Module 3. Probability distributions of random variables.

Overview: In this module you will be introduced to probability distributions. You will learn how to represent a random variable by its probability density function (pdf) and how to perform calculations using the pdf. This probability theory provides a theoretical basis for statistical inferences which will be addressed in later modules. You need to know the basic language and calculations for later derivations of statistical procedures, which have applications in many fields, including bioinformatics.

This module consists of several lessons. In Lesson 1 you will be introduced to probability density functions (pdf). In Lesson 2, you will be shown various properties of the pdf. Lesson 3 will focus on calculating the mean and variance of a pdf, as well as linear transformations of random variables. In lesson 4 you will be introduced to some specific distributions, namely; normal, binomial, and Poisson distributions. The pdf of these and other common distributions are preprogrammed in R, and can be used directly. Finally, in lesson 5 we introduce the generation of random variables from these distributions. The generated samples are useful for Monte Carlo simulations, which the next module will cover in detail. The Monte Carlo methods are very useful for statistical inferences and will be used extensively in later modules.

As in previous modules, these lessons include examples, script for how to perform the calculations in R, and opportunities for practice.

Learning Objectives

By the end of this module, you should be able to:

- Calculate the probability of a set given a **probability density function (pdf)**
- Distinguish between **discrete** and **continuous** random variables using the sum or integral in the calculation
- Calculate the **mean** and **variance** of a random variable given a pdf
- Calculate mean and variance of the **linear transformation** of a random variable
- Determine the set probability, mean and variance for a **variety of distributions**
- **Generate** random variables for simple Monte Carlo simulations.

Readings

Statistics Using R with Biological Examples: Pages 71-107

Applied Statistics for Bioinformatics Using R: Pages 31-43

Unit 3.1: Defining the probability density function.

Concepts:

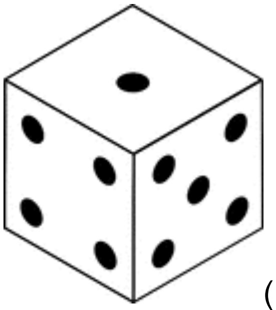
Random variable: the numerical value assigned to the outcome of an experiment.

Random experiment (in probability): a procedure that can be infinitely repeated and has a well-defined set of possible outcomes.

http://en.wikipedia.org/wiki/Experiment_(probability_theory)

Probability density function (pdf) f(x) describes the probability associated with the outcome value x.

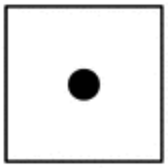
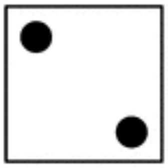
Example (A): roll a die. This is an experiment, whose outcome is the face on top. This is not a random variable yet: the outcome is the “face”, not a number. We define a random variable X as the number of dots in the face on top for a roll of the die. So X takes numerical values: 1,2,3,4,5,6.



(Image from http://etc.usf.edu/clipart/)

The pdf f(x) here describes the chances of each possible outcome to happen in one roll of the die. The following table contains pdf for two different random variable X1 and X2 corresponds to two different dice.
(A1) For a fair die, the random variable X1, all 6 faces have equal probability (chance) of 1/6.
(A2) For random variable X2, we consider an unfair die such that the outcome “1” is five times more likely than each of the other five outcomes.

These pdf’s can be represented using a table in the next page.

Outcome	Value of X1 or X2 (=x)	(A1) For a fair die pdf f(x) =P(X1=x)	(A2) For an unfair die pdf f(x) =P(X2=x)
	1	1/6	1/2
	2	1/6	1/10
	3	1/6	1/10
	4	1/6	1/10
	5	1/6	1/10
	6	1/6	1/10

Example (B): Y = the type of nucleotide residue at a random chosen position on a DNA sequence. Using 1,2,3,4 to represent the four types A,T,C,G respectively. The possible values of Y: 1,2,3,4.

Say, 30% of nucleotides in whole DNA sequence are adenine (A), 30% are thymine (T), 20% each are cytosine (C) and Guanine (G) respectively.
Then the pdf f(y)=P(Y=y) is: f(1)=0.3, f(2)=0.2, f(3)=0.2, and f(4)=0.3.
We can also represent this pdf using a table below.

x	1	2	3	4
f(x)	0.3	0.2	0.2	0.3

Mathematical Abstraction:
Given the pdf f(x), we know everything about the distribution of the random variable X. There is no need to know where the random variable comes from. The only mathematically relevant information in the above three distributions are:
(A1) f(x)=1/6, for x=1,2,3,4,5,6.
(A2) f(x)=1/2 for x=1, and f(x)=1/10 for x=2,3,4,5,6.
(B) f(x)=0.3 for x=1,4 and f(x)=0.2 for x=2,3.

From the pdf, we can calculate the probability that the random variable X falls into a set A, by summing up the probability of the all the outcomes in A.

Example 1: Given the following pdf of random variable X:
f(x)=0.3 for x=1,4 and f(x)=0.2 for x=2,3.
Then P(X<3) = P(X=1) + P(X=2) = f(1) + f(2) =0.3 + 0.2 =0.5.
And P(1.8< X ≤4) = P(X=2)+ P(X=3) + P(X=4) = f(2)+f(3)+f(4)= 0.2+0.2+0.3 =0.7.

Notation: Summation is often denoted with the "Sigma" notation. In the above example, we can write $\sum_{x=2}^{4} f(x) = f(2) + f(3) + f(4)$. Generally, that means, $P(A) = \sum_{x \in A} f(x)$.

Using the "Sigma" notation, we generally calculate $P(A) = \sum_{x \in A} f(x)$.

Example 2: Given the pdf of random variable X: f(x)= x/8 for x=1,3,4.

Then $P(X > 1.5) = \sum_{x>1.5} f(x) = f(3) + f(4) = \frac{3}{8} + \frac{4}{8} = \frac{7}{8}$.

Example 3: Given the pdf of random variable Y: $f(y) = \frac{y^2}{14}, \quad y = -2, 1, 3$

Then $P(Y > 1.5) = \sum_{y>1.5} f(y) = f(3) = \frac{9}{14}$.

Types of random variables: Discrete versus **continuous** random variables.
A **discrete random variable**'s possible values are a list of distinct values. See examples (A1), (A2) and (B). A **continuous random variable** can take any value in an interval, like the one in the following example.

Example (C) X = the weight of a lightweight male boxer (in pounds). According to the boxing rules, a lightweight male boxer must weigh between 130 and 135 pounds. So the possible values of X: any number x such that 130≤ x ≤135. Assume the weight is equally likely to fall on the interval [130,135], then pdf f(x)=1/5, 130≤ x ≤135.

The interpretation of the pdf f(x) for a continuous random variable is different from the pdf for a discrete random variable. For continuous random variable X, the probabilities are not assigned to specific discrete values but are instead assigned to intervals of values only. That is, P(X=x) = 0 for any specific x value. We can only talk about probabilities such as P(a<X<b) = probability for X to fall into interval (a,b). And the pdf decides the probability of set A through an integral instead of a sum: $P(A) = \int_{x\in A} f(x)dx$.

