Interpreting Results from Clinical Research

Dr Suman Kumar (@sumankumarpram1), Command Hospital (EC), Kolkata

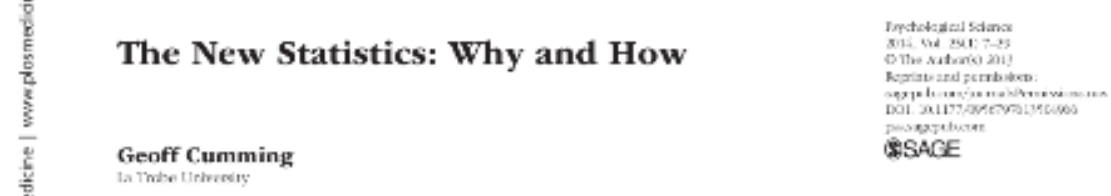
01 April 2015

Background

80% of results of all published research are "FALSE"

There is increasing concern that most current published research findings are false. (Ioannidis, 2005, abstract)

It is time for researchers to avail themselves of the full arsenal of quantitative and qualitative statistical tools. . . . The current practice of focusing exclusively on a dichotomous reject-nonreject decision strategy of null hypothesis testing can actually impede scientific progress. . . . The focus of research should be on . . . what data tell us about the magnitude of effects, the practical significance of effects, and the steady accumulation of knowledge. (Kirk, 2003, p. 100)



Huge problem of results not holding good on replication

- Wastage of resources: patients, time, money
- Wrong, sometimes fatal clinical decisions **(Ethical issues)**

Reasons for the “FALSE” results

- Human emotion

Reasons for the “FALSE” results

- Human emotion
  - Bias by researcher: Hiding data, analyses and results

Reasons for the “FALSE” results

- Human emotion
  - Bias by researcher: Hiding data, analyses and results
  - Bias by editorial staff: Preferential publication of positive results

Reasons for the “FALSE” results

- Human emotion
  - Bias by researcher: Hiding data, analyses and results
  - Bias by editorial staff: Preferential publication of positive results
  - Positive results are linked with professional growth

Reasons for the “FALSE” results

- Human emotion
  - Bias by researcher: Hiding data, analyses and results
  - Bias by editorial staff: Preferential publication of positive results
  - Positive results are linked with professional growth
- Lack of replication of studies

Reasons for the "FALSE" results

- Human emotion
  - Bias by researcher: Hiding data, analyses and results
  - Bias by editorial staff: Preferential publication of positive results
  - Positive results are linked with professional growth
- Lack of replication of studies
- **Mis-interpretation of statistical results**

Reasons for the "FALSE" results

- Human emotion
  - Bias by researcher: Hiding data, analyses and results
  - Bias by editorial staff: Preferential publication of positive results
  - Positive results are linked with professional growth
- Lack of replication of studies
- **Mis-interpretation of statistical results**
  - Commonest one: *P VALUE*

The problem is so rampant that . . .

Basic and Applied Social Psychology
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hbas20

Editorial

David Trafimow [a] & Michael Marks [a]
[a] New Mexico State University
Published online: 12 Feb 2015.

The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP.

Question 1. *Will manuscripts with p-values be desk rejected automatically?*

Answer to Question 1. No. If manuscripts pass the preliminary inspection, they will be sent out for review. But prior to publication, authors will have to remove all vestiges of the NHSTP (*p*-values, *t*-values, *F*-values, statements about "significant" differences or lack thereof, and so on).

R-bloggers
R news and tutorials contributed by (573) R bloggers

Worry about correctness and repeatability, not p-values

April 5, 2013
By John Mount

Essay

Why Most Published Research Findings Are False

PLoS Medicine | www.plosmedicine.org 0696 August 2005 | Volume 2 | Issue 8 | e124

John P. A. Ioannidis

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles should be interpreted based only on *p*-values. Research findings are defined

Actual meaning of “p value”

Ronald Fisher

From Wikipedia, the free encyclopedia

Sir **Ronald Aylmer Fisher** FRS[2] (17 February 1890 – 29 July 1962) was an English statistician, evolutionary biologist, mathematician, geneticist, and eugenicist. Fisher is known as one of the chief architects of the neo-Darwinian synthesis, for his important contributions to statistics, including the analysis of variance (ANOVA), method of maximum likelihood, fiducial inference, and the derivation of various sampling distributions, and for being one of the three principal founders of population genetics. Anders Hald called him "a genius who almost single-handedly created the foundations for modern statistical science",[3] while Richard Dawkins named him "the greatest biologist since Darwin".[4]

Sir Fisher

Born	17 February 1890 East Finchley, London, England
Died	29 July 1962 (aged 72) Adelaide, South Australia

- Actual meaning: Probability that the evidence in question or more extreme belongs to the population depicted by null hypothesis given null hypothesis is true (complicated concept!!)

p value < 0.05: THE MAGICAL EXPRESSION

- DISCRIMINATOR FOR ASSESSING SIGNIFICANCE: The ill defined and often non relevant statistical significance.

p value $<$ 0.05: THE MAGICAL EXPRESSION

- DISCRIMINATOR FOR ASSESSING SIGNIFICANCE: The ill defined and often non relevant statistical significance.

- **"p value of difference in mean in reduction in fasting blood sugar levels between drug A (mean 34 mg/dl) and drug B (mean 36 mg/dl) is 0.001 ($<$ 0.05)"**

p value $<$ 0.05: THE MAGICAL EXPRESSION

- DISCRIMINATOR FOR ASSESSING SIGNIFICANCE: The ill defined and often non relevant statistical significance.

- **"p value of difference in mean in reduction in fasting blood sugar levels between drug A (mean 34 mg/dl) and drug B (mean 36 mg/dl) is 0.001 ($<$ 0.05)"**
  - *Drug B is significantly better than drug A*

p value < 0.05: THE MAGICAL EXPRESSION

- DISCRIMINATOR FOR ASSESSING SIGNIFICANCE: The ill defined and often non relevant statistical significance.

- **"p value of difference in mean in reduction in fasting blood sugar levels between drug A (mean 34 mg/dl) and drug B (mean 36 mg/dl) is 0.001 ($<$ 0.05)"**
  - *Drug B is significantly better than drug A*

- **"p = 0.056"**

p value < 0.05: THE MAGICAL EXPRESSION

- DISCRIMINATOR FOR ASSESSING SIGNIFICANCE: The ill defined and often non relevant statistical significance.

- **"p value of difference in mean in reduction in fasting blood sugar levels between drug A (mean 34 mg/dl) and drug B (mean 36 mg/dl) is 0.001 ($<$ 0.05)"**
  - *Drug B is significantly better than drug A*

- **"p = 0.056"**
  - *Drug B is not better than drug A and researcher does his best to reduce p value to less than 0.05*

p value $<$ 0.05: THE MAGICAL EXPRESSION

- DISCRIMINATOR FOR ASSESSING SIGNIFICANCE: The ill defined and often non relevant statistical significance.

- **"p value of difference in mean in reduction in fasting blood sugar levels between drug A (mean 34 mg/dl) and drug B (mean 36 mg/dl) is 0.001 ($<$ 0.05)"**
  - *Drug B is significantly better than drug A*

- **"p = 0.056"**
  - *Drug B is not better than drug A and researcher does his best to reduce p value to less than 0.05*

- **"p = 0.01 is better discriminator than p = 0.046"**

p value < 0.05: THE MAGICAL EXPRESSION

- DISCRIMINATOR FOR ASSESSING SIGNIFICANCE: The ill defined and often non relevant statistical significance.

- **"p value of difference in mean in reduction in fasting blood sugar levels between drug A (mean 34 mg/dl) and drug B (mean 36 mg/dl) is 0.001 ($<$ 0.05)"**
  - *Drug B is significantly better than drug A*

- **"p = 0.056"**
  - *Drug B is not better than drug A and researcher does his best to reduce p value to less than 0.05*

- **"p = 0.01 is better discriminator than p = 0.046"**
  - *Lesser the p value, better it is as discriminator*

Aims of this presentation

- Provide alternatives to p value
- Provide more informative ways to interpret and report results from research

Scenario to be discussed in presentation

- Comparative intervention trial
- Intervention A vs Intervention B
- Outcome of interest: **proportion of developing a given outcome** within a **period of time**
- Our aim is to **compare** Intervention A and Intervention B
  - Difference in proportion *(Risk difference)*
  - Ratio of proportion *(Risk ratio)*
  - Ratio of odds *(Odds ratio)*

Q 1: Comparability of populations

Is population being tested in trial comparable to our population?

- Patient characteristics
- Environment around patients (in hospital and around the place of living)
- Equality of Supportive care
- Similarity in proficiency of measurement of variables (molecular techniques) and outcomes

Q 2: Understanding Effect Size

Effect Size

- **Most important** number we should care of
- Population characteristic
  - Usually we can only estimate it from the sample
- **One population**
  - Mean/median of WBC
  - Proportion surviving at the end of 1 year (OS)
  - Incidence rate (Hazard) of relapse over 1 year
  - Cumulative incidence of relapse over 1 year
- **Two populations** (comparision)
  - **Difference (Absolute and relative)**
  - Ratio (Hazard ratio, Odds ratio, Risk ratio)

Example (Difference in proportions)

Example 1

Intervention A (standard of care) and intervention B are given over a period of **1 month**. At the end of **1 year**, 50% of patients in intervention A and 60% of patients in intervention B arm are in remission.

Example 2

Intervention A (standard of care) and intervention B are given over a period of **1 year**. At the end of **5 years**, 2% of patients in intervention A and 1% of patients in intervention B arm relapse.

