

Skriftlig tentamen i Undersökningsmetodik inom fortsättningskursen i statistik, onsdagen den 1 december 2004

Skrivtid: 4 timmar

Hjälpmedel: miniräknare, bifogat formelblad.

Ange ditt fullständiga personnummer (dock ej ditt namn) på varje inlämnat pappersark.

Kom ihåg att en motiverad lösning/lösning med angivna formella uttryck kan ge poäng även om beräkningarna inte är helt korrekta.

Lös vänligen inte mer än en uppgift per inlämnat pappersark.

Skrivningsgenomgång: fredagen den 10 december 2004, kl. 12 i sal B705.

Uppgift 1

I. Ange fördelar med att noga definiera undersökningens mål, begrepp, variabler och målpopulation vid planering av en undersökning. (3p)

II. Hur kan man stödja respondenten så att denne håller sig bättre inom vår angivna referensperiod vid inrapportering av fakta? (3p)

III. Finns det anledningar att ibland ställa ledande frågor? Motivera svaret samt formulera en ledande fråga och beskriv den situation i vilken det kan vara motiverat att använda den. (4p)

Uppgift 2

I. Vi tänker oss en mängd U av N element från vilken vi väljer ut med likformig sannolikhet och utan återläggning en mindre mängd s , bestående av n element; $1 < n < N$. Ordningen av elementen i s betraktar vi som oväsentlig (vilket som helst urval med samma element betraktas som ett och samma urval oavsett ordningen).

a) På hur många olika sätt kan vi välja s ur U ?

b) På hur många olika sätt kan vi välja s ur U , om vi lägger till villkoret att vi vill att ett fixt element i i U alltid ska vara med i s ?

c) Hur definieras obundet slumpmässigt urval utan återläggning (OSU/uå) matematiskt?

d) Sätt upp på sedvanligt sätt, och beräkna, urvalssannolikheten π_i för att ett fixt element i i populationen vid OSU/uå blir utvalt i urvalet. (Sedvanligt sätt: antal gynsamma utfall dividerat med antal möjliga utfall; här har alla utfall lika sannolikhet varför vi får göra så.) (5p)

II. MSE

- a) Hur definieras medelkvadratfel (mean squared error - MSE) av en estimator matematiskt?
- b) Hur definieras bias av en estimator matematiskt?
- c) MSE av en estimator kan representeras som summa av två ofta förekommande termer. Härled detta uttryck från matematiska definitionen i a. (5p)

Uppgift 3

I. Minsta stickprovsstorlek

- a) Visa, genom härledning från uttrycket för variansskattning av en andel vid OSU/uå (på formelbladet), hur man bestämmer minsta stickprovsstorlek som satisfierar ett precisionskrav vid en andelsskattning, $\hat{p} \pm B$. Vi får anse att populationen är tillräckligt stor så att ändlighetskorrektionen inte behöver tas med, samt att $n - 1$ får ersättas med n .
- b) Vad blir, för fallet i a, minsta stickprovsstorleken om $B = 0.05$. (2p)

II. Gratistidningen "Metro City" hävdar, grundat i en stickprovsundersökning, att 7 av 10 Stockholmare någon gång har erfaren trängsel i tunnelbanan. Förutsatt att undersökningen genomfördes korrekt (inget bortfall, korrekt mätning, osv), vad måste ha varit tidningens minsta stickprovstorlek för att påstendet "7 av 10" skulle gälla, dvs att både "6 av 10" och "8 av 10" med hög sannolikhet är uteslutna som gällande i populationen? (Med andra ord, vad måste deras minsta stickprovsstorlek ha varit för att ett 95%-igt konfidensintervall centrerat på punktskattningen $\hat{p} = .7$ inte går längre ner än .65 och inte heller längre upp än .75.) Ändlighetskorrektionen bortser vi ifrån samt approximerar $n - 1$ med n ; vi utgår ifrån att urvalsmetoden var ett OSU/uå bland Stockholmare. (5p)

III. I en population av yngre vuxna människor var man, av någon anledning, intresserad av längden. Specifikt, om det fanns en skillnad mellan män och kvinnor med avseende på längd. En totalundersökning genomfördes. Hela populationen blev uppmätta på korrekt sätt, inget bortfall förekom. Resultatet var att, i snitt, mäns längd var 175.34 cm (med standardavvikelse 1.2345 cm), medan kvinnornas längd var 171.14 cm (med standardavvikelse 0.8754 cm). Vad är sannolikheten att män i snitt är längre än kvinnor i denna population? Motivera svaret. (Längden kan anses, inom varje kön, vara normalfördelad; för normalfördelning gäller, med sedvanlig $\alpha = .05$, att $z_{1-\alpha/2} = 1.96$, alltså för en 95%-ig konfidensgrad.) (3p)

Uppgift 4

En stickprovsundersökning, oberoende av den i uppgift 2/II, gjordes med syfte att utreda trafiksituation i Stockholm. Målpopulationen var invånare i åldern 16-84; det fanns 531705 individer i denna population. Bland annat ville man skatta andel i målpopulationen som någon gång under oktober 2004 åkte tunnelbana.

Ett OSU/uå av storlek 1100 drogs; man fick tag i alla respondenter och samtliga svarade. Inget mätfel eller annan sorst fel tycks ha förekommit.

- a) Det erhållna resultatet visade att 770 av respondenterna svarade jakande på undersökningsvariabeln. Skatta andel av målpopulationen som någon gång under oktober 2004 åkte tunnelbanan samt ge tillhörande 95%-igt konfidensintervall. (5p)

b) I samma undersökning var man intresserad även av frågeställningen liknande den i uppgift 2/II. Däremot insåg man här att respondenter kan ha erfarit trängsel i tunnelbanan enbart om de åkt tunnelbana under referensperioden; således var man här intresserad av att skatta andel tunnelbaneåkande (de i målpopulationen som någon gång under oktober 2004 åkte tunnelbana) som erfarit trängsel. Mätningen visade att 616 respondenter svarade jakande på denna hoppfråga (alltså, frågan besvarades inte av de som inte åkte tunnelbana någon gång under oktober 2004). Skatta ifrågavarande parameter samt ge tillhörande 95%-igt konfidensintervall. (5p)

Uppgift 5

SL ville skatta hur mycket de förlorar på att folk fuskåker i ett distrikt där bara bussar trafikerar och man aldrig utför några biljettkontroller. Man förlorar 800 kr per en passagerare som, hade hon/han blivit ertappad utan biljett, skulle behöva betala. Målpopulationen är alla som befinner sig på en av distriktets SL-bussar vid en viss tidpunkt. (Vi förenklar situationen med antagandet att tidpunkten har en positiv längd men att ingen av- eller påstigning sker under den och inte heller något inköp av biljett.)

SL utförde alltså under den tidpunkten en kontroll på 8 ut av sina 78 bussar som vid tillfället trafikerade distriktet, varpå följande data kunde sammanställas:

Buss	Antal passagerare	Antal fuskåkare
1	31	6
2	14	2
3	20	3
4	16	2
5	42	5
6	29	4
7	27	4
8	13	1

Skatta totalsumman av förlusten för SL på att folk fuskåker i distriktet under den aktuella tidpunkten. Räkna in även det belopp som de undersökta fuskåkare borde ha betalt (ingen av dem hade kontanter på sig vid tillfället). Ge tillhörande 95%-igt konfidensintervall. (10p)