Exercise 5

Applied Statistics, IT University of Copenhagen

T=Theoretical exercise, R=R exercise

Preparation

- Read pages 33–44, 50–55, and 60–61 from Verzani (2014).

Problems

1. Octrahedral Die (T)

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

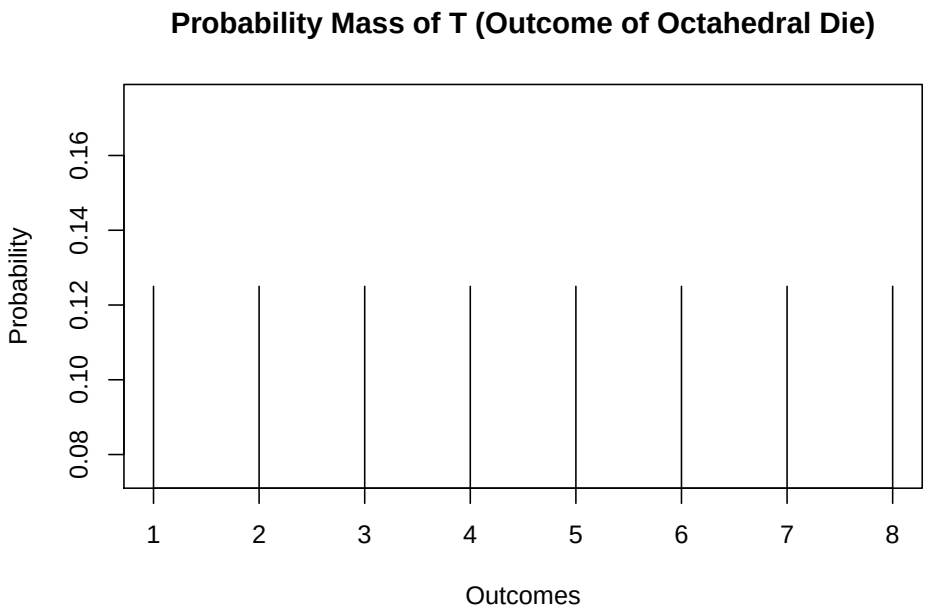
(a) Describle the probability distribution of T, that is, list the outcomes and the corresponding probabilities.

Answer\ The task describes the probabilistic experiment of the roll of a fari octahedral die that is modeled with the *continuous random variable* T describing the outcome of a single roll of this die. We can define the sample space and the probability mass function p associating probabilities to all possible single events in the sample space as follows:

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

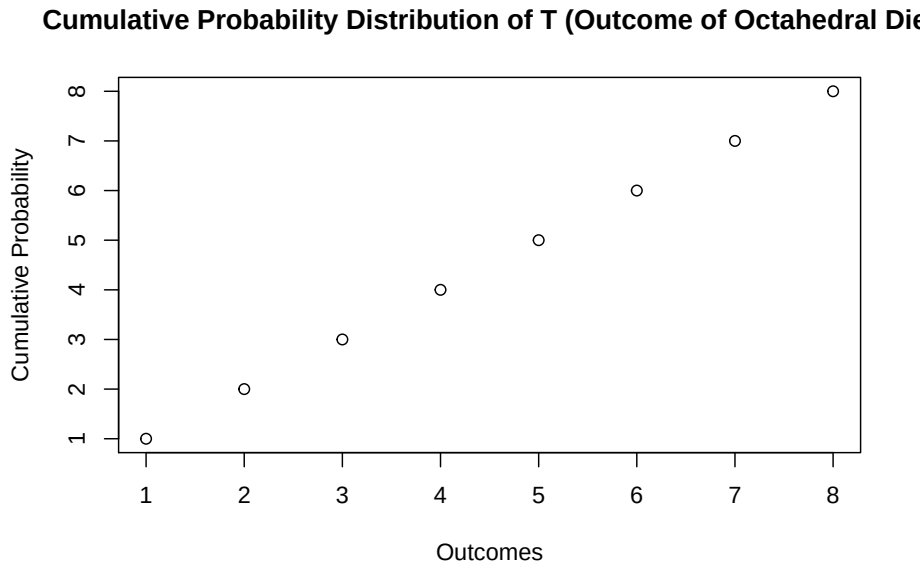
$$p(x) = \begin{cases} \frac{1}{8} \text{ for all x } \in \Omega \\ 0 \text{ else} \end{cases}$$

```
x <- 1:8
y <- c(rep(1/8, 8))
plot(x, y, main='Probability Mass of T (Outcome of Octahedral Die)', type='h', xlab='Outco
```



