Chapter 12. Much ado about nothing?

Having demonstrated the flimsiness of the main argument concerning the investigation report, let us nevertheless resume the affair from the beginning. What amazed the investigators, as they expressed on numerous occasions, was the small sampling error of the duplicate or triplicate counts of basophils that were recorded in the laboratory notebooks of E. Davenas – they were “*made up*” said W. Stewart. The same criticism was made about the results of the article of *Nature* (corresponding to the experiments performed in Israel). Why – regardless of any calculation – did the results often seemed “too good”? Here again, some mathematical considerations could help to consider this issue in a different way.

The misleading consequences of asymmetry

First of all, what does “too good” counts mean? Let us suppose that we perform three successive counts from a tube that contains (with certainty) 100 basophils (per volume unit). It is, according to the statistical terminology, like a box from which we obtain random samples. We take three successive samples: we find 99, 101 and 113. We can calculate the mean, the variance and the standard deviation (it is the famous sampling error that is the square root of the variance). The calculation gives a mean of 104.3 with a standard deviation of 7.6 and a variance of 57.3. Generally we express this result in the following manner: mean ± standard deviation = 104.3 ± 7.6 (n=3).

What we try to assess is the number of basophils in the tube. We have here an approximation of this. We conceive that the more the number of samples is the higher, the higher our confidence in this result. The standard deviation (sampling error) gives us an evaluation of the variability between the various counts. As regards an enumeration we know that, with some conditions (see previous chapter), the law which *a priori* applies is the law of small counts. As we have seen, variables distributed according to the law of small numbers have a variance which is equal to the mean.

Let us resume the calculation obtained above from three samples. Its variance (57.3) is lower than its mean (104.3). Was it therefore “made up”? Not necessarily, because this variance itself fluctuates at random. But within what limits? It is what we are going to evaluate. We will suppose that we are in ideal conditions and that only random is responsible for the results. In other words, we suppose that there is no statistical noise added to the law of small counts.

We use the following procedure: we take 3 samples and then we calculate the mean and the variance of these 3 counts. We reproduce the same operation until we obtain 1000 means of 3 values and their respective 1000 variances.

Among the 1000 means and the respective 1000 variances, what is the percentage of variances (s^2) which will be superior to the mean (m) and what will be the percentage of the variances lower than the mean, that is:

(1) Percentage of triplicate counts with $m > s^2$ (equivalent to $s^2/m < 1$)?

(2) Percentage of triplicate counts with $m < s^2$ (equivalent to $s^2/m > 1$)?

The first answer that comes to mind is: 50% for (1) and 50% for (2). Our intuition suggests us that, due to the law of large numbers, we shall be able to obtain aproximately as many values on one side as on the other one.

Are we so sure of this result? Let us perform a computer simulation with random numbers generated according to the law of small numbers. We thus obtain 1000 counts in triplicate (generated with a mean equal to 100 and consequently with an overall variance equal also to 100). We then calculate the ratio s^2/m.

	Count 1	Count 2	Count 3	s^2/m
Series 1	93	90	117	2.19
Series 2	97	108	112	0.571
Series 3	104	107	108	0.041
Series 4	112	115	110	0.056
Series 5	99	84	105	1.219
Series 6	129	95	97	3.402
Series 7	110	76	97	3.12
…….				
Series 1000	99	99	94	0.086

We can now graphically show as a cloud of points these 1000 values of s^2/m and study class distribution:

