ON STATISTICAL CRITERIA: THEORY, HISTORY, AND APPLICATIONS

JOAKIM EKSTRÖM

Abstract. A statistical criterion is a convention by which certain values are considered relatively probable and others considered relatively improbable. Statistical criteria play a crucial role in the theory of statistics and were originally introduced by Daniel Bernoulli and later independently proposed by Karl Pearson and Ronald Fisher. This article discusses the theory and history of statistical criteria, in particular the density criterion and the distance criterion. Applications for statistical hypothesis generation and testing are discussed. The pedagogical value of statistical criteria is illustrated through a concise and simple explanation of statistical classification. This article also contains discussions on Gauss' least squares conjecture and Fisher's maximum likelihood.

1. Introduction

The founder of the original academy, Plato, was adamant that hypotheses only can be verified through logical derivation within an axiomatic system. Yet, at present a substantial portion of results published by the members of the scientific community are based on experiments, empirical observations, and data. Statistical criteria play a crucial, albeit under appreciated, role in the advancement of science through empirical observation. The present article aims to discuss thoroughly what they are and the role that they play, in terms of theory and history, and illustrate their use through some applications.

In September 1657, Dutch polymath Christiaan Huygens published *De Ratiociniis in Ludo Aleæ*, a thirteen page text on the mathematically correct valuation of games, lottery tickets, and the like. The text, while largely recreational in nature, was the first published work on probability theory (Hald, 1990). At some point in the following decades, Jakob Bernoulli read Huygens' text and subsequently imagined an entirely different use of the theory; combining it with empirical observations to produce a theory of the correct valuation of empirical evidence. The impact that Huygens text had on Bernoulli's thinking was so great that the text was reprinted it in its entirety as part of Bernoulli's last, greatest work, *Ars Conjectandi* (1713).

Bernoulli's idea of combining empirical observations with the concept of probability, later independently proposed by Karl Pearson (1892), launched a new era in the history of science. It provided the members of the scientific community with an alternative method for testing hypotheses, a method which as of present is utilized in nearly every scientific discipline. The essence of the idea is, in Pearson's wording, that a hypothesis is considered verified, or Pearson-verified, if it is demonstrated as overwhelmingly probable. And it is in the evaluation of probability that statistical criteria play a crucial role.

Evaluating the probability of a hypothesis corresponds to evaluating the probability of a proposition $x \sim \mathcal{F}$; whether the value x is an observation of a random variable with probability distribution $\mathcal{F}$ (see Section 3 for a detailed discussion). The probability of the proposition, $\text{Prob}(x \sim \mathcal{F})$, can be evaluated directly through $\text{Prob}(u = x) = \mathbb{P}(\{x\})$, where u is a random variable with distribution $\mathcal{F}$, and, accordingly, $\mathbb{P}$ the probability measure under $\mathcal{F}$. However, often this expression is identically zero and therefore of no use. For example, if $x \in \mathbb{R}^p$ and $\mathbb{P}$ is absolutely continuous with respect to the Lebesgue measure, then $\mathbb{P}(\{x\}) = 0$ for all x, and hence the expression is useless. This conundrum gives rise to the need for a convention on which values that are considered relatively probable vis-à-vis other values, i.e. a statistical criterion.

A statistical criterion is a convention by which certain values are considered relatively probable and others considered relatively improbable. At present, there are two statistical criteria in widespread use: the density criterion, first proposed by Daniel Bernoulli (1778), and the distance criterion, first proposed by Karl Pearson (1900).

The density criterion utilizes probability density. If the probability density at a value is relatively high then the value is considered relatively probable, and if the density at the value is relatively low then the value is considered relatively improbable. The density criterion has been formalized as follows.

Proposition (The density criterion)**.** *Suppose x is a value and $\mathcal{F}$ a probability distribution with density function f. If the density at x, $f(x)$, is low, then the proposition $x \sim \mathcal{F}$ is deemed improbable.*

A density function maps all elements of its domain into the non-negative real numbers. Since the non-negative real numbers is a totally ordered set, all values can be ordered in terms of their density. Use of the density function for the purpose of a statistical criterion also yields many other desirable properties, which are discussed in Section 4.

The distance criterion utilizes Mahalanobis distance. Originally, Pearson (1900) used his chi-distance but that distance has since merged into the more general Mahalanobis distance. If a value is relatively close to the distribution reference point, which is typically the median of the distribution, then the value is considered relatively probable. If the value is relatively far from the reference point, then the value is interpreted as being on the relative periphery of the distribution and is hence deemed improbable. The distance criterion has been formalized as follows.

Proposition (Pearson's distance criterion)**.** *Suppose x 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 $d(x, m)$ is great, then the proposition $x \sim \mathcal{F}$ is deemed improbable.*

Like a density function, a distance maps its domain into the non-negative real numbers. Therefore all values can be ordered in terms of their Mahalanobis distance to the distribution reference point. Use of the Mahalanobis distance for the purpose of a statistical criterion also yields a number of desirable properties, see Section 4.

Since both density functions and distances map to the non-negative reals, the two statistical criteria are easily interchangeable. More precisely, a statistical criterion can be used as a modular, exchangeable component of a statistical analysis, both for statistical hypothesis testing and statistical hypothesis generation. As an example, Section 3 discusses Pearson's statistical hypothesis test under the density criterion. Under certain conditions the two statistical criteria yield identical results in statistical hypothesis generation and testing, and such conditions are detailed in Section 5.

The present article derives properties of the two statistical criteria, reviews history, proves Gauss' least squares conjecture and discusses Fisher's maximum likelihood. Additionally, the inherent pedagogical value of the two statistical criteria is illustrated through an example: a concise and simple explanation of the method called statistical classification.

2. Origins and history

The first proposal of a density criterion found in the literature is D. Bernoulli (1778). Bernoulli proposed using as a point estimate, rather than the arithmetic mean, whichever value is determined to be most probable under the density criterion. Bernoulli assumed that the real valued observational errors were statistically independent and had the semicircle probability distribution, whose density function f is given by

$$f(x) = c\mathbb{1}_{[-r,r]}(x)\sqrt{r^2 - x^2},$$

where c a normalizing constant, r the greatest possible observational error and $\mathbb{1}_A$ the indicator function of the set A. In appearance, the graph of the density function is akin to a semi-circle; hence the name. If there are three or more observations, then the arithmetic mean is generally not the most probable value under the density criterion.

