LANGEBIO - BIOSTATISTICS

OCTAVIO MARTÍNEZ DE LA VEGA

7. Continuous Random Variables

So far we have been studying random variables which result from *counting*. However, in many cases we *measure* something as result of our experiment. That is, we obtain a result in a continuous scale. In those cases, we will be assuming that our variables can take any value in a given interval of the real numbers ($\Re$), that is any value in an infinite (non-numerable) set. In that case we will not have a probability function, but instead a *density* function. Let's give a definition.

7.1. **Continuous Random Variable (CRV).** Let X be a random variable with distribution function

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

Then if there exist a continuous function $f_X(\bullet)$ such that

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du$$

for all $x \in \Re$, then X is said to be a *Continuous Random Variable* with density function $f_X(\bullet)$.

Lets see some examples, using R to illustrate.

7.2. **Example. Uniforme Distribution.** Lets introduce first a handy function; the *indicator* function.

$I_A(x) = 1$ if $x \in A$ and $I_A(x) = 0$ if $x \notin A$

that is $I_A(\bullet)$ is a function that *indicate* (with "1") if x exist or not in a given set A. In general, A can be an interval, say $A = (a, b)$ where $a < b$.

Now, assume that there exist a random variable X which can take any value in the closed interval $[a, b]$, and assume that the density of this variable is *uniform*, that is

$$f_X(x) = \frac{1}{b-a} I_{[a,b]}(x)$$

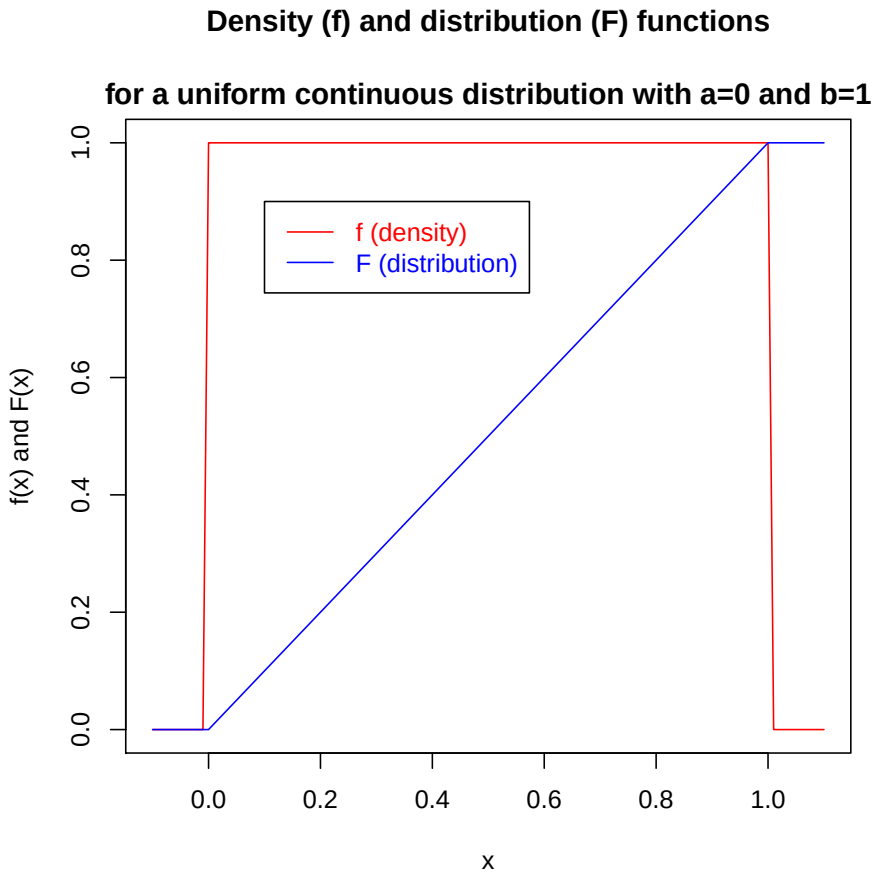
It is easy to see that

Date: April 2012.

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du = (\frac{x-a}{b-a} I_{[a,b]}(x)) + I_{(b,\infty)}(x)$$

