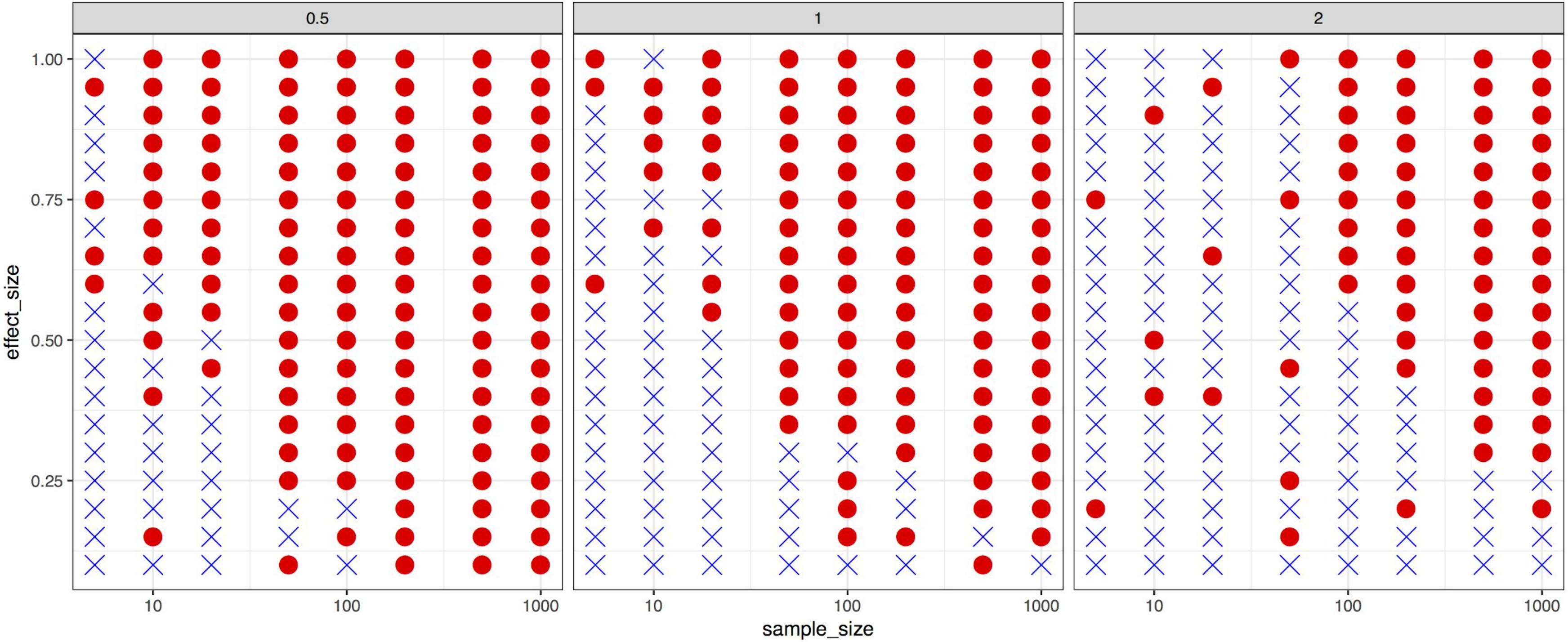
Day 04

Statistical power

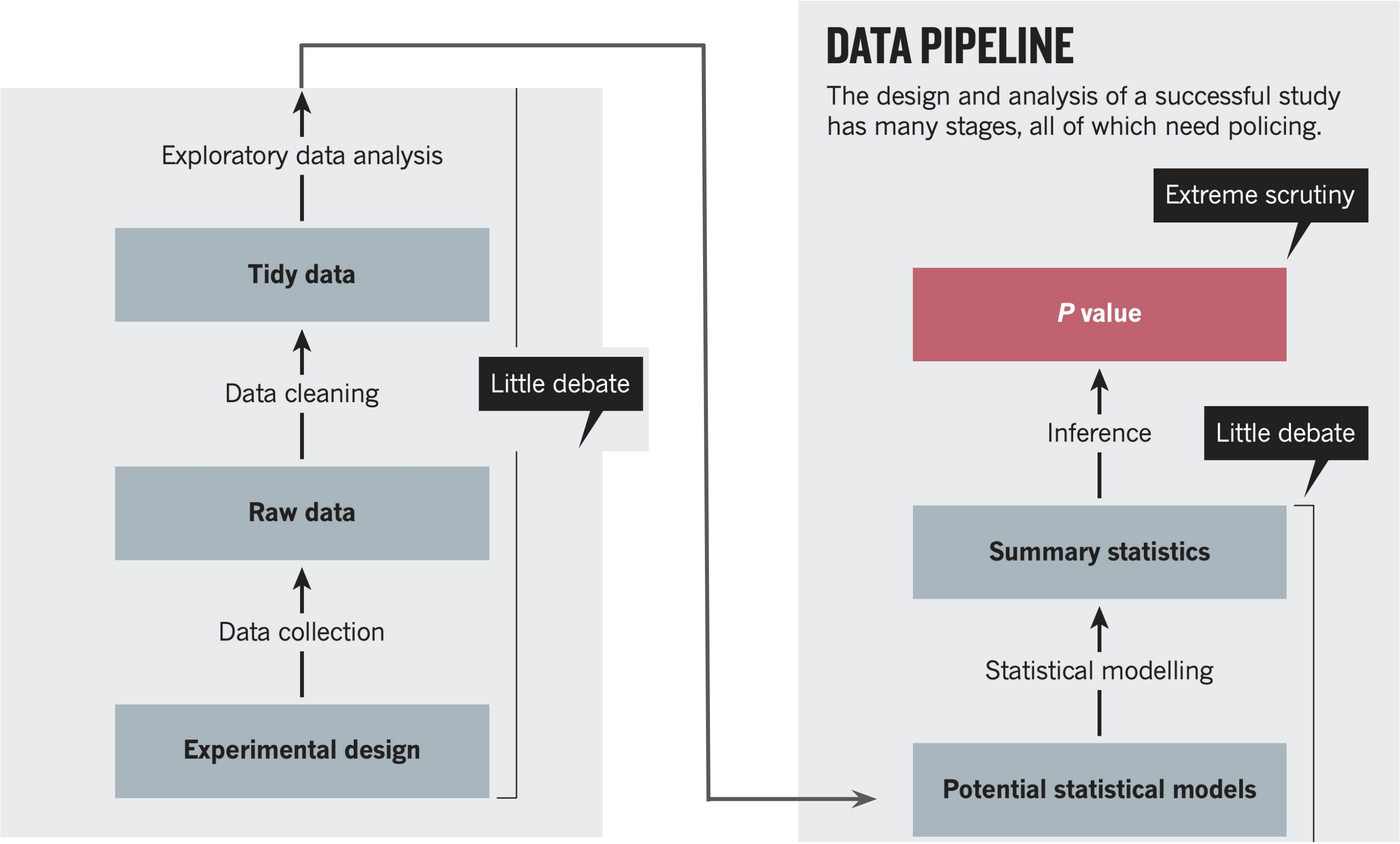
- Statistical power
- Dependence on sample size, effect size, and significance threshold

P-value

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



P-values are just the tip of the iceberg!



DATA PIPELINE

The design and analysis of a successful study has many stages, all of which need policing.

Experimental design

Data collection

Raw data

