Første sett med obligatoriske oppgaver i STK1100 våren 2014

Dette er det første settet med obligatoriske oppgaver i STK1100 våren 2014. Oppgavesettet består av tre oppgaver. Den første oppgaven inneholder både teoretiske punkter, diskusjonspunkter og punkter der du bør bruke MATLAB. Den andre oppgaven er en ren teorioppgave. Den tredje oppgaven er mest teori men også litt bruk av MATLAB. Det er valgfritt om du vil skrive besvarelsen for hånd eller om du vil bruke et tekstbehandlingsprogram. Uansett skal resultater og figurer fra MATLAB-kjøringene i oppgavene 1 og 3 tas med i besvarelsen på en hensiktsmessig måte. Hvis flere studenter samarbeider om å løse oppgavene, må likevel hver student levere sin *selvstendige* besvarelse. Det må gå fram av besvarelsen hvem du har samarbeidet med. Se ellers "Regelverk for obligatoriske oppgaver" som er gitt på kursets hjemmeside.

Besvarelsen leveres ved ekspedisjonen til Matematisk institutt, 7. etasje, Niels Henrik Abels hus.
Frist for innlevering er torsdag 6. mars kl. 14.30.

Oppgave 1

I filen `dod.txt` under "Data" på hjemmesiden til STK1100, vår 2014, er det gitt en tabell over ettårige dødssannsynligheter for norske kvinner og menn basert på dødelighetsstatistikk for 1993.

Du skal i denne oppgaven gjøre en del beregninger med utgangspunkt i denne dødelighetstabellen.

Hvordan du får tallene i tabellen inn i MATLAB kan avhenge litt av operativsystem og versjon av programmet. Her en variant som virket på Linux:

- Klikk på linken til tabellen for å få fram tallene. Lagre filen med navnet `dod.txt` f.eks. på dette området `~/matlab` (som du kanskje må lage først).
- Siden det er noe tekst i filen før selve dataene, kan du ikke bruke `load` for å lese inn dataene. Du må i stedet bruke den mer generelle funksjonen `textread`. Gå inn i MATLAB og les inn dataene med kommandoen

  ```
  X = textread('matlab/dod.txt', '', 'headerlines', 7);
  ```

  (Opsjonen `headerlines` angir størrelsen på filens "header", altså hvor mange linjer som skal hoppes over før dataene leses inn.)
- Det er greit å ha hver kolonne i tabellen i en egen variabel. Det kan du gjøre ved kommandoene

  ```
  alder = X(:,1); menn = X(:,2); kvinner = X(:,3);
  ```

- Hvis du har problemer med å lese inn dataene, kan du spørre en medstudent, en av vaktene på PC-stua eller en av lærerne i kurset.

(*a*) Dødelighetstabellen gir ettårige dødssannsynlighetene pr. 1000 individer. Divider tallene i tabellen med 1000 slik at du får sannsynligheter pr. individ. Ta vare på de ettårige dødssannsynlighetene i variablene `qmenn` og `qkvinner`.

(*b*) Lag plott av de ettårige dødssannsynlighetene som funksjon av alder for både kvinner og menn. Diskuter hvordan dødeligheten endrer seg med alder og hvordan den er forskjellig for de to kjønn. Her er noen råd som kan være til nytte når du plotter:

- For at du lett skal kunne sammenligne dødeligheten til kvinner og menn er det best å gi begge i samme plott. Bruk kommandoen `hold on` etter første plotting for at det neste plottet ikke skal erstatte eksisterende plott. For at nye plott igjen skal erstatte eksisterende plott, bruk `hold off`.
- Da dødeligheten stiger sterkt med alder, kan det være en fordel å plotte denne på logaritmisk skala. For å plotte med logaritmisk y-akse, bruk `semilogy`. Funksjonen `semilogy` brukes akkurat som `plot`, men bruker en logaritmisk y-akse (skriv `help semilogy` og `help plot` for flere detaljer).
- Prøv om du kan få til andre ting som gjør plottene mer informative. For eksempel kan du bruke ulike symboler for kvinner og menn (se hjelpeteksten for kommandoen `plot`) og gi informativ tekst langs aksene.

