ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2019-66871

Null Distribution of P-values and an Empirical 'Uniformitization' Proposal

P-değerlerinin Yokluk Dağılımları ve Ampirik Bir 'Üniformatizasyon' Önerisi

Mehmet KOCAK[a]

[a] Department of Preventive Medicine,
Division of Biostatistics,
The University of
Tennessee Health Science Center,
Memphis, USA

Received: 18.06.2019
Received in revised form: 26.06.2019
Accepted: 27.06.2019
Available Online: 20.12.2019

Correspondence:
Mehmet KOCAK
The University of
Tennessee Health Science Center,
Department of Preventive Medicine,
Division of Biostatistics, Memphis,
ABD/USA
mkocak1@uthsc.edu.

ABSTRACT Objective: In study designs, the statistical power to detect a desired effect size with a specified Type-1 error is computed with the assumption that the p-value distribution under the null hypothesis follows Uniform[0.1]. However, even small departures from this assumption may inflate or deflate the statistical power beyond expectations. In this study, we illustrated the departure of the p-value distribution from Uniform[0,1] for common tests and we proposed an empirical correction to the null p-value distribution. **Material and Methods:** Using statistical simulation techniques, we illustrated the p-value distributions of numerous commonly used hypothesis tests under the null hypothesis, quantified their departures from Uniform[0,1], and proposed a p-value correction algorithm called 'Uniformitization'. We then graphically illustrated and discussed the level of correction with this Uniformitization approach in the corresponding p-value distribution. **Results:** Other than Z-test as expected and the Student t-test to most degree, all other tests we used showed non-ignorable departures from Uniform[0.1]. Our Uniformitization approach corrects the p-value distribution and brings them much closer to Uniform[0,1] especially for continuous response. Although still substantial, the correction level is limited with binary and survival response variables due to the discrete nature of these outcome variables. **Conclusions:** The requirement that the null-distribution of p-values be Uniform[0,1] is an indispensable one to make sure that the obtained statistical power is really where it should be, and our Uniformitization approach provides such corrections in the null distribution of p-values when they deviate from what is theoretically assumed.

Keywords: P-value Distribution; statistical power correction; null hypothesis;
type-1 error rate; type-1 error correction

ÖZET Amaç: Bilimsel çalışma tasarımlarında, istatistiksel güç, hedeflenen bir etki büyüklüğünü tespit etmek için, Birinci Tip hatanın, yokluk hipotezi altında Uniform[0.1] dağıldığı varsayımı altında hesaplanır. Bununla birlikte, bu varsayımdan küçük uzaklaşmalar bile istatistiksel gücü beklentilerin ötesinde şişirebilir veya azaltabilir. Bu çalışmada, sıkça kullanılan testlerin, yokluk hipotezi altında Uniform[0,1] dağılımından uzaklaşmaları belirledik ve yokluk p-değeri dağılımına ampirik bir düzeltme algoritması önerdik. **Gereç ve Yöntemler:** İstatistiksel simülasyon yöntemlerini kullanarak, araştırmalarda sıkça kullanılan çok sayıda istatistiksel hipotez testinin, yokluk hipotezi altındaki p-değer dağılımlarını gösterdik, Uniform[0,1] dağılımından uzaklaşmaları belirledik ve 'Üniformatizasyon' olarak adlandırdığımız bir p-değeri düzeltme algoritması önerdik. Daha sonra, bu 'Üniformatizasyon' yöntemi ile, p-değeri dağılımında elde edilen düzeltmenin seviyesini grafiksel olarak gösterip, tartıştık. **Bulgular:** Beklendiği gibi, Z-testi ve Student-t-testi dışında, kullandığımız diğer tüm testler, Uniform[0.1] dağılımından göz ardı edilemez uzaklaşmalar gösterdi. Üniformatizasyon yaklaşımımızın, p-değeri dağılımlarında beklenen düzeltmeyi yapıp, onları özellikle sürekli hedef değiskenler için Uniform[0,1]'a yaklaştırdığı gözlendi. Bu düzeltme seviyesinin, hâla önemli olmakla birlikte, ikili ve sağkalım yanıt değiskenleri için, onların ayrık yapıları nedeniyle, sınırlı kaldığı gözlendi. **Sonuç:** P-değerlerinin yokluk dağılımının Uniform[0,1] olması şartı, elde edilen istatistiksel gücün gerçekten olması gereken yerde olmasını sağlamak için vazgeçilmezdir ve bizim Üniformatizasyon yaklaşımımız, p-değerlerinde yokluk hipotezi altında teorik olarak beklenen dağılımdan uzaklaşmalar olduğunda, gereken düzeltmeyi sağlamaktadır.

Anahtar Kelimeler: P-değeri dağılımı; istatistiksel güç düzeltmesi; yokluk hipotezi;
birinci tip hata oranı; birinci tip hata düzeltmesi

Merriam-Webster dictionary defines '*science*' as 'knowledge or a system of knowledge covering general truths or the operation of general laws especially as obtained and tested through scientific method' and in turn, it defines '*scientific method*' as 'principles and procedures for the systematic pursuit of knowledge involving the recognition and formulation of a problem, the collection of data through observation and experiment, and the formulation and testing of hypotheses' (https://www.merriam-webster.com/).

