

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 *konsumtprisindex* (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

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

- Beräkna ett 90 % konfidensintervall för den genomsnittliga volymen. Ange tydligt vilka antaganden som du gör. (8p)
- Ge en tolkning av konfidensintervallet i a)-uppgiften. (4p)
- 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

- 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)
- 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 stor-säljande 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 föredrog den nya varianten.

- Testa om proportionen som föredrar den nya varianten är större än 50 % på signifikansnivå 5 %. Ange tydligt vilka antaganden som du gör. (10p)
- 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 erhöles:

$X =$ Produktens vikt:	442	448	447	439
------------------------	-----	-----	-----	-----

$Y =$ Emballagets vikt:	80	86	91	83	90
-------------------------	----	----	----	----	----

- a) Beräkna ett 99 %-igt konfidensintervall för produktens förväntade vikt μ_X . Ange tydligt vilka antaganden som du gör. (10p)
- b) Beräkna ett 99 %-igt konfidensintervall för den förväntade totala vikten $\mu_X + \mu_Y$. Anta att produkternas och emballagens varianser är lika, dvs. $\sigma_X^2 = \sigma_Y^2$. 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)

$$E(c) = c$$

$$V(c) = 0$$

$$E(X + c) = E(X) + c$$

$$V(X + c) = V(X)$$

$$E(aX) = aE(X)$$

$$V(aX) = a^2V(X)$$

$$E(aX + bY + c) = aE(X) + bE(Y) + c \quad V(aX + bY + c) = a^2V(X) + b^2V(Y) + 2abCov(X, Y)$$

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

Stickprovsvarians: $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} (\sum_{i=1}^n x_i^2 - n\bar{x}^2)$

Stickprovskovarians: $s_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} (\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y})$

Binomialfördelningen: $f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$

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

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

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \geq z_{\alpha}$$

$$\bar{x} \pm t_{\alpha/2}^{(f.g.)} \frac{s}{\sqrt{n}}$$

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \geq t_{\alpha}^{(f.g.)}$$

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

$$Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \geq z_{\alpha}$$

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - 0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \geq z_{\alpha}$$

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}^{(f.g.)} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$T = \frac{\bar{X}_1 - \bar{X}_2 - 0}{s_p \sqrt{1/n_1 + 1/n_2}} \geq t_{\alpha}^{(f.g.)}$$

$$\text{där } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - 0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \geq z_{\alpha}$$

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

$$\bar{d} \pm t_{\alpha/2}^{(f.g.)} \frac{s_d}{\sqrt{n}}$$

$$T = \frac{\bar{D} - 0}{S_D/\sqrt{n}} \geq t_{\alpha}^{(f.g.)}$$

$$\frac{y}{n} \pm z_{\alpha/2} \sqrt{(y/n)(1-y/n)/n}$$

$$Z = \frac{Y/n - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \geq z_{\alpha}$$

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$Z = \frac{Y_1/n_1 - Y_2/n_2 - 0}{\sqrt{\left(\frac{Y_1+Y_2}{n_1+n_2}\right)\left(1 - \frac{Y_1+Y_2}{n_1+n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \geq z_{\alpha}$$

$$\chi^2 = \sum \frac{(n_i - n\pi_i)^2}{n\pi_i} \geq \chi_{\alpha}^2(f.g.)$$

$$\chi^2 = \sum \sum \frac{(n_{ij} - n_{i \cdot} n_{\cdot j} / n)^2}{n_{i \cdot} n_{\cdot j} / n} \geq \chi_{\alpha}^2(f.g.)$$

LÖSNINGSFÖRSLAG (kortfattade)

Uppgift 1.

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

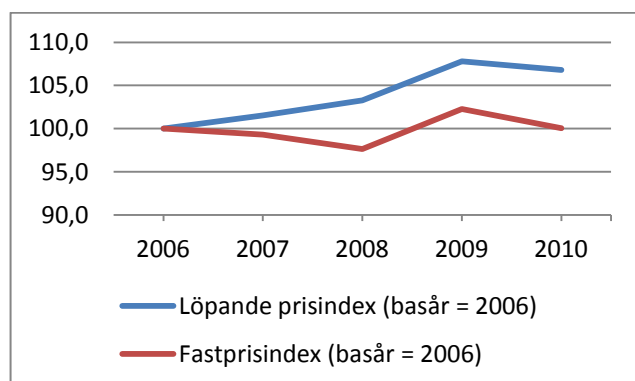
År	Prisbasbelopp	Beräkning	Index
2006	39700	$100 \cdot 397 / 397$	100,0
2007	40300	$100 \cdot 403 / 397$	101,5
2008	41000	$100 \cdot 410 / 397$	103,3
2009	42800	$100 \cdot 428 / 397$	107,8
2010	42400	$100 \cdot 424 / 397$	106,8

- b) Omräkning av KPI till basår = 2006:

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

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

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



Uppgift 2.

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

- a) Ett (approximativt) 90 % KI ges av

$$\bar{x} \pm z_{0,05} \frac{s}{\sqrt{n}} = 3,4 \pm 1,6449 \cdot \frac{0,2}{\sqrt{271}} = 3,4 \pm 0,019984 \approx 3,4 \pm 0,02$$

- b) 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.

- c) Anta att $\sigma^2 \approx s^2 = 0,04$.

$$z_{0,05} \frac{\sigma}{\sqrt{n}} \leq \frac{0,01}{2} \Leftrightarrow \frac{1,6449^2 \cdot 0,2^2}{0,005^2} \leq n \Leftrightarrow 4329,114 \leq n$$

Välj $n = 4330$.

