doi.org/10.26398/IJAS.0030-016

REMEMBERING GINI'S OPENING LECTURE AT THE FIRST SCIENTIFIC MEETING OF THE ITALIAN STATISTICAL SOCIETY ON "THE DANGERS OF STATISTICS"

Fortunato Pesarin[1]

Department of Staistical Sciences, University of Padua, Italy

Abstract. *In his opening lecture at the first scientific meeting of the Italian Statistical Society (SIS), Pisa-Italy 09-10-1939, Professor Corrado Gini presented a penetrating discussion on some of the possible dangers related to the widespread use of statistical tools, especially with regards to the possible ambiguous interpretation of associated conclusions. Since then, due to availability of relatively cheap and powerful computers, efficient software, and diffuse systems for automatic data collection, the possibility of dangers has largely increased. In this presentation, we would like to discuss, among the many, a list of potential dangers including some that could not even be imagined at the time of Gini's lecture.*

Keywords: *Inferential extension, Permutation tests, Pre-test analysis, Selection-bias samples*

1. INTRODUCTION

In his opening lecture at the first scientific meeting of the Italian Statistical Society (SIS), held in Pisa-Italy the 9th of October 1939, Professor Corrado Gini (1884 - 1965) strikingly and accurately discussed some of the possible dangers connected to the widespread use of statistical tools, especially with regards to the possible ambiguous interpretation of associated conclusions. Since then, due to availability of relatively cheap and powerful computers, efficient software, and diffuse systems for automatic data collection, the possibility of dangers has largely increased. Here, we would like to discuss, among the many, a list of potential dangers including some that could not even be imagined at the time of Gini's lecture.

Although not exclusively, to some extent statistical analyses at that time were mainly interested in relatively simple problems, most of which had been satisfactorily and adequately solved by means of substantially heuristic and intuitive methods. Nowadays statisticians are mostly interested in much more complex problems, the solutions of which can hardly be done by heuristic and/or intuitive

[1] Corresponding author: Fortunato Pesarin, email: pesarin@stat.unipd.it

approaches. In fact, the heuristic-intuitive approach may often lead to simplistic or unsatisfactory solutions. To avoid this, before going ahead proceeding with any analysis one must refer to a well suited and well discussed theory and carefully examine the related methodology. Moreover, conditions under which statistical tools are valid must always be accurately checked and explicitly taken into evidence in applications, so that the results of his/her analysis have the proper interpretation. Outside its conditions of validity no statistical tool can confer any clear credibility to results and to associated conclusions. With billions of variables per sample unit [e.g., as with functional (continuous or almost continuous curves), financial, shape and "*omics*" data, etc.] on a limited number of units, no simple method can be reliable for all the related statistical problems.

Moreover, as Gini well emphasized, it is quite rare that a given problem can be satisfactorily analyzed by examining only one aspect. Generally, more than one aspect is of interest and the analysis is then conducted on each of them. From a methodological point of view this *multi-aspect analysis* implies that several different tools are applied to the same data set, and so the consequent partial results are necessarily dependent. A dependence that, on the one hand, in most cases is difficult to work with and/or to model; from the other, it must be taken into consideration while expressing the associated global statistical conclusions.

This presentation is organized as follows: in Section 2 we wish to discuss on how can we extend inferential results from sample data to target population(s) when selection-bias sample data are analyzed; Section 3 presents a criticism on the use of a number of tests on the same data and choosing the *best*, so giving rise to the p-hacking; Section 4 relates on the use of pre-tests of normality and homoscedasticity before Student's t; several other sources of possible malpractices are presented in Section 5 and some concluding remarks are in Section 6; a list of references is reported at the end.

2. EXTENDING INFERENCES

2.1 WHEN NUISANCE PARAMETERS ARE ESTIMATED

We borrow from Gini's opening lecture (1939) the notion that when one extends and interprets inferential results using one (or more) estimated nuisance parameter(s), or even a functional (a function of all parameters, such as the effect of a treatment), "*an element of uncertainty* [...] *arises from the fact that, in such calculation, we substitute the known approximate value* [...] *to its -unknown- precise value*" (translation from Italian to English is our own).

This situation almost always arises when the underlying model depends on

some unknown parameters or functionals where only a few of them are of interest for actual inferences and all others are viewed as nuisance. A common example of this situation is Student's t-test, even when applied within its exact conditions of validity, where the parameter of interest is the mean μ: a random sample $\mathbf{X} = (X_1, \ldots, X_n)$, $n \geq 2$, from a normal distribution $\mathscr{N}(\mu, \sigma^2)$ with unknown variance σ^2. For testing $H_0 : \mu = \mu_0$ versus, for instance, $H_1 : \mu > \mu_0$ the uniformly most powerful similar invariant (UMPSI) test statistic $T = \frac{\bar{X} - \mu_0}{\hat{\sigma}}\sqrt{n}$ is used, where $\bar{X} = \sum_i^n X_i/n$ and $\hat{\sigma}^2 = \sum_i^n (X_i - \bar{X})^2/(n-1)$. The reference null distribution of this statistic is central Student's t with $n-1$ degrees of freedom.

Once we have achieved the inferential conclusion by only working with the summary statistics $\bar{X}$ and $\hat{\sigma}^2$, for instance rejecting H_0, the subsequent problem to solve becomes: *to which population is that inference valid*?

To answer this question, suppose that two experimenters, E_1 and E_2, would like to make inferences about their populations' means. E_1 works with population $P_1 \sim \mathscr{N}(\mu_1, \sigma_1^2)$ and E_2 with $P_2 \sim \mathscr{N}(\mu_2, \sigma_2^2)$. Imagine that both experimenters obtain the same summary statistics from their n-dimensional sample data: $\bar{X}_1 = \bar{X}_2 = \bar{X}$ and $\hat{\sigma}_1^2 = \hat{\sigma}_2^2 = \hat{\sigma}^2$, and so the same value for the test statistics: $T_1 = T_2 = T$. Thus, both would make the same inferential conclusion about the hypothesized μ_0, which is then valid for both populations P_1 and P_2. So, the rationale is: *when σ is unknown and inferences are based on its estimate $\hat{\sigma}$, such inferences can be extended to all normal populations that assign positive probability density to $\hat{\sigma}$, i.e.* $\mathrm{d}P^{(n)}(\hat{\sigma})/\mathrm{d}\hat{\sigma} > 0$.

Thus, unless the population from which random data came from is well and unambiguously established independently and previously of the experiment is carried out and/or the summary statistics $\bar{X}$ and $\hat{\sigma}^2$ are obtained, *generally the inferential extension is not just to one, but to a whole family of populations*. For simple problems, where the parent population is easy to be precisely described before data collection, this is a sort of standard situation and no ambiguity arises to which population that inference is proper. It is not always so, as for instance with some studies where $C > 1$ centres are involved and the analysis implies working by only using summary statistics on C sub-populations (Liu et al., 2015). In most of such situations, since data are typically collected by means of different protocols and selection-bias procedures, the global inference based on the summary statistics $[(\bar{X}_c, \hat{\sigma}_c^2),\ c = 1, \ldots, C]$, each pair being specific of a vaguely defined population distribution, relates to a sort of an hypothetical and undefined mixture of sub-populations.

2.2 WHEN DATA ARE FROM SELECTION-BIAS SAMPLES

2.2.1 SOME GENERAL NOTIONS

It is common knowledge that in almost all clinical trials, due to economic, legal, organizational, procedural, safety, and ethical reasons, the subjects to which two or more treatments are randomly assigned are not a random sample from the target population, that to which the study is addressed to. In fact, *subjects are typically selected from those who are elicited* by several criteria and/or protocols *and who comply with the trial*. Thus, the underlying selected population (as well as the associated distribution of any variable of interest) is unknown and usually extremely difficult, if not impossible, to model properly. As a consequence of the many selection criteria the selected population is different from the target one.

