Bad Stats are Miscommunicated Stats

Pierre Dragicevic, INRIA

www.aviz.fr/badstats

Stats

Your stats are wrong!

- Violation of statistical assumptions
- Use of too small samples
- No correction for multiple comparisons
- etc., etc.

The problem

- Stats have never been an exact science
- Many “deadly sins” only yield a moderate inflation of Type I error rates (should we really care?)
- Yes, serious mistakes are made. But:
  - Mistakes are part of the scientific process
  - Their cost is low if they are easy to detect
  - Perhaps 5% of articles have seriously flawed analyses
  - On the other hand, maybe 90% of articles have flaws of a different kind that are much harder to detect because **they are not recognized as such**

What are stats for?

The dominant mental model

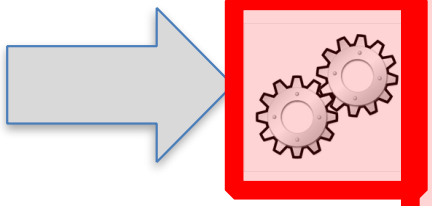
Experiment data

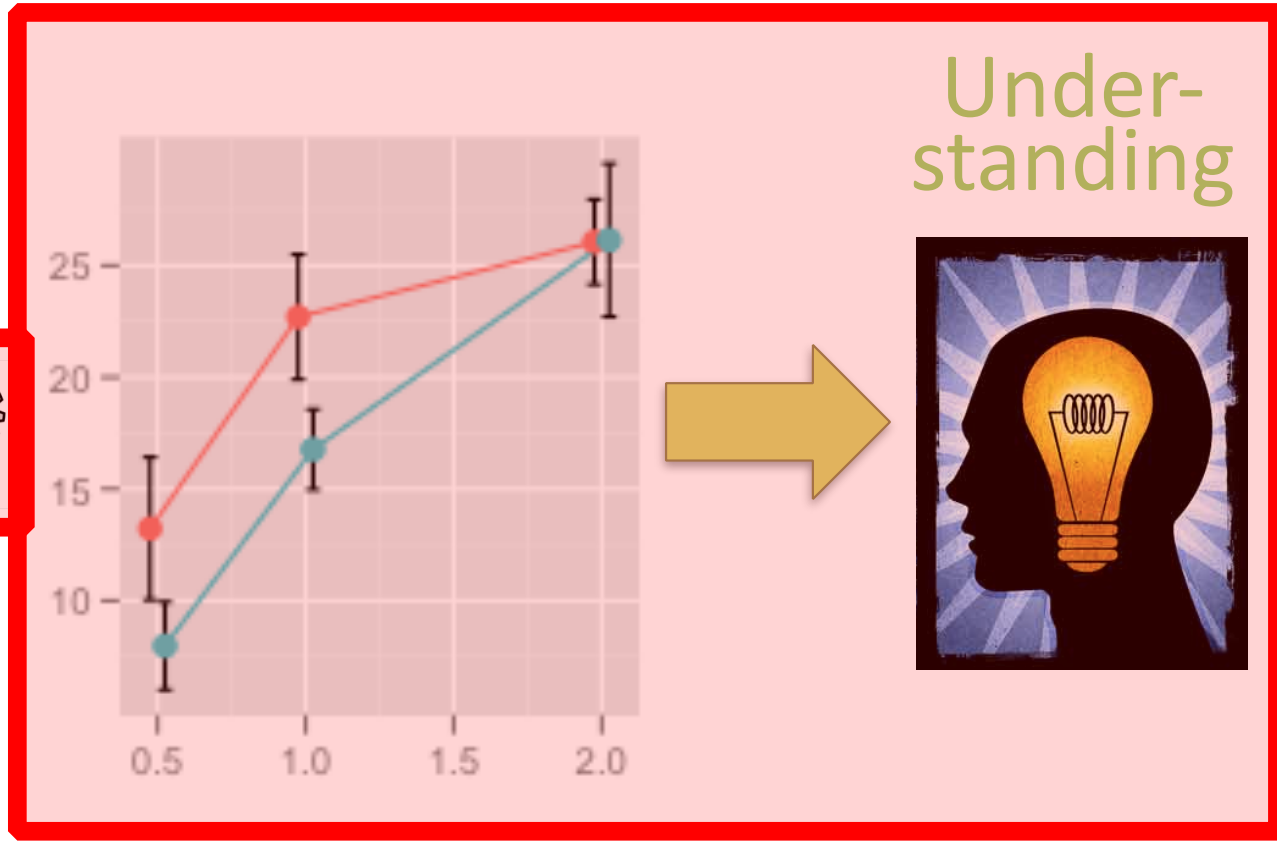
There is an effect

There is no effect

What stats really are

Experiment data

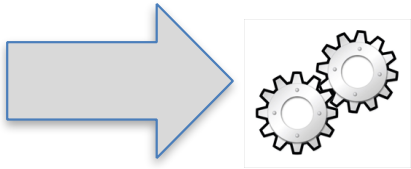


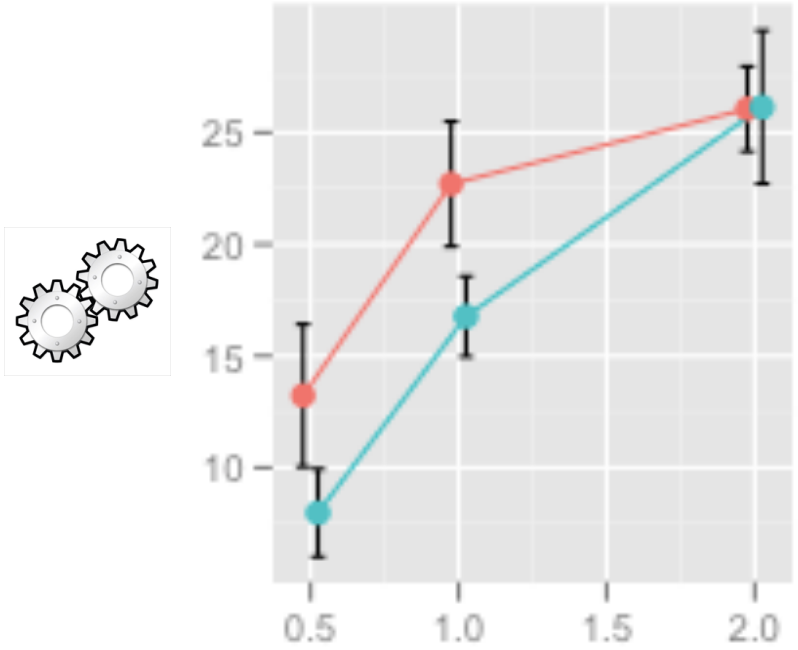


Under-standing

What stats really are

Experiment data