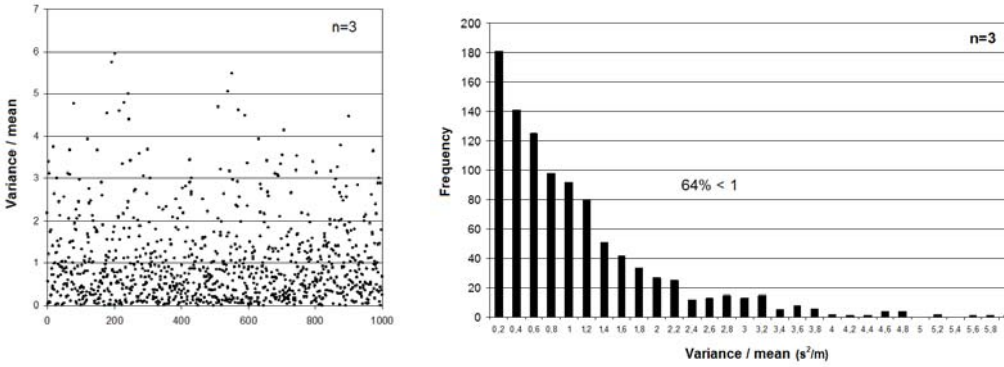

Figure 12.1 Distribution of the ratio variance/mean of small samples (n=3).
NB. For this histogram and the next ones, each value of x-axis corresponds to the upper limit of the interval.

Our intuition was thus completely wrong since we observe that the law of distribution of s^2/m is asymmetric. We find that the distribution on each side of the mean is not 50/50 but 64/36. Furthermore, the most likely values are smaller values of s^2/m! *The mean of the 1000 ratios s^2/m is nevertheless close to 1* according to the law of small counts.

Our intuition (and our poor knowledge of the statistical laws) first led us to confuse the *mean* of a variable and its *mode* (that is the value which is the most frequent). The mean corresponds to the mode only in the case of symmetric laws of probability. The paradoxical conclusion (often poorly understood because not intuitive) is that – due to the asymmetry of the ratio variance/mean for the small samples – *the variance is more frequently lower than the mean.* It is a fundamental result. No doubt that it will make statisticians and mathematicians smile because it is probably an obvious fact for them. I am not certain that this "obvious fact" was shared within the team of the investigators and – let us be honest – among the members of the team of Clamart.

Let us pursue our exploration and see what happens in the case of duplicate counts:

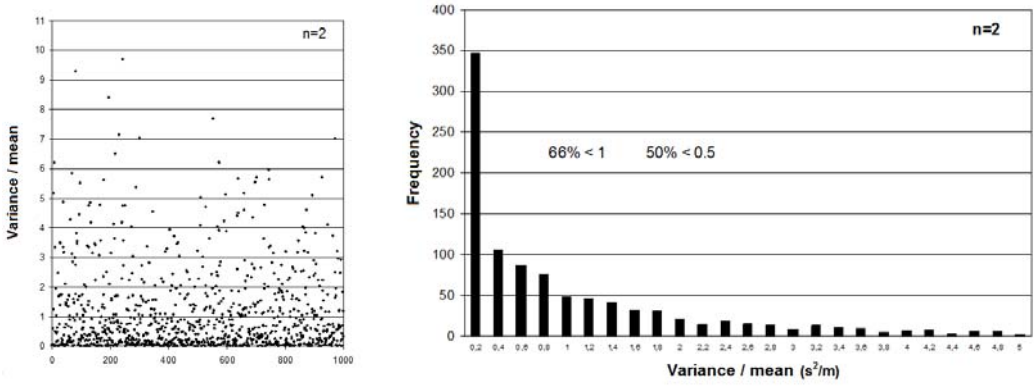

Figure 12.2. Distribution of the ratio variance/mean of small samples (n=2)

The difference is even higher: in 66 % of the cases, the variances are lower than the average. And in half of the cases the ratio variance/mean is lower than 0.5. We end now our exploration with n = 10 (Figure 12.3).

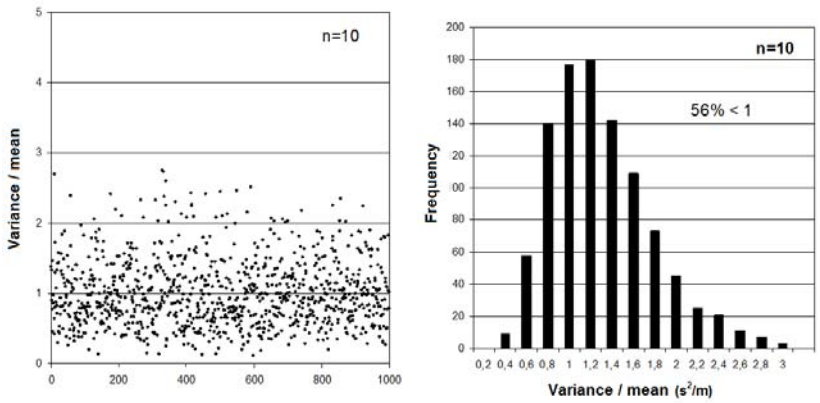

Figure 12.3. Distribution of the ratio variance/mean of small samples (n=10).

