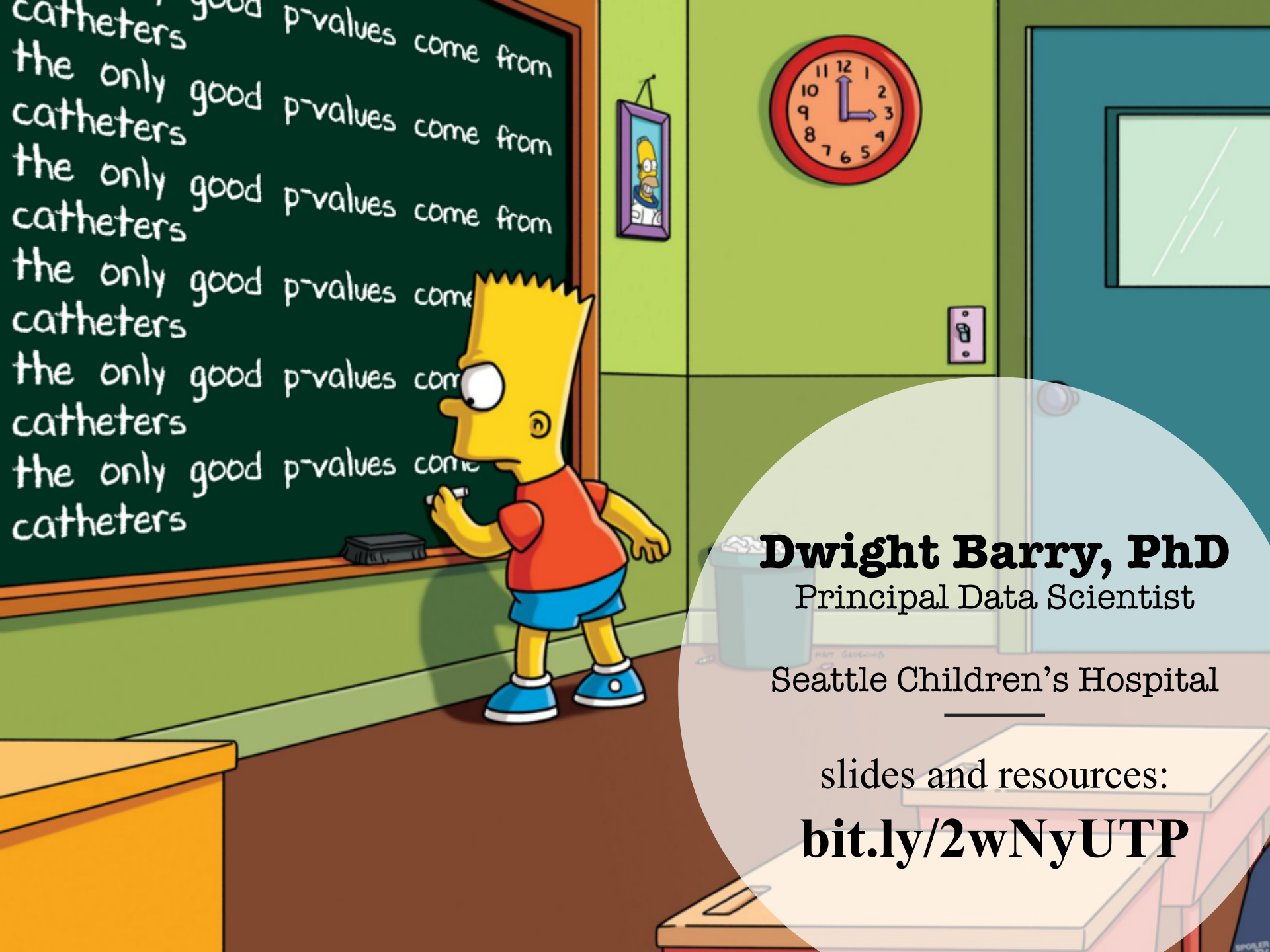


Dwight Barry, PhD
Principal Data Scientist

Seattle Children's Hospital

slides and resources:

bit.ly/2wNyUTP

Singing During Brain Surgery, Kira Performs to Preserve Her Passion

https://pulse.seattlechildrens.org/kira-sings-her-way-through-brain-surgery/

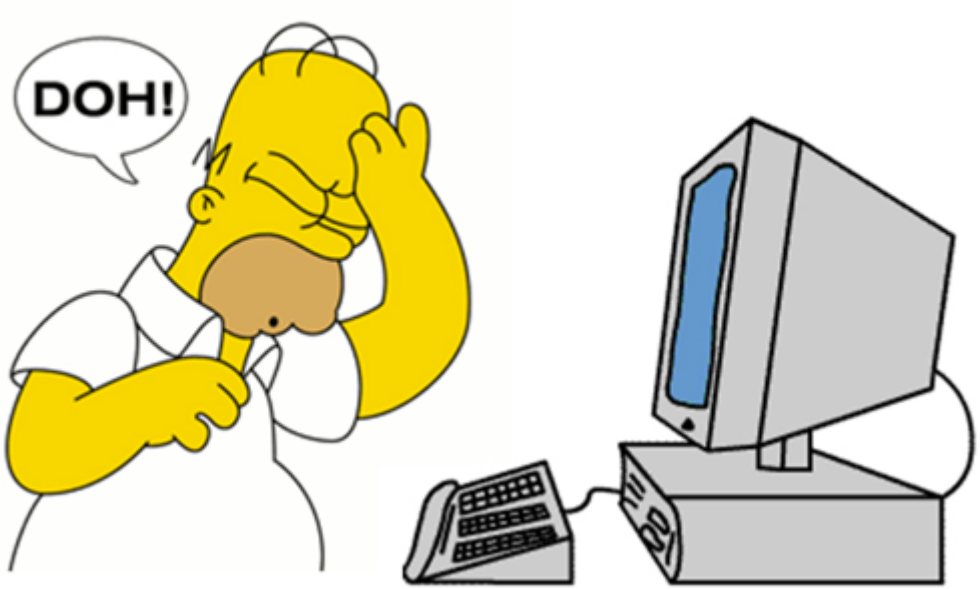
The (practical) question:

Does time from Dx to Tx matter for cognitive outcomes in preschool children?

Shurtleff et al. 2015
bit.ly/2XqqA7Q

The (practical) answer:

$$p = 0.06$$



The (practical) answer:

$p = 0.06$

"**While we did not detect statistically significant differences** … we found clinically important improvements for the short duration group; FSIQ scores improved by 13 on average."

The (practical) problem:

Sample size.

- ***n* = 6** for patients <6 months from Dx to Tx (“Short”)
- ***n* = 5** for patients >18 months from Dx to Tx (“Long”)

The contrast we'll explore here:

Difference in post-surgery Full Scale IQ (FSIQ) score[1] from baseline[2].

[1] Using the Wechsler test

[2] FSIQ score prior to surgery

The evidence:

Dx → Tx >18 months ("Long")

n	mean	sd	median	mad	min	max
5	**-1.4**	11.63	-7	5.93	-11	16

Dx → Tx <6 months ("Short")

n	mean	sd	median	mad	min	max
6	**12**	7.51	10.5	8.9	4	22

The evidence:

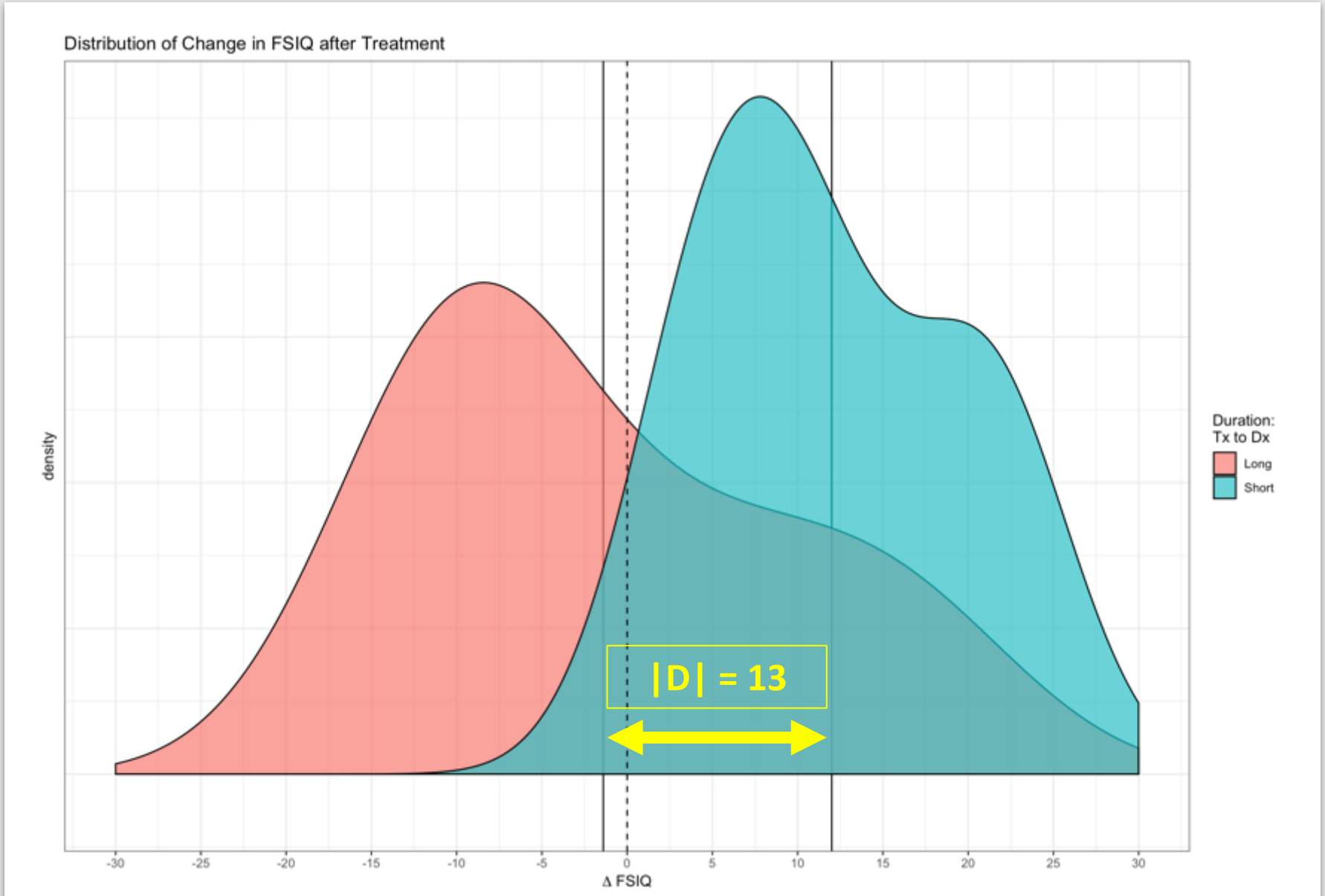
Absolute Difference in Mean FSIQ for Short vs. Long Groups

