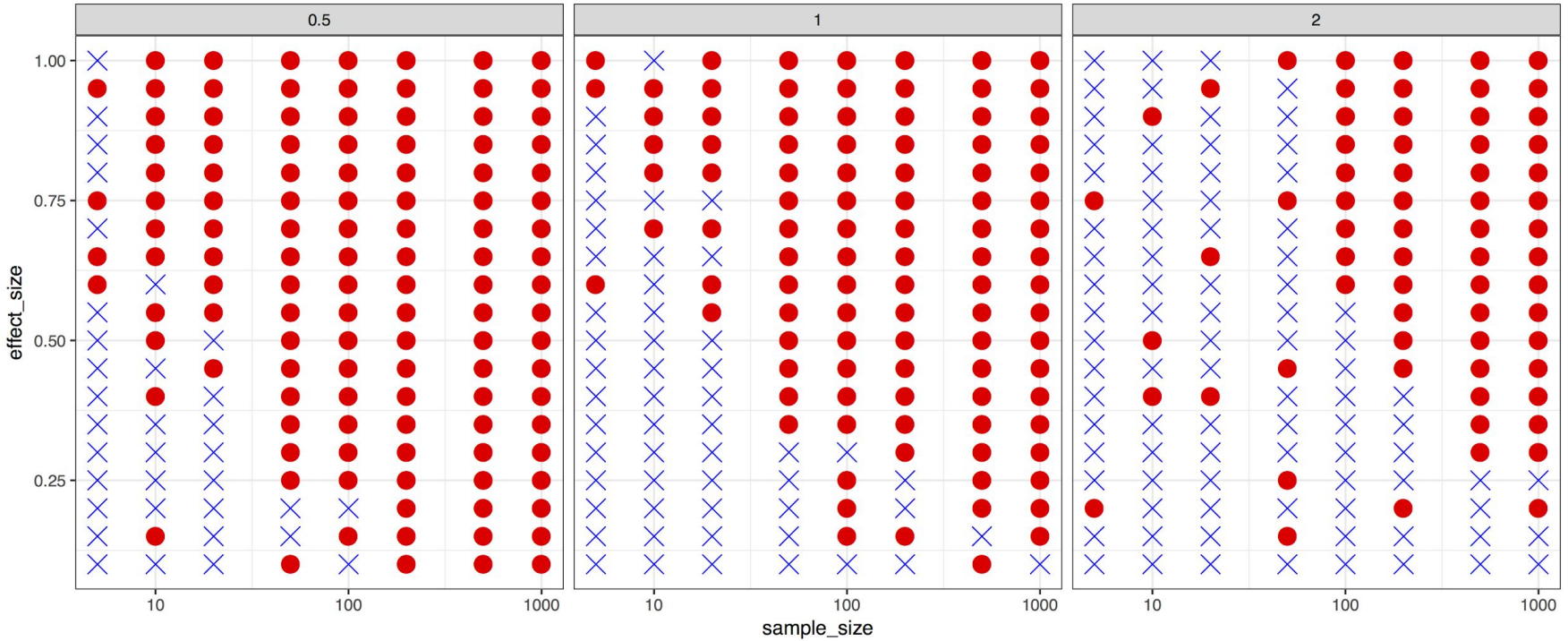
Day 04

Statistical power

- Statistical power
- Dependence on sample size, effect size, and significance threshold

P-value

- P-values are dependent on: sample_size, effect_size, within-group variance



Statistical power of a study

The statistical power of a study is the probability that it can distinguish an effect of a given size from random chance.

