robust estimation

a review

by
T.-E. Moschona
Mathematician (M.Sc.)

Partially based upon the M. Sc. thesis

Statistical procedures are widely used in the social sciences, as they provide techniques for drawing conclusions from quantitative data. The first part of this work is concerned with the robustness — insensitivity of procedures to departures from the assumptions— of some classical tests. The second part is a brief and partial review of the modern theory of robustness of a location parameter.

I. the robustness of classical tests for means and variances

For many experimenters the most commonly used statistical tests are those for comparing sample means and sample variances. Sometimes, they are used mechanically, with little regard to whether or not the required assumptions are satisfied. We will give a brief review here of the effects on them, of departures from the underlying assumptions, such as the effect of non-normality, the effect of inequality of variances and the effect of violation of the independence assumption. Any test or estimate which performs well under the above effects is usually referred to as «robust».

1. *The one sample t-test*

Let $$\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i \qquad s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

be the mean and variance of a sample of n observations $x_1, x_2, \ldots, x_n$ from a population with mean μ and unknown finite variance σ^2. Suppose we wish to test the null hypothesis $H_0:\mu=\mu_0$ against $H_1: \mu\neq\mu_0$, where μ_0 is a specified value.

Let $$t=\frac{\sqrt{n}\,(\bar{x}-\mu_o)}{s}$$

be the test statistic. Then if the parent population is normal and the observations are independently sampled, under H_0 t will follow a Student's t-distribution with n-l degrees of freedom, and the one-sample t-test of H_0 against H_1 at the $100\alpha\%$ level of significance is to reject H_0 it $|t|>t_{n-1;\alpha/2}$ where $t_{n-1;\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the t-distribution with n-l d.f.

1.1. The effect of non-normality

If the parent population is not normal, the statistic t no longer follows the t-distribution. However, if n is large and the population has a finite variance, it follows from the Central Limit Theorem (CLT) and the fact that s^2 converges stochastically to σ^2, that the statistic t has an asymptotic standard normal distribution. Thus, for sufficiently large samples the t-test is robust against non-normality. But, in the case of small samples, the effect on non-normality is quite important. Table 1 shows the approximate percentage increase in the critical value of t at the 21/2% (one-tail) significance level, when sampling from some «typical» non-normal populations.

Thus, positive kurtosis (γ_2) has little effect. Extreme negative kurtosis affects the critical value of t for samples up to 30 but the moderate negative kurtosis of a rectangular population ceases to have effect out if the sample size becomes greater than 15. The moderate skewness (γ_1) of the skew-normal population is effective up to samples of size 50 and the extreme skewness of the exponential population is effective up to samples of size 80. Various corections to the standard normal Tables have been proposed by Geary, Gayen, Tiku and Srivastava, when population normality cannot be assumed, based on Charlier Differential Series, Edgeworth expansions or Hermite polynomials.

1.2. The effect of departure from the independence assumption

Suppose that the observations x_i $i=1,\dots,n$ have a multivariate normal distribution with $E(x_i)=\mu$, $\mathrm{var}(x_i)=\sigma^2$ and ρ as the serial correlation coefficient. Then we find that for large n the statistic t will be asymptotically distributed as $N(0,1+2\rho)$ instead of $N(0,1)$. Thus, the effect of serial correlation on the true significance level of a two-sided t-test having a nominal significance level of 5% can be very serious, which is shown in Table 2.

2. *The two-sample t-test*

Consider two populations 1,2 with means μ_1, μ_2 and variances σ_1^2, σ_2^2 respectively, from which we obtain samples of size n_1, n_2. We wish to test the H_0: $\mu_1-\mu_2=0$ against H_1: $\mu_1-\mu_2\neq 0$. Then if the two populations are normal with the same variance σ^2 and all values are independently sampled, the two-sample t-test is based on the statistic

$$t=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

which under H_0 follows a t-distribution with n_1+n_2-2 d.f., where s^2 is the pooled estimator of σ^2, and the procedure for a test at the 100α% level of significance is to reject H_0 if $|t|>t_{n_1+n_2-2;\alpha/2}$.