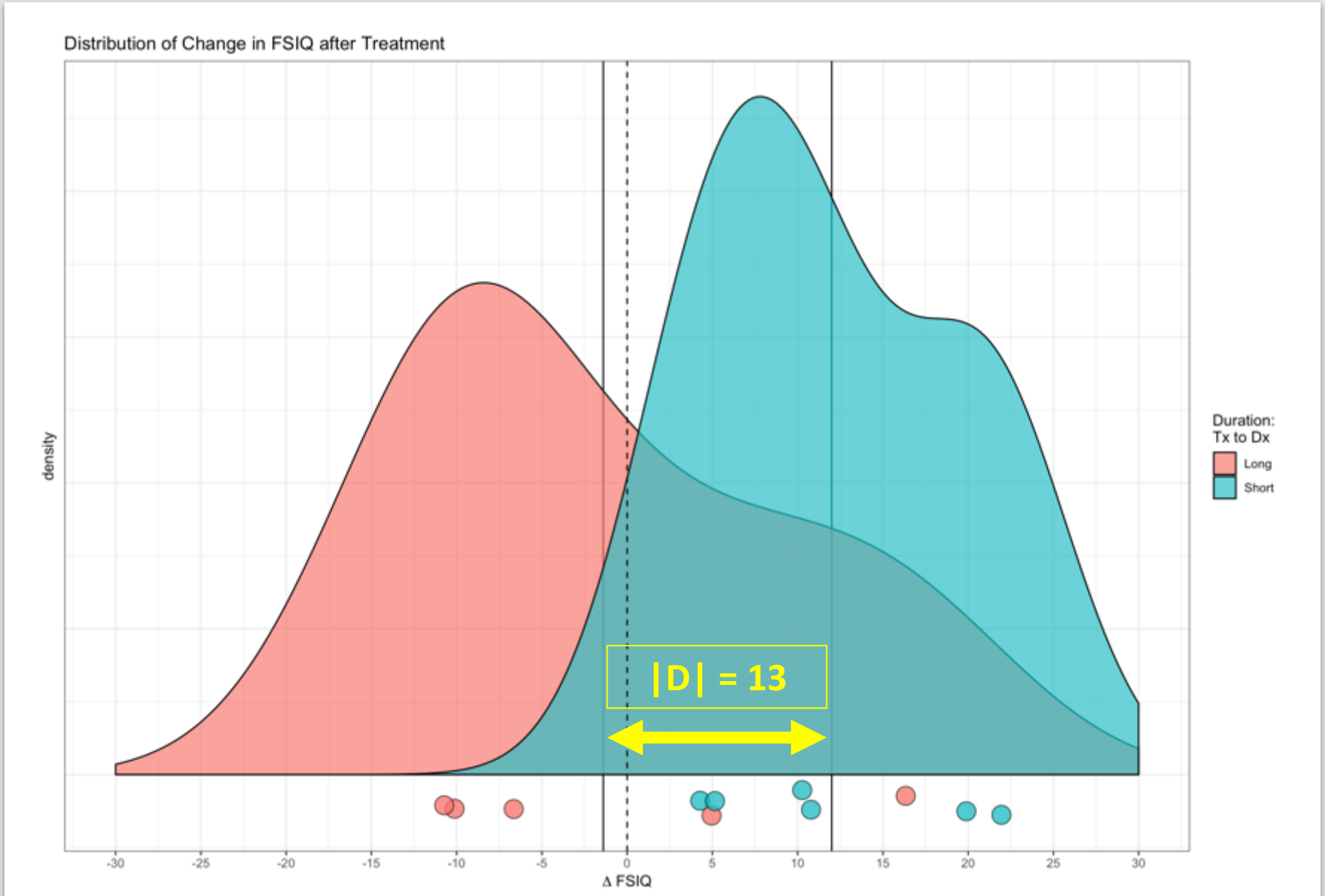
$|D| = 13$

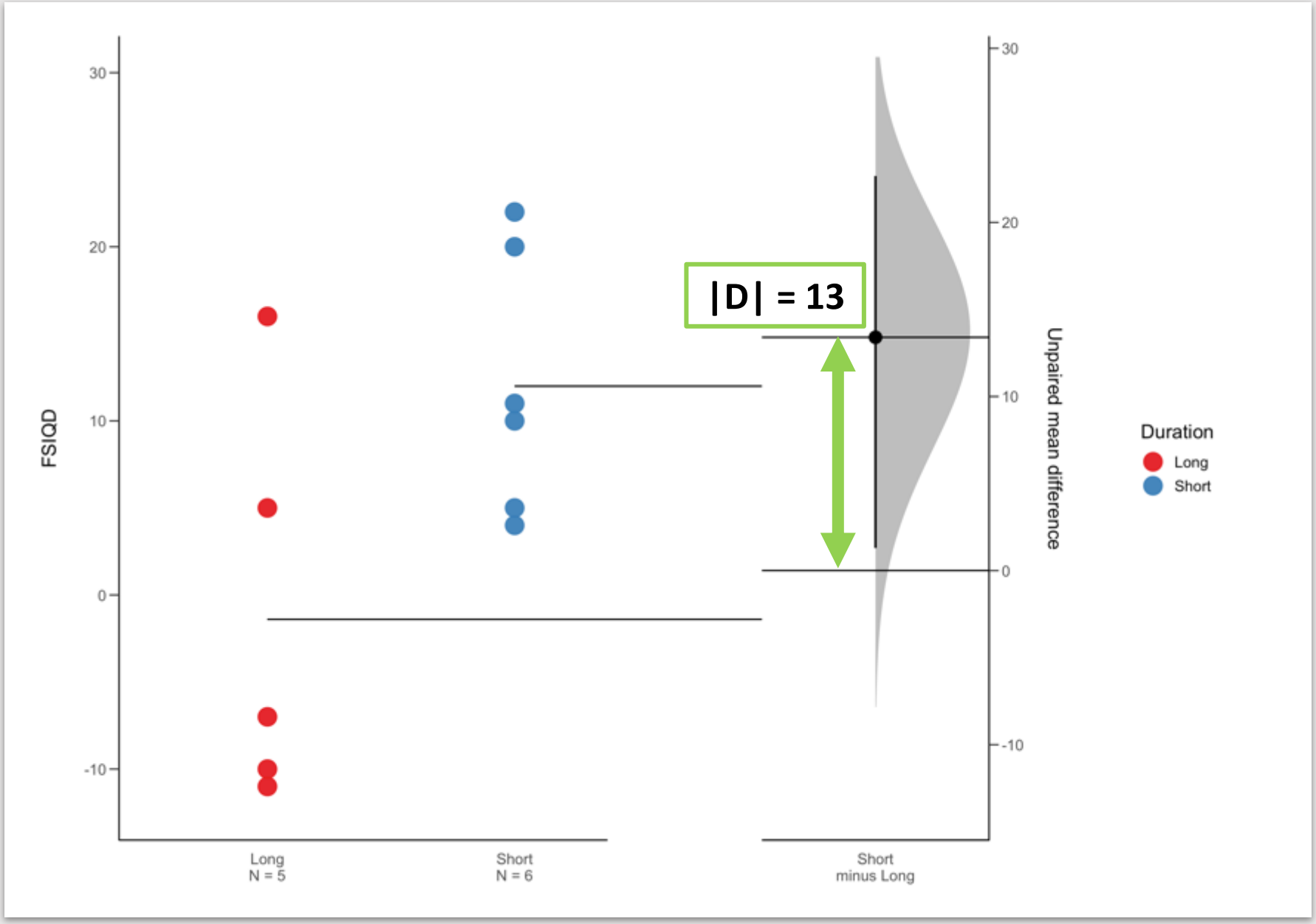
Percent with Same or Improved FSIQ Score After Surgery

***Long*: 40%**

***Short*: 100%**


Distribution of Change in FSIQ after Treatment
density
Δ FSIQ
-30
-25
-20
-15
-10
-5
0
5
10
15
20
25
30
|D| = 13
Duration:
Tx to Dx
Long
Short


Distribution of Change in FSIQ after Treatment
density
Δ FSIQ
-30
-25
-20
-15
-10
-5
0
5
10
15
20
25
30
|D| = 13
Duration:
Tx to Dx
Long
Short


FSIQD
30
20
10
0
-10
Long
N = 5
Short
N = 6
|D| = 13
Unpaired mean difference
30
20
10
0
-10
Short
minus Long
Duration
Long
Short

Inferential Statistical Philosophies

- Frequentism
- Bayesianism
- Likelihood/Information-Theoretic
- EDA
- Causal Analysis*

* not mentioned in this talk

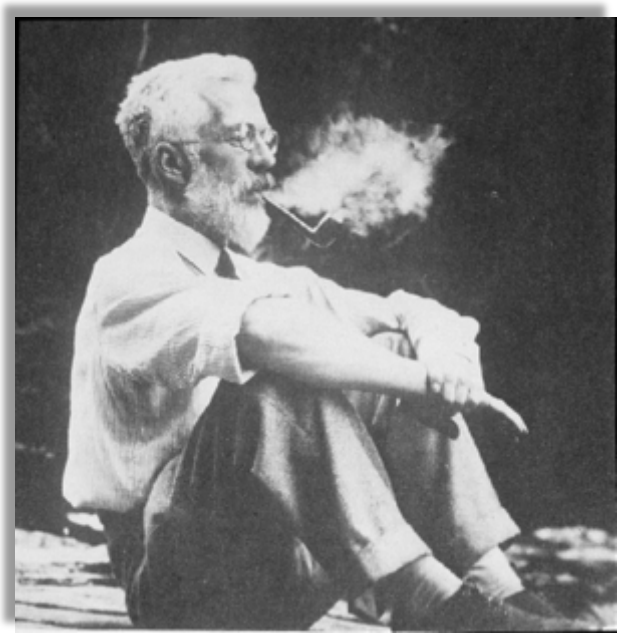
The RSA
@theRSAorg

An algorithm is not a fact. It's an opinion embedded in math.