To test whether a new experiment on a given phenomenon adds significantly to what is already known about that phenomenon (i.e., the accumulated knowledge over time), Pearson (1900) tokened a new statistical term called 'p-value'.[1] Having what we already know about a given phenomenon, p-value gives us the magnitude of the probability that our new observation and any more extreme versions of it still originate from what we already know (i.e., our null belief or our initial belief). A researcher is free to choose any threshold for this p-value to claim that what he or she just observed adds, or does not add, significantly to the null belief although traditionally the threshold of 0.05 has been used and promoted, albeit unnecessarily, over decades.

In a general hypothesis testing framework, a null belief is expressed (e.g., '*5-year overall survival for this patient population is 68%*'), a new claim is formed to challenge the null belief (e.g., '*5-year overall survival for this patient population is 75% if patients are treated with this new therapy*'), a sample size is determined to significantly detect this proposed difference from the null belief with certain statistical power, say, 90%, with say, 5% Type-I error rate (e.g., '*213 patients treated with the new therapy and followed up to 10-years will provide 90% power to significantly detect a 5-year overall survival increase from 68% to 75% at significance level of 5%*'). Upon conducting the study, the researcher then formally tests the new 5-year survival estimate against the null belief (i.e., 68%) and determines the magnitude of the evidence in terms of a p-value.

It is easy to prove, and definitely intuitive, that p-value from Z-test or T-test follows *Uniform* distribution on the unit interval [0,1] (i.e., $p_{Z\text{-test}}$~Uniform[0,1]). The nature of p-value as a random variable and its stochastic characteristics were discussed by Dempster et al. (1965), Sackrowitz and Samuel-Cahn (1999), and Murdoch et al (2008).[2-4] Although generally overlooked, the distribution of p-values representing the magnitude of evidence for binary, count, survival, or other non-continuous, and non-traditional, endpoints against the null belief may have departures from *Uniform* [0,1]. This is also true for any non-parametric or semi-parametric test statistics. If this issue of departure from the expected null distribution is not addressed properly, all cascading operations on p-values including, for example, to control the false discovery rate (FDR) will be negatively impacted.[5]

In this study, we investigate the impact of the departure of the p-value distribution from *Uniform*[0,1] on statistical power and we propose an empirical correction to the p-value distribution so that both the Type-1 error and Power are retained at the desired level. In Section 2, we provide examples of departures of null p-value distributions from Uniform[0,1] distribution through simulations. We propose an empirical corrective approach, called Uniformitization, for the null p-value distributions in Section-3, followed by examples of how Type-I errors (i.e., null p-value distributions) are recovered to the desired level using our proposed approach in Section 4. We end with discussions and conclusions in Section 5.

MATERIAL AND METHODS

We first illustrate p-value distributions from various tests of statistics, primarily for continuous, binary, count, survival endpoints and for meta-analysis of p-values. Under each of these scenarios, we generated 10,000 simulation runs, produced the p-value distributions, and assessed the level of type-1 error rate retention at certain thresholds. Any departure of type-1 error distribution from Uniform(0,1) has a direct impact on the statistical power, either over-estimating or under-estimating it.

DEPARTURE OF NULL P-VALUE DISTRIBUTION FROM *UNIFORM*[0,1] FOR GAUSSIAN RESPONSE

Pretending we have a two-arm randomized clinical trial, where patients in both arms have the same continuous response distribution:

- N-per arm=20, 50, 100, 200, 500, 1000
- Tests (H_0: μ_1=μ_2): Z-test, T-test, and Wilcoxon-Mann-Whitney Test
- Response variable distribution: Gaussian, Non-Gaussian (Exponential)

We present the null p-value distribution for Z-test, T-test (Pooled variance, Cochran and Satterthwaite versions), Wilcoxon test with Z- and T-approximations in Figure-1 below for the Gaussian response case:

We can see the stability of the p-value distribution regardless of the sample size; interestingly, we also observe that for larger sample size Wilcoxon Test with Z- or T-approximation is actually much closer to the nominal P-value Probability Density Function (PDF) level of 0.05, even compared to Z-test. Also, for small sample sizes, p-value distribution is below the 0.05 level for all tests, which in turn inflates the statistical power.

The empirical p-value distribution is illustrated for the non-Gaussian case (Exponential distribution in this case) in Supplementary Figure-1, where we observe similar results with higher variability as expected for all tests.

These simulations also revealed that Z-test retained Type-I error rate to the traditional significance level of 0.05 very closely as expected while Cochran T-test Wilcoxon T-approximation test are underestimating the Type-1 error rate for small sample sizes in all these continuous response scenarios (Table 1).

