NON-I.I.D. GENERALIZATIONS OF THE MATCHED PAIRS T-TEST

Charles D. Coleman, U.S. Census Bureau, 4700 Silver Hill Rd., Stop 8800
Washington, DC 20233-8800 (ccoleman@census.gov)

Abstract:

This paper generalizes the matched-pairs t-test for difference in mean to nonidentically distributed data. A variety of conditions generate a Central Limit Theorem, which makes this test asymptotically standard normal, while the finite-sample distribution is analytically unobtainable. Initially, independence is assumed to show that this test can be constructed when the differences are not identically distributed. Independence also assures the consistency of the bootstrap and wild bootstrap for estimating significance levels. Dependence is introduced using the Martingale Central Limit Theorem, near epoch dependence and mixing conditions. The test is then applied to biological and population estimates data.

Key Words: Non-i.i.d., biostatistics, nonparametric, bootstrap, means, population estimates.

1 Introduction

This article generalizes the matched-pairs t-test to nonidentically distributed data, using central limit theory for non-i.i.d. random variables. This is a generalization of the nonparametric mean bias test of Coleman (1999) to general error structures. Since no parametric assumptions are made, the finite-sample distribution is unobtainable. The test is initially developed assuming independence. Simple moment conditions guarantee asymptotic standard normality. These conditions also enable the wild bootstrap able to estimate significance levels. A strengthening of a moment condition does the same for the bootstrap. These, when used, obviate the need to specify a test distribution. Dependence can occur through the satisfaction of the Martingale Central Limit Theorem, mixing or near epoch dependence. Mixing conditions and near epoch dependence are both types of asymptotic independence. Mixing conditions, which operate on singly infinite sequences, require asymptotic homoscedasticity to create a test statistic. Near epoch dependence operates on doubly infinite sequences and leads to an asymptotically standard normal test statistic, under appropriate assumptions.

1.1 The Model

The data are initially assumed to consist of n independent nonidentically distributed observations of two variables x_i and y_i, where i indexes the units of observation. The differences are $d_i = x_i - y_i$, with respective standard deviations $\sigma_i < \infty$. The null hypothesis is $\mathbf{E}(x_i) = \mathbf{E}(y_i)$ for all i. Equivalently, this is $\mathrm{H}_0 : \mathbf{E}(d_i) = 0$ for all i. The alternative hypothesis is $\mathrm{H}_\mathrm{A} : \mathbf{E}(x_i) \neq \mathbf{E}(y_i)$ or $\mathbf{E}(d_i) \neq 0$ for at least one i.[1] This is similar to the Sign Test in that the data need not be i.i.d. and the alternative consists in the departure from the null hypothesis for at least one observation instead of the whole sample. The differences can have several interpretations: they may simply be the differences between a set of paired measurements, estimation errors, or differences in estimates of unknown parameters. Section 2 develops the test statistic when the differences are independent. Section 3 shows how to use the bootstrap and wild bootstrap to estimate significance levels in this scenario. Section 4 looks at dependent errors and the conditions necessary to satisfy a Central Limit Theorem. Section 5 does some empirical examples from biology and cross-sectional population estimates. Section 6 concludes this paper.

2 Independence

This Section constructs the matched-pairs t-test for independent, nonidentically distributed data. The first assumption is a moment condition to guarantee the existence of a Central Limit Theorem (White, 1984, p. 112):[2]

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

[1] The two-sided alternative is used in the empirical examples in this paper. However, one-sided alternatives can be used with no change in the arguments, as this paper only looks at the distribution of the test statistic under H_0.

[2] Other assumptions can generate other Central Limit Theorems, such as Lindeberg's.

Assumption 1: $E|d_i|^{2+\delta} < \Delta < \infty$ for all i and some δ and $\Delta > 0$.

Note that Assumption 1 is satisfied whenever the d_i have finite, bounded supports.[3] The next assumption is a technical one, which places a positive lower bound on the average variances of the d_i. It has the effect of preventing the asymptotic collapse of the distribution of the mean to a degenerate distribution.

Assumption 2: $\overline{\sigma}_n^2 > \delta' > 0$ for all n sufficiently large, where $\overline{\sigma}_n^2$ is the average variance.