$$F(x) = \begin{cases} \frac{x}{8} \text{ for x on the interval of } \Omega \\ 0 \text{ else} \end{cases}$$

```
x <- 1:8
y <- c()
for (i in x) {
  append(y, i/8)
}
plot(x, y, main='Cumulative Probability Distribution of T (Outcome of Octahedral Die)', xl
```



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

Answer

Expected Value/ Mean:

$$E[T] = \sum_{i=1}^{8} i\frac{1}{8} = \frac{i}{8} = \frac{9}{2} = 4.5$$

Variance:

Variance is a measure of how spread out the values in a distribution are. In our example, a low variance means the sums that we roll will usually be very close to one another. By contrast, the variance is large when the sums that we roll are frequently distant values. The way that we calculate variance is by taking the difference between every possible sum and the mean. Then we square all of these differences and take their weighted average.

$$Var[T] = E[(T-E)^2] = E[T^2]-E[T] = \sum_{i=1}^{8} i^2 \cdot \frac{1}{8} - E[T] = (\frac{1}{8}+\frac{4}{8}+\frac{9}{8}+\frac{16}{8}+\frac{25}{8}+\frac{36}{8}+\frac{49}{8}+\frac{64}{8}) - 4.5 = 25.5-4.5 =$$

$$Var[T] = E[(T-E)^2] = \sum_{i=1}^{8} \frac{(i-4.5)^2}{8} = 2(\frac{49}{32} + \frac{25}{32} + \frac{9}{32} + \frac{1}{32}) = 5.25$$

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

Let X be a continuous random variable with the density function

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

Compute the expectation and variance of X.

Answer

Expected Value/ Mean:

From the probability density function given for X, we can easily derive that the expected value is 0, since the PDF is symmetric around the y-axis, such that $f_X(x) = f_X = (-x)$.

$$E[X] = 0$$

Variance:

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 $\alpha, \beta \in \mathbb{R}$,

$$\mathrm{E}\left[\alpha f(X) + \beta g(X)\right] = \alpha\,\mathrm{E}\left[f(X)\right] + \beta\,\mathrm{E}\left[g(X)\right]. \tag{2}$$

Consider the cases where

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

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

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

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

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.

The distribution function F_X is the antideriative of the probability density function f onb tge interval $0 \leq x \leq 1$.

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

(b) Let $Y = \sqrt{X}$. Determine the distribution function F_Y. We perform a change of variable as follows

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

(c) Determine the probability density of Y. The probability density function of Y is the derivative of the distribution function. Therefore:

$$f_Y(x) = \frac{dF_Y}{dx} F_Y(x) = \frac{dF_Y}{dx} F_X(x^2) = \frac{dF_Y}{dx} 4x^2 - 4x^6 = 8x - 24x^5$$

