Stockholms universitet
Statistiska institutionen

Skriftlig tentamen avseende kursen **Introduktion till statistik för statsvetare**

****

Onsdag 27 april 09:00 – 14:00, 2016

Skrivtid: 5 timmar.
Hjälpmedel: Miniräknare, samt vidhäftat formelblad.
Genomgång: 2016-05-16 klockan 15:00 till 16:00 i sal B705.

Tentamen består av fem uppgifter, totalt 50 poäng. För full poäng på en uppgift/deluppgift krävs att tydliga, fullständiga och välmotiverade lösningar samt svar inlämnas.

Lycka till! / Per Marcus

****

1. (10 p.)

På en enhet för psykisk hälsa görs under en månad 30 undersökningar om patienternas intelligens (IQ). Antalet undersökningar per vecka varierar.

a. Ange vad som här är variabel och vad som är frekvens. För variabeln, ange variabeltyp och datanivå. (2 p.)

b. När medelvärde och median beräknades för variabeln visade det sig att de erhållna värdena inte var lika. Redogör varför det kan komma sig. (2 p.)

c. Ange tre olika exempel på spridningsmått, samt diskutera vilket av spridningsmåtten som här kan vara att föredra. (3 p.)

d. Ett 95 % konfidensintervall skapades för intelligensen och intervallet [108;161] erhölls. Redogör för vad detta intervall innebär och vilka slutsatser som kan dras från detta intervall. *Observera att varje observation kan ses som oberoende från de andra.* (3 p.)

2. (10 p.)

En undersökning skall genomföras vid Stockholms Universitets Studentkår för att se hur stort behovet är av parkeringsplatser för studenter vid Stockholms universitet. Du får som uppgift att genomföra denna undersökning.

a. Definiera målpopulation och rampopulation samt redovisa för vad dessa begrepp innebär. (4 p.)

b. Beskriv vad täckningsfel är och hur det påverkar undersökningen, samt vad som skulle kunna vara täckningsfel i den undersökning som Du skall genomföra. (3 p.)

c. En av de frågor som Du ställer till respondenterna är:
"Parkerade du din bil på SU vecka 14?"
Du vet att Kalle (en av respondenterna) har svarat ja på frågan, men att han i själva verket parkerade bilen på SU vecka 13. Han har alltså antingen ljugit eller misstagit sig. Vilken typ av mätfel är detta? Beskriv felet och hur det kan påverka undersökningen. (3 p.)

3. (10 p.)

En urvalsundersökning för att undersöka inställningen hos boende i Bagarmossen till de lekplatser som finns kring centrum där.

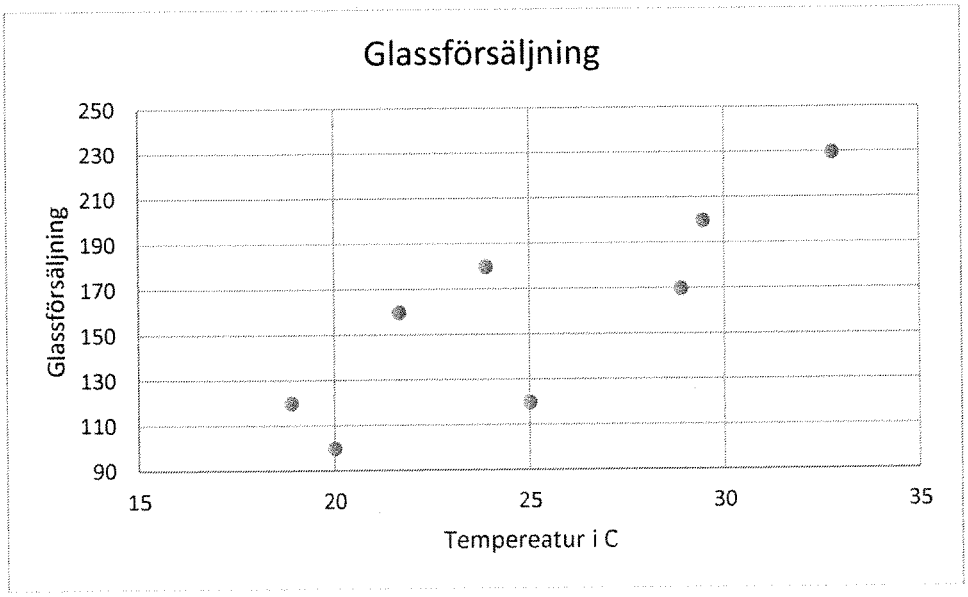
a. Redogör för vilka beståndsdelar som totalfelet innehåller. (4 p.)

b. Ge exempel för de systematiska fel som redovisas av Dig i fråga a. *Max ett exempel per systematiskt fel.* (4 p.)

c. En urvalsundersökning kan i vissa fall ge bättre precision än en totalundersökning. Redogör för när en urvalsundersökning är bättre i detta avseende. (2 p.)

4. (10 p.)

