STAT 428: Homework 1: Testing Rmd and homework submission

Du, Iris(Yuting), yutingd3

Table of Contents

Exercise 1:

For this homework, it will be easiest for you to save this .Rmd file as a template and make necessary additions and/or deletions. *(Make sure you use the naming conventions for file names and contents)*

Your tasks!

- Save this .Rmd file using the naming conventions for homework assignments.
- Open the file in RStudio
- Change the name of the author above to yours
- Add any collaborators anytime you have worked with someone on this assignment
- Click Knit HTML

- After doing all the above steps, add an appropriate section marked `Solution` for Exercise 1, remove any unnecessary text that should not be included in the report and a sentence to your solution section so that your report states that you have followed the tasks listed in Exercise 1.
- Continue doing the remaining exercises, editing the document in RStudio to write out your solutions, appropriately marked as Solutions.
- Finally, review the formating and submission instructions in the homework policy document and **Submit your homework**.

Solution 1:

This is the solution to question 1.

Exercise 2:

This exercise has several parts about RMarkdown syntax. You may look into the Rmarkdown tutorial in Compass as you work through this exercise.

1. Write two sentences below about yourself.
2. Copy and paste the above two sentences. Make one sentence **bold** and the other *italic*. Separate the two sentences by a blank line.
3. What's your favorite website?
   (a) Create a hyperlink anchored to text 'My favorite website' to your favorite website.
   (b) Create a hyperlink anchored by the URL to your favorite website.
4. What character is used to denote headers?
5. Write a quoted statistics or probability joke. *(Remember, to acknowlege the reference to where you found the joke)*. *If you are having a hard time finding one, just quote the paragraphs below:*

- Girlfriend: Our love is like a Poisson distribution, rare and special. Out of all the men in the world, we found each other.
- Boyfriend: Hmm, I think I'd describe it more like a geometric distribution. I failed with all the other women in our class but I knew there would eventually be a success…you!

6. Write Computer Type for running 2+3 in R.
7. Write a bulleted list of top 3 reasons you chose this course.
8. Write an itemized list of your top three fears in this course. You may use sub-lists if necessary!
9. Find an image on the web, or create one using R, of the hypergeometric distribution. Insert the image here.
10. Can you find a way to resize the image in RMarkdown?
11. Write three separate code chunks named A, B, C where

- code chunk A shows only results but no code

- code chunk B shows only the code but does not run it
- code chunk C neither shows the code nor the results but does run the code *(You may include simple code for adding 2 and 3 or more elaborate code like we did in class.)*

12. Using Latex, write the equation for the likelihood function $L(\theta)$ that corresponds to n i.i.d Bernoulli random variables X_i.
13. Describe in one sentence your experience doing this exercise.

Solution 2:

1. I am Iris Du. And I am a big fan of Lay Zhang.

2. **I am Iris Du.**

And I am a big fan of Lay Zhang.