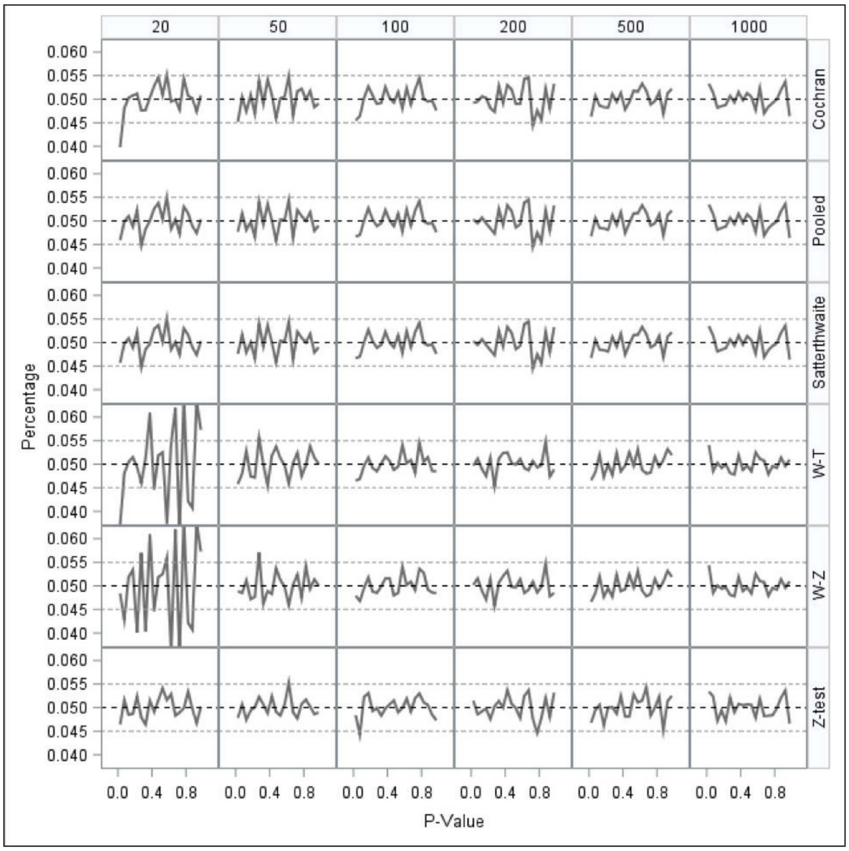

FIGURE 1: P-value Empirical Probability Density Function for Gaussian response scenario with an increment of 0.05 in [0.1] range.

TABLE 1: Type-1 error rates (null p-values) of Two-sample test for continuous response (Departures greater than 0.5%, that is, 0.005, is highlighted). All null p-values below must be considered as referenced to 0.05, which represents 5% Type-1 Error rate.

		Sample Size					
Two-Sample Tests		20	50	100	200	500	1000
Continuous-Gaussian	Cochran	0.043	0.047	0.049	0.049	0.050	0.050
	Pooled	0.049	0.050	0.050	0.050	0.050	0.050
	Satterthwaite	0.049	0.050	0.050	0.050	0.050	0.050
	W-T	0.040	0.047	0.048	0.050	0.050	0.049
	W-Z	0.049	0.050	0.049	0.050	0.050	0.049
	Z-test	0.050	0.051	0.050	0.050	0.050	0.050
Continuous-NonGaussian	Cochran	0.040	0.046	0.049	0.050	0.050	0.049
	Pooled	0.047	0.049	0.051	0.051	0.050	0.049
	Satterthwaite	0.045	0.048	0.051	0.051	0.050	0.049
	W-T	0.040	0.047	0.050	0.050	0.050	0.049
	W-Z	0.049	0.050	0.051	0.051	0.050	0.049
	Z-test	0.051	0.051	0.052	0.051	0.050	0.049

DEPARTURE OF NULL P-VALUE DISTRIBUTION FROM *UNIFORM*[0,1] FOR BINARY RESPONSE WITH BINARY PREDICTOR

Pretending we have a two-arm randomized clinical trial, where patients in both arms have the same binary response distribution:

- N per arm=20, 50, 100, 200, 500, 1000
- Success Rate: 0.01, 0.05, 0.10, 0.20, 0.50
- Response variable distribution: Bernoulli
- Tests: Pearson's Chi-Square, Likelihood Ratio, Mantel-Haenszel, Score, and Wald tests, and Fisher's Exact Test.[6]
- Additional Tests: Firth Penalized Likelihood versions of Likelihood Ratio (LR), Score, and Wald tests

For binary response scenario, we present the empirical PDF of p-value for Chi-Square, Fisher's Exact, Wald, and First-adjusted Wald tests in Figure-2 for sample size of 20 per study arm.

For rare events, we see a high accumulation of unit p-value and we observe that the PDF starts getting closer to the nominal 0.05 line as the probability of success increases, where Pearson's Chi-Square and Wald-test p-values seem to have the least variability around the 0.05 line of theoretical PDF.

For large sample sizes, we have much closer empirical PDF to the 0.05 line as seen in Supplementary Figure-2, while we still have the zig-zag patterns for rare events.

For the binary response in two-sample comparison, we see that all tests under-estimate Type-I error for rare event-events with small sample size (Table 2). Interestingly, Likelihood Ratio Test over-estimate the Type-I error rate for sample sizes 100 and above for rare events. Specifically, we see that the continuity-adjusted Chi-Square test and Fisher's Exact test always underestimates the Type-I error rate. Similarly, Wald and Firth-adjusted Wald test also underestimate Type-I error rates for small sample sizes. The same conclusion is true for the Wald test for small sample sizes when the association of a binary response with a continuous predictor is investigated.

