Projektopgave 1
"Sandsynlighedsregning og statistik"

Michael Andersen – michael@diku.dk
Henrik Jensen – henrikgjensen@gmail.com
Ulrik Bonde – bonde@diku.dk

30. november 2009

Opgave 1

I bilag A.1 er kildekoden til vores **R**-program inkluderet. Funktionen `assignment1()` løser den første opgave. Vi bruger et *for-loop* til at gå gennem en vektor, hvori vi har defineret de tre antalsparametre. Koden er i øvrigt flot indenteret og selvdokumenterende.

Centralt i funtionen `assignment1()` har vi metoden fra **R** kaldet `dbinom(x, n, p)`. Denne metode er defineret som binomialfordelingens sandsynlighedsfunktion som er givet ved

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \tag{1}$$

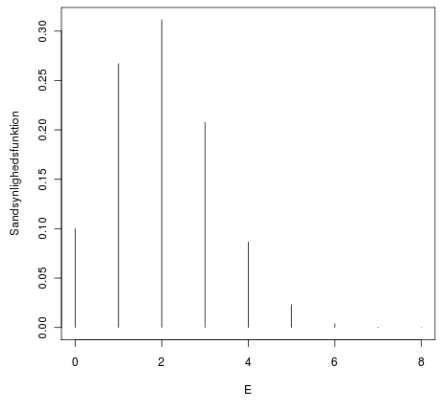
Hvis x er en vektor vil **R** returnere en vektor med resultatet fra ligning 1. Vi kan da plotte resultatet med metoden `plot()` i **R**.

I programmets løkke sættes for hver iteration et nyt filnavn til plottet. Vi eksporterer plottet til et png-billede. I figur 1 ses resultatet fra programmet.

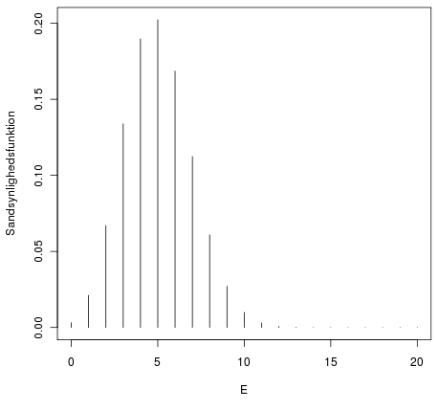
Opgave 2

På samme måde som ovenstående opgave, bruger vi et for-loop til at simulere 1000 udfald af binomialfordelinger med antalsparameter 8, 20 og 100 og sandsynlighedsparameter $\frac{1}{4}$. Metoden kan findes i kildekoden og er kaldt `assignment2_3_5()` da de resterende programmeringsopgaver afhænger af de simuleringer vi kommer frem til.

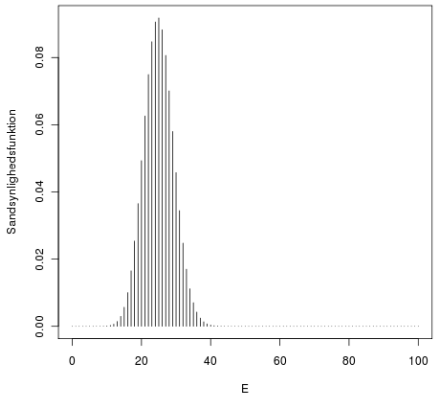
En simulering laves ved at bruge metoden `rbinom(n, size, prob)` i **R**, hvor n er antallet af observationer, $size$ er antalsparameteren og $prob$ er sandsynligheden for et gunstigt udfald.



