Experiments with more than one factor:

A factor is a single treatment variable

Why experiment with > 1 factor?

- Efficiency
- Investigate possible **interactions** between factors

Factorial Design:

- **Every** combination of treatments is investigated

What if you can't do experiments?

- Observational studies can have many features of experimental studies except randomization

- But **randomization** is arguably the most important benefit of experiments

- You can still minimize bias even without randomization by using **matching** and **adjustment**

Usually four major (related) questions:

1. **What is the effect size you are looking for?**
   - Larger sample to find small effect
2. **What is the significance level?**
   - Of you want to be more strict (smaller significance), you will need larger sample size
3. **How much variability?**
   - If you have reason to believe that your data is very noisy, you'll need a larger sample size
4. **How much power?**
   - If there really is an effect of a specified magnitude in the overall population, how sure do you want to be that you will find it?

You will often see statements such as the following:

"We chose to study 321 subjects in each group in order to have 80% power to detect a 33% reduction in the recurrence rate from a baseline rate of 30% with a significance level of 0.05"

If I use n subjects, what information can I learn?

Sample Size: *a minimum n to determine the effect of the treatment (each treatment will need to be this size)*

Plan for precision:

- Predetermined Confidence interval
- For 95% confidence interval:
  - As narrow as possible (usually means a larger sample size)

