LANGEBIO - BIOSTATISTICS

OCTAVIO MARTÍNEZ DE LA VEGA

Apology: I am working on these notes while learning a new markup language (LaTeX), thus I do not pay much attention to grammar or spelling. Pardon me for any idiocies in my English!

3. Probability Definitions

In everyday parlance we say that something is "probable" when it is likely to happen, as in "*It is probable that it will rain today*". In this section we will see the formal definitions of *probability*. Then we will need a bit of set theory to be able to board elemental Probability Theory. We will then review some theorems and examples of it's application, trying to put emphasis in Biology.

3.1. **Definition. Classical or *a priory* Probability.** If a *random experiment* can occur into n mutually exclusive and equally likely ways, and if n_A of them have the attribute "A", then we say that the probability of A is n_A/n and we can write $P[A] = n_A/n$.

This is we divide the number of favorable cases between all possible cases. Note that to apply this definition we must be sure that all cases are *mutually exclusive* (i.e., they cannot happen at the same time) and *equally likely* (we do not have reason to think that one is more likely than any other).

3.1.1. *Example with a dice.* What is the probability of obtaining an odd number if you throw a dice?.

A dice have six sides, numbered 1, 2, 3, 4, 5, 6. If the dice is *honest* (not-charged to a particular side) and if we call A the event that we obtain an odd number (1, 3 or 5), then, applying the previous definition we get

$$P[A] = n_A/n = 3/6 = 1/2$$

3.1.2. *Example in Genetics.* The gene that determine albinism in human is recessive and the locus is in a somatic chromosome. A couple of normal skin color has an albine child. What is the probability that their next offspring is also albine?

Date: April 2012.

A putative solution: There are two non-overlapping (mutually exclusive) results: "*normal*" or "*albine*". Thus defining "C" as "*The next child will be albine*" we could write (using the above definition):

$$P[C] = 1/2$$

Is that correct?

No, it is not. Because if both parents are of normal skin color, then they have the normal allele, say "A" as well as the allele for albinism, say "a"; that is, both are heterocygous of genotype "Aa", and then we have the cross "Aa" x "Aa". And the result of all possibilities is (given in a Punnett square):

	A	a
A	AA Normal	Aa Normal
a	Aa Normal	aa Albine

Thus, the correct answer for the question is off course:

$$P[C] = 1/4$$

The confusion aroused because we considered the two possibilities, Normal and Albine, as *equally likely*, when in fact they are not. Thus, one need to be careful!

3.2. **Definition. Frequencialist or *a posteriori* probability.** Let be a random experiment that can be repeated indefinitely under the same conditions. Assume that you can classify the results into two non-overlapping (mutually exclusive) categories as "A" or "$Not - A$". Now, let "n" be the number of times that the experiment is performed and "n_A" the number of times that the result has the attribute "A". Then, the probability of "A" is defined as the limit of "n_A/n" when n is increased without indefinitely. We can write:

$$P[A] = \lim_{n \to \text{inf}} \frac{n_A}{n}$$

3.2.1. *Example with a coin.* Assume that we can throw a coin indefinitely, and call "H" the event that the "head" side is upside at the end. If the coin in *honest*, i.e., well balanced, then we could apply our *a priory* definition and just say that

$$P[H] = 1/2$$

Now, the problem with our second (*a posteriori*) definition is that, of course we can not throw indefinetely a coin (we have better things to do). However we could "simulate" what happen with n_A/n when the number of experiments grows very large. To this aim we can program a function which gives "at random" 1 or 0 with the same probability; for example:

```
> my.coin <- function(n){1*(runif(n)<=0.5)}
```

This function will give a vector of size n -Number of times that the coin is thrown, which will contain "1" in the i-th element if the coin was "head" and 0 otherwise (if it was "tail"); let's try it with a small number of experiments, say $n = 10$:

