A Semiparametric Approach to the One-Way Layout

Benjamin Kedem
Department of Mathematics & ISR
University of Maryland
College Park, MD

"Give me a place to stand and rest my lever on, and I can move the Earth", (Archimedes, 287-212 B.C.)

Fokianos, K. (2004). Merging information for semiparametric density estimation. *JRSS*, B, **66**, 941-958.

Fokianos, K. and Kaimi, I. (2004). On the effect of misspecifying the density of ratio model. Submitted.

(•) Fokianos, Kedem, Qin, Short (2001). A semiparametric approach to the one-way layout. *Technometrics*, 43, 56–65.

Gilbert, P.B. (2000). Large sample theory of maximum likelihood estimates in semiparametric biased sampling models. *Annals of Statistics*, **28**, 151-194.

(•) Gagnon, R. (2005). Certain Computational Aspects of Power Efficiency and State Space Models. PhD Dissertation, Mathematics Department, University of Maryland, College Park.

Gilbert, P.B., Lele, S.R., Vardi, Y. (1999). Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika*, 86, 27-43.

Kay, R., and Little, S. (1987). Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika*, **74**, 495-501.

(•) Kedem, B. and Fokianos, K. (2002). *Regression Models for Time Series Analysis*. Wiley, New York.

(•) Kedem, B. and Gagnon, R. (2004). Time Series Prediction via density ratio modeling. Submitted

(•) Kedem, B., Wolff, D.B., and Fokianos, K. (2004). Statistical Comparison of Algorithms. *IEEE Trans. Instrum. Meas.*, Vol. 53, 770-776.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman and Hall/CRC, New York.

Patil, G.P., Rao, C.R., and Zelen, M. (1988). Weighted distributions. In *Encyclopedia of Statistical Sciences*, 9, 565-71, Kotz, S. and Johnson, N.L. eds.. Wiley, New York.

(•) Qi, Y. (2002). *Classification of Microarray Data*. MA Thesis, Mathematics Department, University of Maryland, College Park.

Qin, J. (1993). Empirical likelihood in biased sampling problems. *Annals Statistics*. 21, 1182-1186.

Qin, J., and J.F. Lawless (1994), Empirical likelihood and general estimating equations. *Annals of Statistics*, **22**, 300-325.

(•) Qin, J., and Zhang, B. (1997). A goodness of fit test for logistic regression models based on case-control data. in *Biometrika*, **84**, 609-618.

Rao, C.R. (1997). *Statistics and Truth*. World Scientific, New-Jersey.

Introduction

Statistical techniques based on normal theory have been central to the development and teaching of statistics. All along, however, there have been numerous studies of the consequences of departure from the normal assumption, and of transformation methods that produce nearly normal data. As examples we mention in particular the work of Miller (1986) who discusses situations where the normal and other assumptions break down, and the well known normalizing Box-Cox (1964) transformation.

The present work formulates an approach to analysis of variance that relaxes the normal assumption.

Consider the classical one-way ANOVA with $m = q+1$ independent normal random samples:

$$
\begin{aligned}
x_{11}, \ \ldots \ , \mathbf{x_{1n_1}} &\sim g_1(x) \\
&\vdots \\
x_{q1}, \ \ldots \ , x_{qn_q} &\sim g_q(x) \\
x_{m1}, \ \ldots \ , x_{mn_m} &\sim g_m(x)
\end{aligned}
$$

$$g_j(x) \sim \mathrm{N}(\ \mu_\mathrm{j}, \ \sigma^2), \ \ \mathrm{j} = 1, ..., \mathrm{m}.$$

The problem is to test the hypothesis

$$\mathrm{H}_0 : \mu_1 = \cdots = \mu_\mathrm{m}.$$

Then, holding $g_m(x)$ as a reference:

$$\frac{g_j(x)}{g_m(x)} = \exp(\ \alpha_j + \ \beta_j x),\ j = 1, ..., q, \qquad (1)$$

where

$$\alpha_j = \frac{\mu_m^2 - \ \mu_j^{\mathbf{2}}}{2\ \sigma^2}, \quad \beta_j = \frac{\mu_j - \ \mu_{\mathbf{m}}}{\sigma^2}, \ \ j = 1, ..., q,$$

we see that

$$\mathsf{H}_0: \boldsymbol{\mu}_1 = \cdots = \mu_{\mathsf{m}} \Leftrightarrow \mathsf{H}_0: \beta_1 = \cdots = \beta_{\mathsf{q}} = 0.$$

Generalization:

- Dispose of the normal assumption.
- Directly assume exponential distortion of a reference $g_m(x)$.
- Replace x by $h(x)$.

That is, assume

$$\frac{g_j(x)}{g_m(x)} = \exp(\alpha_j + \beta_j h(x)),\ j = 1, ..., q. \quad (2)$$

Test *equality of distributions* $g_j = g_m$, $j = 1, ..., q$ by

$$\mathbf{H}_0: \beta_1 = \cdots = \beta_q = 0.$$

Similar Models:

($\star$) Kay and Little (1987), ($\star$) Qin and Zhang (1997), Neyman (1937), Cox (1966), Anderson (1972, 1982), Prentice and Pyke (1979), Efron and Tibshirani (1996).

Example.

Special case of (2): Multinomial logistic regression.

- RV $\mathbf{y}$ s.t. $P(y = j) = \pi_j, \quad \sum_{j=1}^{m} \pi_j = 1.$

- Assume: For $j = 1, ..., m,$

$$P(y = j|x) = \frac{\exp(\alpha_j^* + \beta_j h(x))}{1 + \sum_{k=1}^{q} \exp(\alpha_k^* + \beta_k h(x))}$$

- Define: $f(x|y = j) = g_j(x), \quad j = 1, ..., m$

Then (2) holds with

$$\alpha_j = \alpha_j^* + \log[\pi_m/\pi_j], \quad j = 1, ..., q$$

Related Example.

Model (2) resembles Neyman (1937) smooth goodness of fit test: The null probability density function is embed in an order k alternative,

$$g(x;\theta,\beta) = C(\theta,\beta)\exp\left\{\sum_{i=1}^{k}\theta_i h_i(x,\ \beta)\right\} f(x;\beta)$$

where $\{h_i(x;\beta)\}$ are complete and orthonormal with respect to $f(x;\boldsymbol{\beta})$.

The Problem

Let $g \equiv g_m$, and consider:

- $g_j(x) = \exp(\alpha_j + \beta_j h(x))g(x), \quad j = 1, ..., q.$
- **Combined Data:** $\mathbf{t} \equiv (\mathbf{x_1}, ..., \mathbf{x_q}, \mathbf{x_m})'$.
- $n \equiv n_1 + \cdots + n_q + n_m$.

Use the **combined data** $\mathbf{t}$ to:

$\aleph_1$. Estimate the cdf $G(x)$ semiparametrically.

$\aleph_2$. Estimate $\boldsymbol{\alpha}=(\alpha_1, ..., \alpha_q)'$, $\boldsymbol{\beta}=(\beta_1, ..., \beta_q)'$.

$\aleph_3$. Test the hypothesis H_0: $\beta_1 = \cdots = \beta_q = 0$.

Estimation and Large Sample Results.

Follow Qin and Lawless (1994), Qin and Zhang (1997). MLE of $G(x)$ can be obtained by maximizing the likelihood over the class of step cdf's with jumps at the observed values $t_1, ..., t_n$. Accordingly, if $p_i = dG(t_i)$, $i = 1, .., n$, the empirical likelihood becomes,

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}, G) = \prod_{i=1}^{n} p_i \prod_{j=1}^{n_1} \exp(\ \alpha_1 + \ \beta_1 h(x_{1j})) \cdots \prod_{j=1}^{n_q} \exp(\ \alpha_q + \beta_q h(x_{qj}))$$

1. Get p_i

Fix $\boldsymbol{\alpha}, \boldsymbol{\beta}$. Maximize $\prod_{i=1}^{n} p_i$ subject to the m constraints:

$$\sum_{i=1}^{n} p_i = 1, \ \sum_{i=1}^{n} p_i[w_j(t_i) - 1] = 0, \quad j = 1, ..., q,$$

where

$$w_j(t_i) = \exp(\ \alpha_j + \ \beta_j h(t_i)), \ j = 1, ..., q.$$

Use Lagrange multipliers $\lambda_0 = n, \ \lambda_j = \nu_j n$. Then

$$\mathbf{p_i} = \frac{1}{n_m} \cdot \frac{1}{1 + \rho_1 w_1(t_i) + \cdots + \rho_q w_q(t_i)} \qquad (3)$$

where

$$\rho_j = n_j / n_m, \quad j = 1, ..., q.$$

2. Get $\boldsymbol{\alpha}, \boldsymbol{\beta}$

Profile log-likelihood up to a constant as a function of $\boldsymbol{\alpha}, \boldsymbol{\beta}$ only:

$$l = \sum_{j=1}^{n_1} [\alpha_1 + \beta_1 h(x_{1j})] + \cdots + \sum_{j=1}^{n_q} [\alpha_q + \beta_q h(x_{qj})] - \sum_{i=1}^{n} \log[1 + \rho_1 w_1(t_i) + \cdots + \rho_q w_q(t_i)] \quad (4)$$

