Oppgave 1: OL

45 langrennsløpere skal gå 30 km klassisk kvinner under et OL. Av disse er fire norske: Therese, Lotta, Tiril og Ragnhild. Blant de 45 deltakerne, trekkes 10 ut tilfeldig (alle har samme sjanse for å bli trukket) for å gjennomføre en dopingtest.

A.

(i) Deltakerne trekkes én og én. Hva er sannsynligheten for at Therese trekkes først?

(ii) Dersom Therese ikke er blant de første 9 som trekkes, hva er sannsynligheten for at hun blir trukket i siste trekk?

(iii) Hva er sannsynligheten for at Therese blir trukket overhodet?

(i) 1/45 (ii) $\frac{1}{45-9} = \frac{1}{36}$ (iii) 10/45

B. Vi ønsker å finne sannsynligheten for at det trekkes en norsk kvinne i nøyaktig trekk 1,3,5 og 7. Siden det er tungt å finne analytisk, bruker vi en simulering. Beskriv hvordan du kan lage en simulering i R der du finner denne sannsynligheten.
Du kan svare med R-kode, pseudo-kode eller en nøye beskrevet algoritme.

```
lopere = c(rep(TRUE,4),rep(FALSE,41))
mean(replicate(1e7,all(sample(lopere,10)[c(1,3,5,7)])))
```

Studentene bør få nær full uttelling dersom de beskriver en simulering, også dersom den mangler enkelte elementer som «all».

Alternative løsninger som bruker noen grad av analytisk resonnement rundt sannsynligheter før Monte Carlo får også full uttelling:

Lager først en vektor med alle de 45 mulige løperne, med fire 1'ere for norske, 41 0'er for ikke-norske. I hver trekning trekker vi 7 verdier uten tilbakelegging og sjekker om trekningen er 1010101. Gjentar mange ganger og finner andel hvor det er sant:

```
lopere <- cc(rep(1,4),rep(0,41))
mean(replicate(1e7,all(sample(lopere,7,replace = F) == c(1,0,1,0,1,0,1))))
```

Ingen uttelling for å sette opp det teoretiske svaret uten å beskrive en simulering.

Oppgave 2: Ulikhet

Hvert år måler SSB inntektsulikheten i Norge ved å samle inn data på inntekt etter skatt. Et vanlig mål på ulikhet er P90/P10, som angir det 90. over det 10. persentilet i inntektsfordelingen. Tenk deg at du laster ned mikrodata fra SSB for inntekt etter skatt (målt i 1000 kroner) på individnivå i R, og produserer følgende deskriptive statistikk:

```
> quantile(innt$inntekt[innt$aar==2005], p = c(.1, .9))
 10%  90%
 135  663
> quantile(innt$inntekt[innt$aar==2020], p = c(.1, .9))
  10%  90%
  230  1 154
```

A. Hva er det 90. persentilet i inntektsfordelingen (etter skatt) i 2020? Forklar med ord hva det angir.

90% av befolkningen har en inntekt på 154 000 kroner eller mindre etter skatt

B. Har inntektsulikheten i Norge, målt ved P90/P10 gått opp eller ned i løpet av perioden 2005-2020?

Opp

C. Hvordan ville du beregnet P80/P20 for 2005, og forklar hvordan denne skiller seg fra P90/10 tolkningsmessig. Vil du forvente at P80/P20 har en høyere eller lavere verdi enn P90/P10?

```
v=quantile(innt$inntekt[innt$aar==2005], p = c(.2, .8))
p80_20 = v[2]/v[1]
print(p80_20)
```

Tolkningsmessig er q80 ikke så langt opp som q90 og vice versa, så forskjell mellom mindre ekstreme grupper. Dermed vil P80/P20 vil være lavere enn P90/P10.

En gruppe forskere ønsker å se nærmere på inntekten til de med aller høyest inntekt ifølge offisielle inntektsdata. De gjennomfører en stor anonym undersøkelse der de trekker tilfeldig 1000 personer fra det 99.persentilet i inntektsfordelingen, og ber dem rapportere urealiserte inntekter (tilbakeholde overskudd i bedrifter, verdiøkninger på f.eks. bolig, aksjer og andre verdipapirer, tjenester fra konsumkapital osv.). Forskerne produserer følgende deskriptive statistikk (der tallverdier er angitt i 1000 kroner):

```
dstats = function(x) {
      c(mean=mean(x), sd=sd(x), n=length(x))
}
> dstats(svartinnt$svartinntekt)
     mean            sd          n
     1000           600          1000
```

D. Foreslå en estimator for forventet urealisert inntekt blant personer i det 99. persentilet i inntektsfordelingen, og beregn et 90% konfidensintervall for forventet urealisert inntekt. Du kan anta at det ikke er en opphopning av personer som rapporterer akkurat null urealisert inntekt. Vær eksplisitt på alle antakelser du gjør og hvordan du tolker konfidensintervallet. Du kan få bruk for deler av utskriften fra R nedenfor.

```
> qnorm(c(.90, .95, .975), mean=0, sd=1)
[1] 1.281552 1.644854 1.959964
> qt(c(.90, .95, .975), df=1000)
[1] 1.282399 1.646379 1.962339
> qt(c(.90, .95, .975), df=10)
[1] 1.372184 1.812461 2.228139
```

Gjennomsnittet tilnærmet normalfordelt ifølge sentralgrenseteoremet.