Specifically, for the interval [a,b], $P(a \leqslant X \leqslant b) = \int_a^b f(x)dx$.

Example 4: Given pdf:

$$f(x) = \frac{1}{5}, \quad 130 \leqslant x \leqslant 135.$$

Then

$$P(131 < X \leqslant 133) = \int_{131}^{133} \frac{1}{5} dx \quad = 0.4 .$$

Example 5: Given pdf:

$$f(y) = \frac{y^2}{9}, \quad 0 \leqslant y \leqslant 3$$

Then

$$P(Y > 1.5) = \int_{y>1.5} f(y)dy \quad = \int_{1.5}^{3} \frac{y^2}{9} dy = 0.875 .$$

The integral in the last line is calculated in R by

```
> integrate(function(x) x^2/9, lower = 1.5, upper = 3)
0.875 with absolute error < 9.7e-15
```

There are two types of random variables: discrete and continuous.
Probability density function (pdf) f(x) describes the probability associated with the outcome value x.

For discrete random variables, $P(A) = \sum_{x \in A} f(x)$.

For continuous random variables, $P(A) = \int_{x \in A} f(x)dx$.

<u>Lesson one ends here.</u>

Unit 3.2: Properties of the pdf, and calculations in R.

The pdf f(x) totally describe the behaviors of the random variable X.

For discrete random variables, $P(A) = \sum_{x \in A} f(x)$.

For continuous random variables, $P(A) = \int_{x \in A} f(x) dx$.

Not every function f(x) can be a pdf. A pdf has to satisfy the following two properties.

(1) f(x) ≥0.

(2) P(S)=1, where S is the set of all possible values of X. That is, $P(S) = \sum_{x} f(x) = 1$ for discrete random variable X; $P(S) = \int_{x} f(x) dx$ for continuous random variable X.

Example 1. Is the following function a pdf? $f(x) = 0.5(x^2 - 1), \quad x = 0,1,2,4$.
This does not satisfies either property (1) nor property (2).
Particularly, f(0) is negative, so it cannot be a pdf.

Example 2. Is the following function a pdf? $f(x) = 0.5(x^2 - 1), \quad x = 2,4$.
Both f(2) and f(4) are positive, so the property (1) is satisfied.
However, for S={2,4}, P(S) =f(2)+f(4) = 1.5+7.5 =9 ≠1, violates the property (2).

This cannot be a pdf.

Example 3. Is the following function a pdf? $f(x) = \frac{1}{18}(x^2 - 1), \quad x = 2,4$.
Both f(2) and f(4) are positive, so the property (1) is satisfied.
For S={2,4}, P(S) =f(2)+f(4) = 3/18 +15/18 =1, so the property (2) is satisfied.
This is pdf for a discrete random variable.
Notice that the formula are for the values in S only. Outside the range, f(x)=0.
In this example, $f(0) = 0 \neq \frac{1}{18}(0^2 - 1)$ since x=0 is not a possible value.

A pdf satisfies the following two properties.

(1) f(x) ≥0.

(2) P(S)=1, where S is the set of all possible values of X. That is, $P(S) = \sum_x f(x) = 1$ for discrete random variable X; $P(S) = \int_x f(x)dx$ for continuous random variable X.

Example 4. Is the following function a pdf? $f(x) = \frac{1}{18}(x^2 - 1), \quad 2 \le x \le 4$.

The set of possible values S=[2,4] is an interval. So IF it is a pdf, it is the pdf for a continuous variable.

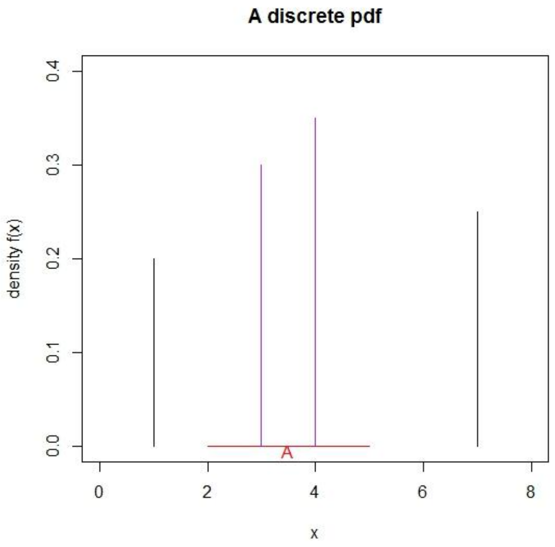
The property (1) holds since f(x)>0 for any x between 2 and 4.

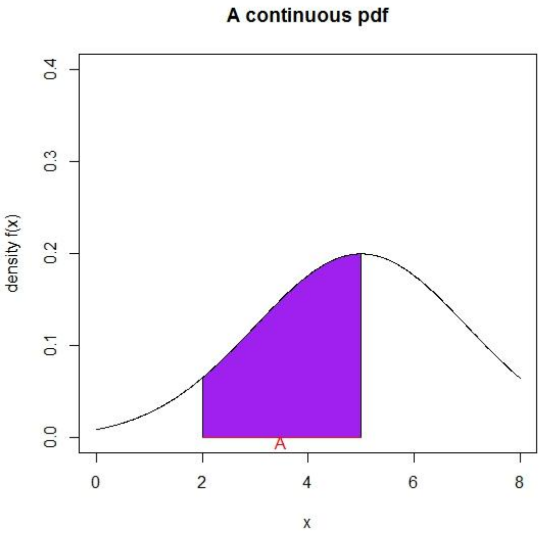
The property (2) is violated, however, since $P(S) = \int_{x=2}^{4} \frac{1}{18}(x^2 - 1)dx = \frac{50}{54} \neq 1$.

Hence, this is NOT a pdf.

IMPORTANT NOTES:

The pdf has different interpretations for discrete versus continuous random variables: $P(A)=\sum_{x\in A} f(x)$ **(discrete) and** $P(A)=\int_{x\in A} f(x)dx$ **(continuous)**. Therefore it is critically important that you distinguish the type of random variables in the calculation of probability. (Some books including Seefeld & Linder call pdf “probability distribution function” instead for discrete random variables. We do not make the distinction of names here, the same as in Krijnen’s textbook. The important issue is that usage of pdf is different in discrete versus continuous random variables.)





The pdf f(x) for discrete X only has nonzero values at a list of x values, while the pdf f(x) for continuous X have nonzero values over intervals.

Graphically, the pdf for a discrete random variable (the left picture above) contains just some “sticks” at the list of possible x values. The probability $P(A)=\sum_{x\in A} f(x)$ sums over the lengths of those sticks falling in set A. In contrast, the pdf for a continuous random variable (the right picture above) is a curve. The probability $P(A)=\int_{x\in A} f(x)dx$ is the area under the curve for set A.

To distinguish the type of any given pdf, we pay attention to its range of possible values. The range of a discrete random variable is a list, while the range of a continuous random variable is the union of intervals.

Example 5:

X has pdf $f(x)=\frac{x}{8}, \quad x=1,3,4$.

