# ON PEARSON-VERIFICATION AND THE CHI-SQUARE TEST

JOAKIM EKSTRÖM

Abstract. Karl Pearson's seminal article On the criterion is reviewed, formalized in modern notation and its method is extended beyond the normal distributions using the Mahalanobis distance as a vehicle. The extension yields a simple method, firmly rooted in history, for computing exact p-values and acceptance regions for hypotheses under a large class of probability distributions and of arbitrary dimension. As a by-product it is for example shown that the so-called method of acceptance intervals by percentiles is a univariate special case of Pearson's method. The article's content is discussed in context of other sources, in particular Pearson's The Grammar of Science, yielding a holistic approach to verification/falisfication of hypotheses through empirical observation.

## 1. Introduction

Karl Pearson was a man who took on a mission to rectify science, and particularly the way in which scientists argued using empirical evidence. Based on his writings it seems as Pearson was driven by the frustration he evidently felt with the lack of standards with respect to the use and valuation of empirical evidence. Pearson's feeling of disappointment with the haphazard and ad hoc manner in which some of his contemporaries argued for their hypotheses is prominently expressed in the storied *"statistics on the table, please"* quote (see, e.g., Stigler, 1999).

Channeling his frustration into action, though, *On the criterion* (Pearson, 1900) proposes a standard for valuation of empirical evidence which nowadays is referred to as statistical hypothesis testing. The method, consisting of Pearson's distance criterion, the chi-square statistic and the p-value, is at present one of the most used methods in science for verification/falsification, i.e. testing, of hypotheses based on empirical observations, and is the standard employed by the United States Food and Drug Administration and its international counterparts, for example.

Though, according to English statistician and historian Plackett (1983) the article (i.e. Pearson, 1900) has not been well understood within the statistical community. The lack of perfect clarity about what is arguably the present de facto standard for valuation of empirical evidence has likely in various ways effected the scientific community negatively. Karl Pearson would in retrospect likely not have regretted presenting his criterion in a pedagogically more elaborate manner.

The present article discusses Pearson (1900) in context of other sources, such as *The Grammar* (Pearson, 1911), and puts the content into a formalized framework using modern notation. Furthermore, the concepts are naturally extended beyond the normal distributions using the Mahalanobis distance as a vehicle.

## 2. Pearson's philosophical basis

*The Grammar of Science* is a good source for context to *On the criterion* (Pearson, 1900). In it, Pearson distinguishes between what he refers to as the World of Conceptions and the World of Perceptions, a viewpoint that bears many similarities with Plato's worlds of forms and senses. Pearson argues that deductive reasoning, cause and effect applies to the world of conceptions only, and that to our world applies only *routine in perceptions*. The measure of the degree of routine is probability, by Pearson seemingly used synonymously with relative frequency. Pearson's reasoning leads to the concise statement: *Proof in the field of perceptions is the demonstration of overwhelming probability*. The statement is formalized, slightly paraphrased, below.

**Postulate** (Pearson's verification postulate)**.** *Verification by means of empirical evidence is the demonstration of overwhelming probability.*

As an example, consider the statement $A$ *implies* $B$. Proving the statement using logical deduction amounts to showing that the combination $(A, \neg B)$ yields contradictions between fundamental axioms and/or assumptions, such as $a \neq a$. Verification of the statement in Pearson's sense is something completely different. Pearson-verification of $A$ *implies* $B$ amounts to demonstrating that the combination $(A, \neg B)$ is overwhelmingly improbable, or equivalently: given $A$; $B$ is overwhelmingly probable. For clarity, the following terminology is formally defined. Note that in the present article the terms hypothesis, statement, proposition and conjecture are used synonymously, meaning a claim which has yet to be verified/falsified, i.e. tested.

**Definition 1.** A proposition is *Pearson-verified* if it is overwhelmingly probable. A proposition is *Pearson-falsified* if its negation is Pearson-verified.

In many situations, it is for various reasons simpler to demonstrate that the negation of a proposition is overwhelmingly improbable, i.e. Pearson-falsify the negated proposition, than to Pearson-verify the proposition. By elementary probability theory, if a proposition has probability $p$, its negation, the logical complement, has probability $1-p$ and hence the duality between the probability of a proposition and the improbability of its negation. Note also that a third case exists in which it can neither be demonstrated that a proposition is overwhelmingly probable nor overwhelmingly improbable. In that third case, consequently, the proposition is neither Pearson-verified nor Pearson-falsified and the empirical evidence must consequently deemed inconclusive.