2.1. The effect of non-normality

For sufficiently large samples, the two-sample t-test is robust against non-normality (the other assumptions being satisfied). In the case of small or moderate sample sizes, the results of studies made by Bartlett, Geary, showed that in the case of unequal sample sizes drawn from different populations, the deviation of the distribution of t from its normal theory law may be considerable. But in cases of equal sample sizes, even skewness in the parents are of little effect and if the parents are symmetrical the test will be robust even for differing sample sizes.

2.2. The effect of inequality of variances

If the parent populations are normal, with unequal variances, we can see that is distributed approximately for large sample sizes:

$$N\left(0,(\Theta+R)(R\Theta+1)^{-1}\right)\quad,\quad R=\frac{n_1}{n_2}\quad,\quad \Theta=\frac{\sigma_1^2}{\sigma_2^2}$$

TABLE 1

Parent population	Sample size 5	 15	 30	 50	 80
U-shaped (γ_1=0, γ_2=1. 6 9)	100%	0%	0%	0%	0%
Rectangular (γ_1=0, γ_2=-1.2)	100%	50%	0%	0%	0%
Exponential (γ_1=2, γ_2= 6)	100%	50%	40%	20%	0%
Skew-normal (γ_1=0. 92, γ_2=0. 9)	100%	25%	10%	0%	0%

TABLE 2

ρ :	−0.4	−0.3	−0.2	−0.1	0	0.1	0.2	0.3	0.4
Sign. Level:	1.10 ‹	0.002	0.011	0.028	0.050	0.074	0.09	0.12	0.14

Hence, apart from the case of equal sample sizes, (R=1) the effect of inequality of variances is very serious on the type I error, prob. Table 3 shows the effect of inequality of variances and unequal sample sizes on the nominal type I error prob. of 0.05, for large n_1, n_2.

TABLE 3

	Θ						
	0	1/5	1/2	1	2	5	∞
1	0.05	0.05	0.05	0.05	0.05	0.05	0.05
R 2	0.17	0.12	0.08	0.05	0.029	0.014	0.006
5	0.38	0.22	0.12	0.05	0.014	0.002	1.10^{-5}
∞	1.00	0.38	0.17	0.05	0.016	1.10^{-5}	0

3. *The analysis of variance*

Consider k populations with means μ_1, μ_k and finite variances $\sigma_1^2, \ldots\ldots, \sigma_k^2$ respectively, from which we obtain samples of size n_i, i=1,..,k. We wish to test the null hypothesis that all the means are equal against the general alternative that not all of them are. Then if the k populations are normal, have the same variance σ^2 and all the observations are independent, the usual ANOVA F-test is based on the statistic F.

$$F=\frac{B/(k-1)}{W/(N-k)}\,, \qquad B=\sum_{i=1}^{k} n_i(\bar{x}_{i.}-\bar{x}_{..})^2\,,$$

$$N=\sum_{i=1}^{k} n_i\,, \qquad \bar{x}_{..}=\frac{1}{N}\sum_i\sum_j x_{ij}\,, \qquad \bar{x}_{i.}=\frac{1}{n_i}\sum_j x_{ij}$$

$$W=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i.})^2\,,$$

which under the null hypothesis follows an F distribution with k-1 and N-k d.f., and the procedure for a test at the 100α% level of significance is to reject H_0 if $F>F_{k-1,\,N-k;\alpha}$.

3.1. The effect of non-normality

If all populations are not normal (the other assumptions holding) B/σ^2 and W/σ^2 no longer follow χ^2 distributions nor they are independent, and as the sample sizes increase the situation of the latter gets worse. In spite of this feature, the distribution of the ANOVA F-ratio is less dependent on the underlying distribution because of the correlation between the numerator and the denominator of it, and turns out to be robust against non-normality (assuming equal variances).

3.2. The effect of inequality of variances

The deviations which occur in the ANOVA F-test due to the above effect are usually not very serious when the sample sizes are equal but very serious when they are not. Table 4 shows the effect of unequal variances on the prob. of type I error at nominal 5% level.