(*c*) I punkt (*a*) beregnet du ettårige dødssannsynligheter for kvinner og menn. De ettårige overlevelsessannsynlighetene er gitt som én minus de ettårige dødssannsynlighetene. Beregn de ettårige overlevelssannsynlighetene for kvinner og menn og ta vare på disse i variablene `pmenn` og `pkvinner`.

(*d*) Sannsynligheten for at en nyfødt skal bli minst k år kalles overlevelsessannsynligheten til alder k år. Denne er gitt som produktet av de ettårige overlevelsessannsynlighetene for alle aldre fra 0 år til og med $k-1$ år. Forklar hvorfor dette er tilfellet. Beregn overlevelsessannsynlighetene til alder $k = 1, 2, \ldots, 100$ for kvinner og menn og ta vare på disse i variablene `smenn` og `skvinner`. Bruk funksjonen `cumprod` til å gjøre dette (`cumprod` står for "cumulative product").

(*e*) Det er hensiktsmessig å ha overlevelsessannsynlighetene til aldrene $k = 0, 1, 2, \ldots, 99$ i stedet for til de k-verdiene vi beregnet i punkt (*d*). Du oppnår dette ved å lage en ny kolonnevektor hvor det første elementet er 1 og de resterende elementene er de 99 første elementene fra vektorene beregnet i forrige punkt. For eksempel `smenn=[1;smenn(1:99)]`. (Merk at det med dette blir overensstemmelse mellom kolonnen for alder og kolonnene med overlevelsessannsynlighetene.)

(*f*) Lag plott av overlevelsessannsynlighetene fra punkt (*e*) som funksjon av alder for både kvinner og menn. Diskuter hvordan overlevelssannsynlighetene avhenger av alder og hvordan de er forskjellig for de to kjønn. Bestem i denne forbindelse spesielt median levealder, dvs. den alderen det er 50% sannsynlig å overleve til.

(*g*) La den tilfeldige variabelen X betegne levealderen til en nyfødt jente/gutt. Punktsannsynlighetene $p(k) = \Pr\{X = k\}$ for $k = 0, 1, 2, \ldots, 99$ gir sannsynlighetene for de ulike verdiene X kan anta. (Vi ser i dette punktet bort fra muligheten for å bli 100 år eller mer.) Du får beregnet disse punktsannsynlighetene ved å multiplisere (elementvis) kolonnen med ettårige dødssannsynligheter fra punkt (*a*) og kolonnen med overlevelsessannsynligheter fra punkt (*e*).

Forklar hvorfor dette er tilfellet og foreta beregningene. Ta vare på punktsannsynlighetene for kvinner og menn i nye variable og lag plott av disse. Kommenter plottene.

(h) Beregn forventet levealder for norske kvinner og menn. Hvordan er disse sammenlignet med median levealder? Kommenter.

(i) Dødeligheten har endret seg siden 1993 og en tilsvarende fil for 2009 ligger på `dod2009.txt` der første kolonne angir alder og 2. og 3. kolonne dødssansynligheter pr. 1000 for henholdsvis menn og kvinner (merk at det ikke er noen header på denne filen så innlesningen må gjøres litt anderledes). Gjennomfør beregninger av levetider for 2009 tilsvarende de du allerede har gjort for 1993 og angi hovedtrekkene i forskjellene.

(j) Gitt at det nå ser ut til å være variasjoner fra år til år med hensyn på dødelighet, hvilke problemer kan det lage i forhold til beregning av sannsynlighetsfordeling til X?

