STATS 200: Introduction to Statistical Inference

Lecture 10: Experimental design

Steps of a statistical study

A “typical” statistical study might consist of the following steps:

1. Identify/formulate the question of interest
2. Design an experiment or study to collect data that addresses this question
3. Clean, visualize, and explore the data
4. Draw an inference from the data to answer the original question

Thus far, our focus has been on Step 4. (We discussed briefly ideas such as hanging histogram plots and QQ plots for Step 3.)

Today we’ll discuss some aspects of Step 2, in the context of two-sample hypothesis testing.

Main questions for today

- How can we eliminate or minimize the influence of confounding factors?

  (Many examples and case studies are discussed in Rice, Section 11.4.)
- How can we reason about the size of the study needed to identify an effect of interest?
- How can we design the experiment so as to maximize the chance of identifying this effect?

Case study: Peer grading experiment in STATS 60

- Context: Grading student homework assignments in large classes is time-consuming and costly, perhaps prohibitively so in Massive Open Online Courses (MOOCs) with thousands or tens of thousands of students.
- Possible solution: Have students grade each other (peer grading).
- Question of interest: Can peer grading actually increase student learning?

Case study: Peer grading experiment in STATS 60

- Context: Grading student homework assignments in large classes is time-consuming and costly, perhaps prohibitively so in Massive Open Online Courses (MOOCs) with thousands or tens of thousands of students.
- Possible solution: Have students grade each other (peer grading).
- Question of interest: Can peer grading actually increase student learning?

Justice Anthony Kennedy, in Supreme Court case *Owasso v. Valvo*: "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."

A simple design

Divide the STATS 60 students (300 over the course of two quarters) into two groups, “peer-grading” and “control”. Have only the students in the peer-grading group grade their peers, and compare learning (e.g. test scores) between the two groups at the end of the quarter.

A simple design

Divide the STATS 60 students (300 over the course of two quarters) into two groups, “peer-grading” and “control”. Have only the students in the peer-grading group grade their peers, and compare learning (e.g. test scores) between the two groups at the end of the quarter.

Problem: Student performance is influenced by many confounding factors—their class year, previous coursework and knowledge of statistics, etc.

Simple solution: Randomly assign students to the two groups, so that confounding factors tend to be balanced between the groups.

A simple design

For this design, we might use a two-sample t-test:

Let $X_1, \ldots, X_n$ be final exam scores of the peer-grading group, $Y_1, \ldots, Y_m$ those of the control group. Supposing that $X_1, \ldots, X_n \sim \mathcal{N}(\mu_X, \sigma^2)$, $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_Y, \sigma^2)$, test

$$H_0 : \mu_X = \mu_Y$$
$$H_1 : \mu_X > \mu_Y$$

using the two-sample T-statistic $T = \dfrac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$. S_p^2 is the pooled sample variance discussed in Lecture 8.

A simple design

For this design, we might use a two-sample t-test:

Let $X_1, \ldots, X_n$ be final exam scores of the peer-grading group, $Y_1, \ldots, Y_m$ those of the control group. Supposing that $X_1, \ldots, X_n \sim \mathcal{N}(\mu_X, \sigma^2)$, $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_Y, \sigma^2)$, test

$$\begin{aligned} H_0 &: \mu_X = \mu_Y \\ H_1 &: \mu_X > \mu_Y \end{aligned}$$

using the two-sample T-statistic $T = \dfrac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$. S_p^2 is the pooled sample variance discussed in Lecture 8.

What is the chance that we identify a significant effect (reject H_0)?

Calculating the power

Here $n + m = 300$ is fairly large, so we expect S_p^2 to be a very accurate estimate of σ^2. Let's assume for simplicity that we know σ^2, and perform the test using

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}.$$

Calculating the power

Here $n+m=300$ is fairly large, so we expect S_p^2 to be a very accurate estimate of σ^2. Let's assume for simplicity that we know σ^2, and perform the test using

$$Z = \frac{\bar{X}-\bar{Y}}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}.$$

Recall $\bar{X}-\bar{Y} \sim \mathcal{N}(\mu_X - \mu_Y, \sigma^2(\frac{1}{n}+\frac{1}{m}))$.

Under H_0, $Z \sim \mathcal{N}(0,1)$, so a one-sided test rejects when $Z > z(\alpha)$.

Calculating the power

Here $n + m = 300$ is fairly large, so we expect S_p^2 to be a very accurate estimate of σ^2. Let's assume for simplicity that we know σ^2, and perform the test using

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}.$$

Recall $\bar{X} - \bar{Y} \sim \mathcal{N}(\mu_X - \mu_Y, \sigma^2(\frac{1}{n} + \frac{1}{m}))$.

