LANGEBIO - BIOSTATISTICS

OCTAVIO MARTÍNEZ DE LA VEGA

6. Random Variables

In our daily lab life we do experiments of very different kind. However we almost never stop to think about their random nature; we do not think much about the "noise" that is always present and, like a fog, in many cases difficult the interpretation of our results. Other characteristic that is almost always present is that we summarize our results as numbers. We count, or measure, quantities of interest. Here we will see a more formal concept that helps with the analysis of our results: The concept of random variable.

Before that, it is important to remember what is the mathematical definition of a *function*.

6.1. **Function.** In mathematics, a function is a relation between a set of inputs and a set of potential outputs with the property that each input is related to exactly one output.

An example of such a relation is defined by the rule $f(x) = x^2$, which relates an input x to its square, which are both real numbers. The output of the function f corresponding to an input x is denoted by $f(x)$ (read "f of x"). If the input is 3, then the output is 9, and we may write $f(3) = 9$.

I took it verbatim from: *http:en.wikipedia.orgwikiFunction_%28mathematics%29*

As we have seen, in R we can define **functions** which have an input (the argument) and an output (the result). Here we are interested in a particular kind of functions: The random variables.

6.2. **Definition. Random Variable (RV).** A *Random Variable* is a **function** that assign to each one of the elements of Ω a real number.

Very simple, to each possible result of a random experiment, say $\omega \in \Omega$, the RV assign a *single* real number. Here we will denote RB with capital letters, for example X, Y, etc.

The name "Random Variable" is a bit misleading, because this relation is NOT random, neither it is a VARIABLE, but anyway the use is to call it like that.

Let see some examples.

6.2.1. *Examples of Random Variables.*

- 1. Random Experiment: "Select an individual from the cross $AaxAa$", $\Omega = \{AA, Aa, aa\}$. Some random variables defined in Ω:
  X : Number of alleles A which the individual has.
  Y : $Y(AA) = Y(Aa) = 0$, $Y(aa) = 1$.

Date: April 2012.

Z : $Z(AA) = 0$, $Z(Aa) = 1$, $Z(aa) = 0$.

- 2. Random Experiment: "Throw two regular dices and look at the numbers that are above." (we think that we can distinguish the two dices; for example one is white and the other is red),
  $\Omega = \{(1,1), (1,2), \ldots, (6,6)\}$
  or
  $\Omega = \{(x,y)|x \in \{1,2,\ldots,6\}, y \in \{1,2,\ldots,6\}\}$
  Some RV defined in Ω:
  W : Sum of the results of the first and the second dices ($W = x + y$)
  M : The maximum of the results of the two dices.
  Z : The product of the two results: $Z = xy$.

7. Statistics

At last! We are going to be gin to see some Statistics after our prelude with set and probability theory.

We are going to give a loose definition of Statistics, in particular of *Inferencial* Statistics. We have defined our universe of the Discourse, Ω, the elements that live there or *sampling points*, which form *events* and finally have also defined *Probability functions* and *Random variables*. In a very general way Statistic is useful to:

- Model real experiments (know what to expect under some assumptions)
- **Estimate** the parameters of interest of these experiments.
- *Describe* the relation between (random) variables.
- Take decisions about interesting hypotheses.
- ...

7.1. **Discrete Random Variable (DRV).** Let X be a RV which can take values $x_1, x_2, \ldots,$. Then we define

$$p(x_i) = P[X = x_i]$$

as the probability function of X.

7.1.1. *Examples of DRVs.* For the previous example(s), let write the probability functions:

- 1. Random Experiment: "Select an individual from the cross $AaxAa$", $\Omega = \{AA, Aa, aa\}$. Some random variables defined in Ω:
  X : Number of alleles A which the individual has.
  $P[X = 0] = 1/4$; $P[X = 1] = 1/2$; $P[X = 2] = 1/4$.
  Y : $Y(AA) = Y(Aa) = 0$, $Y(aa) = 1$.
  $P[Y = 0] = 3/4$; $P[Y = 1] = 1/4$.
  Z : $Z(AA) = 0$, $Z(Aa) = 1$, $Z(aa) = 0$.
  $P[Z = 0] = 1/2$; $P[Z = 1] = 1/2$.
- 2. Random Experiment: "Throw two regular dices and look at the numbers that are above." (we think that we can distinguish the two dices; for example one is white and the other is red),
  $\Omega = \{(x,y)|x \in \{1,2,\ldots,6\}, y \in \{1,2,\ldots,6\}\}$

To practice, let obtain a representation of Ω in R:

```
> Om <- matrix(NA, nrow=6, ncol=6, dimnames=list(c(1:6), c(1:6))) # A Matrix with "NA's"
> Om # Note that we obtain 6*6 = 36 combinations.
   1  2  3  4  5  6
1 NA NA NA NA NA NA
2 NA NA NA NA NA NA
3 NA NA NA NA NA NA
4 NA NA NA NA NA NA
5 NA NA NA NA NA NA
6 NA NA NA NA NA NA
```

Some RV defined in Ω:

W : Sum of the results of the first and the second dices ($W = x + y$)

Let's obtain all the sampling points of Ω:

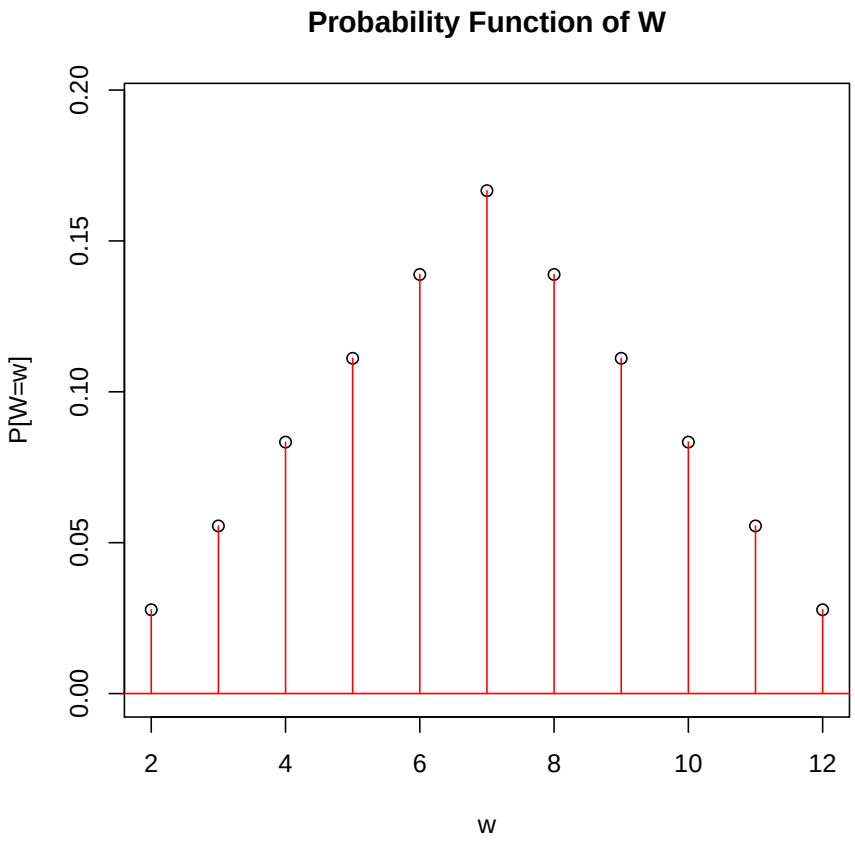
```
> OmW <- matrix(rep(c(1:6),6)+rep(c(1:6), each=6),
+ nrow=6, ncol=6, dimnames=list(c(1:6), c(1:6)))
> OmW
  1 2 3  4  5  6
1 2 3 4  5  6  7
2 3 4 5  6  7  8
3 4 5 6  7  8  9
4 5 6 7  8  9 10
5 6 7 8  9 10 11
6 7 8 9 10 11 12
```

We see that W is a DRV which can take all the values from 2 to 12. Now, lets obtain the probability function, assuming that we are dealing with a *honest* pair of dice; that is assuming that all 36 sampling points have the same probability: 1/36.

```
> pW <- data.frame(c(2:12), rep(NA, 11))
> names(pW) <- c("w", "P[W=w]")
> for(i in 2:12){
+ pW[i-1, 2] <- sum(1*(OmW == i)) / 36
+ }
> pW
    w     P[W=w]
1   2 0.02777778
2   3 0.05555556
3   4 0.08333333
4   5 0.11111111
5   6 0.13888889
6   7 0.16666667
7   8 0.13888889
8   9 0.11111111
9  10 0.08333333
```

```
10 11 0.05555556
11 12 0.02777778
> # There is an easy way to do the same
> table(OmW) # Produce a "table" with the frequencies
OmW
 2  3  4  5  6  7  8  9 10 11 12
 1  2  3  4  5  6  5  4  3  2  1
> table(OmW) / 36 # Gives the probabilities.
OmW
         2          3          4          5          6
0.02777778 0.05555556 0.08333333 0.11111111 0.13888889
         7          8          9         10         11
0.16666667 0.13888889 0.11111111 0.08333333 0.05555556
        12
0.02777778
> # Let's see a graph of the probability function
> plot(pW[,1], pW[,2], xlab="w", ylab="P[W=w]",
+ ylim = c(0, 7/36), main="Probability Function of W")
> abline(h=0, col="red")
> for(i in 2:12){
+         segments(i, 0, i, pW[i-1,2], col="red")
+ }
```

Note that I was very careful to trace the red line, indicating that the probability takes a value of zero when w takes a value intermediate between the whole numbers...



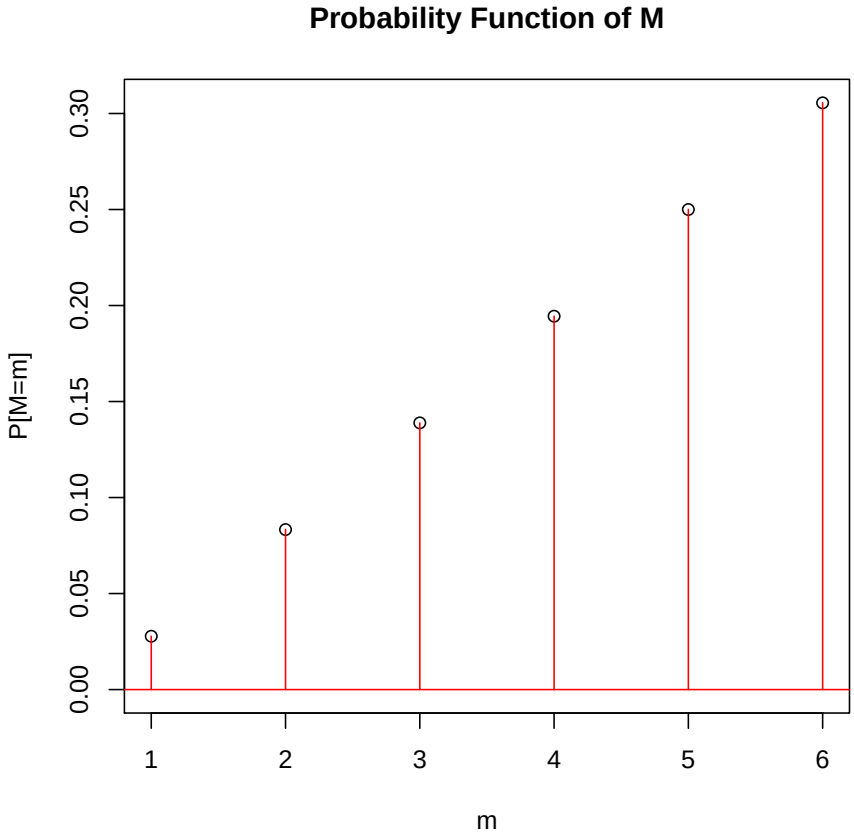
M : The maximum of the results of the two dices.

```
> # This is a bit tricky
> d1 <- matrix(rep(c(1:6), 6), nrow=6, ncol=6)
> d2 <- matrix(rep(c(1:6), each=6), nrow=6, ncol=6)
> OmM <- d1*(d1>=d2) + d2*(d1<d2)
> attributes(OmM)$dimnames <- list(c(1:6), c(1:6))
> OmM # The Omega points for M
  1 2 3 4 5 6
1 1 2 3 4 5 6
2 2 2 3 4 5 6
3 3 3 3 4 5 6
4 4 4 4 4 5 6
5 5 5 5 5 5 6
6 6 6 6 6 6 6
> # Now let's calculate the probability
> pM <- data.frame(c(1:6), rep(NA, 6))
```

```
> names(pM) <- c("m", "P[M=m]")
> for(i in 1:6){
+ pM[i, 2] <- sum(1*(OmM == i)) / 36
+ }
> pM
  m     P[M=m]
1 1 0.02777778
2 2 0.08333333
3 3 0.13888889
4 4 0.19444444
5 5 0.25000000
6 6 0.30555556
```

And let's see the graph of the probability function for M

```
> plot(pM[,1], pM[,2], xlab="m", ylab="P[M=m]",
+ ylim= c(0, 11/36), main="Probability Function of M")
> abline(h=0, col="red")
> for(i in 1:6){
+         segments(i, 0, i, pM[i,2], col="red")
+ }
```



Z : The product of the two results: $Z = xy$.

