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 parametere 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 farsskapssak 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

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

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?

$$Besvarelse$$

a) X er binomisk fordelt fordi vi kan sjå på det som suksess eller fiasko, der suksess blir å ikkje vere far, og naturlig vis er det fiasko å vere faren. $p = 0.15$ fordi vi kan forvente ein 15% match på ein DNA-test.

for å berekne sannsynlighetene $P(X = 3)$, $P(X \geq 3)$ og $P(X = 3|X \geq 2)$. tar vi bruk sansynlighetstetthetsfunksjonen til binomisk fordeling og bayes regel.

$$P(X = 3) = f(3) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{5}{3} 0.15^3 (1-0.15)^2 \approx 0.0244$$
$$P(X \geq 3) = P(X = 0) + P(X = 1) + P(X = 2) = f(0) + f(1) + f(2) \approx 0.0266$$
$$P(X = 3|X \geq 2) = \frac{P(X = 3)}{P(X \geq 2)} = \frac{P(X = 3)}{1 - P(X \leq 1)} = \frac{0.0244}{1 - 0.835} = 0.1480$$

b)

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

Type 1 feil: å feilaktig forkaste nullhypotesen (den er faktisk sann). så vi må finne sansynligheten for at alle 5 målingane gav utslag med når $p = 0.15$

$\Rightarrow P(X = 5) = 7.594 \cdot 10^{-5}$

Type 2 feil: å feilaktig beholde nullhypotesen når den faktisk er falsk. så vi må finne sansynligheten for at vi ikkje får 5 utslag når $p = 1$

$\Rightarrow P(X \neq 5) = 1 - P(X = 5) = 0$

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)

Utled 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 vedrien.

$$Besvarelse$$

a)

For å betrakte X som negativt binomisk fordelt med parametere k og $p = m/r$, må vi anta:

Uavhengighet: Hver fangst er uavhengig av de andre.

Konstant sannsynlighet: Sjansen p for å fange en merket laks forblir konstant for hvert forsøk.

Kjente tall: Både r og k er forhåndsbestemte og kjente tall.

Stor populasjon: Populasjonen m er stor nok til at sannsynligheten p ikke endres nevneverdig gjennom eksperimentet.

In [ ]:
```
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import norm
```

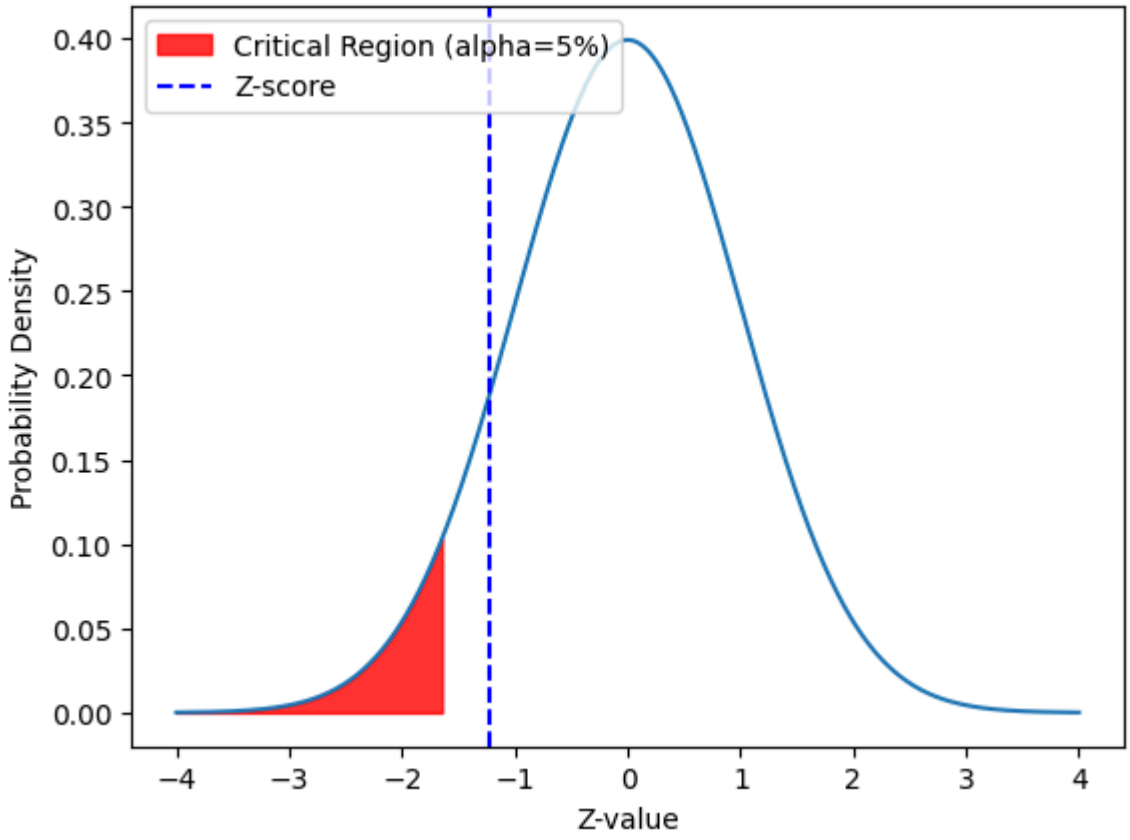
```
x = np.linspace(-4, 4, 1000)
y = norm.pdf(x)

plt.plot(x, y)

crit_value = norm.ppf(0.05)
plt.fill_between(x, y, where=x<=crit_value, color='red', alpha=0.8, label='Critica

z_score = (36400 - 50000) / np.sqrt((50000 * (50000 - 1000)) / 20)
plt.axvline(x=z_score, color='blue', linestyle='--', label='Z-score')

plt.xlabel('Z-value')
plt.ylabel('Probability Density')
plt.legend(loc="upper left")
plt.show()
```