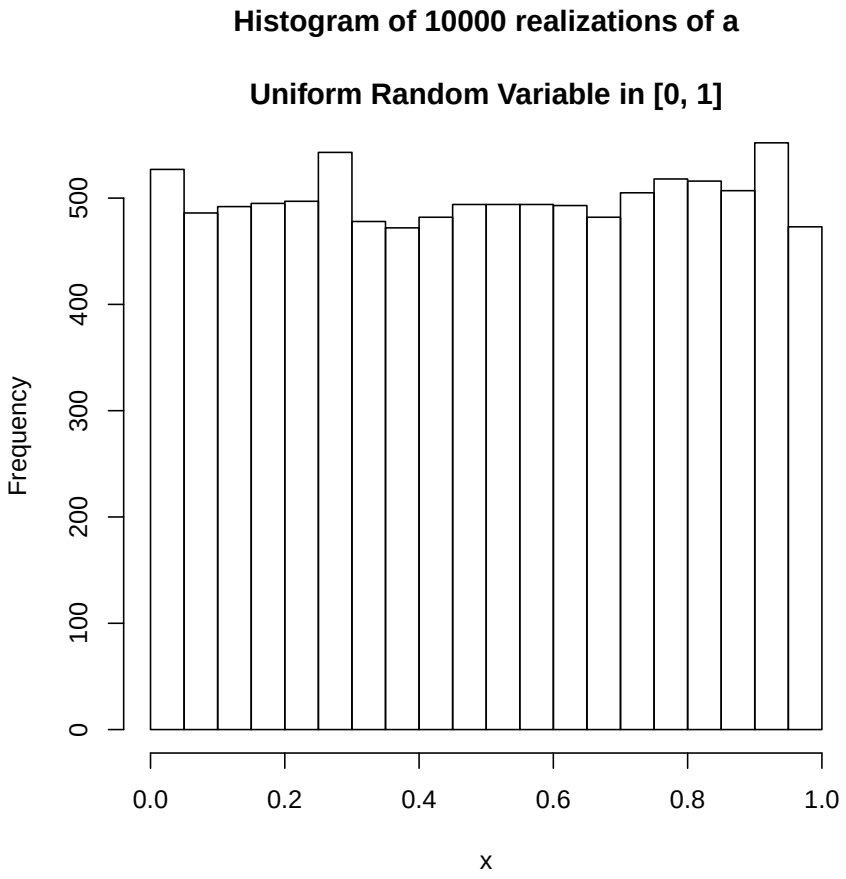
Let's see some examples in R (you can see help for the uniform distribution with "? runif")

```
> # The values for default in R are a=0, b=1; try:
> dunif(-10); dunif(0); dunif(0.5); dunif(0.6); dunif(100)

[1] 0

[1] 1

[1] 1

[1] 1

[1] 0

> # Plotting both, density and distribution functions
> plot(c(-10:110)/100, dunif(c(-10:110)/100),
+ type="l", col="red", xlab="x", ylab="f(x) and F(x)",
+ main="Density (f) and distribution (F) functions\n
+ for a uniform continuous distribution with a=0 and b=1")
> points(c(-10:110)/100, punif(c(-10:110)/100),
+ type="l", col="blue")
> legend(x=0.1, y=0.9, legend=
+ c("f (density)", "F (distribution)"),
+ col=c("red", "blue"), lty=1,
+ text.col=c("red", "blue"))
```



For obvious reasons this is also called the ”rectangular” distribution. Note that an implication of this distribution is that the probability of X taking any value in a sub-interval depends only on the length of the sub-interval. Let simulate some data from this distribution and see its histogram.

```
> hist(runif(10000), xlab="x",
+ main="Histogram of 10000 realizations of a\n
+ Uniform Random Variable in [0, 1]")
```



7.3. **The Normal Distribution.** One of the most important continuous distribution is the "Normal" or "Gauss" distribution (see for example

$http : en.wikipedia.orgwikiNormal\_Distribution$).

This distribution is important, among other reasons, because is the limit distribution of averages of other distributions. It depends on two parameters, μ, a parameter of centrality or position and σ, a parameter of dispersion.

The density function of a normal distribution is given by

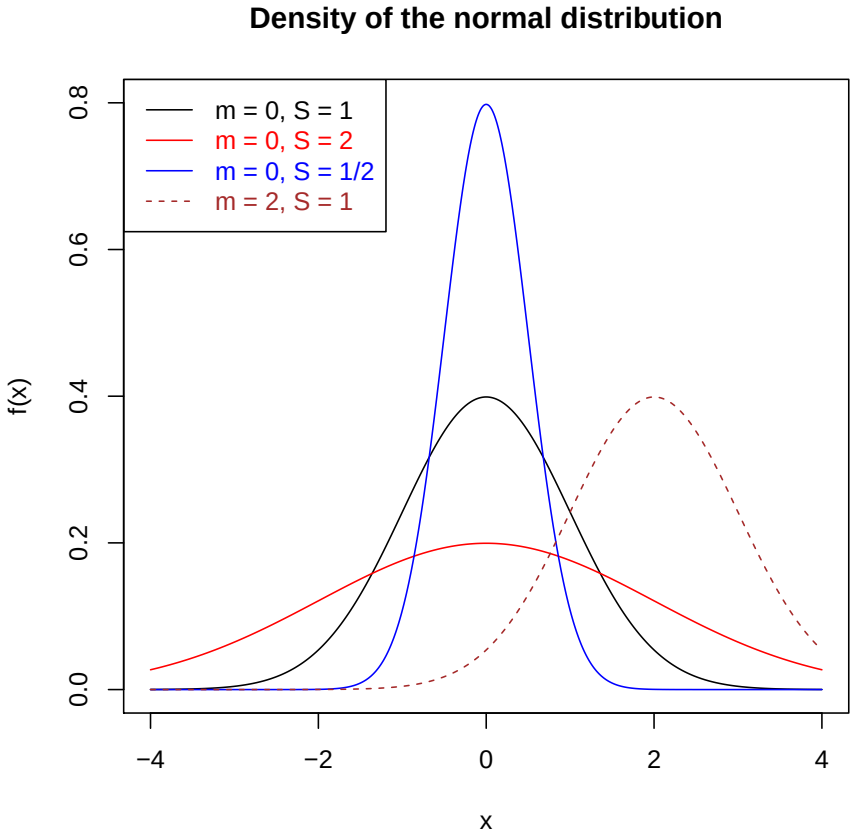
$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where $\sigma > 0$ and μ is any real number. If X has a normal distribution then we write

