**Introduction to biostatistics – Mat 2379B**

*September 3, 2014*

## Lecture 1

1.1. **A brief introduction.** Probability theory, statistics, and data mining software (namely, R) form three constituents of our course, and we will start our discussion with probability.

Probability theory is a mathematical framework for dealing with the phenomenon of uncertainty, or randomness. Randomness is inherently present in sciences - either because there are too many factors that affect the outcome of a given process, or even because randomness is instrinsically present in it. Probability theory was developed in the XVII century and is linked to the names of Huygens, Pascale, Fermat, and afterwards Laplace, Jacob Bernoulli, Gauss, and, in the XX century, Kolmogorov and others. The origins of probability were not quite so honorable: the aim was the analysis of games of chance. Hence, tossing a coin or a dice still remain among the most common examples illustrating the theory.

1.2. **Outcome of a random experiment.** We will be dealing with *random experiments* and their *outcomes*. Rather than trying to define these notions, I will just list a few examples.

1.2.1. *Tossing a coin.* The random experiment consists in tossing a coin (a favourite pastime of probability theorists). There are two possible outcomes of such an experiment: heads and tails. In a mathematical theory, they are usually marked by 1 and 0.

1.2.2. *Tossing a coin four times.* Now let our random experiment consist in tossing the same coin not once, but four times in a row. The outcome in this case is a sequence (string) of four values, heads / tails (or: 1 and 0). For instance, 0010, or 1011 are possible outcomes. Altogether, there are $2^4 = 16$ possible different outcomes of this experiment:

| | | | |
|---|---|---|---|
| 0000 | 0001 | 0010 | 0011 |
| 0100 | 0101 | 0110 | 0111 |
| 1000 | 1001 | 1010 | 1011 |
| 1100 | 1101 | 1110 | 1111 |

1.2.3. *A family with three children.* We are studying all the families with three children, and we are interested in their gender. The outcome of our random experiment in this case is a sequence of three children.

Let us mark a boy with B, a girl with G. For instance, if the oldest child is a girl, followed by two boys, we will write this as GBB. In this case, there are $2^3 = 8$ possible combinations, where we write the older children first:

$$GGG, GGB, GBG, GBB, BGG, BGB, BBG, BBB.$$

FIGURE 1.1. An outcome of the random experiment.

1.2.4. *Electrocardiography.* The random experiment in this case is a familiar test recording patient's heart's electric activity. The outcome is an electrocardiogram (ECG), that is, the graph of a certain function, namely, the function of voltage (depicted along the vertical axis) against time (the horizontal axis).

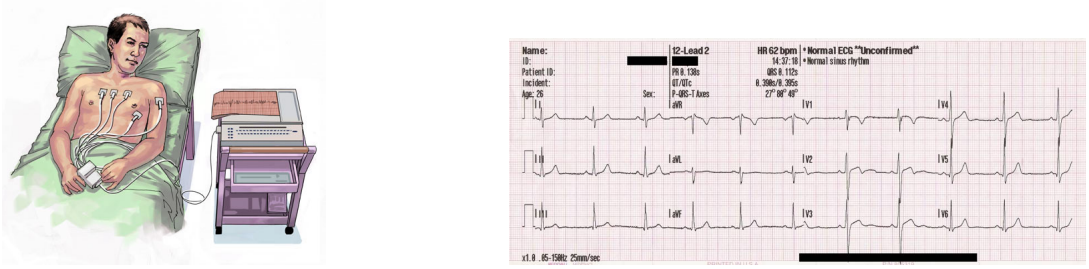


FIGURE 1.2. Our random experiment (left) and its outcome (right).

1.2.5. *DNA microarray analysis.* A DNA microarray (microchip) is used to determine the values of the genome sequence of a patient at predetermined sites (usually between 300,000 to a million out of over 3 billion). These sites (called SNPs, single nucleotide polymorphisms; pronounced "snips") are typically chosen because they admit the largest known variation in the general population and therefore are viewed as statistically significant. Thus, the outcome of our experiment is am extremely

long string of nucleotides $A, T, G, C$, having fixed length $d$ which could be anywhere between $300,000$ and $1,000,000$:

$$...ACAAGATGCCATTGTCCCCCGGCCTCCTGC...$$

1.3. **Sample space.** The sample space (also called *outcome space,* or *space of elementary events*), corresponding to a given random experiment, is simply the collection (set) of all possible outcomes of the experiment. We will denote the sample space $S$.

Here are some examples.

1.3.1. *Tossing a coin once.* Here there are only two possible outcomes, heads / tails, so the sample space is really simple, it consists of two elements:

$$S = \{heads, tails\},$$

or, in a more mathematical fashion,

$$S = \{0, 1\}.$$

The fact that $S$ has two elements is written:

$$|S| = 2.$$

1.3.2. *Tossing a coin four times.* Here, as we have seen, the number of possible outcomes is 16, so the sample space consists of sixteen elements:

$$|S| = 16.$$

They can still be all listed:

$$S = \{0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, 1001, 1010, 1011,$$

$$1100, 1101, 1110, 1111\}.$$

Here is a convenient mathematical notation. If we have a set $A$ and a natural number $d$, then $A^d$ denotes the collection of all strings of elements of $A$ of length $d$.