With samples of 10 values, there is a trend for a more symmetric distribution (but we have still 56% of values lower than 1). For samples above 30, the distribution is symmetric (Gaussian).

To sum up, one can notice that the distribution of the variances of small samples (n=2 or 3) is strongly asymmetric (here we used a Poisson's distribution, but a Gaussian distribution would give similar results). The consequence is that if we try to verify the fairness of countings using the variances of small samples (as it is often the case), we risk to conclude that the results are "too good". Here is for example a computer simulation of 10 counts of basophils corresponding to the law of small countings:

These wells are supposed to contain the same number of basophils (100). Every result is given as mean ± standard deviation.

Well 1 :	117 ± 6	Well 6 :	99 ± 3
Well 2 :	92 ± 2	Well 7 :	110 ± 8
Well 3 :	101 ± 13	Well 8 :	106 ± 16
Well 4 :	95 ± 6	Well 9 :	96 ± 8
Well 5 :	94 ± 3	Well 10 :	93 ± 5

Mentally, we calculate the variances by taking the square of the standard deviation. We notice that except for wells 3 and 8, the variance is very often smaller than the mean and frequently very small. We begin to be suspicious. Indeed we learnt at school that with this kind of counts the variance must be equal to the mean and that it is precisely a method to verify that countings are without bias. Were the values "made-up"? For the variances superior to the mean, we could imagine that the volumes were not quite exact or any other explanation (statistical noise). But for the variances lower than the mean, the only explanation is that "order has been introduced". Let us remind that this reasoning is made with values obtained from a computer simulation (they have been not selected).

Application to the results of Israel of February-March, 1987 (case with n=3)

Armed with our updated knowledge, we now resume these controversial basophil counts and calculate the ratio s^2/m (variance/mean) of Table 1 of the article of *Nature* (namely the 4 experiments made in Israel by E. Davenas with theirs results in Appendix 2) and then we calculate the distribution of the values:

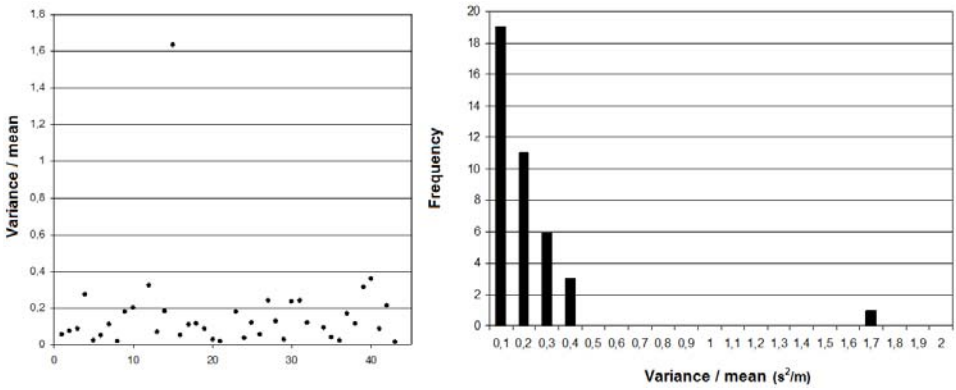

Figure 12.4. Distribution of the ratio variance/mean of the Israeli experiments of February-March 1987 (small samples with n=3).

One can observe once gain the same strong asymmetric distribution with the highest probability for the smallest values. It is difficult "to build up" such results. The reader can try to simulate results by inventing triplicate counts and he/she will notice that it is not easy to obtain such a distribution, especially if one does not think about the aspect of distribution that must be asymmetric. It is – for those who would have doubted – an argument in favour of the "sincerity" of the countings performed in Israel.

