1 Variance Reduction (continued)

1.3 Stratification

The idea in **stratified** sampling is to **split up the domain** $\mathcal{D}$ of $\boldsymbol{X}$ into separate regions, take a sample of points from each such region, and **combine the results** to estimate $E(f(\boldsymbol{X}))$. Intuitively, if each region gets its fair share of points then we should get a better answer. We might be able to do better still by **oversampling within the important strata** and **under-sampling** those in which f is nearly constant.

Our goal is to estimate $\mu = \int_{\mathcal{D}} f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$. We partition $\mathcal{D}$ into **mutually exclusive and exhaustive regions** $\mathcal{D}_j$, for $j = 1, \ldots, J$. These regions are the **strata**. We write $\omega_j = P(\boldsymbol{X} \in \mathcal{D}_j)$ and to avoid trivial issues, we assume $\omega_j > 0$ for all j. Next, let $p_j(\boldsymbol{x}) = \omega_j^{-1} p(\boldsymbol{x}) 1_{\boldsymbol{x}\in\mathcal{D}_j}$, the conditional density of $\boldsymbol{X}$ given that $\boldsymbol{X} \in \mathcal{D}_j$. To **use stratified sampling**, we **must know** the probabilities ω_j of the strata, and we must also know how to sample $\boldsymbol{X} \sim p_j$ for $j = 1, \ldots, J$. These conditions are quite reasonable. When we are defining strata, we naturally prefer ones we can sample from.

Let $\boldsymbol{X}_{ij} \overset{i.i.d.}{\sim} p_j$ for $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$ be sampled independently. The **stratified sampling estimate** of μ is

$$\widehat{\mu}_{\text{strat}} = \sum_{j=1}^{J} \frac{\omega_j}{n_j} \sum_{i=1}^{n_j} f(\boldsymbol{X}_{ij}).$$

We choose $n_j > 0$ so that $\widehat{\mu}_{\text{strat}}$ is properly defined. Unless otherwise specified, we make sure that $n_j > 2$, which will allow the variance estimate below to be applied.

Note that the **stratified estimator** of μ is unbiased as

$$\begin{aligned}
E(\widehat{\mu}_{\text{strat}}) &= \sum_{j=1}^{J} \omega_j E\left(\frac{1}{n_j}\sum_{i=1}^{n_j} f(\boldsymbol{X}_{ij})\right) \\
&= \sum_{j=1}^{J} \omega_j \int_{\mathcal{D}_j} f(\boldsymbol{x})p_j(\boldsymbol{x})d\boldsymbol{x} \\
&= \sum_{j=1}^{J} \int_{\mathcal{D}_j} f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \\
&= \int_{\mathcal{D}} f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \\
&= \mu.
\end{aligned}$$

Let $\mu_j = \int_{\mathcal{D}_j} f(\boldsymbol{x})p_j(\boldsymbol{x})d\boldsymbol{x}$ and $\sigma_j^2 = \int_{\mathcal{D}_j} (f(\boldsymbol{x}) - \mu_j)^2 p_j(\boldsymbol{x})d\boldsymbol{x}$ be the jth stratum **mean** and **variance**, respective. The variance of the stratified sampling estimate is

$$Var(\widehat{\mu}_{\text{start}}) = \sum_{i=1}^{J} \omega_j^2 \frac{\sigma_j^2}{n_j}.$$

An immediate consequence is that $Var(\widehat{\mu}_{\text{start}}) = 0$ for integrands f that are constant within strata $\mathcal{D}_j$ for all $j = 1, 2, \ldots, J$.