Thus, our sample space $S$ can be written as

$$S = \{0, 1\}^4,$$

the collection of all strings of zeros and ones having length 4.

1.3.3. *Family with three children.* Likewise, in our random experiment resulting in the birth of three children in a family, the sample space contains 8 different outcomes:

$$S = \{GGG, GGB, GBG, GBB, BGG, BGB, BBG, BBB\}.$$

We can write, using the above convention:

$$S = \{G, B\}^3.$$

1.3.4. *Electrocardiography.* The sample space corresponding to this experiment is infinite, because it consists of all functions that can serve as cardiograms for a potential patient! There are infinitely many such functions (notwithstanding the fact that the total number of patients currently alive is only finite).[1]

1.3.5. *DNA microarray analysis.* In this random experiment, the sample space is finite, but really huge. Let us fix the dimension, $d$ (e.g., $d = 870,000$). Then $S$ consists of all possible strings of letters $A, T, G, C$ of length $870,000$:

$$S = \{A, T, G, C\}^{870,000}.$$

The number of elements in this sample space is enormous, it is many orders of magnitude greater than the number of elementary particles in the known Universe.

1.4. **Events.** A *random event,* or simply an *event,* is some collection of outcomes. In other words, an event $A$ (corresponding to a given random experiment) is a *subset* of the sample space. This fact is denoted

$$A \subseteq S.$$

1.4.1. *Tossing the coin four times.* Here, for instance, we have the event $A$ consisting in the fact that on the second tossing, the coin fell heads up. This event consists of all outcomes with 1 in the second position:

$$A = \{0100, 0101, 0110, 0111, 1100, 1101, 1110, 1111\}.$$

It is a *subset* of the outcome space $S = \{0, 1\}^4$ in the sense that it contains some, but not necessarily all, possible outcomes.

1.4.2. *Some events in a family with three children.* Here we have the event "the family has at most one boy":

$$A = \{GGG, BGG, GBG, GGB\}.$$

1.4.3. *Impossible event.* In the same situation, consider the event "the family has two boys and two girls". Of course there is nothing impossible in such a combination — but it becomes impossible if we restrict ourserves to the families with three children. This even is called *impossible,* and mathematically, it corresponds to the *empty set* $\emptyset$, that is, a set containing no elements!

$$A = \emptyset.$$

[1]Moreover, this infinity is of a larger order than the infinity of all natural numbers... it is, as we say, *uncountable.*

1.4.4. *Sure, or certain, event.* A certain event corresponds to the entire sample space, that is, it imposes no restrictions on the outcome at all:

$$A = S.$$

For instance, in the case of a family with three children, the event "there is at least one child" is a sure event.

1.4.5. *Tossing the coin once.* Let us list *all* possible events corresponding to this experiment. There are just four different subsets of the set $\{0, 1\}$:

- The impossible event, $\emptyset$,
- The sure event, $A = S = \{0, 1\}$,
- The event "the coin went heads up", $A = \{1\}$,
- The event "the coin went tails up", $A = \{0\}$.

As we see, $4 = 2^2$. More generally, if the set has $k$ elements, then the number of *all subsets* of this set is $2^k$.

For instance, the collection of all events corresponding to the random experiment of tossing a coin four times is quite large: there are

$$2^{16} = 65536.$$

of them. Thus, even relatively simple experiments can leads to a great many events!

## 1.5. **Correspondence between set theory and logic.**

1.5.1. *Inclusion.* Let us consider the example of a family with three children. Denote $A$ the event "there are exactly two girls",

$$A = \{GGB, GBG, BGG\}.$$

ans $B$ the event "there is at most one boy,"

$$B = \{GGG, BGG, GBG, GGB\}.$$

Then $A$ is a subset of $B$, that is, every outcome contained in $A$ is contained in $B$ as well:

$$A \subseteq B.$$

One can say also that if a family has exactly two girls, then it has at least one boy. Thus, the inclusion of events

$$A \subseteq B$$

corresponds to the logical implication "if $A$, then $B$", or "$A$ implies $B$". The logicians write it as

$$A \Rightarrow B.$$

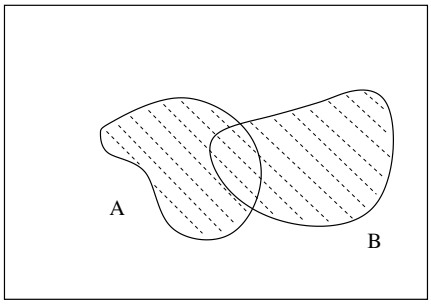


FIGURE 1.3. The union, $A \cup B$, of the events $A$ and $B$.

1.5.2. *Union.* The union of two events, $A$ and $B$, is the event which contains all outcomes from $A$ and all outcomes from $B$. From the logical viewpoint, this corresponds to the event "$A$ or $B$". The union is denoted $A \cup B$. See the figure 1.3.

Consider two events: $A$ "there are exactly two girls", and $B$, "there are exactly two boys". Their union, $A \cup B$, is the event "either the family has two girls or two boys", and can be otherwise expressed as "the family has both a boy and a girl".

1.6. **Recommended reading.** None yet: it seems no part of the textbook elaborates in detail the notions in our Lecture 1.