A natural question of practical importance is, of course, at which point a probability should be deemed overwhelming. Pearson (1900) argued explicitly against any preset threshold value, favoring that the determination should be *"based on the general order of magnitude of the probability, and not on slight differences in its value."* However, Fisher (1925) and Neyman & Pearson (1933) discussed using fixed values for the determination of what is overwhelmingly probable, and the use of preset threshold values has since become the norm. The use of so-called statistical significance levels is formalized in the following definition.

**Definition 2.** A proposition is *Pearson-verified at statistical significance level* $\alpha$, for some given real number $\alpha$, if it is as or more probable than $1 - \alpha$. A proposition is *Pearson-falsified at statistical significance level* $\alpha$ if its negation is Pearson-verified at statistical significance level $\alpha$.

Fisher (1925) and Neyman & Pearson (1933) used statistical significance levels .05 and .01 in examples. The statistical significance level .001 has also become a conventional choice. With respect to Definition 1, whether a probability of .95 should be deemed overwhelming is perhaps arguable.

While the present article does not aim to give an exhaustive account of the history of inductive philosophy, it is noteworthy that Jakob Bernoulli in *Ars Conjectandi* (1713) defined concepts that are nearly identical to those discussed so far. A proposition is *morally certain*, Bernoulli explains, if "*its probability comes so close to complete certainty that the difference cannot be perceived*". A completely certain proposition is one that is proved by means of logical deduction and its probability is then represented by one. Bernoulli proposes statistical significance levels .01 and .001. As an interesting trivia, Bernoulli proposes that the statistical significance level should be set by the state: "*It would be useful, accordingly, if definite limits for moral certainty were established by the authority of magistracy. For instance, it might be determined whether* 99% *of certainty suffices of whether* 99.9% *is required.*" Bernoulli further argues that relevant probabilities can be empirically determined by repeated observation and for the sake of the argument Bernoulli states the theorem of his which is nowadays referred to as the law of large numbers. However, Bernoulli's hypothesis testing theory seems to have been overlooked by his peers and by history. Hald (1990) speculates that it might be because Bernoulli failed to provide any convincing example. Fisher (1925) credits the invention of the hypothesis test to Pearson (1900), which supports the proposition that Bernoulli's concepts at that time had been lost in history. This about Jakob Bernoulli and Ars Conjectandi.

Pearson (1911) exemplifies an application of Definition 1 as follows, slightly paraphrased. If men's past experience has shown that a certain set of causes $A$ are on repetition followed by the same effect $B$ a million times to one, say, the implication $A \implies B$ is Pearson-verified. Using Definition 2, the implication is Pearson-verified at statistical significance level .001, the highest conventional statistical significance level. Colloquially, one could say that the concept of Pearson-verification is constructed so that the exception to the rule is automatically discounted as an erroneous observation or a so-called outlier.

However, in many situations direct application of Definition 2 is not possible. A notable example is when continuous probability distributions apply. This led Pearson to propose his distance criterion and develop what has subsequently become known as the chi-square test.

## 3. Pearson's chi-distance criterion

Pearson (1900) proposes a criterion for whether a value can be reasonably supposed to have arisen from random sampling, using his exact wording. The criterion is based on distance, and from it flow the chi-square statistic, p-value and acceptance region. The present section reviews *On the criterion* (Pearson, 1900), and the subsequent Section 4 is a more formalized extension of Pearson's method beyond normal distributions.

Suppose $x_1, \dots, x_n$ are real-valued observations of some phenomenon with Gauss-Pearson decomposition $x_i = \mu_i + u_i$, $i = 1, \dots, n$, where $\mu_1, \dots, \mu_n$ are the ideal parts of the observed phenomenon and $u_1, \dots, u_n$ are the random parts. For notational convenience, let the arrow accent $\vec{x}$ denote the sequence $(x_1, \dots, x_n)$ and allow sequences to be added through component-wise addition, i.e. vector addition.