Under-standing

What stats really are

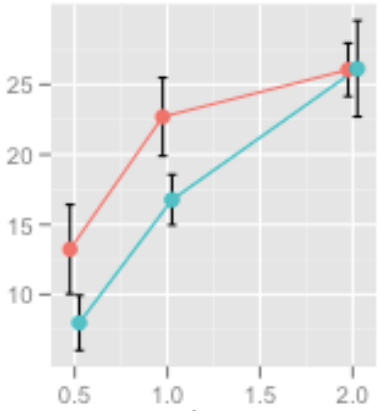
Under-
standing

Investigator

What stats really are

Under-
standing

Investigator



Publication

Understanding

Peers

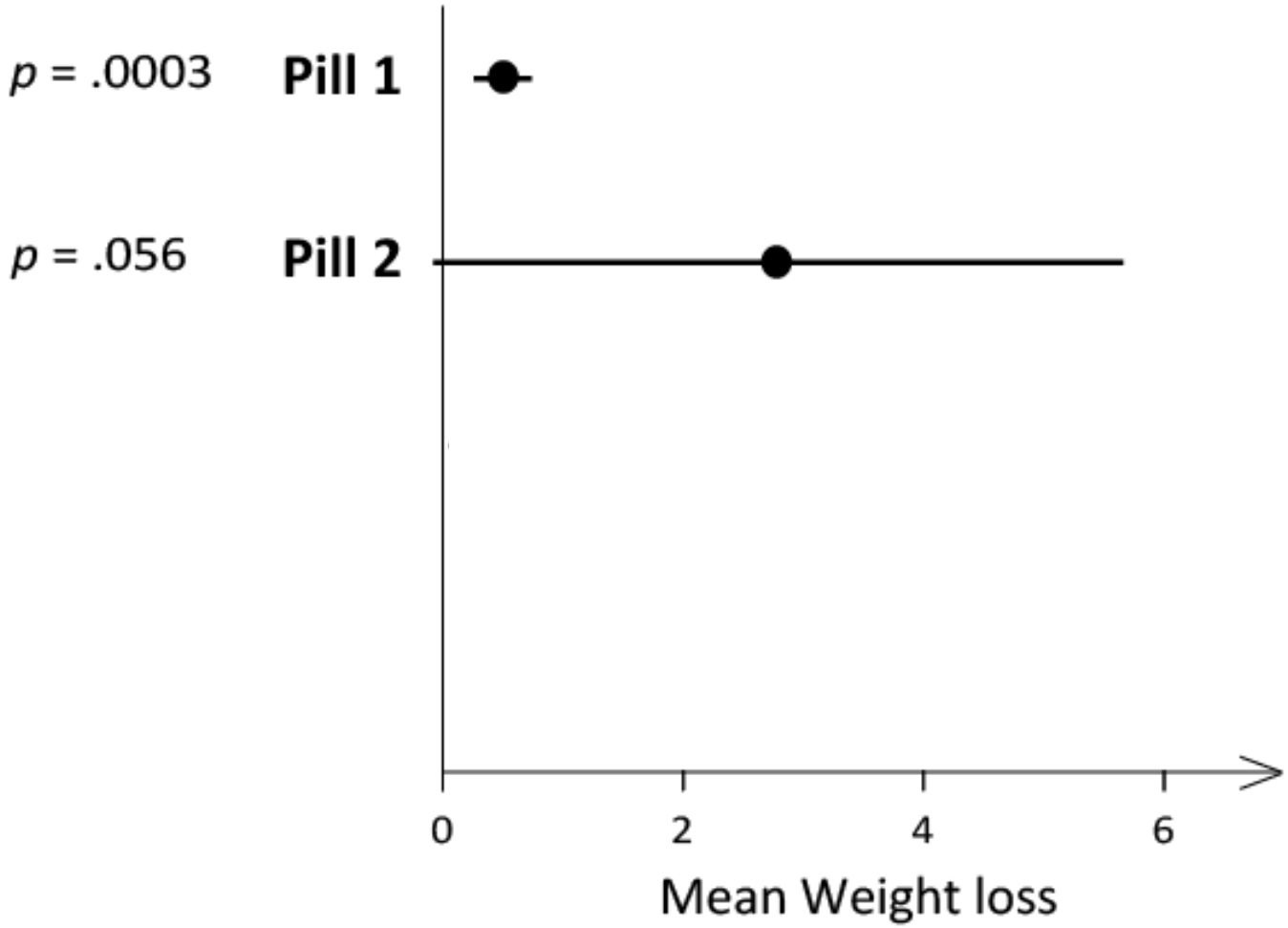
Bad stats are miscommunicated stats

- Reporting the wrong things
  - Irrelevant information
  - Misleading information
  - Often both
- Dichotomous thinking
  - « The tyranny of the discontinuous mind » (Dawkins, 2011)
  - Lots of useful information thrown away
  - Misleading: illusion of objectivity, certainty, exactness

What is good stats communication?

Suppose your best friend wants to loose weight

Which weight-loss pill would you recommend?

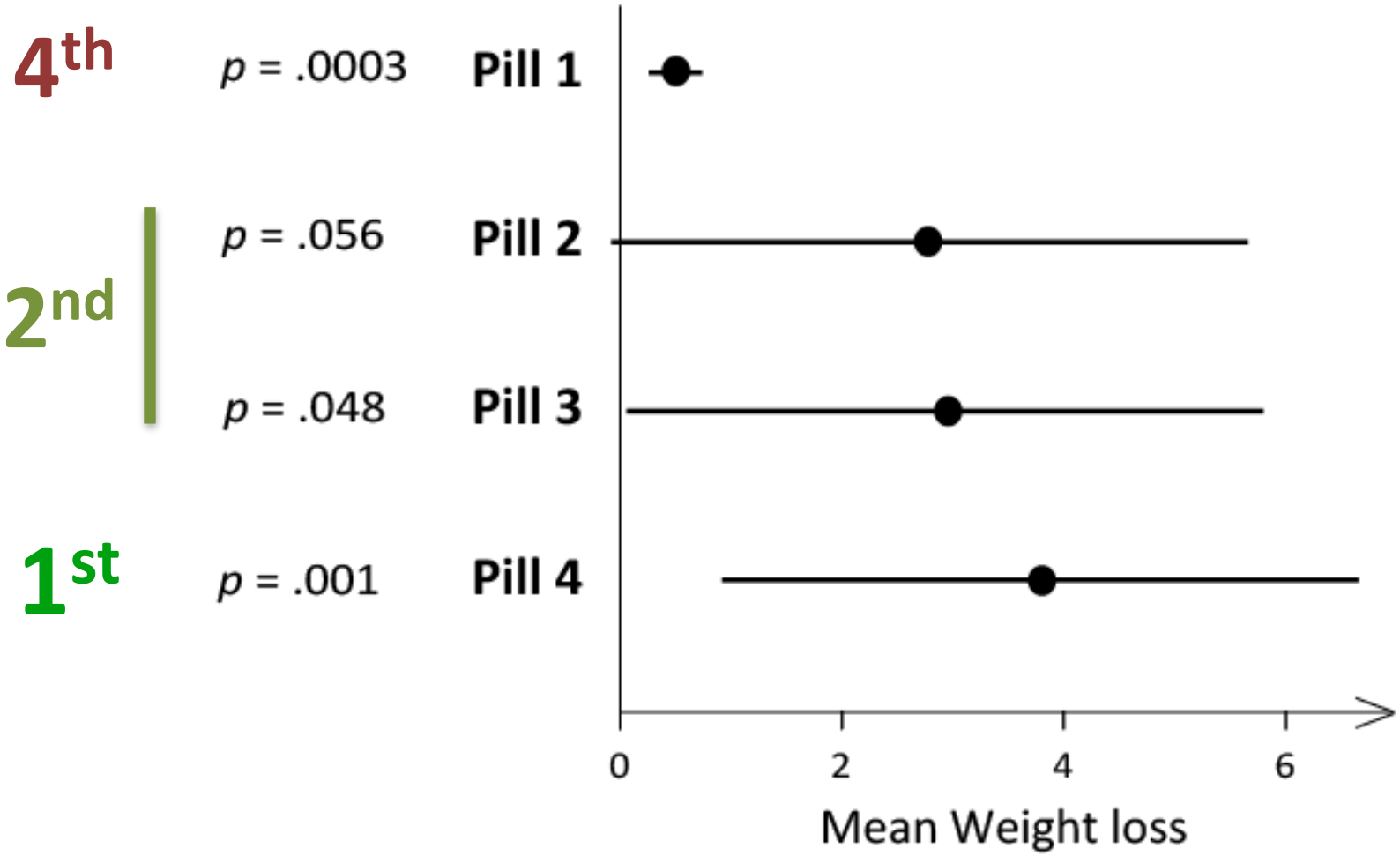


Error bars are 95% CIs

p-values are based on a null hypothesis of no effect

Adapted from
(Ziliak and McCloskey, 2009)

Which weight-loss pill would you recommend?

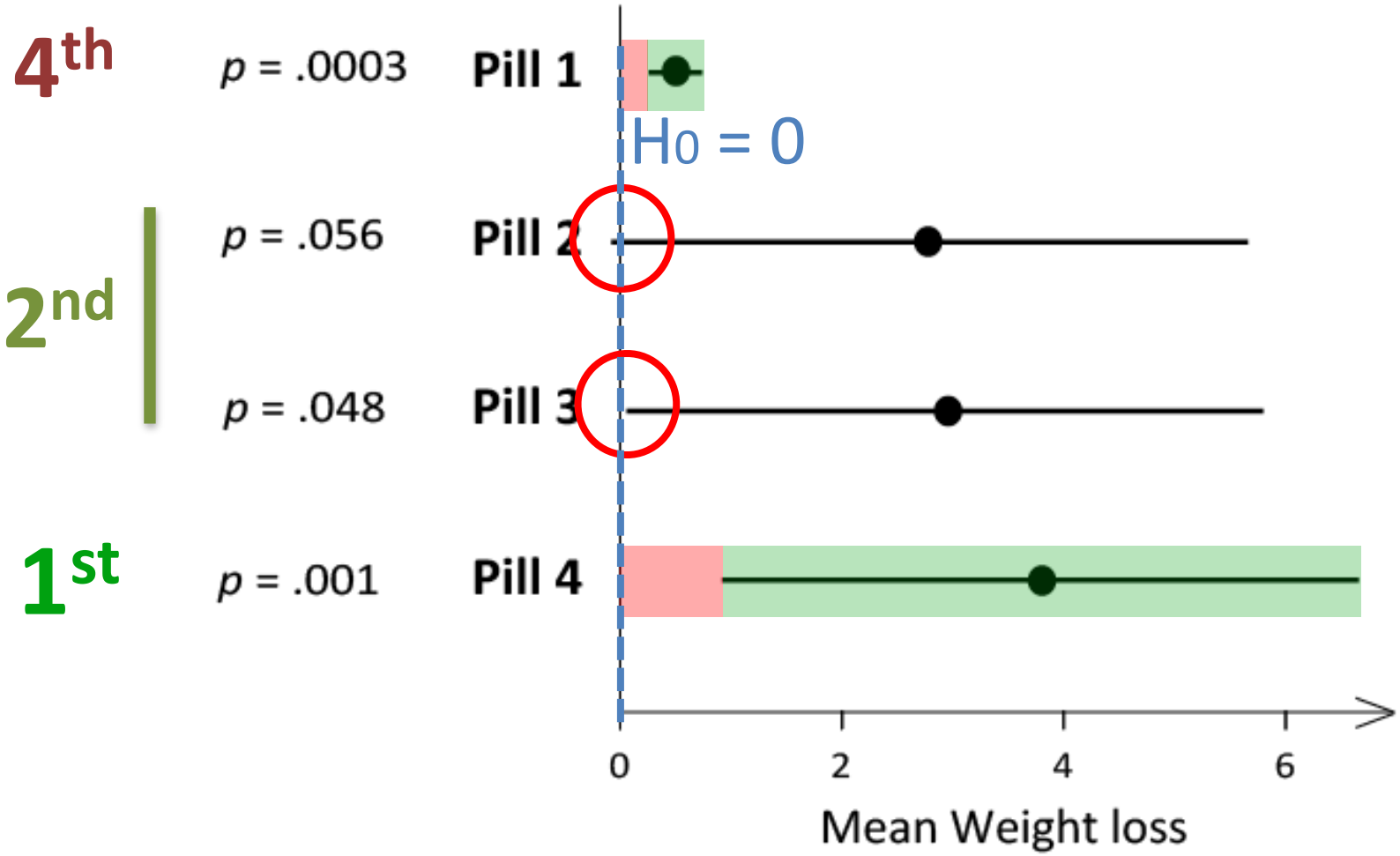


Error bars are 95% CIs

p-values are based on a null hypothesis of no effect

Adapted from
(Ziliak and McCloskey, 2009)

Which weight-loss pill would you recommend?