Y has pdf: $f(y)=\frac{y}{8}, \quad 0\leqslant y\leqslant 4$.

Since the range of X = {1,3,4} is a list, X is discrete. The set {X>2.5} includes two values X=3 and X=4.

$$P(X>2.5)=\sum_{x>2.5} f(x)=f(3)+f(4)=\frac{3}{8}+\frac{4}{8}=0.875.$$

Since the range of Y = [0, 4] is an interval, Y is continuous. The set {Y>2.5} includes the whole interval (2.5, 4].

$$P(Y>2.5)=\int_{y>2.5} f(y)dy \quad =\int_{2.5}^{4}\frac{y}{8}dy=0.609375 \text{ using R.}$$

```
> integrate(function(x) x/8, lower=2.5,upper=4)
0.609375 with absolute error < 6.8e-15
```

Notice: A continuous variable have zero probability for any specific values. Therefore, a set {a<X<b} has the same probability as the set {a≤X≤b}. The “<” and “>” are the same as “≤” and “≥” when calculating probability of sets. For discrete random variable, attention is needed to distinguish between “<” and “>” from “≤” and “≥”.

For discrete random variable, we needed to distinguish between “<” and “>” from “≤” and “≥”when calculating probability of sets.
For continuous random variables, the “<” and “>” lead to the same as “≤” and “≥”

Example 5 (continued): X has pdf $f(x)=\frac{x}{8}, \quad x=1,3,4$. Y has pdf:

$f(y)=\frac{y}{8}, \quad 0\leqslant y\leqslant 4$.

Discrete X: $P(2.5 < X\leqslant 4)=f(3)+f(4)=0.875$ but $P(2.5 < X < 4)=f(3)=0.375$.

Continuous Y: Both $P(2.5 < Y\leqslant 4)$ and $P(2.5 < Y < 4)$ equal $\int_{2.5}^{4}\frac{y}{8}dy=0.609375$.

Coding the pdf in R.

Notice that for the pdf given above, there is always a range of possible values S associated with it. We need to code that into R with the formula for pdf.
For discrete random values, S is a list and we need to code it as such in R.

Example 6:

X has pdf $f(x)=\frac{x}{8}, \quad x=1,3,4$.

How do we code this pdf in R?

Answer:
X is discrete. Its range S={1,3,4}. In R we define this using a list.

```
> X_range <- c(1,3,4)
```

Then the pdf f(x) only takes value x/8 if x falls in this range S.

```
> f.x <- function (x) x/8
> f_x <- function (x) f.x(x)*(x %in% X_range)
```

NOTICE that we coded the pdf into f_x which contains the range info. The function f.x is NOT the pdf here.

We can check that this is indeed a pdf by checking the properties (1) and (2).
Property (1): is f(x)≥0 for all x values in the range S?

```
> all(f_x(X_range)>=0)
[1] TRUE
```

Property (2): is P(S)=1?

```
> sum(f_x(X_range))==1
[1] TRUE
```

The probability of any other set can be calculated directly by sum over probabilities of all cases in the set.
For example to calculate P(2<X<5), we sum the pdf over those values of 2<X<5

```
> sum(f_x(X_range)*(2<X_range & X_range<5))
[1] 0.875
```

Coding the pdf in R.

For continuous random values, S is an interval. We code it in R with logical operator to check if the value falls in the interval S.

Example 7:

Y has pdf: $f(y)=\frac{y}{8}, \quad 0\leqslant y\leqslant 4$.

How do we code this pdf in R?

Answer:

Y is continuous.

Then the pdf f(y) only takes value y/8 if y falls in the range S=[0,4].

```
> f.x <- function (x) x/8
> f_y <- function (x) f.x(x)*(0<=x & x<=4)
```

We can check that this is indeed a pdf by checking the properties (1) and (2).

Property (1): is f(x)≥0 for all x values in the range S? Since S is an interval [0,4], we cannot really check this exhaustively for all x values as in the discrete case. We will only do an approximate check for x=0, 0.01, 0.02, …, 4.

```
> all(f_y(seq(0,4,by=0.01))>=0)
[1] TRUE
```

Property (2): is P(S)=1? P(S) is checked by the integral.

```
> integrate(f_y, lower=0, upper=4)$value==1
[1] TRUE
```

The probability of any other set can be calculated directly by integrating the pdf over the set.

For example to calculate P(2<X<5), we do

```
> integrate(f_y, lower=2, upper=5)$value
[1] 0.75
```

Hint: Although we just integrated over the whole interval (2,5) here, usually we may want to take the range S into consideration. Since the range S=[0,4], in fact we only need to integrate over (2,4] instead. The later integral may be more accurate if the interval being asked has a big proportion outside S.

<u>Lesson two ends here.</u>

Unit 3.3: Mean and Variance. Linear Transformation.

The *mean* or *expected value* of a random variable is the long term average value that one “expect” to find if the random variable process can be repeated infinite times. It is a measure for the central location of the probability distribution. We use the notation E(X) to denote the mean of a random variable X. The “E” comes from the “expected value”.

Example 1: Assume that the types of nucleotides in human DNA sequences are: 30% A, 30% T, 20% C and 20% G. An algorithm assigns score 0, 1, 2, 3 to the four types A, T, C, G respectively. What is the mean score for a randomly selected nucleotide base on the DNA sequence?

Denote X as the score for one randomly selected nucleotide base. Then the pdf of X is given by: f(x)=0.3 for x=0,3 and f(x)=0.2 for x=1,2.
On average, we would expect to get the score:
$E(X)=0(0.3)+1(0.2)+2(0.2)+3(0.3)=1.5$.
In fact this is $E(X)=0f(0)+1f(1)+2f(2)+3f(3)$.

In general, for a discrete random variable, $E(X)=\sum_x xf(x)$ **;**

For a continuous random variable, $E(X)=\int_{-\infty}^{\infty} xf(x)dx$ **.**

Example 2:

X has pdf $f(x)=\frac{1}{x!}e^{-1}, \quad x=0,1,2,....$

Y has pdf: $f(y)=e^{-y}, \quad y>0$.

Since range of X is a list {0,1,2,…}, X is discrete and

$$E(X)=\sum_{x=0}^{\infty} xf(x)=\sum_{x=0}^{\infty}\frac{x}{x!}e^{-1}=(0+1+\frac{2}{2}+\frac{3}{6}+\frac{4}{24}+...)e^{-1}=(1+1+\frac{1}{2}+\frac{1}{6}+...)e^{-1}=1.$$

Since range of Y is an interval (0,∞), Y is continuous and

$$E(Y)=\int_{-\infty}^{\infty} yf(y)dy=\int_0^{\infty} ye^{-y}dy=1.$$

The value is calculated by R command

```
integrate(function(y) y*exp(-y), lower=0, upper=Inf)
```

Calculating the mean from pdf in R.

The mean is calculated either by sum or by integral. So knowing the type of the random variables, we can carry out the calculation in R based on given pdf.

Example 3: X has pdf $f(x)=\frac{x}{8}, \quad x=1,3,4$.