Score equations for $j = 1, ..., q$:

$$\frac{\partial l}{\partial \alpha_j} = -\sum_{i=1}^{n} \frac{\rho_j w_j(t_i)}{1 + \rho_1 w_1(t_i) + \cdots + \rho_q w_q(t_i)} + n_j = 0$$

$$\frac{\partial l}{\partial \beta_j} = -\sum_{i=1}^{n} \frac{\rho_j h(t_i) w_j(t_i)}{1 + \rho_1 w_1(t_i) + \cdots + \rho_q w_q(t_i)} + \sum_{i=1}^{n_j} h(x_{ji}) = 0$$

3. Get $\widehat{G}$

The solution of the score equations gives the maximum likelihood estimators $\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}$, and consequently by substitution also $\widehat{p}_i$. Thus,

$$\widehat{p}_{\mathbf{i}}=\frac{1}{n_m} \cdot \frac{1}{1+\sum_{j=1}^{q} \rho_j \exp(\widehat{\alpha}_j + \widehat{\beta}_j h(t_i))}. \quad (5)$$

Therefore,

$$\widehat{G}(t) = \sum_{i=1}^{n} I(t_i \leq t)\widehat{p}_{\mathbf{i}} \quad (6)$$

Asymptotic Distribution of $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$

Define:

$$\nabla \equiv \left(\frac{\partial}{\partial \alpha_1}, ..., \frac{\partial}{\partial \alpha_q}, \frac{\partial}{\partial \beta_1} ..., \frac{\partial}{\partial \beta_q} \right)'$$

$$\rho_m \equiv 1, \;\; w_m(t) \equiv 1$$

$$E_j(t) \equiv \int h(t) w_j(t) dG(t)$$

$$Var_j(t) \equiv \int h^2(t) w_j(t) dG(t) - E_j^2(t)$$

$$\begin{aligned} A_0(j, j') &\equiv \int \frac{w_j(t) w_{j'}(t) dG(t)}{1 + \sum_{k=1}^q \rho_k w_k(t)} \\ A_1(j, j') &\equiv \int \frac{h(t) w_j(t) w_{j'}(t) dG(t)}{1 + \sum_{k=1}^q \rho_k w_k(t)} \\ A_2(j, j') &\equiv \int \frac{h^2(t) w_j(t) w_{j'}(t) dG(t)}{1 + \sum_{k=1}^q \rho_k w_k(t)} \end{aligned}$$

for $j, j' = 1, ..., q$.

$$\mathbf{V} \equiv Var\left[\frac{1}{\sqrt{n}}\nabla l(\boldsymbol{\alpha},\boldsymbol{\beta})\right]$$

$$\mathbf{S} \equiv \lim_{n\to\infty} -\frac{1}{n}\nabla\nabla' l(\boldsymbol{\alpha},\boldsymbol{\beta})$$

where $\mathbf{V},\mathbf{S}$ are $2q \times 2q$ matrices.

Remark:

The entries in $\mathbf{S}$ are obtained by a repeated application of the facts

$$\int dG(t) = 1, \quad \int w_j(t)dG(t) = 1, \quad j = 1,...,q.$$

Remark:

It should be noted that due to profiling, the matrix $\mathbf{S}$ is not the usual information matrix but it plays a similar role.

The entries in $\mathbf{V} \equiv Var\left[\frac{1}{\sqrt{n}}\nabla l(\boldsymbol{\alpha},\boldsymbol{\beta})\right]$ are:

$$\begin{aligned}
\frac{1}{n}Var\left(\frac{\partial l}{\partial \alpha_j}\right) &= \frac{\rho_j^2}{1+\sum_{k=1}^{q}\rho_k}[A_0(j,j) - \sum_{r=1}^{m}\rho_r A_0^2(j,r)] \\
\frac{1}{n}Cov\left(\frac{\partial l}{\partial \alpha_j}, \frac{\partial l}{\partial \alpha_{j'}}\right) &= \frac{\rho_j\rho_{j'}}{1+\sum_{k=1}^{q}\rho_k}[A_0(j,j') - \sum_{r=1}^{m}\rho_r A_0(j,r)A_0(j',r)] \\
\frac{1}{n}Cov\left(\frac{\partial l}{\partial \alpha_j}, \frac{\partial l}{\partial \beta_j}\right) &= \frac{\rho_j^2}{1+\sum_{k=1}^{q}\rho_k}[A_0(j,j)E_j(t) - \sum_{r=1}^{m}\rho_r A_0(j,r)A_1(j,r)] \\
\frac{1}{n}Cov\left(\frac{\partial l}{\partial \alpha_j}, \frac{\partial l}{\partial \beta_{j'}}\right) &= \frac{\rho_j\rho_{j'}}{1+\sum_{k=1}^{q}\rho_k}[A_0(j,j')E_{j'}(t) - \sum_{r=1}^{m}\rho_r A_0(j,r)A_1(j',r)] \\
\frac{1}{n}Var\left(\frac{\partial l}{\partial \beta_j}\right) &= \frac{\rho_j^2}{1+\sum_{k=1}^{q}\rho_k}[-A_2(j,j) + 2A_1(j,j)E_j(t) \\
&- \sum_{r=1}^{m}\rho_r A_1^2(j,r)] \\
&+ \frac{\rho_j}{1+\sum_{k=1}^{q}\rho_k}Var_j(t) \\
\frac{1}{n}Cov\left(\frac{\partial l}{\partial \beta_j}, \frac{\partial l}{\partial \beta_{j'}}\right) &= \frac{\rho_j\rho_{j'}}{1+\sum_{k=1}^{q}\rho_k}[-A_2(j,j') + A_1(j,j')(E_j(t)+E_{j'}(t)) \\
&- \sum_{r=1}^{m}\rho_r A_1(j,r)A_1(j',r)]
\end{aligned}$$

The entries in $\mathbf{S} \equiv \lim_{n\to\infty} -\frac{1}{n}\nabla\nabla' l(\boldsymbol{\alpha},\boldsymbol{\beta})$ are:

$$
\begin{aligned}
-\frac{1}{n}\frac{\partial^2 l}{\partial \alpha_j^2} &\rightarrow \frac{\rho_j}{1+\sum_{k=1}^q \rho_k}\int \frac{[1+\sum_{k\neq j}^q \rho_k w_k(t)]w_j(t)}{1+\sum_{k=1}^q \rho_k w_k(t)} dG(t) \\
-\frac{1}{n}\frac{\partial^2 l}{\partial \alpha_j \alpha_{j'}} &\rightarrow \frac{-\rho_j \rho_{j'}}{1+\sum_{k=1}^q \rho_k}\int \frac{w_j(t)w_{j'}(t)}{1+\sum_{k=1}^q \rho_k w_k(t)} dG(t) \\
-\frac{1}{n}\frac{\partial^2 l}{\partial \alpha_j \beta_j} &\rightarrow \frac{\rho_j}{1+\sum_{k=1}^q \rho_k}\int \frac{[1+\sum_{k\neq j}^q \rho_k w_k(t)]h(t)w_j(t)}{1+\sum_{k=1}^q \rho_k w_k(t)} dG(t) \\
-\frac{1}{n}\frac{\partial^2 l}{\partial \alpha_j \beta_{j'}} &\rightarrow \frac{-\rho_j \rho_{j'}}{1+\sum_{k=1}^q \rho_k}\int \frac{h(t)w_j(t)w_{j'}(t)}{1+\sum_{k=1}^q \rho_k w_k(t)} dG(t) \\
-\frac{1}{n}\frac{\partial^2 l}{\partial \beta_j^2} &\rightarrow \frac{\rho_j}{1+\sum_{k=1}^q \rho_k}\int \frac{[1+\sum_{k\neq j}^q \rho_k w_k(t)]h^2(t)w_j(t)}{1+\sum_{k=1}^q \rho_k w_k(t)} dG(t) \\
-\frac{1}{n}\frac{\partial^2 l}{\partial \beta_j \beta_{j'}} &\rightarrow \frac{-\rho_j \rho_{j'}}{1+\sum_{k=1}^q \rho_k}\int \frac{h^2(t)w_j(t)w_{j'}(t)}{1+\sum_{k=1}^q \rho_k w_k(t)} dG(t)
\end{aligned}
$$

Fact: Assume

$$g_j(x) = \exp(\alpha_j + \beta_j h(x)) g(x), \quad j = 1, ..., q,$$

with true parameters

$$\boldsymbol{\alpha}_0 = (\alpha_1, ..., \alpha_q)', \quad \boldsymbol{\beta}_0 = (\beta_1, ..., \beta_q)'.$$

Then under regularity conditions $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}$ are both consistent and asymptotically normal (See Sen and Singer 1993, Ch. 5),

$$\sqrt{n} \begin{pmatrix} \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{pmatrix} \Rightarrow \mathrm{N}\left(\mathbf{0}, \mathbf{S}^{-1}\mathbf{V}\mathbf{S}^{-1}\right). \qquad (7)$$

We sometimes write $\boldsymbol{\Sigma} = \mathbf{S}^{-1}\mathbf{V}\mathbf{S}^{-1}$.