In an editorial comment, L. Euler noted that analytically solving for the most probable value amounts to finding roots of polynomials of a degree nearly twice the number of observations, and hence Bernoulli's method was at the time not practically feasible. In the eighteenth century polynomial roots could not be effortlessly found through computer assisted numerical optimization.

Gauss (1809) applied the density criterion for the determination of the most probable Kepler orbit given observations of a heavenly body. Unlike Bernoulli, Gauss assumed that the observational errors were normally distributed, an assumption which makes analytical optimization considerably simpler. Specifically, solving for the value most probable under the statistical criterion amounts to minimizing a sum of squares, a minimum which Legendre (1805) had shown equals the solution of his system of normal equations.

While acknowledging that the arbitrary normal distribution assumption constituted a weakness of his work, Gauss claimed that the method of least squares, which arose as a result, also yields the most probable value under other distributional assumptions. Section 6 of the present article discusses the claim, named *Gauss' least squares conjecture*, including specification of necessary and sufficient conditions. Throughout the nineteenth century, the method of least squares was accepted as a practical and principally sound method for generating estimates, regardless of probability distribution (see, e.g., Airy, 1875).

The first proposal of the distance criterion is Pearson (1900), which uses the statistical criterion for the purpose of the statistical hypothesis test defined in the same article. Pearson (1900) contains few explanatory wordings, but Mahalanobis (1936), which generalizes Pearson's chi-distance, uses a parallel with Galilean transformations in physics to explain the distance: by transforming arbitrary probability distributions into the standard normal distribution, the probability distributions can be evaluated within a frame of reference, and thus much of the complexity is circumvented; see Ekström (2011a) for a discussion and a generalization beyond the normal distributions.

Fisher (1912) proposes the density criterion anew. Its statistic was later termed likelihood (see, e.g, Fisher, 1922), and the method has become most well known under related names. Fisher applies the criterion in a way that is different from that of Bernoulli (1778) and Gauss (1809). In short, Bernoulli and Gauss maximize the value of the density function while Fisher maximizes the density function of the value; Section 7 of the present article discusses this difference in application.

Throughout the literature there are numerous instances in which the probability density of a value is interpreted, or explained, as the probability of the same value. That interpretation, however, stretches the truth to quite an extent; a probability density arises from differentiation of the probability measure, but is not a measure itself. In short, interpreting a probability density as a probability is factually erroneous, and can easily set the stage for a number of misunderstandings.

3. Pearson-verification under the density criterion

The present section discusses Pearson's statistical hypothesis test (1900) under the density criterion, as opposed to the distance criterion which Pearson originally proposed. Definitions are modified to accommodate the density criterion.

The fundamentals of Pearson's statistical hypothesis test are the following. Suppose $x_1, \ldots, x_n$ are observations of the phenomenon of interest, that the observations are elements of some topological space denoted $\mathbb{X}$, and that the observations have Gauss-Pearson decomposition $x_i = \mu_i + u_i$, for $i = 1, \ldots, n$, where μ_i and u_i are the ideal and random parts, respectively, of the observation x_i, and $+$ is a binary operation such that $(\mathbb{X}, +)$ is a group. The Euclidean space, $\mathbb{R}^p$, was used in Pearson (1900), i.e. $\mathbb{X} = \mathbb{R}^p$. For convenience, let the arrow accent denote sequences of length n, e.g. $\vec{x} = (x_1, \ldots, x_n)$, which are added through component-wise addition. Thus the observation $\vec{x} \in \mathbb{X}^n$, sometimes referred to as the sample point, has Gauss-Pearson decomposition $\vec{x} = \vec{\mu} + \vec{u}$. Also, the notation $\mathcal{L}(\vec{u})$ is convenient, which denotes the probability distribution, or the *law*, of the random variable $\vec{u}$. The distribution $\mathcal{L}(\vec{u})$ is assumed ex ante known.

Next, consider the hypothesis that the ideal part of the observed phenomenon, $\vec{\mu}$, equals some given sequence $\vec{v} \in \mathbb{X}^n$. The hypothesis yields the representation $\vec{x} = \vec{v} + \vec{e}$, where $\vec{e}$ is the representation residual. Hence there are two representations of the observation, and the two constitute the following system:

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

Since $\mathbb{X}^n$ is a group, algebraic manipulation of the system yields

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

Therefore, if the proposition $\vec{e} \sim \mathcal{L}(\vec{e})$ is Pearson-falsified (at statistical significance level α) then it follows through logical deduction, specifically contraposition, that the hypothesis, $\vec{\mu} = \vec{v}$, is Pearson-falsified (at statistical significance level α). Note that if e is an observation of a random variable then the notation $e \sim \mathcal{L}(u)$ is shorthand for the more cumbersome *e is an observation of a random variable with distribution $\mathcal{L}(u)$*.

Pearson (1900) proposes the distance criterion for the purpose of determining whether it is probable that the representation residual $\vec{e}$ has distribution $\mathcal{L}(\vec{u})$, or in his own wording: whether it can be reasonably supposed that $\vec{e}$ has arisen from random sampling. Pearson's article and his statistical hypothesis test is reviewed in modern notation and rigor in Ekström (2011b), making detailed discussions in the present text superfluous. In short, though, Pearson (1900) defines the chi-statistic $\eta = d(\vec{e}, \vec{m})$, where d is the Mahalanobis distance under $\mathcal{L}(\vec{u})$ and $\vec{m}$ the distribution's reference point, and evaluates its relative size through the p-value, $P = \mathbb{P}(B_\eta(\vec{m})^c)$, where $\mathbb{P}$ is the probability measure under $\mathcal{L}(\vec{u})$ and B the Mahalanobis ball.

In the following, Pearson's statistical hypothesis testing framework is studied under, and modified to accommodate, the density criterion. Under the density criterion, the analogue of the chi-statistic is the following.

Definition 1. Assuming that the observations have Gauss-Pearson decomposition $\vec{x} = \vec{\mu} + \vec{u}$, where the random part $\vec{u}$ has an ex ante known distribution $\mathcal{L}(\vec{u})$ with density function f, the *density statistic* ζ under the hypothesis $\vec{\mu} = \vec{v}$ is defined

$$\zeta = f(\vec{e}),$$

where $\vec{e}$ is the residual of the representation $\vec{x} = \vec{v} + \vec{e}$.

If the density statistic, ζ, is small, then by the density criterion the proposition $\vec{e} \sim \mathcal{L}(\vec{u})$ is deemed improbable, and if it is large the proposition is not deemed improbable. In particular, if the density at the hypothesis representation residual, $\vec{e}$, is zero then the proposition is deemed to be at the extremity of improbability. Determination of whether the density statistic is small or large is made through the p-value.