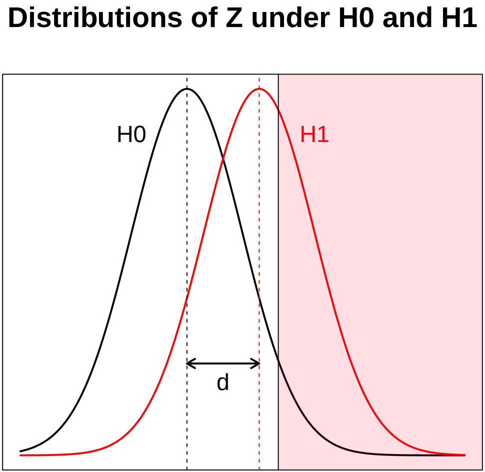
Under H_0, $Z \sim \mathcal{N}(0,1)$, so a one-sided test rejects when $Z > z(\alpha)$.

Under H_1, $Z \sim \mathcal{N}(d,1)$ where

$$d = \frac{\mu_X - \mu_Y}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}.$$

Calculating the power

The power of the test increases with d:



Calculating the power

The separation

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

is determined by:

- The real difference in mean test scores $\mu_X - \mu_Y$
- The standard deviation of test scores (i.e. “noise” level) σ
- The sample sizes n and m

The quantity $\frac{\mu_X - \mu_Y}{\sigma}$ is called the **effect size**—it measures the size of the mean difference in terms of the number of standard deviations of the noise.

Calculating the power

The power of the test is

$$\mathbb{P}_{H_1}[Z > z(\alpha)] = \mathbb{P}_{H_1}[Z - d > z(\alpha) - d] = 1 - \Phi(z(\alpha) - d),$$

where $\Phi(x)$ is the standard normal CDF, and we used the fact that $Z - d \sim \mathcal{N}(0, 1)$ under H_1.

Calculating the power

The power of the test is

$$\mathbb{P}_{H_1}[Z > z(\alpha)] = \mathbb{P}_{H_1}[Z - d > z(\alpha) - d] = 1 - \Phi(z(\alpha) - d),$$

where $\Phi(x)$ is the standard normal CDF, and we used the fact that $Z - d \sim \mathcal{N}(0,1)$ under H_1.

Subject to the constraint of $n + m = 300$ total students, d is maximized when we choose $n = m = 150$ students per group. The effect size identified by the study (in retrospect) was 0.11. So

$$d = \frac{\mu_X - \mu_Y}{\sigma}\sqrt{\frac{n}{2}} = 0.95.$$

At level $\alpha = 0.05$, the above power is only 0.244! In other words, had we done this experiment, we would have only had a 24% chance of rejecting H_0 at level $\alpha = 0.05$.

Typical *p*-value

We can also think in terms of the *p*-value we would have obtained. If the test statistic we observed were Z, then the *p*-value would be the upper tail probability

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

(P and Z here are both random, depending on the outcome of the experiment.)

Typical *p*-value

We can also think in terms of the *p*-value we would have obtained. If the test statistic we observed were Z, then the *p*-value would be the upper tail probability

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

(P and Z here are both random, depending on the outcome of the experiment.)

Under H_1, $Z \sim \mathcal{N}(d, 1)$, so the median value of Z is d. Since $x \mapsto 1 - \Phi(x)$ is monotone (decreasing), the median value of P is $1 - \Phi(d)$. So a "typical" *p*-value from this experiment would have been $1 - \Phi(d)$. For $d = 0.95$, this *p*-value is 0.17.

Both of these calculations indicate that the study would be **underpowered**—the effect size is too small to be detected with statistical significance, if the sample size is 300 students.

How many samples?

Suppose we would like the power to be much larger, say 0.9, under a level $\alpha = 0.05$ test. How many students would we need? If we have n students in each of the peer-grading and control groups, set

$$0.9 = 1 - \Phi(z(\alpha) - d) = 1 - \Phi\left(z(0.05) - 0.11\sqrt{\frac{n}{2}}\right)$$

and solve for n:

How many samples?

