Method of Moments Estimation for Ransomised Response Surveys

2023-11-09

What is a moment?

For a random variable X, the r^{th} moment μ_r is given by $\mathbb{E}[X^r]$.

For a discrete random variable with sample space $\mathcal{X}$, we may calculate this as:

$$\mu_r = \mathbb{E}[X^r] = \sum_{x \in \mathcal{X}} x^r \Pr(X = x).$$

For a continuous random variable with probability density function $f_X(x)$, we may calculate this as:

$$\mu_r = \mathbb{E}[X^r] = \int_{x \in \mathcal{X}} x^r f_X(x) \mathrm{d}x.$$

What is the method of moments?

Suppose we have observed values $x_1, \ldots, x_n$, which we wish to model as realisations from a collection of n independent and identically distributed (i.i.d) random variables $X_1, \ldots, X_n$.

We might assume a parametric model for the probability density function $f_X(x; \theta)$, so that the distribution of the X_i's is entirely described by one or more unknown parameter(s) θ.

The method of moments gives us a way to estimate the parameters θ based on the observed values $x_1, \ldots, x_n$. We do this by equating the moments of the sample and of the assumed parametric model.

Example 1: Exponential distribution

Suppose that $X_1, \dots, X_n \overset{\text{i.i.d}}{\sim} \text{Exp}(\lambda)$ with probability density function

$$f_X(x) = \lambda \exp\{-\lambda x\} \mathbb{I}[x > 0],$$

and that we observe a sample of values $x_1, \dots, x_n$ with mean $\bar{x} = 3.25 = \frac{13}{4}$.

To estimate the single parameter λ for this model, we can equate the first moment of the Exponential model with the first moment of our sample.

You might already know the expectation of an exponential variable is λ^{-1}. If not, you might derive it for yourself:

$$\begin{aligned}
\mu_1 = \mathbb{E}[X] &= \int_{-\infty}^{\infty} x f_X(x) \mathrm{d}x \\
&= \int_{-\infty}^{\infty} x\lambda \exp\{-\lambda x\} \mathbb{I}[x > 0] \mathrm{d}x \\
&= \int_{0}^{\infty} x\lambda \exp\{-\lambda x\} \mathrm{d}x \\
&= [-x \exp\{-\lambda x\}]_0^{\infty} - \int_0^{\infty} -\exp\{-\lambda x\} \mathrm{d}x \\
&= (0 - 0) + \frac{1}{\lambda} \int_0^{\infty} \lambda \exp\{-\lambda x\} \mathrm{d}x \\
&= \frac{1}{\lambda}.
\end{aligned}$$

Equating this first moment of the probabiltiy model to the first sample moment, $\frac{1}{n}\sum_{i=1}^n x_i = \bar{x} = \frac{13}{4}$, we can find the method of moments estimate $\hat{\lambda}$ for λ.

This gives us

$$\frac{1}{\hat{\lambda}} = \bar{x} \quad \Rightarrow \quad \hat{\lambda} = \bar{x}^{-1} = \frac{4}{13}.$$

Example 2 - Gamma Distribution.

Suppose instead that we had a two parameter model for our data generating process, such as the gamma distribution. In this case, we proceed in much the same way, but to estimate the *two* parameters of this model we must equate the first *two* moments of the sample and the model, and then solve those equations simultaneously.

If $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\alpha, \beta)$, then

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp -\beta x \mathbb{I}[x > 0].$$

We can calculate the first two moments of the gamma distribution directly using this probability density function. Alternatively, we might obtain these using the known closed forms for the mean and variance of a Gamma random variable:

$$\mu_1 = \mathbb{E}[Y] = \frac{\alpha}{\beta};$$

$$\begin{aligned} \mu_2 = \mathbb{E}[Y^2] &= \text{Var}(Y) + \mathbb{E}[Y]^2 \\ &= \frac{\alpha}{\beta^2} + \left(\frac{\alpha}{\beta}\right)^2 \\ &= \frac{\alpha + \alpha^2}{\beta^2}. \end{aligned}$$

Equating these to the sample moments we get the system of simultaneous equations that we need to solve:

$$\bar{x} = \frac{\alpha}{\beta} \qquad \frac{1}{n}\sum_{i=1}^n x_i^2 = \frac{\alpha + \alpha^2}{\beta^2}. \tag{1}$$

Suppose $\hat{\alpha}$ and $\hat{\beta}$ satisfy both of these equations. By rearrangement of the first equation we can express $\hat{\alpha} = \hat{\beta}\bar{x}$. Substituting this into the second equation we find that:

$$\frac{1}{n}\sum_{i=1}^n x_i^2 = \frac{\hat{\beta}\bar{x} + (\hat{\beta}\bar{x})^2}{\hat{\beta}^2} \quad \Rightarrow \quad \hat{\beta} = \frac{\bar{x}}{\sum_{i=1}^n x_i^2 - \bar{x}^2}.$$

This can then be substituted back into the fist equation to find that

$$\hat{\alpha} = \frac{\bar{x}^2}{\sum_{i=1}^n x_i^2 - \bar{x}^2}.$$

And so, we have found the method of moments estimators $\hat{\alpha}$ and $\hat{\beta}$ for our two model parameters α and β.