TABLE 4

No. of samples	Ratio of sample variances	Sample sizes	Probab. of type I error
3	1;2;3	5,5,5	.056
		3,9,3	.056
		7,5,3	.092
		3,5,7	.040
3	1;1;3	5,5,5	.059
		7,5,3	.11
		9,5,1	.17
		1,5,9	.013
5	1;1;1;1;3	5,5,5,5,5	.07
		9,5,5,5,1	.14
		1,5,5,5,9	.02

Thus, for two of three lines of the Table, if not for all three, when the sample sizes are equal, the deviations can be considered bearable. However, this cannot be said for the other eight lines where the sample sizes are unequal. A common practice in using the ANOVA F-test has been to test first the assumption of equality of variances, before conducting the test. But this is not recommended. The result is to mask differences when they exist if $\gamma_2<0$ and to find when none exist if $\gamma_2>0$. When the samples are of unequal size and differences in variances might occur, it would seem logical to replace the ANOVA F-test by an alternative criterion proposed by Welch and James in 1951. The last one is robust to inequality of variances and almost certainly to non-normality also.

While tests on means are robust or moderate robust, tests on variances are extremely non-robust. For that, various alternatives have been proposed.

4. *The χ^2 test on a single variance*

Consider a population with unknown mean and finite variance from which we take a sample of size n. We wish to test the null hypothesis $H_0:\sigma^2=\sigma_0^2$ against the general alternative, where σ_0^2 is a specified value. Then if the parent population is normal and the observations are independent, the test statistic used is

$$V=\frac{(n-1)s^2}{\sigma_o^2}$$

which under H_0 has a χ^2 distribution with (n-1) d.f. and the usual procedure for a test at the 100 α% level is to reject H_0 if $V>\chi^2_{(n-1);\alpha/2}$ or if $V<\chi^2_{(n-1);1-\alpha/2.}$

4.1 The effect of non-normality

It can be easily proved that the size of the test based on the statistic V will be different from the nominal level of significance if γ_2 differs greatly from zero. Thus, the effect of non-normality is very serious, and the test not robust. Table 5 shows for a sig. level of 0.05, the actual probability of type I error for the usual two-tailed test.

5. *Tests for the equality of several variances*

Suppose we have k independent samples from populations with unknown means and finite variances $\sigma_1^2,.....\sigma_K^2$, and that we wish to test $H_0:\sigma_1^2=...=\sigma_K^2$ against the general alternative that not all the σ_i^2 are equal. The assumptions made here are that the populations are normal and that all the observations are independent.

5.1 The effect of non-normality

The standard test for the equality of two variances is the F test, based on the ratio of the sample variances, which is extremely non-robust to non-normality. The standard test for the equality of several variances is Bartlett's test, which is also not robust (see Box (1953)). The alternatives which have been proposed to them are Scheffé's χ^2-test, the Box test, the Box-Andersen test, the Jackknife test and the Levene's test (see Miller (1968)). We will only give a short comparison of them. For large samples the efficacies for the Box-Andersen test, Levene's, and Jackknife are identical and asymptotically more efficient than all the other tests. For small samples, Monte-Carlo studies have been carried out to compare their power functions. The results were that the F-test is non-robust. Its actual significance level under the null hypothesis is much smaller than the nominal value α for short-tailed distributions (uniform) and much larger than α for long-tailed distributions (double-exponential). Under H_0 the Bartlett test gives too few significant results for the uniform distribution (γ_2=-1.2) and too many for the double exponential. The Box-Andersen test and the Jackknife with m=1 (where m is the size of the n groups in which we divide the data) have about the same power and together with Scheffé's χ^2-test are the most powerful tests. Box-Andersen has slightly better power for α=0.05 level tests while the Jackknife is slightly better for α=.01 level tests. The Box test is tobust with respect to sign. level but its power is not as good as Box-Andersen or the Jackknife test.

In connection with part II the main effect of actual contamination is not considered for all the previous tests.

