INTRODUCTION TO THE PRINCIPLES AND PRACTICE OF CLINICAL RESEARCH

Overview of Hypothesis Testing

Paul Wakim, PhD
Chief, Biostatistics and Clinical Epidemiology Service
Clinical Center, National Institutes of Health
U.S. Department of Health and Human Services

3 November 2015

NIH Clinical Center
America's Research Hospital

Outline

- Fundamentals of Hypothesis Testing
- Superiority vs. Non-Inferiority vs. Equivalence
- Multiple Comparisons (Multiplicity Adjustment)
- Bottom-Line Key Points

Outline

- Fundamentals of Hypothesis Testing
- Superiority vs. Non-Inferiority vs. Equivalence
- Multiple Comparisons (Multiplicity Adjustment)
- Bottom-Line Key Points

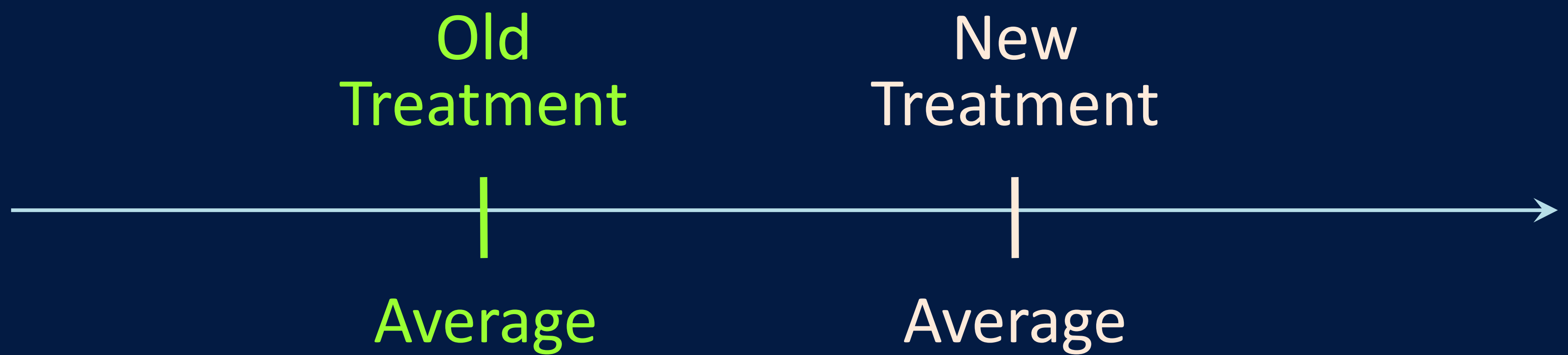
Question

Without ____?______ , there is no need for Statistics

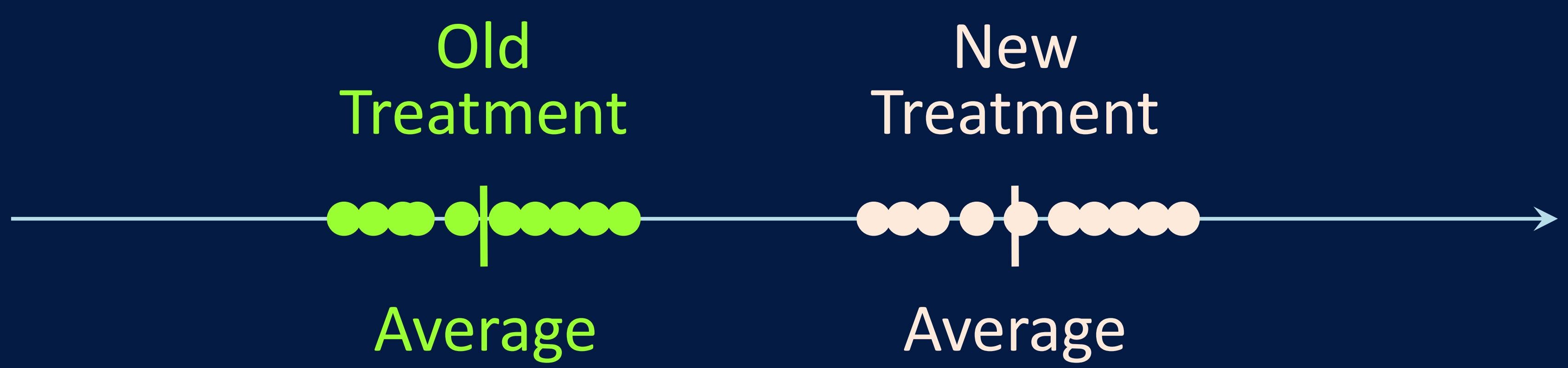
Answer

Without variability, there is no need for Statistics

Variability

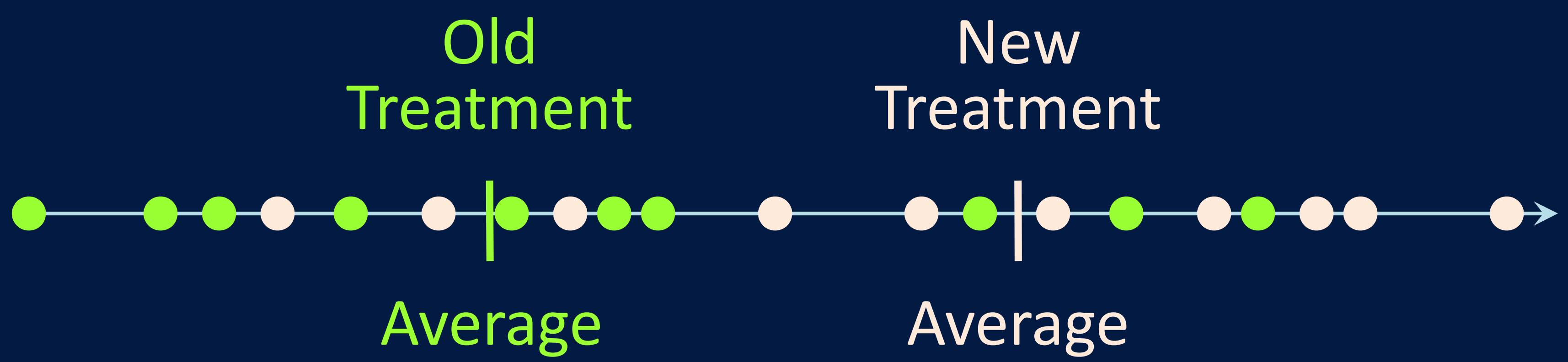


Variability



Variability

Variability



Statistical Inference

1) Draw a sample from the population of interest

2) Analyze the sample data

3) Make conclusion about the population based on results from the sample

Typical Setting of Statistical Inference (non-Bayesian)

Null Hypothesis (H_0)
Experimental = Control
or
Experimental – Control = 0

Alternative Hypothesis (H_1 or H_A)
Experimental ≠ Control
or
Experimental – Control ≠ 0

Question: Is there enough evidence to reject H_0 – the hypothesis of no difference?