$$X \sim \mathcal{N}(\mu, \sigma)$$

Let's see the graph of some normal densities using R.

```
> # Mu = 0, Sigma = 1 the "standard" density
> plot(c(-400:400)/100, dnorm(c(-400:400)/100, mean = 0,
+ sd = 1), type="l", col="black", xlab="x", ylab="f(x)",
+ ylim=c(0, 0.8),
+ main="Density of the normal distribution")
> # Mu = 0, Sigma = 2 larger dispersion
> points(c(-400:400)/100, dnorm(c(-400:400)/100, 0,
+ 2), type="l", col="red")
> # Mu = 0, Sigma = 1/2 larger dispersion
> points(c(-400:400)/100, dnorm(c(-400:400)/100, 0,
+ 1/2), type="l", col="blue")
> # Mu = 2, Sigma = 1 larger mean
> points(c(-400:400)/100, dnorm(c(-400:400)/100, 2,
+ 1), type="l", col="brown", lty = 2)
> legend("topleft", legend=c(
+ "m = 0, S = 1",
+ "m = 0, S = 2",
+ "m = 0, S = 1/2",
+ "m = 2, S = 1"),
+ text.col = c("black", "red", "blue", "brown"),
+ col = c("black", "red", "blue", "brown"),
+ lty = c(1,1,1,2))
```



7.4. **Theorem. Central Limit.** One of the most beautiful and powerful theorems of statistics is the *Central Limit* theorem. It states that in *many cases* the average, $\overline{x}$, of a "large" number of random variables with almost **any** distribution will have approximately the normal distribution.

This theorem is beautiful because is fairly general, and for the same reason powerful. It let us make approximate inferences without knowing the *true* distribution of the data, that in many cases could be unknown. In symbols, let $X_i, i = 1, 2, \ldots, n$ be a set of independent identically distributed variables, and we define the *mean* or arithmetic average as:

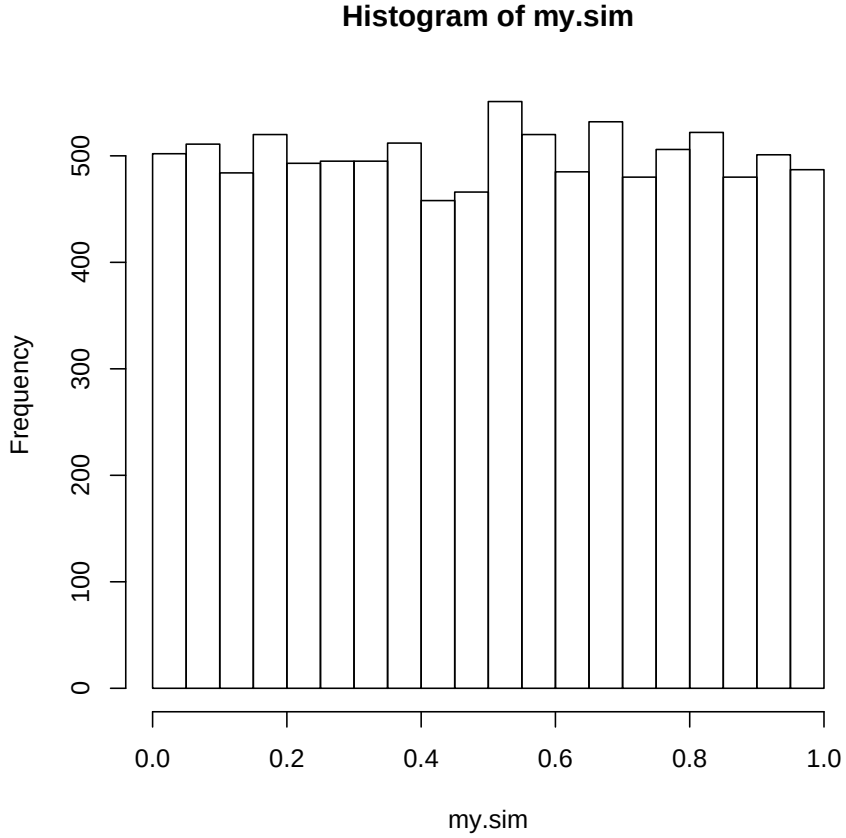
$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

