Statistica & Applicazioni - Vol. VII, n. 1, 2009, pp. 3-14

ON THE USE OF INFERENTIAL CONFIDENCE INTERVALS

Donata Marasini*
Sonia Migliorati**

SUMMARY

Inferential confidence intervals (CI) represent a widely used technique for testing the equality of the means of two Normal distributions by comparing two particular CI's relative to the each mean. They allow simple graphic interpretations and result as being more informative than the traditional testing technique through the confidence levels. Yet they are unable to convey the information relative to the compatibility of the null hypothesis with the observed data usually provided by the p-*value. The present paper identifies a new measure of such compatibility in the context of inferential CI's and extends the technique to the non Normal case. Moreover, in case the null hypothesis is accepted, the problem of estimating the common mean is dealt with by means of the bounds of specific inferential CI's. The proposed procedures are then applied to real data relative to foreign populations originating outside the European Union and sampled from the Italian territory.*

Keywords: *Equality of means, P-value, Overlapping intervals, Chebyshev's inequality.*

1. INTRODUCTION

In the course of sample surveys or experimental trials, it is possible to encounter situations where the estimates of two unknown means result as being extremely different, even if they are expected to be equal or highly similar.

For example, in studies regarding communities of resident foreigners in Italy, the following situation can occur (Blangiardo, 2000): a large sample drawn from the total population of Africans, excluding the Nigerian population, provided a sample mean in relation to a given characteristic of interest, while a much smaller sample of the Nigerian population provided a completely different sample mean. This difference raises the suspicion that the populations of African countries in general do not share similar characteristics with the specific population of Nigerians, as had been previously assumed, at least not in the case of the particular characteristic measured in the two samples. Alternatively, assuming that similarity between the two groups is a given, it can therefore be suspected that the estimate obtained from the smaller sample is not entirely reliable.

* Dipartimento di Statistica - Università degli Studi di Milano-Bicocca - via Bicocca degli Arcimboldi, 8-20126 MILANO (e-mail: donata.marasini@unimib.it).
** Dipartimento di Statistica - Università degli Studi di Milano-Bicocca - via Bicocca degli Arcimboldi, 8-20126 MILANO (e-mail: sonia.migliorati@unimib.it).

When a situation of this type occurs, the first requirement is to verify the equality of the two means, so that, if the hypothesis is accepted, a technique can then be identified which can effectively ''adjust'' such means in order to make them closer, or even substitute them with a single value able to provide given guarantees.

To deal with the above described problem this paper proposes the inferential confidence interval (CI) technique (Goldstein and Healy, 1995; Tryon, 2001), widely used in several types of statistical applications, especially in the fields of health sciences, psychology, and environmental studies (Schenker and Gentleman, 2001; Cumming and Finch, 2005; Bigby and Gadenne, 1996; Hunter and Schmidt, 2004).

The wide use of inferential CI's can be attributed to the fact that this technique provides graphic interpretations which are simple and easy to use, moreover they result as being more informative than traditional tests, as will be shown shortly. The procedure consists of building two intervals, one for each of the means, so that they will not overlap with a given probability if the hypothesis of equality between the two means is true. The hypothesis testing proceeds naturally, accepting the hypothesis if the two intervals based on the sample data have one or more values in common, rejecting the hypothesis if this is not the case.

Therefore, given the significance level of the test α, the confidence level $(1-\gamma)$ of the intervals is automatically assigned and it is equal to both the CI's necessary for the test. This double source of information is not available in the traditional approach, where a single confidence interval for the difference between the means is built at level $(1-\alpha)$ and it determines the acceptance or rejection of the hypothesis. Then, two intervals for the means can be constructed separately, each having an assigned confidence level which is completely independent of α. Nonetheless, the traditional approach has the apparent advantage of being more informative through the *p*-value, which measures the compatibility of the null hypothesis with the observed data.

Actually, it seems very reasonable to establish a connection between the *p*-value and the proportion of the overlap of the two CI's, and several empirical rules, as well as rules referring to particular cases, have been proposed in literature (Cumming and Finch, 2005; Payton *et al.* 2003). As will be shown in the next section, it is possible to identify a general and simple technique to create a measure analogous to the traditional *p*-value, which functions in the context of inferential CI's.

Once the testing procedure has led to the acceptance of the null hypothesis, the use of the inferential CI's allows the identification of an appropriate estimate which represents a sort of balance between the two sample means and has also the advantage of belonging to the overlap between the two inferential CI's, where the unknown mean is supposed to be.