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?

```
library('MASS')
data(Cars93)
names(Cars93)
```

```
##  [1] "Manufacturer"       "Model"              "Type"
##  [4] "Min.Price"          "Price"              "Max.Price"
##  [7] "MPG.city"           "MPG.highway"        "AirBags"
## [10] "DriveTrain"         "Cylinders"          "EngineSize"
## [13] "Horsepower"         "RPM"                "Rev.per.mile"
## [16] "Man.trans.avail"    "Fuel.tank.capacity" "Passengers"
## [19] "Length"             "Wheelbase"          "Width"
## [22] "Turn.circle"        "Rear.seat.room"     "Luggage.room"
## [25] "Weight"             "Origin"             "Make"
```

```
head(Cars93)
```

```
##   Manufacturer   Model    Type Min.Price Price Max.Price MPG.city MPG.highway
## 1        Acura Integra   Small      12.9  15.9      18.8       25          31
## 2        Acura  Legend Midsize      29.2  33.9      38.7       18          25
## 3         Audi      90 Compact      25.9  29.1      32.3       20          26
## 4         Audi     100 Midsize      30.8  37.7      44.6       19          26
## 5          BMW    535i Midsize      23.7  30.0      36.2       22          30
## 6        Buick Century Midsize      14.2  15.7      17.3       22          31
##               AirBags DriveTrain Cylinders EngineSize Horsepower  RPM
## 1                None      Front         4        1.8        140 6300
## 2 Driver & Passenger      Front         6        3.2        200 5500
## 3        Driver only      Front         6        2.8        172 5500
## 4 Driver & Passenger      Front         6        2.8        172 5500
## 5        Driver only       Rear         4        3.5        208 5700
## 6        Driver only      Front         4        2.2        110 5200
##   Rev.per.mile Man.trans.avail Fuel.tank.capacity Passengers Length Wheelbase
## 1         2890             Yes               13.2          5    177       102
## 2         2335             Yes               18.0          5    195       115
## 3         2280             Yes               16.9          5    180       102
## 4         2535             Yes               21.1          6    193       106
## 5         2545             Yes               21.1          4    186       109
## 6         2565              No               16.4          6    189       105
##   Width Turn.circle Rear.seat.room Luggage.room Weight  Origin          Make
## 1    68          37           26.5           11   2705 non-USA Acura Integra
## 2    71          38           30.0           15   3560 non-USA  Acura Legend
## 3    67          37           28.0           14   3375 non-USA       Audi 90
## 4    70          37           31.0           17   3405 non-USA      Audi 100
## 5    69          39           27.0           13   3640 non-USA      BMW 535i
## 6    69          41           28.0           16   2880     USA Buick Century
```

```
# type of cylinder variable
typeof(Cars93$Cylinders)
```

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

```
# summary on cylinders
summary(Cars93$Cylinders)
```

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

```
# cars with 8 cylinders
eight.cylinders <- subset(Cars93, Cylinders==8)
paste(eight.cylinders$Manufacturer, eight.cylinders$Model, eight.cylinders$Type)
```

```
## [1] "Cadillac DeVille Large"    "Cadillac Seville Midsize"
## [3] "Chevrolet Caprice Large"   "Chevrolet Corvette Sporty"
## [5] "Ford Crown_Victoria Large" "Infiniti Q45 Midsize"
## [7] "Lincoln Town_Car Large"
```

```
# mean and standard deviation horsepower of cars with 8 cylinders
print(paste(mean(eight.cylinders$Horsepower), sd(eight.cylinders$Horsepower)))
```

```
## [1] "234.714285714286 54.4264466526899"
```

```
# mean and standard deviation of horsepower for cars with 6 cylinders
six.cylinders <-subset(Cars93, Cylinders==6)
print(paste(mean(six.cylinders$Horsepower), sd(six.cylinders$Horsepower)))
```

```
## [1] "175.58064516129 32.3334441855987"
```

