THE HEAT EQUATION, STEIN'S LEMMA AND AN EXPECTATION IDENTITY: WITH APPLICATIONS

by

A. DasGupta
Purdue University

L. R. Haff
Univ. of California, San Diego

and

W. E. Strawderman
Rutgers University

Technical Report #96-27

Department of Statistics
Purdue University
West Lafayette, IN USA

June 1996

◇

THE HEAT EQUATION, STEIN'S LEMMA AND AN EXPECTATION IDENTITY: WITH APPLICATIONS*

by

A. DasGupta
Purdue University

L. R. Haff
Univ. of California, San Diego

and

W. E. Strawderman
Rutgers University

Abstract

This article proves an expectation identity for multivariate normal distributions and gives a series of applications. Although the identity looks different and serves a set of purposes different from Stein's, we establish that it is in fact equivalent to Stein's in one dimension. Among the applications we give are a lower and an upper bound on the mean squared error of an estimate of the mean, and formulae for computing expectations and variances of statistics. We give convincing evidence that the variance formula is useful as a competitor of the delta theorem variance approximation as a predata approximation as opposed to post data variance estimates coming from resampling methods. We indicate applications to Bayesian computing, and give a number of other applications of probabilistic nature. We also point out extensions of our results and potentials of our general approach in other cases.

1. INTRODUCTION

The purpose of this article is to present an expectation identity for multivariate normal distributions. The special $N(\mu, tI)$ distributions in p dimensions will be presented, although the theory is extendable to general covariance structures. Here $t > 0$ and I denotes the identity matrix in p dimensions. The basic identity is proved by using the multidimensional heat equation (see Powers (1979)) and Green's second formula. We give a (somewhat tricky) proof that it is in fact equivalent to Stein's(1981) famous identity, but as we shall see it looks different and it appears to serve a different set of purposes than Stein's. After the identity is given, we give a series of applications and other formulae that

* Research supported by NSF grants DMS 93–07727 and 94–00476.

arise out of the identity.

In section 2, we present the identity and show that in one dimension it is equivalent to Stein's identity. Using our identity, we then give a lower and an upper bound on the mean squared error of an estimate of a normal mean. We give an example illustrating the bounds and we show that the bounds are quite tight. The bounds decompose the risk, but not in the familiar bias square plus variance form. It is a different kind of decomposition and it needs a different set of computations than Stein's unbiased risk estimate does. Following the bounds we refine them to an exact mean squared error expression which looks like the bound plus a remainder. The remainder involves the sequence of one dimensional Laguerre polynomials, but the hope is that the remainder can be ignored and the main terms give a good approximation to the exact.

In section 3, we give a formula for the mean and the variance of a statistic arising essentially from integrating the basic identity. Each formula involves an integral, which may not be computable in closed form, but the formula will always produce a numerical value. Note that Stein's identity does not give the variance of a function; it will give the mean squared error. We give a number of examples illustrating the use and the apparent sharp approximations it produces even when the underlying distribution is not exactly a normal. The contemplated application in classical statistics is to give a predata variance approximation for a statistic rather than a post data estimate which bootstrap and Jackknife estimates give. In this sense, the formula serves a purpose similar to what the delta theorem variance approximation does. See Bickel and Doksum (1977). Our examples indicate that in moderate samples this approximation is more satisfactory than the delta theorem method and since it is predata, it does not have the disadvantages of resampling estimates such as possible severe fluctuations. A Bayesian example is also provided that indicates remarkable sharpness in approximating a posterior mean.

Section 4 lists a variety of other consequences, some of which are of probabilistic nature. For instance, the following question is answered: if X has the $N(\underset{\sim}{\mu}, tI)$ distribution for known t, and $\underset{\sim}{\mu}$ is given the improper Jeffrey prior, for which parametric functions the mle and the Bayes estimate are the same? The basic identity shows that this is true for only harmonic functions of $\underset{\sim}{\mu}$.

Section 5 provides a discussion of extensions. The extensions indicated are for general covariance structures and for location-scale mixtures of normals. We also indicate that our general approach of identifying a family of densities that satisfy a partial differential equation has greater potential; for example, we show that the use of the wave equation should lead to a parallel set of identities and formulae. We also show that in principle, the formula for the mean of a statistic that we present can be used to find approximations of integrals by combining it with a Tauberian theorem of Wiener on denseness of the closure of translations, a consequence of lack of any zeroes of the characteristic function of the normal distribution.

The principal contributions of the article are the following:

a. We give an identity, and although it looks different from Stein's, we prove the equivalence. This we believe gives more insight into Stein's lemma as well. It could be argued that the real reason Stein's lemma holds for the normal case is that its density satisfies the heat equation. Diaconis and Zabell (1991) indeed show that Stein's lemma is a characterizing property of the normal density.

b. We give many applications; some have a decision theory flavor, others are more classical. In particular, we give mean squared error bounds and a competitor to the delta theorem variance approximation. The mean squared error bounds give a different decomposition than the bias - variance decomposition. We also show the usefulness for Bayesian calculations.

c. We give a number of probabilistic corollaries.

d. We show that the general approach of finding solutions to partial differential equations that are probability densities could be a useful approach.