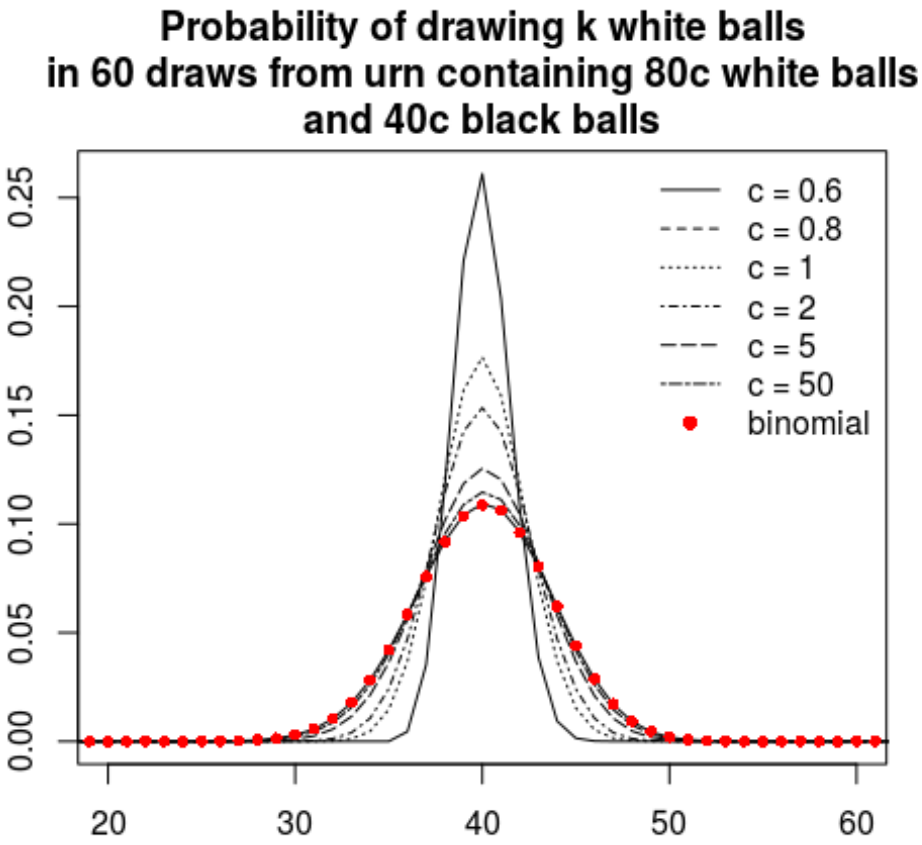
(a) My favorite website
(b) https://www.bilibili.com/

4. The head is denoted by #

5. there are three kinds of lies: lies, damned lies, and statistics.[Attributed by Mark Twain to Benjamin Disraeli , http://www.workjoke.com/statisticians-jokes.html]

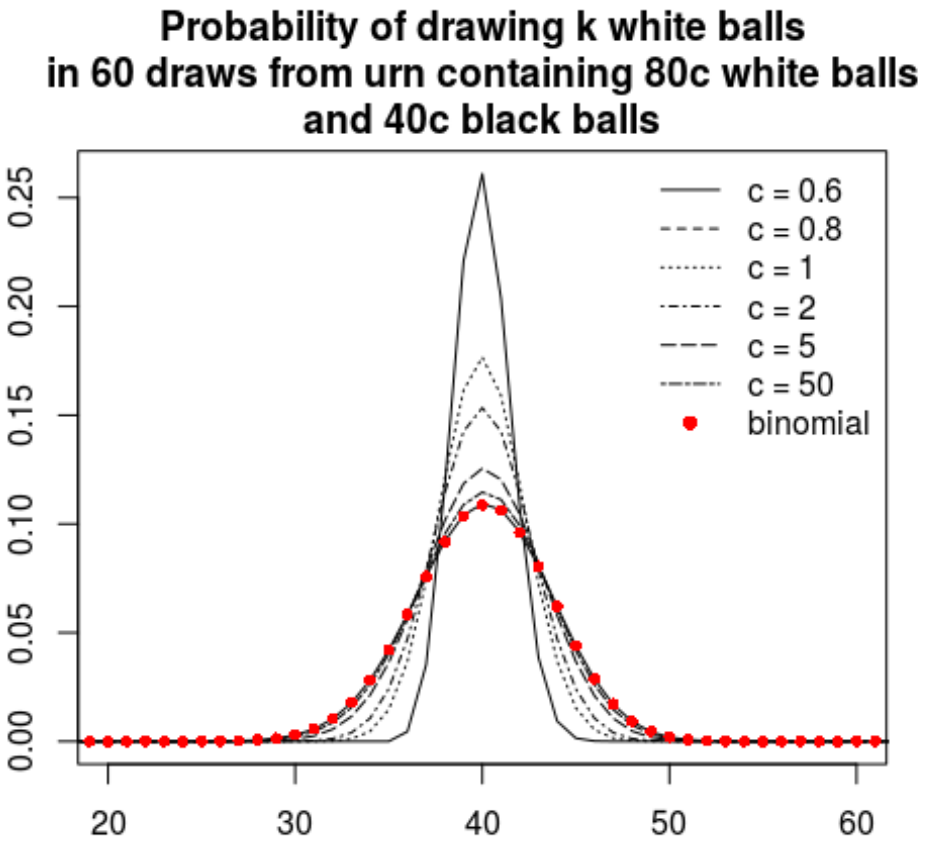
```
2+3
```

- I would like to improve my R coding skill
- I would like to know the way to gernate random variables by computer
- I love the schedule of this course
    1. The team collaboration.
    2. The statitical knowledge.
    3. The in-class participation.



