende av varandra och att antal ord som sägs under en dag är approximativt normalfördelat i båda populationerna. Vi ska använda testfunktionen

   $$F = \frac{S_1^2}{S_2^2}$$

   som är F-fördelad med $n_1 - 1$ frihetsgrader i täljaren och $n_2 - 1$ frihetsgrader i nämnaren då nollhypotesen är sann. Utifrån den givna informationen finner vi att

   $$F = \frac{S_m^2}{S_k^2} = \frac{8\,633^2}{7\,301^2} = 1.398$$

   där, pga F-tabellens begränsningar, den största av varianserna ställts i täljaren, dvs det är männen som är population 1 och kvinnorna population 2. Vi jämför detta värde med F-fördelningen med 185 frihetsgrader i täljaren och 209 frihetsgrader i nämnaren. Denna kombination av frihetsgrader finns inte med i F-tabellen men via en (grov) interpolation bör den kritiska punkten vara mindre än

   $$F_{120,120,5\%} = 1.35$$

   och eftersom

   $$F_{obs} = 1.398 > 1.35 = F_{120,120,5\%} > F_{185,209,5\%}$$

   har vi hamnat i det kritiska området och nollhypotesen förkastas. Vi har på 10%-nivån funnit tillsynes tydliga tecken på att spridningen i antal ord som sägs under en dag skiljer sig mellan män och kvinnor (i den undersökta populationen).

   (b) Låter vi μ_m och μ_k representera det genomsnittliga antalet ord som sägs per dag för män respektive kvinnor (i den undersökta populationen) kan våra hypoteser formuleras på följande sätt.

   $$\begin{aligned} H_0 &: \mu_m = \mu_k \\ H_1 &: \mu_m < \mu_k \end{aligned}$$

vilka vi tänker undersöka med ett hypotestest på 5% signifikansnivå. Vi förutsätter att båda stickproven är OSU dragna oberoende av varandra. Eftersom båda urvalen är förhållandevis stora behöver vi nu *inte*, vilket var ett krav i *a*-uppgiften, förutsätta att antalet ord som sägs under en dag är approximativt normalfördelat i båda populationerna. De båda populationerna bör rimligtvis kunna anses vara mycket stora varför ändlighetskorrektion bortses från. Eftersom vi i *a*-uppgiften fann att det med relativt stor säkerhet gäller att $\sigma_A \neq \sigma_B$ används (i och med att stickproven är stora) testfunktionen

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

som approximativt är $N(0,1)$ då nollhypotesen är sann. I och med att vi utför testet med p-värdesmetoden och använder en signifikansnivå på 5% ska nollhypotesen förkastas om p-värdet understiger 5%. Testfunktionen får värdet

$$z = \frac{15\,669 - 16\,215}{\sqrt{\frac{8\,633^2}{186} + \frac{7\,301^2}{210}}} = -0.675$$

Utifrån Z-tabellen får vi att

$$p\text{-värde} = \Pr(Z < -0.675) \approx \mathbf{0.25}$$

Eftersom p-värdet överstiger den uppsatta signifikansnivån på 5% accepteras nollhypotesen. Det är således på 5% signifikansnivå inte statistiskt säkerställt att män pratar mindre än kvinnor (vad det gäller antal ord per dag).

4. Går det att statistiskt säkerställa att att effektiviteten hos de tre behandlingsmetoderna skiljer sig åt? I och med att det är kvalitativa variabler samtidigt som att det är fler än två behandlingsmetoder som ingår i jämförelsen måste det bli ett χ^2-test med test för oberoende. För att undersöka detta ställer vi upp hypoteserna

   H_0 : Ingen skillnad i effektivitet mellan behandlingsmetoderna vad det gäller att få rökare att sluta röka (inget samband)

   H_1 : Det finns skillnader i effektivitet mellan behandlingsmetoderna vad det gäller att få rökare att sluta röka (finns ett samband)

   och testar detta på 5% signifikansnivå. Om rökarna i stickprovet kan betraktas som ett slumpmässigt urval av rökare (som vill sluta röka) och 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 $(3-1)\cdot(2-1)=2$ frihetsgrader då nollhypotesen är sann. Enligt χ^2-tabellen får vi därmed beslutsregeln att förkasta nollhypotesen först om

$$\chi^2_{\text{obs}} > \chi^2_{2,0.05} = 5.991$$

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: Utfall Columns: Metod

	Inhalator	Plåster	Tuggummi	All
Röker	95	263	191	549
	96.79	253.87	198.34	
Slutat röka	27	57	59	143
	25.21	66.13	51.66	
All	122	320	250	692

Hur har vi då funnit de förväntade frekvenserna? Eftersom nollhypotesen förutsätts vara sann ska det inte vara någon skillnad i effektivitet (andel som lyckas sluta röka) hos de olika behandlingsmetoderna varför vi exempelvis finner förväntat antal som inte lyckas sluta röka bland de som fått använda nikotininhalator via

$$E_{\text{Röker, Inhalator}} = \frac{122\cdot 549}{692} = 96.79$$

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

$$\chi^2_{\text{obs}} = \frac{(95-96.79)^2}{96.79} + \frac{(27-25.21)^2}{25.21} + \cdots + \frac{(59-51.66)^2}{51.66} = 3.06$$

Eftersom

$$\chi^2_{\text{obs}} = 3.06 < 5.991 = \chi^2_{2,0.05}$$

har vi hamnat i accpetansområdet och accepterar därmed nollhypotesen. Det är alltså på 5% signifikansnivå *inte* statistiskt säkerställt att det finns skillnader i effektivitet mellan behandlingsmetoderna vad det gäller att få rökare att sluta röka. Om vi istället använder p-värdesmetoden ser vi först att

$$\chi^2_{\text{obs}} = 3.06 < 4.605 = \chi^2_{2,0.1}$$

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 enligt Minitab 21.6%.

5. Vi betraktar nu slumpvariabeln

$$X = \text{Födelsevikt hos ett slumpmässigt vald barn}$$

som enligt den givna informationen åtminstone approximativt kan betraktas som $N(3\,369, 567)$ där enheten är gram.

(a) Vi söker nu

$$\Pr(X < 2\,500) = \Pr\left(Z < \frac{2\,500 - 3\,369}{567} = -1.53\right) = 1 - \Pr(Z < 1.53) \approx \mathbf{0.063}$$

Vi tolkar detta som att ungefär 6.3% av alla nyfödda barn klassificeras som för tidigt födda enligt denna definition.

(b) Enligt Tabell 5.2.B gäller att

$$z_{0.9} = -1.2816$$

vilket innebär att den första decilen, dvs d_1, ges av

$$d_1 = 3\,369 - 1.2816 \cdot 567 = \mathbf{2\,642.3}$$

Enligt denna definition betraktas nyfödda barn med en födelsevikt på 2642 gram eller mindre som för tidigt födda.

(c) Låt

$$Y = \text{Antal för tidigt födda barn i urvalet (enligt definition i } b\text{-uppgiften)}$$

Sannolikheten att ett slumpmässigt valt nyfött barn klassificeras som för tidigt född är 0.1. Ett nyfött barn klassificeras antingen som för tidigt fött eller inte för tidigt fött. Vi utgår från att olika barn är för tidigt födda oberoende av varandra. Då vidare vår slumpvariabel räknar antalet för tidigt födda barn i urvalet följer att Y är $Bi(80, 0.1)$. Eftersom

$$np(1-p) = 80 \cdot 0.1 \cdot 0.9 = 7.2 > 5$$

gäller att normalapproximation av binomialfördelningen är tillåten och eftersom

$$\begin{aligned} E(Y) &= np = 80 \cdot 0.1 = 8 \\ Var(Y) &= 80 \cdot 0.1 \cdot 0.9 = 7.2 \end{aligned}$$

har vi att Y approximativt är $N\left(8, \sqrt{7.2}\right)$. Med kontinuitetskorrektion finner vi nu den sökta sannolikheten till

$$\Pr(Y \geq 10) \approx \Pr\left(Z \geq \frac{9.5 - 8}{\sqrt{7.2}} = 0.56\right) = 1 - \Pr(Z < 0.56) = \mathbf{0.288}$$

Den exakta sannolikheten enligt binomialfördelningen är 0.277 vilket innebär att approximationen är hyfsad.

TENTAMENSSKRIVNING PÅ KURSERNA

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

2015-04-25

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 en större kommun genomfördes en filmfestival. Efter festivalen ringde man runt till 200 slumpmässigt valda personer av kommunens invånare. Dessa fick bland annat svara på frågan hur många av festivalens filmer de tittat på.

Antal filmer	Antal personer
0	144
1	16
2	20
3	13
4	4
5	2
6	1

(5) **A** Beräkna medelvärde och standardavvikelse för antal filmer personerna i urvalet sett. Ge en ordentlig förklaring av innebörden av den framräknade standardavvikelsen.

(7) **B** Åskådliggör materialet grafiskt genom att konstruera ett lådagram.

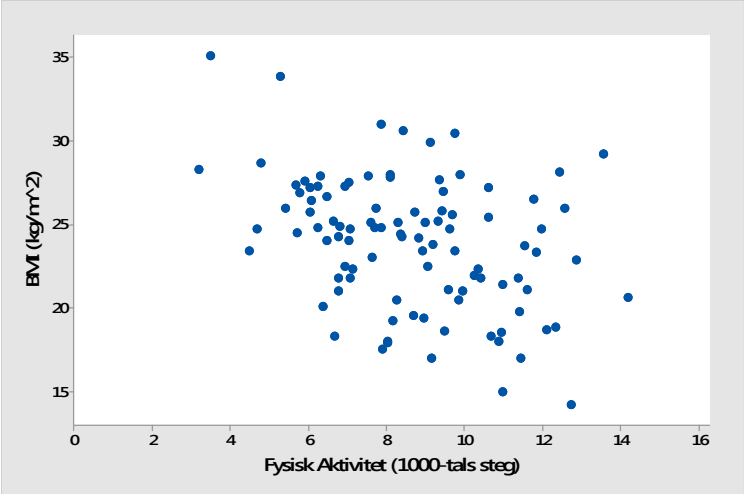
(8) **C** Konstruera ett 90 % konfidensintervall för det genomsnittliga antalet filmer kommunens invånare sett under festivalen.

(12) **D** Om det går att statistiskt säkerställa att mer än var fjärde kommuninvånare besökt filmfestivalen (på så sätt att de tittat på *åtminstone* en film) beslutar kommunledningen att festivalen ska få utökad budget nästa år. Kommer festivalen att få utökad budget nästa år? Ställ utifrån frågeställningen upp hypoteser och utför ett fullständigt hypotestest på 5 % signifikansnivå.

