p-values: a leading cause of the lack of replicability in Science?

Gonzalo García-Donato

Valencia (FISABIO) - Mayo 2017

Universidad de Castilla-La Mancha (Spain)

Table of contents

CRISIS? WHAT CRISIS? (Supertramp 1975)

Introductory words

The problem

The underlying problem behind p-values is testing, meaning that we want to *measure the evidence* (however it means) that data, $\boldsymbol{x}$, gives in favour (or against) certain theories or hypotheses H.

The problem

The underlying problem behind p-values is testing, meaning that we want to *measure the evidence* (however it means) that data, $\boldsymbol{x}$, gives in favour (or against) certain theories or hypotheses H.

• When uncertainty is present, Statistics are called for to solve this fundamental, ambitious and difficult problem in Science.

The problem

The underlying problem behind p-values is testing, meaning that we want to *measure the evidence* (however it means) that data, $\boldsymbol{x}$, gives in favour (or against) certain theories or hypotheses H.

• When uncertainty is present, Statistics are called for to solve this fundamental, ambitious and difficult problem in Science.

• In Statistics, testing is embedded in a probabilistic framework

$$\boldsymbol{x} \sim \boldsymbol{f}(\boldsymbol{x} \mid \boldsymbol{\theta}),$$

and hypotheses are made equivalent to algebraic sentences of the type

$$H : \boldsymbol{\theta} \in \Theta_H$$

where Θ_H is a subset of the parametric space.

The problem

The underlying problem behind p-values is testing, meaning that we want to *measure the evidence* (however it means) that data, $\boldsymbol{x}$, gives in favour (or against) certain theories or hypotheses H.

• When uncertainty is present, Statistics are called for to solve this fundamental, ambitious and difficult problem in Science.

• In Statistics, testing is embedded in a probabilistic framework

$$\boldsymbol{x} \sim \boldsymbol{f}(\boldsymbol{x} \mid \boldsymbol{\theta}),$$

and hypotheses are made equivalent to algebraic sentences of the type

$$H : \boldsymbol{\theta} \in \Theta_H$$

where Θ_H is a subset of the parametric space.

Testing and replicability

Suppose that we conclude that $\boldsymbol{x}$ provides strength evidence in favour of certain H. A replicability issue appears if H is not similarly endorsed by a repetition $\boldsymbol{x}^\star$ of the experiment.

Typical situation

Example: two treatments

If μ_i is the mean of time to recover of certain disease for treatment i then $H : \mu_1 - \mu_2 > 0$ express the hypothesis that treatment 2 is better than treatment 1 (in the sense that, in average, less time to recover is needed).

Typical situation

Example: two treatments

If μ_i is the mean of time to recover of certain disease for treatment i then $H : \mu_1 - \mu_2 > 0$ express the hypothesis that treatment 2 is better than treatment 1 (in the sense that, in average, less time to recover is needed).

• The above example represents a very typical situation where we have two hypotheses (one normally being the complement of the other) and we test one against the other

$$H_0 : \boldsymbol{\theta} \in \Theta_0, \;\; H_1 : \boldsymbol{\theta} \in \Theta - \Theta_0.$$

Hypothesis testing: revealing observations

1. Testing and model selection

Hypotheses define different statistical models (say f_H) with a common parametric form but differing in the location of the parameters.

For instance:

$$x \sim N(\mu, \sigma^2), \quad H_0 : \mu = 0, \; H_1 : \mu \neq 0$$

is equivalent to

$$M_0 : N(0, \sigma^2), \; M_1 : N(\mu, \sigma^2).$$

1. Testing and model selection

Hypotheses define different statistical models (say f_H) with a common parametric form but differing in the location of the parameters.

For instance:

$$x \sim N(\mu, \sigma^2), \quad H_0 : \mu = 0,\ H_1 : \mu \neq 0$$

is equivalent to

$$M_0 : N(0, \sigma^2),\ M_1 : N(\mu, \sigma^2).$$

• That is why in the specialized literature you can find articles in the area of testing with names like "model selection" or "model choice".

1. Testing and model selection

Hypotheses define different statistical models (say f_H) with a common parametric form but differing in the location of the parameters.

For instance:

$$x \sim N(\mu, \sigma^2), \quad H_0 : \mu = 0, \; H_1 : \mu \neq 0$$

is equivalent to

$$M_0 : N(0, \sigma^2), \; M_1 : N(\mu, \sigma^2).$$

• That is why in the specialized literature you can find articles in the area of testing with names like "model selection" or "model choice".

Naive approach to testing:

To measure the degree of compatibility ("better fit") of hypotheses/models with data.

Drawback: lack of natural interpretation in terms of evidence.

2. Impossible theories

A particular (and unfortunately ubiquitous in applied studies) testing situation contains hypotheses where Θ_H consists on only one point $\boldsymbol{\theta}_0$ (e.g. $\mu_1 = \mu_2$, or $\beta = 0$). These are called *point* or *precise* hypotheses.

2. Impossible theories

A particular (and unfortunately ubiquitous in applied studies) testing situation contains hypotheses where Θ_H consists on only one point $\boldsymbol{\theta}_0$ (e.g. $\mu_1 = \mu_2$, or $\beta = 0$). These are called *point* or *precise* hypotheses.

• Traditionally, the hypothesis defending point hypotheses is called the *null* while the other is called the *alternative*.

2. Impossible theories

A particular (and unfortunately ubiquitous in applied studies) testing situation contains hypotheses where Θ_H consists on only one point $\boldsymbol{\theta}_0$ (e.g. $\mu_1 = \mu_2$, or $\beta = 0$). These are called *point* or *precise* hypotheses.

• Traditionally, the hypothesis defending point hypotheses is called the *null* while the other is called the *alternative*.

• Point hypotheses are a disaster for mathematicians as we are trying to compare statistical models of different complexities and this is like trying to compare "peras y manzanas". The idea of "better fit" is useless as obviously the complex models always fits data better.

