Mathematics and Statistics 8(2): 181-186, 2020
DOI: 10.13189/ms.2020.080214 http://www.hrpub.org

The Way of Pooling p-values

Fausto Galetto

Independent Researcher, Polytechnic University of Turin, Italy

Received December 17, 2019; Revised January 17, 2020; Accepted February 7, 2020

Abstract Pooling p-values arises both in practical (in any science and engineering applications) and theoretical (statistical) issues. The p-value (sometimes p value) is a probability used as a ''statistical decision quantity'': in practical applications, it is used to decide if an experimenter has to believe that his/her collected data confirm or disconfirm ''his/her hypothesis'' about the "reality" of a phenomenon. It is a real number, determination of a Random Variable, uniformly distributed, related to the data provided by the measurement of a phenomenon. Almost all statistical software provides p-values when statistical hypotheses are considered, e.g. in Analysis of Variance and regression methods. Combining the p-values from various samples is crucial, because the number of degrees of freedom (df) of the samples we want to combine is influencing our decision: forgetting this can have dangerous consequences. One way of pooling p-values is provided by a formula of Fisher; unfortunately, this method does not consider the number of degrees of freedom. We will show other ways of doing that and we will prove that theory is more important than any formula which does not consider the phenomenon on which we have to decide: the distribution of the Random Variables is fundamental in order to pool data from various samples. Manager, professors and scholars should remember Deming's profound knowledge and Juran's ideas; profound knowledge means "understanding variation (type of variation)" in any process, production or managerial; not understanding variation causes cost of poor quality (more than 80% of sales value) and do not permits a real improvement.

Keywords Statistical Hypotheses, p-value, Random Variable, Distribution, Statistic

1. Introduction

Let's consider a "continuous" probability distribution $F(t)=P[T\leq t]$, T being a random variable. $U=F^{-1}(T)$ is a random variable uniformly distributed: $G(u)=P[U\leq u]$. *p-value is the probability 1-G(u)=P[U>u]*, WHEN F(t) is related to a statistical test THEN the p-value is a result of a statistical test.

The p-value (sometimes p value) [1, 2] is a ''statistical decision quantity'': it is used, in practice, to decide if we have to believe that our data confirm or disconfirm ''our hypothesis'' about the "reality" of a phenomenon.

It is a real number, determination of a Random Variable related to the data provided by the measurement of a phenomenon.

In this section, we provide the general ideas about Hypothesis Testing, while in section 2 we describe the p-values.

We consider here only the parametric version of the method.

We ask the reader to look at figure 1, which shows a typical situation that anybody can be confronted with.

For example, let's assume that we want to decide about the following simple case:

We want to test if people weigh more than 75 kg, assuming that we want to risk only α=5% and assuming that the standard deviation of the weight is 8 kg.

Connect this case to figure 1 which refers to any general setting related "Hypothesis Testing". There you see the (above mentioned) risk α (not its value 5%) related to the Hypothesis H_0 (left unspecified in the case) and the hypothesis H_1={(not shown the value) people weigh (in mean) more than 75 kg}; regarding the Probability Model the information we have is limited to the standard deviation is 8 kg.

Before arriving to the decision based on the data in our hand, we describe the general framework of figure 1.

Any statistical decision is referred to a Probability Model assumed (either on technical or theoretical reasons) to rule the data we are going to collect. The Probability Model is generally specified by the Cumulative Distribution which depends on various "parameters". To explain figure 1, we consider only 1 parameter.

Let π be the parameter we want to "test"; previous to any

collection of data we should state two Hypotheses and a probability α, named the "*significance level*":

1. The "Null Hypothesis", named H_0, where we assume, before any collection of data, a value for the parameter π; we indicate it with the symbol π_0; π_0 is a number, while π is the symbol of the parameter: we write H_0: $[\pi=\pi_0]$
2. The "Alternative Hypothesis", named H_1, where we assume, before any collection of data, *another* value for the parameter π; we indicate it with the symbol π_1; π_1 is a number different from π_0, while π is the symbol of the parameter: we write H_1: $[\pi=\pi_1]$
3. The probability α, the "*significance level*" that we assume, before any collection of data and analysis of the data, is the ''probability that we *accept* of being *wrong* IF we, after the collection and the analysis of the data, claim *the Null Hypothesis* H_0: $[\pi=\pi_0]$ *is Rejected*, when actually (and nobody knows it!) *the Null Hypothesis* H_0: $[\pi=\pi_0]$ *should not be Rejected* ''.

