Stat 200: Introduction to Statistical Inference **Winter 2023**

Lecture 3: The method of moments

Lecturer: Dominik Rothenhäusler *January 18*

Note: *These lecture notes were written by Art Owen. If you like the material, he gets the credit! These notes have not been subjected to the usual scrutiny reserved for formal publications. They are meant as a memory aid for students who took stat 200 at Stanford University. They may be distributed outside this class only with the permission of Art Owen. Also, Stanford University holds the copyright.*

Abstract

These notes are mnemonics about what was covered in class. They don't replace being present or reading the book. Reading ahead in the book is very effective.

In reviewing probability we emphasized moments. Turns out there's a method for using them.

3.1 Preamble

Now we do statistics. I used the example of an MD working in a neighborhood who sees cholesterol numbers $X_1, X_2, \ldots, X_n$ for n patients. That is a fine record of the past but what does it tell about future patients? We let $Y_i = 1_{X_i > 240}$ as an indicator variable for a patient with a worrisome cholesterol level.

Here is how we usually work in statistics. We consider first that the data we got are a random sample from some distribution F. We suppose next that F belongs to a known parametric family such as the normal distributions $N(\mu, \sigma^2)$ with PDF

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad x \in \mathbb{R}$$

or Poisson with probability mass function (PMF) $p(x; \lambda) = e^{-\lambda}\lambda^x/x!$ for $x = 0, 1, 2, \ldots$. In general $X \sim F$ with PDF $f(x; \theta)$ (PMFs are similar).

The past data are independent and identically distributed (IID) from this distribution. So are the future values and they're independent of the past ones. The only thing we don't know is θ. If we knew θ we would know the distribution of these random variables completely.

At a high level, here are the tasks this framework brings:

1. find some estimate of θ of the form $\hat{\theta} = T(X_1, \ldots, X_n)$
   that $\hat{\theta}$ is a function of the data (called a 'statistic'). It is therefore subject to the laws of probability.
2. what good (or bad) properties does our estimate have? Mainly, in what ways is it close to the true θ?
3. given two estimates can we decide which is better?
4. in this framework, do we think that some particular value θ_0 is compatible with the data we have seen? We will then test a hypothesis that $\theta = \theta_0$. Often we set things up so that $\theta_0 = 0$ is meaningful scientifically and then we want to test whether the true θ could really be 0.
5. we may also want to test whether the data are compatible with our chosen distribution family $f(x; \theta)$. These are 'goodness of fit tests'. Maybe we were wrong about that assumption.

© Stanford University 2023

6. We often want a confidence interval. Let $L(X_1, \ldots, X_n)$ and $U(X_1, \ldots, X_n)$ be two statistics. They form a 95% confidence interval for θ if

$$\Pr(L(X_1, \ldots, X_n) \leqslant \theta \leqslant U(X_1, \ldots, X_n)) = 0.95.$$

Those are the main tasks we will look at: estimation, testing, and forming confidence intervals. We usually work with IID X_i. Some of the methods and problems we consider can extend beyond the case of IID X_i.

3.2 Method of moments

Let's think of our friend the MD with $Y_i \in \{0, 1\}$ for $i = 1, \ldots, n$. Since cholesterol levels X_i are IID the Y_i are IID Bern(p) for some unknown p. Suppose that of $n = 200$ patients there were 15 with high cholesterol. I.e., $\sum_{i=1}^n Y_i = 15$.

In the method of moments we find $\mathbb{E}(Y; p)$ under our parametric model and equate it to $\bar{Y} = (1/n)\sum_{i=1}^n Y_i$. We estimate that the population mean equals the sample mean. The Bern(p) distribution has mean p. The data have mean $15/200 = 0.075$. Equating them gives us $\hat{p} = .075$ as our estimate of p. This almost seems too easy. We have 7.5% high cholesterol in the sample so we guess it is 7.5% in the population. Then again without other information, why would one estimate it to be say 7% or 8% when the sample proportion was 7.5%?

In this case, $n\hat{p} \sim \text{Bin}(n, p)$. So we almost know the exact distribution of $\hat{p}$. The sticking point that we will have to get around later is that the distribution of our estimate $\hat{p}$ depends on the true p which we don't know. We face many seemingly circular arguments like this in statistics.

3.2.1 Normal data

Now suppose that $X_i \sim N(\mu, \sigma^2)$. Then $\mathbb{E}(X) = \mu$ so we use $\hat{\mu} = \bar{X}$ as before. But we still need an estimate of σ^2. We get that by using two moments, solving

$$\mathbb{E}(X) = \bar{X}, \quad \text{and} \quad \mathbb{E}(X^2) = \overline{X^2}.$$

Here $\overline{X^2} = (1/n)\sum_{i=1}^n X_i^2$ and more generally $\overline{X^r} = (1/n)\sum_{i=1}^n X_i^r$. Rice's notation is $\mu_r = \mathbb{E}(X^r)$ and $\hat{\mu}_r = \overline{X^r}$. The new equation is $\mathbb{E}(X^2) = \mu^2 + \sigma^2$. Solving two moment equations gives us $\hat{\mu} = \bar{X}$ as before and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2, \quad \text{i.e.,} \quad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Many students will already be familiar with a different estimator

$$s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2.$$

This will come up for us later.