Data cleaning

Tidy data

Exploratory data analysis

Little debate

Potential statistical models

Statistical modelling

Summary statistics

Inference

Little debate

P value

Extreme scrutiny

JT Leek, RD Peng http://www.nature.com/news/statistics-p-values-are-just-the-tip-of-the-iceberg-1.17412

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

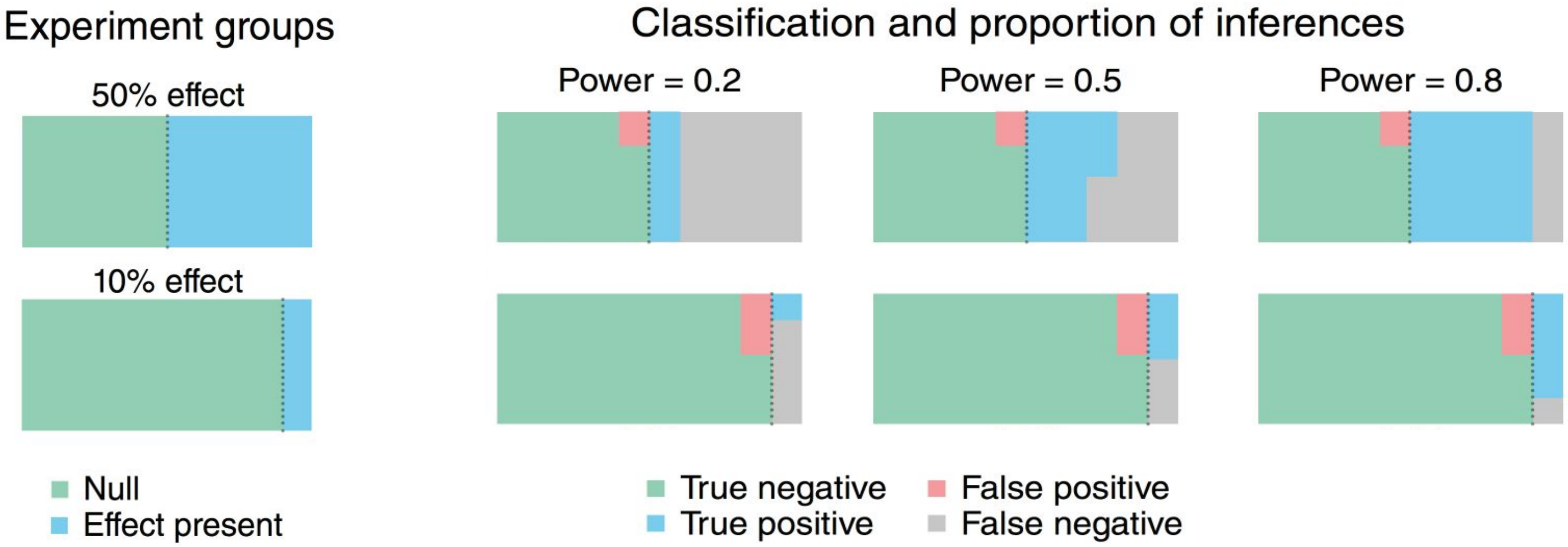
Many studies are underpowered ⟶ Waste of resources & Unethical

