STK1100

Obligatorisk oppgavesett 1 - vår 2025.

Innleveringsfrist

Torsdag 6. mars 2025, klokka 14:30 i Canvas (canvas.uio.no).

Instruksjoner

Du velger selv om du skriver besvarelsen for hånd og scanner besvarelsen eller om du skriver løsningen direkte inn på datamaskin (for eksempel ved bruk av LaTeX). Besvarelsen skal leveres som én PDF-fil. Scannede ark må være godt lesbare. Besvarelsen skal inneholde navn, emne og oblignummer.

Det forventes at man har en klar og ryddig besvarelse med tydelige begrunnelser. Husk å inkludere alle relevante plott og figurer. **Vær oppmerksom på at det ikke er mulighet for å levere en revidert besvarelse dersom den første besvarelsen ikke blir godkjent**. Samarbeid og alle slags hjelpemidler er tillatt, men den innleverte besvarelsen skal være skrevet av deg og reflektere din forståelse av stoffet. Oppgi gjerne hvilke hjelpemidler du har brukt! Er vi i tvil om du virkelig har forstått det du har levert inn, kan vi be deg om en muntlig redegjørelse.

I oppgaver der du blir bedt om å programmere, må du legge programkoden inn i PDF-en sammen med resten av besvarelsen.

For å få adgang til avsluttende eksamen i STK1100 må man

- ha fått godkjent begge de obligatoriske oppgavesett 1 og 2 i samme semester (inneværende semester eller tidligere, se regelverk nedenfor).

Søknad om utsettelse av innleveringsfrist

Ved udokumenterte årsaker kan du benytte egenmelding for å få innvilget fire kalenderdager utsettelse på en obligatorisk oppgave. Send i så fall e-post til faglærer (geirs@math.uio.no) **før** innleveringsfristen går ut.

Ved dokumenterte årsaker kan du søke om inntil én uke utsettelse totalt på en obligatorisk oppgave på bakgrunn av gyldig dokumentasjon. Se lenken nedenfor for prosedyre for dette.

www.uio.no/studier/admin/obligatoriske-aktiviteter/mn-math-oblig.html

Her finner du også fullstendige retningslinjer for innlevering av obligatoriske oppgaver

Spesielt om dette obligatoriske oppgavesettet våren 2025

Hvis du har fått registrert oppmøte på gruppeøvelsene minst 4 av 6 uker før innleveringsfristen, så vil du ikke behøve å levere oppgavene 1 og 2 i oblig 1. Ansvarlige for gruppeøvelsene vil registrere oppmøte for de som har valgt denne varianten. Husk likevel at det kan være god trening å også gjøre disse oppgavene for å få tilbakemelding fra gruppelærerne!

Det anbefales på det sterkeste at du bruker Python til å gjøre beregningene i oppgavene 3 og 4. Hvis du bruker et annet programmeringsspråk, kan vi ikke hjelpe deg hvis du får problemer. Det ligger eksempel-kode, som vil være til stor hjelp, på uio's jupyterhub. Hvis du trenger hjelp til å løse oppgavene, kan du få det på de åpne gruppene i STK1100.

Krav for godkjenning

- Du har kun *ett* forsøk.
- Du må ha gjort et hederlig forsøk på alle deloppgaver i oppgavene 1, 2 og 3, med mindre du har fritak fra å levere 1 og 2. I oppgave 4 holder det at du prøver på a), b) og c), hvis du finner resten av oppgaven vanskelig. Du kan se på de resterende delpunkter som en "utfordringsoppgave".
  - Hvis du trenger råd og tips, bruk gruppelærerne!
- For programmeringsoppgavene må du angi hvilke kommandoer du har brukt for å komme fram til svarene dine. Kode må legges ved.

LYKKE TIL!

Oppgave 1

Matematikkbygningen (Niels Henrik Abels hus) har 12 etasjer (vi ser bort fra underetasje og kjeller). Fire personer går inn i den samme heisen i 1. etasje, og forlater heisen i de enkelte etasjene uavhengig av hverandre. Videre antar vi at sannsynlighetene for at en bestemt person skal gå av i hhv. 2., 3., ... , 12. etasje er like store.