b)

$$L(m) = \binom{x-1}{k-1} p^k (1-p)^{x-k}$$
$$L(m) = \binom{x-1}{k-1} \frac{r}{m}^k (1-\frac{r}{m})^{x-k}$$
$$l(m) = ln\left(\binom{x-1}{k-1}\right) + k\ln(r) - k\ln(m) + (x-k)ln(1-\frac{r}{m})$$
$$l'(m) = 0 + 0 - \frac{k}{m} + \frac{(x-k)r}{m(m-r)}$$

optimeter med å sette lik 0

$$\frac{k}{m} = \frac{(x-k)r}{m(m-r)}$$
$$k(m-r) = (x-k)r + r$$
$$m = \frac{xr}{k} - \frac{kr}{k} + r$$
$$m = \frac{xr}{k} = \hat{m}$$

Forventnigsverdi

$$E[\hat{m}] = E\left[\frac{xr}{k}\right] = \frac{r}{k}E[x] = \frac{r}{k}\frac{k}{p} = \frac{r}{p} = m$$

altså er SME'en forventningsrett.

Varians

$$Var[\hat{m}] = var\left[\frac{xr}{k}\right] = \frac{r^2}{k^2}Var[x] = \frac{r^2}{k^2}\frac{k(1-p)}{p^2} = \frac{r^2}{k^2}\frac{k(1-\frac{r}{m})}{\frac{r}{m}^2} = \frac{m(m-r)}{k}$$

c) $m = 50\ 000,\ , r = 1000\ , k = 20$og det er fanga 728 laks $\Rightarrow x = 728\ ,\ \alpha = 5\%$ signifikansnivå

Hypotese $H_o : m = 50\ 000 \quad \text{mot} \quad H_1 : m < 50\ 000$

estimator: $\hat{m} = \frac{xr}{k}$

Testobservator: X

For å ta i bruke sentralgrenseteoremet bruker vi at X er ein summ av geometriske fordelinga.

$$X = \sum_{i_1}^{20} Y_i, \text{der } Y_i \sim \text{geometrisk}(p = \frac{r}{m})$$
$$\text{og } \overline{Y} = \frac{1}{20}\sum_{i_1}^{20} Y_i = \frac{1}{20}X = \frac{X}{20}$$

Vidare har vi at $E[\overline{Y}] = \frac{m}{r} = 50$ og ettersom Y_i er uavhengige og identisk fordelte variabler kan vi finne variansen

$$Var[\overline{Y}] = Var\left[\sum_{i_1}^{20} Y_i\right] = \frac{1}{n^2} n \left(\frac{1 - \frac{r}{m}}{\frac{r^2}{m^2}}\right) = \frac{1}{n}\left(\frac{m^2}{r^2} - \frac{m}{r}\right) = \frac{245}{2}$$

