Combining p-Values: an Overview, with a Discussion on Uniformity and on Mixtures of Uniform and Beta(1,2) or Beta(2,1)

Maria de Fátima Brilhante, Dinis Pestana, and Fernando Sequeira

Abstract Quite often the only information available to a researcher who wants to perform meta-analysis syntheses is the reported p-values from different studies. An overview on how to tackle the problem of combining p-values in order to shed some light on the plausibility of a global null hypothesis is discussed here. Some recent developments regarding generalized p-values and random p-values are also addressed for the purpose of combining statistical evidence, as well as the role that mixtures of uniform and Beta(1,2) or Beta(2,1) can have in the field of Meta-Analysis.

1 Introduction

In the investigation of some issue, eventually by several research teams, testing H_0 *vs*. H_A is performed n times. We assume that those experiments were independently conducted, and hence *assuming the null hypothesis* H_0 *to be true* the observed p-values $p_1, ..., p_n$ are observations of independent replicas $P_1, ..., P_n$ of $U \frown Uniform(0,1)$.

Those several tests can point out to conflicting interpretations, or in some cases be "inconclusive", i.e. the observed p-value is non-significant. In fact, the goal of a statistical test is in general to conclude that there exists evidence to reject the null hypothesis, and there is even a say "*to reject the null hypothesis is a strong decision, to maintain the null hypothesis is a weak decision*"; it is well-known that in the absence of significative p-values the research output isn't in general published, a source of concern on publication bias in meta analysis [23]. On the other hand, in most published work, reported p-values are the sole usable information for future meta analyses.

Maria de Fátima Brilhante
Universidade dos Açores, Departamento de Matemática and
Universidade de Lisboa, CEAUL, Portugal
e-mail: fbrilhante@uac.pt

Dinis Pestana
Universidade de Lisboa, CEAUL and DEIO–FCUL,
Instituto de Investigação Científica Bento da Rocha Cabral, Portugal
e-mail: dinis.pestana@fc.ul.pt

Fernando Sequeira
Universidade de Lisboa, CEAUL and DEIO–FCUL
e-mail: fjsequeira@fc.ul.pt

Section 2 is a brief overview on combining p-values, either raw or transformed. In Section 3 we discuss generalized and random p-values, cf. [16, 18], since it is in the nature of scientific research to try to falsify the null hypothesis, and therefore H_A true for some trials should be expected. Obviously, if H_A is true, the P_k are no longer uniform.

In Section 4 we discuss computational inflation of the sample of p-values, using some algebra that, starting from the available p-values, generates new uniform random variables independent from the initial sample. Naïvely we had expected that sample augmentation would increase power, i.e. more proneness to reject the null hypothesis if the alternative is true. In fact, the reverse happens: there is a noticeable loss of power, that we exhibit in Section 4, and partially explain in Section 5.

Even when H_0 is true, the reported p-values can be non-uniform if good statistical practice is violated. For instance, if the result is different from what the experimenter expects, he may decide to repeat the experiment, and to report the best of the two results — this is, for instance, a possible explanation for Mendel's "too good" results disputed by Fisher, cf. [24]. The best of two results will be either the minimum or the maximum, whose sampling distribution under H_0 will be either $B(1,2)$ or $Beta(2,1)$. Therefore when meta-analysing p-values, it can be expected that the appropriate model is $X_m,\ m \in [-2,2]$,

$$X_m = \begin{cases} U & X \\ 1-\frac{|m|}{2} & \frac{|m|}{2} \end{cases} \tag{1}$$

with $U \frown Uniform(0,1)$ and $X \frown Beta(1,2)$ if $m \in [-2,0)$, or $X \frown Beta(2,1)$, if $m \in (0,2]$. Section 5 discusses in some depth this family, and some surprising results when its members are used for computationally augment the sample of p-values.

In the concluding section we comment on questions of sample size, and on the bias that cumulative meta analysis can cause.

2 Combining p-Values

Meta analysis of p-values has been done much earlier than meta analysis has been recognised as an important field of statistics [12, 13, 14]; as under H_0 the P_k are independent and identically distributed (iid) uniform random variables, Tippett [26] used the minimum $P_{1:n} \frown Beta(1,n)$ and Fisher [11] used $T = -2\sum_{k=1}^{n} \ln(P_k) \frown \chi^2_{2n}$ to test the combined null hypothesis $H_0^*:\ \forall k \in \{1,...n\}\ H_{0,k}$ is true *vs*. the composite alternative $H_A^*:\ \exists k \in \{1,...n\}$ such that H_A^* is true.

Since the pioneering results of Tippett and Fisher, several combined tests have been discussed in the literature, cf. [22, 23], none of them being the best choice in all situations. Fisher's test is, however, the most advisable choice in many situations [19, 20], although this choice is by no means consensual. For instance in the context of social sciences research, Mosteller and Bush [21] clearly prefer to use Stouffer's method [25]: denoting by Φ the standard normal distribution function (df), use $T = \frac{\Phi^{-1}(P_k)}{\sqrt{n}} \underset{|_{H_0^*}}{\frown} Normal(0,1)$.

A rational combined procedure should of course be monotone, in the sense that if one set of p-values $p = (p_1,...,p_n)$ leads to rejection of the overall null hypothesis H_0^*, any set of componentwise smaller p-values $p\dagger = (p_1^\dagger,...,p_n^\dagger),\ p_k^\dagger \le p_k,\ k = 1,...,n$, must also reject H_0^*; and, in fact, any monotone combined test procedure is admissible, i.e. provides a most powerful test against some alternative hypothesis for combining some collection of tests, and is therefore optimal for some combined testing situation whose goal is to harmonize eventually conflcting evidence, or to pool inconclusive evidence.