Simulation results: $m = 3, h(x) = x$, 500 runs.

Case 1: Uniform, no distortion.
$g_1 \sim U(0,1),\ g_2 \sim U(0,1),\ g \sim U(0,1)$,
$n_1 = n_2 = n_3 = 200$.

	α_1	α_2	β_1	β_2
True	0.0000	0.0000	0.0000	0.0000
Est.	0.0063	-0.0006	-0.0127	0.0017
se	(0.0081)	(0.0081)	(0.0163)	(0.0162)

Case 2: Normal, distortion.
$g_1 \sim N(2,1),\ g_2 \sim N(3,1),\ g \sim$ N$(0,1)$,
$n_1 = 200, n_2 = 300, n_3 = 100$.

	α_1	α_2	β_1	β_2
True	-4.5000	-2.0000	3.0000	2.0000
Est.	-4.5541	-2.0356	3.0427	2.0360
se	(0.0172)	(0.0096)	(0.0115)	(0.0097)

Hypothesis Testing.

Under $\mathsf{H}_0 : \boldsymbol{\beta} = \mathbf{0}$, all the moments are taken with respect to the reference g.

Define the $q \times q$ matrix $\mathbf{A}_{11}$ whose jth diagonal element is

$$\frac{\rho_j[1 + \sum_{k \neq j}^{q} \rho_k]}{[1 + \sum_{k=1}^{q} \rho_k]^2}.$$

For $j \neq j'$, the jj' element is

$$\frac{-\rho_j \rho_{j'}}{[1 + \sum_{k=1}^{q} \rho_k]^2}.$$

The elements are bounded by 1 and the matrix is nonsingular,

$$|\mathbf{A}_{11}| = \frac{\prod_{k=1}^{q} \rho_k}{[1 + \sum_{k=1}^{q} \rho_k]^m} > 0.$$

Define

$$\mathrm{E}(\mathrm{t}^{\mathrm{k}}) \equiv \int \mathrm{h}^{\mathrm{k}}(\mathrm{t})\mathrm{dG}(\mathrm{t}),\ \mathrm{Var}(\mathrm{t}) \equiv \mathrm{E}(\mathrm{t}^2) - \mathrm{E}^2(\mathrm{t}).$$

Then

$$\mathbf{S} = \begin{pmatrix} 1 & \mathrm{E}(t) \\ \mathrm{E}(t) & \mathrm{E}(t^{\mathbf{2}}) \end{pmatrix} \otimes \mathbf{A_{11}}$$

with $\otimes$ denoting the Kronecker product. It follows that $\mathbf{S}$ is nonsingular,

$$|\mathbf{S}| = \{\mathrm{Var}(\mathrm{t})\}^{\mathrm{q}} |\mathbf{A}_{11}|^2$$

and,

$$\mathbf{S}^{-1} = \frac{1}{Var(t)} \begin{pmatrix} \mathrm{E}(t^{\mathbf{2}}) & -\mathrm{E}(t) \\ -\mathrm{E}(t) & 1 \end{pmatrix} \otimes \mathbf{A}_{11}^{-1}.$$

On the other hand, $\mathbf{V}$ is singular,

$$\mathbf{V} = \mathsf{Var}(t) \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{11} \end{pmatrix}$$

as is

$$\mathbf{\Sigma} = \mathbf{S^{-1}VS^{-1}} = \frac{1}{\mathsf{Var}(t)} \begin{pmatrix} \mathsf{E}^2(t) & -\mathsf{E}(t) \\ -\mathsf{E}(t) & 1 \end{pmatrix} \otimes \mathbf{A}_{11}^{-1}.$$

Luckily the right component is nonsingular and we finally have from (7),

$$\sqrt{n}\widehat{\boldsymbol{\beta}} \Rightarrow \mathbf{N}\left(\mathbf{0}, \frac{1}{\mathsf{Var}(t)}\mathbf{A}_{11}^{-1}\right). \tag{8}$$

It follows under $\mathsf{H}_0 : \boldsymbol{\beta} = \mathbf{0}$

$$(\star) \quad \mathcal{X}_1 = n\mathsf{Var}(t)\widehat{\boldsymbol{\beta}}'\mathbf{A}_{11}\widehat{\boldsymbol{\beta}} \tag{9}$$

is approximately distributed as $\chi^2(q)$, and H_0 can be rejected for large values of $\mathcal{X}_1$. In practice we need to estimate $\mathsf{Var}(t)$ from the combined data $\mathbf{t}$:

$$\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{h^2(t_i)\hat{p}_i} - \left(\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{h(t_i)\hat{p}_i}\right)^{\mathbf{2}}$$

3. Testing the Linear Hypothesis

We can go further and test the more general linear hypothesis $H_0 : \mathbf{H}\boldsymbol{\theta} = \mathbf{c}$ with $\mathbf{H}$ a $p \times 2q$, $p < 2q$, predetermined matrix of rank p, $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_q, \beta_1, \ldots, \beta_q)'$, and $\mathbf{c}$ is a vector in R^p. Then, using (7), we have under the hypothesis

$$\sqrt{n}(\mathbf{H}\hat{\boldsymbol{\theta}} - \mathbf{c}) \Rightarrow \mathbf{N}(\mathbf{0}, \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}')$$

Thus, the random variable

$$(\star) \quad \mathcal{X}_{\mathbf{2}} = n(\mathbf{H}\hat{\boldsymbol{\theta}} - \mathbf{c})'(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}')^{-1}(\mathbf{H}\hat{\boldsymbol{\theta}} - \mathbf{c}) \quad (10)$$

has an asymptotic chi-square distribution with p degrees of freedom provided the inverse exists (Sen and Singer 1993 p. 239). A consistent estimator of $\boldsymbol{\Sigma}$ can be obtained by replacing all the parameters by their maximum likelihood estimates.

It should be pointed out that in general the results obtained from (9) and (10) are different, since in (9) we substitute the exact value $\boldsymbol{\beta} = \mathbf{0}$ in $\boldsymbol{\Sigma}$, while (10) requires the maximum likelihood estimate of $\boldsymbol{\theta}$ instead.

Power comparison with the t-test as a function of β_1. $m = 2$, nominal level=0.05, $n_1 = n_2 = 30$. The reference distributions are N(0,1), LN(0,1), Gamma(3,1), respectively.

	Normal			Lognormal		
β_1	χ_1	t-Test	W-test	χ_1	t-Test	W-test
0.1	0.093	0.087	0.086	0.147	0.027	0.066
0.2	0.140	0.133	0.133	0.173	0.093	0.100
0.3	0.247	0.240	0.200	0.247	0.140	0.226
0.4	0.387	0.367	0.326	0.360	0.287	0.346
0.5	0.473	0.447	0.466	0.506	0.347	0.493
0.7	0.793	0.787	0.700	0.760	0.587	0.706
0.8	0.840	0.813	0.806	1.000	0.687	0.813
1.0	0.987	0.987	0.966	1.000	0.860	0.953