then, for large enough n we have

$$\overline{X} \Rightarrow \mathcal{N}(\mu, \sigma)$$

Let's see how good is the approximation when the original variables are, for example, from the uniform distribution.

```
> my.sim <- runif(10000) # Uniform random variables between 0 an 1
> summary(my.sim) # Basic statistics
     Min.   1st Qu.    Median      Mean   3rd Qu.
0.0002222 0.2488000 0.5064000 0.4995000 0.7489000
     Max.
0.9997000
> hist(my.sim) # An Histogram
```



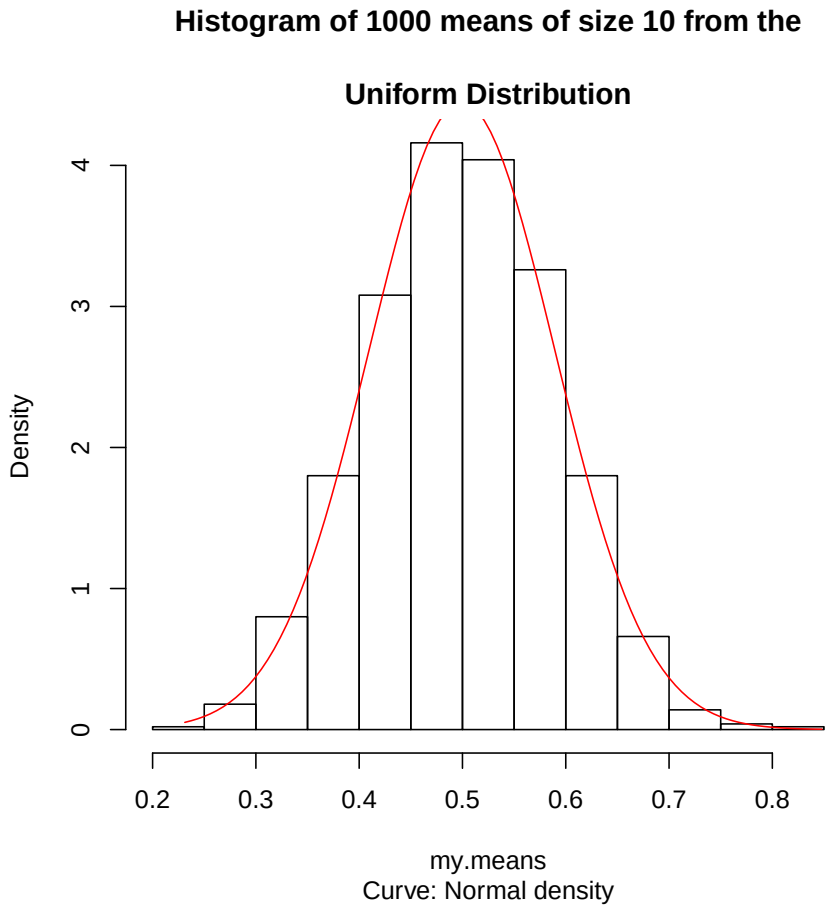
Now, let's obtain 1000 values for the averages of 10 values from the numbers just generated.

```
> my.means <- rep(NA, 1000) # An empty vector
> my.means[1] <- mean(my.sim[1:10]) # First mean
```

```
> j <- 10*(1:1000)
> for(i in 1:999){
+         my.means[i+1] <- mean(my.sim[c((j[i]+1):j[i+1])])
+ }
> summary(my.means) # A summary of the means
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2312  0.4385  0.4993  0.4995  0.5621  0.8419
```

```
> hist(my.means, freq=FALSE,
+ main="Histogram of 1000 means of size 10 from the\n
+ Uniform Distribution",
+ sub="Curve: Normal density") # Histogram in density
> # Compare with the normal density
> my.points <- min(my.means) + c(0:100)*((max(my.means)-min(my.means)))/99
> points(my.points, dnorm(my.points, mean(my.means),
+  sd(my.means)), type="l", col="red")
```



We can see that even with a relatively small sample size ($n = 10$) the convergence of the means from the uniform to the normal distributions is very good. As mentioned, many statistical inference methods are based in this theorem.

