Exercise 5 Solution

Applied Statistics, IT University of Copenhagen

1. Octrahedral Die (T)

Let T be the outcome of roll of fair octahedral die.

(a) Describle the probability distribution of T, that is, list the outcomes and the corresponding probabilities.

Solution: An octahedral die has 8 sides therefore the sample space is:

$$\Omega = \{1, 2, 3, 4, 5, 6, 7, 8\}$$

Since it is a fair die, every side is equally probable:

$$P(T = t) = \frac{1}{8}$$

for $t \in 1, \cdots, 8$

(b) Determine the expected value and variance of T.

Solution: Following the formula in p. 90, Dekking et al. the expectation of a discrete random variable X:

$$E[X] = \sum_i a_i P(X = a_i) = \sum_i a_i p(a_i)$$

We In this case we get:

$$E[X] = \sum_{a=1}^{8} \frac{a}{8} = \frac{1}{8} \sum_{a=1}^{8} a = \frac{1}{8}[1 + 2 + 3 + \cdots + 8] = \frac{36}{8} = 4.5$$

And the variance of a random variable X (p. 97, Dekking et al.):

$$Var(X) = E[(X - E[x])^2] = E[X^2] - (E[X])^2$$

Here we find

$$E[X^2] = \frac{1}{8} \sum_{a=1}^{8} a^2 = \frac{1}{8}[1^2 + 2^2 + 3^2 + \cdots + 8^2] = \frac{204}{8} = 25.5$$

$$Var(X) = E[X^2] - (E[X])^2 = 25.5 - (4.5)^2 = 5.25$$

2. Expectation and Variance of a Continuous Random Variable (T)

Let X be a continuous random variable with the density function

$$f_X(x) = \begin{cases} x+1 & \text{if } -1 \leq x < 0 \\ -x+1 & \text{if } 0 \leq x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Compute the expectation and variance of X.

Solution: The expectation of a continous random variable X with probability density function f is defined by (p. 91, Dekking et al.):

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

$$E[X] = \int_{-1}^{0} x(x+1)dx + \int_{0}^{1} x(-x+1)dx = \int_{-1}^{0} x^2 + xdx + \int_{0}^{1} x - x^2 dx \quad (1)$$

$$= \left[\frac{1}{3}x^3 + \frac{1}{2}x^2\right]_{-1}^{0} + \left[\frac{1}{2}x^2 - \frac{1}{3}x^3\right]_{0}^{1} \quad (2)$$

$$= \frac{1}{3} - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} = 0 \quad (3)$$

The variance is defined by (p. 97, Dekking et al.):

$$Var(X) = E[X^2] - E[X]^2$$

Since we already know that $(E[X])^2 = 0$ we only need to calculate $E[X^2]$:

$$E[X^2] = \int_{-1}^{0} x^2(x+1)dx + \int_{0}^{1} x^2(-x+1)dx \quad (4)$$

$$= \int_{-1}^{0} x^3 + x^2 \, dx + \int_{0}^{1} x^2 - x^3 \, dx \quad (5)$$

$$= \left[\frac{1}{4}x^4 + \frac{1}{3}x^3\right]_{-1}^{0} + \left[\frac{1}{3}x^3 - \frac{1}{4}x^4\right]_{0}^{1} \quad (6)$$

$$= -\frac{1}{4} + \frac{1}{3} + \frac{1}{3} - \frac{1}{4} = \frac{1}{6} \quad (7)$$

So $Var(X) = \frac{1}{6}$

3. Linearity of the Expectation Operator (T)

Show that the expectation operator is linear; that is, for functions $f, g : \mathbb{R} \to \mathbb{R}$, applied on the random variable X, and any scalars $\alpha, \beta \in \mathbb{R}$,

$$\mathrm{E}\left[\alpha f(X) + \beta g(X)\right] = \alpha \,\mathrm{E}\left[f(X)\right] + \beta \,\mathrm{E}\left[g(X)\right].$$

Consider the cases where

(a) X is a discrete random variable taking values $a_1, a_2, \ldots \in \mathbb{R}$,

Solution: The fact that the expectation operator is linear in the discrete case follows from the fact that a sum is linear.

$$\mathrm{E}\left[\alpha f(X) + \beta g(X)\right] = \sum_i p(a_i)(\alpha f(a_i) + \beta g(a_i)) \tag{8}$$

$$= \alpha \sum_i p(a_i) f(a_i) + \beta \sum_i p(a_i) g(a_i) = \alpha \,\mathrm{E}\left[f(X)\right] + \beta \,\mathrm{E}\left[g(X)\right] \tag{9}$$

(b) X is a continuous random variable taking values on the real axis.