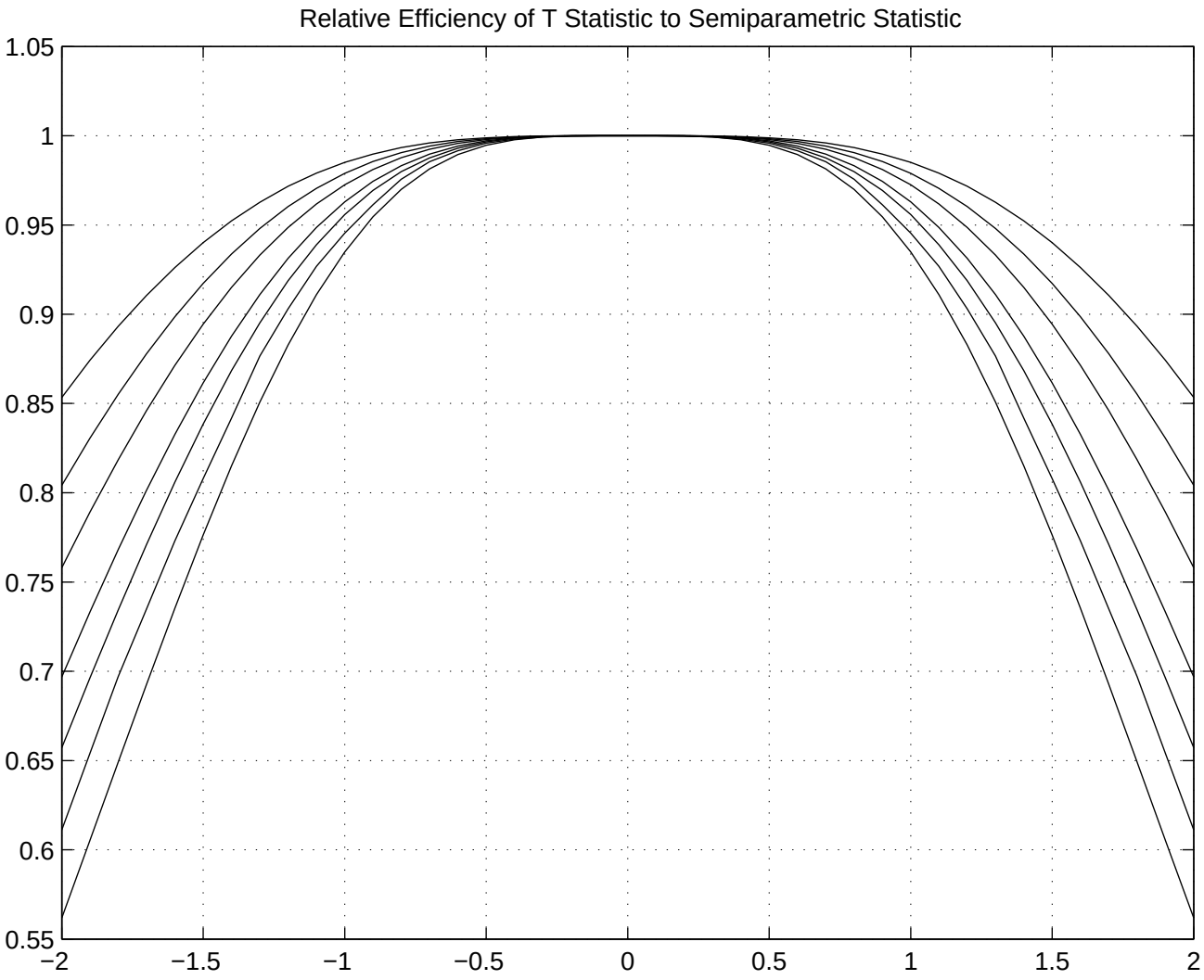
	Gamma		
β_1	χ_1	t-Test	W-test
0.1	0.100	0.047	0.046
0.2	0.127	0.067	0.060
0.3	0.140	0.087	0.120
0.4	0.240	0.153	0.160
0.5	0.253	0.160	0.213
0.7	0.493	0.380	0.393
0.8	0.513	0.340	0.480
1.0	0.693	0.527	0.520

Power comparison with the F-test as a function of β_1, β_2. $m = 3$. The reference distributions are N(0,1), LN(0,1), Gamma(3,1), respectively.

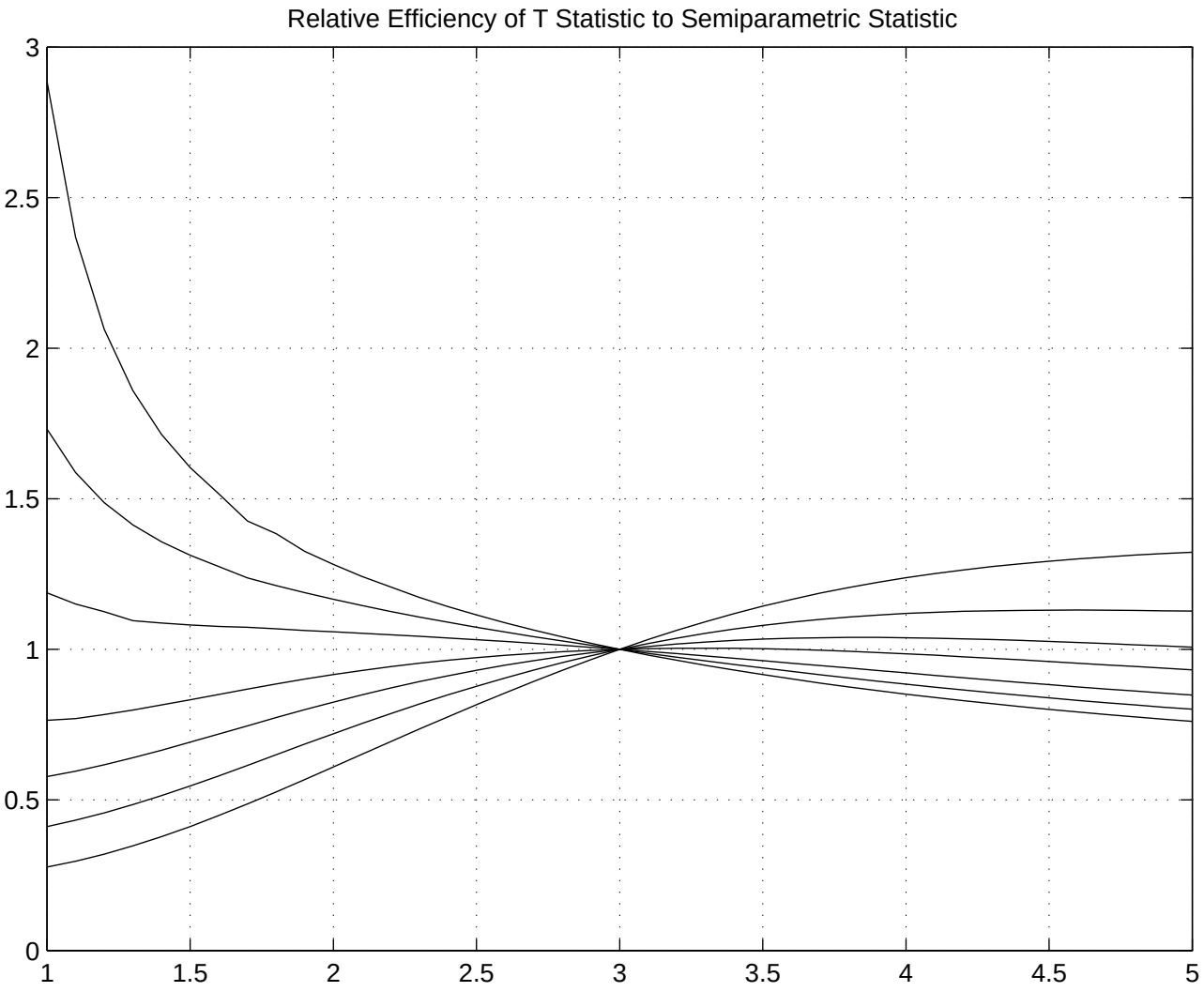
$n_i = 15\ \forall i$		Normal (0.05)			Lognormal (0.05)		
β_1	β_2	χ_1	F	K–W	χ_1	F	K–W
0.2	0.2	0.087	0.060	0.026	0.067	0.047	0.046
0.1	0.4	0.253	0.193	0.106	0.180	0.127	0.173
0.2	0.5	0.313	0.260	0.140	0.247	0.113	0.200
0.5	0.5	0.300	0.260	0.240	0.293	0.107	0.220
0.7	0.5	0.420	0.360	0.346	0.393	0.200	0.333
$n_i = 30\ \forall i$		Normal (0.01)			Lognormal (0.01)		
0.2	0.2	0.053	0.047	0.046	0.073	0.027	0.066
0.1	0.4	0.207	0.147	0.080	0.153	0.073	0.106
0.2	0.5	0.207	0.180	0.160	0.240	0.093	0.140
0.5	0.5	0.307	0.273	0.300	0.287	0.067	0.246
0.7	0.5	0.447	0.380	0.373	0.413	0.160	0.400

$n_i = 15\ \forall i$		Gamma (0.05)		
β_1	β_2	χ_1	F	K–W
0.2	0.2	0.093	0.060	0.053
0.1	0.4	0.167	0.087	0.060
0.2	0.5	0.233	0.093	0.100
0.5	0.5	0.246	0.093	0.146
0.7	0.5	0.287	0.140	0.113
$n_i = 30\ \forall i$		Gamma (0.01)		
0.2	0.2	0.047	0.000	0.020
0.1	0.4	0.067	0.020	0.060
0.2	0.5	0.127	0.033	0.026
0.5	0.5	0.113	0.060	0.053
0.7	0.5	0.227	0.120	0.106