8. Expectation and Variance of Random Variables

Statistics is about summarizing the information given by the data. We use, for example, the mean and standard deviation to have an idea of the tendency of a group of data. Here we will define the expectation and variance of a random variable, two functions that can help to summarize the behavior of these.

8.1. **Definition. Expectation.** Let X be a discrete random variable with probability function $f_x(x)$. Then its expectation X, denoted by $E[X]$ is defined as

$$E[X] = \sum_i x_i f_X(x_i) = \sum_i x_i P[X = x_i]$$

where the sum is performed for all values of x_i for which $P[X = x_i] > 0$. If that sum does not converge, then $E[X]$ is not defined.

If X is a continuous random variable, with density function $f_X(x)$, then we define

$$E[X] = \int_{i=-\infty}^{i=\infty} x f_X(x) dx$$

if this integral converges.

8.2. **Properties of the Expectation.** Let X be a discrete random variable with finite $E[X]$. Then, if a and b are some constants we have:

$$E[a + bX] = a + bE[X]$$

That is, if we have a linear function of X, say $Y = a + bX$, we can obtain $E[Y]$ using the formula above.

To see this, note that by definition

$$E[a + bX] = \sum_{i=1}^{n} (a + bx_i) P[X = x_i]$$

$$E[a + bX] = (\sum_{i=1}^{n} aP[X = x_i]) + (\sum_{i=1}^{n} bx_i P[X = x_i])$$

$$E[a + bX] = a(\sum_{i=1}^{n} P[X = x_i]) + b(\sum_{i=1}^{n} x_i P[X = x_i])$$

But $\sum_{i=1}^{n} P[X = x_i] = 1$ and $(\sum_{i=1}^{n} x_i P[X = x_i]) = E[X]$, which proves the above equality. This is valid even if n is ∞, when $E[X]$ converges.

The same property is valid if X is a continuous random variable.

8.3. **Definition. Variance.** Let X be a discrete random variable with probability function $f_x(x)$. Then its variance, denoted by $V[X]$ is defined as

$$V[X] = E[(X - E[X])^2$$

Note than in the above definitions the value of $E[X]$ is assumed to exist and to be a constant.

8.4. **Properties of the Variance.** Let X be a discrete or continuous random variable with finite $E[X]$. Then,

$$V[X] = E[X^2] - (E[X])^2$$

To see this, note that by definition
$V[X] = E[(X - E[X])^2]$
$V[X] = E[X^2 - 2XE[X] + E[X]^2]$
But expectation is a lineal operator and so
$V[X] = E[X^2] - 2E[X]E[X] + E[X]^2 = E[X^2] - E[X]^2$

If a and b are some constants we have:

$$V[a + bX] = b^2V[X]$$

To see this note that by using the above property,
$V[a + bX] = E[(a + bX)^2] - (E[a + bX])^2$
And using the fact that $E[\bullet]$ is a lineal operator we obtain
$V[a + bX] = E[a^2 + 2abX + b^2X^2] - (a + bE[X])^2$
$V[a + bX] = a^2 + 2abE[X] + b^2E[X^2] - (a^2 + 2abE[X] + b^2E[X]^2)$
and finally
$V[a + bX] = b^2E[X^2] - b^2E[X]^2 = b^2V[X]$

8.5. **Example.** $E[\bullet]$ **and** $V[\bullet]$ **in R.** The name *expectation* for the function $E[\bullet]$ comes to the fact that if we obtain many realizations of the random variable X and obtain its average we will obtain a number that, likely, will be *close* to $E[X]$. Lets program in R these functions for the discrete case, and test those functions with some random variables.

```
> expect <- function(x, p){sum(x*p)} # Easy!
> varian <- function(x, p){expect(x^2, p) - (expect(x, p)^2)}
```

Now, in these functions x **must** be a vector containing all the possible values of X for which we have $P[X = x] > 0$ and p **must** be the vector containing the corresponding probabilities.

Let's test the functions with some random variables. Let X be a random variable defined by its probability function as:

$P[X = 0] = 1/2$, $P[X = 1] = 1/2$. We want $E[X]$ and $V[X]$:

```
> expect(x=c(0,1), p=c(1/2, 1/2)) # E[X]
[1] 0.5
> varian(x=c(0,1), p=c(1/2, 1/2)) # V[X]
[1] 0.25
```

Now, we can simulate many instances (realizations) of that variable and see which are the values of it average (mean) and (sampling) variance.

```
> my.sample <- sample(x=c(1,0), size=1000, replace=T)
> mean(my.sample) # The arithmetic average
[1] 0.498
> var(my.sample) # The sampling variance
[1] 0.2502462
```

