CHAPTER 16

NOTIONS OF STATISTICS FOR THE CLINICIAN

Pierre Duhaut, Claire Andrejak

Clinicians are often resistant to statistics. However, they are essential to be able to observe all the events not necessarily visible 'to the naked eye', and to be able to differentiate the frequency of the events in which we are interested, compared to a random occurrence.

Statistics are useless to demonstrate the action of antibiotics on meningococcal meningitis, or on tuberculosis. They will be necessary to measure the incidence of cerebrovascular accidents, serious pathologies if any, under antihypertensive treatment compared to untreated patients. The difference between the two situations is that in the first one is interested in a quick and individual action, almost immediately visible. In the second, we are primarily interested in a trend in a group, which will result in a long-term improvement in the prognosis of only part of the population treated: the field of statistics is that of the evaluation of trends perceptible only at the level of samples or populations, but which can also benefit, on a clinical level, to some of the patients treated or observed if not to all.

This chapter reviews the most common clinical questions and their translation into language and statistical tests.

Biostatistics and methodology are inseparable: the data collected must be analyzed… and must therefore be collected in such a way as to be analyzable. The construction of a study must provide for the type of analysis to be done according to the question asked, and the way of collecting the data must obey two sometimes contradictory imperatives, which will have to be reconciled:

- Describe reality as closely as possible,
- Make these data compatible with an existing statistical test allowing the most exact comparison between the groups studied.

I. QUESTIONS ASKED, VARIABLES TO COLLECT, ANALYSIS TO DO

Let's start with a banal clinical question and an effective database to make the point more intelligible:

Thrombocytosis is frequently encountered in clinical medicine. An old aphorism asserts that a thrombocytosis greater than 1.10^6/mm^3 corresponds 'in all cases' to essential thrombocythemia (ET) (myeloproliferative syndrome). The etiological diagnosis can however be difficult to make, including by molecular biology (the JAK2 mutation is only present in approximately 60% of cases), and the marrow biopsy may not show suggestive myelofibrosis. A 'gold standard' could be the diagnosis carried out after evolution of the patient and elimination - if possible - of all the reactive causes.

All thrombocytosis greater than 600,000/mm^3 in a hospital were combined in a series. This database must first be described, and the predictive factors of diagnoses can then be analyzed.

II DICHOTOMOUS VARIABLES (PROPORTIONS)

One thousand and forty-seven patients were included: it is necessary to know the distribution of the sexes, the number of patients with reactive (RT) or essential thrombocytosis (ET) (by defining each class as exclusive of the other). These variables are said to be **dichotomous** because they can only take two values (masculine-feminine, essential-reactive, true-false, yes-no). They are also **categorical**, with ***no hierarchy*** between the two values. They make it possible to calculate the proportion of women and men affected by one or the other of the major causes of thrombocytosis.

Application :

Our series includes 509 women and 538 men. The diagnosis of RT was made in 357 women (70.1% of 509) and 461 men (85.7% of 538). Is this rate *significantly different* between men and women? In other words, is 85.7% - 70.1% equal to (or close to) 0?

We just posed the null hypothesis: we will conclude, if the difference between the two rates is not significantly different from 0, that the two rates are close to each other, similar, and that there is no there is no difference between male and female patients. On the contrary, if the difference is significantly different from 0, the two proportions, and therefore the two groups, will be estimated to be different.

The data can be expressed in a 2x2 table:

Table 1: *number of patients actually observed in each category*

	Women	Men	Total
ET	152 (29.9% of 509)	77 (14.3% of 538)	229 (21.9% of 1047)
RT	357 (70.1% of 509)	461 (85.7% of 538)	818 (78.1% of 1047)
Total	509 (48.6% of 1047)	538 (51.4% of 1047)	1047

1. Chi-square test

The chi-square test allows the comparison of proportions. Like any statistical test, its principle is based on the null hypothesis: if there is no difference between men and women, then men and women make up a single group in which 229 patients (21.9%) had essential thrombocythemia, and 818 (78.1%) had reactive thrombocytosis. The chi-square test will measure, for each cell, the number of patients of difference between the patients *observed* ('O' as given by the study carried out) and expected ('E') if the proportions in the group as a whole, and in each subgroup, were equal. The table then becomes:

Table 2: *number of patients expected in each cell)*

	Women	Men	Total
ET	509 x 21.9% (expected proportion of ET in 1047 patients) = 111.5 (expected number = E)	538 x 21.9% (expected proportion of ET in 1047 patients) = 117.5 (Expected number = E)	229
RT	509 x 78.1% (expected proportion of RT in 1047 patients) = 397.5 (expected number = E)	538 x 78.1% (expected proportion of RT in 1047 patients) = 420.5 (expected number = E)	818
Total	509	538	1047

This new table has several notable features compared to the previous one:

- Totals have not changed.
- When the number of patients expected in cell 1 is fixed, the numbers expected in cells 2, 3 and 4 are determined without any freedom: the sum of the rows, and the sum of the columns, must remain constant (because fixed by the number of patients actually observed in each row and each column!).
- Therefore, the determination of the number of patients expected could only be done with a single degree of freedom, that of cell 1.

The sum of the deviations between observed and expected could be written: Σ (O - E).

It will be noted that this sum is equal to 0: indeed, the patients withdrawn from a cell have been added to the neighboring cell: the deviation of the first cell is the exact opposite of the second, that of the third, the opposite of the fourth.

To counter this drawback, each deviation is squared. The sum therefore becomes: $\Sigma (O - E)^2$.

However, an absolute difference cannot on its own account for a difference: the difference between 10000 and 10002 is equal to that between 2 and 4. The increase is 1/5000 in the first case, and 50% in the second... the difference must therefore be related to a denominator. Calculating the chi-2 brings it back to the number of expected events, and the chi-2 formula becomes:

$$\text{chi-2} = \frac{\Sigma (O - E)^2}{E}$$

In our example,

$$\text{chi-2} = \frac{\Sigma (O - A)^2}{A} = \frac{(152-111.5)^2}{111.5} + \frac{(77-117.5)^2}{117.5} + \frac{(357-397.5)^2}{397.5} + \frac{(461-420.5)^2}{420.5} = 37$$

It is then possible to consult the chi-square distribution tables for a degree of freedom and estimate what the probability is that 37 is similar to 0.

This probability is well below 1/1000 (in fact, less than 10^{-7}!)

The probability threshold for admitting the null hypothesis is usually set, by convention, at 5%. The null hypothesis is accepted above the threshold, rejected below. The null hypothesis in our example having less than 1 in 10 million chance of being true, ($p < 10-7$, < 0.05) is rejected. The two groups are not similar, and we deduce that they are probably different. There is a small contortion of logic here, because the whole basis of the calculation rests on the axiom of the equality of the two groups, and it is only on the basis of this axiom that the method of calculation is appropriate: however, the basis of validity of the calculation is not respected… We have simply shown that the two groups probably were not similar, and deduce from this that they are probably different.

Statistics do not establish truth: they seek to circumscribe an uncertainty, and this must always be kept in mind when interpreting the results. However, when the p-value is this low, one can very reasonably think that the two groups are different. When the p-value is close to 0.05 and changing a few patients from group makes it go above or below 0.05, the discussion remains open!

Warning!

- The simple chi-square calculation described above is valid only if the number of expected events is greater than 5 in all cells of the table. Otherwise, an exact calculation using geometric distributions must be performed: that is the ***exact Fisher test***, which can always be used in doubtful cases to give the calculation maximum rigor.

- Several authors wanted to make the calculation more severe (therefore, to bring the sum of the chi-2 closer to 0), by introducing correction factors into the formula. This is the case with Yates' chi-2, which subtracts ½ from each (A-O) before squaring it.

