LANGEBIO - BIOSTATISTICS

OCTAVIO MARTÍNEZ DE LA VEGA

5. Random Experiments and Elemental Probability Theory

We have not seen any Statistics so far, and we will not yet in this section. Here we will review the fundaments of the Theory of Probability, which in turn is the base for Statistics. As in the previous sections, we will see Definitions, Theorems (without a formal proof) and then examples, illustrating the application.

We begin with an undefined concept, **random experiment**. Think about a random experiment as something well defined that can be repeated many times under the *same contidions* and which can give different *results*. Examples of **random experiments** are: Throwing a coin, throwing a dice, obtaining a descendent in an Arabidopsis cross, measuring the PH in a solution, counting the number of ants in a square meter of land, counting the number of fruits of a plant, measuring the length of a fruit, ...

Please, ask questions if you have any doubt; if not, I will assume that I am a great teacher and you are understanding perfectly everything.

5.1. **Definition. Sampling space.** The *sampling space* for a random experiment is defined as the set of all possible results of the experiment and will be denoted here by Ω.

5.1.1. *Example. A sampling space in human genetics.* The *sampling space* for a random experiment is defined as the set of all possible results of the experiment and will be denoted here by Ω.

- Let the random experiment be defined as "*To obtain a descendent from the cross* Aa *x* aa".
- The possible results of this experiment are the genotypes Aa and aa, then
- The sampling space is defined as $\Omega = \{(Aa), (aa)\}$.

5.2. **Definition. Event and Space of Events.** An event is a subset of the sampling space Ω.

For a finite Ω, the class of all events associated with a Sampling Space is defined as the *Space of Events*.

Formally, if we denote by ϵ the Space of Events, we will ask that this collection fulfills the following conditions:

- $\Omega \in \epsilon$; that is, the Sample Space is element of the Space of Events.

Date: April 2012.

- If $A \in \epsilon$ then $A^c \in \epsilon$; that is if a given event belongs to the Space of Events then it's complement is also part of it. In other words, the Space of Events is closed under complement.
- If $A \in \epsilon$ and $B \in \epsilon$ then $(A \cup B) \in \epsilon$; that is, if two events are part of the Space of Events, then their union is also part of the Space of Events. In other words, the Space of Events is closed under union.

5.2.1. *Example. Space of Events in human Genetics.* Consider the sampling space of the previous example (5.18.1), then

- The sampling space is defined as $\Omega = \{(Aa), (aa)\}$.
- The Space of Events of this experiment is given by all subsets of Ω, say
  $\epsilon = \{(\ ), (Aa), (aa), \{(Aa), (aa)\}\ \}$
  or in a more compact notation as
  $\epsilon = \{\phi, (Aa), (aa), \Omega\}$

5.3. **Theorem. About the impossible event.**

$$\phi \in \epsilon$$

that is, the impossible event is part of the Space of Events.

Proof (as an example). We know that $\Omega \in \epsilon$ (first condition above), and also that $\Omega^c \in \epsilon$ (second condition, above), but we know that $\Omega^c = \phi$, thus it follows that $\phi \in \epsilon$

5.4. **Theorem. The Space of Events is closed under intersection.** If $A \in \epsilon$ and $B \in \epsilon$ then $(A \cap B) \in \epsilon$; that is, if two events are part of the Space of Events then their intersection is also part of it.

5.5. **Definition. Probability Function.** Consider a random experiment with sampling space Ω and space of events ϵ. Then a *Probability Function*, denoted by $P[\bullet]$, is a function that fulfills the following conditions (axioms) for any event $\in \epsilon$:

- 1. $P[A] \geqq 0$ (That is, *there are not negative probabilities*).
- 2. $P[\Omega] = 1$ (Our sampling space is the *sure* event).
- 3. If A_1 and A_2 are mutually exclusive events, that is if $A_1 \cap A_2 = \phi$, then

$$P[A_1 \cup A_2] = P[A_1] + P[A_2]$$

That is for two mutually exclusive events, the probability of their union is the sum of their individual probabilities.