2. Impossible theories

A particular (and unfortunately ubiquitous in applied studies) testing situation contains hypotheses where Θ_H consists on only one point $\boldsymbol{\theta}_0$ (e.g. $\mu_1 = \mu_2$, or $\beta = 0$). These are called *point* or *precise* hypotheses.

• Traditionally, the hypothesis defending point hypotheses is called the *null* while the other is called the *alternative*.

• Point hypotheses are a disaster for mathematicians as we are trying to compare statistical models of different complexities and this is like trying to compare “peras y manzanas”. The idea of “better fit” is useless as obviously the complex models always fits data better.

• How to mix “better fit” and “complexity” we do not know but we all agree in

Ockham’s razor

For a similar fit choose the simplest explanation.

2. Impossible theories (II)

Many statisticians think that, in the real world, precise hypotheses do not exist: do you really think that the two average times to recover could ever be *exactly* ($\epsilon = 0$!) the same?

2. Impossible theories (II)

Many statisticians think that, in the real world, precise hypotheses do not exist: do you really think that the two average times to recover could ever be *exactly* ($\epsilon = 0$!) the same?

• What researchers have in mind with precise hypotheses is "$\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ do not differ by more of certain small quantity". (e.g. in the treatment of a headache, 10 minutes?)

2. Impossible theories (II)

Many statisticians think that, in the real world, precise hypotheses do not exist: do you really think that the two average times to recover could ever be *exactly* ($\epsilon = 0$!) the same?

• What researchers have in mind with precise hypotheses is "$\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ do not differ by more of certain small quantity". (e.g. in the treatment of a headache, 10 minutes?)

Keep in mind!

For practical purposes, null hypotheses are expressed as precise hypotheses (say $H_0 : \mu_1 - \mu_2 = 0$ in the treatments example). But, if what is true is $\mu_1 - \mu_2 = \epsilon$ (for certain small ϵ) then H_0 would still be true.

2. Impossible theories (II)

Many statisticians think that, in the real world, precise hypotheses do not exist: do you really think that the two average times to recover could ever be *exactly* ($\epsilon = 0!$) the same?

• What researchers have in mind with precise hypotheses is "$\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ do not differ by more of certain small quantity". (e.g. in the treatment of a headache, 10 minutes?)

Keep in mind!

For practical purposes, null hypotheses are expressed as precise hypotheses (say $H_0 : \mu_1 - \mu_2 = 0$ in the treatments example). But, if what is true is $\mu_1 - \mu_2 = \epsilon$ (for certain small ϵ) then H_0 would still be true.

• Laziness; the difficulty in interpreting complex statistical parameteres; and the availability of magic simple tools have lead researchers to neglect the need to specify clearly what we mean by "close to $\boldsymbol{\theta}_0$".

2. Impossible theories (II)

Many statisticians think that, in the real world, precise hypotheses do not exist: do you really think that the two average times to recover could ever be *exactly* ($\epsilon = 0!$) the same?

• What researchers have in mind with precise hypotheses is "$\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ do not differ by more of certain small quantity". (e.g. in the treatment of a headache, 10 minutes?)

Keep in mind!

For practical purposes, null hypotheses are expressed as precise hypotheses (say $H_0 : \mu_1 - \mu_2 = 0$ in the treatments example). But, if what is true is $\mu_1 - \mu_2 = \epsilon$ (for certain small ϵ) then H_0 would still be true.

• Laziness; the difficulty in interpreting complex statistical parameteres; and the availability of magic simple tools have lead researchers to neglect the need to specify clearly what we mean by "close to $\boldsymbol{\theta}_0$".

• Implicitly or explicitly the responsability of such details is left, by default, to Statistical methods while it seems obvious that such information is not contained in the data nor in the model assumed.

2. Impossible theories (II)

Many statisticians think that, in the real world, precise hypotheses do not exist: do you really think that the two average times to recover could ever be *exactly* ($\epsilon = 0$!) the same?

• What researchers have in mind with precise hypotheses is "$\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ do not differ by more of certain small quantity". (e.g. in the treatment of a headache, 10 minutes?)

Keep in mind!

For practical purposes, null hypotheses are expressed as precise hypotheses (say $H_0 : \mu_1 - \mu_2 = 0$ in the treatments example). But, if what is true is $\mu_1 - \mu_2 = \epsilon$ (for certain small ϵ) then H_0 would still be true.

• Laziness; the difficulty in interpreting complex statistical parameteres; and the availability of magic simple tools have lead researchers to neglect the need to specify clearly what we mean by "close to $\boldsymbol{\theta}_0$".

• Implicitly or explicitly the responsability of such details is left, by default, to Statistical methods while it seems obvious that such information is not contained in the data nor in the model assumed.

There are limits even for statistics!

3. Estimation and testing

- We are always tempted to solve a testing problem via construct an interval for the parameter and reject or not depending on its agreement with hypotheses.

3. Estimation and testing

- We are always tempted to solve a testing problem via construct an interval for the parameter and reject or not depending on its agreement with hypotheses.
- This puts testing and estimation in the same box.

3. Estimation and testing

- We are always tempted to solve a testing problem via construct an interval for the parameter and reject or not depending on its agreement with hypotheses.
- This puts testing and estimation in the same box.

Take the problem above

If an estimation tool is used to sove a testing problem, you are preassuming that M_1 is correct.

3. Estimation and testing

- We are always tempted to solve a testing problem via construct an interval for the parameter and reject or not depending on its agreement with hypotheses.
- This puts testing and estimation in the same box.

Take the problem above

If an estimation tool is used to sove a testing problem, you are preassuming that M_1 is correct.

p-values

Significant testing (p-values)

Significant testing (p-values)

- A main tool to solve testing problems with precise hypotheses is the p-value, popularized by R. Fisher almost 100 years ago.

