Contributions from the study of the history of statistics in understanding students' difficulties for the comprehension of the Variance

Michael Kourkoulos, Constantinos Tzanakis
Department of Education, University of Crete, 74100 Rethymnon, Crete, Greece
mkourk@edc.uoc.gr tzanakis@edc.uoc.gr

1. Introduction

Since the 1980's didactical studies point out that, students encounter important difficulties to understand variation and its parameters (e.g. Mevarech 1983, Loosen et al. 1985, Huck et al. 1986, Batanero et al. 1994, Shaughnessy 1992, 1999). Nevertheless, as Baker (2004a, p.16,) Reading (2004) and others remark, no much attention was given to variation in didactical research before the end of the 90's. Only recently there have been some systematic studies on the development of students' conception of variation. (e.g. Torok and Watson 2000, Watson et al. 2003, Baker 2004(b), Reading 2004, Reading and Shaughnessy 2004, Canada 2006, Garfield & Ben-Zvi 2007, especially pp.382-386).

The research work of deLmas & Liu point out that college students have important conceptual difficulties for understanding and coordinate even the simpler of the underlying foundational concepts of the standard deviation and that considerable and well organized teaching work is needed in order to ameliorate the comprehension and coordination of these concepts (deLmas & Liu 2005). For understanding the variance and the standard deviation (s.d.) students' reasoning has to correspond to the highest of the levels of the developmental hierarchy established by Reading & Shaughnessy (2004), which concerns the description of variation. This is also compatible with Mooney's corresponding developmental hierarchy (Mooney 2002, pp. 36-37)[1].

Today's students are not the only ones for whom the variance appear to be a complex and difficult notions. The historical analysis of Statistics points out that a long, multifarious and conceptually complex path had been followed before a deep understanding of variance was achieved (Stigler 1986, Porter 1986, Tzanakis & Kourkoulos 2006). Examining didactically the historical development of the concept of variance can be useful in its teaching for several reasons of a more general value, but which in the case of Statistics are especially valuable: This historical development was related to several different domains, and the students may appreciate their interrelation and that fruitful research in a scientific domain does not stand in isolation from similar activities in other domains. In addition, it is possible to identify the motivations behind the introduction of the concept of variance, through the study of examples that served as prototypes in its historical development and which may help students to understand it, when they are didactically reconstructed. In fact, history provides a vast reservoir of relevant questions, problems and expositions which may be valuable both in terms of their content and their potential to motivate, interest and engage the learner. Didactical activities designed and/or inspired by history may be used to get students involved into, hence become more aware of, the creative process of "doing mathematics". As we describe later (section 5), students may do "guided research work" in this context. Moreover, the historical analysis may help to appreciate conceptual difficulties and epistemological obstacles that are worth of more attention since they may bear some similarity with students' difficulties; hence, to provide clues for explaining some of the students' difficulties (cf. Tzanakis & Arcavi 2000, section 7.2). Such a historical

[1]Reading & Shaughnessy's hierarchy concerns the types of description and measures of variation used by the students, classified according to their cognitive complexity. Moreover a refinement of this hierarchy, based on SOLO taxonomy (Biggs & Collis 1991, Pegg 2003), is proposed by Reading (Reading 2004). Mooney's developmental hierarchy is part of a broader classification of students' statistical thinking, presented in Mooney 2002; it concerns mainly the correctness and validity of students' descriptions and measures of spread.

approach can be particularly fruitful for a complex notion, like variance, in a domain (Statistics) for which teaching large populations and didactical researches are relatively recent (Baker 2004a ch4).

This paper aims to make clear that the historical analysis of the development of basic statistical concepts related to variation, reveals the importance of physical examples in this context, implicitly suggests their possible didactical relevance and points out that examples from social sciences are definitely more complicated, hence, they should be selected and treated with adequate carefulness, especially in introductory statistics courses. Therefore, in section 2 we present some didactically relevant selected elements of the historical development of the concept of variance and in the next sections we comment on them from a didactical point of view, also using data from our previous experimental teaching work (Kourkoulos Tzanakis 2003 a,b Kourkoulos et al. 2006, Tzanakis & Kourkoulos 2006).

2. Historical aspects of the development of the statistical concept of variance

During the18th century, probabilistic thinking and the treatment of data in astronomy and geodesy followed distinct paths. The convergence and synthesis of these paths, culminating with the works of Gauss and Laplace (from 1809 to 1812), required important developments in both domains, as well as, overcoming deep conceptual barriers (Stigler 1986, part I, Kolmogorov & Yushkevich 1992, ch.4, Maistrov 1974, §§III9, III.10).

The discovery of the normal distribution by De Moivre as an approximation to the binomial distribution, Laplace's work on the approximation of probability distributions, culminating in 1810 with the proof of the central limit theorem, and his works on inverse probability and error functions -aiming at statistical inferences- (Smith 1959 pp.566-575, Laplace 1886/1812 pp.309-327, English translation in Smith 1959 pp.588-604), are key elements concerning the evolution in probabilities necessary for the convergence and synthesis mentioned above (Stigler 1986 chs.2-4).

On the other hand, the important development of methods to combine observations in the second half of the 18th century, culminating in 1805 with Legendre's publication of the least-squares method, was the essential element in the evolution of data treatment in astronomy and geodesy necessary for to the aforementioned convergence to become possible. The development of these methods was enriched by important insights in mechanics and mathematics and by extended acquaintance with the characteristics of the data under consideration. Before Gauss and Laplace's synthesis, there was no appeal to formal probability theory in developing and establishing these methods,[2] although, some limited but essential intuitive probabilistic notions were used. Since observed measures that contain random measurement errors had to be combined, it was considered reasonable to assume that (i) equilibrium centers of sets of observed measures (i.e. averages, centers of gravity) are the most likely values of the correct measures; (ii) positive errors should (most probably) compensate negative ones; and (iii) a line of best fit should minimize the total amount of (weighted, or not) errors' absolute values. These intuitive probabilistic ideas were enhanced by their compatibility with fundamental mechanical models, and scientists' acquaintance with their data characteristics; they were further established by the success of these methods in main problems of astronomy and geodesy (Stigler 1986, ch.1).

In the evolution of probability, the use of variance and standard deviation (s.d.) appears closely connected to the normal distribution. De Moivre was the first to use a parameter equal to twice the s.d. in his work of 1733, where the normal distribution appears as an approximation to the binomial distribution (Smith 1959, pp.566-575).[3] Because of this approximation, in this work he measured distances from the center of the

[2]Though important works have been done in probability concerning error functions before 1810 (Stigler 1986, ch.3; Henry 2001 pp.51-52 table 9), it had not been possible to use them in methods of treating real data in astronomy and geodesy (Stigler 1986, ch.1).

[3] This was, also, the first appearance of the normal distribution.

symmetrical binomial as multiples of $\sqrt{n}$ (where n is the total number of trials). In the 2nd edition of his *The Doctrine of Chances* (1738) he goes further and explains clearly that $\sqrt{n}$ is the unit that should be used for measuring the distances from the center of the distribution and he introduced the term *modulus*[4] for this unit $\sqrt{n}$) (Smith 1959 p.572, Stigler 1986, ch.2, particularly pp.80-85). The interest of using the s.d. (or a multiple of it) as a dispersion parameter was increased as other probability distributions were found to be approximated well by the normal distribution (mainly by Laplace; see Stigler 1986 ch.3; but Lagrange's memoir of 1776 played also a significant role, *ibid* pp.117, 118). This approach culminated with Laplace's formulation and proof of the central limit theorem in 1810 (Laplace 1898a/1810), and the Gauss-Laplace synthesis (1809-1812), which determined a very large category of probabilistic phenomena in which the natural way for measuring distances from the center is by using the s.d. (or a multiple of it) as unit of measurement. [5]

An interesting relevant parameter that enjoyed popularity during the 19th century is the "probable error", which is equal to 0,6745s.d. It was introduced by Bessel before 1820 (Stigler 1986 p.230 footnote 5), and played the role of s.d. in many works of this period. The probable error is that multiple of the s.d. that would correspond to the distance from the mean to a quartile if the distribution were normal. An interesting characteristic of the "probable error" is that, although determined by the s.d., it conserves, through the assumption of normality, a close conceptual relation with the interquartile range, which is another basic aggregate of dispersion, easier to understand than the s.d.

In the combination of observations in geodesy and astronomy, a first[6] significant use of squared distances appeared in Legendre's work of 1805[7]. In this work he also explained that his use of the sum of squared distances leads to a general method for treating problems concerning the combination of inconsistent observations (the method of least squares). Legendre used three main arguments to convince for the importance of his method: (i) the method of least squares satisfies the criterion of minimizing the total amount of weighted errors[8], a criterion then generally accepted; (ii) the solution thus found establishes "a kind of equilibrium among the errors" and reveals the center around which the results of observations arrange themselves; (iii) it is a general and easy-to-apply method. [9]

---

[4]The term *modulus* was used later by Bravais (1846) as a term for the scale parameter of a normal distribution and by Edgeworth for the square root of twice the variance; hence Edgeworth's modulus was equal to De Moivre modulus divided by $\sqrt{2}$ (Stigler 1986, p.83; Stigler 1999 p.103; Walker 1931). The term "standard deviation" was introduced by Pearson at the end of the 19th century (Porter 1986 p.13; Baker 2004, p.70; David 1995).

[5]In the 18th century works, error distributions were proposed, which used squared distances from the center of the distribution: Lambert (1765) examined the error function $\varphi(x)= 1/2\ \sqrt{(1- x^2)}$ (flattened semicircle). Lagrange in his 1776 memoir examined a family of distributions of the mean error, $\varphi(x)$, that are proportional to the quantity p^2-x^2. The error function $\varphi(x)= a^2-x^2$ is also examined by Daniel Bernoulli, in his 1778 memoir (Stigler 1986, ch.3 pp.110, 117, Baker 2004, pp.75, 76). These geometrically motivated distributions, satisfy the basic criteria requested in that period for an error distribution, namely (i) $\varphi(x)$ is symmetric with zero average; (ii) $\varphi(x)$ decreases to the right and left of the average; and (iii) $\varphi(x)=0$ beyond a certain distance from the average, or at least it is very small and tends to 0. As the examined distributions satisfied these criteria, at that period they could be considered as legitimate candidates of good error distributions. However, neither the initial works, nor later ones reveal any significant domain, or categories of practical situations in which the use of these error distributions leads to efficient treatments.

[6]There was a priority dispute between Gauss and Legendre (Stigler 1999 ch.17; Stigler 1986 145,146). Although this issue may not be entirely settled, it seems clear that both Gauss and Legendre conceived the method independently and that Legendre's significant contribution was that he realized the generality and power of the method and formulated it in a way that attracted the attention of the scientific community (Stigler 1999, p.331).

[7]*Sur la méthode des moindres carrés*, reproduced in part in Smith 1959, pp.576-579.

[8]Considering that the weighting coefficients are equal (or proportional) to the errors.

[9]Legendre remarks that there is some arbitrariness in any chosen way to let errors influence aggregate equations (Stigler 1986 p.13). This suggests that he thought that there was no absolutely indisputable reason for choosing the criterion of least squares, although after this remark, he defended firmly his method: "Of all the principles that can be

Before Legendre's least-squares method, there were other important works in the 2nd half of the 18th century on the treatment of inconsistent observations in astronomy and geodesy, where simpler measures were used for measuring deviations (errors): first order relative and absolute deviations (Boscovich's method, presented in his works of 1757, 1760, 1770), as well as, weighted deviations (Laplace's "method of situation" in 1799). An earlier, but also influential method was Mayer's method of 1750 (amended by Laplace in his work of 1787). According to this method, in situations in which more initial (linear) equations than unknowns exist, and these equations are inconsistent (because they are obtained from observed values having errors of measurement), equations were weighted in a simple way (each equation was multiplied by 1, 0 or -1) and then added, in order to obtain an aggregate equation; the final solution was found by solving a system of such aggregate equations[10] (Stigler 1986, 31-55).