The paper is organised as follows. Section 2 introduces inferential CI's and their characteristics in the Normal framework, proposing a measure of compatibility of the null hypothesis with the observed data derived from the *p*-value. Section 3 deals with the non Normal case, while Section 4 proposes ad hoc adjustments for the two estimates, giving a generalization and a formal justification of the method already present in literature (Marasini and Migliorati, 2006; Blangiardo, 2003). The last section is devoted to real data applications.

2. HYPOTHESIS TESTING VIA INFERENTIAL CI'S

Let X_1 and X_2 be two independent Normal random variables (r.v.'s) with mean θ_i and (known) variance σ_i^2 $(i = 1, 2)$ and suppose the hypothesis $H_0 : \theta_1 = \theta_2$ has to be tested on the basis of two random samples of size n_1 and n_2.

The testing procedure using inferential CI's is implemented in the following steps.

Given that the sample means $\bar{X}_1$ and $\bar{X}_2$ are Normal with variance $\sigma^2_{\bar{x}_i} = \sigma_i^2/n_i$, $(i = 1, 2)$, the confidence intervals have the following form:

$$\bar{x}_i \pm k\sigma_{\bar{x}_i} \tag{1}$$

$(i = 1, 2)$, where k is the solution of the equation:

$$P[|\bar{X}_i - \theta_i| \leq k\sigma_{\bar{x}_i}] = (1 - \gamma) \tag{2}$$

meaning that $k = z_{1-\gamma/2}$ represents the $(1 - \gamma/2)$ standard Normal quantile.

The two intervals (1) do not overlap if one of the two inequalities is verified:

$$\bar{x}_1 + k\sigma_{\bar{x}_1} < \bar{x}_2 - k\sigma_{\bar{x}_2} \qquad\qquad \bar{x}_2 + k\sigma_{\bar{x}_2} < \bar{x}_1 - k\sigma_{\bar{x}_1}$$

which means that:

$$|\bar{x}_1 - \bar{x}_2| > k(\sigma_{\bar{x}_1} + \sigma_{\bar{x}_2}). \tag{3}$$

Assuming that the null hypothesis is true and considering that $(\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2})$ is the variance of the r.v. $(\bar{X}_1 - \bar{X}_2)$, then the probability of rejecting the null hypothesis can be expressed as:

$$P[|\bar{X}_1 - \bar{X}_2| > k(\sigma_{\bar{x}_1} + \sigma_{\bar{x}_2})] = P\left[\frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}}} > k\frac{\sigma_{\bar{x}_1} + \sigma_{\bar{x}_2}}{\sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}}}\right] = \alpha$$

where

$$k\frac{\sigma_{\bar{x}_1} + \sigma_{\bar{x}_2}}{\sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}}} = \frac{k}{e} = z_{1-\alpha/2} \tag{4}$$

and

$$e = \frac{\sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}}}{\sigma_{\bar{x}_1} + \sigma_{\bar{x}_2}} \tag{5}$$

takes values in the interval $[1/\sqrt{2}\,,\, 1)$.

Besides, the following version of (5) will be useful:

$$e = \frac{\sqrt{v + 1}}{\sqrt{v} + 1} \tag{6}$$

where $v = \sigma^2_{\bar{x}_1}/\sigma^2_{\bar{x}_2}$ expresses the relative variability of the two sample means.

From (4) and (2) it is now possible to determine the confidence level $(1-\gamma)$ which leads to a significance level α of the test, i.e.:

$$(1-\gamma) = 2\,\Phi\big(z_{1-\alpha/2}\,e\big) - 1 \tag{7}$$

given that $k = z_{1-\gamma/2} = z_{1-\alpha/2}\,e$.

Hence the intervals (1) become:

$$\bar{x}_i \pm z_{1-\alpha/2}\,e\,\sigma_{\bar{x}_i}. \tag{8}$$

If the intervals (8) do overlap, then the null hypothesis should be accepted while it should be rejected if this is not the case, being equal to α the probability of rejecting the (true) null hypothesis.

Furthermore, the null hypothesis should also be accepted when the two intervals are contiguous, as in the case of:

$$z_{1-\alpha/2}e = \frac{|\bar{x}_1 - \bar{x}_2|}{\sigma_{\bar{x}_1} + \sigma_{\bar{x}_2}} \tag{9}$$

as expressed in (3) taking into account (4) and (5).