For error estimation, we can proceed as follows. Denoting $f(X_{ij})$ by Y_{ij}, we have

$$\widehat{Var}(\widehat{\mu}_{\text{start}}) = \sum_{j=1}^{J} \omega_j^2 \frac{s_j^2}{n_j},$$

where

$$\widehat{\mu}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij} \quad \text{and} \quad s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (Y_{ij} - \widehat{\mu}_j)^2.$$

A CLT based 99% confidence interval for μ is

$$\widehat{\mu}_{\text{start}} \mp 2.58\sqrt{\widehat{Var}(\widehat{\mu}_{\text{start}})}.$$

A natural choice for **stratum sample sizes** is **proportional allocation**, $n_j = n\omega_j$. In our analysis, we will suppose that all the n_j are integers. We can usually choose n and D_j to make this so, or else accept small non-proportionalities due to rounding. For proportional allocation, $\widehat{\mu}_{\text{strat}}$ reduces to the ordinary sample mean

$$\widehat{\mu}_{\text{strat}} = \frac{1}{n} \sum_{j=1}^{J} \sum_{i=1}^{n_j} f(\boldsymbol{X}_{ij}),$$

and $Var(\widehat{\mu}_{\text{start}})$ becomes

$$Var(\widehat{\mu}_{\text{strat}}) = \frac{1}{n} \sum_{j=1}^{J} \omega_j \sigma_j^2.$$

Note that the variance of $f(\boldsymbol{X})$ can be decomposed as follows:

$$\sigma^2 = Var(f(\boldsymbol{X})) = \sum_{j=1}^{J} \omega_j \sigma_j^2 + \sum_{j=1}^{J} \omega_j (\mu_j - \mu)^2.$$

In the above, the variance of $f(X)$ is **decomposed into within- and between-stratum** components. Therefore,

$$Var\left(\widehat{\mu}_{\text{strat}}\right) = \frac{1}{n} \sum_{j=1}^{J} \omega_j \sigma_j^2 \leq \frac{\sigma^2}{n}.$$

It shows that stratified sampling with proportional allocation cannot have larger variance than simple Monte Carlo sampling. A **good stratification** scheme is one that **reduces the within–stratum** variance.

1.4 Conditioning

Sometimes we can **do part of the problem in closed form**, and then do the **rest of it by Monte Carlo**. For example, suppose that we want to find $\mu = \int_0^1 \int_0^1 f(x,y)dxdy$, where $f(x,y) = e^{yg(x)}$. It is easy to integrate out y for fixed x, yielding $h(x) = (e^{g(x)} - 1)/g(x)$. Then we have a **one dimensional** problem, which may be simpler to handle. In general, suppose that $\boldsymbol{X} \in \mathbb{R}^k$ and $\boldsymbol{Y} \in \mathbb{R}^{d-k}$ are random vectors and that we want to estimate $E(f(\boldsymbol{X}, \boldsymbol{Y}))$. The natural estimate is $\widehat{\mu} = (1/n)\sum_{i=1}^n f(\boldsymbol{X}_i, \boldsymbol{Y}_i)$, where $(\boldsymbol{X}_i, \boldsymbol{Y}_i) \in \mathbb{R}^d$ are independent **samples** from the **joint distribution** of $(\boldsymbol{X}, \boldsymbol{Y})$. Now, let $h(\boldsymbol{x}) = E(f(\boldsymbol{X}, \boldsymbol{Y})|\boldsymbol{X} = \boldsymbol{x})$. We **might also estimate** μ by

$$\widehat{\mu}_{\text{cond}} = \frac{1}{n}\sum_{i=1}^n h(\boldsymbol{X}_i)$$

where $\boldsymbol{X}_i$ are independently sampled from the distribution of $\boldsymbol{X}$. The justification for the method is that $E(f(\boldsymbol{X}, \boldsymbol{Y})) = E(E(f(\boldsymbol{X}, \boldsymbol{Y})|\boldsymbol{X})) = E(h(\boldsymbol{X}))$. The **method is called conditioning**, or **conditional Monte Carlo**. The main requirement for conditioning is that we must be able to compute $h(\cdot)$. Note that

$$Var(\widehat{\mu}_{\text{cond}}) = \frac{1}{n}Var(h(\boldsymbol{X})) = \frac{1}{n}Var\left(E(f(\boldsymbol{X}, \boldsymbol{Y})|\boldsymbol{X})\right) \le Var(\widehat{\mu}),$$

as

$$Var(f(\boldsymbol{X}, \boldsymbol{Y})) = E(Var(f(\boldsymbol{X}, \boldsymbol{Y})|\boldsymbol{X})) + Var(E(f(\boldsymbol{X}, \boldsymbol{Y})|\boldsymbol{X})).$$