Instead of Tippett's $P_{1:n}$, any other order statistics $X_{k:n}$ can be used, [29]; under validity of the null hypothesis, $X_{k:n} \frown Beta(k, n+1-k)$. Another statistic that can be used is the the geometric mean $\mathscr{G}_n$ (cf. [22]), whose probability density function (pdf) under validity of the overall null hypothesis has the simple expression $f_{\mathscr{G}_n}(x) = \frac{n(-nx\ln(x))^{n-1}}{\Gamma(n)}\mathbb{I}_{(0,1)}(x)$. Therefore the df is

$$F_{\mathscr{G}_n}(x) = \frac{\Gamma^*(n, -n\ln(x))}{\Gamma(n)}\mathbb{I}_{[0,1)}(x) + \mathbb{I}_{[1,\infty)}(x)$$

where $\Gamma^*(n,z)$ is the incomplete Gamma function $\Gamma^*(n,z) = \int_z^\infty x^{n-1}\mathrm{e}^{-x}\mathrm{d}x$.

As

$$\mathscr{G}_n = \left(\prod_{k=1}^n U_k\right)^{1/n} = \exp\left[\tfrac{1}{n}\ln\left(\prod_{k=1}^n U_k\right)\right] = \exp\left(-\tfrac{X}{2n}\right),$$

with $X \frown \chi^2_{2n}$, the $1-\alpha$ quantile for $\mathscr{G}_n$ is easily expressed in terms of the α quantile of χ^2_{2n}:

$$g_{n,1-\alpha} = \mathrm{e}^{-\frac{\chi^2_{2n,\alpha}}{2n}}.$$

The geometric mean is clearly preferable to the arithmetic mean since, aside from having a very cumbersome pdf $f_{\overline{P}_n}(x) = \frac{n}{\Gamma(n)}\left[\sum_{j=0}^n (-1)^j\binom{n}{j}(nx-j)^{n-j}\mathbb{I}_{[\frac{k}{n},\frac{k+1}{n})}(x)\right]\mathbb{I}_{[0,1)}(x)$, the overall test based on the arithmetic mean isn't consistent, in the sense that it can fail to reject the overall test null hypothesis although the result of one of the partial tests is extremely significant.

Fisher [11] and Stouffer [25] are the simplest test statistics using suitable transformations of the P_k; other transformations of random variables can be used, a popular choice being the logistic transformation $\ln\left(\frac{P_k}{1-P_k}\right) \frown Logistic(0,1)$: as $T = -\sum_{k=1}^n \ln\left(\frac{P_k}{1-P_k}\right)/\sqrt{n\,\frac{\pi^2(5n+2)}{3(5n+4)}} \approx t_{5n+4}$, reject H_0^* at the significance level α if $T(\text{obs.}) > t_{5n+4,1-\alpha}$.

From this brief overview, Tippett's decision rule "reject H_0^* at significance level α if the minimum observed p-value $p_{1:n} < 1-(1-\alpha)^{1/n}$", or using Fisher's decision rule: "reject H_0^* at significance level α if $-2\sum_{k=1}^n \ln(p_k) > \chi^2_{2n,1-\alpha}$" are two simple rules illustrating respectively the direct use of the P_k and using transformed P_k. Moreover, Fisher's method is often an efficient way of using all the information available, while Tippett's test uses only drastically restricted information, and therefore they represent two extreme cases. For those reasons, in the next sections, we shall use those two tests to assess the results of computationally augmenting samples to test uniformity.

3 Random and Generalized p-Values

Bayarri and Berger [1], as many other authors, state that "p-values are often perceived as measurements of the degrees of surprise in the data, relative to a hypothesized model". In general, the "hypothesized model" considered is the one indicated in the null hypothesis, but occasionally this is not convenient.

In fact, sometimes there are convincing reasons to think that the true hypothesis is H_A — as in the case of tests combining p-values —, and therefore when performing the test the surprise deals with the alternative.

In recent years, relevant advances on modeling p-values in more general settings than the validity of the null hypothesis have been made, for instance work on random and on generalized p-values, cf. [27, 28]. The main features of those developments, discussed focusing on meta-analyzing p-values, are presented in [7].

Most meta-analysis syntheses are done for combining effect sizes from different studies. However, the pooling of statistical evidence into a common estimate, usually a weighted combination of the individual effect sizes estimates $\tilde{\theta}_k$, $k = 1,\ldots,n$, only makes sense if the null hypothesis of homogeneity, i.e. $H_0 : \theta_1 = \ldots = \theta_n = \theta$, is not rejected. Hartung *et al.* [16] describe some common methods for combining effect sizes from different experiments. For example, a standard test for homogeneity of means in meta-analysis syntheses is Cochran's asymptotic chi-square test. Unless for some very special cases, e.g. testing

homogeneity of means for gaussian populations with known variances or a common unknown variance, these methods rely on approximations.

Finding exact methods whose distributions are free from nuisance parameters when these are present can be quite challenging, if not impossible, using standard statistical procedures. If nuisance parameters are involved, the chances of not computing p-values exactly are quite high, and using approximate p-values in pooling statistical evidence can be a problem.

Tsui and Weerahandi [27] extended the conventional definition of p-value by introducing the concept of generalized p-value, in order to eliminate the former's dependence on nuisance parameters. To compute a generalized p-value a generalized test variable has to be found first, i.e. a random variable $T = T(X;x,\theta,\eta)$, where $X = (X_1,\ldots,X_n)$ is a random sample, $x = (x_1,\ldots,x_n)$ the observed sample, θ the parameter of interest and η the nuisance parameter (real or vector).

