# Exposing the P value fallacy to young residents

Piersante Sestini and Stefania Rossi

School of Medicine, University of Siena

5th International Conference
of Evidence-Based Health Care Teachers & Developers
Taormina, October 29, 2009

- The fallacy underlying the widespread use of the P value as a tool for inference in critical appraisal is well known, still little is done in the education system to correct it

    - Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999;130:995-1004.
    - Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol. 2008;45:135-40

- We used a questionnaire to ask 16 residents in respiratory diseases (age range 26-38 yrs, average 30) about their use and knowledge of the P value.

- We devised a simple intervention to improve understanding of the p value and to promote the use of the Bayes Factor

    - Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med. 1999;130:1005-13

# The *P* value fallacy

- The fallacy underlying the widespread use of the P value as a tool for inference in critical appraisal is well known, still little is done in the education system to correct it

  - Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999;130:995-1004.
  - Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol. 2008;45:135-40

- We used a questionnaire to ask 16 residents in respiratory diseases (age range 26-38 yrs, average 30) about their use and knowledge of the P value.

- We devised a simple intervention to improve understanding of the p value and to promote the use of the Bayes Factor

  - Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med. 1999;130:1005-13

## The *P* value fallacy

- The fallacy underlying the widespread use of the P value as a tool for inference in critical appraisal is well known, still little is done in the education system to correct it
    - Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999;130:995-1004.
    - Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol. 2008;45:135-40

- We used a questionnaire to ask 16 residents in respiratory diseases (age range 26-38 yrs, average 30) about their use and knowledge of the P value.

- We devised a simple intervention to improve understanding of the p value and to promote the use of the Bayes Factor

    - Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med. 1999;130:1005-13

## The *P* value fallacy

- The fallacy underlying the widespread use of the P value as a tool for inference in critical appraisal is well known, still little is done in the education system to correct it
  - Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999;130:995-1004.
  - Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol. 2008;45:135-40

- We used a questionnaire to ask 16 residents in respiratory diseases (age range 26-38 yrs, average 30) about their use and knowledge of the P value.

- We devised a simple intervention to improve understanding of the p value and to promote the use of the Bayes Factor
  - Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med. 1999;130:1005-13

# Does the P value matters to you?

| | |
|---|---|
| Do you know enough about the P value? | 93% |
| Average patients per day | $8 \pm 1$ |
| Average articles read per week | $4 \pm 1$ |
| Do you see the P as the main parameter of a result? | 75% |
| Average drug reps contacts per week | $3.2 \pm .3$ |
| Do the drug reps present sponsored research? | 100% |
| Do the reps present the P value as the main result? | 93% |

## Multiple answers accepted

1. If there is no difference, the probability of getting this or a more extreme result is 5%
2. The probability that there is no difference is 5%
3. The probability that the groups are different is 95%
4. The minimum difference between the groups is 5%
5. If repeating the experiment, the probability of obtaining again the same result is 5%
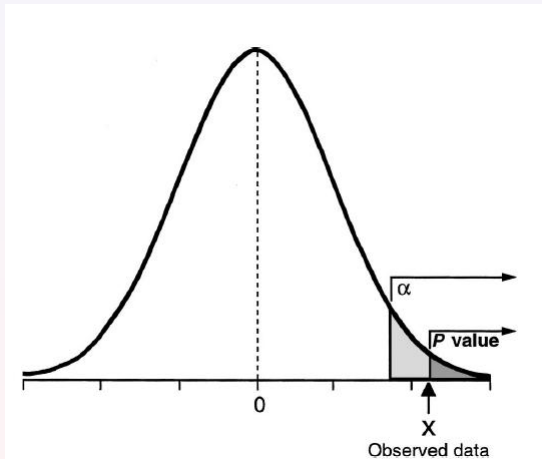6. I have no idea

### Multiple answers accepted

1. If there is no difference, the probability of getting this or a more extreme result is 5%

2. The probability that there is no difference is 5%

3. The probability that the groups are different is 95%

4. The minimum difference between the groups is 5%

5. If repeating the experiment, the probability of obtaining again the same result is 5%

6. I have no idea

The probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.

## Possible answers

| | |
|---|---|
| The probability if there is no difference is 5% | **2** |
| The probability that there is no difference is 5% | **7** |
| The probability that there is a difference is 95% | **11** |
| The minimum difference is 5% | 1 |
| The probability of obtaining again this result is 5% | 2 |
| I have no idea | 2 |

## What a P of 0.05 means...

- Imagine that you came home, and the light is on.

- You want to guess whether someone is in when the light is on

- A P of 0.05 is like knowing that when nobody is in (the null hypothesis), the light is on only 5% of the times

- Does this mean that there is a 95% probability that somebody is in?

## What a P of 0.05 means...

- Imagine that you came home, and the light is on.

- You want to guess whether someone is in when the light is on

- A P of 0.05 is like knowing that when nobody is in (the null hypothesis), the light is on only 5% of the times

- Does this mean that there is a 95% probability that somebody is in?

### What a P of 0.05 means...

- Imagine that you came home, and the light is on.
- You want to guess whether someone is in when the light is on
- A P of 0.05 is like knowing that when nobody is in (the null hypothesis), the light is on only 5% of the times
- Does this mean that there is a 95% probability that somebody is in?

## What a P of 0.05 means...

- Imagine that you came home, and the light is on.
- You want to guess whether someone is in when the light is on
- A P of 0.05 is like knowing that when nobody is in (the null hypothesis), the light is on only 5% of the times
- Does this mean that there is a 95% probability that somebody is in?