(a) $n = 8$ og $p = 0.25$



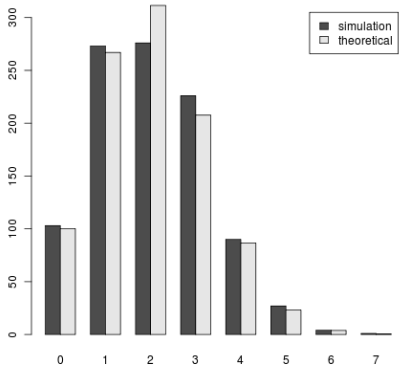
(b) $n = 20$ og $p = 0.25$



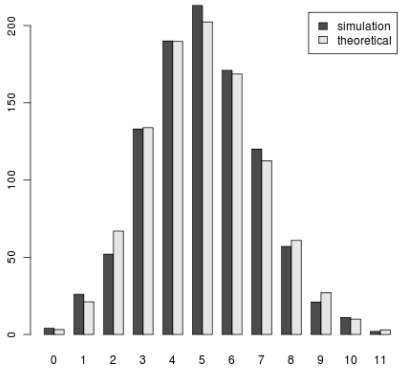
(c) $n = 100$ og $p = 0.25$

Figur 1: Resultater fra binomialfordelinger med antalsparameter 8, 20 og 100 og sandsynlighedsparameter $\frac{1}{4}$

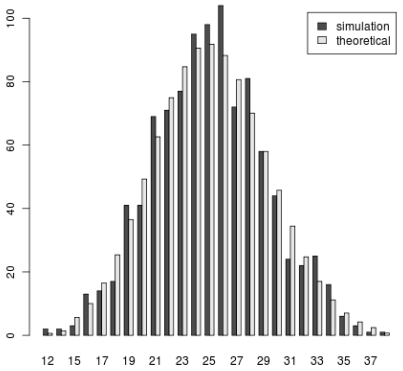
Vi ønsker nu at plotte de observerede værdier for at se dem i forhold til de teoretiske værdier. Disse kan ses i figur 2. Det ses at de simulerede værdier ligger tæt op ad de teoretiske værdier. Værdierne er ikke helt ens, men de holder sig i nærheden. Plottene viser kun de observerede værdier, hvilket kan forklare symmetrien i disse plots. Ved at sammeligne med figur 1 ses det at der kun er gjort meget små observationer af de udfald med lille sandsynlighed. Det mest usædvanlige plot ses i figur 2c, hvor $x = 24$, $x = 25$ og $x = 26$ har en lidt større observation end den teoretiske. En nærmere analyse med et konfidensinterval er ikke blevet foretaget for at fastslå om de virkelig *er* bemærkelsesværdige. De resterende observationer viser ikke denne tendens.



(a) $n = 8$ og $p = 0.25$



(b) $n = 20$ og $p = 0.25$



(c) $n = 100$ og $p = 0.25$

Figur 2: Barplot af simulerede værdier og de teoretiske værdier.

Opgave 3

Metoden `assignment2_3_5()` printer ved kørsel gennemsnittet og den empiriske varians for de simulerede kørsler ud. Dette gøres ved metoder `mean()` og `var()` i **R**. Endvidere bruges metoden `paste()` til at manipulere tekststrenge. Vi får følgende udskrift:

```
[1] Mean for simulation with size 8 is: 2.029
[1] Var for simulation with size 8 is: 1.60376276276276
[1]
[1] Mean for simulation with size 20 is: 4.997
[1] Var for simulation with size 20 is: 3.55654754754755
[1]
```

```
[1] Mean for simulation with size 100 is: 25.087
[1] Var for simulation with size 100 is: 17.6550860860861
```

Middelværdien for en binomialfordeling $\mathbf{X}$ med antalsparameter n og sandsynlighedsparamter p er givet som $\mathrm{E}(\mathbf{X}) = np$ og variansen er givet ved $\mathrm{Var}(\mathbf{X}) = np(1-p)$. For $\mathbf{X} \sim \text{bin}\,(n,p)$, hvor $n = 8$ og $p = 0.25$, får vi da

$$\mathrm{E}(\mathbf{X}) = np = 8 \cdot 0.25 = 2 \tag{2}$$
$$\mathrm{Var}(\mathbf{X}) = np(1-p) = 8 \cdot 0.25(1-0.25) = 1.5 \tag{3}$$

For $\mathbf{X} \sim \text{bin}\,(n,p)$, hvor $n = 20$ og $p = 0.25$, får vi

$$\mathrm{E}(\mathbf{X}) = 20 \cdot 0.25 = 5 \tag{4}$$
$$\mathrm{Var}(\mathbf{X}) = 20 \cdot 0.25(1-0.25) = 3.75 \tag{5}$$

