Sensorveiledning utsatt eksamen V23 ECON2130

Informasjon om oppgavesettet

Hver oppgave består av flere deloppgaver. Noen av deloppgavene bygger på hverandre. Hvis du ikke får løst en deloppgave, men trenger informasjon derfra for å komme videre, kan du gjøre en antakelse om den informasjonen du mangler og eksplisitt bruke denne antakelsen ved behov.

Oppgave 1 – 15%

Hver deloppgave har lik vekt.

Yatzy er et spill der man får poeng for ulike terningskombinasjoner. Du spiller Yatzy med fem terninger, hvilket vil si at du kaster de fem terningene samtidig.

a)

Hva er sannsynligheten for at du får nøyaktig tre seksere?

$$\Pr(X=3) = \binom{5}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2 = \frac{5!}{3!\,2!}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2 = 10 * \frac{1}{6*6*6} * \frac{5*5}{6*6} \approx 0.032$$

b)

Hva er sannsynligheten for at du får nøyaktig tre seksere, gitt at én av terningene viser en sekser?

Tilsvarer sannsynligheten for nøyaktig to seksere i løpet av fire forsøk.

$$P(X=2) = \binom{4}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^2 = \frac{4!}{2!\,2!}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^2 = 6 * \frac{1}{6*6} * \frac{5*5}{6*6} \approx 0.116$$

c)

Forklar hvordan du kunne brukt R til å finne den tilnærmede sannsynligheten for at summen av de fem terningene blir 17, gitt at summen er minst 10.

Kurset har ikke fokusert på hva som er mest mulig effektiv kode, og alle fornuftige framgangsmåter bør få uttelling. God forklaring bør belønnes. Én mulig løsning:

```
terning <- 1:6
## Simulér først 5 terningkast mange ganger:
sim <- replicate(10000, sum(sample(terning, 5, replace=TRUE)))
## Ta bort de som ikke summerer seg til minst 10 vha. firkantklamme
mulige <- sim[sim>=10]
## Tell deretter opp de som har sum nøyaktig lik 17
gunstige <- mulige[mulige==17]
## Sjekk til sist hvor stor andel dette utgjør av de med minst 10:
length(gunstige)/length(mulige)

## Kan også løses direkte:
mean(sim[sim>=10]==17)
```

Oppgave 2 – 25%

Hver deloppgave har lik vekt.

a)

Forklar hva følgende R-kode gjør:

```
> x <- rnorm(100, 2, 4)

> mean(x)
[1] 2.223409
```

Første linje trekker 100 observasjoner fra N(2, 16). Linje to finner gjennomsnittet i utvalget, som er på rundt 2.22.

b)

Konstruer et 95% konfidensintervall for μ med utgangspunkt i informasjonen fra deloppgave a. Gi en kort forklaring av framgangsmåten din.

Siden utvalget er trukket fra $N(2,16)$ vet vi at populasjonsstandardavviket er kjent og lik $\sqrt{16} = 4$. Vi kan derfor bruke kritiske verdier fra normalfordelingstabellen.

$$\bar{X} \pm 1.96 * \sigma/\sqrt{n}$$

$$2.22 \pm 1.96 * 4/10$$

$$[1.436\,, 3.004]$$

c)

Dekker konfidensintervallet ditt μ? Svaret må begrunnes for å gi uttelling.

Ja – dataene er trukket fra en fordeling med $\mu = 2$, og vårt KI dekker 2.

Oppgave 3 – 30%

Hver deloppgave har lik vekt.

Vi har tre stokastiske variabler $A \sim N(0,1)$, $B \sim N(2,3)$ og $C = 2A + B$.

a)

Vi antar at A og B er uavhengige av hverandre. Forklar kort hva denne antakelsen innebærer. Finn forventningen og variansen til A, B og C.

De to første kan leses av direkte:

$$E(A) = 0, Var(A) = 1$$

$$E(B) = 2, Var(B) = 3$$

For C bruker vi regneregler for forventning og varians:

$$E(C) = E(2A + B) = E(2A) + E(B) = 2E(A) + E(B) = 2 * 0 + 2 = 2$$

$$Var(C) = Var(2A + B) = 2^2 Var(A) + Var(B) + 2Cov(A, B)$$

Uavhengighet betyr at vi kan anta $Cov(A, B) = 0$

$$= 4 * 1 + 3 + 0 = 7$$

b)

Hva er $P(C > 3)$? Forklar framgangsmåten din.

