

TENTAMEN I STATISTIKENS GRUNDER 2

2012-11-20

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.

En förpackningsmaskin används för att paketera socker. Man antar att vikten är normalfördelad med standardavvikelsen 30 gram. Från ett stickprov bestående av 25 påsar fick man ett stickprovsmedelvärde på 505 gram. Man bildade sedan ett konfidenstervall för μ som angavs till (495,2 ; 514,8).

- Vad är konfidensgraden för det angivna intervallet? Ange tydligt vilka antaganden som gäller. (8p)
- Med samma konfidensgrad som du kom fram till i a) ovan, hur stort stickprov skulle det krävas om man vill att totala längden på intervallet är högst lika med 5? (8p)
- Om man inte vill ändra stickprovsstorleken men ändå vill få ett kortare intervall, vad kan man göra? (4p)

Uppgift 2.

Vid en flygplats räknades antalet passagerare som gick igenom säkerhetskontrollen per minut. Man mätte antalet passagerare per minut under 200 slumpvis utvalda 1-minutersperioder under en vecka.

X = antal passagerare per minut	0	1	2	3	4	5	6	7	8	≥ 9
Antal 1-minutersperioder	16	50	51	43	27	9	3	0	1	0

- Beräkna en väntevärdesriktig skattning för det genomsnittliga antalet passagerare per minut. (2p)
- Anta att fördelningen för X = antalet passagerare kunder per minut är Poisson. Beräkna eller använd tabell för att ange sannolikheterna för $X = 1, 2, \dots, 8$ samt för $X \geq 9$. Utgå ifrån skattningen av väntevärdet λ du fick i a) ovan. (6p)
- Undersök med hjälp av skattningen och ovanstående data om fördelningen för X är Poisson. Genomför ett formellt hypotestest med signifikansnivån 5 % och ange tydligt testets alla steg. (12p)

Uppgift 3.

Ett experiment genomfördes för att jämföra människors reaktionstid på ett rött signalljus mot grönt ljus. Åtta försökspersoner som deltog ombads att så fort som möjligt slå till en strömbrytare när en signallampa av vardera färgen lyste upp varvid deras reaktionstider registrerades. Följande data registrerades:

Försöksperson i	1	2	3	4	5	6	7	8
Rött ljus	0,30	0,23	0,41	0,53	0,24	0,36	0,38	0,51
Grönt ljus	0,43	0,32	0,58	0,46	0,27	0,41	0,38	0,61

- Undersök om reaktionstiden vid rött ljus är kortare än vid grönt ljus. Genomför ett formellt hypotestest med signifikansnivån 5 %. Ange tydligt testets alla steg och vilka antaganden du gör. (10p)
- Förklara kortfattat begreppet p-värde och ange sedan ett ungefärligt p-värde för resultatet av testet i a). Det går bra att t.ex. ange ett intervall för p-värdet. (10p)

Uppgift 4.

Du har fått ärva lite pengar och kontaktar din bank för att diskutera olika placeringsalternativ. Du ställs inför tre möjligheter. Den första är ett s.k. certifikat med en garanterad avkastning på 5 tkr (tkr = tusen kronor) oavsett vad som händer. Det andra alternativet är en aktiefond med låg risk; vinsten beräknas bli 35 tkr om börsen går upp men om börsen går ner beräknas det bli en förlust på 10 tkr. Det tredje alternativet är en utvecklingsfond med lite högre risk; går börsen upp kan du tjäna 55 tkr men går det ner så kan du förlora 25 tkr.

- Vilket investeringsalternativ bör väljas om maximinkriteriet används? (6p)
- Vilket investeringsalternativ bör väljas om minimax regret kriteriet används? (8p)
- Vilket investeringsalternativ bör väljas om sannolikheten för att börsen ska gå upp bedöms vara 0,5? (6p)

Uppgift 5.

En grupp läkare studerade svältrelaterade sjukdomar i två hårt drabbade områden i Afrika. Två stickprov från område A respektive område B gav följande resultat: 520 av 1300 undersökta förskolebarn i område A och 385 av 1100 i område B befanns vara kroniskt undernärda.

