

Innlevering 5

Oppgave 1 *

I situasjoner der det er uklart hvem som er den biologiske faren til et barn kan farskapet avklares ved å sammenligne DNA-prøver fra barnet med mulige fedre. For en mulig far gjøres dette ved å sammenligne n ulike deler av DNA-strukturen til mannen med de samme n deler av DNA-strukturen hos barnet. De n undersøkte delene av DNA-strukturen skal vi anta er uavhengige.

Hos et barn og en tilfeldig valgt mann (som ikke er biologisk far) er det for hver enkel del av DNA-strukturen som undersøkes en sannsynlighet $p = 0.15$ for at delen er sammenfallende hos barnet og mannen. Anta videre at en biologisk far alltid har alle de undersøkte delene av DNA-strukturen sammenfallende med barnets (dvs. vi ser bort fra mutasjoner o.l.), slik at hver undersøkte del av DNA-strukturen hos biologisk far og barn er sammenfallende med sannsynlighet $p = 1$.

I denne oppgaven skal vi anta at $n = 5$ deler av DNA-strukturen sammenlignes. Vi lar X være antall sammenfallende deler av DNA-strukturen hos et barn og en tilfeldig mann (som ikke er biologisk far).

Deloppgave a)

Begrunn av X er binomisk fordelt med parametrene n og $p = 0.15$.

Regn ut sannsynlighetene $P(X = 3)$, $P(X \geq 3)$ og $P(X = 3 | X \geq 2)$.

Her er deloppgave a) slutt

I en farskapssak blir en mann erklært å være biologisk far dersom alle de n undersøkte delene av DNA-strukturen er sammenfallende hos mannen og barnet. Dette kan vi se på som en hypotesetest der vi tester

$$H_0 : p = 0.15 \text{ (ikke far)} \quad \text{mot} \quad H_1 : p = 1 \text{ (far)}.$$

og forkaster H_0 (dvs. erklærer at mannen er far til barnet) dersom $X = n$.

Deloppgave b)

Hva er sannsynligheten for å gjøre en type I-feil i denne testen?

Hva er sannsynligheten for å gjøre en type II-feil i denne testen?

Oppgave 2 *

På et av de mange oppdrettsanleggene på Frøya ønsker oppdretteren å vite hvor mange (oppdretts)laks han har i en av merdene. Du er blitt engasjert som konsulent for å bistå. I en merd finnes det m laks, der m er ukjent og skal estimeres. Vi skal undersøke følgende metode: Vi fanger først r laks og merker disse før de legges ut i merden igjen. Siden det er vi som velger hvor mange laks vi vil merke er r selvfølgelig et kjent tall. Etter en stund begynner vi å fange en og en laks, og for hver laks vi fanger sjekker vi om laksen er merket eller ei og legger den ut i merden igjen før vi fanger neste laks. Slik holder vi på inntil vi k ganger har fanget en laks som er merket. Merk at siden k er et tall vi velger vil også dette være et kjent tall. La X være antall laks vi fanger (i den andre fasen hvor vi fanger en og en laks) frem til vi har fanget k merkede laks.

Deloppgave a)

Hvilke antagelser og eventuelt tilnærmelser må vi gjøre for å kunne betrakte X som negativt binomisk fordelt med parametere k og $p = r/m$?

Her er deloppgave a) slutt

Videre i oppgaven forutsetter vi at de antagelser og tilnærmelser som ligger til grunn for resultatet i a) er oppfylt. Dessuten, selv om antall laks m i merden må være et heltall, skal vi videre i oppgaven regne som om m er et reelt tall.

Deloppgave b)

Utleid sannsynlighetsmaksimeringsestimatoren for m , $\hat{m}$.

Finn forventningsverdi og varians til $\hat{m}$.

Her er deloppgave b) slutt

Hvis k er tilstrekkelig stor vil X være tilnærmet normalfordelt. Dette følger fra sentralgrenseteoremet fordi X da kan skrives som en sum av mange uavhengige og identisk fordelte variabler. Du trenger ikke å bevise dette, men du kan benytte resultatet til å løse den siste deloppgaven.

