# LECTURE 12: INTERVAL ESTIMATION AND SIMULATION TECHNIQUES

MECO 7312.
INSTRUCTOR: DR. KHAI CHIONG
NOVEMBER 10, 2021

## 1. Interval estimation

In *point estimation*, we report a point value for the unknown parameter.

In *interval estimation*, we report a range/interval of values for the unknown parameter. What range of values should we report?

For example, suppose we are interested in the population mean. A good point estimator is the sample mean $\bar{X}$. There are many interval estimators, such as $[4, 6]$, $[\bar{X} - 1, \bar{X} + 1]$, $[\bar{X} - 5, \bar{X} + 2]$.

Why should we report an interval, why not just the point estimate $\bar{X}$? An interval estimate comes with additional confidence that our assertion is correct.

**Definition 9.1.1:** Consider a model where $\vec{X} = X_1, \ldots, X_n$ has the joint density $f(x_1, \ldots, x_n|\theta)$. An interval estimate for the parameter $\theta$ is a *pair of functions* $L(\vec{X})$ and $U(\vec{X})$ such that $L(\vec{X}) \leq U(\vec{X})$ for all $\vec{X}$. When the observed data is $x_1, \ldots, x_n$, the inference $L(x_1, \ldots, x_n) \leq \theta \leq U(x_1, \ldots, x_n)$ is made.

Note:

- Both $L(\vec{X})$ and $U(\vec{X})$ are random variables, so that $C(\vec{X}) \equiv [L(\vec{X}), U(\vec{X})]$ is a random interval.
- $[L(\vec{X}), U(\vec{X})]$ is a two-sided interval. Sometimes, we seek $(-\infty, U(\vec{X})]$ or $[L(\vec{X}), \infty)$, which are one-sided intervals.

Suppose $X_1, \ldots, X_4 \sim$ i.i.d $N(\mu, 1)$, and we want to estimate the population mean $\mu$. When we use the point estimator $\bar{X}$, the probability that it is correct is actually zero, since $P(\bar{X} = \mu) = 0$. However with an interval estimator, we now have a non-zero probability of being correct. The probability that $\mu$ is covered by the interval $[\bar{X} - 1, \bar{X} + 1]$ is:

$$\begin{aligned} P(\mu \in [\bar{X}-1, \bar{X}+1]) &= P(\bar{X}-1 \leq \mu \ \cap \ \mu \leq \bar{X}+1) \\ &= P(\mu - 1 \leq \bar{X} \leq \mu + 1) \\ &= P(-1 \leq \bar{X} - \mu \leq 1) \\ &= 0.9544 \end{aligned}$$

Where we know that $\bar{X} - \mu \sim \mathcal{N}(0, \frac{1}{4})$. Note that $\mu$ is a constant here, and $\bar{X}$ is the random variable. Therefore in the derivation above, we rewrite the probability explicitly in terms of $\bar{X}$.

Reporting an interval of values comes with a notion of confidence or guarantee.

### 1.1. Coverage probability

Let $X_1, \ldots, X_n \sim$ i.i.d $f(x|\theta)$, where $\theta$ is the unknown parameter of interest.

**Definition 9.1.4:** the **coverage probability** of an interval estimator is $P_\theta\left(\theta \in [L(\vec{X}), U(\vec{X})]\right)$. This is the probability that the random interval $[L(\vec{X}), U(\vec{X})]$ covers the true $\theta$. The probability above is computed using the pdf $f(x|\theta)$, hence its dependence on $\theta$.

In the expression for the coverage probability, $\theta$ is fixed and not random, but $L(\vec{X})$ and $U(\vec{X})$ are random variables. So $P_\theta\left(\theta \in [L(\vec{X}), U(\vec{X})]\right)$ means $P_\theta\left(L(\vec{X}) \leq \theta \cap U(\vec{X}) \geq \theta\right)$.

One problem with the coverage probability is that it can vary depend on what $\theta$ is.

**Definition 9.1.5:** For an interval estimator $[L(\vec{X}), U(\vec{X})]$ of a parameter $\theta$, the **confidence coefficient** $\equiv \min_\theta P_\theta\left(\theta \in [L(\vec{X}), U(\vec{X})]\right)$.

The confidence coefficient does not depend on $\theta$.

Usually, we use the term **confidence interval** to refer to a combination of an interval estimate, along with a measure of confidence (such as the confidence coefficient). Hence, a confidence interval is a statement like "$\theta$ is between 1.5 and 2.8 with probability 80%."

### 1.2. Example

$X_1, \ldots, X_n \sim$ *i.i.d.* $U[0, \theta]$, and $Y_n \equiv \max(X_1, \ldots, X_n)$. Consider two interval estimators

(i) $[aY_n, bY_n]$, where $1 \leq a < b$

(ii) $[Y_n + c, Y_n + d]$, where $0 \leq c < d$.

What is the confidence coefficient of each?