Power = True positive rate = Sensitivity = Recall

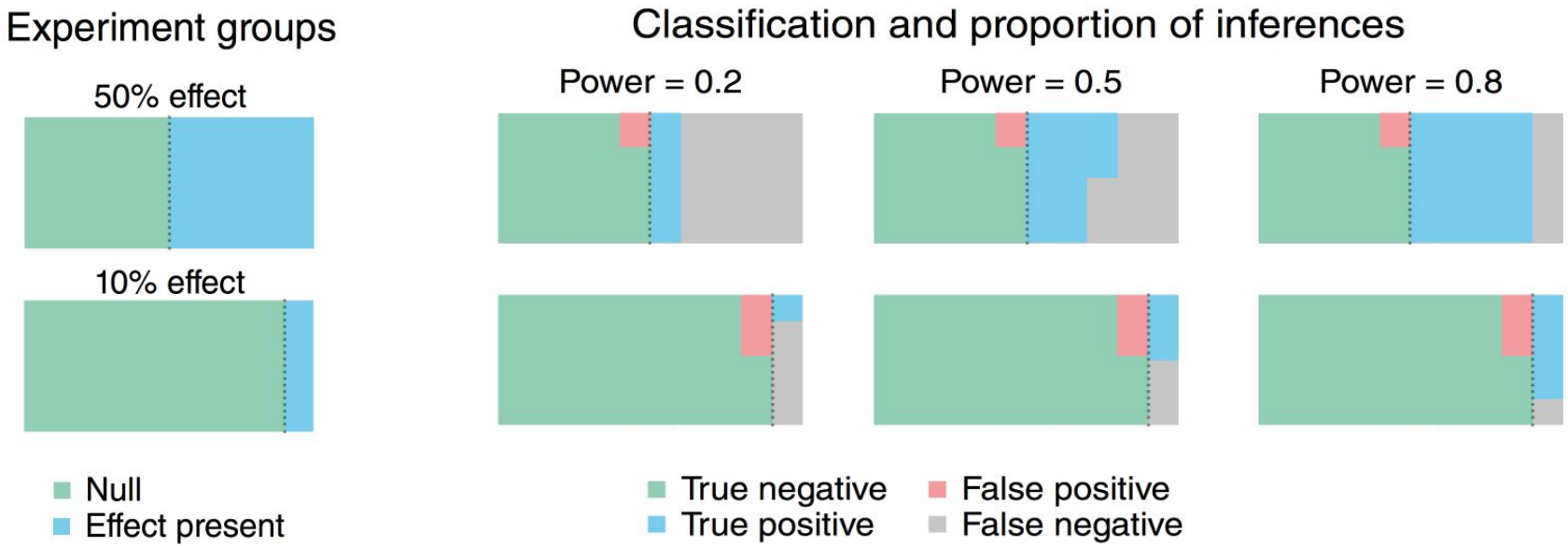
Many studies are underpowered → Waste of resources & Unethical

