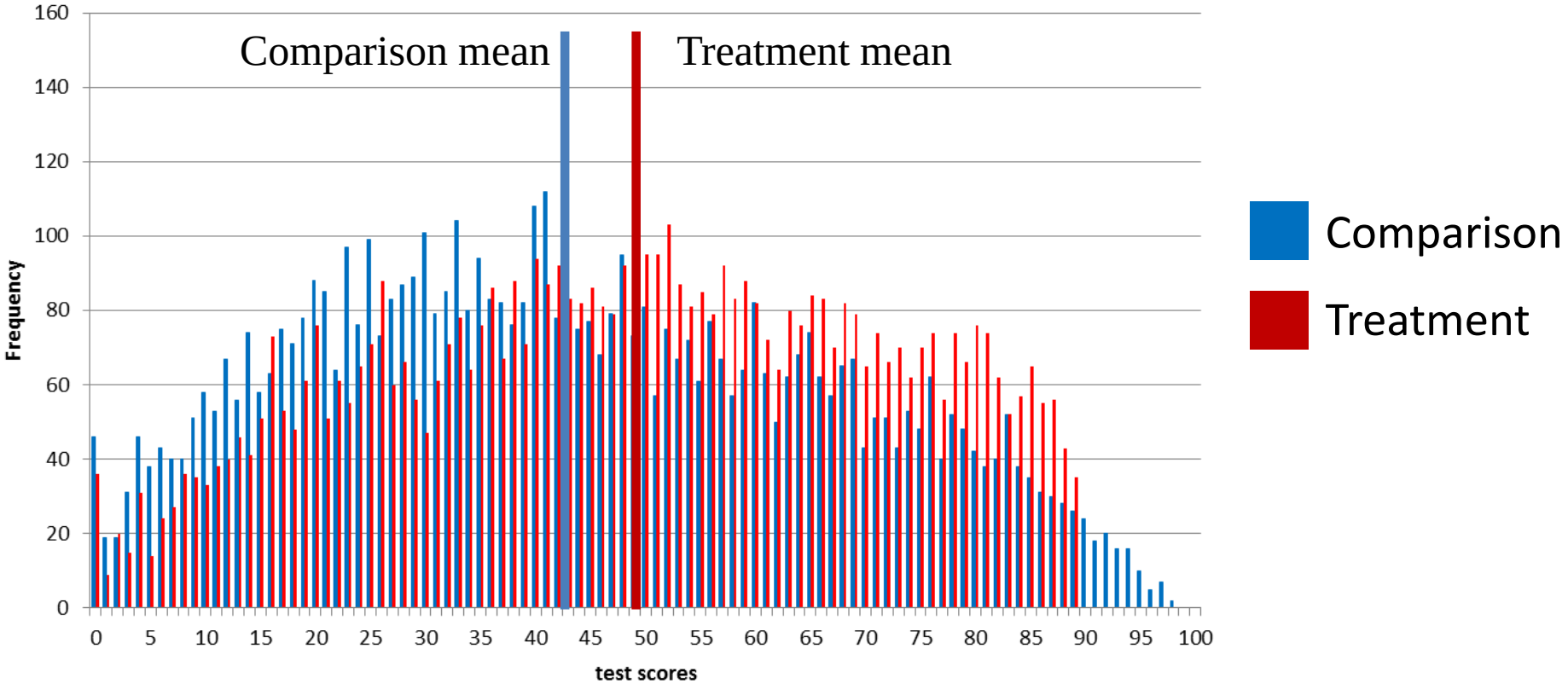
Calculating Power for RCTs

(Based on Glennerster and Takavarasha, 2013, Chapter 6)

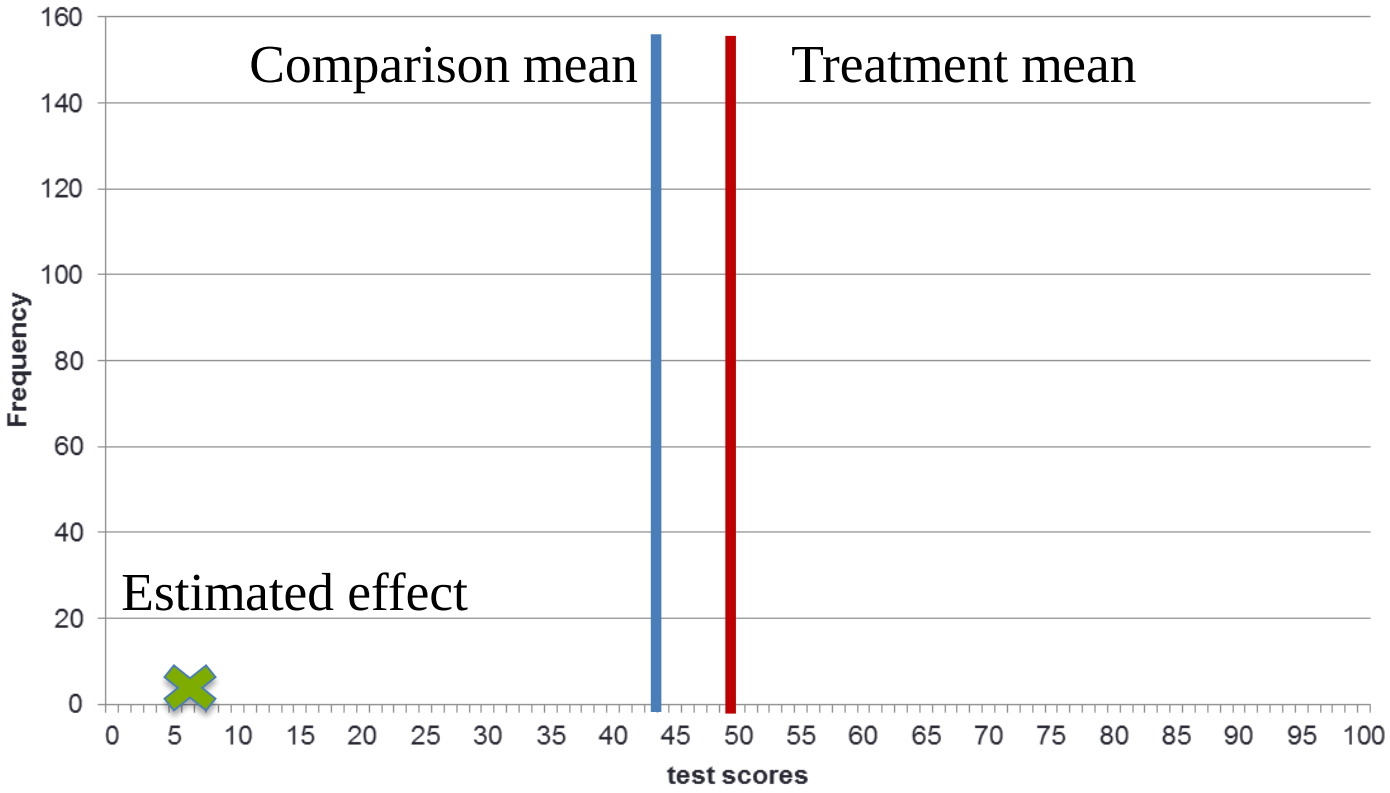
Outline

- Define power
- Determinants of power
- Power in clustered RCTs
- Calculating power in practice

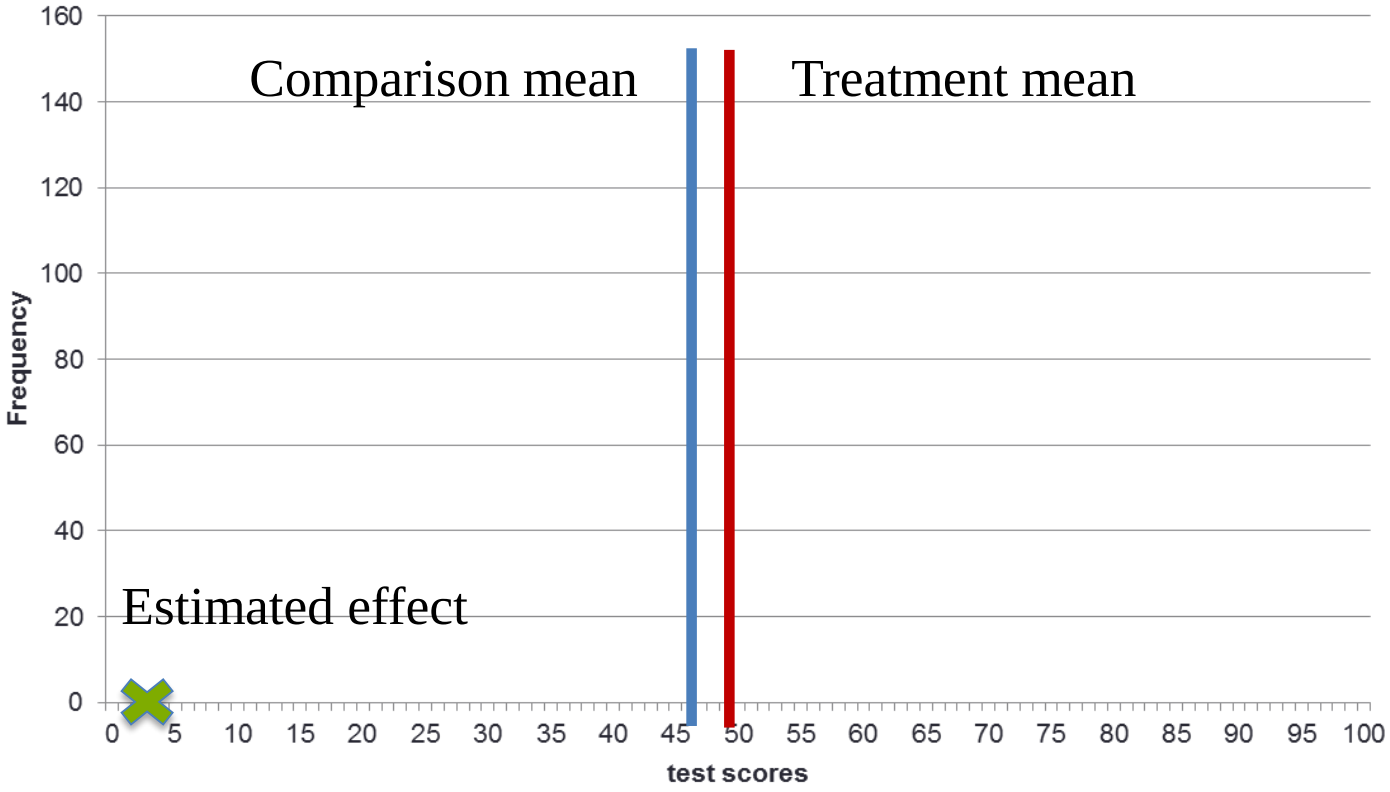
One experiment, 2 samples, 2 means



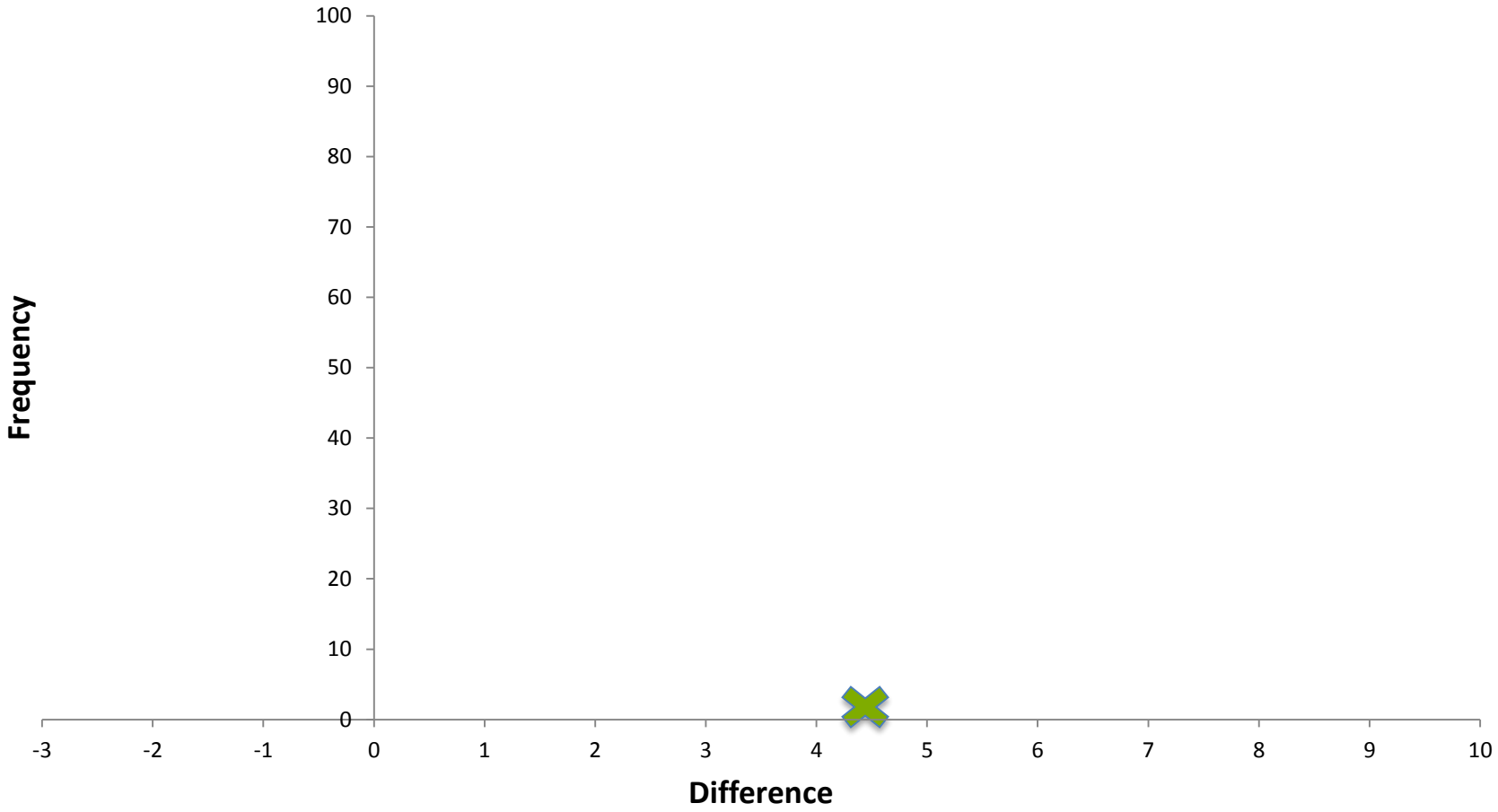
Difference between the sample means



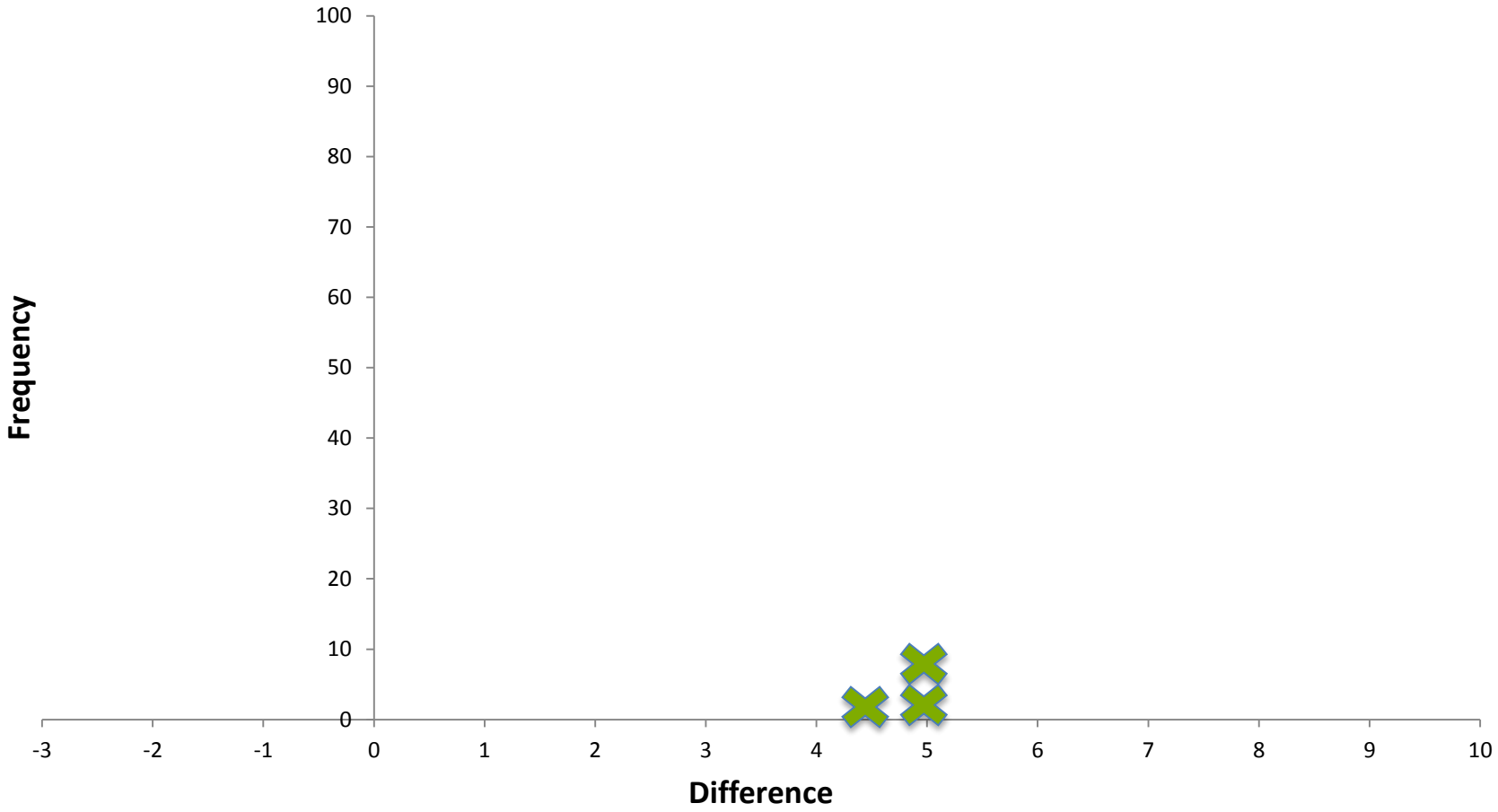
What if we ran a second experiment?