To simplify the discussion, consider a trial with two levels of a treatment: level 1, the old drug; level 2, the new drug. If the two drugs had exactly the same effect, the corresponding null hypothesis would be $H_0 : X_1 \stackrel{d}{=} X_2$, expressing the fact that two underlying (and essentially unknown) response distributions P_1 and P_2 are equal. This hypothesis implies that *the* $n = n_1 + n_2$ *observed data* $\mathbf{X}_j = (X_{ji},\ i = 1,\ldots,n_j),\ j = 1,2$, *are exchangeable* (i.e., permutable) *between two treatments* (or samples, or groups). Moreover and equally important, H_0 true implies that *the pooled data set* $\mathbf{X} = (\mathbf{X}_1,\mathbf{X}_2) = \{X_i = X(i),\ i = 1,\ldots,n;n_1,n_2\}$ *is a set of sufficient statistics for the underlying common distribution P, whatever it is* (continuous, discrete, ordered categorical, nominal categorical, unidimensional, multidimensional, or even infinite dimensional). The notation $\mathbf{X} = \{X(i),\ i = 1,\ldots,n;n_1,n_2\}$ means that first n_1 listed elements belong to the first sample and the rest to the second: $\mathbf{X}_1 = \{X_{1i} = X(i),\ i = 1,\ldots,n_1\}$ and $\mathbf{X}_2 = \{X_{2i} = X(i),\ i = n_1 + 1,\ldots,n_2\}$, respectively.

If instead the new drug would likely give, for instance, greater values than the old, the alternative hypothesis would be $H_1 : X_1 \stackrel{d}{<} X_2$, expressing the fact that the distribution representing responses of new drug stochastically dominates that of the old. And so, under H_1 two cumulative distributions satisfy $P_1(x) \geq P_2(x)$ on all real points x and the inequality is strict in a set of positive probability for both distributions; whereas under H_0 it is $P_1(x) = P_2(x)$ on all real points.

Referring to a typical experiment, we call with Δ_1 and Δ_2 the effects of two drugs. It is worth noting that, in this context, it is generally not appropriate to assume that the difference of two effects $\Delta = \Delta_2 - \Delta_1$ is simply *additive* and *fixed* as in $\Delta \stackrel{d}{=} \delta$, where δ is an unknown constant. In general Δ can be either fixed or random, and in the latter case not necessarily independent on the so-called natural

errors ε, those one has when the related responses are expressed as in $X = \mu + \varepsilon$ (see also point 5 in Section 5). In such situations, while modelling the alternative behavior it is generally too hard to justify an assumption such as $P_1(x) = P_2(x-\Delta)$, in the sense that treatment may also affect the underlying dispersion or even other aspects characterizing both distributions. In this case, the problem is a sort of restricted specification for the Behrens-Fisher problem in which the heteroscedasticity is only under H_1, when two effects are different, and not under H_0.

To the best of our knowledge, the only available exact inferential solutions to the latter problem are essentially nonparametric. They involve using a conditional permutation testing procedure, where the conditioning is taken with respect to the pooled observed data set $\mathbf{X}$, and permutations involve either the observed plain data $\mathbf{X}$ or their rank transformations (Pesarin, 2001). The data set $\mathbf{X}$, being sufficient in H_0 for the unknown common distribution P, is also assumed as the leading term for the *reference conditional sample space*, denoted here with the symbol $\Pi(\mathbf{X})$. It is worth observing that:
(i) $\Pi(\mathbf{X})$ contains the data set $\mathbf{X}$ and all its distinct permutations $\mathbf{X}^*$, i.e.

$$\Pi(\mathbf{X}) = \left\{ \bigcup_{\mathbf{u}^* \in \Pi(\mathbf{u})} [X(u_i^*),\ i = 1,\ldots,n; n_1, n_2] \right\},$$

where $\Pi(\mathbf{u})$ is the set of all permutations $\mathbf{u}^*$ of unit labels $\mathbf{u} = (1,2,\ldots,n)$;
(ii) each point $\mathbf{X}^* \in \Pi(\mathbf{X})$ is also sufficient in H_0 for P since, due to the assumed exchangeability, it is $f_P(\mathbf{X}) = f_P(\mathbf{X}^*)$, f_P being the generalized density corresponding to $P^{(n)}$;
(iii) hence, $\Pi(\mathbf{X})$ is a sufficient space that corresponds to the orbit of points of sample space $\mathscr{X}^n$ containing the same information on P as that contained in $\mathbf{X}$.

2.2.2. MAIN PROPERTIES OF PERMUTATION TESTS

R.A. Fisher (1936), who is considered the author of the permutation testing approach, wrote: "[…] *the statistician does not carry out this very simple and very tedious process* (if carried out by hand), *but his conclusions have no justification beyond the fact that they agree with those which could have arrived at by this elementary method* […]".

The great computational complexity needed to conduct permutation tests seems to have led Fisher to base statistical inference on the likelihood concept. So, Fisher seems considering the role of traditional parametric testing to provide an approximation of the null permutation distribution of a test.

Due to cheap and powerful computers and efficient software, permutation testing methods have increased in number of applications and in solving com-

plex multivariate problems. Prior to formally discussing the inferential extension problem, we wish presenting the main and necessary properties of permutation tests without formal proofs, that can be found in the book by Pesarin and Salmaso (2010-a).

Without loss of generality, assume that large values of test statistics $T : \mathscr{X}^n \to \mathscr{R}^1$ are evidence against H_0.

Property 1. *Sufficiency of* $\Pi(\mathbf{X})$ *for P under* H_0 *implies that the null conditional probability given* $\Pi(\mathbf{X})$ *of every event A, member of a suitable collection* $\mathscr{A}$*, is independent of P; that is,* $\forall P$ *and* $\forall A \in \mathscr{A}$,

$$\Pr\{\mathbf{X}^* \in A; P|\Pi(\mathbf{X})\} = \Pr\{\mathbf{X}^* \in A|\Pi(\mathbf{X})\}.$$

Thus, since the number $M^{(n)} = \sum_{\Pi(\mathbf{X})} \mathbf{1}[\mathbf{X}^* \in \Pi(\mathbf{X})]$ of points in $\Pi(\mathbf{X})$ is finite for finite sample size n, the null conditional probability of any $A \in \mathscr{A}$ is calculated as

$$\Pr\{\mathbf{X}^* \in A|\Pi(\mathbf{X})\} = \frac{\sum_{\mathbf{X}^* \in A} f_P(\mathbf{X}^*)\mathrm{d}\mathbf{X}^*}{\sum_{\mathbf{X}^* \in \Pi(\mathbf{X})} f_P(\mathbf{X}^*)\mathrm{d}\mathbf{X}^*} = \sum_{\Pi(\mathbf{X})} \frac{\mathbf{1}(\mathbf{X}^* \in A)}{M^{(n)}},$$

because for the assumed exchangeability it is $f_P(\mathbf{X}^*)\mathrm{d}\mathbf{X}^* = f_P(\mathbf{X})\mathrm{d}\mathbf{X}$ for every $\mathbf{X}^* \in \Pi(\mathbf{X})$. A consequence of the latter relation is that *all data permutations, i.e. all elements of* $\Pi(\mathbf{X})$, *are conditionally equally likely under* H_0.

Note that in carrying out the calculations for this conditional probability it is not necessary to call upon the *hypothetical repeated sampling principle* (Cox and Hinkley, 1974). That is, it is not necessary to examine the whole sample space $\mathscr{X}^n$, which in turn has often a merely virtual existence. And so it is not necessary to consider all sample points of $\mathscr{X}^n$ that could have been realized. In fact, since only the observed data $\mathbf{X}$ are taken into consideration, the determination of $\Pr\{A|\Pi(\mathbf{X})\}$ requires the *complete enumeration of* $\Pi(\mathbf{X})$ which has an *objective existence* being made by all permutations $\mathbf{X}^*$ of $\mathbf{X}$ and thus it is fully known just after the data are available.