Plotten nedan beskriver temperatur utomhus och försäljning av glass. Regressionsmodellen skattades till $\hat{y} = -29{,}22 + 7{,}55x$.



a. Vilka av de två variablerna är den beroende och vilken är den oberoende enligt regressionsmodellen ovan? (2 p.)

b. Redogör för vilken av de skattade värdena som är interceptet samt vilken som beskriver lutningen. Tolka dessa två parameterskattningar. (3 p.)

c. För den givna modellen är korrelationskoefficienten beräknad till $0{,}84$ och R^2 till $0{,}70$. (5 p.)

- Tolka korrelationskoefficienten samt redogör för vad den innebär i detta fall.

- Redogör för vad R^2 betyder samt vad den mäter.

5. (10 p.)

När en urvalsundersökning görs väljs observationer olika beroende på vilken urvalsmetod som väljs.

a. Redogör kortfattat för nedanstående urvalsmetoder. Redogör även om urvalen är sannolikhetsurval eller inte, samt vad som definierar ett sannolikhetsurval. (8 p.)

- Obundet Slumpmässigt Urval

- Stratifierat urval

- Snöbollsurval

- Systematiskt urval

b. Beskriv vad en hjälpvariabel är och hur en sådan kan användas för att få bättre urval. Redogör även vad som händer med sannolikhetsurval om en "dålig" hjälpvariabel används (*det räcker med ett exempel*). (2 p.)

1. Beskrivande statistik

1.1 Medelvärde, varians, standardavvikelse

Ett statistiskt material består av n observationer

$$x_1, x_2, \ldots, x_n$$

Medelvärdet är

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x}{n} \tag{1.1.1}$$

Variansen är

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1} = \frac{\sum x^2 - (\sum x)^2 / n}{n-1} \tag{1.1.2}$$

Standardavvikelsen är

$$s = \sqrt{s^2} \tag{1.1.3}$$

När materialet redovisas i en **frekvenstabell**, där värdet x_i förekommer med frekvensen f_i, är medelvärdet och variansen

$$\bar{x} = \frac{\sum f_i x_i}{n} \tag{1.1.4}$$

$$s^2 = \frac{\sum f_i (x_i - \bar{x})^2}{n-1} = \frac{\sum f_i x_i^2 - (\sum f_i x_i)^2 / n}{n-1} \tag{1.1.5}$$

Räkneregler

Om $y = a + bx$, där a och b är konstanter, är

$$\bar{y} = a + b\bar{x} \tag{1.1.6}$$

$$s_y^2 = b^2 s_x^2 \tag{1.1.7}$$

1.2 Regression, korrelation

Regressionslinjen är $y=a+bx$.

Regressionskoefficienten är

$$b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{\sum xy - \sum x \sum y / n}{\sum x^2 - (\sum x)^2 / n} \tag{1.2.1}$$

$$a = \bar{y} - b\bar{x} \tag{1.2.2}$$

Korrelationskoefficienten är

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} = \tag{1.2.3}$$

$$= \frac{\sum xy - \sum x \sum y / n}{\sqrt{\left[\sum x^2 - (\sum x)^2 / n\right]\left[\sum y^2 - (\sum y)^2 / n\right]}} = b \frac{s_x}{s_y}$$

Residualvariansen är

$$s_e^2 = \frac{n-1}{n-2} s_y^2 (1 - r^2) = \frac{1}{n-2} \left(\sum y^2 - a \sum y - b \sum xy \right) \tag{1.2.4}$$

1.3 Prisindex

Laspeyres index är

$$\frac{\sum p_t q_0}{\sum p_0 q_0} \cdot 100 \tag{1.3.1}$$

Paasches index är

$$\frac{\sum p_t q_t}{\sum p_0 q_t} \cdot 100 \tag{1.3.2}$$

Statistiska institutionen

Stockholms universitet

Rättningsblad

Datum: 27/4-2016

Sal: Uggleviksalen

Tenta: Statistik för statsvetare

Kurs: Introduktion till statistik för statsvetare

ANONYMKOD: SFS 0003

☒ Jag godkänner att min tenta får läggas ut anonymt på hemsidan som studentsvar.

OBS! SKRIV ÄVEN PÅ BAKSIDAN AV SKRIVBLADEN

Markera besvarade uppgifter med kryss

1	2	3	4	5	6	7	8	9	Antal inl. blad
X	X	X	X	X					3
Lär.ant. 10	10	10	8	9,5					

EV

POÄNG	BETYG	Lärarens sign.
47,5p	A	

SU, STATISTIK

Skrivsal: Brunnsviksalen Anonymkod: (52) Blad nr: 1

SFS0003

1. Undersökning om patienternas intelligens (IQ)

a) Variabeln är patienternas intelligens (IQ). Denna variabel är kvantitativ, då den antar numeriska värden. Samtidigt är variabeln kontinuerlig, då denna kan anta alla värden inom ett intervall.
Datanivån som föreligger är intervallnivå, då variabeln är numerisk men det finns inte en absolut nollpunkt. Det finns inte en person med IQ = 0. I dåre går det att använda sig av summa och abstraktion för att göra jämförelser, men det går inte att bilda kvot.
→ Variabeln är kontinuerlig då den kan anta oändligt många värden inom ett intervall.
Frekvensen är antal undersökta patienter. 2p