Significant testing (p-values)

• A main tool to solve testing problems with precise hypotheses is the p-value, popularized by R. Fisher almost 100 years ago.

p-value (p for short)=

the probability of obtaining the observed data, or more extreme, if the null hypothesis is true.

• Nothing wrong with p-value itself (it is just a number!) it is how we use and interpret it.

What we teach to our students

- Wikipedia (valor-p); spanish version: el valor p nos muestra la probabilidad de haber obtenido el resultado que hemos obtenido si suponemos que la hipótesis nula es cierta (...)

What we teach to our students

- Wikipedia (valor-p); spanish version: el valor p nos muestra la probabilidad de haber obtenido el resultado que hemos obtenido si suponemos que la hipótesis nula es cierta (...)si el valor p es inferior al nivel de significación, lo más verosímil es que la hipótesis de partida sea falsa.

What we teach to our students

- Wikipedia (valor-p); spanish version: el valor p nos muestra la probabilidad de haber obtenido el resultado que hemos obtenido si suponemos que la hipótesis nula es cierta (...)si el valor p es inferior al nivel de significación, lo más verosímil es que la hipótesis de partida sea falsa.
- Johnson and Bhattacharyya 1992: *The p-value gauges the strength of evidence against H_0 on a numerical scale.*

What we teach to our students

- Wikipedia (valor-p); spanish version: el valor p nos muestra la probabilidad de haber obtenido el resultado que hemos obtenido si suponemos que la hipótesis nula es cierta (...)si el valor p es inferior al nivel de significación, lo más verosímil es que la hipótesis de partida sea falsa.
- Johnson and Bhattacharyya 1992: *The p-value gauges the strength of evidence against H_0 on a numerical scale.*
- Velez and García 1993: *El p-valor indica el apoyo que la hipótesis nula recibe de las observaciones.*

What we teach to our students

- Wikipedia (valor-p); spanish version: el valor p nos muestra la probabilidad de haber obtenido el resultado que hemos obtenido si suponemos que la hipótesis nula es cierta (...)si el valor p es inferior al nivel de significación, lo más verosímil es que la hipótesis de partida sea falsa.
- Johnson and Bhattacharyya 1992: *The p-value gauges the strength of evidence against* H_0 *on a numerical scale.*
- Velez and García 1993: *El p-valor indica el apoyo que la hipótesis nula recibe de las observaciones.*
- Agulló, Carratalá, and Gimeno 1999: *Como el p-valor es muy pequeño, la evidencia muestral en contra de* H_0 *es fuerte.*

ASA statement on Statistical Significance and p-values (2016)

- Researchers often wish to turn a p-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The p-value is neither.

What we teach to our students

- Wikipedia (valor-p); spanish version: el valor p nos muestra la probabilidad de haber obtenido el resultado que hemos obtenido si suponemos que la hipótesis nula es cierta (...)si el valor p es inferior al nivel de significación, lo más verosímil es que la hipótesis de partida sea falsa.
- Johnson and Bhattacharyya 1992: *The p-value gauges the strength of evidence against H_0 on a numerical scale.*
- Velez and García 1993: *El p-valor indica el apoyo que la hipótesis nula recibe de las observaciones.*
- Agulló, Carratalá, and Gimeno 1999: *Como el p-valor es muy pequeño, la evidencia muestral en contra de H_0 es fuerte.*

ASA statement on Statistical Significance and p-values (2016)

- Researchers often wish to turn a p-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The p-value is neither.
- By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

The issues with p-values are not new

Many statisticians have alerted in the past about the dangers associated with significant testing and p-values:

Hogben 1957

We can already detect signs of such deterioration in the growing volume of published papers...recording so-called significant conclusions which an earlier vintage would have regarded merely as private clues for further exploration.

The issues with p-values are not new

Many statisticians have alerted in the past about the dangers associated with significant testing and p-values:

Hogben 1957

We can already detect signs of such deterioration in the growing volume of published papers...recording so-called significant conclusions which an earlier vintage would have regarded merely as private clues for further exploration.

Lang, Rothman, and Cann 1998; Epidemiology editorial

p-value continues to be used mistakenly as a measure of the importance and credibility of study results.

The issues with p-values are not new

Many statisticians have alerted in the past about the dangers associated with significant testing and p-values:

Hogben 1957

We can already detect signs of such deterioration in the growing volume of published papers...recording so-called significant conclusions which an earlier vintage would have regarded merely as private clues for further exploration.

Lang, Rothman, and Cann 1998; Epidemiology editorial

p-value continues to be used mistakenly as a measure of the importance and credibility of study results.

The issues with p-values are not new

Many statisticians have alerted in the past about the dangers associated with significant testing and p-values:

Hogben 1957

We can already detect signs of such deterioration in the growing volume of published papers...recording so-called significant conclusions which an earlier vintage would have regarded merely as private clues for further exploration.

Lang, Rothman, and Cann 1998; Epidemiology editorial

p-value continues to be used mistakenly as a measure of the importance and credibility of study results.

-In our days, the focus has placed on the idea that p-values are a major cause of the "reproducibility crisis" of Science.

The issues with p-values are not new

Many statisticians have alerted in the past about the dangers associated with significant testing and p-values:

Hogben 1957

We can already detect signs of such deterioration in the growing volume of published papers...recording so-called significant conclusions which an earlier vintage would have regarded merely as private clues for further exploration.

Lang, Rothman, and Cann 1998; Epidemiology editorial

p-value continues to be used mistakenly as a measure of the importance and credibility of study results.

-In our days, the focus has placed on the idea that p-values are a major cause of the "reproducibility crisis" of Science.

J. O. Berger 2015

...few people actually understand what a p-value means; and the rampant misinterpretation of p-values is largely responsible for the well-documented lack of reproducibility of science.

Journals policies

These considerations have lead to several journals to regulate in some sense the use of such tools.

Journals policies

These considerations have lead to several journals to regulate in some sense the use of such tools.

ASA statement on Statistical Significance and p-values (2016)

The widespread use of "statistical significance" (generally interpreted as $p \leq 0.05$) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.