A generalized test variable for the parameter θ has to satisfy three properties: i) the observed value of $T(X;x,\theta,\eta)$, i.e. $T(x;x,\theta,\eta)$, is free of θ and η; ii) when θ is specified, the distribution of $T(X;x,\theta,\eta)$ is free of η; iii) for fixed x and η, $\mathbb{P}(T \leq t;\theta)$ is a monotonic function of θ for any given t. Thus, for the right one-sided test $H_0 : \theta \leq \theta_0$ *vs.* $H_A : \theta > \theta_0$, the generalized p-value is given by

$$p_G = \mathbb{P}(T(X;x,\theta,\eta) \geq T(x;x,\theta,\eta)|\theta = \theta_0)$$

which securely won't depend on η. The advantage of using generalized p values instead of ordinary p values in testing problems with nuisance parameters, is that the former enables the problems to be solved exactly.

Another important aspect when dealing with p-values is to recognize that they are conditional on the data gathered from a particular experiment, i.e. they are the observed value of some test statistic T. As pointed out by Kulinskaya *et al.* [18], if we want to compare p-values from different experiments, or combine them in meta-analysis syntheses, we must deal with p-values as random variables, specially when there is some evidence that the alternative H_A is true.

For the particular case of large values of a continuous test statistic T giving evidence in favor of H_A, the observed p-value is $p = 1 - F_0(t)$, where t is the observed value of T and F_0 denotes the df of T under H_0. Therefore, the random p-value associated with T is, in this case, the random variable $P = 1 - F_0(T)$. From the previous definition it follows immediately that $P \frown Uniform(0,1)$ under the null hypothesis. On the other hand, the df of P under some alternative θ is

$$\mathbb{P}_\theta(P \leq p) = 1 - F_\theta(F_0^{-1}(1-p)), \quad 0 < p < 1,$$

where F_θ denotes the df of T under such alternative. So, if there is evidence that H_0 is false, the correct approach to combine statistical evidence should be under H_A, not under H_0. Taking into consideration these ideas, Brilhante [7] obtained explicit expressions for the pdf of P related to Fisher's test statistic, for small sample sizes, when the alternative to uniformity is the pdf of the random variable defined in (1).

4 Generating Pseudo-p-Values and Loss of Power Using Computationally Augmented Samples

Aside from the combination of p-values techniques to test the overall composite hypothesis H_0^* *vs.* H_A^*, as described in Section 2, a test of goodness-of-fit by the standard uniform would settle the matter. However in meta-analytic syntheses the size n of the p-values sample is in general rather small, and therefore the power of the test uncomfortably low.

In 2009, Gomes *et al*. [15] thought that computationally inflating the set of p-values would increase power. The algorithms they used to create *pseudo-p-values* have been, given a starting sample $(p_1,...,p_n)$ of p-values, to compute

- $p_{n+k} = \min\left(\frac{u_k}{p_k}, \frac{1-u_k}{1-p_k},\right),\ k = 1,..n$, where the u_k are uniform pseudo-random numbers;
- $p_{2n+k} = p_k + p_{n+k} - \lfloor p_k + p_{n+k} \rfloor,\ \ k = 1,..n$, where the floor function $\lfloor a \rfloor$ denotes the largest integer not greater than a.

The rationale for that is the following theorem:

Theorem 1. *Let $U \frown Uniform(0,1)$ and X with support $[0,1]$ be independent random variables. Then*

$$\min\left(\frac{U}{X}, \frac{1-U}{1-X},\right) = V \frown Uniform(0,1)$$

with V and X independent;

$$U + X - \lfloor U + X \rfloor = W \frown Uniform(0,1)$$

with W and X independent.

To the authors surprise, power *decreased* instead of increasing, thus inflating the sample led to worse performance. Figure 1, resulting from extensive simulation, shows the decrease of power when the sample is increased from $(p_1,...,p_n)$ to $(p_1,...,p_{2n})$, and then further increased to $(p_1,...,p_{3n})$, for some values of n (the results are for a two-sided test). This unexpected result will be explained in Section 5.

Remark 1: Since high or low p-values are now both considered extreme cases, the two-sided Tippett's test statistic is based on the midrange statistic, i.e. $\mathscr{M}_n = \frac{X_{1:n}+X_{n:n}}{2}$, whose pdf under uniformity is $f_{\mathscr{M}_n}(x) = n2^{n-1}(\frac{1}{2} - |x - \frac{1}{2}|)^{n-1}\, \mathrm{I}_{(0,1)}(x)$. Therefore, Tippett's decison rule is reject the null hypothesis at significance level α if $\mathscr{M}_n(obs) < m_{n,\alpha/2}$ or $\mathscr{M}_n(obs) > m_{n,1-\alpha/2}$, where the α quantile for $\mathscr{M}_n$ is

$$m_{n,\alpha} = \begin{cases} \left(2^{1-n}\alpha\right)^{1/n} & ,\ 0 < \alpha < \frac{1}{2} \\ 1 - \left[2^{1-n}(1-\alpha)\right]^{1/n} & ,\ \frac{1}{2} \leq \alpha < 1 \end{cases}.$$

