One–sample one–sided Student t test under contaminants

Ryszard Zieliński

Summary. A paradoxical behavior of the t test under $\varepsilon-$contamination is presented. The paradox consists in that under a fixed distribution of contaminants an increasing of the probability of the appearance of a contaminant may decrease the violation of the size of the test! A simple explanation of the phenomenon is given. It is revealed which contaminants make the test conservative and which make it liberal: it appears that, in spite of the established opinion, conservatism or liberalism of the test depends not so much on the tails of the contaminating distribution as on where its support is located.

AMS subject classification: 62 F 04, 62 F 35

Keywords: $t-$test, $\varepsilon-$contamination, conservatism, liberalism, contaminants, outliers

---

Partially supported by Grant KBN 2 P301 010 04

1. INTRODUCTION. The problem how does the Student t test behave under nonnormal distributions is as old as the test itself. The prevailing opinion is that the test is conservative for a sample from a long-tailed distribution or, at least, that its conservatism or liberalism depends on the tails of the parent distribution (as usual the conservatism means that the size of the test, or the Type I error, is smaller than the assumed significance level). The literature of the subject is abundant: see Hotelling (1961), Efron (1969), Johnson (1978), Cressie (1980), Benjamini (1983) for the most fundamental results.

The paper treats a somewhat different problem: we are interested in how much is the size of the test changing if to a Gaussian sample some contaminants are added. It appears that it is not long-tailedness but rather the location of the contaminating distribution which makes the test conservative or liberal.

Let X be a normally distributed random variable with an unknown variance and consider the problem of testing the hypothesis $H : EX = 0$ vs. $K : EX > 0$. Let $X_1, X_2, \ldots, X_n$ be a sample from the parent distribution $N(0,1)$; the cumulative distribution function of $N(0,1)$ will be denoted as usual by Φ. Let

$$t = \frac{\bar{X}}{S}\sqrt{n-1},$$

where $\bar{X} = \sum_1^n X_i/n$, $S^2 = \sum_1^n (X_i - \bar{X})^2/n$, and, given a (small) positive α, let k by the critical value of the one-sided Student t test:

$$Prob(t > k) = \alpha.$$

Now suppose that the sample comes from $(1-\varepsilon)\Phi(x) + \varepsilon F(x)$ for some $\varepsilon \in [0, 1/2]$, where $F(x)$ is "another distribution". Our interpretation of such situation is: $X_1, X_2, \ldots, X_n$ are independent; for every $i = 1, 2, \ldots, n$, the observation X_i comes from the parent distribution Φ with probability $1-\varepsilon$ or from the parent distribution F with probability ε. In the latter case X_i is said to be a contaminant (other terms are also used – see e.g. Beckman and Cook (1983)). Let

$$\alpha(\varepsilon, F) = Prob(t > k)$$

be the Type I error when X_i is distributed according to $(1-\varepsilon)\Phi+\varepsilon F$, so that $\alpha(0,F)=\alpha(\varepsilon,\Phi)=\alpha$ for every ε and F. If F is a symmetric distribution (i.e. $F(x)=1-F(-x)$) then $\alpha(\varepsilon,F)$ is the size of the t test "under contaminants". The distribution F represents the magnitude of the contaminant and ε represents the amount of the contaminants in the sample.

The aim of the paper is to demonstrate and to explain some peculiarities concerned with the behavior of the function $\alpha(\varepsilon,F)$ under fixed F.

2. Contamination Let us begin with the well known Tukey (1960) choice of $F(x)=\Phi(x/\sigma)$, $\sigma>1$. Results of 100,000 simulation experiments for $\alpha=0.01$, $\alpha=0.05$, $\alpha=0.1$, $\sigma=12$, and $n=10$ are presented in Fig. 1. All simulation results presented in the paper concern the case that $n=10$; for other values of n the results similar.

The results look rather strange: under a fixed distribution of contaminants an increasing of the probability of the appearance of a contaminant may decrease the violation of the size of the test!

One can argue that if the contaminants do not differ substantially from the proper observations (like in Tukey contamination where contaminants concentrate around zero) then the size of the test would not differ substantially from the assumed value of the significance level, possibly slightly oscillating around it when ε varies.

Consider the following model ("mean-shift model") with the contaminating distribution F defined as follows:

$$(*)\qquad F_\mu(x)=\frac{1}{2}\Phi(\frac{x-\mu}{\sigma})+\frac{1}{2}\Phi(\frac{x+\mu}{\sigma})$$

For large μ this can be considered as a suitable model for "true" outliers. It appears that under contamination (*) the Student t test behaves similarly as under the Tukey contamination. Some numerical results for $\mu=24$ and $\sigma=0.1$, based on 100,000 simulations, are presented in Fig.2 which is designed in full analogy to Fig.1.

3. Fixed number of contaminants in the sample. Denote by $P_i(F), i=1,2,\ldots,n$, the probability $Prob(t>k)$ under exactly i contaminants with the distribution F. Then

$$\alpha(\varepsilon,F)=\sum_{i=0}^{n}\binom{n}{i}\varepsilon^i(1-\varepsilon)^{n-i}P_i(F).$$

It appears that for some i, (i+1) outliers influence the size of the test to a lesser extent than i outliers do. As an example consider $P_i(*)$ for $\mu = 24$ and $\sigma = 0.1$: some numerical results, based on 100,000 simulations, are presented in Fig. 3.

