Boletín de Estadística e Investigación Operativa
Vol. 25, No. 1, Febrero 2009, pp. 4-13

ESTADÍSTICA

Give q-value a chance

Julián de la Horra
Departamento de Matemáticas, Facultad de Ciencias
Universidad Autónoma de Madrid
✉ julian.delahorra@uam.es

Abstract

The p-value is, possibly, the most used tool in applied statistics, although it has some important shortcomings. The q-value was introduced by Storey (2003) as a byproduct of the positive false discovery rate. In this article, it is shown that the q-value could be considered as a good alternative to the p-value, because the q-value exhibits better properties than the p-value and, moreover, the q-value is easier to interpret than the p-value.

Keywords: Coherence, hypothesis testing, positive false discovery rate, p-value, q-value.
AMS Subject classifications: 62F03, 62F15.

1. Introduction

The p-value is, possibly, the most used tool in applied statistics since 1925, when R. A. Fisher published his *Statistical Methods for Research Workers*. The p-value is automatically given by all the statistical packages, and it is very easy to use for taking a decision in hypothesis testing. These are the main reasons for its popularity.

Nevertheless, to the same time, the p-value is one the most seriously criticized notions in statistics. This is because of its shortcomings. Some of the most important discussion papers on the p-value are Cox (1977), Shafer (1982), Berger and Delampady (1987), Berger and Sellke (1987) and Casella and Berger (1987).

The p-value is shortly explained in Section 2, and some of its main shortcomings are shown in Section 3, by means of two simple examples.

The concept of q-value was introduced by Storey (2003) as a byproduct of the positive false discovery rate, but it is interesting by itself because the q-value exhibits better properties than the p-value and, moreover, the q-value is easier to interpret than the p-value. The notion of q-value is explained and analyzed in Section 4. Finally, the main conclusions are given in Section 5.

2. The p-value

Let us consider a random sample $(X_1, ..., X_n)$ from an amount of interest in a population, X, with a probability density $f(x|\theta)$, where $\theta \in \Theta$ is an unknown parameter. In the simplest situation, we want to test the null hypothesis $H_0 : \theta = \theta_0$ versus the alternative hypothesis $H_1 : \theta = \theta_1$. For doing that, we have to choose a test statistic, $T = T(X_1, ..., X_n)$. For the sake of simplicity, we will assume that this test statistic is measuring (in some suitable way) the discrepancy between the sample we have obtained, $(x_1, ..., x_n)$, and the null hypothesis: the larger is the value of the test statistic the larger is the discrepancy between the sample and the null hypothesis.

When the sampling has been carried out, and a sample $(x_1, ..., x_n)$ has been obtained, we will denote by t the current value of the test statistic, that is, $t = T(x_1, ..., x_n)$.

The p-value for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, when $T = t$ has been obtained, is defined as

$$p\text{-}value_{\theta_0}(t) = Pr(T \geq t|\theta_0).$$

Of course, the p-value can also be defined when we are considering a general hypothesis testing. So, let us assume that we want to test the null hypothesis $H_0 : \theta \in \Theta_0$ versus the alternative hypothesis $H_1 : \theta \in \Theta_1 = \Theta - \Theta_0$.

In this case, the p-value is usually defined as

$$p\text{-}value_{\Theta_0}(t) = \sup_{\theta \in \Theta_0} Pr(T \geq t|\theta).$$

From a technical viewpoint, the p-value may be meant as "the probability of obtaining a sampling result less compatible with H_0 than the current sample, given that the null hypothesis is true". When the concept of significance level has been previously introduced, the p-value may also be meant as "the smallest value of the significance level for which the sample we have obtained will lead to rejection of H_0". The problem with these phrases is that, although technically correct, they are rather complicated to understand for users. I have always suspected that, for users, these phrases are like a famous phrase in a film by Marx Brothers: "the party of the first part shall be known in this contract as the party of the first part". Despite this, the p-value is a very popular tool for users. This generalized use among users has been possible because of three reasons:

1. First of all, simple interpretations, suitable for users, were given. One of the most usual interpretations of the p-value is that "it measures the degree to which the data $(x_1, ..., x_n)$ support the null hypothesis H_0". See, for instance, Lehmann (1975; p. 11).

2. Then, a simple rule for obtaining a decision through the p-value was given. This rule is as follows.

   First, we choose a significance level, say α=0.05, and then:

   (a) When the p-value is less than 0.05, we say that a significant sampling result against H_0 has been obtained, and we have found evidence enough for rejecting H_0.