NOTE: The axioms of probability tell us only the kind of functions that can be considered as a valid probability function, but these axioms DO NOT tell us which function is the "correct" one!. In other words, to assign probabilities we need the definitions that we enunciated before, the *a priori* (equally likely) or *a posteriori* (limit of the relative frequency) definitions. In almost all "real" cases we will need to *estimate* the probabilities, something that is part of the theory of Statistics. Thus we must assign probabilities to every possible result of the experiment, that is to every sampling point $\omega \in \Omega$.

5.5.1. *Example. Assignation of Probabilities.* Consider the random experiment of obtaining one descendent from the genetic cross $AaxAa$. Then it is clear that

$$\Omega = \{AA,\ Aa,\ aa\}$$

Now consider the following probability functions:

- 1. $P[AA] = 1$
- 2. $P[AA] = 1/3$, $P[Aa] = 1/3$, $P[aa] = 1/3$
- 3. $P[AA] = 0.1$, $P[Aa] = 0.8$, $P[aa] = 0.1$
- 4. $P[AA] = 1/4$, $P[Aa] = 1/2$, $P[aa] = 1/4$

Note that ALL four assignations are *valid* probability functions, however only one is *correct* from the Genetics point of view.

5.6. **Theorem. Properties of $P[\bullet]$.** Consider a random experiment with sampling space Ω and space of events ϵ. In al cases we assume that the events (represented by capital letters) are elements of ϵ.

- i. $P[\phi] = 0$ (we can call ϕ the *impossible event*).
- ii. If $A_1, A_2, \ldots A_n$ are disjoint events then

$$P[A_1 \cup A_2 \cup \ldots A_n] = P[A_1] + P[A_2] + \cdots P[A_n]$$

  Or, in a more compact notation

$$P[\cup_1^n A_i] = \sum_{i=1}^{i=n} P[A_i]$$

  (The probability of the union of mutually exclusive events is the sum of the probabilities of each event)
- iii. $P[A^c] = 1 - P[A]$
- iv. $P[A] = P[A \cap B] + P[A \cap B^c]$
- v. $P[A \cup B] = P[A] + P[B] - P[A \cap B]$
- vi. If $B \subset A$ then $P[B] \leqq P[A]$.

5.6.1. *Example. Probabilities (continuation).* Consider again the previous example () where we have:

$$\Omega = \{AA,\ Aa,\ aa\}\ P[AA] = 1/4,\ P[Aa] = 1/2,\ P[aa] = 1/4$$

Let define the following events:

- B = *An heterocigote is obtained*
- C = *We obtain an individual with at least one A allele*
- D = *A homocigote is obtained*

Construct a Venn diagram with Ω containing the sampling points, AA, Aa and aa, as well as the events defined above.

Is it true that:

- B and C are disjoint?
- B is a subset of C ($B \subset C$)?
- $C \cup D = \Omega$?

- $C - B = C \cap D$?
- $B^c = C$?
- $P[B] = P[D]$?
- $B \cap C = D^c$?
- $B \cap C \cap D = \phi$?
- $P[(B^c \cap D) \cup B] = 1$?

5.6.2. *Example. A Finite Sample Space with Equally Likely Points.* Imagine that a monkey is sited in a computer which have a keyboard with only the 26 lowercase letters plus the space, " ". Further, imagine that the monkey is typing in such a way that each key has the same probability of being push.

What is the probability that the monkey produce the word "hell" given that it produced a four letter word?

Can we construct a function to mimic the behavior of the monkey?

Here our sampling space, Ω, is the set of the 27 symbols (considering the space as a symbol), and each one of the symbols has the same probability of being extracted, 1/27. The experiment is to "*obtain a four letter word*" (Note: In words we take into account the order of the symbols).