These widely used methods were the conceptual background that allowed Legendre to conceive his method. Thus, the emergence of the least-squares method appears as a natural evolution of previously existing methods of data treatment, rather than as a jump, or discontinuity in their evolution due mainly to one man's genius. Within the conceptual context formed by the previously existing methods, the least squares method appears as another way of weighting errors, whose important advantages were initially supported by Legendre with theoretical and practical arguments (and later on, by the Gauss-Laplace synthesis and the accumulated experience from its use). [11]

The simplicity and generality of this method, the interest in the results of the treated examples, and Legendre's arguments and clarity of presentation were decisive for his method to attract the interest of scientists in astronomy and geodesy from the outset[12]. The method was gradually disseminated in continental Europe and England so that, by the end of 1825, it had become a standard and widely used tool

---

proposed for this purpose, I think there is none more general, more exact or easier to apply, than that which we have used in this work; it consist of making the sum of the squares of the errors *minimum*. By this method, a kind of equilibrium is established among the errors, since it prevents the extremes from dominating, is appropriate for revealing the state of the system which more nearly approach the truth." (Smith 1959 p.577). Then he explained that (i) if there is a perfect match, the method will find it and (ii) the arithmetic mean is a special case of the solutions found with this method. After that he explained that the center of gravity of several equal masses in space, as well as, the center of gravity of a solid body, are also a special case of the solutions found with the method; then, by analogy to the center of gravity he concluded: "We see, then, that the method of least squares reveals to us, in a fashion, the center aboud which all the results furnished by experiments tend to distribute themselves, in such a manner as to make their deviations from it as small as possible." (*ibid* p.579). It is explicit in these quotations that Legendre considered an analogy between the properties of the solution obtained by his method and properties of mechanical equilibrium (Stigler pp 11-15, 55-61). Conceptually, this analogy was an important convincing element, especially in this post-Newtonian era, where, for example, the basis of the theoretical framework of astronomy and geodesy was Newtonian and classical mechanics.

[10]The underlying principle of the method is that the system of aggregate equations is more stable (less sensitive to measurement errors) than the systems obtained from the initials equations, if adequate weightings of initial equations are chosen. (Thus, weighting was chosen by taking under consideration simple criteria of mechanical equilibrium, at least for the values of the more important of the involved statistical variables, as well as, other criteria specific to the examined situation). In this early method, errors' measures are not expressed in the mathematical treatment and properties of errors' distribution are not directly discussed. Annihilation of the influence of errors is realised through the use of equilibrium criteria, often of an ad hoc and context-dependent character. In fact, at the conceptual level, in this method the key issue is stability, rather than the explicit discussion and treatment of instability factors. In this respect, Boscovich method constitute a significant advance, since the measure of errors is explicit, properties of errors' are clearly expressed and constitute the key point of the whole treatment for obtaining aggregate equations.

[11]That a few years earlier (1801), Gauss used the same method for the determination of the orbit of Ceres, the first asteroid ever discovered, is an additional indication that the least-squares method was the outcome of a natural evolution of pre-existing methods of data treatment in geodesy and astronomy (Gauss 1996/1821 p.III; cf. footnote 6 and references therein).

[12]In the same year (1805) the method was presented in the *Traité de géodésie* by L. Puissant and the next year it was presented in Germany by von Lindenau in von Zach's astronomical journal (Stigler 1986 p.15).

in both disciplines, although there was some resistance and explicit objections[13].

We have noticed that in Legendre's initial work there is no explicit appeal to probability for founding the method of least squares (Stigler 1986 pp.11-15, 55-61), despite the fact that important works had already been done on error functions, inverse probability and statistical inference (Stigler 1986, ch.3). For an interpretative probabilistic framework to explain this method, two key elements were still needed: (a) Gauss' brilliant result in 1809, that the normal distribution is the adequate choice of an error function under apparently quite plausible conditions (Gauss 1996/1809, pp.65-76)[14] [15], and (b) Laplace's central limit theorem, which allowed him to provide in his works of 1810, 1811 and 1812 better explanations for Gauss' choice and to point out a large family of situations where the normal distribution was an appropriate error function[16] (Laplace 1898b/1810, 1898c/1811, 1886/1812 pp.309-327, Stigler 1986 pp.139-148).

Gauss and Laplace's works (from 1809 to 1812) constitute, both a main synthesis of the distinct evolutionary paths followed in the 18th century in probabilities and in the treatment of data in astronomy and geodesy, and the next step in understanding the significance of the sum of squares distances, hence of variance.

Concerning the use of the sums of absolute deviations and the sums of squared deviations it is worth noting the following from the works of Laplace and Gauss:

In his works of 1810, 1811, 1812, Laplace considered that the basic criterion for selecting a best estimate value for an unknown parameter sought is: to chose as an estimate that value which minimize the posterior expected absolute error (that he called "l'erreur moyen à craindre" - "the mean error to be feared"); and it seems to have no doubt about his criterion[17]. n these works, he explained that in the examined cases, involving normal distribution, the least square method leads to the same estimate value as his criterion. In the work of 1810 he adds explicitly that in the examined cases this estimate is also the "most probable" (it corresponds to the mode of the posterior distribution), and thus it also satisfies this criterion of choosing a best estimate, which was used by Daniel Bernoulli, Euler and Gauss; however he continues that in the general case his criterion is more appropriate (Laplace 1898b/1810 p 352).

Gauss in his work of 1821 agrees with Laplace's idea that the best estimate should be the one which

---

[13]The Mayer - Laplace method is simpler and demands much less labour than the least-squares method, therefore it enjoyed popularity until the mid 19th century, even though it is less accurate (Stigler 1986, p. 38-39).
As late as 1832, Bowditch was recommending Boscovich's method, which involves first-order relative and absolute deviations, over least squares, because it attributes less weight to defective observations (Stigler 1986, p.55).

[14] Gauss relied on the widely spread idea at that time that the arithmetic mean was a very good way for combining observations' results. He admitted as an axiom that the most probable value of a single unknown quantity observed several times under the same circumstances is the arithmetic mean of the observations, and he proved that, if so, the probability law of the errors of observations has to be a normal distribution. He proved then that in the more general case, this errors' distribution leads to the method of least squares as the method that provides the most probable estimates of the sought parameters (Stigler 1986 pp.140-143).

[15]In 1808 R. Adrain has also obtained the normal distribution as an appropriate error function, but his work went largely unnoticed (Maistrov 1974, pp.149-150, Stigler 1978).

[16]E.g. in his work of 1810 (Laplace, 1898b/1810) he explained that when the measurements' errors are aggregates (e.g. sums, or averages) of a large number of commonly distributed elementary errors, the normal distribution approximates their distribution via the central limit theorem, and strengthened the conclusion by proving that in this case, the solution provided by the least-squares method was not only the most probable, but also the most accurate one, (in the sense that it minimizes the posterior expected error; Stigler 1986 pp.143-146, 201-202; Maistrov 1974 p.147; Kolmogorov & Yushkevich 1992, p.225).

[17]For example in his work of 1810 he writes « Pour déterminer le point de l'axe des abscisses où l'on doit fixer le milieu entre les résultats des observations *n, n′, n″,*...nous observerons que ce point est celui où l'écart de la vérité est un minimum ; or, de même que, dans la théorie de probabilités, on évalue la perte à craindre en multipliant chaque perte que l'on peut éprouver par sa probabilité, et en faisant une somme de tous ces produits, de même on aura la valeur de l'écart à craindre en multipliant chaque écart de la vérité, ou chaque erreur, *abstraction faite du signe*, par sa probabilité, et en faisant une somme de tous ces produits. » (our emphasis; Laplace 1898b/1810 p. 351).

minimizes "l'erreur moyen à craindre" (Gauss 1996/1821 pp.11-13), however it determines differently "l'erreur moyen à craindre (*m*)": for Gauss, *m* should be the square root of the expected squared error ($m^2 = \int_{-\infty}^{+\infty} \varphi(x)x^2dx$, x being an error and $\varphi(x)$ the "the relative facility of error x"- in modern terminology, the probability density distribution of error x to occur).[18] This difference in the definition of "the mean error to be feared" in fact changed the criterion for the best estimate value.

Gauss discussed his choice in some length for "the mean error to be feared" (ibid p 12):

- Initially he admits that there is an element of arbitrariness in the determination of "the mean error to be feared". ("Si l'on objecte que cette convention est arbitraire et ne semple pas nécessaire nous en convenons volontiers. La question qui nous occupe a, dans sa nature même, quelque chose de vague et ne peut être bien précisée que par un principe jusqu'à un point arbitraire." (ibid p 12))
- Then he establishes an analogy, pretty much like Laplace (footnote 17), between the determination of a quantity through observations and a game of chance where there is "a loss to fear and no gain to expect". In the context of this analogy, each error, positive or negative, corresponds to a loss of the truth and thus the expected loss is the sum of products of possible losses multiplied by their respective probability to occur. But what loss corresponds to each error? According Gauss, it is this point, which is unclear and needs to be settled by a partially arbitrary convention. ("Mais quelle perte doit on assimiler à une erreur déterminée? C'est ce qui n'est pas clair en soi; cette détermination dépend en partie de notre volonté.", *ibid* p.12). The only restriction he considered is that each positive or negative error, should correspond to a loss and not to a gain, and he concludes that among all functions that satisfy this condition "it seems natural to choose the simplest, which is doubtlessly, the square of the error" ("il semble naturel de choisir la plus simple, qui est sans contredit, le carré de l'erreur"); *ibid* p 12.
- Then Gauss considers Laplace's choice for that function: the errors' absolute value ("l'erreur elle même prise positivement" - the error itself taken positively). He considered that Laplace's choice was equally arbitrary to his own, but admitted that it was also equally legitimate, and concluded that his choice was recommended because of the "generality and the simplicity of its consequences".
-What he meant by the "generality and the simplicity of its consequences" becomes clear later in the text. Based on probabilistic arguments, Gauss concluded that if his definition of "the mean error to be feared" (and thus his criterion of best estimate) is accepted, then the method of least squares provides "les combinaisons le plus avantageuses des observations" (the most advantageous combinations of observations) even in the cases that there is a small number of observations and errors' probability law is any such law, and not necessarily a normal distribution (*ibid* pp. 21-26). So he extended the cases for which the method of least squares can be considered as preferable beyond those in which the normal distribution is involved, directly or through the central limit theorem. This generalised result offers a unified solution, and thus simplifies the whole problem of the treatment of observations, if his criterion is accepted; this result is known as the Gauss-Markov theorem (Stigler 1986, p 148).

Gauss argumentation, that the square of errors is a simple function and thus adequate to be used to define "the mean error to be feared", may be a convincing argumentation for an educated audience, but that it is simpler than the absolute value of the error it is much less convincing; for example, Laplace thought differently on this issue (see also footnote 13). Most likely, it was the aposteriori appreciation of the expected squared error (the "generality and the simplicity of its consequences") that mainly convinced Gauss to prefer the use of squared errors than the use of absolute errors.

[18]He writes "Nous ne limitons pas, du reste, cette dénomination [l'erreur moyen à craindre] au résultat immédiat des observations, mais nous l'entendons, au contraire, à toute grandeur qui peut s'en déduire d'un manière quelconque" (*ibid* p13). So, "l'erreur moyen à craindre", thus defined, is used not only for the errors of directs observations but also for aggregate errors as well as posterior errors' distributions.

The theory of errors and the techniques of analyzing data for the treatment of problems in geodesy and astronomy, developed in the 18th century, were not a sufficient basis for the elaboration of adequate tools for the statistical treatment of problems in social sciences. A difficult and laborious evolution for almost a century, and overcoming important conceptual barriers were necessary for the elaboration of such tools (Stigler 1986 ch 8-10, Porter 1986, part 3, Stigler 1999 part I). Galton's work on heredity and the related conceptual framework on linear regression that he established[19] constituted a major breakthrough[20] that opened the way to the works of Edgeworth, Pearson and Yule, who elaborated adequate conceptual frameworks and the first efficient tools for the statistical treatment of problems in social sciences. It is characteristic of the importance of the conceptual difficulties of the statistical treatment of such problems that only in 1897, on the basis of theoretical arguments, Yule proposed a generalised method[21] of linear regression for problems in social sciences based on the use of least squares.

