Distributions in R

Introduction

In this R guide, we will see how to plot some probability distributions, calculate probabilities, quantiles and generate sample from these distributions. Please, when you use the R commands listed in this guide, think first on what you want R to do and which information you should provide in order to perform that task. Please do not copy blindly the commands, pay attention in the language and grammar used. Remember also to use the R function `help()` ton find out information about a particular command. Finally, but maybe more important, discuss and think on each result that you get in this guide, as this will help you to understand distributions and random variables.

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

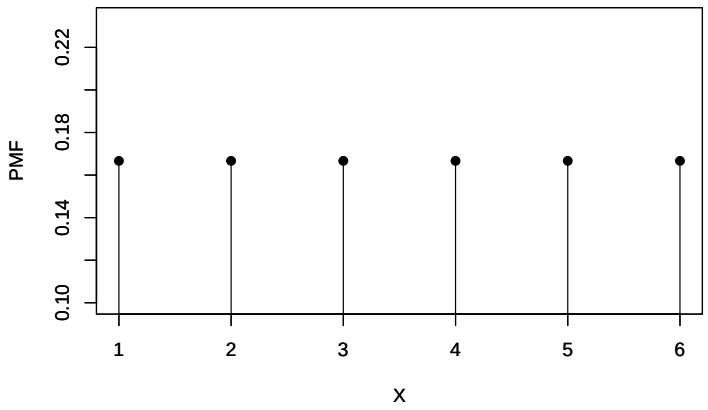
Discrete distributions

Let X be a random variable representing the outcome of rolling a fair die. To plot the probability mass function of X in R do the following (consider writing the commands in an R script and save it for future use):

```
X = c(1,2,3,4,5,6)
prob = c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)
```

The vector `X` contains all the possible values that X can take, and the vector `prob` contains the probabilities associated with each of these values.

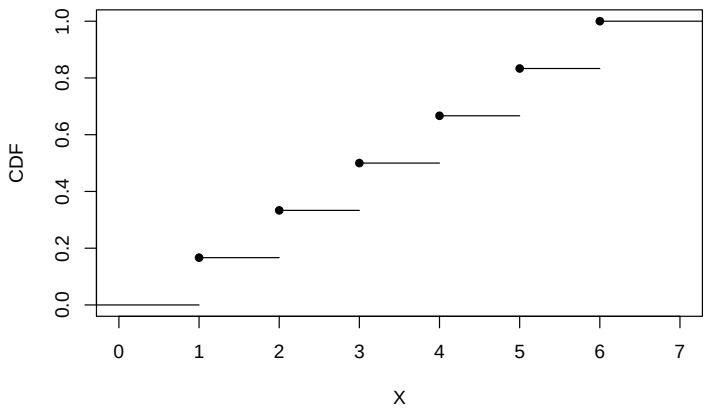
```
plot(X,prob,type="p",xlab="X",ylab="PMF", pch=19)
par(new=T)
plot(X,prob,type="h",xlab="X",ylab="PMF", pch=19)
abline(h=0)
```



To plot the cumulative distribution function, first we need to calculate the cumulative probabilities. This can be done using the command `cumsum`:

```
cdf = c(0,cumsum(prob))
```

```
cdf.plot = stepfun(X,cdf,f=0)
plot.stepfun(cdf.plot,xlab="X",ylab="CDF",verticals=FALSE,do.points=TRUE,main="",pch=16)
```



Exercise: let X be a random variable representing the total numbers of heads obtained when tossing two coins. Plot in R the probability mass function and the cumulative distribution function of X.

