SF 298 MASTER COPY KEEP THIS COPY FOR REPRODUCTION PURPOSES

REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
		Technical Report

4. TITLE AND SUBTITLE

Title on Technical Report

5. FUNDING NUMBERS

DAAH04-93-G-0201

6. AUTHOR(S)

Author(s) listed on Technical Report

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(ES)

Florida State University
Tallahassee, Florida 32306

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

ARO 31029.5-MA

11. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Abstract on Report

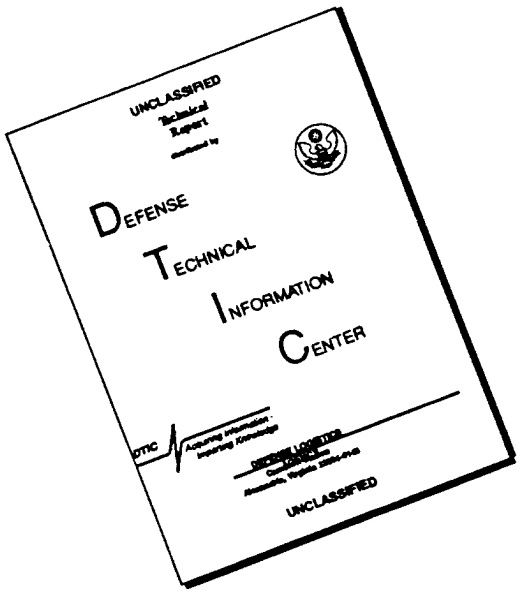
14. SUBJECT TERMS

15. NUMBER IF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

Bahadur slope of the t-statistic for a contaminated normal[†]

Narasinga R. Chaganty
Department of Mathematics
Old Dominion University

Jayaram Sethuraman[1]
Department of Statistics
The Florida State University

March 1996

FSU Techincal Report Number **M 907**

USARO Technical Report Number **D-137**

[†]Research partially supported by the U. S. Army research office grant number [1]DAAH04-93-G-0201. The United States Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

AMS 1991 subject classification: Primary 62F03, 62F05, 62G35. Secondary 60F10.
Key words and phrases: Bahadur slope, Large Deviations, Robustness, Tukey model.

Bahadur slope of the t-statistic for a contaminated normal[†]

NARASINGA R. CHAGANTY[1] AND JAYARAM SETHURAMAN[2]

Old Dominion University and Florida State University

ABSTRACT

In this paper we derive the Bahadur slope of the t-statistic based on a random sample from contaminated normal distribution, using some results in large deviation theory. We also present a table of Bahadur slopes at various alternatives at several levels of contamination.

1. INTRODUCTION. To study robustness of standard tests of location in a normal model, one generally studies their properties under the Tukey model (see Tukey(1960)) of contaminated normal alternatives, namely, the probability distributions $P_{(\epsilon,\,\theta,\sigma)}$ with probability density function (pdf)

$$f_{(\epsilon,\,\theta,\sigma)}(x) = (1-\epsilon)\,\phi(x;\theta,1) + \epsilon\,\phi(x;\theta,\sigma) \tag{1}$$

for $0 < \epsilon < 1$, where $\phi(x;\theta,\sigma)$ is the pdf of a normal distribution with mean θ and variance σ^2.

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from $f_{(\epsilon,\,\theta,\sigma)}(x)$ and that we wish to test the null hypothesis $\theta = 0$ using the t-statistic $T_n = \overline{X}_n/S_n$, where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. The robustness of this t-test as measured by Pitman

[†]Research partially supported by the U. S. Army research office grant numbers [1]DAAL03-91-G-0179, [2]DAAH04-93-G-0333. The United States Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

AMS 1991 subject classification: Primary 62F03, 62F05, 62G35. Secondary 60F10.
Key words and phrases: Bahadur slope, Large Deviations, Robustness, Tukey model.

efficiency has been studied in the famous Princeton study by Andrews et al. (1972). In this paper we derive the large deviation rate of T_n under $P_{(\epsilon,\, 0,\sigma)}$ which allows us to obtain the Bahadur slopes of the t-test under a general alternative $P_{(\epsilon,\, \theta,\sigma)}$. Following the practice of other authors, we set σ equal to 3, and give the Bahadur slopes for various values of ϵ and θ in Table 1. This table gives an indication of the region of robustness of the t-test as measured by the Bahadur slope. The robustness of the t-test, in the sense of Bahadur efficiency, is gleamed by comparing the slope at the contaminated distribution $P_{(\epsilon,\, \theta,3)}$ with the slope at the uncontaminated distribution $P_{(0,\, \theta,3)}$. As expected, Table 1 shows that there is adequate robustness in a region of small values of ϵ . Furthermore, for a fixed θ the slope is a decreasing function of ϵ and for a fixed ϵ the slope is an increasing function of θ.

