

# Skriftlig innlevering 4, Oppgave 1 (av 4)

## Oppgave 1 \*

Ola vil måle tyngdeakselerasjonen  $g$  ved å slippe en stein utfor et høyt loddrett stup. Han holder steinen i ro før han slipper den og lar hans gode venn Petter måle hvor lang tid det tar fra han slipper steinen til den treffer bakken. Han antar at steinen er tilstrekkelig tung til at effekten av luftmotstand er neglisjerbar. Han lar  $s$  betegne høyden på stupet og lar  $t$  betegne tiden steinen bruker på fallet. Fra fysikkundervisningen på videregående skole husker han at sammenhengen mellom  $s$ ,  $g$  og  $t$  er gitt som

$$s = \frac{1}{2}gt^2.$$

Dersom han og Petter greier å måle  $s$  og  $t$  uten feil vil man dermed ha at

$$g = \frac{2s}{t^2}.$$

Ola og Petter går nå på NTNU og vet at det er urealistisk å anta at de greier å måle  $s$  og  $t$  uten feil. De velger derfor å betrakte verdiene de måler for disse to størrelsene som realisasjoner av stokastiske variabler, og betegner disse stokastiske variablene med henholdsvis  $S$  og  $T$ . Den tilhørende utregnede tyngdeakselerasjonen blir dermed en realisasjon av den stokastiske variabelen

$$G = \frac{2S}{T^2}.$$

Ut fra hvordan målingene av  $s$  og  $t$  gjøres anser de to gode vennene det naturlig å anta at  $S$  og  $T$  er uavhengige stokastiske variabler. De lar dessuten  $\mu_s$  og  $\mu_t$  betegne de sanne (men ukjente) verdiene for henholdsvis  $s$  og  $t$ , og antar at  $E[S] = \mu_s$  og  $E[T] = \mu_t$ . Dessuten innfører de notasjonen  $\sigma_s^2 = \text{Var}[S]$  og  $\sigma_t^2 = \text{Var}[T]$ .

Ola og Petter målte  $s$  til å være 241.3 (meter), og  $t$  til å være 7.02 (sekunder).

## Deloppgave a)

Bruk feilforplantningsformelen til å finne formler for tilnærmede verdier for  $\mu_g = E[G]$  og  $\sigma_g^2 = \text{Var}[G]$ . Du skal oppgi svarene som funksjon av  $\mu_s$ ,  $\mu_t$ ,  $\sigma_s^2$  og  $\sigma_t^2$ .

Ola og Petter er spesielt interessert i usikkerheten  $\sigma_g^2$ . Ut fra måten de gjør målingene av  $s$  og  $t$  på finner de det rimelig å sette  $\sigma_s = 2$  (meter) og  $\sigma_t = 1$  (sekund).

Dessuten approksimerer de ved å sette  $\mu_s$  lik den målte verdien for  $s$  og  $\mu_t$  lik den målte verdien for  $t$ . Hvilken numeriske verdier får de da for den approksimerte variansen  $\sigma_g^2$  og for tilhørende approksimert standardavvik  $\sigma_g$ ?

## Her er deloppgave a) slutt

Ola og Petter har akkurat lært at stokastisk simulering også kan benyttes for å approksimere  $\mu_g$  og  $\sigma_g$ , og ønsker å bruke dette til å finne alternative approksimasjoner av de to størrelsene.

Når de skal benytte stokastisk simulering er det ikke tilstrekkelig bare å kjenne forventningsverdier og varianser til  $S$  og  $T$ , man må da også anta noe om hvilke fordelinger  $S$  og  $T$  har. For enkelthets skyld antar de at både  $S$  og  $T$  er normalfordelte, de setter  $\mu_s$  og  $\mu_t$  lik tilhørende målte verdier for  $s$  og  $t$  og bruker de samme verdier for  $\sigma_s$  og  $\sigma_t$  som i deloppgave a).

## Deloppgave b)

Skriv pythonkode for å simulere  $n = 10\,000$  realisasjoner av  $G$ , og bruk disse til å approksimere  $\mu_g$  og  $\sigma_g$ . Sammenlign med verdiene du fant i deloppgave a) og diskuter.