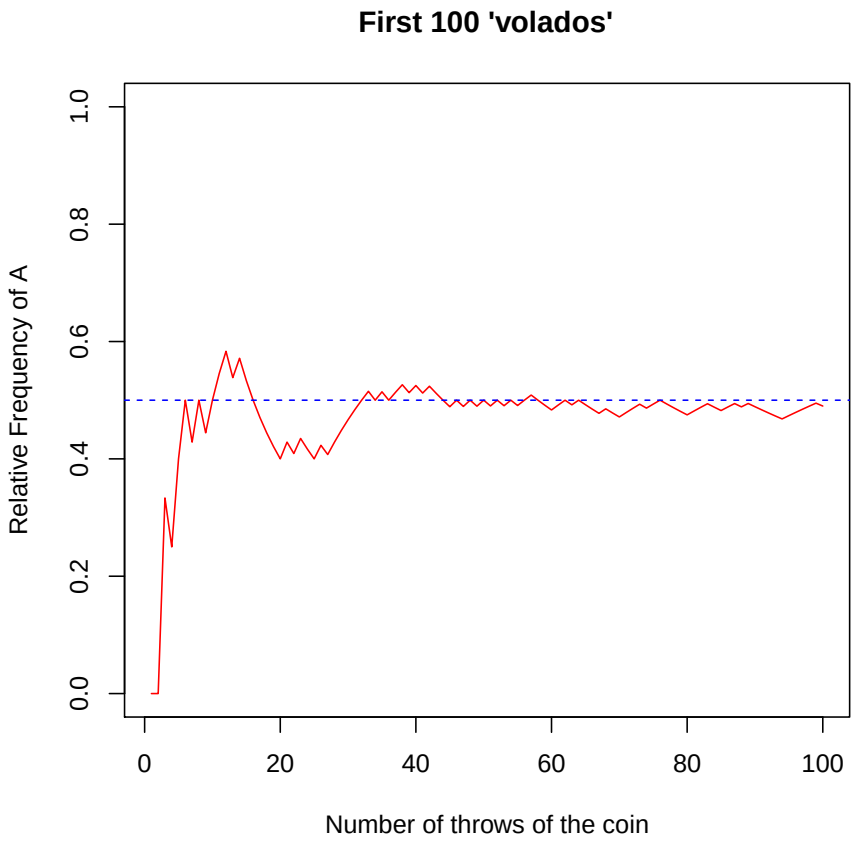
```
> my.coin(10)
 [1] 0 0 1 1 1 1 1 0 0 0
```

Now, let's do a really large number of coin throws, say $n = 10,000$ and put the result in a vector called *coin*.

```
> coin <- my.coin(10000)
> coin[1:20] # The first 20 values of the 10,000
 [1] 0 0 1 0 1 1 0 1 0 1 1 1 0 1 0 0 0 0 0 0
```