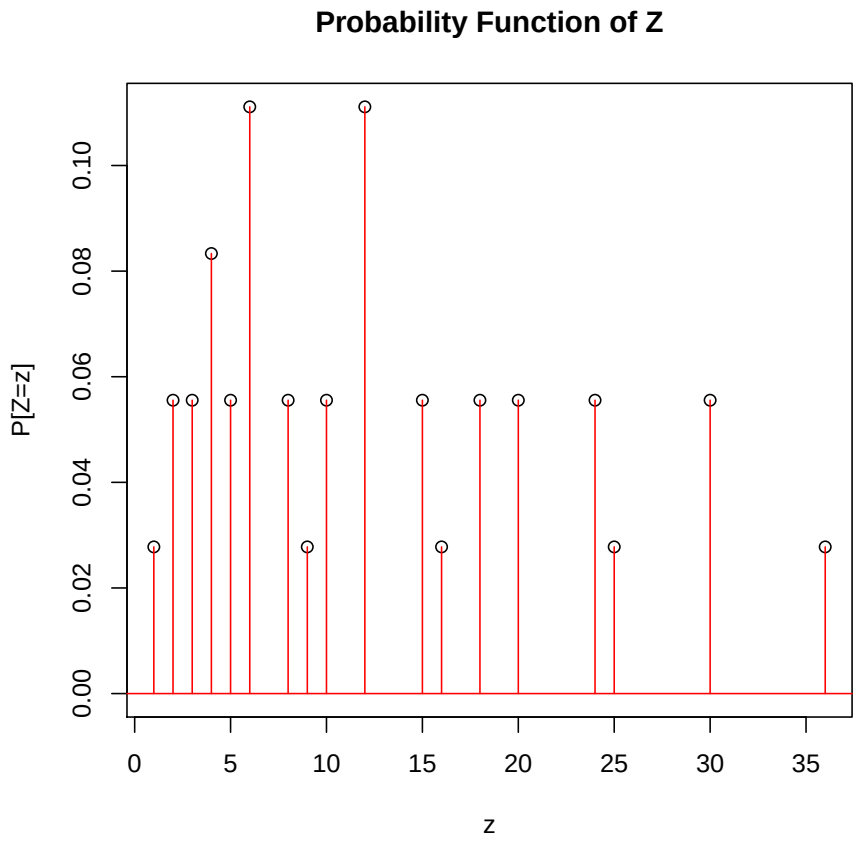
```
> OmZ <- matrix(c(1:6)%*%t(c(1:6)), nrow=6, ncol=6, dimnames=list(c(1:6), c(1:6)))
> OmZ
  1  2  3  4  5  6
1 1  2  3  4  5  6
2 2  4  6  8 10 12
3 3  6  9 12 15 18
4 4  8 12 16 20 24
5 5 10 15 20 25 30
6 6 12 18 24 30 36
> Rz <- sort(unique(as.vector(OmZ))) # Values of z
> # With probability larger than zero (not all 1, 2, ... 36)
> Rz
 [1]  1  2  3  4  5  6  8  9 10 12 15 16 18 20 24 25 30
[18] 36
```

```
> pZ <- data.frame(Rz, rep(NA, 18))
> names(pZ) <- c("z", "P[Z=z]")
> for(i in 1:18){
+ pZ[i, 2] <- sum(1*(OmZ == Rz[i])) / 36
+ }
> pZ # The probability
```

```
    z     P[Z=z]
1   1 0.02777778
2   2 0.05555556
3   3 0.05555556
4   4 0.08333333
5   5 0.05555556
6   6 0.11111111
7   8 0.05555556
8   9 0.02777778
9  10 0.05555556
10 12 0.11111111
11 15 0.05555556
12 16 0.02777778
13 18 0.05555556
14 20 0.05555556
15 24 0.05555556
16 25 0.02777778
17 30 0.05555556
18 36 0.02777778
```

And let's see the graph of the probability function for Z

```
> plot(pZ[,1], pZ[,2], xlab="z", ylab="P[Z=z]",
+ ylim= c(0, max(pZ[,2])), main="Probability Function of Z")
> abline(h=0, col="red")
> for(i in c(1:18)){
+         segments(Rz[i], 0, Rz[i], pZ[i,2], col="red")
+ }
```



7.2. **The Binomial Distribution.** Assume that a random experiment can result in two distinct outcomes, and let name them "Success" (S) or "Failure" (F). We have a sampling space

$$\Omega = \{S, F\}$$

Imagine that this experiment can be repeated many times and that we have a constant probability $P[S] = p$ and $P[F] = 1 - p$ for some value $0 < p < 1$. This is called a *Bernoulli* trial or experiment.

Now, assume that you repeat the experiment k times and define the random variable

$$X : \textit{Number of successes in } k \textit{ trials}$$

Find the probability function for X.

Let's begin with a simple case where $k = 3$ and then try to generalize to any k.

First, note that X can take values in the set $\{0, 1, 2, 3\}$. How can we write the sample space of this experiment?, well lets write it as

$$\Omega = \{(SSS), (SSF), (SFS), \ldots, (FFF)\}$$

(Ω has 8 sample points, $2 * 2 * 2 = 2^3$). We can also write it as

$$\Omega = \{(r_1, r_2, r_3) | r_i \in \{S, F\}\}$$

where r_i is the $i - th$ try. Now, note that

$P[X = 0] = P[(FFF)]$,

$P[X = 1] = P[(FFS) \cup (FSF) \cup (SFS)] = P[(FFS)] + P[(FSF)] + P[(SFS)]$, because these three are distinct sampling point;

$P[X = 2] = P[(SSF) \cup (SFS) \cup (FSS)] = P[(SSF)] + P[(SFS)] + P[(FSS)]$ and finally

$P[X = 3] = P[(SSS)]$.