Together, Assumptions 1 and 2 imply the Liapounov Central Limit Theorem: $z_n = \sqrt{n}\,\overline{d}_n/\overline{\sigma}_n$ converges to a standard normal distribution, where $\overline{d}_n$ is the mean difference and $\overline{\sigma}_n = \sqrt{\overline{\sigma}_n^2}$. Whenever plim $s_n^2 = \overline{\sigma}_n^2$,[4] Slutzky's Theorem (Jurečková and Sen, 1996, pp. 56–57) shows that t can be constructed for sample data by substituting s_n for $\overline{\sigma}_n$ in the definition of z_n: $t = \sqrt{n}\,\overline{d}_n/s_n$.[5,6]

One may conjecture that dividing the d_i by their actual or estimated standard deviations accelerates convergence to normality. This is done in the empirical examples in Subsection 5.2.

3 The Bootstrap and Wild Bootstrap under Independence

The bootstrap and wild bootstrap can estimate significance levels in finite samples. This is especially useful when t is not close to normality. Liu (1988) proves that these procedures are consistent for independent, nonidentically distributed data when the variables satisfy the condition $n^{-1}\sum_{i=1}^{n}(\mu_i - \overline{\mu})^2 \to 0$ where $\mu_i = \mathbf{E}d_i$ and $\overline{\mu} = n^{-1}\sum_{i=1}^{n}\mu_i$. This condition is trivially satisfied by the null hypothesis. Moreover, if we can assume the strengthening of Assumption 1 to Assumption 3 and knowing that $n^{-1}\sum_{i=1}^{n}|\mu_i - \overline{\mu}|^3 \to 0$, then the bootstrap corrects for skewness.

[3] To see this, we can choose some number $N > \max_i |d_i|$. Then, $|d_i|^{2+\delta} < N^{2+\delta}$ for all i and all $\delta > 0$. We can thus choose $\Delta = N^{2+\delta}$ in Assumption 1.

[4] Sufficient conditions include $E|d_i|^{4+\delta} < \Delta' < \infty$ for all i and the same δ in Assumption 1 and some $\Delta' > 0$, and $\lim_{n\to\infty}\overline{\sigma}_n^2 = \overline{\sigma}^2 < \infty$.

[5] Slutzky's Theorem enables the substitution of the sample standard deviation for the population standard deviation in any univariate CLT so as to preserve asymptotic standard normality. Thus, it enables parameters to be tested outside of i.i.d. normal settings. Its use will be implied in the construction of t-tests from other CLTs in subsequent Sections.

[6] Coleman (1999) refers to t as "z," in an allusion to its asymptotic normality.

Assumption 3: $E|d_i|^{3+\delta} < \Delta < \infty$ for all i and some δ and $\Delta > 0$.

The procedure for estimating p^*, the estimated bootstrap significance level, is described first.[7] The wild bootstrap uses a similar procedure, with a different resampling mechanism. First, t and $\overline{d}$ are calculated from the data.[8] Then, a number of replications, B, is chosen. This paper uses $B = 100{,}000$. For each of these replications, indexed by b, the components of the original vector of observations $\mathbf{d} = (d_1, \ldots, d_n)$ are sampled n times with replacement to create the n-dimensional vector $\mathbf{d}_b^* = (d_{1b}^*, \ldots, d_{nb}^*)$. Then, for each b, t_b^* is calculated as $t_b^* = \sqrt{n}\left(\overline{d}_b - \overline{d}\right)/s_b^*$, where $\overline{d}_b$ and s_b^* are the mean and sample standard deviation of $\mathbf{d}_b^*$, respectively.[9] Then, the two-sided p^* equals the proportion of $|t_b^*| > |t|$.[10] (Hall 1988, p. 151) The one-sided p^* likewise equals the proportion of $t_b^* > t$ ($t_b^* < t$), if the alternative hypothesis is that at least one of the $\mu_i > (<)\ 0$.