Let $(f > t)$ denote the set $\{x \in \mathbb{X}^n : f(x) > t\}$, for some $t \in \mathbb{R}$. In the case $\mathbb{X}^n = \mathbb{R}^2$, the set can be visualized by thinking of the graph of f as a landscape, which then is flooded up to level t; the set $(f > t)$ corresponds to the area of the islands that are above water. Under the density criterion, $(f > \zeta)$ is the set of points more probable than ζ while its complement is the set of points as or less probable than ζ. Consequently, the p-value under the density criterion is defined as follows.

Definition 2. In the notation and context of Definition 1, the *p-value* P is defined

$$P = \mathbb{P}((f > \zeta)^c),$$

where $\mathbb{P} : \mathcal{B}(\mathbb{X}^n) \to \mathbb{R}$ is the probability measure under $\mathcal{L}(\vec{u})$.

While the p-value under the distance criterion can be expressed as a percentile of the chi-square distribution, it is in general more difficult to express the p-value under the density criterion in closed form. However, the p-value under the density criterion can easily be approximated numerically, for example through monte carlo integration. The definition of acceptance regions under the density criterion is immediate.

Definition 3. In the notation and context of Definition 2, the *acceptance region at statistical significance level* α, A, is defined

$$A = (f > t),$$

where t is the solution of the equation $\mathbb{P}((f > t)^c) = \alpha$.

Solving the equation $\mathbb{P}((f > t)^c) = \alpha$ for t is relatively easy since the left hand side is a non-decreasing real-valued function of $t \in \mathbb{R}$ that can be numerically approximated. If the equation does not have a solution, which it need not have, it is prudent to choose the greatest smaller statistical significance level $\tilde{\alpha}$ for which a solution exists, or equivalently let the acceptance region be defined by the greatest t that satisfies the inequality $\mathbb{P}((f > t)^c) \leq \alpha$.

The following are desirable properties of acceptance regions under the density criterion. At all non-zero statistical significance levels every acceptance region is contained in the support of the probability measure, $\mathbb{P}$. Conversely, if the representation residual $\vec{e}$ is not an element of the support of $\mathbb{P}$, then the proposition $\vec{e} \sim \mathcal{F}$ is Pearson-falsified at every non-zero statistical significance level. Additionally, if the distribution, $\mathcal{L}(\vec{u})$, has one or more point probabilities, i.e. if the probability measure, $\mathbb{P}$, has a singular part, then those points are included in every acceptance region since the density at those points are infinity.

Acceptance regions under the two statistical criteria are identical if $\mathcal{L}(\vec{u})$ is strictly unimodal elliptical with a density function, as per Theorem 7, Section 5. Notably, Pearson (1900) assumes that the random part is normally distributed, and as a result the choice of statistical criterion, i.e. whether to use the chi-statistic or the density statistic, is immaterial under Pearson's assumptions.

4. Comparison of statistical criteria

The existence of two statistical criteria upon which statistical hypothesis tests can be based naturally leads to the question: Is one statistical criterion better than the other? Unfortunately, the question does not have a simple answer; each of the two have both desirable and undesirable properties. The present section aims to summarize some of the properties of the statistical hypothesis test under the two statistical criteria.

Under the distance criterion acceptance regions are connected, while under the density criterion acceptance regions need not be. Connectedness under the distance criterion follows from the homogeneity property of Mahalanobis balls and the fact that continuity of Mahalanobis transformations is necessary for uniqueness of the distance. A counter-example that shows that acceptance regions under the density criterion need not be connected is the distribution that is uniform on the set $\cup_{n=1}^{\infty}[n, n+2^{-n}] \subset \mathbb{R}$. In fact, for this distribution, each acceptance region under the density criterion has infinitely many connected components.

Under the density criterion, acceptance regions are contained in the support of the density function, while acceptance regions under the distance criterion need not be. This property follows immediately from the fact that acceptance regions under the density criterion are defined by $(f > t)$ for non-negative t. A counter-example that shows that acceptance regions under the distance criterion not need be contained in the support of the density function is the distribution of the preceding paragraph, since acceptance regions under the distance criterion are connected.

Because of the homogeneity property of Mahalanobis balls, acceptance regions under the distance criterion are preserved under suitable transformations T in the sense that if A_1 is the acceptance region for U and A_2 is the acceptance region for $T(U)$ then $A_2 = T(A_1)$. This property is in many cases convenient, and is also intuitively desirable. This property does not hold under the density criterion; a counter-example is the transformation $x \mapsto x^2$ applied to a standard uniform random variable.

The following property was discussed by Neyman & Pearson (1933), and is referred to by names such as best critial region or most efficient/powerful test. An accurate and more informative name of the property is that acceptance regions under the density criterion have *least Lebesgue measure*. The following theorem details this property; for enhanced readability the proof is in the Appendix.

Theorem 1. *Suppose f is a probability density function, λ the Lebesgue measure, and let the probability measure $\mathbb{P}$ be defined by $\mathbb{P}(A) = \int_A f d\lambda$. If $A = (f > t)$, for some t, and B is a set satisfying $\mathbb{P}(A) = \mathbb{P}(B)$, then $\lambda(A) \leq \lambda(B)$.*

Note that when the density function is defined with respect to a measure other than the Lebesgue measure, then Theorem 1 holds with the Lebesgue measure substituted for the other measure. The event in which a random variable with a density function different than f attains an element of the acceptance region is by Neyman & Pearson (1933) referred to as an error of second type, and the probability that it does not attain such an element is referred to as statistical power. The following corollary implies that acceptance regions under the density criterion have greatest statistical power in a certain sense.

Corollary 2. *In the notation of Theorem 1, if V is a random variable with constant probability density on $A \cup B$, then* $\operatorname{Prob}(V \in A) \leq \operatorname{Prob}(V \in B)$.

Under the distance criterion, the p-value can be expressed as a value of the noncentral chi-square distribution function (see Ekström, 2011b), while the p-value under the density criterion in general must be numerically approximated. This fact is a theoretical and practical advantage that the distance criterion has relative to the density criterion.

An advantage that the density criterion has relative to the distance criterion is that it is a simpler concept. Under the distance criterion there are sometimes issues relating to existence and uniqueness of the Mahalanobis distance. Use of the distance criterion also requires specification of a distribution reference point, which use of the density criterion does not. On the other hand, the reference point is ensured to be an element of every acceptance region under the distance criterion while it need not be an element of acceptance regions under the density criterion.