Etter en høststorm undrer oppdretteren på om det er rømt noen oppdrettslaks fra en av de andre merdene på oppdrettsanlegget. Dagen før stormen var det $m = 50,000$ laks i den aktuelle merden. Det blir derfor bestemt å estimere antall laks i merden etter høststormen. For dette benyttes det $r = 1000$ og $k = 20$, og det viste seg at man måtte fange 728 laks.

Deloppgave c)

Har oppdretteren noen grunn til å bekymre seg over rømt laks på et (tilnærmet) $\alpha = 5\%$ signifikansnivå? Formuler hypotesene H_0 og H_1 , velg en testobservator, bestem en beslutningsregel og finn hva konklusjonen blir med den observerte verdien.

Oppgave 3

SAR (Synthetic Aperture Radar) er en målemetode som benyttes for å kartlegge jordoverflaten fra satellitt. Teknikken går kort fortalt ut på at man sender ut radarstråler fra satellitten og observerer hvor mye av denne strålingen som reflekteres tilbake. Hvor mye av radarstrålingen som reflekteres, avhenger av egenskapene til jordoverflaten på den aktuelle posisjonen og dermed kan man skille mellom ulike overflatetyper. En observasjon gjøres egentlig ved at man tar flere målinger (såkalte "looks") og summerer disse. Ut fra fysiske lover for radarstråler er det kjent at en observasjon, X , vil være gammafordelt med parametre $\alpha = a$ og $\beta = r/a$, dvs. sannsynlighetstettheten er gitt ved

$$f(x) = \frac{a^a}{r^a \Gamma(a)} x^{a-1} \exp\left\{-\frac{ax}{r}\right\}, \quad x \geq 0,$$

der a er antall "looks" og refleksivitetsparameteren r er en størrelse som beskriver refleksjonsegenskapene til jordoverflaten der observasjonen gjøres. Ut fra kjente formler for forventningsverdi og varians for gammafordelingen vet vi dermed også at tilhørende forventningsverdi og varians er gitt ved

$$E[X] = r \quad \text{og} \quad \text{Var}[X] = \frac{r^2}{a}.$$

Vi skal nå anta at vi har n observasjoner fra et homogent område (dvs. verdien på r er den samme for alle de n observasjonene). La $X_1, X_2, \dots, X_n$ være de n observasjonene og anta at de er uavhengige. Fra disse observasjonene er vi interessert i å estimere r . Antall "looks", $a = 5$, antar vi kjent. Som estimator for r skal vi benytte

$$\hat{r} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

Deloppgave a)

Finn forventningsverdi og varians til estimatoren $\hat{r}$.

Forklar hvordan du kan bruke sentralgrenseteoremet til å konkludere at

$$\frac{\hat{r} - r}{r} \sqrt{na}$$

er tilmærmet standard normalfordelt når n er stor (nok).

Her er deloppgave a) slutt.

Vi fokuserer nå på et område hvor det er kjent at det for et år siden var en overflatetype som hadde refleksjonsparameter $r_0 = 12.5$. For å undersøke om overflaten har endret seg siden i fjor på en slik måte at refleksjonsparameteren har økt, gjøres det $n = 20$ målinger med $a = 5$ looks fra det aktuelle området. De $n = 20$ observerte verdiene er som følger:

7.98, 10.82, 15.88, 17.00, 24.22, 12.20, 8.17, 16.53, 7.46, 14.31, 34.55,
19.46, 20.21, 13.58, 10.98, 4.42, 24.92, 30.29, 23.45, 23.36

Vi ønsker nå å benytte disse verdiene til å undersøke om det er grunnlag til å påstå at verdien til refleksjonsparameteren har økt siden i fjor.

Videre i oppgaven skal vi forutsette at $n = 20$ observasjoner er tilstrekkelig til at normalapproximasjonen diskutert i deloppgave a) er god, slik at du kan benytte denne i dine videre beregninger.