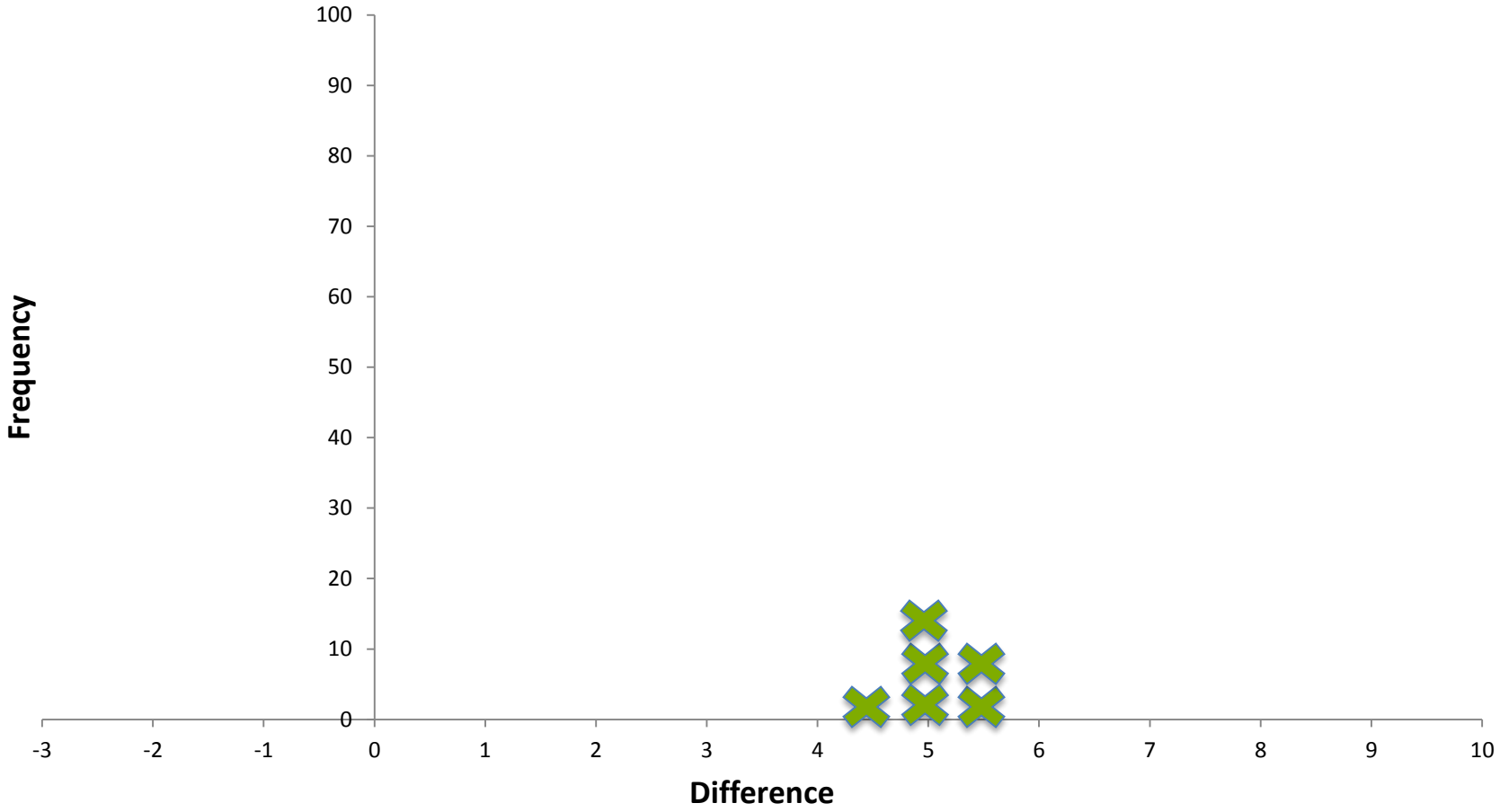
Many experiments give distribution of estimates



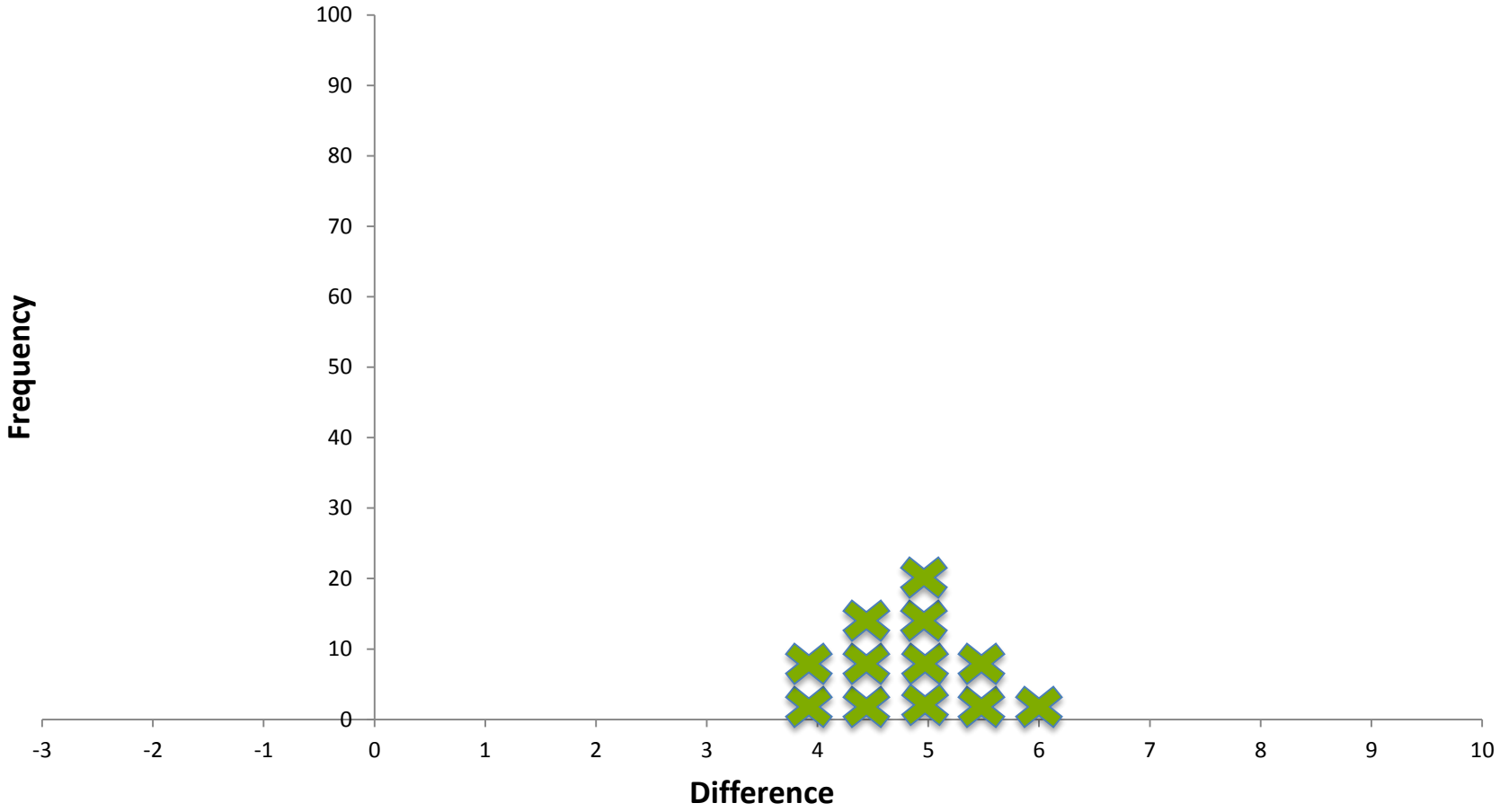
Many experiments give distribution of estimates



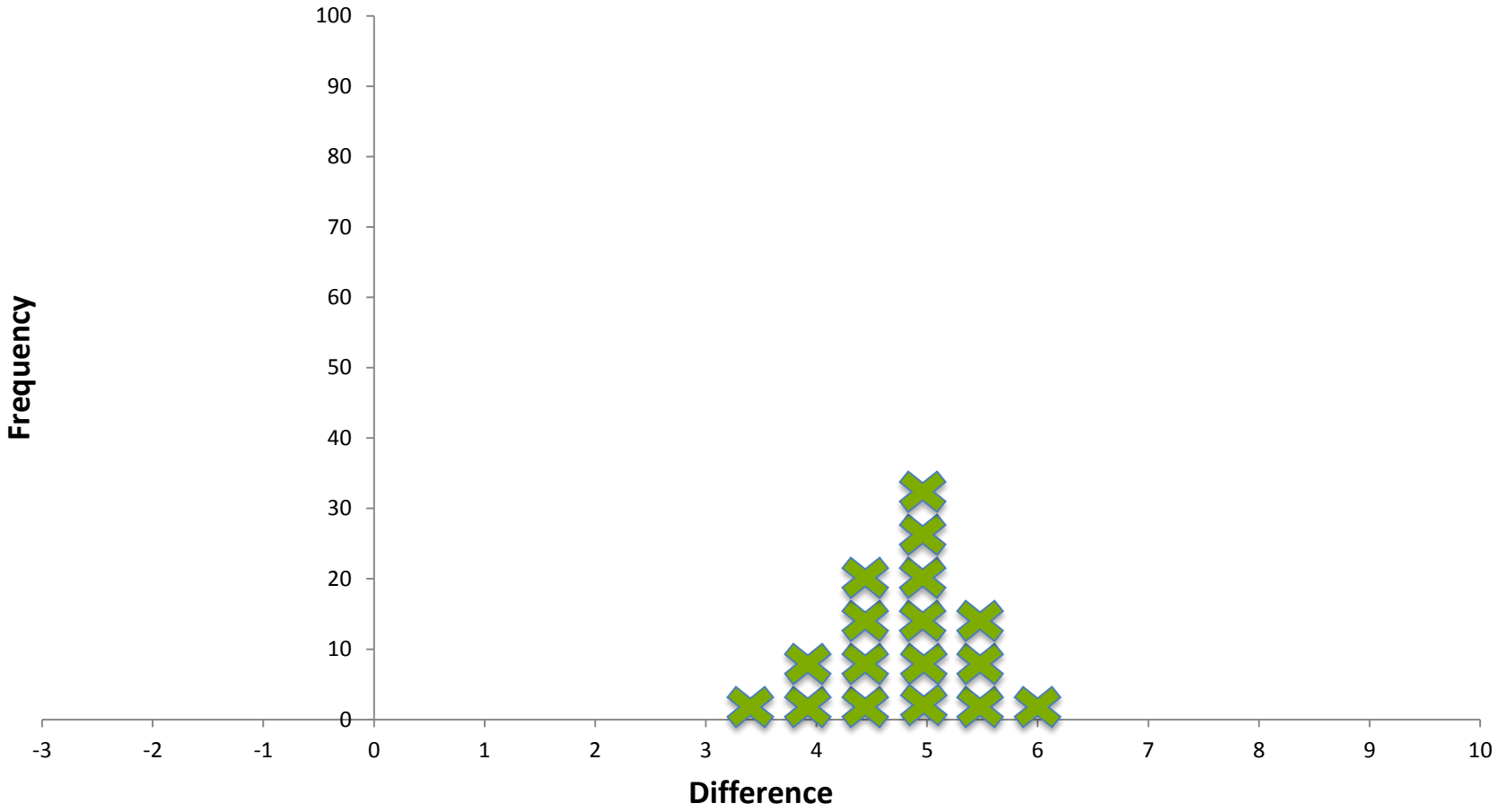
Many experiments give distribution of estimates



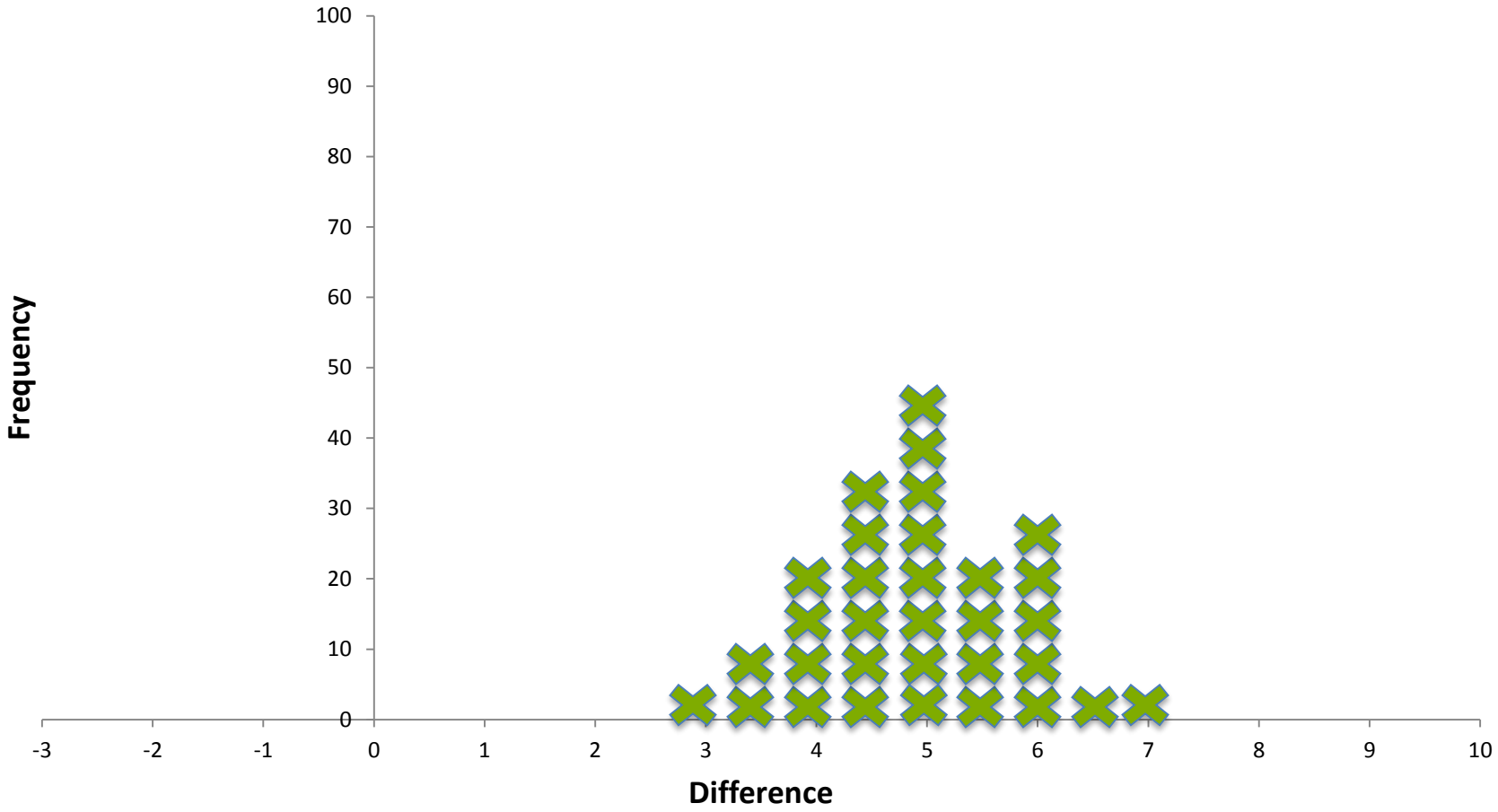
Many experiments give distribution of estimates



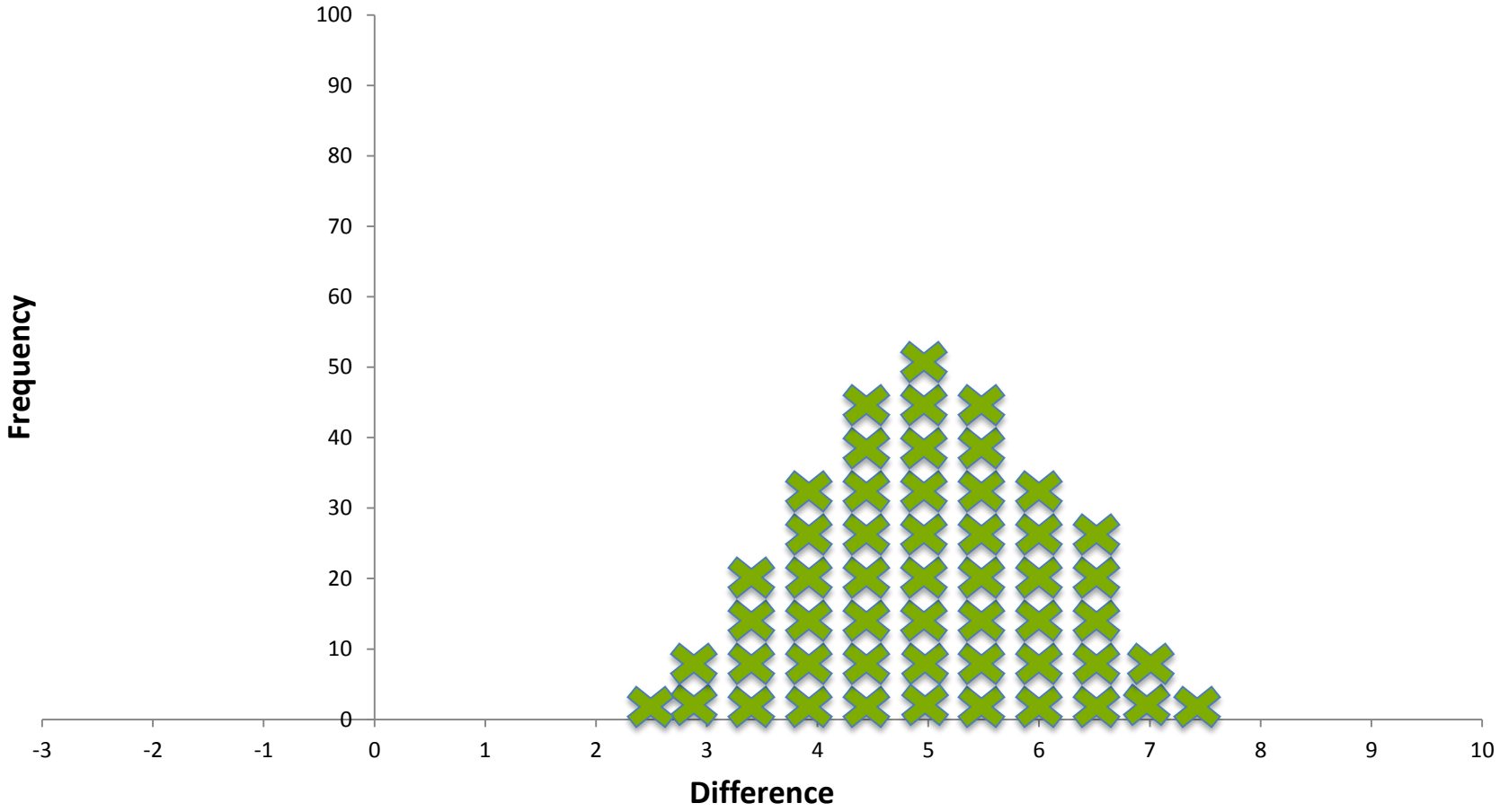
Many experiments give distribution of estimates