II. the modern theory of robust estimation, and robustness of a location parameter

Statistics like other bodies of scientific knowledge and technique, can undergo revolutions in both thought and practice. «There is no such thing as a universally applicable and acceptable method of going about doing science or statistics» (Irvine, Miles and Evans, 1979). By 1960 statisticians had recognized that:

— one never has a very accurate knowledge of the true underlying distribution;

TABLE 5

γ_2:	−1.5	−1	−0.5	0	0.5	1	2	4	7
Prob:	9.10^{-5}	0.006	0.024	0.05	0.08	0.11	0.17	0.26	0.36

—the parametric models are never strictly true. The main reasons for that are: (i) the occurrence of «gross errors» either as clear outliers or as «hidden contamination» (5-10% wrong values exist, almost in any data set); (ii) that the parametric model is only an approximation of the true underlying chance mechanism. Samples of thousand of data in astronomy, which should follow the normal law of error are mildly but definitely leptokurtic. Geary in 1947 was proposing to write in all new text-books that: «Normality is a myth; there never was and never will be a normal distribution».

— the performance of some classically optimal procedures is inefficient under small deviations from a strict parametric model. (e.g. the arithmetic mean being worse than the median in the presence of very mild outliers).

Robust estimation comes to consider all the above. «As the main aim of robust estimation, we can consider building in safeguards against unsuspectibly large amounts of gross errors, putting a bound on the influence of hidden contamination and questionable outliers, isolating clear outliers for separate treatments and still being nearly optimal at the strict parametric model. Robust are all the old methods, but modified a little, with sensible looking at the data» (Hampel, 1973).

1. *Some examples of robust estimators of location*

There are three main methods for constructing estimators of location (see Huber (1968)): (i) maximum likelihood type estimators; (ii) estimators based on linear combinations of order statistics; (iii) estimators derived from rank tests.

Maximum likelihood estimators are the solutions T of an equation of the form

$$\sum_{j=1}^{n} \psi\left[\frac{x_j - T}{s}\right] = 0$$

where ψ is an odd function and s is a measure of spread, either estimated independently or simultaneously from another equation. One estimator of this type is *Huber's M-estimators*, where

$$\psi(z) = \begin{cases} -k & z < -k \\ z & -k \leqslant z \leqslant k \\ k & z > k \end{cases} \qquad 0 < k < \infty$$

Another estimator of this type, due to Hampel, is the *three-parts descending M-estimators*, where

$$\psi(z) = \text{sign}(z) \begin{cases} |z| & 0 \leqslant |z| \leqslant a \\ a & a \leqslant |z| < b \\ \dfrac{c - |z|}{c - b}\, a & b \leqslant |z| < c \\ 0 & |z| \geqslant c \end{cases}$$

a=2.5, b=4.5, c=9.5 (estimate 25A) or a=1.2, b=3.5, c=8.0 (estimate 12A)

Estimators based on linear combination of order statistics are the *g-trimmed mean* (0<g<1/2) which deletes the gn smallest and largest observations of the ordered sample before taking the mean. Thus the effect of outliers may be considerably reduced. It is defined as:

$$T = \frac{1}{n(1-2g)} \left\{ \sum_{[ng]+2}^{[n(1-g)]} x_{(i)} + \right.$$

$$\left. + \left(1 - ng + [ng]\right)\left(x_{([ng]+1)} + x_{([n(1-g)]+1)}\right) \right\}$$

The 0-trimmed mean is the sample mean which is the worst estimate when outliers exist. The 0.50-trimmed mean is the median, the most robust estimator being least affected by gross errors but non- robust when grouping effects are considered. Another estimator of method (ii) is the *g-Winsorized mean* which is calculated by taking the mean of those values which result from replacing the g most extreme values at each end by the next most extreme value, and taking the mean of the modified sample. It is defined as:

$$T = \frac{1}{n}\left([gn]\, x_{[gn]} + \sum_{i=[gn]+1}^{n-[gn]} x_{(i)} + [gn]\, x_{n-[gn]+1} \right)$$

Winsorization treats extreme values as though they were not so extreme and the effect of those values is not eliminated as happens with trimming.

