STAT/Q SCI 403: Introduction to Resampling Methods **Spring 2025**

Lecture 2: Monte Carlo Simulation

Instructor: Marina Meilă. Course notes courtesy of Yen-Chi Chen

2.1 Monte Carlo Integration

Assume we want to evaluate the following integration:

$$\int_0^1 e^{-x^3} dx.$$

What can we do? The function e^{-x^3} does not seem to have a closed form solution so we have to use some computer experiment to evaluate this number. The traditional approach to evaluate this integration is to use so-called the *Riemann Integration*, where we choose points $x_1, \cdots, x_K$ evenly spread out over the interval $[0, 1]$ and then we evaluate $f(x_1), \cdots, f(x_K)$ and finally use

$$\frac{1}{K}\sum_{i=1}^{K} f(x_i)$$

to evaluate the integration. When the function is smooth and $K \to \infty$, this numerical integration converges to the actual integration.

Now we will introduce an alternative approach to evaluate such an integration. First, we rewrite the integration as

$$\int_0^1 e^{-x^3} \cdot 1 dx = \mathbb{E}\left(e^{-U^3}\right),$$

where U is a uniform random variable over the interval $[0, 1]$. Thus, the integration is actually an expected value of a random variable e^{-U^3}, which implies that evaluating the integration is the same as *estimating the expected value*. So we can generate IID random variables $U_1, \cdots, U_K \sim \mathsf{Uni}[0, 1]$ and then compute $W_1 = e^{-U_1^3}, \cdots, W_K = e^{-U_K^3}$ and finally use

$$\bar{W}_K = \frac{1}{K}\sum_{i=1}^{K} W_i = \frac{1}{K}\sum_{i=1}^{K} e^{-U_i^3}$$

as a numerical evaluation of $\int_0^1 e^{-x^3} dx$. By the Law of Large Number,

$$\bar{W}_K \xrightarrow{P} \mathbb{E}(W_i) = \mathbb{E}\left(e^{-U_i^3}\right) = \int_0^1 e^{-x^3} dx,$$

so this alternative numerical method is statistically consistent.

In the above example, the integration can be written as

$$I = \int f(x)p(x)dx, \tag{2.1}$$

where f is some function and p is a probability density function. Let X be a random variable with density p. Then equation (2.1) equals

$$\int f(x)p(x)dx = \mathbb{E}(f(X)) = I.$$

Namely, the result of this integration is the same as the expected value of the random variable $f(X)$. The alternative numerical method to evaluate the above integration is to generate IID $X_1, \cdots, X_N \sim p$, N data points, and then use the sample average

$$\widehat{I}_N = \frac{1}{N}\sum_{i=1}^{N} f(X_i).$$

This method, the method of evaluating the integration via simulating random points, is called the integration by *Monte Carlo Simulation*.

An appealing feature of the Monte Carlo Simulation is that the statistical theory is rooted in the theory of sample average. We are using the sample average as an estimator of the expected value. We have already seen that the bias and variance of an estimator are key quantities of evaluating the quality of an estimator. What will be the bias and variance of our Monte Carlo Simulation estimator?

The bias is simple–we are using the sample average as an estimator of it expected value, so the $\textbf{bias}(\widehat{I}_N) = 0$. The variance will then be

$$\begin{aligned}
\mathsf{Var}(\widehat{I}_N) &= \frac{1}{N}\mathsf{Var}(f(X_1)) \\
&= \frac{1}{N}\left(\mathbb{E}(f^2(X_1)) - \underbrace{\mathbb{E}^2(f(X_1))}_{I^2}\right) \\
&= \frac{1}{N}\left(\int f^2(x)p(x)dx - I^2\right).
\end{aligned} \tag{2.2}$$

Thus, the variance contains two components: $\int f^2(x)p(x)dx$ and I^2.

Given a problem of evaluating an integration, the quantity I is fixed. What we can choose is the number of random points N and the *sampling distribution* p! An important fact is that when we change the sampling distribution p, the function f will also change.

For instance, in the example of evaluating $\int_0^1 e^{-x^3}dx$, we have seen an example of using uniform random variables to evaluate it. We can also generate IID $B_1, \cdots, B_K \sim \mathsf{Beta}(2,2)$, K points from the beta distribution $\mathsf{Beta}(2,2)$. Note that the PDF of $\mathsf{Beta}(2,2)$ is

$$p_{\mathsf{Beta}(2,2)}(x) = 6x(1-x).$$

We can then rewrite

$$\int_0^1 e^{-x^3}dx = \int_0^1 \underbrace{\frac{e^{-x^3}}{6x(1-x)}}_{f(x)} \cdot \underbrace{6x(1-x)}_{p(x)} dx = \mathbb{E}\left(\frac{e^{-B_1^3}}{6B_1(1-B_1)}\right).$$

What is the effect of using different sampling distribution p? The expectation is always fixed to be I so the second part of the variance remains the same. However, the first part of the variance $\int f^2(x)p(x)dx$ depends how you choose p and the corresponding f.