b) Om erhållna värden av medelvärdet och medianen inte är lika, innebär det att det finns extrema observationer som har påverkat medelvärdet. Om medelvärdet är högre än medianen, finns det en eller fler högextrema observationer som påverkade medelvärdet. Om medelvärdet är lägre än medianen, finns det en/fler lågextrema observationer som påverkade medelvärdet.
När det inträffar, betyder det att det finns en sned fördelning, och då kan medelvärdet bli missvisande. I dessa fall är det fördelaktigt att använda sig av medianen, ett lägesmått som visar på det mittersta värdet efter en rangordning, och inte påverkas av extrema observationer. 2p

c) Spridningsmått: Tre exempel på spridningsmått är variationsvidd, kvartilavstånd och standard avvikelsen.
Variationsvidden står för skillnaden mellan det största och det minsta värdet. Detta spridningsmått skulle kunna användas i undersökningen för att ta reda på vad skillnaden är mellan den patient som har högst IQ, och den patient som har lägst IQ. Dock får man inte mer information om spridningen än så.
Standard avvikelsen står för hur variationerna i genomsnitt avviker från det aritmetiska medelvärdet. Detta mått brukar vara att föredra, då den bidrar med mer information om spridningen, och kan användas för att analysera den statistiska signifikansen och pröva hypoteser.
Dock anser jag att kvartilavståndet är det mest lämpliga spridningsmåttet vid den här undersökningen. Detta därför att det finns en sned fördelning av observationerna och standard avvikelsen, lika så som medelvärdet, kan bli missvisande.
Kvartilavståndet är ett spridningsmått som beskriver inom vilket intervall 50% av de mest centrala observationerna befinner sig i. 3p

1. d) 95% Konfidensintervall för intelligensen [108; 161]

Konfidensintervall är ett specifikt intervall där det sanna värdet av medelvärdet befinner sig i. Det erhållna konfidensintervallet i undersökningen innebär att det sanna värdet av medelvärdet befinner sig mellan 108 och 161, d.vs $108 \geq \bar{x} \leq 161$. Dessutom finns 95% säkerhet på att det sanna värdet av medelvärdet befinner sig i just detta intervall.

3p 10p

2. Undersökning för att se hur stort behovet är av parkeringsplatser för studenter vid Stockholms universitet.

a) Målpopulation och rampopulation:

Målpopulationen är de individer/element som man är intresserad av att undersöka. Målpopulationen i studentkårens undersökning är studenter vid Stockholms Universitet som använder sig av parkeringsplatser på universitetet.

Rampopulation är den ramen/registern som man använder sig av för att kunna täcka sin målpopulation. En passande ram för studentkårens undersökning skulle kunna vara en register av alla studenter som har bil/motorcykel. Alternativ en ram där man kan se vilka studenter som har Körkort.

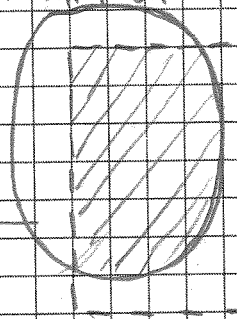
Eftersom det är sannolikt att denna information inte finns, kan man använda sig av register över alla studenter (exempelvis Ladok), välja ut ett urval, och ha med filtreringsfrågor i frågeformuläret när dessa delas ut. På så sätt, kan man identifiera vilka som har bil, vilka som använder sig av parkeringsplatserna och sortera bort eventuell övertäckning. 4p

b) Täckningsfel är ett systematiskt fel som uppstår när ramen inte överensstämmer med målpopulationen. Det finns två typer av täckningsfel: övertäckning och undertäckning.

Övertäckning skulle kunna uppstå i studentkårens undersökning som det finns studenter som har bil men inte använder sig av parkeringsplatser på universitetet. Det är alltså de individer som inte är i vår målpopulation men som ändå dyker upp i rampopulationen. Det går att motarbeta detta fel, exempelvis genom att ställa filtreringsfrågor i frågeformuläret och sortera bort övertäckningen.

Undertäckning uppstår när individer i målpopulationen inte täcks av rampopulationen, d.v.s. de som inte är med i registret. Ett exempel skulle kunna vara nya studenter som inte hunnits registrera/uppdaterats i Ladok, och som har bil och använder sig av parkeringsplatserna. Undertäckningsfel är mer problematiskt då det är information som vi skulle behöva, och inte har. Det blir kostsamt att täcka undertäckningsfel.

3p



SU, STATISTIK

Skrivsal: Brunnsviksalen Anonymkod: SFS0003 Blad nr: 2

2. c) Teleskoppeffekt: När en respondent placerar en händelse felaktigt i tiden, uppstår ett mätfel som kallas för teleskoppeffekt. Kalle har angett i enkäten att han parkerade sin bil på Su vecka 14, men egentligen var det vecka 13. Kalle har antingen ljugit eller misstagit sig och nu har vi fått ett mätfel i vår undersökning. → Det kan vara avsiktligt eller av misstag.