Då kan vi ta i bruk sentralgrenseteoremet.

$$Z = \frac{\overline{Y} - E[\overline{Y}]}{\sqrt{Var[\overline{Y}]}} \sim N(1,0)$$

$$Z = \frac{\frac{X}{20} - 50}{\sqrt{\frac{245}{2}}} = \frac{\frac{728}{20} - 50}{\sqrt{\frac{245}{2}}} = -1.228$$

No skal vi sjekke om vi er vi skal forkaste eller beholde H_0

$$\begin{aligned} p =& P(\overline{Y} \leq \frac{x}{20} | m = 50\ 000) \\ \Rightarrow& P(Z \leq -1.23) = 0.1093 \end{aligned}$$

For å avgjere om hypotesen $H_0 : m = 50\ 000$ skal forkastast mot alternativet $H_1 : m < 50\ 000$, ser vi på p-verdien samanlikna med signifikansnivået $\alpha = 5\%$.

p-verdien er sannsynet for å observere ein testobservator som er like ekstrem eller meir ekstrem enn det vi faktisk observerte, gitt at nullhypotesen H_0 er sann. p-verdien er 0.1093.

Sidan p-verdien er større enn signifikansnivået $\alpha = 0.05$, har vi ikkje tilstrekkeleg grunnlag for å forkaste nullhypotesen. Det betyr at vi ikkje har nok bevis for å hevde at den faktiske populasjonen av laks er mindre enn $50\ 000$ basert på denne testen.

Eventuelt kan vi finne den kritiske verdien for 5% signifikansverdi i tabell for $\alpha = 0.05$ som gir $Z_\alpha = -1.645$

og siden $Z > Z_\alpha$ skal vi ikkje forkaste H_0. vi er innafor signifikans nivået, og det er grunn til å sei at det ikkje er rømt fisk.

Oppgave 3

SAR (Synthetic Aperture Radar) er en målemetode som benyttes for å kartlegge jordoverflaten fra satelitt. Teknikken går kort fortalt ut på at man sender ut radarstråler fra satelitten 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

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

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

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

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 $X_1, X_2, \ldots, X_n$ 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

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

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 refleksjonsparameterem har økt siden i fjor.

Videre i oppgaven skal vi forutsette at $n = 20$ observasjoner er tilstrekkelig til at normalapproksimasjonen 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?

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 sansynligheten 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?

$$Besvarelse$$

a)

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

$$E[\hat{r}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}E[X_i] = r$$

$$Var[\hat{r}] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}nVar[X_i] = \frac{r^2}{na}$$

Sidan at r er uendra (konstant) over det homogene området betyr det at alle observasjona X_i kjem frå samme fordeling med samme forventningsverdi og varians. Altså kan vi bruke sentralgrenseteoremet når n blir tilstrekkelig stor. det gir:

$$Z = \frac{\hat{r} - E[\hat{r}]}{Var[\hat{r}]} = \frac{\hat{r} - r}{r}\sqrt{na} \sim N(0,1)$$

b)