Three interrelated issues were the main reasons for the conceptual difficulties encountered in the statistical treatment of such problems:

(i) Social phenomena are susceptible to important influences by a very large number of factors; hence, the initial classification of social data becomes a main issue[22].

(ii) Unlike geodesy or astronomy, there were no theories of social phenomena incorporating coherently and efficiently all (or most of) the influencing factors, that would permit to determine a priori which of them are important and which have only a secondary influence that could be fairly assimilated to random disturbances. Hence, it was not possible to compare the results of data statistical analysis with relatively reliable a priori expectations.

(iii) Unlike the phenomena in many areas of physics, or psychometrics, social phenomena are very

[19]The full presentation of Galton's development of the concept of regression is given in his 1889 book *Natural Inheritance*. However this work is based on a series of previous works dating from 1874, or even earlier. The main part of this development was achieved by 1886 (Stigler 1986 ch 8; 1999 ch.9).

[20]Stigler (1986 p.281) qualified Galton's work as "…perhaps the single major breakthrough in statistics in the second half of the 19th century" (cf. Stigler 1999 p.176).

[21]Yule's method can also be applied to bivariate and multivariate distributions that are not normal. Until that time (1897), Edgeworth and Pearson's significant works in two or more dimensions, concerned only distributions that fit, at least approximately, bivariate or multivariate normal distributions (prior to 1897, Pearson had done important work on distributions that deviate significantly from normality, but his work was restricted to one dimension only concerning what was called "skew curves"; Stigler 1986, chs 9, 10).

[22]The large number of these factors made impossible to use all of them in the initial classifications necessary to the collection and the statistical data analysis of social phenomena. Therefore, the selection of a relatively small number of factors to be used in the initial classifications was considered to be a main issue, since it could greatly influence the conclusions of the statistical analysis. In 1827, Baron de Keverberg argued that since a very large number of factors are expected to have an important influence on the ratio of population to births in the Low Countries, and since a classification based on all these factors is impossible to be used for an adequate sampling process, any inference based on such data is untrustworthy. Moreover, Quetelet's results of his first crude efforts to use incomplete samples for inference purposes seemed to corroborate Keverberg's critique. These results combined with this critique convinced Quetelet that only complete census and large amounts of data allow for reliable conclusions (Stigler 1986, pp.163-169). On the other hand, in 1843, A.A. Cournot, argued that using probabilities is irrelevant to the treatment of social data. The key point of his reasoning was that: Following the factors that a priori can influence the examined social phenomenon, a large number of subdivisions of the data is relevant. Because of the importance of this number, it is very likely that in one or more of these subdivisions chance alone will produce important differences between the observed and the a-priori ratios (probabilities). So, in the limited number of subdivisions that will be effectively examined, we have not any objective criterion to decide which of the differences indicated as significant by the calculus of probabilities are really due to chance alone (Stigler 1986, pp.195-201). Keverberg and Cournot's reserves were shared by many contemporary social scientists and the whole of their arguments constituted a main problem, the solution of which did not come early; in fact, we cannot say that it is completely settled even today (Stigler 1986, pp. 200, 359-361).

rarely subject to experimental treatment; hence, in most cases it is not possible to vary some factors, while keeping the others constant in order to evaluate their importance and influence. This was not only a difficulty for the creation and evolution of social theories, but also an additional important difficulty for elaborating and evaluating statistical methods for treating social data.

Related to these issues is the fact that, the aggregates of central tendency and variation of social data have not the status of approximations to measures of "real objects" of central importance in the situation under consideration, as in the case of geodesy, or astronomy; such aggregates often represent only data tendencies[23]. Therefore, their good understanding required at least the understanding of the basic characteristics of this type of data, which required a wide experience of data treatment and analysis[24]. But this analysis, in turn, necessitated also the use of statistical aggregates, whose meaning is well understood. In fact, what took place was a slow and difficult interactive evolution (where an improved understanding of data's characteristics permitted an improved understanding of the existing aggregates, as well as, the introduction of new aggregates (e.g. correlation coefficients), these improvements allowed for a better understanding of data characteristics, and so on).

Critical information for the fundamental significance of variance came also from another domain, namely, physics. In the 19th century, the parallel development of the kinetic theory of gases and statistical mechanics through the work of Maxwell, Boltzmann, Gibbs and others gave further insights into the importance of variance of the molecular distribution of velocity and provided its physical interpretation as the macroscopic temperature of a system of molecules. (Tzanakis & Kourkoulos 2006, pp.289-290; Sklar 1993 §§2.II, 2.III, Brush 1983 §§1.11, 1.13) [25]. More specifically, Maxwell derived the normal distribution as the distribution of velocities in an ideal gas, in analogy to Gauss' derivation of the law of errors and

[23]E.g. a mean can approximate the unknown position of a celestial body, a regression line can be an approximation to the trajectory of such a body, and the variance can be a measure of the inaccuracy of observations. (Of course, often the approximated objects (e.g. the trajectory of a celestial body) have the status of real objects within a specific theoretical framework (for example, Newtonian theory). Changing the theoretical framework can provoke changes of what "real objects" are. Empirical, statistical treatments can also provoke changes of what is considered to be real object. An interesting example is Laplace's initial consideration within Newtonian mechanics, of the shape of the earth as an ellipsoid, and his reconsideration of that shape when he found that the statistical line of best fit still had important deviations from the observed data. This reconsideration had led him to revise also what it was initially considered as measurements' errors (Stigler 1986, pp.50-55, particularly p.53).

On the contrary, for social phenomena, it was not easy to find (or even to construct theoretically) "real objects" that correspond to statistical aggregates of central importance. A characteristic example is Quetelet's conception of "average man", which can be considered as an effort to conceive a palpable social analogue of the statistical centre of gravity for multidimensional social data. It was highly criticized and finally rejected, at least by the scientific community, but still, as Stigler remarks because of its simplicity and apparent clarity it remains very alive in popular use (Stigler 1986, pp.169-174, 201; Porter 1986, pp.171-172, 175, 182, 188).

Furthermore, for simpler statistical aggregates, like the mean of the distribution of one variable, real elements having a value close to the value of the aggregate may exist (e.g. one, or more individuals in a population may have an income very close to its mean value for that population). However, the role and importance of these real elements is neither comparable to the role of the aggregate in the situation described by the distribution, nor to the aforementioned role of real objects in astronomy and geodesy. Therefore, these real elements are inadequate for interpreting the examined aggregate.

[24]The existing statistical methods developed in astronomy and geodesy, and even direct analogies with the treatment of data analysis in these domains, was a very important help. Nevertheless, the insufficiency of possibilities for extrinsic evaluation and interpretation based on social theory, or experimental design, raised the issue of the evaluation and interpretation based on the intrinsic treatment of social data, as main elements for understanding its characteristics and structure. (Stigler 1986 pp 199-201, 221-223, 358-361)

[25]In addition, the first general proof of the (weak) Law of Large Numbers by Chebyshev (in 1866; Maistrov 1974, pp.198-201), using his inequality, was a key result in which variance was explicitly used (the inequality had been published earlier, in 1853 by Bienaymé in a paper on the least squares method; *ibid* p.201). At the same time, it was a probabilistic proposition that did not require exact knowledge of the distribution; by using the variance, it provided general conditions for the stability of averages, hence the regularity of randomness (Kolmogorov & Yushkevich 1992, p.259; Tzanakis & Kourkoulos 2006, pp.288).

influenced by the ideas of Quetelet as presented by J. Herschel. This made clear that the variance of the microscopic velocity distribution is proportional to the mean kinetic energy of the microscopic particles. Since, on the basis of previous more rough statistical ideas of Waterston, Krönig and Clausius, it was known that this energy was proportional to the product of the volume of the gas and its pressure, which in turn was proportional to the (absolute) temperature of the gas by the ideal gas law, variance acquired a direct macroscopic physical interpretation as being proportional to the (absolute) temperature of the gas (Porter, 1986, ch.5 particularly pp.118-119; Brush 1983, p.59; Sklar,1993, §2.II.2). In fact, it was exactly because of the physical meaning of velocity, that Maxwell's derivation makes no use of Gauss' assumption that the mean and the most probable values of the sought distribution coincide (Jeans, 1954/1904, 55-57). More generally, there was a close interplay between ideas in physics and the social sciences that influenced the conception and development of statistical methods for both (Porter 1986 chs.5, 7). For example, it is not accidental that Quetelet was educated in mathematics and astronomy and Lexis in mechanics, or that Maxwell and Boltzmann were influenced in their ideas in statistical physics by similar ideas in the social sciences (Porter 1986 pp.42ff, 242, 118-119, 208).

3. Didactical remarks

3.1. The historical elements presented in section 2 reveal that the variance is a multifarious notion of great conceptual complexity. This is compatible with the results of didactical research, which, although at an initial stage, points out that variance is a complex and difficult concept for students (Mevarech 1983, Loosen et al 1985, Batanero et al 1994, Reading & Shaughnessy 2004, Reading 2004, deLmas & Liu 2005). In fact, the examination of the complexity and multifarious character of this concept suggests that the expectations from introductory courses in statistics and probability should be restricted: a relatively complete understanding of variance is not a realistic task for first courses in statistics and /or probabilities, but rather a task of a longer and more complete statistical education. This raises the question what can be taught of variance in introductory courses in statistics and probability. The remarks on this and the next sections concern this question for undergraduate students.[26]

3.2. The good computational properties and the resulting (relative) easiness of treating the sum of squared distances from a center, or from a regression line, were one of the main reasons that, historically, contributed to gradually impose these sums as a principal tool to measure and treat variation, both in statistics and probability. The importance of these properties is more pronounced in the case of bivariate (or multivariate) distributions, but is also apparent for distributions of one variable, and their understanding is a basic element for understanding variance. Many of these properties can initially be taught in an introductory course on descriptive statistics, or on Exploratory Data Analysis (EDA)[27]. Of course, these properties must be reconsidered in the context of probability theory and inferential statistics but introducing them initially in a context with no formal reference to probability can facilitate their understanding, since it is an approach, which does not accumulate their mathematical difficulties with conceptual difficulties inherent in probability theory.

It is worth noting that an important preparatory work can be done independently of statistics courses, in domains where appropriate activities can be realised concerning the properties of the sums of squared

[26]Most of the remarks done in this section are valid for introductory courses to upper high school students, as well.

[27] EDA is a relatively new approach of data analysis in which the focus is on meaningful investigations of data sets, using multiple representations and graphing techniques and just a bit of probability theory, or inferential statistics. (Tukey 1977). EDA, or its variations, was adopted by several statistics educators and researchers on statistics education, because of the possibilities that it offers for elaborating an introductory teaching of statistics with more data analysis and less theory and recipes. EDA was considered as an opportunity to broaden the context of descriptive statistics, to give students a richer and more genuine experience in meaningful contexts of what statistics is really about (Shaughnessy et al. 1996, Ben-Zvi & Arcavi 2001, Jones et al. 2001, Baker 2004 p.11).

distances, such as equilibrium, motion of systems of (microscopic or macroscopic) bodies, collision and explosion problems in physics and the associated momentum and energy calculations, extrema problems, geometrical problems etc.

3.3. It is possible to teach variance in introductory courses on probability and descriptive statistics, or EDA. The historical analysis points out that these two possibilities offer significantly different insights to this concept and indicate that their subsequent synthesis can be very fruitful.[28]

Concerning the **introductory teaching on probabilities**, Baker formulates the hypothesis that the basic idea of the interquartile range is conceptually simpler than those of variance and s.d. (Baker 2004a pp. 69, 70).[29] [30] Our previous research, with students of Sciences of Education (Tzanakis& Kourkoulos 2006, §3.3) points out that many undergraduate students, with little knowledge of statistics and probabilities, consider that the inerquartile range is also a clearer and more interesting parameter than variance and s.d. Their main argument for this evaluation is that the interquartile range (combined with a quartile value) allows to know where precisely lies the middle half of the probability distribution, whereas, they believe that, variance (and the average absolute deviation) does not give any precise information on where some part of the probability distribution lies.