The exact distribution of T_n^2 under $P_{(\epsilon,\, \theta,\sigma)}$ has been derived in Lee and Gurland (1977). We will derive the large deviation rate of T_n under $P_{(\epsilon,\, 0,\sigma)}$ and the Bahadur slope under the alternative $P_{(\epsilon,\, \theta,\sigma)}$ in Section 2.

2. LARGE DEVIATION RATES AND BAHADUR SLOPES. We refer to the excellent monograph of Varadhan (1984) for an introduction to the theory of large deviations and to the monograph of Bahadur (1971) for the concept of Bahadur slopes and efficiencies. One needs a strong law under the alternative and a large deviation result under the null hypothesis to obtain Bahadur slopes. It is easy to see from the usual strong law of large numbers that

$$T_n \;\rightarrow\; m(\epsilon,\theta,\sigma) = \frac{\theta}{\sqrt{(1-\epsilon)+\epsilon\sigma^2}}, \tag{2}$$

with probability one under $P_{(\epsilon,\, \theta,\sigma)}$. We need to obtain a result of the form

$$\frac{1}{n}\log P_{(\epsilon,0,\sigma)}\,(T_n \geq m) \rightarrow -\gamma(m), \tag{3}$$

where $\gamma(m)$ is continuous in m. The function $\gamma(m)$ is usually referred to as the large deviation rate function. It then follows that the Bahadur slope is given by

$$c(\epsilon,\theta,\sigma) = 2\,\gamma(m(\epsilon,\theta,\sigma)). \tag{4}$$

We now proceed with the derivation of $\gamma(m)$. Note that the event $\{T_n^2 > m^2\}$ is equal to

the event $\{W_n > 0\}$ where W is the quadratic form $W = X' A X/n$ with $A = J - n a I$, $a = m^2/(1+m^2)$, I is the identity matrix and J is a matrix of ones. Since the distribution of T_n is symmetric under $P_{(\epsilon, 0, \sigma)}$, we have

$$P(T_n \geq m) = \frac{1}{2} P(W_n \geq 0). \tag{5}$$

(From here onwards, P without a suffix corresponds to the probability under $P_{(\epsilon, 0, \sigma)}$.) The left hand side of (5) can be appropriately approximated by using the moment generating function (mgf) of W_n which is given by

$$\begin{aligned} M_n(t) &= E[\exp(tW_n)] \\ &= \sum_{k=0}^{n} \binom{n}{k} (1-\epsilon)^k \epsilon^{(n-k)} |I - \frac{2t}{n} A_k A|^{-1/2} \end{aligned} \tag{6}$$

where $A_k = \text{diag}(\overbrace{1, \ldots, 1}^{k}, \overbrace{\sigma^2, \ldots, \sigma^2}^{n-k})$. Let $p = k/n$ and $q = 1 - p$. Using a matrix determinant formula, (see Appendix), we can show that

$$\begin{aligned} M_{np}(t) &= |I - \frac{2t}{n} A_k A|^{-1/2} \\ &= (f_1(t))^{-np/2} (f_2(t))^{-nq/2} \left(\frac{p f_2(t) f_3(t) + q f_1(t) f_4(t)}{f_1(t) f_2(t)} \right)^{-1/2} \end{aligned} \tag{7}$$

where $f_1(t) = 1 + 2at$, $f_2(t) = 1 + 2at\sigma^2$, $f_3(t) = 1 - 2t(1-a)$ and $f_4(t) = 1 - 2t\sigma^2(1-a)$. Thus the mgf of W_n is given by

$$M_n(t) = \sum_{k=0}^{n} \binom{n}{k} (1-\epsilon)^k \epsilon^{(n-k)} M_{np}(t) \quad \text{for } t_*(p) < t < t^*(p), \tag{8}$$

where $t_*(p)$, $t^*(p)$ are the roots of the quadratic equation $p f_2(t) f_3(t) + q f_1(t) f_4(t) = 0$.

From the above formula for the mgf $M_n(t)$, we can conclude that the distribution of W_n is a mixture distribution. More precisely, let K be a binomial random variable with parameters n and $(1 - \epsilon)$. Given $K = k$, let W_{nk} be a random variable with mgf given by M_{np}, where $p = k/n$. From (8) we can see that W_n is equal in distribution to W_{nK}. This observation coupled with a theorem of Varadhan, see Theorem 2.2 in Chaganty (1993), is useful to derive the large deviation rate function for the random variable W_n. Theorem 1 below shows that the conditions in Varadhan's theorem are indeed satisfied in our problem.