From the points 1, 2, 3, using the Theory, we can, before any collection and analysis of the data, find two items:

- A "formula", named *''Test Statistics'' s,* that will provide us with a number, after the analysis of the data
- And an interval of the real line (real numbers) **C**, named Critical Region (or Rejection Region) such that we Reject "*the Null Hypothesis* H_0: $[\pi=\pi_0]$" IF $s \in \mathbf{C}$.

Let's assume that we collect the data and analyse them, according to the Theory (probability and statistic)**,** and compute the number **s** (computed from the Decision Function); IF $\mathbf{s} \in \mathbf{C}$, then we, according to the Theory, should Reject *the Null Hypothesis* H_0: $[\pi=\pi_0]$; IF $\mathbf{s} \notin \mathbf{C}$, we Accept [but really we do not reject] *the Null Hypothesis* H_0: $[\pi=\pi_0]$.

NOTICE that the **Test Statistics s** is the "determination (=estimate)" of the *estimator* ***S***, which is a Random Variable!

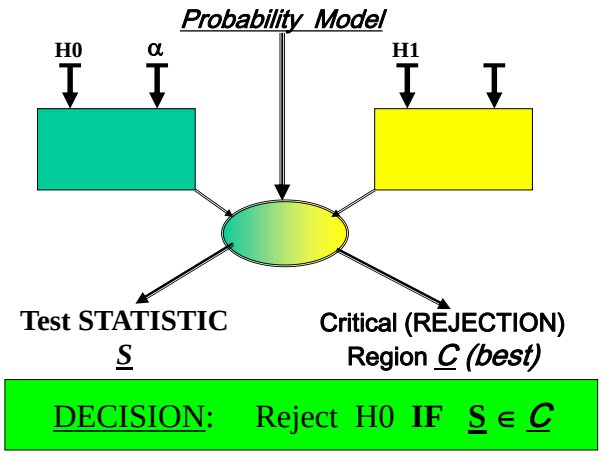

Figure 1. Test of Hypothesis flowchart

In figure 1 it is clearly shown that, in order to take a decision, we need the *Probability Model* (on the top of the figure); the model has to be *suitable* to the analysis of our data for the *parameter* we want to test! Moreover, above the yellow box, you see the Hypothesis H1 and not the risk β; this depends on the fact that in various cases β is not stated and one wants to compute the power of the test 1-β.

For example, in the above case, the parameter we want to test is the "mean weight" π; we have to assume a *Probability Model* depending from the parameter π (and other parameters).

Deming and Juran [3, 4, 5] ("It is a hazard to copy". "It is necessary to understand the theory of what one wishes to do or to make". "Without theory, experience has no meaning". "A figure without a theory tells nothing". «The result is that hundreds of people are learning what is wrong. I make this statement on the basis of experience, seeing every day the devastating effects of incompetent teaching and faulty applications ») and Galetto [6] (Quality of methods for quality is important, *EOQC Conference*, Vienna, 1989) have been very clear about the need of being scientific in Quality Management and decisions. J. Juran praised F. Galetto's paper during his presentation at Vienna Conference.

Hence now let formally define, in our case, what we need for arriving to the scientific decision:

- "*Null Hypothesis*" H_0:{people weigh (in mean) 75 kg=π_0}
- "*Alternative Hypothesis*" H_1:{people weigh (in mean) more than 75 kg}=$[\pi=\pi_1>\pi_0]$; it is named ''one-tailed alternative hypothesis''. Notice that many times it can be ''two-tailed alternative hypothesis''
- "*significance level*"=α=0.05
- *Probability Model* limited to the standard deviation σ=8 kg.

According to the Theory of Probability and Statistics/Reliability [7-16] we need more information; we set it as follows:

- *Probability Model of the data*, Normal distribution with mean μ and standard deviation σ=8 kg, that is variance σ^2=64 kg^2, N(μ, σ^2) [generally the "true" mean is indicated by μ; in our case is π]
- *Number of items to be measured* (*sample size*): we decide n=25

Notice that the *sample size* n can be *computed* when both risks α (related to H_0) and β (related to H_1) are stated. If β is not stated as in figure 1, the sample size is defined by the experimenter.