Suppose we would like the power to be much larger, say 0.9, under a level $\alpha = 0.05$ test. How many students would we need? If we have n students in each of the peer-grading and control groups, set

$$0.9 = 1 - \Phi(z(\alpha) - d) = 1 - \Phi\left(z(0.05) - 0.11\sqrt{\frac{n}{2}}\right)$$

and solve for n:

$$\begin{aligned}
&\Phi\left(z(0.05) - 0.11\sqrt{\frac{n}{2}}\right) = 0.1 \\
&\Rightarrow z(0.05) - 0.11\sqrt{\frac{n}{2}} = \Phi^{-1}(0.1) = -z(0.1) \\
&\Rightarrow n = 2\left(\frac{z(0.05) + z(0.1)}{0.11}\right)^2 \approx 1416
\end{aligned}$$

We would need $2n \approx 2832$ total students. This amounts to doing this experiment for 5–10 years of STATS 60...

Effect sizes in education

The previous calculations assumed we knew the effect size was 0.11. In reality, we don't know this ahead of time. However, we can compare to what we know:

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

These numbers are from the 2015 Hattie ranking, which lists effect sizes for 195 different educational influences/approaches, determined from aggregating previous experimental studies. In education, an effect size larger than 0.4 is typically considered strong.

A different design to improve power

Main problem: There is too much variation in student performance, compared to the size of the improvement from peer-grading.

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

A different design to improve power

Main problem: There is too much variation in student performance, compared to the size of the improvement from peer-grading.

Idea: Compare each student to himself/herself.

Implementation: Divide STATS 60 course into 2 units*, with a quiz at the end of each unit. Assign each student to do peer-grading for one unit, and no peer-grading for the other unit. (To handle the possible confounding factor that one exam is easier than the other, randomly choose which unit each student does peer-grading.)

In other words, set up an experiment with paired samples rather than two independent samples.

*The real study used 4 units instead of 2.

Calculating the power for paired samples

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.

Calculating the power for paired samples

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.

For this design, we might use a one-sample (a.k.a. paired two-sample) t-test:

Let $D_i = X_i - Y_i$, and reject H_0 for large values of the t-statistic

$$T = \frac{\sqrt{n}\bar{D}}{S}.$$

Here, $\bar{D}$ and S^2 are the sample mean and variance of the D_i's.

Calculating the power for paired samples

Assume $X_i \sim \mathcal{N}(\mu_X, \sigma^2)$ and $Y_i \sim \mathcal{N}(\mu_Y, \sigma^2)$, as before. Since X_i and Y_i correspond to the same student, they are likely very correlated. Let's suppose (X_i, Y_i) is bivariate normal with correlation ρ:

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

Calculating the power for paired samples

Assume $X_i \sim \mathcal{N}(\mu_X, \sigma^2)$ and $Y_i \sim \mathcal{N}(\mu_Y, \sigma^2)$, as before. Since X_i and Y_i correspond to the same student, they are likely very correlated. Let's suppose (X_i, Y_i) is bivariate normal with correlation ρ:

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

Then $D_i = X_i - Y_i$ is normally distributed, with mean $\mathbb{E}[D_i] = \mu_X - \mu_Y$ and variance

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

Calculating the power for paired samples

Since n is large, S^2 should be very close to $\text{Var}[D_i] = 2\sigma^2(1-\rho)$. Let's suppose again for simplicity that we know $2\sigma^2(1-\rho)$, and consider the test statistic

$$Z = \frac{\sqrt{n}\bar{D}}{\sqrt{2\sigma^2(1-\rho)}}.$$

Calculating the power for paired samples

Since n is large, S^2 should be very close to $\mathrm{Var}[D_i] = 2\sigma^2(1-\rho)$. Let's suppose again for simplicity that we know $2\sigma^2(1-\rho)$, and consider the test statistic

$$Z = \frac{\sqrt{n}\bar{D}}{\sqrt{2\sigma^2(1-\rho)}}.$$

We have $\bar{D} \sim \mathcal{N}(\mu_X - \mu_Y, \frac{2\sigma^2(1-\rho)}{n})$.

Under H_0, $Z \sim \mathcal{N}(0,1)$, so a level-α test rejects when $Z > z(\alpha)$.

Under H_1, $Z \sim \mathcal{N}(d,1)$, where

$$d = \frac{\mu_X - \mu_Y}{\sigma}\sqrt{\frac{n}{2(1-\rho)}}$$

Power comparison

$$d = \frac{\mu_X - \mu_Y}{\sigma} \sqrt{\frac{n}{2(1-\rho)}}$$

Compared to having two independent samples of size n (one peer-grading, one control), we gain a factor of $1/\sqrt{1-\rho}$ in d. You can think of this as either reducing the effective variance from σ^2 (in the case of unpaired samples) to $\sigma^2(1-\rho)$ (in the case of paired samples), or as increasing the effective sample size from n (in the case of unpaired samples) to $n/(1-\rho)$ (in the case of paired samples). The factor $1-\rho$ is called the **relative efficiency** of the unpaired design to the paired design.