Inferential Statistical Philosophies

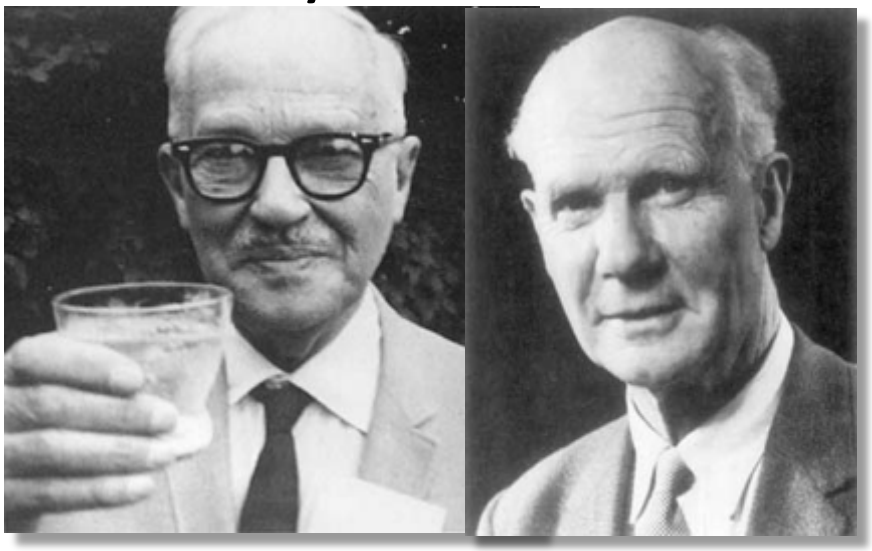
- **Frequentism**
  - *p*-values
  - Error minimization (α/β) and Confidence Intervals*
  - Null Hypothesis Significance Testing (NHST)
- Bayesianism
- Likelihood/Information-Theoretic
- EDA

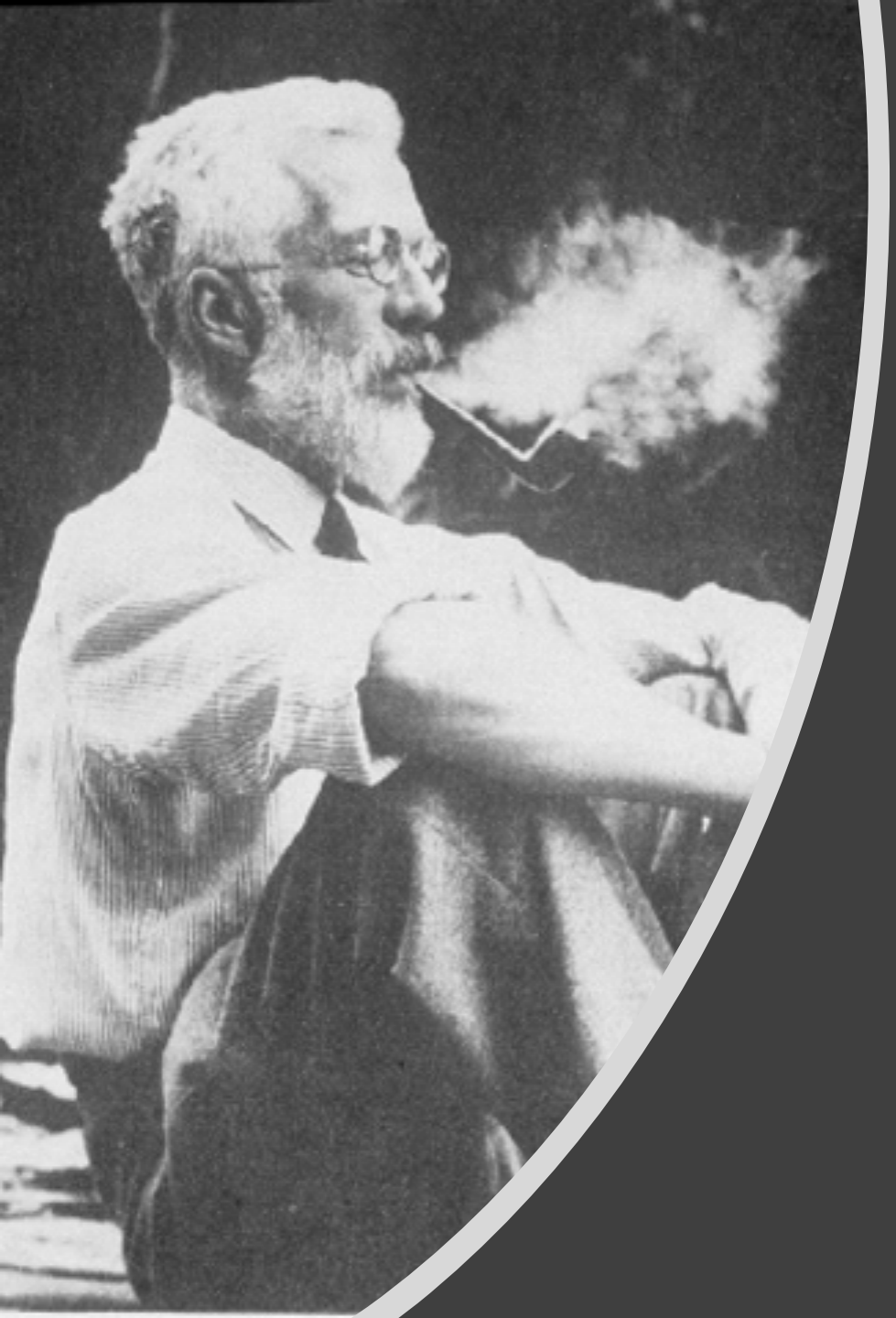
* New term (~2018) for confidence intervals: *compatibility* intervals



Fisher

Neyman & Pearson



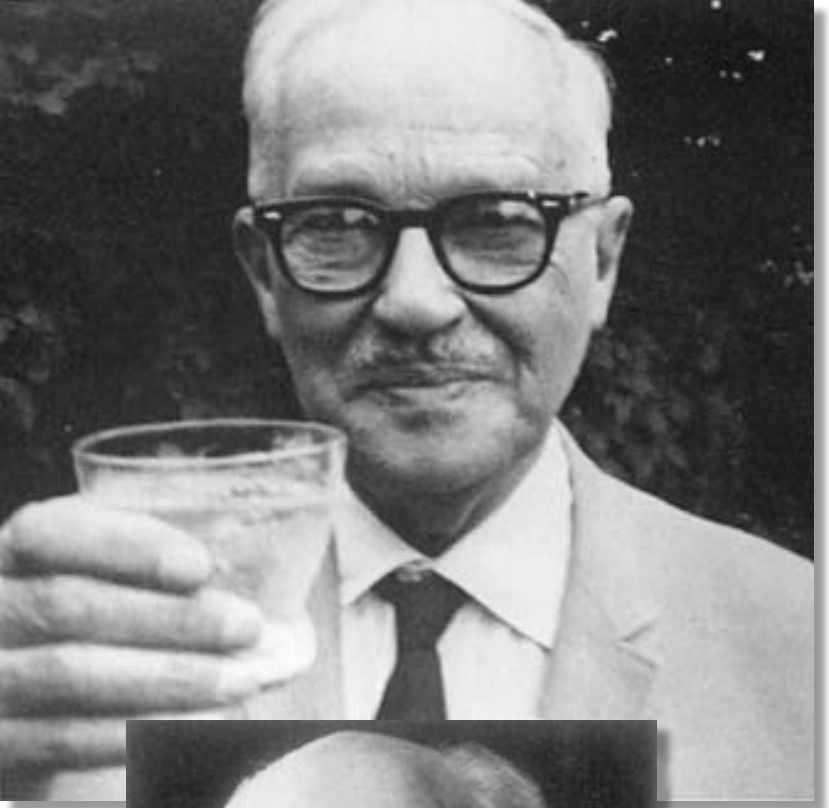


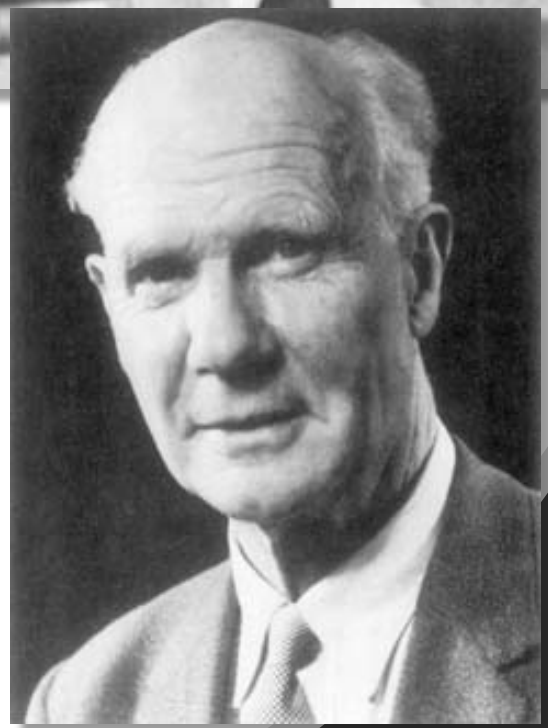
Fisher's point of view: *p*-values as evidence

Technical definition of a *p*-value

p is the probability of seeing data equal to or more extreme than your own results, **given that in reality, the null hypothesis is true**:

$$\textbf{\textit{p}-value} = \Pr(\geq \text{data} \mid \Pr[H_0 = 1.0])$$



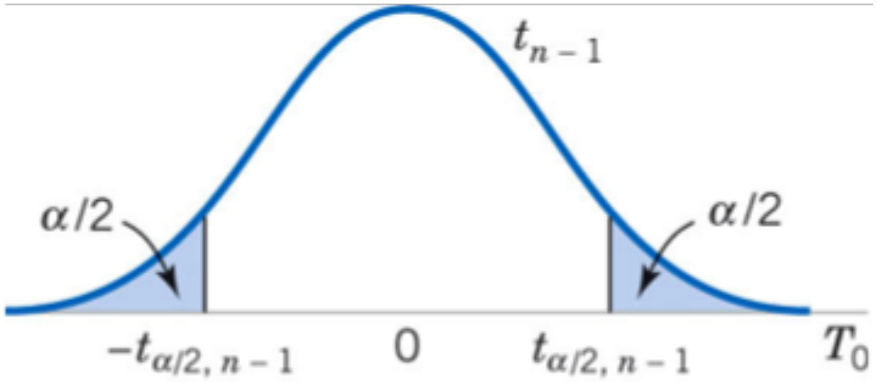


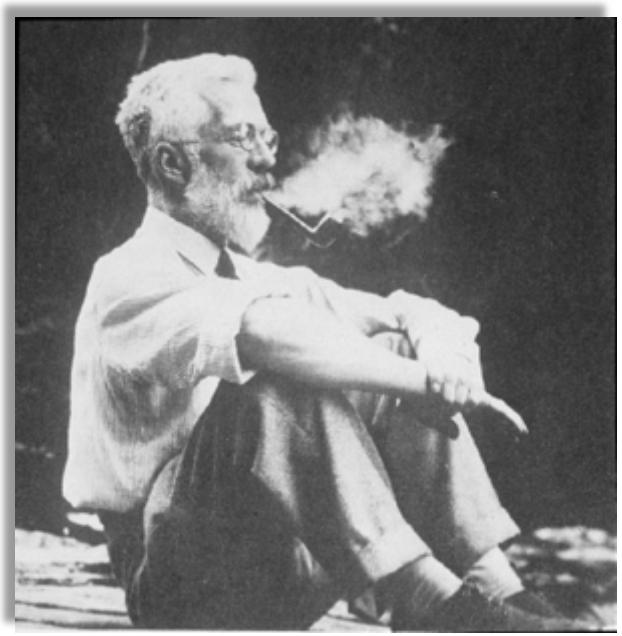
Neyman & Pearson's point of view: error minimization

Reducing *decision* errors

- Set α and β ***before*** starting an experiment.