Now, to obtain $P[X = x]$ we can just work out the probabilities of each one of the sampling points. But note that each one of the individual results (the first, the second and the third) are **independent** and thus, for example:

$P[(FFF)] = (1 - p)(1 - p)(1 - p) = (1 - p)^3$,

$P[(FFS)] = (1 - p)(1 - p)p = (1 - p)^2 p$,

$P[(FSF)] = (1 - p)p(1 - p) = (1 - p)^2 p = P[(FFS)] = P[(SFF)]$.

Thus we can conclude that:

$P[X = 0] = P[(FFF)] = (1 - p)^3$,

$P[X = 1] = P[(FFS)] + P[(FSF)] + P[(SFS)] = 3(1 - p)^2 p$

$P[X = 2] = P[(SSF)] + P[(SFS)] + P[(FSS)] = 3(1 - p)p^2$ and finally

$P[X = 3] = P[(SSS)] = p^3$.

Also, if you remember your algebra from secondary school, you will agree that

$$p^3 + 3p^2(1 - p) + 3p(1 - p)^2 + (1 - p)^3 = (1 + (1 - p))^3 = (1)^3 = 1$$

.

Now, lets generalize for k trials.

First, lets find $P[X = 0]$, that is we want to find

$P[$ Obtaining zero successes in k trials $] = P[(FFF \ldots F)] = (1 - p)(1 - p) \ldots (1 - p)$ (k-times), that is $P[X = 0] = (1 - p)^k$.

Now to obtain $P[X = 1]$ we must consider all the cases where we can obtain 1 success in k trials. How many different ways are for that to happen?. Well, the success could be in the first trial, or in the second or in the $\ldots$ the $k - th$ trial, thus there are k different sample points which give us "one success and $k - 1$ failures". Thus:

$$P[X = 1] = kp(1 - p)^{k-1}$$

Now, to obtain $P[X = 2]$, we need to know how many sample points include exactly 2 successes in k tries. There is a handy formula for that which answer the question "in how many ways can I choose r objects from a total of k", that is call the *combinations of k objects taking r at the time* and is given by:

$$C_r^k = \frac{k(k - 1) \ldots (k - r + 1)}{r(r - 1) \ldots 1} = \frac{k!}{r!(k - r)!}$$

Of course this functions are programed into R; let's see

```
> factorial(c(0:5)) # that is 0!, 1!, ... 5!
```

```
[1]   1   1   2   6  24 120
> choose(3, 1) # How many ways are to choose 1 object from a set of 3?
[1] 3
> choose(3, 2) # How many ways are to choose 2 objects from a set of 3?
[1] 3
> choose(3, c(0:3)) # All the coefficients for k=3
[1] 1 3 3 1
> choose(4, c(0:4)) # All the coefficients for k=4
[1] 1 4 6 4 1
> # Using our formula for combinations:
> factorial(4)/(factorial(2)*factorial(2))
[1] 6
> # Or the R function
> choose(4, 2)
[1] 6
```

Coming back to our problem we find that

$$P[X = x] = C_x^k p^x (1-p)^{k-x} = \frac{k!}{x!(k-x)!} p^x (1-p)^{k-x}$$

That is the formula for the Binomial probability, where k is the number of assays (tries, intents), x is the number of successes and p is the probability of success in a single assay.

Note that this distribution depends on two parameters k (the number of tries) and p (the probability of success in a single try). This probability arises very frequently in Biology (and in general), and we have found its general probability. If X has the Binomial distribution with parameters k and p we will write

$$X \sim \mathcal{B}(k, p)$$

Let's see some examples of the Binomial distribution in R, for this we will work out the following

7.3. **Example. Binomial Distribution in Transcriptomics.** Assume that we are doing the sequencing of a cDNA library of sunflower leaves. The RNA was obtained from leaves of a single plant. Let k be the total number of sequences obtained and X the number of sequences that belong to the RuBisCO (Ribulose-1,5-bisphosphate carboxylase oxygenase) gene.

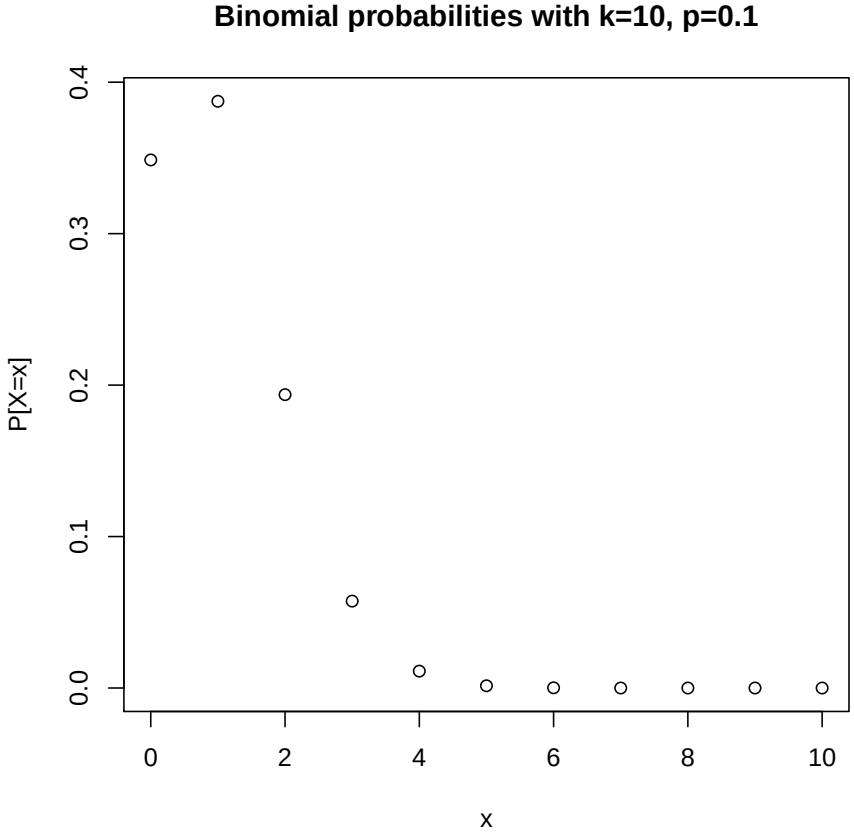
Is it sensible to consider that $X \sim \mathcal{B}(k, p)$? (discuss).