It is worth noting that the recourse to the hypothetical repeated sampling principle is necessarily required by all parametric and semiparametric testing approaches. It is used *once* by the frequentist approach, where the reference is with respect to the whole sample space $\mathscr{X}^n$. It is used *twice* by the traditional Bayesian approach, where the reference is with respect to the product space $\Theta \times \mathscr{X}^n$, Θ being the space of parameters. It is also worth noting that *the permutation testing only requires the existence of an underlying latent likelihood* f_P, provided that $f_P(\mathbf{X}) > 0$ on the observed data. The existence of that likelihood is exclusively required to assure the sufficiency in H_0 of the pooled data set $\mathbf{X}$.

The fact that $\Pr\{A|\Pi(\mathbf{X})\}$ is P-independent in H_0 has several nice conse-

quences. A simple one is: suppose that a first experimenter knows the value of standard deviation $\sigma \in \mathscr{R}_+$ while a second ignore it, using a permutation test both always arrive at the same conditional inference regarding the mean μ. Indeed, suppose the first uses the statistic $T_I^* = (\bar{X}_1^* - \bar{X}_2^*)/\sigma$ and the second $T_{II}^* = \bar{X}_1^*$, they share the same p-value statistic $\forall \mathbf{X} \in \mathscr{X}^n$ and so two procedures are equivalent in terms of inferential conclusions, i.e.

$$\Pr\{(\bar{X}_1^* - \bar{X}_2^*)/\sigma \geq (\bar{X}_1 - \bar{X}_2)/\sigma | \Pi(\mathbf{X})\} = \Pr\{\bar{X}_1^* \geq \bar{X}_1 | \Pi(\mathbf{X})\}.$$

Thus, knowledge of σ, or more generally of any finite set of nuisance entities regarding P, is irrelevant for permutation testing on μ or Δ.

In practice, when sample sizes (n_1, n_2) are not small, the cardinality of $\Pi(\mathbf{X})$ is too large to enumerate all possible permutations. To overcome this issue, the probability $\Pr\{A|\Pi(\mathbf{X})\}$ can be estimated, to any degree of accuracy, by a *conditional Monte Carlo simulation* on $\Pi(\mathbf{X})$ as, for instance, by the four steps algorithm:

• **1.** Randomly take from $\Pi(\mathbf{u})$ one of its equally likely permutations: $\mathbf{u}^* = (u_1^*, \ldots, u_n^*)$.
• **2.** So, $\mathbf{X}^* = \{X(u_i^*),\ i = 1, \ldots, n; n_1, n_2\}$ gives two permuted samples $\mathbf{X}_1^*$ and $\mathbf{X}_2^*$.
• **3.** Independently, repeat B times steps 1 and 2, obtaining the simple random sample $\{\mathbf{X}_b^*, b = 1, \ldots, B\}$ from $\Pi(\mathbf{X})$.
• **4.** Thus, $\hat{P}(A|\Pi(\mathbf{X})) = \sum_{b=1}^{B} \mathbf{1}(\mathbf{X}_b^* \in A)/B$ gives an unbiased and consistent estimate of $\Pr\{A|\Pi(\mathbf{X})\}$.

Property 2. *Assume that the exchangeability condition on data* $\mathbf{X}$ *is satisfied in* H_0*. Then the conditional rejection probability of any randomized test*

$$\phi_R(\mathbf{X}) = \begin{cases} 1 & \text{if} \quad T^o > T_\alpha \\ \gamma & \text{"} \quad T^o = T_\alpha \\ 0 & \text{"} \quad T^o < T_\alpha \end{cases},$$

for which $\mathbb{E}\{\phi_R(\mathbf{X})|\Pi(\mathbf{X})\} = \alpha, \forall \alpha \in (0,1)$, *is* $\mathbf{X}$*-P-invariant for all* $\mathbf{X} \in \mathscr{X}^n$ *and all P, where:* $\mathscr{X}$ *is the sample space for variable* X, $T^o = T(\mathbf{X})$ *is the observed value value of statistic* T *on data* $\mathbf{X}$, $T_\alpha = T_\alpha[\mathbf{X}(0)]$ *is the* α*-size conditional critical value which can be determined by complete enumeration of the permutation sample space* $\Pi(\mathbf{X})$*, and*

$$\gamma = [\alpha - \Pr\{T^o > T_\alpha | \Pi(\mathbf{X})\}] \,/ \Pr\{T^o = T_\alpha | \Pi(\mathbf{X})\}\,.$$

Property 2 states the *uniform similarity property of randomized permutation tests*; and corresponds to the stronger version of the Neyman α-structure of T.

Since in practice the critical value T_α can be determined only if H_0 is known to be true, we use the observed p-value statistic defined as $\lambda^o = \lambda_T(\mathbf{X}) = \Pr\{T(\mathbf{X}^*) \geq T^o(\mathbf{X})|\Pi(\mathbf{X})\}$. The latter is a non-increasing function of T^o and is one-to-one related with the critical value T_α, since $T^o < T_\alpha$ implies $\lambda^o > \alpha$, $T^o > T_\alpha$ implies $\lambda^o < \alpha$, and vice versa. Hence, α works as the critical value for λ^o. Of course, the statistic λ^o coincides with the true p-value of test T if H_0 would be true; then it works as a *p-value-like statistic*.

Thus, the non-randomized version of test is

$$\phi(\mathbf{X}) = \begin{cases} 1 & \text{if} \quad \lambda^o \leq \alpha \\ 0 & " \quad \lambda^o > \alpha \end{cases} .$$

Due to Property 1, under H_0 we have $\mathbb{E}\{\phi(\mathbf{X})|\Pi(\mathbf{X})\} = \Pr\{\lambda_T(\mathbf{X}) \leq \alpha|\Pi(\mathbf{X})\} = \alpha$ for every *attainable* $\alpha \in (0,1)$. In practice, the attainable support of $\lambda_T(\mathbf{X})$ is a subset of the rationals $(\frac{k}{M^{(n)}},\ k = 1,2,\ldots,M^{(n)})$.

Property 3. *Permutation tests for random positive alternatives* $(\Delta \overset{d}{\geq} 0)$ *and based on divergence of symmetric statistics of non-degenerate measurable non-decreasing transformations of the data, i.e.* $T^*(\Delta) = S_1[\mathbf{X}_1^*(\Delta)] - S_2[\mathbf{X}_2^*(\Delta)]$, *where* $S_j(\cdot)$, $j = 1,2$, *are symmetric functions of their entry arguments* $(\cdot)$, *are conditionally unbiased for every attainable* α, *every population distribution* P, *and uniformly for every data set* $\mathbf{X} \in \mathscr{X}^n$. *In particular*

$$\Pr\{\lambda(\mathbf{X}(\Delta)) \leq \alpha|\Pi(\mathbf{X}(\Delta)\} \geq \Pr\{\lambda(\mathbf{X}(0)) \leq \alpha|\Pi(\mathbf{X}(0)\} = \alpha.$$

Thus, the p-value statistic under H_1 *is uniformly stochastically dominated by that under* H_0, i.e. $\lambda(\mathbf{X}(\Delta)) \overset{d}{\leq} \lambda(\mathbf{X}(0)), \forall \mathbf{X} \in \mathscr{X}^n$.