$P\left(\bar{X} - t_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}\right) = \alpha$. Estimerer et KI: $1000 \pm 1{,}64 \times \frac{600}{\sqrt{1000}} = [968{,}9; 1031{,}1]$

Hvis vi trekker veldig mange utvalg og estimerer et konfidensintervall for hvert utvalg, vil 95% av de estimerte konfidensintervallene dekke forventet urealisert inntekt blant personer i det 99. persentilet i (offisiell) inntektsfordeling. Merk at det er feil å si at intervallet dekker den sanne parameterverdien med en sannsynlighet på 90%, da parameteren ikke er stokastisk.

E. Test hypotesen at urealisert inntekt blant personer i det 99.persentilet i inntektsfordelingen er lik null på 10% signifikansnivå.

La X_U være urealisert inntekt. Formulerer følgende hypoteser:

$$H_0: X_U = 0$$
$$H_1: X_U \neq 0$$

Siden konfidensintervallet ekskluderer null, kan vi forkaste hypotesen på 90% signifikansnivå.

Dersom noen gjennomfører hypotesetesten, kan de velge om de gjennomfører Z- test eller T-test pga stort utvalg.

$|Z| = \frac{1000}{600/\sqrt{1000}} = 52{,}7 > z_{0{,}5} = 1{,}64$. Forkaster H_0 på 10% signifikansnivå.

F. Anta at utskriften fra R over var basert på en spørreundersøkelse for 10 personer i stedet for 1000. Vil fremgangsmåten for å teste hypotesen i f), samt konklusjonen din endre seg i dette tilfellet? Forklar.

Må bruke T-test i dette tilfellet (siden utvalget er lite er normalfordelingen er dårlig tilnærming). $|T| = 5{,}27 > t_{0{,}5} = 1.81$. Forkaster fortsatt H_0 på 10% signifikansnivå.

Ekstra uttelling dersom studenten kommenterer at man må bruke T-test.

G. Anta at forskerne ved et uhell hadde sett bort ifra deler av utvalget, og at det allikevel er en opphopning av data på null. Diskuter hvorvidt dette kan ha noen konsekvenser for konklusjonen din i f)?

SGT: Med en skjev fordeling trenger vi mer data for at tilnærmingen skal bli god.

Oppgave 3: Medisin

Det har kommet en ny medisin for lungebetennelse, og myndighetene må vurdere om de skal ta i bruk den nye medisinen, eller fortsette å skrive ut den gamle. Den nye medisinen er anvendt på 71 pasienter, og en forsker finner at gjennomsnittlig sykdomsforløp blant disse er 14,3 dager med et standardavvik på 3. Basert på journalene til tidligere pasienter finner forskeren at gjennomsnittlig sykdomsforløp ved behandling med gammel medisin er 15 dager.

Den nye medisinen er betydelig mer kostbar enn den gamle. Tenk deg at du er økonom i Statens legemiddelverk, og skal gi råd til myndigheten om hvorvidt de bør betale for at medisinen gjøres tilgjengelig for alle pasienter i Norge.

A. Formuler et sett med hypoteser for å teste om den nye medisinen er bedre enn den gamle. Forklar eventuell notasjon du introduserer.

$$H_0: \mu \geq 15$$

$$H_1: \mu < 15$$

der μ angir forventet helbredelsestid for den nye medisinen.

B. Finn testens p-verdi, og gi en tolkning av p-verdien i denne konteksten. Du kan få bruk for utskriften fra R nedenfor.

```
> pnorm((14.5-15)/(3/sqrt(71)))
[1] 0.08010609
> 1-pnorm((14.5-15)/(3/sqrt(71)))
[1] 0.9198939
```

$$P_{\mu=\mu_0}(\bar{X} \leq 14{,}5) = P\left(Z \leq \frac{14{,}5 - 15}{\frac{3}{\sqrt{71}}}\right) = P(Z \leq -1{,}404) = 1 - P(Z \leq 1{,}404) = 0.08$$

Det er 8% sannsynlig å finne at den nye medisinen har en helbredelsestid på maksimalt $14{,}5$ dager, dersom helbredelsestiden i virkeligheten er 15 dager (eller mer).

C. Er p-verdien en stokastisk variabel? Forklar.

Ja. Et annet utvalg vil gi en annen p-verdi.

D. Kan du bruke p-verdien til å avgjøre om du kan forkaste H_0 på 5% og 10% signifikansnivå?

$0{,}10 > p - verdi = 0{,}08 > 0{,}05 \rightarrow$ kan forkaste H0 på 10%, men ikke 5% signifikansnivå.

E. Forklar, ut ifra konteksten over, hva type I og type II feil er, og diskuter hvorvidt den ene type feil er mer alvorlig enn den andre.

Type I: Anta at H0 er sann: den gamle medisinen er minst like god som den nye. Vi gjennomfører en hypotesetest og bestemmer oss for å forkaste H0. Vi avskaffer dermed en veletablert medisin, og tar kostanden med å innføre en ny medisin til tross for at den gamle medisinen egentlig er minst like god. De fleste er enige om at dette er en feil vi ikke ønsker å gjøre. Dette kalles type I feil.

Type II: Anta at H1 er sann: den nye medisinen er bedre enn gamle. Vi gjennomfører testen og bestemmer oss for ikke å forkaste H0. Dermed beholder vi den gamle medisinen selv om den nye egentlig er bedre. Kan argumenteres for at dette er en mindre alvorlig feil å gjøre enn type I feil i denne konteksten, siden helbredelsestiden er svært lik, og kostnaden med å bytte system, samt for ny medisin er høy.

F. Basert på konklusjonen din i C, samt rådataene som testen er basert på, hva ville du anbefalt myndighetene? (Her er vi ute etter en kort diskusjon, uten nødvendigvis en absolutt konklusjon).

Et pluss dersom studenten også tar opp økonomisk signifikans. Er det i praksis stor forskjell på 14,5 og 15 dager?