In summary, each of the two statistical criteria have properties that are desirable. Under certain conditions the acceptance regions yielded by the two statistical criteria are equal, see Section 5. If the two acceptance regions are not equal, an acceptance region with a mix of properties can be constructed by intersecting the acceptance regions yielded by the two statistical criteria; the intersection has a statistical significance level between α and 2α.

5. Conditions for equivalence of the statistical criteria

The contours of the normal distribution's density function concur with the distribution's Mahalanobis spheres. As a consequence, given a normal distribution assumption all p-values and acceptance regions are identical under the density criterion and the distance criterion, making the choice of statistical criterion immaterial. The present section aims to derive more general conditions sufficient for the two statistical criteria to yield equal acceptance regions and p-values.

Assuming that $\mathbb{X}$ is a Euclidean space, a distribution is *spherical* if it is radially symmetric about the origin, 0. In greater detail, if v is a random variable, then $\mathcal{L}(v)$ is spherical if v is equal in distribution to rs where r and s are two statistically independent random variables with ranges $[0, \infty)$ and $\{x \in \mathbb{X} : \|x\| = 1\}$, respectively. In particular, if v has a density function, f, then the density of every ray $f(tx/\|x\|) = f_r(t)$, $x \neq 0$, $t \geq 0$, is identical. The distribution function, F, of r consequently satisfies $F(t) = \int_0^t f_r(s)ds$. Furthermore, a random variable is elliptically distributed, i.e. its distribution is *elliptical*, if it is equal in distribution to an affine transformation of a spherically distributed random variable.

If a spherical distribution has a density function, then its Mahalanobis distance exists. Because the standard normal distribution is spherical it is natural to use a Mahalanobis transformation that is radially symmetric about the origin, i.e. one that acts on all rays $tx/\|x\|$, $x \in \mathbb{X}$, $x \neq 0$ and $t \geq 0$, equally. Using the expression $v = rs$ of the preceding paragraph, such a Mahalanobis transformation T can be written $T(v) = g(r)R(s)$, where R is orthogonal and g a real-valued function. Under necessary Mahalanobis uniqueness conditions g equals the composition $G^{-1} \circ F$, where F is the distribution function of the random variable r and G is the chi distribution function with one degree of freedom (see Ekström, 2011a, Theorem 11). Hence the radially symmetric Mahalanobis distance satisfies

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

where $T(0) = 0$ and $T(x) = G^{-1} \circ F(\|x\|)x/\|x\|$ for $x \neq 0$. If the elliptically distributed random variable w satisfies $w = L(v) + b$, for some linear and injective $L : \mathbb{X} \to \mathbb{X}$ and $b \in \mathbb{X}$, then the Mahalanobis transformation $T \circ L^{-1}(w - b)$ is natural. For example $L(v) = \mathrm{Var}(w)^{1/2}v$ and $b = \mathrm{E}(w)$ satisfy $w = L(v) + b$.

The following three lemmas contain facts related to elliptical distributions, and they are of use in the derivation of the main theorems of this section. For enhanced readability the proofs are in the Appendix.

Lemma 3. *Suppose the distribution $\mathcal{F}$ is spherical and has a density function, and let $B_r(m)$ and $E_r(m)$ denote the Mahalanobis and Euclidean balls respectively, then for every r,*

$$B_r(0) = E_{\hat{r}}(0),$$

where $\hat{r} = F^{(-)} \circ G(r)$ and $F^{(-)}$ is the pseudoinverse defined $F^{(-)}(y) = \inf\{x : F(x) = y\}$.

Lemma 4. *Suppose that the random variable U has density function f and that T is an injective affine transformation, then it holds*

$$T((f > t)) = (\hat{f} > \tilde{t}),$$

where $\hat{f} = |\det(T)|^{-1}(f \circ T^{-1})$ is the density function of $T(U)$ and $\tilde{t} = |\det(T)|^{-1}t$.

Lemma 5. *Suppose $\mathcal{F}$ is an elliptical distribution with a density function, then its density function is constant on Mahalanobis spheres centered at the median.*

A distribution is *strictly unimodal* if it has a density function that has no local optimum on its support except for its mode, m. If $\mathbb{X}$ is a linear space over $\mathbb{R}$, then the condition is equivalent to the density of each ray originating from the mode, $t \mapsto f(m + t(x - m))$, $t \geq 0$ and $x \neq m$, being strictly decreasing. A distribution is *unimodal* if the density of every ray originating from the mode is non-increasing. A distribution which is strictly unimodal is also unimodal but the converse need not hold.

The following lemma contains a property of strictly unimodal distributions; the proof is in the Appendix.

Lemma 6. *Suppose $\mathcal{F}$ is a strictly unimodal distribution with mode m and density function f, then*

$$(f = t) \subset \partial(f > t) = \partial(f < t),$$

for all $0 < t < f(m)$.

Under the distance criterion an acceptance region is defined as a Mahalanobis ball, $B_r(m)$, and under the density criterion an acceptance region is defined as a set $(f > t)$, see Definition 3. The following theorem, a main theorem of the section, details conditions under which interchanging the two statistical criteria will not change the acceptance region, implying that the choice of criterion is immaterial.

Theorem 7. *Suppose $\mathcal{F}$ is a unimodal elliptical distribution with density function f, then for each $t > 0$ there is an r such that*

$$\mathrm{Int}(f > t) = B_r(m),$$

where m is the median of $\mathcal{F}$ and $B_r(m)$ the Mahalanobis ball under $\mathcal{F}$. Furthermore, if $\mathcal{F}$ is strictly unimodal then for each $r > 0$ there is a t such that the equality holds.

The normal distribution, which was assumed in Pearson (1900), is an example of a strictly unimodal elliptical distribution. The next theorem is related to Theorem 7 but applies to statistical hypothesis generation under the two criteria.

Theorem 8. *Suppose $\mathcal{F}$ is a unimodal elliptical distribution with density function f and median m, then for any set $A \subset \mathbb{X}$,*

$$d(a,m) = \inf_{x \in A} d(x,m) \implies f(a) \geq \sup_{x \in A} f(x).$$

Furthermore, if $\mathcal{F}$ is strictly unimodal then

$$f(a) = \sup_{x \in A} f(x) \implies d(a,m) = \inf_{x \in A} d(x,m).$$

In traditional statistics, the subsets $A \subset \mathbb{X}$ are commonly lines, planes, hyperplanes, or graphs of arbitrary functions, sometimes referred to as non-linear models. If the dimension of A is less than that of $\mathbb{X}$, then projection onto A is commonly referred to as dimension reduction. The application of the two statistical criteria for the purpose of statistical hypothesis generation is also discussed in Section 6.