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.

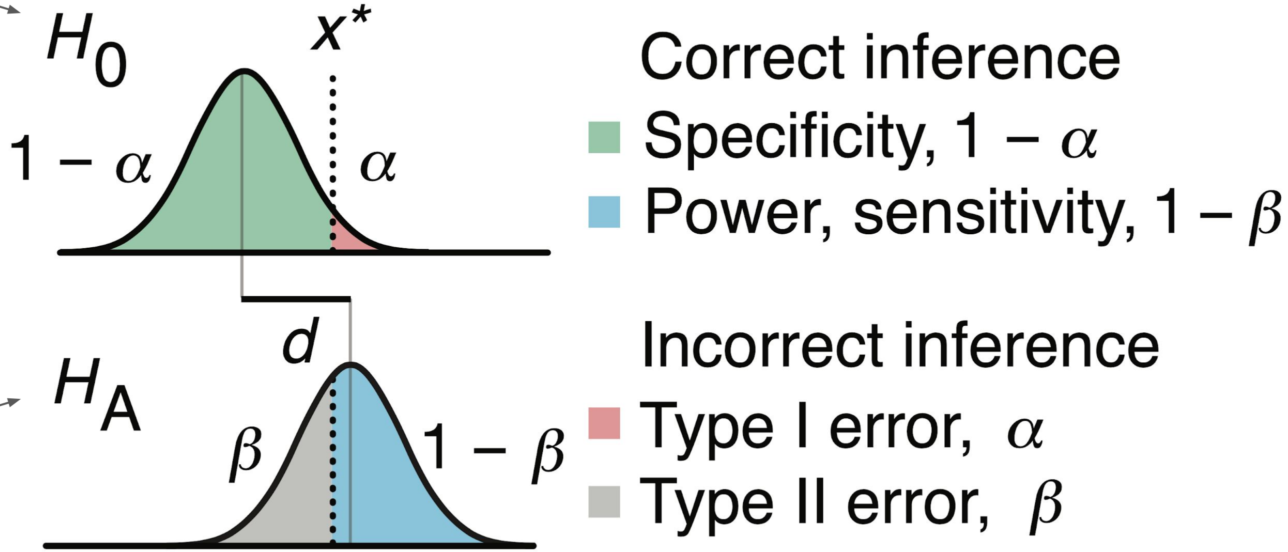
Power = True positive rate = Sensitivity = Recall



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

Statistical power

Null hypothesis



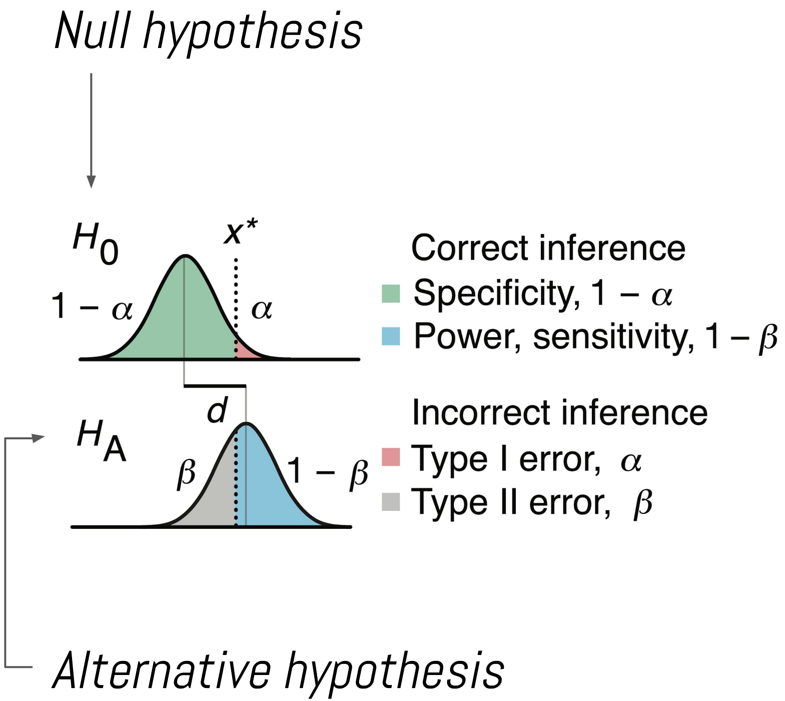
Alternative hypothesis

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.