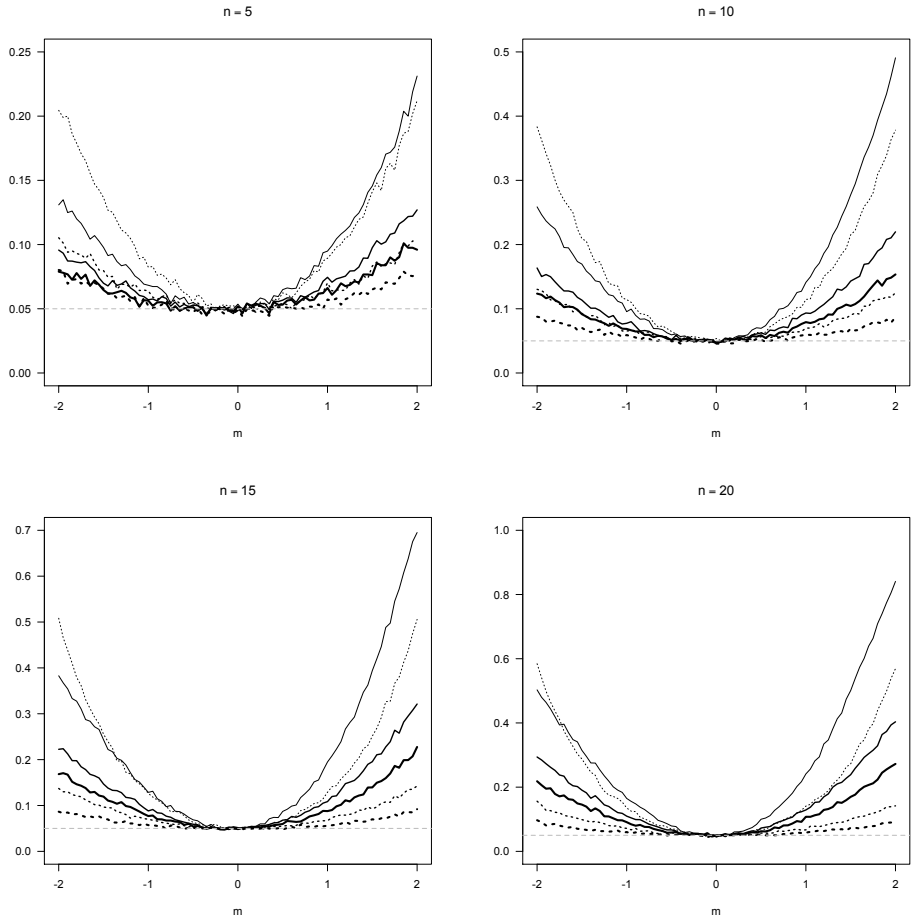
5 Mixtures of Uniform and Beta(1,2) or Beta(2,1) Distributions

One interesting controversy in science arose from Fisher's suggestion that the results reported by Mendel on his pioneering research on the mathematics of genetic inheritance were too good to be true. Fisher didn't boldly accuse Mendel of cheating, but suggests that he didn't control properly the integrity of some of his co-workers. Since there is clear indication in Mendel's writings that when results from the experience were too surprising there was a second data collection, and the most conform (with Mendel's theories) of the two results was reported. Pires and Branco [24] published an ingenious explanation assuming that there had been in fact an important proportion of experiments dealt with this way.

Meanwhile, Brilhante, Pestana and co-workers ([15, 4, 5, 23]) investigated a family of random variables $X_m,\ m \in [-2,2]$, that perfectly fits the situation described. The pdf of X_m is

$$f_{X_m}(x) = \left[1 + m\left(x - \frac{1}{2}\right)\right] \mathbb{I}_{(0,1)}(x) \tag{2}$$

Fig. 1 Power function: two-sided test $H_0: \ m = 0$ *vs.* $H_A: \ m \in [-2,0) \cup (0,2]$ (solid lines for the Fisher test, dashed line for the Tippett test; the thinner lines refers to the original n size sample, the medium to the double size sample, the thicker lines to the $3n$ size augmented sample)



(observe that $X_0 \frown Uniform(0,1)$, $X_{-2} \frown Beta(1,2)$, and $X_2 \frown Beta(2,1)$), that precisely model the sampling distribution of the reported p-values under H_0 in the above described circumstance, when the proportion $\frac{|m|}{2}$ of duplicated experiments is known. More complex convex mixtures of $Beta(k,1)$ or of $Beta(k,p)$ models can arise if the experimenter performs a variable number of experiments, either until he obtains a p-value that fits his expectations, or attains a predetermined maximum number of experiments.

Observe that X_m models not a generalized or a random p-value, these models arise under true H_0 when there is a dubious experimental behaviour.

What is the effect of the $P_k, \ k = 1, ..., n$, being independent replicas of $X_m, \ m \neq 0$, instead of iid standard uniform? The answer is given in the theorem that follows:

Theorem 2. *Let X_{m_1}, X_{m_2} be independent random variables with pdf $f_{m_i}(x) = \left[1 + m_i \left(x - \frac{1}{2}\right)\right] \mathbb{I}_{(0,1)}(x)$, with $m_i \in [-2,2], \ i = 1,2$. Then*

$$V_{m_1,m_2} = \min\left(\frac{X_{m_1}}{X_{m_2}}, \frac{1-X_{m_1}}{1-X_{m_2}},\right) \;=\; X_{\frac{m_1 m_2}{6}} = \begin{cases} U & X_{2\,sign(m_1 m_2)} \\ 1-\frac{|m_1 m_2|}{12} & \frac{|m_1 m_2|}{12} \end{cases}.$$

On the other hand

$$W_{m_1,m_2} = X_{m_1} + X_{m_2} - \lfloor X_{m_1} + X_{m_2} \rfloor = \begin{cases} U & Y \\ 1-\frac{|m_1 m_2|}{12} & \frac{|m_1 m_2|}{12} \end{cases}$$

where $Y \frown Beta(2,2)$.

In the convex mixture $f_{V_{m_1,m_2}}(x) = \left[1+\frac{|m_1 m_2|}{6}\left(x-\frac{1}{2}\right)\right]\mathbb{I}_{(0,1)}(x)$ the mixing coefficient of the uniform component is $1-\frac{|m_1 m_2|}{12}$, greater than the uniform mixing components $1-\frac{|m_i|}{6},\ i=1,2$, of either X_{m_1} or X_{m_2}. While for $X_{m_i},\ i=1,2$, the the absolute value of the slope of the "tilting" of the uniform pdf can be as large as 2, the absolute value of the slope of the pdf of V_{m_1,m_2} will be at most $\frac{2}{3}$.