How to calculate E(X)?

Notice that, first, we need to express the pdf and the range of the random variables in R correctly. X is a discrete random variable, its range is a list {1,3,4}. Hence

```
> X_range<-c(1,3,4)
> f.x <- function (x) x/8
> f_x <- function (x) f.x(x)*(x %in% X_range)
```

To calculate the E(X) we need to sum over all x*f(x) values, that is:

```
> sum(X_range*f_X(X_range))
[1] 3.25
```

Hence E(X)=3.25.

The above R command calculates $E(X)=1f(1)+3f(3)+4f(4)$.

Be careful do not confuse E(X) with $\sum_{x} f(x)=f(1)+f(3)+f(4)$ which equals P(S)=1 and calculated in R as

```
> sum(f_X(X_range))
[1] 1
```

Example 4: Y has pdf: $f(y)=\frac{y}{8}, \quad 0\leqslant y\leqslant 4$. How to calculate E(Y)?

Again we first find the type of Y from its range.
Y is a continuous random variable, its range is an interval [0,4].
So we express its pdf in R as

```
> f_Y<- function(y) y/8*(0<=y & y<=4)
```

The expected value is calculated as

```
> integrate(function(y) y*f_Y(y), lower=0, upper=4)
2.666667 with absolute error < 3e-14
```

So that E(Y)=2.666667.

Notice that this calculation used our knowledge of the range to simplify the expression as $E(Y)=\int_{-\infty}^{\infty} yf(y)dy=\int_{0}^{4} yf(y)dy$. We can also directly calculate $\int_{-\infty}^{\infty} yf(y)dy$ in R:

```
> integrate(function(y) y*f_Y(y), lower=-Inf, upper=Inf)
2.666667 with absolute error < 4.4e-05
```

In this case, we get the exactly the same correct answer for $\int_{-\infty}^{\infty} yf(y)dy$. However, generally we want to restrict the range of the integral since, for some f(x), there may be numerical errors if integrated over (-∞, ∞).

Be careful do not confuse E(Y) with $\int_{-\infty}^{\infty} f(y)dy=\int_{0}^{4} \frac{y}{8}dy$ which equals P(S)=1 and calculated in R as

```
> integrate(function(y) f_Y(y), lower=0, upper=4)
[1] 1
```

Variance:
The mean EX is a measure of the central location of the distribution of X. To measure the "spread" of the distribution, a common measure is the ***variance***, which is the expected value of the square distance from the mean:

$Var(X)=E[(X-EX)^2]$. Using algebra, we can get a commonly used equivalent expression $Var(X)=E(X^2)-(EX)^2$. It also customary to denote the mean and variance of X as μ_X and σ_X^2 respectively.

Hence, **for a discrete random variable,** $\mu_X=E(X)=\sum_x xf(x)$ **and** $\sigma_X^2=Var(X)=\sum_x (x-\mu_X)^2 f(x)$ **or** $[\sum_x x^2 f(x)]-\mu_X^2$ **; for a continuous random variable,**

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

and

$$\sigma_X^2=Var(X)=\int_{-\infty}^{\infty}(x-\mu_X)^2 f(x)dx \text{ **or** } \int_{-\infty}^{\infty} x^2 f(x)dx-\mu_X^2 \text{ .}$$

Also, the square root of the variance is called the standard deviation:

$\sigma_X=std(X)=\sqrt{Var(X)}$ **.** Note that standard deviation has the same unit as that of X, variance does not have the same unit.

We can calculate the mean and variances in R similar to above.

Example 5:

X has pdf $f(x)=\frac{x}{8}, \quad x=1,3,4$.

Y has pdf: $f(y)=\frac{y}{8}, \quad 0\leqslant y\leqslant 4$.

How to calculate Var(X) and Var(Y)?

X is discrete.

```
> X_range<-c(1,3,4)                                    #Range of X
> f.x <- function (x) x/8
> f_x <- function (x) f.x(x)*(x %in% X_range) #pdf of X
> EX<-sum(X_range*f_X(X_range))                        #Calculate E(X)
> VarX<- sum((X_range-EX)^2*f_X(X_range))   #Calculate Var(X)
> VarX     #Print value of Var(X)
[1] 0.9375
```

Hence Var(X)=0.9375.

Y is continuous.

```
> f_Y<- function(y) y/8*(0<=y & y<=4)       #pdf of Y
> EY<-integrate(function(y) y*f_Y(y), lower=0, upper=4)$value #Calculate E(Y)
> VarY<-integrate(function(y) (y-EY)^2*f_Y(y), lower=0, upper=4)$value #Calculate
Var(Y)
> VarY     #Print value of Var(Y)
[1] 0.8888889
```

Hence Var(Y)=0.888889.

Linear transformation.

For a random variable X, a linear transformation Y=aX+b is also a random variable for constants a and b: when X takes value x, Y takes value ax+b. For example: X has pdf $f(x)=\frac{x}{8}, \quad x=1,3,4$. That is, X has three possible values {1,3,4} with probabilities:

x	1	3	4
f(x)=P(X=x)	0.125	0.375	0.5

Then Y=2X+1 also has three possible values {3,7,9} with probabilities

y	3	7	9
f(y)=P(Y=2X+1=y)	0.125	0.375	0.5

Using some algebra, we can get the following useful properties on mean and variances of linear transformation of random variables.
If X is a random variable, a and b are constants, then
$E(aX+b)=aE(X)+b$ and $Var(aX+b)=a^2Var(X)$.

Example 6:

X has pdf $f(x)=\frac{x}{8}, \quad x=1,3,4$.

Y has pdf: $f(y)=\frac{y}{8}, \quad 0\leqslant y\leqslant 4$.

Before, we have calculated E(X)=3.25, Var(X)=0.9325, E(Y)=2.666667 and Var(Y)=0.8888889.
What are the means and variances of 3X+1 and 10Y-6?

Answer: we do not need to find pdf of 3X+1 and 10Y-6, and can directly find the means and variances using the properties above.
$E(3X+1)=3E(X)+1=3(3.25)+1=10.75$,
$Var(3X+1)=9Var(X)=9(0.9325)=8.3925$,
$E(10Y-6)=10E(Y)-6=10(2.666667)-6=20.66667$
and $Var(10Y-6)=100Var(Y)=100(0.8888889)=88.88889$.

Lesson Three ends here.

Unit 3.4: Several common distributions. Calculating their probability from pdf in R.

In this unit, we will learn several common probability distributions whose pdf's are already coded in R.

Normal distribution:

A standard normal distribution has a bell shaped pdf

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad -\infty < x < \infty$$

with mean zero and variance one. We denote a random variable with the standard normal distribution as Z~ $N(0,1)$.

One important feature of normal random variable is that its linear combination is also normal (bell-shaped). Particularly, let X=μ+σZ, then we can calculate that X has a mean μ and standard deviation σ as in the last unit. And X is normal distributed, we denote this as X~ $N(mean = \mu, sd = \sigma)$. The pdf for $N(mean = \mu, sd = \sigma)$ also has an explicit formula

$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$. Hence we can calculate probabilities based on this formula.

