STOCKHOLMS UNIVERSITET HT2012

Statistiska institutionen

Michael Carlson

**TENTAMEN I STATISTIKENS GRUNDER 2**

**2012-11-01**

**Skrivtid:** kl 9.00 - 14.00

**Godkända hjälpmedel:** Miniräknare, språklexikon

**Bifogade hjälpmedel:** Formelblad, statistiska tabeller

* Tentamen består av 5 uppgifter, i förekommande fall uppdelade i deluppgifter. Maximalt antal poäng anges per uppgift.
* För full poäng på en uppgift krävs tydliga, utförliga och väl motiverade lösningar.
* Kontrollera alltid dina beräkningar och lösningar! Slarvfel kan också ge poängavdrag!

**Uppgift 1.**

*Prisbasbeloppet* används för att räkna fram t.ex. maximal lönegaranti, pensioner, pensionsavgifter mm. Beloppet fastställs för helt kalenderår och ska spegla och justera för inflationen i samhället och räknas fram på grundval av förändringarna i det allmänna prisläget med utgångs­punkt i *konsumentprisindex* (KPI). KPI är det mest använda måttet för prisutveckling och används bl.a. som inflationsmått och vid av­talsreglering. KPI avser att visa hur konsumentpriserna i genomsnitt utvecklar sig för hela den privata inhemska konsumtionen, de priser konsumenterna faktiskt betalar. Nedan anges prisbasbelopp och KPI för åren 2006-2010 (källa: SCB).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| År | 2006 | 2007 | 2008 | 2009 | 2010 |
| Prisbasbelopp | 39700 | 40300 | 41000 | 42800 | 42400 |
| KPI (basår = 1980) | 284,22 | 290,51 | 300,61 | 299,66 | 303,46 |

1. Beräkna ett *enkelt* index för prisbasbeloppet för åren 2006-2010 med år 2006 som basår. Åskådliggör indextalen i ett lämpligt diagram. (6p)
2. Räkna om KPI till det nya basåret 2006 och räkna sedan om prisbasbeloppet till fasta priser uttryckta i 2006 års priser. (8p)
3. Beräkna ett reviderat enkelt index för prisbasbeloppet i fasta priser uttryckta i 2006 års priser dvs. de fasta priser som framräknats i b)-uppgiften. Åskådliggör de nya indextalen i ett lämpligt diagram. (6p)

**Uppgift 2.**

Spirometri är en lungfunktionsundersökning som utförs med spirometer, det vill säga en undersökning av hur lungorna fungerar. Forcerad vitalkapacitet (FVC) är den volym luft som en människa kan blåsa ut efter en maximal inandning. Från *n* = 271 patientjournaler beräknades medelvärdet hos kvinnor i åldern 20-30 till 3,4 liter och stickprovvariansen till 0,04.

1. Beräkna ett 90 % konfidensintervall för den genomsnittliga volymen. Ange tydligt vilka antaganden som du gör. (8p)
2. Ge en tolkning av konfidensintervallet i a)-uppgiften. (4p)
3. Hur stort urval skulle krävas för att längden av ett 90 % konfidensintervall skulle bli högst 0,01? (8p)

**Uppgift 3.**

En stor statistikproducent har stora problem med ökande svarsbortfall i ett antal stora och viktiga undersökningar. För att undersöka om olika typer av gåvor kunde öka svarsbenägenheten genomfördes ett experiment där urvalspersonerna i en enkätundersökning slumpmässigt delades in i tre grupper: en grupp fick ingen gåva, en grupp fick en nyckelring och en grupp fick en Trisslott om de deltog. Utfallet av experimentet redovisas i tabellen nedan:

|  |  |  |  |
| --- | --- | --- | --- |
| Grupp: | 1 | 2 | 3 |
|  | Ingen gåva | Nyckelring | Trisslott |
| Svarade på enkäten | 198 | 200 | 220 |
| Svarade ej på enkäten  | 202 | 200 | 180 |