Summen av to normalfordelte variabler (som A og B er) er selv normalfordelt, så vi vet at C følger en normalfordeling, og videre vet vi fra oppgave a at forventningen er 2 og variansen 7, så $C \sim N(2,7)$.

$$P(C > 3) = P\left(\frac{C-2}{\sqrt{7}} > \frac{3-2}{\sqrt{7}}\right) = P\left(Z > \frac{1}{\sqrt{7}}\right) \approx P(Z > 0.38) = 1 - P(Z \leq 0.38) = 1 - 0.6480 = 0.352$$

c)

Det finnes en fjerde variabel D. Vi vet at denne er normalfordelt. Videre får vi opplyst at det er trukket et utvalg på 25 observasjoner fra fordelingen denne variabelen følger, og at dette utvalget har gjennomsnitt 3 og standardavvik 2. Sett opp hypoteser og test nullhypotesen om at D har forventning større eller lik 4. Bruk signifikansnivå 5%.

$$H_0: \mu_D \geq 4$$

$$H_A: \mu_D < 4$$

Vi kjenner ikke populasjonsstandardavviket σ, men får oppgitt utvalgsstandardavviket $S_D = 2$. Siden vi vet at D er normalfordelt vet vi at $\frac{\bar{X}_D - \mu_D}{S_D}$ følger en t-fordeling med $n-1$ frihetsgrader, her 24. Med 5% signifikansnivå og ensidig test får vi kritisk verdi på -1.711.

$$T = \frac{3-4}{2} = \frac{-1}{2}$$

Forkastningsområdet er alle verdier under -1.711. Vi er godt over dette og må beholde nullhypotesen.

Oppgave 4 – 15%

Hver deloppgave har lik vekt.

Du har inntektsdata for to utvalg av tidligere samfunnsøkonomistudenter. Nedenfor følger litt beskrivende statistikk.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 100000  300000  400000  400000  400000 1000000
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 100000  300000  400000  600000 1000000 1000000
```

a)

Gi en kortfattet forklaring av hvordan de to utvalgene skiller seg fra hverandre. Du trenger ikke regne på noe.

Vi ser at de to utvalgene har samme minimums- og maksimumsverdi, og at fordelingen ser lik ut opp til medianen. Fra median og oppover ser vi at utvalg 1 fortsetter med mange observasjoner av 400000, siden tredje kvartil fremdeles er 400000. Det vil si at når vi rangerer utvalget fra laveste til høyeste verdi og deler utvalget i fire like store deler er 400000 både den verdien som skiller gruppe 2 fra 3, og gruppe 3 fra 4, så hele den tredje gruppa må være observasjoner av 400000. Derimot ser vi at utvalg nummer to må ha mange verdier på 1000000 fra median og oppover, ettersom 1 million både er verdien som skiller gruppe 3 fra 4, og maksimalverdien i utvalget. Hele den fjerde gruppa i utvalg 2 må derfor bestå av observasjoner på 1 million.

Denne forskjellen reflekteres også i at gjennomsnittet i utvalg 2 er betydelig høyere enn i utvalg 1.

b)

De to utvalgene har ulike gjennomsnitt. Forklar hva det vil si at gjennomsnittsestimatoren er en forventningsrett estimator for forventningen μ. Kan vi si noe om forventningen til populasjonen(e) de to gruppene er trukket fra basert på utskriften over?

At gjennomsnittsestimatoren er forventningsrett vil si at vi i snitt vil treffe på forventningen dersom vi trekker uendelig mange utvalg og tar gjennomsnittene av disse. Denne egenskapen sier imidlertid ikke noe om hvor langt unna vi treffer hver gang – estimeringsfeilen kan være betydelig. Vi kan derfor ikke uten videre avgjøre om de to utvalgene er trukket fra populasjoner som har samme forventning eller ei. Da måtte vi evt. gjennomført en hypotesetest som kan si noe om hvorvidt det er sannsynlig.

Oppgave 5 – 15%

Hver deloppgave har lik vekt.

En undersøkelse finner at gjennomsnittsinntekten i en gruppe kvinner med universitetsutdannelse er 560 000 kr, mens den i en gruppe av menn med universitetsutdannelse er 570 000.

a)

Forklar hvilken informasjon du trenger for å kunne vurdere om gjennomsnittsestimatoren er normalfordelt.

Enten må vi vite at populasjonene følger en normalfordeling, eller vi må vite at utvalgene er så store at vi kan trekke på sentralgrenseteoremet, som sier at gjennomsnitt av utvalg går mot normalfordeling når utvalgsstørrelsen går mot uendelig.

b)

Vi antar at de to gjennomsnittene er normalfordelte og uavhengige av hverandre. En medstudent har satt opp følgende testobservator:

$$Z = \frac{(570000 - 560000)}{\sqrt{\frac{10000^2}{100} + \frac{12000^2}{120}}} \approx 6.742$$

Sett opp en nullhypotese som denne testobservatoren er egnet til å kaste lys over. Forklar hva du må anta for å kunne bruke kritiske verdier fra normalfordelingstabellen, og ta stilling til om nullhypotesen kan forkastes på 1% signifikansnivå eller ikke.

Her ser vi at det er testet hvorvidt $\hat{\theta} = \bar{X}_M - \bar{X}_K$ er signifikant ulik 0, som passer til hypotesesettet

$$H_0 : \hat{\theta} = \bar{X}_M - \bar{X}_K = 0$$

$$H_A : \hat{\theta} = \bar{X}_M - \bar{X}_K \neq 0$$

For å kunne bruke kritiske verdier fra normalfordelingstabellen må vi anta at hhv. 10000 og 12000 representerer populasjonsstandardavvik, ikke utvalgsstandardavvik. (Imidlertid er utvalgene så store at den praktiske forskjellen på å bruke t- og normalfordelingstabell vil være minimal.) Med 1% signifikansnivå er kritisk verdi ca. 2.58, så vi kan forkaste nullhypotesen med solid margin.