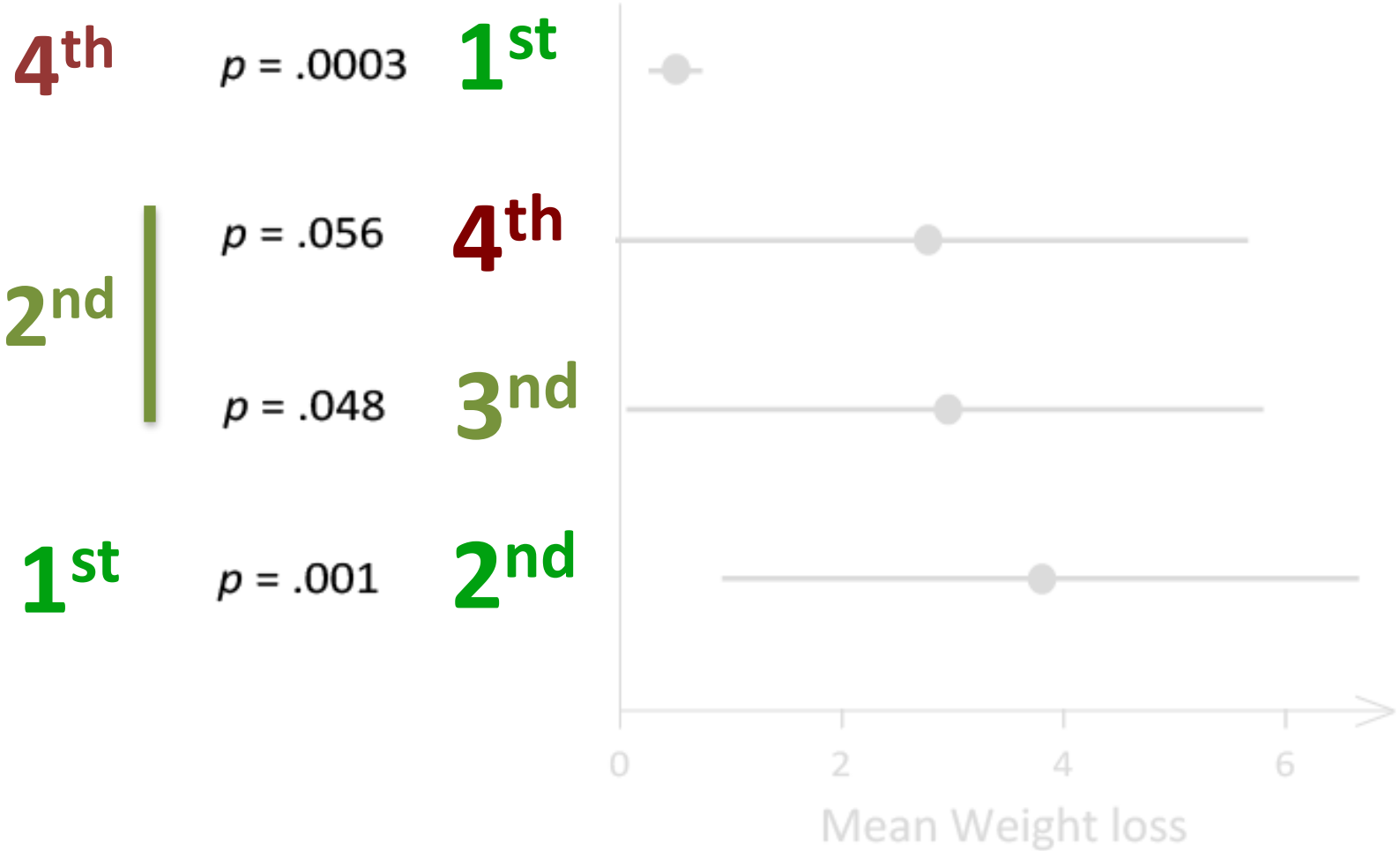


Error bars are 95% CIs

p-values are based on a null hypothesis of no effect

Adapted from
(Ziliak and McCloskey, 2009)

Which weight-loss pill would you recommend?



Error bars are 95% CIs

p-values are based on a null hypothesis of no effect

Adapted from
(Ziliak and McCloskey, 2009)

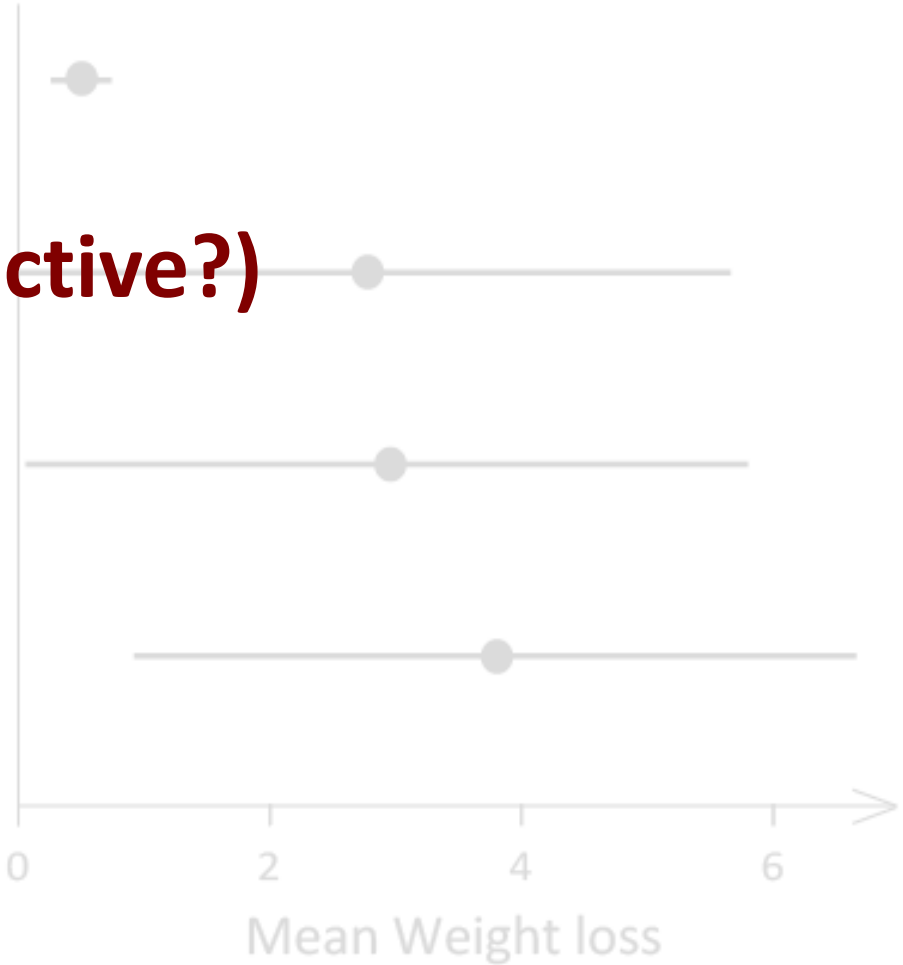
Which weight-loss pill would you recommend?

4th Effective

2nd n.s. (Ineffective?)