2. THE IDENTITY AND BASIC APPLICATIONS

2.1. An Expectation Identity

Lemma 1. Let $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, tI)$ and let $g(\underset{\sim}{x}, \underset{\sim}{\mu})$ be a scalar valued function such that it has two partial derivatives with respect to each coordinate of $\underset{\sim}{X}$. Suppose in addition that g

satisfies the growth conditions:

a $$re^{-\frac{1}{2t}r^2}\int_{\partial B(-\underset{\sim}{\mu},r)}|g(\underset{\sim}{x}+\underset{\sim}{\mu},\underset{\sim}{\mu})|d\sigma \longrightarrow 0 \quad \text{as} \quad r\to\infty,$$

b $$e^{-\frac{1}{2t}r^2}\int_{\partial B(-\underset{\sim}{\mu},r)}||\nabla g(\underset{\sim}{x}+\underset{\sim}{\mu},\underset{\sim}{\mu})||d\sigma \longrightarrow 0 \quad \text{as} \quad r\to\infty,$$

where $\partial B(\underset{\sim}{a},r)$ denotes the boundary of the ball $B(\underset{\sim}{a},r)=\{\underset{\sim}{z}\colon\ ||\underset{\sim}{z}-\underset{\sim}{a}||\le r\}$ and $\int_{\partial B}hd\sigma$ denotes the surface integral of h over ∂B.

Then, $\frac{\partial}{\partial t}Eg(\underset{\sim}{X},\underset{\sim}{\mu})=\frac{1}{2}E(\Delta_x g(\underset{\sim}{X},\underset{\sim}{\mu}))$, where $\Delta_x g$ is the Laplacian of g with respect to $\underset{\sim}{x}$.

Proof: We will write $f(\underset{\sim}{x}|\underset{\sim}{\mu},t)$ for the $N(\underset{\sim}{\mu},tI)$ density. Thus,

$$\begin{aligned}
&\frac{\partial}{\partial t}Eg(\underset{\sim}{X},\underset{\sim}{\mu})\\
&=\frac{\partial}{\partial t}\int g(\underset{\sim}{x},\underset{\sim}{\mu})f(\underset{\sim}{x}|\underset{\sim}{\mu},t)d\underset{\sim}{x}\\
&=\int g(\underset{\sim}{x},\underset{\sim}{\mu})\frac{\partial}{\partial t}f(\underset{\sim}{x}|\underset{\sim}{\mu},t)d\underset{\sim}{x}\\
&=\frac{1}{2}\int g(\underset{\sim}{x},\underset{\sim}{\mu})\Delta_x f(\underset{\sim}{x}|\underset{\sim}{\mu},t) \qquad \text{(Heat equation)}\\
&=\frac{1}{2}\int f(\underset{\sim}{x}|\underset{\sim}{\mu},t)\Delta_x g(\underset{\sim}{x},\underset{\sim}{\mu})+\frac{1}{2}\int\{g(\underset{\sim}{x},\underset{\sim}{\mu})\Delta_x f(\underset{\sim}{x},\underset{\sim}{\mu},t)-f(\underset{\sim}{x}|\underset{\sim}{\mu},t)\Delta_x g(\underset{\sim}{x},\underset{\sim}{\mu})\}d\underset{\sim}{x}\\
&=\frac{1}{2}E\Delta_x g(\underset{\sim}{X},\underset{\sim}{\mu})+\frac{1}{2}\lim_{r\to\infty}\int_{B(\underset{\sim}{0},r)}\{g(\underset{\sim}{x},\underset{\sim}{\mu})\Delta_x f(\underset{\sim}{x}|\underset{\sim}{\mu},t)-f(\underset{\sim}{x}|\underset{\sim}{\mu},t)\Delta_x g(\underset{\sim}{x},\underset{\sim}{\mu})\}d\underset{\sim}{x}\\
&=\frac{1}{2}E\Delta_x g(\underset{\sim}{X},\underset{\sim}{\mu})+\frac{1}{2}\lim_{r\to\infty}\int_{\partial B(0,r)}\int\{g(\underset{\sim}{x},\underset{\sim}{\mu})\frac{\partial f}{\partial n}(\underset{\sim}{x}|\underset{\sim}{\mu},t)-f(\underset{\sim}{x}|\underset{\sim}{\mu},t)\frac{\partial g}{\partial n}(\underset{\sim}{x},\underset{\sim}{\mu})\}d\sigma\\
&\qquad\text{(Green's second formula)} \qquad (2.1)
\end{aligned}$$