8.

9. *A remote image*



10.

11. A remote image

```
## [1] 5
```

```
2 + 3
```

$$L(\theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)}$$

13. The Latex is quite annoying

Exercise 3:

Mark whether the following statements pertaining to this class are True or False. If False, make sure you know what the true statement is. You may need to refer to the Syllabus to answer the following questions.

1. **True/False**: Piazza should be used as much as possible for questions pertaining to course administration and general course information.

2. **True/False**: Email is the fastest way to get questions related to homework as well as general questions answered.
3. **True/False**: The exam dates and location are
   - Exam 1: Week 7, Thursday, Mar 5, in Foellinger
   - Exam 2: Week 15, Friday, April 26, in class
4. **True/False**: Homework and exams can be easily made up after the deadline by sending an email to the instructor or the TA requesting this. Such emails will be answered instantly.
5. **True/False**: The project will be a group project and will have only one deadline: May 03, 2019.

Solution 3:

1. **Ture**
2. **False**
3. **False**
4. **False**
5. **False**

Exercise 4:

Simple R questions: Some of the questions below can be answered with very little or no programming. However, write code that outputs the final answer and does not require any additional paper calculations. For example, suppose I ask for how many numbers are in the vector, `x=c(1,9,2,8,10,12)`. Do not count the numbers in the vector, instead have R count by coding `length(x)`. *('c()' is a function to create vector in R; 'length()' is a function to calculate the length of a vector (refer here or google).)*

R has a built-in matrix of different types of data of every US state, `state.x77`. Use this matrix and R's built in functions and **vector calculation** to perform the following tasks. You will also need `rownames(state.x77)` to get the state names and R's character functions to manipulate them (refer here or google).

(a) What is the longest state name (including spaces) and how long is it?
(b) List all the states that are more then one word. How many are there?
(c) List all the US State names, where all of the upper and lower case `a's` are replaced with a capital `Z`.
(d) Output only the second column of the matrix and store it in the numeric vector `capita`. This vector indicates the per capita income (1974) of every US states.
(e) What is the average per-capita income of the US (1974)?
(f) What is the average per-capita income of the states that have names that are more than one word?
(g) Which state has the highest per-capita income?

Solution 4:

```
rownames(state.x77)[nchar(rownames(state.x77)) ==
max(nchar(rownames(state.x77)))]
```

```
## [1] "North Carolina" "South Carolina"
```

```
max(nchar(rownames(state.x77)))
```

```
## [1] 14
```

```
state_name = rownames(state.x77)
state_name[grep(" ", state_name)]
```

```
##  [1] "New Hampshire"  "New Jersey"     "New Mexico"     "New York"
##  [5] "North Carolina" "North Dakota"   "Rhode Island"   "South
Carolina"
##  [9] "South Dakota"   "West Virginia"
```

```
length(state_name[grep(" ", state_name)])
```

```
## [1] 10
```

```
sub(pattern = "A|a", replacement = "Z", state_name)
```

```
##  [1] "Zlabama"        "Zlaska"         "Zrizona"        "Zrkansas"
##  [5] "CZlifornia"     "ColorZdo"       "Connecticut"    "DelZware"
##  [9] "FloridZ"        "GeorgiZ"        "HZwaii"         "IdZho"
## [13] "Illinois"       "IndiZna"        "IowZ"           "KZnsas"
## [17] "Kentucky"       "LouisiZna"      "MZine"          "MZryland"
## [21] "MZssachusetts"  "MichigZn"       "MinnesotZ"
"Mississippi"
## [25] "Missouri"       "MontZna"        "NebrZska"       "NevZda"
## [29] "New HZmpshire"  "New Jersey"     "New Mexico"     "New York"
## [33] "North CZrolina" "North DZkota"   "Ohio"           "OklZhoma"
## [37] "Oregon"         "PennsylvZnia"   "Rhode IslZnd"   "South
CZrolina"
## [41] "South DZkota"   "Tennessee"      "TexZs"          "UtZh"
## [45] "Vermont"        "VirginiZ"       "WZshington"     "West
VirginiZ"
## [49] "Wisconsin"      "Wyoming"
```

```
captita = as.numeric(state.x77[,2])

captita
```

```
##  [1] 3624 6315 4530 3378 5114 4884 5348 4809 4815 4091 4963 4119
5107 4458 4628
## [16] 4669 3712 3545 3694 5299 4755 4751 4675 3098 4254 4347 4508
5149 4281 5237
## [31] 3601 4903 3875 5087 4561 3983 4660 4449 4558 3635 4167 3821
4188 4022 3907
## [46] 4701 4864 3617 4468 4566
```

```r
mean(captita)

## [1] 4435.8

c = grep(" ", rownames(state.x77))

mean(captita[c])

## [1] 4296.1

state.name[state.x77[,2] == max(state.x77[,2])]

## [1] "Alaska"
```

