STOCKHOLMS UNIVERSITET
Statistiska institutionen
Ulf Högnäs

ÖVNINGSTENTA Statistisk översiktskurs, 4.5 hp

Kurs: ST1801, Statistisk Översiktskurs, 7.5 hp
Tentamensdatum: 2025-xx-xx

Skrivtid: xx.00-xx.00 (5 timmar).
Godkända hjälpmedel: Miniräknare utan lagrade formler och text.

Tentamen består av 6 uppgifter, uppdelade i deluppgifter. Maximalt antal poäng anges per deluppgift.

Svar med fullständiga redovisningar ska lämnas.

- Använd endast skrivpapper som tillhandahålls i skrivsalen.
- För full poäng på en uppgift krävs tydliga och väl motiverade lösningar.
- Om du inte lyckas lösa en deluppgift och behöver det svaret för en senare deluppgift så kan du hitta på värdet för att kunna göra beräkningarna i de efterföljande uppgifterna.

Tentamen kan maximalt ge 100 poäng, och för godkänt resultat krävs minst 50.

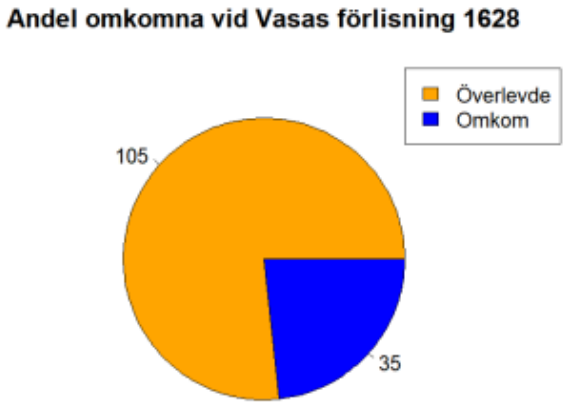
Betygsgränser:

A:	90-100
B:	80-89
C:	70-79
D:	60-69
E:	50-59
Fx:	40-49
F:	0-39

OBS! Fx och F är underkända betyg som kräver omexamination. Studenter som får betyget Fx kan alltså inte komplettera för högre betyg. Lösningsförslag läggs ut på kurshemsidan efter tentamen i samband med rättningen.

`Lycka till!`

1. (a) Enligt en (mycket osäker) uppgift hade regalskeppet Vasa en besättning på 150 personer varav 35 (23.3%) omkom vid förlisningen 1628. Hur många fel kan du hitta i följande Figur 1? (5 p)



Figur 1: Ett felaktigt pajdiagram

Lösning Siffrorna i pajdiagrammet ska ange andelar, inte antal. Det totala antalet i besättningen är 150, men $105+35 = 140$. Bonus: en vanlig konvention säger att den största biten bör vara placerad med sin "vänsterkant" vid klockan 12.

(b) Förklara kortfattat skillnaden mellan intervallskala och kvotskala. Använd exempel om du vill (5 p)

Lösning En variabel på kvotskala har en nollpunkt som gör att kvoten mellan två tal på skalan har en naturlig tolkning. Temperatur mätt i grader Celsius är på intervallskala eftersom kvoten mellan två tal inte betyder något - tio grader är inte dubbelt så varmt som fem grader.

Temperatur mätt i Kelvin är ett exempel på kvotskala. Ett annat exempel är längd. Om Anna är 150 cm lång och Anders är 180 cm lång så är kvoten Anders delat på Anna $180/150 = 1.2$. Tolkningen är att Anders är 20% längre än Anna.

(c) Antag att det bor 200 låg- och medelinkomsttagare i en liten by. En vacker dag flyttar en framgångsrik företagare med miljoninkomst till byn. Hur påverkas medianen? Förklara. (3 p)

Lösning Medianen beräknas genom att ordna inkomster, minsta till största. Sedan väljs mittenvärdet som median. Om det är ett jämnt antal värden beräknas medianen som genomsnittet av de två mittersta talen.

Med 200 invånare blir medianen genomsnittet av den hundrade och hundraförsta inkomsten, ordnat långt till högt. När en mycket rik person flyttar in blir medianen istället den hundraförsta inkomsten. Om invånarna har små och medelstora inkomster så kommer de två mittenobservationerna att ligga ganska nära varandra, så medianen påverkas inte så mycket i detta fall.