Conditioning is a special case of derandomization. The function $f(\boldsymbol{X}, \boldsymbol{Y})$ has **two sources of randomness**, $\boldsymbol{X}$ and $\boldsymbol{Y}$. For any given $\boldsymbol{x}$ and random $\boldsymbol{Y}$ we **replace the random value** $f(\boldsymbol{x}, \boldsymbol{Y})$ by its **expectation** $h(\boldsymbol{x})$, **removing one** of the two sources of randomness.

1.5 Control Variates

Suppose that we want to find $\mu = E(f(\boldsymbol{X}))$ and that we know the value $\theta = E(h(\boldsymbol{X}))$, where $h(\cdot)$ is 'similar' to $f(\cdot)$. The precise **meaning of 'similar'** depends on the problem. For a value $\beta \in \mathbb{R}$, the **regression estimator** of μ is given by

$$\widehat{\mu}_\beta = \frac{1}{n}\sum_{i=1}^n \left(f(\boldsymbol{X}_i) - \beta h(\boldsymbol{X}_i)\right) + \beta\theta = \widehat{\mu} + \beta\left(\theta - \widehat{\theta}\right),$$

where $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n f(X_i)$ and $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n h(X_i)$. The variance of the regression estimator is

$$Var(\widehat{\mu}_\beta) = \frac{1}{n}\left(Var(f(\boldsymbol{X})) - 2\beta Cov(f(\boldsymbol{X}), h(\boldsymbol{X})) + \beta^2 Var(h(\boldsymbol{X}))\right).$$

Let β_{opt} denote the best value of β in the sense that $Var(\widehat{\mu}_\beta)$ is minimum at $\beta = \beta_{\text{opt}}$. Then

$$\beta_{\text{opt}} = \frac{\rho\sigma}{\sigma_h} \quad \text{and} \quad Var(\widehat{\mu}_{\beta_{\text{opt}}}) = \frac{\sigma^2}{n}(1-\rho^2),$$

where $\rho = Corr\left(f(\boldsymbol{X}), h(\boldsymbol{X})\right)$, $\sigma^2 = Var\left(f(\boldsymbol{X})\right)$ and $\sigma_h^2 = Var\left(h(\boldsymbol{X})\right)$. In the regression estimator, any **control variate that correlates** with f is **helpful, even** one that correlates **negatively**.

In practice we do not know β_{opt} and hence we estimate it by

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} (f(\boldsymbol{X}_i) - \widehat{\mu}) \left(h(\boldsymbol{X}_i) - \widehat{\theta} \right)}{\sum_{i=1}^{n} \left(h(\boldsymbol{X}_i) - \widehat{\theta} \right)^2},$$

In general $E(\widehat{\mu}_{\widehat{\beta}}) \neq \mu$, but this bias is usually small. The estimated variance of $\widehat{\mu}_{\widehat{\beta}}$ is

$$\widehat{Var}(\widehat{\mu}_{\widehat{\beta}}) = \frac{1}{n^2} \sum_{i=1}^{n} \left(f(\boldsymbol{X}_i) - \widehat{\mu}_{\widehat{\beta}} - \widehat{\beta} \left(h(\boldsymbol{X}_i) - \widehat{\theta} \right) \right)^2.$$

A 99% **confidence interval** is $\widehat{\mu}_{\widehat{\beta}} \pm 2.58 \sqrt{\widehat{Var}(\widehat{\mu}_{\widehat{\beta}})}$.