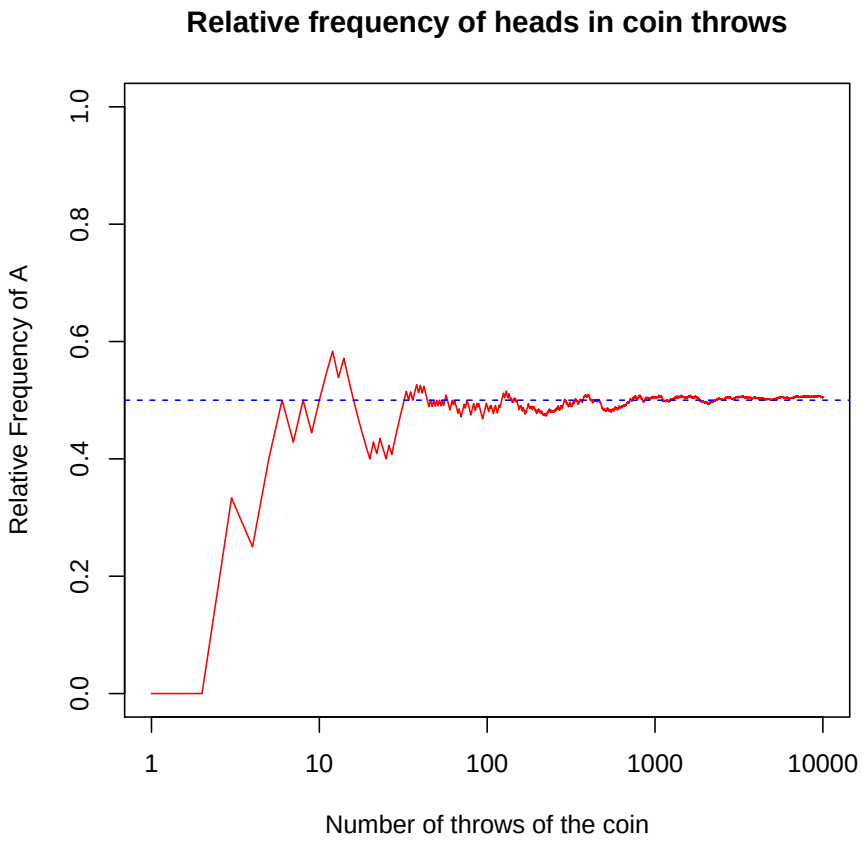
Now, let define a data.frame to keep our coin vector, and calculate the relative frequency of A in the $i-th$" throw of the coin, say, n_i/n for $i = 1, 2, ..., 10000$. The objective of this small exercise is to see if the relative frequency tends to stabilize into a given value (which one it could be?)

```
> coin.d <- data.frame(coin, rep(NA, 10000))
> names(coin.d)[2] <- "P.A"
> coin.d$P.A[1] <- coin.d$coin[1]/1 # Can be 0 or 1
> for(i in 2:10000){ # We open a "loop"
+ coin.d$P.A[i] <- sum(coin.d$coin[1:i])/i
+ } # Close the loop
> # Let see the first 10 rows...
> coin.d[1:10,]
   coin       P.A
1     0 0.0000000
2     0 0.0000000
3     1 0.3333333
4     0 0.2500000
5     1 0.4000000
6     1 0.5000000
7     0 0.4285714
8     1 0.5000000
9     0 0.4444444
10    1 0.5000000
> # And let's plot the first 100 cases...
> plot(c(1:100), coin.d$P.A[1:100], type="l", col="red",
+ xlab="Number of throws of the coin",
+ ylab="Relative Frequency of A", ylim=c(0,1),
+ main="First 100 \'volados\'")
> abline(h=0.5, col="blue", lty=2) # accessory line
```



Now, let's see the full data (logarithm of the number of throws).

```
> plot(log10(c(1:10000)), coin.d$P.A, type="l", col="red", ylim=c(0,1), xaxt = "n",
+ xlab="Number of throws of the coin",
+ ylab="Relative Frequency of A",
+ main="Relative frequency of heads in coin throws")
> abline(h=0.5, col="blue", lty=2)
> axis(side=1, at=c(0:4), labels=10^c(0:4))
```



Finally, note how close of the "theoretical" probability (0.5) is to the relative frequency after 10,000 throws:

```
> coin.d$P.A[10000]
```

```
[1] 0.5049
```

Of course, your results (if you run the code) will be slightly different but, surly, of the same order of magnitude than

```
> abs(coin.d$P.A[10000]-0.5)
```

```
[1] 0.0049
```

Note that this fact, say, the convergence of the relative frequency to a given (and normally unknown) value between zero and one, as the number of realizations of the experiment increases, is a *fact of nature*, not a theorem or a mathematical construct. This is of course a simple result of the regularities that we expect to find in Nature.

This last definition, the probability as a limit of the relative frequency when the numbers of realization of the experiment increases, is the most useful. Normally we will not know these probabilities, and thus will need to estimate them, a theme that is the real core of statistics.

4. Elemental Set Theory

Set Theory happens to be a great tool to help in the understanding of Probability Theory.

We begin with a "*Universal Set*" which will contain all the elements of the discourse, say Ω. Later we will use this universal set to group all results of a *random experiment*. The Ω to use will depend on the problem at hand. For example,

- $\Omega = \{A_1, A_2, ..., A_n\}$, a finite set of alleles.
- $\Omega = \{1, 2, ...\}$, the infinite set of all natural numbers.
- $\Omega = \{x | x \text{ is a natural number}\}$, the infinite set of all natural numbers
- $\Omega = \{x | x \text{ is a real number}\}$, the infinite set of all real numbers
- $\Omega = \{0 < x < 1\}$, the infinite set of all real numbers between zero and one (an interval).