1. Ange väntevärdesriktiga (eng. *unbiased*) punktskattningar för svarsbenägenheten dvs. andelen som svarar, för var och en av de tre grupperna. Förklara kortfattat vad som menas med väntevärdesriktiga punktskattningar. (8p)
2. Undersök med hjälp av dessa data om svarsbenägenheten påverkas av gåvor. Genomför ett formellt hypotestest med signifikansnivån 5 % och ange tydligt vilka förutsättningar som gäller för testet. (12p)

**Uppgift 4.**

Ett godisföretag har tagit fram en ny förbättrad version i ny förpackning av sin stor­säljande halstablett med eukalyptusmak. Man har låtit 900 slumpmässigt utvalda personer oberoende av varandra testa både den gamla och den nya varianten. Av dessa svarade 495 att de fördrog den nya varianten.

1. Testa om proportionen som föredrar den nya varianten är större än 50 % på signifi­kansnivå 5 %. Ange tydligt vilka antaganden som du gör. (10p)
2. Beräkna sannolikheten för att förkasta *H*0 i a)-uppgiften om den sanna propor­tionen är 55 %. Tips: skriv först om den kritiska gränsen i a)-uppgiften som en kritisk gräns för den observerade proportionen. (10p)

**Uppgift 5.**

En firma använder sig av posten för att distribuera en viss produkt. Då produkten är känslig för ovarsam behandling har man från annat håll beställt ett särskilt emballage. Som underlag för portoberäkning vill man få en uppfattning dels om produktens genomsnittliga vikt men även emballagets genomsnittliga vikt. Man har därför vägt fyra färdiga produkter och oberoende därav fem emballage. Följande resultat i gram erhölls:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *X* = Produktens vikt: | 442 | 448 | 447 | 439 |
|  |  |  |  |  |
| *Y* = Emballagets vikt:  | 80 | 86 | 91 | 83 | 90 |

1. Beräkna ett 99 %-igt konfidensintervall för produktens förväntade vikt μ*X*. Ange tydligt vilka antaganden som du gör. (10p)
2. Beräkna ett 99 %-igt konfidensintervall för den förväntade totala vikten μ*X* + μ*Y*. Anta att produkternas och emballagens varianser är lika, dvs. $σ\_{X}^{2}=σ\_{Y}^{2}$. Ange tydligt vilka övriga antaganden som du gör. (10p)

**FORMLER**

Räkneregler för väntevärden och varianser (*a*, *b* och *c* är konstanter och *X* och *Y* är stokastiska variabler)

|  |  |
| --- | --- |
| $$E\left(c\right)=c$$ | $$V\left(c\right)=0$$ |
| $$E\left(X+c\right)=E\left(X\right)+c$$ | $$V\left(X+c\right)=V(X)$$ |
| $$E\left(aX\right)=aE(X)$$ | $$V\left(aX\right)=a^{2}V(X)$$ |
| $$E\left(aX+bY+c\right)=aE\left(X\right)+bE\left(Y\right)+c$$ | $$V\left(aX+bY+c\right)=a^{2}V\left(X\right)+b^{2}V\left(Y\right)+2abCov\left(X,Y\right)$$ |

Ändlighetskorrektion: $\frac{N-n}{N-1}$

Stickprovsvarians: $s^{2}=\frac{1}{n-1}\sum\_{i=1}^{n}\left(x\_{i}-\overbar{x}\right)^{2}=\frac{1}{n-1}\left(\sum\_{i=1}^{n}x\_{i}^{2}-n\overbar{x}^{2}\right)$

Stickprovskovarians: $s\_{xy}=\sum\_{i=1}^{n}\left(x\_{i}-\overbar{x}\right)\left(y\_{i}-\overbar{y}\right)=\frac{1}{n-1}\left(\sum\_{i=1}^{n}x\_{i}y\_{i}-n\overbar{x}\overbar{y}\right)$