Let's construct our machine in R

```
> monkey <- function(n=4){paste(sample(c(letters, " "), n, replace=TRUE), collapse="")}
> monkey() # Trying our new brand monkey machine
[1] "twzq"
```

Now, let's have more fun, producing some text with different sizes of words from 1 to 5, say produce 100 words...

```
> monkey.speech <- c()
> for(i in 1:100){
+         k <- sample(1:5)
+         monkey.speech <- c(monkey.speech, monkey(k))
+ }
> monkey.speech <- paste(monkey.speech, collapse=" ")
> monkey.speech
[1] "rionw eoq jlyt  anp qkr i wxxt z nru gz pdyqr gi f biytq no u yt e t qo xe j s s ff
```

OK, stop playing!. The probability of the monkey producing the word "hell" given that each letter has the same probability is

```
> (1/27)^4 # A small number around 1 in 531441
[1] 1.881676e-06
> 1/531441
[1] 1.881676e-06
```

A further question: About how long do we need to wait in time if we sit until monkey() produces the word "hell", well...

```
> system.time(monkey()) # How long it takes to execute
```

```
   user  system elapsed
      0       0       0
> 531441 * 0.006 # (in my machine it takes 0.006 seconds)
[1] 3188.646
> 531441 * 0.006 / 60 # In minutes (around one hour)
[1] 53.1441
```

In some cases we have partial knowledge of what happen as a result of a random experiment. In this context arise the concept of conditional probability, which is defined below.

5.7. **Definition. Conditional Probability.** Let A and B be two events in the space of events, such that $P[B] > 0$. Then the *Conditional Probability* of A given that B has happened, denoted by $P[A|B]$ is defined as

$$P[A|B] = \frac{P[A \cap B]}{P[B]}$$

and it is not defined if $P[B] = 0$.

5.8. **Theorem. Properties of Conditional Probability.** Let B as above ($P[B] > 0$), then

- $P[A|B] \geqq 0$
- $P[\Omega|B] = 1$
- If A_1 and A_2 are mutually exclusive events, that is if $A_1 \cap A_2 = \phi$, then

$$P[A_1 \cup A_2|B] = P[A_1|B] + P[A_2|B]$$

- $P[\phi|B] = 0$
- If $A_1, A_2, \ldots A_n$ are disjoint events then

$$P[A_1 \cup A_2 \cup \ldots A_n|B] = P[A_1|B] + P[A_2|B] + \cdots + P[A_n|B]$$

  Or, in a more compact notation

$$P[\cup_1^n A_i|B] = \sum_{i=1}^{i=n} P[A_i|B]$$

- $P[A^c|B] = 1 - P[A|B]$
- $P[A|B] = P[A \cap C|B] + P[A \cap C^c|B]$
- $P[A \cup C|B] = P[A|B] + P[B|B] - P[A \cap C|B]$
- If $C \subset A$ then $P[C|B] \leqq P[A|B]$.

5.8.1. *Conditional Probability in Genetics.* Assume that there is a dominant molecular marker M, with recessive allele m which is close to a locus R, which recessive allele r gives resistance to a fungus. Furthermore, assume that the recombination probability between M and R is equal to 0.1

Assume the cross

$$\frac{mR}{Mr} \times \frac{mR}{Mr}$$

and define the experiment of obtaining a descendent from this cross and determine its molecular "phenotype" (M or m) and its phenotype, say Res = Resistant (genotype rr) or Sus = Susceptible (genotypes RR or Rr).

The individuals crossed are both of the same phenotype: (M, Sus) and also they are heterocygous for both loci, having the genes in "repulsion".

Note that the sampling space of this experiment can be written as

$$\Omega = \{(M, Res), (m, Res), (M, Sus), (m, Sus)\}$$

Calculate the probabilities of the following events:

- 1. A : "The individual is Resistant (Res)"
- 2. B : "The individual is M"
- 3. $A|B$
- 4. $A|B^c$
- 5. $A^c|B$
- 6. $A^c|B^c$

We can give a quick response to the first two items above, say $P[A] = 1/4$ and $P[B] = 3/4$, however items 3 and 4 need more careful consideration.