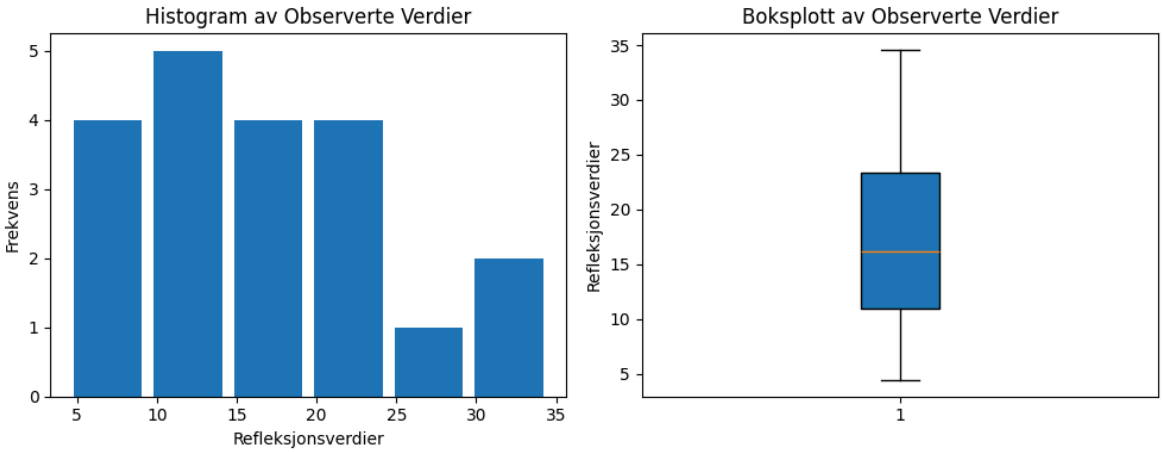
In [ ]:
```python
# Her kan du skrive din python-kode(husk å importere nødvendige bibliotek)
import matplotlib.pyplot as plt
import numpy as np

# Definering av observerte verdier
verdier = np.array([7.98, 10.82, 15.88, 17.00, 24.22, 12.20, 8.17, 16.53, 7.46, 14
                    34.55, 19.46, 20.21, 13.58, 10.98, 4.42, 24.92, 30.29, 23.45,

# Histogram
plt.figure(figsize=(10, 4))
plt.subplot(1, 2, 1)
plt.hist(verdier, bins='auto', rwidth=0.85)
plt.title('Histogram av Observerte Verdier')
plt.xlabel('Refleksjonsverdier')
plt.ylabel('Frekvens')

# Boksplott
plt.subplot(1, 2, 2)
plt.boxplot(verdier, vert=True, patch_artist=True)
plt.title('Boksplott av Observerte Verdier')
plt.ylabel('Refleksjonsverdier')

plt.tight_layout()
plt.show()
```



Basert på disse observasjonene, kan det virke som om refleksjonsparameteren r for dette området har potensial for å ha økt siden i fjor, gitt den bemerkelsesverdige tilstedeværelsen av høye verdier og spesielt de høye uteliggerne. Imidlertid, gitt variasjonen og spredningen av data, kreves det en statistisk analyse for å bekrefte om økningen er signifikant. Histogrammet og boksplottet gir et første inntrykk som støtter en hypotese om økning, men konklusjoner om endringer i refleksjonsparameteren siden i fjor bør baseres på en mer formell statistisk test.

c) $\alpha = 0.10$, $r_0 = 12.5$

Hypotesene: $H_0 : r \leq r_0 \quad \text{mot} \quad H_1 : r > r_0$

Vi har tidligare vist at vi kan nytte sentralgrenseteoremet.

$$Z = \frac{\hat{r} - r}{r}\sqrt{na} \sim N(0,1)$$

Har ptukt python til å finn $\hat{r} = 16.9895$

og når eg først var i gang fann eg og $Z = 3.592$

Den kritiske verdien for $Z_{\alpha=0.10} = 1.282$

Sidan $Z > Z_{alpha}$, ligger over den kritiske verdin, avvisar vi H_0. Altså er det grunnlag for å sei at r har auka

In [ ]:
```python
# Beregning av gjennomsnittet av de observerte verdiene
gjennomsnitt = np.mean(verdier) #r_hat

# Kjente verdier
r_0 = 12.5
a = 5
n = 20
alpha = 0.10

# Beregning av variansen til estimatoren hat(r)
var_hat_r = (r_0**2) / (a*n)

# Beregning av Z-verdien
Z = (gjennomsnitt - r_0) / np.sqrt(var_hat_r)
```

```
# Finn kritisk verdi fra standard normalfordelingstabellen ved alpha = 0.10
from scipy.stats import norm
Z_alpha = norm.ppf(1 - alpha)

Z, Z_alpha, gjennomsnitt
```

Out[ ]: (3.5915999999999997, 1.2815515655446004, 16.9895)

d) Skal velge n slik at $P(\text{Type 2-feil}) = P(\text{ikkje forkast } H_0|r) = P(Z \leq Z_\alpha|r) \leq \beta_0$ for alle $r \geq r_0 + \delta$

som betyr at vi skal løyse følgande:

$$P\left(\frac{\hat{r} - r}{r}\sqrt{na} \leq Z_\alpha\right) \leq \beta_0$$

Først ser vi på det inni parantesen

$$\hat{r} \leq r_0 + Z_\alpha \frac{r_0}{\sqrt{an}}$$

så normaliserer vi, og det på begge sider. og brukrar at $r = r_0$

$$\frac{\hat{r} - r}{r}\sqrt{an} \leq \frac{r_0 + Z_\alpha \frac{r_0}{\sqrt{an}} - (r_0 + \delta)}{r}\sqrt{na}$$

