S&DS 242/542: Theory of Statistics

Lecture 9: Effect size, power, and experimental design

Steps of a scientific study

A typical scientific study might consist of the following steps:

1. Identify and formulate a question of interest
2. Design an experiment to produce or collect data that addresses this question
3. Visualize and perform exploratory analysis of the data
4. Apply an inferential statistical procedure to answer the question of interest

Our focus in this course is mostly on Step 4.

Today we'll discuss some aspects of Step 2 — the design of the experiment — in the context of hypothesis testing.

Main questions for today

- Can we predict in advance whether the study will be able to identify an effect of interest?
- Can we predict the size of the study that would be needed to identify this effect?
- How can our experimental design potentially influence our ability to identify this effect?

Case study: Stanford peer grading experiment

- Context: Grading homework assignments in large classes is time-consuming and costly. It may even be infeasible in Massive Open Online Courses (MOOCs) with thousands or tens of thousands of students.
- An approach that has been suggested is *peer grading*: have students grade each other's assignments.
- Adopted by the Owasso School District in Tulsa County, OK in 2001. Challenged as a violation of student privacy, and brought to the U.S. Supreme Court in 2002.

"Correcting a classmate's work can be as much a part of the assignment as taking the test itself. It is a way to teach material again in a new context, and it helps show students how to assist and respect fellow pupils."

—Anthony Kennedy, *Owasso v. Falvo 2002*

Case study: Stanford peer grading experiment

- Question of interest: In addition to saving cost, does peer grading actually increase student learning?
- In 2014–2015, a statistics course at Stanford University conducted an experiment to answer this question: Divide 300 students in one year of the course into "peer-grading" and "control" groups, and compare the difference in learning between the two groups as measured by exam scores.
- Can we predict, *before* doing the experiment, whether we will discover a significant difference in learning? Can we determine in advance the number of students needed for the study?

Predicting the power

In a hypothesis testing context, "discovering the effect" usually means rejecting H_0 at a desired level of significance.

These questions then pertain to our anticipated *power*: Under an alternative H_1 that we believe may be true (peer grading improves test scores on average by 20%), what is our probability of rejecting H_0 (that peer grading has no effect on test scores)?

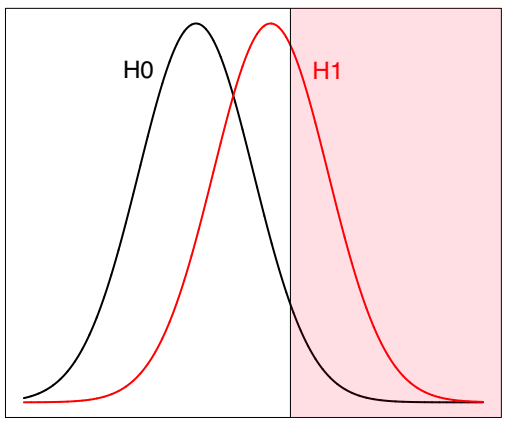
What is the sample size of the study that we would need to make this probability larger than, say, 90%?

Power in the one-sample z-test

Suppose we observe $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2)$, and test

$$H_0 : \mu = 0 \qquad \text{vs.} \qquad H_1 : \mu > 0$$

using the z-statistic $Z = \frac{\sqrt{n}}{\sigma}\bar{X}$. (By the Neyman-Pearson lemma, assuming σ^2 is known, this is the most powerful test.)



In this depiction of the sampling distributions of Z, the *power* is the probability of the rejection region under the red H_1 curve.