The usual hypothesis testing procedure for $H_0 : \theta_1 - \theta_2 = 0$ is based on the $(1-\alpha)$ level CI for the difference of the means:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2}\sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}}$$

leading to acceptance if 0 is included in the interval, to rejection if it is not.

Even though inferential CI's are widely used, the relationship between the level α of the test and the level $(1-\gamma)$ of the two CI's was for a long period erroneously interpreted, testing the hypothesis at level α by means of the intervals of level$(1-\alpha)$. On the contrary, it can be proved (Schenker and Gentleman, 2001; Goldstein and Healy, 1995) that once the confidence level is fixed at $(1-\gamma)$, the resulting test is conservative in the sense that $\alpha \leq \gamma$.

To better clarify this relationship, consider (6) and (7) which demonstrate that the confidence level $(1-\gamma)$ depends on the probability α and on the quantity e through the relative variability v of the two sample means.

Once the level α is fixed, it is possible to verify that $(1-\gamma)$ is an increasing function of e, where:

$$2\,\Phi\left(\frac{z_{1-\alpha/2}}{\sqrt{2}}\right) - 1 \leq (1-\gamma) < 1-\alpha.$$

Furthermore, if $v = 1$ then e reaches its minimum value equal to $1/\sqrt{2}$ and, as a consequence, also $(1-\gamma)$ assumes its minimum value compatibly with α; for example, if $\alpha = 0.05$, then $(1-\gamma) = 0.834$. Given that the minimum value of $(1-\gamma)$ is reached if the two variances are equal, it can be remarked that if the variances are ''close'', a confidence level $(1-\gamma)$ smaller than $(1-\alpha)$ is enough to obtain a given significance level α. If, instead, $v \to 0$ or $v \to \infty$, then $e \to 1$ and $(1-\gamma)$ takes its

maximum value, i.e. $(1-\gamma) \to (1-\alpha)$. Such a situation occurs when one of the two variances is close to 0, thus the most sensible interpretation is that if one of the two variances is ''much'' larger than the other, in order to obtain a level α test (for example 0.05) it is to necessary calculate the CI of level $(1-\gamma) = (1-\alpha)$ (i.e. 0.95).

Notice also that, given a fixed value of e, the probability α is a decreasing function of $(1-\gamma)$, which means that a more reliable test (that is, a decrease of α) corresponds to the need for a more accurate confidence level, expressed as an increase in the level $(1-\gamma)$.

With this established, a measure of compatibility of the null hypothesis with the data (analogous to the p-value) can be identified in the context of inferential CI's. The need to ''translate'' the p-value into the new setting, as previously indicated, is present in literature with proposals mainly based on the proportion of overlap between the intervals (8) and the relationship with the p-value is illustrated through numerical examples.

It is now possible to identify a new type of measure of evidence which is a monotone function of the p-value, thus conveying the same kind of information.

More specifically, the natural measure Δ of overlap between the intervals (8) can be expressed as:

$$\Delta = \begin{cases} \min\{u_1, u_2\} - \max\{l_1, l_2\} & \textit{if} \quad \min\{u_1, u_2\} > \max\{l_1, l_2\} \\ 0 & \text{otherwise} \end{cases}$$

where:

$$\begin{aligned} l_1 &= \bar{x}_1 - z_{1-\alpha/2} e \sigma_{\bar{x}_1}, \quad l_2 = \bar{x}_2 - z_{1-\alpha/2} e \sigma_{\bar{x}_2} \\ u_1 &= \bar{x}_1 + z_{1-\alpha/2} e \sigma_{\bar{x}_1}, \quad u_2 = \bar{x}_2 + z_{1-\alpha/2} e \sigma_{\bar{x}_2} \end{aligned}$$

and:

$$0 \leq \Delta \leq 2 z_{1-\alpha/2} e \min\{\sigma_{\bar{x}_1}, \sigma_{\bar{x}_2}\}.$$

Hence, the following normalized index $\tilde{\Delta}$ can be obtained:

$$0 \leq \tilde{\Delta} = \frac{\Delta}{2 z_{1-\alpha/2} e \min\{\sigma_{\bar{x}_1}, \sigma_{\bar{x}_2}\}} \leq 1.$$

Such an index is equal to 0 if the two intervals (8) do not overlap or are contiguous, it assumes value 1 when the interval related to the sample mean with the lowest variance is contained inside the other. Anyway, $\tilde{\Delta}$ results as being unsatisfactory as $\tilde{\Delta} = 1$ indicates the maximum evidence for the null hypothesis whatever the distance $|\bar{x}_1 - \bar{x}_2|$, which, vice versa, could be large, thus creating uncertainty as to the extent data are compatible with the null hypothesis.