Is Intervention B better than intervention A (standard of care)?. We will use difference in proportion as our ***Effect Size Measure****.*

Relative vs absolute difference in proportion

- **Relative risk difference (RRD)**

Relative vs absolute difference in proportion

- **Relative risk difference (RRD)**
  - *Example 1:* $(0.6 - 0.5)/0.5 = 0.2 = 20\%$

Relative vs absolute difference in proportion

- **Relative risk difference (RRD)**
  - *Example 1:* (0.6 - 0.5)/0.5 = 0.2 = *20%*
  - *Example 2:* (0.02 - 0.01)/0.02 = 0.5 = *50%*

Relative vs absolute difference in proportion

- **Relative risk difference (RRD)**
  - *Example 1:* (0.6 - 0.5)/0.5 = 0.2 = *20%*
  - *Example 2:* (0.02 - 0.01)/0.02 = 0.5 = *50%*

- **Absolute risk difference (ARD)**

Relative vs absolute difference in proportion

- **Relative risk difference (RRD)**
  - *Example 1:* (0.6 - 0.5)/0.5 = 0.2 = *20%*
  - *Example 2:* (0.02 - 0.01)/0.02 = 0.5 = *50%*

- **Absolute risk difference (ARD)**
  - *Example 1:* 0.6 - 0.5 = 0.1 = *10%*

Relative vs absolute difference in proportion

- **Relative risk difference (RRD)**
  - *Example 1:* (0.6 - 0.5)/0.5 = 0.2 = *20%*
  - *Example 2:* (0.02 - 0.01)/0.02 = 0.5 = *50%*

- **Absolute risk difference (ARD)**
  - *Example 1:* 0.6 - 0.5 = 0.1 = *10%*
  - *Example 2:* 0.02 - 0.01 = 0.01 = *1%*

Relative vs absolute difference in proportion

- **Relative risk difference (RRD)**
  - *Example 1:* (0.6 - 0.5)/0.5 = 0.2 = *20%*
  - *Example 2:* (0.02 - 0.01)/0.02 = 0.5 = *50%*
- **Absolute risk difference (ARD)**
  - *Example 1:* 0.6 - 0.5 = 0.1 = *10%*
  - *Example 2:* 0.02 - 0.01 = 0.01 = *1%*
- Usually, **RRD is presented in literature** rather than ARD

Relative vs absolute difference in proportion

- **Relative risk difference (RRD)**
  - *Example 1:* (0.6 - 0.5)/0.5 = 0.2 = *20%*
  - *Example 2:* (0.02 - 0.01)/0.02 = 0.5 = *50%*
- **Absolute risk difference (ARD)**
  - *Example 1:* 0.6 - 0.5 = 0.1 = *10%*
  - *Example 2:* 0.02 - 0.01 = 0.01 = *1%*
- Usually, **RRD is presented in literature** rather than ARD
  - Inflates the effect size, especially when risks are nearer to zero (Example 2)

Relative vs absolute difference in proportion

- **Relative risk difference (RRD)**
  - *Example 1:* (0.6 - 0.5)/0.5 = 0.2 = *20%*
  - *Example 2:* (0.02 - 0.01)/0.02 = 0.5 = *50%*
- **Absolute risk difference (ARD)**
  - *Example 1:* 0.6 - 0.5 = 0.1 = *10%*
  - *Example 2:* 0.02 - 0.01 = 0.01 = *1%*
- Usually, **RRD is presented in literature** rather than ARD
  - Inflates the effect size, especially when risks are nearer to zero (Example 2)
  - For example 2, by using intervention B, there is only 1% decrease in relapse (in absolute term), but 50% reduction in relapse, when compared to intervention A

Clinically relevant effect size

- Needs to be defined by user
- Requires thorough knowledge of subject area, expertise and environment
- Example 1: Say, the disease concern is an indolent and non life threatening disease. Improvement of remission rate to 10% may not be clinically relevant

ARD and NNT

- **Number Needed to Treat (NNT) = 1/ARD**
- Very useful effect size measure
- **Example 1:** $NNT = 10$
- **Example 2:** $NNT = 100$

- We need to treat *10 patients* to get 1 extra remission at the end of *1 year* (Example 1) and *100 patients* to prevent 1 extra relapse at the end of *5 years* (Example 2).

ARD and NNT

- **Number Needed to Treat (NNT) = 1/ARD**
- Very useful effect size measure
- **Example 1:** $NNT = 10$
- **Example 2:** $NNT = 100$

- We need to treat *10 patients* to get 1 extra remission at the end of *1 year* (Example 1) and *100 patients* to prevent 1 extra relapse at the end of *5 years* (Example 2).
  - If I get 10 patients of the disease in Example 2 in my centre in a year, I will have to wait for 10 years to get one extra remission after waiting for 5 years (i.e., from 6^{th} to 16^{th} year)

ARD and NNT

- **Number Needed to Treat (NNT) = 1/ARD**
- Very useful effect size measure
- **Example 1:** $NNT = 10$
- **Example 2:** $NNT = 100$

- We need to treat *10 patients* to get 1 extra remission at the end of *1 year* (Example 1) and *100 patients* to prevent 1 extra relapse at the end of *5 years* (Example 2).
  - If I get 10 patients of the disease in Example 2 in my centre in a year, I will have to wait for 10 years to get one extra remission after waiting for 5 years (i.e., from 6^{th} to 16^{th} year)
  - Is intervention B really better for me at my centre??

Q 3: Estimating Effect Size

Population vs Sample

- We donot know the real Effect Size as it is a population characteristic
- We can only estimate it from **Random Sample** chosen from the underlying population by **carrying out experiments**

Q 4: Quality of Effect Size Estimate

Three qualities

- **Validity** of estimate
  - Difference in estimate and actual effect size (Bias)
- **Magnitude** of estimate
  - Greater the magnitude, better it is.
- **Precision** of estimate (denoted by Confidence Interval)
  - Greater the precision, better it is

Q 4a: Validity of effect size estimate (Problem of CONFOUNDERS)

Are groups intervention A and intervention B equal in all respects other than the interventions?

- **Baseline known confounders** are equal between both groups

Are groups intervention A and intervention B equal in all respects other than the interventions?

- **Baseline known confounders** are equal between both groups
  - If not, are they taken care of statistically (Multiple regression analysis, stratified analysis)?

Are groups intervention A and intervention B equal in all respects other than the interventions?

- **Baseline known confounders** are equal between both groups
  - If not, are they taken care of statistically (Multiple regression analysis, stratified analysis)?
- **Baseline unknown confounders** are equal between both groups

Are groups intervention A and intervention B equal in all respects other than the interventions?

- **Baseline known confounders** are equal between both groups
  - If not, are they taken care of statistically (Multiple regression analysis, stratified analysis)?
- **Baseline unknown confounders** are equal between both groups
  - Only way to take care of is by **appropriate randomisation** (randomised allocation of treatment ensures confounders to be distributed equally in both the groups)

Are groups intervention A and intervention B equal in all respects other than the interventions?

- **Baseline known confounders** are equal between both groups
  - If not, are they taken care of statistically (Multiple regression analysis, stratified analysis)?
- **Baseline unknown confounders** are equal between both groups
  - Only way to take care of is by **appropriate randomisation** (randomised allocation of treatment ensures confounders to be distributed equally in both the groups)
- **Blinding** of allocation of intervention arms, taking care of patients, measuring outcomes, performing statistical analyses

Are groups intervention A and intervention B equal in all respects other than the interventions?

- **Baseline known confounders** are equal between both groups
  - If not, are they taken care of statistically (Multiple regression analysis, stratified analysis)?
- **Baseline unknown confounders** are equal between both groups
  - Only way to take care of is by **appropriate randomisation** (randomised allocation of treatment ensures confounders to be distributed equally in both the groups)
- **Blinding** of allocation of intervention arms, taking care of patients, measuring outcomes, performing statistical analyses
  - To maintain equality among both the groups till publishing the results

(contd . . .)

- Equality of **loss to follow up or cross over** between both groups: numbers and reasons

(contd . . .)

- Equality of **loss to follow up or cross over** between both groups: numbers and reasons
  - Intention to Treat vs Per Protocol analysis

(contd . . .)

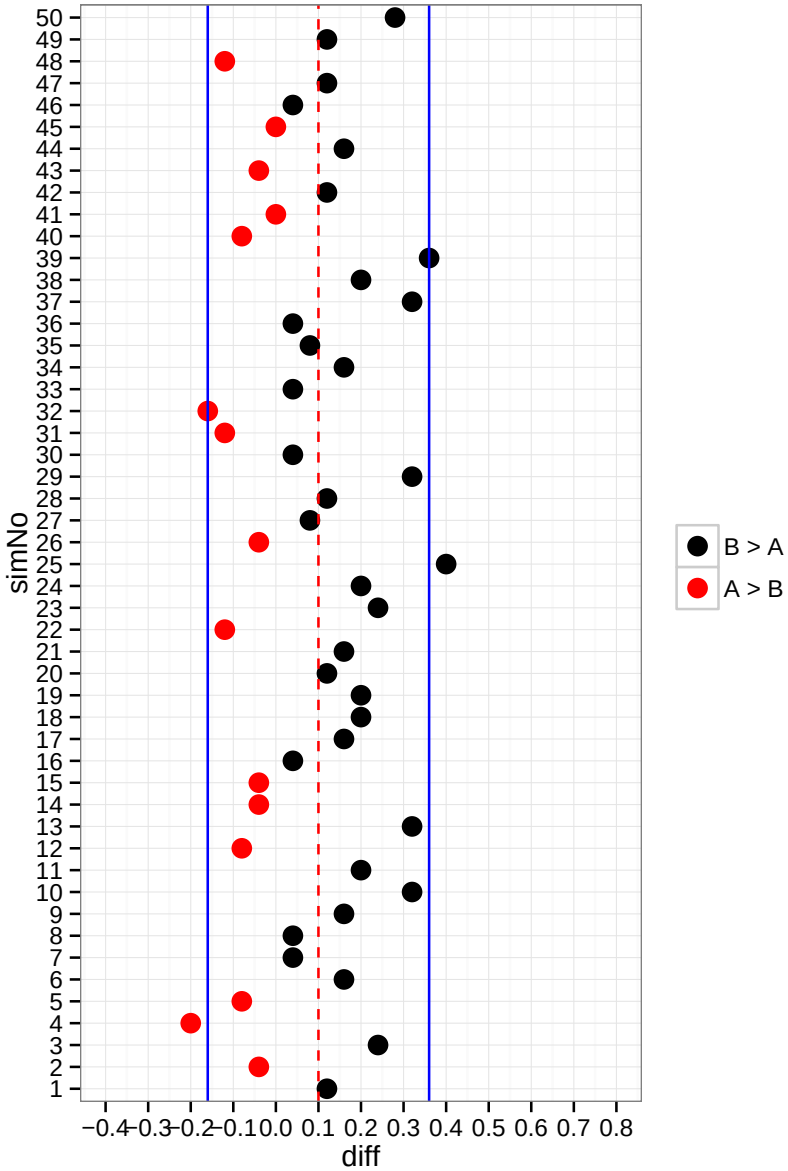
- Equality of **loss to follow up or cross over** between both groups: numbers and reasons
  - Intention to Treat vs Per Protocol analysis
- RCTs yield more valid estimate of Effect Size than observational studies (Cohort, Case Control studies)