- a) Ange ett 90 % konfidensintervall för andelen kroniskt undernärda i område A. Ange tydligt vilka antaganden du gör. (8p)
- b) Ange ett 90 % konfidensintervall för skillnaden mellan område A och B. Ange tydligt vilka antaganden du gör. (8p)
- c) Skulle du med ledning av resultatet i b) och utan att genomföra några ytterligare beräkningar kunna säga att andelarna kroniskt undernärda i område A och B är signifikant skilda på 10 % nivå? Förklara varför. (4p)

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) Under givna antaganden (normalfördelade observationer med känd varians) gäller att konfidensintervallet för μ ska beräknas enligt $\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$. Insättning av givna storheter ger

$$505 \pm z_{\alpha/2} \frac{30}{\sqrt{25}} \text{ eller } 505 \pm 6z_{\alpha/2} \Leftrightarrow 6z_{\alpha/2} = \frac{514,8 - 495,2}{2} = 9,8$$

$$\Leftrightarrow z_{\alpha/2} = 1,6333 \approx 1,63$$

Enligt Tabell 1 är $P(Z < 1,63) = 0,94845$ vilket ger att konfidensgraden för intervallet $1 - \alpha = 1 - 2 \cdot (1 - 0,94845) = 0,8969$. Ett avrundat svar 0,90 är ok.

- b) Använd gärna samma värde på $z_{\alpha/2}$ som härleddes i a):

$$z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \frac{5}{2}; \text{ insättning ger } \left(\frac{9,8}{6}\right)^2 \frac{30^2}{2,5^2} = 384,16 \leq n$$

dvs. minst 385 krävs. Svar baserade på avrundningar av $z_{\alpha/2}$ och α är ok.

- c) Om stickprovsstorlek n och standardavvikelse σ står fast så är det endast $z_{\alpha/2}$ som kan påverka konfidensintervallet längd; man får ändra konfidensgraden $1 - \alpha$ för att ändra intervallet längd.

Uppgift 2.

- a) En väntevärdesriktig skattning för det genomsnittliga antalet passagerare per minut (n_x = antalet observationer el. frekvensen där $X = x$).

$$\bar{x} = \frac{1}{n} \sum_{x=0}^8 x n_x = \frac{1}{200} (0 \cdot 16 + 1 \cdot 50 + \dots + 8 \cdot 1) = 2,3$$

- b) Fördelningen för $X \sim Po(2,3)$. Sannolikheterna för $X = 1, 2, \dots, 8$ samt för $X \geq 9$ (uppgiften krävde inte att man skulle beräkna sannolikheten för $X = 0$ men den tas med här för att man ska kunna verifiera att summan blir 1):

x	0	1	2	3	4	5	6	7	8	≥ 9
$P(X=x)$	0,1003	0,2306	0,2652	0,2033	0,1169	0,0538	0,0206	0,0068	0,0019	0,0006

- c) χ^2 -test för Goodness-of-fit el. anpassningstest.

$H_0 : X \sim Po(\lambda)$, där λ okänt

$H_1 : X \text{ ej } Po(\lambda)$

Signifikansnivå $\alpha = 5 \%$

Testvariabel: $\chi^2 = \sum \frac{(n_x - np_x)^2}{np_x}$

där n_x = observerat antal sådana att $X = x$, $n = 200$ är totala antalet observationer och p_x är den skattade sannolikheten $P(X = x)$ under H_0 och med $\hat{\lambda} = \bar{x} = 2,3$; dvs. $np_x = 200p_x$ = förväntat antal under H_0 och med skattat värde på λ .

Då det förväntade antalet under H_0 , dvs. np_x , enligt tumregeln ska vara större än 5 i varje cell slås cellerna för $X = 6, 7, 8$ samt $X \geq 9$ slås ihop till en cell, $X \geq 6$.

Testvariabeln är approximativt χ^2 -fördelad under H_0 med $7-1-1 = 5$ frihetsgrader, 7 klasser, -1 för att de ska summera till 200 samt ytterligare -1 för att man måste skatta parametern λ .

Förkasta H_0 om $\chi_{obs}^2 \geq \chi_{0,05}^2(5) = [\text{enl. tabell 4}] = 11,070$.