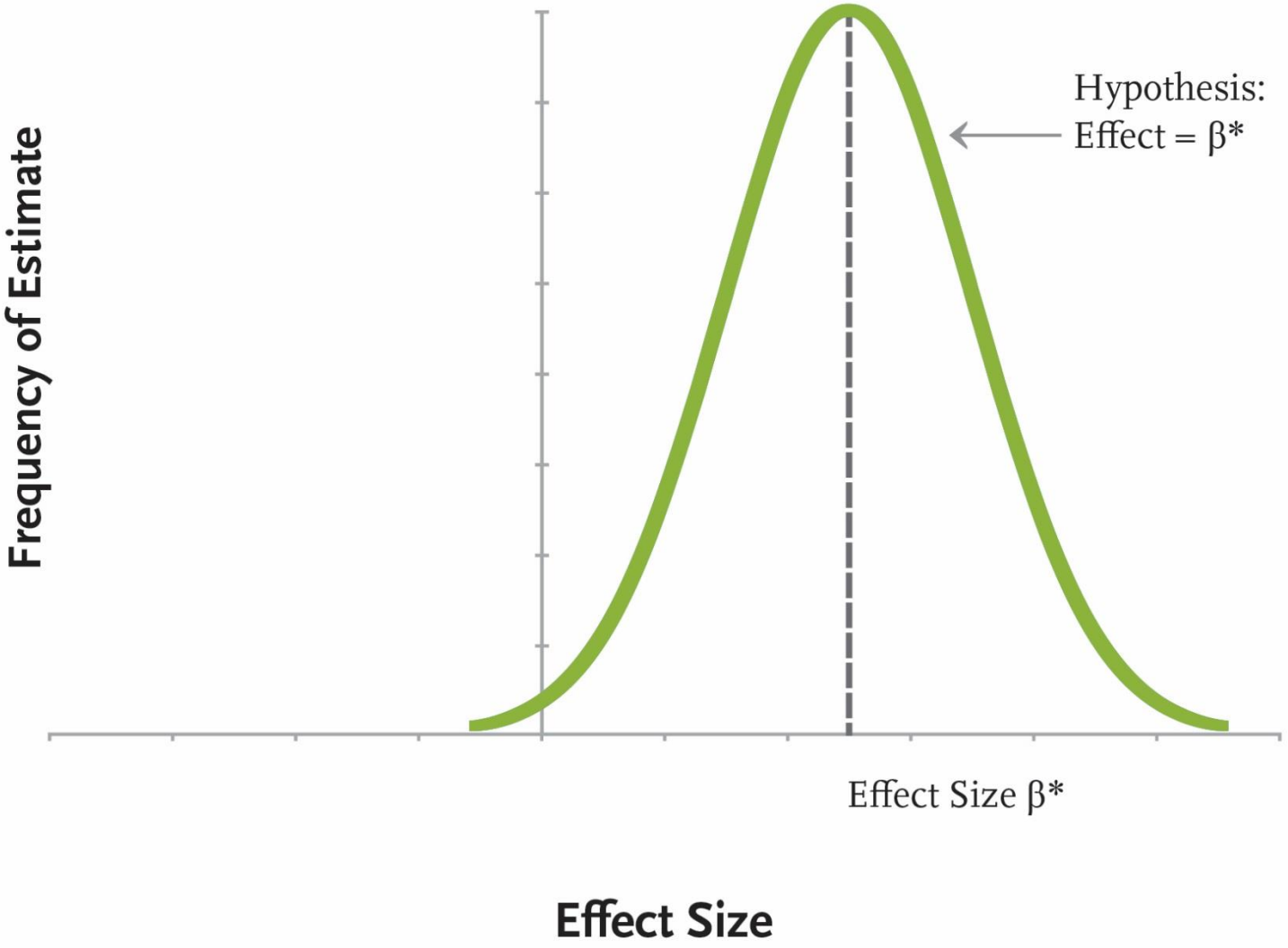
Many experiments give distribution of estimates



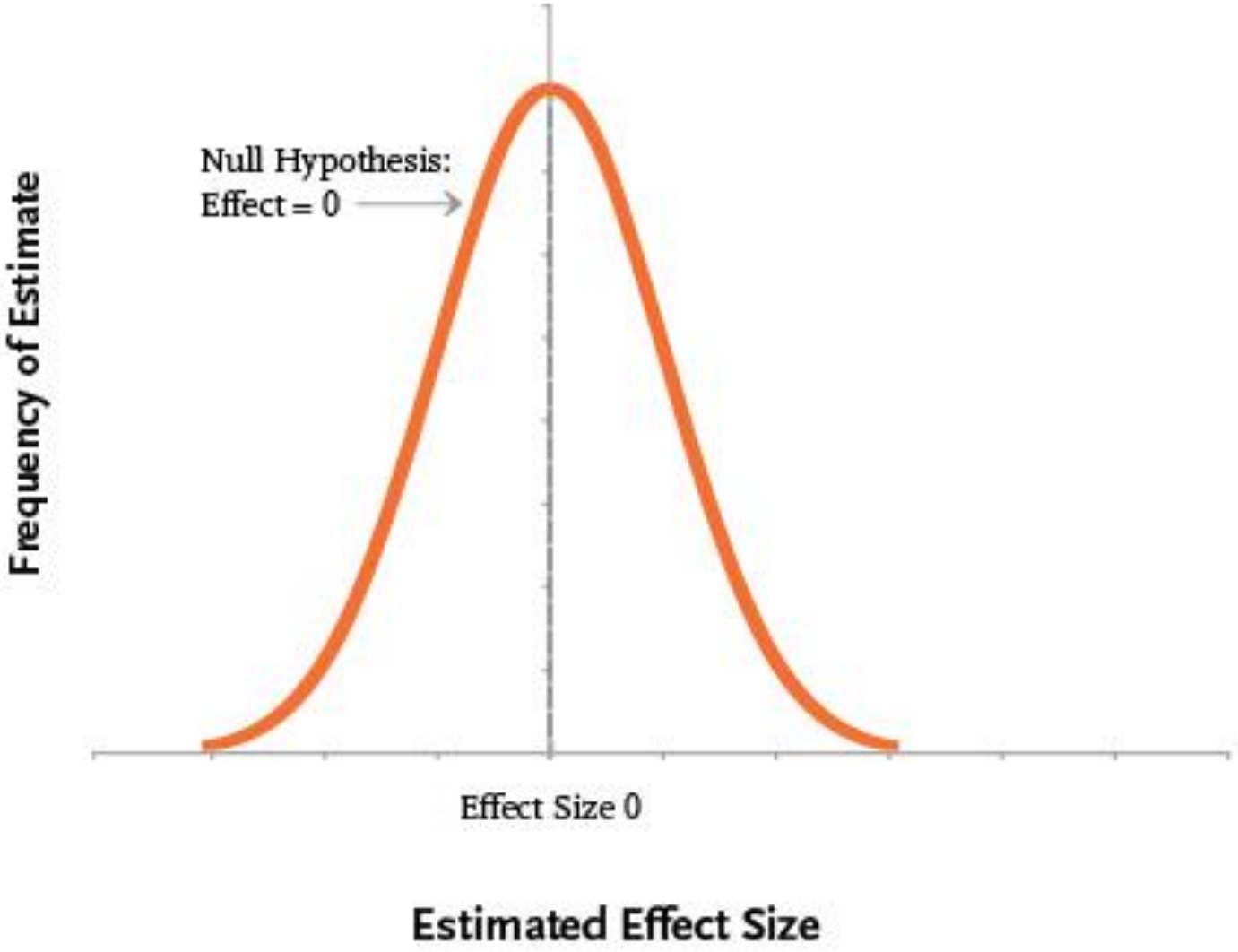
Many experiments give distribution of estimates



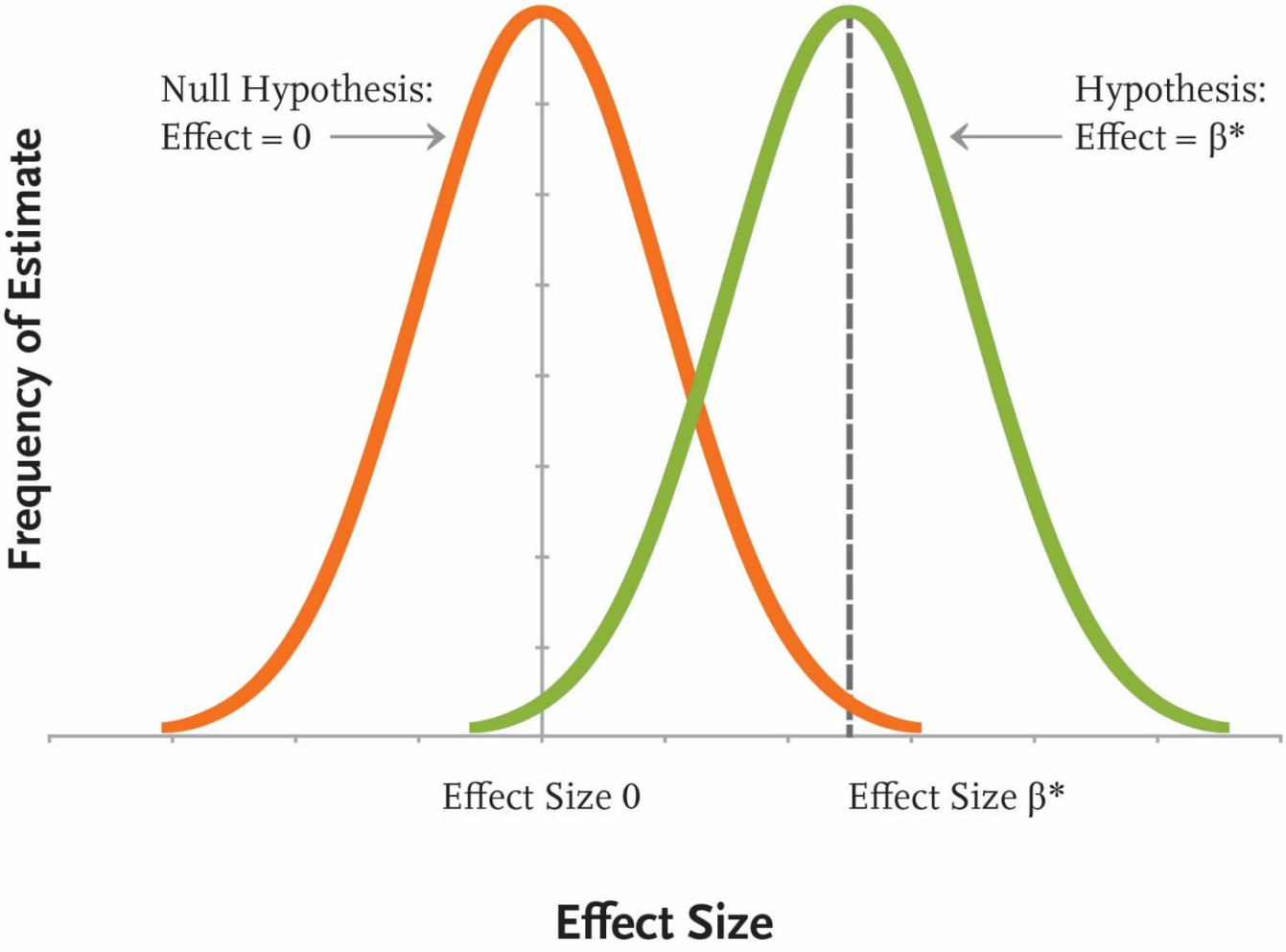
Normal distribution of estimates around true effect



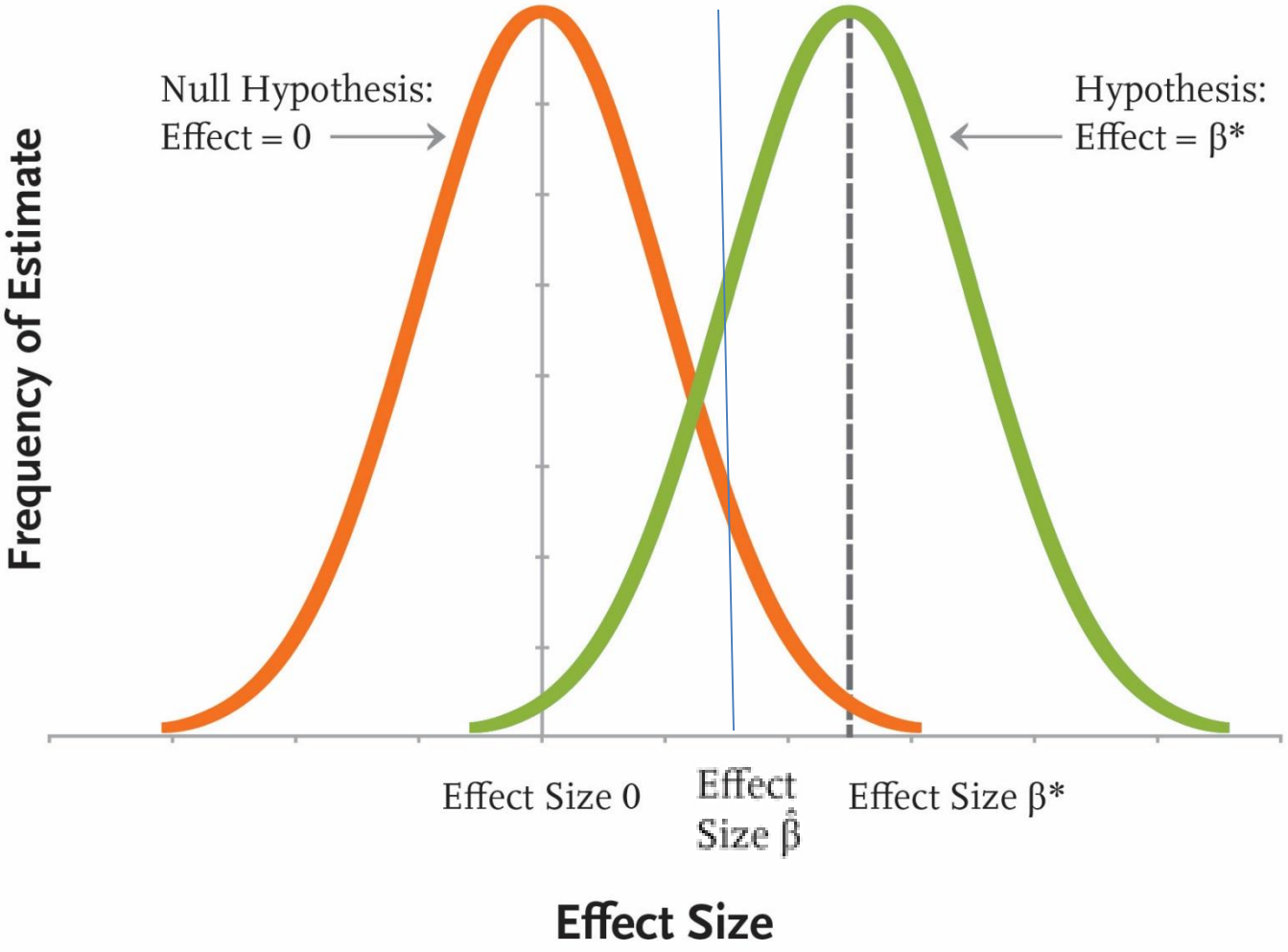
Normal distribution if true effect is zero



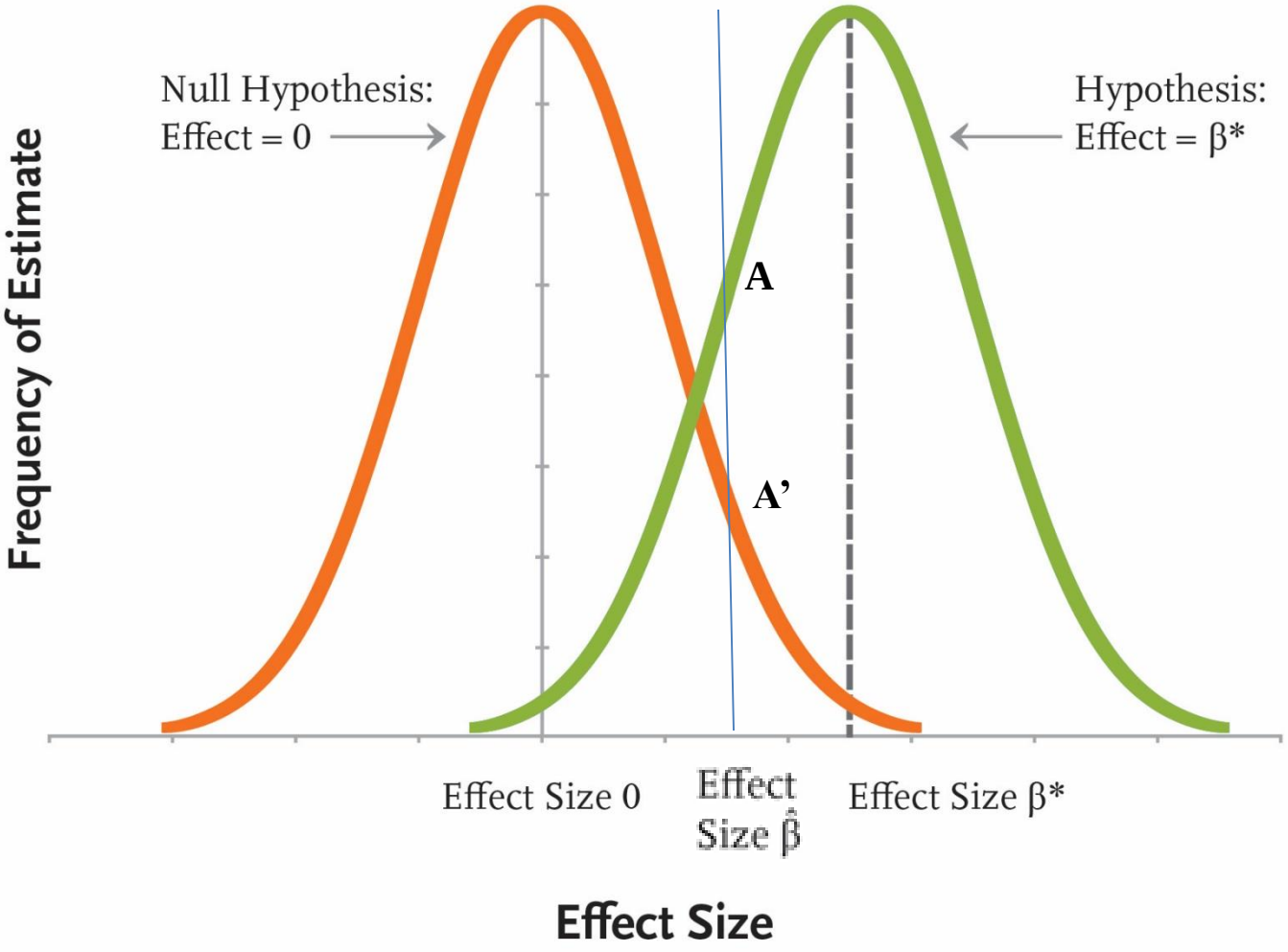
Distributions under two alternatives



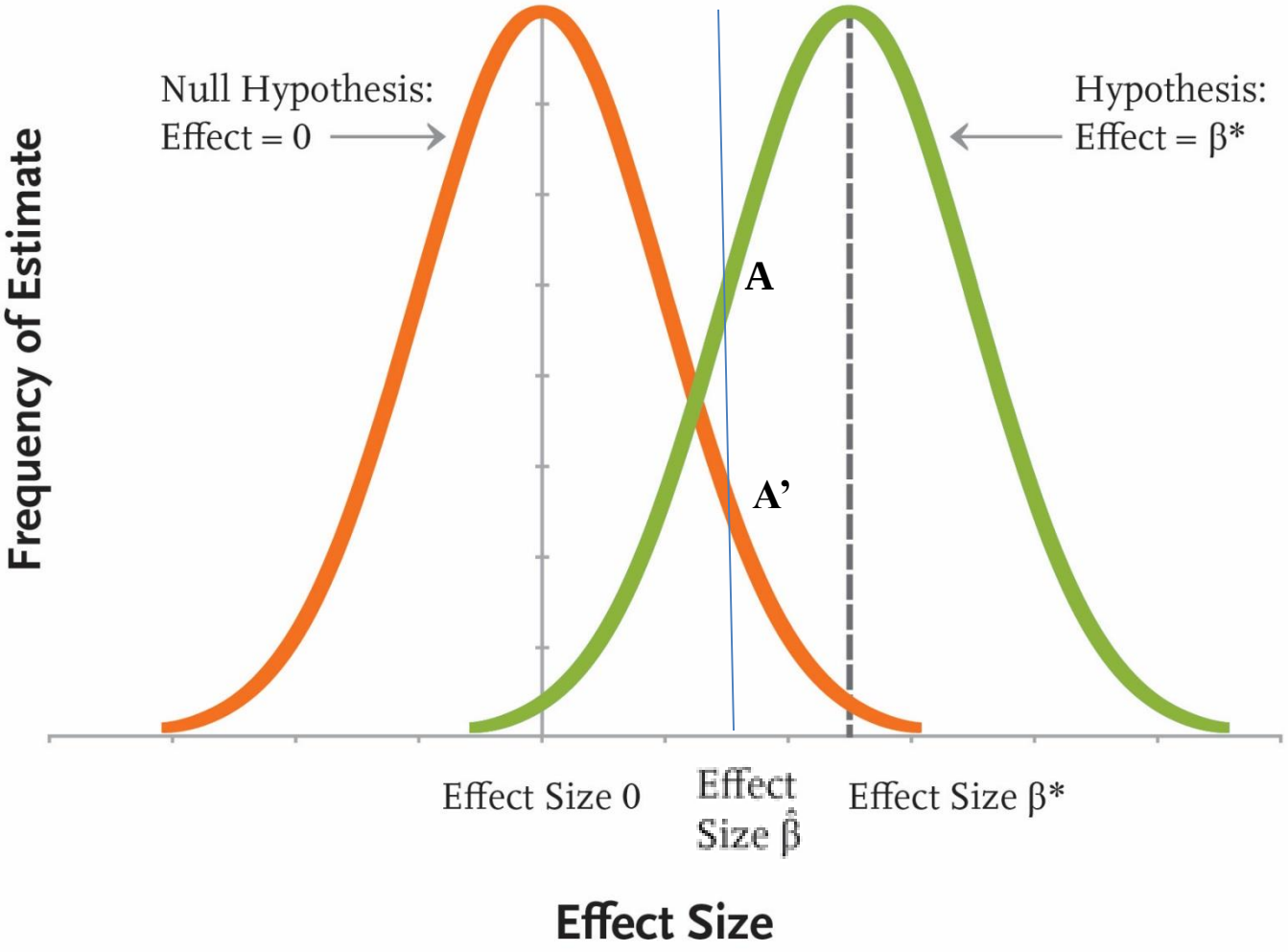
We don’t see these distributions, just our estimate $\hat{\beta}$