From these last two points, according to the Theory, anybody can derive that the *Probability Model for decision*; since we assumed N(μ, σ^2) as *Probability Model of the data*, in figure 1 the *Probability Model for decision* is the Normal distribution with mean μ and variance σ^2/n, N(μ,σ^2/n).

According to the Theory of Probability and Statistics/Reliability [7-16], we know that the test statistic is the mean of the data, indicated with $\bar{x}$, which is the

determination of Random Variable Mean $\bar{X}$.

Then we derive the Critical Region which is $C = \{\bar{x} > \mu + z_\alpha \sigma/\sqrt{n}\}$.

In this case (notice, "only in this case"), we have z_α=1.645; then the Critical Region is C={>77.632}

Now we are allowed to collect 25 data.

IF $\bar{x} = 79.25$ then $\bar{x} \in$C and, in this case, we reject H_0 with 5% significance level.

2. p-values

In section 1, we saw the general ideas of Testing of Hypothesis (see figure1). The decision there depends on the Test Statistic s and on the Critical Region C which is related to the *risk* α (*significance level*): Test Statistic s and Critical Region C were the ''two statistical decision quantities''. The Test Statistic s is a real number, determination of a Random Variable S, depending on the Probability Model.

Now we see a different ''statistical decision quantity'': the p-value (sometimes p value) [1, 2]. It is used to decide if we have to believe that our data confirm or disconfirm ''our hypothesis'' about the "reality" of a phenomenon. It is a real number, determination of a Random Variable related to the data provided by the measurement of a phenomenon.

For the same purpose almost every statistical software provides the ***p*-value** (**probability value**).

In our example, it is the probability that the Random Variable Mean (indicated generally with $\bar{X}$) is greater than the empirical (observed) mean $\bar{x} = 79.25$, that is

$$P[\bar{X}>79.25|\ \sigma_{\bar{X}}^2 = \sigma^2/n].$$

From the example we derive the definition of the ***p*-value**: it is the probability, under the *Null Hypothesis* H_0 (opposed to the *Alternative Hypothesis* H_1 about the distribution of the random variable), that the variate (RV) to be observed takes values equal to or more extreme than the value observed.

The p-value can be viewed as an index of the "strength of evidence" against the *Null Hypothesis* H_0; a small p-value indicates an unlikely event and, hence, an unlikely hypothesis.

Let f(x) be the pdf of the *estimator T* and t the determination of *T*; after the elaboration of the collected data, we can compute the integral

$$\int_t^\infty f(x)\,dx = \pi \qquad (1)$$

which is a real number.

The value of the integral π is the determination of the Random Variable **Π**, related to the **π** "*parameter p-value*".

We find, in our example, from (1), find $P[\bar{X}>79.25|\ \sigma_{\bar{X}}^2 = \sigma^2/n]$=0.00395=p-value (*estimation* of the parameter p-value π), well below the 5% significance level, that is "strong evidence" against H_0.

Many "professionals" think that this p-value ("strength of evidence") is much more informative than α.

The author thinks the opposite.

Let's see why.

IF, with other 25 data, $\bar{x} = 77.75$ then $\bar{x} \in$ C and, AGAIN in this new case, we reject H_0 with 5% significance level, same as before. BUT, from (1), the new p-value (*estimation* of the same parameter p-value π) is 0.0428: the "strength of evidence" is less than before…

What is the <u>*true p-value for*</u> H_0={people weigh (in mean) 75 kg=π_0} <u>*based on both samples*</u>? We do not know.

What is the <u>*true significance level*</u>? α=0.05 for both decisions!

Notice that the example used the normal distribution, and the decision function depended on that.

For example, the same <u>*numerical*</u> procedure <u>*cannot*</u> be used (but Six Sigma Professionals do not know it) for the following case:

1. You have 10 atoms: 5 disintegrate and 5 do not disintegrate.
2. You have 100 atoms: 5 disintegrate and 95 do not disintegrate.

Set α=0.05; how can you test H_0?

The same is for "people dying" (life insurance)!

In the next section we see how to combine two computed p-values.

3. Pooling p-values

To explain the point, we consider the data of table 1, drawn form a reliability test on the same type of systems.

