Oppgavesettet består av 3 oppgaver med flere deloppgaver. Hver deloppgave har samme vekt ved sensur.

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?

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.

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.

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

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?

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
```

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

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)?

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.

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
```

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

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

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).