An appropriate index can be identified on the basis of not only the overlap between the two intervals (8) but also on the larger or smaller distances, analogously to what occurs with the p-value. This last quantity is defined as follows:

$$p = P(|\bar{X}_1 - \bar{X}_2| > \ |\bar{x}_1 - \bar{x}_2| \mid H_0) = 2\left[1 - \Phi\left(\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}}}\right)\right]. \qquad (10)$$

Taking into account (6) and (9), the p-value can be associated with the theoretical confidence level $(1 - \gamma^*)$ which is expressed as:

$$(1 - \gamma^*) = 2\Phi\left(\frac{|\bar{x}_1 - \bar{x}_2|}{\sigma_{\bar{x}_1} + \sigma_{\bar{x}_2}}\right) - 1 = 2\Phi\left(z_{1-p/2}\, e\right) - 1, \qquad (11)$$

i.e. $(1 - \gamma^*)$ characterizes the inferential CI's relative to the hypothetical level p test.

Since $p < \alpha$ leads to rejection of the null hypothesis and $p < \alpha$ if and only if $\gamma^* < \gamma$, γ^* can be interpreted as measure of compatibility of the null hypothesis with the data. Once e is fixed, γ^* results as being a decreasing function of the distance $|\bar{x}_1 - \bar{x}_2|$ and the larger the overlap between the two intervals (8), the higher the values assumed by γ^*, analogously to what occurs with p. Such properties make the new index γ^* more appropriate than $\tilde{\Delta}$. For example, $\tilde{\Delta} = 1$ represents a situation of maximum overlap, then γ^*, being a decreasing function of $|\bar{x}_1 - \bar{x}_2|$, indicates that the less the observed difference, the higher its value (i.e. the compatibility of the null hypothesis with the data), and it is easy to check that if $|\bar{x}_1 - \bar{x}_2| = 0$ then $\gamma^* = 1$. Conversely, $\tilde{\Delta} = 0$ means that the two intervals do not overlap, then if the intervals are contiguous the value of γ^* indicates a higher or lower evidence for the null depending on the lower or higher difference between the observed means; in particular, if $|\bar{x}_1 - \bar{x}_2| \to \infty$ then $\gamma^* \to 0$.

The measure γ^* can be normalized conditionally on the acceptance of the null hypothesis. In such case $\gamma^* > \gamma$ and the difference $(\gamma^* - \gamma)$ has a maximum of $(1 - \gamma)$ which is obtained for $\gamma^* = 1$; therefore the normalized index $\tilde{\gamma}$ can be expressed as:

$$\tilde{\gamma} = \frac{\gamma^* - \gamma}{1 - \gamma}. \qquad (12)$$

Obviously $(1 - \gamma^*)$ can be interpreted as a measure of evidence against the null hypothesis: if $p < \alpha$ (rejection of the null hypothesis) and therefore $\gamma^* < \gamma$, then $(1 - \gamma^*) > (1 - \gamma)$, which indicates that confidence levels higher than the one actually utilized (therefore larger intervals than those actually calculated) would have caused the rejection of the null hypothesis. In other words, $(1 - \gamma^*)$ is the highest confidence level that causes rejection, being thus interpretable as a measure of incompatibility of the null hypothesis with the observed data.

If $p = \alpha$, the following can be obtained from (10):

$$\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}}} = z_{1-\alpha/2}.$$

Therefore, taking into account (9), it follows that the two intervals are contiguous, and, by comparing (6) and (11), also that $(1 - \gamma^*) = (1 - \gamma)$. This means that

among all the CI's with a level smaller than or equal to $(1-\gamma)$ only the level $(1-\gamma)$ interval allows the hypothesis to be accepted.

