Andre sett med obligatoriske oppgaver i STK1100 våren 2014

Dette er det andre settet med obligatoriske oppgaver i STK1100 våren 2014. Oppgavesettet består av tre oppgaver. Den første oppgaven er stort sett en teorioppgave. Den andre oppgaven inneholder både teoretiske punkter, diskusjonspunkter og punkter der du må bruke MATLAB eller annen programvare. Den tredje oppgaven er først og fremst en oppgave som må utføres på datamaskin. Det er valgfritt om du vil skrive besvarelsen for hånd eller om du vil bruke et tekstbehandlingsprogram. Uansett skal resultater og figurer fra kjøringene oppgavene 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 8. mai kl. 14.30.

Oppgave 1

La X være årsinntekten til en tilfeldig valgt person i en befolkningsgruppe. Det er vanlig å anta at X er Pareto-fordelt, det vil si at X har sannsynlighetstettheten

$$f_X(x) = \begin{cases} \theta\kappa^\theta x^{-\theta-1} & \text{for} \quad x > \kappa \\ 0 & \text{ellers.} \end{cases}$$

Her er κ minsteinntekten i den befolkningsgruppen vi betrakter, mens $\theta > 2$ er en parameter som avhenger av lønnsforskjellene i gruppen.

(a) Vis at den kumulative sannsynlighetsfordelingen til X er gitt ved

$$F_X(x) = \begin{cases} 1 - \kappa^\theta x^{-\theta} & \text{for} \quad x > \kappa \\ 0 & \text{ellers.} \end{cases}$$

Bestem median årsinntekt.

(b) Finn $E(X)$.

(c) I dette punktet antar vi at $\kappa = 200\,000$ kroner og $\theta = 2.5$. Hva blir da median årsinntekt og forventet årsinntekt? Hvilken av de størrelsene gir etter din mening best uttrykk for den "typiske årsinntekten" i befolkningsgruppen?

(d) Finn variansen og standardavviket til X.

(e) Basert på den kumulative sannsynlighetsfordelingsfunksjonen, beskriv en metode for å simulere fra Pareto fordelingen. Bruk denne metoden til å beregne (estimere) forvetning, varians og standardavvik til X ved hjelp av stokastisk simulering og sammenlikn med dine tidligere resultater.

(*f*) La $Y = \theta \ln(X/\kappa)$. Bestem sannsynlighetstettheten til Y. Hvilken kjent sannsynlighetstetthet er dette?

Oppgave 2

La $X_1, X_2, \ldots, X_n$ være uavhengige og identisk fordelte stokastiske variabler (dvs. alle X_i-ene har samme fordeling). Sett $\mu = E(X_i)$ og $\sigma^2 = \mathrm{Var}(X_i)$.

Innfør gjennomsnittet

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1}$$

og det standardiserte gjennomsnittet

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma} \tag{2}$$

På forelesningene har vi vist at $E(\overline{X}_n) = \mu$ og $\mathrm{Var}(\overline{X}_n) = \sigma^2/n$.

(*a*) Vis at $E(Z_n) = 0$ og $\mathrm{Var}(Z_n) = 1$.

Formålet med oppgaven er å studere fordelingen til det standardiserte gjennomsnittet (2) for ulike fordelinger av X_i-ene og ulike verdier av n. Spesielt vil vi være interessert i å undersøke hvor godt (eller dårlig!) fordelingen til det standardiserte gjennomsnittet kan tilnærmes med standardnormalfordelingen, og hvordan denne tilnærmingen avhenger av n og av fordelingen til X_i-ene.

Konkret vil vi la $n = 3$, 10 og 30, og vi vil betrakte fordelingene:

- Den uniforme fordelingen med sannsynlighetstetthet

$$f(x) = \begin{cases} 1 & \text{for } 0 \le x \le 1 \\ 0 & \text{ellers} \end{cases} \tag{3}$$

- Eksponensialfordelingen med sannsynlighetstetthet

$$f(x) = \begin{cases} e^{-x} & \text{for } x \ge 0 \\ 0 & \text{for } x < 0 \end{cases} \tag{4}$$

- Bernoulli fordelingen med punktsannsynlighet $p(x) = P(X = x)$ gitt ved

$$p(x) = 1/2 \quad \text{for } x = 0, 1 \tag{5}$$

(*b*) Bestem $\mu = E(X_i)$, $\sigma^2 = \mathrm{Var}(X_i)$ og $\sigma = \sqrt{\mathrm{Var}(X_i)}$ for de tre fordelingene.

Du skal bruke stokastisk simulering i dine undersøkelser. Framgangsmåten er som følger. For gitt fordeling (en av de tre over) og gitt verdi av n (lik 3, 10 eller 30) trekker du ved hjelp av datamaskin n "observasjoner" fra den aktuelle fordelingen og beregner det standardiserte gjennomsnittet (2). Dette gjentas 10 000 ganger. Du får med dette 10 000 "observasjoner" av Z_n.

(*c*) Forklar hvorfor et (normert) histogram av de 10 000 "observasjonene" av Z_n vil ligge nær sannsynlighetstettheten/punktsannsynlighetene til Z_n.

