Intro to Econometrics

Statistical Inference

David Murakami

St Hilda's College;
Department of Economics,
University of Oxford

17th February 2022

Introduction

The key concepts of statistical inference we need to understand in econometrics (or in any statistics course) are:

- Probability model: A mathematical description of random phenomenon that mimics reality.
- Estimator: A function or procedure for estimating a parameter of a probability model using observed data.
- Sampling distribution: The probability distribution of the estimator when the population is repeatedly sampled.
- Estimate: The actual outcome of the estimator from a given sample – that is, the sample you have collected.

The Bernoulli Distribution

We start with a simple probability model for binary data: the Bernoulli distribution. If a random variable Y has a Bernoulli distribution with parameter $\pi(Y \sim Bern(\pi))$ then:

- Y takes binary values (i.e., 0 or 1)

$$Y = \begin{cases} 1, & \text{w.p. } \pi; \\ 0, & \text{w.p. } 1\text{-}\pi. \end{cases}$$

- Y has the probability mass function (pmf):

$$\Pr(Y = y) = \pi^y (1 - \pi)^{1-y}, \quad y \in \{0, 1\}.$$

- Only one parameter, π, describes the entire distribution.

Examples: Tossing a coin (head/tail), voting for a candidate (yes/no), and survival (dead/alive).

Moments of a Bernoulli Distribution

The first two moments of a Bernoulli distribution are:

- The mean, or expectation of Y, equals π because:

$$\begin{aligned}\mathbb{E}[Y] &= \sum_{y=0}^{1} \Pr(Y = y) y \\ &= \Pr(Y = 1) \times 1 + \Pr(Y = 0) \times 0 \\ &= \pi \times 1 + (1 - \pi) \times 0 \\ &= \pi.\end{aligned}$$

- The variance is given by:

$$\begin{aligned}\mathrm{Var}(Y) = \mathbb{E}[Y - \mathbb{E}[Y]]^2 = \mathbb{E}[Y^2 - 2Y\mathbb{E}[Y] + \mathbb{E}[Y]^2] &= \mathbb{E}[Y^2 - \mathbb{E}[Y]^2] \\ &= \sum_{y=0}^{1} \Pr(Y = y) y^2 - \pi^2 \\ &= \pi \times 1 + (1 - \pi) \times 0 - \pi^2 \\ &= \pi(1 - \pi).\end{aligned}$$

Estimator for the Bernoulli Model

- If we have n observations generated independently from a $Bern(\pi)$ random variable then:

$$\hat{\pi} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n} = \frac{\sum_{i=1}^{n} Y_i}{n},$$

  is an estimator of π.
- It is intuitively appealing – the observed proportion of the sample to be collected.

Statistical Inference for the Bernoulli Model

For large enough n (rule of thumb: $n\pi > 5$ and $n(1-\pi) > 5$):

$$\hat{\pi} \overset{d}{\sim} \mathcal{N}\left(\pi, \frac{\pi(1-\pi)}{n}\right).$$

- $\hat{\pi}$ is approximately Gaussian distributed;
- $\mathbb{E}[\hat{\pi}] = \pi$ ($\hat{\pi}$ is unbiased);
- $\text{Var}(\hat{\pi}) = \pi(1-\pi)/n \to 0$ as $n \to \infty$.

What the heck does all of that mean?

- If we repeatedly take samples, of size n, from a Bernoulli random variable, calculate the estimate for each sample, the collection of estimates will have a Gaussian distribution with mean π and variance $\pi(1-\pi)/n$.
- Simulation can help to explain this concept.
- Simulation: The reproduction of randomness by random number generators in computers.
- We will use R to do this and visualise the results – but feel free to have a go at this in your favourite programming language!

Statistical Inference for the Bernoulli Model

Here's what we're going to do:

- Create a true (made up) probability of success, π.
- Create the sample size, n.
- Check $n\pi > 5$ and $n(1-\pi) > 5$.
- Create the number of samples (large number).
- For each sample, simulate n observations of $Y_i \sim Bern(\pi)$ and calculate the mean.
- The collection of these means is an approximation of the sampling distribution.