Deloppgave b)

For å få et første inntrykk at de $n = 20$ observerte verdiene, bruk verdiene til å lage følgende plott i python

- histogram
- boksplott (for dette kan du bruke funksjonen `boxplot` i modulen `matplotlib.pyplot`)

Ut fra hva du ser i plottene, hva tror du om verdien av refleksjonsparameteren i år i forhold til i fjor? Tror du det er grunnlag for å påstå at verdien har økt siden i fjor?

```
In [ ]: # Her kan du skrive din python-kode(husk å importere nødvendige bibliotek
```

Deloppgave c)

Formuler problemstillingen gitt over som et hypotesetestingsproblem. Formuler null- og alternativ hypotese, velg testobservator, bestem beslutningsregel og finn hva konklusjonen blir basert på de observerte verdiene gitt over for signifikansnivå $\alpha = 0.10$.

Her er deloppgave c) slutt.

Anta at man også ønsker å gjøre en tilsvarende undersøkelse av et annet område som også hadde refleksjonsparameter $r_0 = 12.5$ i fjor. Før man gjør observasjoner fra dette området ønsker man å bestemme hvor mange observasjoner n man bør gjøre for at sannsynligheten for type II-feil skal være tilstrekkelig lav. Vi skal anta at man fremdeles bruker observasjoner med $a = 5$ looks og at signifikansnivået fremdeles settes lik $\alpha = 0.10$.

Deloppgave d)

Bestem hvor mange observasjoner, n , man minst må gjøre for at sannsynligheten for type II-feil skal være mindre enn eller lik 0.20 dersom refleksjonsparameteren er større enn eller lik 15.

Her er deloppgave d) slutt.

I deloppgave c) og d) har vi regnet som om størrelsen diskutert i deloppgave a) er standard normalfordelt. Dette er en approksimasjon, noe som innebærer at beslutningsregelen bestemt i deloppgave c) kan ha en sannsynlighet for type I-feil som avviker noe fra 0.10. I neste deloppgave skal du bruke stokastisk simulering til å estimere den faktiske sannsynligheten for type I-feil som man får ved å benytte beslutningsregelen fra deloppgave c).

Deloppgave e)

Skriv en python-funksjon som først simulerer $n = 20$ "observasjoner" fra gammafordelingen som gjelder når H_0 er sann, og deretter benytter disse observasjonene til å evaluere og returnere verdien til den tilhørende testobservatoren. *Merk: For å generere realisasjoner fra en gammafordeling kan du benytte funksjonen `gamma` i modulen `numpy.random`. Se mappen "Fordelinger" i JupyterHub for eksempel på bruk av funksjonen.*

Benytt så denne python-funksjonen, for eksempel $m = 100\,000$ ganger, til å estimere testens eksakte sannsynlighet for type I-feil. Finn også et 95%-konfidensintervall for den eksakte sannsynligheten for type I-feil. Basert på resultatene du har fått her, hva tenker du om kvaliteten av approksimasjonen du gjorde da du i deloppgave c) regnet som om testobservatoren var standard normalfordelt under H_0 ?

```
In [ ]: # Her kan du skrive din python-kode.
```

Oppgave 4

Anta at man har gjennomført en medisinsk studie for å vurdere om en ny medisin for en bestemt sykdom er bedre enn den tradisjonelle medisinen som har vært benyttet for denne sykdommen. Det var $n = 15$ pasienter som deltok i studien. Av disse n pasientene ble $m = 7$ tilfeldig trukket ut og disse m pasientene ble gitt den nye medisinen. De øvrige $n - m = 8$ pasientene fikk den gamle medisinen.

Effekten av behandlingen (medisinen) kan måles ved hjelp av en blodprøve. Vi skal i denne oppgaven anta at en lav blodprøveverdi er en indikasjon på at medisinen som er gitt har hatt en god effekt. Jo lavere blodprøveverdien er, jo bedre har effekten av medisinen vært.

I denne oppgaven ønsker vi å bruke følgende observerte blodprøveverdier til å vurdere om det er grunnlag for å påstå at den nye medisinen er mer effektiv enn den gamle.