In R we can use the function `sample()` to generate random numbers from a discrete uniform distribution. Imagine we want to simulate, using a computer, the experiment of rolling a die. The outcome of this experiment will be an integer number between 1 and 6, with equal probability. Type in the command window of R Studio the following:

```
sample(c(1,2,3,4,5,6), 1, prob=c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6))
```

and press return. Note that your result it may be different from the one of the person sat next to you, as these are random numbers. Repeat the experiment a few times (note that in R studio if you press the "up" arrow of your keyboard the last command used will be displayed) and look at the results that you get each time. Do the results surprise you? If you keep "rolling" the die and annotating the results that you get, do you expect to observe any particular number more than others in the long term? Do you expect to observe similar frequencies for all of them? Do the following and comment on what you observe with a colleague:

```
s=sample(c(1,2,3,4,5,6), 100, replace=TRUE, prob=c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6))
table(s)
barplot(table(s), ylab="Freq")
```

Exercise: Repeat the experiment but this time simulate one thousand rolls and comment on the results with your colleagues.

Functions for the most commonly used distributions, such as the binomial distribution, are implemented in R. Using these functions we can easily determine probabilities and cumulative probabilities for a given set of values, calculate the quantiles and draw random numbers.

Consider question 1 of the tutorial sheet about discrete distributions: X is a discrete random variable that follows a binomial distribution with parameters $n = 5$ and $p = 0.4$, that is $X \sim Bin(5, 0.4)$. To calculate probabilities from a binomial distribution we can use the command `dbinom`. For example, the probability that $X = 0$ when $X \sim Bin(5, 0.4)$, can be calculated in R by doing:

```
dbinom(0, size=5, prob=0.4)
```

```
## [1] 0.07776
```

Does this match your answer? To check all the probabilities required in part (a) of question 1, do the following:

```
dbinom(c(0,1,2,3,4,5), size=5, prob=0.4)
```

```
## [1] 0.07776 0.25920 0.34560 0.23040 0.07680 0.01024
```

We have seen in class that the probability mass function of a discrete random variable must satisfy $\sum_{\forall x_i \in S} x_i P(X = x_i) = 1$. We can easily verify this by using the command `sum`:

```
sum(dbinom(c(0,1,2,3,4,5), size=5, prob=0.4))
```

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

To calculate a cumulative probability we can use the command `pbinom`. For example, to calculate $P(X \leq 2)$, do:

```
pbinom(2, size=5, prob=0.4)
```

```
## [1] 0.68256
```

Similarly, to determine quantiles we can use the command `qbinom`. For example, to find the 50% quantile (median) of a binomial distribution with parameters $n = 50$, $p = 0.7$, do:

```
qbinom(0.5, size=50, prob=0.7)
```

```
## [1] 35
```

Finally, to generate random samples from a binomial distribution we can use the command `rbinom`. For example, to generate 100 random realizations from the $B(n = 50, p = 0.7)$ do:

```
rbinom(100, size=50, prob=0.7)
```

```
##   [1] 36 37 37 36 32 32 33 32 35 35 38 37 34 36 38 32 35 29 31 32 32 31 35
##  [24] 35 29 36 35 38 32 36 38 38 32 35 38 38 34 38 35 35 38 33 32 35 40 34
##  [47] 34 34 35 38 34 35 33 38 39 30 33 27 37 33 36 27 40 38 33 38 41 34 42
##  [70] 29 34 35 31 35 30 33 33 37 33 42 35 39 39 39 31 30 36 35 39 34 36 40
##  [93] 41 37 38 39 38 33 32 30
```

You can perform similar tasks from a Poisson distribution by using the commands: `dpois`, `ppois`, `qpois`, `rpois`. Use the help to learn how to use these commands. For example, type:

```
help(dpois)
```

Exercise: suppose a country experiences 4 tropical storms on average per year. Use R to answer the following questions:

- What is the probability of suffering from only two tropical storms in a given year?
- What is the probability that not more than 2 storms are experienced in a given year?
- Simulate the number of storms that could be experienced in the following 10 years. (Note that this is not a prediction, it is just a simulation -like for example rolling the die).

