STOCKHOLMS UNIVERSITET

Statistiska Institutionen

Mikael Havasi

Tentamen för Introduktion till statistik för statsvetare

Tisdag, 2014-04-30, Klockan 12:00-17:00, Laduviksalen

Tillåtna hjälpmedel: Formelblad (som vidhäftats tentamen) och miniräknare, utan lagrade formler eller text

Utlämning av tentan: torsdag 15/05 kl.11 i B705

Efter utsatt datum och tid för utlämning, kan tentan endast hämtas ut från studentexpeditionen

För poäng krävs tydliga svar! Tentan kan ge 100 poäng totalt. För varje fråga anges det vad respektive fråga motsvarar i poäng.

Lycka till!

1. En stor undersökning om universitetsstudenters alkoholvanor gjordes och man samlade in information hos de 5000 tillfrågade studenterna om hur mycket alkohol som dracks per vecka i form av starksprit (1 glas vin = 4 cl starksprit etc.). Det klassindelade materialet för variabeln "konsumtion av centiliter starksprit per vecka" ser ut som följande:

cl/vecka:	0-4	5-9	10-14	15-19	20-24	>24
Antal pers.:	1288	2019	861	254	154	424

a) Ange vilken variabeltyp och datanivå som här föreligger. Motivera ditt svar. (2p)

b) Illustrera fördelningen i ett lämpligt diagram tillsammans med beskrivande kommentarer. Motivera ditt val av diagram. (3p)

c) Medelvärdet och medianen beräknades för samtliga 5000 observationer (alltså ej för den klassindelade variabeln ovan, utan för alla ursprungliga värden). Medelvärdet blev 10.9 och medianen blev 7. Hur skall dessa två lägesmått tolkas i relation till varandra och till fördelningen? (2p)

d) I en undersökning med samma frågor, gjord på samma målpopulation några år tidigare erhölls medelvärdet 12.3 för den veckoliga konsumtionen av starksprit i cl och det 95% -iga konfidensintervallet [11.6 ; 13.0]. Ett 95 % -igt konfidensintervall för den nuvarande undersökningens medelvärde blev [10.0 ; 11.8]. Hur skulle du tolka ett sådant intervall för ett medelvärde? Vad drar du för slutsatser från intervallen? (4p)

2. För ett visst datamaterial med data på 9735 personer gjorde man en enkel linjär regressionsmodell mellan de två variablerna "Timlön" och "Antal arbetstimmar per vecka", där timlön var den beroende variabeln. Minsta-kvadrat-metoden gav följande modell:

$$\hat{y} = 104.5 + 2.74 * X_1$$

a) Vilken av de två koefficienterna är interceptet och vilken koefficient är lutningskoefficienten? Tolka dessa två koefficienter. (5p)

b) Du får inte tillgång till det kompletta datamaterialet, däremot får du reda på att R^2 beräknats till 0.043. Förklara vad R^2 är, hur det beräknas och vad det givna värdet säger om modellen. (5p)

c) Sedan skapade man även en multipel linjär regressionsmodell där man lade till ytterligare oberoende variabler och fick då ett värde på R^2 som var 0.597.

De oberoende variablerna var:

X_1 = Antal arbetstimmar per vecka

X_2 = Utbildningsnivå, antal års utbildning (t.ex. grundskola = 9 år, gymnasie = 12 år, kandidatexamen = 15 år etc.)

X_3 = Kön (Kvinna=1, Man=0)

X_4 = Privat sektor (Ja=1, Nej=0)

X_5 = Arbetserfarenhet i antal år

Nu blev modellen:

$$\hat{y} = 50.4 + 0.084 * X_1 + 10.14 * X_2 - 20.79 * X_3 + 17.74 * X_4 + 4.74 * X_5$$

Resonera kring den multipla modellen i jämförelse med den enkla linjära modellen. Gör en kritisk analys med hjälp av den information som ges. *Skriv kort och koncist! Max ett halv A4!* (5p)

3. Ett landsting är intresserade av åsikter om åtgärdsprogram för långtidsarbetslösa bland de arbetslösa i länet och vill därför göra ett slumpmässigt urval för att kunna uttala sig om hela gruppen.