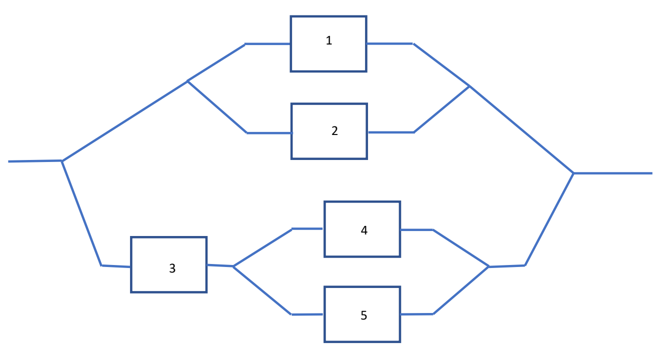
(a) Hva er sannsynligheten for at de 4 personene går av i hver sin etasje?

(b) Hva er sannsynligheten for at minst 2 av de 4 går av i samme etasje?

(c) Anta nå at akkurat 3 av de 4 går av i 8. etasje, der statistikk-seksjonen holder til. Hvor mange mulige ulike grupper på 3 kan dette være?

Heisen har et alarmsystem som består av 5 elementer koblet sammen som vist på figuren nedenfor. Komponent 1 og 2 er koblet i parallell, slik at denne delen av alarmen virker dersom enten 1 eller 2 virker. Komponent 3 er koblet i serie med komponent 4 og 5, som er koblet i parallell. Denne delen av alarmen virker dersom 3 virker samtidig som enten 4 eller 5 virker. De to delene (1,2) og (3,4,5) er til slutt koblet i parallell.

(d) Gitt at alle komponenter virker uavhengig av hverandre, og at sannsynligheten for at hver komponent virker er 0.95, finn sannsynligheten for at alarmen i heisen virker.



Oppgave 2

Et anti-jukse-program avslører at en tekst er AI-generert med sannsynlighet 0.90. Hvis teksten ikke er AI-generert, vil programmet ta feil og fastslå at teksten er AI-generert med sannsynlighet 0.07. Det antas videre at sannsynligheten for at en tilfeldig student benytter AI til å skrive en tekst, er 0.05.

(a) Hvis programmet konkluderer med at studentens tekst er AI-generert, hva er sannsynligheten for at den faktisk er det?

(b) Hvor liten må sannsynligheten for at programmet feilaktig fastslår at teksten er AI-generert være, for at sannsynligheten i a) skal bli over 90%?

Oppgave 3

På forelesning har vi sett på forventet levealder for norske menn. Denne oppgaven vil først og fremst gå ut på å bruke Python for tilsvarende beregninger for kvinner, og å sammenligne resultatene for kvinner og menn. Du kan ta utgangspunkt i Python-skriptet `doedelighet.ipynb` som er tilgjengelig gjennom jupyterhub.uio.no og også på kursets hjemmeside. Datasettet er også tilgjengelig i filen doedelighet.txt.

La X være levealder (i hele år) for en tilfeldig norsk mann, og tilsvarende, la Y være levealder (i hele år) for en tilfeldig norsk kvinne. Med levealder mener vi tid (i hele år) fra fødsel til død. Det vi ønsker å regne ut er forventet gjenstående levealder for menn og kvinner, ved en gitt alder a, som kan skrives som

$$\mathrm{E}(X-a|X \geq a) \quad \text{og} \quad \mathrm{E}(Y-a|Y \geq a)$$

for forskjellige verdier av a. Vi kan da bruke utledningene fra forelesningspresentasjonene til å si at

$$\mathrm{E}(X-a|X \geq a) = E[h(X)]$$

der

$$h(x) = \begin{cases} 0, & \text{hvis } x < a \\ \frac{x-a}{1-F_X(a-1)}, & \text{hvis } x \geq a. \end{cases}$$

Her er $F_X(x) = P(X \leq x)$, den kumulative fordelingsfunksjonen til X. Vi får også tilsvarende uttrykk for $\mathrm{E}(Y-a|Y \geq a)$.