Oppgave 2

Heldigvis kan legene ofte teste om en person har en bestemt sykdom eller ikke. De fleste slike tester har en viss usikkerhet knyttet til seg. Det vil si at testen kan gi positivt utslag selv om personen er frisk, eller at testen gir negativt utslag selv om personen er syk.

Tuberkulose er en alvorlig infeksjonssykdom som ofte rammer lungene, og kan avsløres ved hjelp av røntgenbilder. I Norge er tuberkulose sjelden, men den har en økt hyppighet i mange land. Vi skal se på en undersøkelse fra USA gjort på 50 tallet (Brown og Hollander, 1977).

Røntgenbilder gir ikke et 100 % sikkert svar angående tuberkulose, slik at noen ganger kan man tro at en person har tuberkulose selv om personen ikke hadde det, eller at personen ikke har tuberkulose selv om personen har det.

Undersøkelsen viste at hvis en person led av tuberkulose ble dette oppdaget på røntgen i 73.3 % av tilfellene. Undersøkelsen viste også at selv om en person var frisk, viste røntgenundersøkelsen tegn på tuberkulose i 2.8 % av tilfellene.

I studien fra USA ble andelen av befolkningen som har tuberkulose anslått til å være 1.7 %.

(a) Hva er sannsynligheten for at testen vil vise positiv?

(b) Finn sannsynligheten for at en person er syk, gitt at testen viste positiv.

(c) I dag er tuberkulose en sjeldnere sykdom enn det den var på 50 tallet. La oss anta at andelen av befolkningen som har tuberkulose nå er anslått til være 1 per 1000. Hva er nå sannsynligheten for at en person er syk, gitt at testen viste positiv? Sammenlign med resultatet i (b) og kommenter. Hvorfor tror du de aller fleste som tester positiv i (c) er friske?

Oppgave 3

Vi skal i denne oppgaven se på bruken av DNA-databaser i forbindelse med kriminalsaker. Anta et DNA spor av type $\mathcal{S}$ er funnet på et åsted. Vi vil ta som utgangspunkt at det er en populasjon på $N = 5\,000\,000$ individer som er mulige bidragsytere til sporet. Videre vil vi anta at $n = 30\,000$ individer ligger inne i databasen. Anta også det er $M = 50$ i hele populasjonen som har DNA av type $\mathcal{S}$.

(*a*) La X være antall med spor $\mathcal{S}$ innen databasen. Finn den eksakte sannsynlighetsfordelingen til X. Begrunn svaret og spesifiser hvilke antagelser som ligger bak. Lag et plot av sannsynslighetsfordelingen.

(*b*) Forklar hvorfor en binomisk fordeling kan være en god tilnærming for sannsynlighetsfordelingen til X. Plot denne tilnærmingen sammen med den eksakte fordeling for å demonstrere at tilnærmingen er god.

(*c*) Beregn sannsynlighetsfordelingen for at $X = 1$. Hvordan kan denne sannsynligheten tolkes?

(*d*) Anta at alle individer i populasjonen i utgangspunktet er like sannsynlige som bidragsyter. La A være begivenheten at bidragsyter er et av individene i databasen. Hva er $P(A)$?

(*e*) Finn $P(X = 1|A)$.

Hint: Når vi vet at bidragsyter er i databasen så er det $M - 1 = 49$ igjen som vi ikke vet om er i databasen eller ikke. Argumenter for at vi da er interessert i sannsynligheten for at ingen av disse er i databasen.

(*f*) Finn $P(A|X = 1)$. Argumenter for at dette svarer til sannsynligheten for at individet med matchende DNA profil innen databasen er den riktige. Diskuter hvorfor dette svaret virker intuitivt rimelig.

Merk: Her kan du få litt forskjellig svar avhengig av om du putter inn numeriske verdier for $P(A)$, $P(X = 1|A)$ og $P(X = 1)$ eller om du først regner ut en formel uttrykt ved N, M og n. Dette har med at MATLAB tilsynelatende bruker den binomiske tilnærmingen i utregningene.