Now, let's try our functions with more complicated random variables, for example, Binomial and Poisson. With the Poisson we have a problem, because it can have an infinite number of values. However, for large enough values of the variable, the probabilities become very small and we can obtain a good approximation of the mean and variance. Let's try.

```
> # Binomial with k=10, p=0.5; note x = 0, 1, ..., 10
> expect(c(0:10), dbinom(x=c(0:10), size=10, prob=0.5))
[1] 5
> varian(c(0:10), dbinom(x=c(0:10), size=10, prob=0.5))
[1] 2.5
> # For Poisson with parameter lambda=5
> expect(c(0:100), dpois(x=c(0:100), lambda=5))
[1] 5
> varian(c(0:100), dpois(x=c(0:100), lambda=5))
[1] 5
> # Terms of which size are we neglecting?
> # In the Poisson expectation (and variance) we are
> #    neglecting numbers smaller than
> 100*dpois(x=100, lambda=5)
[1] 5.695402e-89
```

Now, lets try some transformations of random variables.

Assume that there is an infestation of a worm in a plantation of apple trees and the number of bugs per plant is a Poisson random variable of parameter $\lambda = 12.45$. It is known that each worm eats, in its life time, 2.25 g of fruit.

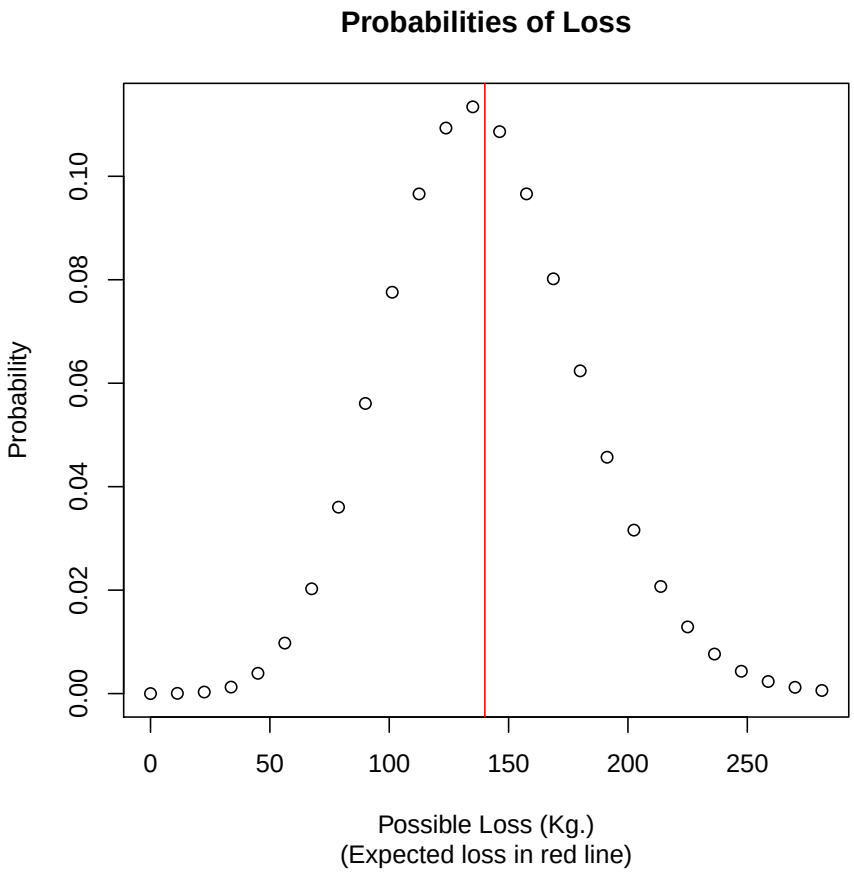
Find the expected loss of production per Ha. if there are 5,000 plants per Ha.

(I assume that you do not know which is the $E[X]$ for a Poisson distribution, but probably you are suspecting the answer by now)

We can define: $Y = 2.25X$, the loss per worm, and then the total loss per Ha. will be $Z = 5000 * 2.25 * X$, and naturally the expected loss will be: $E[Z] = 5000 * 2.25 * E[X] = 11250 * E[X]$

The farmer wants you to give a table with the probabilities of each loss that you can calculate, as well as the ”expected” loss in Kg.

```
> expect(c(0:500), dpois(c(0:500), lambda=12.45)) # E[X]

[1] 12.45

> # Thus, the expected loss is
> 11250*12.45 # in g.

[1] 140062.5

> 11.250*12.45 # in Kg.

[1] 140.0625

> # Trying our theorem E[Z] = 11250*E[X]
> expect(11250*c(0:500), dpois(c(0:500), lambda=12.45)) # E[Z]

[1] 140062.5

> # Let's graph the putative loss in Kg and their probabilities:
> plot(11.25*c(0:25), dpois(c(0:25), lambda=12.45),
+ xlab="Possible Loss (Kg.)", ylab="Probability",
+ main="Probabilities of Loss",
+ sub="(Expected loss in red line)")
> abline(v=11.250*12.45, col="red")
```