Teleskoppeffekt kan resultera i en snedvriden bild av verkligheten om den förekommer i flera observationer. För att motarbeta detta kan man replikera frågan flera gånger vid olika tidspunkter för att säkerställa att respondenterna har angett ett svar som stämmer med verkligheten (1p) (3p)

3. Urvalsundersökning för att undersöka inställningen hos boende i Bagarmossen till de lekplatser som finns kring Centrum där.

a) Totalfelet

$$MSE = (\text{Var}) + [\text{Bias}(\bar{O})^2]$$

Var → Urvalsfel

Bias → Systematiska fel: Mätfel, bortfallsfel, täckningsfel och Bearbetningsfel.

Det totala felet är summan av urvalsfel med systematiska fel / Non sampling errors (Mätfel, bortfallsfel, täckningsfel och bearbetningsfel.)

(1p.)

- Urvalsfel: Avvikelsen mellan det erhållna värdet i det slumpmässiga urvalet och värdet man hade fått i en totalundersökning. Man antar att urvalsfelet i en totalundersökning är lika med noll.

Systematiska fel / Non sampling errors:

- Mätfel: Uppstår när man inte mäter det man vill mäta, alltså fel i mätningen.
- Bortfallsfel: Uppstår när man går miste om information från respondenterna. Detta kan vara objektbortfall, när en respondent inte svarar; eller Partiellt bortfall, när en fråga inte besvaras / missuppfattas.
- Täckningsfel: När ramen inte överensstämmer med målpopulationen. Det kan vara undertäckningsfel eller övertäckningsfel.
- Bearbetningsfel: När fel uppstår i bearbetningen av materialet. Det kan vara fel i kodningen eller att man slår in felaktiga siffror av misstag.

b) Exempel för de systematiska felen i undersökningen i Bagarmossen

- Mätfel: Prestigebias. En av frågorna i enkäten handlade om att ange om man någon gång hade inhandlat narkotika på någon lekplats i Bagarmossen. Andreas valde alternativet "Aldrig" som sitt svar, även om han ett par gånger hade gått till parken för att köpa narkotika. Andreas angav ett svar som han ansåg skulle vara mer "socialt accepterat" men som inte överensstämmer med verkligheten. Det uppstod ett fel i mätningen.

Fortsättning 3b.

- Bortfallsfel: Ett exempel på partiellt bortfall är följande. Peter håller på att svara en enkät om inställningen till lekplatser i Bogärmossen. En av frågorna handlade om hur ofta man besöker lekplatserna i området på en vecka. Peter svarar "10 gånger i veckan" då han måste gå igenom en lekplats varje dag för att ta sig till/från jobbet. Peter ansåg att "ta sig igenom parken" var svaret på "frekvensen man besöker lekplatsen". De som genomförde undersökningen förväntade sig att respondenterna skulle använde närmaste parken för nöjesskull. Dock specifieras det ej i frågan och Peter missuppfattar den. Då vi det ett partiellt bortfallsfel.

1p

- Täckningsfel: I urvalsundersökningen hos boende i Bogärmossen, användes folkbokföringen som rampopulation. Dock uppstod det en undertäckning på individer som nyligen hade flyttat in i området och som inte hade hunnit folkbokföra sig. Det visade sig att det var just dessa individer som mest besökte lekplatserna i området.

- Bearbetningsfel: Sara och Niklas fick tillbaka en del av enkäterna som de hade skickat ut för att samla in datan om inställningen till lekplatser i Bogärmossen. De skulle börja registrera all information i ett dataprogram, som skulle vara till hjälp i bearbetningen av materialet. Dock råkade Sara knappa in felaktiga siffror på ett par ställen, hon var trött och Niklas var inte uppmärksamm nog för att märka detta.

c) Totalundersökning Kontra Urvalsundersökning

En totalundersökning kännetecknas av att inget urvalsfel uppstår. Däremot förekommer systematiska fel som måste bearbetas / skattas för att minimera felet och garantera hög kvalitet i undersökningen.

En urvalsundersökning kan i vissa fall ge bättre precision än en totalundersökning, då det är mindre data att bearbeta och skattningarna kan bli mer precisa. Dessutom brukar totalundersökningar vara dyra att genomföra. I en urvalsundersökning kan man använda sina ekonomiska resurser på att göra bättre skattningar och minimera felet i större utsträckning. Detta strategi kan vara mer gynnsamt än att investera på datainsamlingen och inte ha råd med bra skattningar och motarbetande av exempelvis bortfall.

Man bör alltid tänka på vilken typ av undersökning som är att föredra i linje med hur mycket pengar man disponerar för sin undersökning. Med andra ord, ska man ha det minsta felet för minsta antal kronor.