Historically, the first appearance of *modulus* (see footnote 4) is tightly related to the discovery of the normal distribution as an approximation of the binomial distribution; moreover the dissemination of the use of (multiples of) s.d. as dispersion parameters is related to the large number of cases where, because of the central limit theorem, the normal distribution is an adequate approximation of the unknown distribution. In these cases, using (a multiple of) s.d. as a measure of the distance of values from the mean allows for finding approximately the probability density of a given value, and by integration, the probability of the variable between the mean and that value. This was a main reason for imposing the s.d. as a principal parameter of variation.

There is an interesting similarity between the historical reasons of widely using s.d and variance and students' reserves about the s.d and variance. In both cases, the possibilities offered by a dispersion parameter to determine the relation between a range of values and its corresponding probability, are considered to be an important element for evaluating the efficiency and interest of this parameter. But of course, students' limited knowledge leads them to different evaluations than those established in history.

Teaching should take into consideration this point and provide some adequate explanations, especially in the usual case, where the variance and s.d. are introduced long before the introduction of the central limit theorem. Otherwise, it is possible that, for a long time, students will have the feeling that the s.d. and its use are unjustifiably imposed by the teacher, while, there are other, more efficient dispersion parameters with clearer meaning, or that s.d. is another probabilistic object whose reasons of existence remain obscure. Chebychev's inequality can be taught soon after the introduction of variance and the s.d. and could offer interesting insights to the students: it allows for a quick, albeit partial, answer to students' reserves about the s.d., showing that the s.d. gives information on the corresponding probability of a range of values centered on the mean, when the radius of this range is greater than the s.d. (Tzanakis& Kourkoulos 2006, p.292). Of course, this information is not as precise as the information deduced from the s.d. in cases the normal distribution can be used as an approximation, but it can be applied more generally

[28]Teaching variance in probability courses only, or trying to avoid the use of probabilities for very long, are, of course, approaches that put severe restrictions on the goal of a rather complete understanding of this concept.

[29]This does not mean that the procedure of calculating the interquartile range is simpler than those of calculating the variance (see Schuyten 1991, Batanero et al 1994) or that interquartile range have better computational properties.

[30]The "probable error", presented on p. 3, is an historical aggregate of variation relevant to this subject: it may be considered as expressing a tendency to combine the clarity and simplicity of interquartile range with the efficiency of standard deviation.

and leads to the proof of the (weak) law of large numbers.

4. Introductory statistics courses and data related to social phenomena

4.1. It is interesting to remark that often, in introductory statistics courses, the examples with an extra-mathematical content (i.e. that are not pure numerical) are mainly (or almost exclusively) examples referring to data related to social phenomena (e.g. students' weights, notes, diseases, income etc) [31].

However, the historical analysis points out that the adequate collection and treatment of such data was, in most of cases, a very complex and difficult task. Are the reasons for the difficulties encountered in the historical evolution obsolete, or irrelevant to modern students' learning? At least for the main reasons this is not so (see section 2 pp.7-8). In addition, it is not true that methods have been found to face and easily overcome these sources of difficulties in actual introductory statistics courses. In fact, in the usual initial teaching of statistics the difficulties of collecting and treating data concerning social phenomena are underestimated and often hidden.[32]

4.2 A characteristic example on this subject, concerns the initial design of data collection. In usual introductory statistics courses to students of education and social sciences that we have observed, we have noticed the following behavior of teachers: After the selection of the phenomenon to be examined (e.g. students' weight), the teacher initiates a discussion on the relevant factors to be examined and on the questions to pose. When some factors have been identified (e.g. gender, age, parents' corresponding values) and agreement on their importance has been achieved, the teacher stops the discussion, with no reference to the fact that a large number of other factors are relevant and may have important influence on the examined phenomenon. In other cases, of classroom discussions, students have started to propose a

---

[31] For example, this is so for the introductory course in descriptive statistics in the last-year of Greek General High-School. This course is part of a one-year compulsory mathematics course and lasts for about one third of the teaching time; namely, 24 teaching hours (another 6 hours are used to introduce simple probabilistic concepts and the rest two thirds are devoted to introductory calculus). In the official textbook (in primary and secondary education in Greece there is an official textbook for each course, published by the ministry of education and used in all schools) there are 37 solved examples and applications and 59 unsolved exercises in statistics which are taught (Mathematics and Elements of Statistics, C' class of High School, O.E.D.B., 1999, pp55-104, 126-128)] among these:
(i) 6 examples (exm) and 10 exercises (exr) are purely numerical.
(ii) 11exm & 26 exr concern social phenomena
(iii) 18 exm & 14 exr concern biological phenomena for human populations, also influenced by social factors (e.g. disease or death phenomena, human height distributions).
(iv) 1 exm & 2 exr concern socio-geographical phenomena.
(v) 1 exm & 4 exr have multiple questions that belong to the categories (ii) and (iii).
(vi) 2 exr concern situations that are good approximations of outcomes of simple random experiments (e.g. an urn model), 1exercise concerns a quality control.
There is no exercise referring to physics or geometry, or to natural phenomena that don't involve directly a human activity (for other examples on this subject see footnote 35).

[32] The much extended use in introductory statistics courses of situations and examples related to social phenomena, can be explained by two reasons: (i) the teachers' and curricula designers' confidence on students' feeling of familiarity with these phenomena; (ii) the interest that the study of such phenomena has for the students, teachers, curricula designers and the educational community in general. Usually, familiarity and previous experience with the situations under consideration facilitate their treatment. However, students' feeling of familiarity with social phenomena is due to their familiarity with only some aspects of these phenomena; they are unfamiliar with many more aspects of the same phenomena, as well as, with their real complexity. (Besides, this complexity is not the only source of difficulties; the absence of a theory allowing for reliable a-priori expectations and the lack of possibilities for experimentations are also main sources of difficulties). Hence, students' feeling of familiarity, concerning the easiness of treatment of situations related to social phenomena is misleading. However, studying situations related to social phenomena is indeed an important issue for statistical education. Therefore, the questions relevant to an introductory teaching of statistics are of the kind: To what extent situations referring to such phenomena should be included in this teaching? For example, is it efficient to almost exclude all others types of situations, as it is often the case? What are the difficulties inherent to the treatment of this type of situations? What other types of situations are necessary, or useful to include in introductory courses and for what purposes?

multitude of other factors as relevant (grandparents' corresponding values, nutrition, place of birth etc.). After a while, the teacher intervenes, accepting that there is a large number of such factors, but explaining that they should be limited to a small number only, because otherwise, it would be very difficult to collect, organize and treat the relevant data. In these cases, the basic point that a large number of factors influence phenomena related to the society, is recognized as an element of difficulty, but that's all. None of the questions that naturally emerge concerning this point is discussed; e.g. what are the interest and the validity of a data collection where factors that may have an important influence on the examined phenomenon are ignored? What is expected to be, at least qualitatively, the influence of the large number of significant factors ignored, on the relations among the factors selected to be examined?

This type of teaching approach can lead to students' misunderstanding on the real complexity of the relations among the factors influencing social phenomena, on the role of factors not taken under consideration, and finally, on the meaning of the relations among the factors selected to be examined. Moreover, it can lead to several misunderstandings concerning valid ways to collect data, especially, for sample data supposed to be representative of the characteristics of a larger population.

4.3 A known conceptual difficulty of many students, novice in the study of statistics, is that they tend to underestimate the relation between two factors, or even to deny its existence, if it is not deterministic one (Rubin and Rosbery 1990, Shaughnessy J.M.,1992, Batanero et al., 1994, Noss R. et al., 1999). This difficulty is reinforced when there is a large variation that appears in the graphical plan representation of the corresponding bivariate data. When the two factors are related to social phenomena it is often the case that a large variation appears because of the influence of a large number of other factors, even if the existing relation is relatively strong compared to the relations of other relevant factors (e.g. blood pressure and age of patients). In such cases, not only novice students, but also professionals with limited statistical education have difficulties to understand, or recognize the relation in the presented data, although they know about it by professional experience[33] (e.g. nurses). Because of these difficulties, in many cases, the presentation of the bivariate data functions almost inversely, urging subjects to doubt about their previous opinion concerning the strength of the relation, although this opinion was formed on the basis of their professional experience[34].

4.4 In most cases, statistical aggregates concerning social phenomena have only the status of elements representing data tendencies, whereas, for physical or geometrical phenomena, often, measurements have also the status of approximating measures of **real** objects, which play an important role in the examined phenomena. During the historical evolution, this characteristic was one of the reasons that had made difficult to understand the meaning and the role of statistical aggregates of data concerning social phenomena (see last part of section 2).

Didactical research points out that conceiving statistical aggregates as representatives of data tendencies is one of their most difficult aspects to be understood by the students (Shaughnessy1992, Batanero et al

[33] For the study of an interesting example involving blood pressure, age of patients and nurses understanding see Noss et al. 1999

[34]The points discussed in §§4.2, 4.3 are interrelated. A better approach than that presented in §4.2 concerning the initial design of data collection, could make students aware of the phenomenon that in situations, where there exists a large number of influencing factors, even if, prior to data examination they have strong reasons, to believe that two factors are related, it is possible, and not rare, that a large variation appears in the corresponding bivariate data. This variation can be the result of the influence of other factors and doesn't necessarily signify the absence of statistical relation between the two factors. On the contrary, in the approaches mentioned in §4.2, the easiness by which the possible existence of many influencing factors is not taken under consideration in the initial design of data collection, and the lack of discussion on this subject, can give to the students the impression that the factors ignored do not influence the relations among the factors selected to be examined. This enhances the possibility that students consider a large variation appearing in the bivariate data of a pair of the examined factors, as some kind of chaotic random disturbance, indicating the absence of relation between the examined pair of factors.

1994, Mokros & Russell 1995, Noss et al. 1999). These research results support the hypothesis that, in introductory courses, where only examples of data referring to social phenomena are used, understanding statistical aggregates will be more difficult for the students, than in courses where a significant part of the examples that are used refer to adequate physical, or geometrical phenomena.

For aggregates of central tendency, in particular the arithmetic and the weighted mean, undergraduates and last-year high-school students have significant experience from other domains (everyday life, school life etc); this is a factor that weakens their difficulties to understand, these concepts, at least for the initial level of their understanding. However, students have much less experience of aggregates of variation, so concerning these aggregates the choice of an adequate set of examples to elaborate on is even more important for introductory statistics courses; this is particularly true for the variance, one of the most complex aggregates of variation.

5. Comments on experimental teaching concerning variation parameters

In this section we comment on a teaching of variation aggregates, in particular of variance, based on two experimental courses on descriptive statistics with two groups of students of the Department of Education of the University of Crete (prospective primary school teachers). For each group the corresponding experimental course was their first undergraduate statistics course (Kourkoulos & Tzanakis 2003a, b).

Students worked in small groups (of 3-5 students each in the first course, and 3-4 students each in the second). In these courses the approach of guided research work was used, so that students had to investigate the properties of basic statistical aggregates of central tendency and of variation. In the version of the guided research work that was employed, students had not only to treat problems posed by the teacher, but also to be involved in forming the research questions, and gradually posing their own research questions and problems (closed and open questions, conjectures etc). Posing such questions and problems gradually became an essential part of their work. We should also notice that, in these courses after a (more or less) long period of labor, students' work created a network of questions and problems, which had important implications in all aspects of their research work; in particular, it influenced considerably the metacognitive aspects of their way to conceive and re-conceive the results of their work. In addition, this network was an important factor that often determined the subjects treated in the course. In this context, with the exception of certain examples initially provided by the teacher, students had to find examples of data necessary in their investigations on the properties of statistical aggregates. Often they had also to change existing sets of data, or to produce their own examples of such sets (e.g. when they were exploring the influence on statistical aggregates of different types of changes in the data). They had looked for examples of data in different textbooks on introductory statistics[35] and the Internet. Besides the purely

[35]In the first course, the students looked for examples of data in 9 different textbooks and in the second, they looked for such examples in 7 textbooks: 2 high school textbooks, 4 (3 in the second course) addressed to students of social sciences and 3 (2 in the second course) addressed to engineering students.