$$\bar{Y}_1 - \bar{Y}_2 \pm uncertain$$

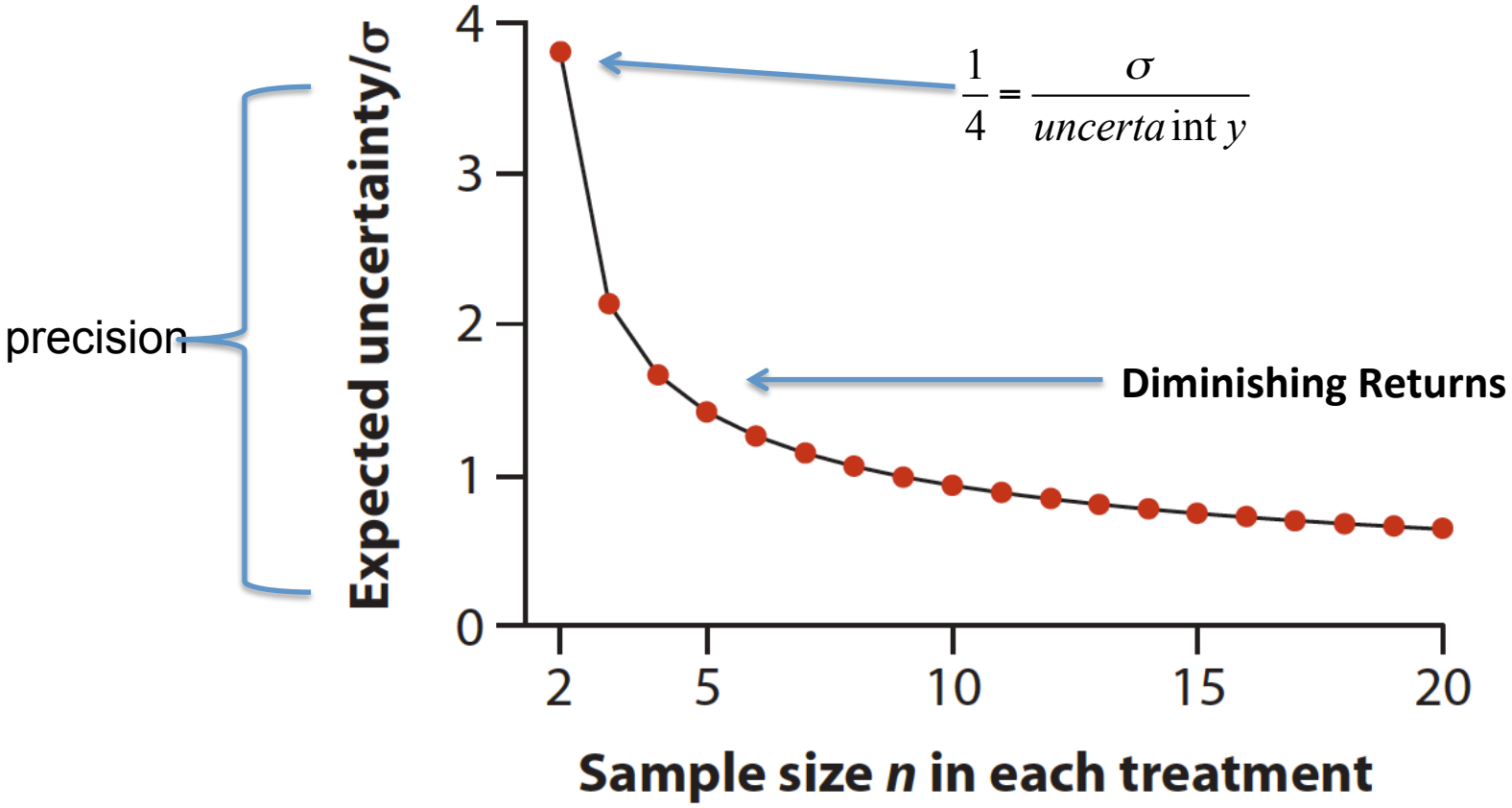
1/2 width of CI

$$n = 8\left(\frac{\sigma}{uncertain}\right)^2$$

Sample Size: *a minimum n to determine the effect of the treatment*

Plan for precision:

- Predetermined Confidence interval
  - As narrow as possible



Sample Size: *a minimum n to determine the effect of the treatment*

Plan for Power:

- Predetermined type I and type II error rates
  - Choosing a sample size that would have a high probability of rejecting if the true difference between the means was at least at great as the specified value D (D is just a minimum value)
- P(reject|H_0 is false) = 0.80

$$n \approx 16\left(\frac{\sigma}{D}\right)^2$$

- P(reject|H_0 is true) = 0.05

Sample Size: *a minimum n to determine the effect of the treatment*

Plan for the worst case scenario:

- Data loss
- Start with more than you need to accommodate individuals who begin the study but who do not end the study

Mistaking the question being asked:

What is the question? Jeffery T.Leek and Roger D. Peng

Science 347**,** 1314 (2015);

DOI: 10.1126/science.aaa6146

P-values are widely mis-interpretated and mis-applied to multiple hypotheses:

https://www.sciencenews.org/article/odds-are-its-wrong

P-value ban from a journal:

https://www.sciencenews.org/blog/context/p-value-ban-small-step-journal-giant-leap-science

What is p-hacking?

Excellent example found here:

http://fivethirtyeight.com/features/science-isnt-broken/

* Remember that there is (somewhat unconscious) bias present in many experiments: Collect data until I get a statistically significant result. If initial results aren't statistically significant, keep collecting the data.
← **Ad Hoc Sequential Sample Size Determination (SSSD)**

Data Dredging:

When you use multiple tests on a data set, the probability of making ***at least one*** *type I error, α, is larger than the significance level suggests* - *each hypothesis test has some probability of error and these errors compound as more tests are conducted*

$P(\text{No type I errors}) = (1-\alpha)^N$, *where N = independent tests*

$P(\geq 1 \text{ type I error}) = 1 - (1-\alpha)^N$

Example:

P(No type I errors)= $(1-0.05)^{100} \approx 0.006$

P(≥ 1 I errors)=(1-0.006)=0.994

Bonferroni

- If your goal is to determine which variable really did respond to treatment rather than just preliminary data exploration
- Only reject H_0 if the **P** value is less than the bonferroni-adjusted α^*
- P(≥ 1 Type I error) = α
- lose power - penalized for ‘asking too many questions of the data”

$$\alpha^* = \frac{\alpha}{Num\_tests}$$

Data Dredging:

α	**Test Number**	**P(No type I errors)**	α^*
0.05	10	0.60	0.005
	100	0.006	0.0005
0.01	10	0.90	0.001
	100	0.37	0.0001

Clinical test can have 10+ tests
Gene location or association (GWAS) can have 100+ tests

False discovery Rate

- Carry out multiple tests at fixed significance level
- Gather all tests that yield statistically significant result (discoveries)
- Estimate the false discovery proportion from the total discoveries

$P(\text{Reject}|H_0 \text{ true})/\{P(\text{reject}|\ H_0 \text{ true}) + P(\text{reject}|\ H_A \text{ true})\}$