Exercise 5:

Definitions

1. Define a random variable X.
2. What are the types of random variables?
3. What is a distribution of a random variable?
4. Define the pmf, pdf, and cdf of a random variable.
5. Define the expected value $E[X]$, variance $V[X]$ and mode of a random variable X.
6. Give examples of two discrete and two continuous random variables. Include their distribution functions, expected value and variance.
7. Plot the above random variables in R.

Sampling distributions and estimation

8. What are i.i.d random variables? What is a random sample?
9. What is a statistic? Can you give two examples?
10. What is an estimator θ? Can you give two examples?
11. What is the bias of an estimator $bias(\hat{\theta})$?
12. What is the variance (Var), standard error (se) of an estimator $\hat{\theta}$?
13. What is Mean Square Error of an estimator $\hat{\theta}$?
14. What is the relation between $MSE(\hat{\theta})$ and $bias(\hat{\theta})$?
15. What do the methods of MOM and MLE help you do? You can explain it by an example.
16. Give an application example that would require you to use one of these estimation techniques.

Solution 5:

1. iRandom variable s a variable whose possible values are outcomes of a random phenomenon.
2. There are two types of random variables, discrete and continuous

3. The probability distribution for a random variable describes how the probabilities are distributed over the values of the random variable.

a. Cumulative distribution function (CDF) will give you the probability that a random variable is less than or equal to a certain real number.

b. Probability mass function (PMF) gives you the probability that a discrete random variable is exactly equal to some real value

c. Probability density function (PDF) of a random variable X, when integrated over a set of real numbers A, will give the probability that X lies in A.

a. $E[X]$ indicates its weighted average. Let X be a random variable assuming the values x1, x2, x3, … with corresponding probabilities p(x1), p(x2), p(x3),….. For any function g, the mean or expected value of g(X) is defined by $E(g(X)) = sumg(xk)p(xk)$.

b. $V[X]$ If X is a random variable with mean $E(X)$, then the variance of X, denoted by $Var(X)$, is defined by $Var(X) = E((X - E(X))^2$.

c. mode of a random variable X The mode of a set of data values is the value that appears most often. If X is a discrete random variable, the mode is the value x (i.e, X = x) at which the probability mass function takes its maximum value. In other words, it is the value that is most likely to be sampled

6. The examples for **discrete random variables**:

a. The result of a dice roll The distribution function is $F(x) = 1/6, x = (0,6)$ The mean is 3.5 the variance is 3.5

b. The Bernoulli experiment with k success follows a bionomial distribution The example will be flip a fair coin 10 times, the probility to see k heads(success), The distribution function

$$ p10(k) = \Pr\{N = k\} = \dbinom{10}{k} 0.5^k(1-0.5)^{10-k} \ ; \ 0 \leq k \leq 10 $$

c. The mean is 5 The variance is 2.5

The examples for **continuous random variables** a. A normal distribution(mean = 0, std = 1) $N(0,1)$ The distribution function

