

On the Inferential Confidence Intervals for Pairwise Comparisons (*)

Intervalli di Confidenza Inferenziali per Confronti a Coppie

Maurizio Carpita

Università degli Studi di Brescia, Contrada S. Chiara 50, e-mail: carpita@eco.unibs.it

Riassunto: Negli ultimi anni si è diffuso l'impiego di grafici che rappresentano, secondo la definizione di Tryon (2001), *intervalli di confidenza inferenziali*. Le due procedure più note di costruzione di tali intervalli per confronti tra medie sono quelle di Goldstein e Healy (1995) e dello stesso Tryon (2001). Nel presente lavoro si propone una procedura basata sulla distribuzione t usata da Welch (1938) e la si compara tramite simulazione, nel caso di due popolazioni normali con varianze ignote e due campioni indipendenti, con le procedure usuali. Si ottiene che, per numerosità piccole o moderate, solo la procedura derivata da Welch ha ampiezza effettiva prossima a quella nominale.

Keywords: Graphical Confidence Intervals, Two Sample Problem, Statistical difference

1. Introduction

Confidence Intervals (CIs) to compare independent sample estimates are nowadays a frequently advocate graphical tool in social research. For example, graphed CIs are regularly used to compare school effects or more generally institutional performances with multilevel models (Browne and Rasbash, 2004). CIs are often recommended as an alternative of the classical *Null Hypothesis Significance Testing* (NHST), that has severe misconceptions and has been widely criticized (Nickerson, 2000). However, the *inference by eye* interpretation of graphically represented CIs is not without drawbacks. For example, it is common thinking that only when two 95% CIs with about the same width do not overlap then the related NHST is not greater than 5%. But, as long as the two CIs overlapping is lower than half as much their average margin error, then the actual size of NHST is still not greater than 5% (Cumming and Finch, 2005, rule 4). It suggests that if we want to use CIs for pairwise comparisons as an inferential graphical tool at the stated significance level, we must reduce their widths. These adjusted intervals have been named *Inferential CIs* (ICIs) by Tryon (2001). They are obtained by reducing the length of the usual CIs such that nonoverlapping ICIs are algebraically equivalent to the related NHST at the stated significance level. The reduction depends on the unknown population variances that must be estimated, so that for small or moderate sample sizes ICIs have an actual size that is not generally equal to the nominal significance level of the related NHST. Recently, the problem of checking for overlap between two ICIs has been investigated by Payton et al. (2003) and Schenker and Gentleman (2001), but only in particular situations or for the asymptotic case. The aim of this paper is to introduce a procedure based on the t distribution used by Welch (1938) for the Behrens-Fisher problem (Brentari et al., 2001) and to compare it, by extensive simulation, with the two procedures proposed by Goldstein and Healy (1995) and Tryon (2001).

(*) This paper is financially supported by MIUR funds PRIN 2004, prot. 2004130924-004.

2. Two plus one procedures to construct ICIs

Consider the case where the means μ_i and μ_j of two populations i and j with variances δ_i^2 and δ_j^2 are compared. Let M_i and M_j their estimators, with standard errors $\sigma_i = \sqrt{\delta_i^2/n_i}$ and $\sigma_j = \sqrt{\delta_j^2/n_j}$, n_i and n_j the samples sizes and m_i and m_j the sample means.

Assuming normality and δ_i^2 , δ_j^2 known, ICIs at level α for the two means are given by

$$CI_i = [m_i \pm \gamma \sigma_i] \quad \text{and} \quad CI_j = [m_j \pm \gamma \sigma_j] \quad (1)$$

where γ is the proper positive constant which makes the nonoverlapping probability under the null hypothesis $\mu_i = \mu_j$ equal to α , i.e. the nominal significance level of the related NHST. Goldstein and Healy (1995) show that in this case the two CIs at the nominal level α are given by

$$CI_i = [m_i \pm z_{\alpha/2} \varepsilon_{ij} \sigma_i] \quad \text{and} \quad CI_j = [m_j \pm z_{\alpha/2} \varepsilon_{ij} \sigma_j] \quad (2)$$

with $\varepsilon_{ij} = \sqrt{\sigma_i^2 + \sigma_j^2} / (\sigma_i + \sigma_j)$ and $z_{\alpha/2}$ the $(1 - \alpha/2)$ quantile of the standard normal distribution.

If the variances δ_i^2 and δ_j^2 are unknown, σ_i and σ_j in (2) have to be estimated by the usual unbiased standard errors s_i and s_j , and ε_{ij} by $e_{ij} = \sqrt{s_i^2 + s_j^2} / (s_i + s_j)$. So, for large samples, ICIs at the nominal level α are

$$CI_i = [m_i \pm z_{\alpha/2} e_{ij} s_i] \quad \text{and} \quad CI_j = [m_j \pm z_{\alpha/2} e_{ij} s_j] . \quad (3)$$

Tryon (2001) extended (3) for small samples by proposing the following ICIs

$$CI_i = [m_i \pm t_{i,\alpha/2} e_{ij} s_i] \quad \text{and} \quad CI_j = [m_j \pm t_{j,\alpha/2} e_{ij} s_j] \quad (4)$$

where $t_{i,\alpha/2}$ and $t_{j,\alpha/2}$ are the $(1 - \alpha/2)$ quantiles of the Student's t distribution with $(n_i - 1)$ and $(n_j - 1)$ degrees of freedom, respectively.