It is of note that, while $\tilde{\Delta}$ is an index based on the length of intervals, $\tilde{\gamma}$ is a probabilistic measure, and while the first of the two indices might be erroneous, the second provides indications which are coherent with those provided by the p-value, as it is an increasing function of it as demonstrated in (11). To better clarify, consider the following Table 1 which reports the level $(1-\gamma)$ confidence intervals and the value of the indices Δ, $\tilde{\Delta}$, γ^*, $\tilde{\gamma}$ in different frameworks. More precisely, $\alpha = 0.05,\ \sigma_1^2 = 2,\ \sigma_2^2 = 2,\ \bar{x}_1 = 1.8,\ \bar{x}_2 = 1.3$ and $n_1 = 20$ remain fixed while the sample size n_2 assumes increasing values, i.e. situations of increasing ''reliability'' for the second sample are taken into account. The p-value and its (conditionally on the acceptance of the null hypothesis) normalized version $\tilde{p} = (p-\alpha)/(1-\alpha)$ are also reported.

TABLE 1. - *CI's and Normalized Measures for Different Sample Sizes*

n_2	$1-\gamma$	CI_1	CI_2	Δ	$\tilde{\Delta}$	γ^*	$\tilde{\gamma}$	p	$\tilde{p}$
20	0.83	(1.36,2.24)	(0.86,1.74)	0.38	0.43	0.43	0.32	0.26	0.23
80	0.86	(1.34,2.26)	(1.07,1.53)	0.19	0.42	0.29	0.17	0.16	0.11
120	0.87	(1.33,2.28)	(1.11,1.49)	0.17	0.44	0.26	0.15	0.14	0.10
500	0.90	(1.27,2.33)	(1.19,1.41)	0.13	0.63	0.19	0.10	0.12	0.08
1655	0.92	(1.24,2.36)	(1.24,1.36)	0.12	1	0.15	0.09	0.12	0.07

It is immediately apparent that, as n_2 increases, the indices $\tilde{\gamma}$ e $\tilde{p}$ indicate a decreasing compatibility of the null hypothesis with the data; conversely $\tilde{\Delta}$ does not provide information coherent with $\tilde{\gamma}$ and $\tilde{p}$. For example, if n_2 increases from 80 to 120 $\tilde{\gamma}$ and $\tilde{p}$ decrease, while $\tilde{\Delta}$ increases. Furthermore, if $n_2 = 1655$ then the value $\tilde{\Delta} = 1$ means that one interval is contained inside the other, even though $\tilde{\gamma} = 0.09$ and $\tilde{p} = 0.07$ are far from indicating there is maximum compatibility.

3. HYPOTHESIS TESTING IN NON NORMAL FRAMEWORKS

If the variances σ_i^2 $(i = 1, 2)$ of the two Normal r.v.'s X_1 and X_2 are unknown but equal, the null hypothesis $H_0 : \theta_1 = \theta_2$ can be tested using Student's distribution with $(n_1 + n_2 - 2)$ degrees of freedom, once the unknown variances have been substituted with their unbiased estimates s_1^2 and s_2^2 (Tryon, 2001). Furthermore, if the unknown variances cannot be assumed as equal, it is necessary to substitute Student's distribution with its Aspin-Welch version, i.e. Student's with degrees of freedom g given by:

$$g = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{s_1^2/\left[n_1(n_1-1)\right] + s_2^2/\left[n_2(n_2-1)\right]}$$

(Cumming and Finch, 2005; Carpita, 2006).

In the absence of information regarding the distribution of X_1 and X_2, but assuming their variances are known, Chebyshev's inequality can be applied, which permits the construction of intervals (1) by means of the following:

$$P[|\bar{X}_i - \theta_i| \leq k_c \sigma_{\bar{x}_i}] \geq 1 - 1/k_c^2 \tag{13}$$

$(k_c > 1;\ i = 1, 2)$ which substitutes (2). The two intervals do not overlap if (3) is verified so that, assuming the null hypothesis is true, Chebyshev's inequality leads to the following:

$$P[|\bar{X}_1 - \bar{X}_2| > k_c(\sigma_{\bar{x}_1} + \sigma_{\bar{x}_2})] \leq \frac{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2}{k_c^2(\sigma_{\bar{x}_1} + \sigma_{\bar{x}_2})^2} = e^2/k_c^2 = \alpha \tag{14}$$

where e is provided by (5) and it can be assumed that $0 < \alpha < 0.5$, as it is usual practice. The quantity $k_c = e/\sqrt{\alpha}$ obtained from (14) results then as being larger than 1 since $\alpha < 0.5 \leq e^2$.