Tror du  $G$  er normalfordelt? For å undersøke dette: Lag et sannsynlighetshistogram av de  $n$  genererte verdiene for  $G$  og plott deretter (i samme plott som sannsynlighetshistogrammet) sannsynlighetstettheten til en normalfordeling med forventningsverdi og standardavvik lik de approksimerte verdiene for  $\mu_g$  og  $\sigma_g$ .

```
In [ ]: # Her kan du skrive pythonkoden din
```

## Oppgave 2 \*

Vi betrakter en skålvekt for veiing av prøver. Prøven med vekt  $\mu$  plasseres på en av skålene og lodd plasseres på den andre skålen til vektens arm er tilnærmet horisontal. Den kjente vekten av loddene er da tilnærmet lik prøvens vekt. Vi antar at den målte vekt er normalfordelt med forventningsverdi  $\mu$  og varians  $\sigma^2$  og at resultatene av forskjellige målinger er uavhengige.

### Deloppgave a)

Anta i denne deloppgaven at man gjør to veiinger av en prøve hvor  $\mu = 10$  gram og  $\sigma = 0.2$  gram. La  $X$  og  $Y$  være målt vekt i de to veiingene.

Hva er sannsynligheten for at gjennomsnittlig målt vekt er større enn 10.2 gram?

Hva er sannsynligheten for at differansen mellom de to målte vektene er større enn

0.4 gram?

### Her er deloppgave a) slutt.

Det er kommet inn to prøver som begge skal veies, prøve A og prøve B. La  $\mu_A$  og  $\mu_B$  betegne de sanne vektene av prøvene. Det er foreslått å estimere vektene av prøve A og B på to forskjellige måter:

Alternativ 1: Vei først prøve A fem ganger, og la  $X_1, X_2, X_3, X_4, X_5$  betegne de målte vektene. Vei deretter prøve B fem ganger, og la  $Y_1, Y_2, Y_3, Y_4, Y_5$  betegne de målte vektene. Bruk så følgende estimatorer for  $\mu_A$  og  $\mu_B$ ,

$$\hat{\mu}_A = \bar{X} = \frac{1}{5} \sum_{i=1}^5 X_i \quad \text{og} \quad \hat{\mu}_B = \bar{Y} = \frac{1}{5} \sum_{i=1}^5 Y_i.$$

Alternativ 2: Vei først fem ganger summen av vekten på prøve A og B ved å plassere begge prøvene på samme vektskål og la  $U_1, U_2, U_3, U_4, U_5$  betegne de målte vektene. Vei deretter fem ganger differansen mellom vektene på prøve A og prøve B, dvs A-B, ved å plassere prøvene på hver sin vektskål, og la  $V_1, V_2, V_3, V_4, V_5$  betegne de målte differansene. Bruk så følgende estimatorer for  $\mu_A$  og  $\mu_B$ ,

$$\tilde{\mu}_A = \frac{1}{2} (\bar{U} + \bar{V}) \quad \text{og} \quad \tilde{\mu}_B = \frac{1}{2} (\bar{U} - \bar{V}),$$

der  $\bar{U} = \frac{1}{5} \sum_{i=1}^5 U_i$  og  $\bar{V} = \frac{1}{5} \sum_{i=1}^5 V_i$ .

### Deloppgave b)

Finn forventningsverdi og varians for de fire estimatorene  $\hat{\mu}_A, \hat{\mu}_B, \tilde{\mu}_A$  og  $\tilde{\mu}_B$ .

*Merk: hver gang man leser av vekten med denne skålvekten har vi samme usikkerhet,  $\sigma$ . Dvs. hver måling har samme usikkerhet.*

Hvilken av de to alternative fremgangsmåtene for å estimere  $\mu_A$  og  $\mu_B$  vil du foretrekke? Gi en begrunnelse for svaret.

## Oppgave 3 \*

Det vurderes å utbedre en lite trafikkert, men farlig veistrekning i Trondheimsområdet. I den sammenheng blir du bedt om å analysere hvor mye trafikk det er på veien. La  $X$  være antall biler som passerer et bestemt punkt på veistrekningen fra kl 16:00 til kl 18:00 på en tilfeldig valgt hverdag. Vi antar at  $X$  er poissonfordelt med parameter  $\lambda$ , dvs

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, 2, \dots$$

Verdien til parameteren  $\lambda$  er ukjent, så for å kunne estimere denne verdien observerer vi antall passerende biler fra kl 16:00 til 18:00 på  $n$  tilfeldig valgte hverdager, og lar  $X_1, X_2, \dots, X_n$  være antall biler som passerte på disse  $n$  dagene. Vi antar at observasjonene er uavhengige og identisk fordelt med samme poissonfordeling som beskrevet over.