http://www.nature.com/doifinder/10.1038/nmeth.2738

Statistical power of a study

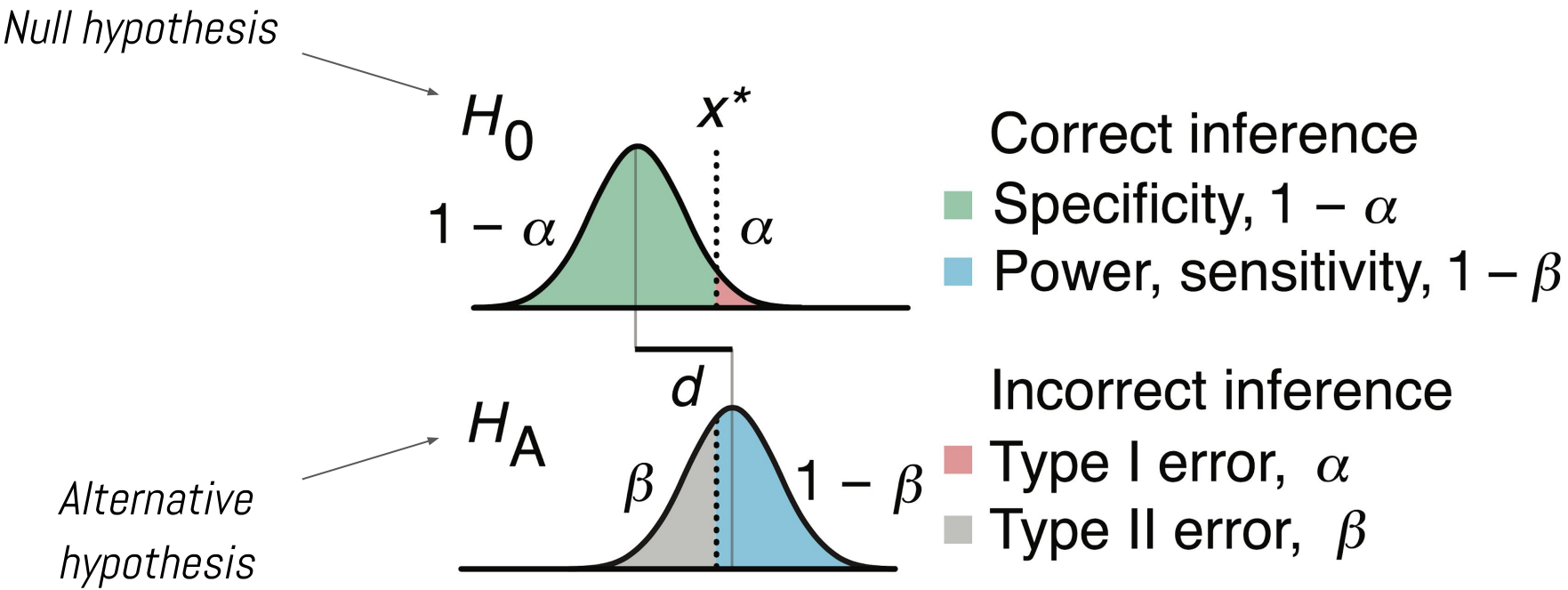
Probability that the study can distinguish an effect of a given size from random chance.

Power = True positive rate = Sensitivity = Recall



http://www.nature.com/doifinder/10.1038/nmeth.2738

Statistical power

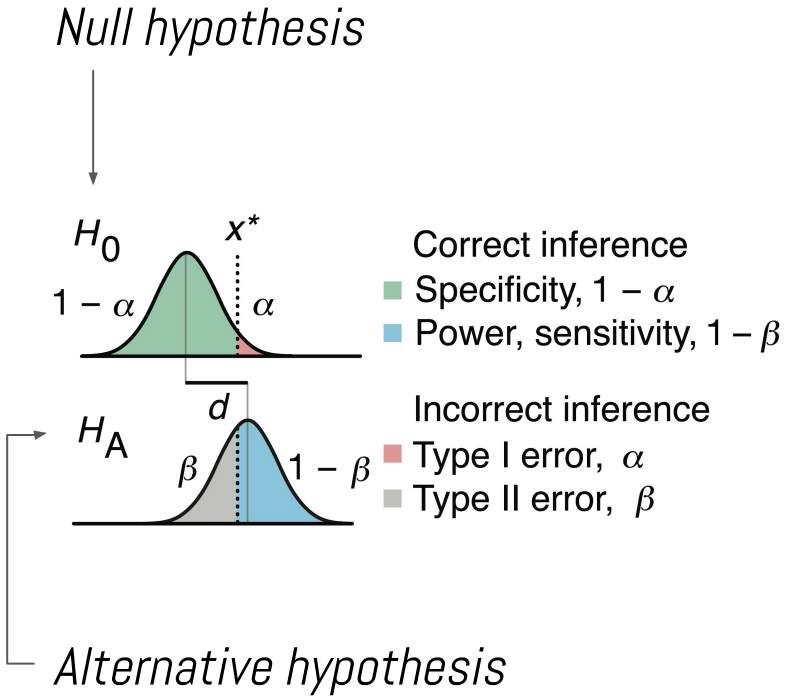


http://www.nature.com/doifinder/10.1038/nmeth.2738

Statistical power

The power is the probability that the test correctly rejects the null hypothesis (H_0) when a specific alternative hypothesis (H_1) is true.

- Pr(reject H_0 | H_1 is true)
- H_1 has to be specific (cannot just be negation of H_0)
- The probability that it will yield a statistically significant outcome.