Solution: The argument for the continous case is excatly the same. (i.e the integral of a sum is the sum of the integrals and you can take constants "outside" integrals)

$$\mathrm{E}\left[\alpha f(X) + \beta g(X)\right] = \int_{-\infty}^{\infty} p(x)(\alpha f(x) + \beta g(x))dx \tag{10}$$

$$= \alpha \int_{-\infty}^{\infty} p(x) f(x) dx + \beta \int_{-\infty}^{\infty} p(x) g(x) dx = \alpha \,\mathrm{E}\left[f(X)\right] + \beta \,\mathrm{E}\left[g(X)\right] \tag{11}$$

4. Transforming a Random Variable (T)

Given is a random variable X with the probability density function f given by $f(x) = 0$ for $x < 0$, and for $x > 1$, and $f(x) = 4x - 4x^3$ for $0 \leq x \leq 1$.

(a) Determine the distribution function F_X.

Solution: Lets start by calculating $F(x)$ in the interval where $f(x) \neq= 0$

$$F_X(x) = \int_0^x (4x - 4x^3)dx = \left[2x^2 - x^4\right]_0^x = 2x^2 - x^4$$

So we conclude that that

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 2x^2 - x^4 & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}$$

(b) Let $Y = \sqrt{X}$. Determine the distribution function F_Y.

Solution:

From p. 103, Dekking gives info on a random variable transforms if we apply a function to it i.e $Y = g(X)$

By Taking the Inverse we find that

$$X = g^{-1}(Y) = Y^2$$

Now following the example in Dekking p 105 we can write.

$$F_Y(y) = P(Y \leq y) = P(\sqrt{X} \leq y) = P(X \leq y^2) = F_X(y^2)$$

We then substitute that into the distribution function and get the expression for F_Y(y):

$$F_Y(y) = 2(y^2)^2 - (y^2)^4 = 2y^4 - y^8$$

(c) Determine the probability density of Y.

Solution: Again, using relation between the distribution function F and the probability density function f (p. 59, Dekking et al.) we simply differentiate $F_Y(y)$

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}(2y^4 - y^8) = 8y^3 - 8y^7$$

5. Accessing Data and Numeric Summaries (R)

(a) Take `Cars93` (MASS) data set. What is the type of the Cylinders variable? What does the summary command do for the Cylinders variable? Get the names of the cars having 8 cylinders. What is the mean horsepower of the cars having 8 cylinders, how about standard deviation? How about those for the cars having 6 cylinders? Is the result what you expect? *Solution:* The type of the Cylinders variable is 3 4 5 6 8 androtary

```
library('MASS')
typeof(Cars93$Cylinders)
```

```
## [1] "integer"
```

```
Cars93$Cylinders
```

```
##  [1] 4      6      6      6      4      4      6      6      6      8
## [11] 8      4      4      6      4      6      6      8      8      6
## [21] 4      6      4      4      4      6      4      6      4      6
## [31] 4      4      4      4      4      6      6      8      3      4
## [41] 4      4      4      4      4      4      4      8      6      6
## [51] 6      8      4      4      4      6      rotary 4      6      4
## [61] 6      4      6      4      4      6      6      4      4      6
## [71] 6      4      4      4      6      6      6      4      4      3
## [81] 4      4      3      4      4      4      4      4      5      4
## [91] 6      4      5
## Levels: 3 4 5 6 8 rotary
```

The summary of Cylinders gives the number of cars with each of the cylinder type (3,4,5,6,8 or rotary)

```
summary(Cars93$Cylinders)
```

```
##      3      4      5      6      8 rotary
##      3     49      2     31      7      1
```