NOTE: To begin, we will assume some very large (un-realistic) probabilities of transcription of a gene, just to exemplify.

Let's work in R for a while. There is a set of functions related with the binomial distribution; to get help about them type "? dbinom".

Assume that the probability of finding a sequence of mRNA from the RuBisCo gene is very large, say 10% or $p = 0.1$ and assume that we sample just 10 sequences. What are the probabilities of finding exactly $X = 0, 1, 2, \ldots, 10$ sequences (reads) of this gene?

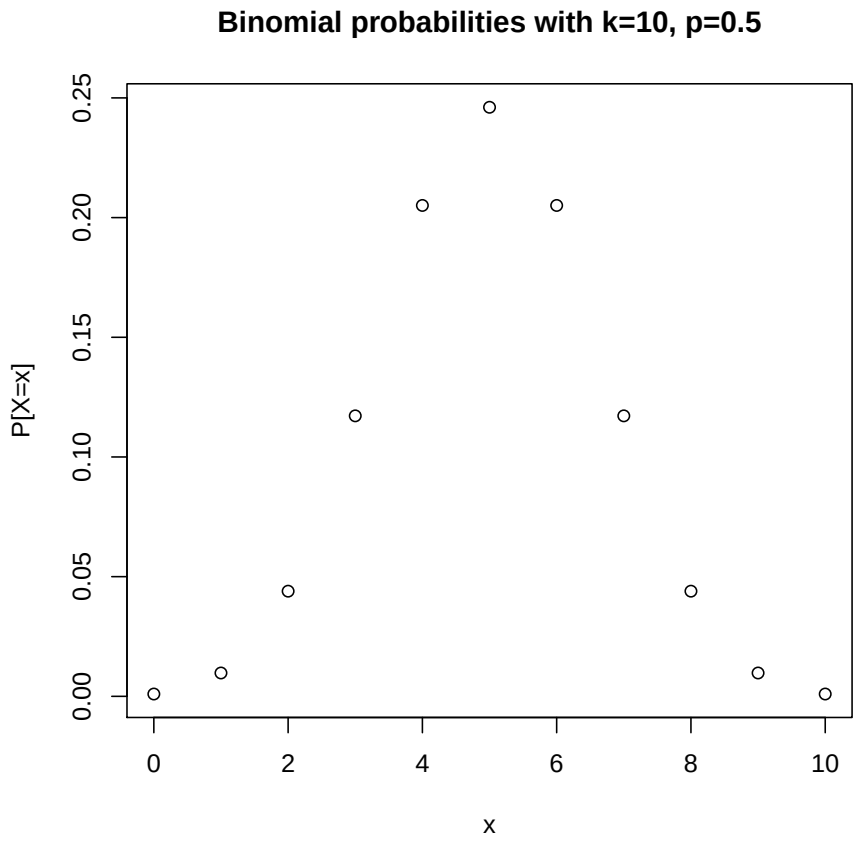
```
> dbinom(c(0:10), size=10, prob=0.1)
 [1] 0.3486784401 0.3874204890 0.1937102445
 [4] 0.0573956280 0.0111602610 0.0014880348
 [7] 0.0001377810 0.0000087480 0.0000003645
[10] 0.0000000090 0.0000000001
> # Let's plot them
> plot(c(0:10), dbinom(c(0:10), 10, 0.1),
+ xlab="x", ylab="P[X=x]",
+ main = "Binomial probabilities with k=10, p=0.1")
```



In the case of throwing a coin 10 times, and defining X as the number of heads we have $p = 1/2$, let see the probability of X in that case

```
> dbinom(c(0:10), size=10, prob=1/2) # Note the simetry
```

```
 [1] 0.0009765625 0.0097656250 0.0439453125
 [4] 0.1171875000 0.2050781250 0.2460937500
 [7] 0.2050781250 0.1171875000 0.0439453125
[10] 0.0097656250 0.0009765625
> # Let's plot them
> plot(c(0:10), dbinom(c(0:10), 10, 1/2),
+ xlab="x", ylab="P[X=x]",
+ main = "Binomial probabilities with k=10, p=0.5")
```



7.4. **Definition. Distribution Function.** Let X be a random variable. Then we define the *Distribution Function* of X as

$$F_X(x) = P[X \leq x]$$

Note that for a discrete random variable we have that

$$F_X(x) = P[X \leq x] = \sum_{X \leq x} P[X = x]$$

that is why we can call the Distribution Function also as the *Accumulative Distribution Function*. In general we have that if $a < b$

$$F_X(a) = P[X \leq a] \leq P[X \leq b] = F_X(b)$$

that is $F_X(\bullet)$ is an increasing function, and also

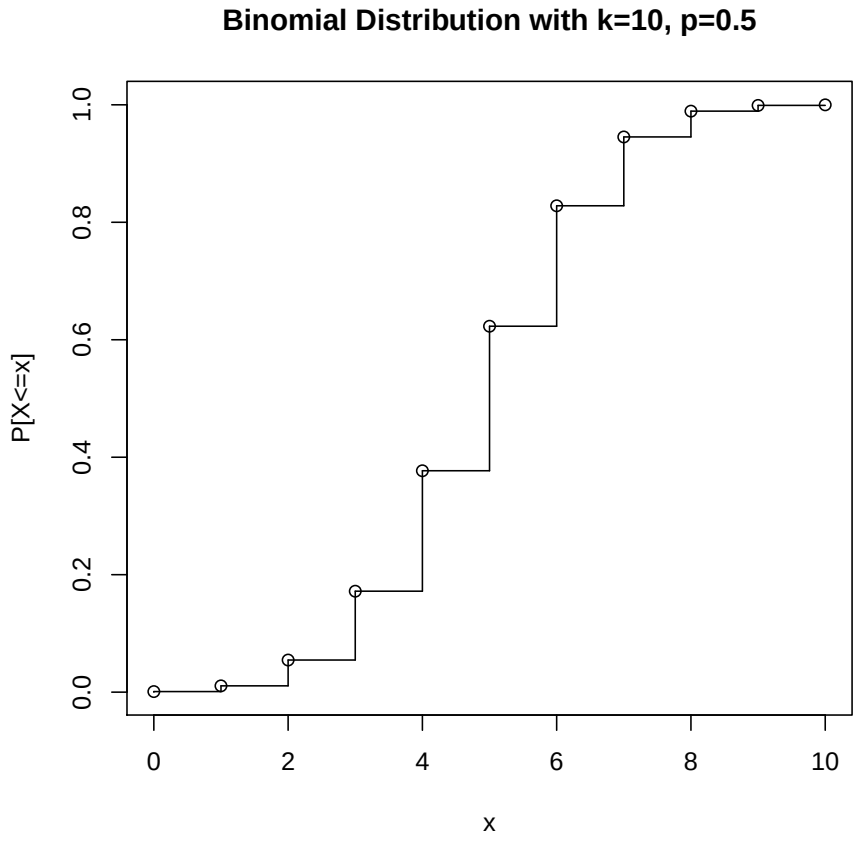
$$\lim_{x \to -\infty} F_X(x) = \lim_{x \to -\infty} P[X \leq x] = 0$$