- The test of Mantel-Haenszel is the chi-square variant for stratified data: we would have applied it in our example if the diagnoses had been given by sex and by age group (20-40 years, 40-60, 60- 80, > 80). The chi-squares are then calculated for each stratum separately, then added up over all the strata.

- It is easy to understand that the value of the chi-2, and therefore the significance of the test, depends on the size of the population analysed: redo the calculation by dividing the numbers of each cell by 10 to reach a total sample size of 105 (frequent in clinical studies) (conserved proportions), and check the value of p!

2. Generalization of the chi-2 test to a table with n rows and m columns:

The sum of the deviations between O and E can always be calculated. Simply, the number of degrees of freedom changes, and becomes equal to (n-1)(m-1): the number E is always fixed without freedom for the last cell of each line, and the last cell of each column. The probability of equality between the sum of the chi-2 and 0 must then be read on the distribution line of the chi-2 at the corresponding number of degrees of freedom (Statistical softwares provide the exactly calculated value of p).

A chi-square significantly different from 0 in a n*m table will not indicate where the difference lies: it can be distributed evenly between the different cells, or, much more frequently, between a few, or even two, table cells. The risk of reaching a number of expected events less than 5 in a multi-cell table increases with the number of cells, which makes the validity of the calculation uncertain: it is better to plan another analysis plan.

III. QUANTITATIVE VARIABLES

1. Mean, variance, standard deviation, median, percentiles, mode.

We now want to describe the ages of the patients in each group, and compare them. Age is a continuous quantitative variable. Its description can be made in the form of mean (sum of all the values divided by the number of subjects), and of variance (sum of all the squares of the deviations between the mean of the sample and the age of each subject, related to the number of subjects: as for the chi-2 test, the differences are squared so as not to cancel each other).

The variance can therefore be written: $V = \Sigma \frac{(\mu - a)^2}{n}$, where

- μ denotes the average age for the entire group,
- a the age of each subject in the group, and
- n the total number of people in the group.

This formula expresses the **ideal variance** of a large population. We usually work in medicine on much smaller patient samples, even in multicenter studies.

A sample can only provide an *estimate, an approximate value*, of the mean and the variance of the total population. For safety, it will be better on a sample to define this variance in a slightly broader way (it will thus be more likely to cover the variance of the total population). To do this, we correct the denominator by replacing n by n-1: the numerical value of the variance increases slightly, and *its formula for a sample* becomes:

$$V = \Sigma \frac{(\mu - a)^2}{n-1}$$

This correction is all the more important as n, the sample size, is small, and all the more insignificant as n is large: the variance of a variable on a sample of 1000 people is more likely to be close to that of the total population, than the variance of the same variable on a sample of 15 people.

Similarly, we want to approximate the mean of the overall population from that of the sample, knowing however that if the sample had been selected differently, its mean would undoubtedly be a little different: it *may be possible* that the mean of glycated hemoglobin of a group of 40 type-2 diabetic patients is *exactly the same* as that of another group of 40 patients. However, it is *more likely* that it is not too far off. We define the notion of 95% confidence interval to express the fact that the mean would be within this interval in 95% of the 100 potential samples of equal size that could be drawn at random from the overall population.

In other words, the average of the glycated hemoglobin level in the general population of type-2 diabetic patients - which we are trying to approximate - undoubtedly has a 95% chance of being included in the interval thus defined from the sample. We can define, in the same way, a confidence interval at 90 or 99 or 99.9%... depending on the precision we want to obtain. The 95% confidence interval is the one most often used in medicine.

The variance is very often expressed in the form of its square root, called ***the standard deviation (SD):***

$$SD = \sqrt{\text{variance}}$$

The interest of the t-deviation is that it allows the distribution of the variable to be assessed fairly quickly: for a sample size greater than 30, 95% of the sample will be between the mean ± 1.96*SD. The value corresponding to 1.96 increases when the sample size decreases: again, this increase reflects the fact that the precision decreases with the sample size, and therefore that the probable dispersion of the values increases.

Warning !

- This way of describing a quantitative variable is only valid when the distribution of values follows a Gaussian curve (normal distribution).

- The average only makes sense if:
  - the values are distributed symmetrically around it, and if
  - it corresponds to the most frequently encountered value (in our example, the largest age group). In fact, establishing an average age of 40 for a series comprising 20 10-year-old children and 20 70-year-old adults would not make it possible to correctly grasp the reality of the patient group: there is no adult of 40 years in this group composed of two very different sub-groups, and calculating an arithmetic mean at 40 years would lead to a false description of reality.

- For the mean and its standard deviation to correctly describe the considered group, the distribution of the examined variable must therefore be symmetric and unimodal, *in other words, the distribution of values must not reflect the existence of two different groups of patients, a biological rate, or any numerical value*. Hodgkin's disease has two peaks of incidence, around 20-25 years old, then around 60-65 years old. To say that the average age of patients is 40 and to treat in the same way 75-year-old patients as 20-year-olds on the pretext of an average tolerance equal to that of 40-year-olds would be medical nonsense.

- It is always useful to draw a graph of the distribution of the data which enables its form to be assessed (symmetrical, unimodal). Most statistical softwares also make it possible to perform a test of normality of the distribution, which can guide subsequent statistical analysis.

When the distribution of the quantitative variable is *not normal (Gaussian)*, other modes of data description and other modes of analysis must be preferred. Quite often in medicine, quantitative variables have a Gaussian distribution *(statistically normal = Gaussian, to be differentiated from biologically normal = within biological standards)* in healthy subjects (example: hemoglobin, leukocytes, platelets ...). This statistical normality very often disappears in the sick subject: leukocytes can vary from 10,000/mm3 to more than 100,000 in chronic myeloid leukemia or acute leukemia, platelets from 400,000 to more than 1,500,000 in essential or reactional thrombocythemia, and one cannot generally extrapolate the normal distribution of the variable in the healthy subject to the asymmetrical distribution, sometimes logarithmic, sometimes difficult to describe, of the variable in sick subjects.

If the distribution is not statistically normal on the visual assessment or on the normality test, using the *median* and the *percentiles* will give a more precise idea of the population considered: the median defines the threshold in absolute value, below which lies 50% of the sample and above which is the remaining 50%. A median leukocytosis of 20,000/mm3 means that 50% of patients have a leukocyte count below 20,000, and 50% of them, higher. The 5th, 10th, or 75th percentiles correspond to the leukocyte level *below which* 5%, 10%, or 75% of patients are found.

Finally, the *mode* is the third description of a quantitative variable: it corresponds to the value most often encountered in the sample. It is relatively little used.

In a statistically normal distribution, mean, median and mode are confounded. In a non-normal distribution, they are usually distinct. A distribution with multiple peaks of equal height may include multiple modes, but will only include one median...and one mean, not representative of the whole sample.

Application :

In our example, the study of the age distribution gives the following results

	Mean	Standard Deviation	Median	Extremes
Women	62.8	19.8	67	18.7-102
Men	55.5	17.5	55.5	18.5-96.5

At first glance, the values given may be compatible with a normal distribution: subtracting or adding two standard deviations from the mean does not lead to aberrant ages (for example: negative), but to ages relatively close to the extremes.

At second glance, if median and mean are similar in men, they differ by more than 4 years in women, which makes one suspect, given the sample size, a disparity, *or a non-normal distribution*.

The question is whether this difference is significant or not.

2. Comparison of means, or analysis of variance:

Are the men and women in our study the same age?

The principle remains the same as for the chi-square test: the null hypothesis, simply, becomes:

The two averages are similar, or μ_{women} - $\mu_{\text{men}} = 0$ (μ_{women} representing the average age of women, and μ_{men} the average age of men).