Null hypothesis

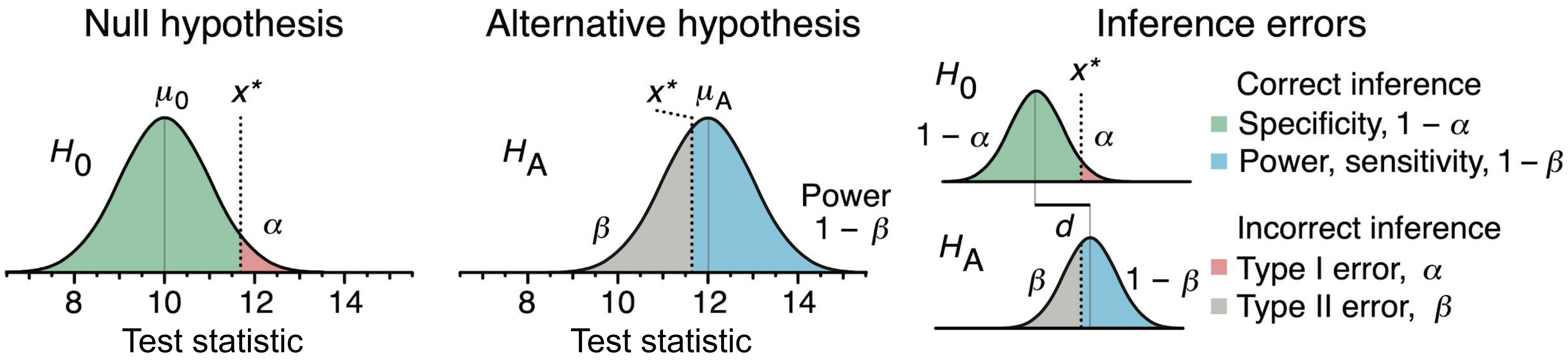
Alternative hypothesis

Power = 1 – β

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

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

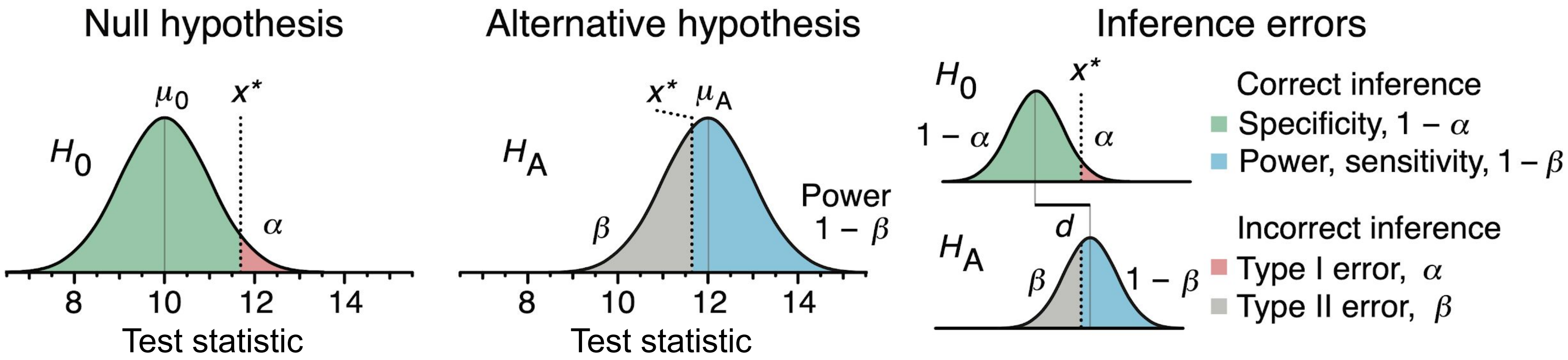
Statistical power



- 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

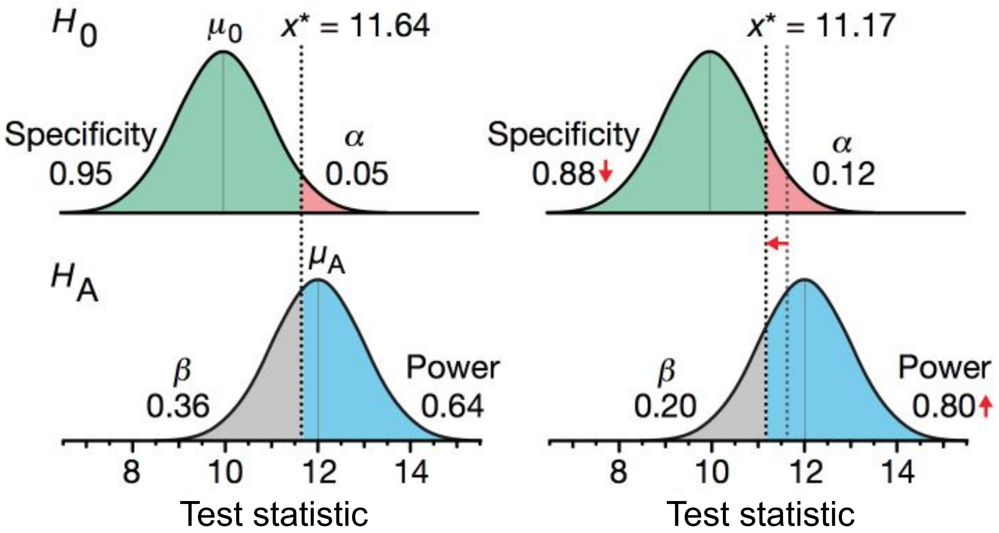
Statistical power

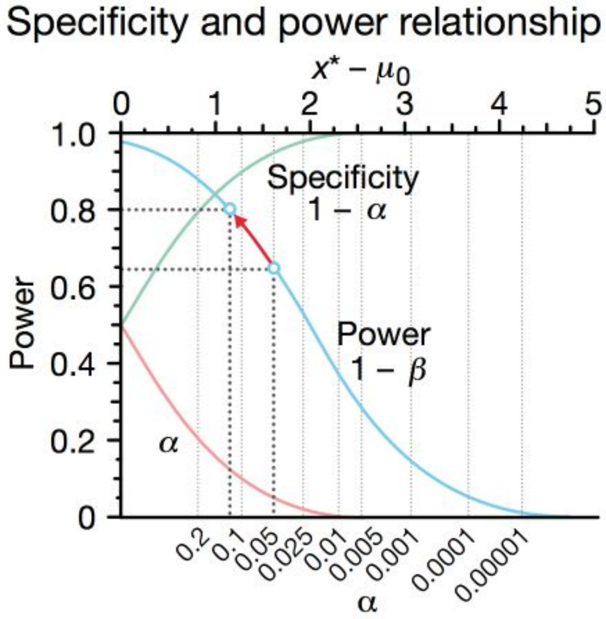


- Typically, α < β: consequences of FP (in an extreme case, a retracted paper) are more serious than those of FN (a missed opportunity to publish).