# Example

## Is somebody in?

|  | No | Yes |  |
|---|---|---|---|
| **Off** | 95 | ? | |
| **On** | 5 | ? | ?/(5+?) |
| P → | 5% | | but you want this↑ |

- The P is the probability of finding the light on when nobody's in (L|P)
- But you want the probability that nobody is in when the light is on (P|L)

## Is somebody in?

|  | No | Yes |  |
|:---:|:---:|:---:|:---:|
| **Off** | 95 | ? |  |
| **On** | 5 | ? | ?/(5+?) |
| P → | 5% |  | but you want this↑ |

- The P is the probability of finding the light on when nobody's in (L|P)
- But you want the probability that nobody is in when the light is on (P|L)

**Is somebody in?**

|  | No | Yes |  |
|:---:|:---:|:---:|:---:|
| **Off** | 95 | ? |  |
| **On** | 5 | ? | ?/(5+?) |
| P → | 5% |  | but you want this↑ |

- The P is the probability of finding the light on when nobody's in (L|P)
- But you want the probability that nobody is in when the light is on (P|L)

Minimum BF= %TP/%FP $= e^{Z^2/2} = e^{1.96^2/2} = 6.8$

### Is somebody in?

|        | No   | Yes            |                  |
|--------|------|----------------|------------------|
| **Off** | 95   | $100 - 32 = 68$ |                  |
| **On**  | 5    | $5 * 6.8 = 32$ | $32/37 = 86\%$   |
|        | 100  | 100            |                  |
| P $\rightarrow$ | 5% | 32%     |                  |

# Example

## Is somebody in?

|       | No   | Yes  |              |
|-------|------|------|--------------|
| **Off** | 95   | 20   |              |
| **On**  | 5    | 80   | 80/85= 94%   |
|       | 100  | 100  |              |
| P ->  | 5%   | 80%  |              |

- In many experiments, the statistical power is selected to have a false negative rate ($\beta$) of 20% and a false positive rate ($\alpha$) of 5% (under certain assumptions *a priori*)

## Is somebody in?

|  | No | Yes |  |
|:---:|:---:|:---:|:---:|
| **Off** | 95 | 20 |  |
| **On** | 5 | 80 | 80/85= 94% |
|  | 100 | 100 |  |
| P -> | 5% | 80% |  |

- In many experiments, the statistical power is selected to have a false negative rate ($\beta$) of 20% and a false positive rate ($\alpha$) of 5% (under certain assumptions *a priori*)

## Is somebody in?

|       | No  | Yes |             |
|-------|-----|-----|-------------|
| Off   | 95  | 2   |             |
| On    | 5   | 8   | 8/13= 62%   |
|       | 100 | 10  |             |
| P ->  | 5%  | 80% |             |

- α and β errors alone do not adjiust for low pre-test probabilities

## Is somebody in?

|  | No | Yes |  |
|---|---|---|---|
| **Off** | 95 | 2 |  |
| **On** | 5 | 8 | 8/13= 62% |
|  | 100 | 10 |  |
| P -> | 5% | 80% |  |

- α and β errors alone do not adjiust for low pre-test probabilities

## Post-intervention

- Many got confused by the abstraction
- In retrospect, a diagnosic example could have been better (e.g. nausea and pregnancy)
- After the course all except two identified the meaning of the P value
- All except one declared that in the future they would use the Bayes Factor instead.
- After 3 months, no one has ever done it

## Post-intervention

- Many got confused by the abstraction
- In retrospect, a diagnosic example could have been better (e.g. nausea and pregnancy)
- After the course all except two identified the meaning of the P value
- All except one declared that in the future they would use the Bayes Factor instead.
- After 3 months, no one has ever done it

## Post-intervention

- Many got confused by the abstraction
- In retrospect, a diagnosic example could have been better (e.g. nausea and pregnancy)
- After the course all except two identified the meaning of the P value
- All except one declared that in the future they would use the Bayes Factor instead.
- After 3 months, no one has ever done it

## Post-intervention

- Many got confused by the abstraction
- In retrospect, a diagnosic example could have been better (e.g. nausea and pregnancy)
- After the course all except two identified the meaning of the P value
- All except one declared that in the future they would use the Bayes Factor instead.
- After 3 months, no one has ever done it

## Post-intervention

- Many got confused by the abstraction
- In retrospect, a diagnosic example could have been better (e.g. nausea and pregnancy)
- After the course all except two identified the meaning of the P value
- All except one declared that in the future they would use the Bayes Factor instead.
- After 3 months, no one has ever done it

- The P value is frequently used by drug reps to promote the results of drug trials and by young residents for interpreting them.

- A simple, short intervention in a group with poor understanding of the P value is able to dramatically improve knowledge.

- However, such interventions are unlikely to have a significant impact, unless major changes occur in the medical community

- The P value is frequently used by drug reps to promote the results of drug trials and by young residents for interpreting them.

- A simple, short intervention in a group with poor understanding of the P value is able to dramatically improve knowledge.

- However, such interventions are unlikely to have a significant impact, unless major changes occur in the medical community

- The P value is frequently used by drug reps to promote the results of drug trials and by young residents for interpreting them.
- A simple, short intervention in a group with poor understanding of the P value is able to dramatically improve knowledge.
- However, such interventions are unlikely to have a significant impact, unless major changes occur in the medical community