Playing the devil's advocate, we notice that the mean of the ratio variance/mean is not 1, but only 0.16. However, by taking into account the results of the article of H. Gérard *et al*, the ratio for approximately 80 basophils should be noticeably lower than 1, approximately 0.34, what is closer to 0.16 without achieving it nevertheless.

It is not impossible that some odd values were counted again, "by precaution", because they were too far from the two other counts. In other words, we cannot objectively rule out a conscious "experimenter effect" on the triplicate counts. On average, this procedure *changes nothing* to the result of the experience because values which are "too far" occur with an equal probability in a direction or in the other one. Given that the "label" of the tested sample which was not known (*blind* experiments), the results could not be biased in favour of an effect of high dilutions. Let us remind that the purpose of these experiments *was not to verify the validity of the law of small counts on repeated counts*, but to assess a possible difference between "active" samples and "inactive" samples with the most precise method.

Application to the results of investigation of Nature of July 1988 (case with n=2)

Let us study now the distribution of the ratio s^2/m for the experiment which had been counted in duplicate and in blind conditions (case where n=2), namely experiment F counted on July 7th (see Appendix).

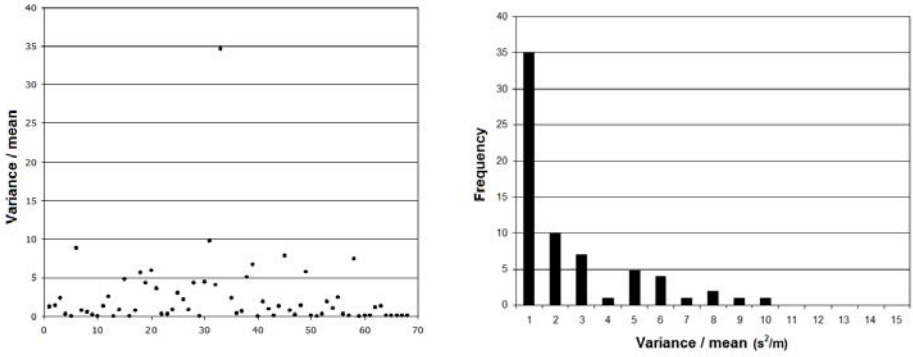

Figure 12.5. Distribution of the ratio variance/mean of the experiment F performed during the investigation of *Nature* in July 1988 (small samples with n=2).

We immediately notice the difference of abscissa of the ratio variance/mean in comparison with the experiments of Israel or with the experiments simulated by a computer for n=2. The problem here is not to explain too small variances (compared to the mean), but to explain *very high variances*! The additional statistical noise is obviously very high (the mean of the ratio variance/average is 2.4). This very high variance could result from errors in pipetting the sample volumes or from heterogeneity of the cell suspension.

Let us repeat, at the risk of annoying the reader, that: 1) these technical steps were managed by W. Stewart; 2) it was precisely this experiment that had been used for sketching the famous expected distribution when the counts were done in blind conditions.

Is there a physical phenomenon which could explain the too low variance of the counts of basophils?

We saw in the previous chapter that if the particles (cells, bacteria…) that are counted tend to repel each other, then their dispersion decreases and the variance of the counts is lower than the expected variance. By which mechanism, could basophils tend to remain at a distance from each other?

To explain this "anomaly", we must find a mechanism which concerns basophils and not the other cells (let us remind that only 1% of all white blood cells are basophils) and if possible only those basophils which are counted, i.e. colored basophils (not activated).

The explanation could be precisely related to the staining of basophils. Indeed, toluidine blue stains basophils in a particular manner: the staining agent is blue, but basophils are colored in red. The phenomenon is named metachromasia. Metachromasia is a property of some staining agents which color tissular structures with a color different from that of the initial staining solution. This property is observed only for some electropositive stains such as toluidine blue. The metachromatic reaction is the hallmark of polycationic structures on which binds the colouring agent. Indeed, basophil granules are mostly composed of a matrix of acidic mucopolysaccharides which is very electronegative. The *high density of negative charges* is responsible for the shift of the emission wavelength of the staining agent (from blue towards red) because of "aggregation" of molecules of staining agent.