Binomialfördelningen: $f\left(x\right)=\left(\begin{matrix}n\\x\end{matrix}\right)p^{x}(1-p)^{n-x}=\frac{n!}{x!\left(n-x\right)!}p^{x}(1-p)^{n-x}$

Poissonfördelningen: $f\left(x\right)=\frac{λ^{x}e^{-λ}}{x!}$

Diverse konfidensintervall och enkelsidiga testvariabler (där *f.g.* = frihetsgrader):

|  |  |
| --- | --- |
| $$\overbar{x}\pm z\_{α/2}\frac{σ}{\sqrt{n}}$$ | $$Z=\frac{\overbar{X}-μ\_{0}}{σ/\sqrt{n}}\geq z\_{α}$$ |
| $$\overbar{x}\pm t\_{α/2}^{(f.g.)}\frac{s}{\sqrt{n}}$$ | $$T=\frac{\overbar{X}-μ\_{0}}{S/\sqrt{n}}\geq t\_{α}^{(f.g.)}$$ |
| $$\overbar{x}\pm z\_{α/2}\frac{s}{\sqrt{n}}$$ | $$Z=\frac{\overbar{X}-μ\_{0}}{S/\sqrt{n}}\geq z\_{α}$$ |
| $$\overbar{x}\_{1}-\overbar{x}\_{2}\pm z\_{α/2}\sqrt{\frac{σ\_{1}^{2}}{n\_{1}}+\frac{σ\_{2}^{2}}{n\_{2}}}$$ | $$Z=\frac{\overbar{X}\_{1}-\overbar{X}\_{2}-0}{\sqrt{{σ\_{1}^{2}}/{n\_{1}}+{σ\_{2}^{2}}/{n\_{2}}}}\geq z\_{α}$$ |
| $$\overbar{x}\_{1}-\overbar{x}\_{2}\pm t\_{α/2}^{(f.g.)}∙s\_{p}\sqrt{\frac{1}{n\_{1}}+\frac{1}{n\_{2}}}$$ | $$T=\frac{\overbar{X}\_{1}-\overbar{X}\_{2}-0}{S\_{p}\sqrt{{1}/{n\_{1}}+{1}/{n\_{2}}}}\geq t\_{α}^{(f.g.)}$$ |
| $$där s\_{p}^{2}=\frac{\left(n\_{1}-1\right)s\_{1}^{2}+\left(n\_{2}-1\right)s\_{2}^{2}}{n\_{1}+n\_{2}-2}$$ |
| $$\overbar{x}\_{1}-\overbar{x}\_{2}\pm z\_{α/2}\sqrt{\frac{s\_{1}^{2}}{n\_{1}}+\frac{s\_{2}^{2}}{n\_{2}}}$$ | $$Z=\frac{\overbar{X}\_{1}-\overbar{X}\_{2}-0}{\sqrt{{S\_{1}^{2}}/{n\_{1}}+{S\_{2}^{2}}/{n\_{2}}}}\geq z\_{α}$$ |

Forts. konfidensintervall och enkelsidiga testvariabler (där *f.g.* = frihetsgrader):