One consequence of Property 3 is that if effects are such that $\Delta' \overset{d}{>} \Delta \overset{d}{>} 0 \overset{d}{>} \Delta''$, then p-value statistics satisfy: $\lambda(\mathbf{X}(\Delta')) \overset{d}{\leq} \lambda(\mathbf{X}(\Delta)) \overset{d}{\leq} \lambda(\mathbf{X}(0)) \overset{d}{\leq} \lambda(\mathbf{X}(\Delta''))$, which shows the *uniform stochastic monotonicity* of p-value statistics with respect to Δ.

Observe that uniform similarity (Property 2) and uniform conditional unbiasedness (Property 3) require data exchangeability (i.e. the randomization of units to treatments) and do not require a random sampling of subjects from a target population. Thus, *they also work with selection-bias sample procedures*. In addition, it is worth mentioning that Property 3 provides for an exact solution of the restricted form of the famous Behrens-Fisher problem where $H_0 : \Delta \overset{d}{=} 0$ and $H_1 : \Delta \overset{d}{>} 0$ (details are in Pesarin, 2001, Chapter 10).

Property 4. *The unconditional* (or population) *power of a permutation test* T *as a function of* Δ, α, T, P, *and* $\mathbf{n} = (n_1, n_2)$ *is defined as*

$$W(\Delta,\alpha,T,P,\mathbf{n}) = \mathbb{E}_{P^n}[\Pr\{\lambda_T(\mathbf{X}(\Delta)) \leq \alpha|\Pi(\mathbf{X})\}] = \int_{\mathscr{X}^n} \Pr\{\lambda_T(\mathbf{X}(\Delta)) \leq \alpha|\Pi(\mathbf{X})\}\, \mathrm{d}P^{(n)}(\mathbf{X}).$$

Of course, $W(\Delta,\alpha,T,P,\mathbf{n}) \geq W(0,\alpha,T,P,\mathbf{n}) = \alpha, \forall\alpha > 0$, *since, by* Property 3**,** *the integrand is* $\geq \alpha$ *for all* $\Delta > 0$, *all* $\mathbf{X} \in \mathscr{X}^n$, *all distributions P, and all sample sizes* $\mathbf{n}$.

Clearly, Property 4 implies unconditional unbiasedness and requires invoking the hypothetical repeated sampling principle since the expectation operator implies examining all points in the sample space $\mathscr{X}^n$.

It is straightforward to see that Property 4 can be extended to composite hypotheses such as $H_0 : \Delta'' \overset{d}{\leq} 0$ versus $H_1 : \Delta \overset{d}{>} 0$, and to see that $\Delta'' \overset{d}{<} 0 \overset{d}{<} \Delta$ implies $W(\Delta'',\alpha,T,P,\mathbf{n}) \leq W(0,\alpha,T,P,\mathbf{n}) = \alpha \leq W(\Delta,\alpha,T,P,\mathbf{n}),\ \forall P$. The latter also provides the *P-uniform monotonicity of unconditional power function* property of T. In doing this it is essential to note that data exchangeability exactly works at point $\Delta \overset{d}{=} 0$, which must not be a point of H_1.

2.2.3 EXTENDING INFERENCES

The uniform similarity and uniform unconditional power (Properties 2 and 4, say)**,** *jointly suffice to weakly extend conditional to unconditional inferences*. To be specific, they provide for the extension of inferential conclusions peculiar to the list of observed units, for example as with: *drug is effective on present sample units* (Lehmann, 2009), to conclusions related to the population P from which units have been drawn, even when there is selection-bias, as with: *drug is effective*. To this end (for a detailed discussion see: Pesarin, 2002; Pesarin and Salmaso, 2010-a), provided that subjects are randomized to treatments and that $\mathrm{d}P^{(n)}(\mathbf{X})/\mathrm{d}\mathbf{X} = f_P(\mathbf{X}) > 0$, it is worth observing:

• for each attainable α and all sample sizes $\mathbf{n} = (n_1, n_2)$, Property 2 implies that the type I error rate of T in H_0 is such that $W(0,\alpha,T,P,\mathbf{n}) = \alpha$, for all samples $\mathbf{X} \in \mathscr{X}^n$ and all distributions P, *independently of how units are selected* (thus including selection-bias samples);

• Property 4 implies that the population power $W(\Delta,\alpha,T,P,\mathbf{n})$ (the unconditional rejection rate) is $\geq \alpha$, for all distributions P and independently of how data are selected from $\mathscr{X}^n$.

Thus, if the conditions needed for permutation testing are satisfied, Properties 2 and 4 are sufficient, though not strictly necessary, for inference extensions. Violation of one or more of these conditions could make the inferential extensions

improper, if not completely wrong. For instance, violation of randomization of units to treatments, as sometimes may occur in observational studies (typical of social and epidemiological surveys), can provide a completely wrong conclusion about the population. Actually, as often occurs with meta-analyses, observing populations with different characteristics that affect the outcome of interest (for instance, observing old people in city A and young people in city B when we are interested in socioeconomic variables) cannot directly provide any correct inferential extension unless the effect of these confounding covariates is removed by a suitable adjustment procedure.

If condition $f_P(\mathbf{X}) > 0$ is not satisfied on some points $\mathbf{X}$ of $\mathscr{X}^n$, since when $f_P(\mathbf{X}) = 0$ data $\mathbf{X}$ are not yet sufficient for P, we can say nothing credible about any of its sampling functionals. In facts, no rational conditional inference is possible because, formally, we cannot know if the related hypothetical conditional rejection rate on those points is at least α. Points with zero density are unobservable with probability one, so that condition implies adopting some specific caution on extensions we are looking for. In general, *the extension* (or extrapolation, or inductive generalization) *of any inference from experimental samples to unobservable populations can only be done using supplementary knowledge and auxiliary assumptions that are not guaranteed by the adopted experimental design*. For instance, extending inferences from experiments on animals to humans requires specific knowledge and/or hypothetical assumptions that are typically external and auxiliary to the observed data $\mathbf{X}$. Such extensions mostly rely on information that lies outside the given experiment and independently of the fact that observed data $\mathbf{X}$ on animals can be numerically compatible also with humans.

For parametric tests, especially when there are nuisance entities to remove, the extension of inferences from conditional to unconditional can generally be done only if data are obtained by well-designed sampling procedures applied to the entire target population distribution P, which must be clearly identified before data collection. When selection-biased data $\mathbf{X}$ are observed and the selection mechanism is not well designed, not well modeled, or its "selection parameters" are not consistently estimated with the data separately from the testing functionals Δ, no parametric approach can be invoked to achieve credible inferential extensions. And so, since those conditions are rarely met in practice, there is no sense to work outside the conditioning and the sufficiency principles of inference. This conditioning strategy implies adopting the permutation testing principle (Pesarin, 2015). Thus, the extension made without considering such objections may become a true malpractice with possible serious consequences.

In any case it is important to emphasize that this kind of weak extension is merely in terms of the presence of a non-null effect, as is typically done by testing analyses. Nothing can be said directly on its precise size at population level both with point and interval estimation. Indeed, the population effect size cannot be deduced by only examining the conditional effect size, the one that is observed on the actual sample, especially if it is selection-biased. In order to achieve *good unconditional estimates*, i.e. to achieve reliable estimates of population effect sizes, either one has to model the selection mechanism and estimate all its coefficients independently from the summary statistics used in the testing process, or one has to refer to a credible set of external and auxiliary information, which often is merely hypothetical or even not at all available. These two conditions are not commonly met in practice. Thus, the extension of conditional effect size to the population level should be done with caution, if not always avoided.