Thus, different choices of p leads to a different variance of the estimator. We will talk about how to choose an optimal p in Chapter 4 when we talk about theory of importance sampling.

2.2 Estimating a Probability via Simulation

Here is an example of evaluating the power of a Z-test. Let $X_1, \cdots, X_{16}$ be a size 16 random sample. Let the null hypothesis and the alternative hypothesis be

$$H_0 : X_i \sim N(0,1), \quad H_a : X_i \sim N(\mu, 1),$$

where $\mu \neq 0$. Under the significance level α, the two-tailed Z-test is to reject H_0 if $\sqrt{16}|\bar{X}_{16}| \geq z_{1-\alpha/2}$, where $z_t = F^{-1}(t)$, where F is the CDF of the standard normal distribution. Assume that the true value of μ is $\mu = 1$. In this case, the null hypothesis is wrong and we should reject the null. However, due to the randomness of sampling, we may not be able to reject the null every time. So a quantity we will be interested in is: *what is the probability of rejecting the null under such μ?* In statistics, this probability (the probability that we reject H_0) is called the *power* of a test. Ideally, if H_0 is incorrect, we want the power to be as large as possible.

What will the power be when $\mu = 1$? Here is the analytical derivation of the power (generally denoted as β):

$$\begin{aligned}
\beta &= P(\text{Reject } H_0 | \mu = 1) \\
&= P(\sqrt{16}|\bar{X}_{16}| \geq z_{1-\alpha/2} | \mu = 1), \qquad \bar{X}_{16} \sim N(\mu, 1/16) \\
&= P(4 \cdot |N(1, 1/16)| \geq z_{1-\alpha/2}) \\
&= P(|N(4,1)| \geq z_{1-\alpha/2}) \\
&= P(N(4,1) \geq z_{1-\alpha/2}) + P(N(4,1) \leq -z_{1-\alpha/2}) \\
&= P(N(0,1) \geq z_{1-\alpha/2} - 4) + P(N(0,1) \leq -4 - z_{1-\alpha/2}).
\end{aligned} \tag{2.3}$$

Well...this number does not seem to be an easy one...

What should we do in practice to compute the power? Here is an alternative approach of computing the power using the Monte Carlo Simulation. The idea is that we generate N samples, each consists of 16 IID random variables from $N(1,1)$ (the distribution under the alternative). For each sample, we compute the Z-test statistic, $\sqrt{16}|\bar{X}_{16}|$, and check if we can reject H_0 or not (i.e., checking if this number is greater than or equal to $z_{1-\alpha/2}$). At the end, we use the ratio of total number of H_0 being rejected as an estimate of the power β. Here is a diagram describing how the steps are carried out:

$$\begin{aligned}
&N(1,1) \overset{\text{generates}}{\longrightarrow} 16 \text{ observations} \overset{\text{compute}}{\longrightarrow} \text{test statistic } \left(\sqrt{16}|\bar{X}_{16}|\right) \overset{\text{Reject } H_0}{\longrightarrow} D_1 = \text{Yes}(1)/\text{No}(0) \\
&N(1,1) \overset{\text{generates}}{\longrightarrow} 16 \text{ observations} \overset{\text{compute}}{\longrightarrow} \text{test statistic } \left(\sqrt{16}|\bar{X}_{16}|\right) \overset{\text{Reject } H_0}{\longrightarrow} D_2 = \text{Yes}(1)/\text{No}(0) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\
&N(1,1) \overset{\text{generates}}{\longrightarrow} 16 \text{ observations} \overset{\text{compute}}{\longrightarrow} \text{test statistic } \left(\sqrt{16}|\bar{X}_{16}|\right) \overset{\text{Reject } H_0}{\longrightarrow} D_N = \text{Yes}(1)/\text{No}(0)
\end{aligned}$$

Each sample will end up with a number D_i such that $D_i = 1$ if we reject H_0 and $D_i = 0$ if we do not reject H_0.

Because the Monte Carlo Simulation approach is to use the ratio of total number of H_0 being rejected to estimate β, this ratio is

$$\bar{D}_N = \frac{\sum_{j=1}^{N} D_j}{N}.$$

Is the Monte Carlo Simulation approach a good approach to estimate β? The answer is–yes it is a good approach of estimating β and moreover, we have already learned the statistical theory of such a procedure!