2p

10p

4. Regressionsmodellen: $\hat{y} = -29,22 + 7,55x$

a) Enligt regressionsmodellen är den beroende variabeln (y) glassförsäljning och den oberoende variabeln (x) är temperatur i C. (2p)

b) Tolkning av parameterskattningar i regressionsmodellen:
Interceptet (a) är $-29,22$. Denna koefficient är startpunkten som utgår från y axeln i regressionsmodellen. Det innebär att glassförsäljning är $-29,22$ utan påverkan av temperaturen i C.
Lutningen (b) är 7,55, den anger hur regressionslinjen lutar sig i modellen. Det innebär att för varje C grad som temperaturen ökar med, så ökar glassförsäljningen med 7,55. Tex om det blir 40 C så ökar glassförsäljning med $30 \geq C$ $7,55 \cdot 40 C$ i snitt. (2p)

SU, STATISTIK

Skrivsal: Brunnsviksalen Anonymkod: SFS0003 Blad nr: 3

4. c) Korrelationskoefficienten = 0,84
$R^2 = 0,70$.

- Korrelation betyder graden av linjärt samband. Korelationskoefficient står för hur starkt sambandet mellan variablerna är. Dessutom visar denna koefficient på om korrelationen är positiv eller negativ. Det brukar vara ett värde mellan −1 och 1, dvs. $-1 \geq r \leq 1$, där $r = -1$ tolkas som ett perfekt negativt samband; och $r = 1$ tolkas som ett perfekt positivt samband.
I vår undersökning, tyder korrelationskoefficienten 0,84 på att det finns ett ganska starkt samband mellan glassförsäljning och temperatur i C. Dessutom visar denna korrelationskoefficient på att sambandet är positivt. När temperaturen ökar, så ökar glassförsäljning. Ett samband som även är synligt i skattningsplottet. Ej kausalt.

- R^2 är determinationskoefficient. Denna koefficient mäter i vilken procent variationer i den beroende variabeln (y) förklaras av variationer i den oberoende variabeln (x). Man kan beräkna determinationskoefficienten genom att höja korrelationskoefficienten upp till två.
I det här fallet, visar vår determinationskoefficient 0,70 att 70% av variationerna i glassförsäljning förklaras av variationerna i temperaturen. Denna koefficient tyder på ett starkt samband mellan variablerna.

4p (8p)

5. Urvalsundersökningar och urvalsmetoder

a) Definition av sannolikhetsurval: Ett sannolikhetsurval är ett sorts urval där slumpen avgör vilka individer/element som kommer in i urvalet. I ett sannolikhetsurval har alla individer/element en sannolikhet (som är större än noll) att vara med i urvalet. Sannolikhetsurval eller slumpmässiga urval, som det också kallas, kan vara obundet eller riktat.

Urvalsmetoder:

- Obundet slumpmässigt urval: OSU är ett sannolikhetsurval som är obundet eller ej riktat. Man låter slumpen avgöra helt vilka individer/element som ska vara med i urvalet. Det är det enklaste sannolikhetsurvalet att genomföra samt det billigaste. Ett problem som kan uppstå är att eftersom slumpen avgör, kan det bli en sned fördelning i urvalet, där det saknas representation av minoriteter, exempelvis. (2p)

- Stratifierat urval: Det är ett sannolikhetsurval som går ut på att man delar sin population i olika strata (grupper), där spridningen mellan individerna inom varje stratum ska vara så liten som möjligt. Däremot ska spridningen mellan de olika stratа vara så stor som möjligt, så att man ska kunna jämföra dem och dra slutsatser.
Viktigt att tänka på är att indelningen av populationen ska ske med hjälp av en hjälpvariabel. Denna måste ha ett samband med undersökningsvariabeln. Man genomför ett vanligt OSU i varje stratum efter att indelningen har skett.

Fortsättning 5a.

- Snöbollsurval: Det är ett icke-sannolikhetsurval där man får reda på respondenter/informanter tack vare andra informanters/respondenters förslag. Det vill säga att de första informanterna/respondenterna leder oss till flera, och de i sin tur föreslår flera, osv.
  Denna urvalsmetod används framförallt vid undersökningar som kan uppfattas som känsliga och där en ram med information om individer är svår att hitta. Ett exempel skulle kunna vara en undersökning om kvinnor som har blivit våldtagna som barn. (2p)

- Systematiskt urval: Det är ett sannolikhetsurval som går ut på att man rangordnar individer/element i sin ram utifrån en hjälpvariabel. Denna hjälpvariabel ska ha ett samband med undersökningsvariabeln.
  Viktigt att tänka på vad det gäller rangordningen är att det ska finnas ekvidistans mellan alla individer/element i ramen efter att man rangordnat dem. Därmed väljer man en siffra från 1-10 och systematiserar sitt urval efter detta val. Tex om man slumpmässigt väljer individ/element 7, då får man ta individer/element 7, 17, 27, 37, 47, 57... tills man fyllt sin urvalsfraktion. (1,5 p)

