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.

Lösningsförslag 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 $P(A \mid B)$. (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)

Lösningsförslag - Uppgift 1

1a (4p)

Den betingade sannolikheten för en viss händelse A givet en annan händelse B är sannolikheten att A inträffar givet att vi *vet* att B har inträffat. Eftersom att A och B är oberoende gäller att $P(A \mid B) = P(A)$.

1b (4p)

Sannolikheten att både A och B inträffar ges av produkten av deras sannolikheter, eftersom att dom är oberoende.

$$P(A \cap B) = P(A) \cdot P(B) = 0.3 \cdot 0.6 = 0.18$$

1c (4p)

För att ta fram sannolikheten att någon av A och B inträffar så tar vi summan av sannolikheterna *minus* sannolikheten att båda inträffar, så att vi inte räknar sannolikheten när båda händer dubbelt.

$$P(A \cup B) = 0.3 + 0.6 - 0.18 = 0.72$$

1d (4p)

En händelse är en kombination/mängd av utfall. Exempelvis, om vi slår en sexsidig tärning är dom olika utfallen 1, 2, 3, 4, 5 eller 6. En *händelse* däremot kan vara att få ett udda antal ögon på tärningen, eller att slå minst 5. Alla utfall är händelser, men inte alla händelser är utfall!

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)

Lösningsförslag - Uppgift 2

2a (3p)

Vi behöver köpa fyra nitlotter, sen en vinstlott. Sannolikheten för detta är

$$0.18 \cdot 0.82^3 \approx 0.10$$

2b (4p)

Antalet lotter inklusive första vinst följer en Geometrisk fördelning med $p = 0.18$.

$$0.82^6 \cdot 0.18 \approx 0.055$$

2c (6p)

Färre än två betyder 0 eller 1. Antalet lyckade av 10 försök följer en binomialfördelning med $n = 10$, $p = 0.18$.

`pbinom(1, 10, 0.18) = 0.44`

2d (6p)

Väntevärdet för binomialfördelningen ges av np. Vi vill att denna ska vara minst 8, så vi sätter lika med åtta och hittar värdet på n.

$$n \cdot 0.18 = 8$$

$$n = 8/0.18 = 44.4444$$

Eftersom att det endast går att ringa ett heltal antal gånger, och vi vill att vätevärdet ska bli *minst* 8 så väljer vi $n = 45$ vilket ger ett väntevärde på 8.1.

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 (μ) och standardavvikelsen (σ) 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 sjuttiotalet. Detta stickprov hade en standardavvikelse på $s = 156$. Genomför ett nytt hypotestest som undersöker om den genomsnittliga vikten skiljer sig åt mellan idag och sjuttiotalet. 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)

Lösningsförslag - Uppgift 3

3a (9p)

$$H_0 : \mu = 2300$$
$$H_1 : \mu \neq 2300$$

Teststatistika

$$T = \frac{\bar{X} - 2300}{s_x/\sqrt{n}}$$

följer en t-fördelning med 7 frihetsgrader. Vi har

$$t_{obs} = \frac{1976 - 2300}{263/\sqrt{8}} \approx 3.48$$

Eftersom $|t_{obs}| < t_{0.005,7} = 3.499$ så förkastar vi inte nollhypotesen. Vi kan inte förkasta att den genomsnittliga vikten på gula fjällrävar är oförändrad.

För att testet ska vara giltigt så behöver populationen vara normalfördelad. Eftersom att vi endast har 8 observationer så kan vi inte lita på att centrala gränsvärdessatsen ger oss en normalfördelning. Observationerna behöver också vara oberoende.

3b (12p)

Innan vi kan genomföra hypotestestet så behöver vi beräkna antalet frihetsgrader. Detta ges av

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} \approx 11.11$$

Vilket ger oss 11.11 frihetsgrader.

Hypoteser:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Teststatistika

$$T = \frac{\bar{X}_1 - \bar{X}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

följer en t-fördelning med 11.11 frihetsgrader. Vi har

$$t_{obs} = \frac{1976 - 2300}{\sqrt{\frac{263^2}{8} + \frac{156^2}{9}}} \approx 3.04$$

Eftersom $|t_{obs}| < t_{0.005,11} = 3.106$ så förkastar vi inte nollhypotesen. Inte heller när vi tar hänsyn till osäkerheten i värdet 2300 så kan vi dra slutsatsen att det skett en förändring i den genomsnittliga vikten hos gula fjällrävar sen sjuttiotalet.

3c (4p)

Nej, detta kan inte hända. Anledningen till att vi skulle kunna få ett signfikant svar i a och inte b är att vi jämför med ett värde 2300 som i själv verket är osäkert och alltså *skulle* kunna vara större.

Det går också bra att resonera utifrån formlerna. Vi kan se att nämnaren i testet som jämför två oberoende grupper alltid är större än motsvarande för ett "vanligt" hypotesttest, så test-statistikans värde kommer vara minde. Vi har även lägre frihetsgrader vilket innebär att vårat kritiska värde kommer vara större och vi kommer alltså vara mindre benägna att förkasta.