Once intervals sharing the form of intervals (1) are constructed using experimental data, i.e.:

$$\bar{x}_i \pm (e/\sqrt{\alpha})\sigma_{\bar{x}_i} \tag{15}$$

the null hypothesis is accepted if they overlap and it is rejected otherwise, being the type 1 error probability (14) not greater than the given level α.

Taking into account (13) and (14), the confidence level $(1 - \gamma_c)$ of the intervals (15) is not less than $(1 - \alpha/e^2)$; moreover, being $e \geq 1/\sqrt{2}$, it is not less than $(1 - 2\alpha)$ for any fixed level α.

Notice that the case of contiguous intervals can be expressed as:

$$\frac{e}{\sqrt{\alpha}} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sigma_{\bar{x}_1} + \sigma_{\bar{x}_2}} \tag{16}$$

and, as stated in Section 2, it leads to the acceptance of the hypothesis.

In the present context as well, it is possible to identify a measure of compatibility of the hypothesis with the data derived from the p-value and expressed in terms of particular confidence levels, although only upper bounds of the values of the indices under study can be dealt with. In this way, with reference to the p-value, denoted by p_c in the non Normal framework, the following is obtained:

$$p_c = P(|\bar{X}_1 - \bar{X}_2| > |\bar{x}_1 - \bar{x}_2||H_0) \leq \frac{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2}{(\bar{x}_1 - \bar{x}_2)^2}.$$

Therefore, as shown in Section 2, it is possible to compute the confidence level $(1 - \gamma_c^*)$ related to a (hypothetical) level p_c test, and interpret the quantity γ_c^* as a measure of compatibility. For such a quantity the following holds:

$$\gamma_c^* \leq \frac{(\sigma_{\bar{x}_1} + \sigma_{\bar{x}_2})^2}{(\bar{x}_1 - \bar{x}_2)^2}.$$

Clearly the missing information regarding the distribution of the r.v.'s introduces a higher uncertainty in connection with the intervals, hence an equal level of significance can only be obtained with wider intervals. This aspect is immediately apparent in the following Table 2, which refers to the same values $\alpha = 0.05$, $\sigma_1^2 = 2$, $\sigma_2^2 = 2$, $\bar{x}_1 = 1.8$, $\bar{x}_2 = 1.3$ and $n_1 = 20$ of Table 1, and provides the intervals (15) and the upper bounds of the quantities γ_c, γ_c^*, p_c.

TABLE 2. - *CI's and Indices for Different Sample Sizes in the non Normal Case*

n_2	γ_c	CI_1	CI_2	γ_c^*	p_c
20	0.1	(0.8, 2.8)	(0.3, 2.3)	1	0.8
80	0.09	(0.746, 2.854)	(0.773, 1.827)	0.9	0.5
120	0.085	(0.715, 2.885)	(0.857, 1.743)	0.793	0.467
500	0.069	(0.598, 3.002)	(1.06, 1.54)	0.576	0.416
1655	0.061	(0.518, 3.082)	(1.16, 1.441)	0.493	0.404

The values of γ_c^* and p_c in Table 2 must interpreted with caution: if the ''small'' values are clear and reliable evidence against the hypothesis, the same does not hold for the ''higher'' ones, which provide no guarantee about the real values due to their nature of upper bounds. It can also be noted that, as expected, the results obtained through Chebychev's inequality tend to coincide with those obtained under the Normality assumption as the sample size increases.

4. THE ESTIMATION PROBLEM

If the hypothesis $H_0 : \theta_1 = \theta_2$ is true, $\bar{x}_1$ and $\bar{x}_2$ represent two estimates of the same parameter, yet it is proposed in literature to adjust only the estimate based on the smaller of the two samples, assuming it to be less reliable.

Suppose for simplicity that $n_1 < n_2$, so that $\bar{x}_1$ is the least reliable estimate and exclude the case in which the intervals have no common values, which would lead to the rejection of the null hypothesis. Then, given the values of α and e, the maximum adjustment for the estimate is obtained when the intervals are contiguous. It follows that the mean $\bar{x}_1$ can be adjusted and moved closer to $\bar{x}_2$ until it reaches the upper bound of its CI if $\bar{x}_1 < \bar{x}_2$, or the lower bound in the opposite case. In other words, the maximum correction consistent with the acceptance of the null hypothesis is strictly related to the identification of the level $(1 - \gamma^*)$ CI's in the case of Normality or level $(1 - \gamma_c^*)$ CI's in the non Normal case.

