

SDAI (ST1101), Tentamen 2, 6 hp

Stockholms universitet, statistiska institutionen

Kurs: Statistik och dataanalys I, 15 hp

Tentamensdatum: 2024-12-13

Skrivtid: kl. 08–13 (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 (17 poäng)

Låt A och B vara två händelser med $P(A) = 0.3$ och $P(B) = 0.25$ och $P(A \cap B) = 0.1$.

- a) Är händelserna oberoende? Är händelserna disjunkta? (6p)
- b) Beräkna $P(A | B)$. (4p)
- c) Vad är sannolikheten att exakt en av A och B inträffar? (4p)
- d) Förklara med ord vad A^c är, och beräkna $P(A^c)$. (3p)

Uppgift 2 (23 poäng)

25% av alla lussekatter som bageri Luci säljer är extra saftiga eftersom att dom innehåller ovanligt många russin. Dessa lussekatter är mycket eftertraktade bland kunderna, men det syns inte på utsidan hur många russin en lussekatt innehåller så det är bara att hålla tummarna när man beställer. Lussekatterna blandas noga innan försäljning, så att varje köp kan ses som oberoende med samma sannolikhet för en saftig lussebulle.

- a) Vad är sannolikheten att en kund behöver köpa 6 lussekatter innan hen får sin första extra saftiga lussekatt? (3p)
- b) Om du köper lussekatter tills du får en extra saftig, hur många lussekatter förväntar du dig att du behöver köpa? (2p)
- c) Om varje saftiga lussekatt innehåller 13 russin, och varje vanliga lussekatt 5 russin, vad är det förväntade totala antalet russin om du köper 8 lussekatter? (7p)
- d) Vad är sannolikheten att du får exakt 3 saftiga lussekatter om du köper 9 stycken? (4p)
- e) Din kusin köper ett slumpmässigt antal lussekatter, där antalet kan beskrivas med en Poissonfördelning med $\lambda = 2$. Vad är sannolikheten att din kusin köper mer än 2 lussekatter? (7p)

Uppgift 3 (22 poäng)

Alkoholkoncentrationen i en flaska av Nisses hemmagjorda glögg kan antas vara normalfördelad med okänt väntevärde μ och okänd standardavvikelse σ . För att få en bättre uppfattning om den förväntade alkoholkoncentrationen (μ) så tar Nisse ett stickprov på 6 flaskor. Alkoholkoncentrationen i dom olika flaskorna hittar du i Tabell 1.

X_1	X_2	X_3	X_4	X_5	X_6
17.6	14.0	16.7	18.8	15.1	16.2

- Skatta μ med en väntevärdesriktig estimator, och förklara vad ordet väntevärdesriktig betyder. (4p)
- Beräkna ett 95%-igt konfidensintervall för μ baserat på dom 6 mätningarna. Tolka konfidensintervallet på ett sätt som visar att du förstår *innebörden*. (7p)
- Genomför ett enkelsidigt hypotestest för att testa om μ är *större* än 15. Använd signifikansnivå $\alpha = 0.05$. Ställ upp hypoteser, beräkna teststatistikan, ange fördelningen under H_0 och utför testet. (7p)
- Nämna minst ett antagande som krävs för att testet ska vara giltigt som *inte* nämns i texten ovan. Resonera om du tror att detta antagande är uppfyllt eller inte. (4p)

Uppgift 4 (19 poäng)

Ett företag som genomför opinionsundersökningar har frågat ett slumpmässigt urval på 300 personer om dom tror att tomten finns på riktigt. Av dom tillfrågade var det 47 som svarade ja, och resterande 253 svarade nej.

- Låt p vara andelen i populationen som tror att tomten finns på riktigt. Beräkna ett approximativt 99% konfidensintervall för p . Vilka antaganden behöver vara uppfyllda för att intervallet ska vara giltigt? Är dessa antaganden uppfyllda? (7p)
- Bland dom som tillfrågades bodde 124 norr om polcirkeln och resterande 176 söder om polcirkeln (se tabell nedan). Använd ett chi2-test för att avgöra om inställningen till tomtens existens är oberoende av var personen bor. Ställ upp hypoteser, utför testet och dra korrekta slutsatser. Använd signifikansnivå $\alpha = 0.05$. (9p)
- Nämnd minst två av dom tre antagandena som krävs för att testet i 4b ska vara giltigt. Är antagandena uppfyllda? (3p)

	Existerar	Existerar ej	Total
Norr om polcirkeln	30	94	124
Söder om polcirkeln	17	159	176
Total	47	253	300

Uppgift 5 (19 poäng)

En forskare har låtit 100 åttondeklassare i Nederländerna genomföra ett IQ-test som testar elevernas verbala IQ (IQ), och sen jämfört resultatet dom fick på IQ-testet med deras resultat på ett språkprov (`language_test`). Medelvärdet på IQ-testet bland eleverna var 105.

Forskaren tror att det finns ett linjärt samband mellan resultaten på dom två proven, och vill använda populationsmodellen nedan.

$$\text{language_test} = \beta_0 + \beta_1 \text{IQ} + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

Skattningen från regressionen hittar du i tabellen nedan.

- Tolka den skattade regressionskoefficienten för variabeln IQ. (3p)
- Studenten Arjan har ännu inte genomfört språktestet, men fick ett resultat på 87 på IQ-testet. Skapa ett nittiofemprocentigt prediktionsintervall för Arjans poäng på språkprovet. (Arjan var *inte* en av dom 100 eleverna vars resultat på dom två testen användes för att estimeras/skatta regressionsmodellen.) (7p)
- Förklara kortfattat skillnaden mellan ett prediktionsintervall och ett konfidensintervall i enkel linjär regression. (4p)
- Förklara vad *feltermerna* och *residualerna* är. På vilket sätt skiljer dom sig åt? (5p)

Measures of model fit

```
-----  
Root MSE      R2    R2-adj  
8.27452  0.33153  0.32471
```

Parameter estimates

```
-----  
                Estimate Std. Error t value  Pr(>|t|)  
(Intercept)    4.5500    4.25108  1.0703  2.8710e-01  
IQ              2.6979    0.38699  6.9716  3.6785e-10
```