They are times to failure (hours) of 5 items in sample 1 (sample size 13) and on 10 items in sample 2 (sample size 17).

It is known that the exponential distribution [7-16] is suitable for the RV "TTF (Time To Failure)".

According to figure 1, we set H_0=[μ_0<100 h] where μ is the MTTF (Mean Time To Failure of the items) and α=0.05; the probability model depends on the exponential distribution; the alternative hypothesis is H_1=[μ_1>100 h].

We have to find the test statistic and the Critical Region.

Table 1. Data collected from reliability tests

sample 1	5.5	26.4	64.4	72.3	84.1	84.1	84.1
	84.1	84.1	84.1	84.1	84.1	84.1	
sample 2	17.4	29.1	30.2	46.6	60.3	61.4	76.4
	87.1	98.3	126.8	126.8	126.8	126.8	126.8
	126.8	126.8	126.8				

Consider firstly sample 1; there are 5 failures and 13 items on test.

According [7-16], the test statistics is the "Total Time On Test" t_1, the determination of the RV $T_1(t)$; "Total Time On Test" is the sum of all the data of the items on test. In this case, the *Probability Model* for decision f(x), in figure 1, is the Erlang distribution with mean 5μ, and the decision is ''Reject H_0'', because the Critical Region is C={>915.3} and t_1=925.3.

The p-value is $0.047=\int_{925.3}^{\infty} f(x)\,dx$. This confirms that we have to ''Reject H_0'', because 0.047<0.05

The "*Total Time on Test" $T(t_g)$* based on g failures is a Random Variable; then the integral

$$\int_{\mathbf{T}(t_g)}^{\infty} f(x)\,dx = \tilde{\Pi} \qquad (2)$$

is a Random Variable.

We call $\tilde{\Pi}$ the "Random Variable p-value".

The p-value 0.047 is the estimate of the π "*parameter p-value*".

Consider secondly sample 2; there are 10 failures and 17 items on test.

According [7-16], the test statistics is again the "Total Time On Test" t_2, the determination of the RV $T_2(t)$; in this case, the *Probability Model* for decision f(x), in figure 1, is the Erlang distribution with mean 10μ, and the decision is ''Accept H_0'', because the Critical Region is C={>1570.5} and t_2=1521.5.

The p-value is $0.063=\int_{1521.5}^{\infty} f(x)\,dx$. This confirms that we have to ''Accept H_0'', because 0.063>0.05

We have two contradictory decisions with risk 5%, because different samples (as obvious) provide different data and any statistical analysis depends on the data.

Notice that we used the same value α=0.05 for both tests.

Since we have two independent comparisons, we have to consider the Bonferroni Correction: in order to avoid a lot of spurious positives, the alfa value needs to be lowered to account for the number of comparisons being performed. In this case, we have two comparisons so that $\alpha_{modified}$=0.025. With this modification, we have to ''Accept H_0'', because 0.047>0.025 and 0.063>0.025; with this ideas the "two decisions" are no longer contradictory.

To decide which of the "two (contradictory) decisions" could be the right one we think to test a new hypothesis H_0^B=[$MTTF_1$=$MTTF_2$] with α=0.05, versus H_0^B=[$MTTF_1$≠$MTTF_2$].

According [7-16] the test statistics is the "Ratio of the Total Time On Test t_1/t_2,", the determination of the RV $T_1(t)/T_2(t)$; in this case the *Probability Model* for decision f(x), in figure 1, is the F distribution [with df=10 for numerator and df=20 for denominator], and the decision is ''Accept H_0^B'', because the Critical Region is C={>2.35} and t_1/t_2=1.22.

The p-value is $0.338=\int_{1.22}^{\infty} f(x)\,dx$. This confirms that we have to ''Accept H_0^B'', because 0.338>0.05

We are again in the fog!

To get out and see the sun, let's decide to pool the two samples and therefore to pool the two p-values.

How can we do it?

Let's follow Fisher's method [2]…

Fisher's method combines extreme value probabilities from each test, p-values, into one test statistic (X^2, chi square) using the formula

$$X_{2k}^2 = -2\sum_1^k \ln(p_i) \qquad (3)$$

where p-value for the i^{th} hypothesis test and k is the number of tests being combined.