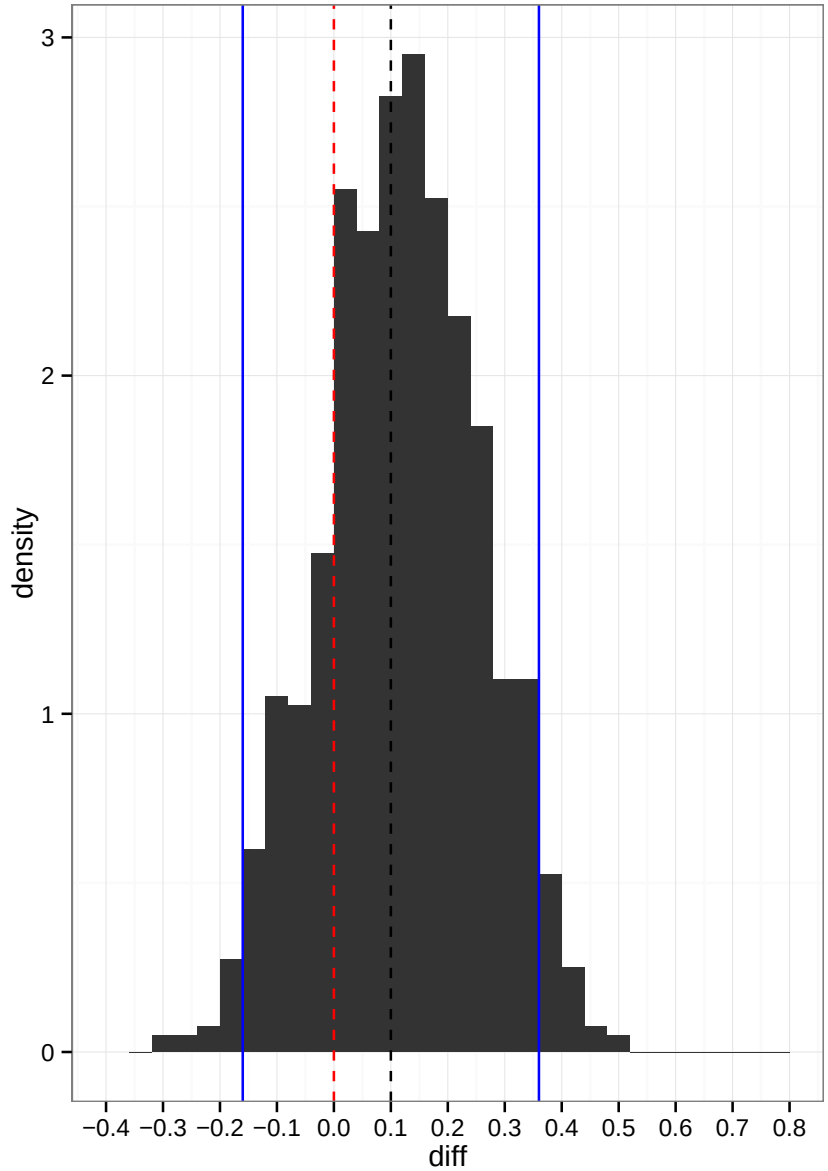
Q 4c: Understanding Precision

Simulation

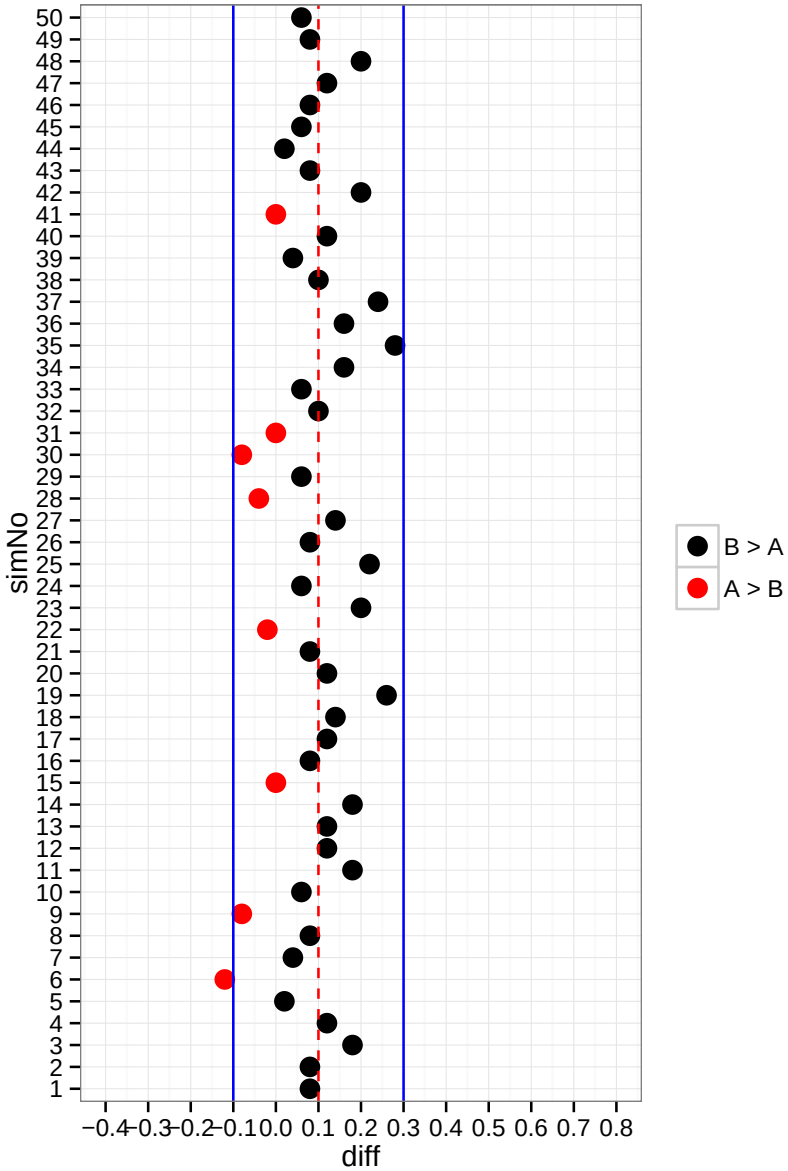
We simulate example 1. Clinically relevant difference between both groups is 0.1. We will draw random samples from population treated with intervention A (prob of remission 0.5) and population treated with intervention B (prob of remission 0.6) 1000 times (equivalent as carrying out 1000 trials). We will compare intervention A and intervention B by difference of proportion.

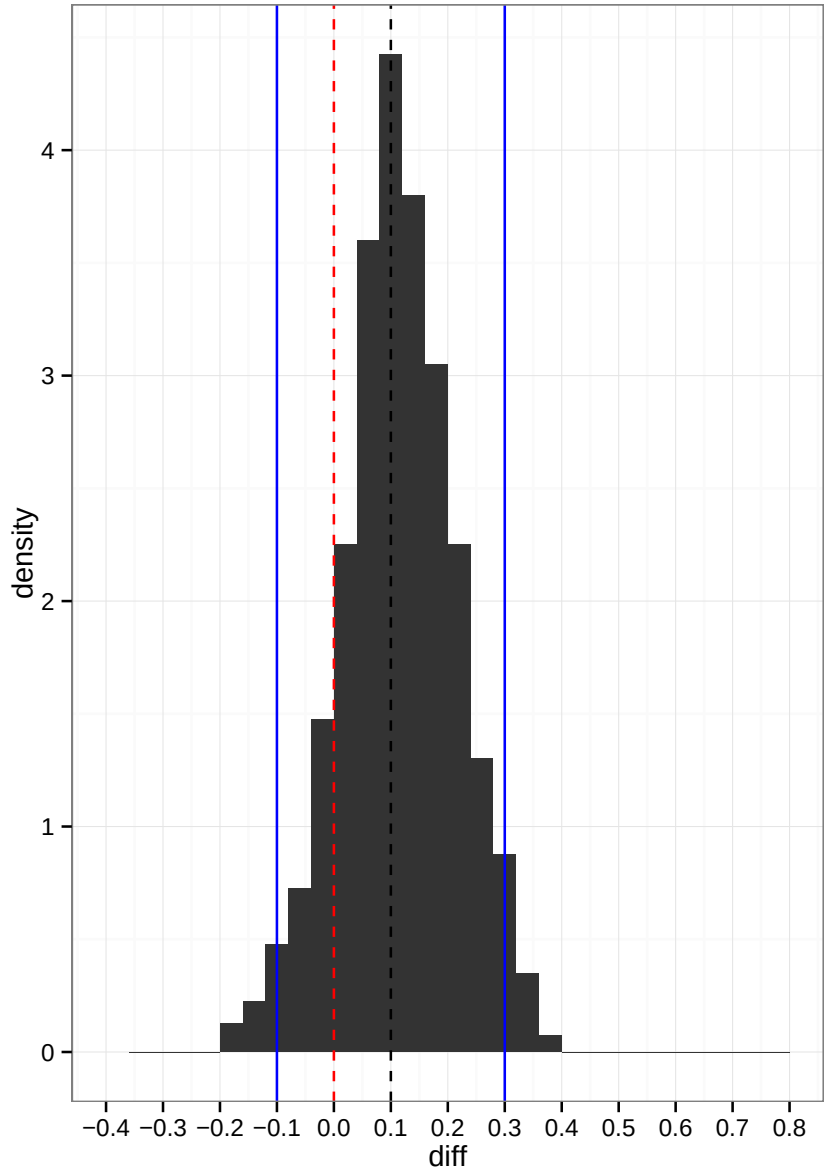
Simulation: probA: 50%, probB: 60%, sample size: 25