$$f_X(x) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-x^2/2\sigma_x^2}$$

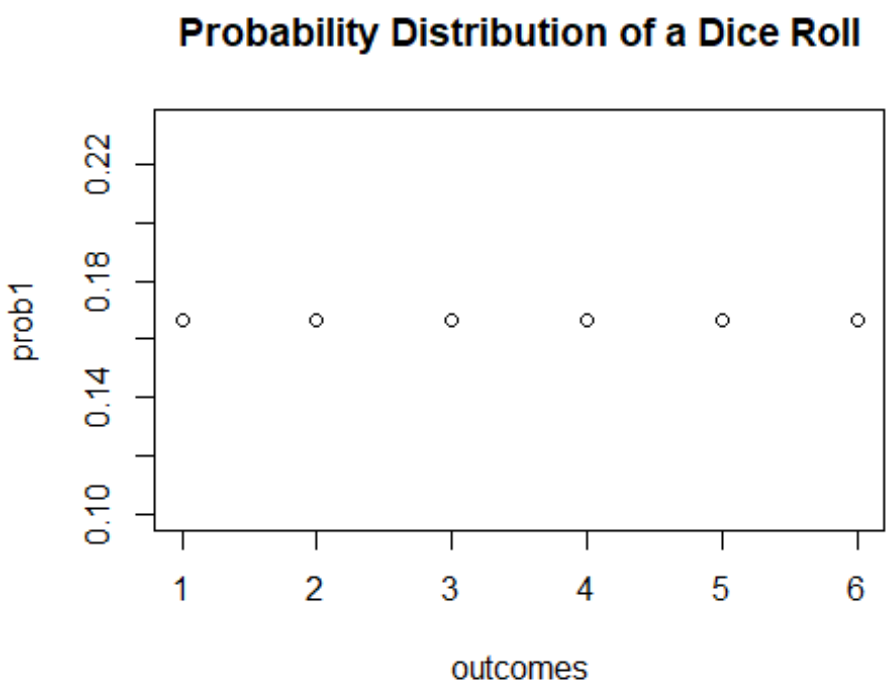
where σ_x^2 is equal to 1. mean is 0, variance is 1

b. Exponential distribution $= 1 $ $f(x) = \lambda\exp(-\lambda x)$ for $x \geq 0$ and $f(x) = 0$ for $x < 0$. mean = 1, variance = 1

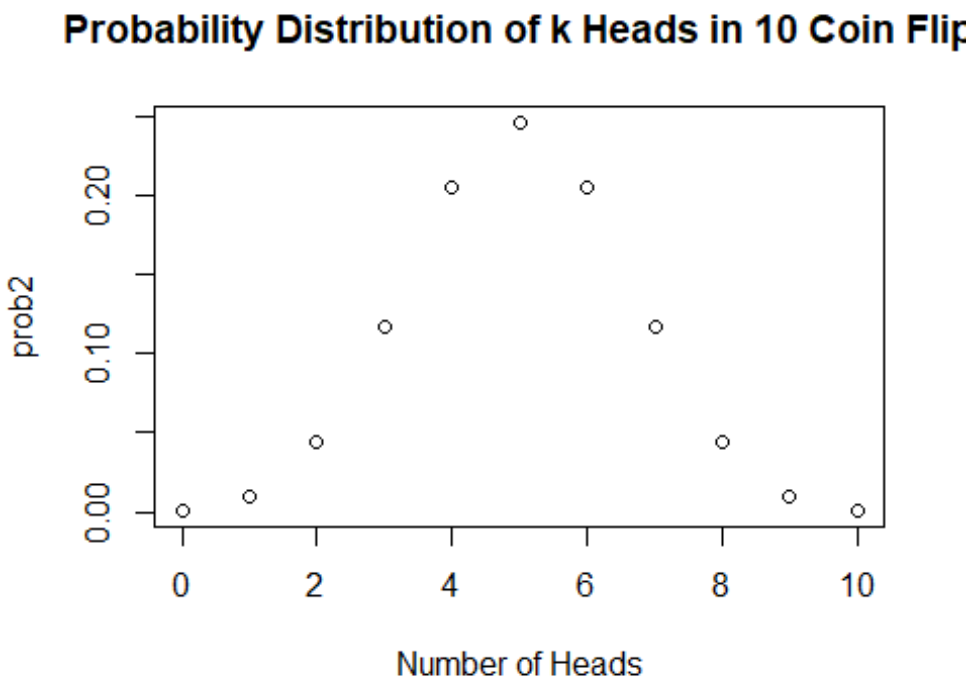
7. **discrete random variable**

```
# the dice roll example
prob1 <- rep(1/6, 6)
```