- But, the balance between α and β depends on the objectives:
  - If FP are subject to another round of testing but FN are discarded, β should be kept low.

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

Compromise between specificity and power



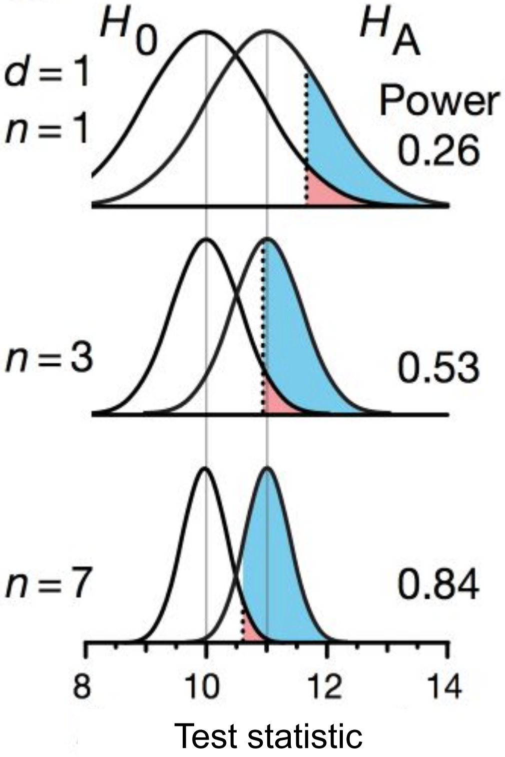


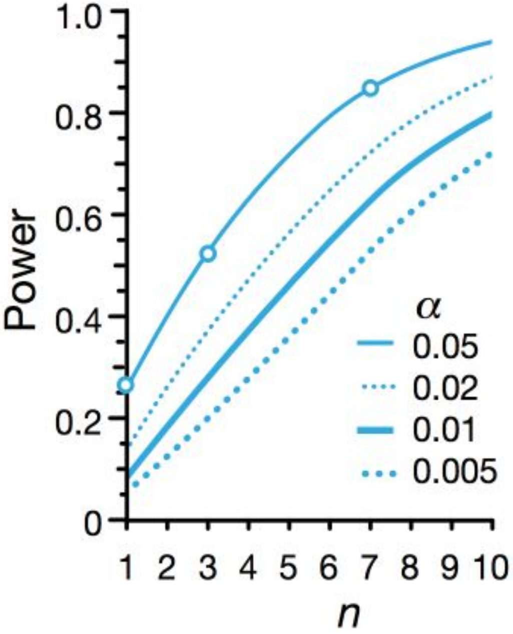
Specificity and power relationship

- 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)