We expect (hope) to reject H_0 in favor of H_A

Correct and Incorrect Conclusion

Conclusion (based on sample)	**Reject H_0** (evidence of difference)		
	Fail to Reject H_0 (no evidence of difference)		

Correct and Incorrect Conclusion

		True (Unknown) State	
		No Difference (H_0 is true)	**Difference** (H_0 is false)
Conclusion (based on sample)	**Reject H_0** (evidence of difference)		
	Fail to Reject H_0 (no evidence of difference)		

Correct and Incorrect Conclusion

		True (Unknown) State	
		No Difference (H_0 is true)	**Difference** (H_0 is false)
Conclusion (based on sample)	**Reject H_0** (evidence of difference)		**Correct Conclusion** (True Positive)
	Fail to Reject H_0 (no evidence of difference)		

Correct and Incorrect Conclusion

		True (Unknown) State	
		No Difference (H_0 is true)	**Difference** (H_0 is false)
Conclusion (based on sample)	**Reject H_0** (evidence of difference)		**Correct Conclusion** (True Positive)
	Fail to Reject H_0 (no evidence of difference)	**Correct Conclusion** (True Negative)	

Correct and Incorrect Conclusion

		True (Unknown) State	
		No Difference (H_0 is true)	**Difference** (H_0 is false)
Conclusion (based on sample)	**Reject H_0** (evidence of difference)	**Type I Error** (False Positive)	**Correct Conclusion** (True Positive)
	Fail to Reject H_0 (no evidence of difference)	**Correct Conclusion** (True Negative)	

Correct and Incorrect Conclusion

		True (Unknown) State	
		No Difference (H_0 is true)	**Difference** (H_0 is false)
Conclusion (based on sample)	**Reject H_0** (evidence of difference)	**Type I Error** (False Positive)	**Correct Conclusion** (True Positive)
	Fail to Reject H_0 (no evidence of difference)	**Correct Conclusion** (True Negative)	**Type II Error** (False Negative)

α (alpha) = probability of making a Type I error

β (beta) = probability of making a Type II error

Alpha (α)
Probability of Making a Type I Error

Non-technical definition (superiority trial):
Chance of concluding that the experimental treatment is more effective when in fact it is not

Technical definition:
Probability of rejecting H_0 when H_0 is true

Different perspectives:
Regulatory agency, pharmaceutical company

Bottom line:
Most commonly used value for α: 0.05 (two-sided)

Beta (β)
Probability of Making a Type II Error

β = Chance of claiming no diff. when a diff. exists
= Probability of *not* rejecting H_0 when H_0 is false

Low β is "good"

Power = $1 - \beta$
= Probability of rejecting H_0 when H_0 is false
= Probability of detecting an effect when it exists

High power is "good"

Power to Detect an Effect

Non-technical definition (superiority trial):
Chance of concluding that the experimental treatment is more effective when in fact it is

Technical definition:
Probability of rejecting H_0 when H_0 is false
(i.e. when H_A is true)

Different perspectives:
Regulatory agency, pharmaceutical company

Bottom line:
Most commonly used value for power:
Early-phase: 0.60 to 0.80 – Late-phase: 0.80 to 0.95

If you don't change the sample size…

α(alpha)↓ ⇔ β(beta)↑ ⇔ Power↓

Experiment

Toss a coin

Null hypothesis (H_0):
It's the regular coin, with *Head* on one side and *Tail* on the other side

Alternative hypothesis (H_A):
It's the other coin, with *Head* on both sides

We assume "equipoise", i.e. the coin is as likely to be a regular coin as to have 2 Heads

Experiment

Hypotheses:

H_0: it's a regular coin

H_A: the coin has 2 Heads

# of Tosses	# of Heads (data)
1	1

Would you reject H_0?

Experiment

Hypotheses:

H_0: it's a regular coin

H_A: the coin has 2 Heads

# of Tosses	# of Heads (data)
1	1
2	2

Would you reject H_0?

Experiment

Hypotheses:

H_0: it's a regular coin

H_A: the coin has 2 Heads

# of Tosses	# of Heads (data)
1	1
2	2
3	3

Would you reject H_0?

Experiment

Hypotheses:

H_0: it's a regular coin

H_A: the coin has 2 Heads

Would you reject H_0?

# of Tosses	# of Heads (data)
1	1
2	2
3	3
4	4

Experiment

Hypotheses:

H_0: it's a regular coin

H_A: the coin has 2 Heads

Would you reject H_0?

# of Tosses	# of Heads (data)
1	1
2	2
3	3
4	4
5	5

Experiment

Hypotheses:

H_0: it's a regular coin

H_A: the coin has 2 Heads

Would you reject H_0?

# of Tosses	# of Heads (data)
1	1
2	2
3	3
4	4
5	5
6	6

Experiment

Hypotheses:

H_0: it's a regular coin

H_A: the coin has 2 Heads

Would you reject H_0?

# of Tosses	# of Heads (data)
1	1
2	2
3	3
4	4
5	5
6	6
7	7

Experiment

Hypotheses:

H_0: it's a regular coin

H_A: the coin has 2 Heads

How many consecutive Heads did it take you to reject H_0?

How does it compare to a p-value of 0.05?

# of Tosses	# of Heads (data)	p-value
1	1	0.500
2	2	0.250
3	3	0.125
4	4	0.063
5	5	0.031
6	6	0.016
7	7	0.008

Hypotheses:
H_0: it's a regular coin
H_A: the coin has 2 Heads

Experiment: 7 tosses
Data: 7 Heads
Result: p-value=0.008
Conclusion: reject H_0

Have we proved H_A?
No

Is 0.008 the likelihood that the results are *due* to chance?
No

Is 0.008 the probability that H_0 is true?
No

Is 0.992 (1–0.008) the probability that H_A is true?
No

0.008 is the probability of getting 7 Heads if it were a regular coin (H_0)