   (b) When the p-value is greater than 0.05, we say that the sampling result is not significant against H_0, and we have not found evidence enough for rejecting H_0.

   Of course, if you have another favorite value for α, you can use it instead of 0.05.

3. Finally, the p-value is automatically given by all the statistical packages.

3. Some shortcomings of the p-value

All we have said in Section 2 is now applied to the following simple example:

Example 3.1. *Let us consider an observation X from a Normal distribution, $N(\theta;1)$. We want to test $H_0 : \theta = 0$ versus $H_1 : \theta = 1$. In this case, we usually take $T = X$ as test statistic. Let us suppose that the value $t = x = 1$ has been observed. We can easily compute the p-value:*

$$p\text{-}value_{\theta_0}(1) = Pr(T \geq 1 | \theta = 0) = Pr(N(0;1) \geq 1) = 0.1587 \simeq 0.16.$$

A p-value of 0.16 is usually considered as a very large p-value (much larger than α=0.05). As a consequence, we would say that the degree of support to H_0 is enough, the sampling result is not significant against H_0, and there is no evidence enough for rejecting H_0.

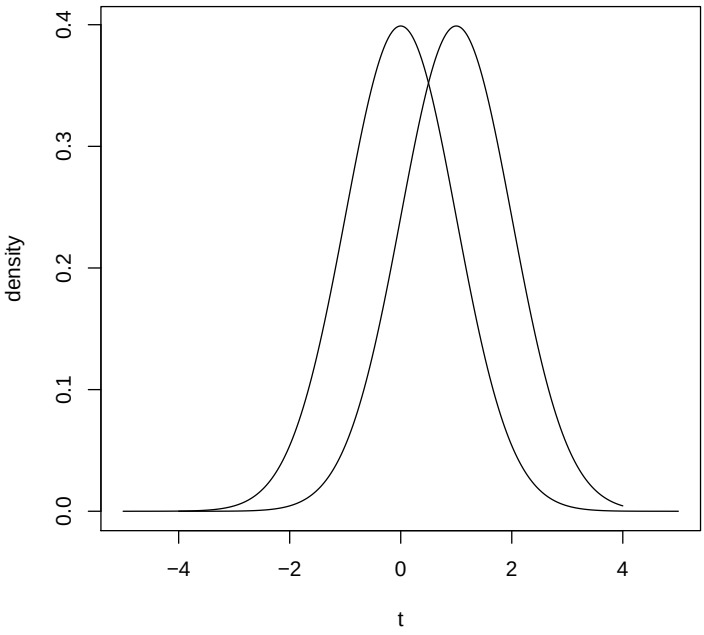
Comments on Example 3.1

It is very interesting to point out some aspects of this example:

1. The p-value ranges in the interval [0, 1], because it is a probability. The value 0.16 is nearer to 0 than to 1. So, why are we saying that 0.16 is usually considered as a very large p-value? The usual answer is that we are calibrating the p-value through a comparison to the significance level. This answer leads to other questions: how to choose the significance level?; why $\alpha = 0.05$?

2. If the p-value is 0.16, it is usually considered that the degree of support to H_0 is enough. But, what are we understanding by support to H_0? More on this, later.

3. Last but not least. If we simply apply our common sense to the example and we have a look at the two densities, $N(\theta = 0; 1)$ and $N(\theta = 1; 1)$, we obviously conclude that the sampling result, $t = 1$, is much more compatible with $N(\theta = 1; 1)$ (H_1) than with $N(\theta = 0; 1)$ (H_0); see Figure 1. Nevertheless, we have said in the example that there is no evidence enough for rejecting H_0! This conclusion seems astonishing. The usual explanation is that we want to be conservative with the null hypothesis. This answer leads to other questions: do we want to be conservative in all the problems?; how conservative do we want to be?



Sampling densities under H_0 and H_1

Now, let us come back to the question in the second comment above: what are we understanding by support to H_0?

As we have said before, the most usual interpretation of the p-value is that it measures "the degree to which the data $(x_1, ..., x_n)$ support the null hypothesis H_0". There is only a little problem: the p-value is not coherent as measure of support for the null hypothesis. This was pointed out by Schervish (1996) in a beautiful paper. His main ideas are next summarized. First of all, we give the definition of *coherence* as it was stated by Schervish (1996), following the idea introduced by Gabriel (1969) for multiple comparisons:

A measure of support for hypotheses is coherent if, when $\Theta_0 \subset \Theta_0'$, *the support given to* Θ_0' *is greater than or equal to the support given to* Θ_0.