(a) For menn, beregn forventet levealder ved fødsel og forventet gjenstående levealder ved alder $a = 25$, 50 og 85 år.

(b) Beregn så tilsvarende forventet levealder ved fødsel og forventede gjenståendene levealder for kvinner. Sammenlign med verdiene du fikk for menn.

(c) Lag til slutt et plott med forventet gjenstående levealder for $a = 0, 1, 2, \ldots, 106$, for kvinner og menn i samme plott. Kommentér resultatene du får.

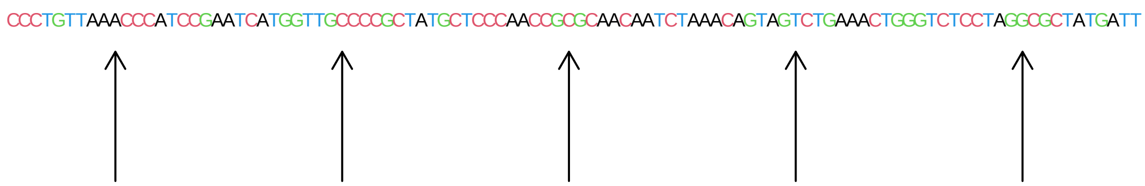
Oppgave 4

Vi skal i denne oppgaven se på bruken av DNA-bevis i forbindelse med kriminalsaker. Utover ren sannsynlighetsregning vil denne oppgaven ha endel utfordringer med hensyn på forståelse av problemstillinger. Det oppfordres derfor sterkt å bruke gruppetimene til å diskutere de ulike delspørsmål med medstudenter og gruppelærere!

Menneskenes DNA består av sekvenser av 4 ulike nukleotider som typisk beskrives ved bokstavene A,G,C og T. Figuren nedenfor viser et eksempel på en liten del av en DNA-sekvens:

DNA kan for eksempel trekkes ut fra blodspor på et åsted. Ofte blir DNA-spor fremstilt som helt sikre bevis. Dette er basert på at alle personer har forskjellige DNA. I kriminal-saker brukes imidlertid kun små deler av den fullstendige DNA-sekvensen, og da kan likheter mellom individer opptre. I slike tilfeller er det viktig å kvantifisere bevisbyrden som ligger i et DNA-spor. Denne oppgaven vil ta opp noen aspekter ved slik kvantifisering. Vi vil her gjøre en kraftig forenkling ved å anta at alle bokstaver A/G/C/T opptrer med like stor sannsynlighet og at det er uavhengighet mellom bokstaver som opptrer på ulike posisjoner.

Anta vi ser på q ulike posisjoner (markører) langs DNA-sekvensen (typisk ulike posisjoner i DNA'et som ligger langt fra i hverandre, slik at uavhengighetsantagelsen kan være rimelig). La $\mathcal{S}$ være en spesifikk kombinasjon av nukleotider på de q posisjonene. I resten av oppgaven vil vi betegne settet av nukleotider for de q utvalgte posisjonene, for *DNA-profilen* til en person (mens det egentlig bare representerer biter av personens DNA). I figuren nedenfor viser pilene 5 ulike markører der da DNA-profilen beskrevet av disse markørene er ACCTG.



(a) Hvor mange mulige typer profiler er mulig med q markører?

Hva er sannsynligheten for at en vilkårlige person har en spesifikk DNA profil $\mathcal{S}$?

Beregn denne sannsynligheten for $q = 5, q = 10$ og $q = 20$.

Hvis vi har en populasjon på $N = 5\,500\,000$ individer, hva blir forventet antall personer som har $\mathcal{S}$ for de ulike verdiene av q?

Merk: Rekkefølgen spiller en rolle her, slik at profilene ACCTG og CATGC er forskjellige.