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

a) 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)

b) Nämn 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)

c) 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)

d) 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)

Lösningsförslag - Uppgift 4

4a (9p)

Vi börjar med våra hypoteser.

$$H_0 : \text{Räknedata följer fördelningen i den andra kolumnen i tabellen.}$$

$$H_A : \text{Räknedata följer inte fördelningen i den andra kolumnen i tabellen.}$$

Vi lägger till två ytterligare kolumner till tabellen. I den första beräknar vi det förväntade antalet om fördelningen från företaget stämmer, och i den andra $(obs - exp)^2/exp$.

Vinst	Sannolikhet	Observerat antal	Förväntat antal	$(Obs - Exp)^2/Exp$
150	0.02	13	10	0.9
20	0.15	69	75	0.48
10	0.3	164	150	1.31
0	0.52	254	260	0.38

Vi kan nu beräkna vår teststatistika, vilken i formelsamlingen ges av

$$\chi^2 = \frac{\sum(Exp - Obs)^2}{Exp}$$

genom att summera den sista kolumnen. Summan av dessa värden blir

$$\chi^2_{obs} = 2.35$$

Det kritiska värdet ges av en χ^2-fördelning med 3 frihetsgrader, eftersom att vi har fyra olika celler. Då vi ska använda $\alpha = 0.05$ så får vi som kritiskt värde

$$\chi^2_{crit} = 7.815.$$

Eftersom att $\chi^2_{obs} < \chi^2_{crit}$ så kan vi *inte* förkasta nollhypotesen.

4b (3)

- Det förväntade antalet observationer i varje cell behöver vara minst 5. Detta antagande är uppfyllt.
- Observationerna behöver vara oberoende. Vi saknar underlag för att svara på om detta antagande är uppfyllt.
- Vi antar att vi har *räknedata*. (Detta är uppfyllt…)

4c (9p)

X är en diskret slumpvariabel (den antar endast fyra olika värden).

Väntevärde beräknas med formeln

$$E(X) = \sum p(x) \cdot x = (-10) \cdot 0.52 + 0 \cdot 0.3 + 10 \cdot 0.15 + 140 \cdot 0.02 = -0.9.$$

Alltså, i genomsnitt förväntar vi oss att gå back 0.9 USD varje gång vi spelar spelet.

För att beräkna variansen behöver vi först beräkna $E(X^2)$

$$E(X) = \sum p(x) \cdot x = (-10)^2 \cdot 0.52 + 0^2 \cdot 0.3 + 10^2 \cdot 0.15 + 140^2 \cdot 0.02 = 459.$$

Variansen blir alltså $459 - (-0.9)^2 = 458.19$.

4d (4p)

Vi håller med Björn baserat på *centrala gränsvärdessatsen* som säger att summan av ett antal oberoende slumpvariabler med ändligt väntevärde kommer bli approximativt normalfördelad. Bra tänkt Björn!

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.

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

Skattningen från regressionen hittar du i tabellen nedan.

a) 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)

b) 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)

c) 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
```

Lösningsförslag - Uppgift 5

5a (5p)

Dom fyra antaganden vi gör är

- linjärt samband mellan responsvaraibeln och förklaringsvariabeln
- normala feltermer
- homoskedasticitet (konstant varians) för feltermerna
- oberoende feltermer

Eftersom att det tydligt framgår att det är ett slumpmässigt urval så är det rimligt att tro att antagandet om *oberoende* är uppfyllt.

5b (5p)

För att testa om variabeln är signifikant förklarande så ska vi testa nollhypotesen $\beta_1 = 0$. Vi har redan s_{b_1} från utskriften, och får att vår teststatistika har värdet

$$T_{obs} = \frac{0.9431 - 0}{0.0766} = 12.31$$

Nästa steg är att ta fram det kritiska värdet, vilket är

$$t_{0.025,23} = 2.069$$

eftersom att vi har angivet att $n = 25$.

Vi kan nu jämföra och ser att eftersom att $T_obs = 12.31 > t_{crit} = 2.069$ så kan vi förkasta nollhypotesen. Kroppsvikten är en signifikant förklarande variabel för hjärnvikten.

5c (8p)

Den genomsnittliga hjärnvikten förväntas öka ett gram för varje kilo som den totala kroppsvikten ökar.

För att skapa ett kofidensintervall så behöver vi det kritiska värdet, vilket är $t_{0.05,23} = 1.714$. Intervallet blir alltså

$$0.943 \pm 1.714 \cdot 0.0766 = (0.811, 1.074)$$

Om vi samlade in fler urval på 25 djur av olika arter så förväntar vi oss att intervall vi skapar på samma sätt om ovan kommer innehålla det sanna värdet på β_1 nio av tio gånger.