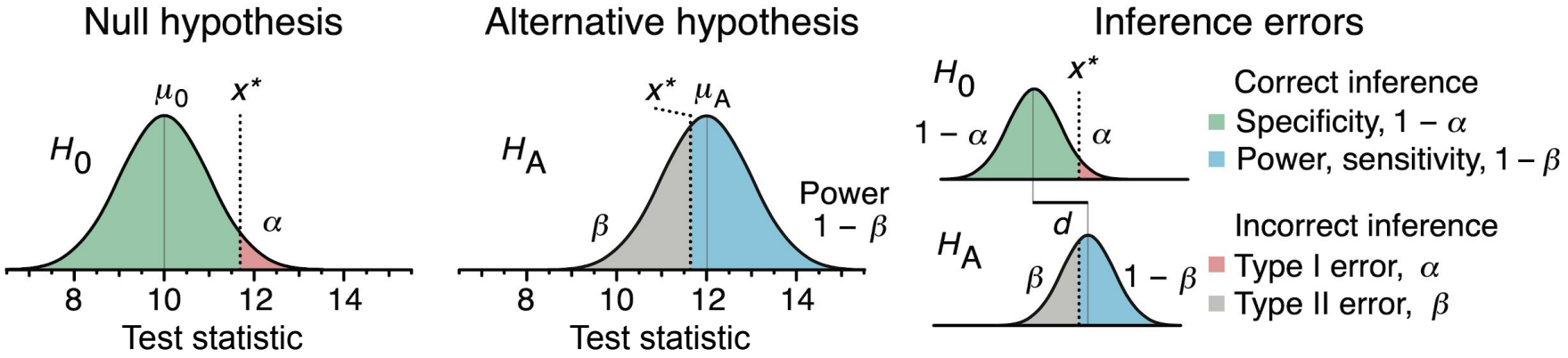


Power = 1 – β

As power increases ⟶ Probability of making type II error (β) decreases.

http://www.nature.com/doifinder/10.1038/nmeth.2738

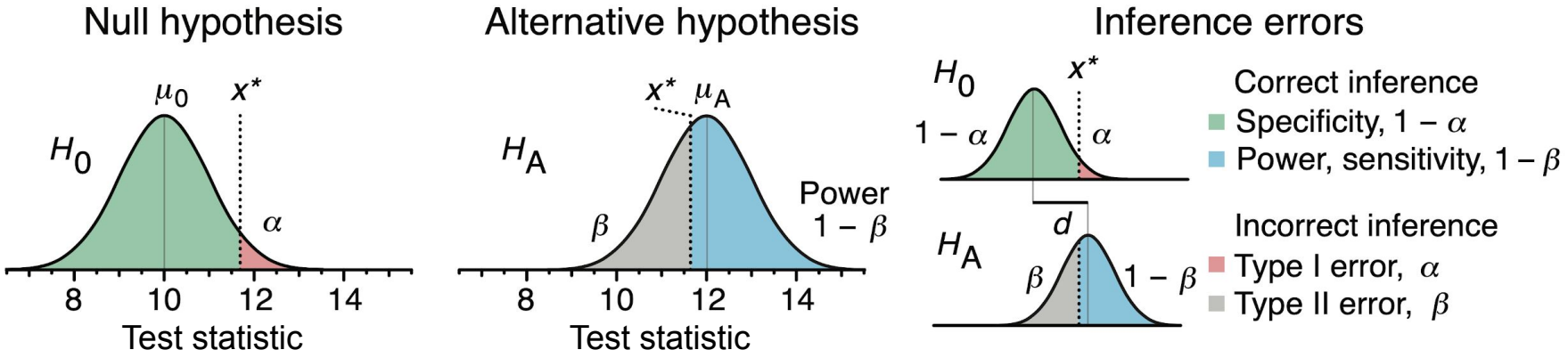
Statistical power



- Values sampled from $H_A < x^*$ do not trigger rejection of H_0 and occur at a rate β.
- Power (sensitivity; TPR) = 1 – β (blue area).
- Good to have low α (FPR) & low β (FNR), but:
  - The α and β rates are inversely related: ↓α → ↑β (& reduces power).