a) Vad innebär mål- och rampopulation? Ge exempel på lämplig mål- och rampopulation i ovanstående situation. (5p)

b) Det totala felet består av fem feltyper. Ange de fem olika feltyperna, förklara vad de innebär och ge exempel i denna situation. *Tips: Strukturera ditt svar! Skriv kort och koncist!* (15p)

4. För ovan nämnda situation i fråga 3, måste också ett slumpmässigt urvalsförfarande väljas. För varje urvalsmetod, förklara hur metoden går till, ge en fördel och en nackdel och ge exempel för den givna situationen. *Tips: Strukturera ditt svar! Skriv kort och koncist!*

a) Obundet slumpmässigt urval (6p)

b) Stratifierat urval (6p)

c) Klusterurval (6p)

d) Systematiskt urval (6p)

5) Redogör *kortfattat* för nedanstående åtta begrepp/metoder samt ge realistiska exempel på respektive begrepp/metod (3p för varje delfråga):

a) Tratt-tekniken

b) Löpande och fasta priser

c) Paaches index

d) Intervallskala

e) Dikotom variabel

f) Säsongsvariation

g) Kohort

h) Chi^2 – test

i) Nonsenssamband

j) Residual

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: 30/4 - 2014

Sal: Laduvikssalen

Tenta: Statistik för statsvetare

Kurs: Introduktion till statistik för statsvetare

ANONYMKOD: SFS 0013

☑ 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					4
Lär.ant. 11	13	20	21	30					

