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.

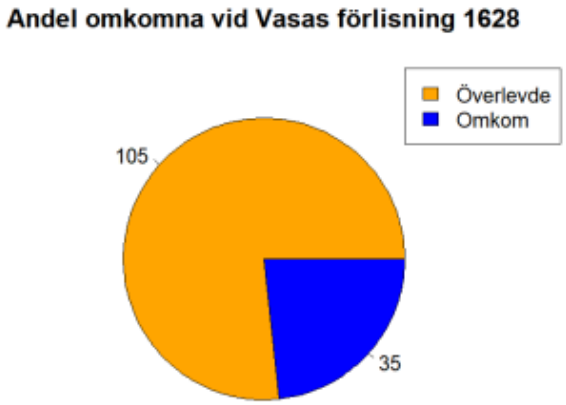
Betygsgränser:

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

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

(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)

(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

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

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

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

   (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)

```
> 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)

   (b) Tolka konfidensintervallet. (3 p)

   (c) Beräkna felmarginalen (5 p)

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

   (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)

   (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)

[†]Dessa siffror är påhittade.

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 4: En modell som skattar utlåningsränta

4. Sammanställningen i Figure 4 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)

   (b) De två variblerna

   - `verified_incomeSource Verified`
   - `verified_incomeVerified`

   beskriver en kategorisk variabel med tre nivåer,

   - Source Verified
   - Verified
   - Not Verified

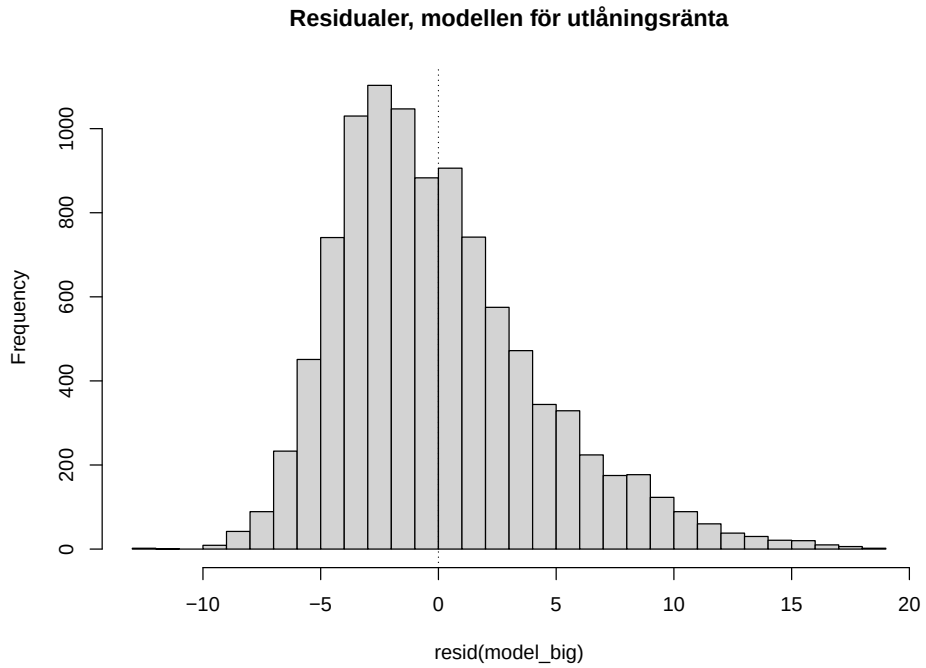
   Förklara kortfattat hur detta fungerar. (5 p)

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

   (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)

   (e) Histogrammet i Figur 5 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)

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



Figur 5: Modellen för utlåningsränta. Residualer.

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)