This simple subtraction, however, is not enough to ensure a correct comparison: one would be ready to recognize that the two means are different if the standard deviation were very small:

For instance :

- 62.8 ± 0.2 would lead to a sample comprising 95% of patients between 62.4 and 63.2 years, and 55.5 ± 0.2 to a sample comprising 95% of patients between 55.1 and 55, 9 years (mean ± 1.96 SD). The age distributions of these two samples do not overlap, and it can therefore be assumed that, although close on average, the two samples are different.

• In our study, the standard deviation is much larger, and leads to an age distribution for 95% of patients between 24 and 101.6 years for women, and 21.2 and 99.8 years for men : the overlap of these two distributions is considerable, and although the means are the same as for the preceding example, one would not spontaneously be inclined to recognize these two distributions as different.

It is therefore essential to take the variance into account: the test for comparing means is in fact *an analysis of variance*. The analysis of variance takes into account, on the one hand, the difference between the means, and on the other hand, the combined variance of the two distributions compared *(= pooled variance*).

This combined variance of the two distributions compared makes it possible to estimate the variance of the distribution of the mean difference: the pooled variance reflects the variance of all the samples compared, and the variance of the difference of the means, that of the mean difference between the two compared populations.

The variance of the difference in means is then used to calculate a confidence interval around the difference in means. If this interval *contains 0*, then the difference is considered similar to 0, and the two means *similar*. If the confidence interval *does not contain 0*, the difference of the means is considered to be far from 0, and the means to be *different*. Depending on the requirement, the confidence interval can be calculated at 95% (corresponding to a p=0.05), 99% (corresponding to a p=0.01), 99.9%... *Significance of the difference in means can be given for any value of p.*

The p value gives the probability for the difference of the means to be equal to 0. We admit, as above, that the means can be considered as different if this probability is less than 5%, if $p < 0.05$.

In summary, a graphical representation of our groups could be:

$\mu1 - 1.96\sigma_1$ $\mu1$ $\mu1 + 1.96\ \sigma_1$

Group 1 |____________|____________|

$\mu2 - 1.96\sigma_2$ $\mu2$ $\mu2 + 1.96\sigma_2$

Group 2 |____________|____________|

$\mu1 - \mu2 – 1.96\sigma_p$ $\mu1 - \mu2$ $\mu1 - \mu2 + 1.96\sigma_p$

|____________________|____________________|

where σ_1 is the standard deviation of the mean in group 1,
σ_2 is the standard deviation of the mean in group 2,
σ_p represents the standard error of the difference of the means, calculated from the pooled variance of the two initial samples, group 1 and group 2.

If the analysis of variance relates to a complete population (extremely rare in medicine), the test used is the ***Z test***, which takes into account the variance (with n in the denominator). If the analysis involves patient samples (usual case), the test to be used is the ***t test***, which takes into account the corrected variance formula (with n-1 in the denominator). We can then refer to the t-test tables, to the line corresponding to the number of total patients -2 (number of degrees of freedom of the t-test), and find the probability that the value of t is different from 0.

Of course, n-1 will be very close to n when n, the sample size, is large. In these situations, the Z test (taking n in the denominator), and the t test (taking n-1 in the denominator) will give similar results.

In our example, the t-test value is 6.7. We would not be surprised if its probability of being close to 0 is very low, given the sample size: it is, in fact, 0.0001. In other words, the probability that the age difference between the group of men and the group of women is equal to 0 is 1/10,000, *in other words, the difference in the means is significantly different from 0 with $p = 0.0001$.* The null hypothesis is rejected, and *we accept that there is a significant age difference between men and women: $\mu_{women} - \mu_{men}$ is significantly different from 0.*

3. Non-parametric test: Wilcoxon rank-sum test:

By looking roughly at the mean, the standard deviation, and the extremes of the ages of women and men, we saw that these values were *compatible* with a normal distribution. However, this summary examination is not sufficient. Let's take a closer look at the data:

Fig 1: Age pyramid in years: group of women

```
 Age (years)        Histogram          Number of patients     Boxplot
                                           #
102.5+*                                    1               |
     .***                                  6               |
     .*******                             14               |
     .******************                  35               |
     .*****************************       57               |
     .*******************************     62            +-----+
     .****************************        56            |     |
     .**********************              44            *-----*
 60  .*************                       26            |  +  |
     .****************                    32            |     |
     .********************                40            |     |
     .***************                     30            +-----+
  40 .**************                      28               |
     .*******                             14               |
     .************                        23               |
     .*********                          18               |
     .*********                          18               |
 17.5+***                                  5               |
      ----+----+----+----+----+----+-
      *  each star can represent 1 to 2 patients
```

The shape of the age pyramid does not seem very Gaussian: there are three peaks, the first corresponding to 23 patients below the 40-year mark, the second to 40 patients below the 60-year mark, the third to 62 patients above. It is possible that there are 3 different groups of patients in this population, and it may be interesting to examine these three groups separately in an exploratory study. The so-called 'boxplot' diagram confirms this impression: the mean

(+) is different from the median (crossbar in the middle of the box), and the median is not located equidistant from the 25th and 75th percentile (lower crossbars and top of the 'box'.

The normality test performed by our statistical program confirms this impression: the null hypothesis (that of normality) is rejected with $p \leq 0.0001$ (the probability that this distribution is normal is less than 1 chance in 10,000).

The look of the age pyramid for men is a bit different than for women, but there are also different peaks. The null hypothesis of normality is also rejected with $p = 0.0001$. (Fig 2)

Fig. 2: Age pyramid for men

```
         Age (years)        Histogram              Number of patients        Boxplot
                                                           #
    97.5+*                                                 1                    |
        .**                                                3                    |
        .*****                                            10                    |
        .***********                                      22                    |
        .***********************                          45                    |
        .**************************                       51                    |
        .**********************                           43                 +-----+
        .**********************                           44                 |     |
    57.5+*****************************                    58                 *--+--*
        .******************************                   59                 |     |
        .***************************                      54                 |     |
        .********************                             40                 +-----+
        .******************                               35                    |
        .***********                                      22                    |
        .*********                                        17                    |
        .*************                                    25                    |
    17.5+*****                                             9                    |
         ----+----+----+----+----+----+
         * each star can represent 1 to 2 patients
```

The analysis of variance is based on the assumption of the normality of the distributions and the equality of the variances between the groups. There are statistical tests that do not require the normality of the distributions or the equality of the variances, and make it possible to compare the groups thus inhomogeneous. This presupposes considering the variables differently and taking into account, instead of the numerical, absolute value of the quantitative variable, *the rank it occupies in the distribution*.

In our example, the patients would thus be classified in each group from the youngest to the oldest, and the test consists in evaluating the general tendency of the ages in one group compared to the other: the question 'is the average age taking into account the variance different in the two groups?' becomes 'is one group overall younger, or older, than the other?'

Application :

The age of the women is then compared to the age of the men by creating 3 categories of pairs each time comprising a single woman and a single man:

1- The pairs for which the women are younger than the men
2- Pairs for which women and men are of equal age
3- The pairs for which the women are older than the men

In our example, the two 18-year women are younger than the 537 out of 538 over-18-year men: there are therefore 2 x 537 pairs (= 1074 pairs) for which a group-1 patient in is younger than a group-2 patient.

The 3 19-year women are younger than the (538 – 4) men aged over 19. There are therefore 3 x 534 (= 1602) additional pairs for which a group-1 patient is younger than a group-2 patient.

We thus continue to count all the pairs for which a group-1 patient is younger than a group-2 patient, and we add them (total number: x).