The wild bootstrap operates not by resampling all of the observations, but by resampling each observation individually in a manner that mimics the underlying heteroscedasticity. For each d_i and b, the wild bootstrap observation d_{ib}^{**} is formed by sampling once from the distribution $w_i = \overline{d} + \left(d_i - \overline{d}\right)u_i$, where the u_i are i.i.d. random variables, such that $\mathbf{E}u_i = 0$ and $\mathbf{E}u_i^2 = \mathbf{E}u_i^3 = 1$.[11] These moment conditions and Assumption 1 are sufficient for the wild bootstrap to be consistent and to correct for skewness. (Liu, 1988) The wild bootstrapped t_b^{**} is constructed from $\mathbf{d}_b^{**}$ and the wild bootstrap significance level p^{**} is constructed from the t_b^{**}, in the same manner as the ordinary bootstrap. In this paper, $u_i \backsim \left(\delta_1 + \nu_{i1}/\sqrt{2}\right)\left(\delta_2 + \nu_{i2}/\sqrt{2}\right) - \delta_1\delta_2$, where ν_{i1} and ν_{i2} are standard normal random variables, $\delta_1 = \sqrt{3/4 + \sqrt{17}/2}$ and $\delta_2 = \sqrt{3/4 - \sqrt{17}/2}$.[12]

[7] This procedure is also known as the "i.i.d. bootstrap."

[8] The subscript on $\overline{d}$ is suppressed in this Section.

[9] If $s_b^* = \mathbf{0}$, then t_b^* is set equal to 999,999 to avoid division by zero. The exact value does not matter, so long as its absolute value is greater than $|t|$. This frequently occurred in very small samples, which could have been handled by permutation tests at the risk of complicating this paper's exposition.

[10] Note that the factor $\sqrt{n}$ could be omitted and the population standard deviation used without altering the results.

[11] Davidson and Flachaire (2000) argue that the third-moment condition is unnecessary.

[12] This is the version of the wild bootstrap used by Mammen (1992) in his simulations. The most popular choice for u_i is $u_i = -\left(\sqrt{5} - 1\right)/2$ with probability $q = \left(\sqrt{5} + 1\right)/\left(2\sqrt{5}\right)$ and $u_i = \left(\sqrt{5} - 1\right)/2$ with probability $1 - q$. Davidson and Flachaire (2000) argue for choosing u_i to equal ± 1, each with probability $1/2$.

Again, 100,000 draws are made.

Some general results are apparent from Section 5 when the bootstrap and wild bootstrap are used. When t is significant or nearly significant at conventional levels, it is usually true that $p^* > p_t > p^{**}$, where p_t is the p-value from the appropriate t distribution. Otherwise, usually, $p^{**} > p^* > p_t$. A possible interpretation is that the wild bootstrap accounts best for the heterogeneity of the underlying distributions. Thus, it tends to produce the strongest evidence for (against) H_0, when H_0 (H_A) is true. Another conclusion is that, for large enough samples, the normal distribution is a suitable test distribution, since p_n, the normal p-value, usually exceeds p^{**}. That is, using the normal distribution is more likely to accept H_0 than the wild bootstrap.

4 Dependent Errors

The data need not be independent for the test to work. Central Limit Theorems still exist for these cases. The simplest one is the Martingale CLT, which requires in addition to Assumption 1: (i) $\mathbf{E}(d_i) = 0$ for all i and (ii) plim $n^{-1}\sum_{i=1}^{n}\left(\sigma_i^2 - d_i^2\right) = 0$. Then, $z_n = \sqrt{n}\,\overline{d}_n/\overline{\sigma}_n$ again converges to a standard normal distribution.[13] Condition (i) is automatically satisfied by H_0. Sequential estimation of time series generally satisfies the Martingale CLT. (Davidson, 1982, pp. 256–257) One may conjecture that it is also satisfied by cross-sectional estimates.

Other forms of dependence with their own CLTs exist. The include mixing conditions and near epoch dependence (NED). The former operates on singly infinite sequences and the latter on doubly infinite sequences. They are both forms of asymptotic independence. Suppose that the errors obey a mixing condition.[14] Then, the Central Limit Theory for these processes requires asymptotic homoscedasticity of the differences or transformed differences.[15] That is, $\lim_{n\to\infty} \overline{\sigma}_n = \overline{\sigma}$, where $\overline{\sigma}$ is a positive constant. t can be constructed per Section 2. CLTs exist for NED variables (e.g., McLeish, 1974) and for the means of NED functions of mixing processes. (Gonçalves and White, 2001) These forms of dependence are generally assumed of time series. A variety of methods exist for testing for dependence in time series. These include runs tests and serial correlation measures.