Now,

$$
\begin{aligned}
&\Big| \int_{\partial B(\underset{\sim}{0},r)} g(\underset{\sim}{x},\underset{\sim}{\mu}) \frac{\partial f}{\partial n}(\underset{\sim}{x}|\underset{\sim}{\mu},t) - f(\underset{\sim}{x}|\underset{\sim}{\mu},t) \frac{\partial g}{\partial n}(\underset{\sim}{x},\underset{\sim}{\mu}) d\sigma \Big| \\
&\leq \int_{\partial B(\underset{\sim}{0},r)} |g(\underset{\sim}{x},\underset{\sim}{\mu})| \, |\frac{\partial f}{\partial n}(\underset{\sim}{x}|\underset{\sim}{\mu},t)| d\sigma + \int_{\partial B(\underset{\sim}{0},r)} f(\underset{\sim}{x}|\underset{\sim}{\mu},t) |\frac{\partial g}{\partial n}(\underset{\sim}{x},\underset{\sim}{\mu})| d\sigma \\
&\leq \int_{\partial B(\underset{\sim}{0},r)} |g(\underset{\sim}{x},\underset{\sim}{\mu})| ||\nabla_x f(\underset{\sim}{x}|\underset{\sim}{\mu},t)|| d\sigma + \int_{\partial B(\underset{\sim}{0},r)} f(\underset{\sim}{x}|\underset{\sim}{\mu},t) ||\nabla_x g(\underset{\sim}{x},\underset{\sim}{\mu})|| d\sigma \\
&\qquad \text{(Schwartz's inequality and the fact that the outer normal } n \text{ has norm 1)} \\
&= \frac{r e^{-\frac{r^2}{2t}}}{t \cdot (2\pi t)^{p/2}} \int_{\partial B(-\underset{\sim}{\mu},r)} |g(\underset{\sim}{x}+\underset{\sim}{\mu},\underset{\sim}{\mu})| d\sigma + \frac{e^{-\frac{r^2}{2t}}}{(2\pi t)^{p/2}} \int_{\partial B(-\underset{\sim}{\mu},r)} ||\nabla_x g(\underset{\sim}{x}+\underset{\sim}{\mu},\underset{\sim}{\mu})|| d\sigma \\
&\qquad \text{(change of variable)}
\end{aligned}
$$

$\longrightarrow 0$ as $r \longrightarrow \infty$ by the growth assumptions a and b. Hence, from (2.1), the lemma follows.

First we show equivalence to Stein's identity (1981), stated below for reference. The proof that the two identities are equivalent is quite tricky. For derivatives of expectations of indicator functions, see Haff and Alcaraz (1991).

2.2. Equivalence to Stein's Lemma

Lemma 2 *(Stein).* Let $X \sim N(\mu, t)$ and let h: $\mathbb{R} \to \mathbb{R}$ be such that $h'(x)$ exists for every x and $E|h'(X)| < \infty$. Then $E(X-\mu)h(X) = tEh'(X)$.

Lemma 3. Let $p = 1$; then Lemma 1 and Lemma 2 are equivalent.

Proof: *Lemma 1 $\Rightarrow$ Lemma 2:* Given a function h as in Lemma 2, define $g(x) = \int_0^x h(u)du$; thus, for $x < 0$, $g(x) = -\int_x^0 h(u)du$.

Now,

$$
\begin{aligned}
Eg(X) &= \int_0^\infty \int_0^x \frac{1}{\sqrt{2\pi t}} e^{\frac{-(x-\mu)^2}{2t}} h(u)dudx - \int_{-\infty}^0 \int_x^0 \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-\mu)^2}{2t}} h(u)dudx \\
&= \int_0^\infty \int_u^\infty \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-\mu)^2}{2t}} h(u)dxdu - \int_{-\infty}^0 \int_{-\infty}^u \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-\mu)^2}{2t}} h(u)dxdu \\
&= \int_0^\infty \{1 - \Phi(\frac{u-\mu}{\sqrt{t}})\} h(u)du - \int_{-\infty}^0 \Phi(\frac{u-\mu}{\sqrt{t}}) h(u)du. \qquad (2.2)
\end{aligned}
$$

Therefore, by Lemma 1,

$$\begin{aligned}\frac{1}{2}Eh'(X) = \frac{1}{2}Eg''(X) &= \frac{d}{dt}Eg(X) \\ &= \frac{d}{dt}\left[\int_0^\infty \{1-\Phi(\frac{u-\mu}{\sqrt{t}})\}h(u)du - \int_{-\infty}^0 \Phi(\frac{u-\mu}{\sqrt{t}})h(u)du\right] \\ &= \int_{-\infty}^\infty \frac{u-\mu}{2t^{3/2}}\phi(\frac{u-\mu}{\sqrt{t}})h(u)du \\ &= \frac{1}{2t}E(X-\mu)h(X), \quad \text{giving Lemma 2.}\end{aligned}$$

Converse: Lemma 2 $\Rightarrow$ Lemma 1: Given a function g as in Lemma 1, define $h(x) = g'(x)$. Thus by Lemma 2,

$$\begin{aligned}tEg'' = tEh'(X) = E(X-\mu)h(X) &= E(X-\mu)g'(X) \\ &= \int_{-\infty}^\infty (x-\mu)\frac{e^{-\frac{(x-\mu)^2}{2t}}}{\sqrt{2\pi t}}g'(x)dx \\ &= \sqrt{t}\int_{-\infty}^\infty \frac{ze^{-\frac{z^2}{2}}}{\sqrt{2\pi}}g'(\mu+z\sqrt{t})dz \\ &\quad \text{(change of variable)} \\ &= 2t\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\frac{d}{dt}(g(\mu+z\sqrt{t}))dz \\ &= 2t\frac{d}{dt}\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}g(\mu+z\sqrt{t})dz \\ &= 2t\frac{d}{dt}\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi t}}e^{-\frac{(x-\mu)^2}{2t}}g(x)dx \\ &\quad \text{(change of variable again)} \\ &= 2t\frac{d}{dt}Eg(X), \text{ giving Lemma 1}\end{aligned}$$

2.3. Mean Squared Error Bounds

Before we give further elaborate consequence of our Lemma 1, we will first use Lemma 1 to give a simple but useful lower and upper bound on the mean squared error of an estimate of μ. We will only present the case $p = 1$ although the case of general p is actually exactly the same. A short discussion of the result will follow its statement. It will be understood that the estimate of μ is such that Lemma 1 applies and all quantities make sense.