The two 18-year-old patients are the same age as the 18-year-old patient: there are therefore two pairs of equal age. We thus count all the pairs of equal age (total number: y)

The woman-patient aged 102 is older than all 538 men, as are the 101- 100-, and 99-year women-patient: this gives 4 x 538 pairs for which the women-patients are older than male-patients, and all the 'older' pairs are counted for the whole series (total number: z).

The number of equal pairs is then divided equally into the "younger pairs" group and the "older pairs" group, which then include x + y/2 pairs for the first, and z + y/2 pairs for the second. Our two groups of patients are thus transformed into two groups of pairs, 'older' and 'younger', which we will compare.

The null hypothesis becomes: the number of pairs is equal in the 'Younger' and 'Older' groups, or (x + y/2) - (z + y/2) = 0, and the comparison amounts to a comparison of proportions. Is the number of ‘younger’ pairs relative to the total number of pairs similar, or different, from the number of “older” pairs relative to the total number of pairs? Comparison of proportions refers to the type of analysis performed by the Chi-square test described above.

This so-called Wilcoxon rank-sum comparison test therefore ignores the distance between ages, and does not give additional weight to extreme ages, unlike the analysis of variance. It is not dependent on whether the distribution is normal or not. It allows the analysis of small samples: an calculated mean on 10 patients is likely not to be very precise, the variance may be important, and the comparison with another group of 10 patients is not very powerful. The rank-sum comparison test will relate to 10 x 10 = 100 pairs, rather than 20 individuals: in the case of small samples, the Wilcoxon rank test will probably be more efficient, but also more rigorous than a comparison of means because it is rare for a distribution to be normal under these conditions, and the analysis of variance could lead to a false estimate of p.

The U-Mann-Whitney test is based on the same principle and arrives, by a different calculation technique, at the same results as the Wilcoxon test. They are interchangeable and statistical software provides one or the other procedure indiscriminately.

Warning:

- Like any test, the Wilcoxon test has limitations. Each group must include at least 10 patients, otherwise the test loses its precision and its value... but wanting to compare too small groups makes us leave the field of statistical analysis!

• In practice, the analysis of variance will give the same results as the Wilcoxon test in terms of p-value when the size of each group exceeds 30, even if the distributions are not Gaussian. It would have been possible in the example above, but only by observing the distribution of the variables does one realize that there are perhaps three age groups of patients.

4. Ordinal variables:

Semi-quantitative variables of a particular type are often used in medicine: colon cancer is classified as Duke stage A, B, or C depending on the degree of invasion of the mucosa, and the severity increases from stage A to B then to C, but C is not three times as serious as A, or twice as serious as B. The same applies to stages I to IV of dyspnea, or to arteriopathy of the lower limbs. These stages cannot add up, multiply, or divide: they simply reflect a hierarchy in the severity of the disease rather than a strictly measurable quantity.

These semi-quantitative variables are called *ordinal*.

One may wish to compare, however, data of this type while preserving their hierarchical, or ordered character: does the administration of a diuretic in this group of patients with heart failure improve the stage of dyspnoea, otherwise said, does it make it pass from a higher order (III or IV for example), to a lower order (II)?

The Wilcoxon test is used for comparing the number of patients with stage I, II, III, or IV dyspnea in a treated and untreated group: each of the stages can be considered as a rank, and the Wilcoxon (or the U-Man-Whitney test) can help to identify the trend towards the less - or more - important stage of dyspnea.

5. Generalization of the analysis of variance to several groups:

It may be interesting to compare the means and variances of a quantitative variable between several groups. In our example, we may want to compare the importance of weight loss between the groups 'psychogenic origin', 'cancer origin', 'nutritional or endocrine origin', and 'other origin'.

There are several ways to look at the problem:

The first is to do a single analysis, and examine whether there is, by means of a single test, a difference *somewhere* within the overall sample of the 4 subgroups. The basic principle is the same as for the analysis of variance between two groups, and the test used is the *F-test (Generalization of the t-test to several groups)*. If the F-test does not show any significant difference within the x samples, we can conclude that there is probably no difference between the group with the weakest, and the group with the strongest, average. Consequently, there is probably no difference between the groups whose means are between the smallest and the largest, and our analysis can stop there.

If, on the other hand, the F-test shows a significant difference, there is *undoubtedly* a difference between the group with the smallest, and the group with the strongest, average... provided that these two groups have a sufficient sample size to arrive at a significant difference. It is also possible that the difference, in fact, lies between two intermediate groups with a larger sample size, or even between several groups, or even in a diffuse way between each group of patients. Like a multi-cell chi-square test, the F-test can detect a difference *somewhere, but cannot locate it*. However, it may be important to know whether, in the face of significant weight loss, it is better to first focus on one etiology or another...

To answer this question, we can consider comparing the groups two by two using a t test. Thus, to compare the 'psychogenic origin' group to the 'cancer origin' group, then to the 'endocrine origin' group, then to the 'other somatic origin' group... and to continue with the 'cancer'-'endocrine' comparisons, 'cancer'-'other', then 'endocrine'-'other' and to exhaust all possible logical combinations. For four groups, there are thus six logical possibilities. For 5 groups, ten. For 6, 15. The systematic comparison of all the groups two by two could appear more rigorous, or more conveying interesting information, than the global analysis of the F-test.

The problem, however, is that it multiplies the chances of showing a significant difference by chance: we accept a difference as significant if its probability of being due to chance is inferior to 5%. Therefore, if 100 random comparisons are made, it is very likely that at least 5 of them will turn out to be significant... by chance, since this is the very limitation of the statistical test. Out of 20 comparisons, at least one may be statistically significant by chance, and this significance would be devoid of any clinical or biological significance.

It is therefore necessary, to avoid falling into the trap of chance and drawing false conclusions from correctly collected but poorly analyzed data, to establish a safeguard. A first solution would be to divide the required p by the number of comparisons made: if 6 comparisons are made, the risk of obtaining a difference at $p = 0.05$ is 5% x 6, or 30%. To reduce this risk to 5%, the easiest way is to divide the required p-value by 6, which reduces to $0.05/6 = 0.0083$ (6 comparisons made, 4 subgroups compared two by two), the p-value required to have less than a 5% chance of being wrong in asserting a significant difference. This adjustment is called the *Bonferroni adjustment*.

Some will claim, however, that this adjustment is too severe, and will not make it possible to recognize, in this exploratory approach without a priori hypothesis on the location of the difference, a difference really existing at the level of $p = 0.05$ between two given subgroups. Other types of adjustment of p, less severe, have been proposed: the adjustment according to *Tukey* (which requires groups of identical size, which is rare in medicine), according to *Scheffé*, which can be applied even if the groups are of different sizes, according to *Neuman-Peul*, which consists of reducing the number of comparisons made (we start by comparing the two extreme groups: if there is no difference, the calculations stop there and only one comparison will have been made. If there is a difference, one of the extreme groups (with the lowest average for example) is then compared with the second opposite extreme group (with an average immediately lower than the group with the highest average). The tests stop there and we will only have made two comparisons: we declare that a priori there should be no other significant differences between the subgroups. If there is a difference, we continue the comparisons between the group with the lower mean and the group with the mean directly lower than the last group tested, and so on. This type of procedure makes it possible not to systematically compare all the subgroups formed, and to stop the comparisons of means as soon as the last difference tested is no longer significant: fewer tests are carried out, and the adjustment of p may be less severe.

There is always a certain trade-off, a certain balance, between rigorous testing and obtaining statistically significant results: the chances of detecting a significant difference are lower with a Bonferroni type adjustment than with a Neuman-Peuls type adjustment, but a significant difference in Bonferroni will be more likely to be really significant than a difference observed according to Neuman-Peuls...