P-values sometimes behave coherently and sometimes do not. A simple example is given by Schervish (1996):

Example 3.2. *Let* X *be an observation from a Normal distribution* $N(\theta; 1)$, *where* $\theta \in \Re$. *Let us assume that the value* $x = 2.18$ *is obtained. Now, let us consider two hypotheses tests:*

$H_0 : \theta \in \Theta_0 = (-0.5,\ 0.5)$ *versus* $H_1 : \theta \notin \Theta_0$;

$H_0' : \theta \in \Theta_0' = (-0.82,\ 0.52)$ *versus* $H_1' : \theta \notin \Theta_0'$.

It is easily obtained (see Schervish (1996)) that $p\text{-}value_{\Theta_0}(2.18) = 0.0502 > p\text{-}value_{\Theta_0'}(2.18) = 0.0498$, *although* $\Theta_0 \subset \Theta_0'$.

Therefore, the p-value cannot be understood as a measure of support for the null hypothesis. Of course, as we have said before, the p-value is a very popular tool in applied statistics, but it must be understood only as a warning: when the p-value is close to zero, possibly the null hypothesis must be rejected.

The p-value is not the only usual concept that cannot be used as a measure of support for the null hypothesis. The same shortcoming was found for the Bayes factor [see Lavine and Schervish (1999)] and for the posterior predictive p-value [see De la Horra and Rodríguez-Bernal (2001)].

We can now summarize the main shortcomings of the p-value:

1. The exact technical meaning of the p-value is quite complicated for being understood by users and, by this reason, it is replaced by a simple interpretation, suitable for users: "the p-value measures the degree to which the data $(x_1, ..., x_n)$ support the null hypothesis H_0". There is only a little problem: the p-value is not coherent as measure of support for the null hypothesis.

2. The p-value must be calibrated for taking a decision about the null hypothesis. This calibration is usually made by comparing the p-value to the significance level. The main problem with the significance level is that we have to choose it in an absolutely arbitrary way.

3. As we have seen in Example 3.1, sometimes, the decision we take from the p-value violates the common sense. The usual explanation is that we want to be conservative with the null hypothesis and, again, the significance level is used in a magical way.

So far, the conclusion seems to be that the p-value has been very used in hypothesis testing because, although it has some important shortcomings, it is

very easy to obtain (through a statistical package) and very easy to use (through a comparison to the significance level), and these features are very important for users in applied statistics.

In the next section, we argue that we could give a chance to a related and easy concept: the q-value.

4. The q-value

The concept of q-value was introduced by Storey (2003) as a byproduct of the positive false discovery rate. The positive false discovery rate is an alternative concept to the family wise error rate, which is the most commonly controlled quantity when multiple hypothesis testing is considered. In the last years, several rates have been introduced and studied for the joint evaluation of multiple hypothesis testing, when a large (or even huge) number of hypotheses are simultaneously considered. As we have said above, the q-value arose from the positive false discovery rate, but it has interest by itself. In fact, the concept of positive false discovery rate is not needed for defining the q-value. The idea of the q-value is next explained.

Let us consider again a random sample $(X_1, ..., X_n)$ from an amount of interest in a population, X, with a probability density $f(x|\theta)$, where $\theta \in \Theta$ is an unknown parameter. In the simplest situation, we want to test the null hypothesis $H_0 : \theta = \theta_0$ versus the alternative hypothesis $H_1 : \theta = \theta_1$. For doing that, we have to choose a test statistic, $T = T(X_1, ..., X_n)$. For the sake of simplicity, we will assume that this test statistic is measuring (in some suitable way) the discrepancy between the sample we have obtained, $(x_1, ..., x_n)$, and the null hypothesis: the larger is the value of the test statistic the larger is the discrepancy between the sample and the null hypothesis. Now, we need to give prior probabilities to θ_0 and θ_1: $\pi_0 = Pr(\theta_0)$ and $\pi_1 = 1 - \pi_0 = Pr(\theta_1)$. These probabilities summarize the confidence we have on the two values of the parameter, before the sample has been obtained. One of the easiest possibilities is to give the same probability to θ_0 and θ_1: $\pi_0 = \pi_1 = 1/2$. This would be suitable when our initial confidences on θ_0 and θ_1 are (approximately) similar.

When the sampling has been carried out, and a sample $(x_1, ..., x_n)$ has been obtained, we will denote by t the current value of the test statistic, that is, $t = T(x_1, ..., x_n)$.

