Stockholms universitet
Statistiska institutionen

Skriftlig tentamen avseende kursen **Introduktion till statistik för statsvetare** (ST131G) 2013-06-05

****

Skrivtid: 5 timmar.
Hjälpmedel: Miniräknare, samt vidhäftat formelblad.
Genomgång: 2013-06-19 klockan 14.00 i sal B307. Obs ! Datum, tid och lokal !!!

****

Tentamen består av fem uppgifter vilka kan ge maximalt fyra poäng vardera, totalt tjugo poäng. Då en uppgift i sin tur består av två eller flera deluppgifter värderas dessa lika. För full poäng på en uppgift/deluppgift krävs att tydliga, fullständiga och välmotiverade lösningar samt svar inlämnas. Lycka till !!! / Peter Claësson

****

1.

Ett urval av studenter i ett visst studentbostadsområde i närheten av ett svenskt lärosäte tillfrågades om hur länge var och en hade köat innan de fick tillgång till sin lägenhet.

Kötid (år)	0	1 - 2	3 - 4
Antal studenter	12	20	8

a) Ange vad som ovan är variabel, vilken variabeltyp och vilken datanivå som föreligger samt vad som är frekvens. Motivera.
b) Illustrera fördelningen ovan i ett lämpligt diagram samt motivera valet av diagram.
c) Såväl medelvärde som median beräknades för fördelningen ovan, varvid man erhöll 1,95 respektive 1,80. Hur ska dessa två värden tolkas?
d) Bland de utvalda studenterna visade det sig att 40% var män. Illustrera även denna fördelning i ett lämpligt diagram, samt motivera valet av diagram.

2.

De faktiska utgifterna av ett särskilt slag vid en stor statlig myndighet har sammanställts för ett antal år. I tabellen nedan ges även konsumentprisindex (KPI) för respektive år.

År	1990	2000	2005	2010
Utgifter (Mkr)	48	40	32	12
KPI	208	260	280	303

a) Beräkna en indexserie i löpande priser (basår 2000) avseende de aktuella utgifterna.
b) Beräkna därefter en motsvarande indexserie i fasta priser (2000 års prisnivå).
c) Illustrera även de två ovan beräknade indexserierna i ett lämpligt diagram. Motivera.
d) Hur stor är den procentuella utgiftsförändringen mellan år 1990 och år 2010?

v g v

3.

För ett antal år sedan gjordes en studie avseende priset på fritidshus. Inom ett visst avgränsat geografiskt område valdes slumpmässigt tio fritidshus, samtliga sålda under året ifråga. Förutom att försäljningspriset studerades för vart och ett av de utvalda fritidshusen noterades även avståndet till kusten. Avsikten var att för aktuella variabler genomföra en regressions- och korrelationsanalys, där man angav försäljningspriset i tusentals kronor och avståndet till kusten i kilometer.

Då en regressionslinje anpassades erhölls följande ekvation: $\hat{y} = 1234{,}0 - 56{,}0x$ (avrundade värden).

a) Efter nödvändiga beräkningar fick man att korrelationskoefficienten blev – 0,7. Hur tolkas detta värde?
b) Rita in linjen i ett lämpligt diagram samt skissera hur sambandet mellan variablerna bör ha sett ut.
c) Tolka de erhållna koefficientvärdena i linjens ekvation, uttryckta i termer av de aktuella variablerna.
d) Ge en prediktion med hjälp av linjen för ett fritidshus på en mils avstånd från kusten, samt kommentera.

4.

Antag att man inom en mycket stor svensk kommun önskar genomföra en omfattande urvalsundersökning bland samtliga invånare (18 år eller äldre), med syfte att ta reda på inställningen till ett planerat vägprojekt.

Precisera vad som här kan anses vara mål- respektive rampopulation. Redogör därefter för, samt ge även exempel på, nedanstående åtta statistiska begrepp/metoder - utifrån det givna antagandet ovan.

a) Undertäckningsfel,
b) Övertäckningsfel,
c) Mätfel,
d) Bortfallsfel,
e) Systematiskt urval,
f) Stratifierat urval,
g) Kvoturval,
h) Urvalsfraktion.

5.