The problem at hand is to test a hypothesis such as $\vec{\mu} = \vec{\nu}$, for some given $\vec{\nu}$, when the random parts have a continuous joint probability distribution. The hypothesis yields the representation $\vec{x} = \vec{\nu} + \vec{e}$, where $\vec{e}$ is the representation residual. Solving for $\vec{e}$ yields $\vec{e} = -\vec{\nu} + \vec{x}$, which under the hypothesis equals $-\vec{\mu} + \vec{x} = \vec{u}$. Hence, the hypothesis implies that the representation residual $\vec{e}$ is an observation from the distribution of $\vec{u}$, which is the basis upon which Pearson's test in constructed. For additional clarity, Pearson uses the logical reasoning: if $A$ implies $B$, then $\neg B$ implies $\neg A$; so if it can be demonstrated that the proposition $\vec{e} \sim \mathcal{L}(\vec{u})$ is overwhelmingly improbable, i.e. Pearson-falsified, then it follows that $\vec{\mu} = \vec{\nu}$, the hypothesis, is Pearson-falsified as well. The notation $\mathcal{L}(\vec{u})$ denotes the probability distribution (or *law*) of the random variable $\vec{u}$.

Pearson (1900) assumes that the random parts are independent and distributed $u_i \sim \mathrm{N}(0, \sigma_i^2)$, $i = 1, \ldots, n$, and that this is known ex ante. Pearson defines the *chi-distance* between $\vec{e}$ and 0, where 0 denotes the zero element of $\mathbb{R}^n$. The chi-distance is nowadays recognized as an obsolete special case of the Mahalanobis distance, so Pearson's chi-statistic is denoted $\eta = d(\vec{e}, 0)$, where $d$ denotes the Mahalanobis distance under the joint distribution of the random parts. Note that Pearson (1900) uses the Greek common $\chi$ for denotation, though the same symbol is nowadays used for many related concepts such as the chi-distance, the chi-distribution and even percentiles of the chi-distribution, and hence, for the purpose of avoiding notational conflicts, the chi-statistic $d(\vec{e}, 0)$ is in the present article denoted by $\eta$.

Pearson then proposes the following. If the distance $d(\vec{e}, 0)$ is small, there is little evidence against the proposition that the residual $\vec{e}$ is an observation from the distribution $\mathcal{L}(\vec{u})$, and consequently there is little evidence against the hypothesis. On the other hand, if the distance $d(\vec{e}, 0)$ is great, then Pearson argues that it is hard to conceive that the residual $\vec{e}$ is an observation from $\mathcal{L}(\vec{u})$, and consequently that the hypothesis must be deemed improbable. The distance $\eta = d(\vec{e}, 0)$ is, for additional clarification, the criterion Pearson proposes for the evaluation of whether the residual $\vec{e}$ can be reasonably supposed to have arisen from random sampling from the distribution $\mathcal{L}(\vec{u})$ of the random part $\vec{u}$.

Whether the distance $d(\vec{e}, 0)$ is great to the extent that the hypothesis must be deemed overwhelmingly improbable, i.e. Pearson-falsified, is determined via the value $P$ which is defined as follows. Let $B_r(m)$ be the Mahalanobis ball with radius $r$ and center point $m$, i.e. $B_r(m) = \{x : d(x, m) < r\}$ where $d$ is the Mahalanobis distance under the distribution of the random part $\vec{u}$. The value $P$, i.e. the p-value, is defined as the probability measure of the complement of $B_\eta(0)$, i.e.

$$P = \mathbb{P}(B_\eta(0)^c),$$

where $\mathbb{P}$ is the probability measure $\mathbb{P}(A) = \int_A f d\lambda$, and where $f$ is the density function of the random part, $\vec{u}$, and $\lambda$ the Lebesgue measure. Note that since Pearson (1900) was published before Henri Lebesgue published his dissertation, Pearson expresses the value $P$ as an iterated integral. The p-value is often interpreted as the probability under the hypothesis of an as, or more, extreme residual than $\vec{e}$, and equivalently a more extreme observation than $\vec{x}$.

Later, it was discovered that the p-value can be computed more easily through the observed radius, i.e. $\eta = d(\vec{e}, 0)$ (by Pearson denoted $\chi$). In fact, the p-value was computed via the squared radius $\eta^2$, the distribution of which has since been named the chi-square distribution.