Journals policies

These considerations have lead to several journals to regulate in some sense the use of such tools.

ASA statement on Statistical Significance and p-values (2016)

The widespread use of "statistical significance" (generally interpreted as $p \leq 0.05$*) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.*

- A radical position is the one taken by the editors of *Basic and Applied Social Psychology* to ban p-values:

BASP editorial (2015)

...authors will have to remove all vestiges of the (null significance test procedure) NHSTP (p-values, t-values, F-values, statements about significant differences or lack thereof, and so on).

...we believe that the $p < .05$ *bar is too easy to pass and sometimes serves as an excuse for lower quality research.*

p-values and replicability

p-values and replicability

The main issue with p-values is that you can be the most honest of the statiscians and still be producing non-replicable results. Why? (a dangerous combination of the following reasons)

p-values and replicability

The main issue with p-values is that you can be the most honest of the statiscians and still be producing non-replicable results. Why? (a dangerous combination of the following reasons)

- The (legitime) goal of researchers to produce *positive* findings (and of companies to have positive endorsements about their products);

p-values and replicability

The main issue with p-values is that you can be the most honest of the statiscians and still be producing non-replicable results. Why? (a dangerous combination of the following reasons)

- The (legitime) goal of researchers to produce *positive* findings (and of companies to have positive endorsements about their products);
- The (legitime) goal of scientific journals to publish positive journals (and of society to embrace positive results);

p-values and replicability

The main issue with p-values is that you can be the most honest of the statiscians and still be producing non-replicable results. Why? (a dangerous combination of the following reasons)

- The (legitime) goal of researchers to produce *positive* findings (and of companies to have positive endorsements about their products);
- The (legitime) goal of scientific journals to publish positive journals (and of society to embrace positive results);
- The easy predisposition of p-values to produce positive results.

p-values and replicability

The main issue with p-values is that you can be the most honest of the statiscians and still be producing non-replicable results. Why? (a dangerous combination of the following reasons)

- The (legitime) goal of researchers to produce *positive* findings (and of companies to have positive endorsements about their products);
- The (legitime) goal of scientific journals to publish positive journals (and of society to embrace positive results);
- The easy predisposition of p-values to produce positive results.

Matthews n.d. (1998)

The plain fact is that 70 years ago R. Fisher gave scientists a mathematical machine for turning baloney into breakthroughs and flukes into fundings.

A high tendency to declare positives

Why are p-values so well-trained to declare positives? (or equivalently to reject null hypotheses). Three possible reasons:

- R1: Because of its definition,
- R2: because of the effect of n,
- R3: because of a distorted interpretation in frequentist terms.

R1. Because of its definition

The BASP editorial states that ”the problem is in traversing the distance from the probability of the finding, given the null hypothesis, to the probability of the null hypothesis, given the finding”.

R1. Because of its definition

The BASP editorial states that "the problem is in traversing the distance from the probability of the finding, given the null hypothesis, to the probability of the null hypothesis, given the finding".

Still the intuition behind p-values seems right:

"Rare" observations under a theory should lead us to suspect about the validity of that theory (hence rejecting the null and declaring a positive).

R1. Because of its definition

The BASP editorial states that "the problem is in traversing the distance from the probability of the finding, given the null hypothesis, to the probability of the null hypothesis, given the finding".

Still the intuition behind p-values seems right:

"Rare" observations under a theory should lead us to suspect about the validity of that theory (hence rejecting the null and declaring a positive).

Nevertheless, with p-values, we impute to the null the rareness of what has been observed plus the rareness of *any* other even more rare unobserved event.

R1. Because of its definition

The BASP editorial states that "the problem is in traversing the distance from the probability of the finding, given the null hypothesis, to the probability of the null hypothesis, given the finding".

Still the intuition behind p-values seems right:

"Rare" observations under a theory should lead us to suspect about the validity of that theory (hence rejecting the null and declaring a positive).

Nevertheless, with p-values, we impute to the null the rareness of what has been observed plus the rareness of *any* other even more rare unobserved event.

Taken to an extreme, the equivalent of this would be to put in jail a guy for having stolen a CD and for not having stolen the rest of the items in the shop.

R1. Because of its definition

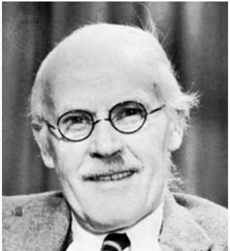
The BASP editorial states that "the problem is in traversing the distance from the probability of the finding, given the null hypothesis, to the probability of the null hypothesis, given the finding".

Still the intuition behind p-values seems right:

"Rare" observations under a theory should lead us to suspect about the validity of that theory (hence rejecting the null and declaring a positive).

Nevertheless, with p-values, we impute to the null the rareness of what has been observed plus the rareness of *any* other even more rare unobserved event.

Taken to an extreme, the equivalent of this would be to put in jail a guy for having stolen a CD and for not having stolen the rest of the items in the shop.



R1. Because of its definition

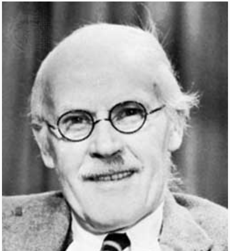
The BASP editorial states that "the problem is in traversing the distance from the probability of the finding, given the null hypothesis, to the probability of the null hypothesis, given the finding".

Still the intuition behind p-values seems right:

"Rare" observations under a theory should lead us to suspect about the validity of that theory (hence rejecting the null and declaring a positive).

Nevertheless, with p-values, we impute to the null the rareness of what has been observed plus the rareness of *any* other even more rare unobserved event.

Taken to an extreme, the equivalent of this would be to put in jail a guy for having stolen a CD and for not having stolen the rest of the items in the shop.



Jeffreys 1961 A hypothesis, that may be true, may be rejected because it has not predicted observable results that have not occurred.

R2. Because of the effect of n

Assume you are testing $H_0 : \theta = 0$.