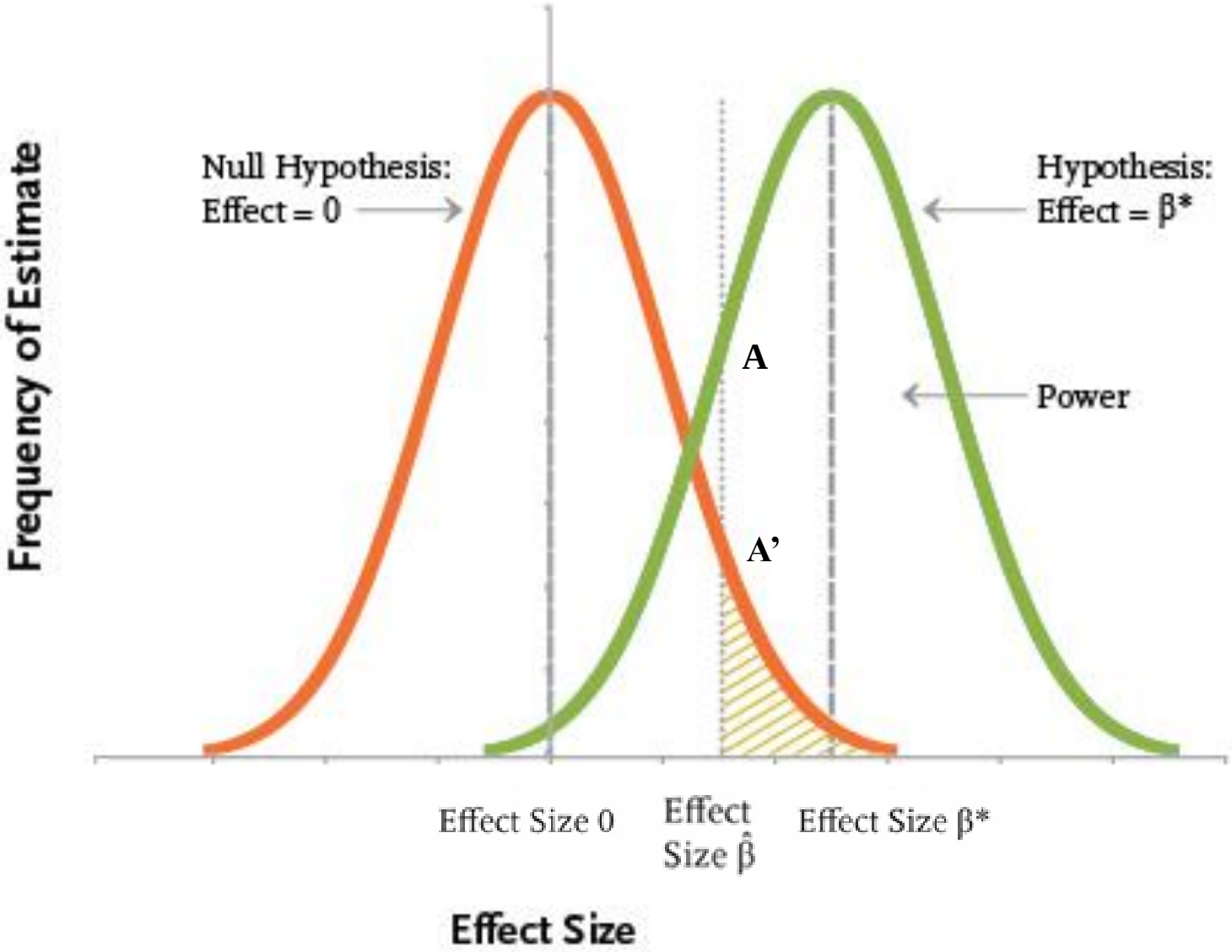
Is our estimate $\hat{\beta}$ consistent with the true effect being β*? With β=0? Which is more likely?



If we observe $\hat{\beta}$ can we rule out true effect = 0?



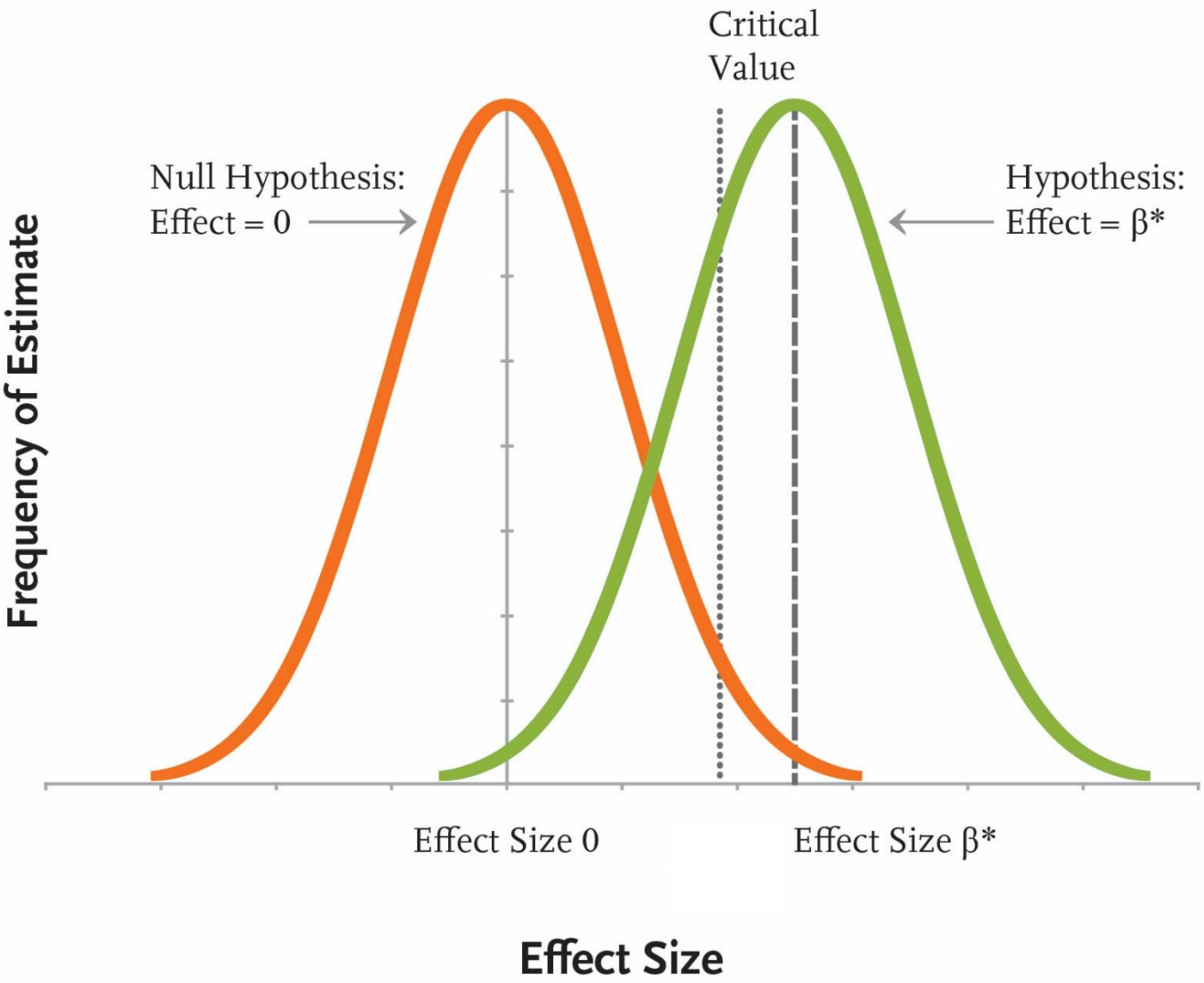
Probability true effect=0 is area to the right of A' for H0, over total area under H0



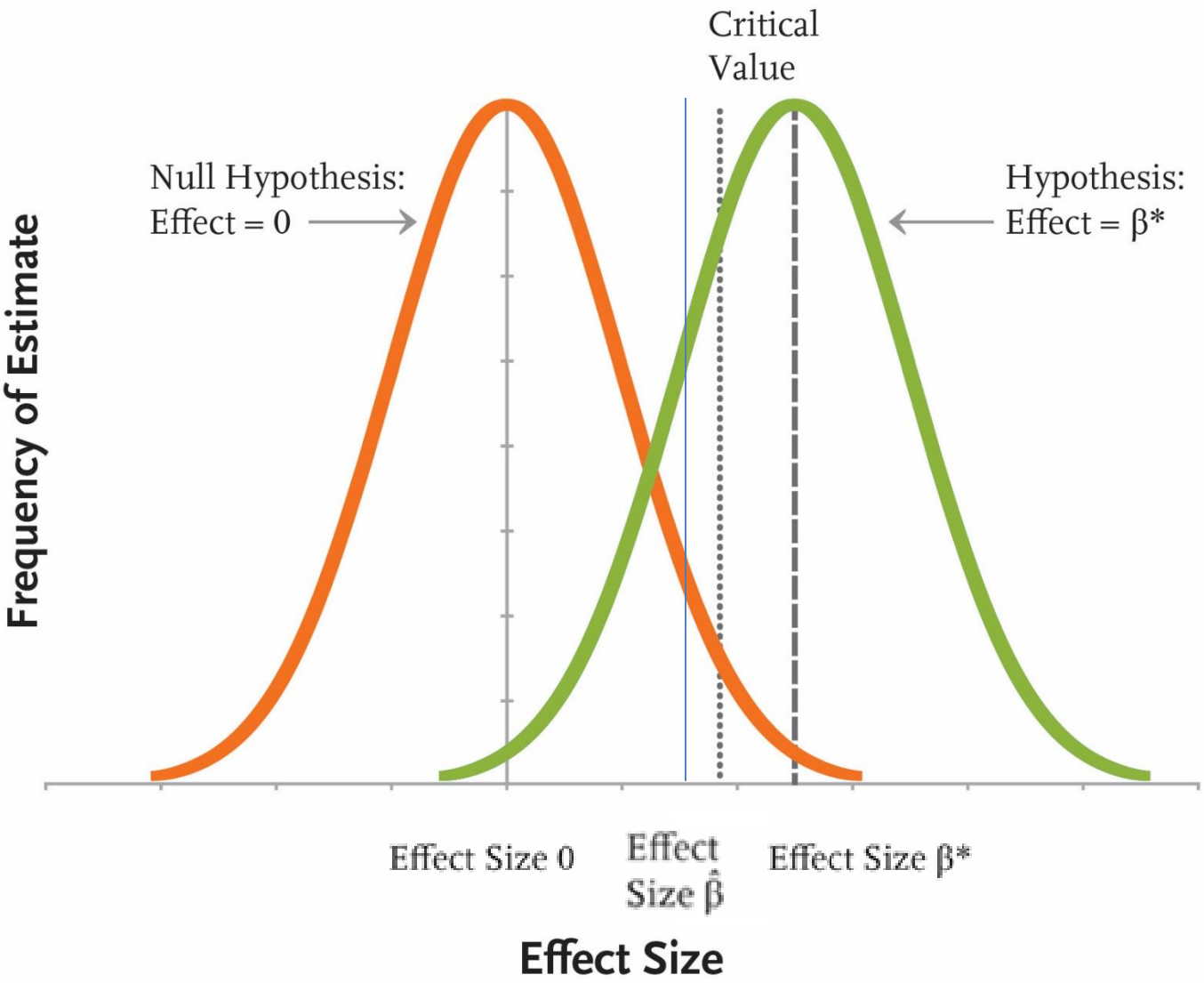
Critical value

- Definition: the critical value is the value of the estimated effect which exactly corresponds to the significance level
- If testing whether bigger than 0, significant at 95% level, it is the level of the estimate where exactly 95% of area under the curve lies to the left