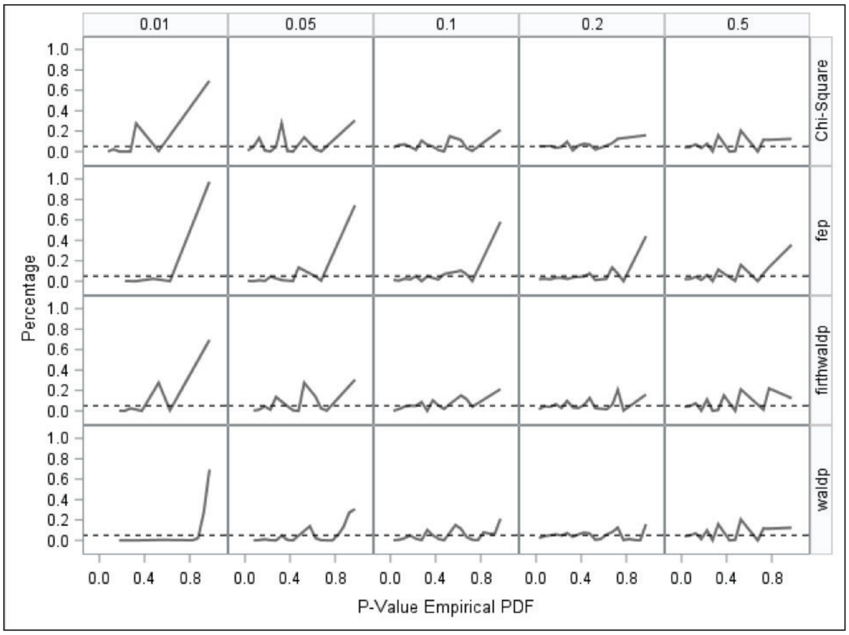

FIGURE 2: Empirical P-value PDF for binary response with sample size of 20 per arm (columns represents the probability of successes).

TABLE 2: Type-1 error rates of Two-sample test for binary response (Departures greater than 0.5% is highlighted).

	0.01					0.2					0.5				
	Sample Size					Sample Size					Sample Size				
	20	50	100	500	1000	20	50	100	500	1000	20	50	100	500	1000
Chi-Square	0	0.002	0.015	0.047	0.058	0.051	0.056	0.050	0.052	0.049	0.044	0.059	0.056	0.053	0.054
contchisqp	0	0	0.000	0.023	0.031	0.018	0.033	0.032	0.042	0.043	0.017	0.035	0.040	0.047	0.048
fep	0	0	0.000	0.024	0.034	0.022	0.033	0.032	0.042	0.043	0.020	0.035	0.040	0.047	0.048
firthlogre	0	0.002	0.003	0.045	0.053	0.050	0.053	0.047	0.052	0.049	0.040	0.059	0.056	0.050	0.054
firthscore	0	0	0.003	0.043	0.048	0.046	0.053	0.047	0.052	0.049	0.044	0.059	0.056	0.053	0.054
firthwaldp	0	0	0	0.021	0.037	0.016	0.040	0.043	0.052	0.048	0.039	0.042	0.043	0.047	0.054
lrp	0.001	0.018	0.060	0.076	0.062	0.056	0.056	0.050	0.053	0.049	0.044	0.059	0.056	0.053	0.054
mhp	0	0.002	0.015	0.047	0.058	0.051	0.053	0.047	0.052	0.049	0.044	0.059	0.056	0.051	0.054
waldp	0	0		0.021	0.044	0.024	0.046	0.046	0.052	0.049	0.040	0.059	0.056	0.050	0.054

DEPARTURE OF NULL P-VALUE DISTRIBUTION FROM *UNIFORM*[0,1] FOR BINARY RESPONSE WITH CONTINUOUS PREDICTOR

Pretending we have a study with a binary response and like to investigate the association of a continuous predictor with the likelihood of our binary response:

- N per arm=50, 100, 150, 200
- Success Rate: 0.10, 0.20, 0.30, 0.40, 0.50
- Response variable distribution: Bernoulli with a Gaussian continuous predictor
- Tests: Likelihood Ratio, Score, and Wald tests from Logistic Regression.

With a continuous predictor for a binary response, we observe that the Score test has the least departure from the theoretical PDF as shown in Figure 3.

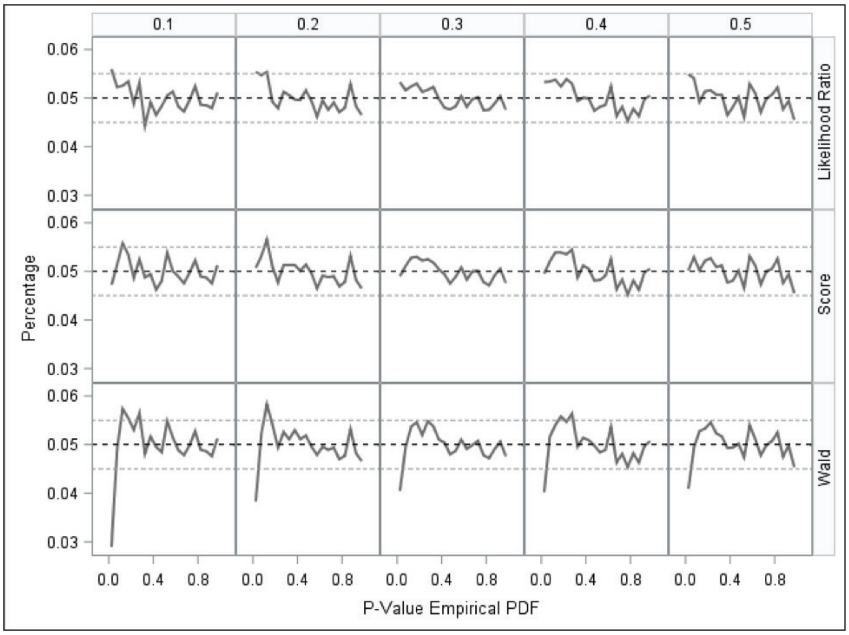

FIGURE 3: Empirical P-value PDF for binary response with continuous predictor with sample size of 50 per arm (columns represents the probability of successes).