Neyman & Pearson (1933) proposed solving for $r$ the equation $\mathbb{P}(B_r(0)^c) = \alpha$, for a given $\alpha$, yielding for the residual $\vec{e}$ an *acceptance region* $B_r(0)$ at statistical significance level $\alpha$. A hypothesis is accepted in the Neyman & Pearson sense if it is not Pearson-falsified, i.e. a hypothesis is Neyman-Pearson accepted if it is either Pearson-verified or if the empirical evidence is inconclusive. For the purpose of formalism the concept is defined, as follows.

**Definition 3.** A proposition is *Neyman-Pearson accepted (at statistical significance level $\alpha$)* if it is not Pearson-falsified (at statistical significance level $\alpha$).

As a final remark, Pearson (1900) introduced a string of concepts of fundamental importance, such as his distance criterion, the chi-distance and the p-value, in a few short pages without any explanatory wordings, and then diverted into a lengthy and technical discussion on computational details. From a pedagogical point of view, Pearson's text is quite possibly one of the worst examples in the history of statistics, a proposition circumstantially supported by Plackett (1983) and Lehmann (1993). The difficulty of understanding the article is compounded by the fact that Pearson included examples with discrete data, which indeed is asymptotically normally distributed however that is not mentioned in the article. Pearson's method was communicated to the scientific community principally via the textbooks of his colleagues Yule (1911) and Fisher (1925).

## 4. Extended framework

Since a distance per definition is real-valued it maps pairs of elements of arbitrary spaces into the reals. The term dimension reduction is sometimes used in statistical settings, and a distance accomplishes this naturally. Furthermore a distance is zero if and only if the two elements are equal, and is resultantly often colloquially interpreted: the greater the distance the less equal the elements. Moreover, the Mahalanobis distance automatically accounts for probability distribution. In all, Pearson's approach has many merits and is well worth extending beyond the normal distribution, something which is easily done through the definition of the Mahalanobis distance given by Ekström (2011).

One of the premises of Pearson's hypothesis test is that the observation $x$ has a Gauss-Pearson decomposition, i.e. that it can be written $x = \mu + u$ where $\mu$ is the ideal part of the observed phenomenon, $u$ is the random part, and $(\mathbb{X}, +)$ some group. It is also presumed, importantly, that the distribution $\mathcal{L}(u)$ is ex ante known.

The hypothesis that the ideal part of the observed phenomenon equals some value $\nu \in \mathbb{X}$, i.e. $\mu = \nu$, then yields a second representation $x = \nu + e$, where $e$ is the representation residual. Hence there are two expressions for the observation,

$$\begin{cases} x = \mu + u, & \text{(the Gauss-Pearson decomposition)} \\ x = \nu + e. & \text{(the hypothesis representation)} \end{cases}$$

Since $\mathbb{X}$ per assumption is a group, it follows

$$\mu = \nu \ \textit{ implies } \ e \sim \mathcal{L}(u), \quad \text{and} \quad e \nsim \mathcal{L}(u) \ \textit{ implies } \ \mu \neq \nu.$$

Consequently, the hypothesis can be falsified through falsification of the proposition $e \sim \mathcal{L}(u)$. In general, though, the latter proposition cannot be falsified by means of logical deduction, however it can be Pearson-falsified, yielding Pearson-falsification of the hypothesis $\mu = \nu$. For this purpose Pearson proposed his distance criterion, which is formalized as follows.

**Proposition** (Pearson's distance criterion)**.** *Suppose $z$ is a value and $\mathcal{F}$ a probability distribution with reference point $m$, and let $d$ denote the Mahalanobis distance under $\mathcal{F}$. If the distance $d(z, m)$ is great, then the proposition $z \sim \mathcal{F}$ is deemed improbable.*

The Mahalanobis distance under a distribution $\mathcal{F}$ is defined

$$d(x, y) = ||T(x) - T(y)||,$$

where $T$ is a transformation that maps a random variable with distribution $\mathcal{F}$ to a standard normal random variable and $|| \cdot ||$ denotes the Euclidean distance. The present definition extends the conventional definition beyond normal distributions. Conditions for existence and uniqueness of the Mahalanobis distance, as well as explicit transformations, are discussed in Ekström (2011).