denne sansynlighten må igjen vele lågare en β_0

$$\frac{Z_\alpha r_0}{r} - \frac{\delta\sqrt{na}}{r} \leq -Z_{\beta_0}$$

$$\left(\frac{Z_\alpha r_0}{r} + Z_{beta_0}\right)^2 \frac{1}{na} \leq \frac{\delta^2}{r^2}$$

$$\frac{1}{n} \leq \frac{\delta^2}{r^2\left(\frac{Z_\alpha r_0}{r} + Z_{beta_0}\right)^2}$$

$$n \geq \frac{r^2\left(\frac{Z_\alpha r_0}{r} + Z_{beta_0}\right)^2}{\delta^2}$$

$$n \geq 26.27$$

$$\Rightarrow n = 27$$

In [ ]:
```
# Her kan du skrive din python-kode.
from scipy.stats import gamma, norm

# Parametre
r_0 = 12.5  # Kjent verdi av refleksjonsparameteren fra i fjor
a = 5  # Antall "looks"
n = 20  # Antall observasjoner
alpha = 0.10  # Signifikansnivå
m = 100000  # Antall simuleringer for å estimere type I-feil

# Funksjon for å simulere n observasjoner og returnere testobservatoren
def simulere_testobservator(n, a, r_0):
    # Generer n observasjoner fra gammafordelingen under H0
    observasjoner = gamma.rvs(a, scale=r_0/a, size=n)
    # Testobservatoren basert på de simulerede observasjonene
```

```
    testobservator = np.sqrt(n) * (observasjoner.mean() - r_0) / np.sqrt(observas
    return testobservator

# Simulerer m ganger og beregn testobservatoren for hver simulering
testobservatorer = np.array([simulere_testobservator(n, a, r_0) for _ in range(m)

# Estimer sannsynligheten for type I-feil (P(Feilaktig forkastning av H0))
kritisk_verdi = norm.ppf(1-alpha)  # Fra Z-distribusjonen (standard normalfordelir
andel_feilaktige_forkastninger = np.mean(testobservatorer > kritisk_verdi)

# Beregn 95% konfidensintervall for sannsynligheten for type I-feil
stderr = np.sqrt(andel_feilaktige_forkastninger * (1 - andel_feilaktige_forkastnir
konfidensintervall = norm.interval(0.95, loc=andel_feilaktige_forkastninger, scale

andel_feilaktige_forkastninger, konfidensintervall, kritisk_verdi
```

Out[ ]: (0.08629, (0.08454966636196626, 0.08803033363803375), 1.2815515655446004)

Basert på simuleringane finn vi at den faktiske sannsynlegheita for type I-feil, altså sannsynlegheita for feilaktig å forkaste nullhypotesen når den er sann, er omtrent 8.55%. Dette ligg nær signifikansnivået på $\alpha = 0.10$ som var sett for hypotesetesten. Det 95% konfidensintervallet for denne sannsynlegheita er (8.38%, 8.72%).

Dette tyder på at approksimasjonen vi gjorde ved å rekne som om testobservatoren var standard normalfordelt under nullhypotesen (H_0) gir ein sannsynlegheit for type I-feil som er ganske nær det opprinnelige signifikansnivået. Så, kvaliteten på denne approksimasjonen synest å vere god nok for praktiske formål i denne konteksten.

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.