Du kan enkelt foreta stokastisk simulering ved hjelp av MATLAB. Vi beskriver framgangsmåten for den uniforme fordelingen (3) når $n = 3$:

- Først setter du $n = 3$ og trekker $n \times 10\,000$ uniformt fordelte "observasjoner" som du tar vare på i matrisen `X`:

  ```
  n = 3;
  X = unifrnd(0,1,n,10000);
  ```

  Elementene i én kolonne i `X` gir deg de $n = 3$ "observasjonene" i én simulering, mens de 10 000 kolonnene gir deg resultatene av 10 000 simuleringer.

- Du beregner så gjennomsnittet av de $n = 3$ "observasjonene" i hver kolonne i `X` og tar vare på gjennomsnittene i `meanX` ved

  ```
  meanX= mean(X);
  ```

  Radvektoren `meanX` vil inneholde 10 000 "observasjoner" av $\overline{X}_n$ gitt ved (1).

- Du beregner endelig de standardiserte gjennomsnittene Z_n gitt ved (2) og tar vare på disse i `Z` ved

  ```
  Z = sqrt(n)*(meanX-μ)/σ;
  ```

  (Merk at uttrykket er skrevet generelt. Du må erstatte μ og σ med de verdiene du fant for disse i punkt (*b*) for den uniforme fordelingen.)

(*d*) Utfør kommandoene gitt ovenfor. Lag et (ikke normert) histogram med klassebredde 0.25 av de standardiserte gjennomsnittene ved kommandoen `hist(Z,-3:0.25:3)`. Kommenter histogrammets utseende.

For å sammenligne den empiriske fordelingen til Z_n med standardnormalfordelingen, vil vi beregne de relative frekvensene av "observasjoner" av Z_n i intervallene $(-\infty, -2.5)$, $[-2.5, -2.0)$, $[-2.0, -1.5)$, $[-1.5, -1.0)$, $[-1.0, -0.5)$, $[-0.5, 0)$, $[0, 0.5)$, $[0.5, 1.0)$, $[1.0, 1.5)$, $[1.5, 2.0)$, $[2.0, 2.5)$ og $[2.5, \infty)$ og sammenligne disse med sannsynlighetene for at en standardnormalfordelt variabel vil falle i de samme intervallene.

(*e*) Bruk Table A.3 i appendikset i læreboka eller MATLAB til å bestemme sannsynlighetene for at en standardnormalfordelt variabel vil falle i intervallene gitt over.

(*f*) Beregn de relative frekvensene av verdier av Z_n i de samme intervallene. Du kan gjøre dette ved kommandoene:

```
int = [-Inf, -2.5, -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2, 2.5, Inf];
ant = histc(Z, int);
relfrekv = ant(1:12)/10000;
```

Sammenlign de relative frekvensene med sannsynlighetene du fant i punkt (*e*). Kommenter.

(g) Gjør punktene (d) og (f) om igjen for den uniforme fordelingen når $n = 10$ og når $n = 30$. Kommenter de resultatene du nå får, og sammenlign med de du fant i (d) og (f).

(h) Gjennomfør en tilsvarende simuleringsstudie for eksponensialfordelingen (4) som den du gjorde for den uniforme i punktene (d), (f) og (g). Kommenter de resultatene du nå får for ulike verdier av n. (Bruk `exprnd(1,n,10 000)` for å trekke $n \times 10\,000$ "observasjoner" fra eksponentialfordelingen med parameter 1.)

(i) Utfør en tilsvarende simuleringsstudie for Bernoulli-fordelingen (5) som for de to andre fordelingene. Kommenter de resultatene du nå får for ulike verdier av n. (Bruk `binornd(1,0.5,n,10000)` for å trekke $n \times 10\,000$ "observasjoner" fra Bernoulli fordelingen med "suksessannsynlighet" 1/2.)

(j) Sammenlign til slutt de resultatene du har fått for de tre fordelingene.

Oppgave 3
Overlevelsestiden til 9 mus er gitt nedenfor. Datasettet er hentet fra Efron and Tibshirani [1993].

52 104 146 10 51 30 40 27 46

La X_i være i-te overlevelsestid og anta $X_1, \ldots, X_9$ er uavhengige identisk fordelte med sannsynlighetfordeling F. Vi er interessert i både forventning og median for fordelingen F. De naturlige estimater er empirisk gjennomsnitt og median:

$$\overline{x} = 56.22, \quad median(x_1, \ldots, x_9) = 46.0.$$

Det empiriske standardavviket til $x_1, \ldots, x_9$ er 42.48 som indikerer en stor variabilitet i dataene. Vi er interessert i egenskapene til disse estimatorene.

(a) Utfør ikke-parametrisk Bootstrapping til å finne standard feil og forventningsskjevhet for de to estimatorene. Kommenter resultatene, spesielt at det ser ut som forventningsskevheten til gjennomsnittet er svært lavt.

(b) Lag histogram av dine simuleringer. Kommenter. Prøv spesielt å forklare den merkelige formen på histogrammet relatert til medianen.

(c) Anta nå at X_i ene er log-normal fordelte. Utfør en parametrisk Bootstrapping. Kommenter forskjeller fra resultatene du har fått tidligere.

References

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap.* Chapman & Hall, New York, 1993.