With respect to the reference point of the distribution, given the interpretation of the p-value as the probability of a more extreme value than the one observed, the reference point $m$ must be the least extreme value of the distribution. Under the normal distribution assumption, Pearson (1900) sets the reference point as the zero element which simultaneously equals the mean, the median and the mode of the distribution. In many cases the median is indeed a both natural and convenient choice of reference point, however if the observation is a distance (cf. the chi-square distribution) it is reasonable to consider zero to be the least extreme value. Admittedly, the different possible choices of reference points cause an ambiguity which is undesired.

If $x_1, \dots, x_n \in \mathbb{X}$ is a sample of $n$ observations, then simply let $\vec{x} \in \mathbb{X}^n$ be the sequence $(x_1, \dots, x_n)$, and let sequences be added through component-wise addition. Then $(\mathbb{X}^n, +)$ is a group and $\vec{\mu}$ and $\vec{u}$ are the ideal and random parts, respectively, of the Gauss-Pearson decomposition of $\vec{x}$. Thus the arrow accent can be added to all elements, i.e. $\vec{x}$, $\vec{\mu}$, $\vec{\nu}$, $\vec{e}$, $\vec{m}$, et cetera, or equivalently the space can be redefined $\tilde{\mathbb{X}} = \mathbb{X}^n$.

In the extended framework, Pearson's concepts are defined as follows.

**Definition 4.** Presuming the observations have Gauss-Pearson decomposition $\vec{x} = \vec{\mu} + \vec{u}$, where the random part $\vec{u}$ has ex ante known distribution $\mathcal{L}(\vec{u})$ such that the Mahalanobis distance exists, *Pearson's chi-statistic*, denoted $\eta$, under the hypothesis $\vec{\mu} = \vec{\nu}$ is defined

$$\eta = d(\vec{e}, \vec{m}),$$

where $\vec{e}$ is the residual of the representation $\vec{x} = \vec{\nu} + \vec{e}$, $\vec{m}$ is the reference point of the distribution $\mathcal{L}(\vec{u})$, and $d$ is the Mahalanobis distance under $\mathcal{L}(\vec{u})$.

**Definition 5.** In the notation and context of Definition 4, the *p-value*, $P$, is defined

$$P = \mathbb{P}(B_\eta(\vec{m})^c),$$

where $\mathbb{P} : \mathcal{B}(\mathbb{X}^n) \to \mathbb{R}$ is the probability measure under $\mathcal{L}(\vec{u})$, $B_\eta(\vec{m})$ is the ball with radius $\eta$ and center point $\vec{m}$, i.e. $\{z \in \mathbb{X}^n : d(z, \vec{m}) < \eta\}$, and $d$ is the Mahalanobis distance under $\mathcal{L}(\vec{u})$.

The following results can in many cases simplify the computation of the p-value considerably.

**Lemma 1.** *Suppose that $z$ is a random variable of dimension $p$ with a distribution $\mathcal{F}$ such that the Mahalanobis distance exists, and that $m$ is an element in the range of $z$. Then, $d(z,m)^2$ has a non-central chi-square distribution with $p$ degrees of freedom and non-centrality parameter $||T(m)||^2$, where $T$ is the transformation used for the Mahalanobis distance $d$.*

*Proof.* By construction the random variable $T(z)$ is standard normal with dimension $p$. It is then immediate from the definition of the non-central chi-square distribution that $||T(z) - T(m)||^2$ is such distributed with parameters as stated. □

**Theorem 2.** *In the notation of Definition 5 it holds that*

$$\mathbb{P}(B_r(\vec{m})^c) = \mathbb{Q}([0, r^2)^c),$$

*where $\mathbb{Q} : \mathcal{B}(\mathbb{R}) \to \mathbb{R}$ is the probability measure defined $\mathbb{Q}(A) = \int_A g d\lambda$, where $\lambda$ is the Lebesgue measure and $g$ is the density function of a non-central chi-square distribution with $np$ degrees of freedom and non-centrality parameter $||T(\vec{m})||^2$ where $T$ is the transformation used for the Mahalanobis distance.*

*Proof.* Note the tautology $z \in B_r(m) \iff d(z,m) < r$, of which the latter is equivalent to $d(z,m)^2 < r^2$ since the quadratic function defined on the non-negative reals is a bijection. The statement then follows by Lemma 1. □

*Remark* 1. By Theorem 2, Definition 5 can be expressed in the following way, which is likely familiar to many,

$$\text{p-value} = 1 - Q(\eta^2),$$

where $Q$ is the distribution function of the distribution specified in Theorem 2.