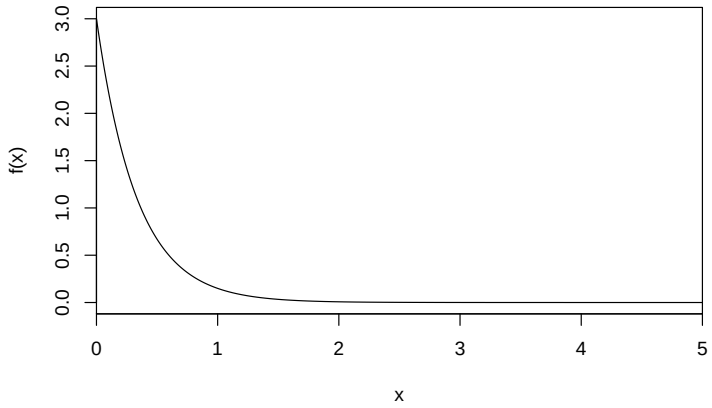
Use the help to investigate about the use of the Geometric distribution in R.

```
help.search("Geometric Distribution")
```

Constinuous distributions

Let X be a random variable with an exponential distribution and parameter $\lambda = 3$, i.e. $X \sim Exp(\lambda = 3)$. To plot the probability density function of X in R do the following (consider writing the commands in an R script and save it for future use):

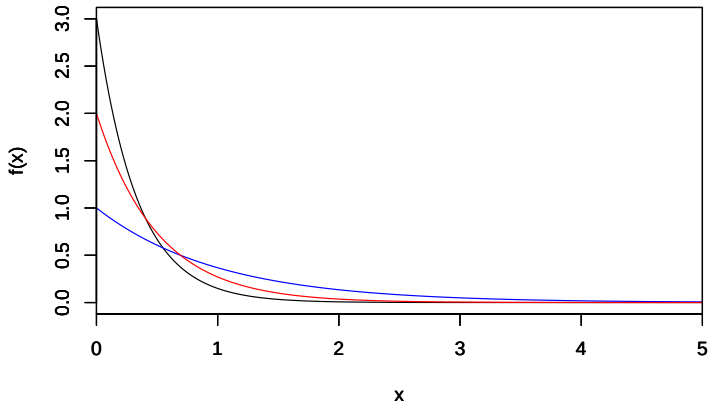
```
x=seq(0,5,0.01)
f1=dexp(x,rate=3)
plot(x,f1,type='l', ylim=c(0,3), xaxs='i', ylab='f(x)')
```



Study carefully what is doing each command introduced in R.

To investigate the effect of different parameter values on the distribution of X, add to the previous plot two additional exponential density functions with parameter values $\lambda = 1, 2$.

```
plot(x,f1,type='l', ylim=c(0,3), xaxs='i', ylab='f(x)')
par(new=T)
f2=dexp(x,rate=1)
plot(x,f2,type='l', ylim=c(0,3), xaxs='i', ylab='f(x)', col='blue')
par(new=T)
f3=dexp(x,rate=2)
plot(x,f3,type='l', ylim=c(0,3), xaxs='i', ylab='f(x)', col='red')
```



Exercise: The command `abline(v = c)` will add a vertical line at c to your plot. Use this command to add three vertical lines to your plot: one at each mean value of the densities (recall that if $X \sim Exp(\lambda)$, then $E(X) = 1/\lambda$). Draw each line with the same colour as its density. If you do not know how to use the command use the help: `help(abline)`.

Using the command `pexp()` we can evaluate cumulative probabilities for the exponential distribution. For example, let X be a random variable with exponential distribution and parameter $\lambda = 1$, that is $X \sim Exp(\lambda = 1)$, to compute $F(3) = P(X < 3)$ in R, we do:

```
pexp(3, rate=1)
```

```
## [1] 0.9502129
```