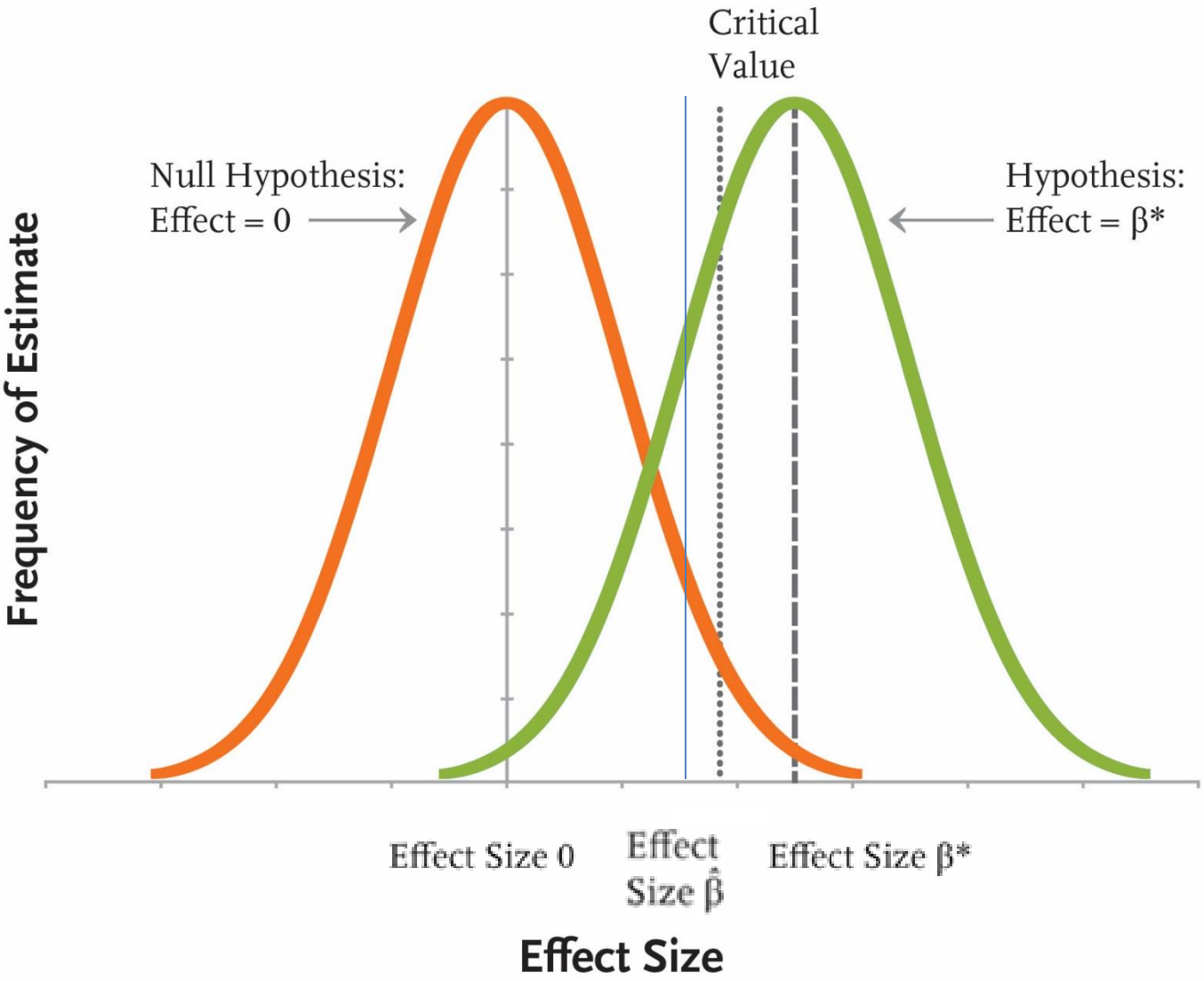
95% critical value for true effect>0



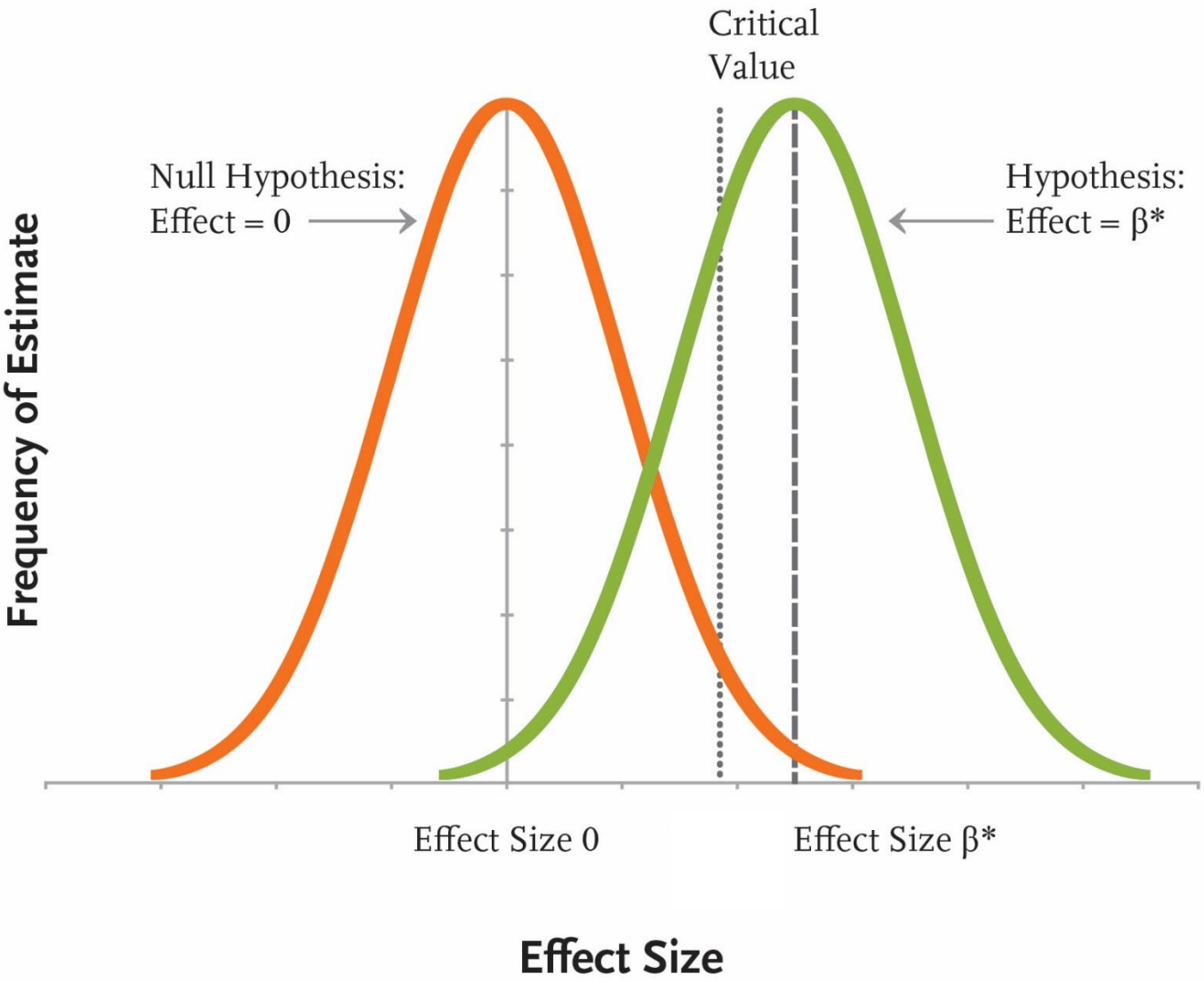
In this case $\hat{\beta}$ is < critical value so….



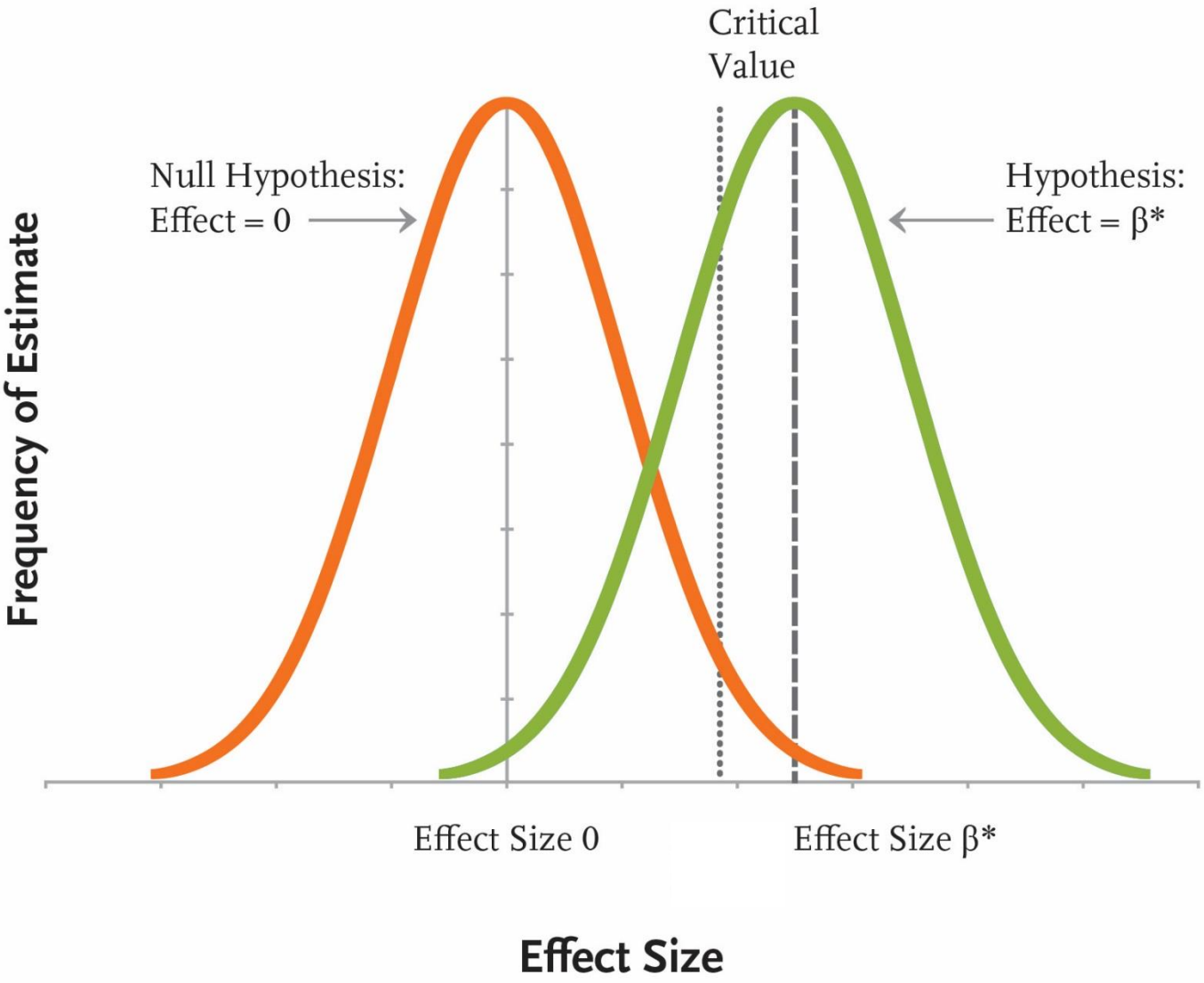
…..we cannot reject that true effect=0 with 95% confidence



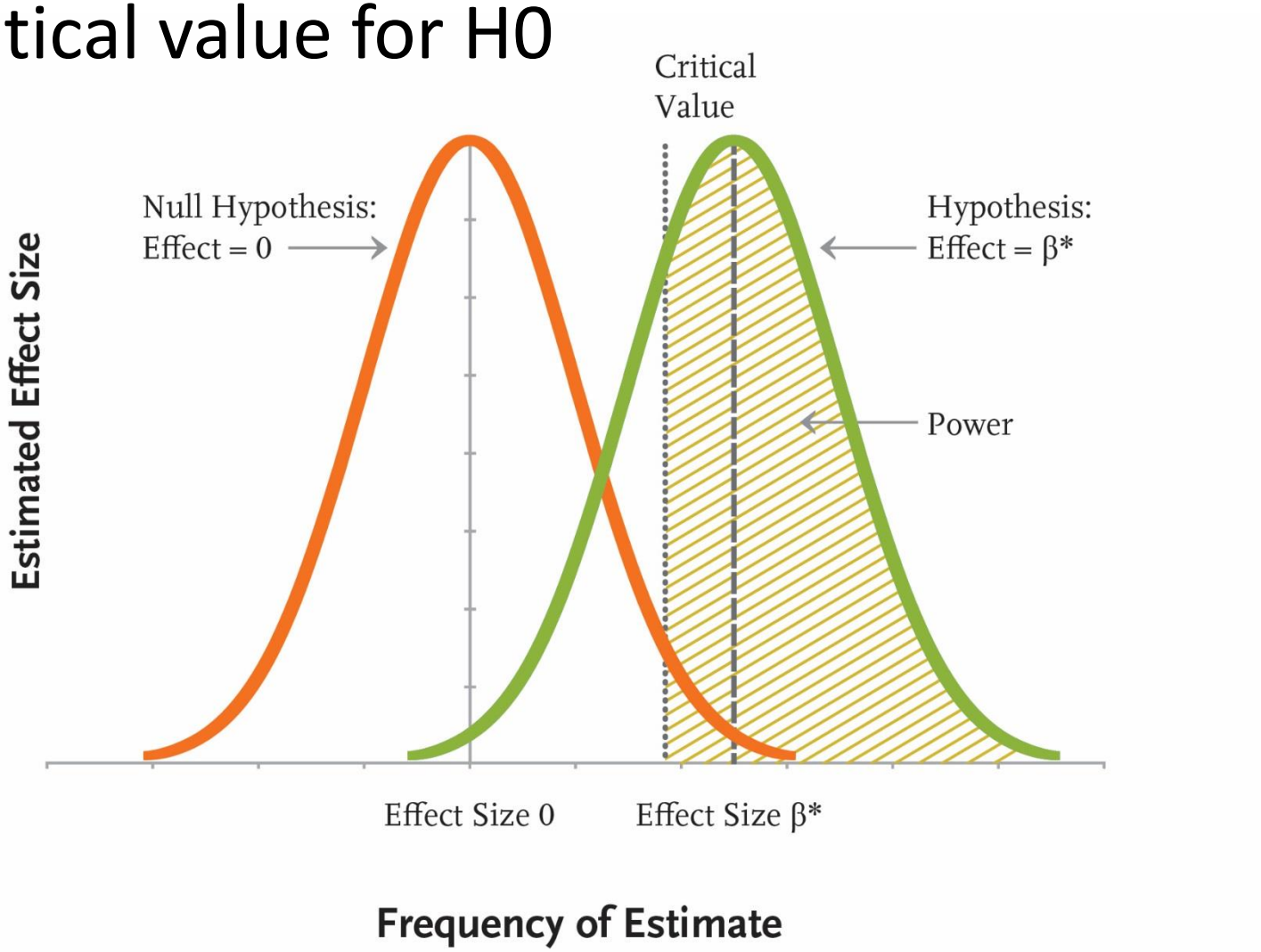
What if the true effect=β*?



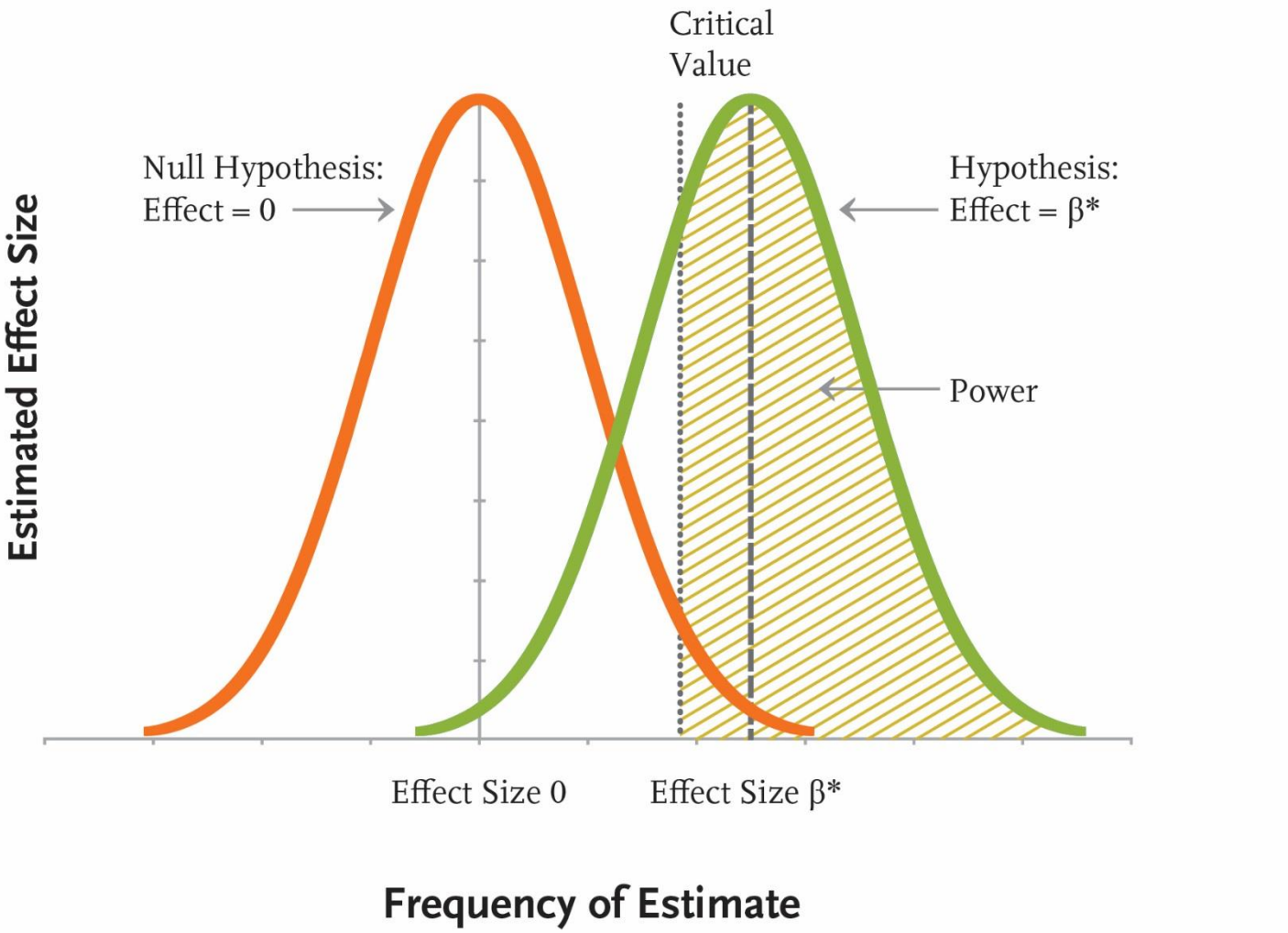
How often would we get estimates that we *could* distinguish from 0? (if true effect=β*)



Chance of getting estimates we can distinguish from 0 is the area under H β* that is above critical value for H0



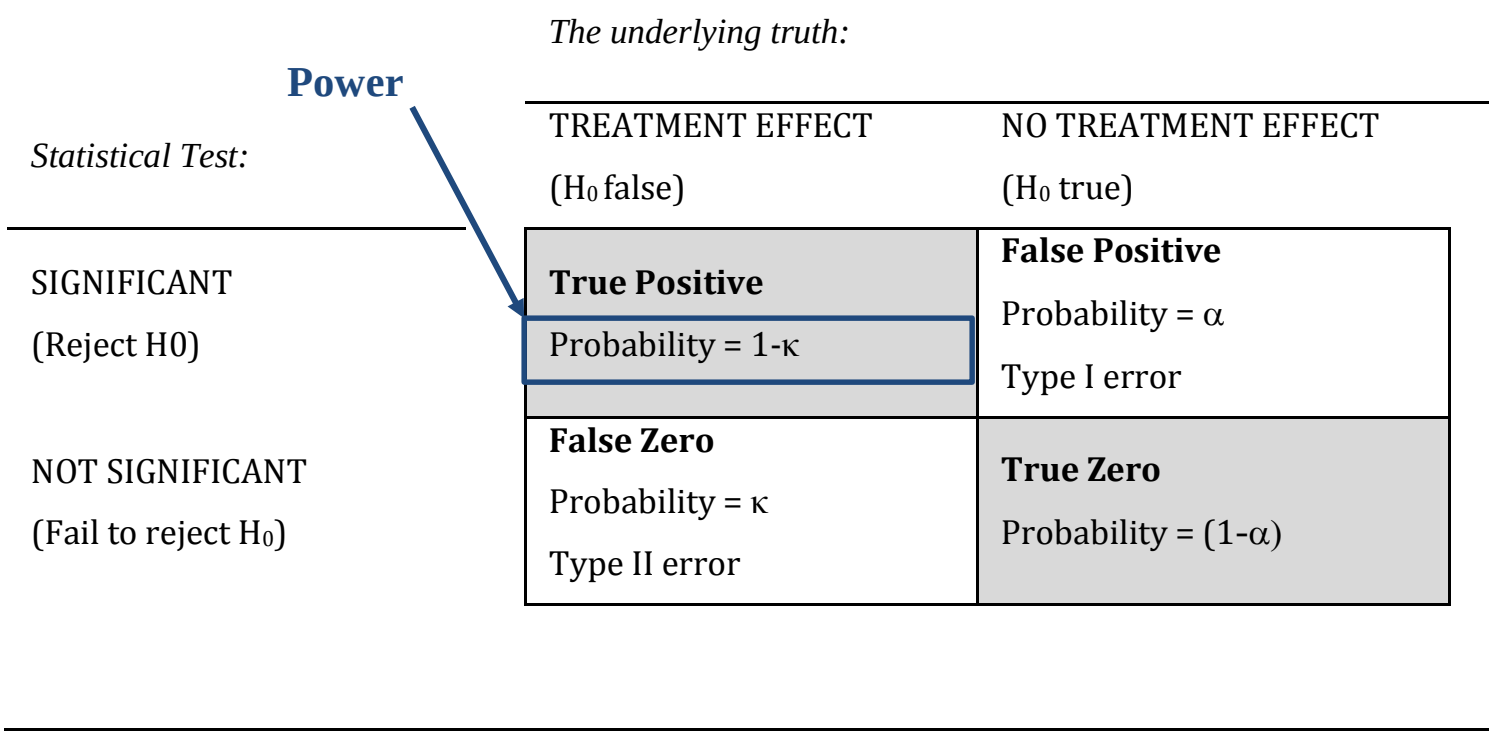
This is power



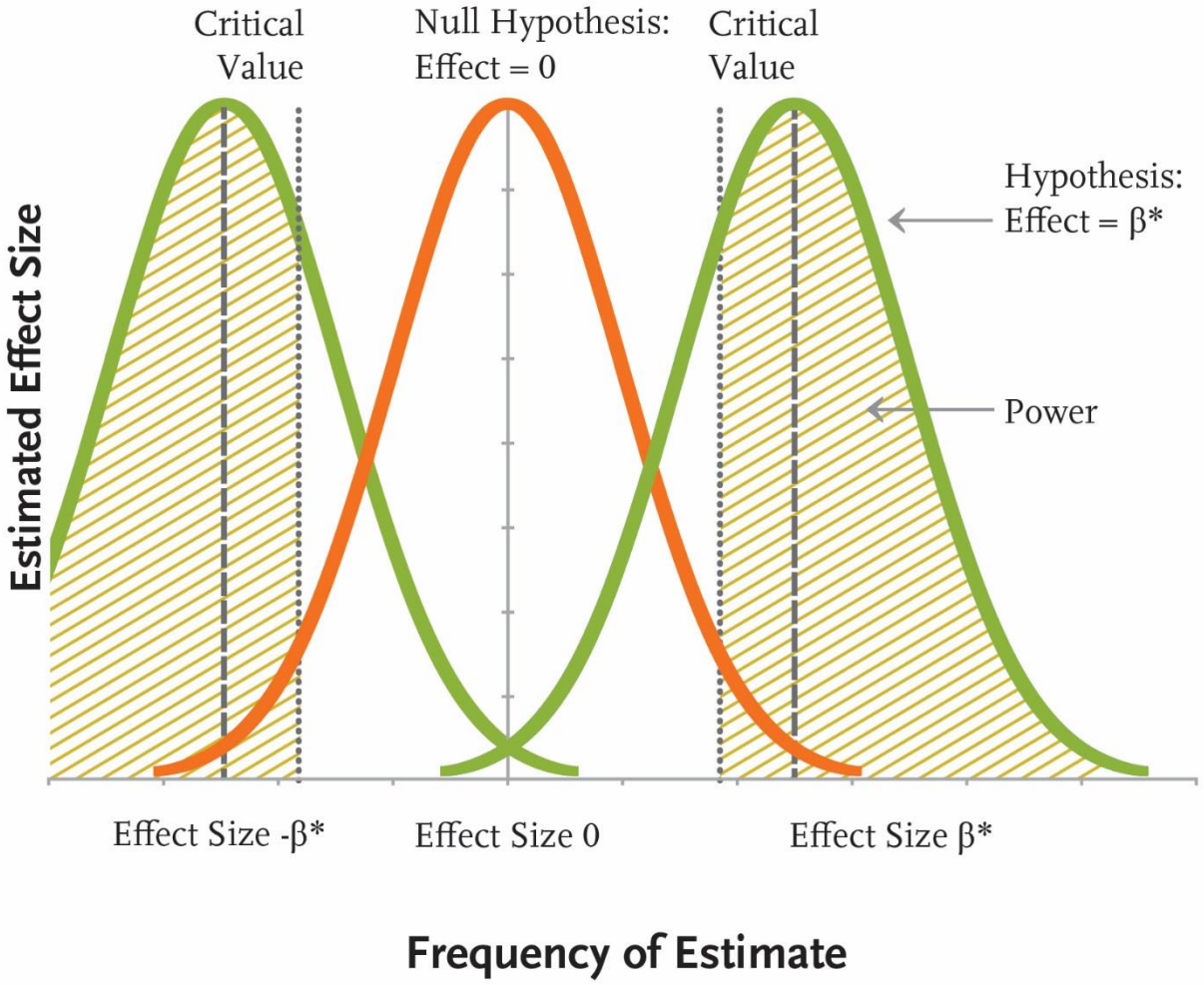
Definition of Power

- Statistical power is the probability that, if the true effect is of a given size, our proposed experiment will be able to distinguish the *estimated* effect from zero
- Traditionally, we aim for 80% power. Some people aim for 90% power