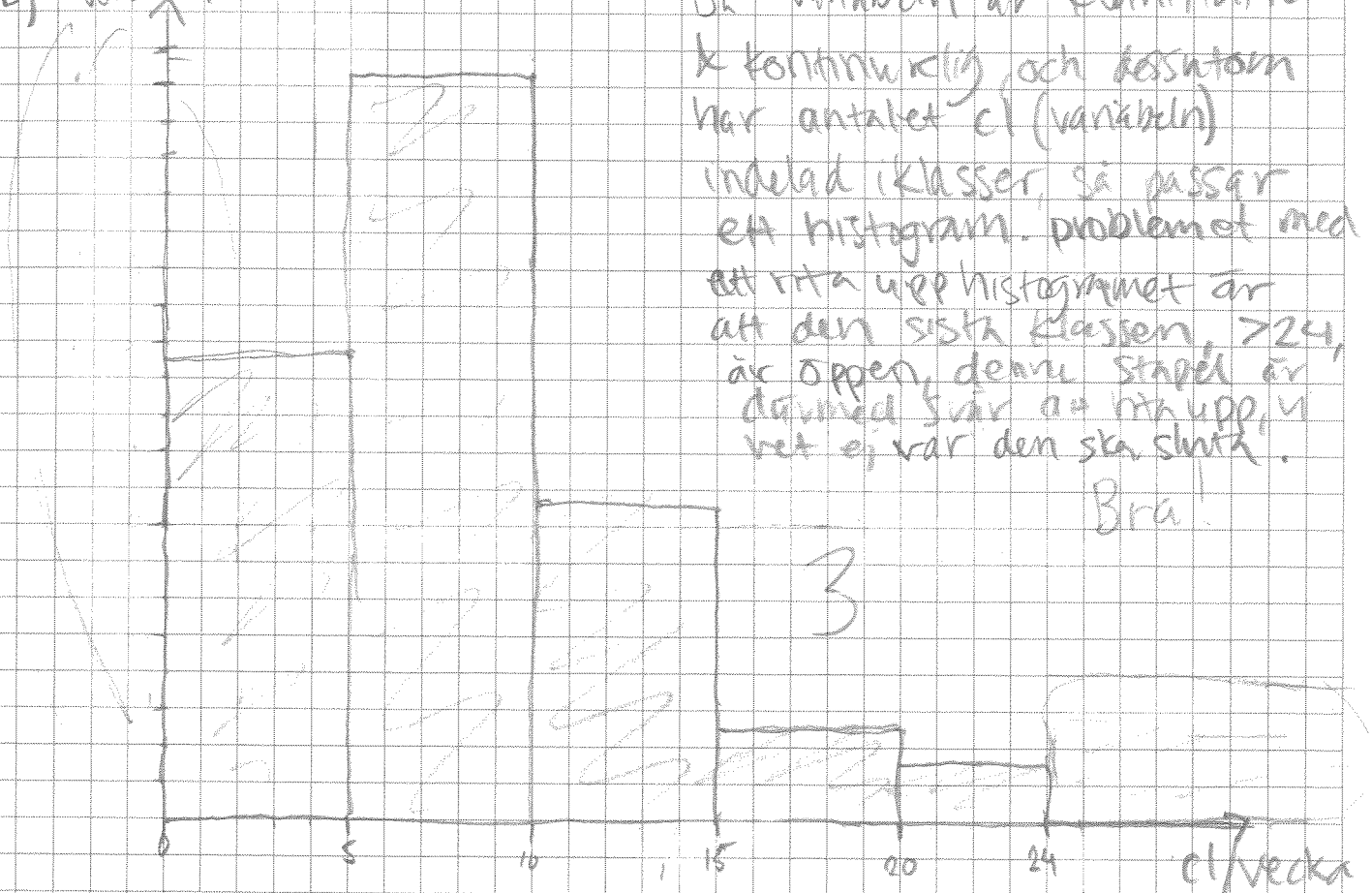
POÄNG	BETYG	Lärarens sign.
95	A	

SU, STATISTIK

Skrivsal: Ladmk Anonymkod: SFS0013 Blad nr: 1

1) a) kvantitativ kontinuerlig variabel. Då variabeln (cl) är numerisk och kan anta alla värden inom intervallet.
o kvotskala då centiliter har en absolut nollpunkt, (man kan säga att 2cl är dubbelt så mycket som 1cl.) 2

b) antal pers



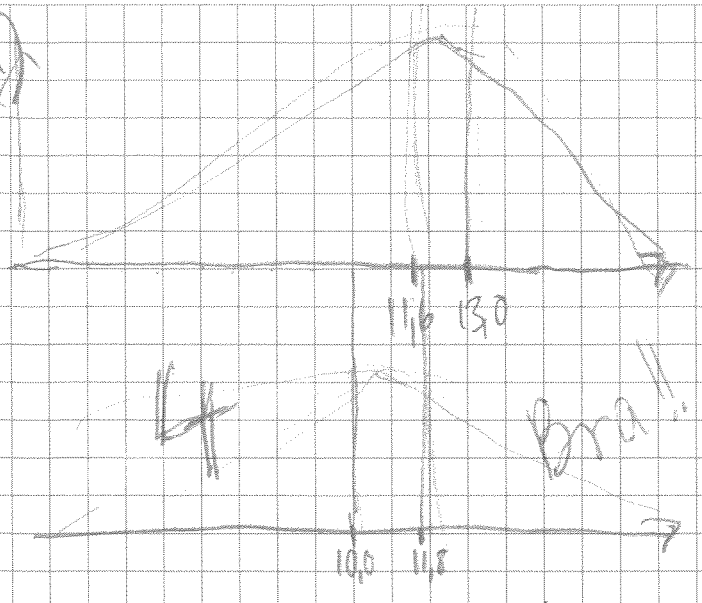
Då variabeln är kvantitativ & kontinuerlig och dessutom har antalet cl (variabeln) indelad i klasser så passar ett histogram. problemet med att rita upp histogrammet är att den sista klassen, >24, är öppen, denna stapel är därmed svår att rita upp, vi vet ej var den ska sluta.
Bra! 3

$\bar{x} = 10,9$ median: 7

c) medelvärdet är summan av alla observationer delat på antalet observationer. Medianen är den mittersta observationen efter att man ordnat dess i storleksordning.
I detta fall är medelvärdet missvisande, det har blivit förskjutet på grund av de högre extrem värdena. medianen 7 i passar därför bättre, det ser vi på histogrammet där klassen 5-9 har högst frekvens. 2

SD

d)



Konfidensintervallet med 95% innebär att det sanna medelvärdet med 95% säkerhet ligger inom intervallet 11,6 och 13,0 för det äldre resultatet och för det nya resultatet inom intervallet 10,0 och 11,8

Jämför man de båda intervallen kan man påstå att medelvärdet rav konsumtionen av starksprit minskat. Men då de båda intervallen överlappar 11,6 - 11,8 finns det en möjlighet att det samma medelvärdet (med 95% säkerhet) inte har förändrats.