Using the t distribution obtained by Welch (1938) applying the two moments fitting method, we define other ICIs. Our procedure is based on the following approximation of the nonoverlapping probability:

$$\alpha = Pr[|M_i - M_j| > t(S_i + S_j)] = Pr[|W| > t/E_{ij}] \cong 2[1 - F_{\beta}(t/E_{ij})]$$

where $F_{\beta}(w)$ is the cumulative t distribution, proposed by Welch (1938) for the random variable $W = (M_i - M_j)/S_{ij}$, with degrees of freedom β estimated by

$$b = \frac{(s_i^2 + s_j^2)^2}{s_i^4/(n_i - 1) + s_j^4/(n_j - 1)} = \frac{(1 + k_{ij}^2)^2}{1/(n_i - 1) + k_{ij}^4/(n_j - 1)}$$

where $k_{ij} = s_j/s_i$ is the estimated standard errors ratio $\kappa_{ij} = \sigma_j/\sigma_i$. If $w_{\alpha/2}$ is the $(1 - \alpha/2)$ quantile of $F_b(w)$ and ε_{ij} is estimated by e_{ij} , from $w_{\alpha/2} = t/e_{ij}$ we easily obtain $t = w_{\alpha/2} e_{ij}$. Therefore, for applications we propose the following ICIs at the nominal level α

$$CI_i = [m_i \pm w_{\alpha/2} e_{ij} s_i] \quad \text{and} \quad CI_j = [m_j \pm w_{\alpha/2} e_{ij} s_j] . \quad (5)$$

Note that the quantile $w_{\alpha/2}$ is an estimate because it depends on b .

Using an extensive simulation, in the next paragraph we compare the performance of the procedures based on the three ICIs.

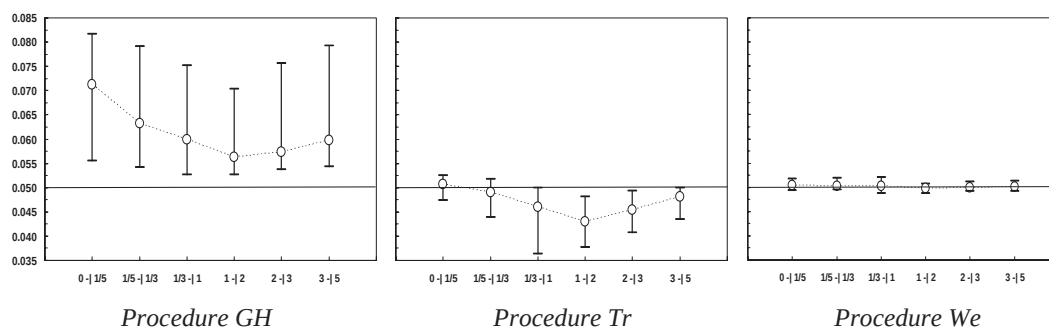
3. Simulation results

In this section the ICIs procedures based on (3), (4) e (5) are respectively named *GH* (Goldstein and Healy), *Tr* (Tryon) and *We* (Welch). To evaluate the three actual levels, we compute the relative frequencies of nonoverlapping of the two ICIs under normality and $\mu_i = \mu_j = 0$, for many configurations of small and moderate sample sizes n_i, n_j with selected values of variances δ_i^2, δ_j^2 properly chosen in order to represent various practical situations. We consider the cases in which both sample sizes are at least 10 (Cumming and Finch, 2005) and the standard errors ratio κ_{ij} is at most 5 (Payton *et al.*, 2003). We set $\delta_i^2 = 1$ and $n_i \leq n_j$, and let the two sample sizes vary from 10 to 50 in steps of 10 (i.e. $5 \cdot 4/2 + 5 = 15$ different pairs) and the variances ratio $\rho_{ij} = \delta_j^2 / \delta_i^2$ from 1 to 25 in steps of 1, and their reciprocals (i.e. $25 + 24 = 49$ different values). Therefore the ratio of standard errors κ_{ij} ranges between 0.089 and 5, and $15 \cdot 49 = 735$ is the number of the resulting configurations. In order to summarize our results, we grouped the configurations into 6 classes with respect to the values of κ_{ij} , as shown in the first column in Table 1. Simulations were performed at the usual nominal level $\alpha = 0.05$; in order to achieve high precisions for all the configurations we have performed 500.000 replications. To check the convergence of simulation results, we used the exact ICIs (2) as benchmark: in all cases the frequencies of nonoverlapping ranged between 0.049 and 0.051. For each class we compute *Min*, *Mean* and *Max* of the nonoverlapping frequencies for the three procedures based on the ICIs: results are reported in Table 1 and represented with *Min-Mean-Max* intervals in Figure 1. For every considered configuration the *procedure GH* always has nonoverlapping frequencies over the nominal level α , with the overall *Mean* equal to 0.062 (from 0.056 for the 4th class to 0.071 for the 1st class). The *Max* 0.082 (larger than 64% of the fixed level α) is observed for the case with $\kappa_{ij} = 0.09$ ($n_i = 10, n_j = 50$ and $\rho_{ij} = 1/23$); generally, the greatest frequencies (larger than 50% of the fixed level α) are observed for those configurations with (i) small sample sizes and one variance greater than the other or (ii) the lowest sample size is associated with the highest variance. With sample size $n_j \geq n_i = 30$ these frequencies are between 0.053 and 0.060, and with $n_i = n_j = 50$ they range from 0.053 to 0.056, for all the 49 different values of the ratio ρ_{ij} . On the contrary, *Procedure Tr* is substantially conservative, with nonoverlapping frequencies appreciably lower than the stated level α and the overall *Mean* equal to 0.048 (from 0.043 for the 4th class to 0.051 for the 1st class).