Proposition 1. The mean squared error of an estimate $\delta(X)$ of μ satisfies the bounds:

$$(\delta(\mu)-\mu)^2 + t \cdot \inf_x \mathcal{D}_\delta(x) + \frac{t^2}{2} \cdot \inf_x \delta^{(3)}(x) \leq E(\delta(x)-\mu)^2 \leq (\delta(\mu)-\mu)^2 + t \cdot \sup_x \mathcal{D}_\delta(x) + \frac{t^2}{2} \sup_x \delta^{(3)}(x)$$

where

$$\mathcal{D}_\delta(x) = (\delta'(x))^2 + (\delta(x)-x)\delta''(x) \tag{2.3}$$

Discussion of Proposition 1: Note that for all linear estimates $ax+b$, the lower and the upper bound coincide and the proposition therefore gives the exact mean squared error. Since estimates arising out of most common methods are approximately linear in the normal case (including Bayes estimates whenever the sample size is moderate), one would expect Proposition 1 to give useful envelopes bounding the mean squared error. We shall see one example. Note another aspect of the proposition: the decomposition of the mean squared error is not in the familiar (bias)^2+ variance form; it is a different sort of decomposition. Note also that in both bounds, the last two terms are free of μ and are thus correction terms to $(\delta(\mu)-\mu)^2$. These bounds serve a purpose different from Stein's unbiased estimate of risk. They require computation of two maximas and two minimas and the estimate δ at the desired μ. Stein's method needs numerical integration of the unbiased estimate while the bounds need numerical differentiation. So they have different flavors. Bounds on risks of Bayes procedures are also considered in other places; see LeCam (1982).

We will now prove Proposition 1. The proof uses both Lemma 1 (our identity) and Lemma 2 (Stein's identity) (the two identities are equivalent, of course: we mean both identities are used in the proof in their familiar forms).

Proof of Proposition 1: Use $g(x,\mu) = (\delta(x)-\mu)^2$ in Lemma 1. Then $g''(x,\mu) =$

$2\{(\delta'(x))^2 + (\delta(x) - \mu)\delta''(x)\}$. By Lemma 1 and the fundamental theorem of calculus,

$$\begin{aligned}
E_{\mu,t}(\delta(X) - \mu)^2 &= E_{\mu,t}g(X,\mu) \\
&= g(\mu,\mu) + \int_0^t \{\frac{d}{ds}E_{\mu,s}g(X,\mu)\}ds \\
&= (\delta(\mu) - \mu)^2 + \frac{1}{2}\int_0^t \{E_{\mu,s}g''(x,\mu)\}ds \\
&= (\delta(\mu - \mu)^2 + \int_0^t E_{\mu,s}\{(\delta'(x))^2 + (\delta(x) - \mu)\delta''(x)\}ds \\
&= (\delta(\mu) - \mu)^2 + \int_0^t E_{\mu,s}\{(\delta'(x))^2 + (\delta(x) - x)\delta''(x) + (x - \mu)\delta''(x)\}ds \\
&= (\delta(\mu) - \mu)^2 + \int_0^t E_{\mu,s}\{(\delta'(x))^2 + (\delta(x) - x)\delta''(x) + s\delta^{(3)}(x)\}ds \\
&\qquad \text{(by using Stein's identity on the third term)} \\
&\leq (\delta(\mu) - \mu)^2 + t \cdot \sup_x\{(\delta'(x))^2 + (\delta(x) - x)\delta''(x)\} + \frac{1}{2}t^2 \cdot \sup_x \delta^{(3)}(x),
\end{aligned}$$

as stated. The lower bound is similarly obtained.

We will now see one illustration of the mean squared error bounds of Proposition 1.

Example 1. Suppose $X_1, \ldots, X_n$ are random samples from the $N(\mu, 1)$ distribution and μ is given a Double Exponential prior with density $\frac{1}{2}e^{-|\mu|}$. In this case, evaluation of the bounds of Proposition 1 is relatively easy as the Bayes estimate $\delta(\overline{X})$, as is well known, can be evaluated in closed form. For specificity, take $n = 16$ (neither too large nor too small). Computation on Mathematica yields:

$$\begin{aligned}
\sup_x \mathcal{D}_\delta(x) &= 1, & \inf_x \mathcal{D}_\delta(x) &= .413736, \\
\sup_x \delta^{(3)}(x) &= 4.2967, & \inf_x \delta^{(3)}(x) &= -2.1782.
\end{aligned}$$

Substituting $t = \frac{1}{n} = \frac{1}{16}$ gives the bounds:

$$(\delta(\mu) - \mu)^2 + .0216 \leq E(\delta(\overline{X}) - \mu)^2 \leq (\delta(\mu) - \mu)^2 + .0709.$$

We think these are quite tight bounds and the whole computation took approximately 7 minutes on Mathematica from beginning to end.

2.4. Exact Mean Squared Error

In continuation and as a refinement of Proposition 1, we now present an expansion for the mean squared error of an estimate $\delta(X)$ of the mean. The expansion is in the form of three terms resembling the bounds in Proposition 1 plus a remainder. It serves the purpose of comparison of the bounds to a similar looking formula for the exact. Again, we present only the case $p = 1$ although the case of the general dimension is exactly the same. It will as usual be understood that $\delta(x)$ is such that Lemma 1 applies with $g(x, \mu) = (\delta(x) - \mu)^2$.