Toluidine blue could thus be seen as a marker of structures having an important density of negative charges; in this view, granules of unstained basophils have lost their electronegative charges (see Appendix 1). As we all know, charges with same sign repel each other. Consequently, during the few minutes when cells settle in the chamber of the hemocytometer, the *repulsive electrostatic forces* that originate from basophils would tend to slighthly repell other basophils. The distances of one basophil with its closest neighbours would consequently be more regular than allowed by chance. And this would be especially the case as the concentration of basophils is high as reported in the article of H. Gérard *et al* because the intensity of the electrostatic force decreases with the distance.

Other mechanisms could be suggested to explain the observations of H. Gérard *et al.*[1] Although there is at present no certainty on the reasons of this deviation from the law of small counts, this allows nevertheless illustrating the idea that it is sometimes simplistic to apply a mathematical law on a complex physical or biological phenomenon without precaution.

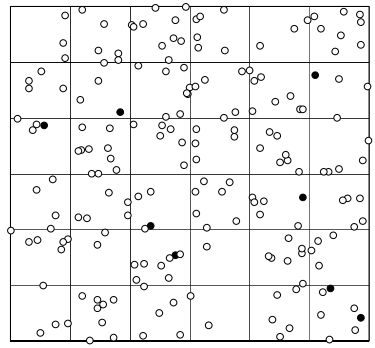

Figure 12.6. If the position of each basophil (black balls) is independent of the positions of the other basophils and white cells (white balls), then the law of small counts apply when one counts a series of samples taken from the same population. However, if a repulsion force (or a mechanism leading to an identical effect) is present, then the distances between basophils are more regular than expected according to the law of small numbers. The consequence is that the variance of the counts decreases because order has been introduced. This is what is suggested by empirical data obtained with basophil counts. It could involve an electrostatic force (force of Coulomb) taking its source from the sulphated glycoprotein matrix of the basophil granules which possess numerous electronegative charges. It is precisely because the density of electronegative charges of this matrix is high that the phenomenon of metachomasia occurs with toluidine blue when basophils are stained. Since such a repulsive force would decrease with the distance, this would explain why this phenomenon would be especially visible with high cell concentrations, as noticed by the authors of the article of H. Gérard *et al.*

Why an effect related to high dilutions was not highlighted in the last three blind experiments of the investigation?

First, let us be clear, it is possible that, even performed in better conditions, the experiments in the center of the debate would have been negative. Only someone totally uninvolved in experimental biology and medicine could be surprised by this fact. Of course, it sometimes happens that one tests "massive" hypotheses for which the use of statistical methods is not necessary. In general, a bench experiment is infrequently a long quiet river. And, as explained by J. Benveniste:

> "[…] All which seems to have interested *Nature*'s people, it is that the experiment could, once, not succeed. But we knew that already! We did not need them to know that! And I have the impression that their purpose was to push the system to its limit, to create problematic working and achievement conditions to obtain, finally, a failed experiment." [2]

To guard against various biases of interpretation, it is important to decide *a priori* (that is before knowing the result) what are acceptable experimental conditions. For example, in the case of high dilutions of anti-IgE, the experience accumulated during three working years, allowed defining – among other conditions – that it was necessary to get correct controls and first peak (i.e., included in predefined limits) even before considering the results with high dilutions. It is surprising to notice that the experts were flabbergasted when they found out (or pretended to find out) that in some experiments), no degranulation with antiserum anti-IgE was observed:

> "We were surprised to learn that the experiments do not always "work". […] It also appears that some bloods that "do not degranulate" are often encountered; we were informed that, in this event, data are recorded but not included in the analyses prepared for publication." [3]

Let us suppose that we test the effect of a medicine on a population of patients. It is quite possible – it is even the rule – that the medicine is ineffective in some patients. We are not surprised by this fact. It is for these reasons that statistics are used to analyze the results, not on an individual basis, but on the entire populations of patients. We are here in the same scenario. What is important is to know if, *overall,* on the whole set of experiments, a statistically significant effect is obtained in the presence of high dilutions. For J. Benveniste, the experience accumulated by his team during several years, including numerous blind experiments with an appropriate statistical analysis, had a weight that was far greater than these three negative experiments performed in poor experimental conditions and, furthermore, with a unique series of anti-IgE dilutions. The purpose here was not to demonstrate a mathematical theorem for which a unique counter-example is enough to invalidate it.