Table 1: Summary statistics for the nonoverlapping frequencies computed by simulation

Classes of κ_{ij}	Simulations		Min			Mean			Max		
	Num.	%	GH	Tr	We	GH	Tr	We	GH	Tr	We
0 – 1/5	137	18.6	0.056	0.047	0.049	0.071	0.051	0.050	0.082	0.053	0.052
1/5 – 1/3	159	21.6	0.054	0.044	0.050	0.063	0.049	0.050	0.079	0.052	0.052
1/3 – 1	91	12.4	0.053	0.036	0.049	0.060	0.046	0.050	0.075	0.050	0.052
1 – 2	89	12.1	0.053	0.038	0.049	0.056	0.043	0.050	0.070	0.048	0.051
2 – 3	113	15.4	0.054	0.041	0.049	0.057	0.045	0.050	0.076	0.049	0.051
3 – 5	146	19.9	0.054	0.043	0.049	0.060	0.048	0.050	0.079	0.050	0.051
Total	735	100.0	0.053	0.036	0.049	0.062	0.048	0.050	0.082	0.053	0.052

Figure 1: Intervals for the nonoverlapping frequencies computed by simulation



The *Min* 0.036 (lower than 28% of the fixed level α) is observed for $(n_i = n_j = 10; \rho_{ij} = 1)$; the smaller frequencies (lower than 20% of the fixed level α) are observed for the 5 configurations with $n_i = n_j = 10$ and ρ_{ij} ranging between $1/3$ and 3 . When $n_j \geq n_i = 30$ these frequencies are between 0.046 and 0.051. *Procedure We* performs very well also with small samples, with nonoverlapping frequencies reaching the nominal level α for all the configurations. The *Mean* is equal to 0.050 for all the 6 classes, with *Min* 0.049 and *Max* 0.052. Since the *procedure We* has practically the same level α of the procedure with known variances based on (2), the power of the first can be fruitfully compared with the power of the second one for the same 735 configurations. We carried out this simulation for $\mu_j = 0.5, 1, 1.5$ with 200.000 replications for each μ_j ; the (*Max*; *Mean*)% losses of the power for *procedure We* in these three cases are (17; 5,3)%, (12; 3,6)% and (14; 1,9)%, respectively.

References

- Brentari E., Carpita M., Dancelli L. (2001), An improvement in the Welch approximate *t* test, *Metron*, LIX, 1-2: 99-112.
- Browne W., Rasbash J. (2004), Multilevel modelling, in Hardy M. and Bryman A., *Handbook of Data Analysis*, Chapter 20: 459-480.
- Cumming G., Finch S. (2005), Inference by eye - confidence intervals and how to read pictures of data, *American Psychologist*, 60, 2: 170-180.
- Goldstein H., Healy M.J.R. (1995), The graphical presentation of a collections of means, *Journal of the Royal Statistical Society, Series A*, 158, 1: 175-177.
- Nickerson R. (2000), Null hypothesis significance testing: a review of an old and continuing controversy, *Psychological Methods*, 5: 241-391.
- Payton M.E., Greestone M.H., Schenker N. (2003), Overlapping confidence intervals or standard error intervals: what do they mean in terms of statistical significance?, *Journal of Insect Science*, 3, 34: 1-6.
- Schenker N., Gentleman J.F. (2001), On judging the significance of differences by examining the overlap between confidence intervals, *The Amer. Stat.*, 55, 3: 182-186.
- Tryon W.W. (2001), Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical test, *Psychological Methods*, 6, 4: 371-386.
- Welch B.L. (1938), The significance of the difference between two means when the population variances are unequal, *Biometrika*, 29: 350-362.