Proposition 3. Let $\alpha = -\frac{1}{2}$ and let L_n^α, $n \geq 0$, denote the sequence of Laguerre polynomials $L_n^\alpha(z) = \frac{e^z z^{-\alpha}}{n!} \frac{d^n}{dz^n}(e^{-z} z^{n+\alpha})$. See Szego (1975). Then, the mean squared error of an estimate $\delta(X)$ satisfies

$$E(\delta(X) - \mu)^2 = (\delta(\mu) - \mu)^2 + t \cdot E\mathcal{D}_\delta(X) + t^2 \cdot E\delta^{(3)}(X) + R(\mu, t) \tag{2.4}$$

where as in Proposition 1, $\mathcal{D}_\delta(x) = (\delta'(x))^2 + (\delta(x) - x)\delta''(x)$, and $R(\mu, t)$ is a remainder term involving the Laguerre polynomials L_n^α.

Proof: The argument is essentially exactly the same as in Proposition 1. The important thing is to understand how the Laguerre polynomials arise.

The proof uses the formula for $e_1(\mu, t)$ in Proposition 2 in the next section with $g(x, \mu) = (\delta(x) - \mu)^2$ and then uses the identity $\Gamma(\alpha, z) = e^{-z} z^\alpha \cdot \sum_{n \geq 0} \frac{L_n^\alpha(z)}{n+1} = e^{-z} z^\alpha (1 + \sum_{n \geq 1} \frac{L_n^\alpha(z)}{n+1})$ (pp. 942 in Gradshteyn and Ryzhik (1980)). The term $e^{-z} z^\alpha \cdot \sum_{n \geq 1} \frac{L_n^\alpha(z)}{n+1}$ contributes to the remainder and the terms $t \cdot E\mathcal{D}_\delta(x) + t^2 \cdot E\delta^{(3)}(X)$ come from the rest, if one copies the proof of Proposition 1.

3. MEAN-VARIANCE FORMULAE AND APPLICATIONS

3.1. Description

In this section we will see how Lemma 1 leads to a general formula for the mean of a statistic. Applying the same formula to the square of the statistic and combining, one gets a variance formula also. In cases where the integral in the variance formula cannot be implemented in closed form, numerical integration will give a numerical value for the variance. Whenever the variance can be computed in closed form by other means, that

should be the recommended method, of course. This is true of our Example 3 that follows. After presenting the formulae, we will give examples illustrating them. Common methods for variance approximation in classical statistics include the delta theorem estimate or resampling variance estimates which typically are strongly consistent for the delta theorem estimate; see Shao and Tu (1995). Our method of variance approximation is pre-data like the delta theorem estimate. But our experience indicates it is a more satisfactory approximation in moderate samples than the delta theorem approximation. The same rationale will also be used in a Bayesian example.

3.2. The Formulae

Proposition 2. For any function $g(\underset{\sim}{x}, \underset{\sim}{\mu})$ satisfying the conditions of Lemma 1,

$$Eg(\underset{\sim}{X}, \underset{\sim}{\mu}) = g(\underset{\sim}{\mu}, \underset{\sim}{\mu}) + e_1(\underset{\sim}{\mu}, t)$$

and

$$\text{var } g(\underset{\sim}{X}, \underset{\sim}{\mu}) = e_2(\underset{\sim}{\mu}, t) - 2g(\underset{\sim}{\mu}, \underset{\sim}{\mu})e_1(\underset{\sim}{\mu}, t) - e_1^2(\underset{\sim}{\mu}, t),$$

where

$$\begin{aligned} e_1(\underset{\sim}{\mu}, t) &= \frac{1}{4\pi^{p/2}} \int \Delta_x g(\underset{\sim}{x}, \underset{\sim}{\mu}) \cdot ||\underset{\sim}{x} - \underset{\sim}{\mu}||^{2-p} \\ &\quad \cdot \Gamma(\frac{p}{2} - 1, \frac{||\underset{\sim}{x} - \underset{\sim}{\mu}||^2}{2t}) d\underset{\sim}{x} \qquad (3.1) \\ e_2(\underset{\sim}{\mu}, t) &= \frac{1}{2\pi^{p/2}} \int \left\{ ||\nabla_x g(\underset{\sim}{x} - \underset{\sim}{\mu})||^2 + g(\underset{\sim}{x}, \underset{\sim}{\mu}) \Delta_x g(\underset{\sim}{x}, \underset{\sim}{\mu}) \right\} ||\underset{\sim}{x} - \underset{\sim}{\mu}||^{2-p} \\ &\quad \cdot \Gamma(\frac{p}{2} - 1, \frac{||\underset{\sim}{x} - \underset{\sim}{\mu}||^2}{2t}) d\underset{\sim}{x} \qquad (3.2) \end{aligned}$$

and $\Gamma(\alpha, y)$ denotes the incomplete gamma function $\Gamma(\alpha, y) = \int_y^\infty e^{-u} u^{\alpha-1} du$, and $\nabla_x g$ denotes the gradient vector of g with respect to $\underset{\sim}{x}$.