6. Special topic: Gauss' least squares conjecture

In *Theoria Motus Corporum Coelestium* (1809), Gauss sought to determine which Kepler orbit is most probable under the density criterion given observations of a heavenly body, and it was shown that the method of weighted least squares maximizes density under an uncorrelated, mean zero joint normal distribution. Gauss acknowledged that the arbitrary normal distribution assumption constituted a weakness of his work, but claimed that the method of least squares yields the most probable Kepler orbit also when the observational errors are not normally distributed. Specifically, the claim is that whether the density function of the observational errors is exactly equal to the normal density function is of no importance in practice (Gauss, 1809, §178), and therefore the principle of least squares must, everywhere, be considered an axiom (Gauss, 1809, §179).

The claim that minimizing squares also maximizes density is referred to as Gauss' least squares conjecture. The fact that Gauss did not present any supporting evidence has been noted by historians and even raised a few eyebrows. In his defense, though, it should be noted that §§174-5 of Gauss (1809) specify that the observational errors can be assumed to have a unimodal distribution that has an even density function, i.e. one that satisfies $f(-x) = f(x)$. In the univariate case, the latter condition implies that the distribution is spherical. In the general multivariate case, however, the condition does not suffice for the distribution to be median zero elliptical. Nevertheless, Gauss' assumptions are remarkably close to the necessary and sufficient conditions derived in the present section. For ease of reference, Gauss' least squares conjecture is formalized below with the corresponding claim relative to the distance criterion incorporated in parentheses.

Conjecture (Gauss' least squares conjecture)**.** *The method of generalized least squares is optimal under the density (distance) criterion, i.e. it yields a value that maximizes (minimizes) the density (chi) statistic.*

Suppose $\mathbb{X}$ is a Euclidean space. Each median zero, elliptically distributed random variable w with a density function can be expressed as an injective linear transformation $L(v)$ of some spherically distributed random variable v. The linear transformation L is sometimes given, otherwise $L(v) = \text{Var}(w)^{1/2}v$ can be taken, or any non-zero scalar multiple thereof. Further, note that if $\|\cdot\|$ denotes the Euclidean norm then $\|L^{-1}(x)\| = \|x\|_\star$ is also a norm. Given the median zero elliptical distribution $\mathcal{L}(w)$, minimizing $\|\cdot\|_\star$ is referred to as the method of generalized least squares.

Because of mathematical convenience, the distance criterion version of Gauss' least squares conjecture is treated firstly, facilitating treatment of Gauss' original least squares conjecture thereafter. The following theorem details sufficient conditions for a value that is optimal under the method of generalized least squares to be optimal also under the distance criterion; the proof is in the Appendix.

Theorem 9. *Suppose $\mathcal{F}$ is a median zero elliptical distribution with density function f, then for any $A \subset \mathbb{X}$*

$$\|a\|_\star = \inf_{x \in A} \|x\|_\star \implies d(a, 0) = \inf_{x \in A} d(x, 0),$$

where d denotes the Mahalanobis distance under $\mathcal{F}$. Furthermore, if $a \in \text{Int}(\text{supp}(f))$, then

$$\|a\|_\star = \inf_{x \in A} \|x\|_\star \iff d(a, 0) = \inf_{x \in A} d(x, 0).$$

The following theorem is similar to Theorem 9, but details the method of generalized least squares vis-à-vis optimization under the density criterion.

Theorem 10. *Suppose $\mathcal{F}$ is a unimodal median zero elliptical distribution with density function f, then and only then for any $A \subset \mathbb{X}$*

$$\|a\|_\star = \inf_{x \in A} \|x\|_\star \implies f(a) \geq \sup_{x \in A} f(x).$$

Furthermore, if $\mathcal{F}$ is strictly unimodal and $A \cap \text{Int}(\text{supp}(f)) \neq \varnothing$, then

$$f(a) = \sup_{x \in A} f(x) \implies \|a\|_\star = \inf_{x \in A} \|x\|_\star.$$

The following corollary summarizes the above theorems in the context of Gauss' least squares conjecture.

Corollary 11. *The distance criterion version of Gauss' least squares conjecture holds if the distribution of the random part is median zero elliptical and has a density function. Gauss' least squares conjecture, the original density criterion version, holds if and only if the distribution of the random part is median zero elliptical, unimodal and has a density function.*

Gauss (1809) contains an additional claim: that if observations are statistically independent then the method of generalized least squares is optimal under the density criterion. However, even if statistically independent random variables are elliptically distributed their joint distribution is typically not elliptical, and thus by Corollary 11 the claim is false. The normal distribution is a notable exception, in which the joint distribution of statistically independent random variables is elliptical.

7. Special topic: Fisher's maximum likelihood

Fisher (1912) proposes the density criterion anew, but suggests a way of applying the statistical criterion that is different from that of Bernoulli (1778) and Gauss (1809). This section discusses the differences between the two ways of applying the density criterion.

The problem considered by Fisher (1912) is determination of the probability distribution of a sample of statistically independent and identically distributed real-valued observations $x_1, \ldots, x_n \in \mathbb{R}$. This particular problem is well-suited for illustration of the differences between the two ways of application. In the following, Bernoulli and Gauss' way of applying the density criterion is discussed firstly, and then compared to Fisher's way of applying the density criterion.

The premise of Theoria Motus (Gauss, 1809) is Gauss' observation postulate, i.e. that each observation consists of two parts: the true value of the observed phenomenon and an observational error. The empirical distribution function of the sample is an unbiased observation of the postulated true distribution function. The empirical distribution function is sometimes seen as an element of the Euclidean space $\mathbb{R}^n$, but in the present discussion regarded as an element of Skorokhod's topological space D of right-continuous functions with left-hand limits. Let y denote the empirical distribution function, and $y = \mu + u$ its Gauss-Pearson decomposition, where μ is the ideal part, i.e. the true distribution function, and u the random part, i.e. the observational error. Given any ideal part, the distribution of the random part, $\mathcal{L}(u)$, can be derived analytically, or approximated through simulation or through Donsker's theorem.