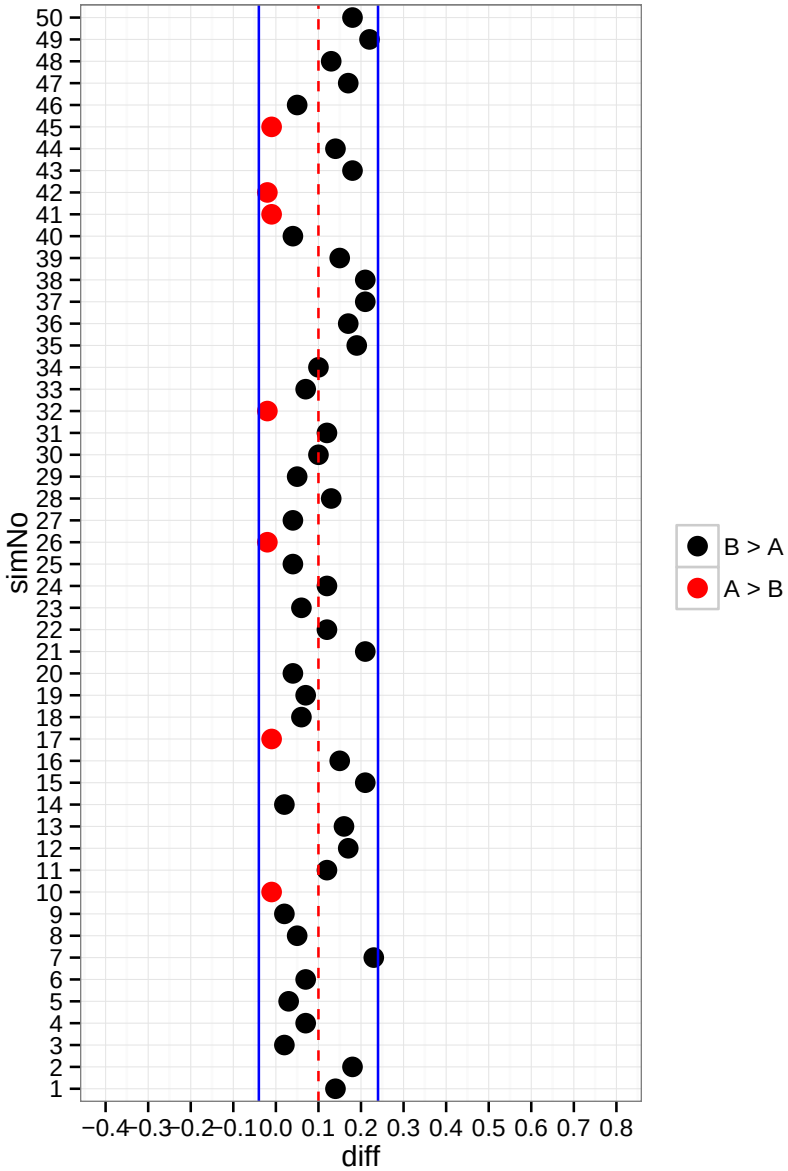


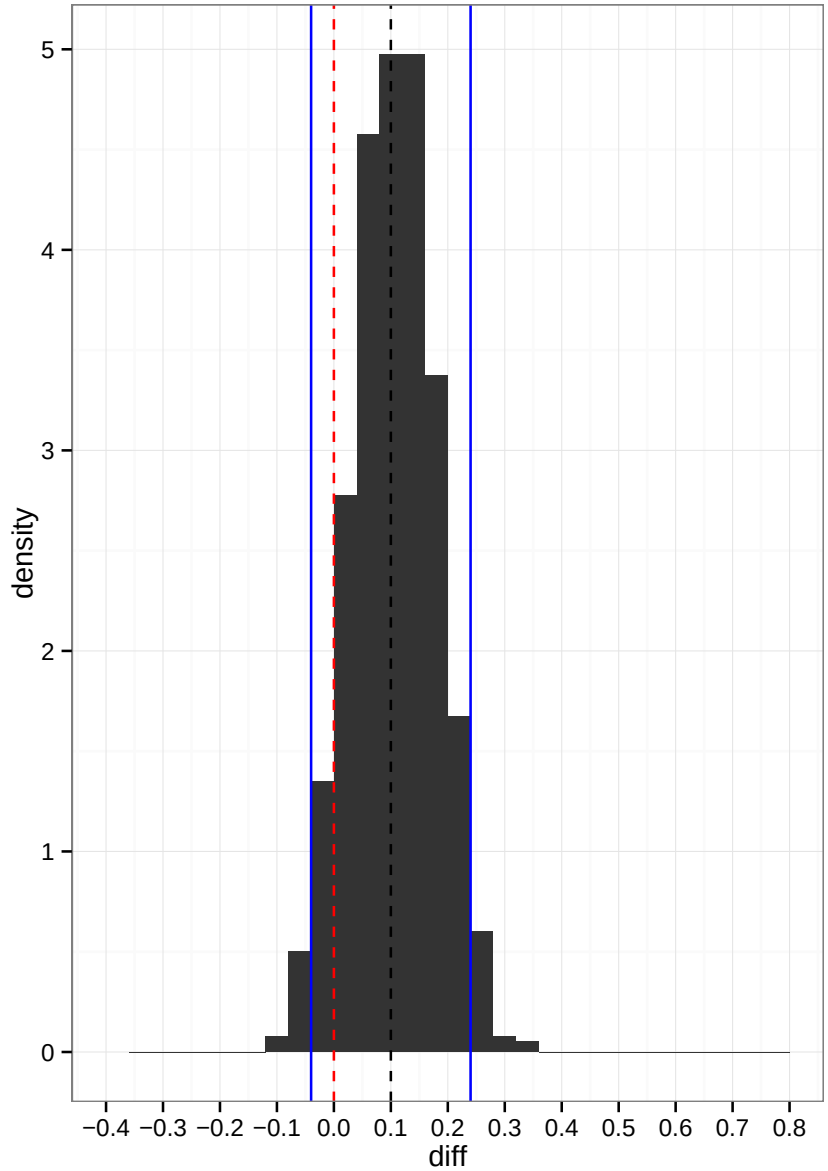
Simulation: probA: 50%, probB: 60%, sample size: 50