Proof: Since $\frac{d}{dt} Eg(\underset{\sim}{X}, \underset{\sim}{\mu}) = \frac{1}{2} E \nabla_x g(\underset{\sim}{X}, \underset{\sim}{\mu})$ by Lemma 1, by the Fundamental theorem of calculus,

$$\begin{aligned} Eg(\underset{\sim}{X}, \underset{\sim}{\mu}) - g(\underset{\sim}{\mu}, \underset{\sim}{\mu}) &= \frac{1}{2} \int_0^t \int_{\mathbb{R}^p} \nabla_x g(\underset{\sim}{x}, \underset{\sim}{\mu}) \frac{1}{(2\pi s)^{p/2}} e^{-\frac{1}{2s}(\underset{\sim}{x} - \underset{\sim}{\mu})'(\underset{\sim}{x} - \underset{\sim}{\mu})} d\underset{\sim}{x} ds \\ &= \frac{1}{2\pi^{p/2}} \int_{\mathbb{R}^p} \nabla_x g(\underset{\sim}{x}, \underset{\sim}{\mu}) \int_0^t \frac{e^{-\frac{1}{2s}(\underset{\sim}{x} - \underset{\sim}{\mu})'(\underset{\sim}{x} - \underset{\sim}{\mu})}}{(2s)^{p/2}} ds d\underset{\sim}{x} \qquad (3.3), \end{aligned}$$

where the application of Fubini's theorem is justified due to the hypothesis of the proposition. A change of variable to $u = \frac{(\underset{\sim}{x}-\underset{\sim}{\mu})'(\underset{\sim}{x}-\underset{\sim}{\mu})}{2s}$ in the inner integral gives the expression for $Eg(\underset{\sim}{X}, \underset{\sim}{\mu})$. Application of the same expression to the function $g^2(\underset{\sim}{X}, \underset{\sim}{\mu})$ gives the second moment and hence the expression for var $g(\underset{\sim}{X}, \underset{\sim}{\mu})$.

3.3. Examples

We will now give some illustrations of Proposition 2.

Example 2. Suppose $X_1, \ldots, X_n$ are iid with the common density $\frac{1}{2}e^{-|x|}$, $-\infty < x < \infty$. Suppose we are interested in the variance of the statistic $\sin \overline{X}$. The exact sampling distribution of $\overline{X}$ is extremely complicated and so an approximation of some sort to the variance is necessary. The choice of the double Exponential sampling density is merely an artifact though. A standard variance estimate is the one given by the delta theorem, which in this case will be $(\cos 0)^2 \cdot \frac{2}{n} = \frac{2}{n} = t$ (say).

Now, var $\sin \overline{x} = E(\sin \overline{X})^2$ (as $E \sin \overline{X} = 0$). Now if we apply our Lemma 1 pretending as if $\overline{X} \sim N(0, t)$, then we will get

$$
\begin{aligned}
&\frac{d}{dt} \text{ Var } \sin \overline{X} \\
&= \frac{d}{dt} E(\sin \overline{X})^2 \\
&= \frac{d}{dt} E(g(X) | X \sim N(0,t)) \text{ (with } g(X) = \sin^2 x) \\
&= \frac{1}{2} E(\frac{d^2}{dx^2} \sin^2 x | x \sim N(0,t)) \\
&= E(\cos 2x | x \sim N(0,t)) \\
&= \frac{2}{\sqrt{2\pi t}} \int_0^\infty (\cos 2x) \cdot e^{-\frac{x^2}{2t}} dx \\
&= \frac{1}{\sqrt{2\pi t}} \int_0^\infty (\cos z) \cdot e^{-\frac{z^2}{8t}} dz \\
&= \frac{1}{\sqrt{2\pi t}} \cdot \frac{\sqrt{\pi} \cdot \sqrt{8t}}{2} {}_1F_1(\frac{1}{2}; \frac{1}{2}; -2t) \\
&\qquad \text{(see pp. 495 in Gradshteyn and Ryzhik (1980))} \\
&= {}_1F_1(\frac{1}{2}; \frac{1}{2}; -2t) = e^{-2t} \text{ (pp. 1059 in Gradshteyn and Ryzhik (1980)).}
\end{aligned}
$$

Therefore, on integration, Var $\sin \overline{X} \approx \frac{1}{2}(1 - e^{-2t})$ [NOTE: We effectively showed $e_1(\mu, t) = \frac{1}{2}(1 - e^{-2t})$]. This is a fundamentally better approximation in the sense the delta theorem

approximation is linear (and unbounded in t), and the second approximation is always in the correct range (i.e., we know $E(\sin \overline{X})^2 \leq 1$). Also note that very interestingly, if the approximation $\frac{1}{2}(1 - e^{-2t})$ is expanded and just the first term is kept, we get the delta theorem approximation, namely, t.

We did a simulation for the case $n = 16$, a moderate value. Simulation of size 5000 gave the variance of $\sin \overline{X}$ to be .110747. The delta theorem approximation is $t = \frac{2}{16} = .125$ and Lemma 1 give the approximation $\frac{1}{2}(1-e^{-2t}) = .1106$, a considerably more satisfactory approximation.

Example 3. Suppose $X_1, \ldots, X_n$ are random samples from the $N(\theta, 1)$ distribution and it is desired to estimate $P(X \leq 0) = 1 - \Phi(\theta) = g(\theta)$. θ is assigned a central t prior with 3 degrees of freedom. A common Bayesian estimate will be the posterior mean of $g(\theta)$. For specificity, assume $n = 16$ and $\overline{X} = 1$.