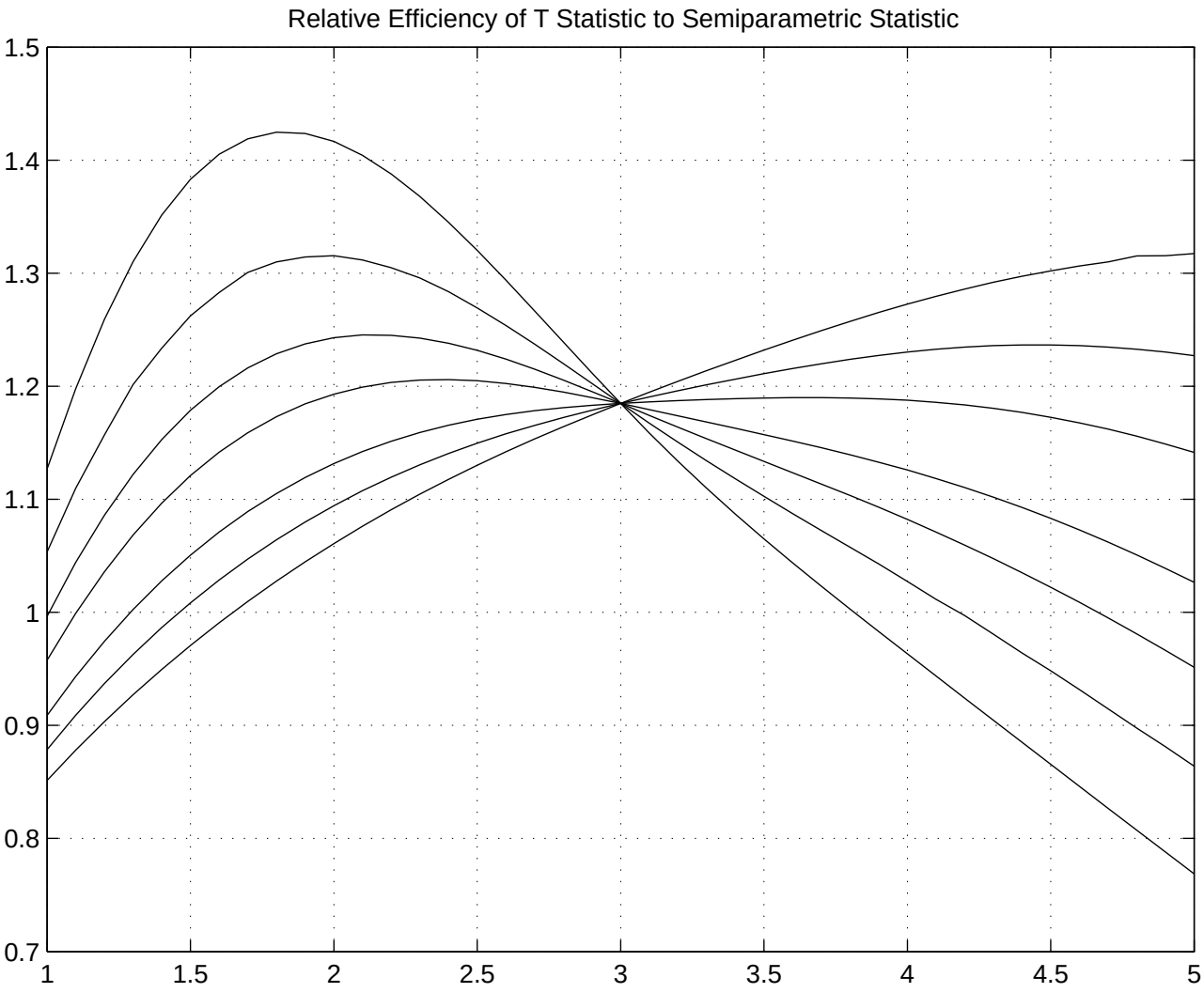
Relative power efficiency: t-test vs semiparametric test. Normal case.



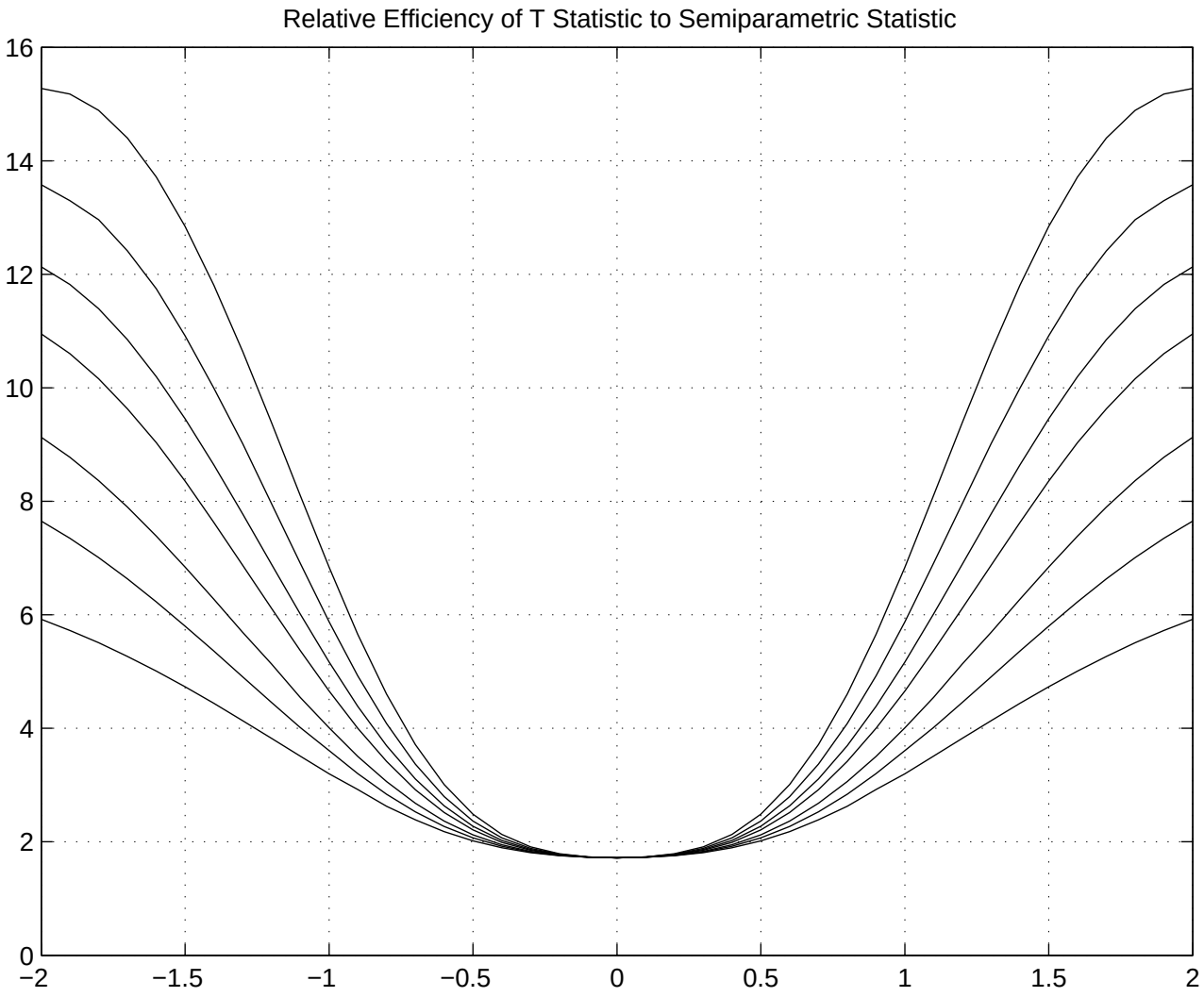
Relative power efficiency: t-test vs semiparametric test. Gamma SCALE case.



Relative power efficiency: t-test vs semiparametric test. Gamma SHAPE case.



Relative power efficiency: t-test vs semiparametric test. Log-Normal case.



Application to Radar Meteorology

The methodology described above is applied here to reflectivity data obtained from two different radars (or "algorithms") at two different time periods. Data obtained by random sampling large data sets.

Kwajalein radar: S-band (10 cm) KPOL radar, located on Kwajalein Island at the southern end of the Kwajalein Atoll, Marshall Islands.

Brown Radar: C-band radar aboard NOAA ship Ronald H. Brown (RHB) at sea near Kwajalein Island.

The data obtained during the first period are referred to suggestively as **Kwajalein1, Brown1**, while those from the second period are called **Kwajalein2, Brown2**.

