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

Lösningsförslag - Uppgift 1

1a (4p)

Den betingade sannolikheten för händelsen A givet att B har inträffat betecknas $P(B \mid A)$ och beräknas med formeln

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{0.15}{0.4} = 0.375$$

1b (4p)

Det finns olika sätt att lösa denna deluppgift. Den lättaste lösning är att poängtera att $P(B \mid A) \neq P(B)$, alltså är A och B inte oberoende. En annan lösning är att visa att $P(A \cap B) \neq P(A) \cdot P(B)$.

1c (4p)

Sannolikheten att åtminstone en av A och B inträffar är $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.4 + 0.5 - 0.15 = 0.75$.

1d (4p)

Utfallsrummet är vad vi kallar mängden av alla möjliga utfall, för en vanlig tärning är det värdena $S = \{1, 2, 3, 4, 5, 6\}$.

Ett utfall är ett *specifikt* värde i utfallsrummet, så tex är 3 ett utfall.

En händelse är en samling av utfall. Ett exempel på en händelse när vi slår en tärning är "udda antal prickar" (den händelsen består av utfallen $\{1, 3, 5\}$). Varje utfall är också en händelse! Exempelvis är 3 både ett utfall och en händelse. Men motsatt riktning är generellt inte sant: "udda antal prickar" är en händelse, men inte ett utfall.

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)

Lösningsförslag - Uppgift 2

2a (3p)

$$0.2 \cdot 0.8 \cdot 0.8 \cdot 0.2 = 0.0256$$

2b (6p)

Sannolikheten för minst två träffar beräknar vi som $1 - P(\textit{ingen träff}) - P(\textit{en träff})$

$$P(\textit{ingen träff}) = 0.8^{15} = 0.03518437$$

$$P(\textit{en träff}) = {}_{15}C_1 \cdot 0.8^{14} \cdot 0.2^1 = 0.1319414$$

Sannolikheten för minst två träffar blir alltså $1 - 0.1319414 - 0.03518437 = 0.8328742$. (Färre decimaler är OK!)

2c (5p)

$$\frac{e^{-15}15^{14}}{14!} = 0.1024359$$

Sannolikheten att Ramona gör exakt 14 kast är alltså ungefär 0.11.

2e (7p)

Vi vet att Ramona gör 15 kast på lördag, vilket motsvarar 7.5 kronor. Det totala antalet kronor hon får (vilket vi kan kalla Y) kommer alltså bli en linjärkombination av hur många kast hon gör på söndag, enligt formeln

$$Y = 7.5 + 0.5X$$

där X är antalet kast på söndag. Väntevärdet av en linjärkombination vet vi är

$$\mathrm{E}(Y) = \mathrm{E}(7.5 + 0.5X) = 7.5 + 0.5\,\mathrm{E}(X) = 7.5 + 0.5 \cdot 14$$

där den sista likheten följer av att väntevärdet för en Poisson med $\lambda = 15$ är 15. Ramona förväntas alltså tjäna ihop 15 kronor per helg.

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 μ och okänd standardavvikelse σ. 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 μ med en väntevärdesriktigt estimator. Förklara vad *väntevärdesriktig* betyder. (4p)

b) Beräkna ett konfidensintervall för μ med 95% konfidensnivå. (4p)

c) Vad menar vi när vi säger att intervallet täcker det sanna värdet av μ 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 slutsatser. Ange vilka antaganden som måste vara uppfyllda för att testet ska gälla. (10p)

Lösningsförslag - Uppgift 3

3a (4p)

För att skatta μ med en väntevärdesriktig estimator kan vi använda medelvärdet av observationerna, dvs $\bar{x} = \frac{1}{n}\sum_{i=5}^{5} x_i = \frac{30.6+32.0+33.7+34.8+36.1}{5} = 33.44$. En väntevärdesriktig estimator är en estimator vars väntevärde är lika med det sanna värdet av parametern som ska skattas. Detta betyder att om vi skulle upprepa experimentet många gånger och beräkna medelvärdet varje gång, så skulle medelvärdet av dessa medelvärden närma sig det sanna värdet av parametern. En väntevärdesriktig skattare är alltså en skattare som inte är för hög eller för låg i genomsnitt.

3b (4p)

Ett konfidensintervall för μ med 95% konfidensnivå ges av

$$\bar{x} \pm t_{\alpha/2,n-1}\frac{s_x}{n}$$

Innan vi kan beräkna intervallet så behöver vi beräkna standardavvikelsen s_x och $t_{\alpha/2,n-1}$. Stickprovsmedelvärdet har vi redan beräknat i a-delen, och det är $\bar{x} = 33.44$. Standardavvikelsen beräknas som

$$\begin{aligned} s_x &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \sqrt{\frac{1}{4}((30.6-33.44)^2 + (32.0-33.44)^2 + (33.7-33.44)^2 + (34.8-33.44)^2 + (36.1-33.44)^2)} \\ &= 2.19 \end{aligned}$$

3c (4p)

När vi säger att intervallet täcker det sanna värdet av μ med 95% säkerhet menar vi att om vi upprepar experimentet många gånger och beräknar ett nytt konfidensintervall varje gång, så kommer 95% av dessa intervall att innehålla det sanna värdet av μ. Det är inkorrekt att använda ordet *sannolikhet* eftersom det sanna värdet av μ är en okänd konstant som antingen innefattas i intervallet eller inte. Det är alltså inte en slumpvariabel som vi kan tala om sannolikhet för.