The posterior distribution of θ is not normal; however, as in Example 2, let us explore how our method will work if we pretend as if it was normal and use Proposition 2 to calculate the expectation of $g(\theta)$ under this normal distribution. In the notation of this article, for μ and t we use the actual posterior mean and variance of θ. These are $\mu = .941347$ and $t = .0603341$. Thus, since $g''(\theta) = \theta\phi(\theta)$, Proposition 2 says

$$Eg(\theta) = 1 - \Phi(\mu) + \frac{1}{4\sqrt{\pi}} \int \theta\phi(\theta)|\theta - \mu|\Gamma(-\frac{1}{2}, \frac{(\theta-\mu)^2}{2t})d\theta, \tag{3.4}$$

which works out to .180313. From the exact posterior distribution, the mean of $g(\theta)$ is equal to .180315. This extremely close match again gives evidence that even for moderate sample sizes, Lemma 1 and Proposition 2 have excellent potential for giving close approximations to means and variances of statistics.

Example 4. We now give another example to give further evidence that use of the formula in Proposition 2 gives a sharp approximation to the true variance of complicated statistics. We take the sample correlation coefficient r in a sample of size n from a Bivariate normal distribution. Its exact density is given by

$$t(r) = \frac{2^{n-3}}{\pi(n-3)!}(1-\rho^2)^{\frac{n-1}{2}}(1-r^2)^{\frac{n-4}{2}} \cdot \sum_{j \geq 0}(\Gamma\frac{n-1+j}{2})^2 \frac{(2\rho)^j}{j!}r^j, \quad -1 \leq r \leq 1; \tag{3.5}$$

see Tong (1990). Diaconis and Efron (1983) present many calculations on the density of r.

A very common test statistic for inference about ρ is $h(r) = \tanh^{-1} r = \frac{1}{2}\log\frac{1+r}{1-r}$. The exact variance of $h^2(r) = g(r)$ (say) is seen to be .00431361 for $n = 25$ when $\rho = 0$: this is found from (3.5). On the other hand, asymptotic theory says $\sqrt{n}r \xrightarrow{\mathcal{L}} N(0,1)$ when $\rho = 0$. Therefore, we could use our variance formula in Proposition 2, using $\mu = 0$ and $t = \frac{1}{n} = .04$. Numerical integration in Proposition 2 variance formula gives .0048691. Again, the remarkable sharpness is quite encouraging.

4. FURTHER APPLICATIONS

The purpose of this section is to give a number of other consequences of Lemma 1. Some of them are of a probabilistic nature. All of these will be stated in a single proposition. A brief discussion of the relevance and interest of the parts of the proposition will then follow; for instance, we will say why part a should be of interest in Bayesian statistics.

Proposition 4. Let $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, tI)$ and suppose the conditions of Proposition 2 hold. Then the following are true:

a If $g(\underset{\sim}{x}, \underset{\sim}{\mu})$ is harmonic (in x), then $Eg(\underset{\sim}{X}, \underset{\sim}{\mu}) = g(\underset{\sim}{\mu}, \underset{\sim}{\mu})$ for all t. Conversely, if $Eg(\underset{\sim}{X}, \underset{\sim}{\mu}) = g(\underset{\sim}{\mu}, \underset{\sim}{\mu})$ for all t, then g is harmonic.

b If g is subharmonic, then always $g(\underset{\sim}{\mu}, \underset{\sim}{\mu}) \leq Eg(\underset{\sim}{X}, \underset{\sim}{\mu})$ and furthermore the difference $Eg(\underset{\sim}{X}, \underset{\sim}{\mu}) - g(\underset{\sim}{\mu}, \underset{\sim}{\mu})$ exactly equals the nonnegative quantity $e_1(\underset{\sim}{\mu}, t)$.

c For any g, $|Eg(\underset{\sim}{X}, \underset{\sim}{\mu}) - g(\underset{\sim}{\mu}, \underset{\sim}{\mu})| \leq \frac{t}{2} \cdot$ ess sup $(\Delta_x g)$

d For any $k \geq 1$, $\frac{d^k}{dt^k} Eg(\underset{\sim}{X}, \underset{\sim}{\mu}) = \frac{1}{2^k} E(\Delta^k g(\underset{\sim}{X}, \underset{\sim}{\mu}))$, whenever everything makes sense.

e If g is subharmonic, then Eg is nondecreasing in t.

f If a given parametric function $h(\underset{\sim}{\mu}, t)$ has an unbiased estimate $g(\underset{\sim}{X})$, then $\frac{\partial}{\partial t} h(\underset{\sim}{\mu}, t)$ also has an unbiased estimate; furthermore, this unbiased estimate is $\frac{1}{2}\Delta g(\underset{\sim}{X})$.

Discussion of Proposition 4: Part a says moment estimates are unbiased only for harmonic functions (for Bayesians, a more interesting statement is that the Jeffrey prior results in the mle only for harmonic functions). One can see that this is connected to the fact that if $B(t)$ is a standard Brownian motion, then $g(B(t)) - t$ is a martingale if and

only if g is harmonic. Part b can be regarded as a generalization of Jensen's inequality for convex functions in one dimension. More than the inequality itself, it is interesting that the difference can be exactly written down. Part c can be regarded as a bound on the bias of moment estimates and the bound says if either g is nearly harmonic or t is small, then the bias is small. Part d is a consequence of repeated use of the basic Lemma 1. Part e is an unusual result; it says, in particular, that in one dimension convex functions have expectations increasing with the variance t. Part f could be useful if $h(\underset{\sim}{\mu}, t)$ is some kind of an accuracy measure (like mean squared error) and one likes to know how the accuracy decreases with an increase in t. Then, knowing an estimate of the accuracy, part f says how to estimate its derivative. The proofs of parts a–f are straightforward consequences of the basic Lemma 1 and Proposition 2. We will omit the details.