Lets solve the problem in R:

```
> # First, calculate the probabilities for the gametes
> # MR, Mr, mR and mr
> p.g <- c(0.1/2, (1-0.1)/2, (1-0.1)/2, 0.1/2) # Why?
> names(p.g) <- c("MR", "Mr", "mR", "mr")
> p.g # Show the probabilities
  MR   Mr   mR   mr
0.05 0.45 0.45 0.05
```

Now, using matrix multiplication we can obtain the probabilities of each one of the 16 possible combination of gametes, say:

```
> p.c <- p.g%*%t(p.g)
> class(p.c) # The "class" to which p.c belongs
[1] "matrix"
> attributes(p.c) # And its attributes
$dim
[1] 4 4
```

```
$dimnames
$dimnames[[1]]
NULL

$dimnames[[2]]
[1] "MR" "Mr" "mR" "mr"
> p.c # The matrix produced
         MR     Mr     mR     mr
[1,] 0.0025 0.0225 0.0225 0.0025
[2,] 0.0225 0.2025 0.2025 0.0225
[3,] 0.0225 0.2025 0.2025 0.0225
[4,] 0.0025 0.0225 0.0225 0.0025
> attributes(p.c)$dimnames[[1]] <- names(p.g)
> p.c # Nicer now!
       MR     Mr     mR     mr
MR 0.0025 0.0225 0.0225 0.0025
Mr 0.0225 0.2025 0.2025 0.0225
mR 0.0225 0.2025 0.2025 0.0225
mr 0.0025 0.0225 0.0225 0.0025
> sum(p.c) # Adding up all elements of the matrix
[1] 1
```

This last object, *p.c*, give us the matrix of probabilities of each one of the (16) gamete's combinations. Let's put those into a full table, giving also the phenotype of each combination.

Table of Genotypes, Phenotypes and Probabilities

Gamete	MR 0.05	Mr 0.45	mR 0.45	mr 0.05
MR 0.05	(M,Sus) 0.0025	(M,Sus) 0.0225	(M,Sus) 0.0225	(M,Sus) 0.0025
Mr 0.45	(M,Sus) 0.0225	(M,Res) 0.2025	(M,Sus) 0.2025	(M,Res) 0.0225
mR 0.45	(M,Sus) 0.0225	(M,Sus) 0.2025	(m,Sus) 0.2025	(m,Sus) 0.0225
mr 0.05	(M,Sus) 0.0025	(M,Res) 0.0225	(m,Sus) 0.0225	(m,Res) 0.0025

Now, adding the probabilities of the corresponding genotypes we can find the probabilities for each phenotype, say

```
> sum(p.c[1,], p.c[2, c(1,3)], p.c[3, c(1, 2)], p.c[4, 1]) # For (M,Sus)
[1] 0.5025
> sum(p.c[2, c(2,4)], p.c[4, 2]) # For (M,Res)
[1] 0.2475
> sum(p.c[3, c(3,4)], p.c[4, 3]) # For (m,Sus)
[1] 0.2475
> p.c[4,4] # For (m,Sus)
[1] 0.0025
```

Thus, we have the probabilities of each one of the elements of Ω, say $P[(M, Sus)] = 0.5025$, $P[(M, Res)] = 0.2475$, $P[(m, Sus)] = 0.2475$ and $P[(m, Res)] = 0.0025$

Note that the probabilities of the events A and B can be obtained by adding the probabilities of the elements of Ω, that are of course disjoint, because they are sample points, say:

$$P[A] = P[\text{"The individual is Resistant (Res)"}]$$
$$= P[(M, Res)] + P[(m, Res)] = 0.2475 + 0.0025 = 0.25$$
and
$$P[B] = P[\text{"The individual is M"}]$$
$$= P[(M, Res)] + P[(M, Sus)] = 0.2475 + 0.5025 = 0.75$$

(as we speculated before)