Redogör för de åtta statistiska begreppen/metoderna nedan, samt ge realistiska exempel på respektive.

a) Nominaldata,
b) Diskret variabel,
c) Frekvenspolygon,
d) Tratt-teknik,
e) Nonsenskorrelation,
f) Fastbasindex,
g) Bias,
h) Konfidensintervall.

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: 5/6-2013

Sal: Brunnsvikssalen

Tenta: Statistik för statsvetare

Kurs: Introduktion till statistik för statsvetare

ANONYMKOD: [redacted]

☑ 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					6
Lär.ant. 3.5	2.75	3	2.25	1					

POÄNG	BETYG	Lärarens sign.
12.5	D	

SU, STATISTIK

Skrivsal: Brunnsviksalen Anonymkod [redacted] Blad nr: 1

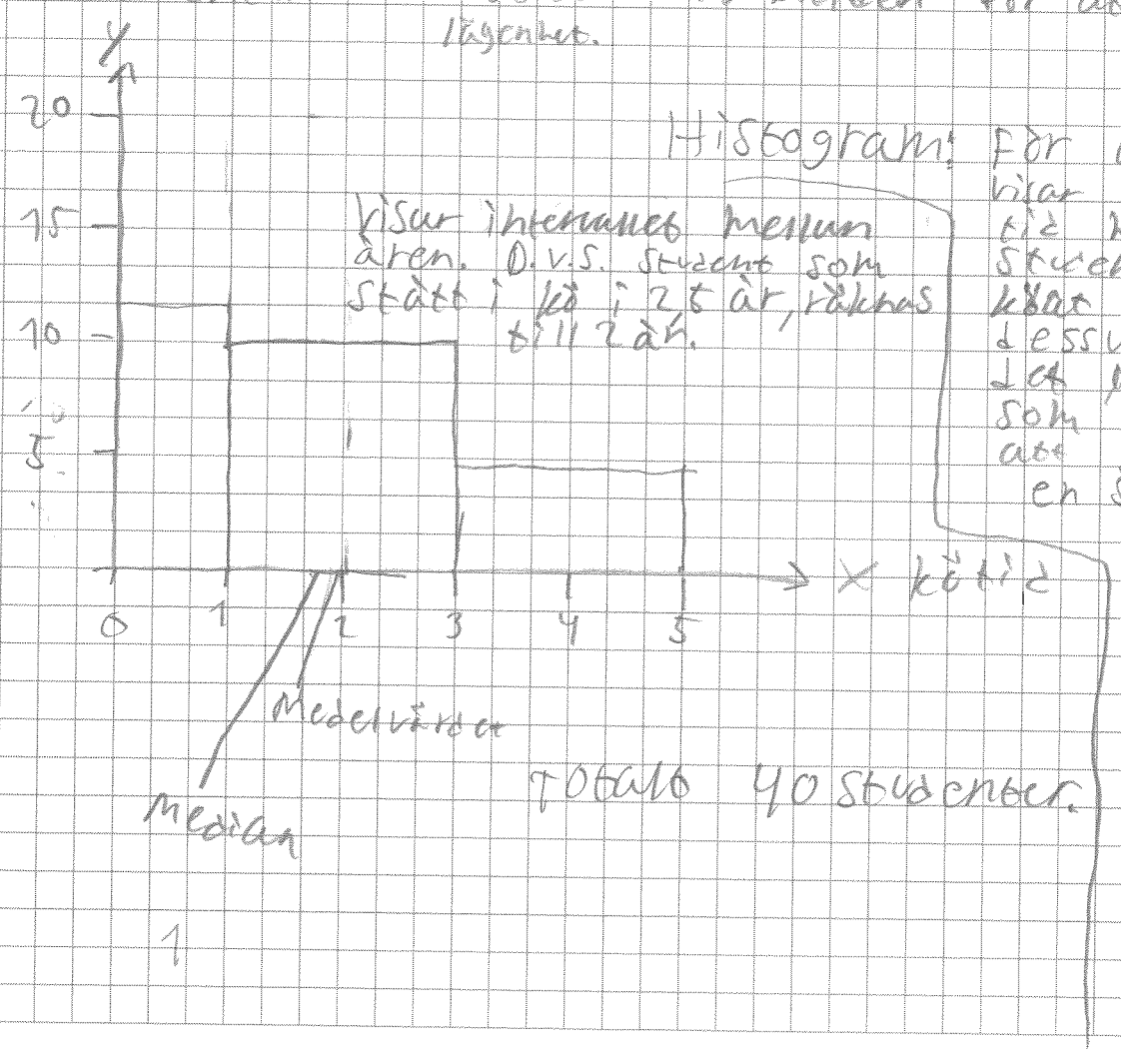
1.