(8) **E** Vi fortsätter nu med situationen i D-uppgiften. Låt oss betrakta det aktuella hypotestestet *innan* resultatet av undersökningen sammanställdes, dvs vi har ännu inte några resultat från undersökningen. Anta att den faktiska andelen kommuninvånare som besökt festivalen är 29 %. Vad är med denna förutsättning styrkan för testet i D-uppgiften? *Anmärkning*. För full poäng måste situationen beskrivas grafiskt.

Uppgift 2

Minskningen av fysisk aktivitet anses vara en viktig bidragande orsak till den ökade förekomsten av övervikt och fetma hos den vuxna befolkningen. Eftersom förekomsten av fysisk inaktivitet bland universitetsstudenter liknar den hos den allmänna vuxna befolkningen, fokuserar många undersökningar på att få en tydligare förståelse för universitets-studenternas fysiska aktivitet. Som en del av en studie har forskare tittat på sambandet mellan fysisk aktivitet (FA) mätt med en stegräknare och Body Mass Index (BMI). Varje deltagare bar en stegräknare i en vecka, och det genomsnittliga antalet steg per dag (i tusental) registrerades. Vidare mättes bl a BMI (i kg per kvadratmeter, kg/m^2). Vi studerar här ett urval av 100 kvinnliga studenter och nedan finner du både en grafisk beskrivning av sambandet och även den Minitabutskrift som återger den skattade regressionsekvationen.



Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
3.65488	14.85%	13.99%	10.81%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	29.58	1.41	20.95	0.000
FA	-0.655	0.158	-4.13	0.000

(5) **A** Tolka på ett begripligt sätt (dvs utan att använda statistiska facktermer) de båda regressionskoefficienterna i ord. Du ska alltså tolka både interceptterm och riktningskoefficient. Ange en orsak till varför tolkningen av interceptet bör tas med en nypa salt.

(2) **B** Uppskatta med hjälp av regressionsmodellen det BMI vi förväntar oss för personer med en fysisk aktivitet på 8000 steg per dag.

(5) **C** Räkna ut korrelationskoefficienten och tolka denna genom att ange vad det beräknade värdet är en indikation på.

Uppgift 3

Som en del av en undersökning rörande barns hälsa studerades självrapporterad längd och faktisk längd hos pojkar i åldrarna 12 till 16 år. Målet med detta var att testa hypotesen att pojkar i denna ålder tenderar att överskatta sin egen längd. Nedan finner du resultat samt en sammanfattning från Minitab för ett urval av tio pojkar.

Pojke	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Själv	173	180	167	178	180	155	165	163	149	160
Faktisk	172.5	177.5	164.8	173.5	181.4	153.9	163.8	160.2	151.2	158.5

	N	Mean	StDev	SE Mean
Själv	10	167.00	10.71	3.39
Faktisk	10	165.73	10.16	3.21
Difference	10	1.270	1.972	0.624

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

(10) **B** Anta att normalfördelningsantagandet som diskuterades i A-uppgiften kan anses vara uppfyllt. Utför ett fullständigt hypotestest enligt *p*-värdesmetoden som utnyttjar detta antagande. Använd en signifikansnivå på 5 %

(12) **C** Anta nu att normalfördelningsantagandet som diskuterades i A-uppgiften inte anses vara realistiskt. Utför ett fullständigt hypotestest på 5 % signifikansnivå där du använder det test som utifrån förutsättningarna utnyttjar informationen på bästa sätt.

(6) **Uppgift 4**

I en artikel om hur ögonvittnen tenderar att blanda samman det de sett drar John Allen Paulos paralleller med ett känt sannolikhersproblem. I detta problem har vi tre mynt där två är vanliga välbalanserade mynt medan det tredje är felbalanserat på så sätt att sannolikheten för att det visar Krona är 0.75. Det går dock inte att se någon skillnad på mynten. Anta att vi slumpmässigt väljer ett av mynten. Vi singlar detta mynt tre gånger och myntet visar Krona vid samtliga tre kast. Bestäm, utifrån all tillgänglig information, sannolikheten att det valda myntet är det felbalanserade.

Uppgift 5

Du står i en biljettkassa för att köpa biljetter till en föreställning. Det finns 60 biljetter kvar till denna föreställning och att det står fyrtio personer framför dig i kön. Problemet är man kan köpa upp till tre biljetter per person och att vissa av de som står framför i kön troligtvis kommer att köpa mer än en biljett. En uppskattning av hur många biljetter en köande person köper (utom du själv som bara skall köpa en biljett) framgår av tabellen nedan.

Antal biljetter	1	2	3
Sannolikhet	0.5	0.4	0.1

(6) **A** Betrakta de tio personer som står först i kön. Bestäm sannolikheten att åtminstone två av dessa personer köper tre biljetter. För full poäng måste en slumpvariabel konstrueras och dess sannolikhetsfördelning anges och motiveras.

(3) **B** Beräkna väntevärde och standardavvikelse för antal biljetter som en slumpmässigt vald person i biljettkön köper.

(7) **C** Bestäm sannolikheten att du lyckas komma över en biljett.

1. Vi börjar med att utvidga den befintliga tabellen.

Antal filmer (x)	f	F	fx	fx^2
0	144	144	0	0
1	16	160	16	16
2	20	180	40	80
3	13	193	39	117
4	4	197	16	64
5	2	199	10	50
6	1	200	6	36
	200		**127**	**363**

$$1.5 \cdot (1 - 0) = 1.5$$

varför uteliggare är observationer under

$$0 - 1.5 = -1.5$$

och över

$$1 + 1.5 = 2.5$$

(a) Utifrån sammanfattningen fås att

$$\overline{x} = \frac{127}{200} = \mathbf{0.635}$$

$$s = \sqrt{\frac{363 - \frac{127^2}{200}}{199}} = \mathbf{1.19}$$

Dessa utvalda personer har alltså under festivalen i genomsnitt sett ungefär 0.6 filmer. Alla har dock inte varit lika flitiga besökare vilket framgår av standardavvikelsen. Den genomsnittliga besökaren avviker med ungefär 1.2 filmer från snittet.

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

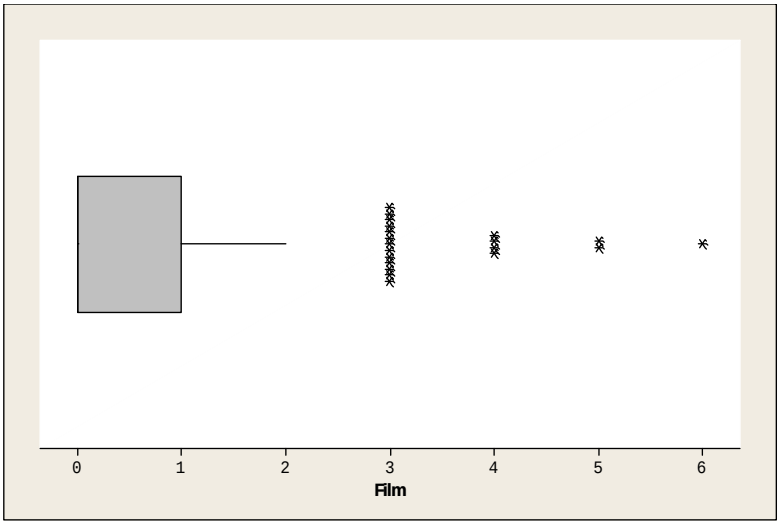
$$q_1 = \left(\text{Värdet på observation } \frac{200+1}{4} = 50.25\right) = 0$$

$$md = \left(\text{Värdet på observation } \frac{200+1}{2} = 100.5\right) = 0$$

$$q_3 = \left(\text{Värdet på observation } \frac{3 \cdot (200+1)}{4} = 150.75\right) = 1$$

Ett och ett halvt kvartilavstånd ges avI och med att den stora massan inte sett någon film har vi ett mycket stort antal uteliggare. Lådagrammet får följande

utseende.



(c) Här låter vi

$$\mu = \text{Genomsnittligt antal filmer kommunens invånare sett under festivalen}$$

och uppgiften är att utifrån den givna stickprovsinformationen konstruera ett 90% konfidensintervall för μ. Vi förutsätter att urvalet är ett OSU och eftersom stickprovet är stort behöver vi inte förutsätta att antal filmer kommunens invånare sett under festivalen är normalfördelad i populationen/kommunen. Vidare gäller att populationen/kommunen kan antas vara stor vilket betyder att vi kan bortse från ändlighetskorrektion och använda konfidensintervallet

$$\overline{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Vi konstaterar att det sökta tabellvärdet är $z_{0.05} = 1.6449$. Nödvändig stickprovsinformation är redan beräknad i a-uppgiften vilket innebär att det sökta intervallet blir

$$0.634 \pm 1.6449 \cdot \frac{1.19}{\sqrt{200}}$$

eller

$$\mathbf{0.50 \leq \mu \leq 0.77}$$

Med 90% säkerhet befinner sig det genomsnittliga antalet filmer kommunens invånare sett under festivalen någonstans mellan 0.50 och 0.77 filmer.

(d) Vi låter nu

$$p = \text{Andelen kommuninvånare som sett åtminstone en film}$$

Utifrån frågeställningen formuleras hypoteserna på följande sätt

$$\begin{aligned} H_0 &: p = 0.25 \\ H_1 &: p > 0.25 \end{aligned}$$

vilka ska testas med en signifikansnivå på 5%. Vi förutsätter (precis som i c-uppgiften) att urvalet är ett OSU och eftersom

$$np_0\left(1-p_0\right) = 200 \cdot 0.25 \cdot 0.75 = 37.5 >> 5$$

är stickprovet så stort att normalapproximation av binomialfördelningen är tillåten. Vidare gäller (precis som i c-uppgiften) att populationen/kommunen kan antas vara stor vilket betyder 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 urvalet blev andelen kommuninvånare som sett åtminstone en film

$$\widehat{p} = \frac{56}{200} = 0.28$$

vilket alltså ger ett visst stöd åt att mer än var fjärde kommuninvånare besökt festivalen. Frågan är hur övertygande resultatet är? Vi sätter in i testfunktionen

$$z = \frac{0.28 - 0.25}{\sqrt{\frac{0.25 \cdot 0.75}{200}}} = 0.98$$

vilket ger ett p-värde på

$$p\text{-värde} = \Pr\left(Z > 0.98\right) = 0.164 > 0.05$$

Eftersom testet utförs på 5%-nivån accepteras nollhypotesen. Vi kan således inte utifrån detta resultat med tillräckligt stark övertygelse säga att mer än var fjärde kommuninvånare besökt filmfestivalen. Festivalen kommer således troligtvis *inte* att få utökad budget nästa år.