Uppgift 3.

- a) π_i = svarsbenägenheten i respektive grupp $i = 1, 2, 3$. Väntevärdesriktiga punktskattningar för π_i :

$$p_1 = 198/400 = 0,495; \quad p_2 = 200/400 = 0,500; \quad p_3 = 220/400 = 0,550$$

Om väntevärdet för en punktskattning är just den parameter den avser att skatta så är den väntevärdesriktig, t.ex. i detta fall $E(P_i) = \pi_i$

- b) χ^2 -test för homogenitet (oberoende).

H_0 : Svarsbenägenhet beror ej på gåva, dvs. $\pi_1 = \pi_2 = \pi_3$

H_1 : Svarsbenägenhet beror på gåva, dvs. ej alla π_1, π_2, π_3 är lika

Signifikansnivå $\alpha = 5 \%$

Testvariabel $\chi^2 = \sum \sum \frac{(n_{ij} - n_i \cdot n_j / n)^2}{n_i \cdot n_j / n}$ som är approximativt χ^2 -fördelad med 2 frihetsgrader. Förkasta H_0 om $\chi_{obs}^2 \geq \chi_{0,05}^2(2) = [\text{enl. tabell}] = 5,991$.

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

$\chi_{obs}^2 = 2,963 < 5,991$. Slutsatsen är att H_0 inte kan förkastas, det finns inget signifikant samband mellan gåvor och svarsbenägenhet på 5 % nivån.

Uppgift 4.

- a) $n = 900$ oberoende observationer, $Y = \text{antal "Ja-svar"} , Y \sim \text{Bin}(900, \pi)$. Enligt CGS är Y/n approximativt normalfördelad med väntevärde π och varians $\pi(1 - \pi)/n$ ty stickprovet kan anses vara tillräckligt stort, $n \geq 30$.

$$H_0 : \pi = 0,5$$

$$H_1 : \pi > 0,5$$

Signifikansnivå $\alpha = 5 \%$. Testvariabel är

$$Z = \frac{Y/n - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}} = \frac{Y/900 - 0,5}{\sqrt{0,5(1 - 0,5)/900}} = \frac{Y}{15} - 30$$

som approximativt är $N(0,1)$ -fördelad. Förkasta H_0 om $z_{obs} > z_{0,05} = 1,6449$ enl. tabell. $z_{obs} = 495/15 - 30 = 3 > 1,6449$. Slutsatsen är att H_0 förkastas, andelen som anser att den nya varianten är bättre är signifikant större än 50 % på 5 % nivån.

- b) Kritiskgräns i a) kan skrivas som

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

Sannolikheten att förkasta H_0 givet att $\pi = 0,55$:

$$\begin{aligned} P\left(\frac{Y}{n} > 0,527415\right) &= P\left(\frac{Y/n - 0,55}{\sqrt{0,55(1 - 0,45)/900}} > \frac{0,527415 - 0,55}{\sqrt{0,55(1 - 0,45)/900}}\right) \\ &= P(Z > -1,361927) \approx \Phi(1,36) = [\text{enligt tabell}] = 0,91309 \end{aligned}$$

Uppgift 5.

- a) Antagande: alla X_i 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

$$\bar{x} \pm t_{0,005}^{(f.g.)} \frac{s_x}{\sqrt{n_x}}$$

$n_x - 1 = 3$ frihetsgrader då vilket ger $t_{0,005}^{(3)} = 5,841$ enl. tabell. $\bar{x} = 444$ och $s_x^2 = 18$.

Insättning ger

$$\bar{x} \pm t_{0,005}^{(3)} \frac{s_x}{\sqrt{n_x}} = 444 \pm 5,841 \cdot \frac{4,242641}{\sqrt{4}} = 444 \pm 12,4$$

eller (431,6 ; 456,4)

- b) Samma antagande som i a) samt att alla Y_j är oberoende och normalfördelade med samma väntevärde μ_Y och varians σ_Y^2 , samt att $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ dvs. X_i och Y_j har samma varians. Anta också att stickproven är oberoende sinsemellan.

$\bar{X} + \bar{Y}$ är då normalfördelad med väntevärde $\mu_X + \mu_Y$ och varians $\sigma^2(1/n_x + 1/n_y)$ vilket enkelt kan verifieras med formlerna för linjärkombinationer.

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

$$\frac{(\bar{X} + \bar{Y}) - (\mu_X + \mu_Y)}{S_p \sqrt{1/n_x + 1/n_y}} \sim t(n_X + n_Y - 2) = t(7)$$

är då också t -fördelad med $n_x + n_y - 2 = 7$ frihetsgrader då vi har (lite slarvigt formulerat) något normalfördelat i täljaren och något χ^2 -fördelat i nämnaren.

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

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2} = \frac{3 \cdot 18 + 4 \cdot 21,5}{7} = 20$$

och insättning ger ett 99 %-igt KI för den förväntade totala vikten $\mu_X + \mu_Y$ enligt

$$(\bar{x}_1 + \bar{x}_2) \pm t_{0,005}^{(7)} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (444 + 86) \pm 3,499 \cdot 4,472136 \sqrt{\frac{1}{4} + \frac{1}{5}}$$

$\approx 530 \pm 10,5$ eller $(519,5 ; 540,5)$.