b) Hjälpvariabel: En hjälpvariabel är en variabel som används för att dela in eller rangordna individer/element i ett sannolikhetsurval, exempelvis vid stratifierat urval och systematiskt urval.
  Hjälpvariabeln som man väljer ska alltid ha ett samband med den variabeln man undersöker, för att kunna säkerställa att man har en passande fördelning i sitt urval. Ett exempel är om man ska undersöka inkomstskillnader i en viss population. Då kan man välja en hjälpvariabel där man indelar/rangordnar sin population från låginkomsttagare till höginkomsttagare.
  Om man väljer en "dålig hjälpvariabel", det vill säga en variabel som inte har att göra med undersökningsvariabeln, då blir urvalet ett vanligt OSU, och då har man slösat tid och resurser i stratifiering/systematisering av sin population, utan framgång.
  För att fortsätta med exemplet om inkomstskillnader kan man anta att man väljer "antal hundar" som hjälpvariabel. Eftersom denna hjälpvariabel inte har något samband med undersökningsvariabeln, så blir stratifieringen/systematiseringen misslyckad och urvalet förblir ett vanligt OSU.

(2p)

9,5

Statistiska institutionen

Stockholms universitet

Rättningsblad

Datum: 27/4-2016

Sal: Uggleviksalen

Tenta: Statistik för statsvetare

Kurs: Introduktion till statistik för statsvetare

ANONYMKOD: SFS-0005

☒ Jag godkänner att min tenta får läggas ut anonymt på hemsidan som studentsvar.

OBS! SKRIV ÄVEN PÅ BAKSIDAN AV SKRIVBLADEN

Markera besvarade uppgifter med kryss

1	2	3	4	5	6	7	8	9	Antal inl. blad
X	X	X	X	X					7
Lär.ant. 10	8	10	9,5	9					

POÄNG	BETYG	Lärarens sign.
46,5	A	

SU, STATISTIK

Skrivsal: Vegelviksalen Anonymkod: SFS-0005 Blad nr: 1

1)

a) Variabeln är patienternas intelligens (IQ), då det är det som vi vill undersöka. Variabeln är kvantitativ då den antar numeriska värden. Vidare ligger den på intervalldatanivå då det går att rangordna & peka ut meningsfulla skillnader mellan variabelvärdena. Den har även en godtycklig nollpunkt då man inte kan ha ett IQ under 0. Den är godtycklig då IQ är ett visst mått på intelligens som vi människor uppfunnit, ungefär som vår tideräkning som också har en godtycklig nollpunkt på år 0.
Frekvensen är antalet undersökningar som görs per vecka. (2p)

b) Medelvärdet är det värde man får om man adderar alla variablers värden & dividerar med antalet observationer. Medianen är det mittersta värdet i ett rangordnat material. Medelvärdet är känsligt för extremvärden vilket gör att medelvärdet kan bli något skevt om det i detta fall förekommer extremvärden i detta fall. Medianen är

å andra sidan inte känsligt för extremvärden & påverkas således inte av dem i lika stor utsträckning som medelvärdet. Att de centrala värdena, alltså medelvärdet & medianen, inte var lika har således att göra med att de delar sett till olika aspekter av en genomsnitt & därför att medelvärdet kan bli påverkats av extremer (2p.

c) Variationsbredd = Skillnaden mellan det största & minsta värdet

Standardavvikelse = Hur värdena i genomsnitt avviker från medelvärdet

Kvartilavstånd = Skillnaden/avståndet mellan den första & tredje kvartilen; $Q_1 - Q_3$

Variationsbredden säger endast något om det största & minsta värdet vilket blir något intetsägande i detta fall då det utelämnar nästerande värden som också är av intresse. Därav utesluts variationsbredden.

I & med det som jag diskuterade i föregående fråga (om hur medelvärdet är känsligt för extremvärden) så kan standardavvikelsen i detta fall ge en något missvisande bild av spridningen eftersom den tänger i hop med medelvärdet. Kvartilavståndet är att föredra i detta fall då det mätet tänger i hop med medianen (som ej är känsligt för extremer) & kan således ge en bra bild av spridningen av de mittersta observationerna. (3p

d) Ett konfidensintervall är en skattning av osäkerheten kring de skattade parametervärdena som tagits fram genom ett stickprov. Det visar alltså med hur stor sannolikhet det sanna värdet återfinns inom intervallet. Ett 95 % konfidensintervall innebär att vi med 95 % sannolikhet kan säga att det sanna parametervärdet återfinns inom intervallet. I detta intervall [108; 161] så kan vi alltså med 95 % säkerhet anta att det i 95 % av alla stickprov, så kommer det sanna parametervärdet ligga mellan 108 & 161.
I detta fall kan man alltså säga att ex. det sanna populationsmedelvärdet för intelligens hos patienterna på det skattade populations-medelvärdet för intelligens kommer inom intervallet [108; 161] med 95 % säkerhet att överensstämma med verkligheten. (10p)

(2) a) Målpopulation: Det vi i slutet vill undersöka. I detta fall kan målpopulationen utgöras av samtliga studenter som studerar vid Stockholms Universitet. Rampopulation är det vi kan undersöka, alltså det vi kan nå genom något slags register eller förteckning. Ramen skulle i detta fall kunna vara registrerade studenter vid Stockholms Universitet. 3p

b) Täckningsfel uppstår när målpopulationen & ramen inte överensstämmer med varandra. Exempel är undertäckning som innebär att individer som tillhör målpopulationen inte finns med i registret & således har 0 inklusionssannolikhet att komma med i urvalet. Övertäckning är även det ett täckningsfel, och uppstår när individer finns med i ramen men inte i målpopulationen vilket gör att man får med folk i sin undersökning som egentligen inte är av intresse. Täckningsfel som kan uppstå i den första undersökningen om studenter vid Stockholms Universitet & bärandet av parkeringsplatser är av både undertäckning & övertäckning.