The dangers of this situation often appear in surveys in which sample units are recruited in a classroom, or in a street, or by post, or by telephone, or solicited to voluntarily participating by TV spots, etc., where most of such units do not respond (it is quite frequent that the rate of respondents is around 10%, or even less). This method of recruiting participants is one way that typical selection-biased samples arise. The acquired information cannot be regarded as representative of the target population, because the selection mechanism is not -or cannot even be- taken into consideration, due to known difficulty modeling it. Pre-electoral surveys of last decades conducted on most countries are popular examples of very wrong extensions to target populations. Consequently, results of this kind of surveys should be regarded with caution, if not always suspected to induce misinformation. In this respect it is worth mentioning that Gini (1951) expresses quite a similar point of view by saying "[…] *replacing a complete study with an incomplete one* […] *I took care to point out the dangers to which one is exposed if a study, which is representative of a phenomenon regarding some characteristics, is extended to other characteristics as well*" (translation from Italian to English is our own).

3. ON THE *p*-HACKING

When using available commercial software it is common practice for most statisticians and users (not all, however) to analyze a data set by using a list of different methods, especially test statistics. And in their report to choose the best one; e.g. that provided with the smallest p-value or that presenting *the most convenient result*. This kind of practice, although not always wrong from a pure methodological

view-point, is generally adopted without proper care, and so the associated final conclusions could be misleading since it may induce users to adopt inferences without any real control of inferential errors. For instance, one does not know how far away from α is its true type I error rate. Actually, the practice of referring to the marginal null distribution of the most convenient p-value leads to a very wrong and sometimes a dishonest practice. This, because the many different test statistics are jointly informative, each on a possible different aspect of the effect of a treatment, and being functions of the same data they are necessarily dependent. This dependence is generally much more complex than linear, and that must be taken into consideration when making inferential conclusions. Such inferences imply referring to the multdimensional null distribution of all those test statistics, which is rarely known within a parametric setting. On the contrary, it can be known, often in quite easy to check conditions, within the permutation approach.

Of course, it is not always wrong to examine a data set from a multiplicity of different viewpoints. For instance, in terms of testing of hypotheses, it is typically unknown if there exists one *best* test statistic for H_0 against H_1 and to know how it is computed and how it works. If the likelihood function is known, the Neyman-Pearson lemma provides for the best test (*most powerful*) only for two-point (simple) parametric hypotheses, e.g. as in $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. But when the hypotheses are composite, to construct *good tests* it is well-known that some more stringent conditions are required. So that, in the general situation, looking for the best test is quite often an unsolvable problem. Thus, it can be of interest for users to examine the data set $\mathbf{X}$ using several test statistics, each specific to one aspect to put into evidence. This is especially true outside the regular exponential family of distributions where, if the null hypothesis is true, the whole data set $\mathbf{X}$ is *minimal sufficient for the underlying distribution P*. In such a case, as $\mathbf{X}$ is n-dimensional, there cannot exist any unidimensional statistic $T : \mathscr{R}^n \to \mathscr{R}^1$ furnished with the property of summarizing the whole information contained in $\mathbf{X}$. So, no parametric method can aspire to be uniformly better than other competitors, parametric or nonparametric. At most one has to stay within the notion of admissibility for testing as is generally provided by the *nonparametric combination* (NPC) of a list *of dependent permutation tests*.

In order to reduce the loss of information associated with using only one single overall statistic, it is possible to take account of a list of statistics suitable for concurrent viewpoints, each fitted for summarizing information on a specific partial aspect of interest for the problem, and so to find solutions within the so-called *multi-aspect methodology*. The latter is based on the NPC (Marozzi, 2004,

2007; Pesarin, 2001; Pesarin and Salmaso, 2010-a; Salmaso and Solari, 2005), where the underlying dependence is processed in a nonparametric way without the necessity of estimating the related unknown dependence coefficients. The multi-aspect methodology has several nice features. Indeed, with some complex problems it is often useful to consider a list (at most countable) of different sub-problems each provided with a proper permutation test (for instance, with some continuous responses the analysis may consider: the area under the curve, the area under the curve over a given set of thresholds, a set of Fourier or wavelet coefficients, a set of functional principal components, and so on).

As a matter of fact, with two partial tests (T_1, T_2), not one-to-one related, the summarized information on P via the NPC is not uniformly smaller than that summarized by each partial test separately. Really, since the resulting NPC test is admissible, the result is that no partial test is uniformly better than the combined one. One more feature of multi-aspect testing occurs when, by chance, the underlying unknown distribution admits a *best* solution, T_1 say, based on a unidimensional sufficient statistic. Of course, T_2 cannot add further information to that summarized by T_1. However, their NPC is asymptotically equivalent to T_1, which then becomes the leading test for the whole NPC procedure. As a consequence, by the NPC no available information on P summarized by the list of test statistics is lost asymptotically. So, the multi-aspect idea is suitable in most ordinary decision problems where only one point of view is generally not sufficient for a complete analysis.

A very important application of multi-aspect and the NPC is with problems where the number of observed variables per unit is larger than sample sizes or, even more intriguing, when they can be expanded to the infinite while that of units remain fixed: n much less than the number V of observed variables. In particular, this notion gives rise to the so-called *finite-sample consistency* (full details are in Pesarin and Salmaso, 2010-a,b).

4. ON THE USE OF PRE-TESTS

Some handbooks of software instructions and of statistical methods written for practitioners, for instance, suggest that when analyzing a two-sample univariate problem, one should first check for normality (e.g. by means of the Shapiro-Wilk test), then to check for homoscedasticity, (e.g. by the F-test), and to proceed with the Student's t-test for testing on equality of means if neither pre-test rejects its null sub-hypothesis.

The reference null distribution of the t-test is (central) Student's t with proper

degrees of freedom if both parent distributions of the data are normal with the same means and variances. However, if two distributions are normal with unequal means and same variances the t-test is non-central Student's t distributed. If the two distributions are normal with different variances and unknown ratio, both null and alternative distributions are essentially unknown. In the literature this is known as the Behrens-Fisher problem, for which there are thousands of contributions devoted to deriving reliable approximations for such distributions. Among these, a reasonably good one is attributed to Welch (1938), where for the null it is used a Student's t with random degrees of freedom depending on the ratio of two variance estimates. If the distributions are non-normal, no parametric solution based on the likelihood ratio behavior can be set-up.

The underlying problem is then quite intriguing. For instance, when testing homoscedasticity after failing to reject the null hypothesis of normality, one should refer to the null distribution of the F statistic *conditional on acceptance of normality by the Shapiro-Wilks test*. Let $SW(\mathbf{X})$ denote the Shapiro-Wilks statistic computed on the data $\mathbf{X}$. Then the appropriate probability statement for this conditional test is

$$\Pr\{F(\mathbf{X}_1,\mathbf{X}_2) \le F_{g_1,g_2}(\alpha) | [SW^C(\mathbf{X}_1,\mathbf{X}_2) \le SW^C(\alpha)]\}$$

where, since normality should be tested separately on the two samples, SW^C is a suitable combination of $SW(\mathbf{X}_1)$ and $SW(\mathbf{X}_2)$ and $SW^C(\alpha)$ is the appropriate critical value for SW^C. However, it is worth noting that this conditional statement for the F-test is essentially unknown and possibly very different from the central F-distribution with appropriate degrees of freedom

$$\Pr\{F(\mathbf{X}_1,\mathbf{X}_2) \le F_{g_1,g_2}(\alpha) | [X_j \sim \mathscr{N}(\mu_j,\sigma^2), j=1,2]\}.$$

Somewhat more intriguing is the reference null distribution of the t statistic upon acceptance of both SW^C and F-tests, because

$$\Pr\{t(\mathbf{X}_1,\mathbf{X}_2) \le t_g(\alpha) | [SW^C(\mathbf{X}_1,\mathbf{X}_2) \le SW^C(\alpha)] \bigcap [F(\mathbf{X}_1,\mathbf{X}_2) \le F_{g_1,g_2}(\alpha)]\}$$

is likely impossible to determine and absolutely different from the central Student's t distribution, given by

$$\Pr\{t(\mathbf{X}_1,\mathbf{X}_2) \le t_g(\alpha) | [X_j \sim \mathscr{N}(\mu,\sigma^2), j=1,2]\}.$$

To the best of our knowledge, at the moment no one knows how to obtain such conditional null distributions, since failure to reject normality does not imply two distributions are truly normal, just as failure to reject homoscedasticity does not imply two distributions are actually homoscedastic. As regards to pretesting for normality, we remember having read a suggestion such as: *use the less*

powerful test if you wish to accept normality so as to avoid using ranks while testing for central tendency. In our opinion this is a very misleading and intrinsically dishonest suggestion.