A similar result applies to $f_{W_{m_1,m_2}}(x) = \left[1-\frac{|m_1 m_2|}{12}+\frac{|m_1 m_2|}{2}x(1-x)\right]\mathbb{I}_{(0,1)}(x)$. So, both V_{m_1,m_2} and W_{m_1,m_2} are closer to the uniform than either X_{m_1} or X_{m_2}. Observe that, in particular, when $m_1 = 0$ or $m_2 = 0$ (i.e., X_{m_1} or X_{m_2} uniform), the result will be uniform.

This and similar results reflect the absorbing behaviour of the uniform when we perform some algebra using variables with support $[0,1]$, that is a consequence of the maximal entropy property of the standard uniform among random variables in that class ([4]).

This also explains why the computational inflation of the sample of p-values in Section 4 lowers the power of uniformity tests. In the worst possible case — using auxiliary uniform pseudo-random numbers — the generated pseudo-p-values will be uniform, and hence the whole sample will be much more difficult to distinguish from an uniform sample, even when in fact the alternative hypothesis is true.

6 Further Issues

6.1 Estimation of m

In Section 5 we have assumed that the mixing parameter m in models such as the one discussed in [24] is known, but in realistic settings this is not so. Estimating m is far from simple, since the usual methods (maximum likelihood, minimum chi-square, moments) often produce inadequate or even inadmissible estimates.

A simulation evaluation of quantile regression methods, a method using spacings, regression methods, and fitting crossed expected values $\mathbb{E}[X^k(1-X)^\ell]$ is under progress.

6.2 Generating pseudo-p-values and other ways of testing uniformity

Further algorithms to generate pseudo-p-values can be used. For instance, if U_1, U_2 are independent replicas of $U \frown Uniform(0,1)$, $\dfrac{U_{1:2}}{U_{2:2}}$ will be uniform; however, the very sensible question of dependence arises, and therefore the via we exploit using Theorem 1 and Theorem 2 must be preferred. Johnson *et al.* [17] (pp.

313–314) discuss other algorithms to generate directly uniform order statistics and spacings generated in the random division of the unit interval.

Other ways of investigating uniformity of the p sample of p-values use other computational methods to inflate information. Brilhante *et al.* [6] devised a complex Sukhatme's type algorithm to test uniformity:

Let $X = (X_1, X_2, \ldots, X_n)$ be a random sample from the absolutely continuous positive random variable X with pdf f_X, and $(X_{1:n}, X_{2:n}, \ldots, X_{n:n})$ the corresponding vector of ascending order statistics. For convenience we assume that left-endpoint $\alpha_X = 0$ and we define $X_{0:n} = \alpha_X = 0$.

The joint pdf of the spacings $S_k = X_{k:n} - X_{k-1:n}$, $k = 1, \ldots, n$, is

$$f_{(S_1,S_2,\ldots,S_n)}(s_1, s_2, \ldots, s_n) = n!\, f_{(X_1,X_2,\ldots,X_n)}(s_1, s_1 + s_2, \ldots, s_1 + \cdots + s_n)$$

($s_k > 0$, $k = 1, \ldots, n$, and if the right-endpoint ω_X is finite, $\sum_{k=1}^{n} s_k < \omega_X$; in this case we can consider the rightmost spacing $S_{n+1} = \omega_X - X_{n:n}$, but this can be expressed as a function $\omega_X - \sum_{k=1}^{n} S_k$). Hence the joint pdf of the ascending reordering of those n spacings is

$$f_{(S_{1:n},S_{2:n},\ldots,S_{n:n})}(y_1, y_2, \ldots, y_n) = (n!)^2 f_{(X_1,X_2,\ldots,X_n)}(y_1, y_1 + y_2, \ldots, y_1 + \cdots + y_n),$$

where $0 < y_1 < y_2 < \ldots < y_n$ and $\sum_{k=1}^{n} y_k < \omega_X$.

Now define

$$W_k = (n + 1 - k)(S_{k:n} - S_{k-1:n}), \quad k = 1, \ldots, n,$$

(similar to Sukhatme's transformation, as defined in David and Nagaraja [8], but applied to ascendingly ordered spacings) again with the convention $S_{0:n} = 0$.

The joint pdf of $(W_1, W_2, \ldots, W_n)$ is

$$f_{(W_1,W_2,\ldots,W_n)}(w_1, w_2, \ldots, w_n) = n!\, f_{(X_1,X_2,\ldots,X_n)}\left(\tfrac{w_1}{n}, \tfrac{2w_1}{n} + \tfrac{w_2}{n-1}, \ldots, w_1 + \cdots + w_n\right),$$

$w_k > 0$, $k = 1, \ldots, n$, (observe that the k-th argument is

$$\frac{kw_1}{n} + \frac{(k-1)w_2}{n-1} + \cdots + \frac{(k+1-j)w_j}{n+1-j} + \cdots + \frac{w_k}{n+1-k},\ k = 1, \ldots, n),$$

and the joint pdf of the vector of partial sums $Y_k = \sum_{j=1}^{k} W_j$, $k = 1, \ldots, n$, is

$$f_{(Y_1,Y_2,\ldots,Y_n)}(y_1, y_2, \ldots, y_n) = n!\, f_{(X_1,X_2,\ldots,X_n)}\left(\tfrac{y_1}{n}, \ldots, \sum_{j=1}^{k} \tfrac{(k+1-j)(y_j - y_{j-1})}{n+1-j}, \ldots, y_n\right)$$

with $0 < y_1 < y_2 < \ldots < y_n$ and the convention $y_0 = 0$.