Power in the one-sample z-test

The z-test rejects H_0 when $Z = \frac{\sqrt{n}}{\sigma}\bar{X} > z^{(\alpha)}$. This ensures

$$\mathbb{P}[\text{Type I error}] = \alpha$$

To compute its power:

Under H_1: $\bar{X} = \frac{X_1 + \dots + X_n}{n} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

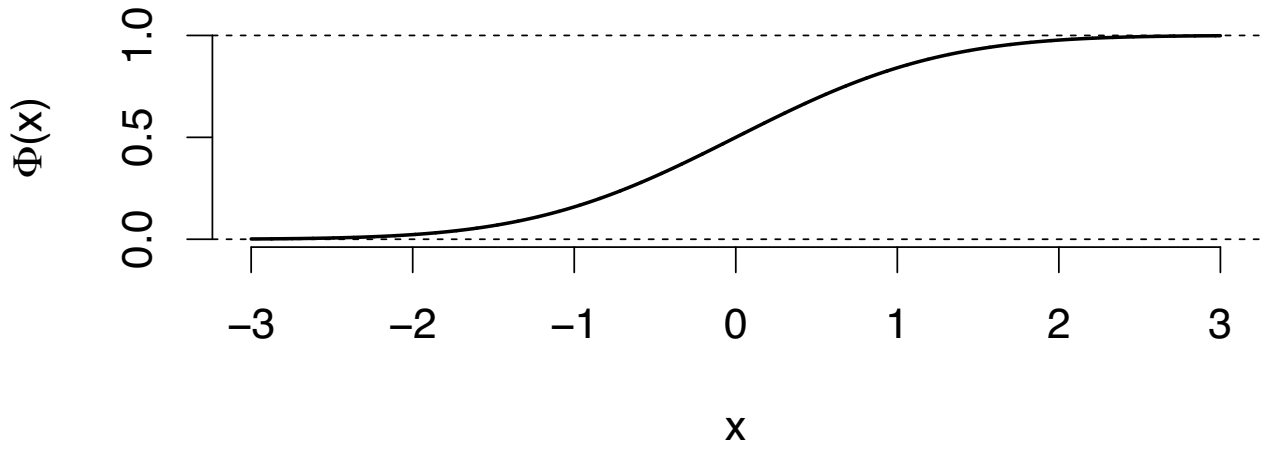
$\Rightarrow Z = \frac{\sqrt{n}}{\sigma}\bar{X} \sim \mathcal{N}(\underbrace{\frac{\sqrt{n}\mu}{\sigma}}_{:=\Delta}, 1)$

$$\text{Power} = \mathbb{P}_{H_1}[Z > z^{(\alpha)}] = \mathbb{P}_{H_1}[\underbrace{Z - \Delta}_{\sim \mathcal{N}(0,1) \text{ under } H_1} > z^{(\alpha)} - \Delta]$$

$$= 1 - \Phi(z^{(\alpha)} - \Delta) = \Phi(\Delta - z^{(\alpha)})$$

(Φ: CDF of $\mathcal{N}(0,1)$)

Factors that determine power



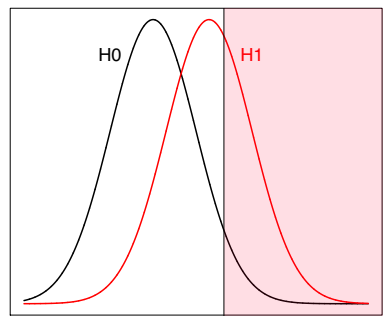
$$\text{Power} = \Phi\left(\sqrt{n}\,\frac{\mu}{\sigma} - z^{(\alpha)}\right)$$

This is influenced by:

- The sample size n. Larger n gives more power, and power increases to 1 as $n \to \infty$.
- The significance level α. A less stringent test (larger α) corresponds to smaller $z^{(\alpha)}$, and hence more power.
- The **effect size** μ/σ: difference in mean of the data between H_0 and H_1, scaled by the noise standard deviation.

Factors that determine power

Power also depends on the choice of test and test statistic: Less powerful test statistics have more overlap between their sampling distributions under H_0 and H_1.