Now we can proceed to calculate the probabilities asked in the problem, say
$P[A|B] = P[$"The individual is Resistant (Res) given that it is M"]
and
$P[A|B^c] = P[$"The individual is Resistant (Res) given that it is m"]
We have that, by definition,

$$P[A|B] = \frac{P[A \cap B]}{P[B]}$$

then, in this case

$$P[A|B] = \frac{P[(M, Res)]}{P[(M)]} = \frac{0.2475}{0.75} = 0.33$$

and again by definition of conditional probability,

$$P[A|B^c] = \frac{P[A \cap B^c]}{P[B^c]}$$

then,

$$P[A|B^c] = \frac{P[(m, Res)]}{P[(m)]} = \frac{0.0025}{0.25} = 0.01$$

Items 5 and 6 ($P[A^c|B]$ and $P[A^c|B^c]$) are let as homework.

5.9. **Theorem. Total Probability.** Let

$$B_1, B_2, \ldots B_n$$

be a set of disjoint events that fulfill the conditions
1) $B_1 \cup B_2 \cup \cdots \cup B_n = \Omega$
and
2) $P[B_i] > 0$ for all i; $i = 1, 2, \ldots n$.
(Note: the collection $B_1, B_2, \ldots B_n$ is called a *partition* of the sample space Ω).
Then:

$$P[A] = P[A|B_1]P[B_1] + P[A|B_2]P[B_2] + \cdots + P[A|B_n]P[B_n]$$

or in a more compact notation,

$$P[A] = \sum_{i=1}^{n} P[A|B_i]P[B_i]$$

5.9.1. *Example.* In the previous example the *genotypes*: MM, Mm and mm form a partition of the sample space *(check that)*. Use the theorem of Total Probabilities to calculate $P[A]$.

5.10. **Bayes Theorem.** Let

$$B_1, B_2, \ldots B_n$$

be a partition of Ω and A a not null event ($A \neq \phi$), then

$$P[B_k|A] = \frac{P[A|B_k]P[B_k]}{\sum_{i=1}^{i=n} P[A|B_i]P[B_i]}$$

To see this, it is sufficient to note that

$$P[A|B_k]P[B_k] = P[A \cap B_k]$$

and, by the theorem of total probability

$$\sum_{i=1}^{i=n} P[A|B_i]P[B_i] = P[A]$$

and, using the definition of conditional probability

$$P[B_k|A] = \frac{P[B_k \cap A]}{P[A]}$$

5.10.1. *Example: Total Probability and Bayes Theorem.* In a given population it is known that there are 60% of catholics, 10% of protestants 5% of mormons and 25% of atheist. The following table shows the vote preferences of each group

Vote Preferences per Group

Party:	PRA	PRE	PRO	**None**
Catholics	75	15	5	5
Protestants	4	90	5	1
Mormons	2	78	10	10
Atheist	1	9	90	0

Can you predict which party will win the election?

Given that you know that a person voted in favor of the PRE party, which is the probability that he/she is Catholic?

5.11. **Independence of two events.** Two events, say A and B are said to be *independent* if and only if

$$P[A \cap B] = P[A]P[B]$$

5.12. **Consequences of the independence of two events.** If two non-null events, say A and B, are independent, then

$$P[A|B] = P[A]$$

and

$$P[B|A] = P[B]$$

To see that, just apply the definition of conditional probability and the independence of the two events.

5.13. **Homework.**

- 1. Write the *sampling space* (Ω) for some real experiment that you have done or are planning to do.
- 2. Assume that you have a finite sample space, Ω, which contains n elements (sampling points). Which is the size of the corresponding Space of Events (ϵ)?. Hint: Begin with a simple Ω containing 2 elements, follow with one with 3 elements and see if you can find the general formula.
- 3. Consider the following experiment: A coin is thrown until the first "head" appears. Write a description of the event space for this experiment. Is it finite? See if you can assign a reasonable probability function to this experiment.
- 4. Construct a function which output random DNA sequences of a given size, say n.
- 5. Improve your function to allow distinct probabilities for each base.
- 6. (a bit more complicated) Further improve your function to begin with a start codon and continue until it produces a stop codon.