Effective

1st Effective

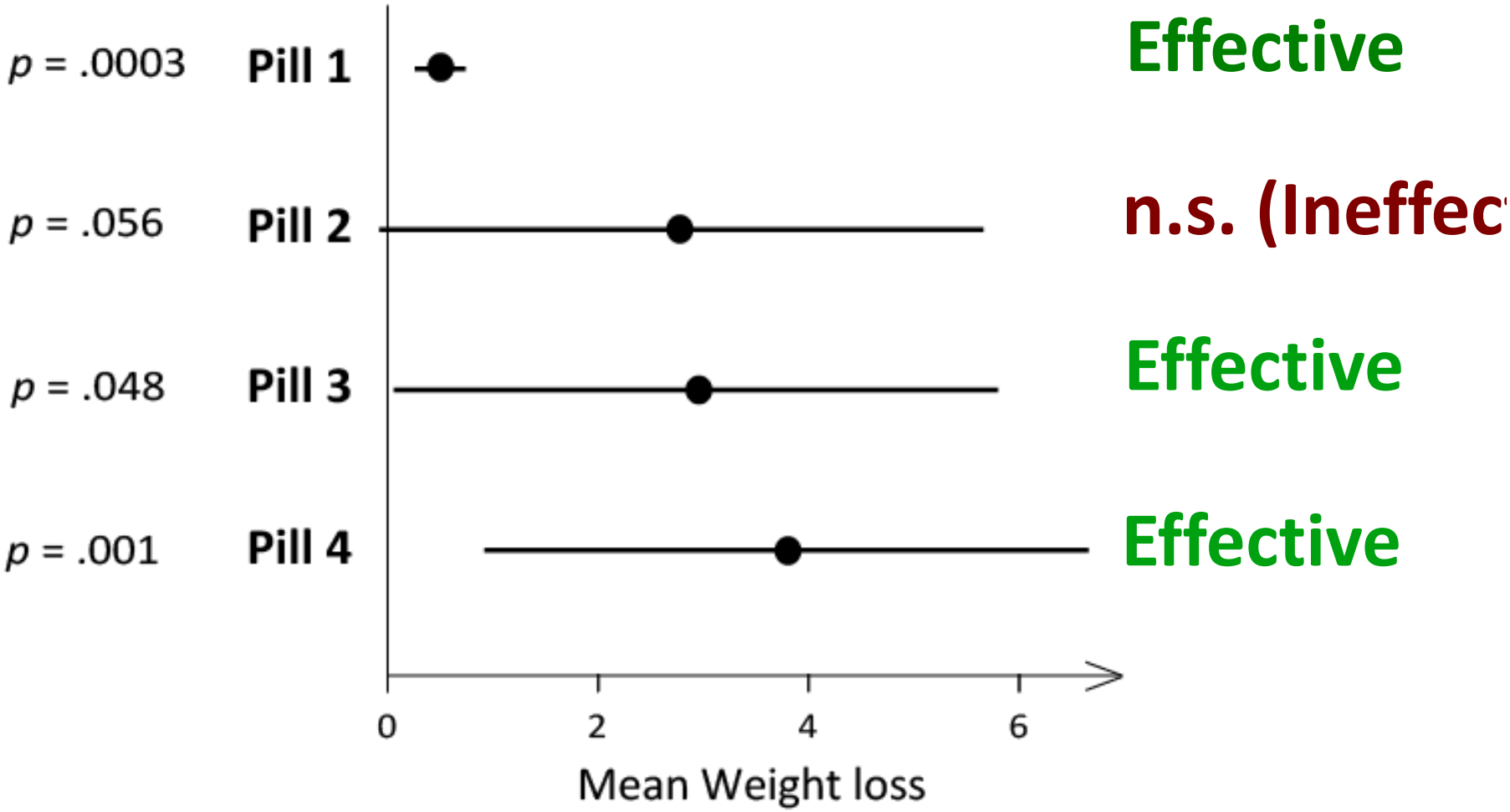


Error bars are 95% CIs

p-values are based on a null hypothesis of no effect

Adapted from
(Ziliak and McCloskey, 2009)

Which weight-loss pill would you recommend?



Error bars are 95% CIs

p-values are based on a null hypothesis of no effect

Adapted from
(Ziliak and McCloskey, 2009)

NHST Criticism

*" [NHST] is based upon a **fundamental misunderstanding** of the nature of rational inference, and is seldom if ever appropriate to the aims of scientific research. "*

(Rozeboom, 1960)

NHST Criticism

*" Statistical significance is perhaps the **least important attribute of a good experiment**; it is never a sufficient condition for claiming that a theory has been usefully corroborated, that a meaningful empirical fact has been established, or that an experimental report ought to be published. "*

(Likken, 1968)

NHST Criticism

*" [there are] more than **300 articles** in different disciplines about the indiscriminate use of NHST […]*

*After review of the debate about NHST, I argue that the criticisms have sufficient merit to support the **minimization or elimination** of NHST. "*

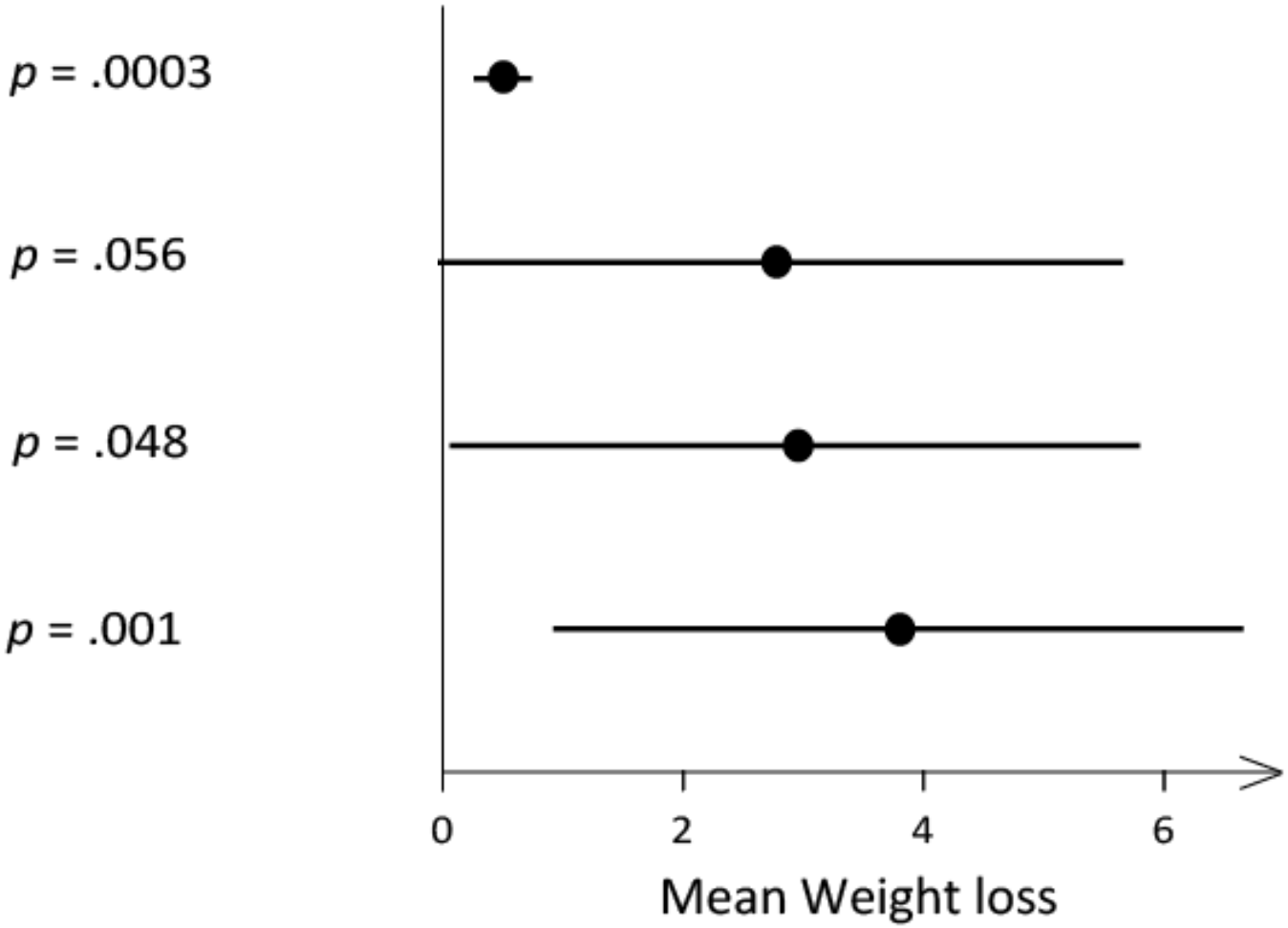
(Kline, 2004)

NHST Criticism

*" No scientific worker has a **fixed level of significance** at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas. "*

(Fisher, 1956)