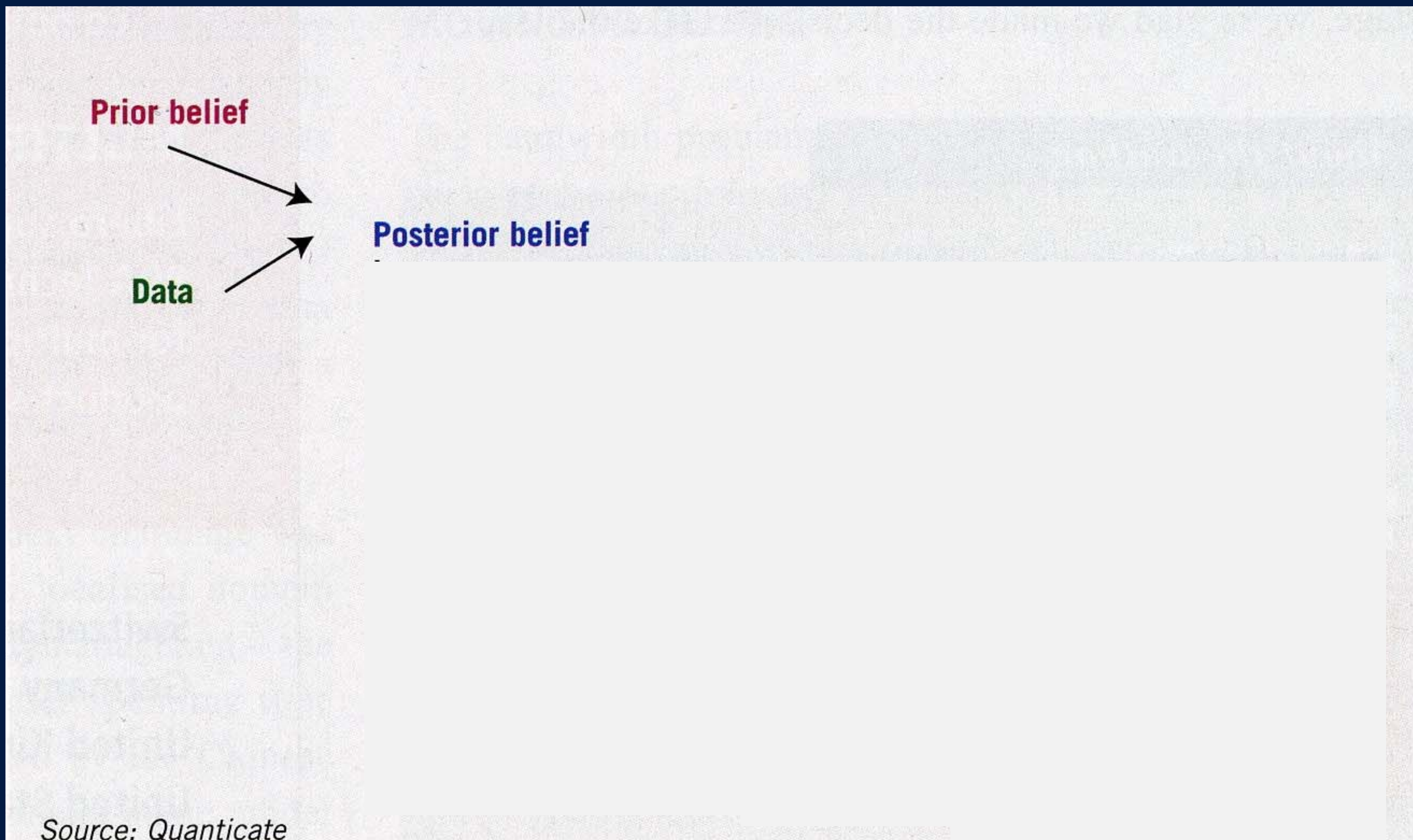
Definition of p-value

The p-value is the probability of obtaining a result as extreme or more extreme than the one obtained, <u>if</u> H_0 were actually true

If p-value $\leq \alpha$ (alpha), reject H_0
If p-value $> \alpha$ (alpha), do not reject H_0

Commonly used alpha levels: 0.05 or 0.01

Bayesian Approach



Underwood, August 2011

Hypotheses:

H_0: it's a regular coin: $P(H) = p = 1/2$

H_A: the coin has 2 Heads: $P(H) = p = 1$

Bayesian Approach

If we start with the belief (prior) that H_0 and H_A are equally likely,

Updating the Distribution of p=P(H) with Data

Hypotheses:

H_0: it's a regular coin: P(H) = p = 1/2

H_A: the coin has 2 Heads: P(H) = p = 1

Bayesian Approach

If we start with the belief (prior) that H_0 and H_A are equally likely,

then after 7 tosses (experiment) and 7 Heads (data),

our updated belief (posterior) is:

we're 0.8% sure P(Head)=1/2 (H_0) – it's the regular coin

and 99.2% sure that P(Head)=1 (H_A) – it's the 2-H coin

Hypotheses:

H_0: it's a regular coin

H_A: the coin has 2 Heads

# of Tosses	# of Heads (data)	Frequentist's p-value
1	1	0.500
2	2	0.250
3	3	0.125
4	4	0.063
5	5	0.031
6	6	0.016
7	7	0.008

Toss #	Result (data)	Bayesian's posterior prob. of regular coin
1	H	0.333
2	H	0.200
3	H	0.111
4	H	0.059
5	H	0.030
6	H	0.015
7	H	0.008

What is the connection between alpha (α) and p-value?

If the p-value is less than α (typically 0.05), the null hypothesis (e.g. of no difference) is rejected, and the result is declared statistically significant at the 5% alpha level

If the p-value is greater than α, the result is not statistically significant at the 5% alpha level

What is the connection between p-value and sample size?

Randomized Controlled Trial

Objective: To compare 2 treatments

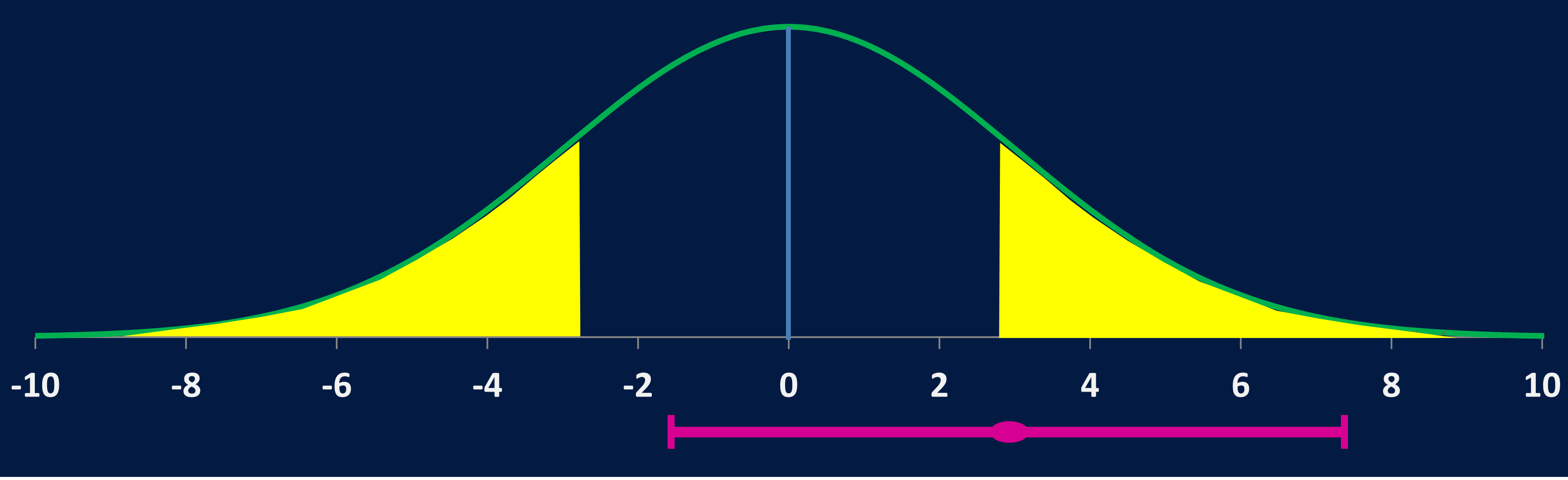
Data are collected. Analysis is done.