The authoritative conclusion of Lehmann (2009) is much better suited. He essentially suggests to *never use any parametric test when a nonparametric competitor is available*. When an optimal parametric test works in its ideal conditions, there exists a nonparametric permutation competitor whose power behavior is asymptotically equivalent (Hoeffding, 1952). In practice the loss of efficiency by using a nonparametric test is generally negligible and vanishes at a fast rate with increasing sample sizes. On the contrary, if the parametric test works outside its ideal conditions, the efficiency of nonparametric competitors can be even infinitely better. Ludbrook and Duddley (1998) make a conclusion in the same vein.

Regarding the two-sample problem, consider two typical settings: (I) the *experimental model*, where subjects are randomized to treatments; (II) the *observational model*, where subjects simply belong to their respective sub-populations and are observed as they are. In (I) the typical null hypothesis states that there is no difference of effects between two treatments, i.e. $H_0 : \Delta_1 \overset{d}{=} \Delta_2 \equiv X_1 \overset{d}{=} X_2$, whereas in the alternative treatments may also have effects on variances or even on other aspects of the underlying distributions. In (II) a typical null hypothesis states the equality of means, medians, or some quantiles, without any reference to the homoscedasticity which then is not assumed.

For experimental problems (I), permutation tests based on rank transformations or on plain data give rise to exact solutions, i.e. uniformly unbiased and so forth. These tests only require the exchangeability condition is satisfied in H_0 but do not require homoscedasticity in H_1, provided that it is either $X_1 \overset{d}{>} X_2$ or $X_1 \overset{d}{<} X_2$ (see Property 3 in section 2.2). Moreover, when conditions for the t-test are exactly met, permutation tests based on plain data are at least asymptotically equivalent to it (Hoeffding, 1952).

For observational problems (II), instead, no exact parametric solution can be invoked (Scheffé, 1943). There are permutation nonparametric solutions (Pesarin, 2001, Chapter 10) which are almost exact, i.e. robust with respect to the ratio of two variances, and asymptotically coincident with the best parametric test under the condition of normality and knowledge of two variances. This holds even in multivariate settings. Thus, one need not check these conditions using pre-tests because, due to Properties 1 to 4, the permutation solutions are intrinsically robust to their violation.

In our opinion, the malpractice of using pre-tests is mostly induced by the great facility to press buttons with some commercial software. The result of failing to condition appropriately on pre-tests leads to misleading p-values and serves to discredit the field of Statistics.

5. SOME OTHER SOURCES OF MALPRACTICE

5.1 HOTELLING'S T^2

With two-sample multivariate problems testing for $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_2 = \mu_1 + \Delta$, $\Delta \neq 0$, it is generally suggested to use Hotelling's T^2 test. When the data are multivariate normal with equal, unknown covariance matrices, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, the T^2 is qualified to be *optimal within the unbiased, similar, invariant tests* (Cox and Hinkley, 1964). These properties are sometimes assumed without any justification for their validity or their usefulness to the actual problem, and without considering the impact of their violation in real problems. This is especially important for the invariance property.

We first report some simulation results for sample sizes $n_1 = n_2 = 10$ and multivariate normal distribution with $\boldsymbol{\Sigma} = \boldsymbol{I}$, supposed unknown, and increasing number V of variables. We compare Hotelling's T^2 and the simplest of its permutation competitors. The latter test is based on the direct combination of V partial tests. The statistic is $T''^*_D = \sum_{h=1}^{V} [\bar{X}^*_{h1} - \bar{X}^*_{h2}]^2 / \hat{\sigma}^2_h$, where $\bar{X}^*_{hj} = \sum_i X^*_{hji}/n_j$, $\hat{\sigma}^2_h = \sum_{ji}(X_{hji} - \bar{X}_{hj})^2/(n-2)$, for samples $j = 1,2$, are the permutation sample means and sample variance of the vth variable, $v = 1,\ldots,V$. Note that all $\hat{\sigma}^2_h$ are invariant over data permutations. The major differences between the two tests are that T^2 is conditional on the minimal sufficient statistics for the common covariance matrix $\boldsymbol{\Sigma}$ (which must be estimated from the data) and parametrically takes account of linear dependence on variables, whereas T''^*_D is conditional on *maximal* sufficient statistics and nonparametrically takes account of all underlying dependences.

The following simulation results on the power of T^2 and T''^*_D are reported from Brombin and Salmaso (2013).

Simulation results are based on $B = 1000$ random permutations, $MC = 1000$ Monte Carlo runs, values for $\alpha = 0.01$ (in normal character) and $\alpha = \mathbf{0.05}$ (in bold face). These results show that:

(i) as V increases, the power of Hotelling T^2 increases up to a maximum and then decreases to a minimum for $V = n - 2$ (after then it cannot be calculated without introducing restrictions on $\boldsymbol{\Sigma}$);

Tab. 1: Power of T^2 and $T^{\prime\prime/*}$, for $n_1 = n_2 = 10$, $\mu_1 = 0$, $\Delta = 0.40$

V	T^2	$T^{\prime\prime/*}$
4	.079 / **.219**	.081 / **.237**
8	.063 / **.234**	.126 / **.347**
12	.037 / **.186**	.176 / **.436**
15	.027 / **.118**	.231 / **.484**
18	.013 / **.067**	.253 / **.543**
19		.244 / **.544**
22		.340 / **.618**
25		.365 / **.656**

(ii) power of T_D'' increases monotonically with V, up to unity;

(iii) power of T_D'' is not invariant with respect to alternatives lying at Mahalanobis distance from H_0 and so in some circumstances it can even be more powerful than T^2, which is the uniformly most powerful among the unbiased, similar and invariant tests (T_D'' is simply unbiased and consistent);

(iv) T_D'' requires homoscedasticity only in H_0 but not in H_1 and does not require multinormality and linear dependence among variables; it only requires monotonic dependence and so it can be applied in many more circumstances than T^2 (Blair et al., 1994), including multivariate ordered categorical and/or mixed variables;

(v) the fact that when $V \to \infty$ the power of T_D'' tends to one has important applications in problems where the number of observed variables per subject is larger than sample sizes, i.e. when $V >> n$ as in most "*omics*" data.

Moreover, in some literature on statistical process control (Montgomery, 2007), as well as for the analysis of "*omics*" and other high-dimensional data (Thulin, 2014, and references therein), T^2 is usually recommended when several variables are considered. Three quite stringent conditions for its correct use are: (a) treatment effects do not influence variances and correlations; (b) all considered variables must have the same degree of importance (for the technological and economical quality assessment, or for the physiological and clinical impact); (c) all deviations from zero must be considered to be equally important. Of course, if that is appropriate for the problem at hand, if the underlying multivariate distribution is assumed to be normal, and if the number V of variables is *small* compared to sample sizes, there is no stringent reason to use something other than the paramet-

ric Hotelling's T^2. But, if for some variables it is of interest to test for restricted (one-sided) alternative, as for instance $\mu_{v2} > \mu_{v1}$ or the like (as when for some variables the kind of alternative is *larger than the target*), or when all variables do not have the same degree of importance for the analysis (e.g., in evaluating the reliability performance of cars, brakes are much more important than internal lamps), or some treatment effects are possibly random, then the nonparametric permutation via the NPC provides for efficient solutions. Forcing the use of T^2 in those enlarged conditions may become a real malpractice.