and

$$\lim_{x \to \infty} F_X(x) = \lim_{x \to \infty} P[X \leq x] = 1$$

Let's see the Distribution Function for $X \sim \mathcal{B}(k, p)$ in R.

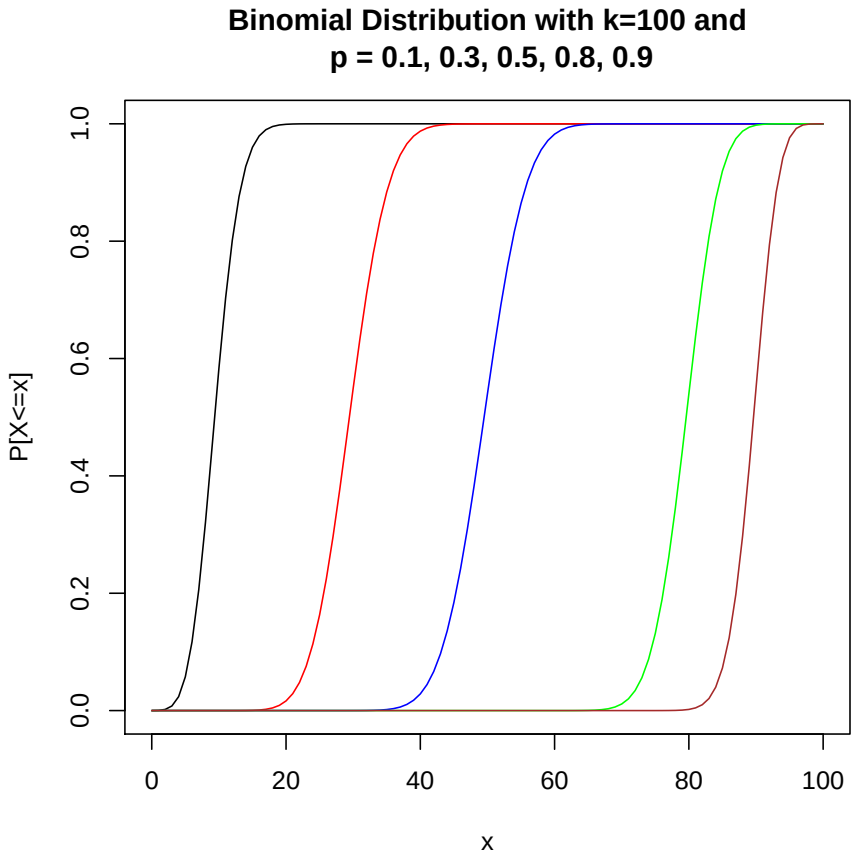
```
> pbinom(c(0:10), size=10, prob=1/2)

 [1] 0.0009765625 0.0107421875 0.0546875000
 [4] 0.1718750000 0.3769531250 0.6230468750
 [7] 0.8281250000 0.9453125000 0.9892578125
[10] 0.9990234375 1.0000000000

> # Let's plot it
> plot(c(0:10), pbinom(c(0:10), 10, 1/2),
+ xlab="x", ylab="P[X<=x]",
+ main = "Binomial Distribution with k=10, p=0.5")
> for(i in 0:9){
+         segments(x0=i, y0=pbinom(i, 10, 1/2), x1=i+1)
+         segments(x0=i, y0=pbinom(i-1, 10, 1/2), x1=i, y1=pbinom(i, 10, 1/2))
+ }
```



Note the form of "stairs" of the function. $F_X(x)$ stay in the same value for x between whole numbers and "jumps" in the following whole number.

A plot with different values of p for $X \sim \mathcal{B}(100, p)$ in R.

```
> plot(c(0:100), pbinom(c(0:100), 100, 0.1),
+ xlab="x", ylab="P[X<=x]", type="l",
+ main = "Binomial Distribution with k=100 and\n p = 0.1, 0.3, 0.5, 0.8, 0.9")
> points(c(0:100), pbinom(c(0:100), 100, 0.3),
+ type="l", col="red")
> points(c(0:100), pbinom(c(0:100), 100, 0.5),
+ type="l", col="blue")
> points(c(0:100), pbinom(c(0:100), 100, 0.8),
+ type="l", col="green")
> points(c(0:100), pbinom(c(0:100), 100, 0.9),
+ type="l", col="brown")
```



7.5. **Definition. Poisson Distribution.** We have seen that when we have a fixed number of trials (k) and a probability of success in each one (p), the Binomial distribution arise naturally as the model for the random variable "Number of successes in k trials". There are many other cases where we count the number of events that happen, for example in a time interval or in a pice of land. For example, the number of telephone calls that an office get in one hour, or the number of fruits that a plant has, or the number of insects in a square meter of land, etc. These cases in many cases fit a probability called Poisson. Let's define such probability.