Undertäckning kan uppstå om ex. studenter som studerat vid SU glömt att registrera sig på kursen, eller om det av någon teknisk fel blivit fel i registreringen så att man inte syns vara dessa där & således inte kommer med i urvalet. Även övertäckning kan uppstå om det även har kunnat blivit något fel i registreringen där ex. vissa personer har registrerat sig på en kurs som studenter fastän de sedan knappt var & inte studerar längre & att registreringen inte blivit uppdaterad. Detta skulle alltså innebära övertäckning då de kommer med i urvalet även inte tillhör målpopulationen då de inte studerar på Stockholms Universitet.

För att sammanfatta frågan så bör stjärna felet med undertäckning & övertäckning att undersökningen bör några missvisande då dem inte helt & hållet speglar den verkliga populationen, målpopulationen som är vår undersökning. (3p)

c) Måtten har att göra med såväl mät instrument, respondenten, intervjuare & mätmetod. I detta fall kan vi anta felaktiga svar beroende på att man vill komma med i undersökningen då man tycker att det är "viktigt" att hans svar kommer med i undersökningen, detta kan då ha att göra med respondenten i fråga. Att respondenten vill komma med i undersökningen då denne anser sin åsikt/svar som intressant

FORTS. Fråga 2.

Tillförlitligheten för undersökningsresultaten. Det kan även ha gjorts mätinstrumentfel & mätmetodfel. Om jag ex. varit otydlig i ett eventuellt frågeformulär så kanske det var då gjorda med det. Eller om det är en intervju så kanske jag Omedvetet påverkar kanske att svara ett visst svar vilket kallas intervjuareffekten. Mätfelet påverkar också undersökningen till det sämre då svaret inte är sammanställbart vilket bidrar till sneddningen resultat i undersökningen. (8p)

(3) a) Totalfelet = Urvalsfel + täckningsfel + bearbetningsfel + bortfallsfel + mätfel.

(Urvalsfel = De fel som uppstår i & med det att enbart ett stickprov från populationen & ändå vill dra slutsatser om det.) (1p)

b) Bortfallsfel ⇒ Det fel som uppstår i undersökningen om vissa svar / individer uteblir. Partiellt bortfall är när en viss variabel uteblir, dvs att det ex. kan bero på att en respondent inte besvarat en viss fråga. Individbortfall uppstår ex. om jag skickar ut postenkäter och vissa individer inte svarar på enkäten.

⇒

Täckningsgrad uppstår då målpopulationen & rampopulationen inte överensstämmer med varandra. Kan ex. vara om jag ska undersöka 16-24 åringar i en viss kommun & ska utgå från Folkbokföringsregistret som ram, & så kommer vissa vara skyddade i adressen & inte finns med i registret. Detta leder till ett undertäckningsfel.

Bortfallsgrad → uppstår vid ex. registreringen om datum eller kodningen av svar osv. Det kan ex. vara så att jag som undersökare ska skicka & föra registreringen av svaren till en dator, så finns det stor sannolikhet att något svar kommer registreras felaktigt, detta blir ett bearbetningsfel pga. av den undersökande faktorn (man är ingen dator).

Mätfel → Har å göra med såväl respondenten, intervjuaren, mätinstrumenta & mätmetoden. Är dessa störningar mellan varandra & enskilda värde. Mätfel kan ex. uppstå om respondenten inte förstår i vad dagar & inte besvarar ett ex. frågeformulär sammanhängande. Säg att det handlar om alkoholkonsumtion & respondenten uppger i att den konsumerar mindre alkohol än vad denne faktiskt egentligen gör. Detta leder då till ett mätfel som kallas prestige-bias.

(7p)

SU, STATISTIK

Skrivsal: Vegetarsalen Anonymkod: SFS-0005 Blad nr: 5

Forts. fråga 3

3 c) Det är svårare att genomföra en undersökning om målpopulationen är extremt stor. Detta då det exempelvis skulle vara väldigt svårt att undersöka ex. alla i hela Sverige. En undersökning för alla i hela Sverige skulle dels innebära extrema kostnader, extremt mycket tid, dels skulle även med stor sannolikhet uppstå stora bortfallsfel då det rent ut sagt skulle vara näst intill omöjligt att få tag i alla plus att många inte ens skulle svara förmodligen. Många systematiska fel skulle alltså uppstå. Då kan det vara bättre med en undersökning där man kan, när man undersöker färre personer, lägga mer tid & pengar på att få svar från dessa samt att bortfallet då kan minskas då man eventuellt skulle kunna ha råd att erbjuda de som svarar en gåva eller liknande. Det detta talar för att en undersökning är att föredra i detta fall då det ger mer precision. (2p) (2p)

4) a) beroende variabeln = glassförsäljning
oberoende variabeln = temperatur i C minus.