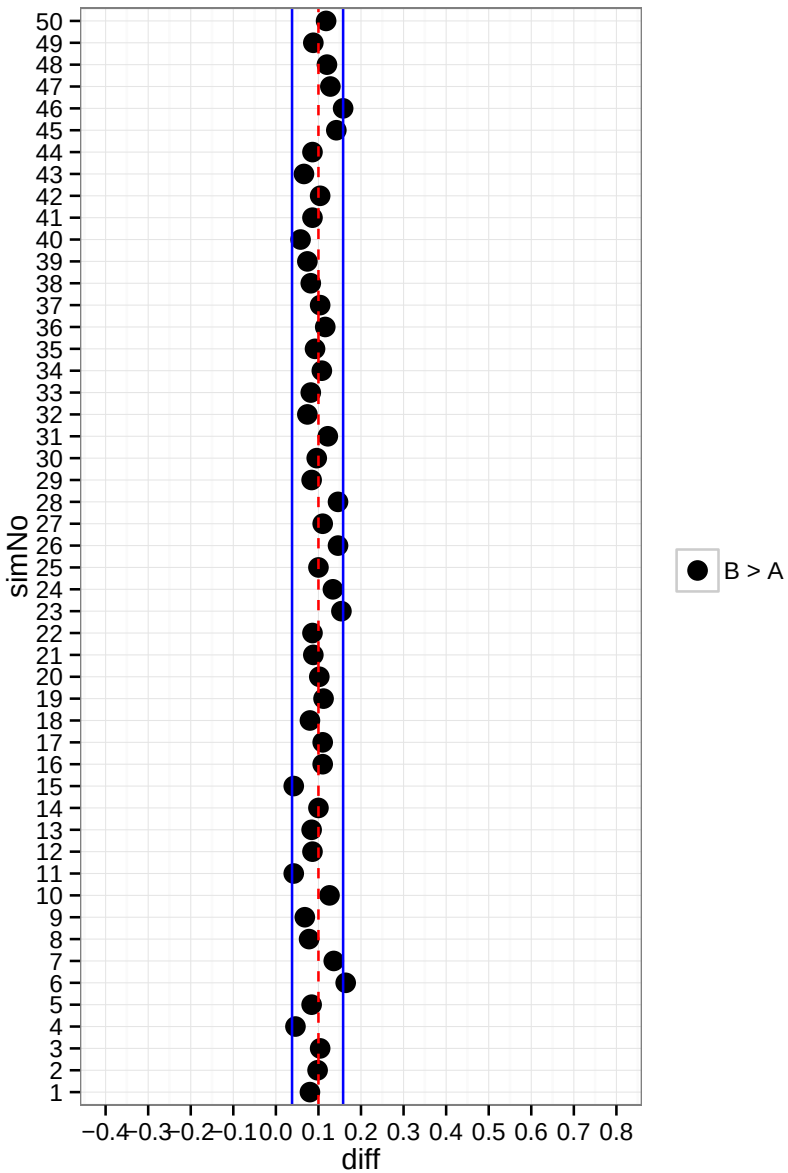


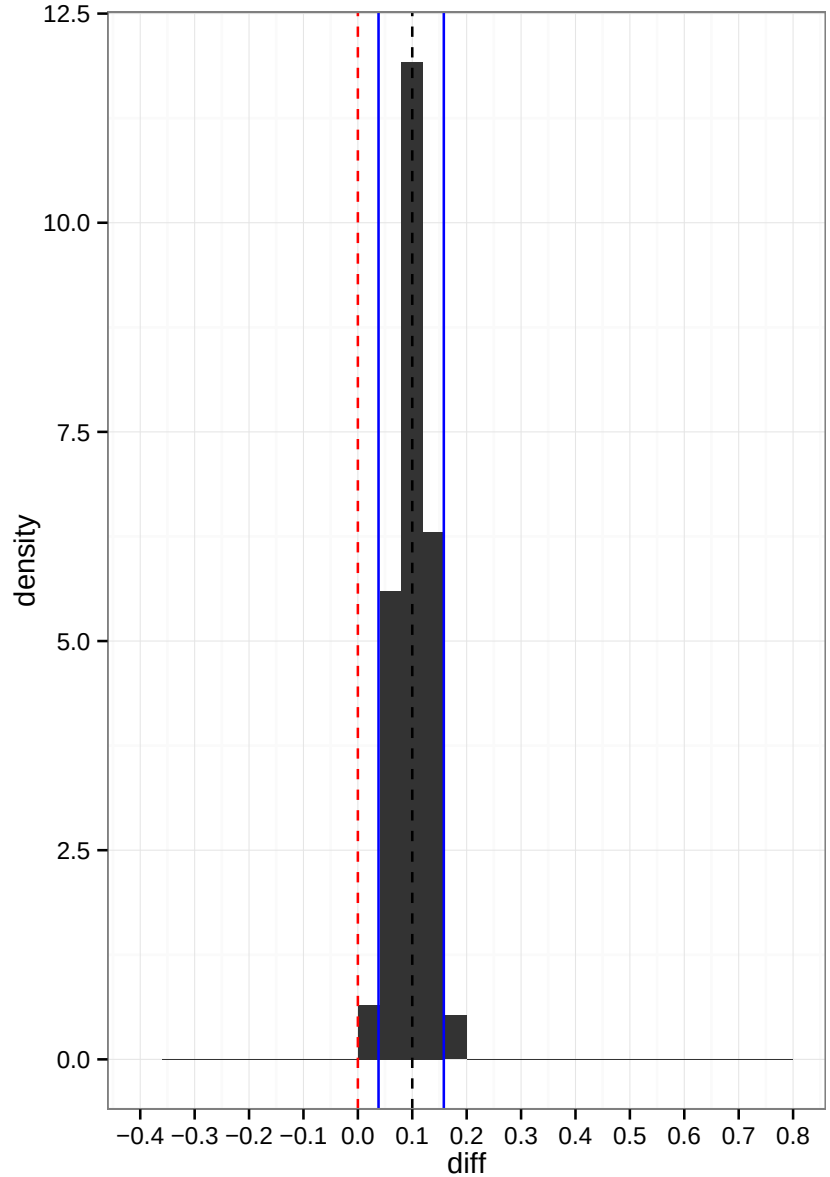
Simulation: probA: 50%, probB: 60%, sample size: 100





Simulation: probA: 50%, probB: 60%, sample size: 500





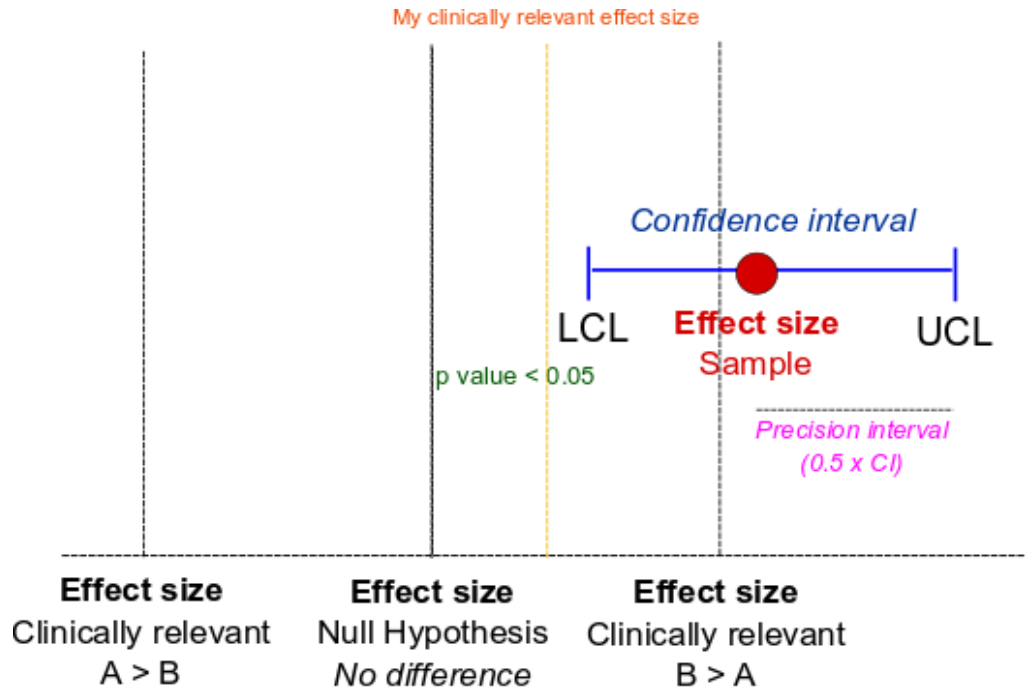
Explanation of simulation

- The **blue** lines, which denote the **bound for mid 95% of all the estimates** is the measure of **Precision**, the width of which is the width of corresponding confidence interval

Explanation of simulation

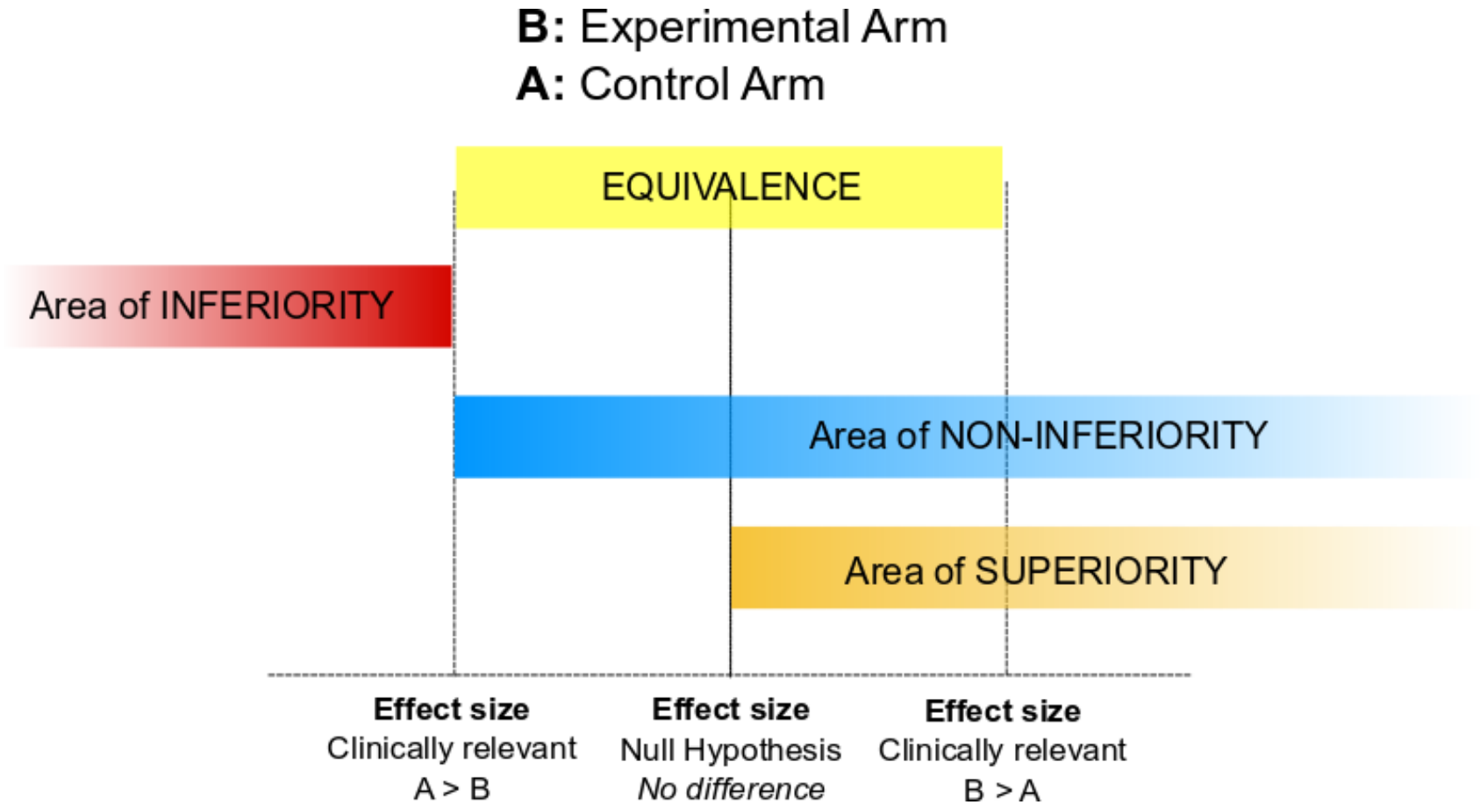
- The **blue** lines, which denote the **bound for mid 95% of all the estimates** is the measure of **Precision**, the width of which is the width of corresponding confidence interval
- The precision increases (width of the distribution decreases) with increasing the sample size