The normal distribution is also called Gaussian distribution, named after the mathematician Karl Gauss.

Example 1: X follows the normal distribution $N(\mu=3,\sigma=4)$. Find $P(2\leqslant X\leqslant 5.5)$.

Solution: $P(2\leqslant X\leqslant 5.5)=\int_{2}^{5.5}\frac{1}{\sqrt{2\pi}4}e^{-\frac{(x-3)^2}{2(16)}}dx=0.333$ in the last step we used R to calculate the integral.

```
> integrate(function(x) exp(-(x-3)^2/32)/4/sqrt(2*pi), lower=2, upper=5.5)
0.3327208 with absolute error < 3.7e-15
```

Remark 1: Before we have the computing power nowadays, it is not easy to do the numerical integration. Generally a statistics textbook would contain a table of pre-calculated probability values for standard normal distribution. It is standard to first transform

$Z=\frac{X-\mu}{\sigma}$, then use the table to find the probability. For this example, we would do

$$P(2\leqslant X\leqslant 5.5)=P(\frac{2-3}{4}\leqslant Z=\frac{X-3}{4}\leqslant\frac{5.5-3}{4})=P(-0.25\leqslant Z\leqslant 0.625)=\int_{-0.25}^{0.625}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=0.333$$

With R doing the numerical integration, *we **no longer need** to transform X into Z*, the standard normal distributed random variable.

Remark 2: R has programed the pdf formula for normal distribution so that you do not have to memorize it. In R, $dnorm(x, mean=\mu, sd=\sigma)$ gives the density function for normal distribution

$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. The density functions in R always start with the letter “d” (for density).

Hence, to calculate the probability in this example, we can simply compute $P(2\leqslant X\leqslant 5.5)$ as

```
> integrate(function(x) dnorm(x,mean=3,sd=4), lower=2, upper=5.5)
0.3327208 with absolute error < 3.7e-15
```

Cumulative distribution function (CDF):
The function F(x)=P(X≤x) is called the cumulative distribution function.

The CDF is also pre-programmed into R for normal distribution, and starts with an initial letter “p” (for probability): $pnorm(x, mean = \mu, sd = \sigma)$.
Notice that the probability of an interval (a,b] can be easily calculated using CDF as $P(a < X \leqslant b) = P(X \leqslant b) - P(X \leqslant a) = F(b) - F(a)$. Hence we can also calculate the probability in the example above from $P(2 \leqslant X \leqslant 5.5) = P(X \leqslant 5.5) - P(X < 2)$

```
> pnorm(5.5,mean=3,sd=4) - pnorm(2,mean=3,sd=4)
[1] 0.3327208
```

Notice here, since X is a continuous random variable, $P(X < 2) = P(X \leqslant 2)$ can be calculated directly from CDF because $P(X = 2) = 0$. For continuous random variables, we do not need to distinguish “<” and “≤”.

Binomial distribution:

One frequently used probability model in biological data analysis is the Binomial distribution. A binomial random variable X is a discrete random variable, counting the number of “successes” in n *independent* Bernoulli trials. A Bernoulli trial comes from experiments with binary outcomes, labeled as “success” and “failure”, with “p” probability of getting success and “1-p” probability of getting failure. The “*independent”* is a statistical concept; here it means that one Bernoulli trial’s outcome is not affected by the outcomes of other Bernoulli trials.

The binomial random variable has two parameters: n (number of total trials) and p (success probability of each trial). The pdf for the X~$binomial(n, p)$ is given by formula $P(X = x) = f(x) = \binom{n}{x} p^x (1-p)^{n-x}$, where $\binom{n}{x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$, called “n choose x”, is the number of combination of x out of n distinct objects.

For more details on the derivation of this formula, read the section on binomial distribution on pages 80-82 of Seefeld & Linder’s book.

To recognize a binomial random variable, notice when you have repeated binary trials each with the same probability of success.

Example 2: If 30% of nucleotides in human DNA are adenine (A), let X be the number of adenine bases in three randomly selected nucleotide bases on the DNA.
Then X follows a binomial distribution: three repeated trials (success means that the base is adenine), each with 30% chance.
That is, X follows $binomial(n=3, p=0.3)$.

We can calculate the probability that there are "at least two adenine bases" as

$$P(X \geqslant 2) = f(2) + f(3) = \binom{3}{2} 0.3^2 (1-0.3)^1 + \binom{3}{3} 0.3^3 (1-0.3)^0 = 3(0.3)^2(0.7) + (1)0.3^3 = 0.216.$$

We can also calculate this in R as

```
> choose(3,2)*0.3^2*(1-0.3)^(3-2)+choose(3,3)*0.3^3*(1-0.3)^(3-3)
[1] 0.216
```

Calculating binomial probabilities in R.

Example 2 (continued): X follows $binomial(n=3, p=0.3)$. We wish to calculate $P(X \geqslant 2)$

In R, the pdf (density) of binomial distribution is given by function $dbinom(x, size=n, p=p)$. Hence we can calculate $P(X \geqslant 2)$ by

```
> set.X<-c(2,3)
> sum(dbinom(set.X,size=3,p=0.3))
[1] 0.216
```

Hence $P(X \geqslant 2) = 0.216$.

Also, the CDF for binomial is given by $pbinom(x, size=n, p=p)$ in R. So we can also use $P(X \geqslant 2) = 1 - P(X < 2) = 1 - P(X \leqslant 1)$ and calculate by

```
> 1-pbinom(1,3,0.3)
[1] 0.216
```

Notice that, for discrete random variables, P(X<x) and P(X≤x) may not be the same since P(X=x) can have positive value. This is different from the case of continuous random variables, where we do not care to distinguish "<" from "≤". So in above calculation we used $P(X \geqslant 2) = 1 - P(X \leqslant 1) = 1 - pbinom(1,3,0.3)$. We can NOT use

$$1 - pbinom(2,3,0.3) = 1 - P(X \leqslant 2)$$

because it would give us instead

$$1 - P(X \leqslant 2) = P(X > 2) = P(X \geqslant 3) \quad \neq P(X \geqslant 2).$$

Example 3: Suppose that the expression values of gene CCND3 (Cyclin D3) can be represented by X which is distributed as $N(\mu = 1.9, \sigma = 0.5)$.

(1) What is the probability that one measurement of this gene's expression will exceed 2?

(2) If we take ten independent measurements of this gene's expression, what is the probability that at least seven out of the ten measurements will have expression values exceeding 2?

Answer:

(1) P(X>2)=0.4207403 can be calculated in R

```
> 1-pnorm(2,mean=1.9,sd=0.5)
[1] 0.4207403
```

(2) Let Y be the number out of ten measurements with expression values exceeding 2. Then Y follows a binomial distribution: ten repeated trials (success means that the expression exceeds 2), each with a chance=0.4207403.
That is, Y follows $binomial(n = 10, p = 0.4207403)$.
So we calculate P(Y≥7)= 1-P(Y≤6) = 0.072 in R

```
> 1-pbinom(6,size=10,prob=0.4207403)
[1] 0.07183252
```

Property of binomial random variable. If X follows the $binomial(n, p)$ distribution, then the mean and variance of X are given by

$$E(X) = np, \quad Var(X) = np(1-p).$$

Example 4: X follows a $binomial(n = 3, p = 0.3)$ distribution.