(e) Detta är en uppgift som måste lösas i två steg. Först måste vi under nollhypotesantagendet, dvs att $p = 0.25$, ta reda på för vilka värden på stickprovsandelen nollhypotesen kommer att förkastas och sedan måste vi under den nya förutsättningen, dvs att $p = 0.29$, ta reda på sannolikheten att detta kommer att inträffa (vilket är testets styrka).

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

$$\frac{\widehat{p} - 0.25}{\sqrt{\frac{0.25 \cdot 0.75}{200}}} > 1.6449$$

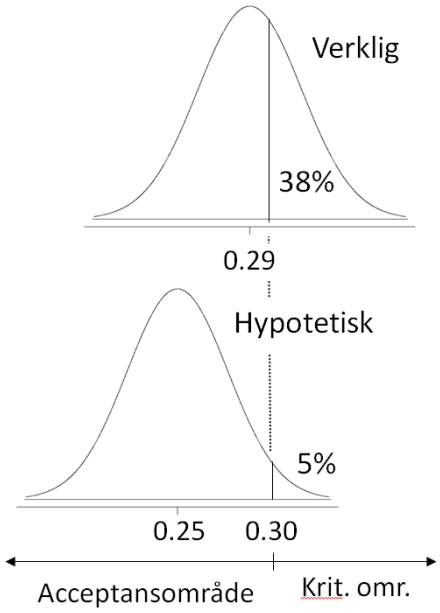
vilket vi översätter till

$$\widehat{p} > 0.25 + 1.6449 \cdot \sqrt{\frac{0.25 \cdot 0.75}{200}} = 0.30036$$

ii. Vad är sannolikheten att vi kommer att förkasta nollhypotesen under den nya förutsättningen att $p = 0.29$? På vanligt normalfördelningsmanér finner vi denna sannolikhet via

$$\Pr\left(\widehat{p} > 0.30036\right) = \Pr\left(Z > \frac{0.300036 - 0.29}{\sqrt{\frac{0.29 \cdot 0.71}{200}}} = 0.31\right) \approx \mathbf{0.38}$$

Testets styrka, dvs sannolikheten att förkasta en felaktig nollhypotes, blir i den här situationen ca 0.38. Chansen att vi under dessa omständigheter kommer att få tillräckligt övertygande bevis om att mer än var fjärde kommuninvånare besökt filmfestivalen är alltså ungefär 38%. Hela situationen beskrivs väl med följande graf



2. Enligt utskriften ges den skattade regressionsekvationen av

$$\widehat{\mu}_{y|x} = a + bx = 29.58 - 0.655 \cdot x$$

(a) Vi tolkar b-koefficienten som att 1 000 extra steg får BMI att *minska* med i genomsnitt 0.655 enheter (dvs kg/m^2). Interceptet, dvs a-koefficienten, anger att personer som inte rör sig alls, dvs för vilka antal steg är noll, har en genomsnittlig BMI på 29.58. Studerar vi spridningsdiagrammet ser vi att de i undersökningen som har rört sig minst har ca 3 000 steg varför en tolkning av a-koefficienten är en extrapolation och bör därför tas med en nypa salt.

(b) Vilket värde på BMI förväntar vi oss för personer med en fysisk aktivitet på 8 000 steg per dag? Insättning i den skattade modellen ger att

$$\widehat{\mu}_{y|x=8} = 29.58 - 0.655 \cdot 8 = 24.34$$

Enligt den skattade modellen har personer med en fysisk aktivitet på 8 000 steg per dag en genomsnittlig BMI på 24.34.

(c) Vi finner korrelationskoefficienten via sambandet med determinationskoefficienten. Vidare ser vi på den skattade b-koefficienten (och även i spridningsdiagrammet) att korrelationskoefficienten är negativ. Det följer därmed att

$$r = -\sqrt{R^2} = -\sqrt{0.1485} = -0.385$$

Korrelationskoefficienten anger graden av linjärt samband mellan de två variablerna. Då korrelationskoefficienten som här är negativ är sambandet negativt vilket innebär att vi här ser tecken på att personer med en högre fysisk aktivitet (med avseende på antal steg) tenderar att ha lägre värden på BMI. Sambandet är vare sig speciellt starkt eller speciellt svagt.

3. Parvisa observationer.

(a) Det är differenserna, dvs skillnaden i självrapporterad och faktisk längd som måste vara approximativt *normalfördelad* i den bakomliggande populationen. Med population avser vi i det här fallet pojkar i åldrarna 12 till 16 år. Om vi, rent hypotetiskt, kunde placera ut alla pojkar i populationen på en skala utifrån skillnaden i självrapporterad och faktisk längd ska den resulterande kurvan vara mycket lik en normalfördelningskurva. Detta är ett nödvändigt antagande eftersom vi i b-uppgiften ska utföra ett hypotestest angående medeldifferensen i populationen och stickprovet är inte tillräckligt stort för att vi utan antaganden kan förutsätta att medeldifferensen i stickprovet är approximativt normalfördelad.

(b) Vi har ett litet stickprov med endast 10 observationer (pojkar) men eftersom *differenserna* antas vara normalfördelade löses många problem. Om vi vidare kan

betrakta pojkarna i undersökningen som slumpmässigt utvalda ur den bakomliggande populationen och inte på något sätt påverkar varandras resultat kan ett parametriskt t-test användas. Låter vi

$$\mu_d = \text{Den genomsnittliga skillnaden i självrapporterad och faktisk längd}$$

formuleras utifrån frågeställningen hypoteserna som

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

där vi här mäter effekten som Självrapporterad–Faktisk vilket betyder att ett positivt värde innebär att den aktuella pojken har överskattat sin längd. Det går givetvis lika bra att ta differenserna i omvänd ordning men då förstås med olikheten i mothypotesen vänd åt andra hållet. Vi ämnar utföra testet med en signifikansnivå på 5% vilket innebär att nollhypotesen ska förkastas först om testets p-värde understiger 5%. Eftersom detta handlar om parvisa observationer börjar vi med att bestämma differenserna, vilket vi som nämnts ovan gör genom att beräkna Självrapporterad–Faktisk så att positiva värden är bra för påståendet att pojkar i denna ålder tenderar att överskatta sin egen längd. Den aktuella testfunktionen rör populationens medelvärde och då den tänkta populationen av pojkar i åldrarna 12 till 16 år rimligtvis kan betraktas som stor kan ändlighetskorrektion bortses från. Med specifika beteckningar för parvisa observationer kan formeln skrivas som

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

I den givna Minitabutskriften finner vi de relevanta värdena till

$$\begin{aligned} \overline{d} &= 1.27 \\ s_d &= 1.972 \end{aligned}$$

varför vi efter insättning av dessa värden i testfunktionen får

$$t_{obs} = \frac{1.27 - 0}{1.972/\sqrt{10}} = 2.04$$

som skall jämföras med t_9-fördelningen. Eftersom

$$t_{9,0.05} = 1.833 < t_{obs} = 2.04 < 2.262 = t_{10,0.025}$$

drar vi slutsatsen att

$$2.5\% < p\text{-värde} < 5\%$$

och då p-värdet understiger den uppsatta signifikansnivån på 5% förkastas nollhypotesen. Det är på 5% signifikansnivå statistiskt säkerställt att pojkar i åldrarna 12 till 16 år, i genomsnitt (med avseende på medelvärde), överskattar sin egen längd.

(c) Vi har ett litet stickprov med endast 10 observationer och eftersom differenserna *inte* kan antas vara normalfördelade måste vi använda ett icke-parametriskt test. Vi förutsätter som i föregående uppgift att pojkarna i undersökningen är slumpmässigt utvalda och inte på något sätt påverkar varandras resultat. Detta tillsammans med det faktum att variabeln mäts på kvotskala innebär att ett *teckenrangtest* kan användas. I detta icke-parametriska test undersöks om fördelningen vad det gäller testresultat är samma i de "båda" populationerna (dvs före och efter genomgången diet), eller om medianen för differenserna är noll, dvs låter vi

$$m_d = \text{Medianskillnaden i självrapporterad och faktisk längd}$$

formuleras utifrån frågeställningen hypoteserna som

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

Vi får

Pojke	1	2	3	4	5
Själv	173	180	167	178	180
Faktisk	172.5	177.5	164.8	173.5	181.4
Själv−Faktisk	0.5	2.5	2.2	4.5	−1.4
Rang	1	8	6.5	10	4
Tecken	+	+	+	+	−

Pojke	6	7	8	9	10
Själv	155	165	163	149	160
Faktisk	153.9	163.8	160.2	151.2	158.5
Själv−Faktisk	1.1	1.2	2.8	−2.2	1.5
Rang	2	3	9	6.5	5
Tecken	+	+	+	−	+

Vi ser att det inte förekommer några ties, dvs pojkar som angett sin faktiska längd. Vi förväntar oss låga rangtal på dom negativa differenserna vilket innebär att vi som testfunktion använder T_-. Det följer att

$$T_- = 4 + 6.5 = 10.5$$

och eftersom

$$T_- = 10.5 > 10 = T_{10,0.05}$$

har vi hamnat i acceptansområdet och accepterar (till skillnad från parametriska testmetoden) nollhypotesen. Det är alltså på 5% signifikansnivå *inte* statistiskt säkerställt att pojkar i åldrarna 12 till 16 år, i genomsnitt (med avseende på median), överskattar sin egen längd. Vi noterar dock att vi hamnat mycket nära den kritiska gränsen.