It is not clear how to use the bootstrap in dependent contexts. Singh (1981, p. 1192) proved the inconsistency of the i.i.d. bootstrap estimator of the variance of an m-dependent sequence. Gonçalves and White (2001) proved a Central Limit Theorem for bootstrapping the means of NED functions of mixing sequences, using versions of the block bootstrap.

5 Empirical Examples

These examples are of two sorts. Subsection 5.1 computes t for an example of paired biological measurements. Subsection 5.2 computes t for population estimates and their decennial census values, with the latter assumed true.[16] The bootstrap and wild bootstrap are applied to the biological data of Subsection 5.1 and to the population estimates of Subsubsection 5.2.2 for the U.S. and the States. In these cases, the normal, t with $n-1$ degrees of freedom, bootstrap and wild bootstrap p-values will be denoted as p_n, p_t, p^* and p^{**}, respectively. All other tests use only the normal p-value. All tests are two-tailed. All computations were done in SAS, with the bootstraps done in SAS/IML.

5.1 Paired Measurements

Dixon and Mood (1946, p. 559) published a Table "Yields of Two Hybrid Lines of Corn." They reported significance of the Sign Test at 5%. Setting the d_i equal to the values of Column A less Column B produces $t = -2.72$ with $p_n = .006$, $p_t = .011$, $p^* = 063$ and $p^{**} = .003$. Except for p^*, which is generally higher than the other p-values, the matched-pairs t-test indicates lower significance levels than the Sign Test in this instance.

5.2 Population Estimates

In the case of population estimates (or any other estimates or forecasts), one sets either the x_i or y_i equal to the true values. Using y_i is more intuitive, as the sign of t is the same as the direction of the estimates' bias. One can also estimate σ_i, normally as a function of the y_i. In the cases below, σ_i is estimated by the regression $\widehat{\sigma}_i = \left|\widehat{d}_i\right| = a + b\,|y_i|$. If this equation is well-specified and b is significant, as is true in both cases below, then one can use $d_i' = d_i/y_i$, the algebraic percentage error, ignoring a

[13] Davidson (1982, pp. 256–257). The crux of this CLT is that the d_i form a martingale difference sequence.

[14] See White (1984, pp. 44–46) for definitions of mixing.

[15] White (1984, p. 124).

[16] This is an assumption because the census contains measurement errors, as it misses some individuals and counts others twice or more. Furthermore, these error rates can vary from census to census.

from the regression, to construct t. Following demographic convention, the mean algebraic percentage error (MALPE) and the sample standard deviation of algebraic percentage errors (SDALPE) are shown below. Thus, $t = \sqrt{n}\,\text{MALPE/SDALPE}$. Significance levels are computed using the normal distribution when the bootstrap is not used. These data come with an important caveat: the errors may not necessarily be independent in all cases.

5.2.1 1970 Washington State County Estimates

These data come from Swanson, Tayman and Barr (2000). In this case, $n = 39$ and $\widehat{\sigma}_i = \left|\widehat{d}_i\right| = 949.09\ (2.84) + 0.019\ (10.84)\ |y_i|$, where the t-ratios are in parentheses. The relevant statistics are MALPE = 2.64588, SDALPE = 5.80553, $t = 17.77$ and $p < .001$.

5.2.2 1990 Census Bureau County Estimates

These estimates are described in Davis (1994). The "true" values are assumed to be the unadjusted 1990 census populations. The error structure was estimated as $\widehat{\sigma}_i = \left|\widehat{d}_i\right| = 475.95\ (7.74) + 0.016\ (74.01)\ |y_i|$ using 3,141 observations. For the U.S. as a whole, this produced the statistics MALPE = 1.77, SDALPE = 4.62, $t = 21.49$ and $p < .00001$ according to all tests. Tables 1–4 summarize these statistics by State, Region, 1990 Size Class and 1980-90 growth rates, respectively. Table 1 contains data for states and reports all p-values when at least one is at least .00001, while a single asterisk next to t in Tables 2–4 denotes significance at 1% using the normal distribution.