Let $lambda > 0$ be a real number and X a random variable with possible values $0, 1, 2, \ldots,$. Then if

$$P[X = x] = \frac{e^{-\lambda}\lambda^x}{x!}$$

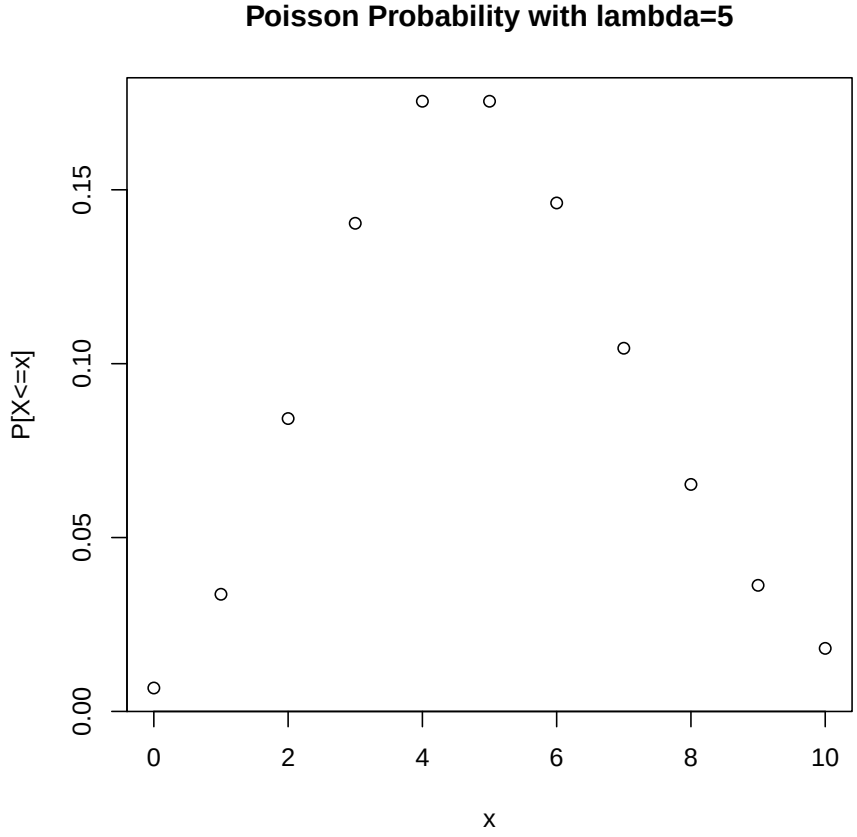
we say that X has a Poisson distribution of parameter λ and we can write $X \sim \mathcal{P}(\lambda)$

The parameter λ is the mean rate at which the phenomenon is happening. For example, if an engineer knows that for a given road pass in average $\lambda = 13.5$ cars per minute, we can calculate the probability that in a given minute pass exactly $x = 0, 1, 2, \ldots,$ cars.

The main assumptions of the Poisson distributions are that the happenings are independent and the rate (λ) is constant.

7.6. **The Poisson distribution in R.** Let's calculate some probabilities in R (use "? dpois" to obtain help about the function related with the Poisson Distribution).

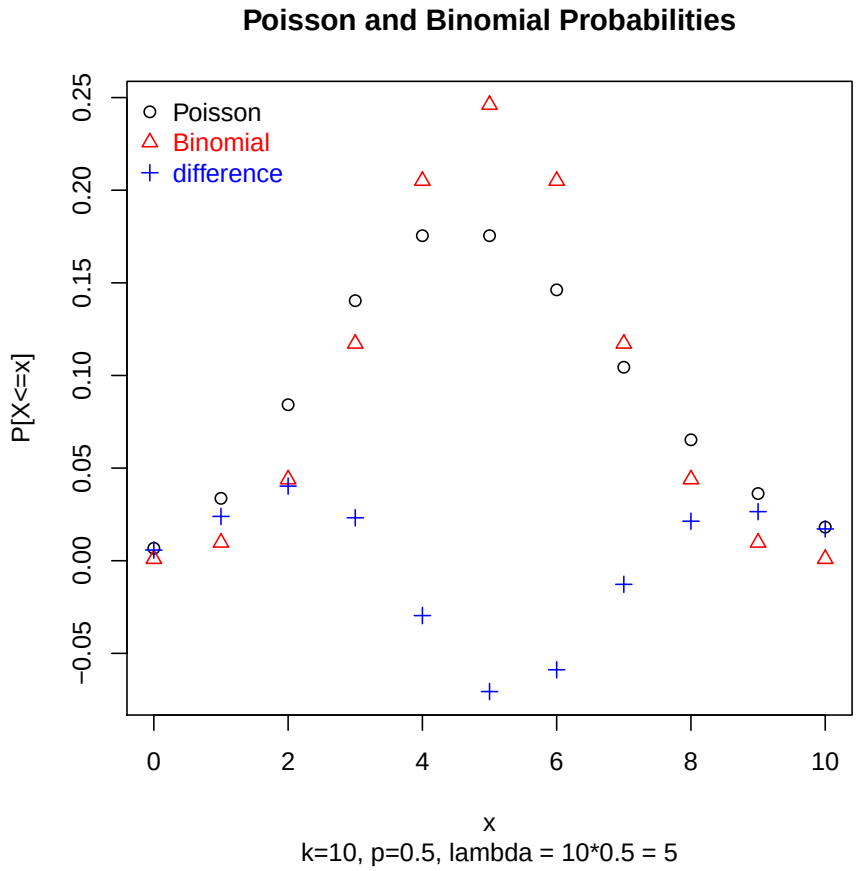
```
> dpois(c(0:10), lambda=5) # A rate of 5
 [1] 0.006737947 0.033689735 0.084224337 0.140373896
 [5] 0.175467370 0.175467370 0.146222808 0.104444863
 [9] 0.065278039 0.036265577 0.018132789
> # Let's plot it
> plot(c(0:10), dpois(c(0:10), 5),
+ xlab="x", ylab="P[X<=x]",
+ main = "Poisson Probability with lambda=5")
```



7.7. **Convergence of the Binomial and Poisson Distributions.** Under certain conditions, the Binomial and Poisson Distributions are alike. The parameters are related in the following way $\lambda \approx kp$. Let's compare (in R) these distributions for some values of the parameters.

$$X \sim \mathcal{P}(\lambda) \text{ and } X \sim \mathcal{B}(k,p) \text{ when } \lambda = kp$$