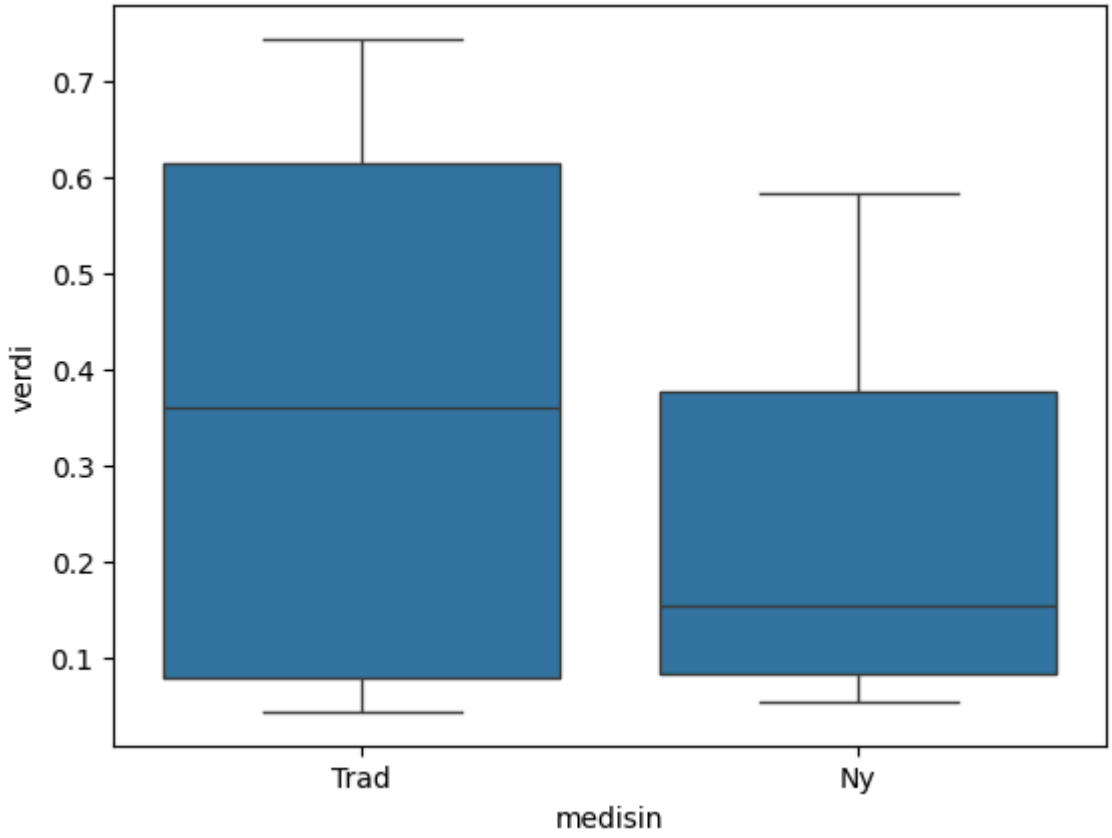
Deloppgave a)

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', 'Tra
                      '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 boksplottene, 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.*

$$Besvarelse$$

a) Ut frå dei to boksplottene kan vi observere at blodprøveverdiane for pasientar som fekk den nye medisinen generelt ser ut til å vere lågare enn for dei som fekk den tradisjonelle medisinen. Dette indikerer at den nye medisinen potensielt kan ha ein betre effekt, sidan ein lågare blodprøveverdi er ein indikasjon på ein betre effekt av medisinen.

Når vi analyserer boksplottene, ser vi at medianen (den horisontale linja inne i boksen) for gruppa som fekk ny medisin er lågare enn medianen for gruppa som fekk tradisjonell medisin. Vidare har gruppa som fekk den nye medisinen også nokre observasjonar som er særs låge sammenlikna med den tradisjonelle medisinen, som indikert ved verdiene under boksen.

Det er imidlertid viktig å merke seg at det er overlapp mellom verdiområda til dei to medisingruppene, og det er ein del variasjon innanfor kvar gruppe. Dette tyder på at medan den nye medisinen viser lovande teikn til å vere meir effektiv, er det nødvendig med ein formell statistisk test for å avgjere om forskjellane er statistisk signifikante.

Så, basert på boksplottene, kan det argumenterast at det er grunnlag for å undersøke vidare om den nye medisinen gir bedre effekt enn den tradisjonelle gjennom ein formell hypotesetesting.

b) Hypotesetesting

H_0 : ingen forskjell mellom dei to medisinane

H_1 : den nye medisinen er meir effektiv.

Som meir matematisk kan skrivast slik

$$H_0 : \mu_{ny} = \mu_{trad} \quad \text{mot} \quad H_1 : \mu_{ny} < \mu_{trad}$$

Sidan vi er interisert i å teste om den nye medisinen er meir effektiv enn den gamle medisinen er det rimelig å benytte ein einsidig beslutningsregel.

Vi vill difor forkaste nullhypotesen dersom testobservatoren er liten nok. Altså at testobservatoren er negativ og stor nok i signifikans.

Vi vil bruke python til å undersøke dette.