4. Vi börjar med att definiera *apriorihändelserna*, dvs

$$\begin{aligned} A_1 &= \text{Det valda myntet är välbalanserat} \\ A_2 &= \text{Det valda myntet är felbalanserat} \end{aligned}$$

vilka representerar utfallet i myntvalet *innan* vi fått information om resultatet av slantsinglingen. Därmed gäller att.

$$\begin{aligned} \Pr(A_1) &= \frac{2}{3} \\ \Pr(A_2) &= \frac{1}{3} \end{aligned}$$

Vi definierar nu händelsen

$$B = \text{De tre slantsinglingarna med det valda myntet blir alla Krona}$$

Sannolikheten för B beror på om det valda myntet är välbalanserat eller felbalanserat vilket innebär att

$$\begin{aligned} \Pr(B \mid A_1) &= \left(\frac{1}{2}\right)^3 = \frac{1}{8} \\ \Pr(B \mid A_2) &= \left(\frac{3}{4}\right)^3 = \frac{27}{64} \end{aligned}$$

Med hjälp av Satsen om total sannolikhet får vi att

$$\begin{aligned} \Pr(B) &= \Pr(B \mid A_1)\Pr(A_1) + \Pr(B \mid A_2)\Pr(A_2) = \\ &= \frac{1}{8} \cdot \frac{2}{3} + \frac{27}{64} \cdot \frac{1}{3} = \frac{43}{192} \end{aligned}$$

Vi söker (den betingade) sannolikheten att det valda myntet är skevt givet att samtliga tre singlingar resulterade i Krona vilket via Bayes sats ges av

$$\Pr(A_2 \mid B) = \frac{\Pr(B \mid A_2)\Pr(A_2)}{\Pr(B)} = \frac{\frac{27}{64} \cdot \frac{1}{3}}{\frac{43}{192}} = \mathbf{\frac{27}{43} \approx 0.628}$$

5. Centrala gränsvärdessatsen

(a) Vi utgår från att vi kan göra samma sannolikhetsbedömning för de tio första i kön som vi gör generellt för de köande, dvs för var och en av dessa tio gäller att sannolikheten att köpa tre biljetter är 0.1. Vi förutsätter vidare att det antal biljetter en köande köper är oberoende av det antal biljetter en annan köande köper. Om vi nu låter

$$U = \text{Antal av de tio första i kön som köper tre biljetter}$$

följer att U är binomialfördelad, $Bi\,(10, 0.1)$. Eftersom denna finns i binomialtabellen finner vi snabbt och lätt att

$$\Pr\,(U \geq 2) = 1 - \Pr\,(U \leq 1) = 1 - 0.7361 = \mathbf{0.2639}$$

(b) Nu låter vi

$$X = \text{Antal biljetter en slumpmässigt vald köande tänker köpa}$$

Utifrån den givna informationen gäller då att

$$E\,(X) = 1 \cdot 0.5 + 2 \cdot 0.4 + 3 \cdot 0.1 = 1.6$$

För att finna standardavvikelsen för X beräknas först variansen som blir

$$\begin{aligned} E\left(X^2\right) &= 1^2 \cdot 0.5 + 2^2 \cdot 0.4 + 3^2 \cdot 0.1 = 3 \\ \text{Var}\,(X) &= 3 - 1.6^2 = 0.44 \end{aligned}$$

vilket ger att

$$\sigma_X = \sqrt{0.44} = 0.6633$$

(c) Låter vi $X_1, X_2, \ldots, X_{40}$ representera antalet köpta biljetter av var och en av de som står framför dig i kön gäller att

$$Y = X_1 + X_2 + \cdots + X_{40}$$

representerar det totala antalet biljetter som köps innan det är din tur. Eftersom vi vill att åtminstone en biljett skall vara kvar när vi kommer fram till kassan söker vi sannolikheten

$$\Pr(Y \leq 59)$$

För att kunna beräkna denna sannolikhet gör vi samma antagande som i *a*-uppgiften, dvs att vi det antal biljetter en köande tänker köpa är oberoende av det antal biljetter en annan köande tänker köpa samt att vi gör samma sannolikhetsbedömning för alla vad det gäller antal biljetter de tänker köpa. Då gäller nämligen att $X_1, X_2, \ldots, X_{40}$ kan betraktas som en samling oberoende och likafördelade slumpvariabler (olfsv) och det faktum att de är så pass många samt att Y är en summa av dessa medför enligt Centrala gränsvärdessatsen att Y är approximativt normalfördelad. För att kunna ta reda på vilken normalfördelning som ska användas för denna approximation måste vi först ta reda på väntevärde och varians för Y. Utifrån det vi fann i *b*-uppgiften följer att

$$\begin{aligned} E(Y) &= n \cdot E(X) = 40 \cdot 1.6 = 64 \\ \sigma_Y &= \sqrt{n} \cdot \sigma_X = \sqrt{40} \cdot \sqrt{0.44} = 4.195 \end{aligned}$$

och således gäller att Y approximativt är $N(64, 4.195)$. Med kontinuitetskorrektion finner vi nu den sökta sannolikheten till

$$\Pr(Y \leq 59) \approx \Pr\left(Z \leq \frac{59.5 - 64}{4.195} = -1.07\right) = \mathbf{0.142}$$

Det är alltså ungefär 14% chans att du får en biljett.

TENTAMENSSKRIVNING PÅ KURSERNA

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

2015-10-29

UPPLYSNINGAR

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

B. **Skrivtid: 8.00-13.00** Skrivningen omfattar 5 uppgifter, om sammanlagt 94 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 tillverkare av batterier till mobiltelefoner påstår att deras standardmodell av batterier har en batteritid som i genomsnitt överstiger 34 timmar, vilket skulle innebära att deras batterier är bättre än andra batterier på marknaden.

En fristående konsumentorganisation genomför en studie för att undersöka detta påstående. Ur den mycket stora produktionen väljs slumpmässigt 60 batterier. Resultatet av undersökningen redovisas i tabellen nedan.

Tabell 1 Batteritid (timmar) (sorterat).

30	31	31	31	31	31	31	32	32	32	32	32	32	32	32	32
33	33	33	33	33	34	34	34	34	34	34	34	34	34	34	34
34	34	34	35	35	35	35	35	35	35	35	35	35	36	36	36
37	37	37	38	38	39	39	39	39	40	40	45				

X: Batteritid Beräkningshjälp: $\sum_{i=1}^{60} x_i = 2\ 071, \quad \sum_{i=1}^{60} x_i^2 = 71\ 953$

(5) **A** Beräkna medelvärde och standardavvikelse för den aktuella variabeln. Ge en ordentlig förklaring av innebörden av den framräknade standardavvikelsen.

(7) **B** Åskådliggör materialet grafiskt genom att konstruera ett lådagram.

(8) **C** Konstruera ett 90 % konfidensintervall för andelen batterier som räcker 35 timmar eller längre.

(12) **D** Går det att statistiskt säkerställa att den genomsnittliga batteritiden överstiger 34 timmar? Besvara frågan med lämpligt hypotestest enligt *p*-värdesmetoden. Använd signifikansnivå 5 %.

(3) **E** Ge en ordentlig tolkning av det p-värde som beräknades i D-uppgiften ovan genom att börja med ”Givet att den genomsnittliga batteritiden är …”. Observera att tolkningen inte ska gälla huruvida nollhypotesen ska förkastas (detta är redan gjort i D-uppgiften).

(8) **F** Vi fortsätter nu med situationen i D-uppgiften. Låt oss betrakta det aktuella hypotestestet *innan* resultatet av undersökningen sammanställdes, dvs vi har ännu inte några resultat från undersökningen. Anta att den faktiska genomsnittliga batteritiden är 35 timmar. Vad är med denna förutsättning styrkan för testet i D-uppgiften? Vid dessa beräkningar kan populationsstandardavvikelsen anses vara känd och lika med standardavvikelsen i stickprovet. *Anmärkning*. För full poäng måste situationen beskrivas grafiskt.

Uppgift 2

Mjölkpriser Sverige.
En liter mjölk kostade år 1950 34 öre, år 2000 6,40 kr och år 2010 8,60 kr.

Källa: Jordbruksverket och SCB.

(4) **A** Hur stor är den årliga genomsnittliga procentuella förändringen av mjölkpriserna (löpande priser)?

1. Under perioden 1950-2000
2. Under perioden 2000-2010

Redovisa procenttalen med en decimal.

(6) **B** Redovisa mjölkpriserna år 1950, år 2000 och år 2010 i 2010 års prisnivå.

År	**1950**	**1980**	**2000**	**2010**
KPI, basår 1949	101	571		
KPI, basår 1980		100	260,7	303,5

(12) **Uppgift 3**

Ett nyhetsbrev rangordnar ett mycket stort antal aktieandelsfonder (mutual funds). Slumpmässiga urval omfattande tio fonder av de som erhållit den högsta ratingen och tio fonder av de som erhållit den lägsta ratingen gjordes. Följande värden är den procentuella avkastning som dessa tjugo fonder uppnådde under det kommande året.

Högst rankade	8.1	12.7	13.9	2.3	16.1	5.4	7.3	9.8	14.3	4.1
Lägst rankade	3.5	14.0	11.1	4.7	6.2	13.3	7.0	7.3	4.6	10.0

Kan det utifrån detta resultat påvisas att de högst rankade fonderna tenderar att ge högre avkastning jämfört med de lägst rankade fonderna? Utför ett fullständigt icke-parametriskt hypotestest på 5 % signifikansnivå.

(12) **Uppgift 4**

I en omfattande konsumentundersökning (i USA) utförd av *Wall Street Journal* och *NBC News* studerades bland annat frågan hur konsumenter i allmänhet bedömer servicenivån hos amerikanska företag. Resultatet i undersökningen ges i tabellen nedan.

Omdöme	Utmärkt	Bra	Godtagbar	Undermålig
Andel svar (%)	8	47	34	11

En butikschef önskar ta reda på huruvida resultatet av konsumentundersökningen kan antas gälla även för kunder på stormarknader i hennes stad. För att få svar intervjuar hon 207 slumpmässigt valda kunder (då de lämnar butiken) i olika delar av staden. Undersökningen gav följande resultat:

Omdöme	Utmärkt	Bra	Godtagbar	Undermålig
Antal svar	21	109	62	15

Hjälp butikschefen genom att utföra ett fullständigt hypotestest där du använder en signifikansnivå på 5 %.

Uppgift 5

Vikten av en slumpmässigt vald halstablett kan betraktas som normalfördelad med medelvärde 0.65 gram och standardavvikelse 0.02 gram.

(3) **A** Anta att du slumpmässigt väljer en halstablett. Bestäm sannolikheten att den valda tabletten väger någonstans mellan 0.6 och 0.7 gram.

(3) **B** Bestäm den vikt sådan att 80 % av alla halstabletter väger mer.

(5) **C** Bestäm sannolikheten att högst nio av tolv slumpmässigt valda halstabletter har en vikt som överstiger den vikt som beräknades i B-uppgiften ovan. Ange en slumpvariabel och dess sannolikhetsfördelning (med motivering) och beräkna därefter den efterfrågade sannolikheten.

(6) **D** En ask bör innehålla 100 halstabletter. För att förenkla påfyllningen av en ask häller man upp tabletter på en vågskål och slutar så snart vikten överstiger 65 gram. Bestäm sannolikheten att asken kommer att innehålla åtminstone 100 tabletter.
Ledning. Anta att vågskålen fylls med tabletter, en åt gången. Vad ska till för att den hundrade tabletten ska läggas i vågskålen?