If X and $X_i \sim N(\mu, \sigma^2)$ then

$$\Pr(X > 240) = \Pr\Big(\frac{X - \mu}{\sigma} > \frac{240 - \mu}{\sigma}\Big) = \Pr\Big(N(0,1) > \frac{240 - \mu}{\sigma}\Big) = 1 - \Phi((240 - \mu)/\sigma)$$

where Φ is the CDF of the $N(0, 1)$ distribution. The MD could also estimate the fraction of high cholesterol patients by $1 - \Phi((240 - \hat{\mu})/\hat{\sigma})$. Later we will look at criteria for choosing from among two or more different estimators.

3.2.2 Geometric distribution

For a geometric distribution with parameter success probability p we have probability mass function $\Pr[X = k] = (1-p)^{k-1}p$ for $k = 1, 2, \ldots$. The geometric distribution describes how many Bernoulli trials are needed to get a success. An alternative definition (also called geometric distribution) starts counting at 0 and counts the number of failures of Bernoulli trials before the first success.

For $X \sim \text{Geo}(p)$ we get $\mathbb{E}[X] = \frac{1}{p}$. Suppose the X_i are IID Geo(p). This gives us the method of moments estimate $\hat{p} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} X_i}$.

3.2.3 Gamma data

The Gamma distribution with shape $\alpha > 0$ is denoted Gam(α) (in these notes). It has PDF $f(x;\alpha) = x^{\alpha-1}e^{-x}/\Gamma(\alpha)$ for $0 < x < \infty$, where Γ is the Gamma function (see lecture 2). If $Y \sim \text{Gam}(\alpha)$ and $X = Y/\lambda$ for $\lambda > 0$ then X has the Gam(α, λ) distribution. This is the definition of Gam(α, λ). The original is then Gam($\alpha, 1$).

Your probability background should enable you to do these two things:

1. show that the PDF of Gam(α, λ) is $f(x;\alpha,\lambda) = f(x/\lambda;\alpha)/\lambda$, and
2. evaluate this PDF.

In class we worked out that the method of moments estimates are

$$\hat{\alpha} = \frac{\bar{X}^2}{\hat{\sigma}^2} \quad \text{and} \quad \hat{\lambda} = \frac{\bar{X}}{\hat{\sigma}^2},$$

where $\hat{\sigma}^2$ is the same as for the normal case.

3.3 How well does M.O.M. work?

Let θ be the true parameter value and $\hat{\theta}$ be our estimate of it. When we want to keep track of the sample size n we write $\hat{\theta}$ as $\hat{\theta}_n$. A very mild requirement is that $\hat{\theta}_n$ should converge to θ as $n \to \infty$. An estimator that cannot get the right answer on unbounded sample sizes is problematic. We want $\Pr(|\hat{\theta}_n - \theta| > \epsilon) \to 0$.

We want that for all $\epsilon > 0$ and for all θ too. We have a set Θ (capital Greek letter θ) containing all possible values of θ. Now our estimator is ***consistent*** if

$$\lim_{n\to\infty} \Pr(|\hat{\theta}_n - \theta| > \epsilon; \theta) = 0, \quad \text{for all } \epsilon > 0 \text{ and all } \theta \in \Theta.$$

Our Θ is the set of all possible values that θ could take. Often $\theta = \mathbb{R}^r$ for some $r \geqslant 1$. Other times only certain values of θ are possible. For $N(\mu, \sigma^2)$

$$\Theta = \{(\mu, \sigma) \mid \mu \in \mathbb{R}, \sigma > 0\}$$

and the Gamma family has $\theta = (\alpha, \lambda) \in \Theta$ where

$$\Theta = \{(\alpha, \lambda) \mid 0 < \alpha < \infty, 0 < \lambda < \infty\}.$$

Later Rice uses Θ (capital of the Greek leter θ) to denote a random variable that takes the value θ. We will keep the two usages separate.

If $\theta = \mathbb{E}(X)$ and $\hat{\theta}_n = \bar{X}$ based on n IID X_i then $\hat{\theta}_n$ is automatically consistent by the law of large numbers. That was easy.