THEOREM **1** *Let K be a binomial random variable with parameters n and $(1-\epsilon)$. Given $K = k_n = np_n$, let W_{nk_n} be a random variable with mgf, $M_{np_n}(t)$, defined in (7). If $p_n \to p$ then*

$$F_n(p_n) = \frac{1}{n} \log P(W_{nk_n} \geq 0) \to F(p) \text{ as } n \to \infty, \tag{9}$$

where $F(p) = -\frac{1}{2}\left[p \log f_1(t^*(p)) + q \log f_2(t^*(p))\right]$, $q = 1 - p$.

Proof: Upper bound: By Chebyshev's inequality it follows that

$$\begin{aligned} \limsup_n \frac{1}{n} \log P(W_{nk_n} \geq 0) &\leq \lim_n \frac{1}{n} \log M_{nk_n}(t) \\ &= -\frac{1}{2}\left[p \log f_1(t) + q \log f_2(t)\right] \end{aligned} \tag{10}$$

for any $0 < t < t^*(p)$. Hence

$$\begin{aligned} \limsup_n F_n(p_n) &= \limsup_n \frac{1}{n} \log P(W_{nk_n} \geq 0) \\ &\leq \inf_{0<t<t^*(p)} -\frac{1}{2}\left[p \log f_1(t) + q \log f_2(t)\right] \\ &= F(p). \end{aligned} \tag{11}$$

Lower bound: Let G_{np_n} denote the distribution function of W_{nk_n}. Let us introduce another random variable V_n with the conjugate distribution function given by

$$dG_{nt_n}(x) = \frac{\exp(x\, t_n)}{M_{np_n}(t_n)}\, dG_{np_n}(x) \tag{12}$$

where $t_n = t^*(p)(1 - \frac{1}{n})$. Now for any $\delta > 0$ we have

$$\begin{aligned} P(W_{nk_n} \geq 0) &= \int_0^\infty dG_{np_n}(x) = M_{np_n}(t_n) \int_0^\infty \exp(-x\, t_n)\, dG_{nt_n}(x) \\ &\geq M_{np_n}(t_n) \int_0^{n\delta} \exp(-x\, t_n)\, dG_{nt_n}(x) \\ &\geq M_{np_n}(t_n) \exp(-n\delta\, t_n)\, P(0 \leq V_n \leq n\delta). \end{aligned} \tag{13}$$

Therefore,

$$\frac{1}{n} \log P(W_{nk_n} \geq 0) \geq \frac{1}{n} \log M_{np_n}(t_n) - \delta t_n + \frac{1}{n} \log P(0 \leq V_n \leq n\delta). \tag{14}$$

Since $p_n \to p$ as $n \to \infty$ it follows from (7)

$$\frac{1}{n} \log M_{np_n}(t_n) \to -\frac{1}{2} [p \log f_1(t^*(p)) + q \log f_2(t^*(p))] = F(p). \tag{15}$$

We will now show that the limiting distribution of V_n/n is a translated gamma distribution. To find the limiting distribution, we first note that the mgf of V_n/n is given by $M_n(s) = M_{np_n}(s_n)/M_{np_n}(t_n)$, where $s_n = t_n + s/n$. It is easy to check that

$$M_n(s) \to M(s) = \exp(-s\,c) \left(\frac{t^*(p)}{t^*(p) - s}\right)^{1/2} \text{ as } n \to \infty, \tag{16}$$

for $s < t^*(p)$, where $c = [ap/(1 + 2at^*(p)) + aq\sigma^2/(1 + 2at^*(p)\sigma^2)]$. Thus V_n/n converges in distribution to $V - c$, where V is a Gamma random variable with shape parameter $1/2$ and scale parameter $1/t^*(p)$. Therefore,

$$P(0 \le V_n/n \le \delta) \to P(c \le V \le c + \delta) > 0 \text{ as } n \to \infty. \tag{17}$$

From (14), (15) and (17) we get

$$\begin{aligned} \liminf_n F_n(p_n) &= \liminf_n \frac{1}{n} \log P(W_{np_n} \ge 0) \\ &\ge F(p) - \delta\, t^*(p). \end{aligned}$$

Since δ is arbitrary we get $\liminf_n F_n(p_n) \ge F(p)$. This completes the proof of the theorem.