In any case, comparisons of multiple means should only be made after adjustment of p, and the interpretation of these differences must take into account the rigor of the adjustment: the p-value < 0.05 does not establish alone a medical or biological truth, but represents an analysis 'clue' to facilitate the reading of the results.

6. Generalization of the Wilcoxon test to several subgroups:

The comparison of a quantitative variable with a non-normal distribution, or of an ordinal variable can be interesting between several subgroups: one may wish to compare the action of a diuretic (group 1), of an inhibitor of converting enzyme (group 2) and a beta-blocker (group 3) in the treatment of heart failure, and to assess how many patients in each of the three groups progress to the NYHA dyspnoea stage(s) (rated from I to IV) lower.

A comparison of the average will not be possible, because an average of the dyspnea stage cannot be calculated. A Wilcoxon test, however, could be performed for two-by-two comparisons.

The Kruskall-Wallis test represents the generalization of the Wilcoxon test for comparing non-Gaussian ordinal or quantitative variables between multiple subgroups. It is also equivalent to the Wilcoxon test for the comparison between two groups.

7. Comparison between two quantitative variables:

It is no longer a question of comparing two means, or the variance of a quantitative variable between two groups (the extent of weight loss in patients with a somatic or psychogenic etiology), but of examining whether there is a relationship between two quantitative variables: is creatinine level a function of body mass? Are serum levels of a potentially nephro- or myelotoxic drug a function of creatinine clearance? Is beta2-microglobulin level a function of lymphoma tumor mass? Is there a relationship between the importance of HIV viremia and the number of circulating CD4 lymphocytes? Is the level of glycosylated hemoglobin a good reflection of glycemic averages?

These questions all come under the notion of correlation, and the relationship can first be apprehended on a graph: the number of viral copies is plotted on the x axis, the number of circulating CD4 lymphocytes on the y, and the shape of the points cloud can be appreciated. If the points are all on a perfect line, there is most certainly a clear correlation between the two variables, with two exceptions: if the points are all on a vertical cloud, this means that the y values (the number of CD4), do not vary according to the viremia, but can take all the values for a given constant viremia. There is therefore no correlation, and the slope of the (vertical) line is equal to + infinity. If the points are all on a horizontal line, the CD4 count remains constant regardless of the value of the viremia: there is no correlation here either, and the slope of the horizontal line is equal to 0.

For there to be a correlation, the slope of the line must first be significantly different from 0 and infinity: the number of circulating CD4 must vary with the number of viral copies, the hemoglobin level glycosylated must vary with the average glycaemia. The slope of the line will be positive (greater than 0), if the increase in one rate is associated with the increase in the other. It will be negative (less than 0) if the increase in one rate is associated with the decrease in the other.

Examples:

Ttere is a positive correlation between the level of glycosylated hemoglobin and blood sugar levels, but the correlation is negative between the number of HIV viral copies and the number of circulating CD4 lymphocytes.

The basic model of the equation of a correlation is therefore the equation of a straight line: y = a.x + b, where a represents the slope of the line, and b the intercept, or the value of y when x is equal to 0. Biological values equal to 0 are rare in medicine, at least for biological parameters baseline (Leucocytes count, blood ionogram, coagulation parameters, etc.) or when the assay techniques are sensitive enough to detect low levels (a TSH strictly equal to 0 is rare with the ultra-sensitive assay technique). The intercept is therefore often a value extrapolated by the equation, and it is not certain that it corresponds to a *biologically observed value.*

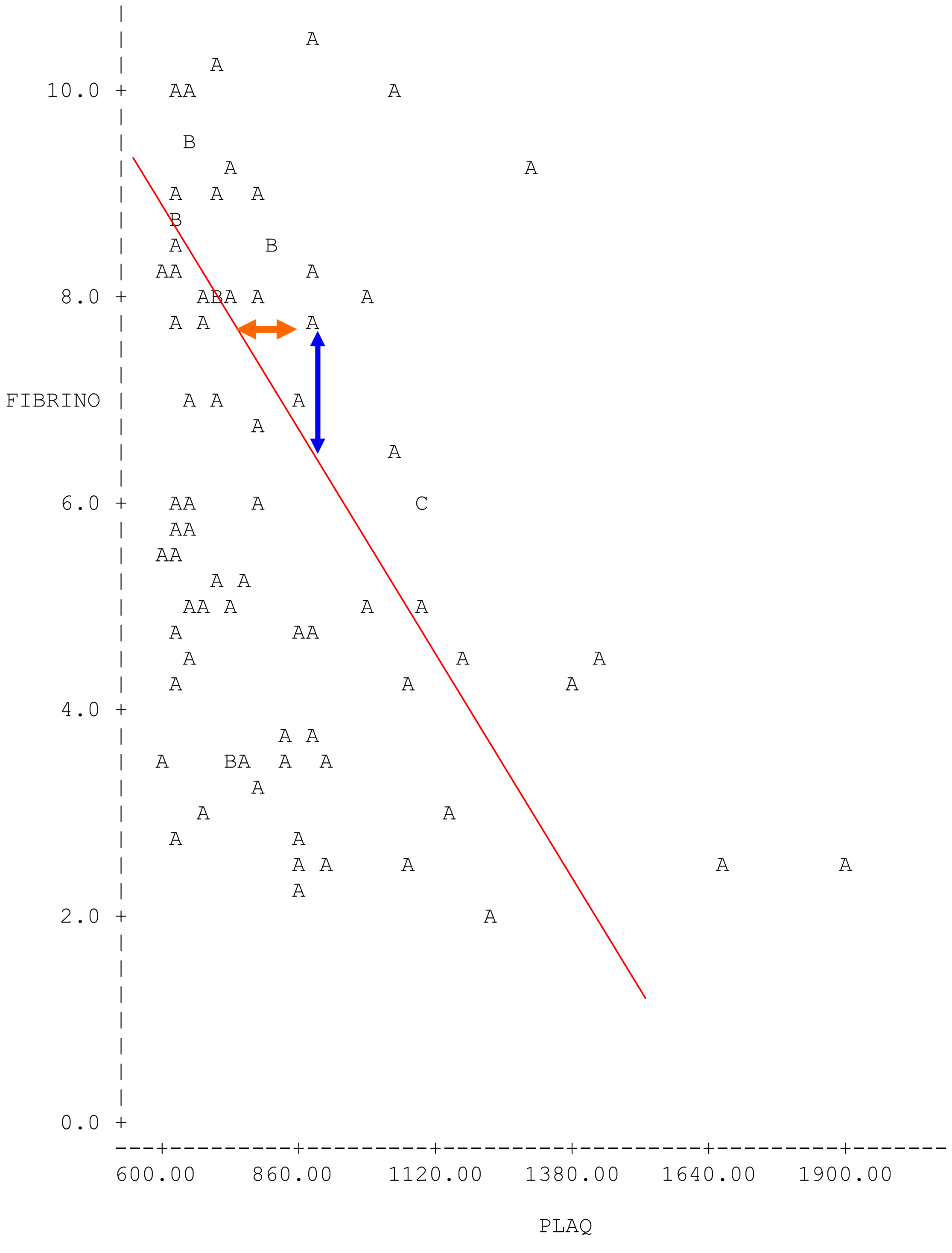
It is rare, however, that in biology or medicine, the points of correlation between two variables can be plotted exactly on a straight line. They most often form a cloud whose correlation line is the bisector. The equation of the line is not enough to satisfactorily describe the observed phenomenon. As in the comparison of means, where the variance describes the dispersion of values around the mean, there is a variance of each of the two variables around the regression line. For a given x (a number of viral copies), the corresponding y (the number of circulating CD4 lymphocytes) can be more or less far below or above the line. The best line, the one that will best represent the correlation, is the one for which the sum of all these distances (the distance of each y from the line) is the smallest possible. Some of these negative distances (when the observed y is located below the line), would artificially reduce the sum of the distances between the observed y, and the line (figuring the y predicted by the model): the solution to this problem consists in adding not the negative and positive raw distances, but the square of these distances. The best straight line describing the phenomenon is the one for which the sum of these squares will be minimal: the technique for producing the straight line is called the technique of the sum of the least squares (least square sum) (Fig. 3).