The estimator $\bar{D}_N$ is just a sample average and each D_j turns out to be a Bernoulli random variable with parameter

$$p = P(\text{Reject } H_0 | \mu = 1) = \beta$$

by equation (2.3). Therefore,

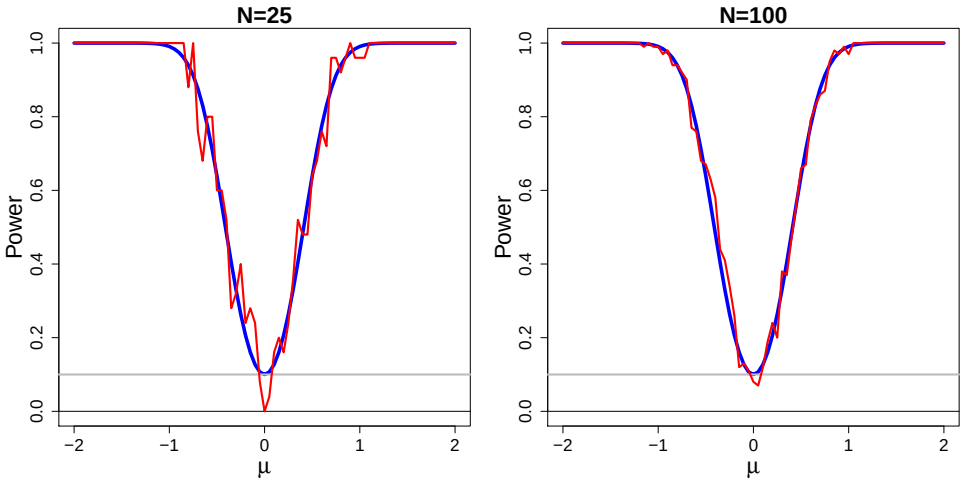
$$\begin{aligned}
\textbf{bias}\left(\bar{D}_N\right) &= \mathbb{E}(\bar{D}_N) - \beta = p - \beta = 0 \\
\textbf{Var}\left(\bar{D}_N\right) &= \frac{p(1-p)}{N} = \frac{\beta(1-\beta)}{N} \\
\textbf{MSE}\left(\bar{D}_N, \beta\right) &= \frac{\beta(1-\beta)}{N}.
\end{aligned}$$

Thus, the Monte Carlo Simulation method yields a consistent estimator of the power:

$$\bar{D}_N \xrightarrow{P} \beta.$$

Although here we study the Monte Carlo Simulation estimator of such a special case, this idea can be easily to generalize to many other situation as long as we want to evaluate certain numbers. In modern statistical analysis, most papers with simulation results will use some Monte Carlo Simulations to show the numerical results of the proposed methods in the paper.

The following two figures present the power β as a function of the value of μ (blue curve) with $\alpha = 0.10$. The red curves are the estimated power by Monte Carlo simulations using $N = 25$ and 100.



$\longrightarrow$ The gray line corresponds to the value of power being 0.10. Think about why the power curve (blue curve) hits the gray line at $\mu = 0$.

2.3 Estimating Distribution via Simulation

Monte Carlo Simulation can also be applied to estimate an unknown distribution as long as we can generate data from such a distribution. In Bayesian analysis, people are often interested in the so-called *posterior* distribution. Very often, we known how to generate points from a posterior distribution but we cannot write down its closed form. In this situation, what we can do is to simulate many points and estimate the distribution using these simulated points. So the task becomes:

given $X_1, \cdots, X_n \sim F$ (or PDF p), we want to estimate F (or the PDF p).

Estimating the CDF using EDF. To estimate the CDF, a simple but powerful approach is to use the EDF:

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x).$$

We have already learned a lot about EDF in the previous chapter.

Estimating the PDF using histogram. If the goal is to estimate the PDF, then this problem is called *density estimation*, which is a central topic in statistical research. Here we will focus on the perhaps simplest approach: histogram. Note that we will have a more in-depth discussion about other approaches in Chapter 8.

For simplicity, we assume that $X_i \in [0,1]$ so $p(x)$ is non-zero only within $[0,1]$. We also assume that $p(x)$ is smooth and $|p'(x)| \le L$ for all x (i.e. the derivative is bounded). The histogram is to partition the set $[0,1]$ (this region, the region with non-zero density, is called the support of a density function) into several bins and using the count of the bin as a density estimate. When we have M bins, this yields a partition:

$$B_1 = \left[0, \frac{1}{M}\right), B_2 = \left[\frac{1}{M}, \frac{2}{M}\right), \cdots, B_{M-1} = \left[\frac{M-2}{M}, \frac{M-1}{M}\right), B_M = \left[\frac{M-1}{M}, 1\right].$$

In such case, then for a given point $x \in B_\ell$, the density estimator from the histogram will be

$$\widehat{p}_n(x) = \frac{\text{number of observations within } B_\ell}{n} \times \frac{1}{\text{length of the bin}} = \frac{M}{n}\sum_{i=1}^{n} I(X_i \in B_\ell).$$

The intuition of this density estimator is that the histogram assign equal density value to every points within the bin. So for B_ℓ that contains x, the ratio of observations within this bin is $\frac{1}{n}\sum_{i=1}^{n} I(X_i \in B_\ell)$, which should be equal to the density estimate times the length of the bin.