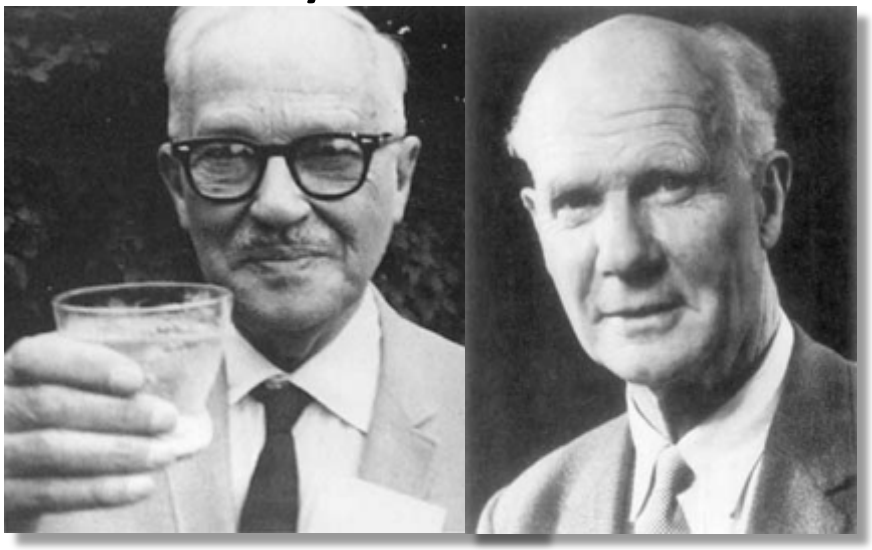
- Compare the test statistic (from your well-powered study, defined by β) with the rejection region (defined by α), & determine whether it falls within (reject H_0) or outside (accept H_0) that region.





Fisher

Neyman & Pearson



Fisher was wrong

1. If probability is long-run frequency, **there is no such thing as a probability of a single experiment**. Either you found the effect, or you didn't.
2. *p* is not a measure of evidence.
3. The logic of *p*-values rests upon a fallacy.

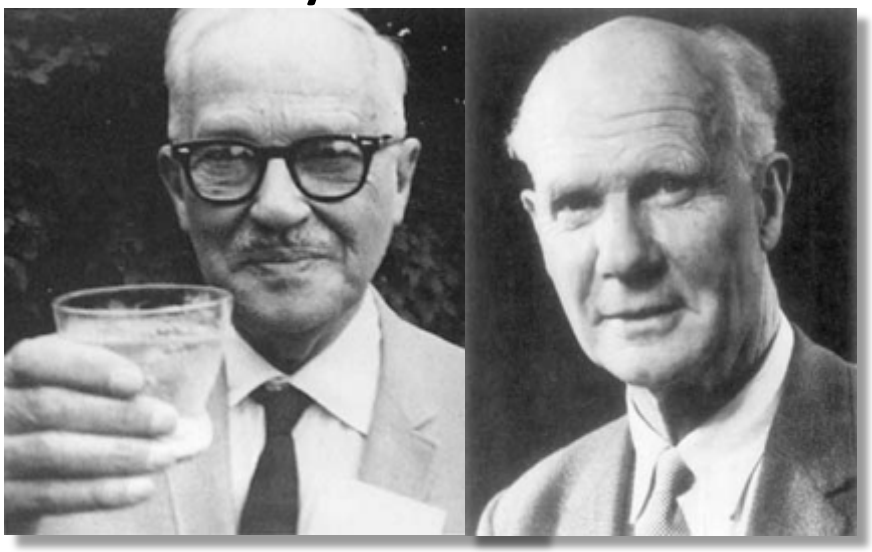
The logic of Fisher's *p*-value

If H_0 is true, then this result would probably not occur.	Most Americans are not members of Congress.
This result has occurred.	This person is a member of Congress.
Therefore, H_0 is probably not true.	Therefore, this person is probably not an American.

Fisher

VS.

Neyman & Pearson



… *in the long run*, an $\alpha = 0.05$ threshold means you can expect that 5 in 100 experiments performed

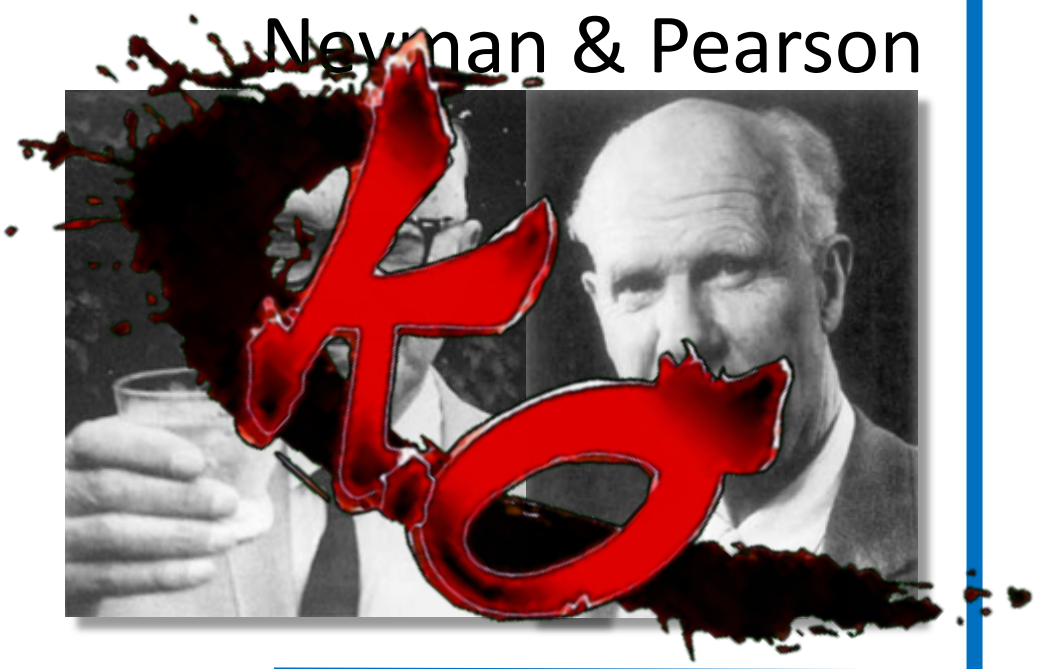
in the same way on the same population

would lead you to falsely reject the null.

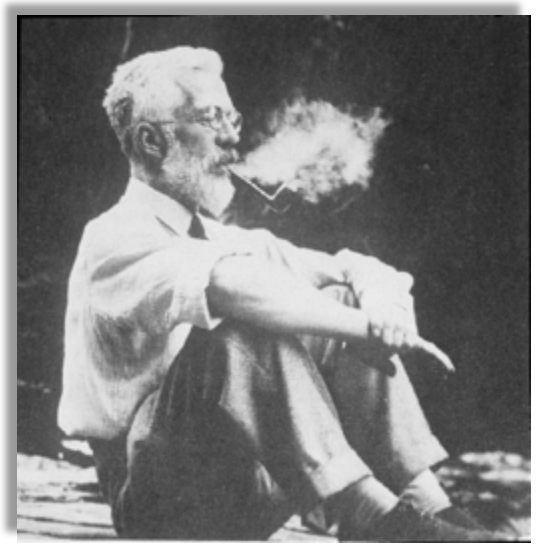
Fisher

VS.

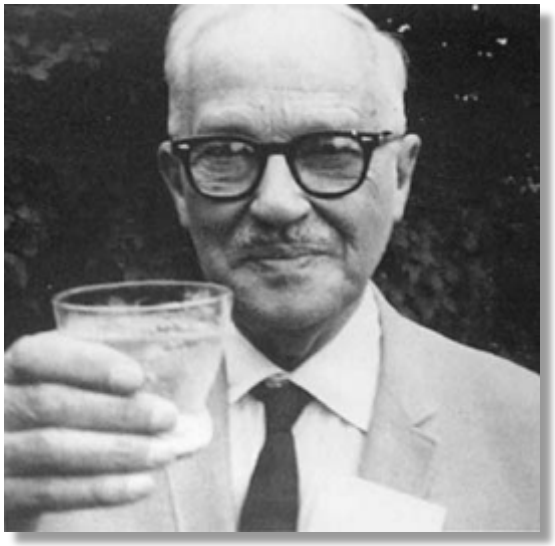
Neyman & Pearson

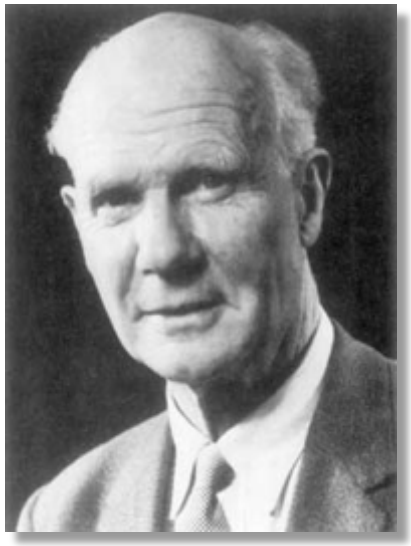


What about $p < \alpha$?



+





=

Null Hypothesis
Significance Testing (NHST)
"statistical significance"

What about p < α?

Fischer:

***p*-value = Pr(≥data | Pr[H_0 = 1.0])**

N-P:

α = probability of falsely rejecting null

∴ Reporting an exact *p* relative to α is completely meaningless.

Any sufficiently crappy research is indistinguishable from fraud.