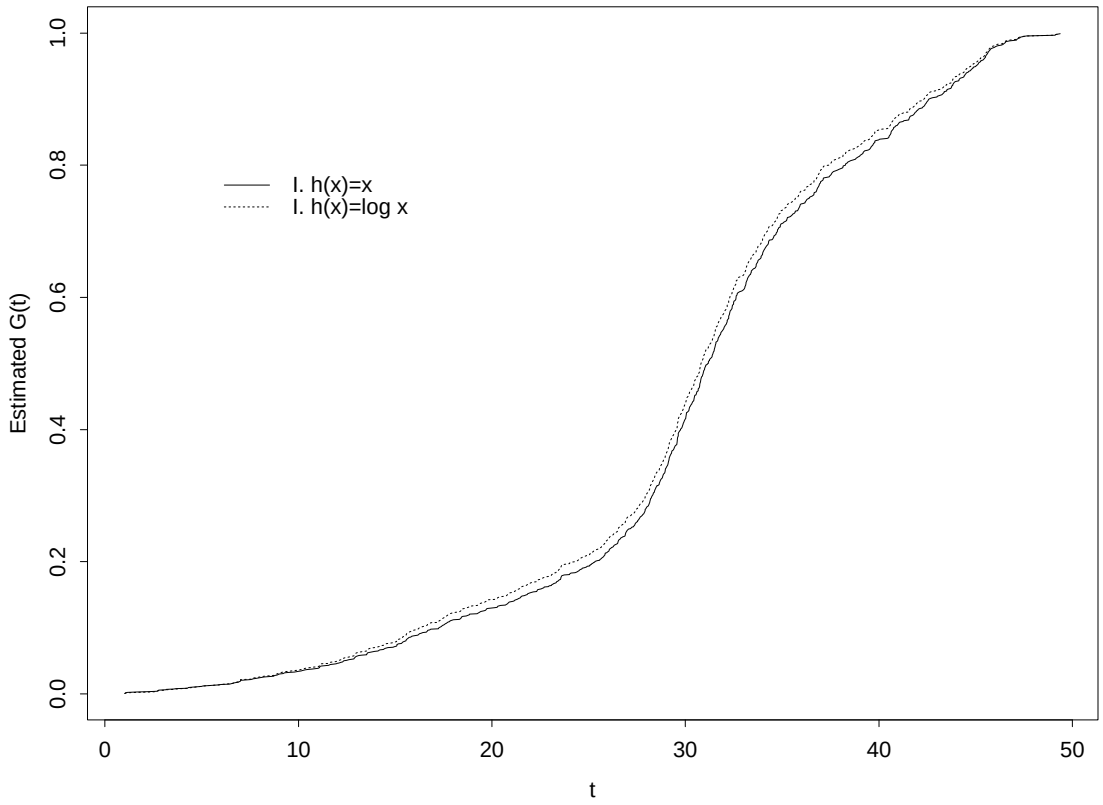
Kwajalein1, Brown1 (Reference)

Typical $\widehat{\alpha}_1, \widehat{\beta}_1$ values from different samples, and typical p-values of the $\mathcal{X}_1$ test statistic. $n_1 = n_2 = 500$.

Small p-values: The hypothesis that the data come from the same radar (algorithm) is rejected quite conclusively.

$h(x)$	$\widehat{\alpha}_1$	$\widehat{\beta}_1$	$\mathcal{X}_1$	p-value
x	0.784	-0.027	14.503	1.399e-03
	0.764	-0.025	12.032	5.226e-03
	1.244	-0.042	33.476	7.216e-09
	0.707	-0.024	12.204	4.768e-04
	0.909	-0.030	17.292	3.206e-05
$\log(x)$	1.319	-0.396	6.520	0.011
	1.544	-0.465	8.245	0.005
	1.908	-0.575	12.744	3.572e-04
	1.871	-0.562	11.202	8.169e-04
	2.050	-0.621	16.510	4.838e-05

Estimated Brown1 reference $G(x)$ with $h(x) = x$ and $h(x) = \log(x)$ obtained by combining data from Kwajalein1 and Brown1. $n_1 = n_2 = 500$.



Brown1, Brown1 (reference).

Typical $\widehat{\alpha}_1, \widehat{\beta}_1$ values from different samples, and typical p-values of the $\mathcal{X}_1$ test statistic. $n_1 = n_2 = 500$.

When both samples come from the same Brown1 radar there is a dramatic decrease in the $\mathcal{X}_1$ values and consequently an appreciable increase in the corresponding p-values. The test recognizes quite decisively the fact the data were generated by the same algorithm.

$h(x)$	$\widehat{\alpha}_1$	$\widehat{\beta}_1$	$\mathcal{X}_1$	p-value
x	-0.078	0.003	0.140	0.709
	0.005	-0.000	0.001	0.980
	-0.139	0.005	0.457	0.499
	0.112	-0.004	0.274	0.601
$\log(x)$	-0.584	0.175	1.723	0.189
	0.095	-0.028	0.042	0.838
	-0.225	0.067	0.250	0.617
	-0.027	0.008	0.003	0.959

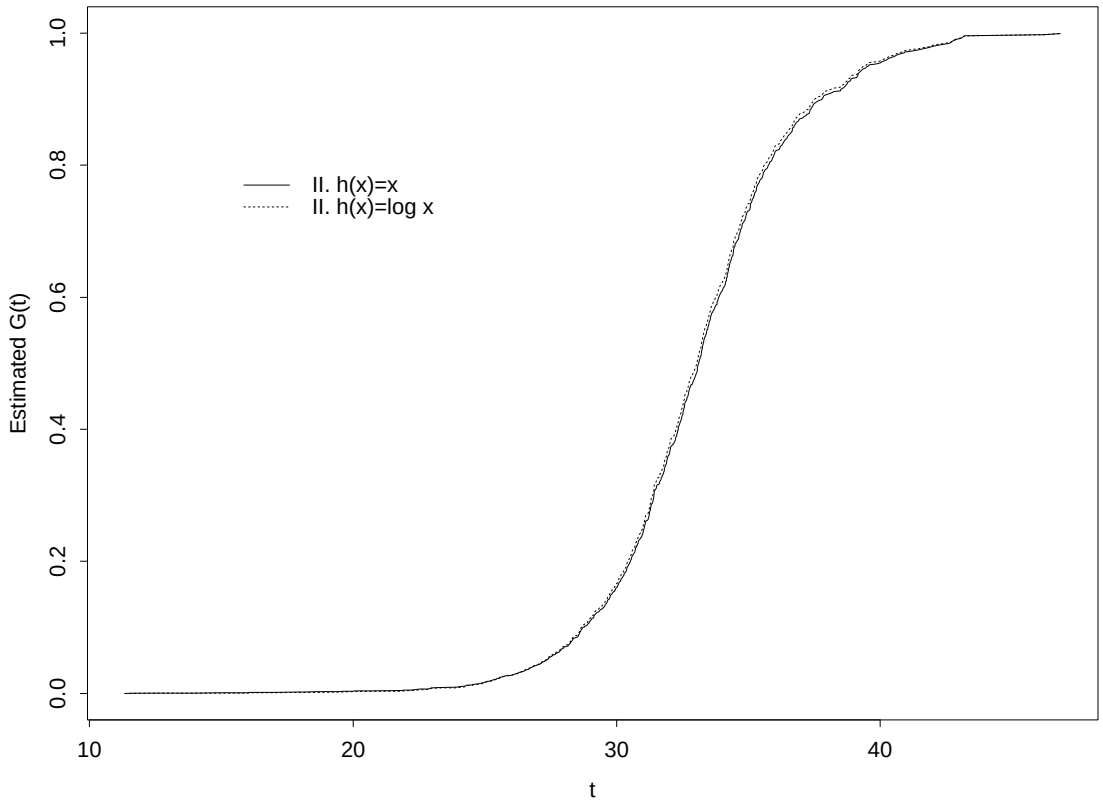
Kwajalein2, Brown2 (reference).

Typical $\widehat{\alpha}_1, \widehat{\beta}_1$ values, and typical p-values of the $\mathcal{X}_1$ test statistic. $n_1 = n_2 = 500$.