4. THE t STATISTIC UNDER CONTAMINANTS Until now we have been discussing the behavior of some properties of the distribution of the statistic t rather than the behavior of t itself. Suppose that one of the observations $X_1, X_2, \ldots, X_n$, say X_1, is replaced by a contaminant, u say, and write the t statistic in the form

$$t(u) = \frac{\frac{1}{n}(u + s_1)}{\sqrt{\frac{1}{n}(u^2 + s_2) - [\frac{1}{n}(u + s_1)]^2}}\sqrt{n-1},$$

where $s_1 = \sum_2^n X_i$ and $s_2 = \sum_2^n X_i^2$. The shape of the function $t(u)$, $-\infty < u < +\infty$, is presented in Fig.4.

The fact that the function $t(u)$ is bounded appears to play the crucial role in the behavior of the t test under contaminants. The function $t(u)$ achieves its maximum (if $s_1 > 0$) or its minimum (if $s_1 < 0$) at $u_0 = s_2/s_1$, the extreme value being equal to

$$sign(s_1) \cdot \sqrt{(n-1)(s_2 + s_1^2)/[(n-1)s_2 - s_1^2]}.$$

If $s_1 = 0$ (observe however that $P\{s_1 = 0\} = 0$) then the function $t(u)$ increases and takes all its values in the interval $(-1, 1)$. If s_1 is close to zero then the maximum of $t(u)$ is close to 1 and every contaminant makes the test conservative (if the critical value of the test is greater than 1 which is a common case). If for a positive constant, say λ, s_2/s_1 is "far to the left of λ" and the outlier u is "far to the right of λ", then $t(u)$ would not exceed the critical value and the test is conservative. On the other hand if the distribution of contaminants is concentrated near s_2/s_1 (if contaminants fall into interval (a, b) in Fig.5) then the test is liberal. But s_2/s_1 is a random variable. Its distribution for $n = 10$, i.e. the distribution of $\sum_2^n X_i^2 / \sum_2^n X_i$ with X_i normal $N(0, 1)$, is presented in Fig. 6. Looking at Fig. 6 one can easily understand why the first outlier $(*)$ with a large value of μ makes the test conservative. To see why the second outlier makes the test liberal look at Fig. 7 where the

distribution of s_2/s_1 with $s_2 = u^2 + \sum_3^n X_i^2$, and $s_1 = u + \sum_3^n X_i$ (solid line) is presented (here u is an $(*)$ outlier with $\mu = 24$ and $\sigma = 0.1$ and numerical results are based on 10,000 simulations): the first outlier shifts s_2/s_1 to the right and the second outlier easily meets that value.

5. Heavy tails. Observe that $t(u) \to 1$ as $u \to +\infty$ and $t(u) \to -1$ as $u \to -\infty$. It follows that if the critical value of the test is greater than 1 then a sufficiently large (positive) outlier makes the test conservative (negative outliers always make the test conservative). If however the critical value is less than 1 (which is a case of a rather theoretical interest) than sufficiently large positive outlier make the test liberal.

6. Exactly one outlier in the sample. The results enable us to answer the following question: what is the maximal size of the α-level t-test when exactly one contaminant in the sample appears: we shall denote that quantity by $MS(\alpha, n)$. The obvious answer is: it is equal to the probability that $s_1 > 0$ and $t(u_0) > k$. After simple calculations we obtain:

$$\begin{aligned} MS(\alpha, n) &= P\left\{ \frac{(n-1)(s_2 + s_1^2)}{(n-1)s_2 - s_1^2} > k^2(\alpha, n) \quad \text{and} \quad s_1 > 0 \right\} \\ &= \frac{1}{2} P\left\{ \frac{s_2}{s_1^2} < \frac{n - 1 + k^2(\alpha, n)}{(n-1)(k^2(\alpha, n) - 1)} \right\} \\ &= \frac{1}{2} P\left\{ F(n-2, 1) < \frac{n}{(n-2)(k^2(\alpha, n) - 1)} \right\} \end{aligned}$$

where $F(n-2, 1)$ is a random variable with F distribution with $n-2$ and 1 degree of freedom. The values of $MS(\alpha, n)$ for some n and α are given in the following table:

n	Significance level α		
	0.10	0.05	0.01
3	0.2627	0.1793	0.0784
5	0.2172	0.1204	0.0340
10	0.2088	0.1033	0.0230
20	0.2091	0.0987	0.0200
∞	0.2114	0.0958	0.0178

If $n \to \infty$ then $MS(\alpha, n) \to 1 - \Phi\left(\sqrt{k_\alpha^2 - 1}\right)$ where $k_\alpha = \Phi^{-1}(1-\alpha)$; the appropriate asymptotic values are given in the last row of the above table.

References.

Beckman, R.J. and Cook, R.D. (1983): "Outliers.....s", *Technometrics*, 25, 119–163

Benjamini, Y. (1983): "Is the t Test Really Conservative When the Parent Distribution is Long–Tailed?", *Journal of the American Statistical Association*, 78, 645–654

Cressie, N. (1980): "Relaxing Assumptions in the One Sample t-Test", *Australian Journal of Statistics*, 22, 143-153

Efron, B. (1969): "Student's t Test Under Symmetry Conditions", *Journal of the American Statistical Association*, 63, 1278–1302

Hotelling, H. (1961): "The Behavior of Some Standard Statistical Tests Under Non-Standard Conditions", *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, I, 319-360

Johnson, J.J. (1978): "Modified t Test and Confidence Intervals for Asymmetrical Populations", *Journal of the American Statistical Association*, 73, 536-544

Tukey, J.W. (1960): "A Survey of Sampling from Contaminated Distributions", In *Contributions to Probability and Statistics.* Olkin, Ed. Stanford University Press, Stanford, Calif.

Ryszard Zieliński
Institute of Mathematics
Polish Acad. Sci.
P.O.B. 137
00–950 Warsaw, Poland
e-mail: rziel@impan.impan.gov.pl