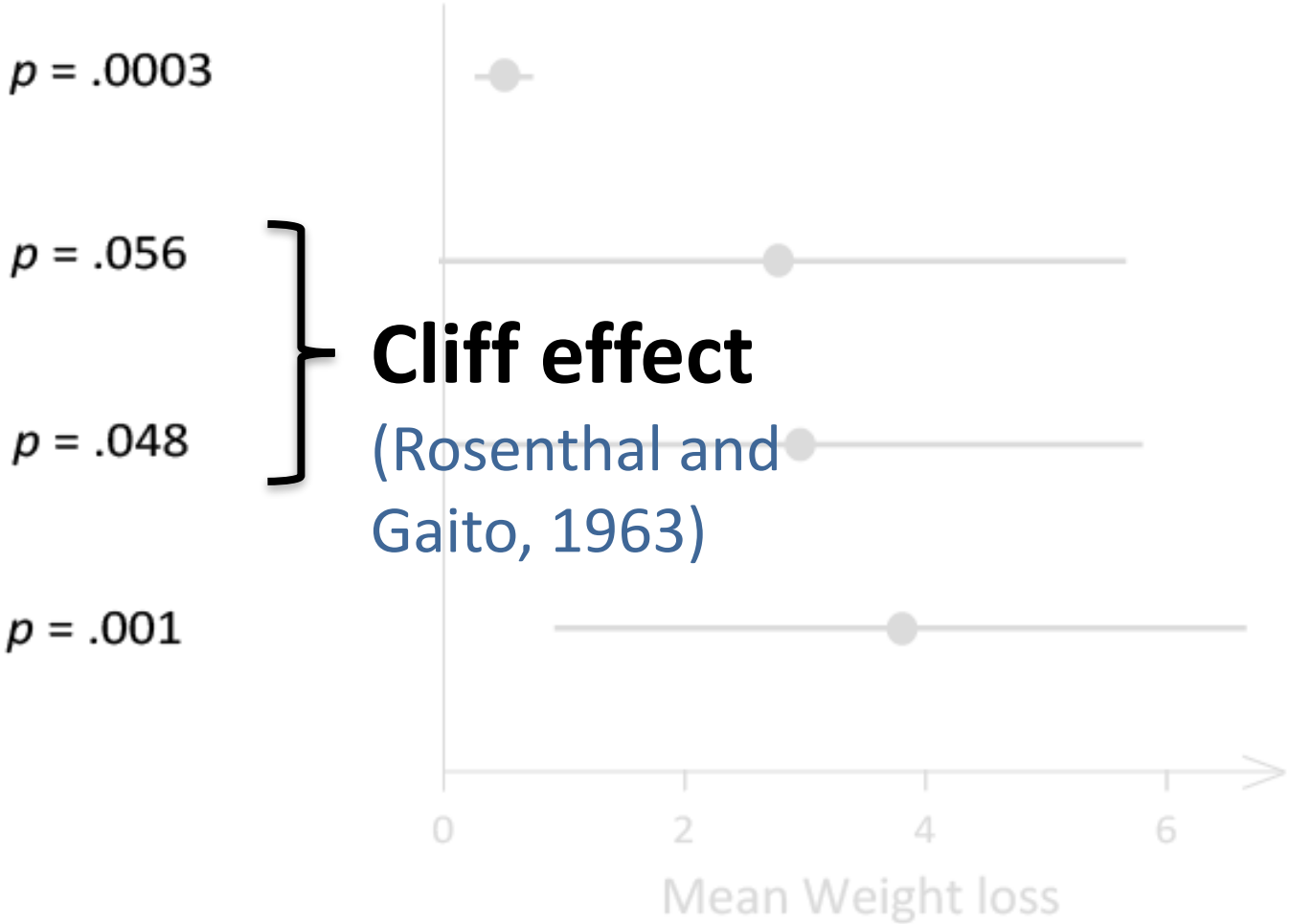
Which weight-loss pill would you recommend?



Error bars are 95% CIs

p-values are based on a null hypothesis of no effect

Which weight-loss pill would you recommend?



Error bars are 95% CIs

p-values are based on a null hypothesis of no effect

Fisher's error

" We have the duty of [...] communicating our conclusions in ***intelligible form****, in recognition of the right of other free minds to utilize them in making their own decisions. "*

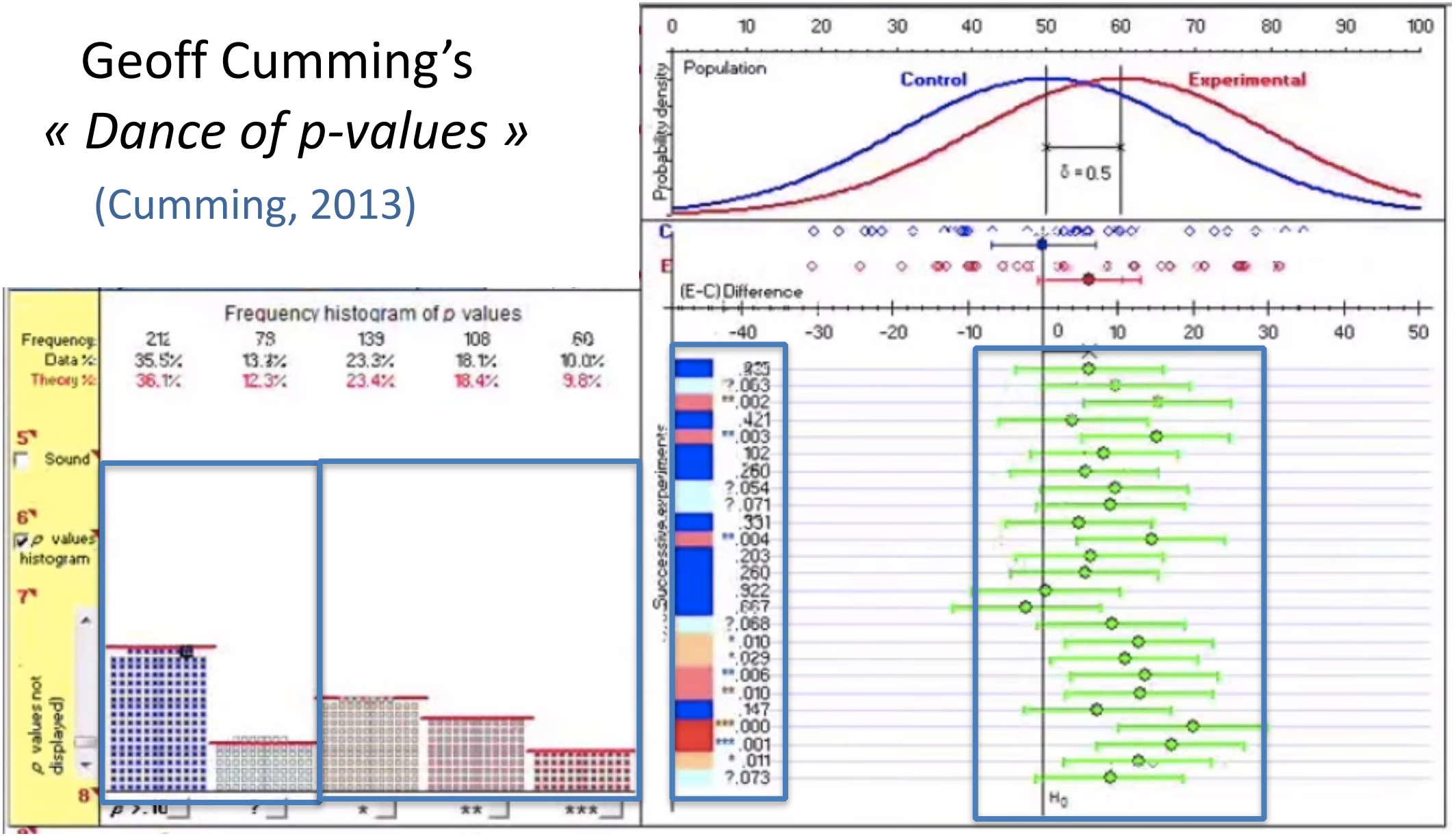
(Fisher, 1955)

p is highly unreliable

Geoff Cumming's

« *Dance of p-values* »

(Cumming, 2013)




$p < .001$
$p < .05$
n.s.

p is highly unreliable

p-intervals
(Cumming, 2008)

p_{obt}	
p_{obt}[a]	Two-sided p interval[c]
.001	(.0000002, .070)
.01	(.000006, .22)
.02	(.00002, .30)
.05	(.00008, .44)
.1	(.00027, .57)
.2	(.00099, .70)
.4	(.0040, .83)
.6	(.0098, .90)

p is highly unreliable

p-intervals
(Cumming, 2008)

p_{obt}	
p_{obt}[a]	Two-sided p interval[c]
.001	(.0000002, .070)
.01	(.000006, .22)
.02	(.00002, .30)
.05	(.00008, .44)
.1	(.00027, .57)
.2	(.00099, .70)
.4	(.0040, .83)
.6	(.0098, .90)

p is highly unreliable

effect of METHOD ($F_{4,44} = 10.1, p < 0.0001$
d $F_{3,33} = 49.1, p < 0.0001$) for both datasets
4) and a significant effect of SCALE for the data
ıt not for SCALE$\geq$ 4 ($F_{2,22} = 2.7, p = 0.0885$).
$= 0.1116$ and $F_{1,11} = 3.9, p = 0.0718$).
ractions of METHOD $\times$ W ($F_{12,132} = 6.1, p <$
$> < 0.0001$ and $F_{6,66} = 10.6, p < 0.0001$) for
SCALE $= 1$ in particular, we have a higher error
his difference vanishes as W increases. The
Mag with other methods. For the remaining
s in the error rates.