Beräkningar:

x	0	1	2	3	4	5	≥ 6	Σ
n_x	16	50	51	43	27	9	4	200
p_x	0,10026	0,23060	0,26518	0,20331	0,11690	0,05378	0,02998	1
np_x	20,052	46,119	53,037	40,662	23,380	10,755	5,995	200
$(n_x - np_x)^2 / np_x$	0,81872	0,32658	0,07823	0,13447	0,56035	0,28638	0,66397	2,8687

$\chi_{obs}^2 = 2,8687 < 11,070$. Slutsatsen är att H_0 inte kan förkastas, fördelningen för X är inte signifikant skilt från en Poissonfördelning på 5 % nivån.

Uppgift 3.

- a) Vi har $n = 8$ personer med två mätningar var, en för rött och en för grönt ljus. Dvs. parvisa observationer X_{ij} där $i = 1, \dots, 8$ och $j = R$ eller G . Vi antar att de parvisa differenserna $D_i = X_{iG} - X_{iR}$ är oberoende och normalfördelade med väntevärde μ_D och standardavvikelse σ_D , båda okända.

$H_0 : \mu_D = 0$ (ingen skillnad mellan rött och grönt)

$H_1 : \mu_D > 0$ (reaktionstiden för rött är kortare än för grönt)

Signifikansnivå $\alpha = 5 \%$

Testvariabel:

$$T = \frac{\bar{D}}{S_D/\sqrt{n}} \sim t(n-1) = t(7)$$

H_0 förkastas om $T_{obs} \geq t_{0,05}^{(7)} = [\text{enl. tabell 3}] = 1,895$

Beräkningar ger $\bar{d} = 0,0625$ och $s_d = 0,076485$. Insättning ger

$$t_{obs} = \frac{0,0625}{0,076485/\sqrt{8}} = 2,311 > 1,895$$

H_0 förkastas, det är en signifikant skillnad mellan reaktionstiderna på 5 % nivån.

- b) Anta att man har ett enkelsidigt test, t.ex. enligt a) ovan med noll- och mothypotes och en testvariabel T . Man har observerat ett stickprov och beräknat ett observerat värde på T som vi kan beteckna t_{obs} .

P-värdet är då sannolikheten att observera $T \geq t_{obs}$ under H_0 .

Ett annat sätt att se det är att vi inte i förväg fastställer en signifikansnivå utan vi antar att den kritiska gränsen sammanfaller med det faktiskt observerade värdet på testvariabeln och i efterhand beräknar signifikansnivån.

Observera att beräkningen av p-värden beror på hur testet är formulerat. Med det alternativa enkelsidiga testet med mothypotesen $H_1 : \mu_D < 0$ beräknas istället p-värdet som sannolikheten $P(T \leq t_{obs})$ och med ett dubbelsidigt test med mothypotesen $H_1 : \mu_D \neq 0$ beräknas p-värdet som sannolikheten $2 \cdot P(T \geq |t_{obs}|)$.

I detta fall observerades $t_{obs} = 2,311$ och p-värdet är sannolikheten $P(T \geq 2,311)$. I tabell 3 kan vi avläsa att sannolikheten (p-värdet) ligger mellan 0,025 och 0,05 ty

$$t_{0,05}^{(7)} = 1,895 < t_{obs} = 2,311 < t_{0,025}^{(7)} = 2,365$$

Med dator kan ett mer exakt p-värde beräknas till 0,027054 el. ca 2,7 %.

Uppgift 4.

- a) För varje handlingsalternativ utgår man ifrån *minsta* nyttan (det sämsta som kan hända) och väljer sedan det alternativ som ger störst minimal nytta, dvs. det bästa bland det sämsta. Beslutsmatris med nytta angiven per alternativ och tillstånd:

Handlings- alternativ	Börsen går		Min. nytta	
	ner	upp		
Certifikat	5	5	5	← Max.
Lågriskfond	-10	35	-10	
Högriskfond	-25	55	-25	

Certifikat väljs som investeringsalternativ.

- b) För varje tillstånd som kan inträffa utgår man ifrån bästa handlingsalternativ, dvs. det som ger störst nytta (det man borde ha valt):
- om börsen går ner är Certifikat bästa alternativ,
 - om börsen går upp är Högriskfond bästa alternativ.