Pearson's hypothesis test is then concluded with the following deduction. If the p-value is small (less than or equal to $\alpha$), then by Pearson's distance criterion and Definitions 1 and 2 the proposition $\vec{e} \sim \mathcal{L}(\vec{u})$ is Pearson-falsified (at statistical significance level $\alpha$), which, in turn, implies that the hypothesis $\vec{\mu} = \vec{\nu}$ is Pearson-falsified (at statistical significance level $\alpha$). The latter is, of course, by definition equivalent to the hypothesis negation $\vec{\mu} \neq \vec{\nu}$ being Pearson-verified (at statistical significance level $\alpha$).

The remainder of the present section regards acceptance regions.

**Definition 6.** In the notation and context of Definition 5, the *acceptance region at statistical significance level* $\alpha$, $A$, is defined

$$A = B_r(\vec{m}),$$

where $r$ is the solution to the equation $\mathbb{P}(B_r(\vec{m})^c) = \alpha$.

Of course, the equation of Definition 6 can via Theorem 2 be restated along the lines of Remark 1, from which it follows that

$$r^2 = Q^{-1}(1 - \alpha),$$

which thus yields an explicit expression for the radius $r$. Moreover, the acceptance region can be expressed as the preimage of a Euclidean ball.

**Theorem 3.** *Suppose $T$ is the transformation used for the Mahalanobis distance and let $B_r(x)$ denote the Mahalanobis ball and $E_r(x)$ the Euclidean ball, then*

$$B_r(x) = T^{-1}(E_r(T(x))).$$

*Proof.* Notice,

$$B_r(x) = \{y : ||T(y) - T(x)|| < r\} = \{y : T(y) \in E_r(T(x))\} = T^{-1}(E_r(T(x))),$$

which shows the statement. □

As a result of Theorem 3, acceptance regions can equivalently be defined as transformations of Euclidean balls. The transformation, which depends on the distribution, accounts for the difference between the distribution at hand and the standard normal distribution. The following corollary applies to the univariate case, i.e. when $\mathbb{X}^n$ is a one dimensional linear space.

**Corollary 4.** *In the notation and context of Definition 6, suppose that the distribution $\mathcal{L}(\vec{u})$ is univariate and absolutely continuous, and let $F$ denote the distribution function of $\mathcal{L}(\vec{u})$ and $\Phi$ the standard normal distribution function. Then it holds that*

$$B_r(\vec{m}) = (F^{-1} \circ \Phi(\hat{m} - r), F^{-1} \circ \Phi(\hat{m} + r)),$$

*where $\hat{m} = \Phi^{-1} \circ F(\vec{m})$.*

*Proof.* For convenience, let the composition $\Phi^{-1} \circ F$ temporarily be denoted $G$. In the univariate absolutely continuous case the Mahalanobis distance equals $|G(x) - G(y)|$ (see Ekström, 2011). Since $G$ is continuous and non-decreasing, application of Theorem 3 yields the stated interval. □

In the univariate case, acceptance regions are often called acceptance intervals. The following notable special cases show that the popular so-called method of acceptance intervals by percentiles is a special case of the method of the present article. And conversely, that the method of the present article is a multivariate extension of the percentile method.

**Corollary 5.** *If the point of reference, $m$, is the median of the univariate distribution $\mathcal{L}(\vec{u})$, then the acceptance region at statistical significance level $\alpha$ reduces to*

$$A = (F^{-1}(\alpha/2), F^{-1}(1 - \alpha/2)).$$

*If the point of reference is the greatest lower bound (least upper bound) of the support of the density function of $\mathcal{L}(\vec{u})$, then the acceptance region at statistical significance level $\alpha$, $A$, reduces to (respectively)*

$$A = [F^{-1}(0), F^{-1}(1 - \alpha)),$$

*and*

$$A = (F^{-1}(\alpha), F^{-1}(1)].$$

*Remark* 2. Taking a sequence converging to the greatest lower bound and least upper bound, respectively, and using the continuity of the composition $\Phi^{-1} \circ F$ shows the latter two cases.

