Oppgave 1

I forbindelse med gjenåpningen av samfunnet etter korona-nedstengingen, er det av stor interesse for både helsemyndighetene og den enkelte å vite om man har vært smittet av koronaviruset. Selv om forskningsresultatene rundt opparbeidet immunitet ikke er klare ennå, er det mye oppmerksomhet rundt utviklingen av serologisker tester som kan påvise antistoffer fra viruset i blod.

Det finnes ulike produsenter av slike serologiske antistoff-tester. Vi skal se på to av de beste som foreløpig finnes på markedet i Europa, og kaller dem her T1 og T2. I følge en fersk sammenlignende studie har de to testene følgende karakteristika:

	Sensitivitet	Spesifisitet
T1	95.5%	98.0%
T2	94.2%	99.8%

Sensitivitet er her sannsynligheten for at testen gir et positivt resultat (testen indikerer antistoffer i blodet), gitt at man faktisk har antistoffer i blodet. *Spesifisitet* er sannsynligheten for at testen gir et negativt resultat (testen indikerer ikke antistoffer i blodet), gitt at man ikke har antistoffer i blodet.

I Norge er det foreløpig en liten andel av befolkningen som har vært smittet av koronaviruset. Et (usikkert) anslag fra Folkehelseinstituttet (FHI) er at ca. 1% så langt har antistoffer i blodet. Vi vil i det følgende se på en populasjon der 1% har slike antistoffer.

a) Finn sannsynligheten for at en tilfeldig valgt testperson får et positivt resultat av test T1.

b) Beregn sannsynligheten for at en tilfeldig valgt testperson faktisk har antistoffer i blodet, gitt en positiv test (T1). Kommenter resultatet.

c) Test T2 har noe høyere spesifisitet enn T1, men lavere sensitivitet. Under samme forutsetninger som tidligere, beregn sannsynligheten for at en tilfeldig testperson faktisk har antistoffer i blodet, gitt en positiv test T2. Sammenlign med T1 og kommenter.

d) Det kan skape problemer om individer som har testet positivt, egentlig ikke har antistoffer i blodet, da smittefaren for dem vil være stor. Hva er sannsynligheten for å ikke ha antistoffer i blodet, gitt at man har testet positivt? Finn denne sannsynligheten for begge testene og kommenter.

e) Hvor høy måtte spesifisiteten ha vært for test T2 for at sannsynligheten i punkt d skal bli mindre enn 5%?

f) 10 uavhengige, tilfeldig valgte personer testes med testen T1, og alle tester negativt. Hva er sannsynligheten for at minst én av disse 10 egentlig har vært smittet av koronaviruset og derfor har antistoffer i blodet?

Kommentar: Denne oppgaven er laget i midten av mai. Den er basert på realistiske tall og opplysninger, men mye kan være endret på få uker.

Oppgave 2

De to kontinuerlige stokastiske variablene X og Y har simultan sannsynlighetstetthet

$$f(x,y) = \begin{cases} kx(x+y) & \text{når } 0 \le x \le 2 \text{ og } 0 \le y \le 2 \\ 0 & \text{ellers} \end{cases}$$

a) Vis at konstanten $k = \frac{3}{28}$.

b) Beregn sannsynligheten for at Y er større eller lik X, dvs. $P(Y \ge X)$.

c) Finn de marginale sannsynlighetstetthetene for X og Y. Er X og Y uavhengige?

d) Finn sannsynlighetstettheten til $U = X + Y$. (*Vink*: Finn først den simultane tettheten til U og V, der $U = X + Y$ og $V = X$.)

Oppgave 3

Anta at $Y \sim N(\mu, \sigma^2)$ og la $X = e^Y$. Da er X lognormalfordelt med parametere μ og σ.

a) Vis at medianen i den lognormale fordelingen er $\eta = e^{\mu}$.
Vis også at $E(X) = E(e^Y) = \eta\, e^{\sigma^2/2}$.
(*Vink*: For å bestemme forventningen, kan du bruke at den momentgenererende funksjonen til Y er $M_Y(t) = E(e^{tY}) = e^{\mu t + \sigma^2 t^2/2}$.)

Anta nå at $X_1, \ldots, X_n$ er uavhengige og lognormalfordelte, og sett $Y_i = \ln(X_i)$ for $i = 1, \ldots, n$. Vi antar videre i punktene b-f at vi kjenner verdien til σ.

b) Vis at

$$\widehat{\mu} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} \ln(X_i)$$

er en forventningsrett estimator for μ og bestem fordelingen til $\widehat{\mu}$.

c) En mulig estimator for medianen η er $\eta^* = e^{\widehat{\mu}}$.
Forklar hvorfor denne estimatorene ikke er forventningsrett.

d) Vis at $\widehat{\eta} = e^{\widehat{\mu} - \sigma^2/(2n)}$ er en forventningsrett estimator for medianen.

Vi antar nå at vi har observert verdiene $x_1, \ldots, x_n$ av de stokastiske variablene $X_1, \ldots, X_n$, og vi setter $\overline{y} = (1/n)\sum_{i=1}^{n} \ln(x_i)$.

e) Vis at

$$\left[e^{\overline{y} - 1.96\sigma/\sqrt{n}},\ e^{\overline{y} + 1.96\sigma/\sqrt{n}} \right]$$

er et 95% konfidensintervall for medianen.