A last-minute correction

We remember that a sentence reporting the results of the 4th experiment had been deleted in the investigation report (it was present in the printer's proofs). This sentence said basically that the effects noticed with the high dilutions were nothing else than statistical fluctuations, but that this explanation did not apply to all results and particularly to the famous 4th experiment.

In fact, this sentence deleted at the last moment represented only a part of a paragraph interesting to reproduce in its totality because it concerned – again – the sampling issue:

> "The control values are used to normalize the readings obtained with reagents at high dilution. Despite the laboratory's convention

> of presenting data as the percentage of degranulation by diluted agents relative to the controls, it appears not to have been considered that the counting error is the statistical sum (square root of the sum of the square) of the sampling error in counting a single well and the estimated error of the mean of the control samples. In the particular case of the first experiment, for example, we estimate the expected sampling error to 14 per cent. It seems clear that many of the peaks reported as significant at Clamart 200 (*sic*) are well within two standard deviations of the line of null "achromasia", even when no account is taken of other sources of error (such as failure to record basophils).
>
> Thus we believe that many of the experiments whose results are regarded as significant are artefacts of statistical noise. But plainly this does not apply to all the data (for example, the fourth experiment of the study." [4]

This passage – even if it was not kept in the published version – confirms the real obsession of the investigators towards the error of sampling. Here, J. Maddox tried to create suspicion (by very technical arguments). Basically, he suggested that the researchers of Clamart selected some results that in fact had emerged from the statistical noise.

However, the approach for the calculation was correct in this case in contrast with the erroneous calculation above. This incoherence could perhaps give some explanation for the calculations with the inaccurate formula that were performed on site by W. Stewart who came to Clamart with a microcomputer – a Macintosh – and from the first day recorded data of the laboratory notebooks of E. Davenas.

J Maddox having finally decided not to publish this passage, we cannot accuse it in bad faith, however he was tempted once again to confirm his prejudice on an experiment which was of poor quality (experiment A; see chapter 9). If J. Maddox had followed the same reasoning on the experiments B and C (see chapter 9) – which he preferred not to show in the investigation report – he should have recognized that these results were not "artefacts of statistical noise". Maybe it is the reason why J. Maddox preferred to delete this passage.

On good usage of the irony

The investigators thus showed a rare insistence on the issue of the error of sampling. Not having discovered the supposed cheater during their investigation, it was the only objective fact that they brought back from from their expedition. Furthermore, not explaining in detail their calculations and the

precise origin of their data and using the authority conferred by *Nature*, it was difficult for the reader of their report to question an argument presented as a theorem.

The "obviousness" of their calculation was apparently not enough for the investigators and they made a mockery of J. Benveniste. They thus noted in the report:

> "Ironically, he is himself one of the three authors of a paper published in 1981, in which such this issue had been addressed in a superficially similar situation (Petiot, J.F., Sainte-Laudy, J. and Benveniste, J. Ann. Biol. Clin. 39, 355 ; 1981) […]. That brief paper deals exclusively with the effect of sampling errors (not other kinds of errors) on the interpretation of measurements of intact basophils after white-cell suspensions had been allowed to react with allergens via their attached IgE molecules." [5]

Irony is sometimes a double-edged weapon. Indeed, it is a pity that the investigators did not read more attentively this "brief article" of Petiot *et al* that they quote with an undisguised pleasure. They would have found out the following information:

> "The experience of Gérard et al as well as our showed us that this estimator of the variance [*i.e., mean of counts of basophils*] is biased. The type-1 risk is thus reduced, certainly lower than the formulated risk, which is the risk of false positive results." [6]

The authors clearly formulated a notion which was already known in the "community" of the users of the test of degranulation, that should have alert the investigators: the variance observed for the counts of basophils is lower in practice than the value calculated with the law of small counts. This sentence did not seem to have aroused the interest of the investigators. They did not indeed hesitate to explain that "the data lack errors of the magnitude that would be expected" and that "repeat observations agree more closely than would be expected from the underlying distribution." [7].