There is 7 cars having 8 cylinders (Assume model name):

```
cylin8 <- subset(Cars93, Cylinders == 8)
cylin8$Model
```

```
## [1] DeVille        Seville        Caprice        Corvette       Crown_Victoria
## [6] Q45            Town_Car
## 93 Levels: 100 190E 240 300E 323 535i 626 850 90 900 Accord ... Vision
```

The mean and standard deviation of horsepower of the cars having 8 cylinders

```
mean(cylin8$Horsepower)
```

```
## [1] 234.7143
```

```
sd(cylin8$Horsepower)
```

```
## [1] 54.42645
```

The same for cars with 6 cylinders

```
sub2 <- subset(Cars93, Cylinders == 6)
mean(sub2$Horsepower)
```

```
## [1] 175.5806
```

```
sd(sub2$Horsepower)
```

```
## [1] 32.33344
```

Yes the mean horsepower is larger for cars with more cylinders.

(b) For the `precip` data set, find the mean and standard deviation of the rain fall over cities. Find all the cities with the average annual rain fall exceeding 50 inches. Which cities are the dryest? Does this match your expectation?

Solution: The mean and standard deviation of the rain fall over cities:

```
mean(precip)
```

```
## [1] 34.88571
```

```
sd(precip)
```

```
## [1] 13.70665
```

The cities with the average annual rain fall exceeding 50 inches:

```
subset(precip, precip > 50)
```

```
##       Mobile       Juneau Jacksonville        Miami  New Orleans     San Juan
##         67.0         54.7         54.5         59.8         56.8         59.2
```

The dryest cities:

```
subset(precip, precip < 10)
```

```
##     Phoenix        Reno Albuquerque     El Paso
##         7.0         7.2         7.8         7.8
```

Yes these are cities in the south.

(c) The `rivers` contains the lengths of the 141 major rivers in North America. Compare the mean and 25% trimmed mean on the data set. What does the result tell you? How big is the standard deviation?

Solution: The fact that after trimming mean dropped significantly shows that there are more short rivers than long ones. High standard deviation value shows that river lengths are widely spread.

```
mean(rivers)
```

```
## [1] 591.1844
```

```
mean(rivers, trim = .25)
```

```
## [1] 449.9155
```

```
sd(rivers)
```

```
## [1] 493.8708
```

6. Flight Overbooking (R)

To maximise the seats occupied during flights, the airlines has the customs to overbook them. Assume that the total number of seats on a flight is 150 and the number of people showing up at the airport is a random variable $X \in 1, 2, \ldots, M$, where all the outcomes are equally probable, and M is the number of bookings made. Assume that each passenger onboard means 500 EUR cash inflow for the airline whereas each refused passenger implies 1000 EUR penalty to the airline. Operating the plane costs 40000 EUR. For how many bookings would you advice the airline to take?

Solution: Let's create two empty lists: CurrentVal for loss/profit and EX for expected loss/profit. Then we write a function which takes number of passengers and returns loss/profit value. Then we iterate from 1 to 300 passengers and populate the list of loss/profit values and expected loss/profit.

Let the profit be defined as h

$$h(X) = 500 \min(150, X) - 1000 \max(0, X - 150) - 40000$$

Now the expetation value of the profit function is

$$E[h(X)] = \sum_{i=0}^{M} p(x_i) h(x_i) = \frac{1}{M} \sum_{i=0}^{M} h(x_i)$$

So now we just need to calculate $E[g(X)]$ for each M

```
CurrentVal <- c()
EX <- c()

h <- function(x) {
  val <- min(150,x)*500 - max(0,x-150)*1000 - 40000
  return(val)
}

for (i in 1:300) {
  CurrentVal <- c(CurrentVal, h(i))
  EX <- c(EX, sum(CurrentVal)/i)
}

max_profit <- max(EX)
cat(which(EX == max_profit), "Booking are recommended", "profit is", max_profit)
```

```
## 183 Booking are recommended profit is 1401.639
```

```
plot(EX, type = "l", xlab = "Number of bookings")
```