Physician's Health Study

An outline for analysis

The Data

One of the most famous oberver-blinded, placebo-controlled clinical trials is the Physicians' Health Study. This was a large prospective study that used physicians as subjects. Among the research questions was whether a low dose of aspirin could reduce the risk of heart attacks. Physicians were randomly divided to receive placebo or aspirin. The data are shown below.

	Heart attack	No heart attack	Total
Aspirin	104	10,933	11,037
Placebo	189	10,845	11,034
Total	293	21,778	22,071

The large numbers of patients was necessary in order to have a reasonably large expected number of heart attacks.

My Goal

. . . in introducing it today is to describe a model for these data which you can also use for your project since you will be analyzing similar data: the Pfizer vaccine trials.

. . . give you a framework for inference

Not my Goal

. . . however, I will not be analyzing the data since it is too similar to your project;

Data analysis: first step

The first step in any analysis is to make visualizations and calculate numerical summaries.

	Heart attack	No heart attack	Total
Aspirin	104	10,933	11,037
Placebo	189	10,845	11,034
Total	293	21,778	22,071

Visualizations?

- segmented barplot
- mosaicplot

Data analysis: first step

The first step in any analysis is to make visualizations and calculate numerical summaries.

	Heart attack	No heart attack	Total
Aspirin	104	10,933	11,037
Placebo	189	10,845	11,034
Total	293	21,778	22,071

Summary: Aspirin is clearly effective in reducing heart attacks: 0.94% ($\frac{104}{11,037}$) versus 1.71% ($\frac{189}{11,034}$). $1 - 0.55$

Of primary interest is aspirin efficacy - which is 45% - meaning it reduces the risk of a heart attack by 45% compared to placebo:

$$0.94\% = 0.55 \times 1.71\%$$

H_0: efficacy = 30% H_1: efficacy > 30%

The FDA requires at least 30% efficacy for a new therapy to be approved.

So it is of interest to construct confidence intervals and perform significance tests to see whether the 45% efficacy rate is statistically different from 30%.

Data Analysis: second step

The next stage is to consider a model for the data and then build out the plan for making inference on the parameter of interest.

I will present two ideas.

- The first model is a natural first thought.
- The second is a better idea and the one I want you to use.

Model 1: Two binomial random variables

Let X denote the number of heart attacks among the n_1 patients randomly assigned to the aspirin group and Y is the number of heart attacks among the n_2 patients assigned to placebo.

We assume

$$X \sim Binom(n_1, \pi_1)$$

independently of

$$Y \sim Binom(n_2, \pi_2).$$

π_1: P(heart attack in aspirin)

π_2: P(heart attack in placebo)

The parameter of interest is the aspirin efficacy $\longrightarrow$ $\pi_1 = a\,\pi_2 \Rightarrow a = \pi_1/\pi_2$

aspirin efficacy: $(1-a) = 1 - \frac{\pi_1}{\pi_2}$

$$\psi = 1 - \frac{\pi_1}{\pi_2} = \frac{\pi_2 - \pi_1}{\pi_2}.$$

Example

In a randomized clinical trial of the Johnson & Johnson vaccine, 49 out 5,000 volunteers assigned to the placebo group got COVID 19 compared to 17 per 5,000 in the vaccinated group.

What is the (estimated) efficacy of the Johnson & Johnson vaccine? Report the answer in plain English. 65%

π_1: P(covid 19 in vaccine)

π_2: P(covid 19 in placebo)

$$\psi = 1 - \frac{\pi_1}{\pi_2}$$

$$\hat{\pi}_1 = 17/5000$$

$$\hat{\pi}_2 = 49/5000$$

$$\hat{\psi} = 1 - \frac{\hat{\pi}_1}{\hat{\pi}_2} = 1 - \frac{17}{49} = 0.65$$

Model 1: Two binomial random variables

How should we carry out inference? We could of course treat the problem as one of comparing two binomial proportions.

$$X \sim Binom(n_1, \pi_1) \text{ independently of } Y \sim Binom(n_2, \pi_2).$$

Slightly tricky because parameter of interest is NOT $\pi_2 - \pi_1$. Instead, we are interested in $\psi = \frac{\pi_2 - \pi_1}{\pi_2}$. $(\pi_2 - \pi_1)/\pi_2$

A natural estimator is $\widehat{\psi} = \frac{\bar{Y} - \bar{X}}{\bar{Y}}$ where $\bar{X} = \frac{X}{n_1}$ and $\bar{Y} = \frac{Y}{n_2}$.

Standard error?

Two binomials to one

Very often, clinical trials are designed to run until a certain number of *events* are observed.

This was the case here: the study was designed to run until 293 heart attacks were observed.

So an alternate model that utilizes this information is preferable.

Model 2: single binomial random variable

Suppose we denote the random variable T as the number in the aspirin group from the $n = 293$ heart attack cases.

	Heart attack	No heart attack	Total
Aspirin	104	10,933	11,037
Placebo	189	10,845	11,034
Total	293	21,778	22,071

The distribution of T can be well approximated by a binomial distribution!

Model 2 - single binomial random variable

Let T denote the number in the aspirin group out of all the $n = 293$ heart attack cases.