---

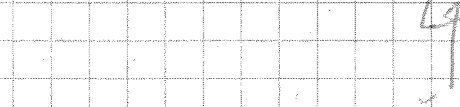
2 $\hat{y} = 104,5 + 2,74 \cdot x_1$

a) 104,5 (a) är interceptet, det vill säga timlönen utan påverkan av antalet arbetstimmar, där linjen skär y-axeln är 104,5 kr

+2,74 (b) är lutningskoefficienten. För varje arbetad timme ökar timlönen med 2,74 kr.

b) $R^2 = 0,043$ är determinationskoefficienten, och kan variera mellan 0 och 1. R^2 anger hur stor andel av y som kan förklaras av x, hur stor andel av observationerna som ligger på den skattade linjen.

0,043 ⇒ 4,3% vilket därmed är en låg andel, det är därmed för y som kan förklaras av x.



2 c) $R^2 = 0,597 = 59,7\%$. Högre andel y kan förklaras av alla oberoende variabler, fler observationer ligger på denna skattade linje.
Med denna modell ser man att antalet arbetade timmar per vecka är det som påverkar lönen minst, kön är det som påverkar mest, nära felfritt av i vilken sektor man arbetar.
Enligt modellen är de med högst timlön lågutbildade kvinnor utan arbetserfarenheter som inte jobbar i privat sektor. Skulle man beräkna R^2 med endast X som KÖN eller/och X som sektor ty skulle R^2 bli närmre 1.

---

3 a) Målpopulationen är de enheter som man idealiskt skulle vilja undersöka, i detta fall vill man ha åsikter från "alla arbetslösa i länet". Ramen är det register vi använder, rampopulationen de enheter som finns inom ramen. I detta fall kan ett register vara alla de som skrivit in sig som arbetslösa på arbetsförmedlingen.
Det som skiljer mål- från rampopulation är ett ev. täckningsfel. (se 3b) 5

b) Det totala felet = Urvalsfel + täckningsfel + bortfallsfel + mätfel + bearbetningsfel.

→

Urvalsfel: Uppstår när vi inte gör ett totalurval. Med slumpmässiga urval kan man beräkna väntesvärde 3
väntevärde/sannolikheten, därför ett bra val av landstinget!

Täckningsfel: Uppstår vid utformningen av ramen.

Övertäckning: När registret (ramen) innehåller enheter som inte ingår i målpopulationen tex personer i arbetsförmedlingens register som inte längre är arbetslösa. Går att motverka genom filterfråga

3 Undertäckning: enheter som finns i målpopulationen men ej i registret. tex arbetslösa som inte ännu registrerat sig hos arbetsförmedlingen. Ses som allvarligare än övertäckning, svår att upptäcka.

Bortfallsfel: Variabelbortfall: vid tex enkäter där vissa frågor inte besvarats, kanske pga missförstånd eller att en viss fråga om arbetslöshet varit känslig att besvara.

3 Individbortfall: den utvalda enheten går inte att få tag på, vill inte medverka eller misslyckas med att medverka - förstår kanske inte språket.

Mätfel: Man lyckas inte mäta det man vill mäta då mätverktyget inte ger samma utfall när 3 försöket upprepas. Kan vara svårt att mäta åsikterna hos de arbetslösa.

bearbetningsfel: När man ska koda de tillfrågades svar råkar man kanske skriva in fel 3 siffra. Denna feltyp är den som minst påverkar det totala felet.

4) a) OSU: Alla enheter har lika stor sannolikhet att komma med, alla arbetslösas namn skrivs upp på varsin lapp och sedan drar man önskat antal bland alla lappar.

\+ teoretiskt enkelt att utföra

\- Om målpopulationen innehåller väldigt många enheter finns risken att de inte blir representerade.

6

b) Strat: De arbetslösa delas in i olika grupper, strata; tex i åldersklasser och sedan görs ett OSU så att alla klasser blir representerade i urvalet.