(2p)

b) $\hat{y} = -29,22 + 7,55x$. Interceptet är -29,22 som står för a. Detta är var regressionslinjen skär y-axeln & är även det värdet som y i genomsnitt antar när $x = 0$. Det går alltså att säga att det är linjens startpunkt. Lutningen är 7,55 & står för b. Detta är även kallat riktningskoefficienten / regressionskoefficienten. Denna parameterskattning visar hur mycket y ändras när X ökar med en enhet. I detta fall har vi en positiv lutning (7,55) som innebär att när X ökar med ett enhet så ökar även y. I detta fall (2,5 p) så kan man se att när temperaturen ökar (x) så ökar även glassförsäljningen (y).

c) Korrelationskoefficienten anger graden av linjärt samband, alltså huruvida två variabler samvarierar eller inte. Perfekt positivt samband är +1 & perfekt negativt samband är -1. I detta fall är korrelationskoefficienten 0,84 vilket tyder på en relativt stark korrelation mellan glassförsäljning & temperatur i C utomhus. I & med att det är en positiv korrelation så innebär det att när värdet på X-axeln ökar/sjunker så ökar/sjunker även värdet på Y-axeln. När temperaturen stiger så ökar även glassförsäljningen. Nämnvärt är dock att detta inte säger något om orsak & verkan utan endast om korrelation.

Fråga 4 forts. c)

c) R^2 står för determinationskoefficienten. Denna anger hur stor del av variationen i y som kan förklaras av variationen i x. I detta fall uppgick R^2 till 0,70 vilket innebär att 70 procent av variationen i grustransport (y) kan förklaras av variationen i temperatur i Celsius (x). 5 p. 9,5 p

(5)

a) Obundet slumpmässigt urval är ett sannolikhetsurval då alla individer i populationen har en inklusionssannolikhet att komma med i urvalet. Obundet slumpmässigt urval innebär att alla individer i populationen har samma inklusionssannolikhet att komma med i urvalet. Detta går att likna vid hur ett lotteri går till. Detta urval är alltså helt obundet & oriktad & blir därefter i många fall som en miniatyr av populationen. 2 p

Stratifierat urval är även det ett sannolikhetsurval då alla individer i populationen har en inklusions-sannolikhet att komma med i urvalet. I detta urval så delar man in sin population i subpopulationer, s.k. strata, där varje stratum representerar en del av populationen. Man stratifierar alltså

Populationen eller någon lämplig stratifieringsvariabel. Sedan drar man ett OSU eller ev. ett proportionerligt urval från varje stratum. (I ett stratifierat urval så eftersträvar man så liten spridning som möjligt inom stratumet men stor spridning mellan respektive stratum) När man sedan dragit ex. ett OSU från varje stratum så får varje stratum sin egen spridning. (2p)

Snöbollsurval är inte ett sannolikhetsurvalsmetod då man medvetet inte har en inklusionssannolikhet att komma med i urvalet. Ett Snöbollsurval går till på så sätt att man först kontaktar en person som anses vara relevant för urvalet & sedan låter denna person föra en till nästa person osv, osv.. Detta urval får alltså således ej anses riktat. (2p)

Systematiskt urval är ett sannolikhetsurval. Detta urval går till så att man först bestämmer sig för hur stor andel av populationen man vill undersöka. Säg 10%. Då väljer man slumpmässigt ut en siffra mellan 1-10. Säg att det blir 4. Då låter vi individer med nummer 14, 24, 34 osv automatiskt ingå i urvalet. Här måste dock individerna vara numrerade i någon förteckning / register för att detta vara möjligt att genomföra denna typ av urval. (2p)

FORTS. fråga 5

5 b) En hjälpvariabel är en variabel som hjälper mig att sortera mina observationer. Om jag ex. genomför ett stratifierat urval (sannolikhetsurval) så är det av stor vikt att ha en bra hjälpvariabel. Säg att jag ska göra ett urval av olika företag, då kan det vara relevant att stratifiera utifrån storleken på dessa företag. Därför väljer jag min hjälpvariabel vara storlek, ex små företag, medelstora & stora företag. Hjälpvariabeln hjälper mig således att sortera för att få mer precision i mina skattningar. Om jag i detta fall inte sorterat mina företag utefter hjälpvariabeln "storlek" så hade det kunnat bli ett missvisande resultat då slumpen skulle kunna orsaka så att ex. fåtterviangar stora företag & jättefå små företag hamnar med i undersökningen. Detta skulle leda till ett icke representativt resultat. Har man en dålig hjälpvariabel att sortera efter så kommer urvalet bli icke representativt & snedvridet. (E 7p