According to Fisher's method: p1=0.047 and p2=0.063

The chi²statistic is the number X²=-2*[ln(0.047)+ln(0.063)]=11.64. Under the null hypothesis H_0=[μ_0<100 h], the related RV statistic is chi²distributed with 4 df (degrees of freedom), so the combined p-value is 0.02 < 0.05.

Hence we should ''Reject H_0''…. The two samples (combined) tell us that the "true mean" should be >100 h.

Let's now follow the Theory given in [7-16].

According [7-16], the test statistics is the "Total Time On Test" t_1+t_2, the determination of the RV $T_1(t)+T_2(t)$; in this case the *Probability Model* for decision f(x), in figure 1, is again the Erlang distribution with mean 15μ, and the decision is ''Reject H_0'', because the Critical Region is C={>2168.65} and t_1+t_2=2446.70.

The p-value is $0.016=\int_{2445.70}^{\infty} f(x)\,dx$. This confirms that we have to '' Reject H_0'', because 0.016<0.05

We got the same decision:

the "true" MTTF, based on the combined sample, is > 100 h, both with the Fisher's method [p-value=0.020] and with the Theory [p-value=0.016].

However, the two p-values are different!

Would always the decision be the same?

We do not know. It is probable that it should be not.

What can we do, then?

Use the Theory, suitable for the data we collected in our samples.

4. Comparison of "Pooling p-values" and Theory

In this section, we test a new "*Null Hypothesis*" H_0={Fisher's method and Reliability Theory provide the same decisions about combining two samples as in § 2, based on exponential distributed data, with 5 failures out of 13 items and with 10 failures out of 17 items}.

The dfs are important both for exponential data and for normal data. We consider these two cases.

Fisher's method misses this important fact about dfs (figure 2, taken from [3]).

Using Fisher's method, two small p-values P_1 and P_2 combine to form a smaller p-value. The yellow-green boundary defines the region where the "combined" p-value is below 0.05. For example, if both p-values are around 0.10, or if one is around 0.04 and one is around 0.25, the "combined" p-value is around 0.05.