With increasing sample size, all three tests approach to the theoretical PDF with similar departure variations.

DEPARTURE OF NULL P-VALUE DISTRIBUTION FROM *UNIFORM*[0,1] FOR SURVıVAL ENDPOINT

Pretending we have a two-arm randomized clinical trial, where patients in both arms have the same survival distribution:

- N-per arm=50, 100, 200, 400, 1000
- Mean time to event: 40 units of time
- Follow-up Rate: 10, 30, 50 units of time
- Response variable distribution: Exponential
- Tests: Likelihood Ratio, Score and Wald tests from Cox Proportional Hazards Model.

For survival endpoint, we see much larger departure with shorter follow-up time as expected (Figure 4).

Type-1 error retention in Survival tests is much more reasonable (Table 3) although Wald test seems to underestimate it for small sample size with shorter survival compared to other tests.

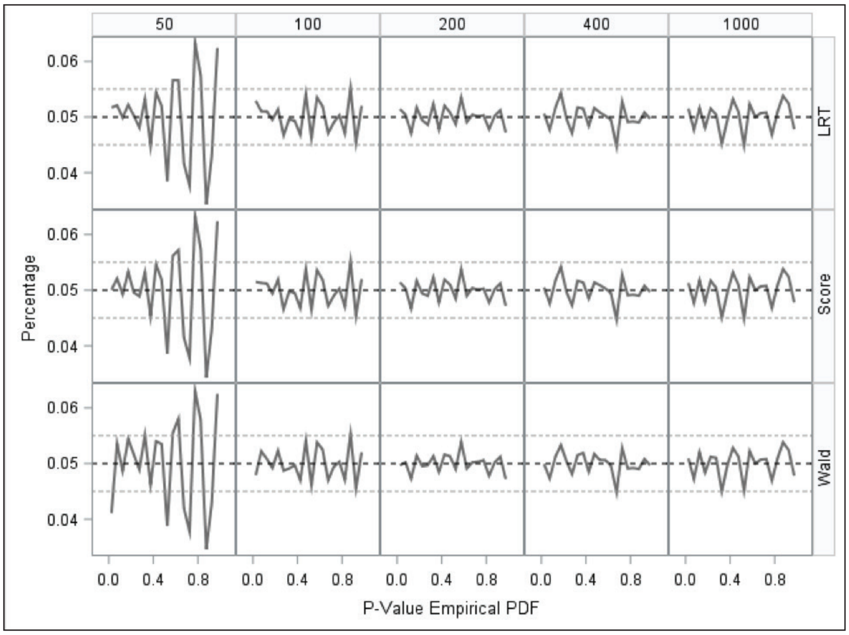

FIGURE 4: Empirical P-value PDF for survival endpoint with follow-up time of 10 unit times (columns represents the sample size).

TABLE 3: Type-1 error rates of Two-sample test for binary response (Departures greater than 0.5% is highlighted).

	Follow-up														
	10					**30**					**50**				
	Sample Size					**Sample Size**					**Sample Size**				
Survival Endpoint	**50**	**100**	**200**	**400**	**1000**	**50**	**100**	**200**	**400**	**1000**	**50**	**100**	**200**	**400**	**1000**
LRT	0.052	0.053	0.051	0.051	0.052	0.051	0.050	0.047	0.052	0.049	0.050	0.047	0.050	0.047	0.049
Score	0.050	0.052	0.051	0.051	0.051	0.052	0.051	0.047	0.052	0.049	0.052	0.048	0.051	0.047	0.049
Wald	0.041	0.048	0.050	0.050	0.051	0.049	0.049	0.046	0.051	0.049	0.049	0.047	0.050	0.047	0.049

DEPARTURE OF NULL P-VALUE DISTRIBUTION FROM *UNIFORM*[0,1] FOR META-ANALYSIS OF P-VALUES

In a meta-analysis of p-values framework, we generated 100,000 null p-value sets.

- Study size: 4, 12

- Response variable distribution: Uniform[0,1]

- Meta P-value Tests: Fisher's, George's, Kocak's, and Strouffer's, tests[7-9]

For Meta-Analysis of p-values, Figure 5 depicts the empirical PDFs of meta-p-values under the null hypothesis.

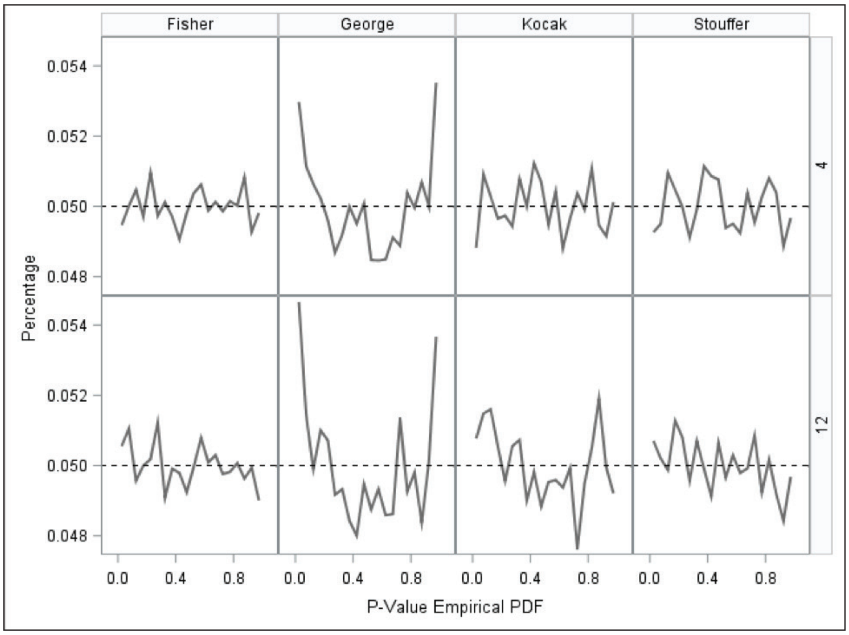

FIGURE 5: Empirical P-value PDF for meta-analysis of p-values endpoint (rows represents the sample size).