(d) I Figur 2 visas en tabell från boken. Med *mortgage* menas bostadslån och med *joint* menas gemensam ansökan. Vad betyder siffran 0.635? (2 p)

Table 4.3: A contingency table with row proportions for application type and homeownership.

	homeownership			
application_type	rent	mortgage	own	Total
joint	0.242	0.635	0.122	1
individual	0.411	0.451	0.138	1

Figur 2: Joint

Lösning Denna tabell visar fördelningen av homeownership, betingat på "application type". Siffran 0.635 är andelen av alla ansökningar i categroin "joint" som tillhör "mortgage".

2. (a) Du singlar slant tre gånger. Vad är sannolikheten att du får *krona* alla tre gångerna? (5 p)

   Lösning Sannolikheten vi en enskild singling är $\frac{1}{2}$. Singlingarna är oberoende, så vi kan få sannolikheten för tre "krona" genom att multiplicera de tre sannolikheterna

$$\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}.$$

   (b) Vad är sannolikheten att du minst en *klave* på tre singlingar? (5 p)

   Lösning Om vi **inte** får minst en "klave" så får vi definitivt tre "krona". Det betyder att

$$\text{P("minst en klave")} + \text{P("tre krona")} = 1$$

   Vi säger att "minst en klave" och "tre krona" är komplement. Vi får

$$\text{P("minst en klave")} = 1 - \text{P("tre krona")} = 1 - \frac{1}{8} = \frac{7}{8}.$$

   (c) Förklara kortfattat varför observationsstudier ofta inte kan användas för att dra slutsatser om kausalitet. Du kan använda ett påhittat exempel. (5 p)

   Lösning Påhittat exempel: I samband med en observationsstudie av svenska vuxna finner man att kaffedrickare är mer stressade en de som inte dricker kaffe. Men orsakar kaffedrickande stress? Vi vet inte. Det kan vara så att stressade personer har en oftare har en livsstil där kaffedrickande ingår.

   (d) Tabellen nedan visar den uppmätta dagstemperaturen klockan 12:00 under fem dagar i Uppsala.

12	14	13	15	16

Tabell 1: Dagstemperaturer i Uppsala

   Här är formeln för **stickprovsvarians**:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

   Beräkna temperaturens standardavvikelse. Visa dina beräkningar.
   (5 p)

Lösning

i. Vi räknar ut medelvärdet

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{70}{5} = 14.$$

ii. Vi gör en tabell för att så ett ordnat sätt räkna ut summans termer.

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	12	-2	4
2	14	0	0
3	13	-1	1
4	15	1	1
5	16	2	4

iii. Nu har vi allt vi behöver för att beräkna variansen.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{4+0+1+1+4}{5-1} = \frac{10}{4} = 2.5.$$

iv. Standardavvikelsen är roten ur variansen

$$s = \sqrt{2.5} \approx 1.58$$

```
> prop.test(c(85, 75), c(530, 550))

        2-sample test for equality of proportions with continuity correction

data:  c(85, 75) out of c(530, 550)
X-squared = 1.0504, df = 1, p-value = 0.3054
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.02024803  0.06827548
sample estimates:
   prop 1    prop 2
0.1603774 0.1363636
```

Figur 3: `prop.test()` i R

3. I en opinionsundersökningen *Ministerförtroende* fick slumpmässigt utvalda väljare frågan "Vilket förtroende har du för följande ministrar?" Den sista mätningen 2024 omfattade 530 personer. I den undersökningen hade 85 personer "ganska stort" eller "mycket stort" förtroende för Johan Persson.[†]

   En tidigare mätning omfattade 550 personer. I den hade 75 personer "ganska stort" eller "mycket stort" förtroende för Johan Persson. Figur 3 visar output from R.

   (a) Figur 3 visar ett konfidensintervall för skillnaden i andel, $p_1 - p_2$. En student som inte studerat statistik säger "jag behöver inget konfidensintervall. Skillnaden i andel är 2.4%". Hen visar följande uträkning:

   $$\frac{85}{530} - \frac{75}{550} \approx 0.0240$$

   Förklara kortfattat skillnaden mellan $p_1 - p_2$ och $\hat{p}_1 - \hat{p}_2$. (5 p)

   Lösning $\hat{p}_1 - \hat{p}_2$ är känd. Det är skillnaden i stickprovsandel. Men vi varje mättillfälle tillfrågades bara cirka femhundra personer av över sex miljoner väljare. Den verkliga skillnaden i andel för de två tidpunkterna, bland alla väljare, är alltså okänd. Denna skillnad som $p_1 - p_2$ syftar på.