One of the 2 high school textbooks was the one mentioned in footnote 31. The other high-school textbook and the textbooks addressed to students of social sciences were close to this one concerning the characteristics of the phenomena to which referred the used data examples. Among the treated examples and the exercises included in these books: 70%-85%, following the book, are related to social phenomena (of the same kind as those presented in the categories ii-iv of footnote 31), 13%-24% refers to purely numerical situations, 2%-7% refers to other types of situations (simple random experiments, quality controls etc), no one included examples or exercises from Physics, or Geometry.

Compared to them, the 3 books addressed to engineering students have some significant differences: Among the treated examples and the exercises presented: 33%-46%, following the book, are related to social phenomena, 28%-36% refers to purely numerical situations, 16%-23% refer to quality control, 5%-10% refer to situations that are (or approximate well) outcomes of simple random experiments (urns, dies, distribution of vowels and consonants in a book etc), 2%- 5% refer to other types of situations; only in one of the 3 books there was one exercise on a physical model (a bar along

numerical examples, the large majority of examples that our students found in these resources were related to social phenomena (mainly of the same types as those present in the categories ii-iv in footnote 31); they also found geographical and weather data. Nevertheless, they had also encountered some examples of data on simpler phenomena; urn models, quality controls, measurements of simple physical objects. Only one example was found (in a textbook for engineers) referring to a physical model, namely a bar, along which weights were attached and the equilibrium point was examined.

However, among the examples of data that students found, those they chose to elaborate, as well as, the examples they produced by themselves, refer to situations of every day life and to educational phenomena, with which they felt familiar (weights and heights of human populations, students' notes, incomes, weddings data etc). The feeling of familiarity as the dominant selection criterion, led students to confine the collection of the examples chosen for elaboration, to a subcategory of situations related to social phenomena, and not to chose simpler ones. This choice, however, did not facilitate their understanding of different aspects of statistical aggregates, e.g. to understand the mean as an equilibrium point (instead, as we shall see in section 6, physical examples like a bar with attached weights on it[36], or a model with springs, can offer important insights to understand this aspect).

Nevertheless, even working with this restricted set of examples, our students succeeded to examine, in a relatively satisfactory depth, basic properties of aggregates of central tendency (mode, median and mean) and the simplest of aggregates of variation (the range, interquartile range and mean absolute deviation[37]). To this end, however, they had to do considerable experimental work and in some cases to reduce the treated situations: to the corresponding purely numerical situations and/or to the corresponding graphical representations[38]. This was due to the fact that in these cases the examined properties were better understood if considered in this way, than in the context of the concrete situations to which the examples of data referred. The analysis of their behavior and difficulties point out that a different collection of the treated examples, and in particular the use of adequate physical models could significantly facilitate their work (Kourkoulos & Tzanakis 2003 a, b, Tzanakis & Kourkoulos 2006).

The difficulties for understanding the variance were more prominent:

Right from the start, many students were reserved for its introduction and had difficulties to understand

---

which weights were attached) and one exercise that concerns measurement errors (repeated measurements of the weight of a bottle).

[36]In their investigations, our students did not choose spontaneously to use examples that refer to the model of the bar with the attached weights. Nevertheless, on teacher's incitement and insistence, students in the second course used some examples of this type.

[37] Initially it was introduced in the teaching activities the mean absolute deviation from the mean, later on students examined the mean absolute deviation of a statistical variable X from a value x_0 and found that the mean absolute deviation become minimum when this value is the median of X. However when they used the term mean absolute deviation without further specification they continue to signify the mean absolute deviation from the mean.

[38]Graphical representations of statistical distributions are basically geometrical. Although they are often approximate and do not follow in a relatively rigorous way, the representation rules that students are used to employ in secondary education for functions in algebra, analysis, or physics, still, they conserve operational properties and characteristics that often can help students very much to understand the properties of these statistical aggregates. Nevertheless, these representations and their use are not themselves free of difficulties and subtleties. Particularly, the deviations between the ways of representing statistical frequency distributions and the way students are used to represent graphically functions in algebra, analysis or physics, can activate important epistemological obstacles. It gets more difficult for the students to overcome these obstacles when teaching does not elaborate explicitly on these deviations, as it is often the case.

In the first course, students were weak in algebra. For them, the use of graphical representations played an even more important role; besides the positive role in understanding properties of the aforementioned statistical aggregates, they helped in generalizing the expression of these properties. Moreover, elaborating on graphical representations often allowed students to find geometrical arguments, explaining the general validity of the observed properties (Kourkoulos & Tzanakis 2003a §§3.1-3.3, pp.7-14).

its meaning. A main reason for these difficulties was that in all situations that students have used in order to obtain examples of distributions for elaboration, the variance had an unclear or even problematic meaning; in most cases, the sums of squares were dimensionally meaningless (squares of weights, squares of grades, squares of money etc). In fact, with the exception of purely numerical examples, the only cases that the sums of squares made sense dimensionally for the students were those involving distributions of lengths (e.g. students' height, travel distances). However, even in these cases, the squares of lengths and theirs sums had unclear meaning in the context of the corresponding situations. Students' expressions are eloquent on this point:

"The squares of the heights of the students… I understand the height, or the weight of a student, but what does it mean the square of a student's height?…I have never heard talking about the square of the height of people."

Referring to a distribution of distances of bus trips, another student remarks:

"I don't understand why to use this sum of squares to measure the dispersion…I see that these squares of distances are certain areas, *but what they have to do with the trips, or the buses? …I mean, there are many quantities really related to the distances of the bus trips, the time of the travel, the fuel, the cost of the travel, but the squares of the distances of the trips?…I don't see what they have to do with the trips, or the buses… I mean that it's only a mathematical artifice; it has nothing to do with the real travel."*

Another student referring to the same situation:

"*I see that when the distances of bus trips are more dispersed, the sum of the squares of the distances from the mean increase, and so… somehow, this sum indicates the dispersion. But why to use these squares of the distances that don't really mean anything for the trips, while we can use the sum of the absolute distances from the mean, or the interquartile range that are simple and we do understand what they measure?* "

The expression of the third student is characteristic of the conceptual difficulties of many other students: Although they can understand that when the dispersion around the mean increases, the sum of the squares of distances from the mean increases too, and therefore, the variance is a parameter that does express the dispersion, but they are reticent to accept this way of measuring dispersion, which involves quantities that are dimensionally ill-defined, and/or have unclear (obscure) meaning in the context of the examined situations. Their resistance against variance is reinforced by the fact that they already disposed others parameters measuring dispersion (mean absolute deviation, interquartile range) that are easier to understand and conceptually simpler than the variance.

Some students have pushed their argumentation on the subject even further, for example:

"*If it is permitted to use such strange quantities like squares of weights, or squares of the height of the students, why not use the cubes or the fourth powers of the weights to find the dispersion?…What I want to say is, why to use these strange quantities and not to use simply the absolute differences from the average and the mean absolute deviation which measures directly the dispersion from the average?*"

In their previous work, students had also encountered cases where they had the feeling that the meaning of a statistical parameter was not fully understandable in the context of the examined situation; e.g., treating distributions of variables with integer values (like the distribution of children per family) they had often calculated mean values that were not integers. Nevertheless, for all others parameters, there was a significant number of situations for which students had the feeling that the parameter has a clear meaning in the given context. For the variance the situation is different. With the exception of purely numerical examples, in all examined situations students had the feeling that variance has an unclear meaning in the given context, or even, that it is dimensionally ill-defined[39]. This can activate an important epistemological

[39]The teacher used a common argument on this point; the use of s.d. instead of variance, resolves, at least partially, the dimensional problem, since the s.d. is dimensionally meaningful (e.g. if the variance is determined using squares

obstacle for many students, since their criteria for the correct definition of a quantity and its coherent integration in the context of a situation remains systematically unsatisfied.[40]

The teacher told students that an essential reason for using variance as a dispersion parameter is that it has important computational properties and that in most cases it is easier to manipulate the variance than the mean absolute deviation from the mean (MAD)[41], but in order to understand this aspect they should work to find its properties.

A cause of difficulty in our students' work, concerning both the initial meaning of variance and its properties, was that in the usual graphical representations of statistical data they did not find interpretative elements that could help them in their investigations.[42] Since, neither the examined real, or realistic data examples, nor their graphical representations offer significant interpretative elements, students' productive work on the properties of the variance was confined mainly on a purely numerical-algebraic level.[43]

Despite these difficulties, students succeeded to find basic properties of variance and understand others explained by the teacher.[44] Comparing these properties with the corresponding ones of the MAD, they realized some of its important computational advantages and that in most cases it was significantly easier to manipulate the variance than the MAD. For example, they considered as an important advantage that, the variance of the union of two or more populations can be obtained from only the variance, the means and the total frequency of each population, whereas, in general the corresponding elements are not sufficient for calculating the MAD in this case. Moreover, they remarked that this phenomenon was not unique and that similar phenomena appear for other properties, since they found that: (a) In the general case they need less information to calculate the change of the variance resulting from the changes of the values of a subset of the total population, than the change of MAD.

(b) For calculating the second moment of a statistical variable X from a point x_0, they needed only the

---

of weights, the s.d. has the dimensions of weight). But students argued that this was not at all satisfactory because the s.d. was obtained through the use of meaningless quantities.

[40]Our students worked in a context of guided research work. In this context, trying to respect their own criteria of correctness and coherence was right from the start an essential element for achieving advancement in their investigations. So, by the time when the variance was introduced, students had developed an increased sensitivity on respecting these criteria that might have strengthened their reticence to accept variance. In conventional statistics courses, students are often led to accept working with elements that they perceive as obscure, or partially understood (this is particularly frequent in courses addressed to "users of statistics", that Shaughnessy qualifies as "ruler-bound receipt-type courses of statistics", Shaughnessy 1992). In such a course, students' reticence against the acceptance of variance is expected to decrease. However, this decrease is an element that expresses students' easiness to accept working with partially understood elements, and not their feeling of satisfactory understanding variance. Teachers should take this point into consideration to avoid an important misinterpretation concerning students' feeling of understanding the variance.

[41] When we use the term MAD, we mean the first absolute moment of a distribution around the mean.

[42]Using graphical representations supported students' work on the statistical parameters examined previously, because, in one or another way, they had always found elements to help them understand the properties of these parameters. Hence, using usual graphical representations of distributions, they looked for the graphical representation of variance, or of sums of squares of distances, but in this case, their research was not fruitful.

[43]Of course students have often used examples referring to real, or realistic situations (pupils' heights, population income etc). However, even in these cases, the productive work on these properties was mainly done in the reduced purely numerical context.

[44]Students in the first course, who were weak in algebra, encountered even more important difficulties than the students in the second course. In the first case, students were able to pose adequate questions for examining the subject; an important help for this was the network of questions and problems' that they have already established, during their work on the statistical parameters examined previously (e.g. they posed the question: *if we have two population (or two sets of values) and we want to calculate the variance of their union, what information we have to know and how we will calculate it?*, in analogy to the corresponding question that they had posed for the parameters that they had previously examined.) However, the aforementioned difficulties combined with their weakness in algebra, often made very difficult for the students to find the answers, and the teacher had to provide considerable help, or even to provide and explain the answer.

distance from x_0 to the mean value of X and the variance of X, whereas, this distance and the MAD are not sufficient to calculate the mean absolute deviation of the variable X from x_0[45].

Properties of this type and their comparisons have convinced students that variance is a useful dispersion parameter and that there are reasons to prefer it to the MAD. Nevertheless, many students still have the feeling that they had not a satisfactory understanding of the meaning of the variance, especially in the context of the real or realistic situations that they had examined.

Concerning this subject, students in the second course, posed the question of the existence of a relation between the variance and the MAD, hoping that such a relation, would link the variance (s^2) to MAD which was better understood and thus, it could help understanding the variance. After persistent research work on this subject, they succeeded to find that, for symmetric distributions, there is a "Pythagorean"-like relation ($s^2 = \text{MAD}^2 + S.R.^2$).[46] Then they found that when *mean = median*, $s^2 = \text{MAD}^2 + (S.R^2 + S.L^2)/2$, and finally that in the general case $s^2 = (\text{MAD}^2 / 4p_1 p_2) + (p_1 S.R^2 + p_2 S.L^2)$. (See Kourkoulos & Tzanakis 2003b).