Then[1]

$$T \sim Binom(n = 293, \pi)$$

where $\pi = P(\text{aspirin}|\text{heart attack})$ and can be shown to be equal to

$$\pi = \frac{n_1\pi_1}{n_1\pi_1 + n_2\pi_2}.$$

When $n_1 \approx n_2$ (as is the case here) we say the randomization is 1:1 and are able to simplify π further:

$$\pi = \frac{\pi_1}{\pi_1 + \pi_2}.$$

[1] You do not need to prove this. See proof at end of slides

Note: we can then write our parameter of interest $\psi = \frac{\pi_2 - \pi_1}{\pi_2}$ in terms of the binomial probability π as follows

$$\psi = \frac{1-2\pi}{1-\pi}.$$

$$= \frac{1 - 2(\pi_1/(\pi_1+\pi_2))}{1 - \pi_1/(\pi_1+\pi_2)} = \frac{\pi_1+\pi_2-2\pi_1}{\pi_1+\pi_2-\pi_1} = \frac{\pi_2-\pi_1}{\pi_2} = 1 - \frac{\pi_1}{\pi_2}$$

$$\psi = 1 - \frac{\pi_1}{\pi_2}$$

Model 2 - single binomial random variable

In a nutshell: Let T be the number from the aspirin group among the 293 heart attacks.

We assume

$$T \sim Binom(293, \pi)$$

We have observed $t = 104$ here.

The parameter of interest is $\psi = \frac{1-2\pi}{1-\pi}$.

Comparative statistical inference

The idea is to come up with (two) different ways to estimate ψ and make inference (confidence interval, P-value) for it.

Some possibilities are below:

- Likelihood inference
- Method of moments estimation
- Other? (anything you can come up with that is sound)

Likelihood inference

$$T \sim Binom(293, \pi)$$

We have observed $t = 104$ here.

The parameter of interest is $\psi = \frac{1-2\pi}{1-\pi}$.

- Write the likelihood function $L(\psi)$ (see problem set 6 from section)
- Maximize it to find the MLE of ψ_0, the true value which generated the data.
- Then use Theorem 25.1 to find the large sample confidence interval estimate for ψ_0 (check plot)
- Conduct a likelihood ratio test of $H_0 : \psi_0 = 0.3$ versus $H_1 : \psi_0 \neq 0.3$.

Method of moments

$$T \sim Binom(293, \pi)$$

We have observed $t = 104$ here.

The parameter of interest is $\psi = \frac{1-2\pi}{1-\pi}$. Note:

$$\pi = \frac{1-\psi}{2-\psi}$$

- M.O.M. estimator satisfies

$$E[T] = t$$
$$\rightarrow 293 \times \pi = t,$$
$$\rightarrow 293 \times \frac{1-\psi}{2-\psi} = t$$

Solve to get estimate/estimator of ψ_0.

For aspirin study: $\hat{\psi}_0 = \frac{293 - 2 \times 104}{293 - 104} = 0.44$

Method of moments

The M.O.M. estimator of ψ_0 will be of the form[2]:

$$\hat{\psi}_0^{mom} = \frac{n - 2T}{n - T}.$$

- You can construct confidence intervals for ψ_0 using the boostrap method. (See Chapter 19)
- You can use empirical P-values for calculating a P-value for testing $H_0 : \psi_0 = 0.3$ versus $H_1 : \psi_0 > 0.3$. (See Chapter 20)

[2] it is your responsibility to verify this to make sure it is correct

Project Info

- Please see the instructions file on the Hub for details of the data and paper you will be writing. It is due on Sunday June 2 in gradescope.
- You are expected to attend classes the remaining two weeks and use class time to work as a group. (If you need more time outside of class, please plan this with your group)
- Class schedule for remaining two weeks (need feedback):
  - Mon 5/20 :
  - Wed 5/22 : group work credit
  - Wed 5/29 :
  - Fri 5/31 : group work credit
- Questions directly related to work on the project (of any kind) should only be asked on Ed publicly.

Why is T binomial?

Recall T denotes the number in the aspirin group from the n heart attack cases.

The binomial distribution for T can be justified by the following argument in terms of the originally defined random variables X and Y.

Since the sample sizes - n_1 and n_2 - in each group are LARGE and the event rates are small, we can say (see chapter 8.2 on Poisson approximation to binomial)

$$X \approx Pois(n_1\pi_1) \qquad \text{independently} \qquad Y \approx Pois(n_2\pi_2)$$

Why is T binomial?

Now let us find the PMF of T.

$$\begin{aligned} P(T=t) &= P(X=t|X+Y=n) \qquad t=0,1,\ldots,n \\ &= \frac{P(X=t \cap X+Y=n)}{P(X+Y=n)}, \\ &= \frac{P(X=t \cap Y=n-t)}{P(X+Y=n)}, \\ &= \frac{P(X=t)\times P(Y=n-t)}{P(X+Y=n)} \qquad X, Y \text{ ind} \\ &= \binom{n}{t}\pi^t(1-\pi)^{n-t} \end{aligned} \tag{1}$$

where $\pi = \frac{n_1\pi_1}{n_1\pi_1+n_2\pi_2}$.