Tradisjonell medisin	0.189	0.743	0.605	0.044	0.091	0.045	0.532	0.642
Ny medisin	0.397	0.583	0.355	0.054	0.155	0.066	0.099	

Vi skal formulere problemet som en hypotesetest og som testobservator skal vi benytte gjennomsnittet av observerte blodprøveverdier for pasienter som fikk ny medisin minus gjennomsnittet av observerte blodprøveverdier for pasienter som fikk den tradisjonelle medisinen.

Deloppgave a)

For å få et første inntrykk at de observerte verdiene, lag et boksplott for observerte blodprøveverdiene for pasientene som fikk tradisjonell medisin, og et boksplott for observerte blodprøveverdier for pasienter som fikk ny medisin. *Hint: Jobben blir gjort ved å kjøre koden under. Du trenger ikke endre på den gitte koden.*

```
In [ ]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

x = [0.189,0.743,0.605,0.044,0.091,0.045,0.532,0.642,
     0.397,0.583,0.355,0.054,0.155,0.066,0.099]
data = pd.DataFrame({'verdi': x, 'medisin':
                    ['Trad', 'Trad', 'Trad', 'Trad', 'Trad', 'Trad', 'Trad', 'Trad',
                     'Ny', 'Ny', 'Ny', 'Ny', 'Ny', 'Ny', 'Ny']})

sns.boxplot(x='medisin',y='verdi',data=data)
plt.show()
```

Ut fra hva du ser i disse to boksploottene, tenker du at det er grunnlag for å påstå at den nye medisinen gir bedre effekt enn den tradisjonelle? Gi argumenter for ditt svar.

Deloppgave b)

Formuler nullhypotese H_0 og alternativ hypotese H_1 for problemstillingen beskrevet i innledningen til oppgaven.

Bestem hvilken beslutningsregel det er rimelig å bruke. *Merk: Du trenger ikke å finne kritisk verdi, bare bestemme om man skal forkaste H_0 når testobservatoren er stor nok, om man skal forkaste H_0 når testobservatoren er liten nok, eller om man bør ha et tosidig forkastningskriterium.*

Skriv og kjør python-kode som bruker stokastiske simulering til å estimere p-verdien i en permutasjonstest når observasjonene er som gitt over. Benytt gjerne $m = 10\,000$ simuleringer for å estimere p-verdien. *Hint: Koden under regner ut observert verdi av testobservatoren og **en** simulert verdi av denne. Du er nødt til å modifisere denne koden slik at den gjør det som det spørres om.*

```
In [ ]: def testStatistic(x,nTrad):
    #x inneholder alle observerte verdier, de nTrad
    #første av disse er for pasienter som som fikk tradisjonell medisin

    #gjennomsnitt av observerte verdier for pasienter som fikk tradisjone
    meanTrad = np.mean(x[0:(nTrad+1)])
    #gjennomsnitt av observerte verdier for pasienter som fikk by medisin
    meanNew = np.mean(x[(nTrad+1):])

    return meanNew - meanTrad #returnerer differansen

from random import sample

# regner ut observert verdi av testobservatoren:
statisticObserved = testStatistic(x = x,nTrad = 8)
print('Observert verdi: ',statisticObserved)

# genererer tilfeldig en permutasjon av (alle) elementene i lista x:
xPermuted = sample(x,len(x))
# regner ut simulert verdi av testobservatoren
statisticSimulated = testStatistic(xPermuted,8)
print('Simulert verdi: ',statisticSimulated)
```

Fasit:

- Oppgave 1a): 0.0244, 0.0266, 0.1480
- Oppgave 1b): $7.594 \cdot 10^{-5}$, 0
- Oppgave 2b): $\hat{m} = \frac{X_r}{k}$, $E[\hat{m}] = m$, $\text{Var}[\hat{m}] = m(m - r)/k$
- Oppgave 2c): $H_0 : m = 50000$. Forkaster ikke H_0 .
- Oppgave 3a): $E[\hat{r}] = r$, $\text{Var}[\hat{r}] = \frac{r^2}{na}$
- Oppgave 3c): Forkast H_0 .
- Oppgave 3d): $n = 27$
- Oppgave 4a): Ja.