In this example, we have produced a correlation graph between the fibrinogen values and those of the platelet count in a population of reactive hyperthrombocytosis (essentially inflammatory), and wanted to test the hypothesis of physiological regulation mechanisms of the risk of thrombosis in these patients.

We first realize that there is a negative correlation (the greater the thrombocytosis, the lower the fibrinogen), with a correlation coefficient of – 0.34690. Moreover, this correlation is significant since the value of p (equal to 0.0016) is well below the classically accepted threshold of 0.05.

Fig. 3 *: CORRELATION ENTRE FIBRINOGENE ET PLAQUETTES*

```
Légende: A = 1 observation, B = 2 observations, etc.
```



The blue arrow represents the distance between the observed value of y (point A on our graph) and its expected (calculated) value on the linear regression line: this is the distance (y- y’). The orange arrow represents the distance between the observed value of x and its calculated value on the linear regression line: this is the distance (x-x’). We understand that if all the observed values of y and x were on the line, then the correlation between the two values would be perfect: y would always be exactly predictable as a function of x, and vice *versa. The correlation coefficient would be equal to 1 in absolute value. The further the observed values y and x 'walk' away from the line, the looser the correlation: the larger the distances (y-y') and (x-x'), the more scattered the point cloud, the weaker the correlation, the closer the correlation coefficient will be to 0.*

The correlation coefficient r (or rho) takes into account all of these distances y-y' and x-x' by relating them to the number of observations. It also takes into account, by its sign, the slope of the regression line: it will be negative if the slope is downward (the higher the platelets, the lower the fibrinogen) and positive if the slope is upward (the more the platelets are higher, the higher the fibrinogen). It therefore translates the sum of all the deviations of the observed points, compared to the calculated points located on the line.

The slope of the line a (or alpha) reflects the increase or decrease in the value of y when x varies: if alpha = 2, the value of y increases twice as fast as the value of x.

The whole procedure (calculating the equation of the line, the variance of x, the variance of y, the correlation coefficient) is called ***linear regression***. Again, a correlation will be said to be significant if the *slope of the line* is sufficiently likely to differ from 0 or from infinity, and if the *regression coefficient* is sufficiently likely to differ from 0. In other words, if the regression is far enough from a horizontal or vertical line, and whether the observed points are close enough to the estimated line. The p-value of the correlation takes these two ingredients into account, and we can thus see significant correlations with less than 5% chance of being wrong, with a low slope but a correlation coefficient close to 1, or at inverse with a significant slope but a low correlation coefficient. Most often, a significant correlation signs a trend, but it is rare that in medicine one can predict knowing x, a value y from the equation of regression (unlike what is done in physics or chemistry).

IV. MULTIVARIATE ANALYSIS: NOTION OF LOGISTIC REGRESSION

All the tests presented so far constitute the tools of univariate analysis: analysis of a variable according to a parameter (group or sub-groups, other variable in simple linear regression). There are many situations in medicine in which univariate analysis quickly marks its limits: we know that arterial hypertension, hypercholesterolemia, diabetes, smoking, stress, are risk factors for cardiovascular disease; but what is the respective weight of each of these risk factors in the occurrence of a myocardial infarction? Are these risk factors independent of each other, or do some of them only 'translate' the others (are smoking and high blood pressure, for example, only expressions, partially or totally, of stress)? Are some of these risk factors synergistic or, on the contrary, antagonistic? The same questions may arise for the known risk factors for breast cancer, lung cancer, the occurrence of a thromboembolic disease, or even diseases whose objective cause is known: Koch Bacillus (KB) is at the origin of tuberculosis, but contact with KB alone is not enough to trigger the disease...before the discovery of KB, the risk factors for multifactorial tuberculosis disease would have included malnutrition, promiscuity, unfavorable socio-economic conditions ... we would now add all the causes of immunosuppression, and even more recently, the way the immune system reacts, and in particular macrophage innate immunity, in contact with KB. The multifactorial disease before the discovery of KB, which became monofactorial with its discovery, becomes de facto multifactorial again with progress in immunology and genetics...

It is possible, to assess the role of each of these risk factors, to 'weigh' them independently, to carry out a succession of univariate analyzes by stratifying by each of them: one could compare, in a study cohort, the incidence of myocardial infarction among smokers and non-smokers; then, among smokers on the one hand, and non-smokers on the other, the incidence of infarction among hypertensive, and non-hypertensive, patients; then, in hypertensive smokers, normotensive smokers, hypertensive non-smokers, and normotensive non-smokers, the incidence of infarction in diabetics on the one hand, and non-diabetics on the other ; then, at.... and so on.

This succession of univariate analyzes makes it possible, of course, to determine whether stress adds an additional risk of heart attack in each of the sub-categories, and makes it possible to measure the importance of this risk: if it is higher in the sub-smoker-hypertensive-diabetic category than in the smoker-hypertensive-non-diabetic subcategory, it is that perhaps stress acts synergistically with one of the first three risk factors... but we will not know not which one.

It goes without saying that the power of the tests decreases with the size of the sample: stratification into subgroups requires a very large initial sample size if the last subgroups must still include a sufficient number of patients... In practice, this strategy, although theoretically satisfactory, is rarely possible. It is also very heavy.

The alternative is represented by logistic regression: without going into mathematical details, the principle of its equation could be written as follows (be careful, this expression is mathematically false, but its overall meaning is correct).

$$Illness = intercept + OR1.\ RF1 + OR2.RF2 + OR3.RF3 + OR4.RF4 + ...ORn.RFn.$$

In this equation, the disease is generally expressed in a binary way: it exists, or it does not. The different risk factors (RF) can be expressed in a binary way (1 or 0), in an ordinal way (1, 2, 3, 4...), or even in the form of a continuous quantitative variable. When it is a binary variable, we can extract from the coefficient assigned to it the odds ratio (OR), representing the relative risk linked to the risk factor considered taking into account the risk corresponding to the other risk factors. In other words, OR1, OR2, OR3, OR4, quantify the risk associated with each risk factor, knowing that the disease is also explained by the other risk factors kept in the equation.

In practice, the risk factors significant at 0.1 in univariate analysis (for which $p < 0.1$) are introduced into the logistic regression model. Risk factors that are not significant in univariate analysis can be forced into the model if this seems to be biologically or medically justified. There are several types of procedures in logistic regression, but the principle consists in first introducing in the model the most significant risk factor; if it does not by itself explain the entire disease, the second risk factor is introduced into the model; if found significant, it remains in the model. If not, it comes out and the third is then introduced and tested. The introduction of the xth risk factor stops when no more significant risk factor is found, in other words, when the introduction of an additional risk factor no longer provides any additional explanation for the occurrence of the sickness.

Example:

If in cardiovascular diseases, hypertension or smoking were only an expression (only a translation) of stress, but did not by themselves explain part of the incidence of myocardial myocardium, they would be eliminated from the model in favor of the stress variable. If, on the contrary, stress did not play a role, and hypertension or smoking acted as confounding factors allowing stress to appear significant in univariate analysis, stress would no longer come out significant from the logistic regression model, which would only keep hypertension and smoking as true risk factors.