The q-value for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ is defined as

$$q\text{-}value_{\theta_0}(t) = Pr(\theta_0|T \geq t) = \frac{\pi_0 Pr(T \geq t|\theta_0)}{\pi_0 Pr(T \geq t|\theta_0) + \pi_1 Pr(T \geq t|\theta_1)} \ .$$

Of course, the q-value can also be defined when we are considering a general hypothesis testing. So, let us assume that we want to test the null hypothesis $H_0 : \theta \in \Theta_0$ versus the alternative hypothesis $H_1 : \theta \in \Theta_1 = \Theta - \Theta_0$. In

this case, we need to give a prior density, $\pi(\theta)$, probability density over Θ. This prior density summarizes the confidence we have on the different values of the parameter, before the sample has been obtained. One of the simplest possibilities is to use noninformative priors. This would be suitable when our initial confidence on the values of the parameter is not very precise.

In this case, the q-value is defined as

$$\begin{aligned} q\text{-}value_{\Theta_0}(t) &= Pr(\Theta_0|T \geq t) = \frac{Pr(\Theta_0; T \geq t)}{Pr(T \geq t)} \\ &= \frac{\int_{\Theta_0} \left[\int_{\{T \geq t\}} f(t|\theta)\pi(\theta)dt\right] d\theta}{\int_{\Theta} \left[\int_{\{T \geq t\}} f(t|\theta)\pi(\theta)dt\right] d\theta} = \frac{\int_{\Theta_0} Pr(T \geq t|\theta)\pi(\theta)d\theta}{\int_{\Theta} Pr(T \geq t|\theta)\pi(\theta)d\theta} \ . \end{aligned}$$

From a technical viewpoint, the q-value may be meant as "the posterior probability of H_0, given that sampling results less compatible with H_0 than the current sample would be obtained". Of course, this phrase is not easy to understand for users but, in this case, the solution is very simple and it has two important advantages:

1. It is not necessary to interpret the definition. We only need to shorten it: the "q-value is an updated probability of the null hypothesis taking into account the sampling result we have obtained". This is very easy to understand for users.

2. A q-value does not need calibration: we may directly determine if this updated probability of the null hypothesis is sufficiently large. We only need to calibrate our initial confidence on the values of the parameter. Of course, this is not easy to do, but we always may use a noninformative prior.

Let us consider again Example 3.1.

Example 4.1. *(Example 3.1 continued)*

Let us consider again an observation X from a Normal distribution, $N(\theta; 1)$. We want again to test $H_0 : \theta = 0$ versus $H_1 : \theta = 1$. Now, we need to elicit prior probabilities. For instance, let us take $\pi_0 = \pi_1 = 0.5$. We take again, as test statistic, $T = X$. If $t = x = 1$ has been observed, we have:

$$\begin{aligned} q\text{-}value_{\theta_0}(1) &= Pr(\theta_0|T \geq 1) = \frac{\pi_0 Pr(T \geq 1|\theta = 0)}{\pi_0 Pr(T \geq 1|\theta = 0) + \pi_1 Pr(T \geq 1|\theta = 1)} \\ &= \frac{(0.5)Pr(N(0;1) \geq 1)}{(0.5)Pr(N(0;1) \geq 1) + (0.5)Pr(N(1;1) \geq 1)} \\ &= \frac{(0.5)(0.1587)}{(0.5)(0.1587) + (0.5)(0.5)} = 0.2409 \simeq 0.24 \ . \end{aligned}$$

In fact, we can compute the q-value, in the same way, for any value of π_0. Table 1 shows the q-values for different prior probabilities.

Table 1: q-values for different prior probabilities.

π_0	0.4	0.5	0.6	0.7	0.8
q-value	0.17	0.24	0.32	0.43	0.56

Comments on Table 1

It is important to point out some aspects of Table 1:

As we have said before, the q-value is the updated probability of the null hypothesis, where this probability has been updated taking into account the sampling result we have obtained. This updated probability may take different values, depending on the prior probability we give to the null hypothesis. The prior probability on H_0 may be interpreted, in a broad sense, as the initial confidence the user has on the null hypothesis. For instance, we can analyze the two following possibilities:

1. If the initial confidences the user has on the null and the alternative hypotheses are (approximately) equal, we can take $\pi_0 = \pi_1 = 0.5$. In this case, the q-value is 0.24 (see Table 1). If this were the case, the updated probability of H_0 would be small and, possibly, the most reasonable decision would be to reject H_0 and, therefore, to accept H_1. This decision agrees with our common sense: the sampling result $t = 1$ is more compatible with H_1 than with H_0.