For example, Homework 5 asks you to show for a nonparametric *sign test* that

$$\text{Power} \approx \Phi\left(\sqrt{\frac{2n}{\pi}}\,\frac{\mu}{\sigma} - z^{(\alpha)}\right)$$

This is smaller than the power of the z-test, as $\sqrt{\frac{2}{\pi}} \approx 0.798 < 1$.

Power in comparing two samples

For designs with a control group, as in the peer-grading study:
Consider $X_1, \ldots, X_n \sim \mathcal{N}(\mu_X, \sigma^2)$, $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_Y, \sigma^2)$, and

$$H_0 : \mu_X = \mu_Y \qquad \text{vs.} \qquad H_1 : \mu_X > \mu_Y$$

The power of the two-sample z-test based on $Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$ is:

Under H_0: $Z \sim \mathcal{N}(0,1)$

Under H_1: $\bar{X} \sim \mathcal{N}(\mu_X, \frac{\sigma^2}{n})$, $\bar{Y} \sim \mathcal{N}(\mu_Y, \frac{\sigma^2}{m})$

$$\Rightarrow \bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_X - \mu_Y, \sigma^2\left(\tfrac{1}{n} + \tfrac{1}{m}\right)\right)$$

$$\Rightarrow Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \mathcal{N}\Big(\underbrace{\frac{1}{\sqrt{\frac{1}{n} + \frac{1}{m}}} \cdot \frac{\mu_X - \mu_Y}{\sigma}}_{:= \Delta}, 1\Big)$$

By same calculation as (single case, $\mathbb{P}_{H_1}[Z > z^{(\alpha)}] = \Phi(\Delta - z^{(\alpha)})$

Power in comparing two samples

$$\text{Power} = \Phi\left(\underbrace{\frac{1}{\sqrt{\frac{1}{n} + \frac{1}{m}}} \cdot \frac{\mu_X - \mu_Y}{\sigma}}_{\text{call this } \Delta} - z^{(\alpha)} \right)$$

This again depends on:

- The sample sizes n and m.
- The significance level α.
- The **effect size** $(\mu_X - \mu_Y)/\sigma$: the difference in mean between the two groups, divided by the common standard deviation.

Optimally splitting the sample

Suppose our budget is determined by the total sample size $N = n + m$. How should we split this between the experiment and control groups?

The power is increasing in

$$\Delta = \frac{1}{\sqrt{\frac{1}{n} + \frac{1}{m}}} \cdot \frac{\mu_X - \mu_Y}{\sigma}$$

and hence decreasing in $\frac{1}{n} + \frac{1}{m}$.

Subject to $N = n + m$, we may check that $\frac{1}{n} + \frac{1}{m}$ is minimized by setting $n = m = \frac{N}{2}$. So to maximize our power, we should take the treatment and control groups to be of equal size.

Predicting the power

If we know the effect size, then we can predict the power before doing the study. In the Stanford peer-grading study, the effect size identified was (in retrospect) 0.11: peer grading improved the mean student test score by about $\frac{1}{9}^{\text{th}}$ of a standard deviation.

Typically we wouldn't know this effect size a priori, but we may have a guess based on previous studies. The 2015 Hattie ranking lists effect sizes for 195 different educational interventions, e.g.:

Classroom discussion: 0.82
Computer assisted instruction: 0.45
Teacher education: 0.12
Charter schools: 0.07

In education, effect sizes around 0.1 are typical, and effect sizes larger than 0.4 are considered very strong.

Predicting the power

Suppose we have $n + m = 300$ total students, and divide them equally into $n = m = 150$ students per group. For an effect size of 0.11 and a two-sample z-test at significance level $\alpha = 0.05$,

$$\Delta = \frac{0.11}{\sqrt{\frac{1}{150} + \frac{1}{150}}} \approx 0.95, \quad \text{Power} = \Phi\left(\Delta - z^{(\alpha)}\right) \approx 0.244.$$

Had we done this experiment, a z-test at level $\alpha = 0.05$ would have a 24% chance of identifying the peer grading effect.