|  |  |
| --- | --- |
| $$\overbar{d}\pm t\_{α/2}^{(f.g.)}\frac{s\_{d}}{\sqrt{n}}$$ | $$T=\frac{\overbar{D}-0}{S\_{D}/\sqrt{n}}\geq t\_{α}^{(f.g.)}$$ |
|  |  |
| $$\frac{y}{n}\pm z\_{α/2}\sqrt{\left(y/n\right)(1-y/n)/n}$$ | $$Z=\frac{Y/n-π\_{0}}{\sqrt{π\_{0}(1-π\_{0})/n}}\geq z\_{α}$$ |
| $$\hat{p}\_{1}-\hat{p}\_{2}\pm z\_{α/2}\sqrt{\frac{\hat{p}\_{1}(1-\hat{p}\_{1})}{n\_{1}}+\frac{\hat{p}\_{2}(1-\hat{p}\_{2})}{n\_{2}}}$$ | $$Z=\frac{Y\_{1}/n\_{1}-Y\_{2}/n\_{2}-0}{\sqrt{\left(\frac{Y\_{1}+Y\_{2}}{n\_{1}+n\_{2}}\right)\left(1-\frac{Y\_{1}+Y\_{2}}{n\_{1}+n\_{2}}\right)\left(\frac{1}{n\_{1}}+\frac{1}{n\_{2}}\right)}}\geq z\_{α}$$ |
|  |  |
|  | $$χ^{2}=\sum\_{}^{}\frac{(n\_{i}-nπ\_{i})^{2}}{nπ\_{i}}\geq χ\_{α}^{2}(f.g.)$$ |
|  | $$χ^{2}=\sum\_{}^{}\sum\_{}^{}\frac{(n\_{ij}-n\_{i∙}n\_{∙j}/n)^{2}}{n\_{i∙}n\_{∙j}/n}\geq χ\_{α}^{2}(f.g.)$$ |

**LÖSNINGSFÖRSLAG (kortfattade)**

**Uppgift 1.**

1. Beräkning av enkelt index för prisbasbelopp med basår = 2006 (diagram, se nedan):

|  |  |  |  |
| --- | --- | --- | --- |
| År | Prisbasbelopp | Beräkning | Index |
| 2006 | 39700 | 100∙397/397 | 100,0 |
| 2007 | 40300 | 100∙403/397 | 101,5 |
| 2008 | 41000 | 100∙410/397 | 103,3 |
| 2009 | 42800 | 100∙428/397 | 107,8 |
| 2010 | 42400 | 100∙424/397 | 106,8 |

1. Omräkning av KPI till basår = 2006:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| År | KPI (1980) | Beräkning 1 | KPI (2006) | Beräkning 2 | Prisbasbelopp i fasta priser |
| 2006 | 284,22 | 100∙284,22/284,22 | 100,00 | 39700/1,0000 | 39700 |
| 2007 | 290,51 | 100∙290,51/284,22 | 102,21 | 40300/1,0221 | 39427 |
| 2008 | 300,61 | 100∙300,61/284,22 | 105,77 | 41000/1,0577 | 38765 |
| 2009 | 299,66 | 100∙299,66/284,22 | 105,43 | 42800/1,0543 | 40595 |
| 2010 | 303,46 | 100∙303,46/284,22 | 106,77 | 42400/1,0677 | 39712 |

1. Beräkning av enkelt index (basår = 2006) för prisbasbelopp i fasta priser (2006 års priser):

|  |  |  |  |
| --- | --- | --- | --- |
| År | Prisbasbelopp i fasta priser | Beräkning | Index |
| 2006 | 39700 | 100∙39700/39700 | 100,00 |
| 2007 | 39427 | 100∙39427/39700 | 99,31 |
| 2008 | 38765 | 100∙38765/39700 | 97,64 |
| 2009 | 40595 | 100∙40595/39700 | 102,25 |
| 2010 | 39712 | 100∙39712/39700 | 100,05 |

**Uppgift 2.**

*n* = 271, $\overbar{x}$ = 3,4 och s2 = 0,04. Fördelningen för de enskilda värdena *Xi* är okänd men fördelningen för $\overbar{X}$ är approximativt *N*(μ, σ2/*n*) enligt CGS ty stick­provet kan antas vara tillräckligt stort, *n* ≥ 30. Här är dock σ2 okänd så vi använder *s*2 istället (eftersom *S*2 är en väntevärdesriktig skattning av σ2 så bör approximationen fungera hyfsat bra med stort *n*). Om man antar att observationerna är normalförde­lade så är $\overbar{X}$ normal­fördelad men då får man motivera valet av *z* isf *t*-fördelningen då σ2 är okänd.