E.g., if $\rho = 0.9$, then the relative efficiency is 0.1, and a paired design with n pairs yields the same power as an unpaired design with two independent samples of size $10n$.

Examples of paired designs

- Before-and-after studies on the same subjects
- Twin studies
- Subject matching by covariates (e.g., in a medical study, matching by age, weight, severity of condition, etc.)

Examples of paired designs

- Before-and-after studies on the same subjects
- Twin studies
- Subject matching by covariates (e.g., in a medical study, matching by age, weight, severity of condition, etc.)

Matching by covariates was also used in the STATS 60 experiment: Rather than randomly choosing, for each student, which unit they did peer-grading, each student was paired with the "most similar" other student based on gender, race, previous statistics background, class year, etc. using a matching algorithm. One student in each pair was then randomly assigned to peer-grade in unit 1, and the other to peer-grade in unit 2.

Pairing by covariates is a special case of a **randomized block design**, which groups subjects into blocks having similar characteristics.

Summary of STATS 60 study

- The estimated (short-term) effect size was 0.11. Despite the small size of the effect, it was found to be statistically significant with *p*-value 0.002.
- Long-term effect was assessed by comparing performance on the questions corresponding to each unit in the final exam; the estimated effect size was 0.12, with *p*-value 0.001.

Conclusion: Peer grading yielded a small but real improvement in student learning.

Details available at: 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.

Addendum: S^2 is close to σ^2 for large n

As $n \to \infty$, the sample variance $S^2 \to \sigma^2$ in probability. Why?

Addendum: S^2 is close to σ^2 for large n

As $n \to \infty$, the sample variance $S^2 \to \sigma^2$ in probability. Why?

Suppose $X_1, \ldots, X_n$ are IID with mean 0 and variance σ^2.

$$\begin{aligned} S^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \\ &= \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right) \\ &= \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \frac{n}{n-1} \bar{X}^2. \end{aligned}$$

Addendum: S^2 is close to σ^2 for large n

As $n \to \infty$, the sample variance $S^2 \to \sigma^2$ in probability. Why?

Suppose $X_1, \ldots, X_n$ are IID with mean 0 and variance σ^2.

$$\begin{aligned} S^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \\ &= \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right) \\ &= \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \frac{n}{n-1} \bar{X}^2. \end{aligned}$$

As $n \to \infty$, $\frac{n}{n-1} \to 1$. Also by the LLN, $\frac{1}{n} \sum_{i=1}^{n} X_i^2 \to \sigma^2$ and

$\bar{X} \to 0$ in probability.

Addendum: S^2 is close to σ^2 for large n

The functions $(x, y) \mapsto x - y$ and $(x, y) \mapsto xy$ are continuous. So if $X_n \to a$ and $Y_n \to b$ in probability, then the Continuous Mapping Theorem implies $X_n - Y_n \to a - b$ and $X_n Y_n \to ab$ in probability.

Addendum: S^2 is close to σ^2 for large n

The functions $(x, y) \mapsto x - y$ and $(x, y) \mapsto xy$ are continuous. So if $X_n \to a$ and $Y_n \to b$ in probability, then the Continuous Mapping Theorem implies $X_n - Y_n \to a - b$ and $X_n Y_n \to ab$ in probability.

Then

$$S^2 = \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \frac{n}{n-1} \bar{X}^2 \to 1 \cdot \sigma^2 - 1 \cdot 0 \cdot 0 = \sigma^2$$

in probability.

Addendum: S^2 is close to σ^2 for large n

The functions $(x,y) \mapsto x - y$ and $(x,y) \mapsto xy$ are continuous. So if $X_n \to a$ and $Y_n \to b$ in probability, then the Continuous Mapping Theorem implies $X_n - Y_n \to a - b$ and $X_n Y_n \to ab$ in probability.

Then

$$S^2 = \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \frac{n}{n-1} \bar{X}^2 \to 1 \cdot \sigma^2 - 1 \cdot 0 \cdot 0 = \sigma^2$$

in probability.

Clearly this also holds if $X_1, \ldots, X_n$ are IID with mean μ and variance σ^2, because S^2 doesn't depend on μ.

We didn't assume that the X_i's are normally distributed—this argument holds as long as $X_1, \ldots, X_n$ are IID with finite variance.