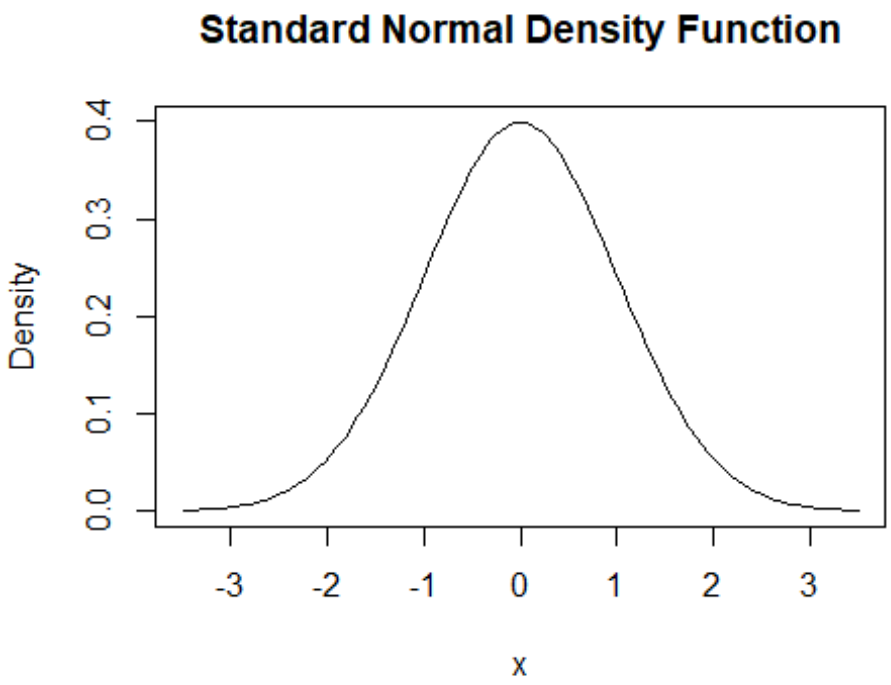
```r
# plot the probabilites
plot(prob1,
     main = "Probability Distribution of a Dice Roll",
     xlab = "outcomes")
```



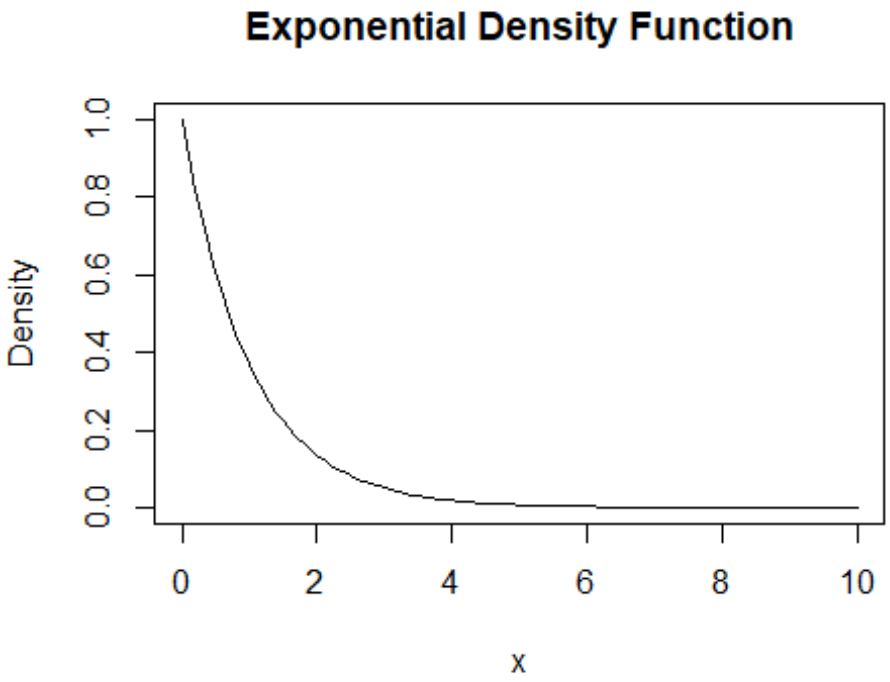

```r
# the coin flip example
k <- 0:10

prob2 <- dbinom(x = k, size = 10, prob = 0.5)

# plot the probabilites
plot(x = k,
     y = prob2,
     main = "Probability Distribution of k Heads in 10 Coin Flip",
     xlab = "Number of Heads")
```




```r
# the normal distribution example
curve(dnorm(x),
      xlim = c(-3.5, 3.5),
      ylab = 'Density',
      main = "Standard Normal Density Function")
```




```r
# the exponential distribution example
curve(dexp(x, rate = 1),
      xlim = c(0,10),
      ylab = 'Density',
      main = "Exponential Density Function")
```