R2. Because of the effect of n

Assume you are testing $H_0 : \theta = 0$.

- Fact 1: The true value of $\theta = \epsilon$ (for certain small $\epsilon \neq 0$) but H_0 is still right.

R2. Because of the effect of n

Assume you are testing $H_0 : \theta = 0$.

- Fact 1: The true value of $\theta = \epsilon$ (for certain small $\epsilon \neq 0$) but H_0 is still right.
- Fact 2: As n increases, the sampling distribution of the maximum likelihood estimator $\hat{\theta}$ concentrates on ϵ (at the rate $1/\sqrt{n}$). This means that, with n large enough we can have $\hat{\theta}$ be "as far as we want" (in probabilistic terms) from 0.

R2. Because of the effect of n

Assume you are testing $H_0 : \theta = 0$.

- Fact 1: The true value of $\theta = \epsilon$ (for certain small $\epsilon \neq 0$) but H_0 is still right.
- Fact 2: As n increases, the sampling distribution of the maximum likelihood estimator $\hat{\theta}$ concentrates on ϵ (at the rate $1/\sqrt{n}$). This means that, with n large enough we can have $\hat{\theta}$ be "as far as we want" (in probabilistic terms) from 0.
- Fact 3: The p-value tends to zero with n even if the null is (in the real world) right.

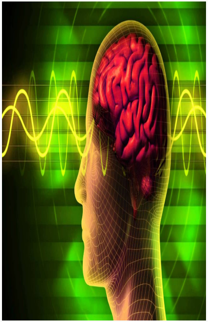
R2. Because of the effect of n

Assume you are testing $H_0 : \theta = 0$.

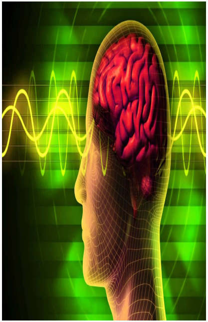
- Fact 1: The true value of $\theta = \epsilon$ (for certain small $\epsilon \neq 0$) but H_0 is still right.
- Fact 2: As n increases, the sampling distribution of the maximum likelihood estimator $\hat{\theta}$ concentrates on ϵ (at the rate $1/\sqrt{n}$). This means that, with n large enough we can have $\hat{\theta}$ be "as far as we want" (in probabilistic terms) from 0.
- Fact 3: The p-value tends to zero with n even if the null is (in the real world) right.

This is why researchers have the impression that, with n large enough, you can always "find" a positive effect.

R2. Because of the effect of n



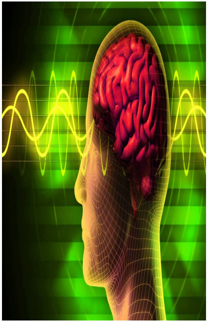
R2. Because of the effect of n



Parapsychology and extra sensory perception (ESP)

- An electronic device (REG) is arranged so that one gets a random sequence of 0s and 1s with theoretically equal probability for each outcome ($\pi = 0.5$).

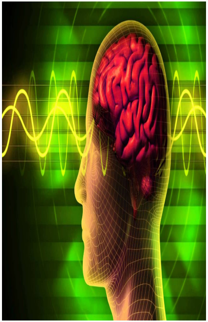
R2. Because of the effect of n



Parapsychology and extra sensory perception (ESP)

• An electronic device (REG) is arranged so that one gets a random sequence of 0s and 1s with theoretically equal probability for each outcome ($\pi = 0.5$).

• A subject then attempts to influence the REG and obtain a sequence of pulses with a distribution that is different from the baseline. In testing terms $H_0 : \pi = 0.5$ against the alternative $H_1 : \pi \neq 0.5$.



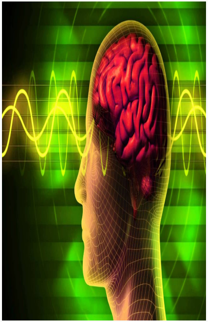
Parapsychology and extra sensory perception (ESP)

• An electronic device (REG) is arranged so that one gets a random sequence of 0s and 1s with theoretically equal probability for each outcome ($\pi = 0.5$).

• A subject then attempts to influence the REG and obtain a sequence of pulses with a distribution that is different from the baseline. In testing terms $H_0 : \pi = 0.5$ against the alternative $H_1 : \pi \neq 0.5$.

- One of such experiment in the late 80's revisited by Jefferys 1990 declared strongly the existence of ESP based on a $p < 0.0003$.

R2. Because of the effect of n



Parapsychology and extra sensory perception (ESP)

• An electronic device (REG) is arranged so that one gets a random sequence of 0s and 1s with theoretically equal probability for each outcome ($\pi = 0.5$).

• A subject then attempts to influence the REG and obtain a sequence of pulses with a distribution that is different from the baseline. In testing terms $H_0 : \pi = 0.5$ against the alternative $H_1 : \pi \neq 0.5$.

- One of such experiment in the late 80's revisited by Jefferys 1990 declared strongly the existence of ESP based on a $p < 0.0003$.
- The experiment had $n = 104,490,000$ and they obtained $s = 52,263,471$ succeses implying an observed proportion of $\hat{\pi} = 0.5001768$.

R3. Because of a distorted interpretation of p in frequentist terms.

- We tend to think that a small p (and as statisticians we have in our DNA what small 'means') implies evidence against the null (in the logical sense that it is far more likely that it comes from the alternative than from the null).

R3. Because of a distorted interpretation of p in frequentist terms.

- We tend to think that a small p (and as statisticians we have in our DNA what small 'means') implies evidence against the null (in the logical sense that it is far more likely that it comes from the alternative than from the null).
- The above would suggest that p summarizes the evidence against the null contained in the data, but this is not true.

R3. Because of a distorted interpretation of p in frequentist terms.

- We tend to think that a small p (and as statisticians we have in our DNA what small 'means') implies evidence against the null (in the logical sense that it is far more likely that it comes from the alternative than from the null).
- The above would suggest that p summarizes the evidence against the null contained in the data, but this is not true.