Now we are going to see a set of definitions and theorems. We will not give proofs of the theorems, but will use them as tools for Probability. You can convince yourself of the veracity of the theorems by using, for example, Venn diagrams.

4.1. **Definition. Subset.** If each element of a set A is also element of the set B, then we say that A is a subset of B, or that A is contained into B, and we write

$$A \subset B$$

The same thing can be written as

$$B \supset A$$

which is read B *contains* A or B is superset of A.

4.2. **Definition. Equality of sets.** We will say that two sets, A and B are equal if and only if $A \subset B$ and $B \subset A$. In that case we write $A = B$ or $B = A$. Note that this implies that both subsets have the same elements.

4.3. **Definition. Empty or null set.** If a set does not have any element we will call it the *null* set, denoting it by ϕ.

4.4. **Definition. Complement.** We define the complement of a set A in Ω as the set of all elements of Ω that does not belong to A. We denote such set as A^c or also as $\Omega - A$

4.5. **Definition. Union.** Lets A and B be two sets in Ω. Then the union of A and B, denoted by $A \cup B$, is defined as the elements that belong to A or B or both.

4.6. **Definition. Intersection.** Lets A and B be two sets in Ω. Then the intersections of A and B, denoted by $A \cap B$, is defined as the elements that belong to A and B.

4.7. **Definition. Difference.** Lets A and B be two sets in Ω. Then the difference between A and B, denoted by $A - B$, is defined as the elements that belong to A but not to B.

4.8. **Theorem. Conmutative Laws.** $A \cup B = B \cup A$ and $A \cap B = B \cap A$

4.9. **Theorem. Associative Laws.**

$$A \cup (B \cup C) = (A \cup B) \cup C)$$

and

$$A \cap (B \cap C) = (A \cap B) \cap C)$$

4.10. **Theorem. Distributive Laws.**

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

and

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

4.11. **Theorem. Complement of the complement.**

$$(A^c)^c = A$$

4.12. **Theorem. About Ω and ϕ with any set A.** $A \cap \Omega = A$; $A \cup \Omega = \Omega$; $A \cap \phi = \phi$; $A \cup \phi = A$;

4.13. **Theorem. About A and A^c.** $A \cap A^c = \phi$; $A \cup A^c = \Omega$; $A \cap A = A$; $A \cup A = A$;

4.14. **Theorem. Morgan's Laws.**

$$(A \cup B)^c = A^c \cap B^c$$

and

$$(A \cap B)^c = A^c \cup B^c$$

4.15. **Theorem. About $A - B$.**

$$A - B = A \cap B^c$$

4.16. **Theorem. A way to describe A as two non-overlapping sets.**

$$A = (A \cap B) \cup (A \cap B^c)$$

$$(A \cap B) \cap (A \cap B^c) = \phi$$

4.17. **Definition. Mutually exclusive sets.** Two subsets of Ω, A and B are *mutually exclusive* or *disjoint* if

$$A \cap B = \phi$$

A set A_1, A_2, ..., is said to be mutually exclusive if

$$A_i \cap A_j = \phi$$

for all pairs $\{i, j\}$.

4.17.1. *Examples of set theory with* ***R****.* In R we have various functions for *Set Operations*; these are: *union(x, y), intersect(x, y), setdiff(x, y) and setequal(x, y).* Ask for help for any of these functions; for example "? union" to understand the basics.