Sampling distributions and estimation

8. A random sample is a sequence of independent, identically distributed (IID) random variables. The term random sample is ubiquitous in mathematical statistics while the abbreviation IID is just as common in basic probability

9. A statistic (singular) or sample statistic is any quantity computed from values in a sample. The two examples are samle mean and sample variance

10. An "estimator" or "point estimate" is a statistic (that is, a function of the data) that is used to infer the value of an unknown parameter in a statistical model. For example, the sample mean($\bar{x}$) is an estimator for the population mean, μ; the sample mean ($\bar{s}$) is an estimator for the population variance, (σ).

11. The bias of an estimator $\hat{\theta}$ of a parameter θ is the difference between the expected value of ˆθ and θ; that is, $Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$.

12. The variance of $\hat{\theta}$ is simply the expected value of the squared sampling deviations, that is $Var(\hat{\theta}) = E[(\hat{\theta} - E(\hat{\theta}))^2]$ t is used to indicate how far, on average, the collection of estimates are from the expected value of the estimates

13. The mean squared error of $\hat{\theta}$ is defined as the expected value (probability-weighted average, over all samples) of the squared errors, that is $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$

14. MSE, mean squared error = variance + square of bias. In particular, for an unbiased estimator, the variance equals the MSE.

15. MOM is short for "method of Moment" and MLE is short for "Maximum Likelihood Estimation". Both methods are used to estimate the population parameters.

a. **MOM** It starts by expressing the population moments (i.e., the expected values of powers of the random variable under consideration) as functions of the parameters of interest. Those expressions are then set equal to the sample moments. The number of such equations is the same as the number of parameters to be estimated. Those equations are then solved for the parameters of interest. The solutions are estimates of those parameters For example, the Bernoulli random variables with parameter p. The first moment about the origin is $E(X_i) = p$. We cam solve this equation to find the parameter p.

b. **MLE** Suppose we have a random sample whose assumed probabilty distribution depends on some unknown parameter θ. Our primary goal here will be to find a point estimator μ which is a good point estimate of θ. For example, the Bernoulli random variables with parameter p. in order to implement the method of maximum likelihood, we need to find the p that maximizes the likelihood $L(p)$. We need to put on our calculus hats now, since in order to maximize the function, we are going to need to differentiate the likelihood function with respect to p. To take the derivative of $ln(L(p))$ (with respect to p) rather than taking the derivative of $L(p)$. Again, doing so often makes the differentiation much easier. Now, all we have to do is solve for p.

16. Suppose that X is a discrete random variable with the following probability mass function: where

$$0 \leq p \leq 1$$

The pmf of X

x	f.x.
0	2p/3
1	p/3
2	2(1-p)/3
3	(1-p)/3

The following 10 independent observations were taken from such distribution :{3,0,2,1,3,2,1,0,2,1}.

The theortical mean value is $7/3 - 2p$ coming from the formula $E(X) = \sum_{i=0}^{3} x\, P(x)$

The sample mean value is 1.5 $\hat{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$

We need to solve the equation $7/3 - 2p = 1.5$ And we finally get the method of moment estimation $\hat{p} = 5/12$