If $X \frown \text{Uniform}(0, \omega_X)$, then

$$f_{(X_1,X_2,\ldots,X_n)}\left(\tfrac{y_1}{n}, \ldots, \sum_{j=1}^{k} \tfrac{(k+1-j)(y_j - y_{j-1})}{n+1-j}, \ldots, y_n\right) = \frac{1}{\omega_X^n} = f_{(X_1,X_2,\ldots,X_n)}(y_1, y_2, \ldots, y_n),$$

and hence $(Y_1, Y_2, \ldots, Y_n) \stackrel{d}{=} (X_{1:n}, X_{2:n}, \ldots, X_{n:n})$.

(Observe that if $\omega_X < \infty$, we can consider $n+1$ spacings, with $S_{n+1} = \omega_X - X_{n:n}$; of course in this situation S_{n+1}, $S_{n+1:n+1}$ and W_{n+1} (where in this case it is convenient to use the transformation

$$W_k = (n + 2 - k)(S_{k:n+1} - S_{k-1:n+1}),$$

as in Johnson *et al.* [17] (p. 305) can be expressed as simple functions of the predecessor members of the corresponding samples. We still get the result that $(Y_1, Y_2, \ldots, Y_n) \stackrel{d}{=} (X_{1:n}, X_{2:n}, \ldots, X_{n:n})$ in case of standard uniform parent X.)

This suggests that uniformity can be investigated testing wether $\{X_{k:n}\}_{k=1}^n$ and $\{Y_k\}_{k=1}^n$ can be considered samples from the same distribution. Unfortunately, under the null hypothesis that the parent distribution is standard uniform,

$$(Y_1, Y_2, \ldots, Y_n) \stackrel{d}{=} (X_{1:n}, X_{2:n}, \ldots, X_{n:n}),$$

but the two vectors are not independent, since we can re-express $Y_k = \sum_{j=1}^k S_{j:n} + (n-k)S_{k:n}$, and consequently $Y_n = X_{n:n}$. Thus, Smirnov two-sample test is of no use in the present situation.

6.3 Size matters!

The temptation to use "many" data — and nowadays automatic data collection, large data sets, resampling techniques, and the use of simulated data easily serve such a purpose — must be used sparingly.

In fact, with too many data the most irrelevant differences will be significant. Consider for instance some 2×2 contingency table appropriate to investigate independence of factors (i.e., table with free margins, resting from a dichotomous cross classification of a sample of size n), say

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	$n(=a+b+c+d)$

The usual test statistic is Pearson's chi-square,

$$X^2_{2,2} = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}\,.$$

For instance, if the table is

33	54	87
40	42	82
73	96	169

then the observed value of the test statistics is 2.025, corresponding to a p-value of 0.155, and at the usual level of significance independence is not rejected. However with the similar table

33×5	54×5	87×5
40×5	42×5	82×5
73×5	96×5	169×5

(strictly similar in the sense that the odds ratios are exactly the same) the observed value of the test statistic is 2.025×5, and the corresponding p-value is 0.0015, leading to straight rejection. It can be argued: OK, it's natural, with bigger size we have stronger evidence.

This is so — but the worrying question is: with a sample size big enough, at the end of the day ANY null hypothesis will be rejected. With too many data anything can be rejected, even truth!

This is a caveat on the abusive use of cumulative meta analysis, computational samples inflation, and in general resampling techniques. Statistics is to be used sparingly, *quantum satis* as in the old recipes.

Appendix

We provide full proof of theorems 1 and 2; the fact that $U+X-\lfloor U+X\rfloor$ is uniform has been first remarked by Feller [10].

Proof of theorem 1

Both variables V and W have support on $[0,1]$. Therefore, for $0<v<1$,

$$\mathbb{P}(V\leq v)=\mathbb{P}(U\leq vX)+\mathbb{P}(1-U\leq v(1-X))=\int_0^1 vxf_X(x)dx+\int_0^1 v(1-x)f_X(x)dx=$$
$$=v\mathbb{E}(X)+v\mathbb{E}(1-X)=v$$

proving that $V\frown Uniform(0,1)$.

The independence between V and X follows from the fact that

$$\mathbb{P}(V\leq v|X=x)=\mathbb{P}(U\leq vx)+\mathbb{P}((1-U)\leq v(1-x))=vx+v(1-x)=$$
$$=v=\mathbb{P}(V\leq v).$$

On the other hand, since $S=U+X$ has pdf

$$f_S(s)=F_X(s)\mathbb{I}_{[0,1)}(s)+[1-F_X(s-1)]\mathbb{I}_{[1,2]}(s)$$

where F_X denotes the df of X, then for $0<w<1$,

$$\begin{aligned}\mathbb{P}(W\leq w)&=\mathbb{P}(0\leq U+X\leq w)+\mathbb{P}(1\leq U+X\leq 1+w)\\&=\int_0^w F_X(s)ds+\int_1^{1+w}(1-F_X(s-1))dx\\&=\int_0^w F_X(s)ds+\int_1^{1+w}ds-\int_1^{1+w}F_X(s-1)ds\\&=\int_0^w F_X(s)ds+w-\int_0^w F_X(x)dx=w\end{aligned}$$

showing that $W\frown Uniform(0,1)$.