There is also a handy vector which contain all the letters, let use it as our universal set, Ω:

```
> letters # To see the set
 [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j" "k" "l" "m" "n" "o" "p" "q" "r" "s" "t"
[21] "u" "v" "w" "x" "y" "z"
> U <- letters # We define our universal set
```

Now, let's define some subsets of Ω

```
> V <- c("a", "e", "i", "o", "u") # The vowels
> O <- c("o", "c", "t", "a", "v", "i") # A non random set!
```

Now, let's play for a while with these sets, for example, let's find if order is important for sets functions

```
> sort(O) # Our O set sorted alphabetically
[1] "a" "c" "i" "o" "t" "v"
> setequal(O, sort(O)) # Are they equal?
[1] TRUE
```

How can we obtain the set of the consonants?. Note that $V^c = \Omega - V$, then

```
> setdiff(U, V) # Those will be the consonants
 [1] "b" "c" "d" "f" "g" "h" "j" "k" "l" "m" "n" "p" "q" "r" "s" "t" "v" "w" "x" "y"
[21] "z"
```

How R represents the null set ϕ?. Let's find out by asking R to give us a null set, for example $V - \Omega$ or, even surer, $V - V$!

```
> setdiff(V, U) # An empty (null) set
character(0)
> setdiff(V, V) # Other one
character(0)
> setequal(setdiff(V, U), setdiff(V, V)) # Are equal?
[1] TRUE
```

Now, let's test the union and intersection functions

```
> union(V, O)
[1] "a" "e" "i" "o" "u" "c" "t" "v"
> intersect(V, O)
[1] "a" "i" "o"
```

Now, let's create another set and try some of the theorems that we have seen in the theory:

```
> M <- c("m", "a", "r", "t", "i", "n", "e", "z")
> # Trying theorem 4.8 = Conmutative Laws
> union(O, M)
 [1] "o" "c" "t" "a" "v" "i" "m" "r" "n" "e" "z"
> union(M, O)
 [1] "m" "a" "r" "t" "i" "n" "e" "z" "o" "c" "v"
> setequal(union(O, M), union(M, O)) # Are the same set?
[1] TRUE
> intersect(O, M)
[1] "t" "a" "i"
> intersect(M, O)
[1] "a" "t" "i"
> setequal(intersect(O, M), intersect(M, O)) # Are the same set?
[1] TRUE
> # Trying theorem 4.9 = Associative Laws
> union(V, union(O, M))
 [1] "a" "e" "i" "o" "u" "c" "t" "v" "m" "r" "n" "z"
> union(union(O, M), V)
 [1] "o" "c" "t" "a" "v" "i" "m" "r" "n" "e" "z" "u"
> setequal(union(V, union(O, M)), union(union(O, M), V))
[1] TRUE
> intersect(V, intersect(O, M))
[1] "a" "i"
> intersect(intersect(O, M), V)
[1] "a" "i"
> setequal(intersect(V, intersect(O, M)), intersect(intersect(O, M), V))
[1] TRUE
> # Trying theorem 4.10 = Distributive Laws
> intersect(O, union(M, V))
[1] "o" "t" "a" "i"
> union(intersect(O, M), intersect(O, V))
[1] "t" "a" "i" "o"
> setequal(intersect(O, union(M, V)), union(intersect(O, M), intersect(O, V)))
[1] TRUE
```

4.17.2. *Homework.*

- 1. Try, using the sets defined here, Theorems 4.11 to 4.16
- 2. Find if the following expressions are theorems or not and demonstrate them using R:

$$((A^c)^c)^c = A^c$$
$$A^c - (B \cap A) = A$$
$$(A \cup B)^c = A^c \cup (A^c \cap B)$$

  HINT: To prove that a given expression is NOT a theorem it is sufficient to give a counterexample; in contrast to really PROVE that something is a theorem is more complicated. Just play for a while.
- 3. Program a function which gives TRUE if two sets are mutually exclusive sets.
- 4. (only for highly motivated students) Program a function where you input a word, for example "eureka" and which give as output the distinct letters of the word; say

```
> separa("eureka")
[1] "e" "u" "r" "k" "a"
> separa("parangaricutirimicuaro")
 [1] "p" "a" "r" "n" "g" "i" "c" "u" "t" "m" "o"
```

  HINT: You can use the functions **strsplit** and **unique**.

Have fun! Remember, no homework is compulsory; do it only if you are interested in learning. In any case at the end of the course we will have a test.