$E(X) = 3(0.3) = 0.9, \quad Var(X) = 3(0.3)(1-0.3) = 0.63$. We can check this with R:

```
> X.range<-(0:3)
> EX<-sum(X.range*dbinom(X.range,size=3,p=0.3))
> VarX<-sum((X.range-EX)^2*dbinom(X.range,size=3,p=0.3))
> c(EX,VarX)
[1] 0.90 0.63
```

Furthermore, if we define a new random variable Y=2X+1, then

$E(Y) = 2(0.9) + 1 = 2.8, \quad Var(Y) = 2^2(0.63) = 2.52$. Notice that Y is NOT a binomial distribution, but we can still calculate its mean and variance using the formula on linear transformations in unit 3.2. To see that Y is not binomial: the range of Y is {1,3,5,7} (Note that we CANNOT write is as {0,1,…,n} for any n).

Generally, a linear transformation of a random variable no longer follows the same type of distribution. The normal distribution above is a special exception: linear transformations of normal random variables are still normal random variables.

Poisson distribution:
Poison distribution is a discrete distribution, with range of possible values as {0,1,2,…}. The Poisson distribution has one parameter λ which is its mean and also is its variance. The Poisson pdf is given by

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0,1,...$$

More detailed descriptions can be found in pages 86-90 of Seefeld & Linder's book.

In R, the pdf and CDF of Poisson(λ) distribution are given by:
$dpois(x,\lambda)$ and $ppois(x,\lambda)$.

Other distributions:
There are several other distributions, all continuous distributions, mentioned in the textbooks. Their information is summarized in the table below.

Table 3.4.1. Several commonly used probability distributions: name (name of density function in R), parameters, range of possible values, pdf (density), mean and variance.

Distributions	parameters	Range	pdf	mean	Variance
Normal $dnorm(x,\mu,\sigma)$	μ,σ	$(-\infty,\infty)$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
Binomial $dbinom(x,size=n,p)$	n,p	0,1,…,n	$\binom{n}{x}p^x(1-p)^{n-x}$	np	$np(1-p)$
Poisson $dpois(x,\lambda)$	λ	0,1,2,…	$\frac{\lambda^x}{x!}e^{-\lambda}$	λ	λ
Gamma $dgamma(x,\alpha,\beta)$	α,β	$(0,\infty)$	$\frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)}$	$\alpha\beta$	$\alpha\beta^2$
Exponential dexp (x,λ)	λ	$(0,\infty)$	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Chi-square $dchisq(x,m)$	m	$(0,\infty)$	$\frac{x^{(m/2)-1}e^{-x/2}}{2^{m/2}\Gamma(m/2)}$	m	2m
Beta $dbeta(x,\alpha,\beta)$	α,β	$(0,1)$	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\beta(\alpha,\beta)}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
T $dt(x,m)$	m	$(-\infty,\infty)$	$\frac{(1+\frac{1}{m}x^2)^{-\frac{m+1}{2}}}{m^{\frac{1}{2}}\beta(\frac{1}{2},\frac{m}{2})}$	0	$\frac{m}{m-2},m>2$
F $df(x,m,n)$	m,n	$(0,\infty)$	$\frac{x^{\frac{m}{2}-1}(1+\frac{m}{n}x)^{-\frac{m+n}{2}}}{(\frac{n}{m})^{\frac{m}{2}}\beta(\frac{m}{2},\frac{n}{2})}$	$\frac{n}{n-2}$	$\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)},n>4$

Lesson Summary

This lesson covers some common probability distributions which are pre-programmed in R. You should know how to call those pdf, and use them to find the set probability, the mean and the variance. Particularly, the Binomial distribution will be used often in the course, and you should be able to recognize a binomial random variable in practice (counts of one outcome in repeated trials each with binary outcomes). Next lesson will teach how to generate random samples from these distributions.

Lesson 5: Generating Random Variables and Monte Carlo Simulation.

Introduction

Objectives

By the end of this lesson you will have had the opportunity to:

- Generate random variables according to the specified probability distributions
- Estimate a set probability by Monte Carlo simulation

Overview

This lesson introduce the definition of the Monte Carlo Method, which will be used extensively throughout the course. We start with generating random samples from probability distributions. We then show a simple Monte Carlo simulation in R. We will consider Monte Carlo simulations in more detail in the next module.

The pdf describes what a repeated sample from the distribution look like. We can also generate samples from the distribution, and using them to check numerically the properties of the distribution.

Using repeated random samples for numerical results is called the ***Monte Carlo Method***, which is widely used complicated computational biology and applied statistics. We will cover the Monte Carlo methods in the next module. Here we first introduce how to generate random samples from probability distributions, and illustrate the concept of Monte Carlo simulations.

R has programmed many common distributions. For these distributions, the random sample can be generated from the routine started with letter “r”. For example, to generate 10 values from the binomial(3,0.5) distribution, we can use the R command 10 values from rbinom(n=10, size=3, p=0.5). To generate 6 values from the Poisson(10) distribution, we use the R command 10 values from rpois(n=6, lambda=10).

<u>Example 1</u> For Binomial(2,0.4) distribution, the random variable has three possible values: {0,1,2} and the pdf is listed in the table:

x	0	1	2
f(x)= P(X=x)	0.36	0.48	0.16

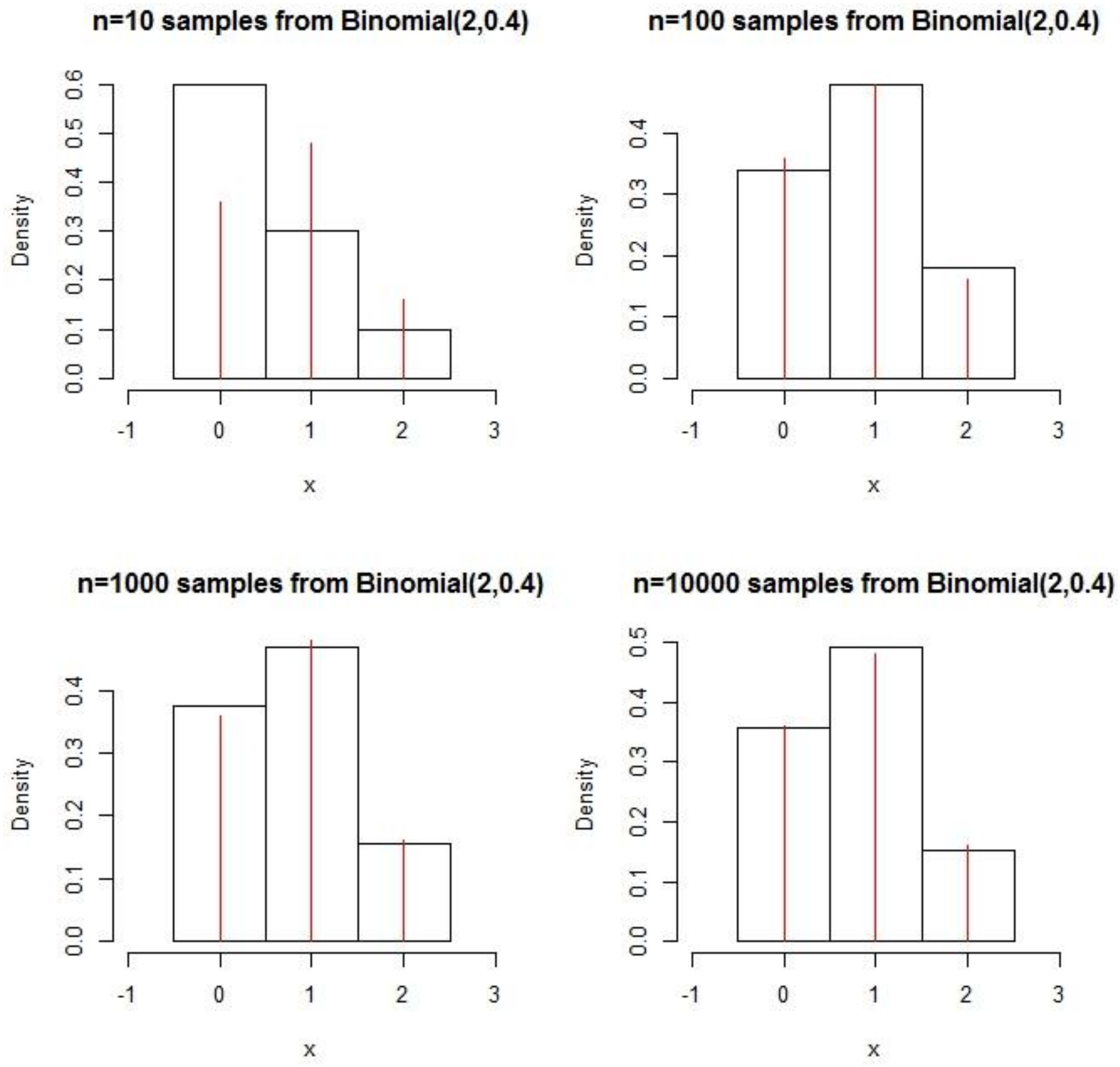
To sample 10 values from binomial(2,0.4) distribution, we use the R command `rbinom(n=10,size=2,p=0.4)`.
Notice that this gives a *<u>random</u>* sample of size 10. Therefore, each time running the command gives different numbers. Following are values for my two runs of the command. You will get different numbers when running it.

```
> rbinom(n=10, size=2, p=0.4)
 [1] 1 0 1 0 0 2 0 0 0 2
> rbinom(n=10, size=2, p=0.4)
 [1] 1 1 0 0 2 1 0 1 0 1
```

When we generate a random sample X_1, X_2, $\cdots$, X_n from this distribution, we can count the proportions of values 0, 1 and 2 in the sample, and draw histogram of the data from those proportions.

By definition, the pdf f(x)=P(X=x) is the long term proportion of the value x among the sample when sample size n increases to infinity. Therefore, observed proportions in the data would approach the theoretical values in the table, when n gets bigger.

We plot in R the histogram of data sets of size n=10, n=100, n=1000, n=10000 from Binomial(2,0.4) distribution, superimposed with pdf. We can see that, as n increases, the proportions in the histogram (the height of the boxes) are getting closer the pdf values (height of the sticks) in the plots.



The proportions in the histogram (the height of the boxes) are getting closer the pdf values (height of the sticks) in the plots, as n increases.

R commands to produce the figure:

```
par(mfrow=c(2,2))
x<-rbinom(n=10, size=2, p=0.4)
hist(x,breaks=(0:3)-0.5,xlim=c(-1,3),freq=F,main="n=10 samples from Binomial(2,0.4)")
points((0:2),dbinom((0:2),2,0.4),col=2,type="h")
x<-rbinom(n=100, size=2, p=0.4)
hist(x,breaks=(0:3)-0.5,xlim=c(-1,3),freq=F,main="n=100 samples from Binomial(2,0.4)")
points((0:2),dbinom((0:2),2,0.4),col=2,type="h")
x<-rbinom(n=1000, size=2, p=0.4)
hist(x,breaks=(0:3)-0.5,xlim=c(-1,3),freq=F,main="n=1000 samples from Binomial(2,0.4)")
points((0:2),dbinom((0:2),2,0.4),col=2,type="h")
x<-rbinom(n=10000,size=2,p=0.4)
hist(x,breaks=(0:3)-0.5,xlim=c(-1,3),freq=F,main="n=10000 samples from Binomial(2,0.4)")
points((0:2),dbinom((0:2),2,0.4),col=2,type="h")
```

Note: for random generated data, each run of the same R command should produce a different set of data. Therefore, if the above R script is run again, the figure produced will be different from the one shown above. However, if you run it several times, you can always see the same general pattern of histogram converges to the pdf for bigger sample size n.

Explanation (video?):

```
par(mfrow=c(2,2))
```

This tells R to put four plots together in one figure in 2 rows and 2 columns.

```
x<-rbinom(n=10,size=2,p=0.4)
```

Generates 10 values from the binomial distribution and store in x.

```
hist(x,breaks=(0:3)-0.5,xlim=c(-1,3),freq=F, main="n=10 samples from Binomial(2,0.4)")
```

Produce the histograms: specifying the breaks at points x= -0.5, 0.5, 1.5, 2.5 (breaks=(0:3)-0.5). This means to do the histogram in three intervals with middle points at x=0, 1, 2. The histogram plots the proportions instead of absolute frequencies (freq=F).

```
points((0:2), dbinom((0:2), 2, 0.4), col=2, type="h")
```

Add the pdf as sticks on the plots (type="h").