a)

Kötid (år)	0	1–2	3–4
Antal studenter	12	20	8

a) Kötiden är variabeln. Den är en kvantitativ diskret variabel. Datanivån är kvotdata eftersom kötiden har en absolut nollpunkt. Kötiden är bakgrundsvariabeln till att få en lägenhet. Studenterna är alltså beroende av kötiden. 0.75

b) Antal studenter är frekvensen eftersom studenterna är beroende av kötiden för att få lägenhet.



Histogram för att lättare visar både över tid hur länge studenterna har köat och essvoom visar det på de åren som är lättast att mista d.v.s. en student som står i kö i 2,5 år fortfarande inte har köat i 3 år.

Totalt 40 studenter.

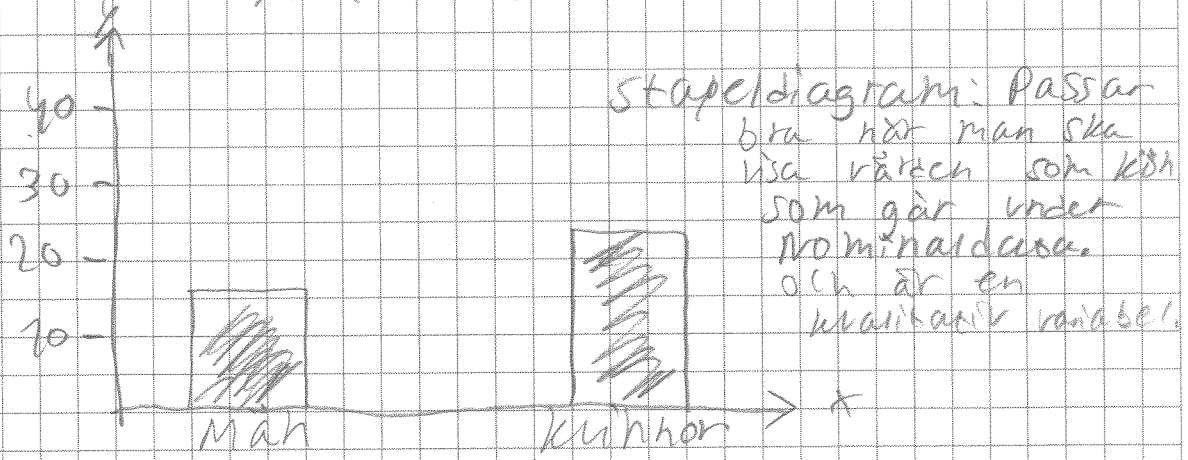
c Medelvärdet är tolkat som medelkvinnan. d.v.s. slumpstudenten har köpt i 1.95 år. Det är uträknat genom att den totala summan har dividerats med antalet observationer.

Medianen är det mittersta värdet i observationen. D.v.s. En sammanställning av alla studenter där det mittersta värdet väljs ut av de totalt 40 studenterna.

0.75

d IAV de 40 studenterna var 40% män

40 · 0,4 = 16 16 män och 24 kvinnor.



stapeldiagram: Passar bra när man ska visa värden som kön som går under nominaldata. och är en kvalitativ variabel.