8.6. **Theorem. Expectation and Variance of a Binomial RV.** Assume that

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

thus we have that

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

We will see that

$$E[X] = kp \text{ and } V[X] = kp(1-p)$$

It is useful to remember the following theorem:

$$(a+b)^k = \sum_{x=0}^{x=k} C_x^k a^x b^{k-x}$$

Now, by definition

$$E[X] = \sum_{x=0}^{x=k} xP[X = x]$$
$$= \sum_{x=1}^{x=k} xP[X = x]$$
$$= \sum_{x=1}^{x=k} \frac{xk!}{x!(k-x)!} p^x (1-p)^{k-x}$$
$$= \sum_{x=1}^{x=k} \frac{k!}{(x-1)!(k-x)!} p^x (1-p)^{k-x}$$

Now, note that all terms of the sum have the term kp; taking out of the sum this common factor we get

$$E[X] = pk \sum_{x=1}^{x=k} \frac{(k-1)!}{(x-1)!(k-x)!} p^{x-1} (1-p)^{k-x}$$

Now let's $r = x - 1$ and $s = k - 1$, thus r will take the values $0, 1, \ldots, k-1$ ($k - 1 = s$) and also note that $k - x = s + 1 - r + 1 = s - r$ thus we can re-write the previous equation as

$$E[X] = pk\{\sum_{r=0}^{r=s} \frac{s!}{r!(s-r)!} p^r (1-p)^{s-r}\}$$

But the expression between brackets is just the sum of all binomial probabilities of a random variable, say R such tahat

$$R \sim \mathcal{B}(s, p)$$

thus (using the binomial theorem if you like) we can see that

$$\{\sum_{r=0}^{r=s} \frac{s!}{r!(s-r)!} p^r (1-p)^{s-r}\} = \sum_{r=0}^{r=s} P[R = r] \equiv 1$$

And thus we have proven that $E[X] = pk$.

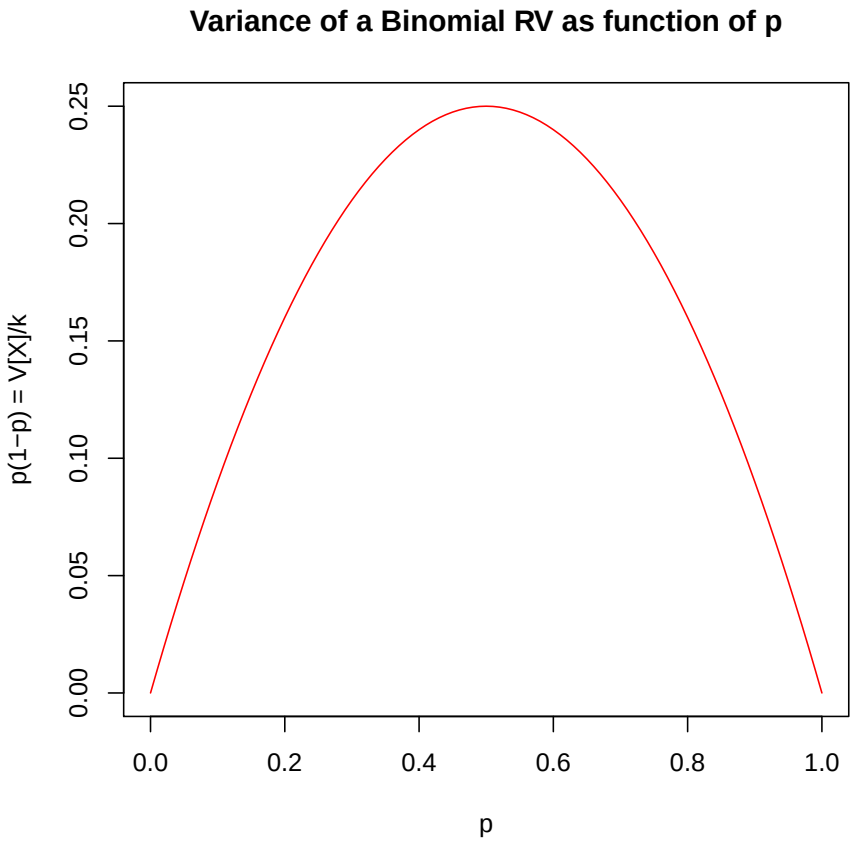
Following the same path, we can see that

$$E[X^2] = kp(1 - p + kp)$$

that completes the demonstration of $V[X] = kp(1-p)$

How is the graph of $V[X]$ as function of p? (for any k constant). Let's see

```
> p <- c(0:100)/100 # p = 0, 0.01, ... 1.
> plot(p, p*(1-p), xlab="p", type="l", col="red",
+ ylab="p(1-p) = V[X]/k",
+ main="Variance of a Binomial RV as function of p")
```



Note that the maximum of $V[X]$ $(1/4k)$ is reached in the point $p = 1/2 = 0.5$. Does it make sense?. Discuss.