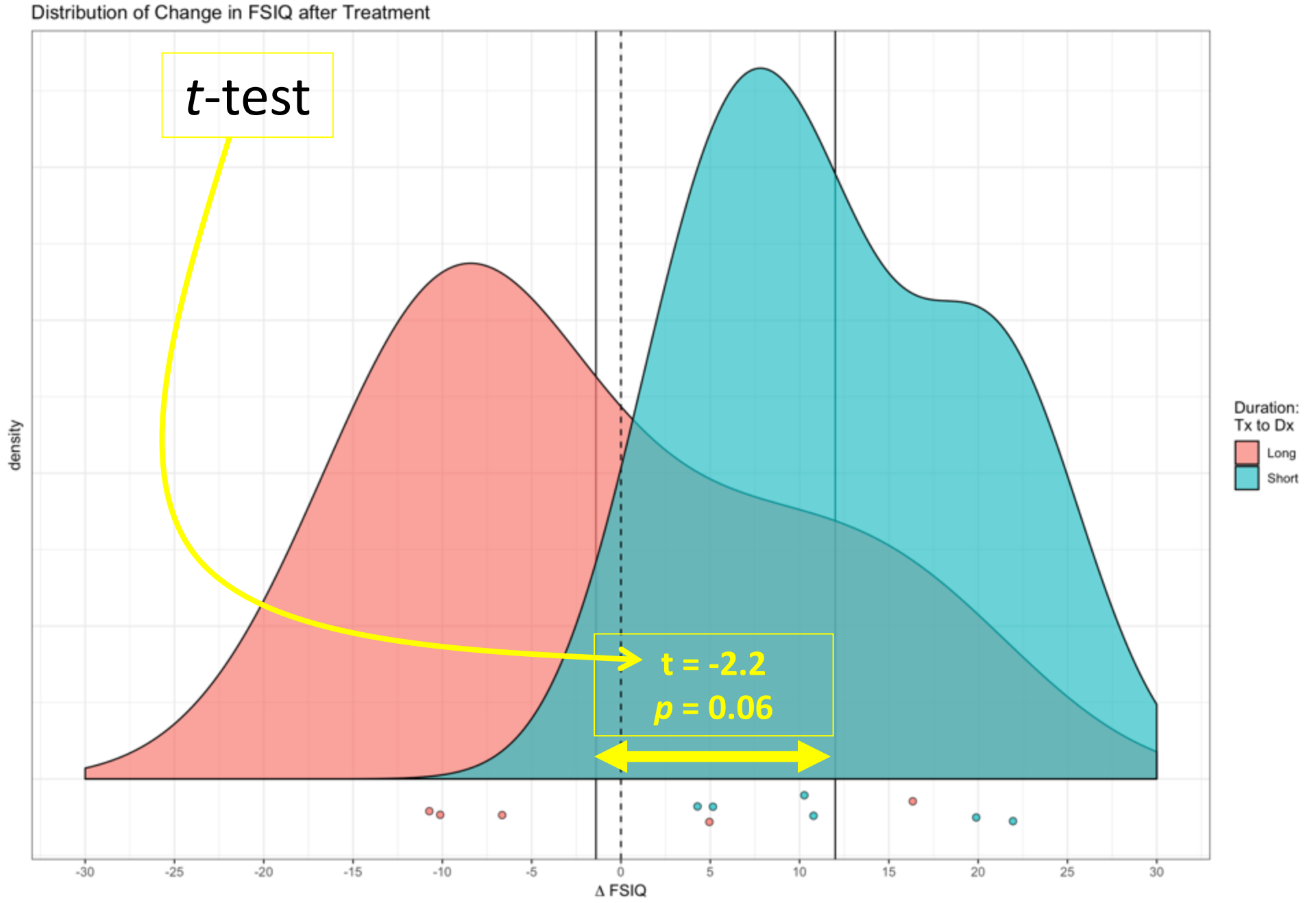
Andrew Gelman, 2016

Statistician, Columbia University

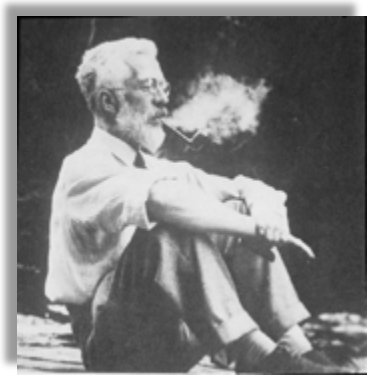
Null Hypothesis
Significance Testing (NHST)

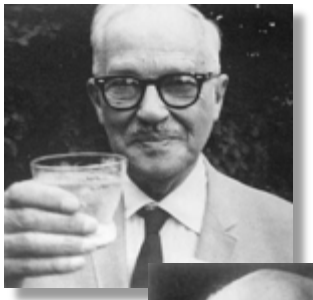
=

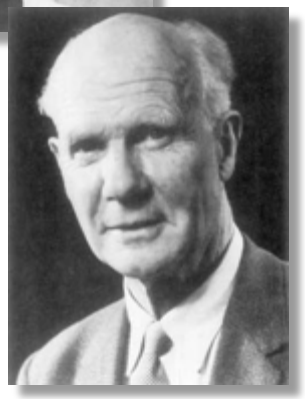
Statistical **H**ypothesis
Inference **T**esting


Distribution of Change in FSIQ after Treatment
t-test
t = -2.2
p = 0.06
density
Δ FSIQ
-30
-25
-20
-15
-10
-5
0
5
10
15
20
25
30
Duration:
Tx to Dx
Long
Short

p = 0.06







- Fisher: "we have weak evidence that our results are different from no effect."

- N-P: "the result is not statistically significant; either there is no effect *or* our study did not have the power to detect an effect."

p-hacking: multiple *comparisons*

Adjustment type	*p*-value
Holm-Bonferroni (conservative)	0.138
False Discovery Rate (liberal)	0.075

Fisher: "what the hell are you even *doing*, you ignorant cretins? This isn't how science works!"

N-P: "the result is not statistically significant; either there is no effect *or* our study did not have the power to detect an effect. Screw you, Fisher; your logic and math is worse than useless."

p-hacking: multiple *options*

Test Type	*p*-value
t-test, equal variances	0.046
t-test, unequal variances	0.064
Permutation test	0.058
Mann-Whitney-Wilcoxon (asymptotic)	0.082
Mann-Whitney-Wilcoxon (exact)	0.093
Regression (F-test)	0.046
Repeated measures ANOVA (*Duration* coefficient)	0.004
ANCOVA (global)	0.002
ANCOVA (*Duration* coefficient)	0.077

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

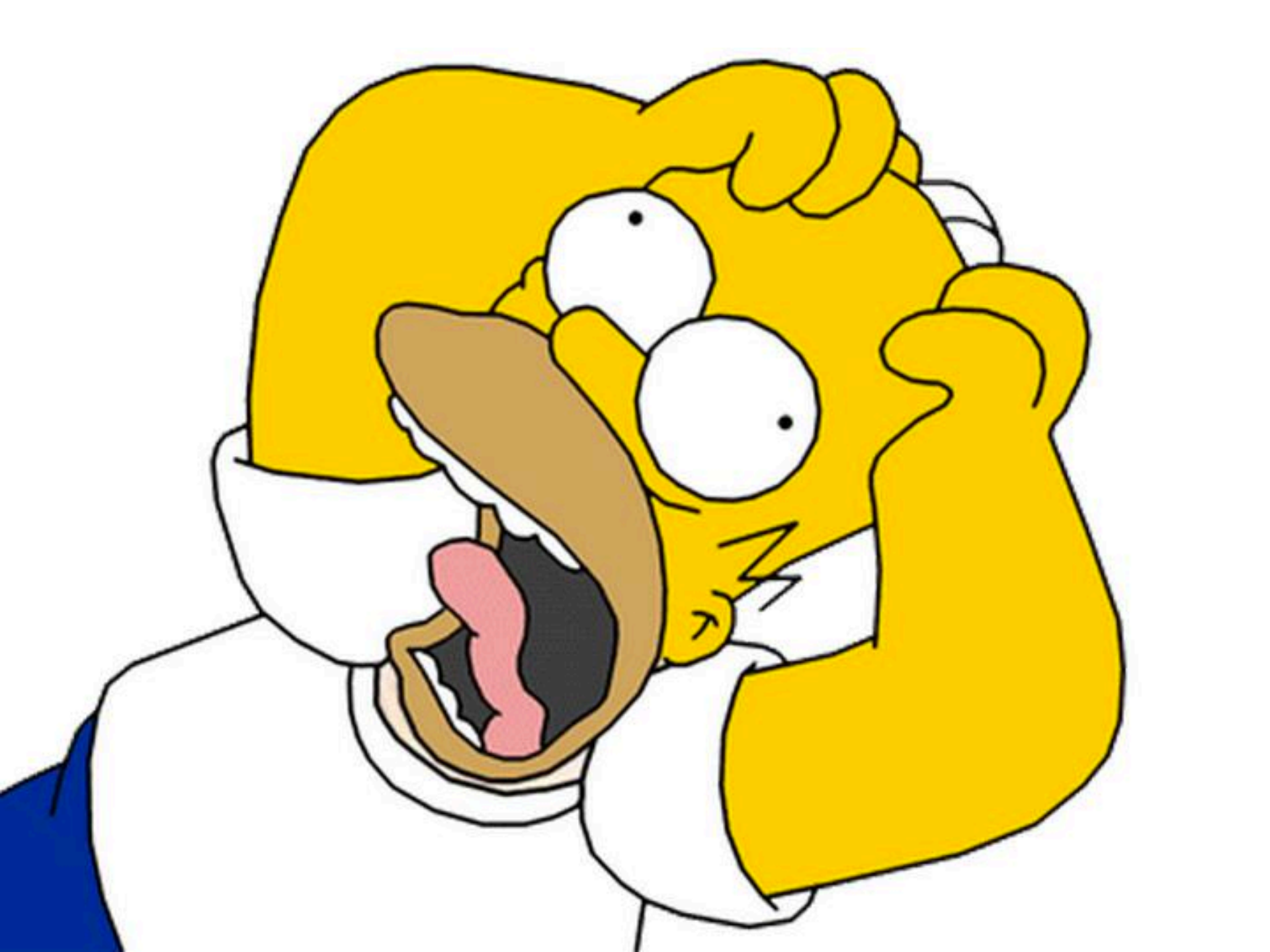
2005, https://doi.org/10.1371/journal.pmed.0020124

A mistake in the operating room can threaten the life of one patient; **a mistake in statistical analysis or interpretation can lead to hundreds of early deaths**.