\+ bra om det finns en snedfördelning, där det finns risk att viktiga mindre grupper inte kommer med

\- finns risk för motstridiga stratifieringar för olika parametrar

5

c) Kluster: Man gör först ett urval bland alla arbetsförmedlingar i olika kommuner och sedan, gör man ett urval bland de utvalda arbetsförmedlingarnas arbetslösa.

\+ Bra om man saknar en fullständig ram

\- tar tid att administrera.

5

d) Syst: arbetsförmedlingens register nummeras, ett startnummer slumpas, tex "6" och sedan väljer man tex var tionde enhet efter det, 6, 16, 26, 36 osv.

+ Bra om register är ordnat på lämpligt sätt för undersökningen

− Inget för periodicitet om register är ordnat på olämpligt sätt för undersökningen, tex ålder, i veckodagar, geografiskt etc.

5

5) a) Vid utformningen av enkätfrågor. Börja tex med breda lättförklarade frågor (ålder, kön) och gå gradvis svårare/djupare frågor (åsikter, känsliga/privata)

3

b) Vid jämförelser av tex prisförändringen på en vara över tid. Löpande priser går inte att jämföra då man inte räknat med inflationen 10kr för biobiljett för 30 år sedan är 120kr idag säger oss inget om den verkliga ökningen. Fasta priser är lösningen på detta. Man beräknar priserna till samma tidpunkt, då går de att jämföra, dvs tex räknar fram vad 10kr på 70-talet har för värde idag.

3

c) Ett räknesätt/teori för hur man försöker lösa problemet med att när en vara ökar i pris så köper man kanske inte lika mycket av den, när man räknar ut index för flera perioder. priset för tidpunkt 1 multipliceras med antalet sålda enheter (kvantiteter) för samma år. produkten divideras sedan med produkten av priset år 0 (basårspunkten) och kvantiteten år 1. Här tar man inte hänsyn till vad kvantiteten säger vid basårspunkten.

3

SU, STATISTIK

Skrivsal: Ladunik Anonymkod: SFS0013 Blad nr: 4

5 d) Kvantitativa variabler som inte har en absolut nollpunkt hamnar på datanivån intervallskala. Dessa variabler ger inte någon användbara kvoter.

3 tex temperatur. 10°C är inte dubbelt så varmt som 5°C, och 0°C är inte "ingen temperatur".

e) En variabel med endast två möjliga utfall, tex Ja/Nej på frågan "har du barn".

3 kallas även 0/1 variabel då svaren kan kodas till just 0 & 1 (Ja=1 Nej=0)

f) En av 4 förändringsfaktorer som kan påverka index. tex försäljningen av glass är som störst under sommar/vår och minst låg under vinter. - försäljning vanligtvis påverkas av säsongen.

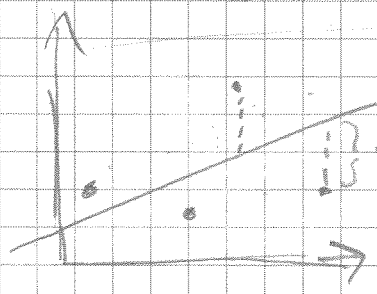
g) Undersökning av en grupp människor under en längre tid. tex alla födda den 29 maj 2010 3 följs under 50 år där man kartlägger deras hälsoutveckling.

h) Ett sätt för att mäta sambandet mellan två kategoriska variabler! tex gift/ogift & villa/ej villa Genom en korstabell beräknar man den förväntade fördelningen om det saknas samband och med chi-två värdet kan för man sedan observationernas värden och ser om det ligger under eller över det kritiska värdet. Varpå man sedan kan förkasta en av 0+1 hypoteserna.

i) Nonsens: Man tror att det finns ett orsakssamband mellan två variabler men egentligen är det en annan variabel som påverkar de båda. tex: tv-abonnemangen ökar samtidigt som allt fler ser dåligt (köper glasögon) Mer tv-tittande leder inte till sämre syn, men ökande förbrukning ekonomin leder till att fler anser sig ha råd att betala för fler kanaler och bekvämt att besöka hos optikern.

3

i)



Avståndet mellan den skattade linjen och den observerade punkten

3

Vid skapandet av linjen vill man att summan av alla kvadrerade residualer ska vara så litet som möjligt, att observationerna ska ligga så nära linjen som möjligt.