http://www.nature.com/doifinder/10.1038/nmeth.2738

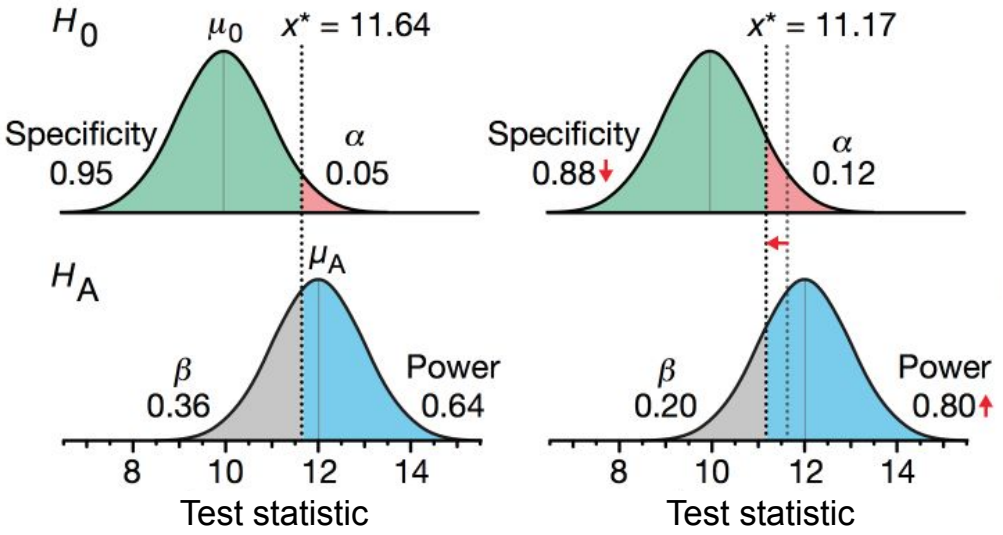
Statistical power

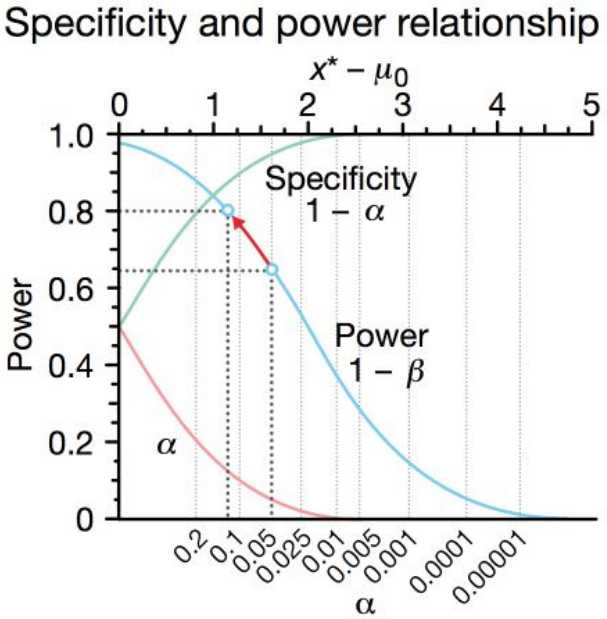


- Typically, $\alpha < \beta$: consequences of FP (in an extreme case, a retracted paper) are more serious than those of FN (a missed opportunity to publish).
- But, the balance between α and β depends on the objectives:
  - If FP are subject to another round of testing but FN are discarded, β should be kept low.