Endelig for $\mathbf{X} \sim \text{bin}\,(n,p)$, hvor $n = 100$ og $p = 0.25$, får vi

$$\mathrm{E}(\mathbf{X}) = 100 \cdot 0.25 = 25 \tag{6}$$
$$\mathrm{Var}(\mathbf{X}) = 100 \cdot 0.25(1-0.25) = 18.75 \tag{7}$$

Det ses at middelværdien for binomialfordelingerne ligger meget tæt på det beregnede gennemsnit for simulationerne. Den empiriske varians ligger ligeledes tæt på variansen af binomialfordelingen, men dog med større procentvis afvigelse end middelværdi mod gennemsnit.

Opgave 4

a)

Vi betragter k uafhængige stokastiske variable med antalsparameter n og sandsynlighedsparameter p, dette vil sige

$$\mathbf{X}_i \sim \text{bin}\,(n,p), \text{ for } i = 1, \ldots, k$$

Der skal nu findes middelværdien $\mathrm{E}(\overline{\mathbf{X}})$, hvor $\overline{\mathbf{X}} = \frac{1}{k}\sum_{i=1}^{k} \mathbf{X}_i$, da $\mathbf{X}_i$ er binomialfordelt med antalsparameter n og sandsynlighedsparameter p er $\mathrm{E}(\mathbf{X}_i) = np$. Derfor følger udregningen

$$\mathrm{E}\overline{\mathbf{X}} = \mathrm{E}\left(\frac{1}{k}\sum_{i=1}^{k} \mathbf{X}_i\right)$$

$$= \frac{1}{k}\sum_{i=1}^{k} \mathrm{E}(\mathbf{X}_i)$$

$$= \frac{1}{k} knp$$

$$= np$$

Vi skal nu finde det samme for variansen, dette gøres påfølgende måde, vi ved at $\text{Var}(\overline{\mathbf{X}}) = \text{E}(\overline{\mathbf{X}}^{\mathbf{2}}) - \text{E}(\overline{\mathbf{X}})^2$ og der vides at $\text{Var}(\mathbf{X}_i) = np(1-p)$.

$$\text{Var}(\overline{\mathbf{X}}) = \text{Var}(\frac{1}{k}\sum_{i=1}^{k}\mathbf{X}_i)$$

$$= \frac{1}{k^2}\sum_{i=1}^{k}\text{Var}(\mathbf{X}_i)$$

$$= \frac{1}{k^2}\sum_{i=1}^{k}np(1-p)$$

Her benyttes sætning 3.7.13.

$$= \frac{1}{k^2}knp(1-p)$$

Derfor er

$$\text{Var}(\overline{\mathbf{X}}) = \frac{np(1-p)}{k}$$

b)

Den empiriske varians, kan findes ved at kigge på opgave 3.26, og beskrives som

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^2$$

Da vi betragter k stokastiske variable fremfor n, har vi altså at

$$s^2 = \frac{1}{k-1}\sum_{i=1}^{k}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^2$$

Vi skal nu finde middelværdien af dette, hvilket vil sige at vi skal kigge på

$$\text{E}(s^2) = \text{E}\left(\frac{1}{k-1}\sum_{i=1}^{k}\mathbf{X}_i - \overline{\mathbf{X}}\right)^2$$

$$= \frac{1}{k-1}\sum_{i=1}^{k}\text{E}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^2$$

$$= \frac{1}{k-1}\sum_{i=1}^{k}\text{E}\left(\mathbf{X}_i^2 + \overline{\mathbf{X}}^2 - 2\mathbf{X}_i\overline{\mathbf{X}}\right)$$

Vi kan ved at bruge sætning 3.7.6 og 3.7.4, dele ovenstående op og tage middel værdien af hvert enkelt led, således at vi har

$$= \frac{1}{k-1} \sum_{i=1}^{k} \left(\mathrm{E}\left(\mathbf{X}_i^2\right) + \mathrm{E}\left(\overline{\mathbf{X}}^2\right) - 2\mathrm{E}\left(\mathbf{X}_i \overline{\mathbf{X}}\right) \right)$$