Bernoulli and Gauss' way of applying the density criterion is to determine the most probable ideal part, out of some subset $A \subset D$, through

$$\arg\max_{z \in A} f(-z + y),$$

where f is the density function of $\mathcal{L}(u)$. Fisher's way of applying the density criterion is quite different; the original sample point $\vec{x}$ is assumed free of observational error and the probability distribution determined through the optimization

$$\arg\max_{\theta\in\Theta} f_\theta(\vec{x}),$$

where $\{f_\theta\}_{\theta\in\Theta}$ is a collection of density functions and Θ its index set. In short, Bernoulli and Gauss apply the density criterion to find the value of greatest density, while Fisher applies the density criterion to find the density function that is greatest at the observed value. Put differently, Bernoulli and Gauss use a known density function to find an unknown value, while Fisher uses a known value to find an unknown density function.

Fisher's assumption of an error free observation makes his way of application difficult to reconcile with the broader body of statistical methods. The methods of Gauss and Pearson, for statistical hypothesis generation and testing, are built on Gauss' observation postulate by which the observation is assumed to consist in part of an observational error that is unavoidable and impossible to eliminate. Fisher's way of application does not allow for a Gauss-Pearson decomposition of the observation since the observation is assumed error free and hence non-random.

While fundamentally different, the two ways of applying the density criterion have co-existed in relative harmony throughout the twentieth century. This fact can to a large extent be explained by a single special case in which the two ways of application happen to yield equal results. Consider the collection $\{f_\theta\}$ of normal density functions with unit variance, indexed by their means. The collection's index set satisfies $\Theta = \mathbb{R}^p = \mathbb{X}$ and it holds that

$$f_\theta(x) = f(x-\theta) = f_x(\theta),$$

where f is the standard normal density function, i.e. the index and the value are elements of the same set, and they are interchangeable. As a result, for any $A \subset \Theta = \mathbb{X}$ it holds

$$\arg\max_{\theta\in A} f_\theta(x) = \arg\max_{\theta\in A} f(-\theta + x),$$

and thus the two ways of applying the density criterion yield a common optimum. Further, by Corollary 11 the method of generalized least squares yields this optimum, regardless of variance, and hence the two ways of application produce identical results under this distributional assumption. Because of the normal distribution assumption's historical prevalence, the special case can to an extent explain why the two different ways of applying the density criterion have co-existed largely without troubles.

With respect to statistical hypothesis testing, the differences between the two ways of applying the density criterion are also considerable. Statistical hypothesis testing under

Bernoulli and Gauss' way of application utilizes the identity

$$\int_{\mathbb{X}} f(x) d\lambda(x) = 1,$$

presuming the density function is defined with respect to λ, the Lebesgue measure. The identity is an immediate consequence of Kolmogorov's axioms. The integral over a subset $A \subset \mathbb{X}$ is interpreted as the probability of the random variable attaining an element of A, a fact that is essential to the construction of p-values and acceptance regions under the density criterion (cf. Section 3). The corresponding integral under Fisher's way of application is

$$\int_{\Theta} f_\theta(x) d\xi(\theta),$$

i.e. integration over the collection $\{f_\theta\}$. In general, it cannot be assumed that the integral, if it exists, is constant over x, and moreover the choice of measure, ξ, is unclear. The meaning of the integral is also difficult to interpret; integration over a subset of the index set Θ cannot be interpreted as a probability, a fact noted by Fisher (1912). Because of these reasons, the integral cannot be used for construction of statistical hypothesis tests under Fisher's way of applying the density criterion. Furthermore, while Bernoulli and Gauss' way of application yields an acceptance region that is a subset of $\mathbb{X}$, the sample space, Fisher's way of application yields an acceptance region that is a subset of Θ, the index set of the collection of density functions, which complicates conceptualization of the acceptance region.

However in another remarkable coincidence, it was discovered that if $\{f_\theta\}$ is again taken to be the collection of normal density functions, with covariance matrix Σ and indexed by their means, then it holds that

$$-2 \log f_\theta(x)/f_x(x) = (x-\theta)^{\mathrm{t}} \Sigma^{-1} (x-\theta) = \eta^2,$$

i.e. a log likelihood ratio expression equals the squared chi-statistic. Consequently, under the hypothesis $\theta \sim \mathrm{N}(x, \Sigma)$ a likelihood ratio expression has a known distribution. Since the index and the value are interchangeable in this case, $f_\theta(x) = f_x(\theta)$, the hypothesis $\theta \sim \mathrm{N}(x, \Sigma)$ is equivalent to $x \sim \mathrm{N}(\theta, \Sigma)$, and it follows through Theorem 7 that the acceptance regions in this case are identical under both ways of application (cf. Section 3).

In general, however, the log likelihood ratio does not have a known distribution, and consequently the type one error probability cannot be controlled. As a fall back option, it is commonly assumed that the above log likelihood ratio expression is chi-square distributed, often using an asymptotic result known as Wilks' theorem as rationale (see Ferguson, 1996, for a detailed list of required assumptions). While the chi-square distribution could be a decent approximation of the log likelihood ratio statistic's distribution in some instances, the lack of control makes the construction unsatisfactory nevertheless.

The possibility of computing exact p-values under the Bernoulli and Gauss way of application is a decided advantage, and it unavoidably raises the question of whether the likelihood ratio test should even be used at all.

In this context a remark on terminology is appropriate. Sometime in between years 1912 and 1922, Fisher started referring to the density statistic by the term likelihood. In *Statistical Methods for Research Workers* (1928), Fisher explains that due to his rejection of the theory of inverse probability, also called Bayesian statistics, he used the term likelihood rather than probability, so to avoid the latter word. The present article uses the term density statistic (cf. Definition 1), a term that is both accurate and informative.

In summary, the way of applying the density criterion suggested by Fisher (1912) is a doable way of determining the probability distribution of a sample of statistically independent and identically distributed observations, in that it produces an expression that can be optimized to yield a density function. But because of its error free observation assumption, it fits the broader body of statistical methods poorly. Fisher's way of application is particularly ill-suited for statistical hypothesis testing since the type one error probability generally cannot be controlled.

8. Example: Statistical classification

The two statistical criteria discussed in the present article have an inherent pedagogical value that reaches beyond theoretical statistics. This section aims to exemplify that value by discussing the statistical technique known as discriminant analysis or statistical classification.