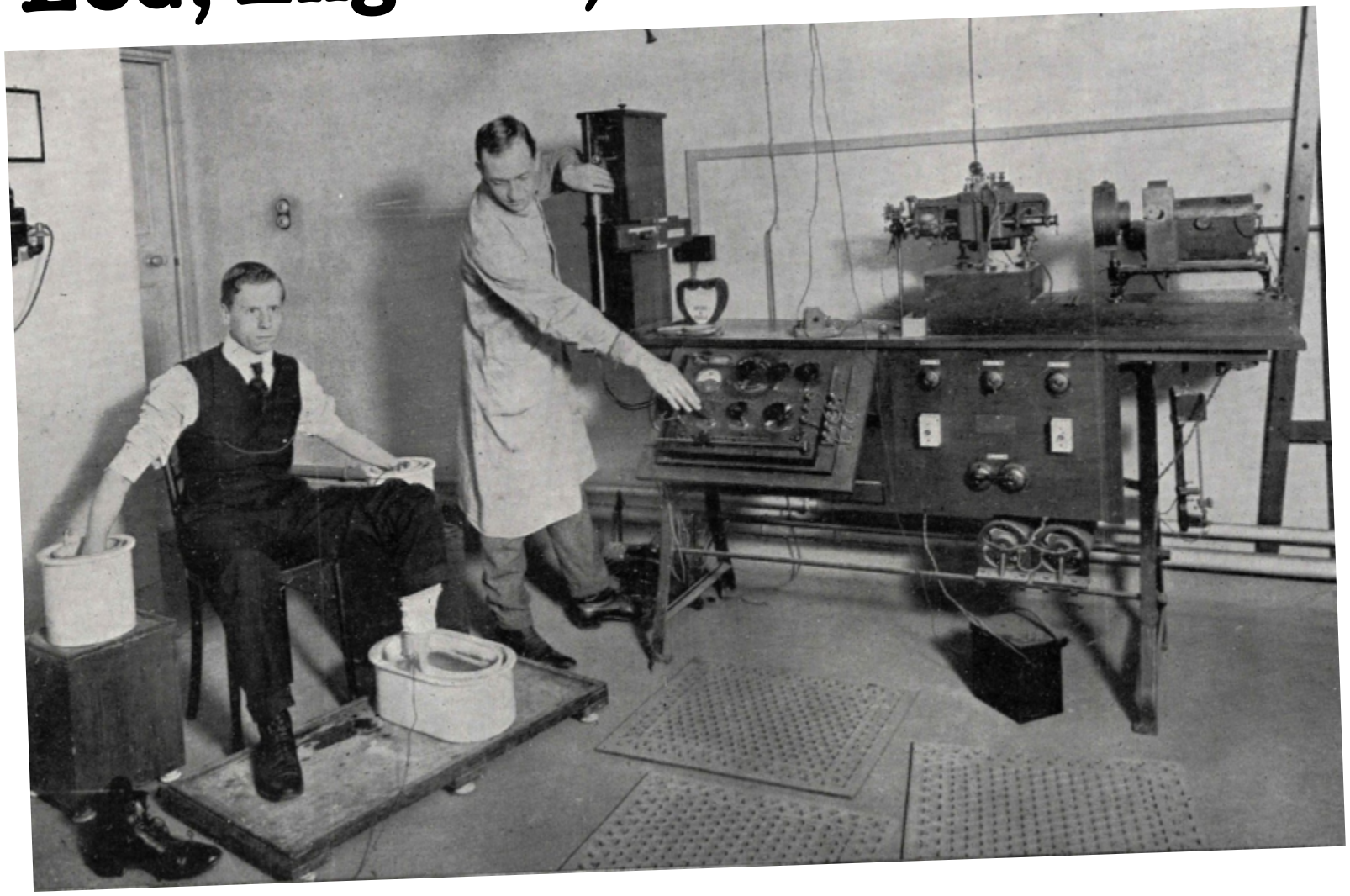
Andrew Vickers

Biostatistician

Memorial Sloan Kettering Cancer Center

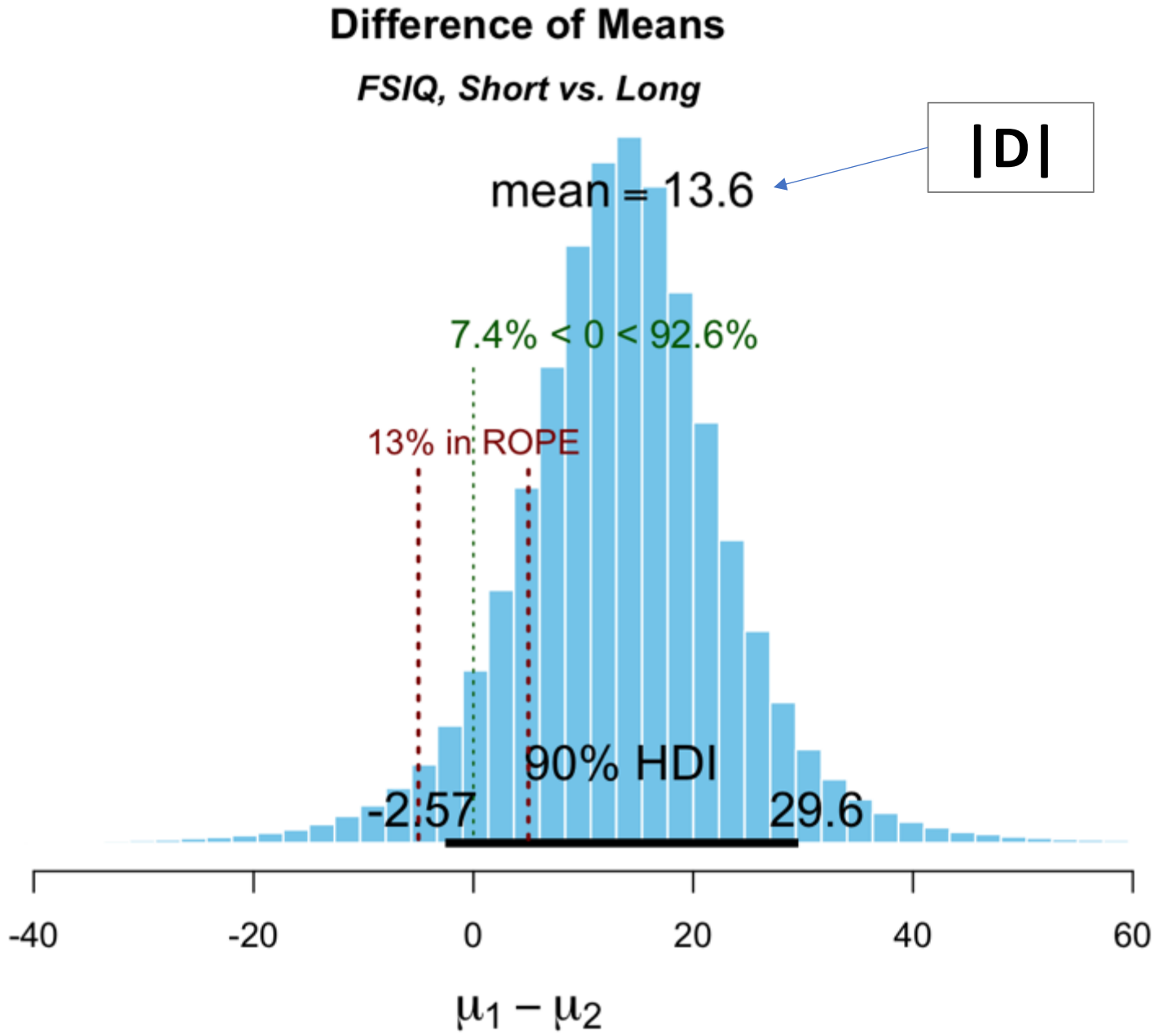


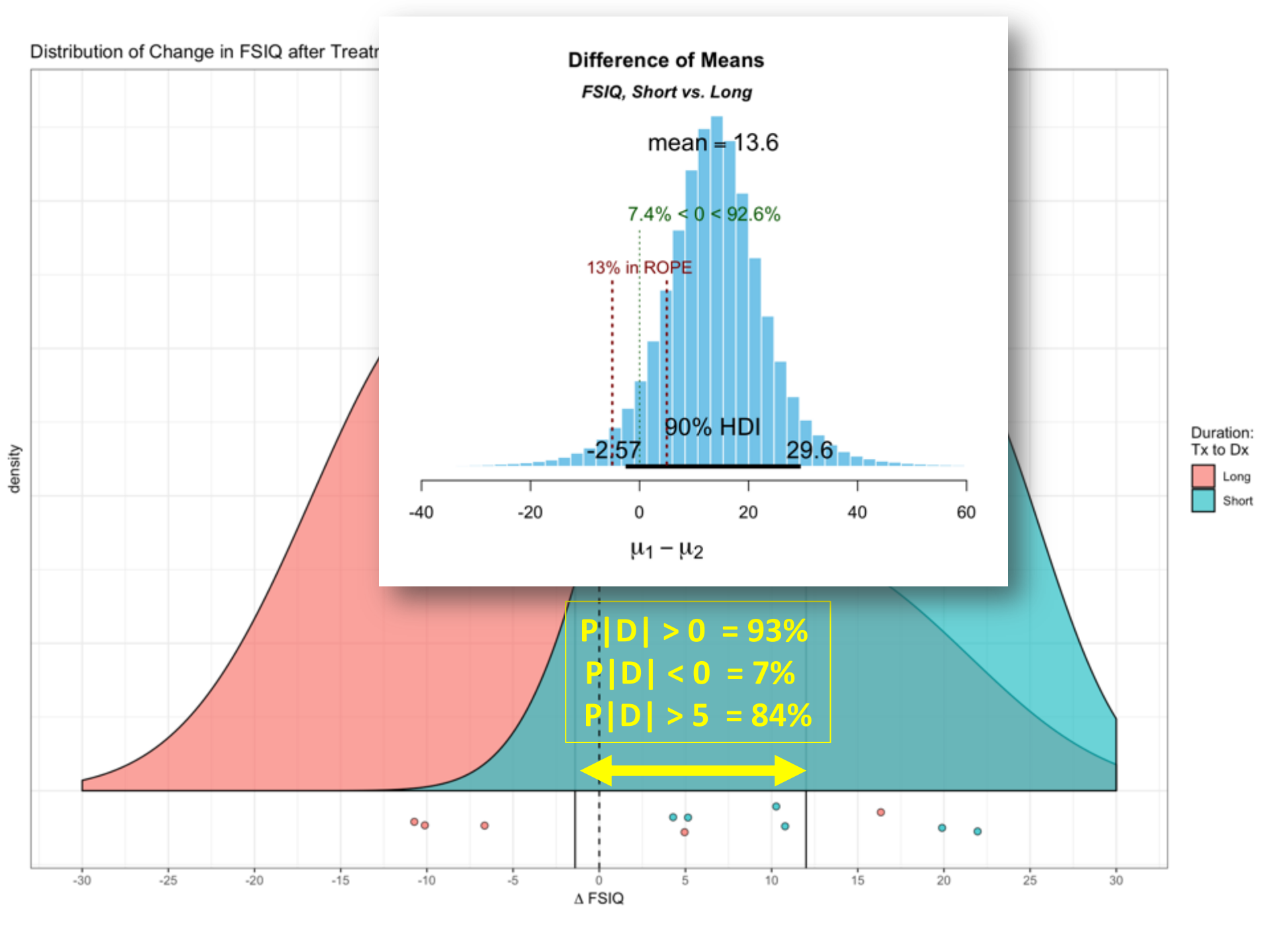
ECG, England, 1916



Inferential Statistical Philosophies

- Frequentism
  - p-values
  - Error minimization (α/β) and Confidence intervals
  - Null Hypothesis Significance Testing
- **Bayesianism**
- Likelihood/Information-Theoretic
- EDA


Difference of Means
FSIQ, Short vs. Long
|D|
mean = 13.6
7.4% < 0 < 92.6%
13% in ROPE
90% HDI
-2.57
29.6
-40
-20
0
20
40
60
μ1 − μ2


Distribution of Change in FSIQ after Treatr
Difference of Means
FSIQ, Short vs. Long
mean = 13.6
7.4% < 0 < 92.6%
13% in ROPE
90% HDI
-2.57
29.6
-40
-20
0
20
40
60
μ1 − μ2
P|D| > 0 = 93%
P|D| < 0 = 7%
P|D| > 5 = 84%
density
Duration:
Tx to Dx
Long
Short
-30
-25
-20
-15
-10
-5
0
5
10
15
20
25
30
Δ FSIQ

🔥Kareem Carr🔥

@kareem_carr

Follow

What toaster should I buy?
Read several guides. Carefully weigh various factors. Examine multiple reviews and pages of user feedback. Decide after weighing all the evidence
Timeline: 2 days

Bayesian

What life saving medicines work?
Google study. Check if $p < 0.05$
Timeline: 2 minutes