The majority of the students considered these relations as a satisfactory answer to the initial question. Nevertheless, some others students pursued their investigation further; by iterating the general formula, they tried to find an approximation of s^2 using only absolute deviations.[47]

Besides the didactical interest of students' research work on this subject, the extended work and the considerable effort that this investigation had demanded, point out the students' great interest in finding relations between MAD and s^2. This interest was closely related to their feeling of insufficient understanding of the meaning of variance.

Remarks on sections 4 and 5

(a) The historical elements presented in section 2 point out that the statistical treatment of problems of social sciences was particularly difficult and a main reason was the complexity of the examined (social) phenomena. Understanding and adapting statistical aggregates and methods in this context was equally difficult; an important source of difficulty was that in this context statistical aggregates represent only data tendencies related to such complex phenomena. On the contrary, in domains of physics such as geodesy and astronomy the examined phenomena were simpler and statistical aggregates often had the status of approximations of measures of real objects that were of central importance for the examined situations. These were critical elements that facilitated the conception of basic statistical aggregates and methods in these domains.

Traditional introductory statistics courses disregard this historical reality and underestimate the

[45]These characteristics of variance and their differences from those of the MAD, all come from the same fundamental characteristic; that the $s^2 \times n$ (the variance multiplied by the corresponding frequency) has properties of an extensive quantity that the MAD×n has not, but the subject was not discussed with the students in such a general perspective.

[46]S.R. is the standard deviation of the "right subpopulation" ("right subpopulation" is that part of the population with values of the variable greater than the mean of the distribution). The S.L., the standard deviation of the "left sub-population" is defined similarly. p_1, p_2 are the relatives frequencies of the right and the left subpopulations of the distribution. This terminology belongs to the students. To formulate these properties it is also assumed that the frequencies of the values exactly on the mean and the median are zero, or negligible.

[47]The first and the second relations point out that for the concerned distributions variance is equal to the square of MAD plus an additional quantity that express the dispersions of the Right and the Left subpopulations around their own means. The third (general) relation can be considered as expressing almost the same thing but with a correction due to the unbalance between the size of the right and the left subpopulations. Furthermore, students have found empirically that in most of cases, if they substitute to the $p_1 S.R^2 + p S.L^2$ of the general relation the weighted mean of the squared MAD(s) of the right and of the left subpopulations $(p_1 MADR^2 + p_2 MADL^2)$ they find a value close of those of the Variance. ($S^2 \approx (\text{MAD}^2 / 4p_1p_2) + p_1 MADR^2 + p_2 MADL^2$. So, at the conceptual level, many of the students have the idea that for a large category of distributions they have arrive to approximate the variance with a rather simple relation that involves only the MAD, MADR and MADL. Students that have not found satisfactory this approximation have continued with the aforementioned iteration.

difficulties that students face concerning the elaboration of data examples related to social phenomena. This is expressed both, in the way examples related to such phenomena are treated and in the fact that often the examples used that are not purely numerical, are (almost) exclusively examples related to social phenomena. In introductory statistics courses more attention should be given to these difficulties concerning both the selection of the proposed examples that are related to social phenomena and the designed elaborations on these examples. Moreover, it is important that a significant part of the elaborated examples refer to situations that are conceptually simpler than those related to social phenomena (e.g. examples concerning adequate physical or geometrical models).

The above comment is particularly worthy concerning the introduction of variance and the s.d. The preceding analysis points out that the meaning of variance in examples referring to everyday-life phenomena (and more generally related to social phenomena), is very often unclear for the students or, even worse, dimensionally ill-defined. Often, introductory statistics courses are confined to the use of examples referring to such situations. This restriction can activate important epistemological obstacles against students' initial understanding of the meaning of variance. Moreover, the absence of examples referring to adequate situations, in which variance has a clear meaning, deprive students of interpretative elements, important for understanding its properties and allowing for a more profound understanding of the subject.

(b) Concerning the use of sums of first order absolute deviations and of squared deviations, there are some significant elements of similarity between our students' behavior and the historical development of the treatment of measurements' errors (in geodesy and astronomy) worth to be noticed: In both cases the use of the sums of absolute deviations appears as an important conceptual predecessor and competitor to the use of the sums of squared deviations. Our students (i) initially considered as conceptually simpler and clearer the meaning of MAD, whereas they considered the meaning of variance as difficult and/or obscure. (ii) By comparing the properties of MAD and of variance, they appreciated the computational advantages of variance and thus understood an important, albeit operationalistic, argument in favor of its use. By searching and finding relations between MAD and variance, they try to ameliorate their understanding of the meaning of variance through its linking to the conceptually simpler MAD; this goal was achieved to a significant extent (and they had the feeling that it was so); (iii) although they accepted the importance of variance as a dispersion parameter, many of them continued to believe that MAD was conceptually simpler (and thus some of them continue to prefer MAD to variance).

In the historical development, (i) The use of squared deviations and their sums emerged as an advantageous way of weighting deviations (errors) in an intellectual environment where weighting errors and equations (with simpler ways of weighting) was a common practice for obtaining aggregate equations. Among the previously used methods of weighted deviations, a principal one was the use of first order absolute deviations (see p.4). (ii) Open discussions on the characteristics of the method using the sums of squared deviations (least squares' method) in comparison to other methods using different ways of weighting errors were an essential element for appreciating both the method's potential and advantages, and its disadvantages. One of the main advantages of the Least Squares' Method (LSM) was the important computational properties of the sums of squared distances from a center; from these properties result the easiness and generality of application of LSM. (iii) Although already by the end of 1825 LSM had become a standard and widely used method in geodesy and astronomy, the method was not totally accepted. A remaining competitor was the use of the sum of first order absolute deviations (see note 13). It is also worth noting Laplace's preference to minimizing the mean absolute deviation as the criterion for finding a "best solution", versus Gauss preference to minimizing the mean squared deviation as such a criterion (see pp.5-7).

(c) Besides the aforementioned elements of similarity between our students' behavior and the historical

development, important differences remain concerning the conception of the use of sums of squared deviations as a way for measuring dispersion. Two of them are relevant to be underlined here: (i) Our students were introduced to this issue in a context using examples related to social phenomena; this was an important source of difficulty concerning their effort to understand variance. Historically, as far as statistics is concerned, the conception of using sums of squared deviations as a way for measuring dispersion was realised in the context of treatment of measurements (and of their errors) in geodesy and astronomy. This context was much simpler than the one related to social phenomena. (ii) The use of sums of squared deviations was introduced to our students by the teacher. Historically, this use emerged as an advantageous way of weighting deviations in an intellectual environment where weighting errors and equations, with simpler ways of weighting, was a common practice for obtaining aggregate equations. So historically, the use of the sums of squared deviations appears as a natural development of previously existing methods and practices.

This last remark suggests an interesting possible approach for introducing the students to the use of sums of squared deviations as a way for measuring dispersion: In the teaching activities situations' examples could be used, in which it is reasonable to consider different ways of weighting deviations from a center (central point or central line) for measuring dispersion, including first order absolute deviations and squared deviations. Students could compare these ways of weighting deviations and the resulting measures of dispersion and look for relations between them, thus achieving a better understanding of the subject. If an adequate teaching method is used (e.g. based on guided research work) students may even propose by themselves the use of the sums of squared deviations as a way for measuring dispersion. An essential question is what kind of situations examples could be used for realizing this approach? Historical analysis suggests that situations involving errors measurements may be adequate for this purpose; however, they are not the only ones. In the next section we will see situations referring to physical models that meet requirements mentioned in (a) above. These situations are also adequate for the approach discussed here.

6. Physical models

Remark (a) above summarizes a need that emerges for the analysis of the previous sections concerning the introductory teaching of variance: Introductory statistics courses should not be confined exclusively to the use of situations' examples related to social phenomena. In these courses it should be used also adequate situations' examples in which variance has a clear meaning, offering to students interpretative elements for understanding variance and its properties and thus facilitating the comprehension of the subject. Such situations can be derived by the rich and intimate relation between statistics and physics (see p.8; Tzanakis & Kourkoulos 2006). In this section we briefly present three elementary physical models that we have identified and it is possible to be used in introductory statistics courses (for further discussion on the historical and epistemological origins of these models (and of some others) as well as on the implications of their didactical use see Tzanakis, Kourkoulos 2006, Kourkoulos et al, Kourkoulos & Tzanakis 2007, Kourkoulos 2008).

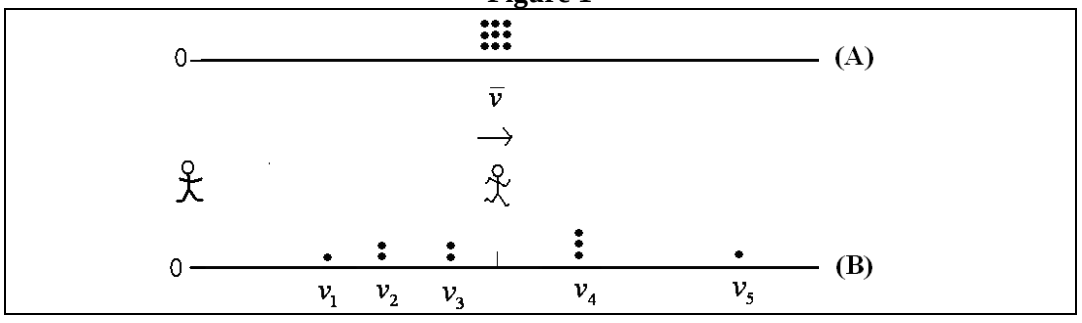
(A) *A system of masses:* A first model can be derived directly from the system of points masses that Legendre used for explaining the plausibility and the meaning of the MLS (footnote 9, Stigler 1986 pp 11-15, 55-61, Smith p.579); by considering that masses are distributed in one dimension. The statistical variable is the position x of the masses; x_i being the position of the mass m_i, which plays the role of the corresponding frequency, and $m_i x_i$ is the moment of m_i about the origin. The mean position $\bar{x}$ is the position of the centre of mass (CM), its variance is proportional to the moment of inertia around the CM, $I_B = \Sigma_i\, m_i(x_i - \bar{x})^2$, hence through the defining relation $I_B = M\, R^2$, the standard deviation equals the gyroscopic radius R of the system, $M = \Sigma_i\, m_i$ being the total mass. If additionally we consider that the axis is a bar to which the masses are attached and that the whole system is within the usual field of gravity,

then $\bar{x}$ is the only one position at which if we attach the bar, the system can be in static stable equilibrium (which is a very clear interpretation of the mean as equilibrium position). If we consider that the bar turns with constant angular velocity around $\bar{x}$ then the variance is also proportional to the angular momentum of the system, which is an additional interpretative element of the variance.
Variance has a rich and clear significance in the context of this model, however using the model in a teaching approach of the variance presupposes some students' familiarity with the physical concepts involved. Thus, it is more adequate to be used with students having such a familiarity (e.g. students' of a department of Physics or of engineering).