Some interpretations of Tables 1–4 are possible. Table 4 shows that, in general, the slower a county's growth, the more likely it was to be overestimated, so that only fast-growing counties were estimated without bias. The significant finding of negative bias in Connecticut in Table 1 may be due to small sample effects and should be disregarded. Interpretations of these results require taking into account the fact the Census Bureau estimates are constrained to an exogenously supplied national control total. The overall bias is due to a high national control, assuming no substantive relative differences between miscounts in the 1980 and 1990 censuses.

Tables 1–4 contain another important lesson for evaluators and users of estimates and forecasts: MALPE, which is commonly reported as a measure of bias, does not contain by itself enough information to determine whether significant bias exists. High (low) values of |MALPE| thus do not necessarily indicate the presence (absence) of bias. For example, Hawaii has a large |MALPE| (5.55), but its larger SDALPE (8.29) and small sample size (5) make its t insignificant. On the other hand, both New York and the Northeast Region have small |MALPE| (0.46 and 0.57, respectively), but are significant at the 10% and 1% levels, respectively.

6 Conclusions and Extensions

This paper has generalized the matched-pairs t-test to non-i.i.d. data with general error structures, effectively expanding upon Coleman's (1999) nonparametric test for mean bias in estimates and forecasts. This test has been applied to examples from biology and population estimates. The population estimates examples were developed under the assumption of independent errors, which may be violated in these cases. These estimates showed clear, consistent trends.

Due to finite sample effects, the true significance values may differ from the nominal ones, contingent on the test distribution. The bootstrap and wild bootstrap can be used to estimate significance levels for independent data. These also avoid the problem of deciding on a null test distribution. The case of dependent, nonidentically distributed data is unclear. The applicability of the bootstrap and wild bootstrap to this kind of data is a subject for future research.

Several Central Limit Theorems for dependent data have also been used to construct the test and show its asymptotic standard normality. The least stringent of them is the Martingale CLT. Current CLTs for mixing processes require asymptotic homoscedasticity. Near epoch dependent processes also possess CLTs. All of these CLTs have been applied to time series, but may be applicable to cross-sectional data as well.

7 References

Coleman, Charles D., (1999) "Nonparametric Tests for Bias in Estimates and Forecasts," in *American Statistical Association: Proceedings of the 1999 Section on Business and Economic Statistics*, 251–256.

Davidson, James, (1982) "Sampling Theory with Dependent Observations, Asymptotic," in Kotz, Samuel, Norman L. Johnson and Cambell B. Read [eds.], *Encyclopedia of Statistical Sciences*, volume 8, 255–257.

State	n	MALPE	SDALPE	t	p_n	p_t	p^*	p^{**}
AK	25	−4.66	9.14	−2.55	.01071	.01749	.02351	.00809
AZ	15	−2.86	6.22	−1.78	.07562	.09735	.09675	.08986
CA	58	1.71	3.82	3.40	.00067	.00123	.00350	.00033
CO	63	1.55	5.28	2.33	.02006	.02335	.02865	.01832
CT	8	−1.80	0.91	−5.61	<.00001	.00080	.01321	.00011
DE	3	3.06	3.11	1.71	.08810	.23022	.33263	.19731
FL	67	2.72	5.72	3.89	.00010	.00024	.00020	.00007
GA	159	1.01	4.95	2.58	.00990	.01081	.01074	.00794
HI	5	5.55	8.29	1.50	.13478	.20911	.37002	.17475
ID	44	0.15	5.39	0.18	.85542	.85627	.85380	.86589
KS	105	1.12	3.86	3.86	.00287	.00357	.00393	.00240
KY	120	1.39	4.10	3.72	.00020	.00030	.00025	.00013
ME	16	0.40	1.85	0.87	.38379	.39751	.39590	.43503
MD	24	−0.54	3.18	−0.83	.40853	.41702	.42606	.46182
MA	14	−0.55	2.77	−0.74	.45778	.47098	.49690	.56513
MS	82	1.53	5.01	2.76	.00571	.00707	.00750	.00472
MT	57	2.01	5.92	2.57	.01030	.01300	.01675	.00741
NV	17	−3.62	5.82	−2.56	.01034	.02080	.04094	.00758
NH	10	1.08	2.78	1.23	.21913	.25029	.26250	.26389
NJ	21	−3.62	2.59	−0.22	.82389	.82614	.82710	.84137
NM	33	1.08	5.65	0.04	.96745	.96770	.96861	.97159
NY	62	−0.13	2.10	−1.73	.08330	.08836	.08922	.08304
OR	36	0.04	3.54	3.33	.00087	.00206	.00208	.00041
RI	5	−1.94	2.24	−1.94	.05285	.12493	.11612	.10164
SC	46	1.49	4.56	2.21	.02681	.03192	.03293	.02688
TX	254	0.25	6.29	0.63	.52879	.52936	.53079	.53746
UT	29	0.50	4.85	0.55	.58151	.58588	.58577	.60695
VA	136	3.23	3.42	3.53	.00041	.00367	.04039	.00081
VT	14	1.23	7.58	1.90	.05779	.05992	.05962	.05369
WA	39	1.07	3.06	2.18	.02917	.03543	.03689	.02875

