NRES 798 — Lab 4

RVs and distributions

Assume that the proportion of female moose in a population is $p = 0.6$. We model moose observations as a random process with a sample space $\{female, male\}$.

Define a random variable as *male* → 0, *female* → 1. This RV has a Bernouilli distribution (sample space $\{0,\ 1\}$), and can be specified by the *probability density function* (PDF) f, with $f(0) = P(\{0\}) = 1 - p = 0.4$, $f(1) = P(\{1\}) = p = 0.6$. Other names for the PDF of a discrete RV are *probability mass function* (PMF), or simply *density.* The textbook uses the term *distribution function*, that should not be confused with its common usage as an alternative name for the *cumulative distribution function* (CDF).

Write down the possible results of a sequence of $n = 4$ observations, something like:
`0000, 0001, 0010, 0011, ...`
Or perhaps better, for cases with 0 females, 1 female, 2 females . . . :
`0000, 0001, 0010, 0100, ...`
By hand. There are suggestions on ways in which something like this could be automated in *R* at the end of the Lab. Anyway, it is usually a good idea to work out the essentials of a problem with pencil and paper before starting to type on the computer. Check that you got the $2^n = 16$ possibilities (think about it, 2 for the first digit, 2 for the second, . . .).

Write down the number of females observed in each case. How many cases of 0, 1, 2, 3, and 4 females? As shown in the textbook, the number of *combinations* of x out of n should be

$$\frac{n!}{x!(n-x)!} = \binom{n}{x} .$$

Verify in *R*. The factorial $n!$ is the product $1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$. It could be computed in *R* directly like this, or as `prod(1:n)`, or as `factorial(n)`. Look up the functions in the *Help*, and check that you get the same numbers either way. In addition, there is the function `choose(n, x)` for $\binom{n}{x}$.

How many combinations are there of 10 towns out of 349? What is the total number of combinations?

Assume that the moose observations are independent. That is, the probability of a sequence x_1, x_2 is $f(x_1)f(x_2)$. Write down, in terms of p, the product for `1011`. Convince yourself that this can be written as $p^3(1-p)$. And that in general, the probability of a sequence with x females is $p^x(1-p)^{n-x}$ (for any n and p).

What is the distribution of the number of females? The number is a RV, say X, because it is derived from the Bernouilli RVs in the sequence. Write down the sample space. Convince yourself that the probability $P(X = x) = f(x)$ of observing x females is the sum of the probabilities of all the ways of observing x females. And that, therefore, the PDF is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \ .$$

Use this equation in *R*, with $n = 4$ and $p = 0.6$, to calculate the PDF value for each sample point. As a vector, `f <- ...`. Does it add to 1?

This is a binomial PDF, for which *R* has the function `dbinom` (d for 'density'). Use `dbinom`, and check that you get the same numbers as before.

Plot `f` (over the vector of sample points). Use `type='h'`.

Calculate the expected value (weighted average) $E[X] = \sum X_i f(X_i)$. One way is to multiply the x and f vectors, and then use `sum`. Display the intermediate steps. There is also a function `weighted.mean`, look it up. Compare to the graph, looks reasonable? It should be np, check.

Calculate the distribution variance $E[(X - E[X])^2]$. Go through step-by-step, displaying the intermediate results. See the book for the general form of the binomial variance, or ask `wolframalpha.com`. Check.

The cumulative d.f. (CDF) is $F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$. Compute it and store in a vector `F`. You can use `cumsum`, look it up and make sure you understand what it does.

The probability of X being either 1 or 2 can be computed as $f(1) + f(2)$. It is also $F(2) - F(0)$ (try it). Why?

R has `pbinom` for the PDF, try it and check that you get the same.

Plot F, use points (the `type` default). Taking the sample space as the whole set of real numbers (it is OK to have sample points with probability 0), what is the value of $F(1.3)$? Of $F(-1)$? Of $F(4.5)$?

To (partially) reflect this, add some lines, plotting F again but this time with `type='s'`. And using `points()` instead of `plot()`, so that it overplots on the old graph. Note that the vertical segments do not correspond to function values, a function has at most one value for any x, in this case the circles (and the horizontal segments if we include non-integers in the sample space).

In hypothesis testing, we are interested in values that have a given probability of being exceeded. That is, we are interested in the inverse $x = F^{-1}(p)$ of $p = F(x)$ (or rather, 1 minus that; think about it!). For that, re-draw the graph swapping the x and y axes, using `type='S'` instead of `s` to get the same shape. In *RStudio* you can switch between the graphs clicking the arrow to compare. That gives us the answer, almost. The problem is that the horizontal portions of the curve are not "valid", remember that they were the verticals in F. The inverse is not defined for probabilities other than those corresponding to the x's in the sample space. So we cheat, and use those points anyway. Actually, we re-define to "the smallest values that have…", but this may be easier to remember. This is sometimes called the "quantile function". Continuous RVs do not have this problem, there the quantile function is simply F^{-1}.

In *R* the binomial quantile function is (obviously!) `qbinom`. Try it and see if you get the same.

Let's now look at the Poisson distributions. For instance, as a model for the number of seedlings in a random 1 m^2 quadrat, assuming that there are $50,000$ seedlings per hectare, or $\lambda = 5$ seedlings / m^2. The PDF for the number of seedlings in the quadrat is

$$f(x) = \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!} \; .$$

Do with the Poisson everything that you did with the binomial. Remember (?) that $\mathrm{e}^y \equiv \exp(y)$; there is a corresponding function `exp` in *R*. You may not be able to represent all the sample space, but use "enough" of it. Do `?Distributions` to see the *R* functions available for the Poisson.

We had seen an example where *Rhexia* could be either present or absent in a town, with probability $p = 0.02$. The number of towns with Rexia out of $n = 349$ was a binomial $X \sim Binom(n, p)$. When p is small and n is large, the binomial is well approximated by a Poisson with parameter $\lambda = np$. Alternatively, we could have model this directly as a Poisson, with *Rhexia* appearing at a rate of λ populations per town.

Calculate the binomial and Poisson PDFs for the *Rhexia* example. Compare graphically. Try different values of n and p, see how far you can push the approximation.

Finally, see how you could use `expand.grid` to generate sequences like those in the moose observations. For instance, define the *factor* (categorical variable) `x <- c(1, 0)`, and do `expand.grid(x, x, x, x)`. It might be useful to know that `apply(frame, 1, fun)` applies a function such as `sum` to the rows of a data frame (with 2 it is applied to the columns). Remember also about coercion, and that it is possible to operate on whole (numeric) data frames or arrays like `0.6 * frame + 0.4 * !  frame`.