1. Batteritid.

   (a) Utifrån sammanfattningen fås att

$$\overline{x} = \frac{\sum x}{n} = \frac{2\,071}{60} = \mathbf{34.5}$$

$$s = \sqrt{\frac{71\,953 - \frac{2\,071^2}{60}}{59}} = \mathbf{2.82}$$

   Dessa utvalda batterier har alltså en genomsnittlig batteritid på 34.5 timmar. Olika batterier räcker dock olika länge vilket framgår av standardavvikelsen. Batteritiderna (för batterierna i urvalet) avviker med i genomsnitt 2.8 timmar från medelvärdet.

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

$$\begin{aligned} q_1 &= \left(\text{Värdet på observation } \frac{60+1}{4} = 15.25\right) = 32 \\ md &= \left(\text{Värdet på observation } \frac{60+1}{2} = 30.5\right) = 34 \\ q_3 &= \left(\text{Värdet på observation } \frac{3\cdot(60+1)}{4} = 45.75\right) = 35.75 \end{aligned}$$

   Ett och ett halvt kvartilavstånd ges av

$$1.5 \cdot (q_3 - q_1) = 1.5 \cdot (35.75 - 32) = 5.625$$

   vilket innebär att uteliggare är observationer/batteritider under

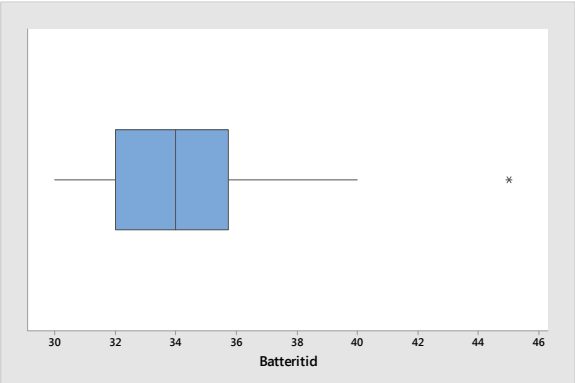
$$32 - 5.625 = 26.375$$

   respektive över

$$35.75 + 5.625 = 41.375$$

   Vi observerar att det endast finns en uteliggare vilket är det batteri med den långa batteritiden på 45 timmar. Dämed dras det undre morrhåret till det minsta värdet, dvs 30 och det övre morrhåret dras till 40. Lådagrammet får således

följande utseende:



(c) Här låter vi

$$p = \text{Andelen batterier med en batteritid på 35 timmar eller längre}$$

och uppgiften är att utifrån den givna stickprovsinformationen konstruera ett 90% konfidensintervall för p. Vi förutsätter (som det är angivet) att urvalet är ett OSU, dvs att batterierna är slumpmässigt valda ur produktionen. I urvalet var det 25 av de 60 batterierna som hade en batteritid på 35 timmar eller längre vilket innebär att $\widehat{p} = 25/60 = 0.4167$. Eftersom

$$n\widehat{p}\,(1-\widehat{p}) = 60 \cdot 0.4167 \cdot 0.5833 = 14.6 > 5$$

är stickprovet så stort att normalapproximation av binomialfördelningen är tillåten. Vidare gäller att produktionen av batterier är så stor att urvalet endast utgör en liten del av populationen av batterier vilket innebär att vi kan använda konfidensintervallet

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

Vi konstaterar att det sökta tabellvärdet är $z_{0.05} = 1.6449$. Efter insättning av stickprovsvärden följer att det sökta intervallet blir

$$0.4167 \pm 1.6449 \cdot \sqrt{\frac{0.4167 \cdot 0.5833}{60}}$$

eller

$$\mathbf{0.312 \leq p \leq 0.521}$$

Med 90% säkerhet är andelen batterier med en batteritid på 35 timmar eller längre någonstans mellan 31.2% och 52.1%.

(d) Vi låter nu

$$\mu = \text{Medelbatteritiden (timmar)}$$

Utifrån frågeställningen formuleras hypoteserna på följande sätt

$$\begin{aligned} H_0 &: \mu = 34 \\ H_1 &: \mu > 34 \end{aligned}$$

vilka ska testas med en signifikansnivå på 5%. Vi förutsätter som ovan att urvalet är ett OSU, dvs att batterierna är slumpmässigt valda ur en stor produktion. Eftersom vi dessutom har att $n = 60 > 30$ följer att vi kan använda testfunktionen

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

som är $N(0,1)$ då nollhypotesen är sann. *Anmärkning.* Ett t-test med 59 (dvs ungefär 60) frihetsgrader är också godkänt att använda här. Eftersom vi ska använda p-värdesmetoden och att testet utförs på 5% signifikansnivå ska nollhypotesen förkastas först om p-värdet understiger 5%. Insättning av våra stickprovsvärden från a-uppgiften ger oss följande värde på testfunktionen

$$z = \frac{34.517 - 34}{2.819/\sqrt{60}} = 1.42$$

vilket utifrån utseendet på mothypotesen innebär att

$$p\text{-värde} = \Pr(Z > 1.42) = 0.078$$

Eftersom

$$p\text{-värde} = 0.078 > 0.05 = \alpha$$

accepteras nollhypotesen. Det är således på 5% signifikansnivå inte statistiskt säkerställt att medelbatteritiden för dessa batterier överstiger 34 timmar.

(e) Givet att den genomsnittliga batteritiden är 34 timmar är sannolikheten att i ett stickprov om 60 slumpmässigt valda batterier få en så hög medelbatteritid som 34.5 timmar eller högre ungefär 7.8%.

(f) Detta är en uppgift som måste lösas i två steg. Först måste vi under nollhypotesantagendet, dvs att $\mu = 34$, ta reda på för vilka värden på stickprovsmedelvärdet nollhypotesen kommer att förkastas och sedan måste vi under den nya förutsättningen, dvs att $\mu = 35$, ta reda på sannolikheten att detta kommer att inträffa (vilket är testets styrka).

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

$$\frac{\overline{x} - 34}{2.819/\sqrt{60}} > 1.6449$$

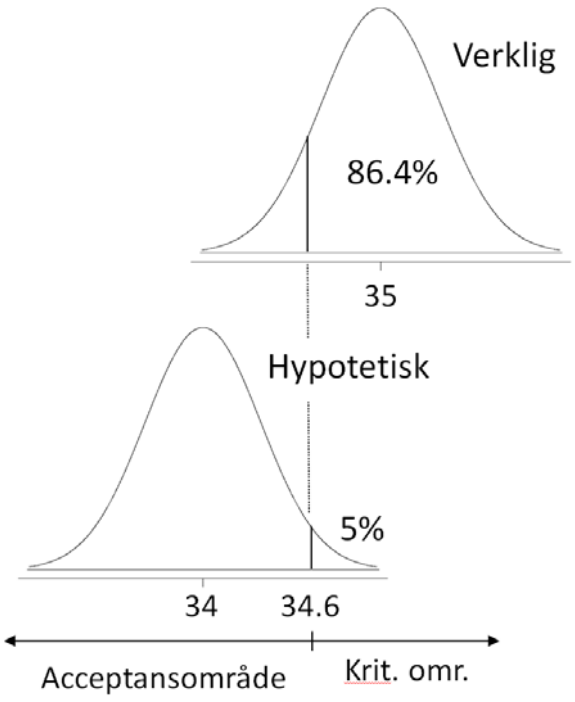
vilket vi översätter till

$$\overline{x} > 34 + 1.6449 \cdot \frac{2.819}{\sqrt{60}} = 34.6$$

ii. Vad är sannolikheten att vi kommer att förkasta nollhypotesen under den nya förutsättningen att $\mu = 35$? På vanligt normalfördelningsmanér finner vi denna sannolikhet via

$$\Pr\left(\overline{X} > 34.6\right) = \Pr\left(Z > \frac{34.6 - 35}{2.819/\sqrt{60}} = -1.10\right) \approx \mathbf{0.864}$$

Testets styrka, dvs sannolikheten att förkasta en felaktig nollhypotes, blir i den här situationen ca 0.86. Chansen att vi under dessa omständigheter kommer att få tillräckligt övertygande bevis om att medelbatteritiden för dessa batterier överstiger 34 timmar är alltså ca 86%. Hela situationen beskrivs väl med följande graf



2. *Mjölkpriser i Sverige.*

   (a) Hur stor var den årliga genomsnittliga procentuella förändringen av mjölkpriserna (i löpande priser) under perioden 1950–2000? Eftersom

   $$\left(\frac{6.40}{0.34}\right)^{1/50} = 1.060$$

   har under denna 50-årsperiod den genomsnittliga årliga ökningen av mjölkpriset varit ca 6.0%. På motsvarande sätt ser vi att

   $$\left(\frac{8.60}{6.40}\right)^{1/10} = 1.030$$

   vilket innebär att den genomsnittliga årliga ökningen av mjölkpriset under period 2000–2010 var ca 3.0%.

   (b) Det finns flera vägar att gå här men vi bör börja med att se till att samtliga involverade KPI-värden har samma basår. Om vi ser till samtliga involverade tidpunkter har 1949 som basår gäller att

   $$\begin{aligned} 1950 &: 101 \\ 2000 &: 571 \cdot 2.607 = 1\,488.597 \\ 2010 &: 571 \cdot 3.035 = 1\,732.985 \end{aligned}$$

   Deflatering görs nu via formeln

   $$P_t^{\text{Fast}} = P_t^{\text{Löp}} \cdot \frac{\text{KPI}_{t_0}}{\text{KPI}_t}$$

   där $t_0 = 2010$. Det följer därmed att

   $$\begin{aligned} P_{1950}^{\text{Fast}} &= P_{1950}^{\text{Löp}} \cdot \frac{\text{KPI}_{2010}}{\text{KPI}_{1950}} = 0.34 \cdot \frac{1\,732.985}{101} = 5.83 \\ P_{2000}^{\text{Fast}} &= P_{2000}^{\text{Löp}} \cdot \frac{\text{KPI}_{2010}}{\text{KPI}_{2000}} = 6.40 \cdot \frac{1\,732.985}{1\,488.597} = 7.45 \\ P_{2010}^{\text{Fast}} &= 8.60 \end{aligned}$$