1

SU, STATISTIK

Skrivsal: Brunnsvikssalen Anonymkod [redacted] Blad nr: 2

Bas år 1990 i löpande

2. År 1990 2000 2005 2010

utgifter 48 40 32 12

KPI 208 260 280 303

a) Bas som tidigare enligt

År	löpande	Deflator löpande
1990	83	208/260 · 0,83 = 0,66 = 66
2000	100	260/260 · 1,0 = 1,0 = 100
2005	125	280/260 · 1,25 = 1,35 = 135
2010	300	303/260 · 3 = 3,5 = 350

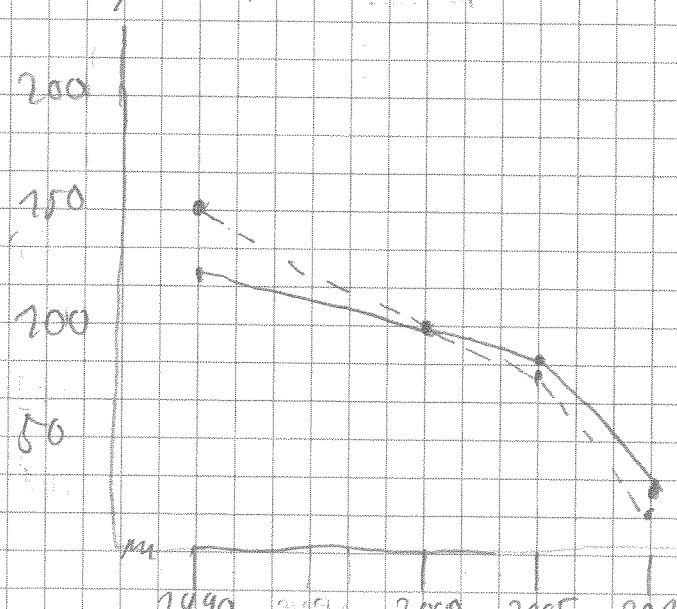
Hade basvärdet som tidigare idag

a) bas som tidigare

År	löpande	Deflator b)
1990	120	150
2000	100	100
2005	80	0,74 = 74
2010	30	0,26 = 26

1 + (1)

b)



0,75

Vad är vad?

d) 1 löpande index är den procentuella förändringen

400% $\frac{48}{12} = 4,000$

1 fast index är den procentuella förändringen

68% $\frac{204}{303} = 0,68$

SU, STATISTIK

Skrivsal: Brunnsvikssalen Anonymkod: [redacted] Blad nr: 3

3) 10 fritidshus slumpmässigt valda - avståndet till kusten studerades.

Avsikt - hitta ett samband mellan pris och avstånd till kusten.

a) Avstånd till kusten (km)

Försäljningspris i tusentals kronor

$\bar{x}$ 1234 tusen

	$X - \bar{X}$
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

0.5

Linjärt

a) -0,7 visar på ett lågt ? negativt samband d.v.s. att sambandet enligt beräkningarna inte är jättehöga. Detta kommer också påverka residualernas avstånd till grundmätningarna. ?

b) 1234,0 − 56,0x visar hur mycket priset sjunker för varje kilometer

1

c) 1234,0 = står för tusen dvs 1234000 kr ???
där −56 000 är att värdet minskar med 56 000 kr
för varje x värde. Husen blir billigare ju längre
från kusten de ligger

0.5

d) 1234,0 − 56,0 · 10 = 674 = 674000 kr

1

I och med detta kommer ett hus enligt dessa
beräkningar kunna vara gratis bara för
att det ligger för långt från kusten.
Det finns andra värden som spelar in så som vilken skick

SU, STATISTIK

Skrivsal: Brunnsviksscen Anonymkod: [redacted] Blad nr: 4

Undertäckning inte när vi till mål

4) Svens kommun vill undersöka

Målpopulationen kan anses vara alla myndiga personer som bor eller är verksamma i kommunen d.v.s. de som även arbetar i kommunen.

Rampopulationen kan vara alla invånare i kommunen. Eftersom det finns register över till/från kommun människor är bosatta i och även vart de bor. ??

a) Undertäckningsfel = Är när man når ut till Rampopulationen men inte till målpopulationen.

Ex. Att man når ut till alla bosatta i kommunen men inte dem som arbetar i den.

0,5

b) Övertäckningsfel = Är när man når man når ut till den målpopulationen men i större utsträckning än vad som var tänkt. men täcker ramen.

Övertäckning. D.v.s att man når ut till fler än vad de var tänkt från början.

(0,5)

Ex. Att man når ut till fler än de som bor och arbetar i kommunen.