Replicability and p

- The probability (in frequentists terms) that a sample, $\boldsymbol{x}$, with (say) $p \approx 0.05$ comes from the null is not small (as we would expect). In fact, it can easily be larger than the probability that it comes from the alternative.

R3. Because of a distorted interpretation of p in frequentist terms.

- We tend to think that a small p (and as statisticians we have in our DNA what small 'means') implies evidence against the null (in the logical sense that it is far more likely that it comes from the alternative than from the null).
- The above would suggest that p summarizes the evidence against the null contained in the data, but this is not true.

Replicability and p

- The probability (in frequentists terms) that a sample, $\boldsymbol{x}$, with (say) $p \approx 0.05$ comes from the null is not small (as we would expect). In fact, it can easily be larger than the probability that it comes from the alternative.
- Consequence: if you repeat the experiment, $\boldsymbol{x}^\star$, it could be quite likely that $\boldsymbol{x}^\star$ comes from the null! –and if it is so, you can see *any* $p^\star$ since

$$p^\star = p(\boldsymbol{x}^\star) \mid H_0 \sim Uniform(0,1).$$

R3. Because of a distorted interpretation of p in frequentist terms.

- We tend to think that a small p (and as statisticians we have in our DNA what small 'means') implies evidence against the null (in the logical sense that it is far more likely that it comes from the alternative than from the null).
- The above would suggest that p summarizes the evidence against the null contained in the data, but this is not true.

Replicability and p

- The probability (in frequentists terms) that a sample, $\boldsymbol{x}$, with (say) $p \approx 0.05$ comes from the null is not small (as we would expect). In fact, it can easily be larger than the probability that it comes from the alternative.
- Consequence: if you repeat the experiment, $\boldsymbol{x}^\star$, it could be quite likely that $\boldsymbol{x}^\star$ comes from the null! –and if it is so, you can see *any* $p^\star$ since

$$p^\star = p(\boldsymbol{x}^\star) \mid H_0 \sim \mathit{Uniform}(0, 1).$$

R3. Because of a distorted interpretation of p in frequentist terms.

- The above observations are part of a series of papers (James O. Berger and Sellke 1987; Sellke, Bayarri, and J. O. Berger 2001) where, for instance, it is proved (mathematically) that for $p \approx 0.05$ the proportion of times the null is true is at least 0.23.

R3. Because of a distorted interpretation of p in frequentist terms.

• The above observations are part of a series of papers (James O. Berger and Sellke 1987; Sellke, Bayarri, and J. O. Berger 2001) where, for instance, it is proved (mathematically) that for $p \approx 0.05$ the proportion of times the null is true is at least 0.23.

James O. Berger and Sellke 1987

• *$p \approx 0.05$ really means that there is at best very weak evidence against H_0.*

• *p-values can be highly misleading measures of the evidence provided by the data against the null.*

R3. Because of a distorted interpretation of p in frequentist terms.

• The above observations are part of a series of papers (James O. Berger and Sellke 1987; Sellke, Bayarri, and J. O. Berger 2001) where, for instance, it is proved (mathematically) that for $p \approx 0.05$ the proportion of times the null is true is at least 0.23.

James O. Berger and Sellke 1987

• $p \approx 0.05$ *really means that there is at best very weak evidence against* H_0.

• *p-values can be highly misleading measures of the evidence provided by the data against the null.*

• The above are mathematical results and hence obscure to many practitioners.

R3. Because of a distorted interpretation of p in frequentist terms.

• The above observations are part of a series of papers (James O. Berger and Sellke 1987; Sellke, Bayarri, and J. O. Berger 2001) where, for instance, it is proved (mathematically) that for $p \approx 0.05$ the proportion of times the null is true is at least 0.23.

James O. Berger and Sellke 1987

• *$p \approx 0.05$ really means that there is at best very weak evidence against H_0.*

• *p-values can be highly misleading measures of the evidence provided by the data against the null.*

• The above are mathematical results and hence obscure to many practitioners.

• Nevertheless, we can experiment these facts using a small program in R (adapted by Hector Perpian from Sellke, Bayarri, and J. O. Berger 2001)

Simulation

Many different experiments $D_1, D_2, \ldots, ..., D_L$ performed to test

$$H_0 : \theta = 0, \;\; H_1 : \theta \neq 0.$$

where θ is the mean of

$$x_1, \ldots x_n \sim N(\theta, \sigma^2).$$

Suppose H_0 is true with certain probability, and under H_1: $\theta \sim N(0, \sigma_P^2)$. Other possibilities:

- $\theta = a$ (fixed value),
- $\theta \sim Un(-a, a)$

R3. Because of a distorted interpretation of p in frequentist terms.

A main message here is that

> *Bayarri's talk; 2013 Knowing that data is "rare" under H_0 is of little use unles one determines whether or not it is also "rare" under H_1*

- In general, $p \approx 0.05$ is as rare under H_0 than under H_1.

p-values, alphas and a possible solution

Classical theories

!

But, if I am not wrong, p values are an ingredient of the classical theory of testing which, of course, has frequentist properties, no?

Classical theories

!

But, if I am not wrong, p values are an ingredient of the classical theory of testing which, of course, has frequentist properties, no?

• What we normally called "classical theory" has in fact two main schools of thinking in testing: the one advocated by Fisher and a second one defended by Jerzy Neyman and Egon Pearson.

Classical theories

!

But, if I am not wrong, p values are an ingredient of the classical theory of testing which, of course, has frequentist properties, no?

• What we normally called “classical theory” has in fact two main schools of thinking in testing: the one advocated by Fisher and a second one defended by Jerzy Neyman and Egon Pearson.

• Both theories (Fisher vs. Neyman-Pearson) are very different (irreconciliable) but surprisingly textbooks have traditionally insisted in telling the story as a unified theory then contributing to the confusion in the interpretation of p values Hubbard and Bayarri 2003.