3. Här har vi en situation med två oberoende stickprov på en kvantitativ variabel varför ett parametriskt t-test skulle kunna bli aktuellt. Dock gäller att stickproven är små och vi bestämmer oss här för att göra ett icke-parametriskt test vilket här blir det som kallas *Mann-Whitney* (eller *Wilcoxons rangsummatest*). Låter vi Md_H och Md_L representera populationsmedianerna vad det gäller den procentuella avkastningen för de högst respektive lägst rankade fonderna följer utifrån frågeställningen att hypoteser ska formuleras som

$$\begin{aligned} H_0 &: Md_H = Md_L \\ H_1 &: Md_H > Md_L \end{aligned}$$

vilka vi tänker undersöka med ett hypotestest på 5% signifikansnivå. Vi förutsätter att de båda stickproven är OSU och oberoende av varandra. I och med att $n_H = n_L = 10$ kan vi själva välja vilket urval som ska användas i analysen. Eftersom vi enligt mothypotesen förväntar oss låga rangtal från de lägst rankade fonderna blir det naturligt att använda den populationen som testpopulation. Nollhypotesen ska förkastas först om

$$R_{obs} < R_{10,10,5\%} = 82$$

Resultatet av rangordningen blir

Högst	Avkastning (%)	8.1	12.7	13.9	2.3	16.1	5.4	7.3	9.8	14.3	4.1
	Rangtal	11	15	17	1	20	6	9.5	12	19	3
Lägst	Avkastning (%)	3.5	14.0	11.1	4.7	6.2	13.3	7.0	7.3	4.6	10.0
	Rangtal	2	18	14	5	7	16	8	9.5.	4	13

vilket innebär att

$$R_1 = 2 + 4 + 5 + 7 + 8 + 9.5 + 13 + 14 + 16 + 18 = 96.5$$

och eftersom

$$R_S = n_L\left(n_L + n_H + 1\right) - R_1 = 10 \cdot (10 + 10 + 1) - 96.5 = 113.5 > R_1$$

följer att

$$R_1 = R_{obs} = 96.5 > 82 = R_{10,10,5\%}$$

har vi hamnat i acceptansområdet och nollhypotesen accepteras vilket innebär att vi på 5% signifikansnivå inte med tillräcklig säkerhet kan påstå att det finns skillnad på Md_H och Md_L, dvs populationsmedianerna vad det gäller den procentuella avkastningen för de högst respektive lägst rankade fonderna I och med att testfunktionens värde (klart) överstiger $R_{10,10,5\%} = 82$ bör p-värdet (klart) överstiga 5%. (Minitab ger p-värdet till 27.3%.)

4. Låter vi π_{Utm}, π_{Bra}, π_{God} och π_{Und} representera hur uppdelningen i de olika kategorierna Utmärkt, Bra, Godtagbar samt Undermålig ser ut bland stormarknadskunder i den aktuella staden ska hypoteserna (baserat på utfallet i den ursprungliga konsumentundersökningen) formuleras som

$$\begin{aligned} H_0 &: \pi_{Utm} = 0.08, \pi_{Bra} = 0.47, \pi_{God} = 0.34, \pi_{Und} = 0.11 \\ H_1 &: \text{Någon annan fördelning} \end{aligned}$$

och eftersom det är fler än två kategorier måste χ^2-metoden användas. Vi utför testet på 5% signifikansnivå. För att kunna utföra testet kompletteras de observerade frekvenserna med de förväntade frekvenserna vilka blir som följer:

Omdöme	Utmärkt	Bra	Godtagbar	Undermålig
Observerade	21	109	62	15
Förväntade	16.56	97.29	70.38	22.77

eftersom vi vid samma utfall som i den ursprungliga konsumentundersökningen förväntar oss att 8% av (de 207) kunderna anger "Utmärkt", 47% av kunderna anger "Bra" osv. Då kunderna i urvalet är slumpmässigt valda samt att ingen av de förväntade frekvenserna understiger 5 kan vi använda testfunktionen

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

som är χ^2-fördelad med $4 - 1 = 3$ frihetsgrader då nollhypotesen är sann. Vi ska förkasta nollhypotesen först om

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

Vi sätter in våra värden och får

$$\chi^2 = \frac{(21 - 16.56)^2}{16.56} + \frac{(109 - 97.29)^2}{97.29} + \frac{(62 - 70.38)^2}{70.38} + \frac{(15 - 22.77)^2}{22.77} = 6.25$$

Eftersom

$$\chi^2_{obs} = 6.25 < 7.815 = \chi^2_{3,0.05}$$

har vi hamnat i acceptansområdet och nollhypotesen accepteras. Det är således på 5% signifikansnivå *inte* statistiskt säkerställt att denna stadens stormarknadskunders bedömning av servicenivå (hos stormarknaderna) skiljer sig från den bedömning konsumenter i allmänhet gör angående servicenivån hos amerikanska företag.

5. Vi betraktar nu slumpvariabeln

$$X = \text{Vikten av en (slumpmässigt vald) halstablett}$$

som enligt den givna informationen kan betraktas som $N(0.65, 0.02)$ där enheten är gram.

(a) Vi söker nu

$$\begin{aligned}\Pr(0.6 < X < 0.7) &= \Pr\left(\frac{0.6-0.65}{0.02} < Z < \frac{0.7-0.65}{0.02}\right) = \\ &= \Pr(-2.5 < Z < 2.5) = \Pr(Z < 2.5) - \Pr(Z < -2.5) = \\ &= 0.9938 - (1 - 0.9938) = \mathbf{0.988}\end{aligned}$$

Vi tolkar detta som att ca 98.8% av halstabletter har en vikt som avviker med högst 0.05 gram från den önskade vikten.

(b) Enligt Tabell 5.2.B gäller att

$$z_{0.8} = -0.8416$$

vilket innebär att den andra decilen, dvs d_2, ges av

$$d_2 = 0.65 - 0.8416 \cdot 0.02 = \mathbf{0.633}$$

Det gäller alltså att 80% av alla tabletter väger 0.633 gram eller mer.

(c) Här inser vi att det underlättar att utgå från komplementhändelsen, dvs att en tablett väger mindre än d_2. Alltså,

$$Y = \text{Antal halstabletter i urvalet som väger mindre än } d_2$$

Sannolikheten att en slumpmässigt vald tablett väger mindre än d_2 är per definition 0.2. En tablett väger antingen mindre än d_2 eller så gör den inte det. Vidare utgår vi från att vikten hos olika tabletter är oberoende av varandra. Därmed följer att Y är $Bi(12, 0.2)$. Via Tabell 5.1 följer således att

$$\Pr(Y \geq 3) = 1 - \Pr(Y \leq 2) = 1 - 0.5583 = \mathbf{0.4417}$$

(d) I ledningen ställs frågan: *"Vad ska till för att den hundrade tabletten ska läggas i vågskålen?"*. För att så ska ske måste de 99 först ditlagda tabletterna ha en sammanlagd vikt som understiger 65 gram. Låter vi $X_1, X_2, \ldots, X_{99}$ representera vikten hos var och en av 99 slumpmässigt valda tabletter bör enligt förutsättningarna gälla att dessa kan betraktas som oberoende och likafördelade slumpvariabler som alla är $N(0.65, 0.02)$. Därmed följer att summan

$$Y = X_1 + X_2 + \cdots + X_{99}$$

också är normalfördelad vilket innebär att vi endast behöver ta reda på väntevärde och standardavvikelse för Y. Eftersom

$$\begin{aligned} \mu_Y &= 99 \cdot \mu_X = 99 \cdot 0.65 = 64.35 \\ \sigma_Y &= \sqrt{99} \cdot \sigma_X = \sqrt{99} \cdot 0.02 = 0.20 \end{aligned}$$

gäller att Y är $N(64.35, 0.20)$. Vi får nu den sökta sannolikheten till

$$\Pr(Y < 65) = \Pr\left(Z < \frac{65 - 64.35}{0.20} \approx 3.27\right) = \mathbf{0.9995}$$

TENTAMENSSKRIVNING PÅ KURSERNA

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

2015-11-28

UPPLYSNINGAR

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

B. **Skrivtid: 9.00-14.00** Skrivningen omfattar 5 uppgifter, om sammanlagt 94 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

Ett flygbolag var intresserade av att få en bättre uppfattning vad det gäller vikten av resenärernas bagage (på en viss rutt). Ett slumpmässigt urval av 68 resenärer gjordes där man för varje resenär vägde både incheckat bagage och handbagage för att få en total bagagevikt. Resultatet av undersökningen är sammanställt i frekvenstabellen nedan.

Vikt(kg)	Antal
5-9	8
10-19	14
20-29	32
30-39	6
40-59	8

(6) **A** Beräkna medel- samt medianvikt för resenärernas bagage.

(3) **B** Förklara kortfattat hur man utifrån genomsnittsmåtten medelvärde och median kan göra en första kontroll över huruvida frekvensfördelningen är symmetrisk. Vad ser vi utifrån beräkningarna i A-uppgiften tecken på i det här materialet?

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

(8) **D** Använd resultatet i detta urval för att konstruera ett intervall som med 95 % säkerhet täcker in medelvärdet för bagagevikten hos resenärer på den aktuella rutten.

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

(12) **F** Är det utifrån detta stickprov statistiskt säkerställt att mer än 60 % av alla resenärer har en bagagevikt som är 20 kg eller mer (efter avrundning till hela kg)? Ställ utifrån frågeställningen upp hypoteser och utför enligt *p*-värdesmetoden ett fullständigt hypotestest på 5 % signifikansnivå.

(3) **G** Ge en ordentlig förklaring av begreppet *signifikansnivå* genom att *utgå från situationen i den här uppgiften*. Enbart en allmän förklaring av begreppet ger inte några poäng.

Uppgift 2

Gör mobiltelefonanvändning under bilkörning att reaktionstiden försämras? Detta var frågeställningen i en undersökning som utfördes vid University of Utah (*D. Strayer and W. Johnston, Psych. Science, vol. 21, pp. 462-466, 2001*).

I undersökningen studerades 64 studenter som slumpmässigt delades in i en mobiltelefongrupp och en kontrollgrupp. Varje deltagare fick under en viss tidsperiod köra bil i en simulator. I denna körsimulator blinkade en lampa antingen grönt eller rött vid slumpmässigt valda tidpunkter och deltagarna fick instruktioner om att bromsa då lampan blinkade rött. De i kontrollgruppen fick under testet lyssna på en radiosändning eller en ljudbok medan de i mobiltelefongruppen hade en telefonkonversation med en person som befann sig i ett annat rum. Den genomsnittliga reaktionstiden (i millisekunder) mättes för varje deltagare. Minitabutskriften nedan ger en sammanfattning av resultatet av undersökningen.