Many textbooks on multivariate analysis, such as Mardia et al. (1979) and Muirhead (1982), contain a chapter on statistical classification. Through utilization of the two statistical criteria, the technique can easily be explained in one sentence: Let x be a value, $\{\mathcal{F}_\theta\}$ a set of distributions with density functions $\{f_\theta\}$ and reference points $\{m_\theta\}$, and suppose that it is ex ante known that x is an observation from one of the distributions, then using the distance criterion x is classified as an observation from the distribution with least distance, $d_{\mathcal{F}_\theta}(x, m_\theta)$, and using the density criterion x is classified as an observation from the distribution with greatest density, $f_\theta(x)$.

Besides the efficient use of words and the impeccable rigor, the utilization of the two statistical criteria also has a certain conceptual beauty. Given a value and a number of distributions, and the ex ante information that the value is an observation from one of the distributions, there is really not much that can be done except for evaluating which distribution is most probable. And for the evaluation a statistical criterion is needed.

It is important that statistical methods are easily understandable because they are routinely used for purposes that have far-reaching consequences. If the methods are easily understandable, then the risks of misunderstandings and unintentional errors are reduced. In the teaching environment, taking the time to explain the two statistical criteria, laying the groundwork so to speak, will simplify explanation of most statistical methods, be they hypothesis generation, hypothesis testing, statistical classification or almost any other statistical method.

Acknowledgements

This work was supported by the Swedish Research Council, project 435-210-565. The author is grateful to Andrew Bray for valuable comments and suggestions for improvements.

References

Airy, G. B. (1875). *Algebraical and numerical theory of errors of observations and the combination of observations*. London: MacMillan and Co.

Bernoulli, D. (1778). Diiudicatio maxime probabilis plurium obseruationum discrepantium atque verisimillima inductio inde formanda. *Acta Acad. Scient. Imper. Petrop., pro Anno 1777, Pars prior*, 3–23. English translation by C. G. Allen, 1961.

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

Ekström, J. (2011a). Mahalanobis' distance beyond normal distributions. *UCLA Statistics Preprints, 624*.

Ekström, J. (2011b). On Pearson-verification and the chi-square test. *UCLA Statistics Preprints, 625*.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. New York: Chapman & Hall.

Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, *41*, 155–160.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. Lond. A*, *222*, 309–368.

Fisher, R. A. (1928). *Statistical Methods for Research Workers, 2nd ed*. Edinburgh: Oliver and Boyd.

Gauss, C. F. (1809). *Theoria Motus Corporum Coelestium in sectionibus conicis solem ambientium*. Hamburg: F. Perthes und I. H. Besser. English translation by C. H. Davis, 1858.

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

Huygens, C. (1657). De ratiociniis in ludo aleæ. In F. van Schooten (Ed.) *Excercitationum Mathematicum, Liber V*, (pp. 521–534). Leiden: Elsevier. English translation by William Browne, 1714.

Legendre, A. M. (1805). *Nouvelles méthodes pour la détermination des orbites des cométes*. Paris: Firmin Didot.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proc. Nat. Inst. Sci., India*, *2*, 49–55.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate Analysis*. London: Academic Press.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. New York: John Wiley & Sons.

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. (1892). *The Grammar of Science, 1st ed*. London: J. M. Dent & Sons Ltd. Reprinted 1937.

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.

Appendix

Proof of Theorem 1. Let $C = A \setminus B$, $D = B \setminus A$ and note that $\mathbb{P}(C) = \mathbb{P}(D)$. Since $C \subset (f > t)$ and $D \subset (f > t)^c$ it holds that $\lambda(C) \leq \mathbb{P}(C)/t = \mathbb{P}(D)/t \leq \lambda(D)$. Thus, $\lambda(A) = \lambda(A \cap B) + \lambda(C) \leq \lambda(A \cap B) + \lambda(D) = \lambda(B)$. □

Proof of Lemma 3. Suppose $x \neq 0$, then the Mahalanobis distance $\|T(x) - T(0)\|$ equals $G^{-1} \circ F(\|x\|)$. Let $B'_r(0)$ be the punctuated Mahalanobis ball with center point 0, then it holds $B'_r(0) = \{x \neq 0 : G^{-1} \circ F(\|x\|) < r\} = \{x \neq 0 : \|x\| < F^{(-)} \circ G(r)\} = E'_{\hat{r}}(0)$, since $G^{-1} \circ F$ is non-decreasing. The balls are empty if and only if $r, \hat{r} \leq 0$, and hence the center point, 0, is an element of both balls when $r, \hat{r} > 0$. This shows the statement. □

Proof of Lemma 4. By the change of variables theorem, $\hat{f} = |\det(T)|^{-1}(f \circ T^{-1})$. Therefore,

$$T((f > t)) = \{T(x) : f(x) > t\} = \{y : f \circ T^{-1}(y) > t\} = ((f \circ T^{-1}) > t) = (\hat{f} > \tilde{t}),$$

which shows the statement. □

Proof of Lemma 5. Every elliptically distributed random variable with a density function is equal in distribution to an injective affine transformation T of a spherically distributed random variable. The homogeneity property of Mahalanobis balls and Lemma 3 yields

$$B_r(m) = T(B_r^{\mathcal{G}}(T^{-1}(m))) = T(B_r^{\mathcal{G}}(0)) = T(E_{\hat{r}}(0)),$$

where $\mathcal{G}$ is the aforementioned spherical distribution, m the median of $\mathcal{F}$ and $E_{\hat{r}}(0)$ a Euclidean ball. By the invariance of domain theorem it follows that $\partial T(A) = T(\partial A)$ for every subset A, and thus $\partial B_r(m) = T(\partial E_{\hat{r}}(0))$. Further, if g denotes the density function of $\mathcal{G}$, then by Lemma 4 $f = c(g \circ T^{-1})$ where c is a normalizing constant. Hence $f(\partial B_r(m)) = c(g \circ T^{-1})(T(\partial E_{\hat{r}}(0))) = cg(\partial E_{\hat{r}}(0))$, which by radial symmetry is a one-point set. □

Proof of Lemma 6. Since $\mathcal{F}$ is strictly unimodal, f does not have any local optimum on $(f > 0)$ except for its mode. Thus every neighborhood of $(f = t)$ intersects both $(f > t)$ and $(f < t)$ and hence every point of $(f = t)$ is a limit point of $(f > t)$ and $(f < t)$. Since the three sets are disjoint, the statement follows. □

Proof of Theorem 7. Every elliptically distributed random variable with a density function can be expressed as an injective affine transformation T of a spherically distributed random variable, and by the change of variables theorem the spherical distribution, $\mathcal{G}$, also has a density function, g. Further, by Lemma 4 $\mathcal{G}$ is (strictly) unimodal if $\mathcal{F}$ is. By radial symmetry and unimodality the set $\mathrm{Int}(g > \tilde{t})$ is a Euclidean ball $E_{\hat{r}}(0)$ with center point zero and some radius $\hat{r}$.