   (b) Tolka konfidensintervallet. (3 p)

   Lösning Vi kan med 95% konfidens säga att skillnaden i andel är mellan -2.02% och 6.83%

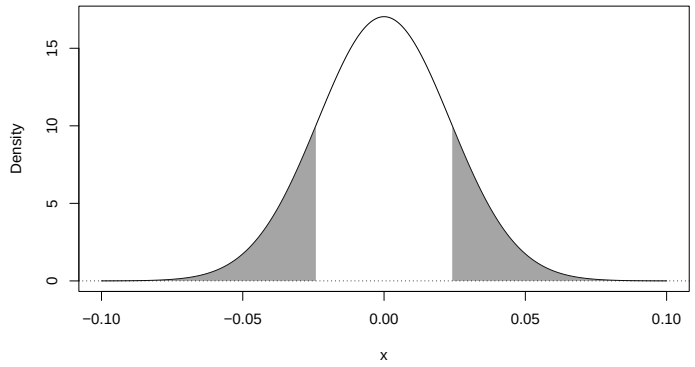
   Alternativ Om vi kunde upprepa denna typ av undersökningar så skulle det resulterande intervallet fånga den verkliga skillnaden med sannolikhet 95%.

   (c) Beräkna felmarginalen (5 p)

   Lösning Felmarginalen är avståndet från punkskattnigen (mitten av K.I.) och kanten av K.I. Konfidensintervallet är alltså två felmarginaler brett. Ett sätt att lösa uppgiften är att beräkna K.I. bredd och dela det med två. Med lite slarvig avrunding,

   $$ME = \frac{0.068 - (-0.020)}{2} = \frac{0.088}{2} = 0.044$$

[†]Dessa siffror är påhittade.



Figur 4: Illustration av p-värdet. Figuren har inte normaliserats.

(d) Vad kan opinionsinstitutet göra för att minska felmarginalen vid nästa undersökning? (5 p)

Lösning De kan minska osäkerheten och därmed felmarginalen genom att fråga fler väljare.

(e) Det p-värde som kan hittas i Figur 3 är relaterat till ett hypotestest med hypoteserna

$$H_0 : p_1 - p_2 = 0$$
$$H_A : p_1 - p_2 \neq 0$$

Förklara kortfattat p-värdets innebörd med detta som exempel. (5 p)

Lösning *Om nollhypotesen är sann, hur sannolikt är det att få en skillnad lika extrem som eller mer extrem än 0.024?* Svaret på detta är p-värdet, dvs 0.3054.

Eftersom det är ett två-sidigt test, så är ´´mer extrem" antingen större än 0.024 eller mindre än -0.024. Sannolikheten illustreras av de gråa svansarna i Figur 4

(f) Låt oss säga nollhypotesen inte kan förkastas i detta fall, men att nollhypotesen i själva verket är falsk. Vad kallas detta typ av fel? (2 p)

Svar Detta kallas **typ II-fel.**

term	estimate	std.error	statistic	p.value
(Intercept)	1.90	0.20	9.56	<0.0001
verified_incomeSource Verified	1.00	0.10	10.05	<0.0001
verified_incomeVerified	2.56	0.12	21.86	<0.0001
debt_to_income	0.02	0.00	7.44	<0.0001
credit_util	4.90	0.16	30.25	<0.0001
bankruptcy1	0.39	0.13	2.96	0.0031
term	0.15	0.00	38.89	<0.0001
credit_checks	0.23	0.02	12.52	<0.0001

Adjusted R-sq = 0.2598
df = 9966

Figur 5: En modell som skattar utlåningsränta

4. Sammanställningen i Figur 5 visar resultaten från en linjär regressionsmodell som skattar den genomsnittliga utlåningsräntan, angiven i procent, baserat en rad förklarningsvariabler hos kunder till låneistitutet *Lending Club*.

   (a) Vad är en dummyvariabel? (5 p)
   Lösning En dummyvariabel är en kategorisk variabel som kodats som 0 & 1. Exempel, kvinna = 1, man = 0.

