

# SDAI (ST1101), Tentamen 2, 6 hp

Stockholms universitet, statistiska institutionen

Kurs: Statistik och dataanalys I, 15 hp

Tentamensdatum: 2024-11-01

Skrivtid: kl. 14–19 (5 timmar).

Godkända hjälpmedel: Miniräknare utan lagrade formler och text.

Bifogade hjälpmedel: Formel- och tabellsamling för statistik och dataanalys I, 15 hp.

Tentamen består av 5 uppgifter uppdelade i deluppgifter. Maximalt antal poäng anges per deluppgift.

Svar med fullständiga redovisningar ska lämnas för fulla poäng.

- Använd endast skrivpapper som tillhandahålls i skrivsalen.
- För full poäng på en uppgift krävs tydliga, utförliga och väl motiverade lösningar.
- Kontrollera alltid dina beräkningar och lösningar! Slarvfel kan ge poängavdrag.
- 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 beräkningar i de efterföljande uppgifterna.
- I beräkningar från R-utskrifter får du utgå från det som är givet.

Tentamen kan maximalt ge 100 poäng. För godkänt resultat krävs minst 50 poäng.

## Betygsgränser

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

Obs! Fx och F är underkända betyg som kräver omexamination. Studenter som får betyget Fx kan alltså inte komplettera för högre betyg.

Lösningförslag läggs ut på athena efter att tentamenstiden är över.

**Lycka till!**

### Uppgift 1 (16 poäng)

Låt  $A$  och  $B$  vara två oberoende händelser med  $P(A) = 0.3$  och  $P(B) = 0.6$ .

- a) Förklara vad som menas med den betingade sannolikheten för en händelse och beräkna  $P(A | B)$ . (4p)
- b) Vad är sannolikheten att både  $A$  och  $B$  inträffar? (4p)
- c) Vad är sannolikheten att åtminstone en av  $A$  och  $B$  inträffar? (4p)
- d) Förklara skillnaden mellan ett utfall och en händelse. Använd gärna ett exempel. (4p)

## Uppgift 2 (20 poäng)

En telefonförsäljare ringer till slumpmässigt valda nummer för att sälja lotter. Försäljaren lyckas sälja lotter i 18% av samtalen. Samtalen kan betraktas som oberoende.

- a) Om telefonförsäljaren genomför fyra samtal, vad är sannolikheten att hen lyckas sälja lotter i endast det första samtalet? (3p)
- b) Vad är sannolikheten att det först är vid det sjunde samtalet telfonförsäljaren säljer sin första lott? (4p)
- c) Innan första fikapausen hinner telefonförsäljaren ringa 10 samtal. Vad är sannolikheten att telefonförsäljaren lyckas sälja lotter i färre än två av dessa samtal? (6p)
- d) Om telefonförsäljaren säljer lotter i färre än 8 samtal per dag i genomsnitt så kan telefonförsäljaren bli avskedad. Vad är det minsta antalet telefonsamtal försäljaren behöver genomföra på en dag för att det *förväntade* antalet samtal som leder till försäljningar ska bli minst 8? (7p)

### Uppgift 3 (21 poäng)

Den genomsnittliga vikten på gul fjällräv har länge antagits ligga på ungefär 2300 gram. På grund av klimatförändringar misstänker dock forskare vid Umeå landsbruksuniversitet att det har skett en förändring av rävarnas genomsnittliga vikt. Utgå ifrån att rävarnas vikt är normalfördelad, men att både väntevärdet ( $\mu$ ) och standardavvikelsen ( $\sigma$ ) är okänd.