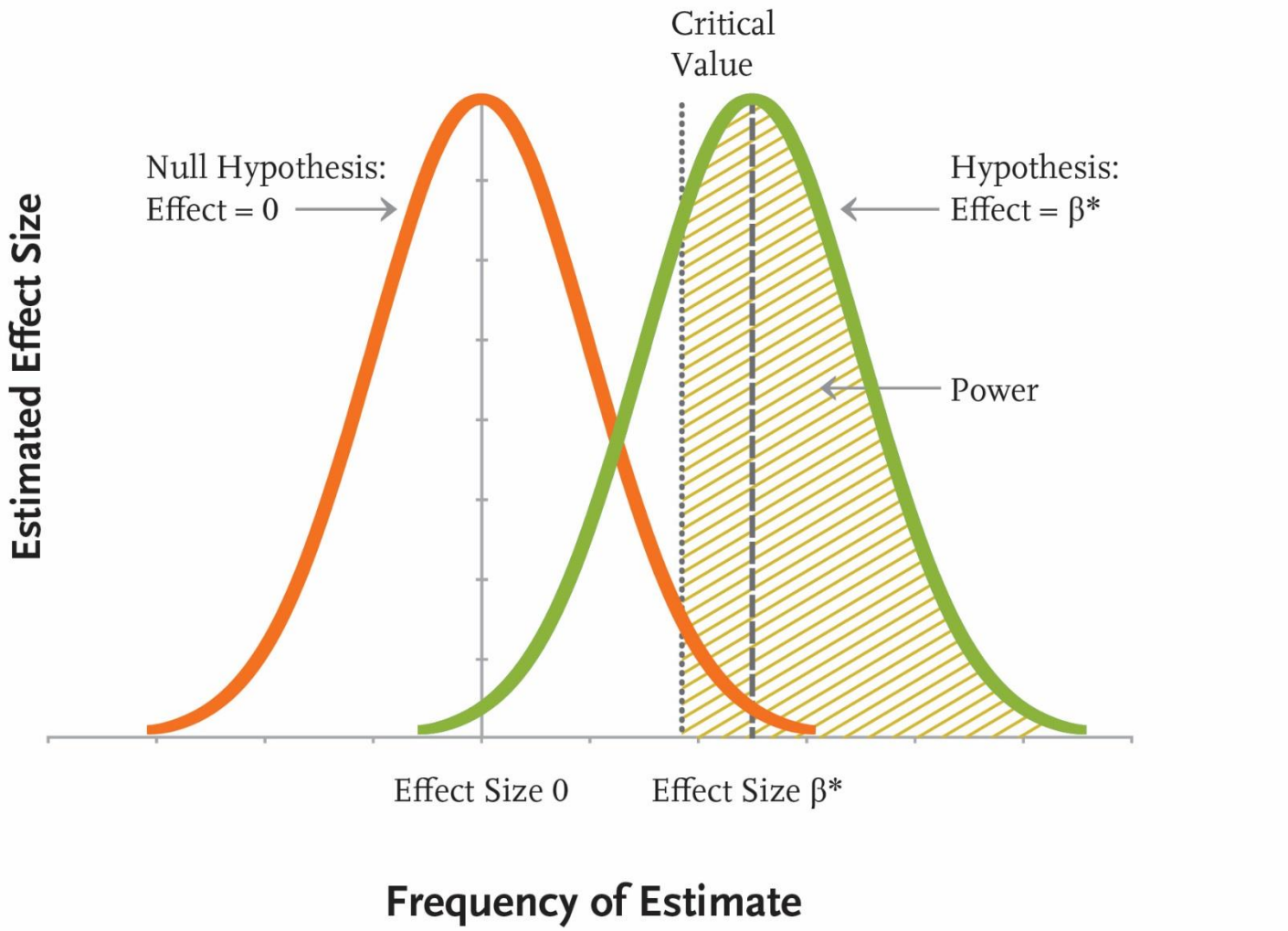
Hypothesis testing and power



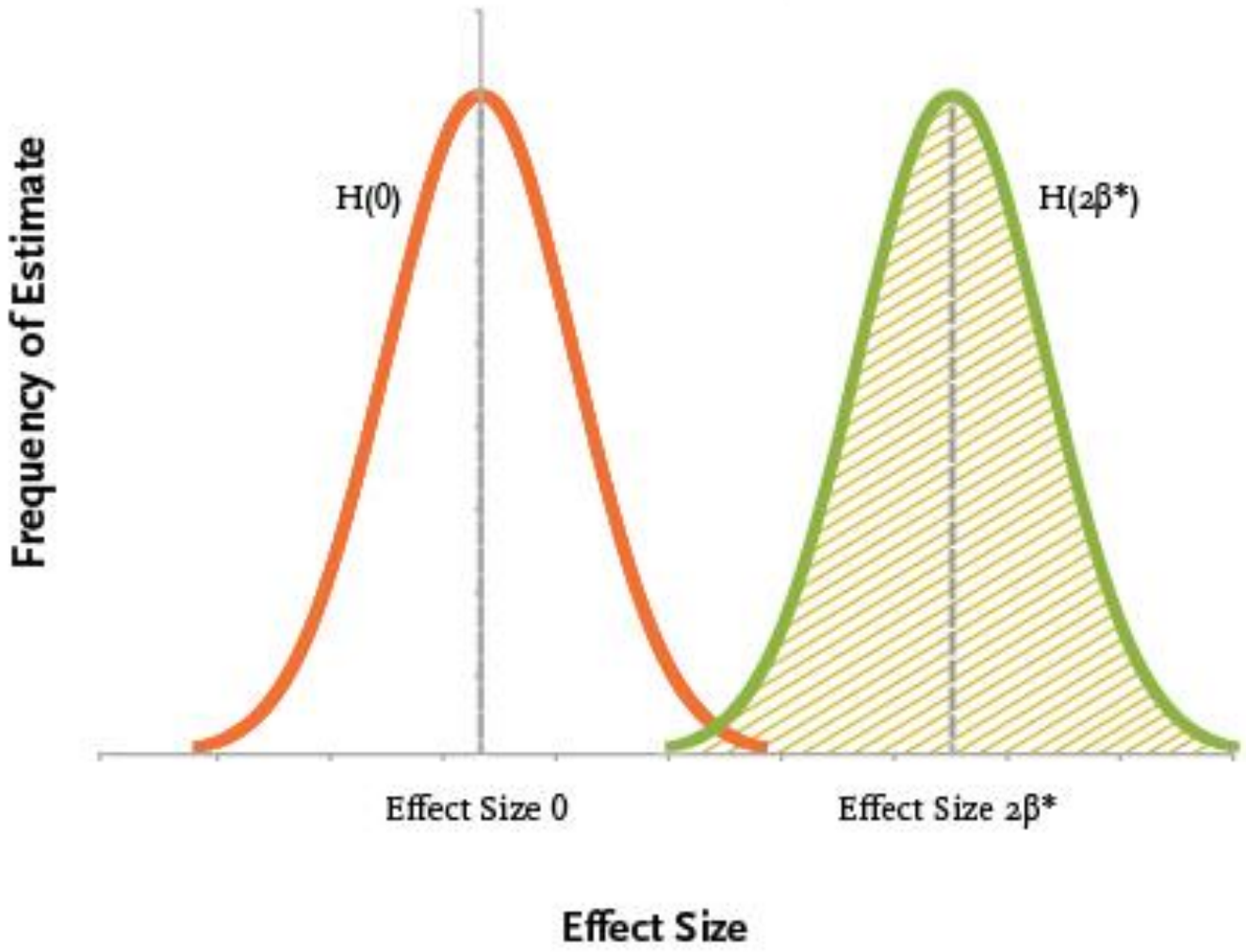
Usually we do a two sided test so examine probability if true effect was - β* or + - β*



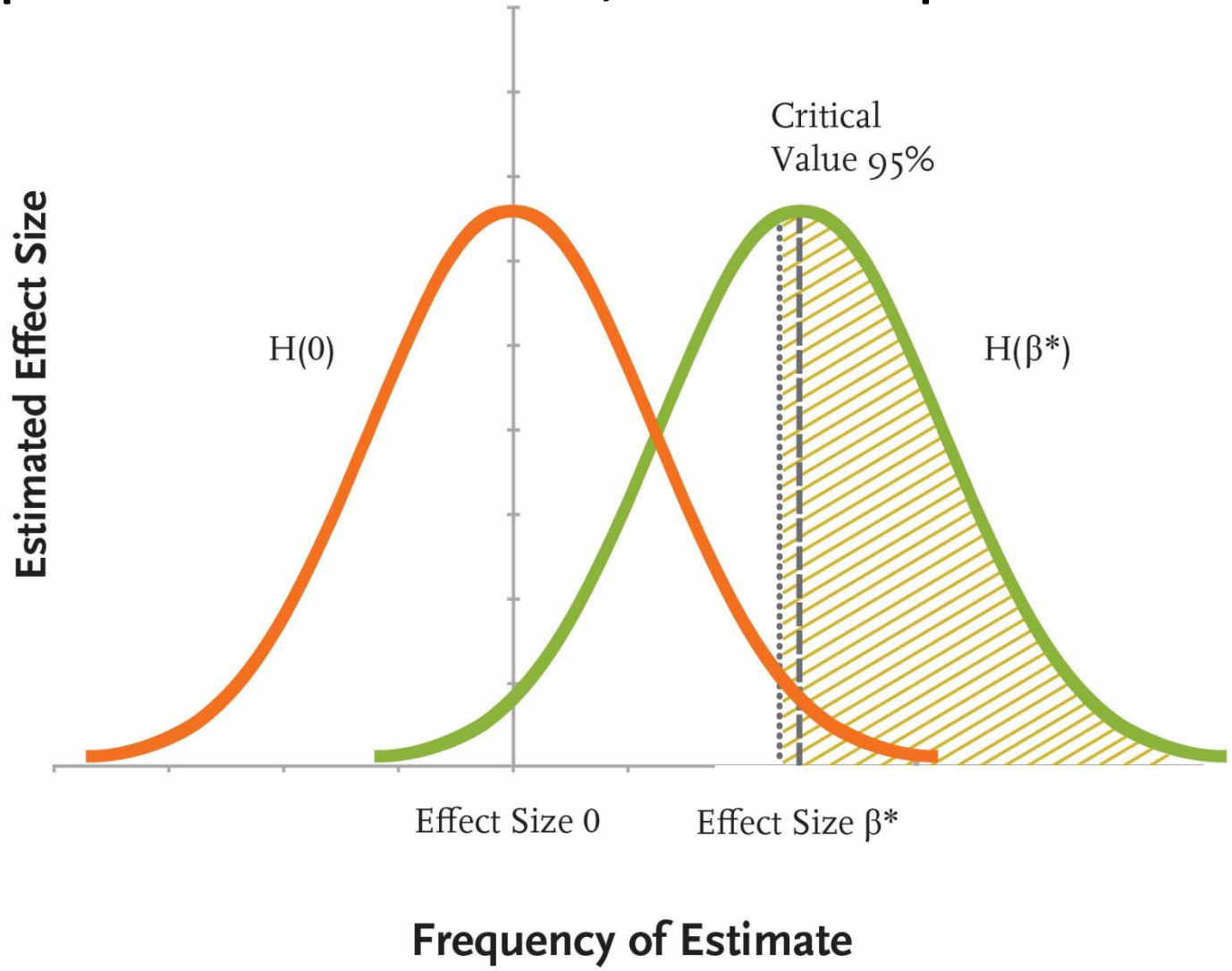
More overlap between H0 curve and Hβ* curve, the lower the power. Q: what effects overlap?



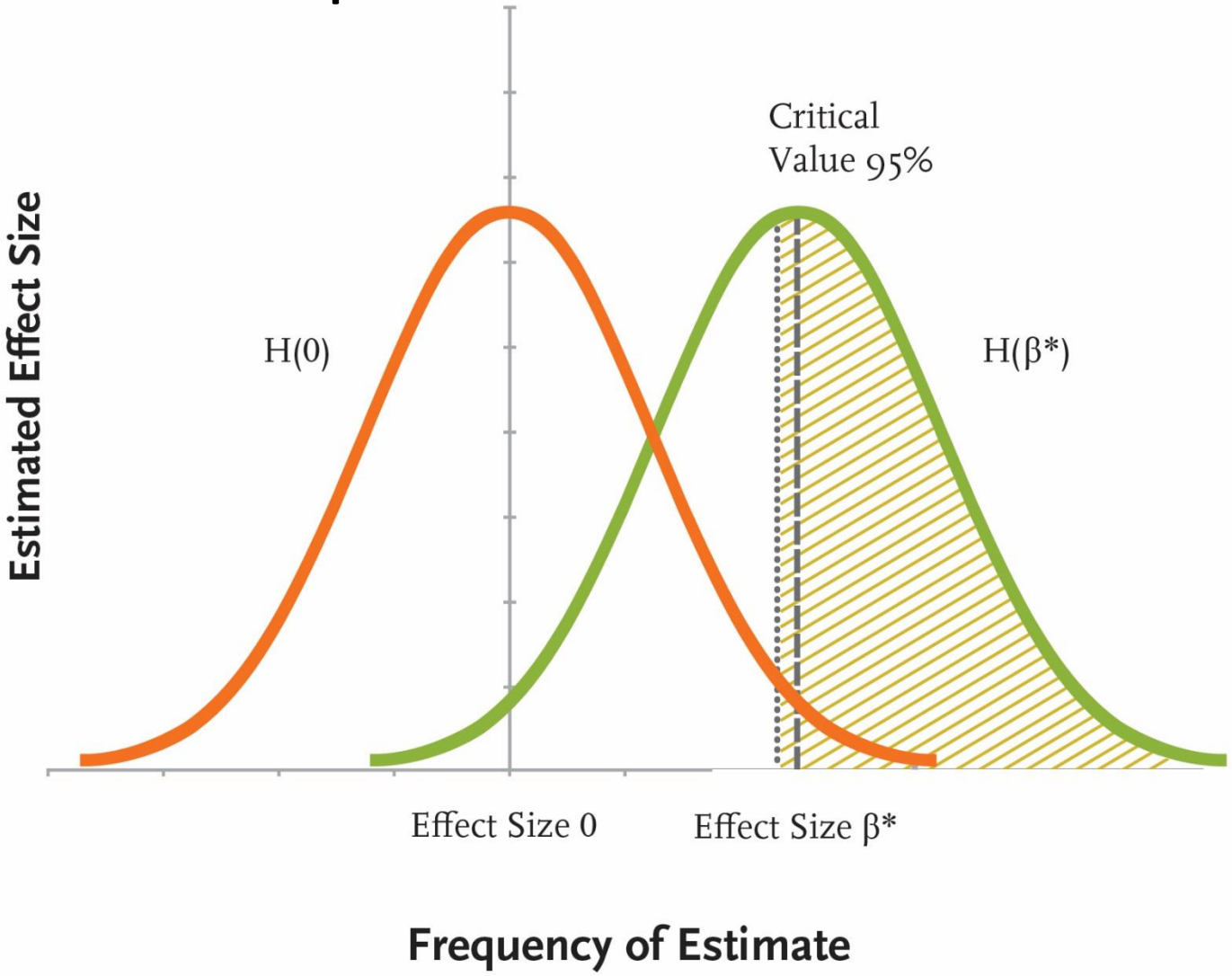
Larger hypothesized effect, further apart the curves, higher the power