The NPC strategy, that works in accordance with Roy's union-intersection method (Roy, 1953; Sen, 2007; Pesarin et al., 2016; also of interest are: Romano, 1990; Wellek, 2010), decomposes a test of hypotheses into a list of K partial tests as $H_0 : \mathbf{X}_1 \overset{d}{=} \mathbf{X}_2 \equiv \bigcap_{k=1}^{K} H_{0k}$ against $H_1 : \mathbf{X}_1 <\overset{d}{\neq}> \mathbf{X}_2 \equiv \bigcup_{k=1}^{K} H_{1k}$, and supposes the existence of a suitable partial permutation test for each sub-hypothesis H_{0k} against H_{1k} (without loss of generality, it is also assumed that large values of each partial test are evidence for the respective sub-alternative). The partial tests are then suitably combined as a function of their statistics, $T_\psi = \psi[(T_k, w_k), k = 1, \ldots, K]$, where $w_k \geq 0$ is the degree of importance assigned to the kth sub-hypothesis. One simple and practical solution for this kind of problem is to use Fisher's combining rule, $T_F = -\sum_k w_k \log \lambda_k$, where $\lambda_k = \Pr\{T_k(\mathbf{X}^*) \geq T_k(\mathbf{X}) | \Pi(\mathbf{X})\}$ is the permutation p-value statistic of partial test T_k (the whole theory and the related methodology of multivariate permutation tests is in Pesarin, 2001; and in Pesarin and Salmaso, 2010). Moreover, if one rejects the global null hypothesis, by using one of the techniques for multiple testing (Bonferroni, Bonferroni-Holm, Simes, etc.) the NPC method easily enables the researcher to find which variable or cluster of variables is mostly responsible for that rejection. For instance, Bonferroni-Holm technique only requires to put the K partial p-values in ascending order: $\lambda_{(1)} \leq \lambda_{(2)} \leq \ldots \leq \lambda_{(K)}$, and so by sequentially declaring significant at size α all partial tests that satisfy the rule: $\lambda_{(k)} \leq \alpha/(K-k+1)$. This similar possibility is not directly available to the T^2, unless the researcher conducts one separate test for each variable, at the cost of an approximate (and often unknown) control of inferential risks. This possibility, however, becomes impossible when, due to its non monotonic power behavior for large K, T^2 fails rejecting H_0 even in the presence of large effects.

5.2 SOME FURTHER POSSIBLE MALPRACTICES

We wish now, without providing details, to list a few among the many further sources of malpractices and/or abuses of statistical tools.

• 1. In many applications, it is common to assume multinormality and homoscedasticity when the number of variables V is large, without proper criticism. For univariate problems there are heuristic suggestions to check normality, but for the multivariate case it is much more difficult to have credible hints of multinormality. Moreover, when covariates are observed, frequently they are assumed (a) to be linearly related with the variables of interest, and (b) the error components are commonly assumed to be independent of covariate values. This is typically assumed without explicit justification, which leads to the use of unknown approximations and possibly bad results.

• 2. The assumption that missing data are missing completely at random, is sometimes invoked. This assumption is made mostly for convenience because, if not, the testing solution may become too difficult or even impossible to attain (Sen, 2007). In the same vein for testing two (or more) survival functions, it is often assumed without proper criticism that censored data are uninformative of the possible effect of a treatment, i.e. the censoring process is independent of treatments. This gives rise to a possible malpractice whenever that uninformativeness is not properly justified.

• 3. It is common to use univariate two-sided tests for the majority of testing applications without a real justification or proper criticism. If there is a treatment effect Δ different from zero on all observed subjects, in general this is either positive or negative, *but not both*. In such problems, it could be of interest to find inductively, via Bonferroni, which of two arms is significantly active, if any. It should be emphasized that two partial tests, for one-sided alternatives $H_{1<} : \Delta < 0$ and $H_{1>} : \Delta > 0$, the respective statistics $T_< = \bar{X}_1 - \bar{X}_2$ and $T_> = \bar{X}_2 - \bar{X}_1$ are negatively related with probability one. This issue appears in both parametric and nonparametric approaches. However, the application of parametric approaches in multidimensional settings may become extremely difficult or even impossible.

• 4. In some applications, for instance when the outcome of interest has a genetic interaction, the effect Δ could be positive on some subjects and negative on others. The related testing problem becomes so complex that no parametric solution is known. However, a satisfactory permutation NPC solution is available (Bertoluzzo et al., 2013). For this kind of experiment, as $0 < \Pr\{\Delta \leq 0\} < 1$, one should test for $H_0 : \Delta \stackrel{p}{=} 0$ against $H_1 : (\Delta \stackrel{d}{<} 0) \bigcup (\Delta \stackrel{d}{>} 0)$, where two subalternatives $H_1^+ : \Delta \stackrel{d}{>} 0$ and $H_1^- : \Delta \stackrel{d}{<} 0$ *can be jointly active*. This problem can be solved by jointly applying two tests, one on negative and one on positive deviations of empirical distribution functions. For instance, one may use two

Kolmogorov-Smirnov tests:

$$\text{and} \quad \begin{array}{l} T^*_{KS-} = \max_{i\le n}[\hat{F}^*_1(X_i) - \hat{F}^*_2(X_i)] \\ T^*_{KS+} = \max_{i\le n}[\hat{F}^*_2(X_i) - \hat{F}^*_1(X_i)], \end{array}$$

followed by their NPC. We emphasize that the two partial tests are dependent and that such dependence is extremely difficult or even impossible to model. That is why the parametric approach is not appropriate for such problems. Of course, to make inference on which partial alternative(s) is (are) significantly active, all that is required is to adjust for multiplicity the partial p-value statistics.

• 5. The assumption of additive effects when observations are obtained through a monotonic increasing function φ of latent values Y is another source of errors. If $Y = \eta + \varepsilon_Y$ is the underlying random variable and Δ the effect, the observed response when $\Delta = 0$ is $X = \varphi(Y)$, and when $\Delta > 0$ it is

$$X(\Delta) = \varphi(Y+\Delta) = \varphi(Y) + \Delta\varphi'(\eta + \varepsilon_Y + h_Y\Delta),$$

φ' being the derivative of φ in a random point and h_Y a convenient point in $(0,1)$. Thus, the effect on X becomes a random quantity dependent on underlying unobservable errors ε_Y even when the underlying effect Δ on Y is fixed: $\Delta_X = \Delta\varphi'(\eta + \varepsilon_Y + h_Y\Delta)$. It is worth noting that parametric solutions require the separate estimate of the effect size Δ_X and of the variance of underlying errors ε_X, and so they cannot be used in such circumstances. Ignoring this notion produces a true malpractice. Permutation methods provide for correct solutions, since they do not require such a separability.

• 6. In testing for difference of heterogeneity with nominal variables in two-sample problems, it is usual to consider the respective *distance from the uniform distribution* and then by building the test statistic as the *difference between these sampling distances*. This, being a comparison between two non-central distributions, cannot assure a proper control of inferential risks (Arboretti et al., 2008). Actually, testing the distance from the uniform distribution on a discrete distribution, considered as the null hypothesis, is relatively easy to attain, usually by means of the chi-squared test. But, as both are typically working under their respective alternatives, the reference distributions of both tests are non-central chi-squared, each with unknown non-centrality parameter. And so, for finite sample sizes the reference distribution of their difference $X^2_1 - X^2_2$, possessing two unknown non-centrality parameters to take into consideration, is unknown.