Vi betragter nu kun det første led, altså $E(\mathbf{X}_i^2)$. Til dette ved vi at der i henhold til 3.7.9 gælder at

$$\mathrm{Var}\left(\mathbf{X}\right) = \mathrm{E}\left(\mathbf{X}^2\right) - \left(\mathrm{E}\left(\mathbf{X}\right)\right)^2$$

Her isolerer vi $\mathrm{E}\left(\mathbf{X^2}\right) = \mathrm{Var}(\mathbf{X}) + \left(\mathrm{E}(\mathbf{X})\right)^2$ fra første del af opgave 4 ved vi at $\mathrm{Var}\left(\mathbf{X}_i\right) = np(1-p)$ og $\left(\mathrm{E}\left(\mathbf{X}_i\right)\right)^2 = n^2p^2$ Derfor har vi at

$$\mathrm{E}\left(\mathbf{X}_i^2\right) = np(1-p) + n^2p^2$$

Det samme gør sig næsten gældende for ledet $\mathrm{E}(\overline{\mathbf{X}}^2)$, her har vi fra første del af opgaven at

$$\mathrm{Var}\left(\overline{\mathbf{X}}\right) = \frac{np(1-p)}{k} \text{ og } \left(\mathrm{E}\left(\overline{\mathbf{X}}\right)\right)^2 = n^2p^2$$

Derfor har vi at

$$\mathrm{E}\left(\overline{\mathbf{X}}^2\right) = \frac{np(1-p)}{k} + n^2p^2$$

Der betragtes nu det sidste led som har formen

$$\mathrm{E}\left(\mathbf{X}_i \overline{\mathbf{X}}\right)$$

$$= \mathrm{E}\left(\mathbf{X}_i \frac{1}{k} \sum_{j=1}^{k} \mathbf{X}_j \right)$$

$$= \frac{1}{k} \sum_{j=1}^{k} \mathrm{E}\left(\mathbf{X}_i \mathbf{X}_j\right)$$

Her kan vi i henhold til sætning 3.7.7 gøre følgende, da $\mathbf{X}_i \perp\!\!\!\perp \mathbf{X}_j$

$$= \frac{1}{k} \left(\sum_{\substack{j=1 \\ j \neq i}}^{k} \left(\mathrm{E}\mathbf{X}_i \mathrm{E}\mathbf{X}_j\right) + \mathrm{E}\left(\mathbf{X}_i^2\right) \right)$$

Vi tager nu resultaterne fra første halv del af opgaven og bruger her.

$$= \frac{1}{k} \left(\sum_{\substack{j=1 \\ j \neq i}}^{k} \left(n^2p^2 + np\left(1-p\right) + n^2p^2\right) \right)$$

$$= \frac{1}{k} \left(\left(k-1\right) n^2p^2 + np\left(1-p\right) + n^2p^2 \right)$$

Disse resultater indsættes nu i vores oprindelige formel

$$\frac{1}{k-1}\sum_{i=1}^{k}\left(\mathrm{E}\left(\mathbf{X}_i^2\right)+\mathrm{E}\left(\overline{\mathbf{X}}^2\right)-2\mathrm{E}\left(\mathbf{X}_i\overline{\mathbf{X}}\right)\right)$$

$$=\frac{1}{k-1}\sum_{i=1}^{k}\left(np(1-p)+n^2p^2+\frac{np(1-p)}{k}+n^2p^2-2\frac{1}{k}\left((k-1)\,n^2p^2+np\,(1-p)+n^2p^2\right)\right)$$

$$=\frac{1}{k-1}\left(\left(knp\,(1-p)+kn^2p^2+np(1-p)+kn^2p^2\right)-2\left((k-1)\,n^2p^2+np\,(1-p)+n^2p^2\right)\right)$$

$$=\frac{1}{k-1}\left(knp\,(1-p)+kn^2p^2+np\,(1-p)+kn^2p^2-2kn^2p^2+2n^2p^2-2np\,(1-p)-2n^2p^2\right)$$