Regarding the independence between W and X:

$$\begin{aligned}\mathbb{P}(W\leq w|X=x)&=\mathbb{P}(0\leq U+x\leq w)+\mathbb{P}(1\leq U+x\leq 1+w)\\&=\mathbb{P}(0\leq U\leq w-x)+\mathbb{P}(1-x\leq U\leq 1-x+w)\\&=\max\{0,w-x\}+\min\{1,1-x+w\}-(1-x)\end{aligned}$$

Two cases are to be considered here:

i) if $w-x<0$, then $1-x+w<1$, thus

$$\mathbb{P}(W\leq w|X=x)=1-x+w-(1-x)=w$$

ii) if $w-x>0$, then $1-x+w>1$, and therefore

$$\mathbb{P}(W \leq w|X=x) = w-x+1-(1-x) = w$$

revealing that W and X are indeed independent random variables.

Proof of theorem 2

Let F_m and f_m denote the df and the pdf of X_m, $m \in [-2,2]$, respectively, and let F_m^* and f_m^* denote the df and pdf of the random variable $1-X_m$, where $F_m^*(x) = 1-F_m(1-x)$ and $f_m^*(x) = f_m(1-x)$.
For $0<v<1$,

$$\begin{aligned}
\mathbb{P}(V_{m_1,m_2} \leq v) &= \mathbb{P}(X_{m_1} \leq vX_{m_2}) + \mathbb{P}((1-X_{m_1} \leq v(1-X_{m_2})) \\
&= \int_0^1 F_{m_1}(vx) f_{m_2}(x)dx + \int_0^1 F_{m_1}^*(vx) f_{m_2}^*(x)dx \\
&= \int_0^1 [\tfrac{m_1}{2}(vx)^2 + (1-\tfrac{m_1}{2})vx][1+m_2(x-\tfrac{1}{2})]dx + \\
&\qquad + \int_0^1 [-\tfrac{m_1}{2}(vx)^2 + (1+\tfrac{m_1}{2})vx][1-m_2(x-\tfrac{1}{2})]dx \\
&= \tfrac{1}{24}[m_1(m_2+4)v^2 + (2-m_1)(m_2+6)v] + \\
&\qquad + \tfrac{1}{24}[m_1(m_2-4)v^2 + (2+m_1)(6-m_2)v] \\
&= \tfrac{m_1m_2}{12}v^2 + (1-\tfrac{m_1m_2}{12})v
\end{aligned}$$

and therefore $V_{m_1,m_2} \stackrel{d}{=} X_{\frac{m_1m_2}{6}}$.

On the other hand, since $S_{m_1,m_2} = X_{m_1} + X_{m_2}$ has pdf

$$f_{S_{m_1,m_2}}(s) = \begin{cases} \frac{m_1m_2}{6}s^3 + \frac{(m_1+m_2-m_1m_2)}{2}s^2 + \frac{(2-m_1)(2-m_2)}{4}s & , s \in (0,1) \\ -\frac{m_1m_2}{6}s^3 + \frac{(m_1m_2-m_1-m_2)}{2}s^2 + \frac{(6m_1+6m_2-m_1m_2-4)}{4}s - \frac{m_1m_2}{6} - m_1 - m_2 + 2 & , s \in [1,2) \\ 0 & , \text{other values} \end{cases}$$

then for $0<w<1$,

$$\begin{aligned}
\mathbb{P}(W_{m_1,m_2} \leq w) &= \mathbb{P}(0 \leq S_{m_1,m_2} \leq w) + \mathbb{P}(1 \leq S_{m_1,m_2} \leq 1+w) \\
&= \tfrac{m_1m_2}{24}w^4 + \tfrac{(m_1+m_2-m_1m_2)}{6}w^3 + \tfrac{(2-m_1)(2-m_2)}{8}w^2 - \\
&\qquad - \tfrac{m_1m_2}{24}w^4 - \tfrac{(m_1+m_2)}{6}w^3 + \tfrac{(m_1m_2+2m_1+2m_2-4)}{8}w^2 + \tfrac{(12-m_1m_2)}{12}w \\
&= -\tfrac{m_1m_2}{6}w^3 + \tfrac{m_1m_2}{4}w^2 + (1-\tfrac{m_1m_2}{12})w \\
&= \tfrac{m_1m_2}{12}(3w^2-2w^3) + (1-\tfrac{m_1m_2}{12})w
\end{aligned}$$

which proves that W_{m_1,m_2} is a mixture of a $Beta(2,2)$ and uniform random variables with mixing proportions $\frac{m_1m_2}{12}$ and $1-\frac{m_1m_2}{12}$, respectively.

Remark 2: While in theorem 1 the pairs of random variables X and V and X and W are independent — an essential feature to use them to augment the random sample —, the variables V_{m_1,m_2} and W_{m_1,m_2} are

correlated with both X_{m_1} and X_{m_2}. In fact, some standard algebra manipulation shows that

$$f_{V_{m_1,m_2}|X_{m_2}=x}(v) = \left[1+m_1\left(v(2x-1)-x+\tfrac{1}{2}\right)\right]\mathbb{I}_{(0,1)}(v)\,,$$

so independence does occur if and only if $X_{m_1} \frown Uniform(0,1)$. On the other hand,

$$f_{W_{m_1,m_2}|X_{m_1}=x}(w) = [1+m_2(w-x-\tfrac{1}{2})]\mathbb{I}_{A_1}(x,w)+[1+m_2(w-x+\tfrac{1}{2})]\mathbb{I}_{A_2}(x,w)\,,$$

where $A_1 = \{(x,w)\in \mathbb{R}^2 : 0<x<w<1\}$ and $A_2 = \{(x,w)\in \mathbb{R}^2 : 0<w\leq x<1\}$. Therefore independence between W_{m_1,m_2} and X_{m_1} occurs if and only if $X_{m_2} \frown Uniform(0,1)$.