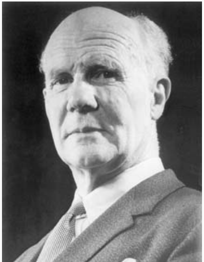
Classical theories

!

But, if I am not wrong, p values are an ingredient of the classical theory of testing which, of course, has frequentist properties, no?

• What we normally called "classical theory" has in fact two main schools of thinking in testing: the one advocated by Fisher and a second one defended by Jerzy Neyman and Egon Pearson.

• Both theories (Fisher vs. Neyman-Pearson) are very different (irreconciliable) but surprisingly textbooks have traditionally insisted in telling the story as a unified theory then contributing to the confusion in the interpretation of p values Hubbard and Bayarri 2003.



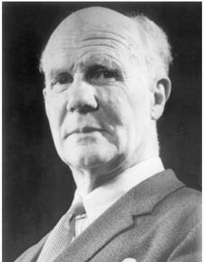
Classical theories

!

But, if I am not wrong, p values are an ingredient of the classical theory of testing which, of course, has frequentist properties, no?

• What we normally called "classical theory" has in fact two main schools of thinking in testing: the one advocated by Fisher and a second one defended by Jerzy Neyman and Egon Pearson.

• Both theories (Fisher vs. Neyman-Pearson) are very different (irreconciliable) but surprisingly textbooks have traditionally insisted in telling the story as a unified theory then contributing to the confusion in the interpretation of p values Hubbard and Bayarri 2003.



J. Neyman- E. Pearson significance testing

Construct a critical region (CR) subject to a prespecified type-I error rate α. If the sample falls in CR then reject H_0 and report α as the error measure.

p-values and α's

Fisher's School	Neyman-Pearson's School

p-values and α's

Fisher's School	Neyman-Pearson's School
No alternative hypothesis (Only H_0)	Two hypotheses: H_0 and H_A

p-values and α's

Fisher's School	Neyman-Pearson's School
No alternative hypothesis (Only H_0)	Two hypotheses: H_0 and H_A
p measures evidence	Choose a hypothesis an report error

p-values and α's

Fisher's School	Neyman-Pearson's School
No alternative hypothesis (Only H_0)	Two hypotheses: H_0 and H_A
p measures evidence	Choose a hypothesis an report error
One can never accept H_0, only fail to reject it	You can accept H_0

p-values and α's

Fisher's School	Neyman-Pearson's School
No alternative hypothesis (Only H_0)	Two hypotheses: H_0 and H_A
p measures evidence	Choose a hypothesis an report error
One can never accept H_0, only fail to reject it	You can accept H_0
Level of significance (threshold for p)	Level of significance (Type-I error)

p-values and α's

Fisher's School	Neyman-Pearson's School
No alternative hypothesis (Only H_0)	Two hypotheses: H_0 and H_A
p measures evidence	Choose a hypothesis an report error
One can never accept H_0, only fail to reject it	You can accept H_0
Level of significance (threshold for p)	Level of significance (Type-I error)
No notion of frequentism	Repetitions of the experiment

p-values and α's

Fisher's School	Neyman-Pearson's School
No alternative hypothesis (Only H_0)	Two hypotheses: H_0 and H_A
p measures evidence	Choose a hypothesis an report error
One can never accept H_0, only fail to reject it	You can accept H_0
Level of significance (threshold for p)	Level of significance (Type-I error)
No notion of frequentism	Repetitions of the experiment

Neyman-Pearson's theory of testing is based on the frequentist vision of probability and is safe in terms of reproducibility. In the app situation, only 5% of the studies rejected at $\alpha = 0.05$ would correspond to a true H_0.

p-values and α's

Fisher's School	Neyman-Pearson's School
No alternative hypothesis (Only H_0)	Two hypotheses: H_0 and H_A
p measures evidence	Choose a hypothesis an report error
One can never accept H_0, only fail to reject it	You can accept H_0
Level of significance (threshold for p)	Level of significance (Type-I error)
No notion of frequentism	Repetitions of the experiment

Neyman-Pearson's theory of testing is based on the frequentist vision of probability and is safe in terms of reproducibility. In the app situation, only 5% of the studies rejected at $\alpha = 0.05$ would correspond to a true H_0.

Here you are only allowed to report the action (accept or reject H_0) and the accompanying (fixed) error α (not very appealing).

p-values and α's

Fisher's School	Neyman-Pearson's School
No alternative hypothesis (Only H_0)	Two hypotheses: H_0 and H_A
p measures evidence	Choose a hypothesis an report error
One can never accept H_0, only fail to reject it	You can accept H_0
Level of significance (threshold for p)	Level of significance (Type-I error)
No notion of frequentism	Repetitions of the experiment

Neyman-Pearson's theory of testing is based on the frequentist vision of probability and is safe in terms of reproducibility. In the app situation, only 5% of the studies rejected at $\alpha = 0.05$ would correspond to a true H_0.

Here you are only allowed to report the action (accept or reject H_0) and the accompanying (fixed) error α (not very appealing).

The p vaue can be used SIMPLY as a convenient tool to avoid the explicit construction of the critical region.