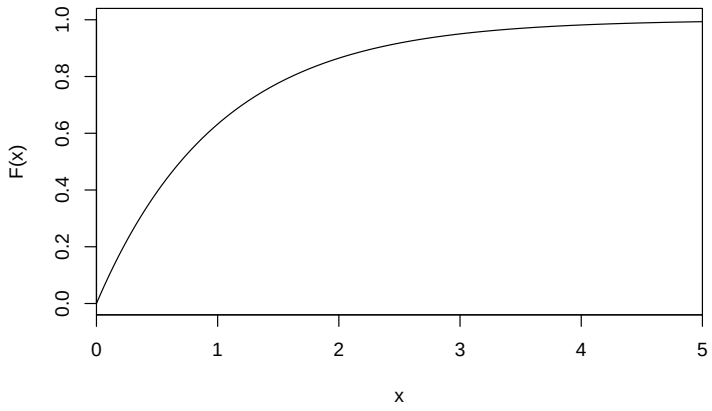
To compute $F(1), F(1.5), F(2)$ and $F(2.5)$, we can do:

```
pexp(c(1,1.5,2,2.5), rate=1)
```

```
## [1] 0.6321206 0.7768698 0.8646647 0.9179150
```

To plot the cumulative distribution function of X, where $X \sim Exp(\lambda = 1)$, do:

```
CDF=pexp(x, rate=1)
plot(x,CDF,type='l', ylim=c(0,1), xaxs='i', ylab='F(x)')
```



Recall the definition of quantiles: the value x_p for which the cumulative distribution function is

$$F(x_p) = P(X < x_p) = p \quad (0 < p < 1)$$

is called the p-quantile. In R we can use the command `qexp()` to find quantiles of the exponential distribution. For example, to find the 0.25 and 0.75 quantiles (also called first and third quartile respectively) of $X \sim Exp(\lambda = 2)$, do:

```
qexp(c(0.25, 0.75), rate=2)
```

```
## [1] 0.1438410 0.6931472
```

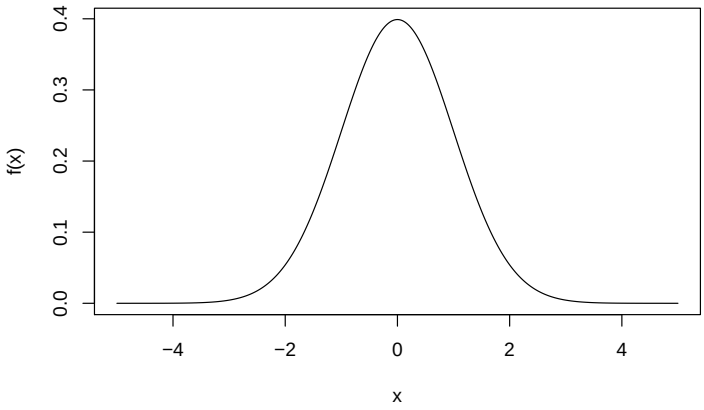
Exercise: mark in the plot of the CDF the median, m, of the distribution. Using R find the value of m, that is the median of a random variable $X \sim Exp(\lambda = 1)$.

The mean and the median can be used to identify if a distribution is positive or negative skew. If the mean is greater than the median the distribution is positively skew; while if the mean is lower than the median the distribution is negatively skew. Looking at plot of the pdf it is clear that the exponential distribution is not symmetric.

Exercise: using R find the median of $X \sim Exp(\lambda = 1)$ (second quartile) and identify if the distribution of X is positive or negative skew.

Analogous commands can be used for the normal distribution. They are: `dnorm`, `pnorm` and `qnorm`. So, for example, to plot the density function of a random variable X with distribution $N(\mu = 0, \sigma = 1)$, do:

```
x=seq(-5,5,0.01)
f1=dnorm(x,mean=0, sd=1)
plot(x,f1,type='l', ylab='f(x)')
```



Exercise: Use R to answer the following questions:

- Let X be a normal random variable with mean equal to 3 and standard deviation equal to 2. Find $P(X < 2.5)$, $P(X > 4)$ and $P(1 < X < 5)$.
- Plot the density function of X, where $X \sim N(-1, 1)$ and add two vertical lines, such that the area between these lines and below the density function is 95%.

- Let X be a normal random variable with $X \sim N(-100, 100)$. Find the first, second and third quartile.