We are now in a position to derive the large deviation rate function for T_n. From Theorem 1 we have,

$$F_n(p_n) = \frac{1}{n} \log P(W_n \ge 0 | K = n\,p_n) \to F(p). \tag{18}$$

whenever $p_n \to p$. Note that

$$\frac{1}{n} \log P(W_n \ge 0) = \frac{1}{n} \log \int \exp(nF_n(p))\, d\mu_n(p) \tag{19}$$

where μ_n is the distribution of K/n. Since the distribution of K is binomial, it is known that the sequence of probability measures $\{\mu_n\}$ obeys the large deviation principle (see Varadhan (1984) for the definition) with rate function

$$h(p) = p \log(p/(1 - \epsilon)) + q \log(q/\epsilon).$$

Using the theorem of Varadhan, see Theorem 2.2 in Chaganty (1993), and (18) and (19) it follows that

$$\frac{1}{n}\log P(W_n \geq 0) \to \sup_{0<p<1}(F(p) - h(p)). \tag{20}$$

From (5) and (20) we get

$$\frac{1}{n}\log P(T_n \geq m) \to -\gamma(m)$$

where $\gamma(m) = \inf_{0<p<1}[-F(p) + h(p)]$.

The rate function $\gamma(m)$ can easily be computed numerically using Newton-Raphson method. In Table 1 we present the Bahadur slope, $c(\epsilon, \theta, \sigma) = 2\,\gamma(m(\epsilon, \theta, \sigma))$, for different values of ϵ and θ when $\sigma = 3$. Note that a large value of $c(\epsilon, \theta, \sigma)$ indicates that the test statistic T_n requires smaller sample size to detect that particular alternative.

Table 1. Slope of the t-statistic $c(\epsilon, \theta, \sigma)$, for the contaminated normal model, when $\sigma = 3$.

$\epsilon \setminus \theta$	0.25	0.50	1.0	1.5	2.0	2.5	3.0
0.00	0.06066	0.22314	0.69314	1.17866	1.60944	1.98100	2.30258
0.05	0.04488	0.17380	0.56738	0.99566	1.39154	1.74208	2.05046
0.10	0.03508	0.14056	0.48860	0.87952	1.24944	1.58306	1.88092
0.15	0.02866	0.11598	0.42936	0.79694	1.14852	1.46908	1.75732
0.25	0.02090	0.08422	0.33264	0.67160	1.00634	1.31238	1.58918

REMARK 1 It is possible to derive, in a similar manner, the Bahadur slope of the t-statistic, for a random sample of n observations with common pdf given by $f(x) =$

$\sum_{i=1}^{L} \pi_i \phi(x;\theta,\sigma_i)$, $\sum_{i=1}^{L} \pi_i = 1$, and $\pi_i > 0$ for all $L \geq 1$. In this case the multinomial distribution plays the role of the binomial distribution in the derivation of the slope. More generally, using the results of Chaganty (1993), we can also establish the large deviation principle for the t-statistic for this model.

3. APPENDIX. In (7) we have used the following determinant formula. Let

$$S = \begin{bmatrix} \overbrace{b\,I + c\,J}^{k} & \overbrace{c\,J}^{(n-k)} \\ e\,J & d\,I + e\,J \end{bmatrix},$$

where b, c, d and e are constants, and as before, I is the identity matrix and J is the matrix of ones. Then we can verify that

$$|\,S\,| = b^k\, d^{(n-k)} \left(1 + \frac{k\,c}{b} + \frac{(n-k)\,e}{d}\right). \tag{21}$$

To obtain the simplification in equation (7), we use the above formula (21) with the substitutions $b = f_1(t)$, $d = f_2(t)$, $c = -\dfrac{2t}{n}$ and $e = -\dfrac{2t\,\sigma^2}{n}$.

Department of Mathematics and Statistics
Old Dominion University
Norfolk, Virginia 23529

Department of Statistics
Florida State University
Tallahassee, Florida 32306

REFERENCES

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). *Robust Estimates of Location: Survey and Advances.* Princeton University Press, Princeton, N. J.

Bahadur, R. R. (1971). *Some limit theorems in statistics.* SIAM, CBMS/NSF Regional Conference in Applied Math. **4** (SIAM, Philadelphia).

Chaganty N. R. (1993). Large deviations for joint distributions and statistical applications. Technical Report #TR93-2, Department of Mathematics & Statistics, Old Dominion University, Norfolk.

Lee, A. F. S. and Gurland, J. (1977). One sample t-test when sampling from a mixture of normal distributions, *Ann. Statist.*, **5**, 803-807.

Tukey, J.W. (1960). A survey of sampling from contaminated distributions, *Contributions to Probability and Statistics*, Ed., I.Olkin, (Stanford University Press, Stanford, CA).

Varadhan, S. R. S. (1984). *Large deviations and Applications.* SIAM, CBMS/NSF Regional Conference in Applied Math. **46** (SIAM, Philadelphia).