Predicting a typical p-value

Instead of fixing the significance level α, we can also ask: What would be a "typical" p-value of the test? For a one-sided z-test that rejects H_0 for large values of Z, the p-value is

$$P = 1 - \Phi(Z).$$

Under H_1, we showed

$$Z \sim \mathcal{N}(\Delta, 1)$$

So the median value for Z under H_1 is Δ, and the median p-value is $1 - \Phi(\Delta)$. For $\Delta = 0.95$, this median p-value is 0.17.

Both these calculations indicate that this test is *underpowered*: The effect size is too small to be reliably detected with a sample size of only 300 students.

How many samples?

Suppose we would like the power to be much larger, say 90%, under a z-test at level $\alpha = 0.05$. We can achieve this by increasing the sample size. How many students would we need in the study?

Assuming an equal split of $m = n$ students in each group, set

$$0.9 = \Phi\left(\frac{0.11}{\sqrt{\frac{1}{n} + \frac{1}{n}}} - z^{(0.05)}\right)$$

and solve for n: We get $n \approx 1416$.

So we would need $2n \approx 2832$ total students. At 300 students per year, this requires running the study over a period of 9–10 years.

Changing the experiment to improve power

Main problem: There is too much variation in scores across students (size of σ), as compared to the mean improvement from peer-grading (size of $\mu_X - \mu_Y$).

This variation may be caused by a number of *confounding factors*: Class year, previous statistics courses, different learning styles, etc.

Differences between students that are attributable to these factors may overwhelm the mean difference arising from peer-grading, leading to the small effect size $(\mu_X - \mu_Y)/\sigma$.

Idea: Change the experimental design to compare students to themselves. This is an example of a **paired design**.

A paired design to improve power

Divide the course into 2 units[1], with a separate quiz at the end of each unit. Randomly assign each student to the peer-grading group for one unit, and the control group for the other unit.

Then compare the performance of each student in the peer-grading unit with *his/her own performance* in the control unit. Many confounding factors that affect the student for one unit are likely to also affect that same student for the other unit.

[1] The real Stanford study used 4 units instead of 2.

Testing and power in the paired design

Why does this help, and how much does it help by?

Suppose there are n students. Let $X_1, \ldots, X_n$ be their quiz scores in the peer-grading unit, and $Y_1, \ldots, Y_n$ their scores in the control unit. Assume $X_i \sim \mathcal{N}(\mu_X, \sigma^2)$ and $Y_i \sim \mathcal{N}(\mu_Y, \sigma^2)$, as before.

However, since X_i and Y_i now correspond to the *same* student, they are likely very correlated. Let's consider a model where (X_i, Y_i) are IID bivariate normal pairs, with correlation ρ:

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix} \right)$$

This is in contrast to the previous unpaired setting, where it was reasonable to model the X_i's and Y_j's as independent because they corresponded to different students.

Testing and power in the paired design

To test

$$H_0 : \mu_X = \mu_Y \qquad \text{vs} \qquad H_1 : \mu_X > \mu_Y$$

consider the **paired differences** $D_i = X_i - Y_i$.

If (X_i, Y_i) is bivariate normal, then $D_i = X_i - Y_i$ is normally distributed, with mean:

$$\mathbb{E}D_i = \mathbb{E}X_i - \mathbb{E}Y_i = \mu_X - \mu_Y$$

and variance:

$$\begin{aligned} \mathrm{Var}[D_i] &= \mathrm{Var}[X_i - Y_i] \\ &= \mathrm{Cov}[X_i - Y_i,\ X_i - Y_i] \\ &= \mathrm{Cov}[X_i, X_i] + \mathrm{Cov}[Y_i, Y_i] - 2\mathrm{Cov}[X_i, Y_i] \\ &= \sigma^2 + \sigma^2 - 2\rho\sigma^2 = 2\sigma^2(1-\rho) \end{aligned}$$