1. Ett (approximativt) 90 % KI ges av

$$\overbar{x}\pm z\_{0,05}\frac{s}{\sqrt{n}} = 3,4\pm 1,6449∙\frac{0,2}{\sqrt{271}} = 3,4\pm 0,019984 ≈ 3,4\pm 0,02$$

1. Om man upprepar proceduren och drar flera stickprov av samma storlek kommer ca 90 % av konfidensintervallen täcka det sanna värdet på μ. Hur det förhåller sig med detta observerade intervall går inte att säga något om.
2. Anta att σ2 ≈ s2 = 0,04.

$$z\_{0,05}\frac{σ}{\sqrt{n}} \leq \frac{0,01}{2} ⇔ \frac{1,6449^{2}∙0,2^{2}}{0,005^{2}} \leq n⇔4329,114 \leq n$$

Välj *n* = 4330.

**Uppgift 3.**

1. π*i* = svarsbenägenheten i respektive grupp *i* = 1, 2, 3. Väntevärdes­riktiga punktskattningar för π*i*:

*p*1 = 198/400 = 0,495; *p*2 = 200/400 = 0,500; *p*3 = 220/400 = 0,550

Om väntevärdet för en punktskattning är just den parameter den avser att skatta så är den väntevärdesriktig, t.ex. i detta fall *E*(*Pi*) = π*i*

1. χ2-test för homogenitet (oberoende).

*H*0 : Svarsbenägenhet beror ej på gåva, dvs. π1 = π2 = π3

*H*1 : Svarsbenägenhet beror på gåva, dvs. ej alla π1 , π2 , π3 är lika

Signifikansnivå α = 5 %

Testvariabel $χ^{2}=\sum\_{}^{}\sum\_{}^{}\frac{(n\_{ij}-n\_{i∙}n\_{∙j}/n)^{2}}{n\_{i∙}n\_{∙j}/n}$ som är approximativt χ2-fördelad med 2 frihetsgrader. Förkasta *H*0 om $χ\_{obs}^{2}\geq χ\_{0,05}^{2}\left(2\right)$ = [enl. tabell] = 5,991.

Beräkningar:

|  |  |  |
| --- | --- | --- |
| Observerat: |  | Förväntat: |
| 198 | 200 | 220 | 618 |  | 206 | 206 | 206 | 618 |
| 202 | 200 | 180 | 582 |  | 194 | 194 | 194 | 582 |
| 400 | 400 | 400 | 1200 |  | 400 | 400 | 400 | 1200 |

$χ\_{obs}^{2}$ = 2,963 < 5,991. Slutsatsen är att *H*0 inte kan förkastas, det finns inget signifikant samband mellan gåvor och svarsbenägenhet på 5 % nivån.

**Uppgift 4.**

1. *n* = 900 oberoende observationer, *Y* = antal ”Ja-svar”, *Y* ~ *Bin*(900,π). Enligt CGS är *Y*/*n* approximativt normalfördelad med väntevärde π och varians π(1- π)/*n* ty stickprovet kan anses vara tillräckligt stort, *n* ≥ 30.

*H*0 : π = 0,5

*H*1 : π > 0,5

Signifi­kansnivå α = 5 %. Testvariabel är

$$Z=\frac{Y/n-π\_{0}}{\sqrt{π\_{0}(1-π\_{0})/n}}=\frac{Y/900-0,5}{\sqrt{0,5(1-0,5)/900}}=\frac{Y}{15}-30$$

som approximativt är *N*(0,1)-fördelad. Förkasta *H*0 om *zobs* > *z*0,05 = 1,6449 enl. tabell. *zobs* = 495/15 – 30 = 3 > 1,6449. Slutsatsen är att *H*0 förkastas, andelen som anser att den nya varianten är bättre är signifikant större än 50 % på 5 % nivån.