NHST

6:50 PM - 20 Jan 2019

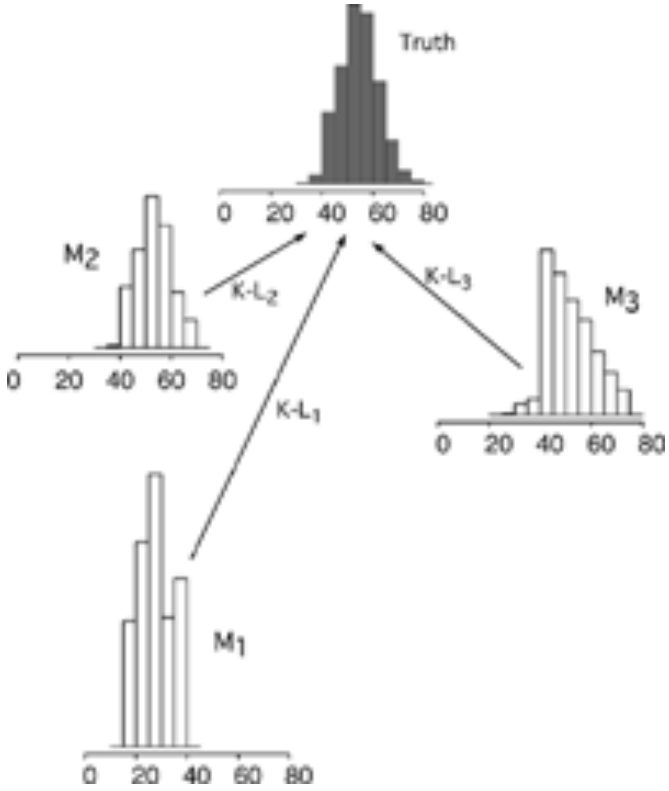
Inferential Statistical Philosophies

- Frequentism
  - p-values
  - Error minimization (α/β) and Confidence intervals
  - Null Hypothesis Significance Testing (NHST)
- Bayesianism
- Likelihood/**Information-Theoretic**
- EDA

I-T Model Set (conceptually)

H_0: **The true difference between population means is 0.**

H_A: **The true difference between population means is ≠ 0.**

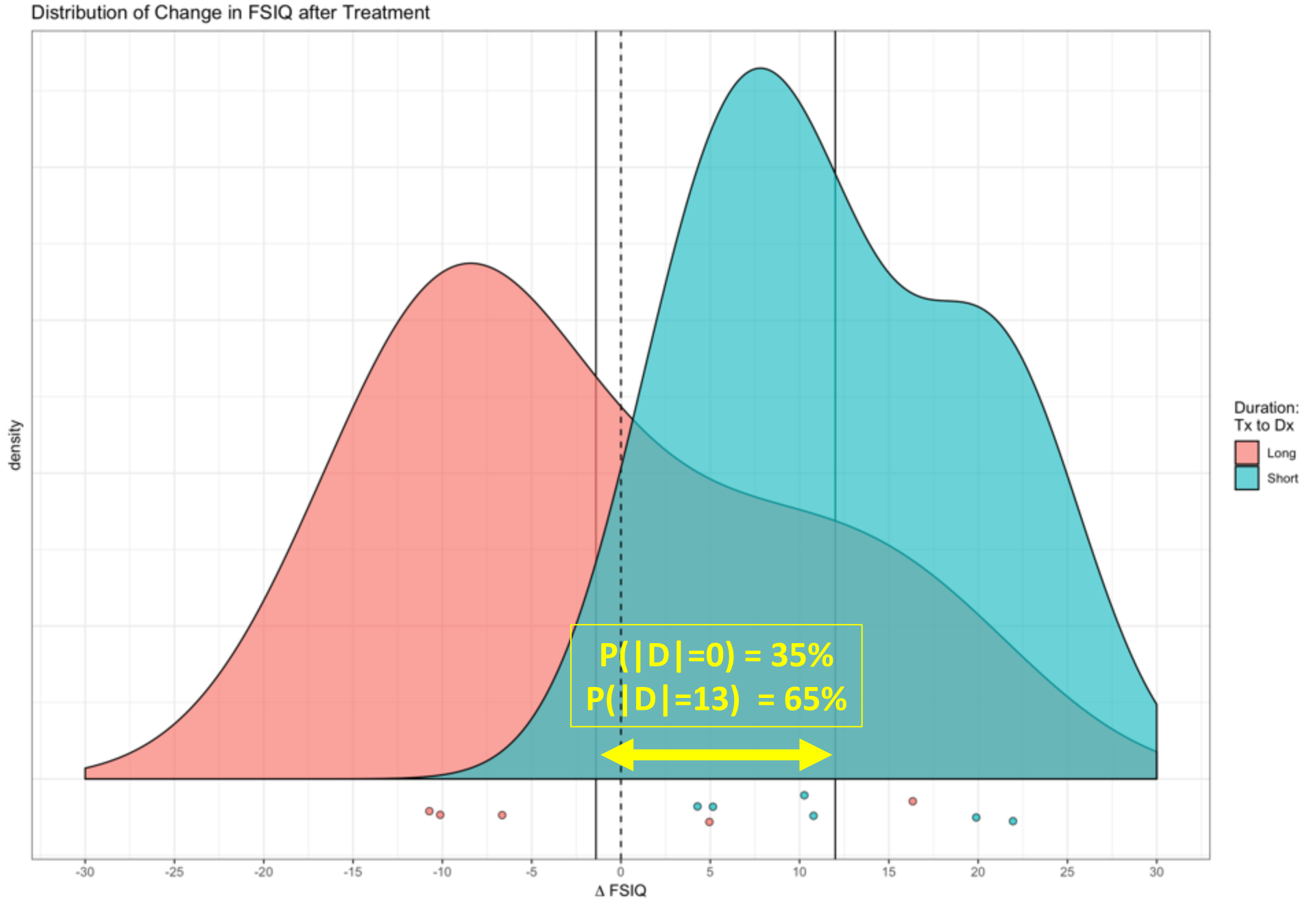


Interpreting the I-T table

Model	k	AICc	Δ AICc	Model Likelihood	AICc Weight	LL	Cuml. Weight
H_A	3	88.11	0.00	1.00	**0.65**	-39.34	0.65
H_0	2	89.32	1.21	0.55	**0.35**	-41.91	1.00

There is a 65% chance that H_A is the best model, and a 35% chance that H_0 is the best model. (Given the model set)

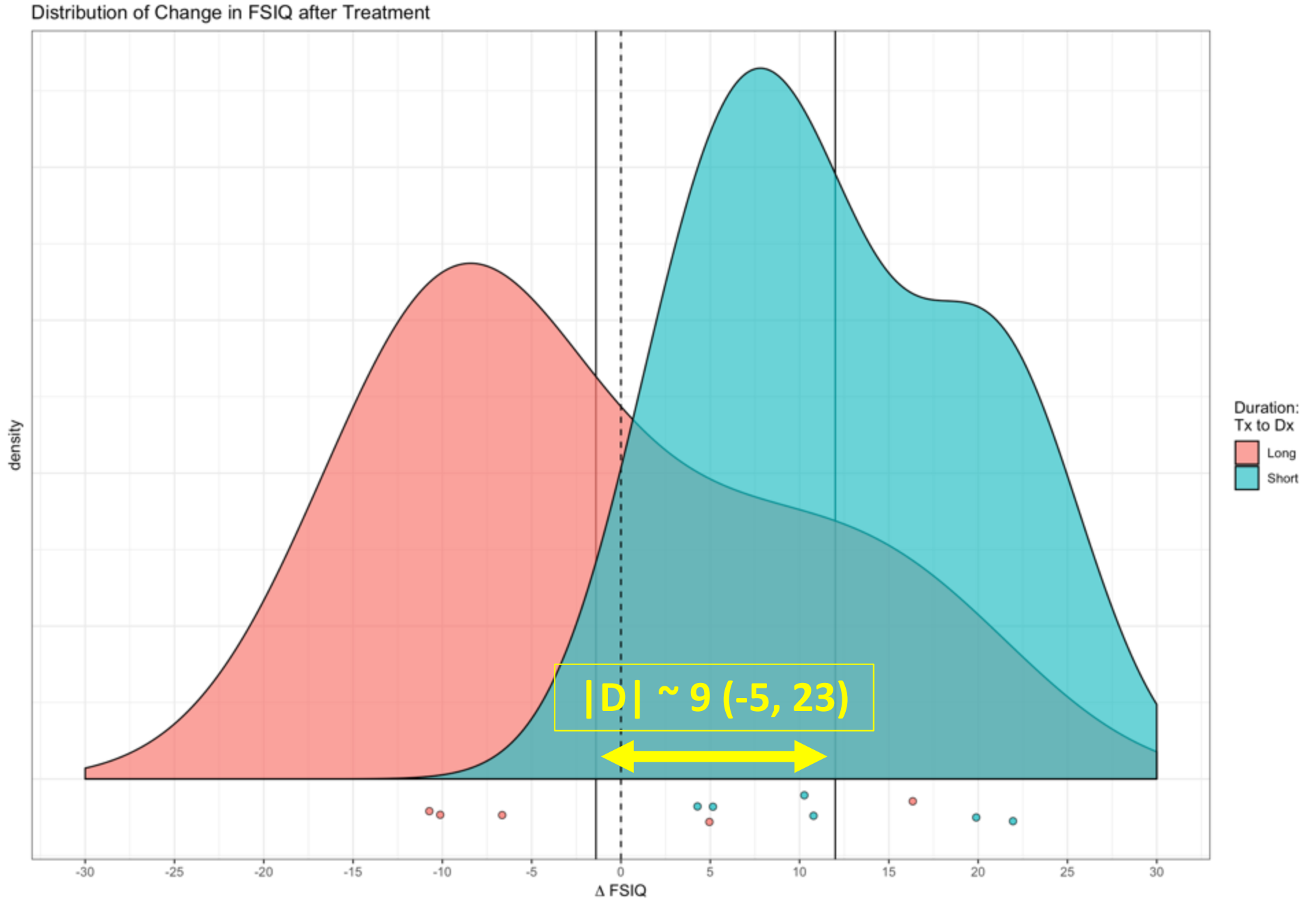
H_A is 1.8 times more likely than H_0. (0.65 / 0.35 = 1.8)


Distribution of Change in FSIQ after Treatment
density
P(|D|=0) = 35%
P(|D|=13) = 65%
-30
-25
-20
-15
-10
-5
0
5
10
15
20
25
30
Δ FSIQ
Duration:
Tx to Dx
Long
Short