Variable	N	Mean	StDev	Q1	Median	Q3
Mobil	32	585.2	89.6	537.5	569.0	621.0
Kontroll	32	533.6	65.4	480.5	530.0	585.8

(12) **A** Undersök med ett fullständigt hypotestest enligt klassisk metod på 10 % signifikansnivå om det går att påvisa någon skillnad i spridning vad det gäller reaktionstid, dvs skillnad i standardavvikelse/varians, mellan de båda grupperna.

(12) **B** *Gör mobiltelefonanvändning under bilkörning att reaktionstiden försämras?* Besvara frågan genom att utföra ett fullständigt hypotestest enligt klassisk metod på 5 % signifikansnivå där du utnyttjar resultatet i A-uppgiften.

Uppgift 3

Historisk prisökning på GB:s mest kända glassar*. Glasspriserna har ökat med 30 procent på 20 år. Priserna på GB Glace storsäljare Magnum, 88:an och Sandwich har dock höjts betydligt mer. (affarsvarlden.se den 9 juni 2014.)*

GB:s storsäljare ”88:an” har varit med länge och prisutvecklingen för en sådan glass under tidsperioden 1984 till 2014 framgår av tabellen nedan tillsammans med KPI för de aktuella årtalen.

År	**1984**	**1994**	**2004**	**2014**
Styckpris (kr)	3,60	8,00	11,00	15,00
KPI	143,2	248,5	279,2	313,5

(7) **A** Räkna om styckpriset på en ”88:an” till det penningvärde som gällde 2014 och konstruera en indexserie över de fasta priserna med 2014 som basår. Samtliga angivna tidpunkter ska vara med i indexserien och indextalen ska redovisas med en decimal.

(4) **B** Enligt artikeln har glasspriserna generellt ökat med 30 procent under den senaste 20-årsperioden medan att priserna på storsäljarna, däribland ”88:an”, har höjts betydligt mer. Beräkna både i löpande och fasta priser hur mycket, procentuellt sett, styckpriset på en ”88:an” har höjts totalt under den senaste 20-årsperioden, dvs mellan 1994 och 2014. Ange dessutom, för samma tidsperiod, med hur många procent KPI har ändrats.

(5) **Uppgift 4**

Från en skylt med texten UPPSALA faller det ner två (slumpmässigt valda) bokstäver. En vänlig analfabet sätter (slumpmässigt) upp de båda bokstäverna på de tomma platserna. Använd Satsen om total sannolikhet för att beräkna sannolikheten att skylten får korrekt text.

Uppgift 5

Ett slumpförsök består av att kasta en vanlig sexsidig tärning fem gånger. Du bildar en slumpvariabel som beräknar totalt antal ögon tärningen visar under dessa fem kast.

(4) **A** Centrala gränsvärdessatsen används ofta för att approximera sannolikheter för vissa typer av sannolikhetsfördelningar. Ge argument både för och emot att normalfördelningen kan användas för att göra approximativa sannolikhetsbedömningar i samband med din slumpvariabel.

(8) **B** Även om det är minst sagt tveksamt här beslutar vi oss trots allt för att använda Centrala gränsvärdessatsen. Beräkna approximativt sannolikheten att du i dina fem kast får högst 15 ögon.

1. Vi börjar med att återge (och utöka) frekenstabellen

Viktklass	f_i	Mitt (x_i)	$f_i x_i$	$f_i x_i^2$	Kum (F_i)
$5-9$	8	7	56	392.0	8
$10-19$	14	14.5	203	2 943.5	22
$20-29$	32	24.5	784	19 208.0	54
$30-39$	6	34.5	207	7 141.5	60
$40-59$	8	49.5	396	19602.0	68
	68		1646	49 287.0	

Här är de faktiska klassgränserna 4.5, 9.5, 19.5, 29.5, 39.5 samt 59.5.

(a) Vi bestämmer först medelvärdet som blir

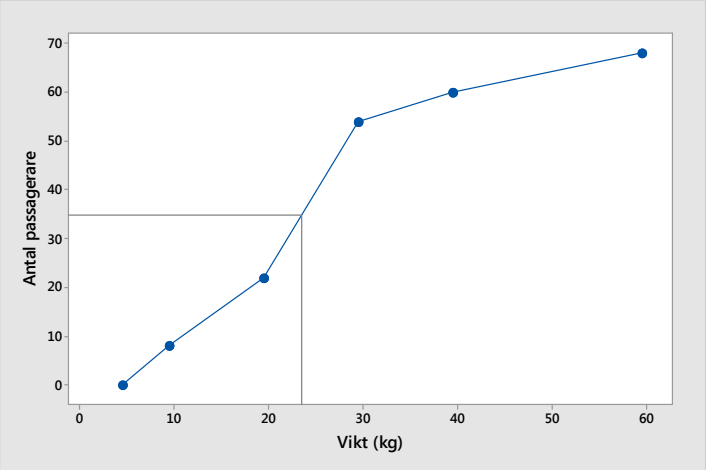
$$\overline{x} = \frac{1646}{68} = \mathbf{24.2}$$

För att bestämma medianen söker vi värdet på observation $\frac{68}{2} = 34$ som finns i klassen $20-29$. Vi interpolerar och finner att

$$md = 19.5 + \frac{34-22}{32} \cdot 10 = \mathbf{23.25}$$

(b) Genom att studera stickprovets medelvärde och median kan vi få en första uppfattning av frekvensfördelningens utseende, dvs om den är symmetrisk eller asymmetrisk. För en helt symmetrisk fördelning är medelvärde och median samma och om detta gäller för vårt stickprov är detta åtminstone en indikation om att fördelningen är symmetrisk. På motsvarande sätt gäller att ju mer de skiljer sig från varandra desto starkare indikation har vi om att fördelningen är asymmetrisk (i riktning mot medelvärdet). Detta beror på att medianen är ett robust genomsnittsmått som inte påverkas av en viss asymmetri medan medelvärdet är känsligt för detta. I det här fallet ger medelvärde och median en indikation om att materialet är något snedfördelat *uppåt* dvs *positively skewed* (en svans åt höger) vilket vi även kan få bekräftelse på genom att studera frekvenstabellen (eller ett histogram).

(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



där vi som angivit ovan använder de faktiska klassgränserna 4.5, 9.5, 19.5, 29.5, 39.5 samt 59.5. Genom att på y-axeln utgå från observation $n/2 = 34$, dvs medianobservationen, och dra en horisontell linje fram till summapolygonen och sedan därifrån dra en lodrät linje ner till x-axeln får vi en uppskattning av medianvärdet. Enligt resultatet i b-uppgiften ska denna bli 23.25 vilket verkar rimligt.

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

$$\mu = \text{Genomsnittlig bagagevikt (på denna rutt)}$$

Vi förutsätter att de 68 passagerarna kan betraktas som ett slumpmässigt urval ur den (hypotetiska) populationen av passagerar på den aktuella rutten. I och med att $n = 68 > 30$ har vi ett tillräckligt stort stickprov för att konstruera konfidensintervall baserat på normalfördelningen utan att behöva förutsätta att den aktuella variabeln, dvs bagagevikt (på denna rutt), är normalfördelad. Vidare gäller att (den hypotetiska) populationen bestående av alla passagerare kan betraktas som stor vilket betyder att vi kan använda konfidensintervallet

$$\overline{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

där vi använt det faktum att t-fördelningen här kan approximeras med z-fördelningen. Det är dock förstås helt okej att för intervallet nedan använda sig av $t_{67,0.025} \approx t_{60,0.025} = 2.00$. Vi inser att stickprovsstandardavvikelsen behöver beräknas vilken vi utifrån den utökade frekvenstabellen finner till

$$s = \sqrt{\frac{49287 - \frac{1646^2}{68}}{67}} = 11.873$$

Eftersom $z_{0.025} = 1.96$ följer efter insättning av våra stickprovsvärden att konfidensintervallet blir

$$24.21 \pm 1.96 \cdot \frac{11.873}{\sqrt{68}}$$

eller som ett intervall

$$\mathbf{21.4} \leq \boldsymbol{\mu} \leq \mathbf{27.0}$$

Med 95% säkerhet befinner sig μ, dvs genomsnittlig bagagevikt (på denna rutt), någonstans mellan 21.4 kg och 27.0 kg.

(e) För att kunna bestämma hur stort stickprov som ska tas för att intervallet ska bli maximalt 4 kg brett måste vi ha en uppfattning om standardavvikelsen i populationen. Stickprovet ovan gav oss en möjlighet att skatta denna via

$$\widehat{\sigma} = 11.873$$

Eftersom $z_{0.025} = 1.96$ och halva bredden $E = 2$ följer att den sökta stickprovsstorleken blir

$$n = \frac{z_{0.025}^2 \cdot \widehat{\sigma}^2}{E^2} = \frac{1.96^2 \cdot 11.873^2}{2^2} = 135.4$$

För att uppfylla kraven krävs alltså ett stickprov om minst **136** passagerare på den aktuella rutten.

(f) Vi låter nu

$$p = \text{Andel passagerare med en bagagevikt på 20 kg eller mer}$$

Utifrån frågeställningen formuleras hypoteserna på följande sätt

$$\begin{aligned} H_0 &: \ p = 0.6 \\ H_1 &: \ p > 0.6 \end{aligned}$$

vilka ska testas på 5% signifikansnivå. Vi förutsätter (precis som i d-uppgiften) att urvalet kan betraktas som ett slumpmässigt urval bland alla passagerare på denna rutt och eftersom

$$np_0\left(1 - p_0\right) = 68 \cdot 0.6 \cdot 0.4 = 16.32 > 5$$

är stickprovet tillräckligt stort för att normalapproximation av binomialfördelningen ska vara tillåten. Vidare gäller (precis som i d-uppgiften) att populationen (av potentiella passagerare på den aktuella rutten) kan antas vara stor vilket betyder att vi kan använda testfunktionen

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

I och med att vi använder en signifikansnivå på 5% gäller att nollhypotesen ska förkastas först om

$$p\text{-värde} < 0.05$$

I urvalet blev andelen passagerare med en bagagevikt på 20 kg eller mer

$$\widehat{p} = \frac{32+6+8}{68} = \frac{46}{68} = 0.6765$$

vilket alltså ger stöd åt att mer än 60% av alla resenärer på denna rutt har en bagagevikt som är 20 kg eller mer. Frågan är hur övertygande resultatet är? Vi sätter in i testfunktionen

$$z = \frac{0.6765 - 0.6}{\sqrt{\frac{0.6 \cdot 0.4}{68}}} = 1.288$$

Då mothypotesen är på formen $H_1 : p > 0.6$ följer via z-tabellen att det sökta p-värdet blir

$$p\text{-värde} = \Pr\left(Z > 1.288\right) \approx \Pr\left(Z > 1.29\right) \approx \mathbf{0.1}$$

Eftersom

$$p\text{-värde} \approx 0.1 > 0.05$$

har vi hamnat i acceptansområdet och därmed accepteras nollhypotesen. Det är alltså på 5% signifikansnivå *inte* statistiskt säkerställt att p, dvs andelen passagerare med en bagagevikt på 20 kg eller mer, överstiger 60%.