## Deloppgave a)

Utled sannsynlighetsmaksimeringsestimatoren (SME) for  $\lambda$ ,  $\hat{\lambda}$ , basert på  $X_1, X_2, \dots, X_n$ .

Er  $\hat{\lambda}$  forventningsrett? Begrunn svaret.

Finn variansen til  $\hat{\lambda}$ .

### Her er deloppgave a) slutt.

Kommunen ønsker å samle inn mer data om hvor trafikkert veien er. Derfor registreres det også hvor mange biler som passerer på veien fra kl 18:00 til kl 20:00 på  $m$  tilfeldig valgte hverdager, og vi lar  $Y_1, Y_2, \dots, Y_m$  være antall biler som da passerer. Vi antar at disse observasjonene er uavhengige og identisk poissonfordelte med parameter  $\lambda/2$ , altså halv intensitet i forhold til fra kl 16:00 til kl 18:00. Vi antar også uavhengighet mellom  $Y_1, Y_2, \dots, Y_m$  og  $X_1, X_2, \dots, X_n$ .

Som estimator for  $\lambda$  basert på  $X_1, X_2, \dots, X_n$  og  $Y_1, Y_2, \dots, Y_m$  er det foreslått

$$\tilde{\lambda} = \alpha \bar{X} + \beta \bar{Y},$$

der  $\alpha, \beta \in \mathbb{R}$  er to konstanter som skal velges, og  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  og  $\bar{Y} = \frac{1}{m} \sum_{i=1}^m Y_i$ .

## Deloppgave b)

Bestem konstantene  $\alpha$  og  $\beta$  slik at estimatoren  $\tilde{\lambda}$  blir best mulig. Svaret skal angis som funksjoner av  $n$  og  $m$ .

## Deloppgave c)

Til slutt i denne oppgaven skal du bruke stokastisk simulering til undersøke fordelingen til estimatoren  $\tilde{\lambda}$  når verdiene til konstantene  $\alpha$  og  $\beta$  velges lik de optimale verdiene du fant i deloppgave b). La  $n = 42$  og  $m = 35$ , og sett  $\lambda = 12$ .

Skriv python-kode som  $r = 1000$  ganger simulerer verdier for  $X_1, X_2, \dots, X_n$  og  $Y_1, Y_2, \dots, Y_m$  og regner ut resulterende verdi for  $\tilde{\lambda}$ . Lag et sannsynlighetshistogram for de  $r$  simulerte verdiene for  $\tilde{\lambda}$ . Bruk for eksempel "bins=25" som argument i pythonfunksjonen "plt.hist".

```
In [ ]: #Her kan du skrive kode for å simulere verdier og lage sannsynlighetshist
```

Greier du ut fra sannsynlighetshistogrammet å vurdere om  $\tilde{\lambda}$  er normalfordelt? Hvor sikker er du i din vurdering?

Lag et normalfordelingsplott (qq-plot) basert på de  $r$  simulerte verdiene for  $\tilde{\lambda}$ . Ser  $\tilde{\lambda}$  ut til å være normalfordelt? *Hint: Du kan bruke pythonfunksjonen "qqplot" i modulen statsmodels for å generere et qq-plot. Under ser du hvordan du kan laste inn modulen statsmodel og hvordan du kan kalle funksjonen qqplot. Her er det antatt at dine simulerte verdier for  $\tilde{\lambda}$  ligger i en vektor med navn "lambdatilde".*

Kan du gi en teoretisk begrunnelse for at  $\hat{\lambda}$  er (i hvert fall tilnærmet) normalfordelt?

```
In [ ]: import statsmodels.api as sm

sm.qqplot(lambdatilde)
plt.show()
```

## Oppgave 4

En kjemistudent har fått i oppgave å estimere konsentrasjonen av koffein i Coca Cola. Han undersøker  $n$  prøver, hver med nøyaktig samme volum, 1.0 dl, Coca Cola. For hver prøve måles antall mg koffein i prøven.

La  $x_i$  betegne målt koffeininnhold, i mg, for prøve nr.  $i = 1, 2, \dots, n$ . Vi antar at  $X_1, X_2, \dots, X_n$  er uavhengige og normalfordelte med (ukjent) forventningsverdi  $\mu$  og kjent standardavvik  $\sigma = 0.19$ , og at målingene  $x_1, x_2, \dots, x_n$  er realisasjoner av de stokastiske variablene  $X_1, X_2, \dots, X_n$ .