8.7. **Theorem. Expectation and Variance of a Poisson RV.** Assume that

$$X \sim \mathcal{P}(\lambda)$$

That is X is a RV with the Poisson distribution of parameter $\lambda > 0$ and thus

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

Then

$$E[X] = \lambda \text{ and } V[X] = \lambda.$$

8.8. **Definition. Independence of two RV.** Let X and Y be two random variables (discrete or continuous), then we say that X and Y are *independent* if and only if for all x and y the following condition is fulfilled:

$$P[X = x \cap Y = y] = P[X = x]P[Y = y]$$

(for discrete variables) or

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

(for continuous variables).

8.9. **Definition. Covariance.** Let X and Y be two random variables (discrete or continuous), then we define the *Covariance* of X and Y as

$$C[X, Y] = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

8.10. **Definition. Covariance of independent variables.** Let X and Y be two independent random variables. Then

$$C[X, Y] = 0$$

To see this, lets take the discrete case. We have that

$$E[XY] = \sum_i \sum_j x_i y_j P[X = x_i \cap Y = y_j]$$

But, because these variables are independent, we can write

$$E[XY] = \sum_i \sum_j x_i y_j P[X = x_i]P[Y = y_j]$$

$$E[XY] = \sum_i x_i P[X = x_i] \sum_j y_j P[Y = y_j]$$

and thus we have that

$$E[XY] = E[X]E[Y]$$

then substituting in the definition of $C[X, Y]$ we obtained the result $C[X, Y] = 0$

Note: Assume that $Y = a + bX$ where X is a random variable. Lets obtain $C[X, Y]$.

Just note that

$$E[Y] = a + bE[X]$$

and

$$E[XY] = E[X(a + bX)] = aE[X] + bE[X^2]$$

By definition,

$$C[X, Y] = E[XY] - E[X]E[Y]$$

and substituting the values found above

$$C[X, Y] = aE[X] + bE[X^2] - E[X](a + bE[X])$$
$$C[X, Y] = aE[X] + bE[X^2] - aE[X] - bE[X]^2$$
$$C[X, Y] = b(E[X^2] - E[X]^2) = bV[X]$$

8.11. **Correlation of random variables.** Let's consider the following function

$$R(X, Y) = \frac{C[X, Y]}{\sqrt{V[X]V[Y]}}$$

$R(X, Y)$ is call the *Pearson's correlation coefficient* by the following reason:

a) Assume that X and Y are independent, then it is easy to see that $C[X, Y] = 0$ (check that)

b) Now assume that $Y = a + bX$, that is Y is a lineal function of X, then we have seen that

$$C[X, Y] = bV[X]$$

and also remembering that

$$V[Y] = V[a + bX] = b^2V[X]$$

thus in this case

$$R(X, Y) = \frac{C[X, Y]}{\sqrt{V[X]V[Y]}}$$
$$R(X, Y) = \frac{bV[X]}{\sqrt{b^2V[X]V[X]}} = 1$$

8.11.1. *Homework.*

- 1. Consider the random variables of example 7.1.1. item 2 (file *Biostatistics2012_4.pdf*):

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

  M : The maximum of the two dices

  and

  Z : The product of the first and the second dices ($Z = xy$)

  1.1 Can you say, just by looking at their tables of probabilities, which of these variables has a larger expectation and variance?

  1.2 For these random variables, find their expectation and variance.

1.3 Are any pair of these variables $\{W, M\}$, $\{W, Z\}$ or $\{M, Z\}$ independent? You do not need to make all the counts, ... Think for a while to find a simple argument.

1.4 For the pair $\{W, M\}$ find the conjoint probability function $P[W = w \cap M = m]$

- 2. Program a function, say **covari**, to find the covariance of two discrete random variables. Can be a bit tricky. If you cannot do it, think what is the difficulty.
- 3. Assume that it is known that the production of maize per Ha in México follows an approximate normal distribution with mean $\mu = 4.3$ Ton. x Ha. and variance $\sigma^2 = 5$. Using this knowledge find the probabilities that an Ha. sampled at random will produce:

  3.1 More than 5 T/Ha.

  3.2 Between 2 and 6 T/Ha.

  3.3 Nothing at all (Note that negative productions could be taken as 0 T/Ha.).

  3.4 More than 8 T/Ha.
- 4. Consider, again the random variable W of the first item in this homework and define the mean of 4 independent realizations of this variable, say

$$\overline{W} = \frac{1}{4}\sum_{i=1}^{i=4} W_i$$

  Find the approximate distribution of $\overline{W}$ in R using simulation and compare it with the normal distribution. How good is the agreement?

Have fun!