Table 1: Statistics by State

Region	n	MALPE	SDALPE	t
Northeast	217	0.57	2.94	2.84*
Midwest	1055	2.45	3.35	23.75*
South	1425	1.74	5.15	12.76*
West	444	0.85	5.68	3.14*

Table 2: Statistics by Region

1990 Population	n	MALPE	SDALPE	t
<2,500	119	3.53	9.14	4.21*
2,500–4,999	180	1.33	5.60	3.19*
5,000–9,999	457	1.77	5.37	7.07*
10,000–19,999	707	2.26	4.49	13.39*
20,000–49,999	836	2.10	3.96	15.30*
50,000–99,999	384	1.78	5.37	7.07*
100,000+	458	0.52	2.87	3.88*

Table 3: Statistics by 1990 Populations

Growth Rate	n	MALPE	SDALPE	t
$< -5\%$	877	3.37	4.15	24.02*
$-4.99\% - 0\%$	548	2.17	3.92	12.97*
$0\% - 4.99\%$	543	1.41	3.86	8.53*
$5\% - 9.99\%$	402	1.16	4.26	5.46*
$10\% - 14.99\%$	220	0.90	4.27	3.12*
$15\% - 24.99\%$	255	0.22	5.26	0.66
$25\%+$	296	-0.22	6.53	-0.59

Table 4: Statistics by 1980-1990 Growth Rates

Davidson, Russell and Emmanuel Flachaire, (2000) "The Wild Bootstrap, Tamed at Last," manuscript, Department of Economics, Queen's Univerisity, Kingston, Ontario, Canada, October.

Davis, Sam T., (1994) "Evaluation of Postcensal County Estimates for the 1980s," Population Division Working Paper No. 5, U.S. Census Bureau, Washington, DC.

Dixon, W.J. and A.M. Mood, (1946) "The Statistical Sign Test," *Journal of the American Statistical Association*, **41**, 557–566.

Gonçalves, Sílvia and Halbert White, (2001) "The Bootstrap of the Mean for Dependent Heterogeneous Arrays," report 2001s-19, Centre Interuniversaire de Recherche et Analyze des Organizations, Montreal, Quebec, Canada.

Hall, Peter, (1988) *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, New York.

Jurečková, Jana and Pranab Kumar Sen, (1996) *Robust Statistical Procedures: Asymptotics and Interrelations*, New York: Wiley.

Liu, Regina Y., (1988) "Bootstrap Procedures under Some Non-i.i.d. Models," *Annals of Statistics*, **16**, 1696–1708.

Mammen, Enno, (1992) *When does Bootstrap Work?: Asymptotic Results and Simulations*, New York: Springer-Verlag.

McLeish, D.L., (1974) "Dependent Central Limit Theorems and Invariance Principles," *The Annals of Probability*, **2**, 620–628.

Singh, Kesar, (1981) "On the Asymptotic Accuracy of Efron's Bootstrap," *The Annals of Statistics*, **9**, 1187–1195.

Swanson, David A., Jeff Tayman and Charles F. Barr, (2000) "A Note on the Measurement of Accuracy for Subnational Demographic Estimates," *Demography*, **37**, 193–201.

White, Halbert, (1984) *Asymptotic Theory for Econometricians*, Academic Press, Orlando, Florida.

8 Acknowledgements

I would like to thank Bev Causey and Leroy Bailey for peer review.