Result: p-value = 0.something

Distribution of the difference between the 2 treatment groups, IF in fact there is no difference

Observed treatment effect = 3

One-sided p-value = 0.16
Two-sided p-value = 0.32



Distribution of the difference between the 2 treatment groups, IF in fact there is no difference

N is increased – everything else remains the same

Observed treatment effect = 3

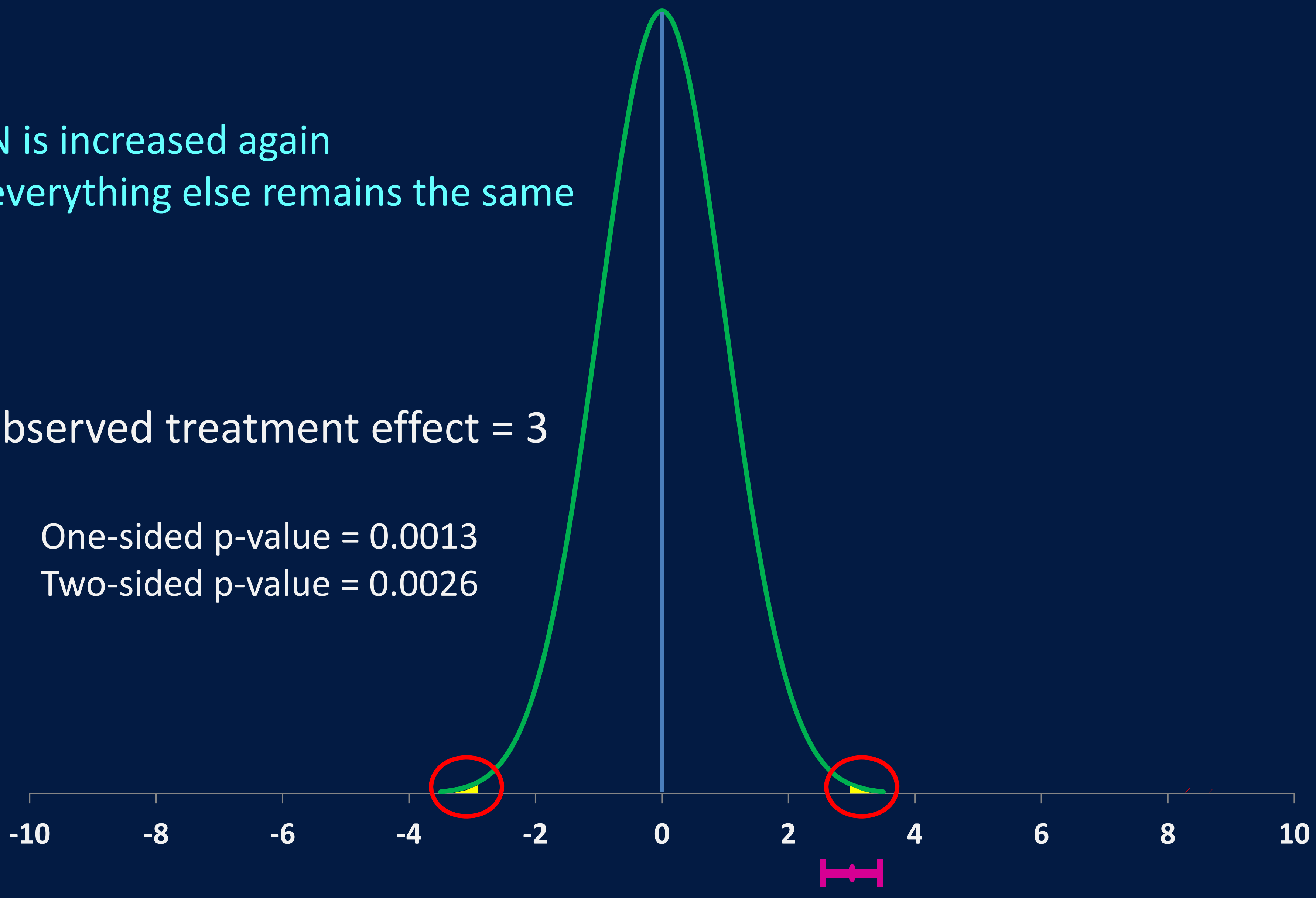
One-sided p-value = 0.067
Two-sided p-value = 0.134

Distribution of the difference between the 2 treatment groups, IF in fact there is no difference