$$=\frac{1}{k-1}\left(knp\,(1-p)-np\,(1-p)\right)$$

$$=\frac{1}{k-1}\left((k-1)\,np\,(1-p)\right)$$

$$=np\,(1-p)$$

Opgave 5

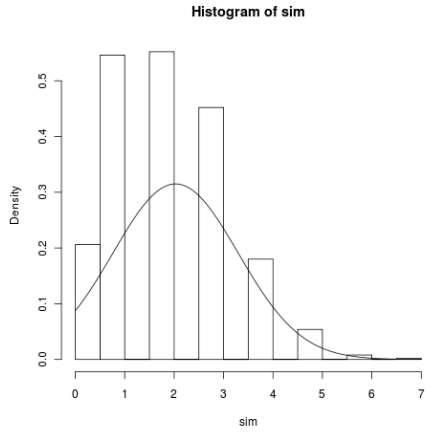
Vi får at vide, at binomialfordelingen tilnærmer sig normalfordelingen for store værdier af n. Dette ses i figur 3. For $n = 8$ ses det at de observerede værdier fra simulationen ikke følger normalfordelingen. Ved $n = 20$ ligger vi tættere på normalfordelingen. For $n = 100$ ligger observationerne stort set på niveau med normalfordelingen, men vi lægger også her mærke til at der i intervallet $[24, 26)$ er et højt antal observationer, som vi så i opgave 2. Også intervallet $[30, 32)$ ligger udenfor normalfordelingen, men her er der et færre antal observationer.

Opgave 6

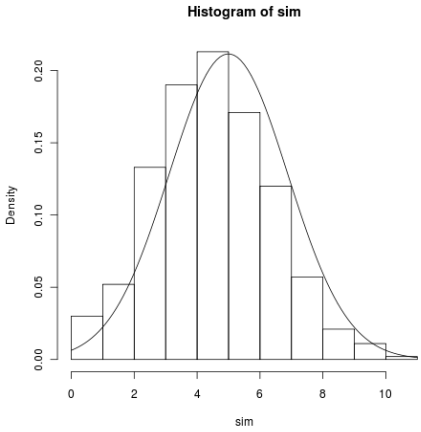
Da der ligger et uendeligt antal handsker på stranden antager den stokastiske variable $\mathbf{X}$ værdier i udfaldsrummet $E = \{0, \ldots, n\}$, $n \to \infty$. Det oplyses at der er lige mange højre og venstre handsker, derfor er sandsynligheds parameteren $p = 0.5$

I det konkrete tilfælde med $n = 14$, har vi altså

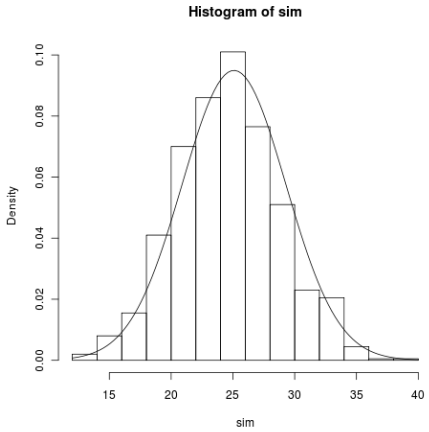
$$\mathbf{X} \sim \mathrm{bin}(14, 0.5)$$



(a) $n = 8$ og $p = 0.25$



(b) $n = 20$ og $p = 0.25$



(c) $n = 100$ og $p = 0.25$

Figur 3: Histogram og normalfordeling

Opgave 7

Der skal beregnes sandsynligheden for at få netop 10 venstrehandsker og 4 højrehandsker. I henhold til den opstillede statistiske model. Dette gøres ved at bruge formlen $\binom{n}{k}p^k(p-1)^{n-k}$.

Derfor har vi at

$$\binom{14}{10}0.5^{10}(1-0.5)^{14-10} = \binom{14}{10}0.5^{14} = 0.061$$

Sandsynligheden for udfaldet 10 venstrehandsker og 4 højrehandsker er derfor ca. 6%.