Now suppose that $\theta = g(\mathbb{E}(X))$ and $\hat{\theta}_n = g(\bar{X})$. If g is continuous at θ then for any $\epsilon > 0$ there exists a $\delta > 0$ such that $|\bar{X} - \mathbb{E}(X)| < \delta$ implies that $|g(\bar{X}) - g(\mathbb{E}(X))| < \epsilon$. This is from the definition of continuity. That means

$$\Pr(|\hat{\theta}_n - \theta| > \epsilon) = \Pr(|g(\bar{X}) - g(\mathbb{E}(X))| > \epsilon) \leqslant \Pr(|\bar{X} - \mathbb{E}(X)| > \delta) \to 0$$

by the LLN. So $\hat{\theta}_n$ is consistent when g is continuous at the true θ. If g is a continuous function of Θ (i.e., continuous everywhere) then the moment estimator $\hat{\theta}_n$ is consistent.

3.3.1 Delta method variance

We will use a Taylor expansion in order to apply the CLT to the method of moments. Rice uses a Taylor expansion on the method of maximum likelihood so we might as well add that in for the method of moments too (to be consistent).

If $\hat{\theta} = \bar{X}$ then a CLT for $\bar{X}$ immediately gives one for $\hat{\theta}$. The same would happen for a linear function $\hat{\theta} = a + b \times \bar{X}$. If $\bar{X} \approx N(\mu, \sigma^2/n)$ then $a + b \times \bar{X} \approx N(a + b\mu, b^2\sigma^2/n)$

If $\bar{X}$ is the average of n IID random variables having mean μ and variance σ^2 then $\bar{X}$ converges to $\mu = \mathbb{E}(X)$ (by the LLN). Let's make a Taylor approximation to $g(\bar{X})$ at μ:

$$g(\bar{X}) = g(\mu) + (\bar{X} - \mu)g'(\mu) + \frac{1}{2}(\bar{X} - \mu)^2 g''(\mu) + \cdots .$$

Therefore

$$g(\bar{X}) - g(\mu) = (\bar{X} - \mu)g'(\mu) + \frac{1}{2}(\bar{X} - \mu)^2 g''(\mu) + \cdots .$$

Now $\mathbb{E}((\bar{X} - \mu)^2) = \sigma^2/n$ so the typical size of $|\bar{X} - \mu|$ is about $\sigma/\sqrt{n}$. Higher powers of $|\bar{X} - \mu|$ are then relatively negligble. So we can work with the approximations

$$g(\bar{X}) - g(\mu) \approx (\bar{X} - \mu)g'(\mu) \approx N(0, \sigma^2 g'(\mu)^2/n).$$

We wrote an infinite Taylor expansion but we could also terminate it using $g''(\mu)$ where μ^* is somewhere between $\bar{X}$ and μ. A more advanced course would take more care about conditions on g than we do here.

There is one situation where the argument above goes wrong. If $g'(\mu) = 0$ then the first term $(\bar{X} - \mu)g'(\mu)$ is no longer dominant. We would then have to find the smallest order derivative of g that is not zero.

Our delta method approximation is then

$$\hat{\theta}_n = g(\bar{X}) \approx N\Big(\theta, \frac{\sigma^2 g'(\mu)^2}{n}\Big)$$

if X_i are independent with mean μ and variance σ^2, $\theta = g(\mathbb{E}(X)) = g(\mu)$ for a function g that is smooth and has $g'(\mu) \neq 0$.

3.3.2 Problems and advantages of moments

Maybe θ has $r = 3$ parameters in it. Then we form and solve equations $\mathbb{E}(X^k) = \overline{X^k}$ for $k = 1, 2, 3$. We could be out of luck of $\mathbb{E}(X^3) = \infty$. Then the method of moments would not deliver estimates for us.

We could have X with a PDF $f(x;\theta)$ where $\theta = \mathbb{E}(X)$ is known to satisfy $\theta > 0$. Yet we might get $\hat{\theta} = \bar{X} < 0$. In some settings we can get a negative estimate $\hat{\sigma}^2$ for a variance. We probably knew that $\hat{\sigma}^2$ would be wrong but still getting $\hat{\sigma}^2 < 0$ remains embarrasing.

We might have a parameter that must be an integer. Suppose your puppy got 0 or 1 or 2 copies of a certain gene from its parents. Call that value θ. Now you get some random variables that are (say) Poi$(10 \times \theta)$. After a bit of algebra you find that the method of moments estimate is $\hat{\theta} = \bar{X}/10$ but it is not 0 or 1 or 2.

Now some advantages. Suppose that $n = 10^{11}$. Then $\hat{\theta}$ will require a lot of summing. You can however spread those sums over tens of thousands of computers all running in parallel (if you have them). More complicated estimates can be harder to parallelize.

A second advantage. Suppose that you really want $\theta = \mathbb{E}(X)$ or $\mathrm{Var}(X)$ or both and you made an estimate assuming that X_i have a PDF with the parametric distribution form $f(x;\theta)$. If you were wrong about that f your $\bar{X}$ will still be consistent for $\mathbb{E}(X)$.