Taking into account (9) and (16), such maximum correction gives rise to the new estimate:

$$\hat{x} = \bar{x}_1 + \frac{\bar{x}_2 - \bar{x}_1}{\sigma_{\bar{x}_1} + \sigma_{\bar{x}_2}} \sigma_{\bar{x}_1} = \frac{\bar{x}_1 \sigma_{\bar{x}_2} + \bar{x}_2 \sigma_{\bar{x}_1}}{\sigma_{\bar{x}_1} + \sigma_{\bar{x}_2}} \tag{17}$$

which is a weighted average of the two observed means. Therefore it is certainly in-

cluded among the sample means and (if the null hypothesis is true) the estimator $\hat{X}$ is unbiased. Notice that the weight of the least reliable estimate $\bar{x}_1$ is represented by $\sigma_{\bar{x}_2}$ which refers to the larger (and thus more reliable) sample, as also occurs ''mutatis mutandis'' for $\bar{x}_2$ and $\sigma_{\bar{x}_1}$; in other terms, a sort of equilibrium between $\bar{x}_1$ and $\bar{x}_2$ is ensured.

The proposal (17) is also strengthened by the fact that, given the null hypothesis has been accepted and thus the overlap of the two level $(1-\gamma)$ inferential CI's is non empty, $\hat{x}$ belongs to such overlap, i.e. where the unique mean $\theta\,(=\theta_1=\theta_2)$ is expected to be.

It is noteworthy that, alternatively to what is proposed in literature, once the null hypothesis is accepted, it seems proper to adjust both the estimates $\bar{x}_1$ and $\bar{x}_2$ and not only the least reliable one. Clearly, the larger the sample, the smaller the adjustment on the corresponding sample mean, as can be deduced from (17).

5. AN APPLICATION ON FOREIGN POPULATION IN ITALY

The context of immigrant people of a given nation represents a very interesting applicative framework. On the one hand, it is natural to assume that subjects belon-

TABLE 3. - *Main African Nationalities*

Egypt				Rest of the Continent			
n_1	$\bar{x}_1$	σ_1^2	p_{SW}	n_2	$\bar{x}_2$	σ_2^2	p_{SW}
94	30.96	38.5	0.816	229	30.72	35.2	0.081
Marocco				Rest of the Continent			
n_1	$\bar{x}_1$	σ_1^2	p_{SW}	n_2	$\bar{x}_2$	σ_2^2	p_{SW}
70	30.23	45.6	0.083	253	30.94	33.4	0.238
Senegal				Rest of the Continent			
n_1	$\bar{x}_1$	σ_1^2	p_{SW}	n_2	$\bar{x}_2$	σ_2^2	p_{SW}
32	31.53	29.7	0.507	291	30.71	36.8	0.105
Tunisia				Rest of the Continent			
n_1	$\bar{x}_1$	σ_1^2	p_{SW}	n_2	$\bar{x}_2$	σ_2^2	p_{SW}
19	30.16	28.3	0.154	304	30.83	36.6	0.095
Nigeria				Rest of the Continent			
n_1	$\bar{x}_1$	σ_1^2	p_{SW}	n_2	$\bar{x}_2$	σ_2^2	p_{SW}
10	32.6	31.2	0.093	313	30.73	36.2	0.067
Cameroon				Rest of the Continent			
n_1	$\bar{x}_1$	σ_1^2	p_{SW}	n_2	$\bar{x}_2$	σ_2^2	p_{SW}
6	30.17	33.4	0.05	317	30.8	36.2	0.047

ging to a given macro-area (for example, immigrants from African nations) have similar characteristics. On the other, it is relatively common that specific micro-areas within the macro-area are made up of such a small number of subjects that the relative information drawn from the sample can provide unreliable results. This type of situation calls for testing the hypothesis of the equality of the means, which, if accepted, is then followed by an estimate of the common mean.

This section focuses on the data relative to foreign populations originating outside the European Union, sampled from the Italian territory in 2003 (Blangiardo, 2003). Given the fact that the sample is drawn every year, the data obtained in the preceding years allow the hypothesis the variances of the r.v.'s are known to be assumed.

Of particular interest is the variable ''age'' restricted by the value interval 10-49 years. In Table 3, the main African nationalities present in the sample are compared with the rest of the continent through sample size, mean age, variances, and the p-value (p_{SW}) from the Shapiro and Wilks Normality test.