### Deloppgave a)

Gi en kort forklaring av de fysiske tolkningene av parameterne  $\mu$  og  $\sigma$  basert på den informasjonen du har fått.

### Deloppgave b)

Med utgangspunkt i at det skal gjøres  $n$  målinger, utled uttrykket for et  $(1 - \alpha) \cdot 100\%$ -konfidensintervall for  $\mu$ . Forklar hvorfor nedre og øvre grense vil variere hver gang forsøket gjentas.

Kjemistudenten har gjort  $n = 12$  målinger og rapporterer at det empiriske gjennomsnittet er  $\bar{x} = \sum_{i=1}^n x_i = 8.2$ . Regn ut et 95% konfidensintervall for  $\mu$ .

### Her er deloppgave b) slutt.

Kjemistudenten er ikke fornøyd med bredden på konfidensintervallet. Han vurderer derfor å gjenta eksperimentet med et nytt antall målinger,  $n$ .

## Deloppgave c)

Bestem den minste gyldige verdien for  $n$  slik at bredden på 95%-konfidensintervallet blir høyst 0.1.

## Oppgave 5

Det er vanlig at gjennomsnittstemperaturen i mars er lavere enn gjennomsnittstemperaturen i april. I 2012 var det mange steder i Norge kaldere i april enn i mars.

Vi lar  $X$  være gjennomsnittstemperatur i mars og  $Y$  gjennomsnittstemperatur i april ved Værnes et gitt år, begge målt i  $^{\circ}C$ . Anta at  $X$  er normalfordelt med ukjent forventningsverdi  $\mu_m$  og varians  $\sigma^2$ , og at  $Y$  er normalfordelt med ukjent forventningsverdi  $\mu_a$  og varians  $\sigma^2$ .

## Deloppgave a)

Anta at vi har data for verdier av  $X$  og  $Y$  fra  $n = 12$  ulike år. Du kan anta at dette er et tilfeldig utvalg. Forklar hvordan du kan grafisk undersøke om antakelse om normalfordeling holder. Hvordan kan du undersøke grafisk om  $X$  og  $Y$  er uavhengige?

Du skal nå gjennomføre den grafiske undersøkelsen med den gitte dataen gitt under. Begrunn om normalantakelsen holder, og om det virker som  $X$  og  $Y$  er uavhengige.

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
$x$	-2.5	0.5	3.3	2.6	-0.7	-4.6	3.3	0.8	1.9	-0.5	1.2	3.8
$y$	4.1	7.2	5.0	7.9	5.8	4.9	5.0	5.9	6.9	4.8	6.7	3.2

```
In [ ]: import numpy as np
import matplotlib.pyplot as plt

x = np.array([-2.5, 0.5, 3.3, 2.6, -0.7, -4.6, 3.3, 0.8, 1.9, -0.5, 1.2, 3.8])
y = np.array([4.1, 7.2, 5.0, 7.9, 5.8, 4.9, 5.0, 5.9, 6.9, 4.8, 6.7, 3.2])

#Normalantakelse

#Uavhengighet mellom X og Y

import scipy as sp
sp.stats.binomial()
```

## Deloppgave b)

Anta at temperaturene fra mars er uavhengige fra et år til et annet. Finn et 99%-konfidensintervall for forventet gjennomsnittstemperatur i mars.

## Fasit:

- Oppgave 1a:  $\sigma_g^2 \approx 7.791$
- Oppgave 2a: 0.0793, 0.1586
- Oppgave 2b:  $E[\hat{\mu}_A] = \mu_A, E[\hat{\mu}_B] = \mu_B, \text{Var}[\hat{\mu}_A] = \text{Var}[\hat{\mu}_B] = \frac{\sigma^2}{5},$   
 $E[\tilde{\mu}_A] = \mu_A, E[\tilde{\mu}_B] = \mu_B, \text{Var}[\tilde{\mu}_A] = \text{Var}[\tilde{\mu}_B] = \frac{\sigma^2}{10}$
- Oppgave 3a:  $\hat{\lambda}$  er forventningsrett,  $\text{Var}[\hat{\lambda}] = \frac{\lambda}{n}$
- Oppgave 3b:  $\alpha = \frac{2n}{2n+m}, \beta = \frac{2m}{2n+m}$
- Oppgave 4b: [8.09, 8.31]
- Oppgave 4c:  $n = 56$
- Oppgave 5b: [-1.506, 3.023]