(g) Testets *signifikansnivå* är risken att göra ett Typ1-fel, dvs att förkasta en korrekt nollhypotes. Denna risk är här bestämd till 5%. En tolkning av signifikansnivån har alltså som utgångspunkt att nollhypotesen är sann vilket här är att andelen passagerare med en bagagevikt på 20 kg eller mer är (inte överstiger) 60%. Alltså; om det är så att andelen passagerare med en bagagevikt på 20 kg eller mer är 60% finns ändå en risk för att vi på grund av slumpmässig variation i vår undersökning får indikationer om att denna andel faktiskt överstiger 60%. Risken för att få så starka indikationer att vi blir övertygade om att denna andel överstiger 60% (trots att den egentligen inte gör det) är i och med den uppsatta signifikansnivån begränsad till 5%.

2. Statistisk inferens i samband med två oberoende stickprov.

   (a) Vi låter

$$\begin{aligned} \sigma_m &= \text{Standardavvikelse för reaktionstid i mobilpopulationen} \\ \sigma_k &= \text{Standardavvikelse för reaktionstid i kontrollpopulationen} \end{aligned}$$

Vi är nu intresserade av att testa hypoteserna

$$\begin{aligned} H_0 &: \sigma_m = \sigma_k \\ H_1 &: \sigma_m \neq \sigma_k \end{aligned}$$

Vi förutsätter att båda stickproven är OSU dragna oberoende av varandra och att reaktionstiden är approximativt normalfördelad i båda populationerna. Vi ska använda testfunktionen

$$F = \frac{S_1^2}{S_2^2}$$

som är F-fördelad med $n_1 - 1$ frihetsgrader i täljaren och $n_2 - 1$ frihetsgrader i nämnaren då nollhypotesen är sann. Utifrån den givna informationen finner vi att

$$F = \frac{S_m^2}{S_k^2} = \frac{89.6^2}{65.4^2} = 1.877$$

där, pga F-tabellens begränsningar, den största av varianserna ställts i täljaren, dvs det är mobilgruppen som utgör population 1 och kontrollgruppen som utgör population 2. Vi jämför detta värde med F-fördelningen med 31 frihetsgrader i både täljare och nämnare. Denna kombination av frihetsgrader finns inte med i F-tabellen men en enkel interpolation ger tillsammans med det faktum att det är en tvåsidig mothypotes att den kritiska punkten ungefär är

$$F_{31,31,5\%} \approx F_{30,30,5\%} = 1.84$$

och eftersom

$$F_{obs} = 1.877 > 1.84 = F_{30,30,5\%} > F_{31,31,5\%}$$

har vi hamnat i det kritiska området och nollhypotesen förkastas. Det är således på 10% signifikansnivå statistiskt säkerställt att standardavvikelse för reaktionstid inte är samma i mobilpopulationen och kontrollpopulationen.

(b) Låter vi

$$\begin{aligned} \mu_m &= \text{Medelreaktionstid i mobilpopulationen} \\ \mu_k &= \text{Medelreaktionstid i kontrollpopulationen} \end{aligned}$$

ska utifrån frågeställningen hypoteserna formulera som

$$\begin{aligned} H_0 &: \mu_m = \mu_k \\ H_1 &: \mu_m > \mu_k \end{aligned}$$

vilka vi tänker undersöka med ett hypotestest på 5% signifikansnivå. Vi förutsätter att båda stickproven är OSU dragna oberoende av varandra. Eftersom

$$n_m = n_k = 32 > 30$$

är båda urvalen förhållandevis stora vilket innebär att vi nu till skillnad från a-uppgiften *inte* behöver förutsätta att reaktionstiden är approximativt normalfördelad i de båda populationerna. De båda populationerna bör rimligtvis kunna anses vara mycket stora vilket tillsammans med att vi i a-uppgiften fann att det med relativt stor säkerhet gäller att $\sigma_A \neq \sigma_B$ ger att vi (i och med att stickproven är stora) använder testfunktionen

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

som approximativt är $N(0,1)$ då nollhypotesen är sann. Det är dock förstås helt okej att använda t-fördelningen där antal frihetsgrader är 56 (enligt Minitab). Eftersom testet utförs med en signifikansnivå på 5% och att mothypotesen är på formen $H_1 : \mu_m > \mu_k$ ska nollhypotesen förkastas om

$$z_{\text{obs}} > 1.6449 = z_{0.05}$$

Testfunktionen får värdet

$$z = \frac{585.2 - 533.6}{\sqrt{\frac{89.6^2}{32} + \frac{65.4^2}{32}}} = 2.63$$

Eftersom

$$z_{\text{obs}} = 2.63 > 1.6449 = z_{0.05}$$

förkastas nollhypotesen. Det är således på 5% signifikansnivå statistiskt säkerställt att medelreaktionstiden i mobilpopulationen överstiger medelreaktionstiden i kontrollpopulationen.

3. Index.

(a) Räknar vi om priset på en "88:an" till 2014 års penningvärde får vi

År	**Fast pris (2014)**
1984	$3.60 \cdot \frac{313.5}{143.2} = 7.88$
1994	$8.00 \cdot \frac{313.5}{248.5} = 10.09$
2004	$11.00 \cdot \frac{313.5}{279.2} = 12.35$
2014	$= 15.00$

så vår indexserie blir nu

År	**1984**	**1994**	**2004**	**2014**
Styckpris (2014 års nivå)	7.88	10.09	12.35	15.00
Index (Basår 2014)	52.5	67.3	82.3	100.0

(b) Vi är nu intresserade av hur mycket, procentuellt sett, styckpriset på en "88:an" har höjts totalt under den senaste 20-årsperioden, dvs mellan 1994 och 2014. Beräknat i *löpande* priser fås kvoten $15.00/8.00 = 1.875$, dvs en höjning med 87.5%. Beräknat i *fasta* priser fås kvoten $15.00/10.09 = 1.486$, dvs en höjning med 48.6%. Motsvarande beräkning för KPI ger $313.5/248.5 = 1.262$, dvs under samma period har KPI ökat med 26.2%.

4. Vi börjar med att göra observationen att sannolikheten att skylten får korrekt text är 0.5 i *nästan* samtliga möjliga fall. Det finns dock två fall där detta inte stämmer och det är om de båda nedfallna bokstäverna är identiska, dvs om båda är A eller båda är P. Vi börjar därför med att definiera *apriorihändelserna*, dvs

$$\begin{aligned} B_1 &= \text{De nedfallna bokstäverna är olika} \\ B_2 &= \text{De nedfallna bokstäverna är identiska (dvs två A eller två P)} \end{aligned}$$

Hur ser då sannolikheterna ut för dessa händelser? Av de sju bokstäverna i UPPSALA kan två väljas på $\binom{7}{2} = 21$ olika sätt och eftersom det sker helt slumpmässigt är alla dessa lika sannolika. Eftersom det endast är två utfall som leder fram till att de båda nedfallna bokstäverna är identiska följer därmed att

$$\begin{aligned} \Pr(B_1) &= 19/21 \\ \Pr(B_2) &= 2/21 \end{aligned}$$

Vi definierar nu händelsen

$$A = \text{Skylten får korrekt text}$$

Sannolikheten för B beror på om de nedfallna bokstäverna är identiska eller inte och det gäller att

$$\begin{aligned} \Pr(A \mid B_1) &= 1/2 \\ \Pr(A \mid B_2) &= 1 \end{aligned}$$

Med hjälp av Satsen om total sannolikhet får vi nu att

$$\begin{aligned} \Pr(A) &= \Pr(A \mid B_1)\Pr(B_1) + \Pr(A \mid B_2)\Pr(B_2) = \\ &= \frac{1}{2} \cdot \frac{19}{21} + 1 \cdot \frac{2}{21} = \mathbf{\frac{23}{42} \approx 0.55} \end{aligned}$$

dvs det är ungefär 55% chans att skylten får korrekt text.

5. Vi bildar slumpvariabeln

$$Y = \text{Totalt antal steg på fem tärningskast} = X_1 + X_2 + X_3 + X_4 + X_5$$

där X_i är utfallet i det i:te kastet.

(a) Nu gäller det att finna argument för att normalfördelningen skall kunna användas för att approximera sannolikheter för Y. Vad som talar för är att det helt korrekt är en summa av oberoende och likafördelade slumpvariabler. Vad som då talar emot är förstås att det enbart är fem sådana slumpvariabler (tärningskast) eftersom man generellt kräver åtminstone 30 observationer för att en sådan summa skall kunna anses som normalfördelad. Nu känner vi dock till den aktuella sannolikhetsfördelningen och denna är symmetrisk (samma sannolikhet för samtliga sex utfall) vilket innebär att det inte krävs lika många observationer för att sannolikhetsfördelningen för summan skall vara normalfördelningslik. Fem kast är antagligen ändå i minsta laget.

(b) För att kunna utföra beräkningen måste vi finna väntevärde och standardavvikelse för den ursprungliga sannolikhetsfördelningen, dvs för ett tärningskast.

$$\begin{aligned} E(X) &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = 3.5 \\ E\left(X^2\right) &= 1 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + \cdots + 36 \cdot \frac{1}{6} = \frac{91}{6} \\ \sigma(X) &= \sqrt{\frac{91}{6} - 3.5^2} = \sqrt{\frac{35}{12}} \end{aligned}$$

Vi finner nu motsvarande parametrar för Y via

$$\begin{aligned} E(Y) &= 5 \cdot 3.5 = 17.5 \\ \sigma(Y) &= \sqrt{5 \cdot \frac{35}{12}} \approx 3.8 \end{aligned}$$

Om vi kastar en tärning fem gånger och lägger ihop resultaten kommer vi i genomsnitt att få 17.5. Vi kommer dock aldrig att få detta värde och i genomsnitt kommer summan att avvika från det med ca 3.8. Om vi här accepterar det faktum att vi enbart har fem observationer gäller enligt Centrala gränsvärdessatsen att Y approximativt är $N(17.5, 3.8)$. Vi finner nu den sökta sannolikheten till

$$\Pr(Y \leq 15) \approx \Pr\left(Z \leq \frac{15.5 - 17.5}{3.8} = -0.52\right) \approx \mathbf{0.3}$$