N is increased again
everything else remains the same

Observed treatment effect = 3

One-sided p-value = 0.0013
Two-sided p-value = 0.0026



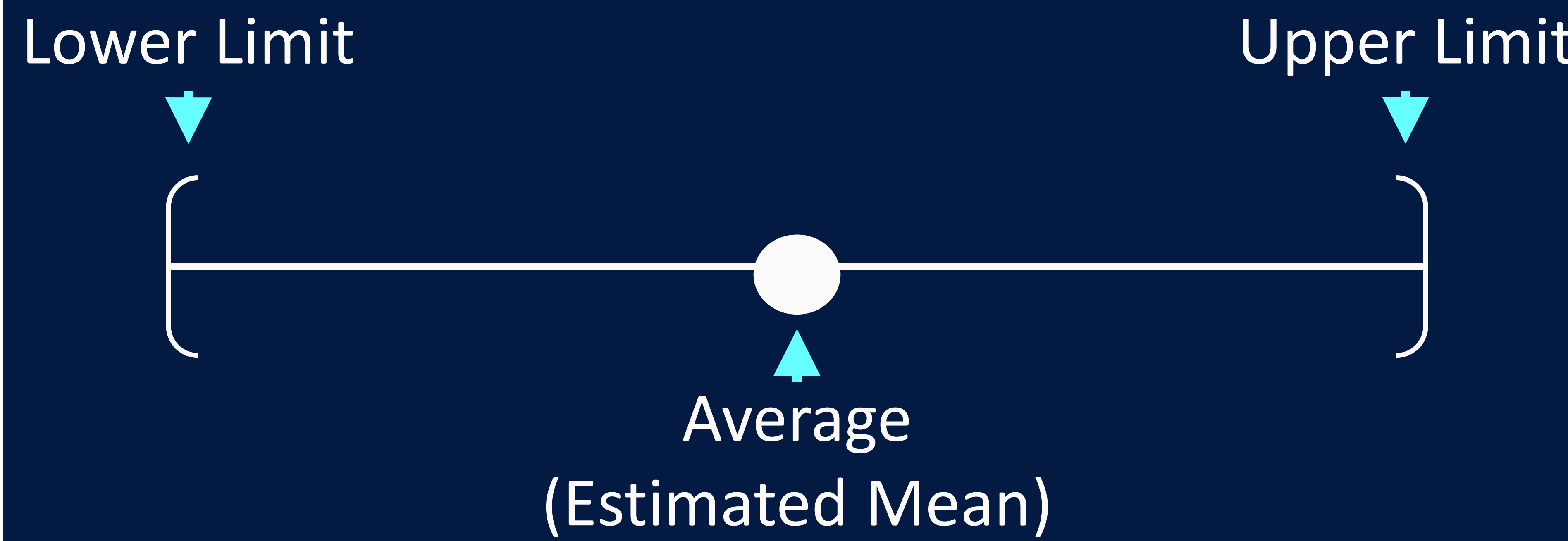
What’s the point?

There are two ways to get statistically significant results… guaranteed!

1. Analyze a very large sample

What is the connection between confidence intervals and hypothesis testing?

95% Confidence Intervals

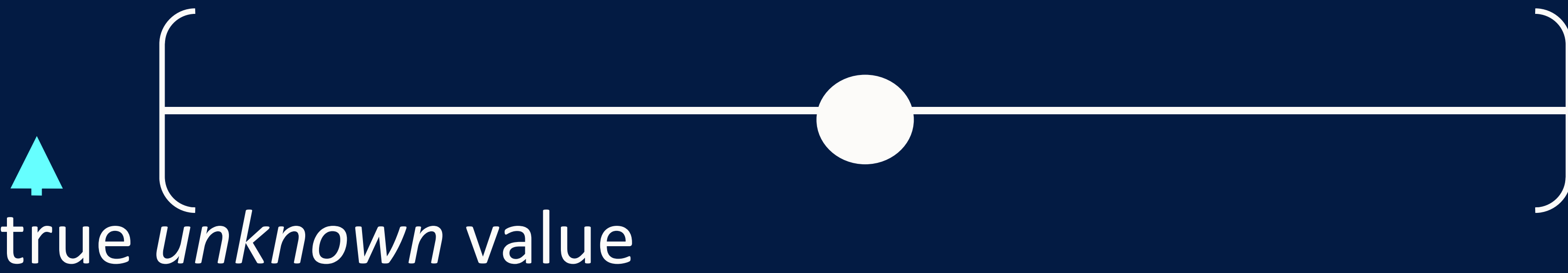


95% Confidence Intervals
(natural perspective)

There is a 95% chance that the true *unknow*n value is inside the confidence interval

We are 95% confident that the true *unknown* value is somewhere within the confidence interval

95% Confidence Intervals
(natural perspective)



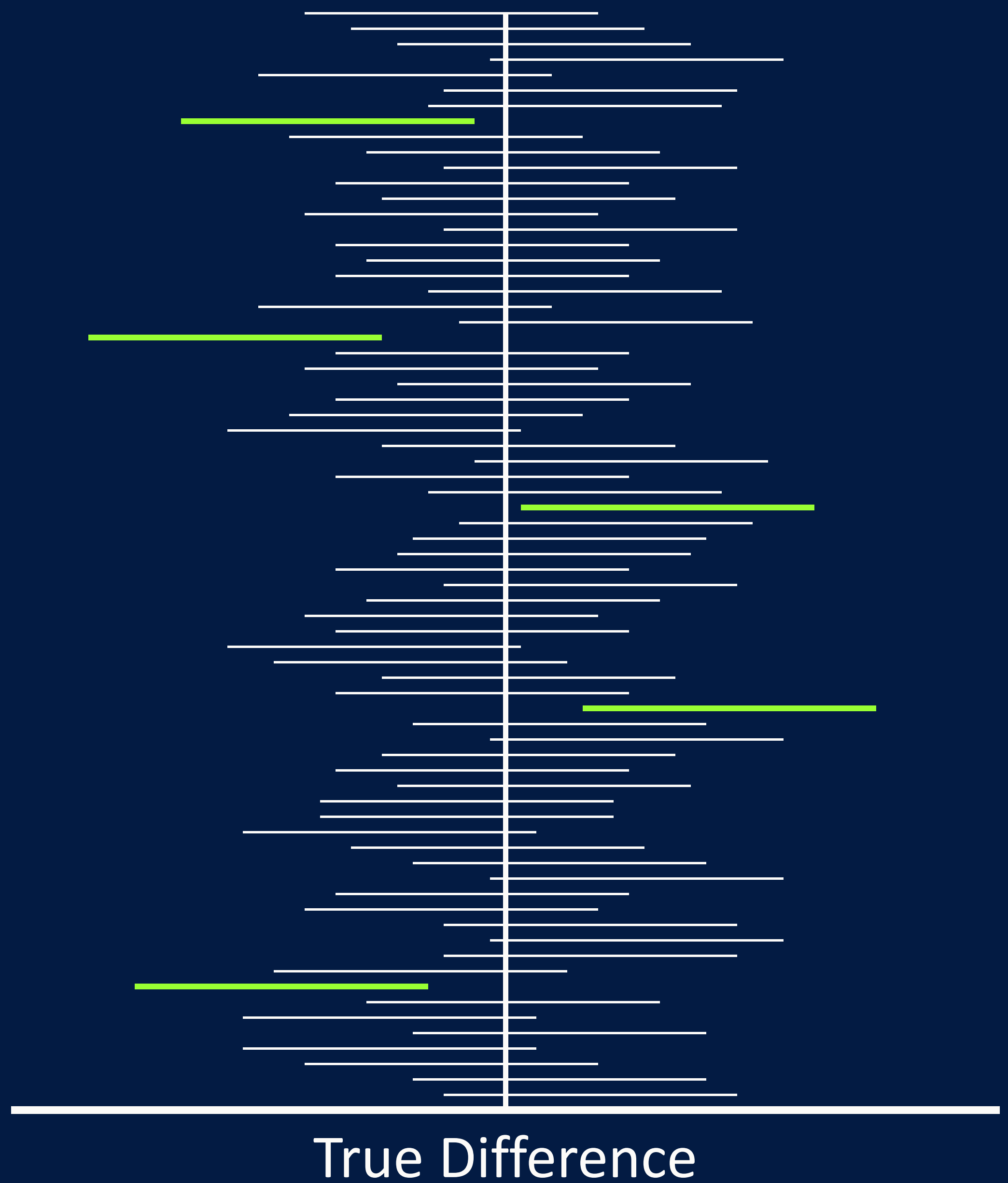
95% Confidence Intervals
(frequentist's pure perspective)

There is a 95% chance that the confidence interval covers the true *unknown* value

95% Confidence Intervals
(frequentist’s pure perspective)

true *unknown* value

95% Confidence Intervals



What is the connection between confidence intervals and hypothesis testing?