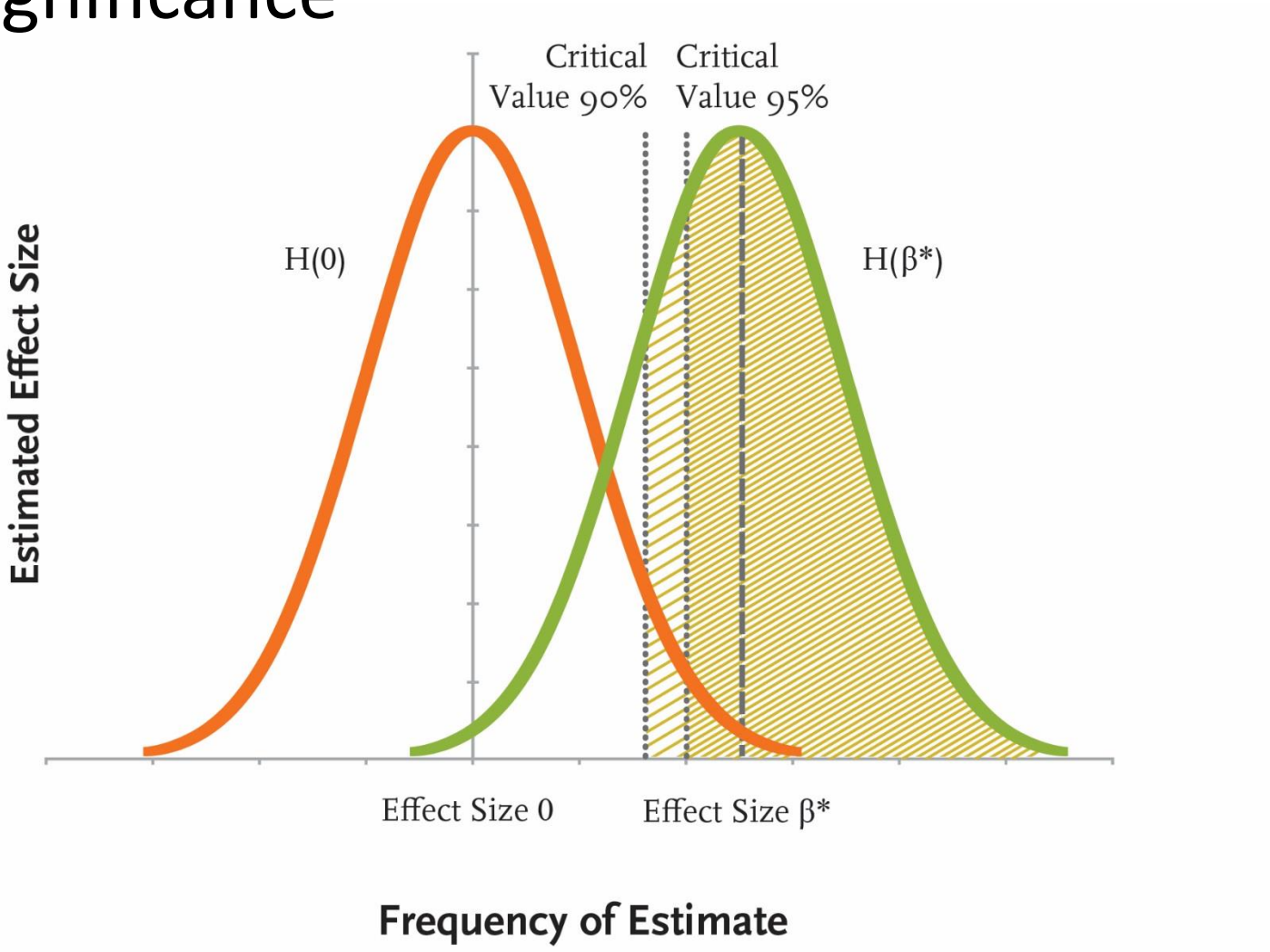
Greater variance in population, increases spread of possible estimates, reduces power



Small sample size, reduces precision of estimate and reduces power



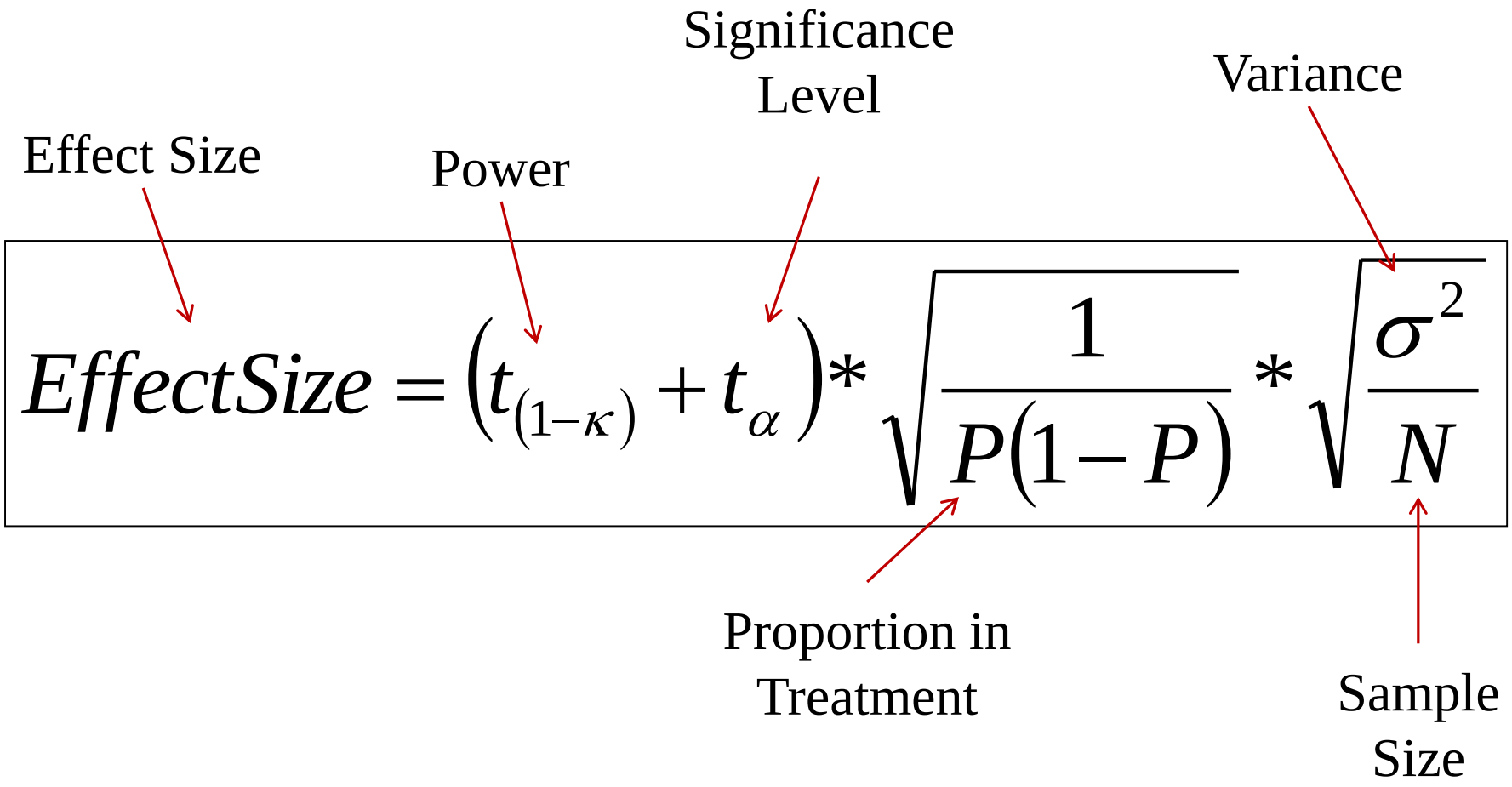
10% significance gives higher power than 5% significance



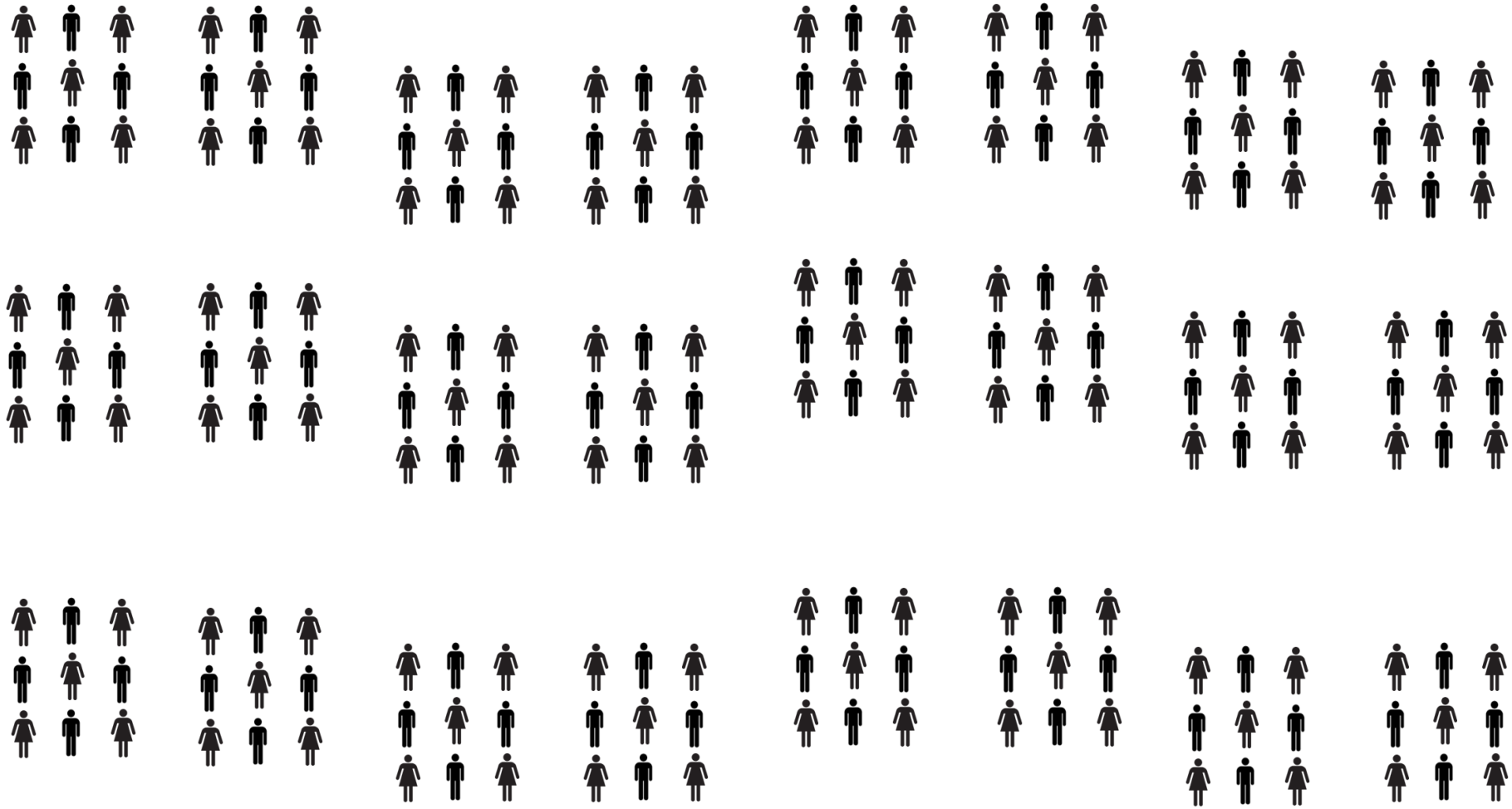
Allocation ratio and power

- Definition of allocation ratio: the fraction of the total sample that allocated to the treatment group is the allocation ratio
- Usually, for a given sample size, power is maximized when half sample allocated to treatment, half to control
- Diminishing marginal benefit to precision from adding sample, so best to add equally