Opgave 8

Vi skal finde sandsynligheden for at finde 10 eller flere handsker til samme hånd i den nænvte model, hvor der er lige mange højre- og venstrehandsker.

Det vil sige at vi er interesseret i sandsynligheden for

$$P(\mathbf{X} \leq 4, \mathbf{X} \geq 10) = P(\mathbf{X} \leq 4) + P(\mathbf{X} \geq 10) = P(\mathbf{X} \leq 4) + (1 - P(\mathbf{X} \leq 9)) = 2 \cdot P(\mathbf{X} \leq 4)$$

Det sidste kan lade sig gøre da sandsynlighedsfunktionen er symmetrisk omkring middelværdien 7

$$P(\mathbf{X} \leq 4) + (1 - P(\mathbf{X} \leq 9)) = 0.0898 + 0.0898 = 0.18$$

Så i henhold til den opstillede statistiske model er der 18% sandsynlighed for at finde 10 eller flere af samme type handske.

Opgave 9

Vi skal nu kommentere på hvor vidt det årvågne strandvandrepars fund er bemærkelsesværdigt eller ej. For at undersøge dette laver vi et 95%-konfidensinterval for p, hvis 0.18 er inkluderet i dette interval er fundet ikke bemærkelsesværdigt. Formlen der skal benyttes er som følger

$$p \pm 1.96 \cdot s(p), \text{ hvor } s(p) = \sqrt{\frac{p(1-p)}{n}}$$

Vi har $p = 0.5$ og $n = 14$, udregningen er derfor som følger.

$$0.5 \pm 1.96 \cdot 0.134 = (0.238, 0.762)$$

Da

$$0.18 \notin (0.238, 0.762)$$

må fundet af de 10 venstrehandsker være bemærkelsesværdigt.

A Source code

A.1 main.r

```
#!/usr/bin/env Rscript

# R statistics for the course "Sandsynlighedsregning og statistik"
# at the Department of Mathematical Sciences, University of Copenhagen

########################################
#
# Project 1
#
########################################

# Assignment 1
assignment1 <- function() {
    # We define n as an array with the values 8, 20 and 100
    n    <- c(8, 20, 100);
    prob <- 0.25;

    # Traverse the n-array
    for (size in n) {
        # Define E
        E <- 0:size;

        # Define the probability function
        sshfunk <- dbinom(E, size, prob);

        # Set the filename for the plot and plot the graph
        name <- paste(size, "_plot_1.png", sep="");
        png(name);
        plot(E, sshfunk, "h", ylab="Sandsynlighedsfunktion");
        dev.off();
    }
}

########################################

# Assigment 2, 3, and 5
assignment2_3_5 <- function() {
    # Define the simulation using a random binom
    n    <- 1000;
    size <- c(8, 20, 100);
    prob <- 0.25;

    for (s in size) {
        # Do a simulation and count the observations of each occurrence
        sim <- rbinom(n, s, prob);
        simulation <- table(sim);

        # Assignment 3
        # Print mean and var
        print(paste("Mean for simulation with size ",s, " is: ", mean(sim), sep=""));
        print(paste("Var for simulation with size ",s, " is: ", var(sim), sep=""));
        print("");

        # Get the values coresponding to an occurrence
        values <- as.numeric(names(simulation));

        theoretical <- 1000*dbinom(values, s, prob);

        # Combine the simulation and the theoretical vectors
        comb <- rbind(simulation, theoretical);

        # Assignment 2
        # Set up the filename for the plot, then draw it
        name <- paste(s, "_sim-theo_plot_2.png", sep="");
        png(name);
        barplot(comb, beside=TRUE, legend.text=TRUE);

        # Assignment 5
```

```
69          # Set up the filename for the histogram, draw and close pipe
70          histname <- paste(s, "_hist_5.png", sep="");
71          png(histname);
72          hist(sim, prob=TRUE);
73          curve(dnorm(x, mean = mean(sim), sd = sd(sim)), add=TRUE);
74          dev.off();
75      }
76  }
77
78  ########################################
79  # Main
80
81  assignment1();
82  assignment2_3_5();
```