Sedan beräknas differensen (*regret* el. ånger) mot dessa bästa alternativ inom varje tillstånd. För varje alternativ markeras det största värdet (max. *regret*). Slutligen väljs det handlingsalternativ som ger minsta maximal *regret*. Beslutsmatris med *regret* angiven per alternativ och tillstånd (bästa alternativ per tillstånd är markerad i fetstil):

Handlings- alternativ	Börsen går		Max. <i>regret</i>	
	ner	upp		
Certifikat	0	50	50	← Min.
Lågriskfond	15	20	20	
Högriskfond	30	0	30	

Lågriskfond väljs som investeringsalternativ.

- c) För varje handlingsalternativ beräknas förväntad nytta och det alternativ som ger störst förväntad nytta väljs. Beräkningar ger:

Handlings- alternativ	Börsen går		Förväntad nytta	
	ner	upp		
Sannolikhet	0,5	0,5		
Certifikat	5	5	5,0	← Max.
Lågriskfond	-10	35	12,5	
Högriskfond	-25	55	15,0	

Högriskfond väljs som investeringsalternativ.

Uppgift 5.

- a) Låt π_A beteckna den okända andelen barn i område A som är kroniskt undernärda. Låt Y_A beteckna antalet kroniskt undernärda barn i ett stickprov från område A bestående av $n_A = 1300$ slumpmässigt valda barn. Y_A kan då antas¹ vara $\text{Bin}(n_A, \pi_A)$. Då n_A är tillräckligt stort (>30) kan binomialfördelningen enligt CGS approximeras med en normalfördelning. Ett 90 % konfidensintervall för π_A ges av

$$\frac{y_A}{n_A} \pm z_{0,05} \sqrt{\frac{(y_A/n_A)(1 - y_A/n_A)}{n_A}}$$

där y_A är det observerade värdet på Y_A . Insättning ger

$$\frac{520}{1300} \pm 1,6449 \cdot \sqrt{\frac{(\frac{520}{1300})(1 - \frac{520}{1300})}{1300}} = 0,4 \pm 0,02235$$

eller (0,378 , 0,422) eller 37,8 % – 42,2 %.

- b) Låt π_B , n_B , Y_B och y_B beteckna motsvarande storheter för ett stickprov draget i område B. Samma antaganden som i a) gäller för dessa. Vi antar dessutom att de två stickproven är sinsemellan oberoende. Ett 90 % konfidensintervall för skillnaden $\pi_A - \pi_B$ ges av

$$\hat{p}_1 - \hat{p}_2 \pm z_{0,05} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

där $\hat{p}_k = y_k/n_k$ för $k = A$ och B . Insättning ger

$$\frac{520}{1300} - \frac{385}{1100} \pm 1,6449 \cdot \sqrt{\frac{\frac{520}{1300}(1 - \frac{520}{1300})}{1300} + \frac{\frac{385}{1100}(1 - \frac{385}{1100})}{1100}} = 0,05 \pm 0,0325$$

eller (0,0175 , 0,0825) eller 1,75 % – 8,25 %.

- c) Nej, man behöver inte testa. Då konfidensintervallet i b) inte täcker värdet noll kan man anta att skillnaden också är signifikant skilt från noll på 10 % nivån.

Men man måste komma ihåg att när man beräknar ett konfidensintervall för skillnaden mellan två andelar så antas inte att andelarna π_A och π_B är lika och därmed inte heller att varianserna för Y_A resp. Y_B är det. Med motsvarande hypotestest utgår man däremot ifrån en nollhypotes där andelarna antas vara lika stora, $H_0: \pi_A = \pi_B$, och då vägs observationerna y_A och y_B ihop för att ge en skattning av variansen; jämför formlerna i formelbladet. Siffermässigt kommer därför resultaten att skilja sig åt något.

¹ Om stickprovsdragningen sker med återläggning är det binomialfördelat men om dragningen sker utan återläggning är det *approximativt* binomialfördelat om populationen är tillräckligt stor. Se kurslitteraturen. Dock inget krav att man ska förklara detta i detalj för att få full poäng.