Physical models in which the variance expresses the dispersion energy of a physical system

(B) *The model of moving particles*: In the mid 19th century considering that a gas is constituted by molecules moving with different velocities (the molecular hypothesis) and taking into account the ideal gas law it was found that the mean kinetic energy of the molecules and (thus the variance of their velocities) is proportional to the absolute temperature of a gas. Therefore, the variance of velocities acquired a fundamental physical meaning and it was the first fundamental idea connecting macroscopic properties of a physical system to its microscopic structure (see p.8). From this fundamental physical modelisation a simple generic model can be derived by considering a set of particles of unit mass moving in the same direction. The distribution of the masses' velocities is such that the mass m_i (constituted by m_i unit mass particles) moves with velocity v_i (e.g. state B in figure 1). The energy of the particles as seen by an immovable observer is $E_{B\ obser0} = \frac{1}{2}\Sigma_i m_i v_i^2$. If we consider an initial state (A) where all masses move with the mean velocity, the energy of the particles in this state is $E_{A\ obser0} = \frac{1}{2}\Sigma_i m_i \bar{v}^2$. The system of masses can change from state A to state B for any internal or external reason that conserves the mean velocity, and equivalently the momentum ($\Sigma_i m_i v_i = \Sigma_i m_i \bar{v}$). If the change from state A to state B is due to an **internal** reason (e.g. an explosion), then momentum conservation holds and thus the conservation of $\bar{v}$ is assured.
The energy necessary for dispersing the masses from situation *A* in which they move together with the same velocity $\bar{v}$ to situation *B* where they move with different velocities is:
$E_{disp} = \frac{1}{2}\Sigma_i m_i v_i^2 - \frac{1}{2}\Sigma_i m_i \bar{v}^2$ but $E_{disp} = \frac{1}{2}\Sigma_i m_i s^2$, where s^2 is the variance of the velocities at state B. In other terms the variance s^2 is proportional to the energy necessary for dispersing the masses from the state in which they move together with the same velocity $\bar{v}$ to state B where they move with different velocities. More precisely, s^2 is (numerically) equal to twice the mean dispersion energy per particle or per unit mass. In the context of this model, not only variance has a clear meaning, but also it is easy to justify why it is interesting and natural to use it as a dispersion measure; since it expresses (is proportional to) the energy necessary for realizing the dispersion phenomenon examined. In other terms, there is a strong causal relation between the variance and the dispersion phenomenon that it measures.

Figure 1



For an observer moving with the mean velocity: $E_{A\ obser\bar{v}} = 0$ and $E_{B\ obser\bar{v}} = \frac{1}{2}\Sigma_i m_i (v_i^2 - \bar{v})^2$. So the energy necessary for dispersing masses' velocities from state A to state B for this observer is

$E'_{disp} = ½ \Sigma_i m_i (v_i^2 - \bar{v})^2$ *-0.* By the basic property of variance, $\overline{(x-\bar{x})^2} = \overline{x^2} - \bar{x}^2$, we have that $E'_{disp} = ½ \Sigma_i m_i s^2 = E_{disp.}$ The consideration of the 2nd observer permits to obtain an independent interpretation for each one of the two basic formulas that express variance; moreover through the basic property of variance reveals an important physical property: although the two observers perceive differently the energy in states A and B, they perceive the same "dispersion energy" (the amount of energy necessary to disperse masses' velocities from state A to state B).

(C) *The springs' model*: Later on, in the first decade of the 20th century, Einstein and Debye considered a solid body as constituted by elementary constituents that behave as microscopic oscillators. When the whole system is at equilibrium, the kinetic and potential energies of each oscillator is quadratic in its velocity and deviation from its mean position and the total energy of the system is proportional to its absolute macroscopic temperature. Furthermore, the variance of the variable corresponding to each degree of freedom (which are practically of infinite number) is also proportional to the absolute temperature of the system; this is the so-called (classical) energy equipartition theorem (see Tzanakis & Kourkoulos 2006). From this basic physical modelisation a simple generic model can be derived by considering a set of springs (in analogy to the oscillators) stretched in the same direction. For simplicity, only the aspect of potential energy is examined, thus the system is considered to be static. Although this model can be used independently (e.g. see Kourkoulos Tzanakis 2007) it is also complementary to model (B); in (B) the dispersing energy is kinetic whereas here it is potential.

We consider a set of springs having one edge attached to a bar and the other edge attached at some distance from the bar, so that the springs are perpendicular to the bar and the bar remains parallel to a fix direction (see figure 2 (B) below). Springs obey Hook's law and have the same spring constant k (in our teaching approach we initially put k=1 for simplicity). If the bar is at the origin, then the force on a spring is kx_i , where x_i is the distance of spring's end point from the origin, and the potential energy of this spring is $½kx_i^2$. The total force that the springs exert on the bar is $k\Sigma_i n_i x_i$ and the total potential energy of the system is $½k\Sigma_i n_i x_i^2$ (where n_i is the number of springs having their endpoint at distance x_i from the origin). Let $\bar{x}$ been the mean distance of springs endpoints from the origin. Consider that in a previous state of the system all springs' endpoints were at distance $\bar{x}$ from the origin and the attachment bar was at the origin (see figure 2 (A) below). In this state of the system the same force is exerted to the bar, $k\Sigma_i n_i x_i = k\Sigma_i n_i \bar{x}$. In this state the total potential energy of the system is $½k\Sigma_i n_i \bar{x}^2$. Therefore $E_{disp} = ½k\Sigma_i n_i x_i^2 - ½k\Sigma_i n_i \bar{x}^2$ is the energy necessary for dispersing the springs' endpoints from $\bar{x}$ to the positions x_i, with the attachment bar being at O; this energy is also equal to $½ k\Sigma_i n_i s^2$, (where s^2 is the variance of the distances of the springs end points from the origin at the final state). Thus s^2 is proportional to the energy necessary for dispersing the springs' endpoints from $\bar{x}$ to the positions x_i. More precisely, $½ks^2$ is the mean dispersion energy per spring.

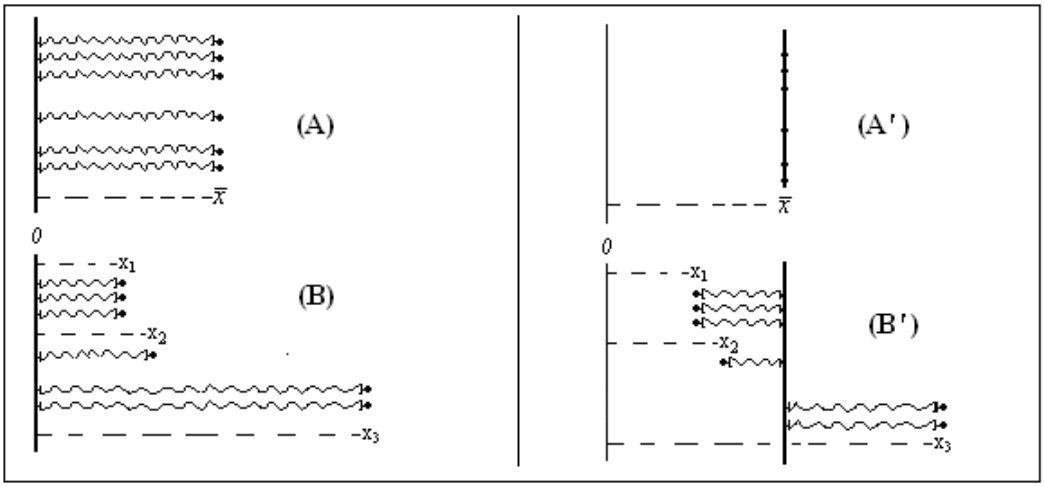


Figure 2

If initially the attachment bar is at $\bar{x}$ and all endpoints are also at $\bar{x}$, both the force exerted to the bar and the potential energy of the system is 0 (figure 2 (A′) above). If after that, the attachment bar remains at $\bar{x}$ but springs' end points are dispersed to the positions x_i, so that n_i is conserved as previously, (see figuren 2 (B′) above) then: (a) The total force exerted on the bar is $k\Sigma_i n_i(x_i - \bar{x})=0$, which explains that $\bar{x}$ is the equilibration position of the bar. (This constitutes a very clear interpretation of the mean as equilibrium position.) (b) The potential energy of the system is $½k\Sigma_i n_i(x_i - \bar{x})^2$. Since the initial potential energy of the system is zero, this is also the energy necessary for dispersing the springs' endpoints from the mean position $\bar{x}$ to the final positions x_i, when the attachment bar remains to $\bar{x}$. This energy is also equal to $½\, k\Sigma_i n_i s^2$.

So the variance (s^2) expresses (is proportional to) the energy necessary for dispersing the springs endpoints from the mean position to their final positions, whether the attachment bar remains at the origin, or at the mean position. In the context of this model too, variance has a clear meaning and it appears natural to use it as a dispersion measure; since it expresses (is proportional to) the energy necessary for realizing the dispersion phenomenon examined.

In experimental introductory teachings of statistics (that followed those mentioned in section 5) that we realized with students of the department of education, we have used not only situations examples related to social phenomena, but also examples referring to physical models for interpreting variance (Tzanakis & Kourkoulos 2006, Kourkoulos et al 2006, Kourkoulos 2008). We have used mainly models (B) and (C), because: (i) students were taught the physical concepts involved in these models, both, in high school and university courses of elementary physics, (ii) moreover, students have a significant informal background of basic concepts involved in these models (velocities, forces, energy) [48].

Using both models in the teaching activities, combined with reference to the relevant physical phenomena and the related historical modelisation in physics, created a strong conceptual image to the students, namely, that the variance expresses the dispersion energy in a variety of fundamental physical phenomena. This image helped students significantly to acquire a first understanding of the meaning of variance and accept its legitimacy as a basic dispersion parameter. Furthermore, these models permit to conceive

[48] Model (A) involves inertia and/or angular momentum for interpreting variance, which were concepts less familiar to our students. So, model (A) was used for interpreting the mean as an equilibrium center, that do not involve these concepts, and only general reference was given concerning the interpretation of variance in the context of this model. However this reference functioned toward enhancing students' idea (mainly created in relation to the use of models B and C) that the variance has a deep physical meaning and it is the natural measure of fundamental physical dispersion phenomena.

variance as equivalent to a mean energy, hence, to endow it with basic properties of energy. This offers important interpretative and explanatory elements on the properties of variance and on related aggregates, importantly facilitating their understanding and helping our students to use them with ease[49].

Moreover, the use of these models facilitated importantly students' acceptance of variance as a basic dispersion parameter in general, including its use in social phenomena, in which contextualizing the variance is unclear, or even problematic (Kourkoulos et al 2006). This was a significant difference compared to our previous experimental teaching work (section 5), in which, besides the pure numerical examples, only situations examples related to social phenomena were used; this was the main reason for which these students faced important epistemological obstacles to understand and accept the variance as a dispersion parameter.

Final Remarks

The didactical considerations of historical elements of statistics presented in this paper underlined some important students' difficulties, to which usual introductory teaching pay little attention concerning the understanding of variance; also it helped in better understanding the depth of students' difficulties on this issue. Furthermore, these considerations, combined with elements on students' behavior, permitted to identify elements that are important to enrich introductory teaching of the subject, in particular concerning the characteristics of the set of situations' examples used in this teaching. Additionally, they permitted to identify approaches and didactical activities that may be fruitful for the introductory teaching of variance: (i) The introductory teaching approach mentioned in p.26 §C is not yet investigated empirically; in our opinion such an investigation is an interesting further research work (ii) The didactical use of physical models (in particular, models B & C in section 6) were investigated in our first experimental teaching works and the results were promising. Further research work is needed for investigating the didactical use of other such models, and in particular of model A (section 6). (iii) With adequate adaptation, the aforementioned physicals models can be used fruitfully in introductory teaching of other important statistical concepts (e.g. extending models A & C in two dimensions permits to use them in an introductory teaching of the Method of Least Squares; see also Kourkoulos Tzanakis 2007; or, by enriching all three models with adequate random aspects of the phenomena involved, it is possible to use them in an introductory teaching of the sum of random variables and of the normal distribution). Exploring the didactical potential of the models concerning the teaching of these concepts is an appealing research perspective.

References

-Baker A., 2004a, *Design research in statistics education: On symbolizing and computer tools*, Utrecht: The Netherlands, CD-Beta Press.

-Baker A., 2004b, "Reasoning about shape as pattern in variability", *SERJ* 3(2), 64-83.

-Batanero C., Godino J.D., Vallecillos A., Green D.E., Holmes P., 1994, "Errors and difficulties in understanding elementary statistical concepts", *International Journal of Mathematical Education in Science and Technology* **25**(4), 527–547.