If the 95% confidence interval does not include the value of the null hypothesis (e.g. of zero difference), the result is statistically significant at the 5% alpha level

If it does, the result is not statistically significant at the 5% alpha level

Outline

- Fundamentals of Hypothesis Testing
- Superiority vs. Non-Inferiority vs. Equivalence
- Multiple Comparisons (Multiplicity Adjustment)
- Bottom-Line Key Points

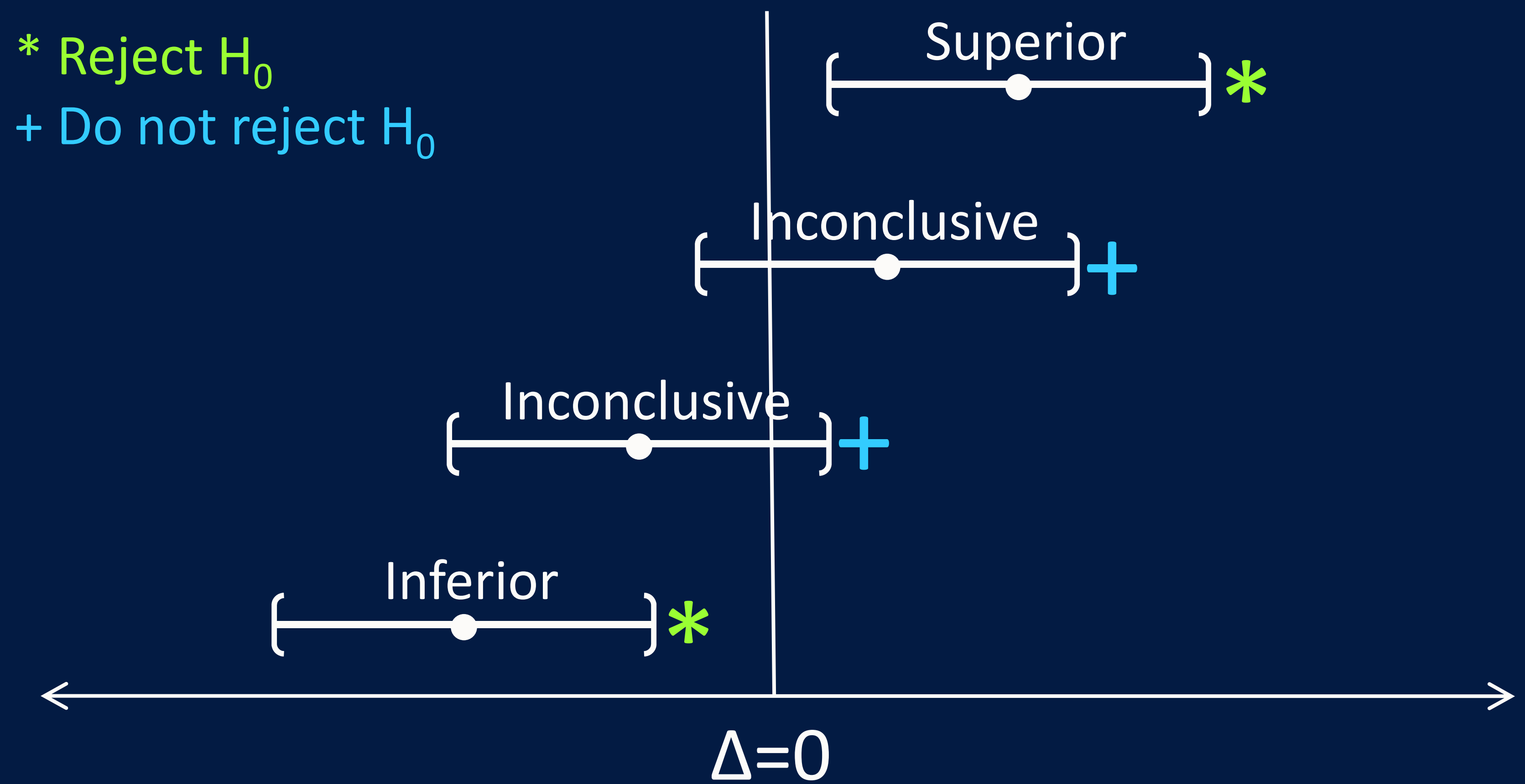
Superiority

Clinical hypothesis:
Experimental treatment is more effective than the control treatment

Statistical hypotheses:
Null hypothesis H_0: Experimental = Control
Alternative hypothesis H_A: Experimental ≠ Control

We expect (hope) to reject H_0 in favor of H_A

Superiority



95% confidence intervals around the difference: Experimental – Control
High numbers (on the right) represent good outcome

Based on Piaggio 2006

Non-Inferiority

Clinical hypothesis:

Experimental treatment is not less effective than the control treatment

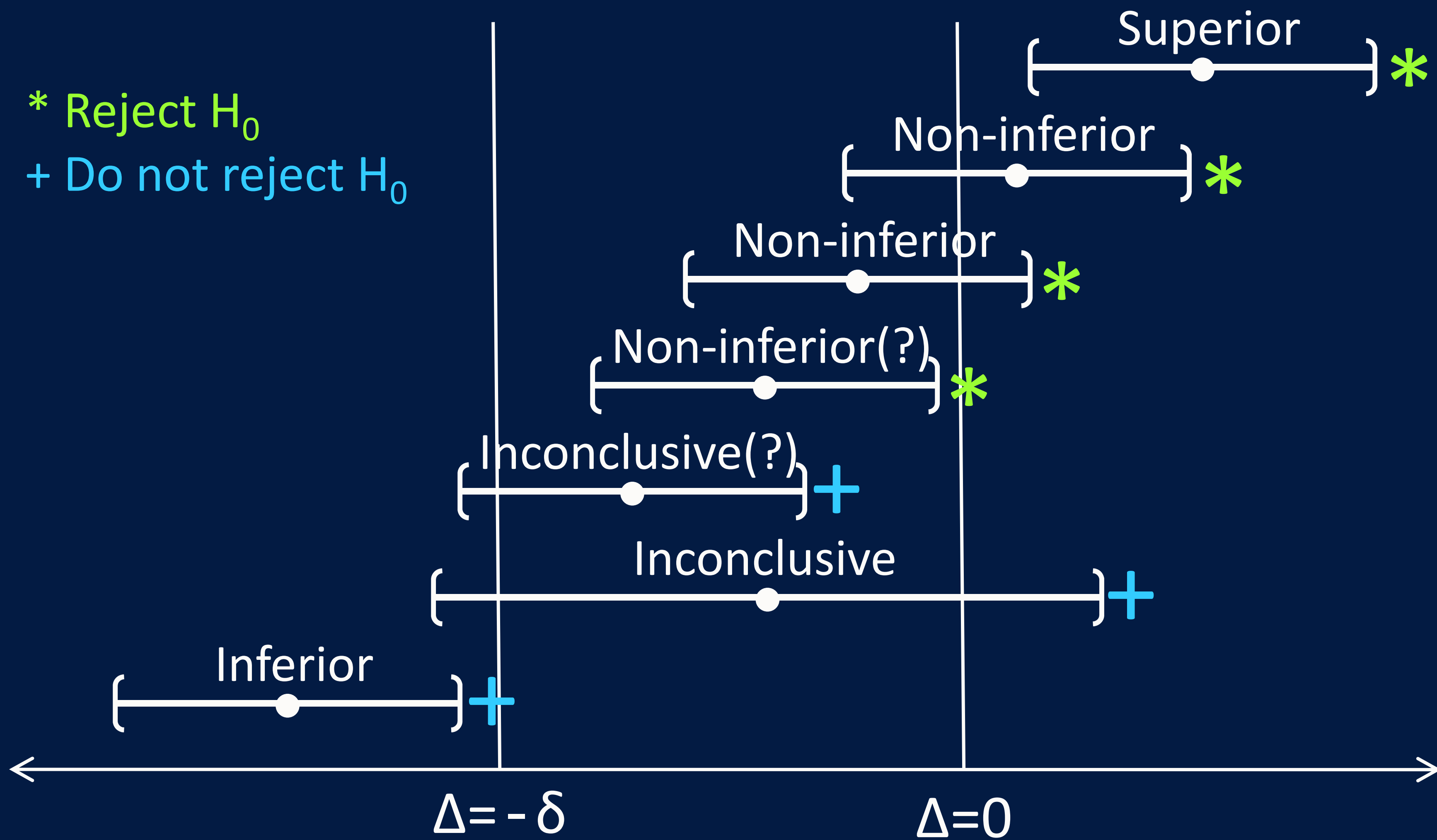
Statistical hypotheses:

Null hypothesis H_0: Experimental < Control – δ

Alternative hypothesis H_A: Experimental $\geq$ Control – δ

We expect (hope) to reject H_0 in favor of H_A

Non-Inferiority



95% confidence intervals around the difference: Experimental – Control

High numbers (on the right) represent good outcome

Based on Piaggio 2006

Equivalence

Clinical hypothesis:

Experimental treatment is as effective as the control treatment

Statistical hypotheses:

Null hypothesis H_0:

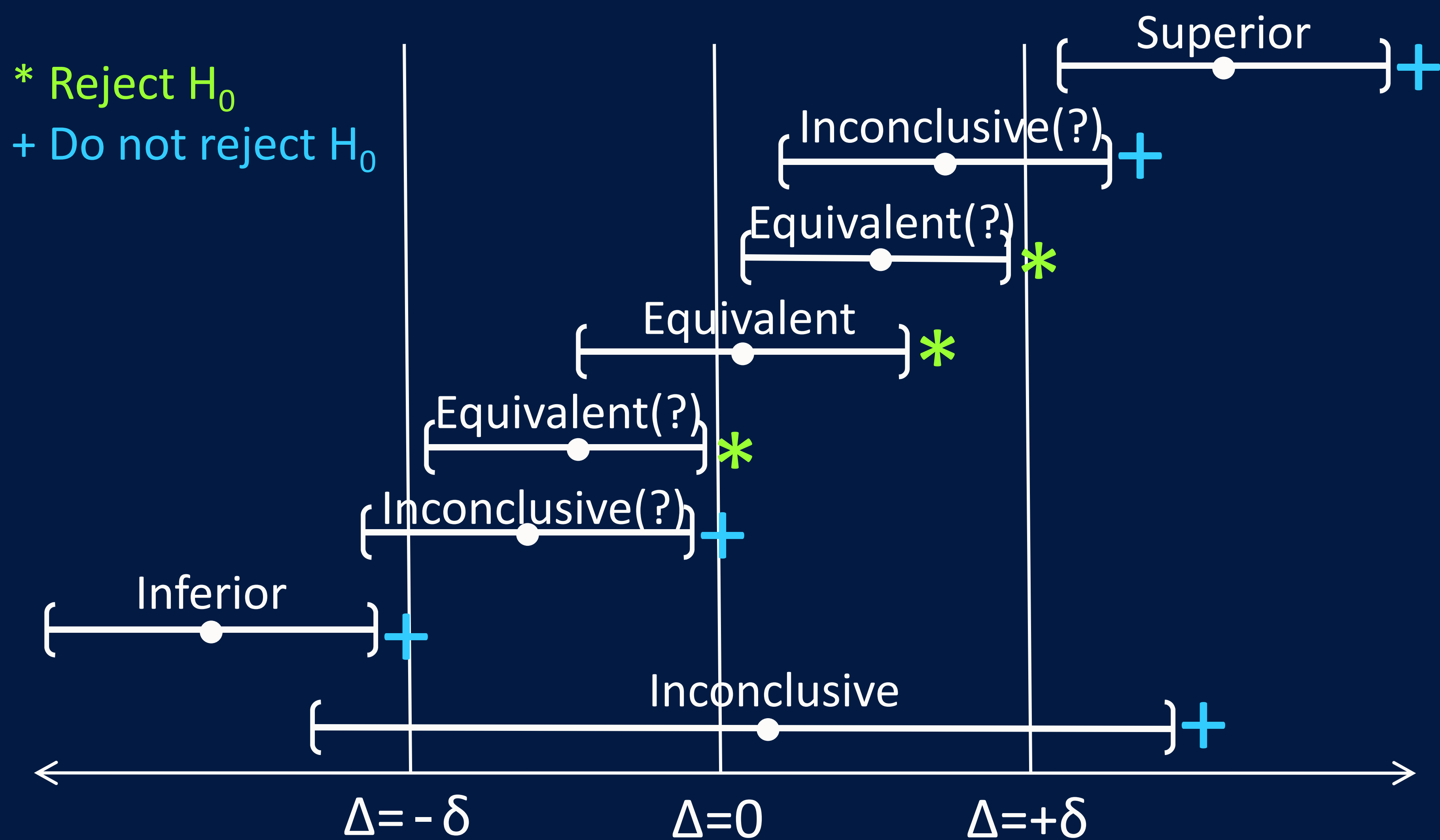
Experimental < Control – δ <u>or</u> Experimental > Control + δ