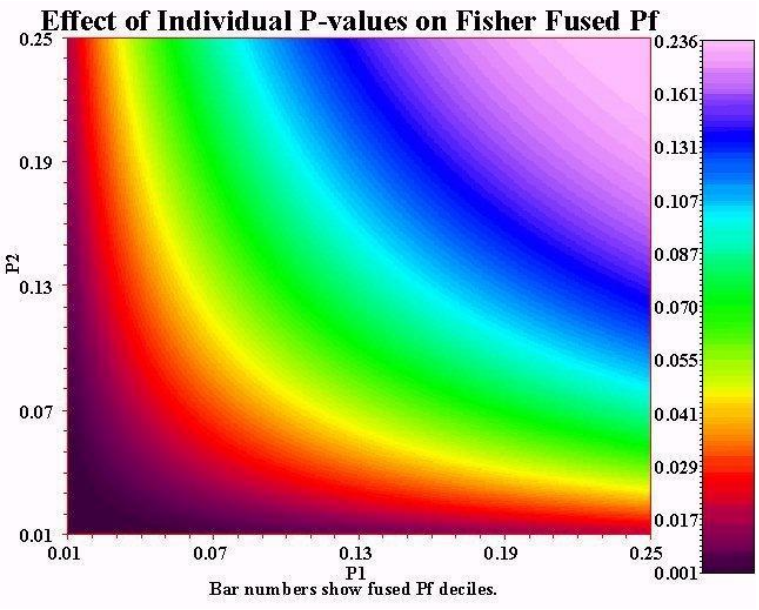

Figure 2. Combining two p-values with the Fisher's method

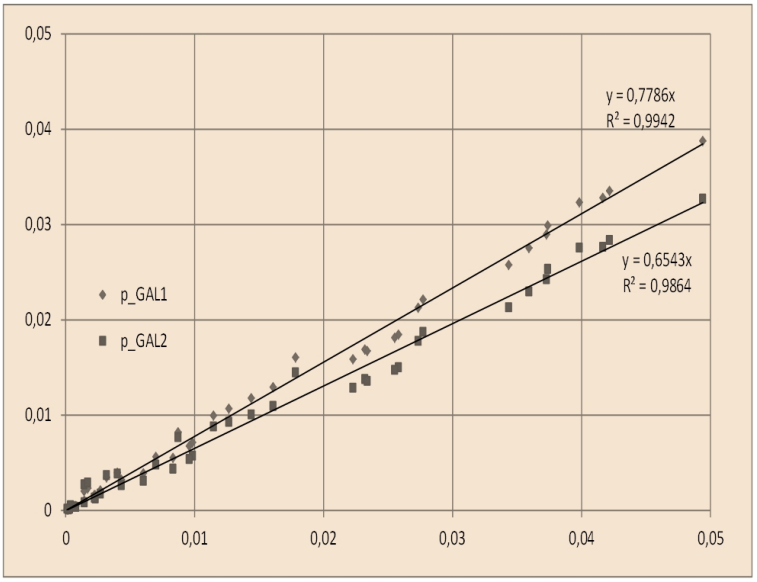

Figure 3. Comparison of Fisher's p-value with Theory for Exponential data (blue points) and Normal data (red points)

The comparison of Fisher's method and the Theory [7-16] is shown in figure 3.

It is clear that Fisher's p-value is higher than the value computed from Theory, taking into account the dfs!

That means wrong decisions depending on the data…

We *should reject the* "*Null Hypothesis*" H_0={Fisher's method and Reliability Theory provide the same decisions about combining two samples as in § 3, based on *exponentially* distributed data, with 5 failures out of 13 items and with 10 failures out of 17 items} and we *should reject the* "*Null Hypothesis*" H_0={Fisher's method and Reliability Theory provide the same decisions about combining two samples as in § 3, based on *Normally* distributed data, with 5 failures out of 13 items and with 10 failures out of 17 items}.

The same for others distributions of the data and different numbers of degrees of freedom.

This is in line with W. E. Deming's "*profound knowledge*" [3, 4] and Juran's ideas [5].

5. Conclusions

We saw that Theory is more important than formulae which do not consider the phenomenon on which we have to decide (distribution of the data).

The number of degrees of freedom of the samples we want to combine is influencing our decision: *forgetting this can have dangerous consequences*.

Manager, professors and scholars should remember

Deming's [3, 4] and Juran's ideas [5], because "Quality of Methods for Quality is important [6].

REFERENCES

[1] Montgomery D. C., 2009, 6th edition, Introduction to Statistical Quality Control, Wiley & Sons

[2] Wikipedia https://en.wikipedia.org/wiki/Fisher%27s_method

[3] Deming W. E., 1986, Out of the Crisis, Cambridge University Press.

[4] Deming W. E., 1997, The new economics for industry, government, education, Cambridge University Press.

[5] Juran, J., 1988, Quality Control Handbook, 4th ed, McGraw-Hill, New York.

[6] Galetto, F., (1989) Quality of Methods for Quality is important, EOQC Conference, Vienna.

[7] Galetto, F., AFFIDABILITÀ vol. 1 Teoria e Metodi di calcolo, CLEUP editore, Padova, 1981, 84, 87, 94.

[8] Galetto, F., AFFIDABILITÀ vol. 2 Prove di affidabilità: distribuzione incognita, distribuzione esponenziale, CLEUP editore, Padova, 1982, 85, 94.

[9] Galetto, F., Qualità Alcuni metodi statistici da Manager, CUSL, 1995/7/9.

[10] Galetto, F., Gestione Manageriale della Affidabilità CLUT, Torino, 2010.

[11] Galetto, F., Manutenzione e Affidabilità CLUT, Torino, 2015

[12] Galetto, F., 2016, Reliability and Maintenance, Scientific Methods, Practical Approach, Vol-1, www.morebooks.de.

[13] Galetto, F., 2016, Reliability and Maintenance, Scientific Methods, Practical Approach, Vol-2, www.morebooks.de.

[14] Galetto, F., 2016, Design of Experiments and Decisions, Scientific Methods, Practical Approach, www.morebooks.de.

[15] Galetto, F., 2017, The Six Sigma HOAX versus the versus the Golden Integral Quality Approach LEGACY, www.morebooks.de.

[16] Galetto, F., 2018, Quality and Quality Function Deployment, a Scientific Analysis, Lambert Academic Publishing ISBN 978-613-9-90898-1