Testing and power in the paired design

Thus, $D_1, \ldots, D_n \overset{IID}{\sim} \mathcal{N}(\mu_X - \mu_Y, 2\sigma^2(1-\rho))$, and we wish to test

$$H_0 : \mu_X - \mu_Y = 0 \qquad \text{vs} \qquad H_1 : \mu_X - \mu_Y > 0$$

This reduces to a one-sample testing problem, and we may perform our test using the one-sample z-statistic[2] $Z = \frac{\sqrt{n}}{\sqrt{2\sigma^2(1-\rho)}}\bar{D}$.

Applying our previous result for the power of the one-sample z-test with $\mu_X - \mu_Y$ in place of μ and $\sqrt{2\sigma^2(1-\rho)}$ in place of σ,

$$\begin{aligned}\text{Power} &= \Phi\left(\sqrt{n} \cdot \frac{\mu_X - \mu_Y}{\sqrt{2\sigma^2(1-\rho)}} - z^{(\alpha)}\right) \\ &= \Phi\left(\frac{1}{\sqrt{1-\rho}}\sqrt{\frac{n}{2}} \cdot \frac{\mu_X - \mu_Y}{\sigma} - z^{(\alpha)}\right)\end{aligned}$$

[2] In practice, we may not know σ or ρ and use instead a 1-sample t-test.

Power comparison

To summarize, the power of the paired 2-sample z-test is

$$\text{Power} = \Phi\left(\frac{1}{\sqrt{1-\rho}}\sqrt{\frac{n}{2}}\cdot\frac{\mu_X-\mu_Y}{\sigma} - z^{(\alpha)}\right)$$

The power of the unpaired two-sample z-test with $m = n$ is

$$\text{Power} = \Phi\left(\sqrt{\frac{n}{2}}\cdot\frac{\mu_X-\mu_Y}{\sigma} - z^{(\alpha)}\right)$$

The difference is this additional factor of $1/\sqrt{1-\rho}$, where ρ is the correlation between the two scores of the same student.

Power comparison

$$\text{Power} = \Phi\left(\frac{1}{\sqrt{1-\rho}}\sqrt{\frac{n}{2}}\cdot\frac{\mu_X-\mu_Y}{\sigma} - z^{(\alpha)}\right)$$

The paired test with $(1-\rho)n$ pairs has the same power as an unpaired test with n individuals per group.

Here $1-\rho$ is called the **relative efficiency** of the unpaired design to the paired design.

For example, if the correlation were $\rho = 0.9$, then the relative efficiency of the unpaired design to the paired design is 10%. This means that an unpaired design with n pairs and a paired design with only $10\% \times n$ pairs would yield the same testing power.

Examples of paired designs

- Before-and-after studies on the same subjects
- Twin studies
- Subject matching by covariates (For example: In a medical study, matching by age, weight, severity of condition, etc.)

Matching by covariates was also used in the Stanford peer-grading experiment: Each student was paired with the "most similar" other student based on previous statistics courses, class year, and several other possible confounding variables.

One student in each pair was randomly assigned to peer-grade in unit 1, and the other to peer-grade in unit 2.

Summary of peer-grading study

- The estimated (short-term) effect size was 0.11. This effect was found to be statistically significant (with p-value 0.002) using a study of only 300 students.
- To understand whether the effect persisted until the end of the course, a longer-term effect was assessed by having questions for both units on the final exam, and comparing the performance of each student between the questions for Unit 1 versus Unit 2. This estimated effect size was 0.12, and also found to be statistically significant (with p-value of 0.001).

Conclusion: Peer grading did improve student learning.

For more details, see: DL Sun, N Harris, G Walther, M Baiocchi, "Peer Assessment Enhances Student Learning: The Results of a Matched Randomized Crossover Experiment in a College Statistics Class," *PLoS One*, 10(12), 2015.