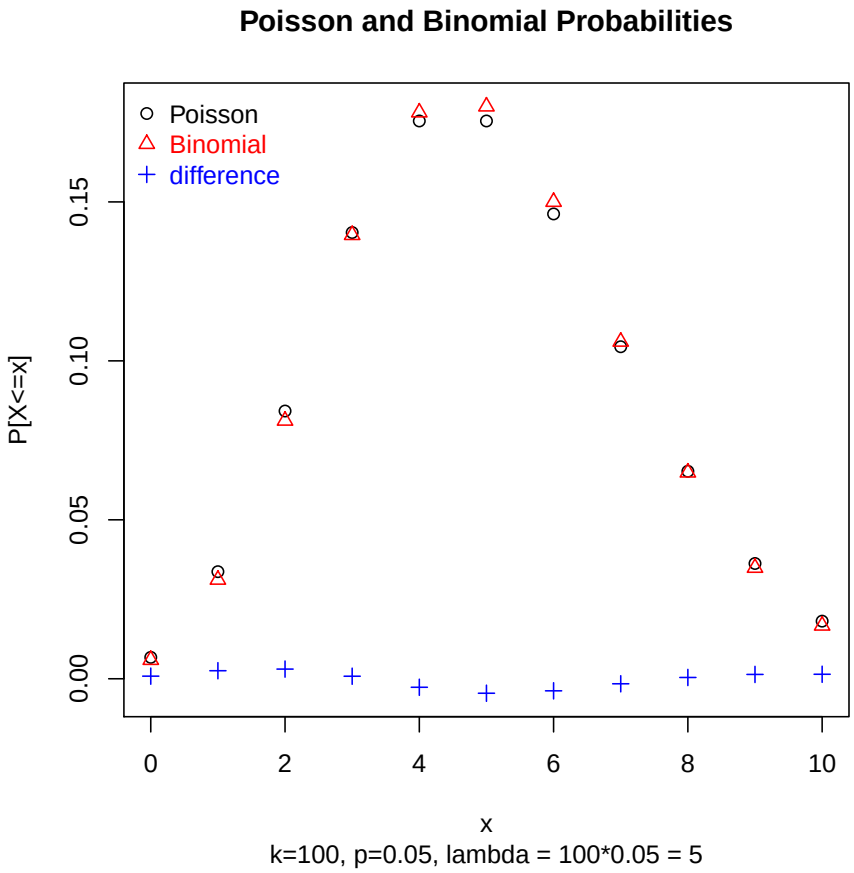
```
> # k = 10, p = 0.5, lambda = 10*0.5 = 5
> poi <- dpois(c(0:10), lambda=5)
> bin <- dbinom(c(0:10), size=10, p=0.5)
> dif <- poi - bin
> # Let's plot it
> plot(c(0:10), poi,
+ xlab="x", ylab="P[X<=x]",
+ ylim = c(min(dif), max(c(poi,bin))),
+ main = "Poisson and Binomial Probabilities",
+ sub="k=10, p=0.5, lambda = 10*0.5 = 5", col="black")
> points(c(0:10), bin, col="red", pch=2)
> points(c(0:10), dif, col="blue", pch=3)
> legend("topleft",
+ legend=c("Poisson", "Binomial", "difference"),
+ pch=c(1,2,3),
+ col=c("black","red","blue"),
+ text.col=c("black","red","blue"),
+ bty = "n")
```



We can see that in that case ($k = 10, p = 0.5, \lambda = 10 * 0.5 = 5$) the approximation is NOT good. Let's try with a larger $k = 100, p = 0.05$) (same $\lambda = 5$). We are going to plot only the "large" probabilities ($x < 11$), and not all the range of X.

```
> # k = 100, p = 0.05, lambda = 100*0.05 = 5
> poi <- dpois(c(0:10), lambda=5)
> bin <- dbinom(c(0:10), size=100, p=0.05)
> dif <- poi - bin
> # Let's plot it
> plot(c(0:10), poi,
+ xlab="x", ylab="P[X<=x]",
+ ylim = c(min(dif), max(c(poi,bin))),
+ main = "Poisson and Binomial Probabilities",
+ sub="k=100, p=0.05, lambda = 100*0.05 = 5", col="black")
> points(c(0:10), bin, col="red", pch=2)
> points(c(0:10), dif, col="blue", pch=3)
```

```
> legend("topleft",
+ legend=c("Poisson", "Binomial", "difference"),
+ pch=c(1,2,3),
+ col=c("black","red","blue"),
+ text.col=c("black","red","blue"),
+ bty = "n")
```



Much better!. In fact, the approximation is good when k is large. Note that the Poisson distribution does not have a maximum for the value of x, thus its Ω is infinite, say the number of events (happenings) can be any whole number (zero included):

$$\Omega = \{0, 1, 2, \dots, \}$$

There are many other probability functions that are important enough to have a proper name, for example Hypergeometric, Negative Binomial, etc. We do not have the time to see more but in a real case you need to look for the proper distribution to model the case under study.

7.7.1. *Homework.*

- 1. Look for help about the Hypergeometric and Negative Binomial distributions (first in R, then you can use for example the Wikypedia). For each one think about an application of these distributions in your area of interest. Plot some examples in R and compare them with the Binomial and Poisson distributions.
- 2a. Assume that there are two unlinked loci $\{A\}$ and $\{B\}$ each one with alleles upper and lower case, that is $\{A, a\}$ and $\{B, b\}$. Also assume that each capital allele gives a unit of tolerance to a pathogen in a discrete scale. Thus, for example the tolerance of the genotype $aabb$ is $T(aabb) = 0$ while $T(Aabb) = T(aaBb) = 1$, $\ldots, T(AABB) = 4$.

  Consider the cross $AaBb$ x $AaBb$ and the random variable T: Tolerance of a individual taken at random from the population. Give the probability function of T (you can do it by hand or using R as a helper)

  2b. Generalize the problem to k loci, say the inner-cross (auto-fertilization) of the genotype $A_1a_1, A_2a_2, \ldots, A_ka_k$, where, again, each upper case allele gives a unit of tolerance and the lowercase is neutral. Find the probability function for T.
- 3. Now consider again the case of the cross in item 2a above, but now we set a third locus $\{C\}$ which allele C is epistatic over the loci $\{A\}$ and $\{B\}$; that is, when a C allele is present in the genotype, then $T = 0$, independently of $\{A\}$ and $\{B\}$. Consider the inner cross $AaBbCc$ (auto-fertilization) and find the probability function for T.
- 4. You know that a gene is transcribing in a given situation with a probability equal to $p = 0.001$. You are going to sequence around a million of sequences from a library. Assume a) Binomial Distribution and b) Poisson Distribution. Simulate 1000 instances of the phenomenon (with each distribution) and show an histogram (in R "hist") of your simulations. Which distribution do you think is better for this situation. Discuss.

Have fun!.