(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?

```
data(precip)
precip[]
```

```
##              Mobile              Juneau             Phoenix         Little Rock
##                67.0                54.7                 7.0                48.5
##         Los Angeles          Sacramento       San Francisco             Denver
##                14.0                17.2                20.7                13.0
##            Hartford          Wilmington          Washington        Jacksonville
##                43.4                40.2                38.9                54.5
##               Miami             Atlanta            Honolulu               Boise
##                59.8                48.3                22.9                11.5
##             Chicago              Peoria        Indianapolis          Des Moines
##                34.4                35.1                38.7                30.8
##             Wichita           Louisville         New Orleans            Portland
##                30.6                43.1                56.8                40.8
##           Baltimore              Boston             Detroit    Sault Ste. Marie
##                41.8                42.5                31.0                31.7
##              Duluth Minneapolis/St Paul             Jackson         Kansas City
##                30.2                25.9                49.2                37.0
##            St Louis         Great Falls               Omaha                Reno
##                35.9                15.0                30.2                 7.2
##             Concord       Atlantic City         Albuquerque              Albany
##                36.2                45.5                 7.8                33.4
##             Buffalo            New York           Charlotte            Raleigh
##                36.1                40.2                42.7                42.5
##             Bismark          Cincinnati           Cleveland            Columbus
##                16.2                39.0                35.0                37.0
```

```
##       Oklahoma City            Portland        Philadelphia            Pittsburg
##                31.4                37.6                39.9                36.2
##           Providence            Columbia        Sioux Falls             Memphis
##                42.8                46.4                24.7                49.1
##            Nashville              Dallas             El Paso             Houston
##                46.0                35.9                 7.8                48.2
##      Salt Lake City          Burlington             Norfolk            Richmond
##                15.2                32.5                44.7                42.6
##      Seattle Tacoma             Spokane          Charleston           Milwaukee
##                38.8                17.4                40.8                29.1
##             Cheyenne           San Juan
##                14.6                59.2
```

```
# mean and standard deviation
mean(precip)
```

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

```
sd(precip)
```

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

```
# cities exceeding 50 inches average annual rain fall
over.50 <- precip[precip > 50]
over.50
```

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

```
# driest cities
sorted.cities <- names(sort(precip))
sorted.cities[c(1:5)] # the five driest cities
```

```
## [1] "Phoenix"     "Reno"        "Albuquerque" "El Paso"     "Boise"
```

(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?

```
data(rivers)
```

```
mean(rivers)
```

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

```
sd(rivers)
```

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

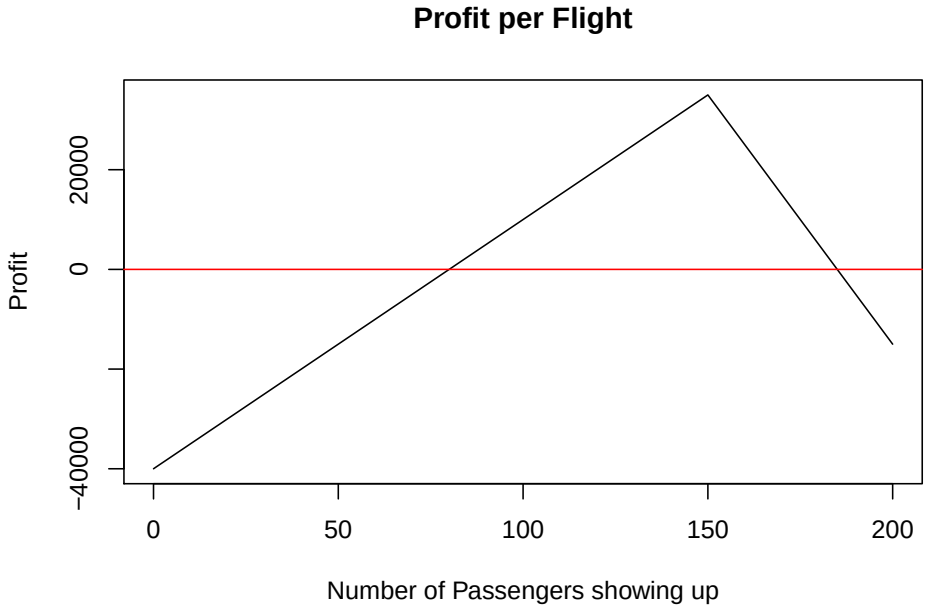
6. Flight Overbooking (R)

To maximize 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?

```
x <- 0:200
y <- c()

for (i in x){
  profit <- min(i,150)*500-max(0, i-150)*1000-40000
  y <- c(y, profit)
}
options(scipen=999)
plot(x, y, main='Profit per Flight', xlab='Number of Passengers showing up', ylab='Profit'
abline(h = 0, col = "red")
```



Verzani, John. 2014. *Using R for Introductory Statistics*. CRC Press.