-Ben Zvi D., Arcavi A. (2001) Junior high School Students' construction of global view of data and data

[49]For example: As we have seen, the basic property of variance, $\overline{(x-\bar{x})^2}=\overline{x^2}-\bar{x}^2$, has a clear meaning in the context of both models, moreover model (C) permits to explain (prove) this property by simple energy consideration and without other algebraic calculations except simple additions and subtractions of equations expressing the energy of some states of the model (Kourkoulos, et al 2006). In the same way, model (B) permits to explain the important property that the variance of the union of two (or more) groups is "the variance of their means plus the mean of their variances" (Tzanakis & Kourkoulos 2006). Both models offer a clear meaning for the 2nd moment of a distribution around a value a and permit to explain its basic property, $M_2(a)=V+(\bar{x}-a)^2$ (Kourkoulos et al 2006). (For the interpretation and explanation of other properties and related aggregates in the context of these models see Kourkoulos & Tzanakis 2007, Kourkoulos 2008).

representations, *Educational Studies in Mathematics* **45**, 35–65.
-Biggs J. B., Collis K. F., 1982, *Evaluating the Quality of Learning: The SOLO Taxonomy*, NY: Academic Press.
-Biggs J. B., Collis K., 1991, "Multimodal learning and the quality of intelligent behavior", in H. Rowe (ed), *Intelligence, Reconceptualization and Measurement*, New Jersey: Laurence Erlbaum Assoc., pp. 57-76.
-Brush S.G., 1983, *Statistical Physics and the Atomic Theory of Matter,* Princeton: Princeton University Press.
-Canada D., 2006, "Elementary Pre-Service Teachers' Conceptions of Variation in a Probability Contex", *SERJ* 5(1), 36-63
-David H.A, 1995, First(?) occurrences of common terms in mathematical statistics *The American Statistician*, **49**, 121-133
-delMas R., Liu Y., 2005, Exploring students' conceptions of the standard deviation, *Statistics Education Research Journal* **4**(1), 55-81, http// www.stat.auckland .ac.nz/serj
-Garfield J. & Ben-Zvi D., 2007, "How Students Learn Statistics Revisited", *International Statistical Review* 75(3), 372–396.
-Gauss, K.F., 1996/1809, "Exposition de la Méthode des moindres carrés (Extrait du Theoria Motus Corporum celestium)", in Gauss, 1855, *Méthode des moindres carres. Memoires sur la combinaison des observation,* translation by J. Bertrand Paris: Mallet-Bachelier, reprinted, 1996, in *Reproduction de textes anciens, nouvelle série no11*, IREM, Université Paris VII, pp 65-76, (This work was originally published in Gauss, 1809, "Theoria Motus Corporum celestium", Hamburg: Perhtes et Besser).
-Gauss, K.F., 1996/1821, 1823 "Théorie de la combinaison des observations qui expose aux moindres erreurs", in Gauss, 1855, *Méthode des moindres carrés. Memoires sur la combinaison des observations,* translation by J. Bertrand Paris: Mallet-Bachelier, reprinted, 1996, in *Reproduction de textes anciens, nouvelle série no11*, IREM, Université Paris VII, part 1 pp 9-26, part 2 pp 26-43 (Original title "Theoria combinationis observationum erroribus minimis obnoxiae", presented at the Royal Society of Göttingen, 1st part presented in 1821, 2nd part presented in 1823).
-Henry M. (ed), 2001, *Autour de la modélisation en probabilités*, Besançon: Presses Universitaires de Franche-Comté
-Huck S., Cross T.L., Clark S. B., 1986, "Overcoming misconceptions about z-scores", *Teaching Statistics* **8**(2), 38-40
-Jeans, J., 1954/1904, *The Dynamical Theory of Gases*, New York: Dover (first published in 1904, Cambridge: Cambridge University Press).
-Jones G.A., Langrall C.W., Thornton C.A., Money E.S., Wares A. et al, 2001, Using students statistical thinking to inform instruction, *Journal of Mathematical Behavior*, 20, pp. 109-144
-Kolmogorov, A.N., Yushkevich A.P., 1992, *Mathematics of the 19th Century*, vol.I, Basel: Birkhäuser.
-Kourkoulos M., Tzanakis C., 2003a, "Graphic representations of data and their role in understanding elementary statistical concepts: An experimental teaching based on guided research work in groups" (in Greek), in M. Kourkoulos, G. Troulis, C. Tzanakis (eds), *Proceeding of the 3rd Colloquium on the Didactics of Mathematics*, Rethymnon, University of Crete, pp.209-228.
-Kourkoulos M., Tzanakis C., 2003b, "Introductory Statistics with problem-solving activities and guided research work, assisted by the use of EXCEL" in T. Triandafyllidis, C. Hadjikyriakou (eds) *Proceedings of the 6th International Conference on Technology in Mathematics Teaching* (ICTMT6), Athens: New Technologies Publications, pp.109-117.
-Kourkoulos M., Mantadakis E., Tzanakis C., 2006, "Didactical models enhancing students understanding of the concept of variance in Statistics" in D. Hughes-Hallett, I. Vakalis, H. Arikan (eds) *Proceedings of the 3rd International Conference on the Teaching of Mathematics* (*at the undergraduate level*)-ICTM3, Istambul: Turkish Mathematical Society, Pub. N.Y. : John Wiley & Sons, on CD-ROM, Paper-151.pdf
-Kourkoulos M., Tzanakis, 2007, "Enhancing students understanding on the Method of Least Squares: An interpretative model inspired by historical and epistemological considerations", in E. Brabin, N. Stelikova, K. Tzanakis (Eds), *Proceedings of ESU5*, Prague (in press)
-Kourkoulos M., 2008, "Didactical investigation of a simple physical model of moving particles for ameliorating the understanding of Variance in Statistics" (in Greek), in M. Kourkoulos, K. Tzanakis (Eds), *Proceedings of 5th ICDM*, Rethymnon, University of Crete, (in press)
-Laplace P.S., 1886, *Œuvres complètes de Laplace,* volume 7, Paris: Gauthier-Villars, published under the auspices of the Académie des Sciences (this volume is a reprint of the «Théorie analytique des probabilités», 3rd edition with supplements, Paris: Courcier, 1820; first edition in 1812).
-Laplace P.S., 1898a, « Mémoire sur les approximations des formulas qui sont fonctions de très grands nombres et sur leurs applications aux probabilités» in *Œuvres complètes de Laplace*, vol. 12, Paris: Gauhtier – Villars, published under the auspices of the Académie des Sciences, pp 301-345 (originally published in *Mémoires de l'Académie des Sciences de Paris, 1ere Série, Tome X 1809,1810*, pp 353-415).
-Laplace P.S., 1898b, «Supplément au mémoire sur les approximations des formulas qui sont fonctions de très grands nombres» in *Œuvres complètes de Laplace*, vol. 12, Paris: Gauhtier – Villars, published under the auspices of the

Académie des Sciences, pp 349-353 (originally published in *Mémoires de l'Académie des Sciences de Paris, 1ere Série, Tome X 1809,1810*, pp 559-565).
-Laplace P.S., 1898c, « Mémoire sur les intégrales définis et leur application aux probabilités et spécialement a la recherche du milieu qu'il faut choisir entre les résultats des observations », in *Œuvres complètes de Laplace,* vol. 12, Paris: Gauhtier – Villars, published under the auspices of the Académie des Sciences, pp. 357-412 (originally published in *Mémoires de l'Académie des Sciences de Paris, 1ere Série, Tome XI (1ere partie) 1810,1811*, pp 279-347).
-Loosen F., Lioen M., Lacante M., 1985, "The standard deviation: some drawbacks of an intuitive approach", *Teaching Statistics* **7**(1), pp.2-5.
-Maistrov, L.E., 1974, *Probability Theory: A historical sketch*, New York: Academic Press.
-Mevarech Z., 1983, "A deep structure model of students' statistical misconceptions", *Educational Studies in Mathematics* **14**, 415–429.
-Mokros J., Russell S.J., 1995, Children's concepts of average and representativeness, *Journal of Research in Mathematics Education,* Vol 26 (1) pp20-39
-Mooney E.S., 2002, "A framework for characterizing middle school student's statistical thinking", *Mathematical Thinking and Learning* **4**(1), 23-63.
-Noss R., Pozzi S. et Hoyles C., 1999, Touching Epistemologies: Meaning of average and variation in nursing practice, *Educational Studies in Mathematics* **40**, pp 25–51.
-Porter, Th. M., 1986, *The rise of statistical thinking: 1820-1900*, Princeton: Princeton University Press.
-Reading C., 2004, "Student description of variation while working with weather", *Statistics Education Research Journal* **3**(2), 84-105, http// www.stat.auckland .ac.nz/serj
-Reading, C., Shaughnessy, M., 2000, "Student perceptions of variation in a sampling situation", in T. Nakahara and M. Kyama (eds), *Proceedings of the 24th Conference of the International Group for the Psychology of Mathematics Education*, Hiroshima: Hiroshima University, Vol 4, pp.89–96.
-Reading C., Shaughnessy J.M., 2004, "Reasoning about variation", in D. Ben-Zvi & G. Garfield (eds), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*, Dordrecht: Kluwer, pp.201-226
-Pegg J., 2003, "Assessment in mathematics: A developmental approach", in J.M. Royer (ed.), *Advances in Mathematical Cognition*, Greenwich, CT: Information Age publishing, pp.227-259.
-Rubin, A. & Rosebery, A.S, 1990, "Teachers' misunderstandings in statistical reasoning: evidence from a field test of innovative material", in A. Hawkins (ed.), *Training Teachers to Teach Statistics*, Voorburg: ISI, pp. 72–89.
-Shaughnessy J.M.,1992, "Research in probability and statistics: reflections and directions", in D.A. Grouws (ed.), *Handbook of Research on Mathematics Teaching and Learning*, New York: Macmillan, pp.465–494.
-Shaughnessy J.M., Garfield J., Greer B., 1996, Data handling, in A.J. Bishop, K. Clements, C. Keitel, J. Kilpatrick, C. Laborde (Eds), *International handbook of Mathematics Education*, Dordrecht: Kluwer, pp 205-237
-Shaughnessy J.M., Watson J., Moritz J., Reading C., 1999, "School mathematics students' acknowledgement of statistical variation", NCTM Precession Research Symposium: *There's more in life than centers*. Paper presented at 77th Annual NCTM Conference, San Francisco, CA.
-Sklar, L., 1993, *Physics and Chance: Philosophical issues in the foundations of Statistical Mechanics*, Cambridge: Cambridge University Press.
-Smith D.E., 1959, *A source book in Mathematics*, New York: Dover
-Stigler S.M., 1978, "Mathematical Statistics in the Early States", *The Annals of Statistic*s, **6** No2, pp. 239-265.
-Stigler S.M., 1986, *The History of Statistics: The measurement of uncertainty before 1900*, Cambridge (MA): Harvard University Press.
-Stigler, S.M., 1999, *Statistics on the table: The history of statistical concepts and methods*, Cambridge (MA): Harvard University Press.
-Torok, R., & Watson, J., 2000, "Development of the concept of statistical variation: An exploratory Study" *Mathematics Education Research Journal, 12*, 147–169.
-Tukey 1977 *Exploratory Data Analysis,* Reading MA, Addison-Wesley
-Tzanakis C., Arcavi A. et al., 2000, "Integrating history of mathematics in the classroom: an analytic survey" in - J. Fauvel & J. van Maanen (eds.), *History in Mathematics Education: The ICMI Study*, Dordrecht: Kluwer. pp.201-240.
-Tzanakis C., Kourkoulos M. 2006, "May history and physics provide a useful aid for introducing basic statistical concepts?" *Proceedings of the HPM Satellite Meeting of ICME-10 & the 4th Summer University on the History and Epistemology in Mathematics Education*, revised edition, F. Furinghetti, S. Kaijser & C. Tzanakis (editors), University of Crete, Greece, pp284-295.
-Walker H.M., 1931, *Studies in the history of statistical methods with special reference to certain educational problems*, Baltimore: Williams and Wilkinson Co.
-Watson, J. M., Kelly, B. A., Callingham, R. A., & Shaughnessy, J. M., 2003, "The measurement of school students' understanding of statistical variation", *IJMEST, 34*, 1-29.