George's test over-estimated the Type-I error rate for both small and large sample cases (Type-I error rate=0.053 and 0.055, respectively) while the other tests retained it close to 0.05.

All the above simulation results point to the fact that the null distribution of p-values are not necessarily Uniform[0,1] expect for Z-test, and the Type-1 error rate may be under- or over-estimated for a given test, which directly inflates or deflates the statistical power. Due to this fact, statistical power comparison among competing statistical tests may not be accurate and fair as the underlying Type-1 error mechanism is not the same. To address this important issue, we propose the following Empirical 'Uniformitization' process before we conduct any test for specific sample sizes for a specific endpoint.

3 An Empirical P-value correction proposal: Uniformitization

The Uniformitization process includes the following steps:

Step-1, Generating Empirical Null Distribution: We generate the null p-value distribution of the specific case under investigation using a large number simulations (e.g., N-simulation runs=1 million) with the sample size at hand, and obtain the p-values. We then identify the percentiles (or lower order quantile estimates such as 0.1th percentile, etc., depending on how precisely we want to correct the Null P-value distribution) from the empirical distribution of p-values. If our test-statistic produces p-values that follow Uniform[0,1], then the estimated percentiles will be equal to the percentiles of Uniform[0,1]; for example, 5th percentile, our traditional Type-I error rate, would be equal to 0.05 under ideal circumstances.

Step-2, Generating Correction Weights: Consecutive percentile estimates from Step-1 provide correction intervals that will be used as weights. For example, say, 1st and 2nd percentiles were estimated to be 0.015 and 0.030, respectively, which indicates that the empirical null distribution of p-values are above the ideal distribution. Therefore, any p-value that falls in the interval [0,0.015] must be projected onto the interval [0,0.01] by multiplying these p-values by 1/1.5 in this case, and any p-value that falls in the interval (0.015,0.03] must be projected to the interval (0.01,0.02], again multiplying these p-values by 1/1.5. These multiplicative weights are nothing but the ratio of the expected size of the p-value bins to the observed size of the corresponding bin. When this process is applied for the entire support of Uniform[0.1], all corrections weights will be obtained for this particular test with these particular design setup. Then, we retain these weights, as they are static for the same problem, and apply them to the actual problem at hand to correct the p-values.

A SAS Macro program to obtain the correction weights and apply it to the list of p-values we wish to correct is provided in Appendix.

```
%macro maketruenull(data=nullp, pname=np, precision=1, outof=100, NullWeight=nullweight, AdjustedPValues=newp, newpname=newp,
PriorNullweight=priornulldata);
options nonotes;
data &AdjustedPValues; set &data; run;
%if "&PriorNullweight"="" %then %do;
proc sql; create table _NullWeight as select distinct 0 as pstart, &precision/&outof as pcut,
count(*)/totaln as pdf, count(*)/totaln as cdf from
(select distinct *, count(*) as totaln from &data) where &pname<=%sysevalf(&precision/&outof); quit;
proc sql noprint; select distinct pdf into :cumcdf from _NullWeight; quit;
%do i=&precision %to %sysevalf(&outof-&precision) %by &precision;
%let cw1=%sysevalf(&i/&outof);
%let cw2=%sysevalf((&i+&precision)/&outof);
proc sql; create table midweight as select distinct &cw1 as pstart, &cw2 as pcut, count(*)/totaln as pdf
from (select distinct *, count(*) as totaln from &data) where &cw1<&pname<=&cw2; quit;
data midweight; set midweight; if pdf=. then pdf=0; run;
proc sql noprint; select distinct pdf into :midcdf from midweight; quit;
%let cumcdf=%sysevalf(&cumcdf+&midcdf);
data midweight; set midweight; cdf=&cumcdf; run;
proc append data=midweight base=_NullWeight force; run; proc sql; drop table midweight; quit;
proc sql; create table &NullWeight as select distinct min(pstart) as pstart, max(pcut) as pend,
sum(pdf) as pdf, cdf from _NullWeight group by cdf order by cdf; quit;
%end;
data &NullWeight; set &NullWeight; binid=_n_; run;
data finalnullweight; set &NullWeight; run;
%end;
%else %do; data finalnullweight; set &PriorNullweight; run; %end;
proc sql noprint; select distinct count(*) into :nofbins from finalnullweight; quit;
%do j=1 %to &nofbins;
proc sql noprint; select distinct pstart, pend, cdf into :pstart, :pend, :cdf
from finalnullweight where binid=&j; quit;
%if &j>1 %then %do; proc sql noprint; select distinct cdf into :prevcdf
from finalnullweight where binid=%sysevalf(&j-1); quit; %end; %else %let prevcdf=0;
data &AdjustedPValues; set &AdjustedPValues; if &pstart.<&pname.<=&pend then &newpname=&prevcdf+(&pname-&pstart)/(&pend-&pstart)*
(&cdf-&prevcdf); run;
%end;/**/
options notes;
%put NOTES: Final Weight Data is created as &NullWeight and/or Weight Data &PriorNullweight data have been used;
%put NOTES: Null-weight adjusted p-values are placed in Final Data &AdjustedPValues;
%mend;
*** Sample Run Syntax ***;
/*
data pvalues; do i=1 to 10000; np=rand('beta', 0.5, 0.5); output; end; run;
%maketruenull(data=pvalues, pname=np, precision=1, outof=100, NullWeight=nullweight, AdjustedPValues=newp100, newpname=newp,
PriorNullweight=);
proc sgplot data=newp100; histogram np/transparency=0.50; histogram newp/transparency=0.50; run;
*/
```