5. EXTENSIONS

Our general approach has potential for some extensions. We will indicate a few briefly.

5.1. Mixtures

Consider the case when g is a pure statistic, i.e., a function of only X. Proposition 2 gives a formula for the mean $Eg(\underset{\sim}{X})$ when $\underset{\sim}{X}$ has $N(\underset{\sim}{\mu}, tI)$ distribution as $Eg(\underset{\sim}{X}) = g(\underset{\sim}{\mu}) + e_1(\underset{\sim}{\mu}, t)$. Therefore, if $\underset{\sim}{X}$ has a location-scale normal mixture as its distribution, then in principle $Eg(\underset{\sim}{X})$ is obtainable also, by simply integrating $g(\underset{\sim}{\mu}) + e_1(\underset{\sim}{\mu}, t)$ with respect to the mixing distribution on $(\underset{\sim}{\mu}, t)$. Thus our stated results could be stated in a more general form.

5.2. Approximation of Integrals

Since the characteristic function of the $N(0,1)$ distribution on the line has no zeroes, the following Tauberian theorem holds, which then has some implication for our results.

Lemma 4 *(Wiener).* Let $\phi \in L_1(\mathbb{R})$ be such that its characteristic function has no zeroes. Then given any $f \in L_1(\mathbb{R})$, and $\varepsilon > 0$, there exist m and constants a_i, μ_i such that

$$\int |f(x) - \sum_{i=1}^{m} a_i \phi(x - \mu_i)| dx < \varepsilon$$

and

$$\sup_x |f(x) - \sum_{i=1}^{m} a_i \phi(x - \mu_i)| < \varepsilon.$$

See Bochner (1936) for a proof. Note that the translates are already dense; there is no need to scale. Let us suppose we have a given L_1 function f and the constants $\{a_i, \mu_i\}_1^m$ have been found. Then, Lemma 4 and Proposition 2 give the following:

Proposition 5. Suppose $g(x)$ is any other function such that it is either bounded or in L_1. Then

$$|\int_{-\infty}^{\infty} g(x)f(x)dx - \sum_{i=1}^{m} a_i\{g(\mu_i) + e_1(\mu_i, 1)\}| \leq c\varepsilon, \tag{4.1}$$

where

$$C = \min(\sup_x |g(x)|, \int_{-\infty}^{\infty} |g(x)|dx).$$

Recall that $e_1(\mu, t)$ is as in Proposition 2 and one can take t as 1 due to the nature of Lemma 4. (4.1) is reminiscent of Gaussian quadrature in numerical analysis; see Powell (1981).

5.3. More General Covariance Structures

A parallel set of theorems can be written down assuming the probabilistic structure $\underset{\sim}{X} \sim N(\underset{\sim}{\mu}, t\Sigma)$ for a general positive definite Σ, as the density of the $N(\underset{\sim}{\mu}, t\Sigma)$ admits the general heat equation.

5.4. Other Differential Equations

The results of this paper point to the fact that whenever a family of probability densities satisfies a partial differential equation, the techniques we have used may lead to expectation identities. To see the potential of the approach of this paper, consider densities $f(x,t)$ of the general form $f(x,t) = u(x-t) + v(x+t)$. Then, under mild derivative conditions, $f(x,t)$ will satisfy the wave equation $\frac{\partial^2 f}{\partial t^2} = \frac{\partial^2 f}{\partial x^2}$. In the lines of our Lemma 1, one will then get an expectation identity $\frac{d^2}{dt^2} E_f g(X) = E_f g''(X)$ and extension to higher dimensions is similar. Such an identity may potentially have another parallel set of applications.

References

Bochner, S. (1936). Lectures on Fourier Analysis, Princeton University Press.

Bickel, P. J. and Doksum, K. (1977). Mathematical Statistics: Basic Ideas and Selected Topics, Holden-Day. Oakland, California.

Diaconis, P. and Efron, B. (1983). Computer intensive methods in Statistics, Scientific American, 116–130.

Diaconis, P. and Zabell, S. (1991). Closed form summation formulae for classical distributions, *Stat. Sc.*, **6, 3**, 284–302.

Gradshteyn, I. S. and Ryzhik, I. M.(1980). Table Of Integrals, Series and Products, Academic Press, New York.

Haff, L. R. and Alcaraz, J. E. (1991). On the differentiation of certain probabilities with applications to statistical decision theory, Proc. 5th Purdue Symposium, S. S. Gupta and J. O.Berger, Springer-Verlag, New York.

Le Cam, L. (1982). Risk of Bayes estimates, Proc. 3rd Purdue Symp, 2, 121–137, S. S.Gupta and J. O. Berger, Academic Press, New York.

Powell, M. J. D. (1981). Approximation Theory and Methods, Cambridge Univ. Press, Cambridge.

Powers, D. (1979). Boundary Value Problems, Academic Press, New York.

Shao, J. and Tu, D.(1995). The Jackknife and Bootstrap,Springer-Verlag, New York.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution, *Ann. Stat.*, **9**, 1135–1151.

Tong, Y. L. (1990). The Multivariate Normal Distribution, Springer-Verlag, New York.