• 7. It is well-known that there are two-sample problems where data from one group are difficult (or too expensive) to obtain and those from the other group are easy (or quite inexpensive), as with some rare diseases. Formally, n_1 is taken

small and fixed and n_2 is large and possibly divergent. In such cases, for instance under normality, additive effects and homoscedasticity, the t-test has the form:

$$t(\Delta) = \frac{\bar{X}_1+\Delta-\bar{X}_2}{\hat{\sigma}} \cdot \sqrt{\frac{n_1 n_2}{n_1+n_2}},$$

whose non-centrality parameter is $\delta_t = \frac{\Delta}{\hat{\sigma}}\sqrt{\frac{n_1 n_2}{n_1+n_2}}$, the limit of which is $\frac{\Delta}{\sigma}\sqrt{n_1}$ for $n_2 \to \infty$. The latter, being finite, implies that adding data from only one group does not improve the inferential performance since the power remains asymptotically bounded below 1. This malpractice gives rise to the false expectation that adding data always improves the testing information.

• 8. In the presence of several p-values, for the global test it is sometimes suggested to use their sum (Edgington and Onghena, 2007; Chang, 2007). This practice is suggested without analyzing the properties of that sum. Indeed, through some counterexamples it is easy to prove its inconsistency (Pesarin and Salmaso, 2010, page 139). Thus, in the presence of a large amount of information on its falsity there is a positive probability to do not reject the null hypothesis even asymptotically. So, giving rise to a malpractice.

6. CONCLUDING REMARKS

We have put into evidence that while using statistical tools there are several risky situations to which statisticians and users are exposed and to which they have to pay due attention. Not always the use of such risky tools becomes a real malpractice, since they are generally found to be correct within their peculiar conditions of validity. Of course, when they are used outside their respective conditions of validity, they may become proper malpractices. And so giving rise to improper statistical conclusions and consequently providing discredit to statistics and even to professional statisticians. Typically, the extensive use of *optimal tools* outside their conditions of optimality often, in applications, gives rise to unjustified interpretations of related results.

According to Gini's suggestions, statisticians should pay the deserved attention while presenting their reports so as to avoid improper inferential risks and related objective dangers generated by false or unjustified conclusions, together with some associated bad reputation due to the induced discredit to statistics.

ACKNOWLEDGEMENT

We would like to express our sincere thanks to K. Ottoboni, the University of California at Berkeley -US- for the many suggestions while reading the manuscript.

Moreover, we wish expressing our thanks to an anonymous referee for careful reading and precise suggestions which result in improving readability of present work.

REFERENCES

Arboretti Giancrisfaro, R., Bonnini, S. and Pesarin, F. (2008). A permutation approach for testing heterogeneity in two-sample categorical data. In *Statistics and Computing*, 19: 209-216.

Basso D., Pesarin F., Salmaso L. and Solari A. (2009). Permutation tests for stochastic ordering and ANOVA: theory and applications in *R. Lecture Notes in Statistics*, N. 194, Springer, New York.

Bertoluzzo, F., Pesarin, F. and Salmaso, L. (2013). On multi-sided permutation tests. In *Communications in Statistics - Simulation and Computation*, 42(6): 1380-1390.

Blair, R.C., Higgins, J.J., Karninski, W. and Kromery, J.D. (1994). A study of multivariate permutation tests which may replace Hotelling's T^2 test in some circumstances. In *Multivariate Behavioral Research*, 29: 141-163.

Brombin, C. and Salmaso, L. (2013). *Permutation Tests in Shape Analysis.* Springer, New York, USA.

Chang M. (2007). Adaptive design method based on sum of p-values. In *Statistics in Medicine*, 26: 2772-2784.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics.* Chapman and Hall, London.

Edgington, E.S. and Onghena, P. (2007). *Randomization Tests* (4th Ed.). Chapman & Hall/CRC, Boca Raton, USA.

Fisher, R.A. (1936). "The coefficient of racial likeness" and the future of craniometry. In *Journal of the Royal Anthropological Institute of Great Britain and Ireland*, 66: 57-63.

Gini, C. (1939). I pericoli della Statistica. In *Atti della I^a Riunione Scientifica delle Società Italiana di Statistica.* Supplemento Statistico ai Nuovi Problemi. Ferrara, V, Serie II, N. 2-3-4, 1-44.

Gini, C. (1951). Caractères des plus récents dévelopments de la méthodologie de la Statistique. In *Statistica*, 11: 2-11.

Good, P. (2005), *Permutation, Parametric, and Bootstrap Tests of Hypotheses* (3rd Ed.). Springer-Verlag, New York.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. In *Annals of Mathematical Statistics*, 23: 169-192.

Lehmann, E.L. (1986). *Testing Statistical Hypotheses.* 2nd Ed. Wiley & Sons, New York, USA.

Lehmann, E.L. (2009). Parametric versus nonparametrics: two alternative methodologies. In *Journal of Nonparametric Statistics*, 21: 397-405.

Liu, D., Liu, R. and Xie, M. (2015). Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness. In *Journal of the American Statistical Society*, 110: 326-340.

Ludbrook, J. and Duddley, H. (1998). Why permutation tests are superior to t and *F* tests in biomedical research. In *American Statistician*, 52: 127-132.

Marozzi, M. (2004). A bi-aspect nonparametric test for the two-sample location problem. In *Computational Statistics and Data Analysis*, 44: 639-648.

Marozzi, M. (2007). Multivariate tri-aspect non-parametric testing. In *Journal of Nonparametric*

Statistics, 19: 269-282.

Montgomery, D.C. (2007). *Introduction to Statistical Quality Control.* Wiley & Sons, New York, USA.

Pesarin, F. (2001). *Multivariate Permutation Tests: with Application in Biostatistics.* Wiley & Sons, Chichester, UK.

Pesarin, F. (2002). Extending permutation conditional inference to unconditional one. In *Statistical Methods and Applications*, 11, 161-173.

Pesarin, F. (2015). Some elementary theory of permutation tests. In *Communications in Statistics - Theory and Methods*, 44: 4880-4892.

Pesarin, F. and Salmaso, L. (2010-a). P*ermutation Tests for Complex Data. Theory, Applications and Software.* Wiley & Sons, Chichester, UK.

Pesarin, F. and Salmaso, L. (2010-b). Finite-sample consistency of combinationbased permutation tests with application to repeated measures designs. In *Journal of Nonparametric Statistics*, 22: 669-684.

Pesarin, F. and Salmaso, L., (2013). On the weak consistency of permutation tests. In *Communications in Statistics - Simulation and Computation*, 42: 1368-1397.

Pesarin, F,. Salmaso, L., Carrozzo, E. and Arboretti, R. (2016). Union-Intersection permutation solution for two-sample equivalence testing. In *Statistics & Computing*, 26: 693–701.

Romano, J.P. (1990). On the behavior of randomization tests without group variance assumption. In *Journal of the American Statistical Association*, 85: 686-692.

Roy, S.N. (1953). On a heuristic method of test construction and its use in multivariate analysis. In *The Annals of Mathematical Statistics*, 24: 220-238.

Salmaso, L. and Solari, A. (2005). Multiple aspect testing for case-control designs. In *Metrika*, 62: 331-340.

Scheffé, H. (1943). On Solution of the Behrens-Fisher problem based on the *t* distribution. In *Annals of Mathematical Statistics*, 14: 1430-1432.

Sen, P.K. (2007). Union-intersection principle and constrained statistical inference. In *Statistical Planning and Inference*, 137: 3741–3752.

Thulin, M. (2014). A high-dimensional two-sample test for the mean using random subsamples. In *Computational Statistics and Data Analysis*, 74: 26-38.

Welch, B.L. (1938) The significance of the difference between two means when the population variances are unequal. In *Biometrika*, 29: 350-360.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority.* Chapman & Hall/CRC, Boca Raton, USA.