Since $\mathcal{G}$ is unimodal, it follows that $\mathrm{supp}(g)$ is convex and further that the distribution function F of its radius is strictly increasing on the interior of the support. Hence F is injective on $\mathrm{Int}(\mathrm{supp}(g_r))$ and, by Lemma 3, $r = G^{-1} \circ F(\hat{r})$ and consequently there is an r for each $\hat{r} \in \mathrm{Int}(\mathrm{supp}(g_r))$ such that $E_{\hat{r}}(0) = B_r^{\mathcal{G}}(0)$. By Lemmas 3 and 4, and the homogeneity property of Mahalanobis balls it then follows

$$\mathrm{Int}(f > t) = T(\mathrm{Int}(g > \tilde{t})) = T(E_{\hat{r}}(0)) = T(B_r^{\mathcal{G}}(0)) = T(B_r^{\mathcal{G}}(T^{-1}(m))) = B_r(m),$$

where the first equality holds by the invariance of domain theorem.

If $\mathcal{F}$ is strictly unimodal then it follows that for each Euclidean ball $E_{\hat{r}}(0)$ there is some $\tilde{t}$ such that $E_{\hat{r}}(0) = \mathrm{Int}(g > \tilde{t})$, and the previous equalities then yield that for every r there is some t such that $B_r(m) = \mathrm{Int}(f > t)$. □

Proof of Theorem 8. Suppose $d(a, m) = \inf_{x \in A} d(x, m) = r$ and note $A \cap B_r(m) = \varnothing$ and $a \in \partial B_r(m)$. Let $f(a) = t$, then $a \notin (f > t)$. By Theorem 7, $\mathrm{Int}(f > t) = B_s(m)$ for some s, and since $a \notin \mathrm{Int}(f > t)$, $s \leq r$ and $A \cap B_s(m) = \varnothing$. Consequently, if $x \in A \cap \mathrm{Cl}(f > t)$ then $x \in \partial B_r(m)$, but by Lemma 5 $f(x) = f(a) = t$. Therefore $A \subset (f \leq t)$ and the first implication follows.

Suppose $\mathcal{F}$ is strictly unimodal and $f(a) = \sup_{x \in A} f(x) = t$. Let $d(a, m) = r$, then $a \in \partial B_r(m)$, and note that $(f > t) \cap A = \varnothing$. By Theorem 7, $\mathrm{Int}(f > t) = B_s(m)$, some s, and thus $\partial(f > t) = \partial B_s(m)$. But $a \in (f = t)$ and by Lemma 6, $(f = t) \subset \partial(f > t) = \partial B_s(m)$ and thus $s = r$. Hence $A \cap B_r(m) = A \cap \mathrm{Int}(f > t) = \varnothing$, which shows the second implication. □

Proof of Theorem 9. Let $w \sim \mathcal{F}$, take $L(x) = \mathrm{Var}(w)^{1/2}x$ and let $\mathcal{G} = \mathcal{L}(L^{-1}(w))$ which is spherical. Let, further, $B_r(m)$, $E_r(m)$ and $E^{\star}_r(m)$ denote the Mahalanobis ball under $\mathcal{F}$, the Euclidean ball and the ball under $\| \cdot \|_{\star}$, respectively. By Lemma 3 and the homogeneity property of Mahalanobis balls, $B_r(0) = L(B^{\mathcal{G}}_r(L^{-1}(0))) = L(E_{\hat{r}}(0))$. Note also that $E^{\star}_r(0) = L(E_r(0))$, and hence $B_r(0) = E^{\star}_{\hat{r}}(0)$.

Let $\|a\|_{\star} = s$ and $d(a, 0) = r$, and suppose $s = \inf_{x \in A} \|x\|_{\star}$. Then $a \in \partial E^{\star}_s(0)$, $a \notin B_r(0)$ and $E^{\star}_s(0) \cap A = \varnothing$. It will be shown that $B_r(0) \cap A = \varnothing$ which yields $r = \inf_{x \in A} d(x, 0)$. Since $B_r(0) = E^{\star}_{\hat{r}}(0)$, $a \notin B_r(0)$, $\hat{r} \leq s$ and hence $B_r(0) \subset E^{\star}_s(0)$ and $B_r(0) \cap A = \varnothing$.

Suppose $a \in \mathrm{Int}(\mathrm{supp}(f))$, then the equality $\hat{r} = F^{(-)} \circ G(r)$, from Lemma 3, is invertible in a neighborhood of a and thus defines a local one-to-one relationship between $\hat{r}$ and r. It follows that $a \in \partial B_r(0)$, and consequently $\hat{r} = s$ and $B_r(0) = E^{\star}_s(0)$. Hence $B_r(0) \cap A = \varnothing$ if and only if $E^{\star}_s(0) \cap A = \varnothing$, which shows the second implication. □

Proof of Theorem 10. The first implication follows from Theorems 8 and 9. With regards to the second implication, note that since $\mathcal{F}$ is elliptical all points of $\partial\mathrm{supp}(f)$ have, by symmetry, equal density. Since $\mathcal{F}$ further is strictly unimodal, that density is less than that of any point of $\mathrm{Int}(\mathrm{supp}(f))$, and thus $A \cap \mathrm{Int}(\mathrm{supp}(f)) \neq \varnothing$ and $f(a) = \sup_{x \in A}(x)$ imply that $a \in \mathrm{Int}(\mathrm{supp}(f))$. The second implication then follows by the same theorems.

To show that unimodality is necessary for the first implication, suppose for some $y \in \mathbb{X}$ and $t > 1$, $f(ty) > f(y)$. Let $A = \{y, ty\}$ and note $\|y\|_{\star} = \inf_{x \in A} \|x\|_{\star}$ while $f(y) < \sup_{x \in A} f(x) = f(ty)$. To show that the distribution necessarily must be median zero elliptical, suppose $y_1, y_2 \in \mathbb{X}$, $\|y_1\|_{\star} = \|y_2\|_{\star}$ and $y_1 \neq y_2$, but $f(y_1) \neq f(y_2)$, say $f(y_1) > f(y_2)$. Let $A = \{y_1, y_2\}$ and note $\|y_2\|_{\star} = \inf_{x \in A} \|x\|_{\star}$ while $f(y_2) < \sup_{x \in A} f(x)$. This completes the proof. □

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

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