Appendix. SAS code for Uniformitization Algorithm

RESULTS

We now present the level of Uniformitization correction in null p-value distributions. We have applied the Uniformitization approach to Wald Test for Survival endpoint with sample size of 20 and follow-up time of 10. The impact of correction is illustrated in Figure 6.

We provide four more examples to illustrate the level of correction in Figure 7.

We clearly see the improvement in the p-value distribution towards its underlying theoretical distribution, Uniform[0,1]. Kocak and Mozhui (2018)used this approach successfully in the application of the Bayesian Test of Periodicity to identify diurnal cyclic genes in the brain.[10]

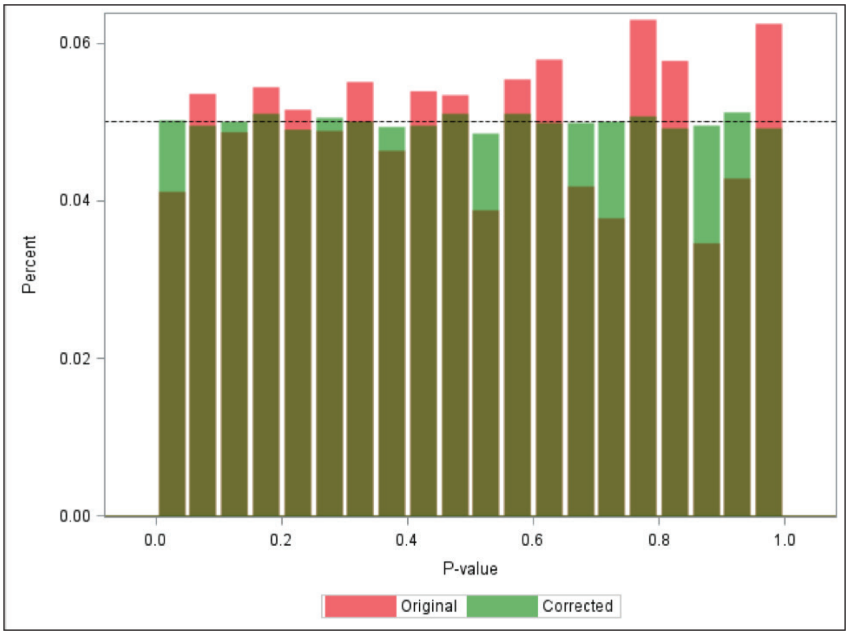

FIGURE 6: Uniformitization correction of Null P-values for Wald Test for survival endpoint (n per arm=20, follow-up=10 unit time).

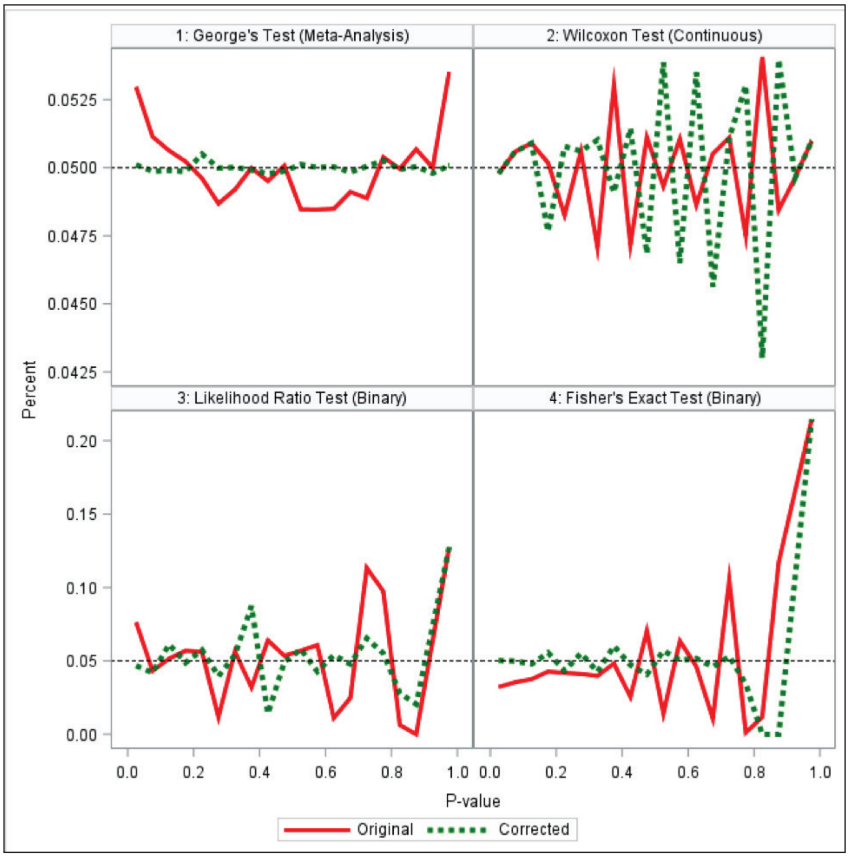

FIGURE 7: Examples of Uniformitization correction (George's test is from the Meta-analysis scenario with N=4, Wilcoxon Test with T-approximation is from the continuous endpoint scenario with N=50, Likelihood Ratio Test was from the binary endpoint scenario with p=0.01 and N=500, and Fisher's Exact test was from the binary endpoint scenario with p=0.2 and N=100).