Anledning till detta kan vara på grund av undersökningen som inte begränsar urvalet till målpopulationen. ?

c) Mätfel = En av de olika feltyperna.

Bortfall, täckningsfel, bearbetningsfel, mätfel, ...

mätfel är det uppstår fel i mätningarna. Att någon svarar dåligt på frågor eller inte förstår frågorna.

Ex. I kommunen kanske man väljer att göra en undersökning där man gör fel i mätningarna. Frågor fel vilket kan människor och genom dessa tör man inte överensstämmande svar i undersökningen.

(0.5)

d) Bortfallsfel = Det uppstår bortfall i undersökningen, man tar inte in antalet svar som man räknade med.

Ex. kommunen väljer att göra ett systematiskt urval där man väljer ut var tionde. 10, 20, 30, 40... Men där 10% av de som slumpmässigt har valts ut inte svarar.

0.25

e) Systematiskt urval = Bundet slumpmässigt urval.

Ex: I kommunen använder man sig av ett register över alla kommunmedlemmar där man slumpmässigt väljer en siffra ex. 4 då tar man 4, 14, 24, 34... Och Hur? Varför? bör de svaren på undersökningen.

0.25

F) Stratifierat urval = Delar upp i strata ex. kön där man slumpmässigt väljer ut vilken? från varje stratum.

Ex: Delar upp kommunmedlemmarna i låg, medel o hög inkomst. Därefter väljer man slumpmässigt ut vilka från varje strata.

0.25

4

g) kvoturval = icke slumpmässigt urval.
Undersökningsansvarige lämnas gränsen för
att avgöra vilka som ska undersökas.

Ex: Undersökningsansvarige står på centrum
tillsammans med 50 kollegor under en vecka
De ska totalt fråga ut - 1000 personer
d.v.s. 20 personer var. Kommunmedelmånarna
är totalt bestående av 60% kvinnor 0
40% män. Därför ska 600 kvinnor 0
400 män frågas ut. Individerna får
den undersökningsansvarige tillsammans med
hans kollegor välja själva. 0.5

h) (Urvalsfraktion = Hur mycket det skämmer
i jämförelse med en totalundersökning.
I fraktionen ska det kunna tas ett skick prov
där det ska kunna jämföras med ett skick prov
i en totalundersökning. Slumpmässigt bundet.

Ex. Urvalsfraktionen ska mäta samma
värden som skulle komma in i en
total undersökning.

Urvalsfraktionens resultat ska motsvara samma
resultat som i en totalundersökning.

100 slumpmässigt valda kommunmedelmånars
svar ska motsvara hela kommunens. 2)

0

5) a) nominaldata = kvalitativ, används när man ska mäta ex kön eller civilstånd. dvs 2 värden som är kvalitativa, men där ... ???
Ex: könsfördelningen i en gymnasieklass.

0.25

b) Diskret variabel = En kvantitativ variabel (med absolut noll punkt) Ex. på detta kan vara antalet barn i en familj som då går under datanivån kvotdata. Definition?

0.25

c) Frekvenspolygon = Flera frekvenser d.v.s. där du har en frekvens som inte kan separeras från en annan. Här ihop med kumulativ-frekvens, där frekvenserna för olika variabelvärden sammanställs till totala frekvens. ?)

0

d) Trädd-teknik = Där information sållas ut genom en hierarkisk struktur.
Man sållar i den tokala informationen för att underlätta sammanställningsarbetet.

Ex: I den tobaksundersökningen vi gjorde i en skola. Vill frågan använda oss av en träd-teknik för att snabbt kunna få en bild av hur resultatet ser ut. ?)

0

e) Nonsens korrelation: Ett samband som egentligen inte bör finnas.
Ex. Att man tittar mycket på tv skulle visa sig leda till att människor får högre IQ.
Det är ett nonsenssamband eftersom detta är ju helt beroende på vad du väljer att tittar på när du tittar på tv. (0.5)

f): Fastbasindex = (När du räknar ut det fasta Indexet som en deflator. Genom detta Sorterar du bort Inflationens påverkan av Indexet. Ex när man räknar ut ett Indexserie genom att använda sig av KPI. ☺) ○

g) Bias = — ○

h) Konfidensintervall = Intervalluppskattning — ○