Understanding Confidence Interval



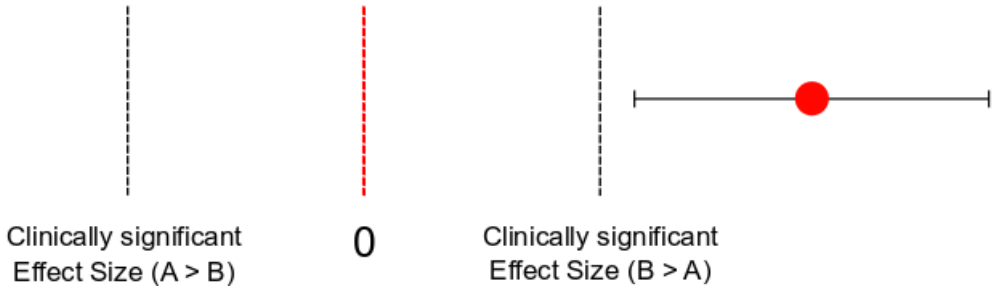
- Width of LCL - UCL, dependent on **variability** of sample and **sample size**
- Any of the points bounded by LCL and UCL can be the **Population Effect Size** (with 95% certainty)

Understanding clinically relevant regions



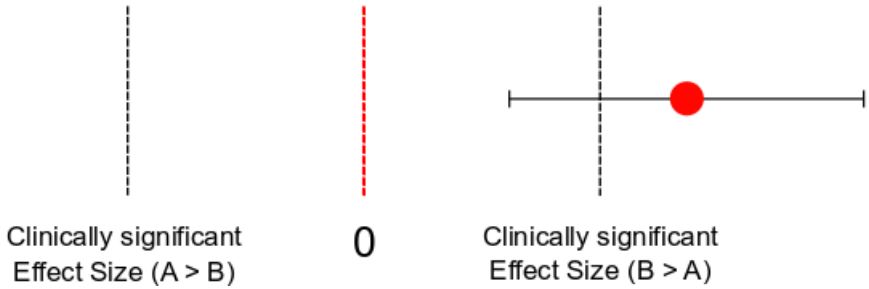
Various positions of CI

Scenario 1



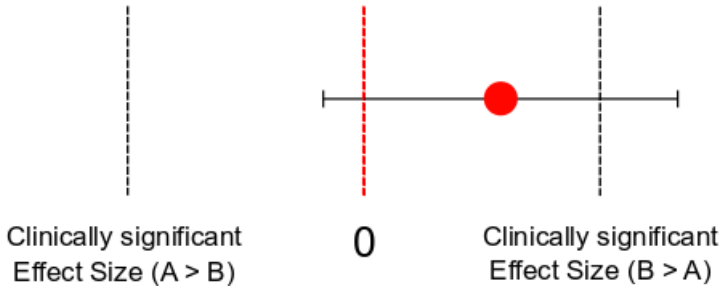
- The population effect size is more than 0 and clinically significant effect size (Intervention B is definitely clinically better to Intervention A)

Scenario 2



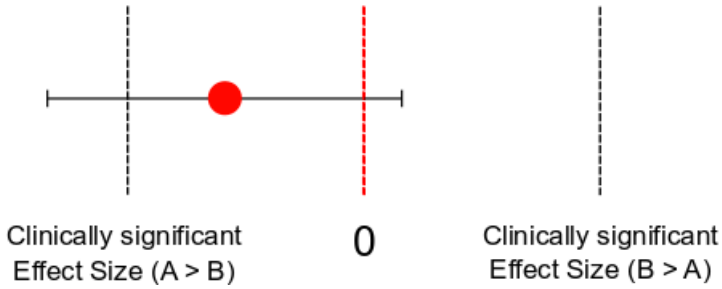
- The population effect size is more than 0 and but may not be more than clinically significant effect size (Intervention B is better than intervention A but may not be clinically relevant).

Scenario 3



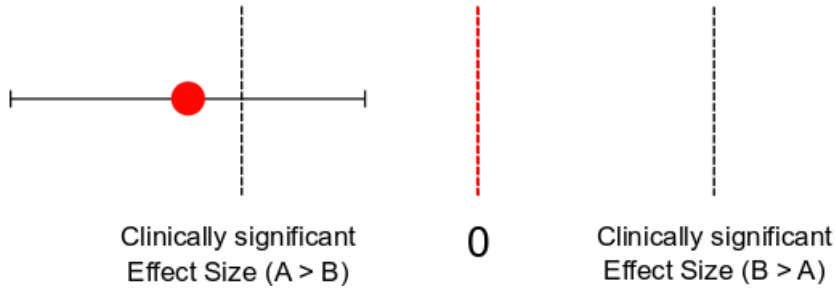
- The population effect size crosses 0, **we cannot say that B is better than A**. We can say that B is **not inferior** to A. We should not say that B is not better than A, we need to be **more precise.**
- Absence of evidence that a fact is true does not mean that fact is not true.

Scenario 4



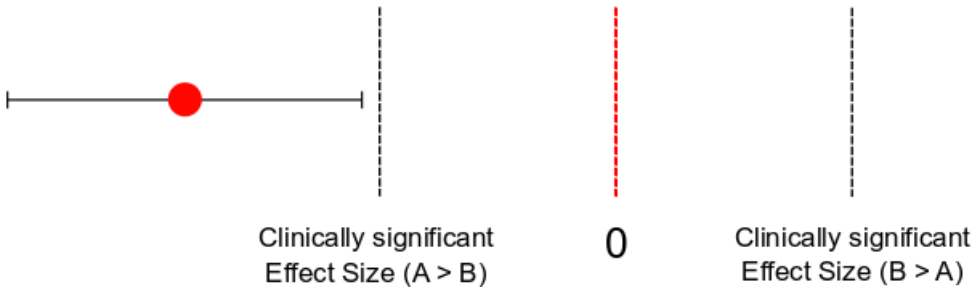
- We are not sure that B is not inferior to A. We are sure that B is not better than A in a clinically relevant manner.

Scenario 5



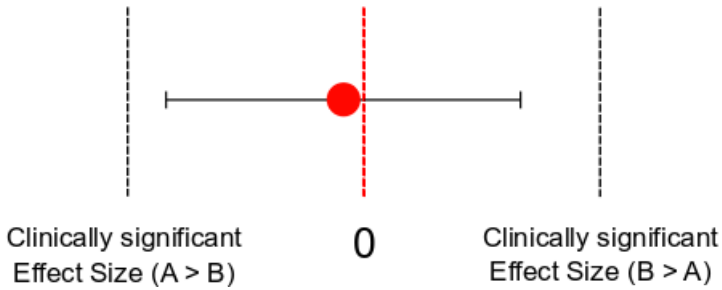
- We are sure that B is not better than A

