# SDAI (ST1101), Tentamen 2, 6 hp

**Stockholms universitet, statistiska institutionen**

**Kurs: Statistik och dataanalys I, 15 hp**

**Tentamensdatum: 2024-04-27**

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.

Lösningsförslag läggs ut på athena efter att tentamenstiden är över.

**Lycka till!**

## Uppgift 1 (17 poäng)

Låt $A$ och $B$ vara två händelser med $P(A) = 0.4$, $P(B) = 0.5$ och $P(A \cap B) = 0.15$.

a) Beräkna den betingade sannolikheten för $A$ givet att $B$ har inträffat. (4p)

b) Är händelserna $A$ och $B$ oberoende? (4p)

c) Vad är sannolikheten att åtminstone en av $A$ och $B$ inträffar? (4p)

d) Använd en vanlig sexsidig tärning som exempel och förklara följande begrepp (5p)

- Utfallsrum
- Utfall
- Händelse

## Uppgift 2 (21 poäng)

Ramona älskar basket, och kastar boll varje lördag och söndag. Hennes sannolikhet att träffa är $p = 0.2$. Vi antar att hennes kast är oberoende av varandra.

a) Om Ramona gör 4 kast, vad är sannolikheten att hon träffar på första och sista kastet, men inte på dom två mittersta? (3p)

b) På lördagar kastar Ramona bollen 15 gånger. Vad är sannolikheten att Ramona träffar minst två gånger om hon gör 15 kast? (6p)

c) På söndagar så kastar Ramona bollen ett slumpmässigt antal gånger, där antalet kast följer en Poisson-fördelning med $\lambda = 15$. Vad är sannolikheten att Ramona kastar bollen exakt 14 gånger? (5p)

d) Ramonas mamma vill gärna uppmuntra henne och betalar därför Ramona 50 Öre (50 Öre = 0.5 kronor) för varje kast hon gör. Vad är väntevärdet och variansen för antalet kronor Ramona tjänar ihop per helg? (7p)

## Uppgift 3 (22 poäng)

Amerikanska fotbollsspelare i NFL tenderar att avsluta sina karriärer i ung ålder. Vi kallar åldern då en spelare avslutar sin karriär för *pensionsåldern*, och antar att denna ålder är normalfördelad med okänt väntevärde $\mu$ och okänd standardavvikelse $\sigma$. Pensionsåldern för 5 slumpmässigt utvalda spelare ges i tabellen nedan. Data är insamlat 2023.

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|
| 30.6 | 32.0 | 33.7 | 34.8 | 36.1 |

a) Skatta $\mu$ med en väntevärdesriktigt estimator. Förklara vad *väntevärdesriktig* betyder. (4p)

b) Beräkna ett konfidensintervall för $\mu$ med 95% konfidensnivå. (4p)

c) Vad menar vi när vi säger att intervallet täcker det sanna värdet av $\mu$ med 95% säkerhet? Varför är det inkorrekt att använda ordet *sannolikhet* istället för *säkerhet*? (4p)

d) Den genomsnittliga pensionsåldern för NFL-spelare på 1930-talet var 37 år. Genomför ett hypotestest för att testa om pensionsåldern för NFL-spelare har minskat baserat på datasetet i tabellen ovan. Använd signifikansnivå $\alpha = 0.05$. Ställ upp hypoteser, utför testet och dra korrekta slutsaser. Ange vilka antaganden som måste vara uppfyllda för att testet ska gälla. (10p)

## Uppgift 4 (24 poäng)

Ett företag använder en maskin som tillverkar vantar och vill säkerställa att kvaliteten på vantarna är hög. För att undersöka kvaliteten så väljer dom ut ett slumpmässigt stickprov på 200 vantar. Varje vante kontrolleras sedan för att se om något av fingrarna är trasiga. Tabellen nedan sammanställer antalet felfria vantar, och antalet vantar med olika många trasiga fingrar, bland dom 200 vantar som kontrollerades.

Tabellen visar även dom *förväntade* andelarna vantar med olika antal trasiga fingrar, baserat på statistik från sjuttiotalet.

| | Antal vantar i stickprovet | Förväntad andel |
|---|---|---|
| Felfria | 129 | 70 % |
| 1 trasigt finger | 17 | 6 % |
| 2 trasiga fingrar | 13 | 6 % |
| 3 trasiga fingrar | 10 | 6 % |
| 4 trasiga fingrar | 18 | 6 % |
| 5 trasiga fingrar | 13 | 6 % |

a) Låt $p$ vara andelen felfria vantar som maskinen producerar i genomsnitt, och $\hat{p}$ andelen i stickprovet. Normalapproximera samplingfördelningen för $\hat{p}$. Ange eventuella antaganden som krävs, och resonera om dessa är uppfyllda. (11p)

b) Enligt statistiken från sjuttiotalet så ska 70% av vantarna (i genomsnitt) vara felfria. Testa på 1% signifikansninvå om andelen felfria vantar skiljer sig från 70% genom att beräkna p-värdet för ett dubbelsidigt test. (5p)

c) Genomför ett chi2-test på 5% signifikansnivå för att undersöka om fördelningen i stickprovet matchar hur det såg ut på sjuttiotalet. Ange antaganden, och resonera om dessa är uppfyllda. (8p)

## Uppgift 5 (16 poäng)

Grönmusslor från Bohuslän är en delikatess som säljs i många restauranger i Sverige. En restaurangägare i Stockholm har märkt att kunderna tenderar att beställa grönmusslor oftare när de är stora. För att undersöka sambandet mellan musslornas ålder och vikt har restaurangägaren samlat in data från 10 musslor och genomfört en regressionsanalys med ålder (i dagar) som förklaringsvariabel och vikt (i gram) som responsvariabel. Den genomsnittliga åldern av dom 10 musslorna är 57.8 dagar.

a) Tolka det skattade värdet `ålder`. (3p)

b) En kund beställer en mussla som är 53 dagar gammal. Skapa ett 99% prediktionsintervall för musslans vikt. Du hittar en skattning av residualstandardavvikelsen under `Measures of model fit`. (6p)

c) Restaurangägaren påstår att prediktionsintervallet är ett intervall för det betingade väntevärdet av vikten givet åldern. Är detta påstående korrekt? Glöm inte att motivera ditt svar! (3p)

d) I enkel linjär regression så gör vi ett antagande om linjäritet. *Vad* är det som ska vara linjärt? Beskriv hur du kan undersöka om antagandet är uppfyllt. (4p)

```
Measures of model fit
-----------------------------------------------
Root MSE       R2   R2-adj
 8.64791  0.97852  0.97584

Parameter estimates
-----------------------------------------------
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)   8.9939   6.281907  1.4317 1.9011e-01
ålder         1.8673   0.097803 19.0921 5.8674e-08
```