STOCKHOLMS UNIVERSITET HT2012

Statistiska institutionen

Michael Carlson

**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 konfi­densin­tervall för μ som angavs till (495,2 ; 514,8).

1. Vad är konfidensgraden för det angivna intervallet? Ange tydligt vilka antagan­den som gäller. (8p)
2. 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)
3. 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 |

1. Beräkna en väntevärdesriktig skattning för det genomsnittliga antalet passa­gerare per minut. (2p)
2. Anta att fördelningen för *X* = antalet passagerare kunder per minut är Poisson. Beräkna eller använd tabell för att ange sannolikheterna för *X* = 1, 2, …, 8 samt för *X* ≥ 9. Utgå ifrån skattningen av väntevärdet λ du fick i a) ovan. (6p)
3. 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 |

1. 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)
2. 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.

1. Vilket investeringsalternativ bör väljas om maximinkriteriet används? (6p)
2. Vilket investeringsalternativ bör väljas om minimax regret kriteriet används? (8p)
3. 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.

1. Ange ett 90 % konfidensintervall för andelen kroniskt undernärda i område A. Ange tydligt vilka antaganden du gör. (8p)
2. Ange ett 90 % konfidensintervall för skillnaden mellan område A och B. Ange tydligt vilka antaganden du gör. (8p)
3. 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 % nivån? Förklara varför. (4p)

**FORMLER**

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

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

Ändlighetskorrektion:

Stickprovsvarians:

Stickprovskovarians:

Binomialfördelningen:

Poissonfördelningen:

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

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  | |
|  |  |

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

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

**LÖSNINGSFÖRSLAG (kortfattade)**

**Uppgift 1.**

1. Under givna antaganden (normalfördelade observationer med känd varians) gäller att konfidensintervallet för μ ska beräknas enligt . Insätt­ning av givna storheter ger

Enligt Tabell 1 är *P*(*Z* < 1,63) = 0,94845 vilket ger att konfidensgraden för inter­vallet 1 – α = 1 – 2∙(1 – 0,94845) = 0,8969. Ett avrundat svar 0,90 är ok.

1. Använd gärna samma värde på *z*α/2 som härleddes i a):

dvs. minst 385 krävs. Svar baserade på avrundningar av *z*α/2 och α är ok.

1. Om stickprovsstorlek *n* och standardavvikelse σ står fast så är det endast *z*α/2 som kan påverka konfidensintervallet längd; man får ändra konfidensgraden 1 – α för att ändra intervallet längd.

**Uppgift 2.**

1. En väntevärdesriktig skattning för det genomsnittliga antalet passa­gerare per minut (*nx* = antalet observationer el. frekvensen där *X* = *x*).
2. Fördelningen för *X* ~ *Po*(2,3). Sannolikheterna för *X* = 1, 2, …, 8 samt för *X* ≥ 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 |

1. χ2-test för Goodness-of-fit el. anpassningstest.

*H*0 : *X* ~ *Po*(λ), där λ okänt

*H*1 : *X* ej *Po*(λ)

Signifikansnivå α = 5 %

Testvariabel:

där *nx* = observerat antal sådana att *X* = *x*, *n* = 200 är totala antalet observationer och *px* är den skattade sannolikheten *P*(*X* = *x*) under *H*0 och med ; dvs. *npx* = 200*px* = förväntat antal under *H*0 och med skattat värde på λ.

Då det förväntade antalet under *H*0, dvs. *npx*, enligt tumregeln ska vara större än 5 i varje cell slås cellerna för *X* = 6, 7, 8 samt *X* ≥ 9 slås ihop till en cell, *X* ≥ 6.

Testvariabeln är approximativt χ2-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 = [enl. tabell 4] = 11,070.

Beräkningar:

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

= 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.**

1. Vi har *n* = 8 personer med två mätningar var, en för rött och en för grönt ljus. Dvs. parvisa observationer *Xij* där *i* = 1,…,8 och *j* = R eller G. Vi antar att de parvisa differenserna *Di* = *XiG* – *XiR* är oberoende och normalfördelade med väntevärde μ*D* och standardavvikelse σ*D* ,båda okända.

*H*0 : μ*D* = 0 (ingen skillnad mellan rött och grönt)

*H*1 : μ*D* > 0 (reaktionstiden för rött är kortare än för grönt)

Signifikansnivå α = 5 %

Testvariabel:

*H*0 förkastas om *Tobs* ≥ = [enl. tabell 3] = 1,895

Beräkningar ger och Insättning ger

*H*0 förkastas, det är en signifikant skillnad mellan reaktionstiderna på 5 % nivån.

1. 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 *tobs*.

P-värdet är då sannolikheten att obser­vera *T* ≥ *tobs* 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 : μ*D* < 0 beräknas istället p-värdet som sannolikheten *P*(*T* ≤ *tobs*) och med ett dubbelsidigt test med mot­hypotesen *H*1 : μ*D* ≠ 0 beräknas p-värdet som sannolikheten 2∙*P*(*T* ≥ |*tobs*|).

I detta fall observerades *tobs* = 2,311 och p-värdet är sannolikheten *P*(*T* ≥ 2,311). I tabell 3 kan vi avläsa att sannolikheten (p-värdet) ligger mellan 0,025 och 0,05 ty

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

**Uppgift 4.**

1. 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 nyttan angiven per alternativ och tillstånd:

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

Certifikat väljs som investeringsalternativ.

1. För varje tillstånd som kan inträffa utgår man ifrån bästa handlingsalternativ, dvs. det som ger störst nytta (det man borde ha valt):
   1. om börsen går ner är Certifikat bästa alternativ,
   2. 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- | Börsen går | | Max. |  |
| alternativ | ner | upp | *regret* |  |
| Certifikat | **0** | 50 | 50 |  |
| Lågriskfond | 15 | 20 | 20 | ← Min. |
| Högriskfond | 30 | **0** | 30 |  |

Lågriskfond väljs som investeringsalternativ.

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

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

Högriskfond väljs som investeringsalternativ.

**Uppgift 5.**

1. Låt π*A* beteckna den okända andelen barn i område A som är kroniskt under­närda. Låt *YA* beteckna antalet kroniskt undernärda barn i ett stickprov från område A bestående av *nA* = 1300 slumpmässigt valda barn. *YA* kan då antas[[1]](#footnote-1) vara *Bin*(*nA*,π*A*). Då *nA* är tillräckligt stort (>30) kan binomial­fördelningen enligt CGS approximeras med en normalfördelning. Ett 90 % konfidensintervall för π*A* ges av

där *yA* är det observerade värdet på *YA*. Insättning ger

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

1. Låt π*B*, *nB* , *YB* och *yB* 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 är oberoende. Ett 90 % konfidensintervall för skillna­den π*A* – π*B* ges av

där för *k* = *A* och *B*. Insättning ger

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

1. 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 *YA* resp. *YB* är det. Med motsvarande hypotestest utgår man däremot ifrån en nollhypotes där andelarna antas vara lika stora, *H*0: π*A* = π*B,* och då vägs observationerna *yA* och *yB* 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. [↑](#footnote-ref-1)