Models can be *averaged* to account for uncertainty

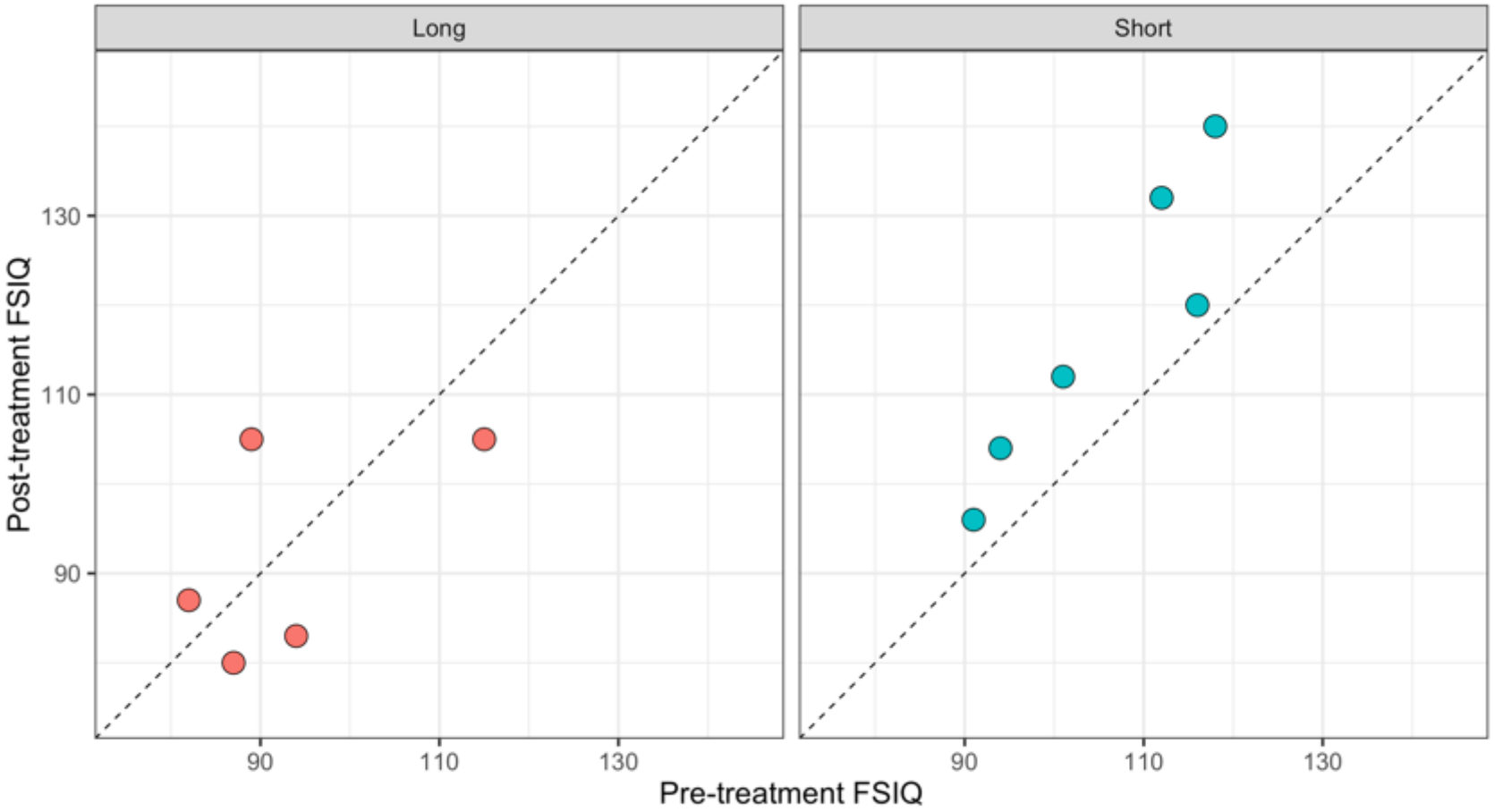
|D|

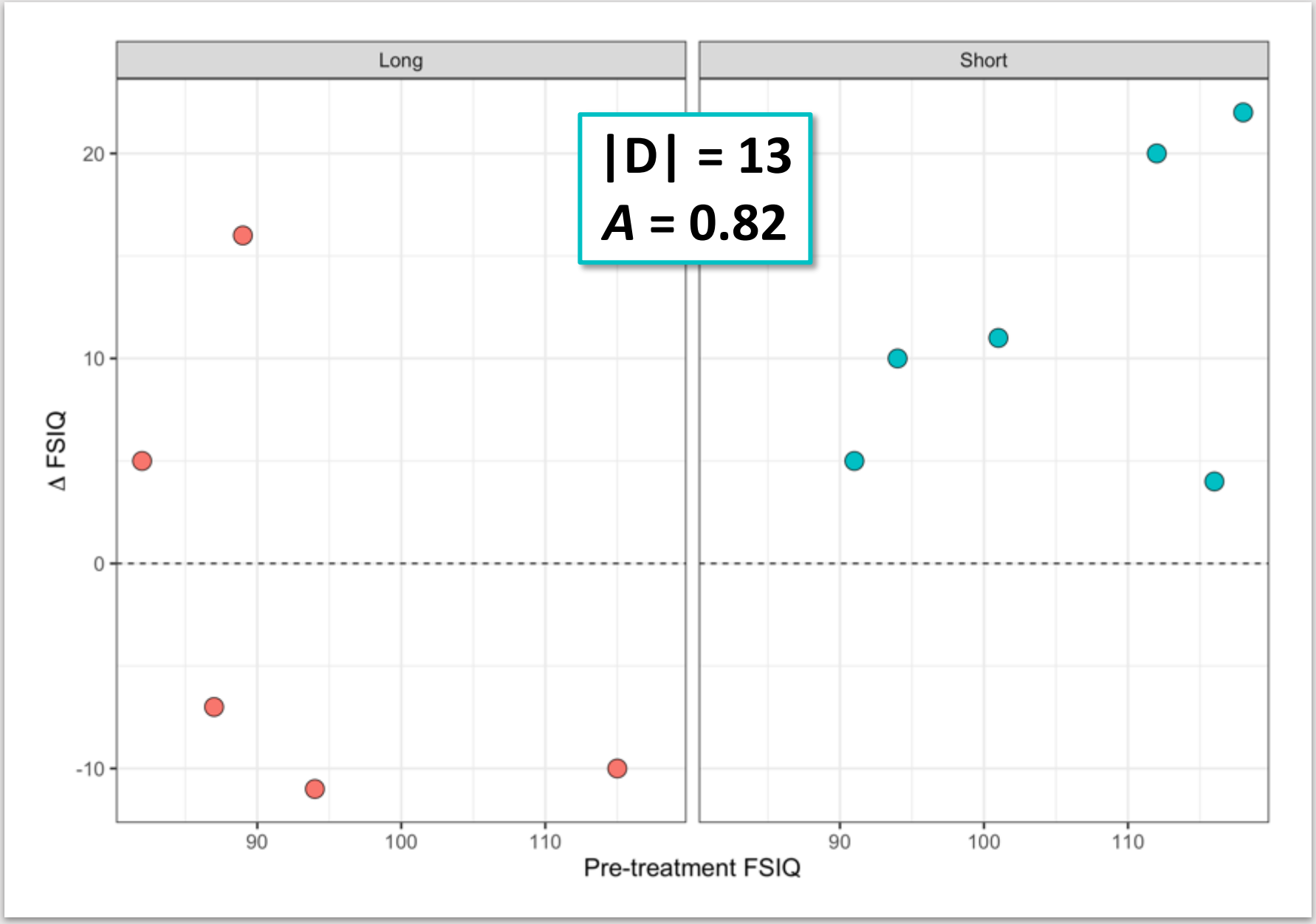
```
## Model-averaged effect size on the response scale based on entire
model set:
##
## Multimodel inference on "Short - Long" based on AICc
##
## AICc table used to obtain model-averaged effect size:
##
##      K AICc Delta_AICc AICcWt Effect(Short - Long) SE
## H1   3 88.11      0.00   0.65                13.4  5.79
## H0   2 89.32      1.21   0.35                 0.0  4.89
##
## Model-averaged effect size: 8.66
## Unconditional SE: 8.44
## 90% Unconditional confidence interval: -5.21, 22.54
```


Distribution of Change in FSIQ after Treatment
density
|D| ~ 9 (-5, 23)
-30
-25
-20
-15
-10
-5
0
5
10
15
20
25
30
Δ FSIQ
Duration:
Tx to Dx
Long
Short

Inferential Statistical Philosophies

- Frequentism
  - p-values
  - Error minimization (α/β) and Confidence intervals
  - Null Hypothesis Significance Testing (NHST)
- Bayesianism
- Likelihood/Information-Theoretic
- **EDA**

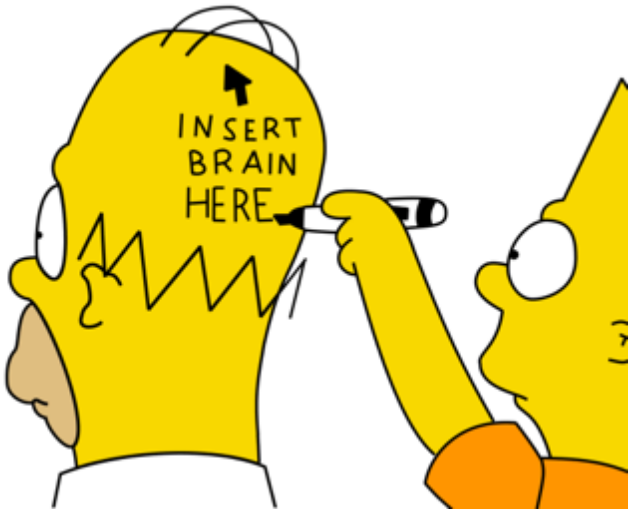

Long
Short
Post-treatment FSIQ
Pre-treatment FSIQ
90
110
130
90
110
130
90
110
130


Long
Short
|D| = 13
A = 0.82
20
10
0
-10
Δ FSIQ
90
100
110
90
100
110
Pre-treatment FSIQ

How best to infer?

- Frequentism
  - ~~Fisher: *p* is an evidence probability~~
  - Neyman-Pearson (N-P): α is an error probability, also need α & β prior to experiment
  - ~~NHST: the corrupt hybrid where you want $p < \alpha$~~*
- Bayesian
- Information-theoretic
- EDA

* **S**tatistical **H**ypothesis **I**nference **T**esting



Statistical pragmatism

- **Use Neyman-Pearson frequentist tools** to obtain decision procedures for long-term error control in ***truly*** replicable, stable population contexts. (Or at least drop *p* and use CIs instead.)

- **Use Bayesian tools** if you need to know what you *should* believe.

- **Use information-theoretic tools** if you need to know what the evidence best supports.

The impact of statistical philosophy on brain surgery:

EDA: Observed Effect size: $|D| = 13$, $A = 0.82$

- **Frequentism:** $|D_{obs}| = 13$ (90% CI: 3, 24)
  - Fisher's *p*-value = 0.06, weak evidence for an effect
  - N-P: unable to reject H_0
  - S.H.I.T.: not "statistically significant" ($p = 0.06$, $p > \alpha$)
- **Bayesian:** posterior $|D_{true}| = 14$ (90% HDI: -2, 30)
  - 93% chance of being > 0
  - 84% chance of being > 5
- **Information-theoretic:** $|D_{avg}| = 9$ (90% CI: -5, 23)
  - 35% chance H_0 is best model
  - H_A is ~2x more likely than H_0

slides and resources:

bit.ly/2wNyUTP

Grumpy Old Health Stats D... · 3/3/19

"If you never use another p-value, you will have improved medicine."

-me, to clinicians

#statstwitter #medtwitter #epitwitter