Acknowledgements This research has been supported by National Funds through FCT — Fundação para a Ciência e a Tecnologia, project PEst-OE/MAT/UI0006/2011.

References

1. Bayarri, M.J., Berger, J.O.: Quantifying surprise in the data and model veridication. In J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith (eds), Bayesian Statistics **6**, pp. 53–82, Oxfor University Press, Oxford and New York (1998)
2. Birnbaum, A.: Combining independent tests of significance. *J. Amer. Statist. Assoc.* **49**, 559–575 (1954)
3. Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R.: *Introduction to Meta-Analysis*, Wiley, Chichester (2009)
4. Brilhante, M.F., Mendona, S., Pestana, D., Sequeira, F.: Using Products and Powers of Products to Test Uniformity, In Luzar-Stiffler, V., Jarec, I. and Bekic, Z. (eds.), Proceedings of the ITI 2010, 32nd International Conference on Information Technology Interfaces, IEEE CFP10498-PRT, pp. 509-514 (2010)
5. Brilhante, M.F., Pestana, D., Sequeira, F.: Combining p-values and random p-values, In Luzar-Stiffler, V., Jarec, I. and Bekic, Z. (eds.), Proceedings of the ITI 2010, 32nd International Conference on Information Technology Interfaces, IEEE CFP10498-PRT, pp. 515–520 (2010)
6. Brilhante, M. F., Malva, M., Mendona, S., Pestana, D., Sequeira, F., and Velosa, S. Uniformity. In Lita da Silva, J.; Caeiro, F.; Natrio, I.; Braumann, C.A. (Eds.): Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and other Statistical Applications, 73-81, Springer, Berlin (2013)
7. Brilhante, M.F.: Generalized p Values and Random p Values when the Alternative to Uniformity is a Mixture of a Beta(1,2) and Uniform. In Oliveira, P. *et al.* (eds), *Recent Developments in Modeling and Applications in Statistics*, Springer, Heidelberg, 159-167 (2013)
8. David, H. A., and Nagaraja, H. N.: *Order Statistics*, 3rd edn. Wiley, New York (2003).
9. Erdös, P., and Rényi, A.: On a central limit theorem for samples from a finite population. Publ. Math. Instit. Hungar. Acad. Sci. **4**, 49–61(1959)
10. Feller, W. . *An Introduction to Probability Theory and its Applications*, II, Wiley (1966)
11. Fisher, R. A.: *Statistical Methods for Research Workers*, 4th ed., Oliver and Boyd, (1932)
12. Glass, G. V.: Primary, secondary, and meta-analysis of research. *Edu. Res.*, **5**, 3-8 (1976)
13. Glass, G.V.: Integrating findings: The meta-analysis of research. *Review of Research in Education*, **5**, 351-379 (1978)
14. Glass, G. V.: Meta-Analysis at 25, http://glass.ed.asu.edu/gene/papers/meta25.html (1999)
15. Gomes, M.I., Pestana, D.D., Sequeira, F., Mendonca, S., Velosa, S.: Uniformity of offsprings from uniform and non-uniform parents. In Luzar-Stiffler, V., Jarec, I. and Bekic, Z. (eds.), Proceedings of the ITI 2009, 31st International Conference on Information Technology Interfaces, pp. 243–248 (2009)
16. Hartung, J., Knapp, G., and Sinha, B. K.: *Statistical Meta-Analysis with Applications*, Wiley, New York (2008)
17. Johnson, N. L., Kotz, S., and Balakrishnan, N.: *Contagious Univariate Distributions* 2, Wiley, New York (1995)
18. Kulinskaya, E., Morgenthaler, S., and Staudte, R. G.: *Meta Analysis. A Guide to Calibrating and Combining Statistical Evidence*, Wiley, Chichester (2008)
19. Littel, R. C., and Folks, L. J.: Asymptotic optimality of Fisher's method of combining independent tests, I. *J. Amer. Statist. Assoc.* **66**, 802–806 (1971)
20. Littel, R. C., and Folks, L. J.: Asymptotic optimality of Fisher's method of combining independent tests, II. *J. Amer. Statist. Assoc.* **68**, 193–194 (1973)
21. Mosteller, F., and Bush, R.: Selected quantitative techniques, in G. Lidsey (Ed.), *Handbook of Social Psychology: Theory and Methods*, vol. I, Addison-Wesley, Cambridge, MA (1954)

22. Pestana, D.: Combining p-values, in M. Lovric (Ed.), International Encyclopaedia of Statistical Science, pp. 1145–1147, Springer Verlag, New York (2011)
23. Pestana, D., Rocha, M.L., Vasconcelos, R., Velosa, S.: Publication Bias and Meta-Analytic Syntheses. In Lita da Silva, J.; Caeiro, F.; Natrio, I.; Braumann, C.A. (Eds.): Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and other Statistical Applications, pp. 347–354, Springer, Berlin (2013)
24. Pires, A.M., Branco, J.A.: A statistical model to explain the Mendel-Fisher controversy. Statistical Science **25** 545–565 (2010)
25. Stouffer, S. A., Schuman, E. A., DeVinney, L. C., Star, S. and Williams, R. M.: *The American Soldier*, vol. I: *Adjustment During Army Life*, Princeton University Press, Princeton (1949)
26. Tippett, L. H. C. : *The Methods of Statistics*, Williams & Norgate, London (1931)
27. Tsui, K., and Weerahandi, S.: Generalized p-values in significance testing of hypotheses. J. Amer. Statist. Assoc. **84**, 602–607 (1989)
28. Weerahandi, S. Exact Statistical Methods for Data Analysis. Springer, New York (1995)
29. Wilkinson, B.: A statistical consideration in psychological research. *Psychological Bulletin* **48**, 156–158 (1951)