p is highly unreliable

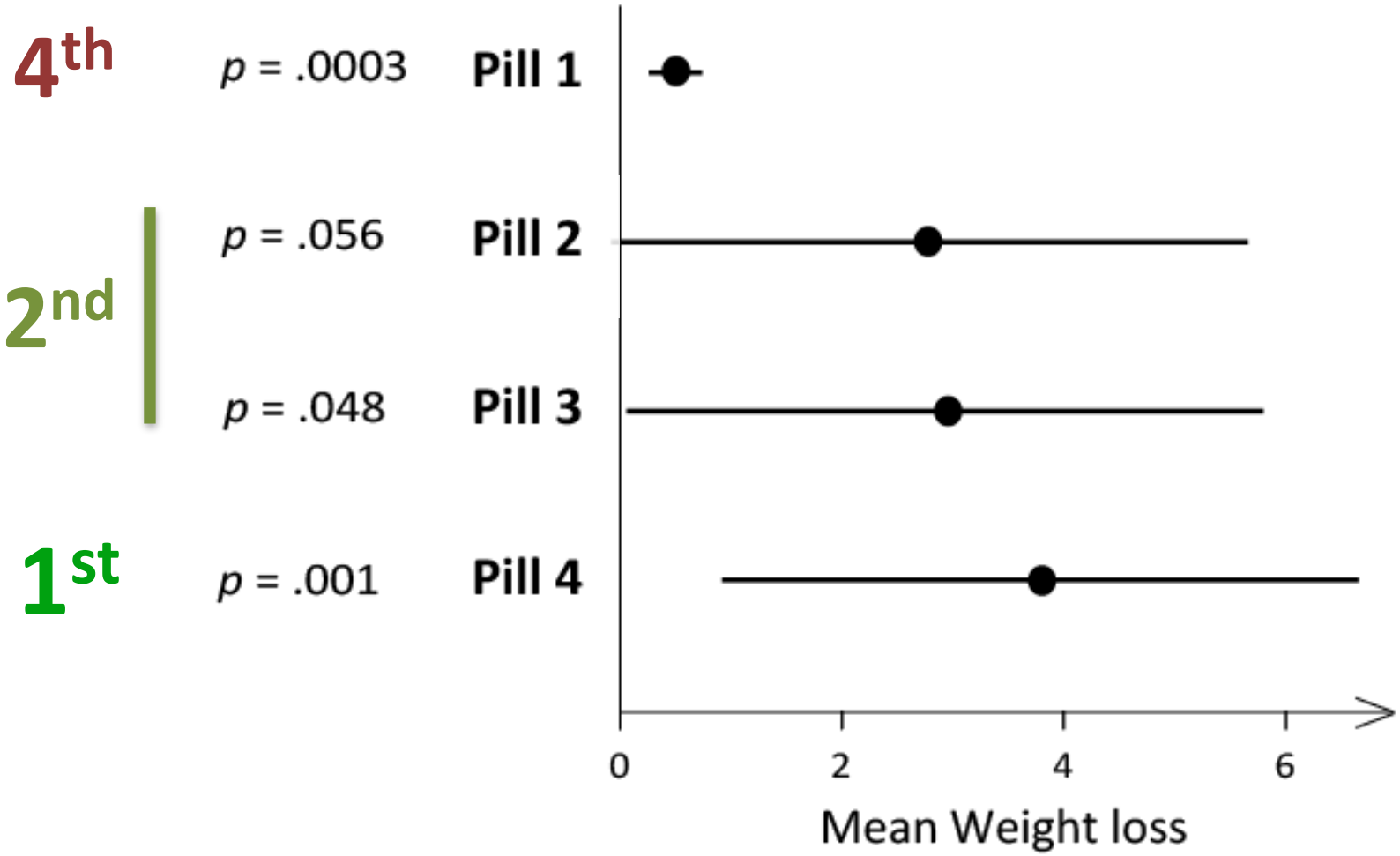
Running an HCI Experiment In Multiple **Parallel Universes**

Pierre Dragicevic Fanny Chevalier Stéphane Huot

INRIA - Université Paris-Sud - CNRS

What to report?

Which weight-loss pill would you recommend?



Error bars are 95% CIs

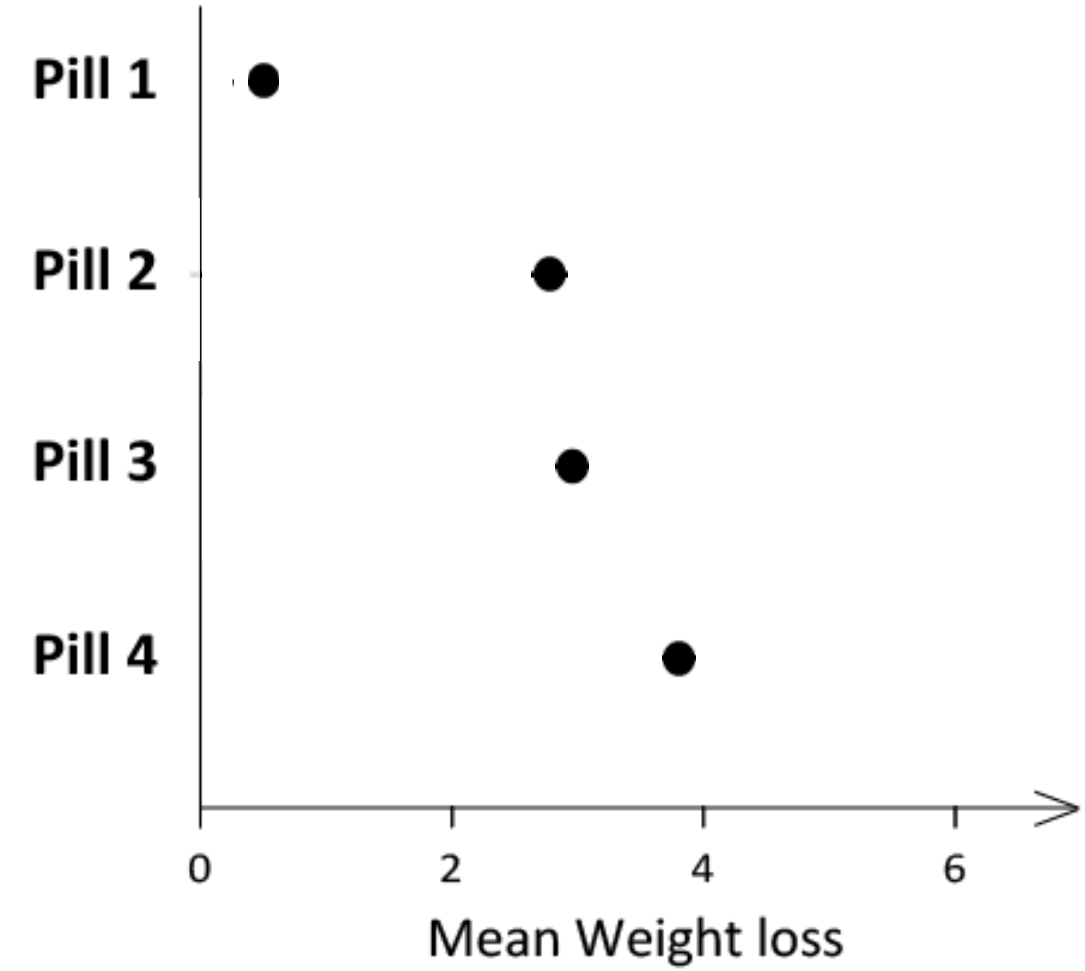
p-values are based on a null hypothesis of no effect

Which weight-loss pill would you recommend?

4th

2nd

1st



Error bars are 95% CIs

p-values are based on a null hypothesis of no effect

What's an effect size?

- Taken broadly, « *the amount of something that might be of interest* » (Cumming, 2011)
- E.g., writing « *vis A yields 1.2 times more insights than vis B* » is reporting an effect size
- Measures like Cohen's *d* are *standardized* effect sizes
- Many recommend reporting simple (unstandardized) effect sizes instead

What's an effect size?

" Only rarely will uncorrected standardized effect size be more useful than simple effect size. It is usually ***far better to report simple effect size*** [...] *"*

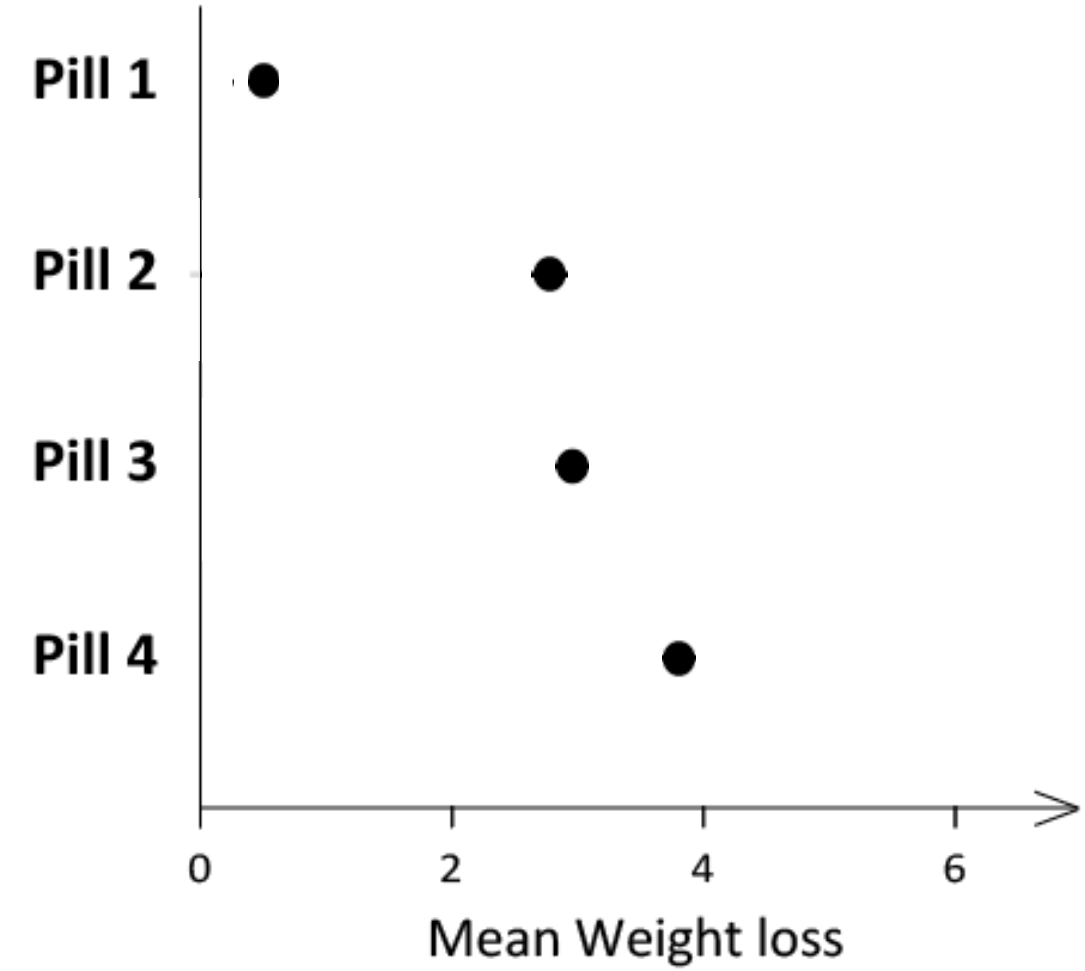
(Baguley, 2009)