Scenario 6



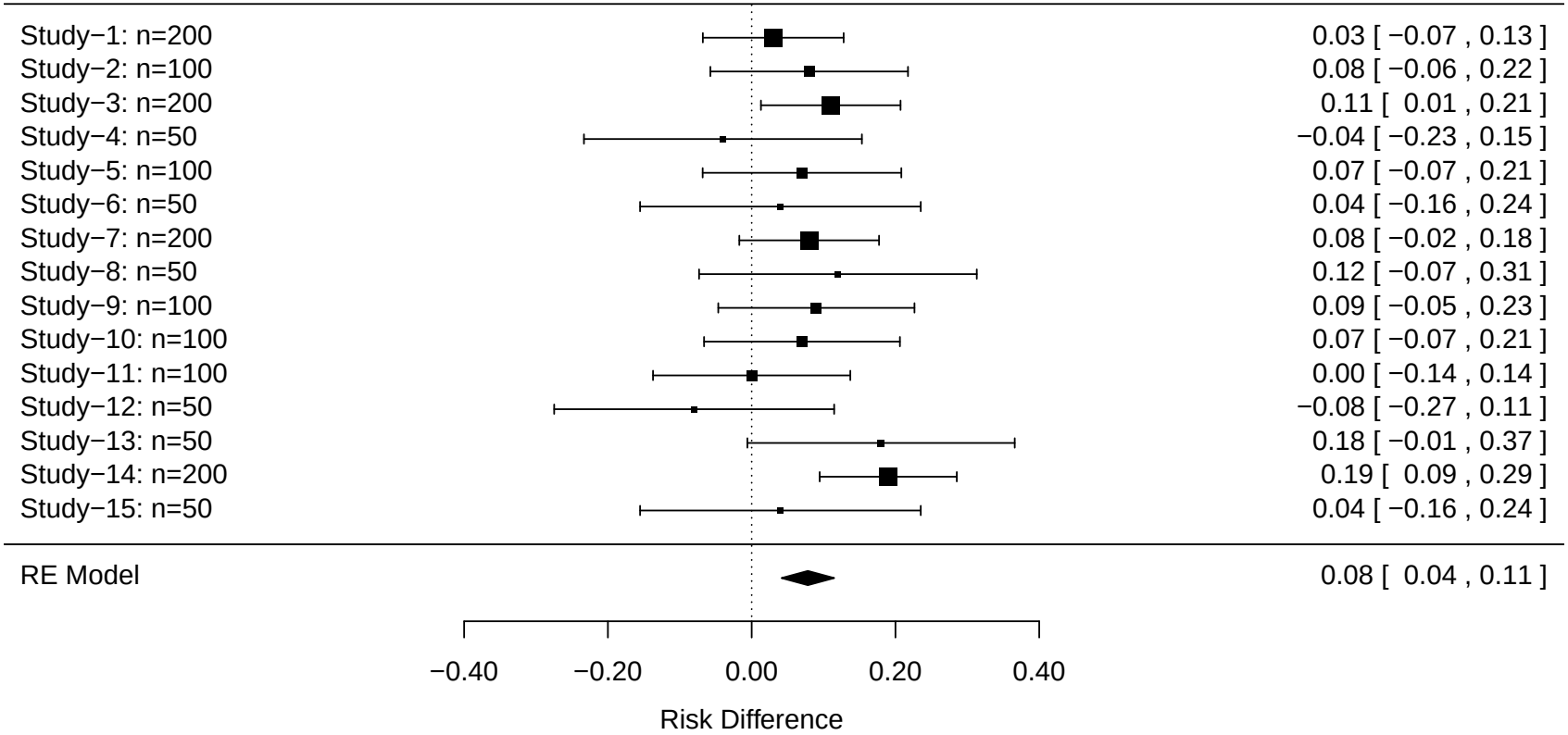
- We are sure that B is inferior to A

Scenario 7



- B is equivalent in effect to A

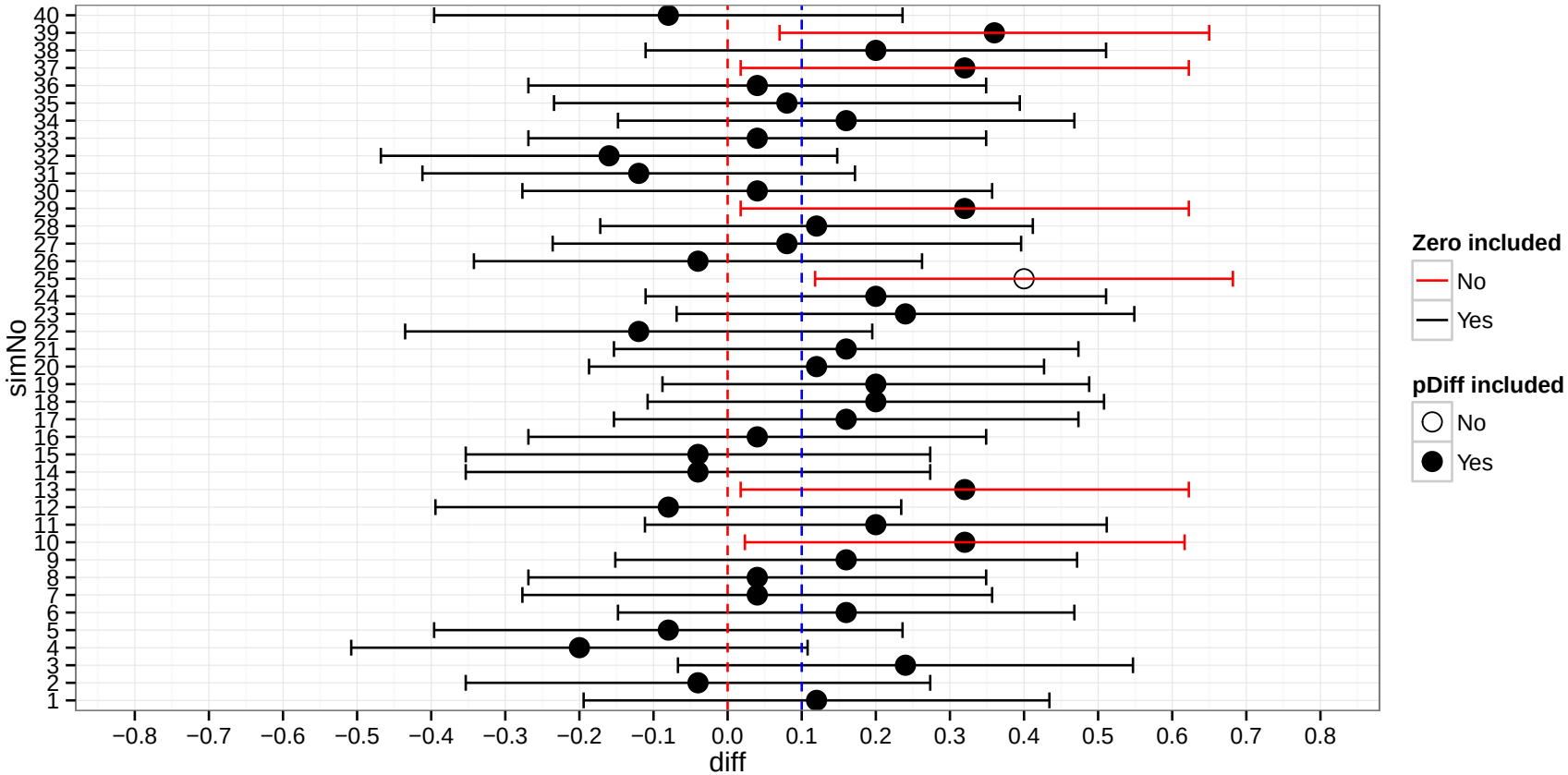
Finally, Importance of Replication of Experiments



- We are **more certain about the population effect size.** Miniscule confidence interval
- Interpretation of effect size depends on us.