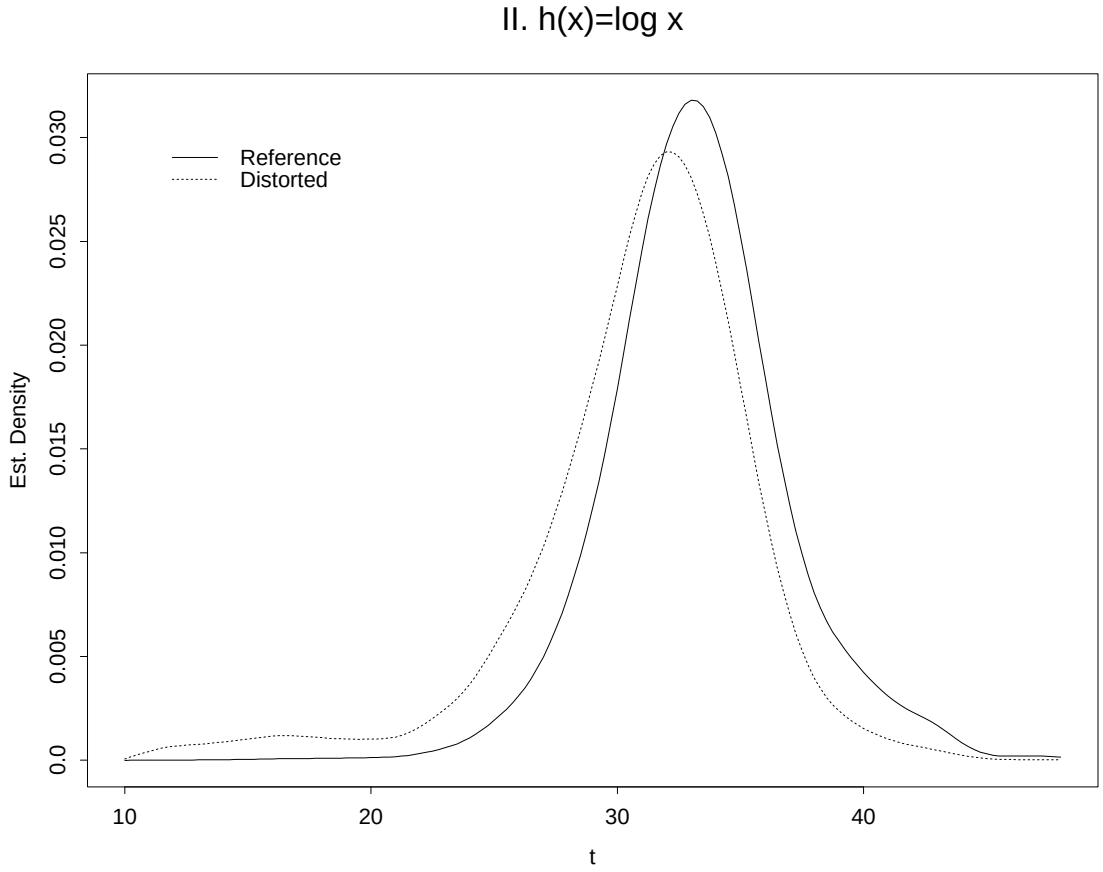
Small p-values: Radars are not the same.

$h(x)$	$\widehat{\alpha}_1$	$\widehat{\beta}_1$	$\mathcal{X}_1$	p-value
x	5.323	-0.164	88.332	0
	3.975	-0.123	52.279	4.815e-13
	4.695	-0.146	74.950	0
	5.016	-0.156	85.325	0
log(x)	14.359	-4.142	54.526	1.534e-13
	18.625	-5.367	79.723	0
	14.880	-4.302	60.788	6.328e-15
	13.580	-3.921	49.771	1.727e-12

Estimated Brown2 reference $G(x)$ with $h(x) = x$ and $h(x) = \log(x)$ obtained by combining data from Kwajalein2 and Brown2. $n_1 = n_2 = 500$.



Smoothed estimated reference $g(x)$ of Brown2 with $h(x) = \log(x)$ and its distortion obtained by combining data from Kwajalein2 and Brown2. $n_1 = n_2 = 500$.



Brown2, Brown2 (reference).

The situation changes dramatically when the data are from the same radar. Then the p-values are high implying great similarity or equivalently lack of distortion.

Typical $\widehat{\alpha}_1, \widehat{\beta}_1$ values, and typical p-values of the $\mathcal{X}_1$ test statistic. $n_1 = n_2 = 500$.

$h(x)$	$\widehat{\alpha}_1$	$\widehat{\beta}_1$	$\mathcal{X}_1$	p-value
x	0.479	-0.014	0.757	0.384
	0.396	-0.012	0.521	0.471
	0.184	-0.006	0.106	0.744
	-0.190	0.006	0.129	0.720
log(x)	-0.515	0.148	0.125	0.724
	-0.248	0.071	0.026	0.872
	0.882	-0.253	0.269	0.604
	-0.900	0.259	0.367	0.545

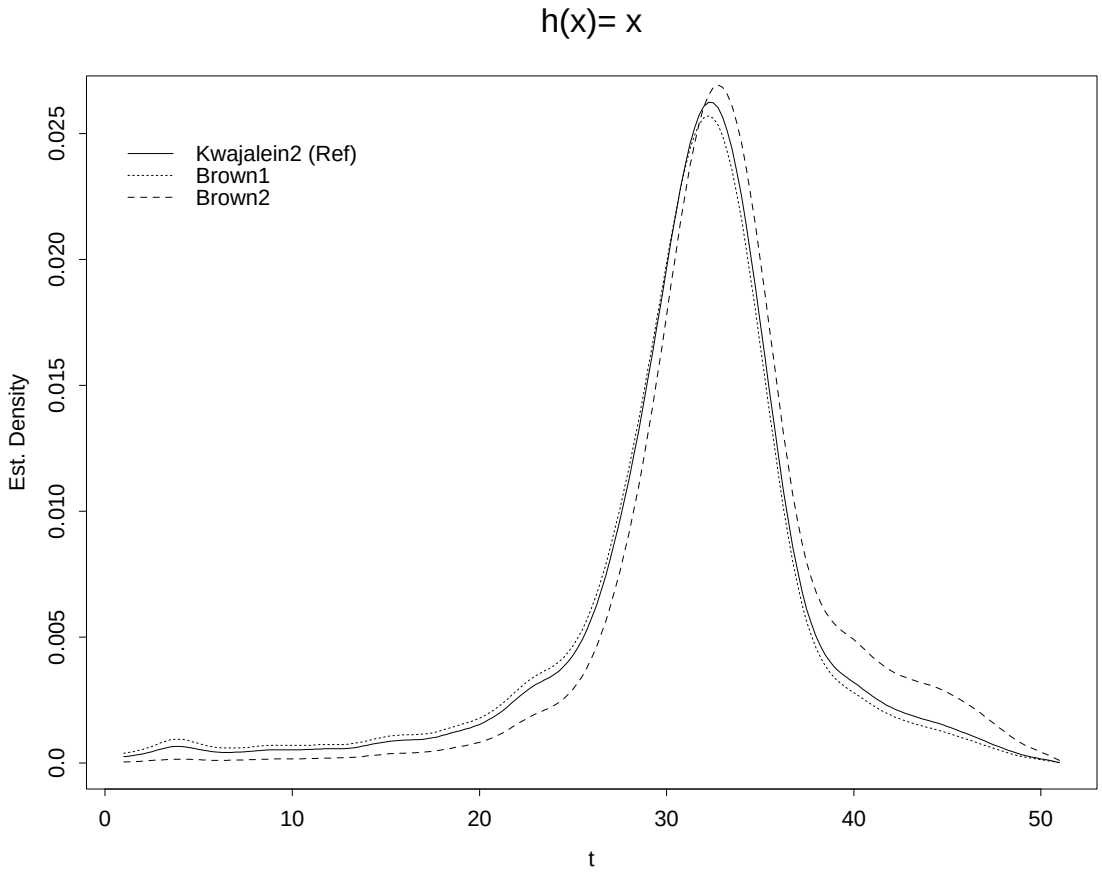
Brown1, Brown2, Kwajalein2 (reference).

Typical $\widehat{\alpha}_i, \widehat{\beta}_i,\ i = 1, 2$, values and typical p-values of the $\mathcal{X}_1$ test statistic. $n_1 = n_2 = n_3 = 500$.