3d (10p)

Hypoteser:

$$H_0 : \mu = 37$$
$$H_A : \mu < 37$$

Teststatistika

$$T = \frac{\bar{X} - 37}{s_x/\sqrt{n}}$$

följer en t-fördelning med 4 frihetsgrader. Vi har ett enkelsidigt test och vill använda $\alpha = 0.05$ så vi får ett kritiskt värde på $t_{0.05,4} = -2.132$. Anledningen att det blir minus är att vi vill avgöra om μ har minskat, och vi kommer förkasta nollhypotesen att μ *inte* har minskat om det observerade värdet är mycket mindre än 37. Alltså, vi kommer förkasta nollhypotesen om vi får en stor negativ teststatistika, så vi bryr oss bara om vänstersvansen. Vid ensidiga test bryr vi oss bara om avvikelser i alternativhypotesens riktning. Vi har

$$t_{obs} = \frac{33.44 - 37}{2.19/\sqrt{5}} = -3.63$$

Eftersom $-3.63 < t_{crit} = -2.132$ så förkastar vi nollhypotesen. Vi kan alltså dra slutsatsen att pensionsåldern för NFL-spelare har minskat.

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)

Lösningsförslag - Uppgift 4

4a (11p)

För att normalapproximera samplingfördelning för $\hat{p}$ behöver vi beräkna väntevärdet och standardavvikelsen. Vi skattar andelen som $\hat{p} = 129/200 = 0.645$ och använder normalapproximeringen

$$\hat{p} \overset{approx}{\sim} N\left(p, \sqrt{\frac{0.645(1-0.645)}{200}}\right) = \mathrm{N}(p, 0.033836)$$

För att normalapproximationen ska vara giltig gäller att

1. Svaren från personerna är oberoende av varandra. Detta antagande är uppfyllt, eftersom att det i uppgiften står att stickprovet är slumpmässigt.
2. Stickprovet utgör mindre än 10% av populationen. Populationen i detta fall är alla vantar som maskinen producerar, och vi antar att den är mycket större än 200.
3. np och nq är båda större eller lika med 10. I detta fall är $np = 129$ och $nq = 71$, så antagandet är uppfyllt.

4b (5p)

För att beräkna p-värdet behöver vi beräkna teststatistikan

$$Z = \frac{\hat{p} - 0.7}{\sqrt{\frac{0.7(1-0.7)}{200}}} = \frac{0.645 - 0.7}{\sqrt{\frac{0.7(1-0.7)}{200}}} = \frac{-0.055}{0.0324037} \approx -1.70$$

eftersom att det är ett dubbelsidigt test så blir p-värdet $2 \cdot P(Z < -1.70) = 2 \cdot 0.0446 = 0.0892$. Eftersom p-värdet är större än signifikansnivån (0.01) så förkastar vi inte nollhypotesen, vi har alltså inte hittat något stöd för att andelen felfria vantar skiljer sig från 70%.

4c (8p)

För att genomföra ett chi2-test behöver vi först beräkna förväntade frekvenser. Dessa beräknas genom att multiplicera den förväntade andelen med antalet vantar i stickprovet. Därefter beräknar vi teststatistikan.

	Antal vantar (O_i)	Förväntad andel	Förväntad frekvens (E_i)	$(O_i - E_i)^2/E_i$
Felfria	129	70 %	140	0.8642857
1 trasigt finger	17	6 %	12	2.083333
2 trasiga fingrar	13	6 %	12	0.08333333
3 trasiga fingrar	10	6 %	12	0.3333333
4 trasiga fingrar	18	6 %	12	3
5 trasiga fingrar	13	6 %	12	0.08333333

Summan blir $0.8642857 + 2.083333 + 0.08333333 + 0.3333333 + 3 + 0.08333333 = 6.447619$. Teststatistikan blir alltså 6.447619. Vi har $6 - 1 = 5$ frihetsgrader, eftersom att vi har 6 celler/bins. Kritiskt värde för χ^2 med 5 frihetsgrader och $\alpha = 0.05$ är 11.070. Eftersom $6.448619 < 11.070$ så förkastar vi inte nollhypotesen, och vi har alltså inte hittat något stöd för att fördelningen i stickprovet skiljer sig från hur det såg ut på sjuttiotalet.

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
```

Lösningsförslag - Uppgift 5

5a (3p)

Musslorna förväntas växa med knappt två gram per dag.

5b (6p)

Prediktionsintervallet för en ny mussla som är 53 dagar gammal ges av

$$\widehat{y}_* \pm t_{\alpha/2,n-2} \cdot \sqrt{\frac{s_e^2}{n} + s_{b_1}^2(x_* - \bar{x})^2 + s_e^2}$$

För att beräkna detta intervall behöver vi ta fram ett antal olika värden.

- $\widehat{y}_* = 8.9939 + 1.8673 \cdot 53 = 107.9608$
- $s_e = 8.64791$
- $s_{b_1} = 0.097803$
- $t_{0.005,10-2} = 3.355$

Detta ger intervallet

$$107.9608 \pm 3.355 \cdot \sqrt{8.64791^2/10 + 0.097803^2(53 - 57.8)^2 + 8.64791^2} \approx 107.96 \pm 30.47 = (77.49, 138.43)$$

5c (3p)

Detta påstående är inte korrekt. Det restaurangägaren beskriver är ett konfidensintervall, inte ett prediktionsintervall. Ett prediktionsintervall tar hänsyn till både osäkerheten i skattningen av parametrarna och osäkerheten i den framtida observationen. Ett konfidensintervall tar endast hänsyn till osäkerheten i skattningen av parametrarna.

5d (4p)

I enkel linjär regression ska sambandet mellan förklaringsvariabeln och responsvariabeln vara linjärt. I detta fall är det ålder som ska vara linjärt relaterat till vikten. Vi kan undersöka detta genom att plotta vikten mot åldern och se om det ser linjärt ut, och inte exempelvis som en banan.