In the first five cases, the Normality hypothesis for the various pairs of populations can be accepted, making it possible to calculate the inferential CI's as proposed in (8). More specifically, given $\alpha = 0.05$, the following Table 4 reports the level $(1 - \gamma)$ provided by (7), the CI's , the values of $\tilde{\gamma}$ from (12) and the estimate $\hat{x}$ of the common mean given by (17):

TABLE 4. - *Inferential CI's in the Normal Case*

	$1-\gamma$	CI_1	CI_2	$\tilde{\gamma}$	$\hat{x}$
Egypt	0.846	(30.048, 31.872)	(30.161, 31.279)	0.783	30.811
Marocco	0.862	(29.034, 31.426)	(30.402, 31.479)	0.471	30.720
Senegal	0.873	(30.060, 33.000)	(30.167, 31.253)	0.466	30.931
Tunisia	0.887	(28.224, 32.096)	(30.280, 31.381)	0.627	30.682
Nigeria	0.906	(29.644, 35.556)	(30.161, 31.299)	0.310	31.032

Table 4 reveals that, although the hypothesis of equality was accepted for the mean age of all the considered nationalities, the measure of compatibility of the null hypothesis with the observed data $\tilde{\gamma}$ was higher for the Egyptians and the Tunisians, while it was much lower for the Nigerians. For the latter, the sample mean age was higher than all the other populations, and the small sample size influenced the interval (CI_1) making it the widest of all the CI's. It is also of note that this occurs even though the interval (CI_2) referring to all the remaining African populations is contained inside CI_1.

In conclusion, this example illustrates that the use of inferential CI's, together with a measure of compatibility of the null hypothesis with the data and an estimate of the common mean in case such hypothesis is accepted, is a procedure which is useful, informative and simple to implement.

RIASSUNTO

Gli intervalli di confidenza inferenziali rappresentano una tecnica molto diffusa che consente di sottoporre a verifica l'ipotesi di uguaglianza delle medie di due popolazioni Normali ponendo a confronto, per l'appunto, due particolari intervalli di confidenza riferiti alle due medie. Tale tecnica ha il vantaggio di consentire una semplice visualizzazione grafica da un lato e di risultare maggiormente informativa rispetto al tradizionale approccio di verifica di ipotesi dall'altro attraverso i livelli di confidenza dei due intervalli. Tuttavia fa perdere l'informazione usualmente contenuta nel p-value relativa alla compatibilità dell'ipotesi nulla con i dati.

Il presente lavoro propone una nuova misura di tale compatibilità nel contesto degli intervalli di confidenza inferenziali ed estende la tecnica al caso di assenza di Normalità. Inoltre, qualora l'ipotesi nulla venga accettata, propone un metodo di stima della comune media basato sugli estremi di particolari intervalli di confidenza inferenziali. Le procedure proposte vengono infine applicate ad un data set reale relativo alla popolazione straniera presente in Italia e proveniente da Paesi non appartenenti all'Unione Europea.

REFERENCES

Bigby M., Gadenne A. (1996). Understanding and evaluating clinical trials. *Journal of the American Academy of Dermatology*, **34**, 555-590.

Blangiardo G.C. (2000). Sample design and implementation, Appendix: Methodological note on sampling technique. In Eurostat, 3/2000/E/n.5, *Push and pull factors of international migration. Country report – Italy*, (pp. 16-22;107-117). European Communities Printing Office, Bruxelles.

Blangiardo G.C. (2003). L'immigrazione straniera in Lombardia 2002, *Rapporto statistico dell'Osservatorio Regionale per l'integrazione e la multi etnicità*. Fondazione I.S.MU e Regione Lombardia, Milano.

Carpita M. (2006). On the inferential confidence intervals for pairwise comparisons. *Proceedings XLIII Scientific Meeting SIS*, 563-566.

Cumming G., Finch S. (2005). Inference by eye. *American Psychologist*, **60**, 170-180.

Goldstein H., Healy M.J.R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society A*, **158**, 175-177.

Hunter J.E., Schmidt F.L. (2004). *Methods of meta-analysis*. Sage Publications, London.

Marasini D., Migliorati S. (2006). Combining information from several groups in estimating characteristics of immigrant people. *Statistical Methods and Applications*, **15**, 107-127.

Payton M.E., Greenstone 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**, 1-6.

Schenker N., Gentleman J.F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, **55**, 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 tests. *Psychological Methods*, **6**, 371-386.