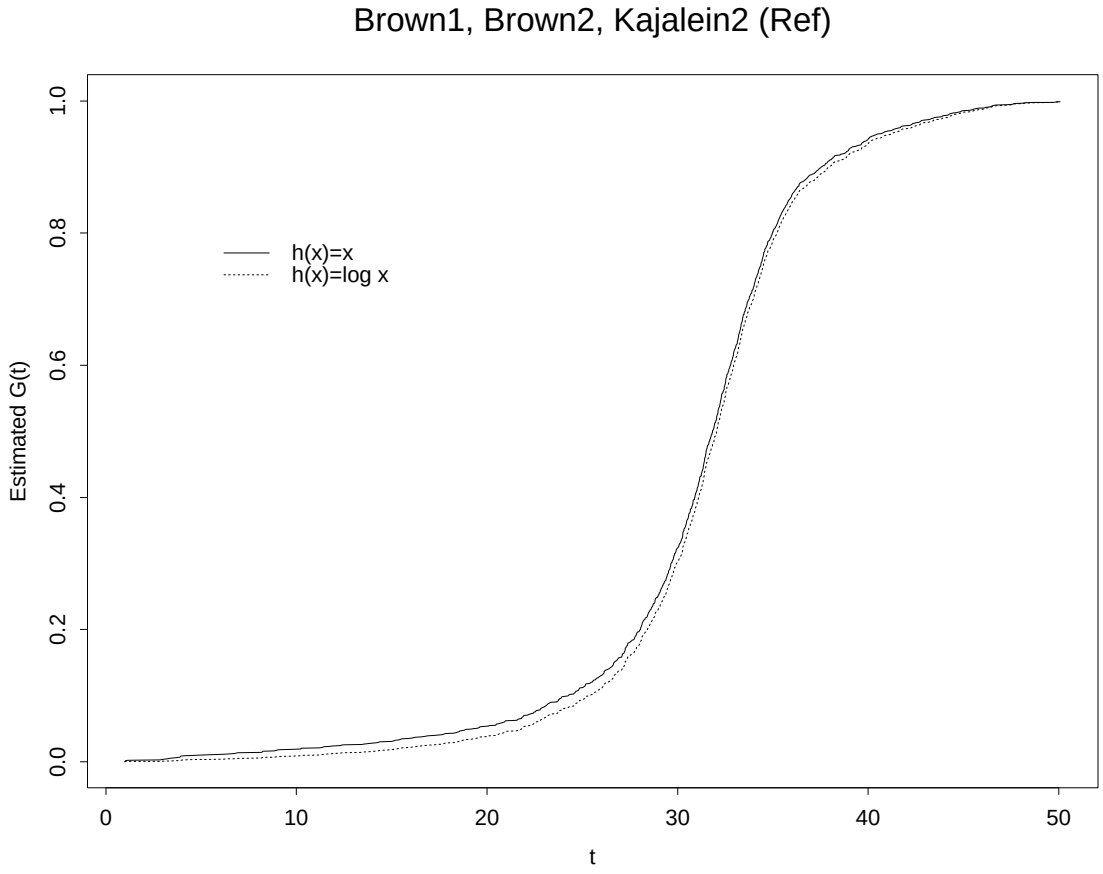
The data produced by the "three radars" are not statistically the same.

$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\mathcal{X}_1$	p-value
$h(x) = x$					
0.534	-1.528	-0.017	0.047	45.6	1.24e-10
0.409	-1.558	-0.013	0.048	43.4	1.24e-10
0.265	-1.711	-0.009	0.053	46.6	7.60e-11
0.087	-1.552	-0.003	0.049	36.4	1.23e-08
0.200	-1.751	-0.006	0.054	44.6	2.08e-10
$h(x) = \log x$					
2.564	-5.266	-0.753	1.519	78.6	0.000000
2.517	-4.877	-0.739	1.409	67.6	2.00e-15
2.322	-4.108	-0.684	1.188	62.5	2.64e-14
2.247	-4.801	-0.660	1.386	76.0	0.000000
2.304	-4.943	-0.678	1.427	75.7	0.000000

The spline–smoothed estimated reference density $g(x)$ of Kwajalein2 obtained from the combined data of the "three radars" with $h(x) = x$ and its two tilted forms.



The corresponding Kwajalein2 cumulative distribution $G(x)$ for both $h(x) = x$ and $h(x) = \log(x)$.



Kwajalein2, Kwajalein2, Kwajalein2 (reference).

For data from the same radar and the same period the p-values increase greatly.

Typical $\widehat{\alpha}_i, \widehat{\beta}_i,\ i = 1, 2$, values and typical p-values of the $\mathcal{X}_1$ test statistic. $n_1 = n_2 = n_3 = 500$.

$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\mathcal{X}_1$	p-value
$h(x) = x$					
0.108	0.049	-0.003	-0.002	0.283	0.868
0.065	-0.003	-0.002	0.000	0.135	0.935
0.227	-0.041	-0.007	0.001	1.896	0.388
0.239	-0.220	-0.008	0.007	4.707	0.095
$h(x) = \log x$					
0.453	2.278	-0.132	-0.665	1.929	0.381
-0.792	-0.223	0.231	0.065	0.250	0.882
-0.359	0.735	0.105	-0.215	0.553	0.758
1.665	1.246	-0.485	-0.363	1.014	0.602

Summary.

By appealing to a semiparametric statistical method, we have shown how to:

(•) Extend one way ANOVA to non-normal data.

(•) Adjust, correct, or calibrate data producing algorithms or instruments given data from several algorithmic sources of which one is taken as a reference.

(•) Combine data from *all sources*, reliable as well as biased or distorted sources, in order to estimate the probability distribution of the data from the reliable source or algorithm.

Remarks:

1. For non-normal data with known $h(x)$, the $\mathcal{X}_1$ test appears to be more powerful than the t and F tests.

2. Experience indicates that for skewed data $h(x) = \log x$ is quite appropriate, while $h(x) = x$ is suitable for symmetrically distributed data.

3. For positive data the problem of choosing $h(x)$ has been considered recently in Fokianos and Kaimi (2004). However the issue of estimating $h(x)$ in general is still open and it will be taken up elsewhere.

Extension: Density Estimation
(Fokianos (2004))

Smoothing the increments of $\widehat{G}_l$ for $l = 1, 2$ and setting $w_1 \equiv \exp(\alpha_1 + \boldsymbol{\beta}_1' \mathbf{h}(x))$, $w_2 \equiv 1$, amounts to the following estimators

$$\widehat{g}_l(x) = \frac{1}{h_n} \sum_{i=1}^{2} \sum_{j=1}^{n_i} \widehat{p}_{ij} \widehat{w}_l(x_{ij}) K\left(\frac{x - x_{ij}}{h_n}\right), \quad l = 1, 2,$$

where h_n is a sequence of window widths such that $h_n \to 0$ as $n \to \infty$ and K is a kernel function.

The following hold:

- $$\int \widehat{g}_l(x)dx = 1.$$

- $$\int x\widehat{g}_l(x)dx = \frac{1}{n_l}\sum_{j=1}^{n_l} x_{lj},$$

- $$\int x^2\widehat{g}_l(x)dx = \frac{1}{n_l}\sum_{j=1}^{n_l} x_{lj}^2 + h_n^2 k_2,$$

Fact: The following estimators

$$\tilde{g}_l(x) = \frac{1}{n_2 h_n} \sum_{i=1}^{2} \sum_{j=1}^{n_i} \frac{w_l(x_{ij})}{\sum_{k=1}^{m} \rho_k w_k(x_{ij})} K\left(\frac{x - x_{ij}}{h_n}\right)$$

satisfy

$$\hat{g}_l(x) = \tilde{g}_l(x) + O_p(n^{-1/2}), \quad l = 1, 2.$$

But

$$\mathrm{E}\left[\tilde{g}_l(x)\right] = g_l(x) + \frac{1}{2} h_n^2 g_l''(x) k_2 + o(h_n^2)$$

and

$$\mathrm{Var}\left[\tilde{g}_l(x)\right] = \frac{1}{n_l h_n} \frac{\rho_l w_l(x) g_l(x)}{\sum_{k=1}^{m} \rho_k w_k(x)} \int K^2(t) dt + o((nh_n)^{-1})$$

as $n \to \infty$ and $h_n \to 0$ such that $nh_n \to \infty$.

The pooled data yields kernel density estimates with the same amount of bias but less variable.

It turns out that under some regularity conditions

1.

$$\begin{aligned}\text{AMISE}\left[\widehat{g}_l(x)\right] &= \frac{1}{4}h_n^4 k_2^2 \int (g_l^{''}(x))^2 dx \\ &+ \frac{1}{n_l h_n}\int \frac{\rho_l w_l(x) g_l(x)}{\sum_{k=1}^m \rho_k w_k(x)} dx \int K^2(t)dt.\end{aligned}$$

2. The asymptotically optimal bandwidth–which is found by minimizing AMISE $[\widehat{g}_l(x)]$–is equal to

$$\begin{aligned}h_n^* &= \left(\int \frac{\rho_l w_l(x) g_l(x)}{\sum_{k=1}^m \rho_k w_k(x)} dx\right)^{1/5} \left(\int K^2(t)dt\right)^{1/5} \\ &\left(\int (g_l^{''}(x))^2 dx\right)^{-1/5} k_2^{-2/5} \rho_l^{-1/5} n^{-1/5}.\end{aligned}$$

3. Assigning h_n^* from the above expression, we obtain that the asymptotic mean integrated square error is equal to

$$\begin{aligned}\text{AMISE}^*\left[\widehat{g}_l(x)\right] &= \frac{5}{4}\left(\int \frac{\rho_l w_l(x) g_l(x)}{\sum_{k=1}^m \rho_k w_k(x)} dx\right)^{.8} \left(\int K^2(t)dt\right)^{.8} \\ &\left(\int (g_l^{''}(x))^2 dx\right)^{1/5} k_2^{2/5} \rho_l^{-4/5} n^{-4/5}.\end{aligned}$$

The new density estimator reduces the AMISE of the classical density estimator.