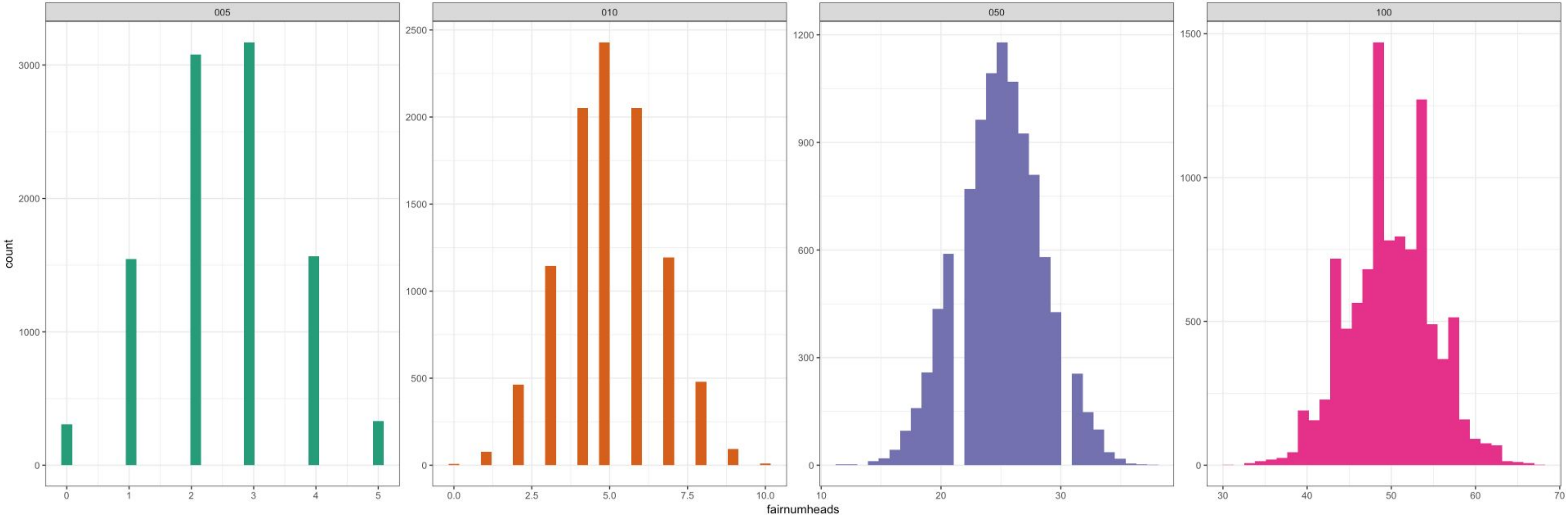
- Measurement error: counting cells vs. estimating level of fatigue/depression
- Experimental design: e.g. in a two-sample setting, optimal to have equal number

Generating a power curve

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

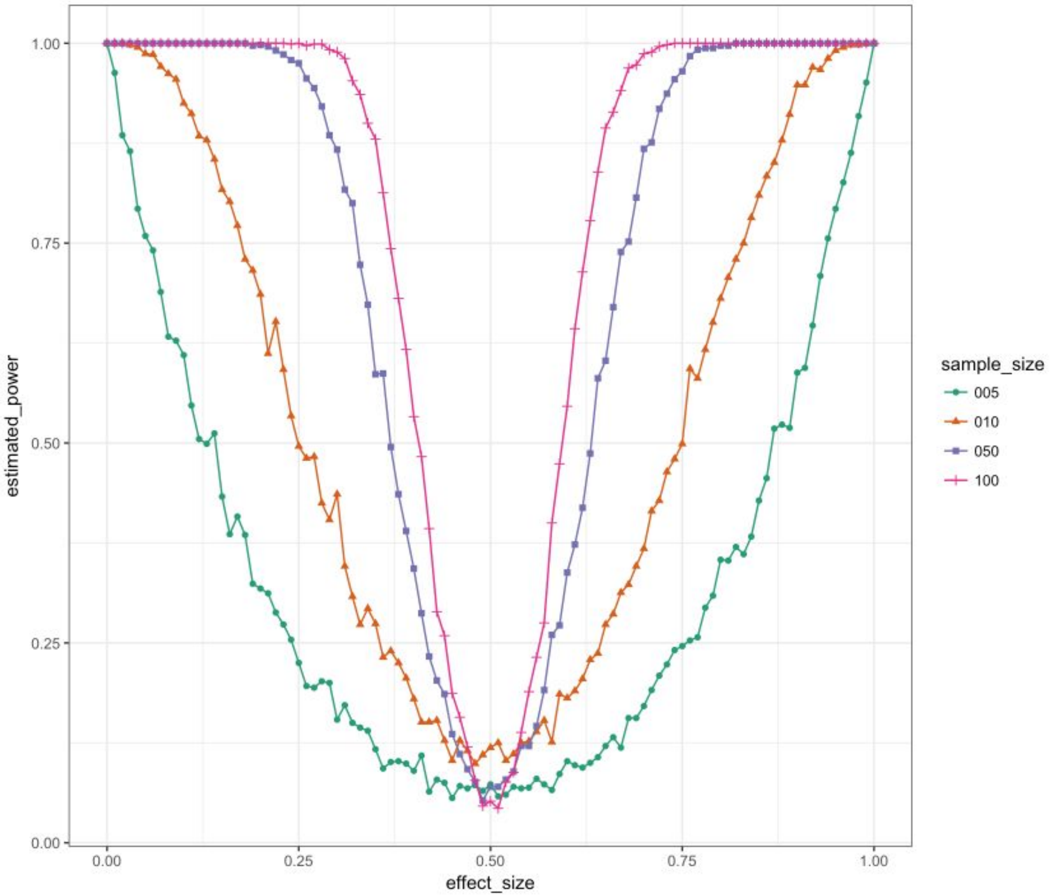
- You are given a coin and asked to detect if it is biased.
- Experiment: Flip the coin **num_flips** and take a call.
- Establish the null hypothesis (**num_permutations** = 10,000) for a given **num_flips** (this is the **sample size**).
- For a given **bias** (i.e. the **effect size**), find out how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