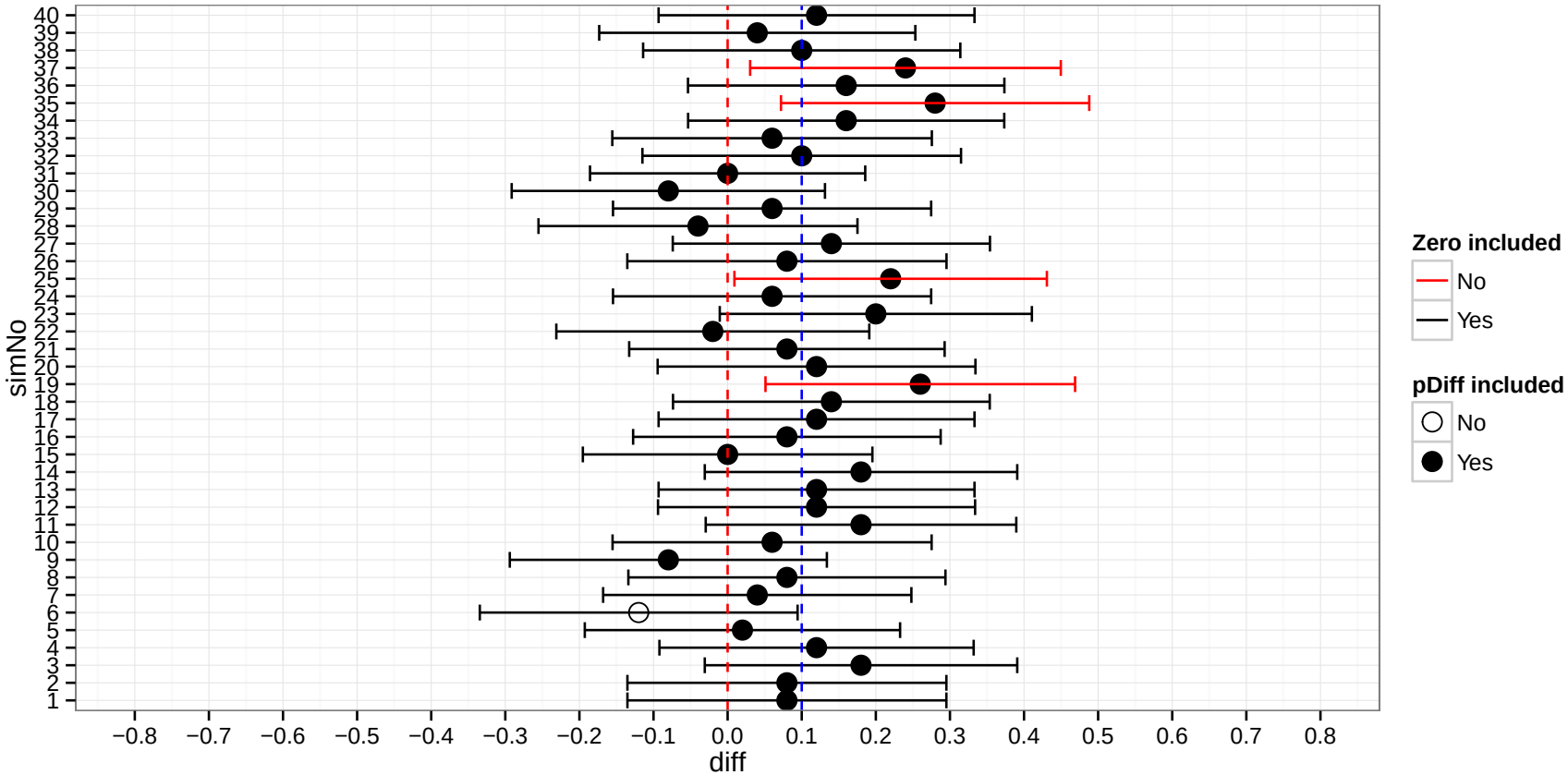
Manipulating CI

Simulation with CI: probA: 50%, probB: 60%, sample size: 25



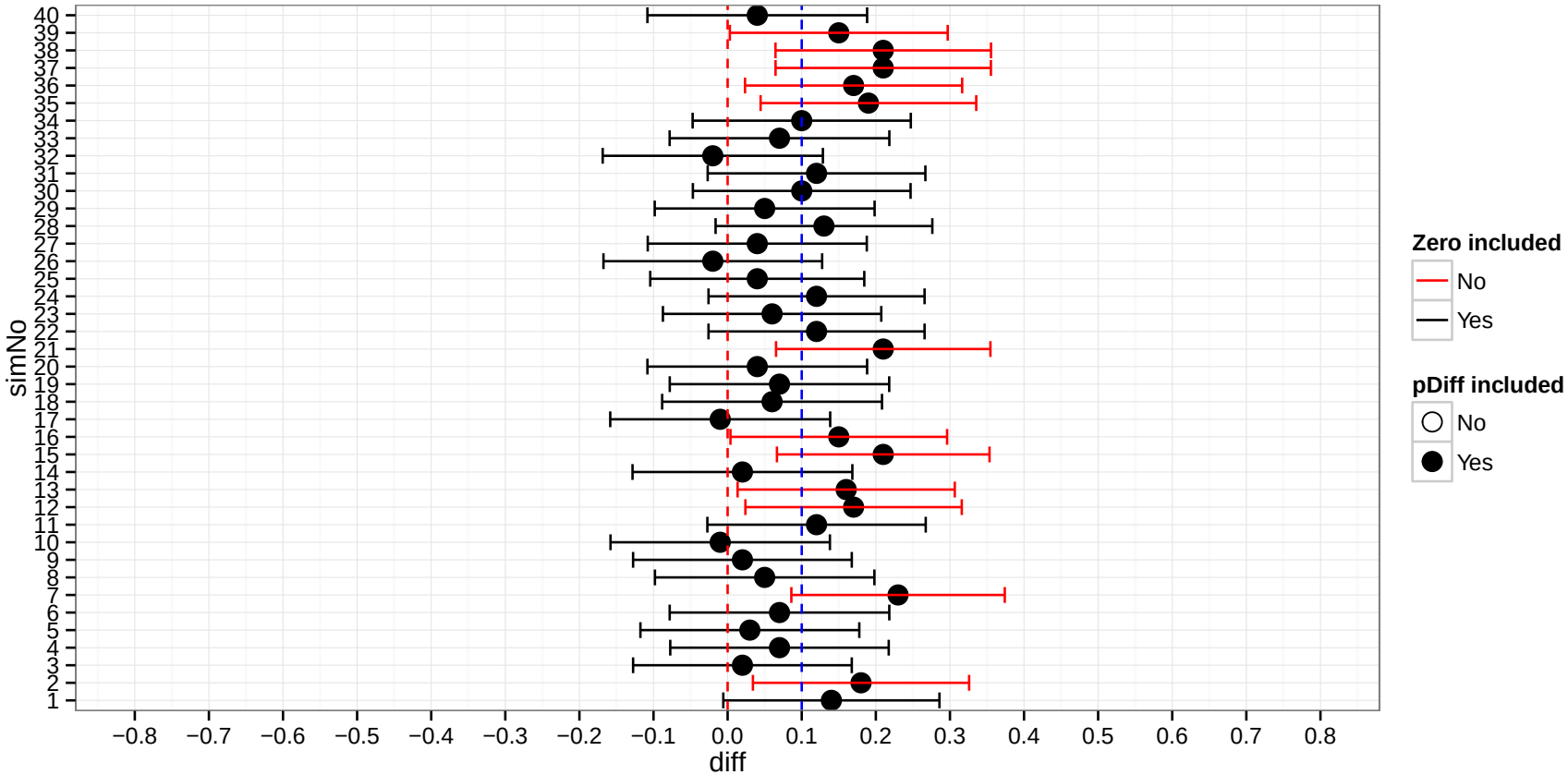
- Proportion of experiments failing to include population effect size in CI (Alpha Error): 0.023
- Proportion of experiments failing to show difference between both groups (Beta Error): 0.918

Simulation with CI: probA: 50%, probB: 60%, sample size: 50



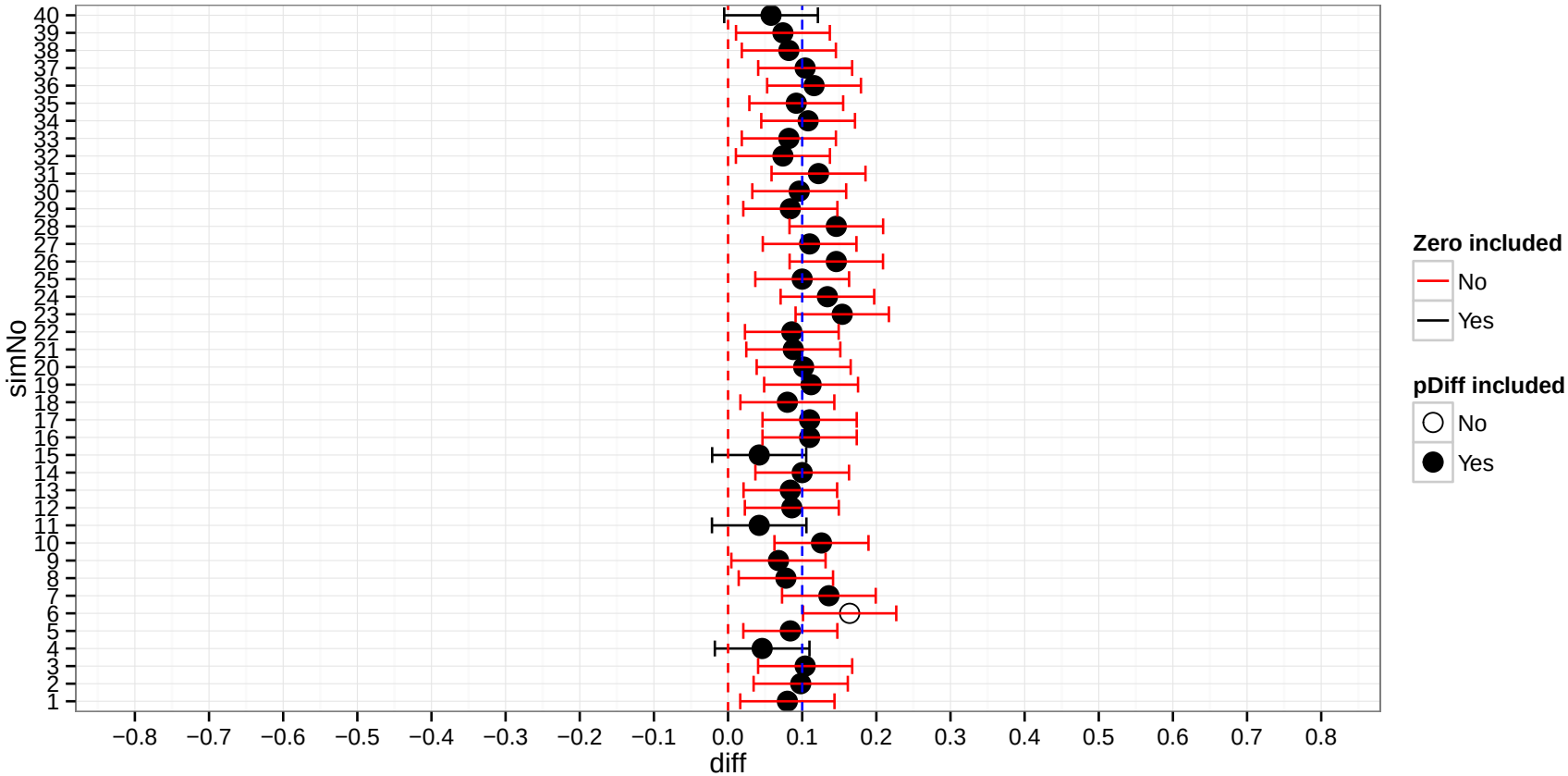
- Proportion of experiments failing to include population effect size in CI (Alpha Error): 0.034
- Proportion of experiments failing to show difference between both groups (Beta Error): 0.857

Simulation with CI: probA: 50%, probB: 60%, sample size: 100



- Proportion of experiments failing to include population effect size in CI (Alpha Error): 0.043
- Proportion of experiments failing to show difference between both groups (Beta Error): 0.713

Simulation with CI: probA: 50%, probB: 60%, sample size: 500

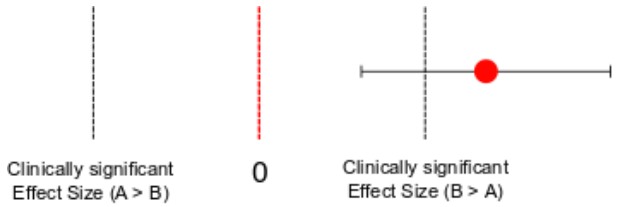


- Proportion of experiments failing to include population effect size in CI (Alpha Error): 0.034
- Proportion of experiments failing to show difference between both groups (Beta Error): 0.128

We learnt about . . .

- **Alpha error:** Proportion of times when CI **fail** to include the **population effect size**
  - Usual value: 0.05
- **Beta error:** Proportion of times when CI include **effect size of null hypothesis (0)**
  - Usual value: 0.20
- **Power** of study *(1 - Beta error)*: Proportion of times when CI **do not** include **effect size of null hypothesis (0)**

Steps to increase power of study



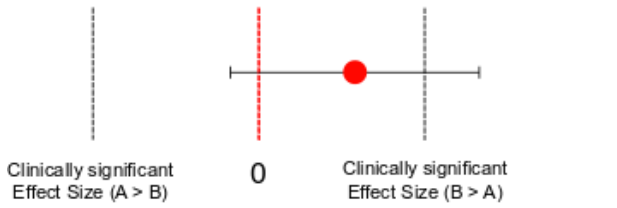
INCREASE PROPORTION

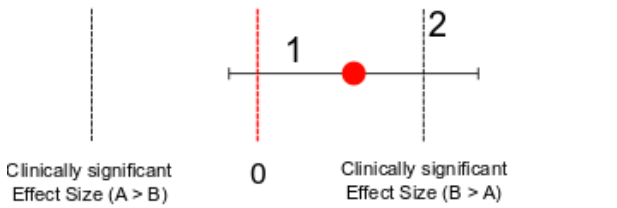
↓

INCREASES POWER

↑

DECREASE PROPORTION





1: Decrease width by
- Increasing sample size
- Decreasing pop variability
- Decreasing confidence levels (say to 90%)

2: Increase gap between 0 and clinically relevant effect size

To summarise, interpretation of study results means

- Similarity of population depicted in study with ours
- Understanding relevant effect size
- Ascertaining equality of groups A and B (Tackling Bias)
- Understanding position and precision of effect size estimate

THANK YOU

Acknowledgements

1. R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`
2. ggplot2, metafor packages.

Presentation available at `https://github.com/sumprain/blog/tree/master/aiims_presentation`