Interestingly, Corollary 5 gives a straight answer to whether a hypothesis test should be one-sided or two-sided, in the terminology of the method of acceptance intervals by percentiles; it depends on the point of reference. Furthermore, it follows that a point of reference in between the median and one of the extremes yields an acceptance interval which is asymmetric.

## 5. Discussion

Karl Pearson sought to create a standard for the use and valuation of empirical evidence in science, and the present article reviews and extends his proposal. Pearson's quest for a commonly agreed upon standard raises two questions for discussion. First, is there at all a need for such a standard? Second, is Pearson's proposed standard, consisting of his verification postulate, the concept of Pearson-verification, the distance criterion, et cetera, a good enough standard for the purpose it is supposed to serve?

Answers to the two questions are of course largely normative in nature, and opinions will likely continue to differ for years to come. As is well known, philosophical discussions have a tendency to make people take uncompromising positions. However, one approach is to look at the course of history, and see whether opinions within the scientific community have tended to converge towards some consensus. And at present, it is likely fair to say that Pearson's method has become a de facto standard for valuation of empirical evidence within the scientific community. P-value, acceptance region and statistical significance level have become some of the most important function words in the modern grammar of science, not least in the health and social sciences. Based on the wide usage of Pearson's method, it seems as if members of the scientific community are generally comfortable with the idea of having a commonly agreed upon standard for the valuation of empirical evidence.

Part of the appeal of Pearson's method is likely that given the Gauss-Pearson decomposition and the ex ante known joint distribution of the random parts, and a preset statistical significance level, Pearson's method is fully automated, a characteristic which, right or wrong, is generally seen as a contributor to scientific fairness and objectiveness. However, the objectiveness can of course in practice be undermined for several reasons, for example if the distribution of the random parts is not ex ante known but merely assumed.

It would be difficult to argue that Pearson's method is best in some sense, and it is possible that Pearson's method eventually will be replaced by some other standard for valuation of empirical evidence. Although even if there is a future such paradigm shift, the concept of Pearson-verification will remain relevant at least for historical reasons. Consequently, the study of the present topic must at least be deemed relevant, however one feels about the appropriateness of Pearson's verification postulate, the concept of Pearson-verification and the distance criterion.

Moreover, the present article extends Pearson's method beyond normal distributions, yielding a simple method, firmly rooted in history, for computing exact p-values and acceptance regions for hypotheses under a large class of probability distributions and of arbitrary dimension. Also, the chi-square test and the method of acceptance intervals by percentiles are shown to be one and the same.

## Acknowledgements

This work was supported by the Swedish Research Council, project 435-210-565.

## References

Bernoulli, J. (1713). *Ars Conjectandi*. Basel: Thurnisiorum Fratrum. English translation by E. D. Sylla, 2006.

Ekström, J. (2011). Mahalanobis' distance beyond normal distributions. UCLA Statistics Preprint.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd. Numerous later editions.

Hald, A. (1990). *History of Probability and Statistics and their Applications Before 1750*. New York: John Wiley & Sons.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *J. Amer. Statist. Assoc.*, *88*, 1242–1249.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc. Lond. A*, *231*, 289–337.

Pearson, K. (1900). On the criterion that a system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Phil. Mag. Ser. 5*, *50*, 157–175.

Pearson, K. (1911). *The Grammar of Science, 3rd ed*. London: Adam and Charles Black.

Plackett, R. L. (1983). Karl Pearson and the chi-squared test. *Internat. Statist. Rev.*, *51*, 59–72.

Stigler, S. M. (1999). *Statistics on the table: the history of statistical concepts and methods*. Cambridge, Mass.: Harvard University Press.

Yule, G. U. (1911). *An introduction to the theory of statistics*. London: Charles Griffin & Co.

UCLA Department of Statistics, 8125 Mathematical Sciences Building, Box 951554, Los Angeles CA, 90095-1554

*E-mail address*: `joakim.ekstrom@stat.ucla.edu`