Which weight-loss pill would you recommend?

4th

2nd

1st



Error bars are 95% CIs

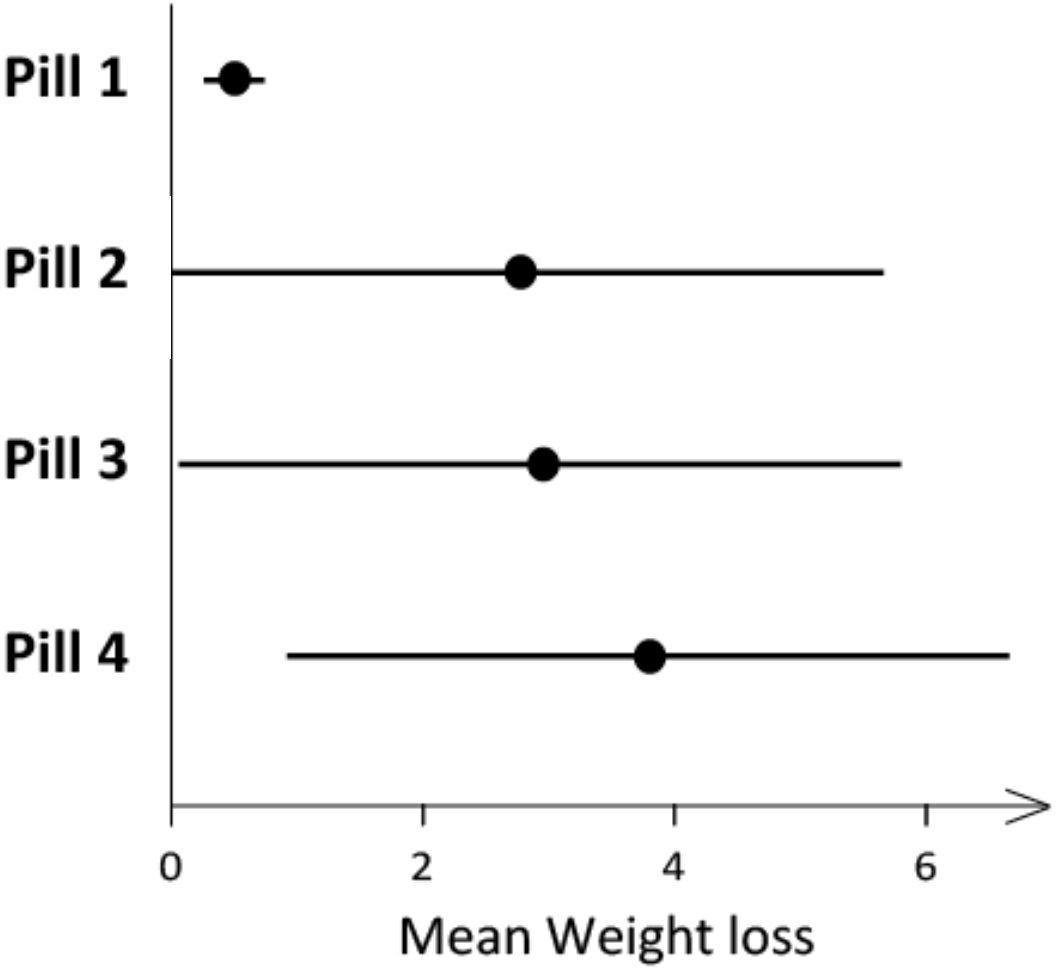
p-values are based on a null hypothesis of no effect

Which weight-loss pill would you recommend?

4th

2nd

1st



Error bars are 95% CIs

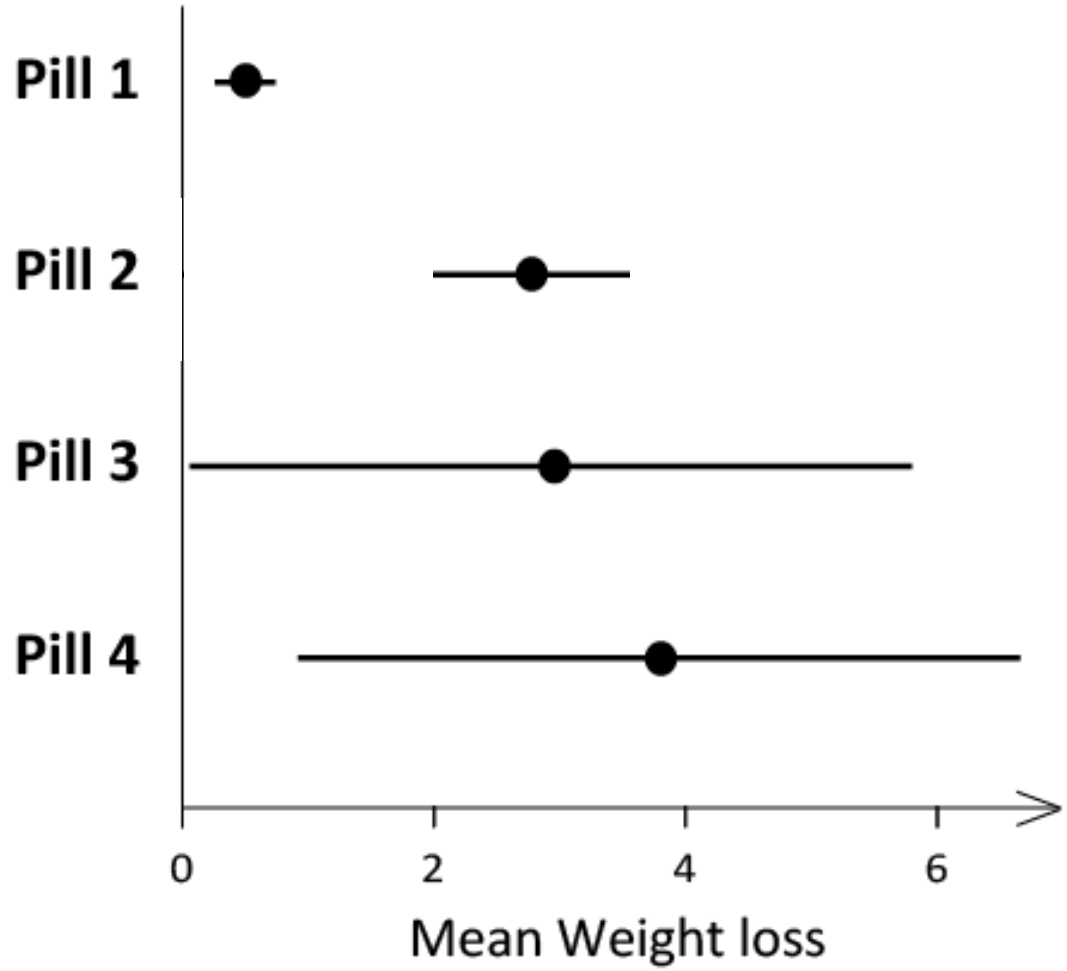
p-values are based on a null hypothesis of no effect

Which weight-loss pill would you recommend?