The odds ratio for each risk factor tested and their confidence interval can be calculated from the parameters of each risk factor integrated into the logistic regression equation. We can also test the interaction between two risk factors. Imagine, in a study of mesothelioma risk factors, that smoking is coded 1 if present, 0 if absent. Asbestos exposure will be coded in the same way. It is easy to create a variable reflecting the simultaneous exposure to the two risk factors: tobacco.asbestos will take the value 1 (1x1) when the two risk factors are present, and 0 (1x0, 0x1, or 0x0) when there will be no simultaneous presence of smoking and exposure to asbestos. We can write the following logistic regression equation:

$$mesothelioma = intercept + OR1.Tobacco + OR2.\ Asbestos + OR3.\ Tobacco.asbestos$$

If the tobacco.asbestos variable is retained in the model as significant with an odds ratio greater than 1, it means that the tobacco-asbestos association adds a significant risk in relation to exposure to tobacco on the one hand, and to asbestos on the other hand. Therefore, there is synergy between tobacco and asbestos.

If the tobacco-asbestos association is retained as significant, but with an odds ratio lower than 1, it means that the association is antagonistic: the combined presence of the two risk factors reduces the overall risk of occurrence of the disease (the antagonists are rather rare in medicine!).

If the association of the two, combined, risk factors is not retained in the model, it is because it does not modify the risk already expressed by the presence of tobacco on the one hand, and asbestos on the other hand. : the two risk factors are neither synergistic nor antagonistic, but add up.

Logistic regression is therefore an extremely powerful and useful tool in medicine or biology. It makes it possible to carry out a fine analysis while avoiding the pitfall of the gigantic sample sizes required by the stratified univariate analysis. It makes it possible to weigh each risk factor, to measure its independence from the others, to test the interactions, to control the confounding elements.

V. ANALYSIS OF THE PROGNOSIS. SURVIVAL CURVES

Survival curves represent the essential tool for the analysis of the prognosis: the 'living/dead' event can of course constitute the descriptive factor used, but any other event translating the prognosis can also be: occurrence or not of a complication, occurrence or not of recovery, occurrence or not of an intercurrent illness: their common characteristic is that each time, the event may have already occurred at the time of data analysis, or may not have occurred (because it has not yet occurred, or because it will not occur), and we then speak of censored data. The survival curves therefore include patients for whom it is not known whether the event studied will occur, or not, one day. It may seem presumptuous to analyze data whose reality in the future we do not grasp: doctors, biologists, statisticians are all human, and the future does not belong to them... it is true. A survival curve therefore never makes it possible to describe what will happen to a patient: it makes it possible at most to describe what happened to previously known patients, and a probability of survival at some point in the evolution of the disease. A survival curve should never be used to tell a patient or his family that his risk of death, his chance of recovery, his risk of a complication occurring is x% at one year: for a given patient, the risk of death is either 100% or 0%. There is no alternative between the fact to live, or not to live. At the present time, no science makes it possible to predict the future, and using a scientific vocabulary, even a calculated approach, to try to approach it never makes it possible to be affirmative at the level of an individual: the field of prognosis is undoubtedly the one in which medicine offers the more uncertainties, and these uncertainties are not limited by analytical techniques.

How do you construct a survival curve? Weaknesses and strengths:

1- Time 0

The only fixed times in a life are those of birth, usually dated with precision, and of death, the time of which can be known... once it has occurred. The only survival curve providing indubitable information would therefore be the one counting down the time between birth and death. To analyze a duration, it is necessary to start if possible from an identical time 0 for all the patients.

When we are interested in the prognosis of a disease, time 0 is more difficult to define. We very often consider time 0 the moment of diagnosis of the disease, and we say, a little lightly, that the prognosis of lung cancer is x% survival at two years. The limits appear in an obvious way: time 0, that of the diagnosis, is not the same for the patient screened in occupational medicine (small asymptomatic round spot) than for the one diagnosed in front of a massive deterioration of the general state with metastases bones and brain. One could, for more precision, stratify the time 0 according to the stage of the cancer. However, screening studies have shown that time 0 was not the same, at the size of an asymptomatic lung round image on a routinely performed X-ray, for a lesion diagnosed in the year of the first screening carried out in a company, year during which tumors that may have been present for 2, 3, 4... years, and the following year, that of the second screening, where only tumors that have appeared in the last twelve months are screened . At 1 centimeter in diameter, a tumor developed in 3 months is probably more aggressive than a tumor of the same diameter evolving for 3 years. The time 0 of the diagnosis of the tumor is not the same: an aggressive tumor at three months can already be advanced, an indolent tumor at three months can still be in its pre-clinical phase...

Time 0, prior to the construction of any survival curve, is therefore defined by the means of observation at our disposal. We will remember before constructing a survival curve that the sun probably exists before its sunrise time, that its sunrise time changes at any point on the surface of the globe, and that, if noon within a time zone will strike at the same time, true noon will be different for each individual depending on their location in the time zone. An approximation of the same nature, but of undoubtedly greater amplitude depending on the pathology, governs the construction of a survival curve.

2- Notion of conditional survival

The ideal would of course be to have the same observation period for all the patients included in the study: we could thus affirm, while retaining the uncertainty of time 0, that at ten years from the diagnosis the survival of the cohort reaches, for example, 50% (which does not mean, once again, that the chances of survival of Mr. P. Dupont at 10 years are 50%). However, patients do not all start their disease at the same time: some will be followed for 10 years at the time of the analysis, others only for 1 year. The first will have had time to heal, to go into remission, to die, the others, not. We speak of censored data, when the measured event has not yet occurred, and of uncensored data, when it has occurred. If death is the event measured, the censored data will correspond to patients alive at the time of the analysis (their date of death is not known).

The cumulative survival probability curve based on the product of the conditional probabilities (Kaplan-Meier curve) is therefore constructed as follows: all patients (100% of them) are alive at the time of diagnosis. The survival curve starts, on the ordinate, from the value 100. At the first death (let's put it 3 months from time 0), 1% of the initial population disappears: the curve descends by a step of 1% to the level of the abscissa '3 months'. Imagine that two deaths occur 6 months from time 0: the curve will then record a downward march of 2% on the basis of the 99 survivors, at the abscissa '6 months'. If only 5 patients have been followed for more than 5 years (either because all the others died, or because they were diagnosed during the last 4 years...) and two patients among these 5 die at 5 years, the downward step will be 2/5, or 40% of the residual population.

This explains that on a survival curve, the downward steps become more and more marked towards important times. It also appears that they are becoming less and less precise, because they relate to an increasingly reduced sample size: *the confidence interval around the value, and therefore the uncertainty about the value, increases when the size of sample decreases, and it inevitably decreases along the survival curve.*

Caution is therefore called for when discussing the probability of survival in medicine, and certain points should always be borne in mind:

- A probability of survival applies to a group of patients, but does not apply to each individual patient. We do not know, because the statistical tool does not answer this question, if the patient diagnosed on this day will live, or will not live, in 5 years: his survival will be 100%, or will not be. His probability of survival will not be equal to that of the group.
- The probability of survival at x years is valid for the group only at time 0: the conditional probability of survival varies as a function of time, and for the example of the 5 patients surviving at 5 years among 100 patients initially included, it will be, at this moment of the curve, 40% for the time to come.
- The future, even surrounded by statistics, remains a very difficult data to approach.

3- Comparisons of survival curves in univariate mode: the log-rank test.

Two survival curves always end up meeting, it is only a matter of time… in the long term, the difference cannot be significant. Consequently, their comparison will depend on the speed with which each event (death, occurrence of such a complication) will occur during follow-up, and on the number of events occurring at a given time. At time t, x% of patients in one group, y% in the other, will survive. The comparison of proportions, at that instant t, can therefore be carried out by a chi-2 test. At the next moment, one of the proportions may have changed: a second chi-square test will then have to be carried out. It will be the same for each change of proportion on one or the other curve, in other words, for each new stair step on one or the other curve.