Alternative hypothesis H_A:

Control – δ ≤ Experimental ≤ Control + δ

We expect (hope) to reject H_0 in favor of H_A

Equivalence



95% confidence intervals around the difference: Experimental – Control

High numbers (on the right) represent good outcome

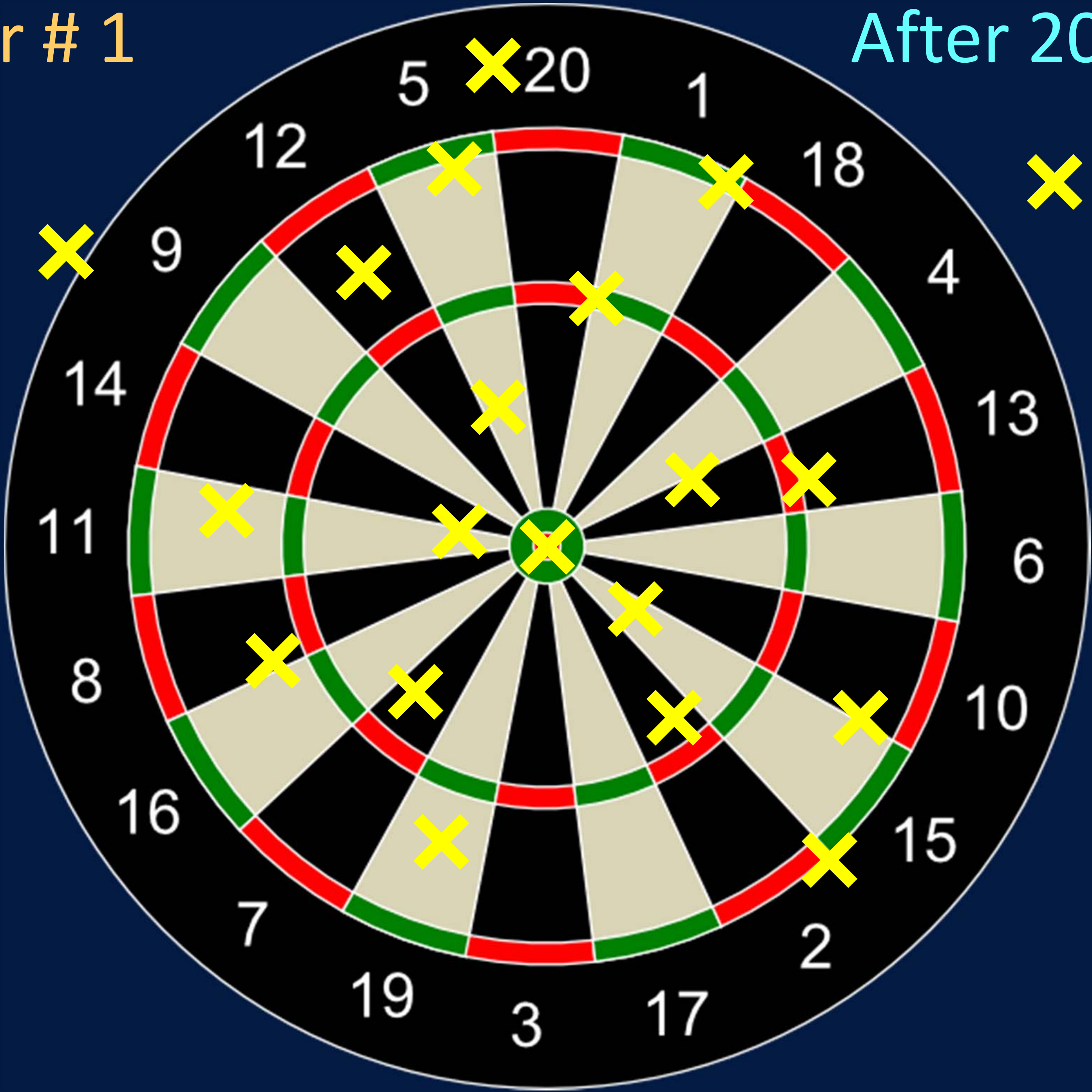
Based on Piaggio 2006

Outline

- Fundamentals of Hypothesis Testing
- Superiority vs. Non-Inferiority vs. Equivalence
- Multiple Comparisons (Multiplicity Adjustment)
- Bottom-Line Key Points

Player # 1

After 20 shots



Player # 2

After 1 shot

When to do multiplicity adjustment?

Formally, whenever there are more than one primary endpoint (or primary hypothesis), more than two treatment conditions, more than one dose vs. placebo, or more than one time point

Informally, whenever there are more than one secondary analysis, including subgroup analyses

The second way….

There are two ways to get statistically significant results… guaranteed!

1. Analyze a very large sample

2. Keep trying different statistical tests on different assessments (outcomes) or on different subgroups of the data

Outline

- Fundamentals of Hypothesis Testing
- Superiority vs. Non-Inferiority vs. Equivalence
- Multiple Comparisons (Multiplicity Adjustment)
- Bottom-Line Key Points

Bottom-Line Key Points

- Statistical inference uses results from a sample from the population of interest to draw conclusions about the population
- The null hypothesis is set up with the hope that it will be rejected
- Alpha (α) is the chance of making a Type I error, i.e. of concluding that there is a difference when in fact there isn't
- Beta (β) is the chance of making a Type II error, i.e. of concluding that there isn't a difference when in fact there is
- Power = $1 - \beta$ = the chance of concluding that there is a difference when in fact there is

Bottom-Line Key Points (cont'd)

- The investigator controls the chance of making a Type I error (alpha) *and* the chance of making a Type II error (beta) via the sample size
- P-value is the probability of obtaining a result as extreme or more extreme than the one obtained, <u>if</u> there were no difference
- Statistical significance does not mean clinical importance
- Confidence intervals are very useful to better understand results
- Multiplicity adjustment is needed with more than one primary hypothesis
- Bayesian approach is gaining popularity as being more intuitive, and is worth considering

The End

Thank you for your attention
I hope this was worth your time

References

Piaggio G et al., *Reporting of Noninferiority and Equivalence Randomized Trials: An Extension of the CONSORT Statement*, JAMA, 2006, 295:1152-1160

Underwood D, *The Profitable Pause*, International Clinical Trials, August 2011, Issue 21, 56-60

Questions / Comments