- a) Forskarteamet fångar in 8 slumpmässigt valda gula fjällrävar och mäter deras vikt. Det visar sig att den genomsnittliga vikten för stickprovet är  $\bar{x} = 1976$  gram, med en standardavvikelse på  $s = 263$ . Utför ett hypotestest på 1% signifikansnivå för att undersöka om den genomsnittliga vikten har ändrats, och alltså skiljer sig från 2300. Ställ upp hypoteser, utför testet och dra korrekta slutsaser. Vilket antaganden om populationsfördelningen krävs för att testet ska vara giltigt? (9p)
- b) Det visar sig att den tidigare genomsnittliga vikten (2300) är baserad på ett begränsat stickprov på 9 rävar taget någon gång på sjuttioalet. Detta stickprov hade en standardavvikelse på  $s = 156$ . Genomför ett nytt hypotestest som undersöker om den genomsnittliga vikten skiljer sig åt mellan idag och sjuttioalet. Använd samma signifikansnivå som i 3a. Ställ upp hypoteser, utför testet och dra korrekta slutsaser. (8p)
- c) Skulle det kunna förekomma att du får ett *icke signifikant* resultat i a och ett *signifikant* resultat i b? Glöm inte att motivera ditt svar! (Detta gäller alla frågor...)(4p)

#### Uppgift 4 (25 poäng)

Ett företag som skapar slot machines för kasinospel påstår att utbetalningen av vinster kan beskrivas med sannolikheterna i nedanstående tabell (kolumn 2). Vi kallar vinsten minus kostnaden för en viss spelomgång för *resultatet*. Varje spel kostar 10 USD, så resultatet blir alltid 10 USD lägre än vinsten.

Vinst	Sannolikhet	Observerat antal
150	0.02	13
20	0.15	69
10	0.3	164
0	0.52	254

- En medelålders spelfantast vid namn Björn vill undersöka om dom givna sannolikheterna faktiskt stämmer och bestämmer sig för att spela 500 gånger. Du hittar resultaten från Björns experiment i kolumn tre i tabellen ovan. Genomför ett goodness-of-fit test för att undersöka om dom observerade antalen följer fördelningen i kolumnen "Sannolikhet" eller inte. Ställ upp hypoteser, utför testet och dra korrekta slutsatser. Använd signifikansnivån  $\alpha = 0.05$ . (9p)
- Nämnd minst två av dom tre antagandena som krävs för att testet i 4a ska vara giltigt. Avgör, i dom fall det är möjligt, om antagandena är uppfyllda baserat på texten ovan. (3p)
- Resultatet varje spelomgång kan beskrivas som en slumpvariabel  $X$  vars sannolikhetsfunktion ges av kolumn 2 i tabellen ovan. Låt  $X$  vara resultatet vid en spelomgång. Är  $X$  diskret eller kontinuerlig? Beräkna väntevärdet och variansen för  $X$ . (9p)
- Slumpvariabeln i c) följer *inte* normalfördelningen. Trots detta anser Björn att det är OK att normalapproximera det totala förväntade resultatet om man spelar 500 gånger. Håller du med Björn? Förklara varför/varför inte. (4p)

## Uppgift 5 (18 poäng)

Ett slumpmässigt urval av 25 djur av olika arter, från råttor och människor till elefanter, har samlats in av en grupp biologer. Biologerna har sedan mätt både djurens kroppsvikt (i kilo) och vikten på deras hjärnor (i gram), med syftet att studera hur sambandet mellan *kroppsvikt* och *hjärnans vikt* ser ut. Biologerna vill använda populationsmodellen nedan för att analysera datasetet.

$$\text{brainweight}_i = \beta_0 + \beta_1 \text{bodyweight}_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

Skattningen från regressionen hittar du i tabellen nedan.

- Ange dom fyra centrala antagandena för linjär regression. Baserat enbart på texten ovan så kan vi dra slutsatsen att i alla fall *ett* av dom fyra antagandena som den linjära regressionsmodellen bygger på antagligen är uppfyllt. Vilket? (5p)
- Testa på 5% signifikansnivå om **bodyweight** är en signifikant förklarande variabel. Du behöver inte ställa upp ett komplett test, det räcker att ange teststatistikans värde, det kritiska värdet, och din slutsats. (5p)
- Tolka regressionskoefficienten för **bodyweight**, alltså  $b_1$ . Skapa ett 90%-igt konfidensintervall och tolka intervallet på ett sätt som gör det tydligt att du förstår innebörden. (8p)

	Estimate	Std. Error
(Intercept)	191.2225651	110.08777308
bodyweight	0.9431658	0.07657678