http://www.nature.com/doifinder/10.1038/nmeth.2738

Compromise between specificity and power

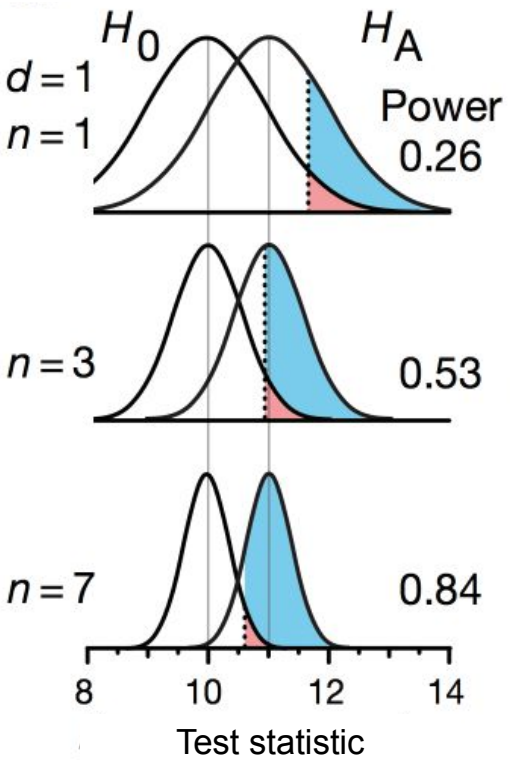


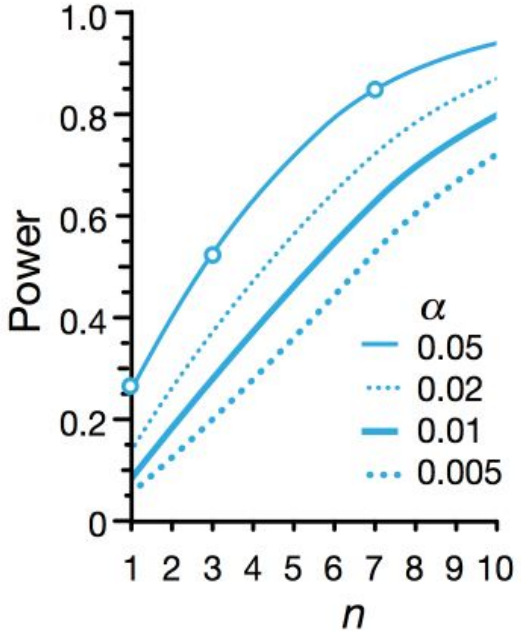


- Decreasing specificity (TNR) increases power (TPR)
- Can we improve our chance to detect increased effect from H_A (increase power) without compromising α (increasing FP)?

http://www.nature.com/doifinder/10.1038/nmeth.2738

Impact of sample size on power





One can control experimental conditions (e.g. using genetically identical orgs under lab conditions; adding precise amount of a drug) to reduce the variation b/w samples & compensate to some extent for small sample sizes.

In practice, because we estimate population σ from the samples, power is decreased and we need a slightly larger sample size to achieve the desired power.

http://www.nature.com/doifinder/10.1038/nmeth.2738

Statistical power depends on a number of factors

Power depends on:

- Statistical significance criterion:
  - Lesser conservative test (larger significance criterion) → More power
- Sample size:
  - Collecting more data → Easier to detect small effects; relates to the efficiency of a given testing procedure, experimental design, or an estimator (sample size required for a given power)
- Size of the effect:
  - Larger the effect → Easier it is to detect; (std. effect size better)