1. Kritiskgräns i a) kan skrivas som

$$\frac{Y}{15}-30> 1,6449 ⇔ \frac{Y}{900}>0,527415$$

Sannolikheten att förkasta *H*0 givet att π = 0,55:

$$P\left( \frac{Y}{n} >0,527415\right)=P\left( \frac{Y/n-0,55}{\sqrt{0,55(1-0,45)/900}} >\frac{0,527415-0,55}{\sqrt{0,55(1-0,45)/900}}\right)$$

= *P*(*Z* > -1,361927) ≈ Φ(1,36) = [enligt tabell] = 0,91309

**Uppgift 5.**

1. Antagande: alla *Xi* är normalfördelade med samma väntevärde μ*X* och varians $σ\_{X}^{2}$. Med okänd varians ges ett 99 %-igt KI för förväntad produktvikt μ*X* då av

$$\overbar{x}\pm t\_{0,005}^{(f.g.)}\frac{s\_{x}}{\sqrt{n\_{x}}}$$

*nx* – 1 = 3 frihetsgrader då vilket ger $t\_{0,005}^{(3)}$ = 5,841 enl. tabell. $\overbar{x}$ = 444 och $s\_{x}^{2}$ = 18. Insättning ger

$$\overbar{x}\pm t\_{0,005}^{(3)}\frac{s\_{x}}{\sqrt{n\_{x}}} = 444\pm 5,841∙\frac{4,242641}{\sqrt{4}} = 444\pm 12,4 $$

eller (431,6 ; 456,4)

1. Samma antagande som i a) samt att alla *Yj* är oberoende och normalfördelade med samma vänte­värde μ*Y* och varians $σ\_{Y}^{2}$, samt att $σ\_{X}^{2}=σ\_{Y}^{2}=σ^{2}$ dvs. *Xi* och *Yj* har samma varians. Anta också att stick­proven är oberoende sinsemellan.

$\overbar{X}+\overbar{Y}$ är då normalfördelad med väntevärde μ*X* + μ*Y* och varians σ2(1/*nx*+1/*ny*) vilket enkelt kan verifieras med formlerna för linjärkombinationer.

Då σ2 är okänd skattas den med $s\_{p}^{2}$ enligt formelbladet. Situationen är med andra ord densamma som om vi skulle betrakta differensen $\overbar{X}-\overbar{Y}$. Summan $\overbar{X}+\overbar{Y}$ eller snarare transforma­tionen

$$\frac{\left(\overbar{X}+\overbar{Y}\right)-\left(μ\_{X} + μ\_{Y}\right)}{S\_{p}\sqrt{{1}/{n\_{x}}+{1}/{n\_{x}}}} \~ t\left(n\_{X}+n\_{Y}-2\right)=t\left(7\right)$$

är då också *t*-fördelad med *nx* + *ny* – 2 = 7 frihetsgrader då vi har (lite slarvigt formu­lerat) något normalfördelat i täljaren och något χ2-fördelat i nämnaren.

Vi observerar $\overbar{y}$ = 86 och $s\_{y}^{2}$ = 21,5. Den poolade variansskattningen beräknas till

$$s\_{p}^{2}=\frac{\left(n\_{x}-1\right)s\_{x}^{2}+\left(n\_{y}-1\right)s\_{y}^{2}}{n\_{x}+n\_{y}-2}=\frac{3∙18+4∙21,5}{7}=20$$

och insättning ger ett 99 %-igt KI för den förväntade totala vikten μ*X* + μ*Y* enligt

$$\left(\overbar{x}\_{1}+\overbar{x}\_{2}\right)\pm t\_{0,005}^{(7.)}∙s\_{p}\sqrt{\frac{1}{n\_{1}}+\frac{1}{n\_{2}}} = \left(444+86\right)\pm 3,499∙4,472136\sqrt{\frac{1}{4}+\frac{1}{5}}$$

≈ 530 ± 10,5 eller (519,5 ; 540,5).