Differences

- We have seen very disturbing *essential* differences among the main schools of statistics to approach testing problems: Neyman-Pearson (α), Fisher (p) and Bayesian (Bayes factors or posterior probabilities). These differences are nicely explained in a historical context in J. O. Berger 2003

Differences

• We have seen very disturbing *essential* differences among the main schools of statistics to approach testing problems: Neyman-Pearson (α), Fisher (p) and Bayesian (Bayes factors or posterior probabilities). These differences are nicely explained in a historical context in J. O. Berger 2003

An example

Suppose a basic test $H_0 : \mu = 0$ with σ known with $n = 10$ and test statistic

$$z = \sqrt{n}\bar{x}/\sigma = 2.3.$$

- Fiher would report: $p = 0.021$,

Differences

• We have seen very disturbing *essential* differences among the main schools of statistics to approach testing problems: Neyman-Pearson (α), Fisher (p) and Bayesian (Bayes factors or posterior probabilities). These differences are nicely explained in a historical context in J. O. Berger 2003

An example

Suppose a basic test $H_0 : \mu = 0$ with σ known with $n = 10$ and test statistic

$$z = \sqrt{n}\bar{x}/\sigma = 2.3.$$

- Fiher would report: $p = 0.021$,
- Neyman-Pearson would report $\alpha = 0.05$,

Differences

• We have seen very disturbing *essential* differences among the main schools of statistics to approach testing problems: Neyman-Pearson (α), Fisher (p) and Bayesian (Bayes factors or posterior probabilities). These differences are nicely explained in a historical context in J. O. Berger 2003

An example

Suppose a basic test $H_0 : \mu = 0$ with σ known with $n = 10$ and test statistic

$$z = \sqrt{n}\bar{x}/\sigma = 2.3.$$

- Fiher would report: $p = 0.021$,
- Neyman-Pearson would report $\alpha = 0.05$,
- Jeffreys would report $Pr(H_0 \mid data) = 0.28$.

Possible solution

Agreement is still possible

Surprisingly there is a simple solution that would satisfy all sorts of statisticians and is very easy to calculate.

Possible solution

Agreement is still possible

Surprisingly there is a simple solution that would satisfy all sorts of statisticians and is very easy to calculate. It is based on:

$$\underline{\alpha} = \frac{1}{1 - \frac{1}{e\, p \log(p)}}$$

Possible solution

Agreement is still possible

Surprisingly there is a simple solution that would satisfy all sorts of statisticians and is very easy to calculate. It is based on:

$$\underline{\alpha} = \frac{1}{1 - \frac{1}{e\, p \log(p)}}$$

- This value was proposed by Sellke, Bayarri, and J. O. Berger 2001 and is a lower bound (but also an accurate approximation) on data-based frequentist type-I error probabilities.

Possible solution

Agreement is still possible

Surprisingly there is a simple solution that would satisfy all sorts of statisticians and is very easy to calculate. It is based on:

$$\underline{\alpha} = \frac{1}{1 - \frac{1}{e\, p \log(p)}}$$

- This value was proposed by Sellke, Bayarri, and J. O. Berger 2001 and is a lower bound (but also an accurate approximation) on data-based frequentist type-I error probabilities.
- So for instance, associated with $p = 0.05$, you obtain $\underline{\alpha} = 0.29$. This means that rejecting H_0 in experiments with $p \approx 0.05$ implies a type-I error probability of at least 29%.

Possible solution

Agreement is still possible

Surprisingly there is a simple solution that would satisfy all sorts of statisticians and is very easy to calculate. It is based on:

$$\underline{\alpha} = \frac{1}{1 - \frac{1}{e\, p \log(p)}}$$

- This value was proposed by Sellke, Bayarri, and J. O. Berger 2001 and is a lower bound (but also an accurate approximation) on data-based frequentist type-I error probabilities.
- So for instance, associated with $p = 0.05$, you obtain $\underline{\alpha} = 0.29$. This means that rejecting H_0 in experiments with $p \approx 0.05$ implies a type-I error probability of at least 29%.
- Alternatively, $\underline{\alpha}$ can be interpreted in a Bayesian fashion:

$$\text{Bayes Factor} = B_{A0} \leq \frac{1}{\underline{\alpha}} - 1, \text{ or equiv. if } Pr(H_0) = Pr(H_A), \text{ then } Pr(H_0 \mid data) \geq \underline{\alpha}$$

- In the example with $p = 0.05$ you can interpret $B_{A0} \leq 2.45$ or $Pr(H_0 \mid data) \geq 0.29$.

Almost finished!

Conclusion

The regular adoption of the "-eplog(p) rule" could be of much help in reducing the impact on replicability issues of p values since it dramatically disminishes their argued prediposition towards declaring positive findings.

References

- Agulló, J., V. Carratalá, and J. Gimeno (1999). *Inferencia estadística para economía y empresa*. Textos docentes. Publicaciones de la Universidad de Alicante.
- Berger, J. O. (2003). "Could Fisher, Jeffreys and Neyman have agreed on Testing?" In: *Statistical Science* 18, pp. 1–32.
- – (2015). "Invited contribution to: "BANNING NULL HYPOTHESIS SIGNIFICANCE TESTING"". In: *The ISBA Bulletin* 22.1, p. 5.
- Berger, James O. and T. Sellke (1987). "Testing a Point Null Hypothesis: The Irreconcialability of P-values and Evidence". In: *Journal of American Statistical Association* 82, pp. 112–122.
- Hogben, L. (1957). *Statistical theory*. Read Books.
- Hubbard, R. and M. J. Bayarri (2003). "Confusion over measures of evidence (p's) versus errors (alpha's) in classical statistical testing". In: *The American Statistician* 57.171-182.
- Jefferys, W.H. (1990). "Bayesian Analysis of Random Event Generator Data". In: *Journal of Scientific Exploration* 4.2, pp. 153–169.

Bibliography II

- Jeffreys, Harold (1961). *Theory of Probability*. 3rd. Oxford University Press.
- Johnson, R.A. and G.K. Bhattacharyya (1992). *Statistics Principles and Methods (2nd ed)*. John Wiley & Sons.
- Lang, J., K. Rothman, and C.I. Cann (1998). "That confounded p-value". In: *Epidemiology* 9.1, pp. 7–8.
- Matthews, R. *http://www2.isye.gatech.edu/ brani/isyebayes/bank/pvalue.pdf*.
- Sellke, T., M. J. Bayarri, and J. O. Berger (2001). "Calibration of P-values for Testing Precise Null Hypotheses". In: *The American Statistician* 55, pp. 62–71.
- Velez, R. and A. García (1993). *Principios de inferencia estadística*. UNED.