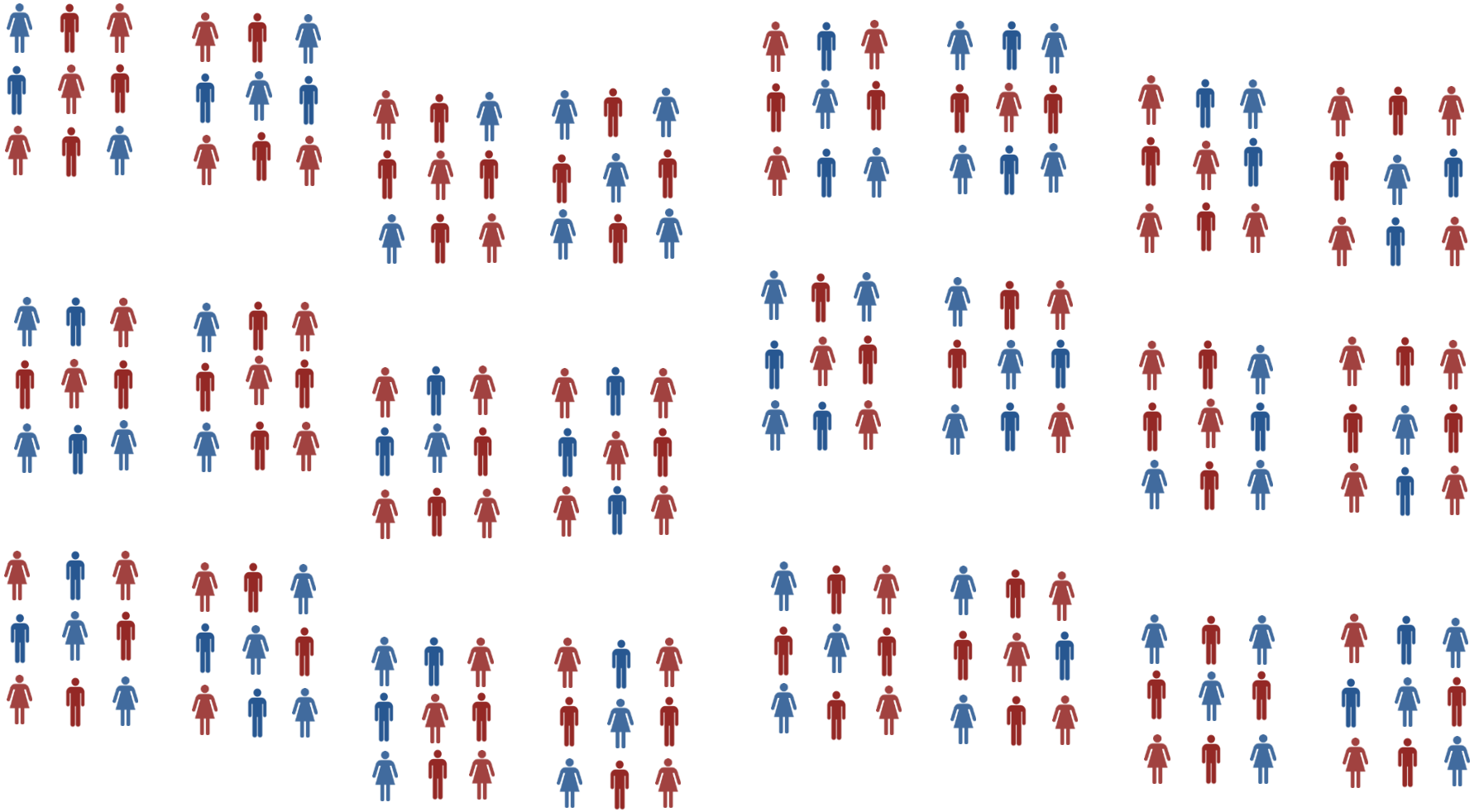
Power equation: MDE



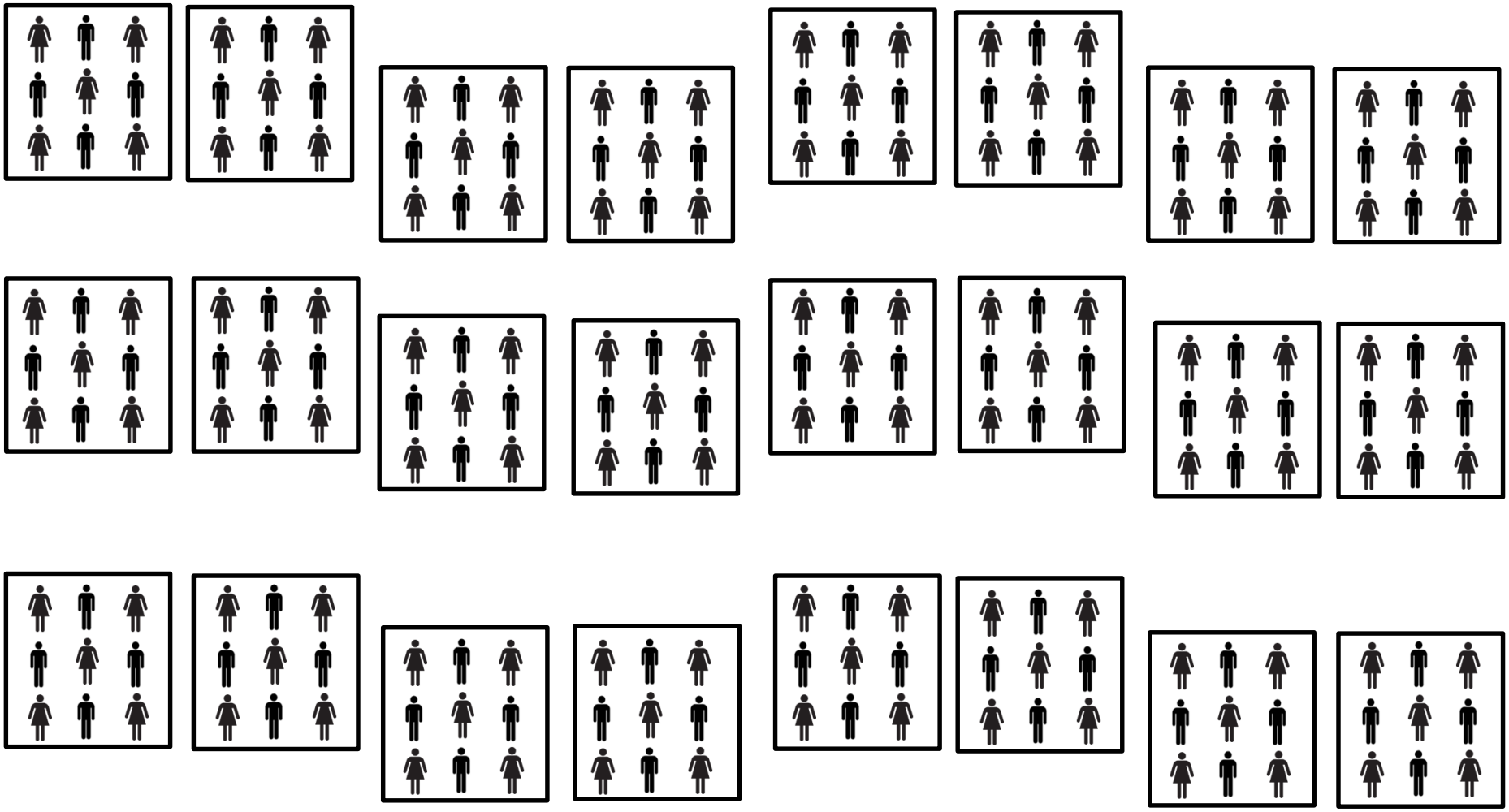
Randomize individuals to T or C



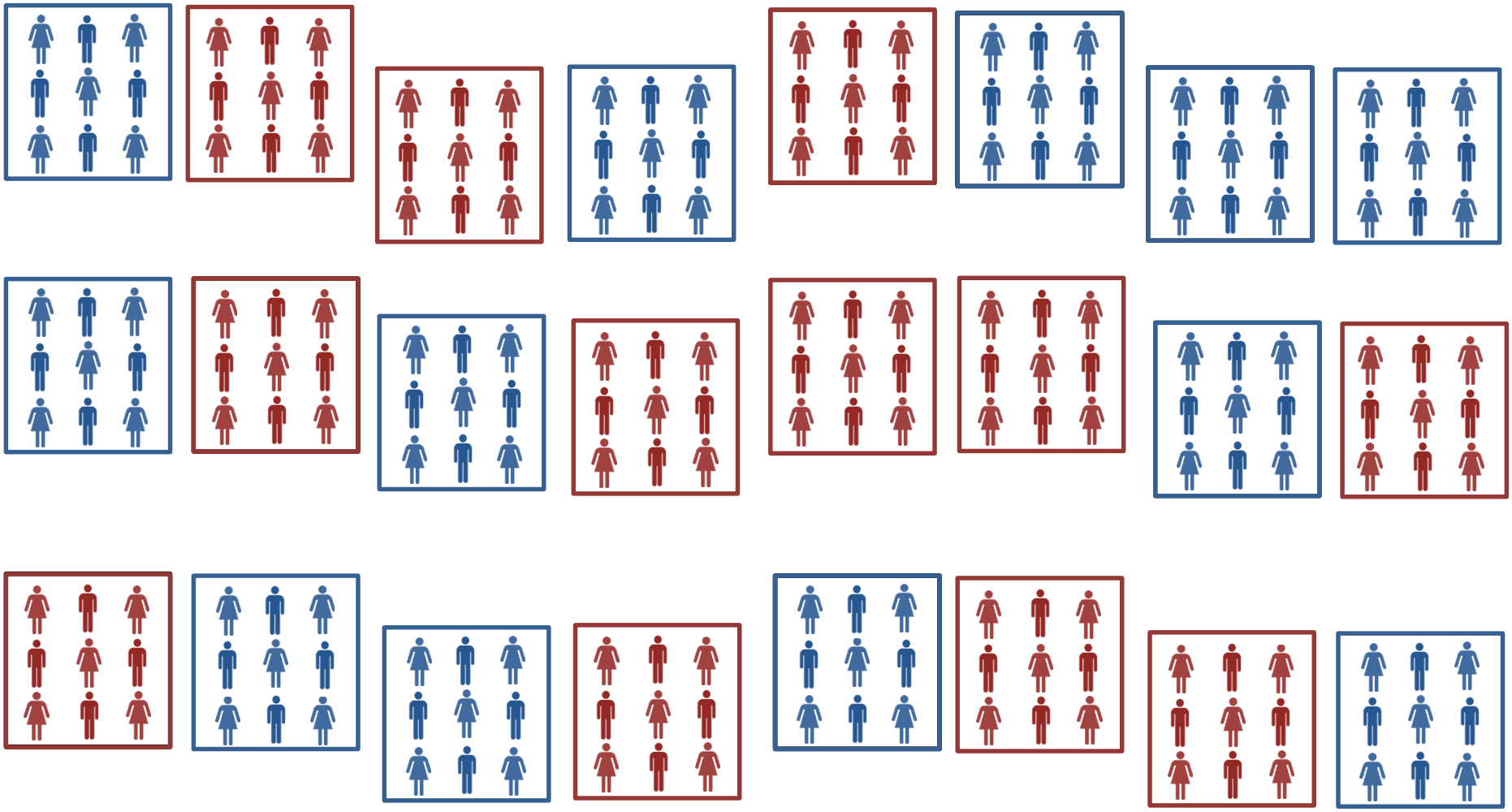
Randomize individuals to T or C



Or randomize clusters: eg classes



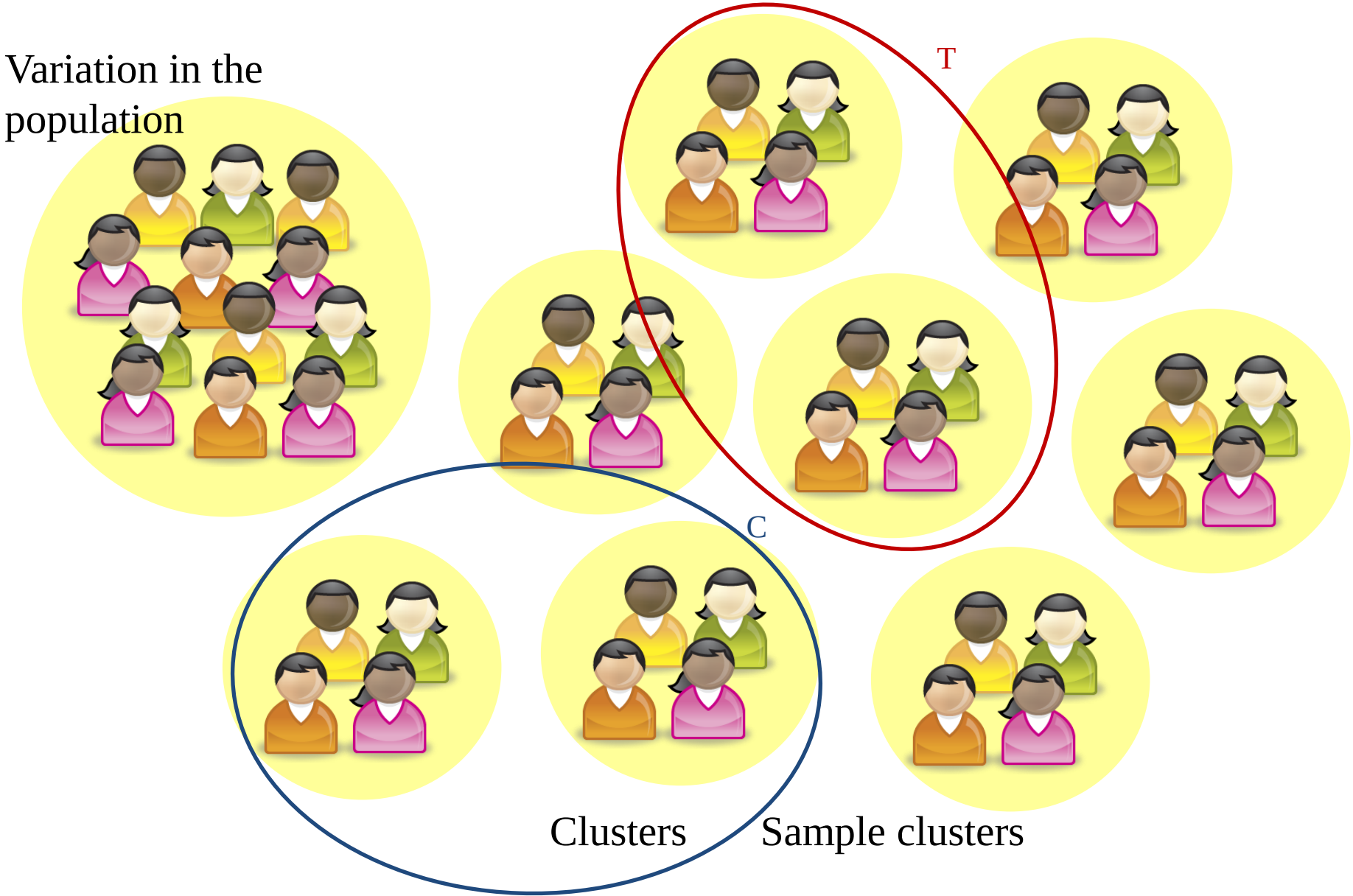
Or randomize clusters: eg classes



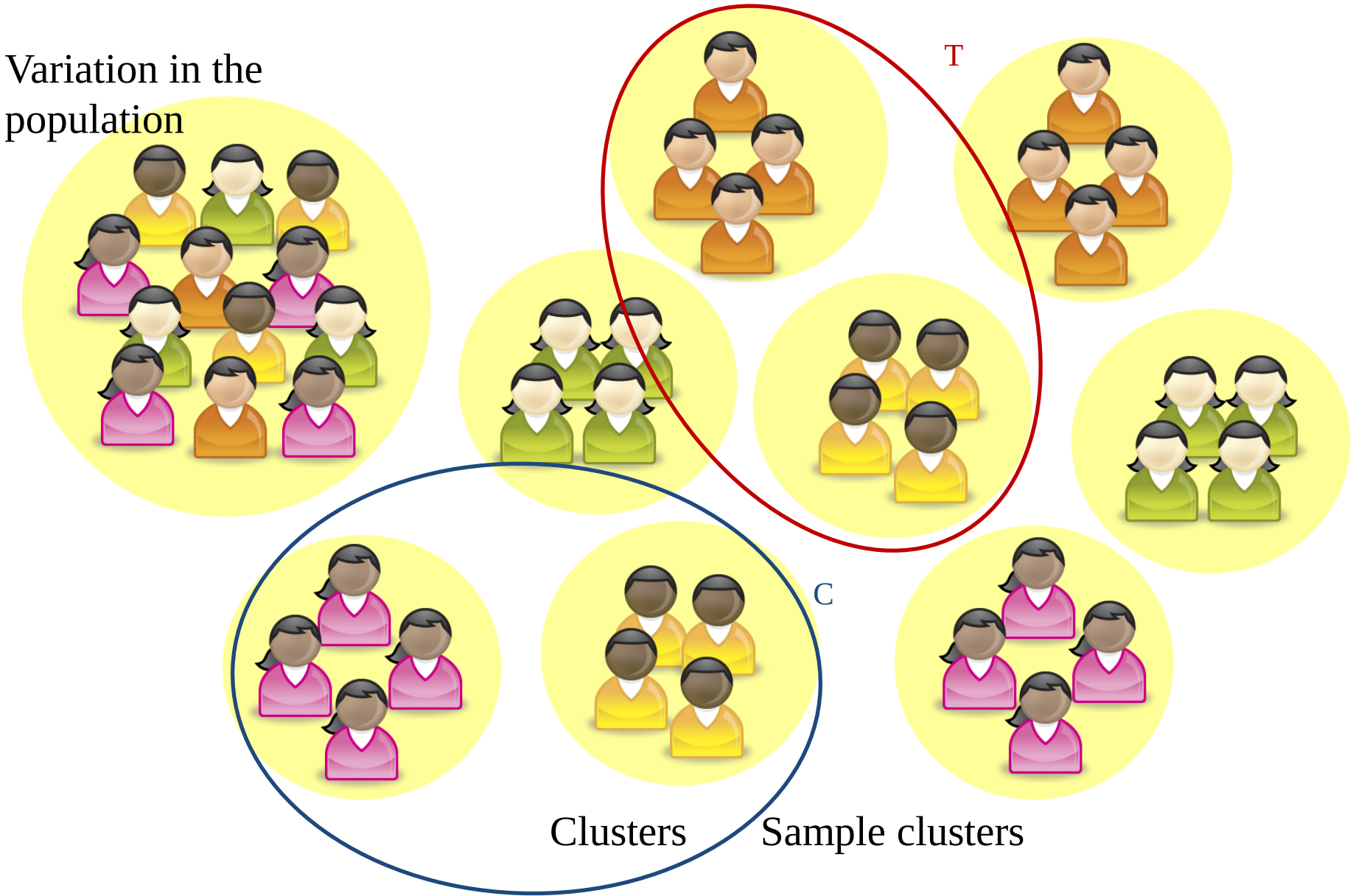
Clustered design: intuition

- We want to know how much rice the average farmer in Sierra Leone usually grows
  - Randomly select 9,000 farmers from across country
  - Randomly select 9,000 farmers from one district
- A few farmers from every district more accurate
  - Some districts better for rice (long run correlations)
  - Some districts may have suffered drought last year (correlated shocks)
- Cheaper to survey farmers in clusters, larger N
- How many per cluster vs how many clusters depends
  - Cost advantage of surveying in clusters
  - Intracluster correlation

Low intracluster correlation



HIGH intracluster correlation



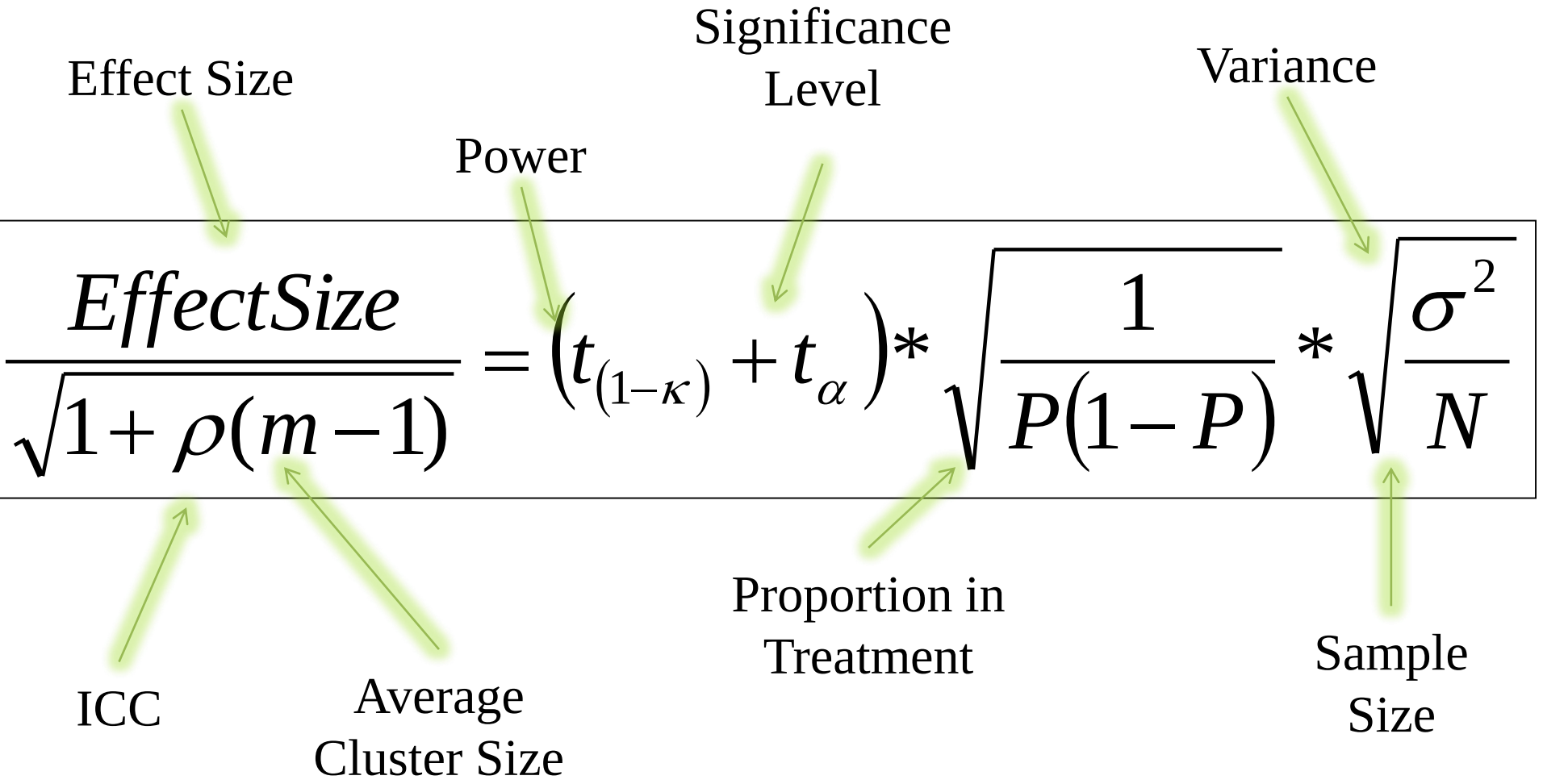
Intracluster correlation

- Total variance can be divided into within cluster variance (τ^2) and between cluster variance (σ^2)
- When variance within clusters is small and the variance between clusters is large, the intra cluster correlation is high (previous slide)
- Definition of intracluster correlation (ICC): the proportion of total variation explained by within cluster level variance
  - Note, when within cluster variance is high, within cluster correlation is low and between cluster correlation is high
- $icc = \rho = \frac{\tau^2}{\sigma^2+\tau^2}$

How does ICC effect power?

- For a given N we have less power when we randomize by cluster (unless intracluster correlation is zero)
- There are diminishing returns to surveying more people per cluster
- Usually the number of clusters is the key determinant of power, not the number of people per cluster

Power with clustering



Calculating power in practice

Calculating power: A step-by-step

1. Set desired power (80%, 90%) and significance (95%)
2. Calculate residual variance (& rho) using pilot data or data from other studies in similar population (eg DHS)
3. Decide number of treatments
4. Set MDE size for T vs C and between treatments
5. Decide allocation ratio
6. Calculate sample size
7. Estimate resulting budget
8. Adjust parameters above (e.g cut number of arms)
9. Repeat

Residual variance

- If a lot of variance in outcomes in our population this makes our estimated effect size less precise
- Some variation can be explained by observables
  - more educated farmers have higher yields
- Using controls in analysis soaks up variance, impact more precisely estimated, more power
- Calculate residual variance by regressing outcome on controls in existing data
- Baseline value of outcome good control
  - In stata can add number or rounds data collection
  - Need estimate of correlation in outcome between rounds

Estimating rho

- Rho must be between 0 and 1
- Depends on context and variable.
- Need big samples to calculate accurately
- Calculate power for different estimates of rho

Malawi: Households produces maize	0.003
Sierra Leone: Households produce cocoa	0.57
Sierra Leone: Average rice yields	0.04
Busia, Kenya: Math and language test scores	0.22
Busia, Kenya: Math test scores	0.62
Mumbai, India: Math and language test scores	0.28

Number of treatment arms

- Different treatment arms help disentangle different mechanisms behind an effect
- Tempting to have large number of treatment arms
- If not enough power to distinguish between arms we learn little
- Usually good to have at least one intensive arm, where a zero would be surprising
- In the analysis will it be useful to pool all the treatment arms, creating an “any treatment” arm?
  - Eg any price vs free
  - Any information vs none

Unequal allocation ratio

- Usually we want equal sample in T vs C because marginal power from additional sample per cell declines
- If budget covers treatment and evaluation, more expensive to add one person to T than C
  - Unequal allocation ratio gives a bigger total N which may be worth it
  - Try different allocation ratios within a given budget and explore tradeoff
- With multiple treatments put more sample behind most important question
  - If going to pool treatments, have bigger control
  - If really care about between treatments need bigger sample in treatment groups

Minimum detectable effect size

- The most important ingredient for calculating power
- MDE is not the effect size we expect or want
- MDE is the effect size below which we may not be able to distinguish the effect from zero, even if it exists
  - Ie below which effect might as well be zero
- Useful questions to ask when determining MDE:
  - Below what effect size would the program not be cost effective
  - How big an effect would this need to be to be interesting? (small deviation from rational not so interesting, big deviation is interesting)
- MDE may be smaller between arms than between T and C
  - Common mistake is powering on T vs C so don't have power to distinguish between arms

Calculating power in stata

- Stata has a new command “power” where you can state sample size and get out power
  - But does not allow for clustering
- Most still use sampsi and sampclus (add ons)
  - Default is power 90%, significance 5%, equal allocation
- To detect an increase in average test scores from 43% to 45% with power of 80%:

  sampsi 0.43 0.45, power(0.8) sd(0.05)

- Stata gives N per cell eg N1=99
  - With multiple arms need to multiply by number of cells (ie number of treatments plus control)
- For binary outcomes, SD determined by mean

Power in stata with clustering

- First calculate sample size without clustering and then add information on cluster
- If above experiment were to be randomized at class level with 60 per class and ICC of 0.2

```
sampsi 0.43 0.45, power(0.8) sd(0.05)

sampclus, obsclus(60) rho(0.2)
```

Power with optimal design

- Optimal design is a free software specifically designed for power calculations
- MDE must be entered in standardized effect size (ie effect size divided by standard deviation)
- OD allows multiple levels of clustering and works with dropdown menus—see JPAL exercise for details