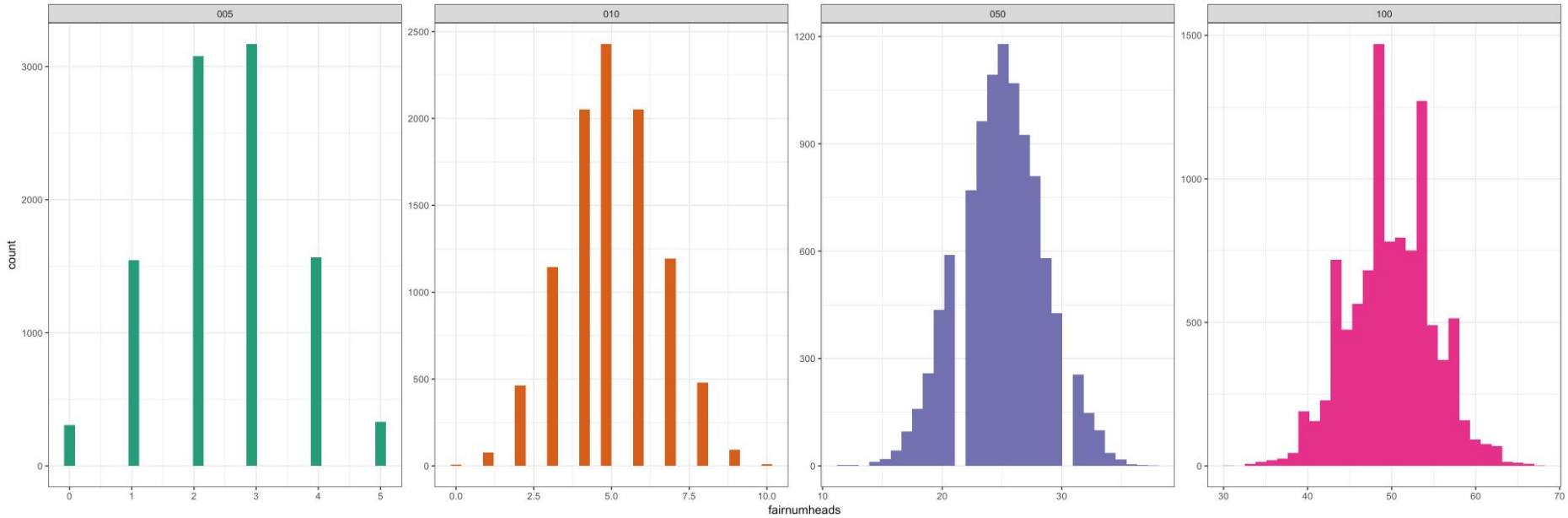
- Measurement error: counting cells vs. estimating level of fatigue/depression
- Experimental design: e.g. in a two-sample setting, optimal to have equal number

Generating a power curve

Let's examine some code to generate a power curve to detect unfair coins:

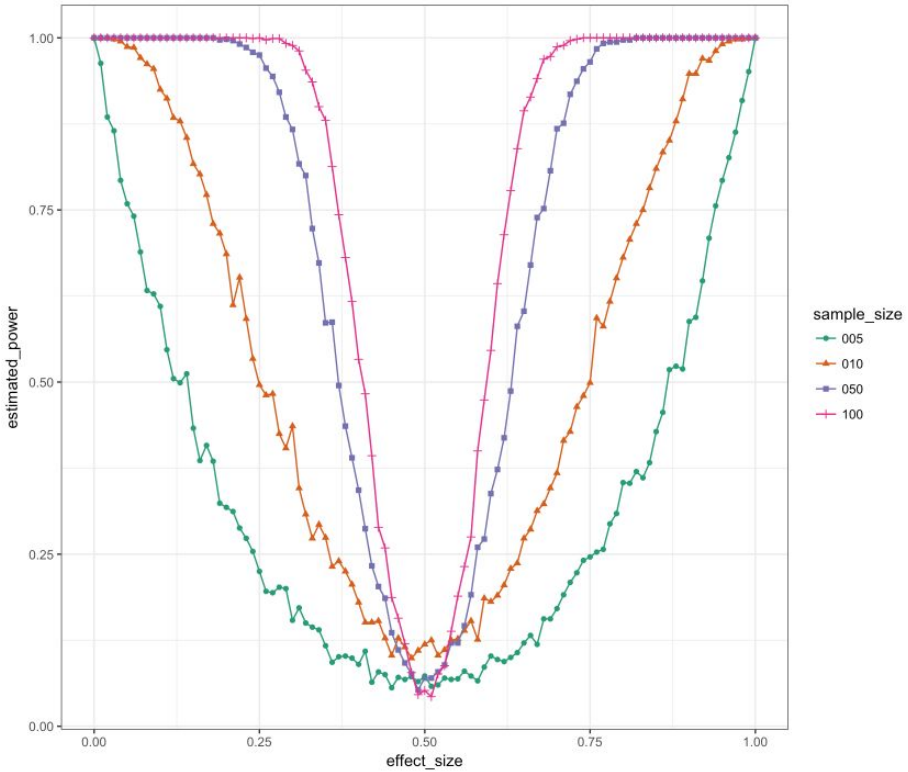
- You are given a coin and asked to detect if it is biased.
- Experiment: Flip the coin **num_flips** and take a call.
- Establish the null hypothesis (**num_permutations** = 10,000) for a given **num_flips** (this is the **sample size**).
- For a given **bias** (i.e. the **effect size**), find out how many times does an experiment like the one above can reject the null hypothesis.