Example 2

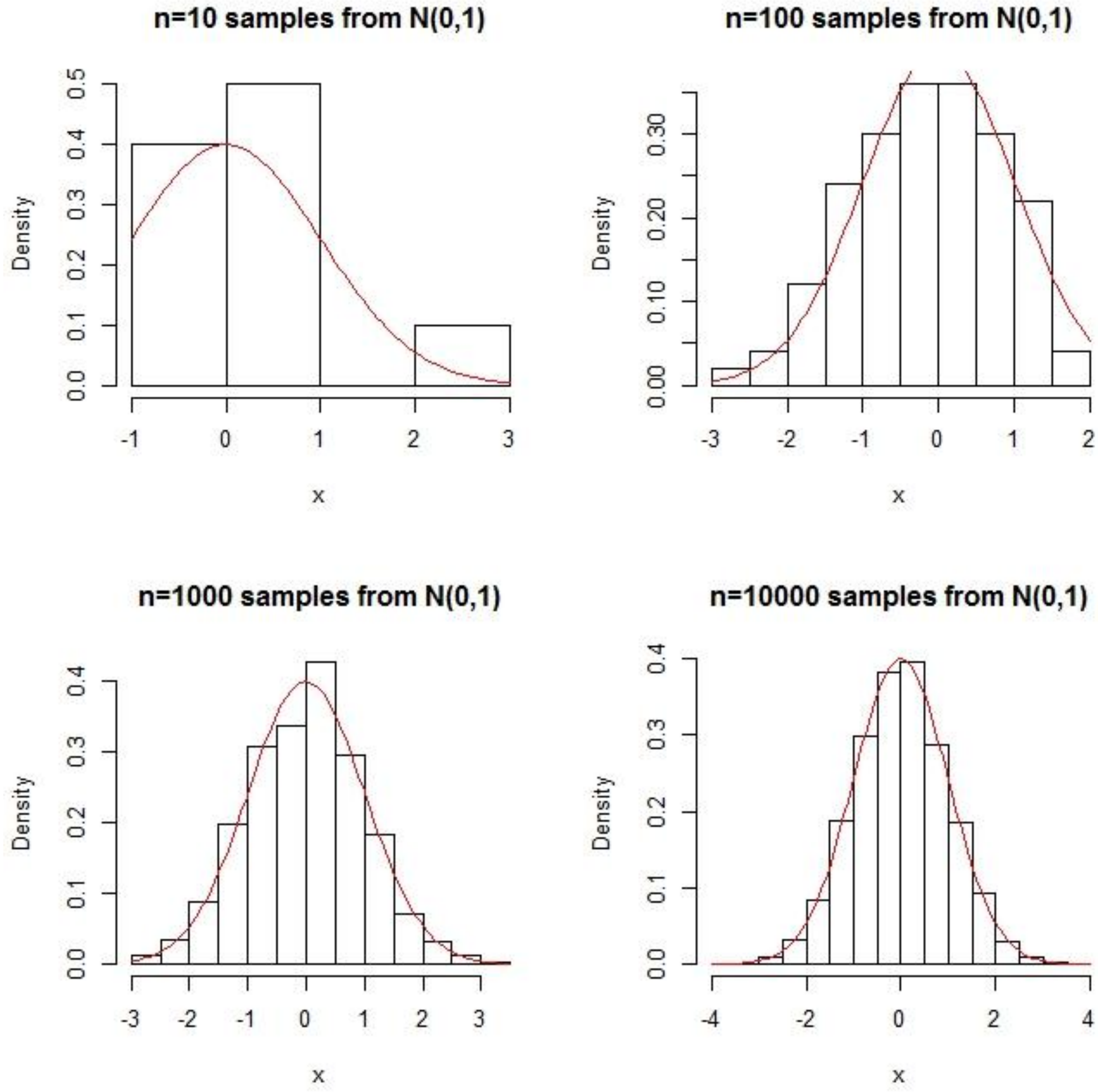
To sample 10 values from the normal distribution $N(\mu = 0, \ \sigma = 1)$ with mean 0 and standard deviation 1, we use the R command

rnorm(n=10, mean=0, sd=1)

This is a continuous distribution, with the area under pdf

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

corresponds the long term probability. Following we produce the histograms of data sets of size n=10, n=100, n=1000, n=10000 from Binomial(2,0.4) distribution, superimposed with pdf. We can see that, as n increases, the histograms are getting closer the pdf curve for continuous distribution.



R script for producing the above figure:

```
par(mfrow=c(2,2)) #put 4 graphs in 2 rows by 2 columns
x<-rnorm(n=10, mean=0, sd=1) #samples 10 number from N(0,1)
hist(x, freq = FALSE, main="n=10 samples from N(0,1)") #histogram, freq=F to get draw in proportions (between 0 and 1) instead of the counts.
curve(dnorm(x,mean=0,sd=1),col=2,add=T) #draw the pdf (dnorm) curve. Overlay to previous graph by “add=True”
x<-rnorm(n=100, mean=0, sd=1)
hist(x, freq = FALSE, main="n=100 samples from N(0,1)")
curve(dnorm(x,mean=0,sd=1),col=2,add=T)
x<-rnorm(n=1000, mean=0, sd=1)
hist(x, freq = FALSE, main="n=1000 samples from N(0,1)")
curve(dnorm(x,mean=0,sd=1),col=2,add=T)
x<-rnorm(n=10000, mean=0, sd=1)
hist(x, freq = FALSE, main="n=10000 samples from N(0,1)")
curve(dnorm(x,mean=0,sd=1),col=2,add=T)
```

You can see this is similar to the previous script.

To generate data from other known distributions, use the following commands:

Poisson $rpois(n,\text{lambda}=\lambda)$; Gamma $rgamma(n,\text{shape}=\alpha,\text{scale}=\beta)$; Exponential $rexp(x,rate=\lambda)$; Chi-Square $rchisq(n,df=m)$, etc.

Monte Carlo Simulations to Check Probabilities

In the earlier lessons, we taught how to calculate probability of an event P(A). The probability P(A) is the long-term frequency in repeated sampling from the given distribution. As we have seen in examples above, for big sample size n, the proportion belonging to event A will get close to the theoretical value P(A). Therefore we can use the proportion from randomly generated samples for a numerical approximation of P(A) if we do not know the theoretical value. Such numerical estimations using randomly generated samples are called as ***Monte Carlo*** methods. We illustrate the idea on a random variable from the F distribution.

<u>Example 3</u>

For X from an F-distribution with degrees of freedoms 4 and 3, we wish to find the $P(0.4 < X < 1.5)$. Using Monte Carlo simulation with sample size of n=10,000, we can approximate the value using R commands below:

```
x<-rf(n=10000, df1=4, df2=3)
mean((0.4<x) & (x<1.5))
```

Run these commands and we get

```
[1] 0.4149
```

Therefore $P(0.4 < X < 1.5) \approx 0.415$ by the Monte Carlo simulation. Notice that this is an approximation, and the values will change if you ran the simulation again. However, for big n, the change will be very small.

The true probability can be calculated from the pdf using R as

```
> integrate(function(x) df(x,df1=4, df2=3), lower=0.4, upper=1.5)$value
[1] 0.4165663
```

So we see that the approximation is accurate to the second decimal space in this case.

In this example, we can calculate the true value exactly. So we can observe how close the Monte Carlo estimate to the true value is. For real applications, generally the true value is from a very complicated model and hard to calculate exactly. The Monte Carlo simulations often is straightforward and can give good estimate of the true value. You will see more Monte Carlo methods in later modules.

Lesson Summary

In this lesson, we taught how to generate data for several common probability distributions. We also introduced the concept of Monte Carlo method. You should be able to find simple Monte Carlo estimations for event probabilities.

We also taught how to plot the data histogram and the density curves in R, and how to overlay the plots together.

Module 3. Summary.

In this module, we focused on how to describe a random variable mathematically, through its pdf. We looked at various types of distributions and their associated expected values and variances. We also taught how to generate random variables, and use them for Monte Carlo simulations. Next week we will look at the Monte Carlo simulations in more detail. Then you will learn more on generating data from multivariate distributions, and learn more probability calculations and the Central Limit Theorem.