4th

2nd

1st

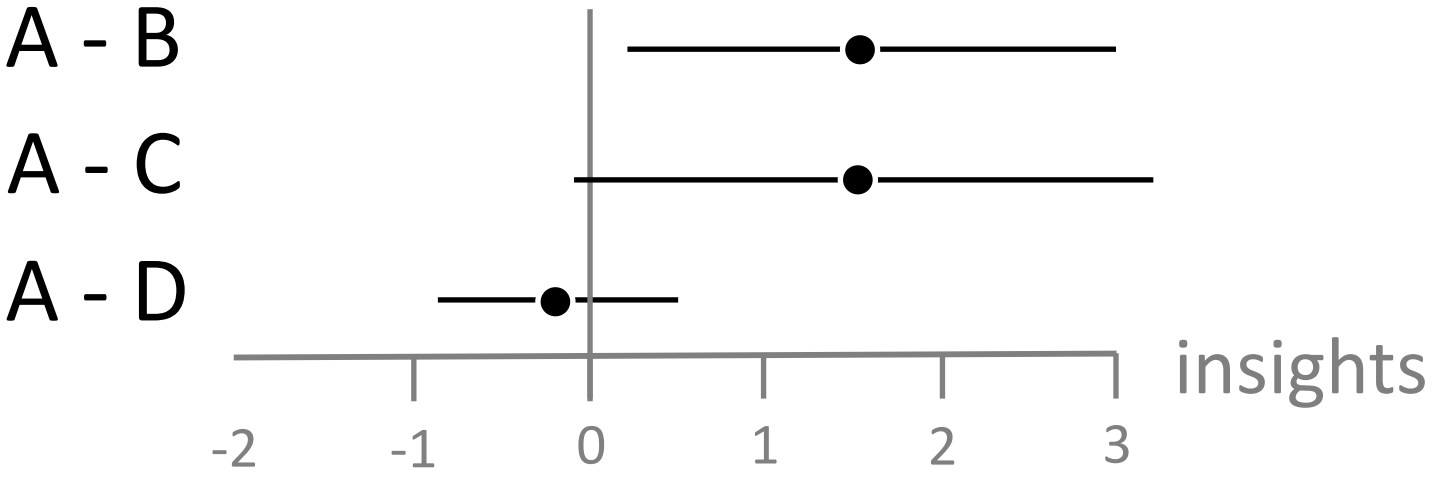


Error bars are 95% CIs

p-values are based on a null hypothesis of no effect

How to interpret CIs?

- « *A range of plausible values for μ. Values outside the CI are relatively implausible.* »
  (Cumming and Finch, 2005)



How to interpret CIs?

*“ It seems clear that **no** confidence interval should be interpreted as a a significance test.”*

(Schmidt and Hunter, 1997)

How to interpret CIs?

"It is best for individual researchers to present point estimates and confidence intervals and ***refrain from attempting to draw final conclusions*** *about research hypotheses ."*

(Schmidt and Hunter, 1997)

Do we want stats to be this?

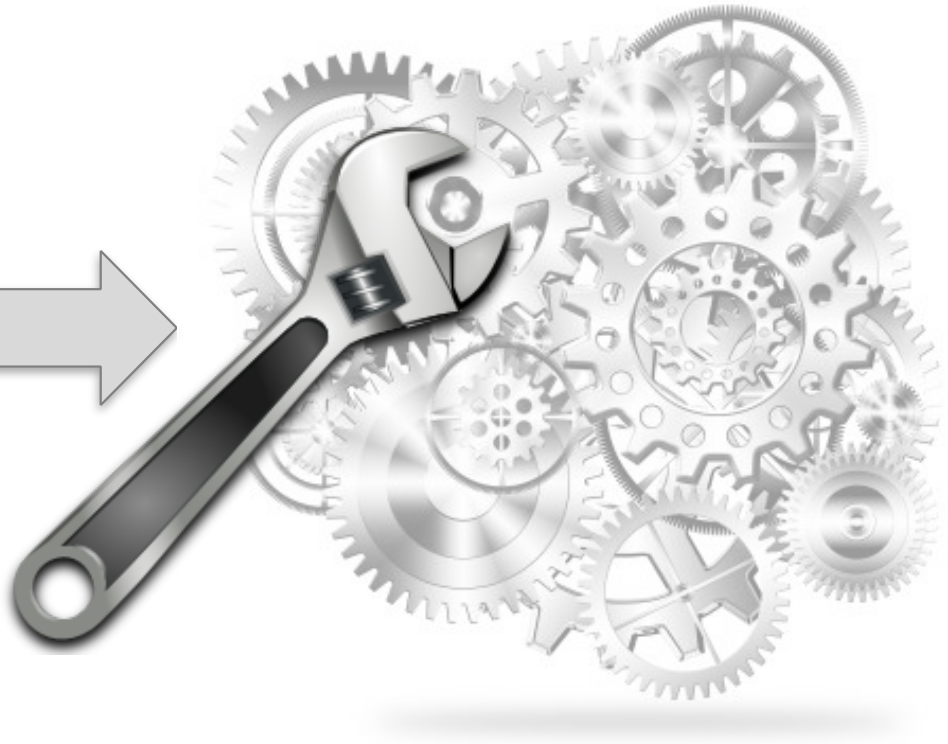
Experiment data

There is an effect

There is no effect

Do we want stats to be this?

Experiment data



There is an effect

There is no effect

Do we want stats to be this?

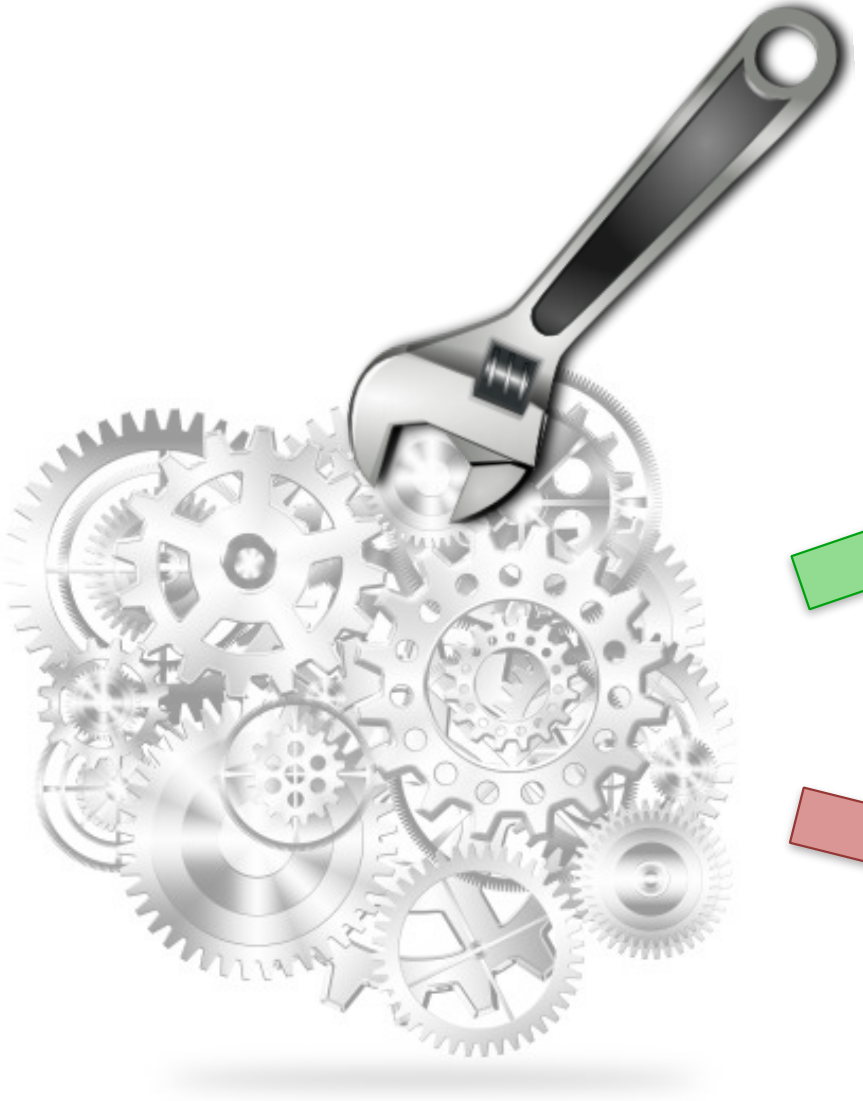
Experiment data

There is an effect

There is no effect

Do we want stats to be this?

Experiment data



There is an effect

There is no effect

Do we want stats to be this?

Experi
data

There is
an effect

There is
no effect

Do we want stats to be this?

Experi
data

There is
an effect

There is
no effect

Do we want stats to be this?

Experiment data

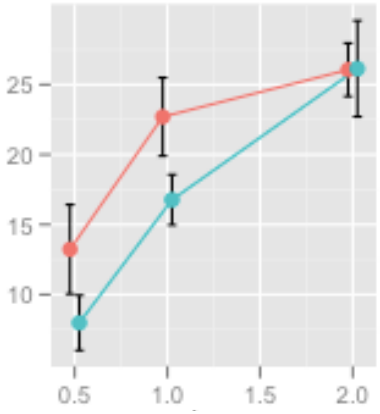
There is an effect

There is no effect

Or that?

Under-
standing

Investigator



Publication

Understanding

Peers

What to report?

- Effect sizes

What to report?

- Effect sizes
- Confidence intervals

What to report?

- Effect sizes
- Confidence intervals
- ~~p-values~~

What to report?

- Be pedagogical, use figures
- Be creative, but honest
- Use your judgment
- Protect yourself against cognitive biases
- Be nuanced, let your peers decide
- Seek transparency
- **Judge papers according to these merits!**

A last quote

“[Sciences] can only be successfully conducted by responsible and independent thinkers [...]

The idea that this responsibility can be delegated to a giant computer programmed with Decision Functions belongs to the phantasy of circles rather remote from scientific research .”

(Fisher, 1973)

www.aviz.fr/badstats