Anta nå at vi jobber med en kriminalsak der vi har observert en DNA-profil $\mathcal{S}$ (bestående av q nukleotider) på et åsted. Bidragsyter til et DNA-spor på et åsted kan i utgangspunktet være hvem som helst innenfor en populasjon på N personer, der vi antar i utgangspunktet alle personene i populasjonen like sannsynlige som bidragsytere. Med bidragsyter mener vi her den som har etterlatt seg DNA-sporet (blod) på åstedet. Vi har også en mistenkt person.

Definer begivenhetene

A =Mistenkt er bidragsyter;

B =Mistenkt har DNA-profil $\mathcal{S}$.

Vi vil anta at man klarer å lese riktig DNA-profil både fra sporet på åstedet og fra en blodprøve av den mistenkte[1].

[1] I praksis kan det være noe usikkerhet her, men vi ignorerer dette for enkelthetsskyld.

(*b*) Argumenter hvorfor det er rimelig å anta at $P(B|A) = 1$.

Beregn også $P(B|A')$.

Hint: Når A ikke inntreffer så svarer begivenheten B til at en tilfeldig person har profil $\mathcal{S}$.

(*c*) Uttrykt ved N og q, hva er sannsynligheten for at den mistenkte er bidragsyter, betinget på at det er matchende profil (mistenkte har DNA-profil $\mathcal{S}$), dvs $P(A|B)$?

Beregn spesifikt sannsynligheten for $N = 5\,500\,000$ og $q = 5, 10$ og 20.

Basert på verdiene du får ut, men også forventningene du beregnet i (*a*), diskuter resultatene.

Anta fremdeles et DNA-spor av type $\mathcal{S}$ er funnet på et åsted. Vi vil nå se på en situasjon der vi *ikke* har noen mistenkte i utgangspunktet. Vi har imidlertid tilgjengelig en database bestående av $n \leq N$ personer der DNA-profilene til disse n individene er kjent. For enkelthetsskyld vil vi anta at de n personene i databasen er trukket tilfeldig ut av de N personene i hele populasjonen[2].
Vi vil spesielt se på en situasjon der vi søker i databasen og ender opp med å finne kun *ett* individ som har samme DNA-profil $\mathcal{S}$ som vi fant i blodsporet på åstedet. Vi ønsker da å kvantifisere bevisbyrden for dette. Dette vil vi gjøre gjennom flere trinn nedenfor.
Definer begivenheten

$C =$ bidragsyter er et av individene i databasen.

(*d*) La X være antall personer med spor $\mathcal{S}$ innen databasen.

Hvis C *ikke* har inntruffet, forklar hvorfor en binomisk fordeling kan være en rimelig sannsynlighetsfordeling for X.

Spesifiser spesielt sannsynligheten for at $X = 1$ gitt C'.

For $n = 30\,000$, plott fordelingen til X for $q = 5, 10, 20$.

Hint: Hvis du plotter fordelingen for alle mulige x-verdier så vil det være vanskelig å se noe mønster. Du bør derfor konsentrere plottet mot regioner der det meste av sannsynlighetsmassen er. Du bør da velge de x-verdiene du plotter punktsannsynlighetene for, med omhu. Bruk forventet antall som utgangspunkt.

(*e*) Argumentér for at $P(C) = \frac{n}{N}$.

Hva blir $P(X = 1|C)$?

Hint: Når du vet C så vet du noe om ett av individene i databasen men ikke noe om de øvrige. Hva betyr $X = 1$ for de øvrige individer?

[2] I de fleste land er slike databaser bygget opp av personer som av en eller annen grunn har vært i kontakt med politiet. Noen av antagelsene som blir gjort i denne oppgaven kan da være noe tvilsomme.

(f) Utled et uttrykk $P(C|X = 1)$.

Argumenter med ord for at dette svarer til sannsynligheten for at individet med matchende DNA-profil innen databasen er bidragsyter til sporet på åstedet.

Se spesielt på $P(C|X = 1)$ når $n = 1$ og $n = N$. Finner du resultatene rimelige?

(g) For $n = 30\,000$, beregn $P(C|X = 1)$ for de samme verdier for $q = 5, 10, 20$.

Diskutér resultatene.