   (b) De två variblerna

   - `verified_incomeSource Verified`
   - `verified_incomeVerified`

   beskriver en kategorisk variabel med tre nivåer,

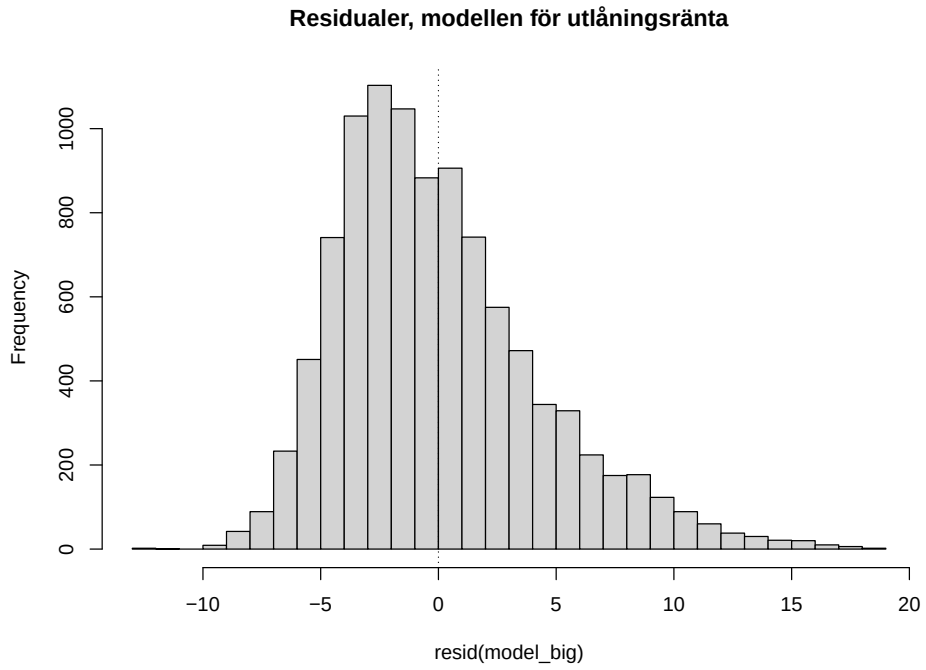
   - Source Verified
   - Verified
   - Not Verified

   Förklara kortfattat hur detta fungerar. (5 p)
   Lösning

Verification	Verified	SourceVerified
Not Verified	0	0
Verified	1	0
Verified	1	0
Not Verified	0	0
Source Verified	0	1

Tabell 2: Tre nivåer kodas med två dummyvariabler.

Om båda variablerna `verified_incomeSource Verified` och `verified_incomeVerified` sätts till noll så betyder det att personen tillhör "not verified". Detta kallas **Base Category**.



Figur 6: Modellen för utlåningsränta. Residualer.

(c) Variabeln `credit_checks` beskriver hur många gånger kundens kreditvärdighet har kontrollerats av något kreditinstitut de senaste tolv månaderna. Tolka koefficienten. (5 p)

Lösning Om vi jämför två sorters kunder, där den ena har en credit check mer än den andra så har kunden med en mer credit check en förväntad ränta om är 0.23 högre än den andra kunden, **Allt Annat Lika!**

(d) Antag att vi vill använda *Backward elimination* för att förbättra modellen. Vi tar bort variabeln `bankruptcy1` och skattar modellen igen. Hur kan vi avgöra om detta har förbättrat eller försämrat modellen? (5 p)

Lösning Vi jämför **Adjusted** R^2. Om den blir lägre så har modellen försämrats, och tvärt om.

(e) Histogrammet i Figur 6 visar fördelningen av modellens residualer. Detta tyder på att ett antagande som ligger till grund för linjär regression inte är uppfyllt. Förklara kort. (5 p)

Lösning Ett viktigt antagande för linjär regression är att felen, dvs de slumpmässiga avvikelserna från "linjen", är normalfördelade. Man kan undersöka detta antagande genom att studera residualerna. Detta är en klart skev fördelning och inte symmetrisk som vi förväntar oss av en normalfördelning.

(f) Vad är multikollinearitet? Ge ett kortfattat exempel. (5 p)

Lösning Multikollineritet uppstår när två förklaringsvaribler är starkt korrelerade. Detta kan orsaka problem i modellen. Ett exempel är bostadsyta och antal sovrum i modellen för huspriser som förekommer i kapitel 10.

För svar på fråga 5, se föreläsningsbilderna från F10!

5. (a) Förklara kortfattat vad som menas med primärdata respektive sekundärdata. Ange en fördel och en nackdel med att använda sekundärdata. (5 p)

   (b) Förklara begreppen målpopulation och rampopulation med ett exempel. (5 p)