2. On the other hand, if the initial confidence the user has on the null hypothesis is much larger than his/her initial confidence on the alternative hypothesis, we could take, for instance, $\pi_0 = 0.8$. In this case, the q-value is 0.56 (see Table 1). If this were the case, the updated probability of H_0 would be large and, possibly, the most reasonable decision would be do not reject H_0. This is the practical situation in those cases in which the user want to be conservative with the null hypothesis: his/her initial confidence on H_0 is large.

From a more sophisticated viewpoint, we could consider, not only prior probabilities, but also a loss function. But the point in this paper is that it is possible to replace the p-value for a similar and better tool: the q-value.

Finally, what about coherence of the q-value? In other words, is it possible to use the q-value as a measure of support for the null hypothesis?

The answer is very clear: it is really easy to prove that q-values are coherent, provided that the same test statistic is used. Let us consider the hypotheses

tests $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \notin \Theta_0$ and $H_0' : \theta \in \Theta_0'$ versus $H_1' : \theta \notin \Theta_0'$, where $\Theta_0 \subset \Theta_0'$. Then:

$$\text{q-value}_{\Theta_0}(t) = Pr(\Theta_0|T \geq t) \leq Pr(\Theta_0'|T \geq t) = \text{q-value}_{\Theta_0'}(t) \ .$$

5. Conclusions

1. The technical definition of the p-value is rather difficult to understand for users. This problem does not arise with the q-value because the q-value is, simply, an updated probability of the null hypothesis.

2. The q-value do not need calibration because, as we have just said, it is a posterior probability of the null hypothesis. The initial confidence of the user on the null hypothesis can be incorporated through the prior distribution.

3. Of course, we need to elicit the prior distribution on the values of the parameter, in a suitable way, and this is not always easy to do, although I think that it is easier to calibrate the initial confidence on the values of the parameter than to calibrate the p-value. Notice that a possible solution is to use a noninformative prior.

4. Q-values may be used as a measure of support for the null hypothesis (if you want to do it) because they are coherent, provided that the same test statistic is used, although this property is less necessary for q-values than for p-values. Notice that p-values are often interpreted as measures of support for H_0, because its exact meaning is difficult to understand for users, and this problem does not arise with q-values.

5. Finally, what about computation? This question is very important because one of the main advantages of the p-value is that it is given by all the statistical packages in an automatic way. I am not a consummated expert in computation, but I am sure that it would be easy to add this possibility for the most usual sampling models. The user only would have to choose his/her prior in a menu including the noninformative prior.

6. Final recommendation: I know that several practical problems must be solved but, please, give q-value a chance.

Acknowledgements

This article is dedicated to Professor Sixto Ríos, my PhD supervisor many years ago. The research has been partially supported by Grants MTM2007-66632 and CCG07-UAM/ESP-1761 (Spain).

References

[1] Berger, J. and Delampady, M. (1987). Testing precise hypothesis (with discussion). *Statist. Sci.* **2**, 317-352.

[2] Berger, J. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of p-values and evidence (with discussion). *J. Amer. Statist. Assoc.* **82**, 112-122.

[3] Casella, G. and Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problems (with discussion). *J. Amer. Statist. Assoc.* **82**, 106-111.

[4] Cox, D. R. (1977). The role of significance tests (with discussion). *Scand. J. Statist.* **4**, 49-70.

[5] De la Horra, J. and Rodríguez-Bernal, M. T. (2001). Posterior predictive p-values: what they are and what they are not. *Test* **10**, 75-86.

[6] Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd.

[7] Gabriel, K. R. (1969). Simultaneous test procedures. Some theory of multiple comparisons. *Ann. Math. Statist.* **40**, 224, 250.

[8] Lavine, M. and Schervish, M. J. (1999). Bayes factors: what they are and what they are not. *Amer. Statist.* **53**, 119-122.

[9] Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*, San Francisco: Holden-Day.

[10] Schervish, M. (1996). P-values: what they are and what they are not. *Amer. Statist.* **50**, 203-206.

[11] Shafer, G. (1982). Lindley's paradox (with discussion). *J. Amer. Statist. Assoc.* **77**, 325-351.

[12] Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Statist.* **31**, 2013-2035.

About the author

Julián de la Horra is Professor in the Department of Mathematics of the "Universidad Autónoma de Madrid". In the last twenty years, his research interest has been focussed on different topics of Bayesian inference as, for instance, Bayesian robustness, posterior predictive p-value and Bayesian model selection.