An estimator based on rank tests is the *Hodges-Lehmann* estimator, derived from the nonparametric Wilcoxon (Mann-Whitney) test, which can be defined as the median of all pairs of observations,

$$M = \underset{i \leqslant j}{\text{med}} \ \frac{(x_i + x_j)}{2}$$

If we want to choose between different robust competitors to a classical procedure, we have to make precise the goals we want to achieve. According to Huber (1972) untortunately we can find five or six conflicting goals:

(i) a robust estimator should possess a high absolute efficiency for all suitably smooth shapes F(the underlying distribution).

(ii) a robust estimator should possess a high efficiency relative to the sample mean, and this for all F.

(iii) a robust estimator should possess a high absolute efficiency over a strategically selected finite set F of shapes e.g. normal, logistic, double exponential, Cauchy, and rectangular shapes.

(iv) a robust estimator should possess a small asymptotic variance over some neighborhood of one shape, in particular of the normal.

(v) the distribution of a robust estimator should change little under arbitrary small variations of the underlying distribution F, and this uniformly in the sample size n.

Concerning the goal (i) Takeuchi (1971) proposed an estimate which involved estimating the minimum variance unbiased linear combination of order statistics from a subsample of size k. The expected value of this linear combination under all permutations was then calculated as the estimate. As for (ii), Bickel (1965) proposed the Hodges-Lehmann estimate as the safest one. Concerning (iii), studies were made by Crow and Siddiqui, and Birnbaum and Miké, but while this goal is attractive for small samples, is dangerous as an independent goal for optimization. As for (iv), Huber (1964) justifies it by arguing that we usually have quite a good idea of the approximate shape of the true distribution, so that we can consider the neighborhood of only one shape. He constructed his M-estimators using as a measure of robustness for asymptotically normal estimators, the supremum of the asymptotic variance when F ranges over some suitable set of underlying distributions, in particular over the set of all $F=(1-\varepsilon)\Phi+\varepsilon G$ (this model arises for instance if the observations are assumed to be normal with variance 1, but a fraction ε of them is affected by «gross errors»). Hampel (1968) introduced the goal (v) studying the stability aspects of robustness. He introduced the notions of influence curve[1] and qualitative robustness.[2] and developed the requirements that a robust estimator should possess:

— they should react little to small perturbations —corresponding to qualitative robustness—and they should be safe in the presence of large contamination (or many gross errors)—corresponding to a high breakdown point.[3]

— they should keep a bound on the maximal relative influence of any fixed amount of contamination—corresponding to a low gross-error sensitivity.[4]

— they should react smoothly to rounding and grouping and they should separate extreme observations from the bulk of the data—meaning a low rejection point.[5]

Hampel (1974) constructed a class of estimators to possess all the above properties and at the same time rather high efficiency at the normal distribution, called the three-parts descending M-estimators (estimates 25A and 12A in Table 6).

2. *Concluding remarks*

The question of which estimator to choose, does not have a simple answer. A year-long research seminar was held in 1971 at Princeton, on the robustness of a location parameter. About 70 estimators were studied, their Monte-Carlo variances, influence curves, etc. under about 20 different distributions. Table 6 shows the Monte-Carlo variances of some estimators, mentioned here, under different distributions. The main results were that the mean is the worst estimate, being very sensitive to outliers, while the three parts descending M-estimators are the best, especially for the case of poorly specified

[1] The influence curve is the derivative of an estimator (functional) T on the space of prob. distributions at some distribution F, and measures the change of the estimate caused by an additional observation x.

$$IC_{T,F}(x) = \lim_{\varepsilon \to o} \frac{T\left((1-\varepsilon)F + \varepsilon\delta_x\right) - T(F)}{\varepsilon}$$

[2] Qualitative robustness — small change of the estimate with small change of the model— is described by continuity of the estimator with respect to the Prochorov metric. The Prochorov distance takes care of a small fraction of arbitrary gross errors, of rounding and grouping effects.

[3] The breakdown point tells us the fraction of gross errors needed until the estimator becomes completely unreliable.

[4] The gross error sensitivity—the supremum of the absolute value of the influence curve—measures the worst approximate effect which a fixed amount of contamination can have on the value of the estimator.