The difference of appreciation about the importance of the sampling error according to the investigators or according to J. Benveniste is rather delicious. J. Benveniste who dismissed any "theoretical" consideration as soon as the facts contradicted it, is in the lineage of Claude Bernard for whom: "the experimental method […] is nothing else but a reasoning with which we submit methodically our ideas to the experience of the facts" or, according to another close sentence, "When we meet a fact which contradicts a prevailing theory, we must accept the fact and abandon the theory, even when the theory is supported by great names

and generally accepted".[8] We are thus in the presence of an approach, which we could define as pragmatic. Pragmatism is rather attributed to the Anglo-American researchers. Indeed, according to this cliché, the latter are not very disturbed about theories when "it works".

In contrast, the Cartesian mind, which requires a theoretical frame for any sort of observation, would be rather privilege of the French tradition. And when there is a contradiction between facts and theory – when "it works" while the theory fails to explain these facts – what makes the Cartesians? According to Descartes: "And the demonstrations are so certain that, even if experience seemed to show us the contrary, we would nevertheless be obliged to place more faith in our reason than in our senses". And, ironically, the investigators that came however across the Channel and across the Atlantic adopted an ultra-Cartesian attitude: as the facts did not fit with their expectations, then, in good Cartesians, they rejected the facts.

A sum up for Chapters 11 and 12

For the reader who skimmed through the previous two chapters, here is a summary of the scientific criticisms that could be directed to the investigators of *Nature*:

1) Poor methodological and scientific practices:

- The investigators made a *calculation error* that minimized the variance of the difference of the duplicate counts;

- Concerning these repeated measures, the investigators did not take into account the fact that the *variance of small statistical samples* (counts in duplicate or triplicate) have more often than not low values because the law of distribution of the variance is asymmetric (while respecting of course the law of small numbers on average). The consequence of this asymmetry is that the results of the counts of small series could seem "too good";

- The investigators systematically highlighted in their report the experiments that nevertheless did not achieve the *criteria of quality* and therefore must not be considered.

2) Knowledge of the research area not taken::

- It was a *well-known* (*and published*) *notion* among researchers who used the test of degranulation that the error of sampling was lower on average than calculated with the law of small numbers, in particular when the cell density was high;

- It is possible that this lower dispersion of the basophil counts could be *explained by a physical phenomenon* which would tend to make basophils repel one another due to, for example, electrostatic charges.

In conclusion, not having succeeded in exposing the one who "played a trick on J. Benveniste", *Nature*'s investigators have fallen back on technical arguments based on statistics because, according to them, the results were "too good". Nevertheless this central argument of the investigation report was also irrelevant.

Notes of end of chapter

[1] M. Schiff (Un cas de censure dans la science, p. 237) suggested that repulsive forces at long distance called forces of Frölich could play a role in the low variance of basophil counts. Such a long-distance force could indeed play a role in the interaction between red blood cells (Rowland et al. A Frölich interaction of human erythrocytes. *Physics Letters* 1981 ; 82A : 436). However, it does not seem that this type of force plays a role in the present case because as regards red blood cells, this strength is observed only if cells are alive. Indeed let us remind that the staining solution for basophils fixes cells by the ethanol.

[2] P. Alfonsi, Au nom de la science, p. 33.

[3] J. Maddox, W. Stewart, J. Randi. "High-dilution" experiments a delusion. *Nature*, July 28th, 1988, p. 287.

[4] Printer's proofs of July 25th, 1988.

[5] J. Maddox, J. Randi, W. Stewart. "High dilution" experiments a delusion. *Nature*, July 28th, 1988, p. 288.

[6] J.F. Petiot, J. Sainte-Laudy, J. Benveniste. Interprétation du résultat d'un test de dégranulation des basophiles humains. *Ann Biol Clin* 1981:39:355–359.

[7] J. Maddox, J. Randi, W. Stewart. "High dilution" experiments a delusion. *Nature*, July 28th, 1988, p. 290.

[8] C. Bernard. Introduction à l'étude de la médecine expérimentale (1865).