Example 3 - Beta Distribution

Suppose instead that we had observations $z_1, \ldots, z_n$ which we model as realisations of $Z_1, \ldots, Z_n$ which are i.i.d. random variables from a beta distribution. This is another two parameter family of probability distributions, characterised by probability density functions of the form

$$f_Z(z; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha - 1}(1 - x)^{\beta - 1}.$$

We can show that

$$\mathbb{E}[Z] = \frac{\alpha}{\alpha + \beta}$$

and

$$\mathbb{E}[Z^2] = \mathrm{Var}(Z) + \mathbb{E}[Z]^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} + \left[\frac{\alpha}{\alpha + \beta}\right]^2 = \frac{\alpha\beta + \alpha^2(\alpha + \beta + 1)}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

So the system of equations that we wish to solve simultaneously for $\hat{\alpha}$ and $\hat{\beta}$ are

$$\sum_{i=1}^{n} z_i = \frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}}$$

and

$$\sum_{i=1}^{n} z_i^2 = \frac{\hat{\alpha}\hat{\beta} + \hat{\alpha}^2(\hat{\alpha} + \hat{\beta} + 1)}{(\hat{\alpha} + \hat{\beta})^2(\hat{\alpha} + \hat{\beta} + 1)}.$$

This system of equations has no closed-form solution but could be solved using numerically to optimisation to produce the method of moments estimates for a particular data set.

This points out that while the method of moments is conceptually simple, actually obtaining the estimates for a given data generating model is not always straightforward.

Example 4 - Binomial Distribution (Direct Response)

Suppose we are estimating some population proportion p of people with a postgraduate degree. Assuming that we can take a representative sample of size m from the population, we might model the number of survey respondents who have a postgraduate degree as $N \sim \text{Bin}(m, p)$.

In that case $\mathbb{E}[N] = mp$ and since we only run the survey once, our sample mean is simply the one observed count.

Equating the fist moments of the sample and the model, we have $n = m\hat{p}$ and so $\hat{p} = \frac{n}{m}$.

Therefore, under this direct-response survey design, the method of moments estimate for the population proportion of people with a postgraduate degree is simply the sample proportion.

Example 5 - Binomial Distribution (Random Response)

Suppose instead that we were interested in the proportion of the population who have illegally downloaded content from the internet. We will denote the true population proportion that we wish to estimate as p.

If we use a direct response surey, some respondents might be reluctant to admit thier bad behaviour, this might cause non-response or dishonest responses from some participants. To alleviate this issue, we can use a randomised response survey to provide respondents with plausible deniability. We will add some element of randomisation to the survey design, so that for an individual respondent who replies in the affirmative we will no longer be certain that they have illegal downloads.

In particular, if a participant has not downloaded content from the internet, they will declare as such with probability 0.75 and with probability 0.25 they will declare that they do indeed have illegal downloads. Therefore, a respondent might reply in the affirmative to this survey **either** if they have truly downloaded illegal content **or** if they have not and happen to do so because of their random allocation. The probability of an affirmative response is given by $\theta = p + (1 - p)/4$.

The expectation of our total number of affirmative responses N could, as in the direct response design, be modelled using a binomial distribution with m trials. However, for this randomised response design the "success" probability is θ rather than p.

Therefore the first moment of our affirmative response count N is

$$\mathbb{E}[N] = m\theta = m\left[p + (1 - p)/4\right],$$

which we can equate to the sample mean (i.e. the observed count)

$$m\left[\widehat{p} + (1 - \widehat{p})/4\right] = n.$$

Rearranging this equality results in the following method of moments estimate for p:

$$\widehat{p} = \frac{4}{3}\left(\frac{n}{m} - \frac{1}{4}\right) = \frac{4n}{3m} - \frac{1}{3}.$$

We can use this expression to get an unbiased estimate of the population proportion of people who illegally download content, based on our deliberately corrupted responses that provide individuals with plausible deniability.

Session Information

R version 4.3.1 (2023-06-16)

Platform: x86_64-apple-darwin20 (64-bit)

locale: en_GB.UTF-8||en_GB.UTF-8||en_GB.UTF-8||C||en_GB.UTF-8||en_GB.UTF-8

attached base packages: *stats*, *graphics*, *grDevices*, *utils*, *datasets*, *methods* and *base*

loaded via a namespace (and not attached): *compiler(v.4.3.1)*, *fastmap(v.1.1.1)*, *cli(v.3.6.1)*, *tools(v.4.3.1)*, *htmltools(v.0.5.6)*, *rstudioapi(v.0.15.0)*, *yaml(v.2.3.7)*, *Rcpp(v.1.0.11)*, *pander(v.0.6.5)*, *rmarkdown(v.2.24)*, *knitr(v.1.43)*, *jsonlite(v.1.8.7)*, *xfun(v.0.40)*, *digest(v.0.6.33)*, *rlang(v.1.1.2)* and *evaluate(v.0.21)*