Generating a power curve



Null distributions for different sample sizes

Generating a power curve



Power curves for different sample sizes

Power analysis

Balancing sample size, effect size, and power is critical to good study design.

- First, set the values of type I error (α) and power ($1 - \beta$) to be statistically adequate:
  - Traditionally 0.05 and 0.80, respectively.
- Then determine sample size (n) on the basis of the smallest effect we wish to measure.
  - If the required sample size is too large → may need to reassess objectives or more tightly control the experimental conditions to reduce the variance.
- When the power is low, only large effects can be detected, and negative results cannot be reliably interpreted.

Typical sample sizes

- Clinical research (behavioral or drug treatments):
  - Need enough participants to represent all subtypes for which treatment might be used.
  - Some issues: lack reliable methods for diagnosis.
  - Rough rule of thumb: at least 100 people.
    - The actual number needed to find a valid effect depends on a range of factors, including the magnitude and frequency of the effect in the general population.

https://www.spectrumnews.org/news/studies-size-matters-let-us-count-ways/

Typical sample sizes

- Brain imaging studies:
  - Historically included 20 or fewer participants. In the past 10 years, closer to 100 participants.
  - Studies that aim to trace developmental trajectories should also track the same few individuals over time, scanning their brains at regular intervals, rather than examining a cross-section of people of different ages at different sites.

https://www.spectrumnews.org/news/studies-size-matters-let-us-count-ways/

Typical sample sizes

- Genetic studies (large no. of variants/genes, each making a small contribution):
  - Rare variants in coding regions: order of thousands of people.
  - Risk variants across the whole-genome: tens of thousands of individuals.
    - Millions of statistical tests, one per variant → increases FPR.
  - GWAS: hundreds of thousands of individuals
    - Common gene variants that contribute to the risk of a condition.

https://www.spectrumnews.org/news/studies-size-matters-let-us-count-ways/

Typical sample sizes

- Preclinical research:
  - Underpowered animal studies for decades (cost and ethical issues).
  - Make up for their low numbers by analyzing a large number of cells or other samples from each animal → 'pseudoreplication.'
  - Can control lab animals' diets, ages and housing conditions, and scale doses or treatments by weight → sample sizes on the order of 10 animals to be acceptable. Should ≥15 per group to identify important biological effects.
  - In the past few years, push for larger numbers in animal studies.

https://www.spectrumnews.org/news/studies-size-matters-let-us-count-ways/

Typical sample sizes

- Biomarker studies (physiological characteristics, such as patterns of eye movements, brain waves or activity, or blood chemistry):
  - Candidate biomarkers have often failed in subsequent studies.
  - Must draw samples from at least 100 individuals.
  - Clinical trials of biomarkers designed to flag people with disease → ≥ 1,000 participants. Researchers should also replicate the efficacy of a biomarker in an independent sample.
  - Some scientists are designing biomarker studies of thousands of participants that combine data from behavioral, imaging and genetic studies.

https://www.spectrumnews.org/news/studies-size-matters-let-us-count-ways/

Typical sample sizes

- Field trials:
  - Variables that are hard to control, and so must include hundreds of individuals to yield meaningful results.
  - Needs more than an appropriate number of participants.
    - Representative mix of sexes and ages.

https://www.spectrumnews.org/news/studies-size-matters-let-us-count-ways/