DISCUSSIONS, CONCLUSIONS, AND FUTURE DIRECTIONS

The null distribution of p-values may be remote from Uniform Distribution, and contrary to the common perception, this may be the case for commonly used tests as well. Even small departure of the p-value distribution of a given test from Uniform[0,1] may result in the issue of under-powering a study, or over-powering the study, both of which are not desirable, and unethical in clinical trials setting. The work by Robins et al. is another example of this issue.[11] Therefore, the clinical trial designer needs to investigate the underlying p-value distribution, especially if non-traditional test statistic is involved, to make sure that Type-1 error rate is retained for the sample size being proposed for the study. Researchers generally form their opinion of the Type-1 error rate retainment based on the large sample asymptotics, which holds true most tests. However, the sample size proposed for a given study due to certain constraints and targeted effect size is small, the behavior of the p-value distribution, thus Type-1 error rate, may not be Uniform[0,1] any longer, and this may potentially under- or over-power the study, the former of which is not acceptable as it informs the scientific community of lack of significance, which is not true, and the latter of which is not acceptable as it indicates that more than needed resources are used.

To help lessen these concerns, we proposed an empirical approach that 'uniformitizes' the pvalue distribution and showed through simulations that it works. If the p-value distribution of a given test is truly Uniform[0.1], even then using the Uniformitization approach does not cause any issue as the correction weights are practically 1.0, as in for the Z-test, for example, which does not need any 'correction' naturally. Such a correction is needed before any procedures like FDR or non-iterative processes such as the one proposed by Netteton et al. so that true signals and false signals can be separated from each other correctly.[12]

It is clear that the null p-value distribution correction is much more critical at the lower-tail of the distribution as traditionally significance levels of 0.01, 0.05, or 0.10 are used in practice. Therefore, the user may choose a much finer grid for correction such as 0 to 1 by 0.001, or even, 0 to 1 by 0.0001. In today's computation power, the task of generating a million random sample or even more in given research design and computing the empirical power is not very costly, especially when we think about the value it brings to the design discussions. Also, it is a process done only once and the correction weights can be used for that particular setting for any future corrections if needed.

We plan to continue this research focusing on the discrete outcome variables as the correction is not ideal due to discreteness and we hope to generate a version of the Uniformitization approach by a way of smoothing. Another natural extension of this is multivariable p-value distributions, which is a case with multivariable regression models. We also plan to study the impact of uniformitization on approaches used for family-wise error rate control. Especially in studies such as genomics, proteomics, metabolomics, the researcher deals with excess amount of p-values coming from relatively smaller samples, and the impact of Uniformitization may be more needed and more pronounced in such applications.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and / or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

This study is entirely author's own work and no other author contribution.

REFERENCES

1. Pearson KX. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 1900;50(302):157-75. [Crossref]
2. Dempster AP, Schatzoff M. Expected significance level as a sensitivity index for test statistics. J Am Stat Assoc. 1965;60(310):420-36. [Crossref]
3. Sackrowitz H, Samuel-Cahn E. P values as random variables-expected P values. Am Stat. 1999;53(4):326-31. [Crossref]
4. Murdoch DJ, Tsai YL, Adcock J. P-values are random variables. Am Stat. 2008;62(3):242-5. [Crossref]
5. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995;57(1):289-300 [Crossref]
6. Fisher RA. Statistical Methods for Research Work. 2nd ed. Edinburg & London: Oliver & Boyd; 1931. p.336.
7. George EO, Mudholkar GS. On the convolution of logistic random variables. Metrika. 1983;30(1):1-13. [Crossref]
8. Kocak M, Zhang G, Narasimhan G, George EO, Pyne S. Differential meta-analysis for testing the relative importance of two competing null hypotheses over multiple experiments. Statistical Genomics in Journal of Indian Society of Agricultural Statistics. 2010;61(1):1-10.
9. Stouffer SA, Suchman EA, DeVinney LC, Star SA, Williams RMJ. The American soldier: adjustment during army life. Am Sociol Rev. 1949;14(4):557-9. [Crossref]
10. Kocak M, Mozhui K. An Application of the Bayesian periodicity test to identify diurnal rhythm genes in the brain. IEEE/ACM Trans Comput Biol Bioinform. 2018 Jul 25. Doi: 10.1109/TCBB.2018.2859971. [Epub ahead of print]. PMID: 30047896. [Crossref] [PubMed]
11. Robins JM, van der Vaart A, Ventura V. Asymptotic distribution of p values in composite null models. J Am Stat Assoc. 2000;95(452):1143-56. [Crossref]
12. Nettleton D, Hwang JG, Caldo RA, Wise RP. Estimating the number of true null hypotheses from a histogram of p values. J Agric Biol Environ Stat. 2006;11(3):337. [Crossref]