[5] The rejection point tells whether an estimator rejects outliers, and at what distance.

TABLE 6

Monte-Carlo variances of $\sqrt{n}\ T_n$ *for selected estimates and distributions, sample size* $n=20$

		N(0,1)	(n−p)N(0,1)+pN(0,9)			18N(0,1)
						+
						2N(0,100)
							(1-ε)N(0,1)+εN/U*	

			p=1	p=3	p=5		ε=0.1	ε=0.25
	mean	1.00	1.40	2.20	3.00	10.90		
	g=0.05	1.02	1.16	1.64	2.27	2.90	1.47	3.84
trimmed	g=0.1	1.06	1.17	1.47	1.93	1.46	1.26	1.81
mean	g=0.15	1.10	1.19	1.44	1.80	1.43	1.26	1.64
	median	1.50	1.52	1.75	2.16	1.80	1.64	1.94
	k=2.0	1.01	1.17	1.66	2.30	1.78	1.30	2.17
Huber	k=1.5	1.04	1.16	1.49	1.96	1.50	1.24	1.74
prop. 2	k=1.0	1.11	1.21	1.44	1.78	1.43	1.26	1.62
Hodges-Lehmann		1.06	1.18	1.50	1.88	1.52	1.26	1.70
Takeuchi		1.05	1.19	1.53	2.02	1.32	1.22	1.60
Hampel 25A		1.05	1.16	1.49	1.94	1.26	1.19	1.59
Hampel 22A		1.20	1.26	1.47	1.78	1.32	1.30	1.56

*N/U denotes the distribution of the quotient of a normal (0,1) variable divided by a uniform (0,1) variable.

and possibly long-tailed situations. In more specified, approximately normal cases use the g-trimmed mean (g=0.1 or 0.15) or Huber M-estimators. Later, Tukey compared rejection rules with robust estimators. He found that when «gross errors» were very clear, hard rejection procedures did as well as robust estimators, but when the distinction between «good» and «bad» observations was more difficult, the rejection procedures were much inferior to robust estimators.

Research is continued to robustify regression, analysis of variance, time-series problems, etc. Some robust methods that good practical statisticians were applying before, are justified now by theory. In connection with computers, Tukey considers robust statistics as the «third-generation statistics» after parametric and non-parametric statistics.

BIBLIOGRAPHY

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

Bickel, P.J. (1965), «On Some Robust Estimates of Location», *Ann. Math. Stat.*, 36,847-858.

Box, G.E.P. (1953), «Non-normality and Tests on Variances», *Biometrica*, 40, 318-355.

Box, G.E.P. (1954a), «Some Theorems on Quadratic Forms Applied on the Study of Analysis of Variance Problems», *Ann. of Math. Stat.*, 25, 290-302.

Geary, R.C. (1947), «Testing for Normality», *Biometrica*, 34, 209-242.

Hampel, F.R. (1968), *Contributions to the Theory of Robust Estimation*, Ph. D. Dissertation, Univ. of California, Berkeley.

Hampel, F.R. (1974), «The Influence Curve and Its Role in Robust Estimation», *J. Amer. Statist. Association*, 69, 383-399.

Huber, P.J. (1964), «Robust Estimation of a Location Parameter», *Ann. Math. Stat.*, 35, 73-101.

Huber, P.J. (1968b), «Robust Estimation», *Mathematical Centre Tracts*, 27, 3-25.

Huber, P.J. (1972), «The 1972 Wald Lecture. Robust Statistics», *Ann. Math. Stat.*, 43, 1041-1067.

Irvine, J., Miles, I., Evans, J. (1979), *Demystifying Social Statistics*, Pluto Press.

Miller, R.G.JR. (1968), «Jackknifing variances», *Ann. Math. Stat.*, 39, 567-582.

Takeuchi, K. (1971), «A Uniformly Asymptotically Efficient Estimator of a Location Parameter», *J. Amer. Stat. Assoc.*, 66, 292-301.

Tukey, J.W. (1962), «The Future of Data Analysis», *Ann. Math. Stat.*, 33, 1-67.