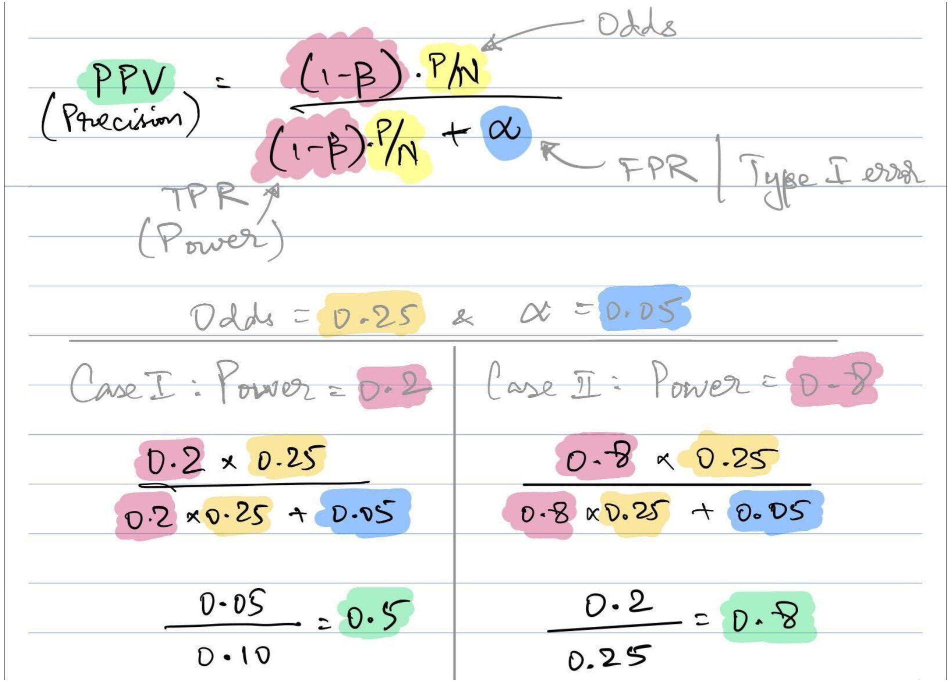
Underpowered studies

- Undermines the purpose of scientific research because it reduces the chance of detecting a true effect.
  - By definition, lower power means that the chance of discovering effects that are genuinely true is low.
  - Low-powered studies produce more FN than high-powered studies.
  - When studies in a given field are designed with a power of 20%, it means that if there are 100 genuine non-null (i.e. real) effects to be discovered in that field, these studies are expected to discover only 20 of them.

http://www.nature.com/articles/nrn3475

Underpowered studies

- Reduces the likelihood that a statistically significant result reflects a true effect.
  - I.e. low positive predictive value (PPV) when an effect is claimed.



$$\underset{\text{(Precision)}}{\text{PPV}} = \frac{(1-\beta)\cdot \overset{\text{Odds}}{P/N}}{\underbrace{(1-\beta)}_{\text{TPR (Power)}}\cdot P/N + \underbrace{\alpha}_{\text{FPR | Type I error}}}$$

$$\text{Odds} = 0.25 \quad \& \quad \alpha = 0.05$$

Case I : Power = 0.2

$$\frac{0.2 \times 0.25}{0.2 \times 0.25 + 0.05}$$

$$\frac{0.05}{0.10} = 0.5$$

Case II : Power = 0.8

$$\frac{0.8 \times 0.25}{0.8 \times 0.25 + 0.05}$$

$$\frac{0.2}{0.25} = 0.8$$

http://www.nature.com/articles/nrn3475

Underpowered studies

- Can lead to an exaggerated estimate of the magnitude of the effect when a true effect is discovered.
  - **Winner's curse** (likely to occur whenever claims of discovery are based on thresholds of statistical significance (for example, $p < 0.05$) or other selection filters).
  - Effect inflation is worst for small, low-powered studies, which can only detect effects that happen to be large.
  - E.g. if the true effect is medium-sized, only those small studies that, by chance, overestimate the magnitude of the effect will pass the threshold for discovery.

http://www.nature.com/articles/nrn3475

Underpowered studies

- Can lead to an exaggerated estimate of the magnitude of the effect when a true effect is discovered.