Now we study the bias of the histogram density estimator.

$$\begin{aligned}
\mathbb{E}\left(\widehat{p}_n(x)\right) &= M \cdot P(X_i \in B_\ell) \\
&= M \int_{\frac{\ell-1}{M}}^{\frac{\ell}{M}} p(u)du \\
&= M\left(F\left(\frac{\ell}{M}\right) - F\left(\frac{\ell-1}{M}\right)\right) \\
&= \frac{F\left(\frac{\ell}{M}\right) - F\left(\frac{\ell-1}{M}\right)}{1/M} \\
&= \frac{F\left(\frac{\ell}{M}\right) - F\left(\frac{\ell-1}{M}\right)}{\frac{\ell}{M} - \frac{\ell-1}{M}} \\
&= p(x^*), \quad x^* \in \left[\frac{\ell-1}{M}, \frac{\ell}{M}\right].
\end{aligned}$$

The last equality is done by the mean value theorem with $F'(x) = p(x)$. By the mean value theorem again, there exists another point x^{**} between x^* and x such that

$$\frac{p(x^*) - p(x)}{x^* - x} = p'(x^{**}).$$

Thus, the bias

$$\begin{aligned}\textbf{bias}(\widehat{p}_n(x)) &= \mathbb{E}\left(\widehat{p}_n(x)\right) - p(x) \\ &= p(x^*) - p(x) \\ &= p'(x^{**}) \cdot (x^* - x) \\ &\leq |p'(x^{**})| \cdot |x^* - x| \\ &\leq \frac{L}{M}.\end{aligned} \tag{2.4}$$

Note that in the last inequality we use the fact that both x^* and x are within B_ℓ, whose total length is $1/M$, so the $|x^* - x| \leq 1/M$. The analysis of the bias tells us that the more bins we are using, the less bias the histogram has. This makes sense because when we have many bins, we have a higher resolution so we can approximate the fine density structure better.

Now we turn to the analysis of variance.

$$\begin{aligned}\mathsf{Var}(\widehat{p}_n(x)) &= M^2 \cdot \mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^{n} I(X_i \in B_\ell)\right) \\ &= M^2 \cdot \frac{P(X_i \in B_\ell)(1 - P(X_i \in B_\ell))}{n}.\end{aligned}$$

By the derivation in the bias, we know that $P(X_i \in B_\ell) = \frac{p(x^*)}{M}$, so the variance

$$\begin{aligned}\mathsf{Var}(\widehat{p}_n(x)) &= M^2 \cdot \frac{\frac{p(x^*)}{M} \times \left(1 - \frac{p(x^*)}{M}\right)}{n} \\ &= M \cdot \frac{p(x^*)}{n} - \frac{p^2(x^*)}{n}.\end{aligned} \tag{2.5}$$

The analysis of the variance has an interesting result: the more bins we are using, the higher variance we are suffering.

Now if we consider the MSE, the pattern will be more inspiring. The MSE is

$$\mathsf{MSE}(\widehat{p}_n(x)) = \mathsf{bias}^2(\widehat{p}_n(x)) + \mathsf{Var}(\widehat{p}_n(x)) \leq \frac{L^2}{M^2} + M \cdot \frac{p(x^*)}{n} - \frac{p^2(x^*)}{n}. \tag{2.6}$$

An interesting feature of the histogram is that: *we can choose M, the number of bins.* When M is too large, the first quantity (bias) will be small while the second quantity (variance) will be large; this case is called *undersmoothing*. When M is too small, the first quantity (bias) is large but the second quantity (variance) is small; this case is called *oversmoothing*.

To balance the bias and variance, we choose M that minimizes the MSE, which leads to

$$M_{\mathsf{opt}} = \left(\frac{n \cdot L^2}{p(x^*)}\right)^{1/3}. \tag{2.7}$$

Although in practice the quantity L and $p(x^*)$ are unknown so we cannot chose the optimal M_{opt}, the rule in equation (2.7) tells us how we should change the number of bins when we have more and more sample size. Practical rule of selecting M is related to the problem of *bandwidth selection*, a research topic in statistics.