Vi har i punktene c-e fokusert på medianen η. Vi ser så på forventningen $E(X)$.

f) Vis at $\widehat{\eta}\, e^{\sigma^2/2}$ er en forventningsrett estimator for $E(X)$. Vis også at vi får et 95% konfidensintervall for $E(X)$ ved å gange nedre og øvre grense av konfidensintervallet i punkt e med $e^{\sigma^2/2}$.

Vi har ovenfor antatt at σ er kjent. Vanligvis vil vi ikke kunne gjøre denne antagelsen. Til slutt vil vi se hvordan vi kan gå fram når σ er en ukjent parameter. Vi vil nøye oss med å se på estimering av medianen $\eta = e^{\mu}$.

Vi vet at

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \tag{1}$$

er en forventningsrett estimator for σ^2. Så ut fra resultatet i punkt d, er det rimelig å bruke $\widetilde{\eta} = e^{\widehat{\mu} - S^2/(2n)}$ som en estimator for medianen.

Vi kan bruke stokastisk simulering til å undersøke egenskapene til $\widetilde{\eta}$ og sammenligne den med estimatoren η^* fra punkt c. Vi går da fram på følgende måte:

Først velger vi verdier for parameterene μ og σ og utvalgsstørrelsen n.

Deretter

(i) trekker vi n verdier $y_1, \ldots, y_n$ fra $N(\mu, \sigma^2)$-fordelingen

(ii) beregner $\widehat{\mu} = \overline{y} = (1/n)\sum_{i=1}^{n} y_i$ og $s^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2/(n-1)$

(iii) beregner estimatene $\eta^* = e^{\widehat{\mu}}$ og $\widetilde{\eta} = e^{\widehat{\mu} - s^2/(2n)}$

Vi gjentar punktene (i)-(iii) B ganger og får B verdier av η^* og $\widetilde{\eta}$. Til slutt beregner vi gjennomsnittsverdi og empirisk standardavvik for de B verdiene av η^* og $\widetilde{\eta}$.

g) Gjør beregningen beskrevet ovenfor for $n = 10$ og $n = 30$ når $\mu = 0.7$ og $\sigma = 1.2$. Bruk $B = 10\,000$. Diskuter hva resultatene sier deg om (mulig) skjevhet og standardfeil for estimatorene η^* og $\widetilde{\eta}$.

Oppgave 4

Det er vanlig å anta at det årlige alkoholforbruket (liter ren alkohol) i en befolkningsgruppe er lognormalfordelt (når vi ser bort fra de som aldri drikker alkohol), og vi vil gjøre denne antagelsen i denne oppgaven.

La X være alkoholforbruket for en tilfeldig valgt ikke-avholdende, voksen norsk kvinne. Vi vil anta at X er lognormalfordelt med $\mu = 0.7$ og $\sigma = 1.2$ (som er en noenlunde rimelig antagelse ut fra tilgjengelig statistikk).

a) Hvor stor andel av (ikke-avholdende) norske kvinner drikker minst 10 liter ren alkohol per år?

b) Bestem medianen til det årlige alkoholforbruket og forventet årlig alkoholforbruk for (ikke-avholdende) voksne norske kvinner. Hvilken av de to verdiene egner seg etter din mening best til å beskrive alkoholforbruket for voksne norske kvinner?

En rusmiddelforsker vil undersøke hvordan alkoholforbruket er blant en gruppe kvinnelige studenter. Vi vil anta at det årlige alkoholforbruket i denne gruppen er lognormalfordelt med (ukjente) parametre μ og σ. Forskeren kartlegger det årlige alkoholforbruket for et tilfeldig utvalg på 30 studenter fra den aktuelle gruppen og finner at det årlige forbruket for de 30 studentene er

```
1.0  3.4  5.0 14.4 11.5  8.2  0.6  2.7 26.8  3.0
1.3 20.2  4.0 14.0  3.3  1.8  1.7  4.6  7.4  7.1
5.2 23.6  1.6  1.1 15.5  3.0  1.9  4.2 27.4  1.5
```

I punktene c-f i oppgave 3 ser vi på estimatorer og konfidensintervall for medianen og forventningsverdien når det forutsettes at σ er kjent. Disse estimatorene og konfidensintervallene kan vi også bruke når σ er ukjent. Vi erstatter da σ med estimatet s, jf. (1). [Estimatoren i punkt 3 d vil da ikke være eksakt forventningsrett og konfidensintervallet i punkt 3 e vil ikke ha konfidensgrad ("confidence level") som er nøyaktig 95%.]

c) Gi et estimat og et tilnærmet 95% konfidensintervall for medianen til det årlige alkoholforbruket for studentgruppen. Gi også et estimat og et tilnærmet 95% konfidensintervall for forventet årlig alkoholforbruk.

d) Bruk parametrisk bootstrap til å bestemme standardfeilen til estimatene i forrige punkt.

e) Diskuter hva resultatene i punktene c og d sier deg om alkoholforbruket til den aktuelle gruppen av kvinnelige studenter sammenlignet med alle voksne norske kvinner.

SLUTT