Scenario:

- An association truly exists with an effect size of ~1.20.
- We're trying to discover it by performing a small study with power of, say, ~20%.

The results of any study are subject to sampling variation and random error in the measurements of the variables and outcomes of interest.

- The study may find an effect <1.20 (e.g., **1.00**) or >**1.20** (e.g., **1.60**).
- Effect of 1.00 or 1.20 will not reach stat. significance because of the small sample size.
- Nominally significant association only when random error creates an effect of 1.60.

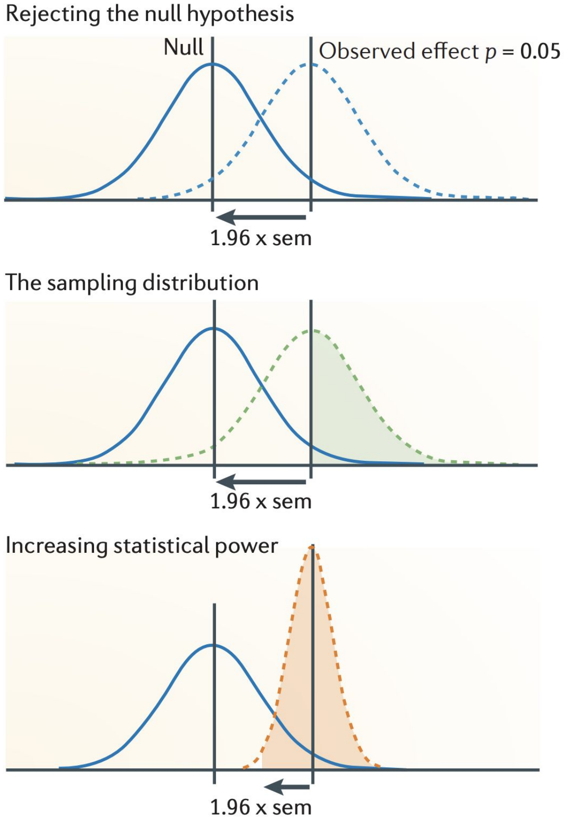
Winner's curse: The 'lucky' scientist who makes the discovery in a small study is cursed by finding an inflated effect.

http://www.nature.com/articles/nrn3475

Underpowered studies

- Hamper replicating research findings.
  - If the original estimate of the effect is inflated, then replication studies will tend to show smaller effect sizes as findings *converge* on the true effect.
  - More replication studies → eventually arrive at the more accurate effect size, but this may take time or may never happen if we only perform small studies.

Common misconception: A replication study will have sufficient power to replicate an initial finding if the sample size is similar to that in the original study.



http://www.nature.com/articles/nrn3475

Underpowered studies

- Has ethical dimensions:
    - Unreliable research is inefficient and wasteful.
        - Even a study that achieves only 80% power still presents a 20% possibility that the animals have been sacrificed without the study detecting the underlying true effect.
        - If the average power of studies is 20–30%, the ethical implications are substantial.
    - Continue data collection once it is clear that the effect being sought does not exist or is too small to be of interest.
        - Studies are not just wasteful when they stop too early, they are also wasteful when they stop too late.

http://www.nature.com/articles/nrn3475

Recommendations

- Perform an a priori power calculation
  - Use existing literature to estimate the size of effect and design your study accordingly.
  - If time or financial constraints mean your study is underpowered, make this clear and acknowledge this limitation (or limitations) in the interpretation of your results.

- Disclose methods and findings transparently
  - If the intended analyses produce null findings and you move on to explore your data in other ways, say so.
  - Null findings locked in file drawers bias the literature, whereas exploratory analyses are only useful and valid if you acknowledge the caveats and limitations.

http://www.nature.com/articles/nrn3475

Recommendations

- Pre-register your study protocol and analysis plan
  - Pre-registration clarifies whether analyses are confirmatory or exploratory, encourages well-powered studies and reduces opportunities for non-transparent data mining and selective reporting.

- Work collaboratively to increase power and replicate findings
  - Combining data increases the total sample size (and therefore power) while minimizing the labour and resource impact on any one contributor. Large-scale collaborative consortia in fields such as human genetic epidemiology have transformed the reliability of findings in these fields.

http://www.nature.com/articles/nrn3475

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/