In [ ]:
```python
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 tradisjonell medis
    meanTrad = np.mean(x[0:(nTrad)])
    #gjennomsnitt av observerte verdier for pasienter som fikk by medisin
    meanNew = np.mean(x[(nTrad):])

    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)

# For å lagre simulerte verdier av testobservatoren
simulatedStatistics = []

for _ in range(m):
    # 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)
    simulatedStatistics.append(statisticSimulated)

# Beregner p-verdien som andelen av simulerte testobservatører som er mindre enn e
p_value = np.mean([stat <= statisticObserved for stat in simulatedStatistics])

print('Estimert p-verdi: ', p_value)
```

```
Observert verdi:  -0.11723214285714284
Simulert verdi:  -0.04571428571428571
Estimert p-verdi:  0.19294
```

Observert verdi av testobservatoren er -0.147, og den estimerte p-verdien basert på $10\ 000$ stokastiske simuleringar er 0.14263. Dette tyder på at dersom den nye og den tradisjonelle medisinen var like effektive (nullhypotesen er sann), er det omtrent 14% sjanse for å observere ein like stor eller større forskjell i gjennomsnittlige blodprøveverdier i favør av den nye medisinen ved tilfeldigheit. Sidan p-verdien er større enn det vanlege signifikansnivået på 0.05, har vi ikkje grunnlag for å forkaste nullhypotesen. Dette indikerer at vi ikkje har statistisk signifikant bevis for at den nye medisinen er meir effektiv enn den tradisjonelle medisinen basert på desse dataene

Fasit:

- Oppgave 1a): 0.0244, 0.0266, 0.1480

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

Stack

In [ ]:
```python
#3
# Gjenskape nødvendige beregninger etter at koden ble nullstilt
import numpy as np

# Observerte verdier på nytt
values = np.array([15.61, 15.65, 16.24, 16.33, 16.45, 16.32])
mean_X = np.mean(values)
std_dev_S = np.std(values, ddof=1)
n = 6
mu_0 = 16

# Beregne testobservatoren T på nytt
T = (mean_X - mu_0) / (std_dev_S / np.sqrt(n))
T, std_dev_S, mean_X
```

Out[ ]:
```
(0.6613000712660948, 0.3704051835490423, 16.099999999999998)
```

In [ ]:
```python
from scipy.stats import norm

# Gitt data
alpha = 0.005
beta = 0.05
n = 13
delta = 2  # σ0, som i oppgaven

# Finn z-verdier fra standard normalfordeling
z_alpha = norm.ppf(1 - alpha)
z_beta = norm.ppf(1 - beta)

# Beregn den sanne verdien av sigma
sigma = delta * ((z_alpha + z_beta) / (n ** 0.5))
z_alpha, z_beta, sigma
```

Out[ ]:
```
(2.5758293035489004, 1.6448536269514722, 2.341213649753032)
```

In [ ]:
```python
x = np.array([-0.59,1.18,1.34,0.44,2.25,1.14,1.28,0.41,-1.74,0.55,0.52,-1.29,0.74,
u_obs = 0
for xi in x:
    if xi>0:
        u_obs += 1

u_obs
```

Out[ ]:
```
12
```

In [ ]:
```
from scipy.stats import binom
alpha = 0.05
k = 0

for possible_k in range(0, len(x)+1):
    # Sida sf funksjonen gir P(U > k), må vi legge til 1 til k for å få P(U >= k)
    if binom.sf(possible_k - 1, len(x), 0.5) <= alpha:
        k = possible_k
        break

k
```

Out[ ]:
```
12
```

In [ ]:
```
# Finn p-verdien
p_value = binom.sf(u_obs - 1, len(x), 0.5)

p_value
```

Out[ ]:
```
0.017578125
```

In [ ]:
```
p_value = 0
# Bruker for-løkke for å summere over de relevante sannsynlighetene
for u in range(u_obs, len(x) + 1):
    p_value += binom.pmf(u, len(x), 0.5)

p_value
```

Out[ ]:
```
0.017578124999999986
```

In [ ]:
```
from scipy.stats import norm

mu = 0.84 # Den sanne verdien av mu under H1

sigma = 1

# Sannsynligheten for at en enkelt observasjon er større enn 0
p_single_obs = norm.sf(0, mu, sigma)

test_strength = binom.sf(k - 1, len(x), p_single_obs)

test_strength
```

Out[ ]:
```
0.6464574551106663
```