The overall comparison of all the survival curves will therefore depend on the result of all the comparisons of proportions carried out at each change in one of the curves analyzed, i.e., in statistical terms, on the sum of all the chi-2 performed adjusted for varying sample size at each step. The log-rank test performs this analysis and summarizes all the added differences: it is a particular form of Mantel-Haenszel test.

In practice, this statistical test measures the differences between the two survival curves compared, for each stair step occurring on one or the other of the survival curves. It then tests the null hypothesis: the sum of these distances is equal to 0. If this is true, the two curves are not very far from each other, and they reflect similar survival. If this is not true, then the two curves are far from each other, and reflect different survivals.

4- Comparison of survival curves in multivariate mode: the Cox model

Several elements can be taken into account in survival: the presence or absence of disease, of course, but also the type of treatment received, compliance with treatment, compliance with the initial protocol, the presence of co-morbidities, the existence of other risk factors, the stage or grade of the disease, the patient's clinical condition measured by a score at the time of diagnosis, etc.

All these elements cannot be represented on a survival curve, but one can imagine that some of them may be more important, for the survival of the patient, than the type of treatment received. If we compare two treatments A and B, two forms of the disease X and Y, it will be important to equalize the possible prognostic factors between the two groups. This equalization is the goal of randomization in a controlled trial, but it may not always be achieved. It is rarely reached when there is no randomization, and taking into account the various factors involved in survival should then call for a staged stratification, and the comparison carried out in subgroups of homogeneous patients for each prognostic factor. This would quickly lead to a multiplication of subgroups with reduced sample size, and a significant loss of power for the study. The results would become difficult to interpret.

Another possibility consists in integrating all the variables that can play a role in the prognosis in a logistic regression analysis model, adapted to the analysis of survival curves (capable of taking into account censored data, and not censored). *This particular logistic regression model is the model published by* ***Cox****, which will be able to give an estimate of the relative risk associated with each prognostic factor*. If this relative risk is significantly different from 1, the considered factor will play a significant role in the prognosis. Otherwise, it may be considered as non-determining. In all cases, the model will give an estimate of the relative weight of each prognostic factor in the occurrence of the measured event (death, occurrence of an iatrogenic or natural complication, etc.), without loss of power secondary to any stratification. On the other hand, only observations without missing data for all the variables analyzed will be taken into account, which can result in a significant reduction in the size of the sample if these variables have not been correctly entered: as for any clinical study, the rigor in the collection of data determines the reliability of the results, regardless of the degree of sophistication of the analysis used.

CONCLUSION

Statistical analysis is essential in medicine and biology to test trends at the level of samples (whether they are groups of patients, cell colonies, groups of genes, families of proteins, etc.) or populations. The trends tested will be valid for the populations tested, but cannot be applied as such to a given individual, in particular for survival studies integrating, in addition to known data (uncensored), data of unknown value at the time of analysis (censored). The analysis is only valid if the data have been correctly collected, are complete, and if they correspond to an ***a priori working hypothesis***. 'Fishing' analysis in a database, looking for statistically significant associations or correlations with no biological or clinical basis foreseen by the investigator, is likely to yield significant results only by chance and should therefore be avoided. It is possible that this type of unplanned analysis is at the origin of a large part of the contradictory results reported in the medical literature, and the controversies then engaged could be based, at least partially, only on effects of chance… and not on the basis of acquiring so-called 'scientific' knowledge. As for any observation method, the conditions of application, the indications and the contraindications of the various statistical tests must be known by the doctors who will de facto be the users of the results. Any result then must be interpreted according to the entire methodology of the study prior to the statistical test, and the conditions under which the test was applied.

A statistical test rarely proves. It tries to circumscribe the fact of chance around the observed events, provided that the model underlying the test describes the multifaceted reality of life fairly well (this can sometimes be quite difficult to assert). On the other hand, it makes it easier to avoid undue assertions (one treatment is superior to, or different from, another) when the difference in frequency of the events observed is too close to 0.

To know more :

- Bouyer J. Statistical methods. Medicine-Biology. ESTEM, INSERM Editions, 2000.

- Bailer JC III, Mosteller F. Medical uses of Statistics, 2nd Edition, 1992, NEJM Books, Boston, Massachusetts.

- Hill C, Com-Nougué C, Kramar A, Moreau T, O’Quigley J, Senoussi R, Chastang C. Statistical analysis of survival data, 2nd Edition, 1996. INSERM Médecine-Sciences, Editions Flammarion.

Table I: *Summary of common statistical tests, their indications and limitations.*

Type of variable	Test indicated	Limits of applicability	Alternative required if limit of the indicated test reached
Dichotomous (proportions), 2 groups	Chi-2 with 1 degree of freedom Alternatives: Yates (Chi-2 corrected) Mantel-Haenszel (Chi-2 for stratified data)	At least one of the cells in the table contains less than 5 expected events	Exact Fisher test
Dichotomous (proportions), multiple groups	Chi-2 with degrees of freedom = (number of columns -1) * (number of rows -1)	Doesn't tell where a difference is if the test is significant, but just says that there is a notion of difference within the data table. Will give false results if one of the table cells contains less than 5 expected events	If one of the cells of the table contains less than 5 expected events, - a Fisher test is theoretically possible, but few computers will have sufficient memory to perform it.... - know that the results are distorted, difficult to interpret, and consider another type of data analysis.
Ordinal, 2 groups	Wilcoxon rank sum test	Each group must contain at least 10 observations	
Ordinal, multiple groups	Kruskall-Wallis	Indicates whether there is a difference somewhere between the different groups. Does not indicate where the difference lies.	
Quantitative, 2 groups	Analysis of variance (comparison of means) Z-test if very large	The distribution of variables must be normal (Gaussian) in both groups.	If the distribution is not normal, use the Wilcoxon test as for the ordinal variables. The Wilcoxon test

	population t-test if sample: usual clinical situation		and the analysis of variance will give similar results if n > 30
Quantitative, multiple groups	Analysis of variance	The distribution of the variables must be normal in the different subgroups.	If the distribution is not normal, use the Kruskall-Wallis test as for the ordinal variables.
	F-test: global analysis	Indicates that there is a difference somewhere between the groups. Does not indicate where the difference lies.	
	Subgroup analysis with adjustment according to Bonferroni, Scheffe, Tukey	Indicates which subgroups differ from each other. The severity of the test depends on the type of adjustment (Bonferroni being the most rigorous)	
Relationship between two quantitative variables	Correlation test according to ***Pearson***	The distribution of the two variables must be normal. Test very sensitive to extreme values (outsiders)	If the distribution is not normal, use the rank correlation test (according to ***Spearman***).
Relationship between a quantitative variable and several quantitative variables	Multiple linear regression	The distribution of all the variables must be normal.	If the distribution is not normal, or for particular distribution variables, there are nonlinear regression models, with or without transformation of the variables
Relationship between a dichotomous variable and several dichotomous, ordinal or quantitative variables	Logistic regression	Allows the estimation of interactions between variables, the control of confounding elements and effect modifiers. Does not require the normal	

		distribution for quantitative variables.	
Construction of a survival curve	Kaplan-Meier model	Establishes the conditional probability of survival. Its accuracy decreases as follow-up progresses and the number of patients followed decreases	
Comparison of two survival curves	log-rank test (equivalent to the Mantel-Haenszel test)	Does not take into account co-factors involved in survival	
Comparison of two survival curves, taking into account co-factors, confounding elements, or effect modifiers	Cox model	Used to determine the respective weight of the different factors involved in survival. Allows to estimate the relative risk for dichotomous factors. Allows you to identify confounding factors or effect modifying factors.	