(i) The coverage probability

$$\begin{aligned} P_\theta(\theta \in [aY_n, bY_n]) &= P_\theta(aY_n \leq \theta \leq bY_n) \\ &= P_\theta\left(\frac{\theta}{b} \leq Y_n \leq \frac{\theta}{a}\right). \end{aligned}$$

From before, we know that density of $Y_n$ is $f(y) = \frac{1}{\theta^n} n y^{n-1}$, for $y \in [0, \theta]$, so that

$$\begin{aligned} P_\theta\left(\frac{\theta}{b} \leq Y_n \leq \frac{\theta}{a}\right) &= \frac{1}{\theta^n} \int_{\frac{\theta}{b}}^{\frac{\theta}{a}} n y^{n-1} dy \\ &= \frac{1}{\theta^n}\left[\left(\frac{\theta}{a}\right)^n - \left(\frac{\theta}{b}\right)^n\right] \\ &= \left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n \end{aligned}$$

Since coverage probability is not a function of $\theta$, then this is also confidence coefficient.

Suppose $n = 100$ and we desire a confidence coefficient of 0.95, then one such interval estimator is $[\max(X_1, \ldots, X_n), 1.03 \max(X_1, \ldots, X_n)]$,[1] which is a rather narrow interval. This interval gets narrower as $n$ increases.

(ii) The coverage probability

$$\begin{aligned} P_\theta(\theta \in [Y_n + c, Y_n + d]) &= P_\theta(Y_n + c \leq \theta \leq Y_n + d) \\ &= P_\theta(\theta - d \leq Y_n \leq \theta - c) \\ &= \frac{1}{\theta^n} \int_{\theta-d}^{\theta-c} n y^{n-1} dy = \frac{1}{\theta^n}\left((\theta - c)^n - (\theta - d)^n\right) \end{aligned}$$

so that coverage probability depends on $\theta$.

But note that $\lim_{\theta \to \infty} \frac{1}{\theta^n}\left((\theta - c)^n - (\theta - d)^n\right) = 0$, so that confidence coefficient is 0.

## 2. Methods of Finding Interval Estimators

General principle: "invert" a test statistic.

Consider the following example: $X_1, \ldots, X_n \sim$ *i.i.d.* from a population with mean $\mu$ and variance $\sigma^2$.

[1] $a = 1$, $0.95 = 1 - b^{-n}$, and $b = (0.05)^{-1/n}$.

Consider the test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ The $t$-test statistic is $Z_n = \frac{\sqrt{n}(\bar{X}-\mu_0)}{\hat{\sigma}}$, and we reject the null when $|Z_n| > c$ for a critical value $c$. For a two-sided $t$-test of size 0.05, the critical value is $c = 1.96$, as such, the decision rule is $\mathbb{1}(|Z_n| > 1.96)$.

That is, the rejection region is chosen such that $P(|Z_n| > 1.96 \,\big|\, \mu = \mu_0) = 0.05$. We then have:

$$\begin{aligned} P(-1.96 \leq Z_n \leq 1.96 \,\big|\, \mu = \mu_0) &= 0.95 \\ P\left(-1.96 \leq \frac{\sqrt{n}(\bar{X}-\mu_0)}{\hat{\sigma}} \leq 1.96 \,\middle|\, \mu = \mu_0\right) &= 0.95 \\ P\left(-\frac{1.96\hat{\sigma}}{\sqrt{n}} \leq \bar{X}-\mu_0 \leq \frac{1.96\hat{\sigma}}{\sqrt{n}} \,\middle|\, \mu = \mu_0\right) &= 0.95 \\ P\left(\bar{X}-\frac{1.96\hat{\sigma}}{\sqrt{n}} \leq \mu_0 \leq \bar{X}+\frac{1.96\hat{\sigma}}{\sqrt{n}} \,\middle|\, \mu = \mu_0\right) &= 0.95 \end{aligned}$$

Because the statement above is true for any arbitrary $\mu_0$, it holds true for the true unknown $\mu$, so we can replace $\mu_0$ with $\mu$. Therefore,

$$\begin{aligned} P\left(\bar{X}-\frac{1.96\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{X}+\frac{1.96\hat{\sigma}}{\sqrt{n}} \,\middle|\, \mu = \mu\right) &= 0.95 \\ P\left(\bar{X}-\frac{1.96\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{X}+\frac{1.96\hat{\sigma}}{\sqrt{n}}\right) &= 0.95 \end{aligned}$$

The interval estimator $\left[\bar{X}-\frac{1.96\hat{\sigma}}{\sqrt{n}}, \bar{X}+\frac{1.96\hat{\sigma}}{\sqrt{n}}\right]$ has a coverage probability of 0.95. That is, we say that the 95% confidence interval for $\mu$ is $\left[\bar{X}-\frac{1.96\hat{\sigma}}{\sqrt{n}}, \bar{X}+\frac{1.96\hat{\sigma}}{\sqrt{n}}\right]$

$\mu$ is a constant, and there is an underlying true value. Therefore, it is not correct to say $\mu$ lies within the interval estimator with probability 0.95. Instead, the interpretation should be: the interval estimator $\left[\bar{X}-\frac{1.96\hat{\sigma}}{\sqrt{n}}, \bar{X}+\frac{1.96\hat{\sigma}}{\sqrt{n}}\right]$ covers the true $\mu$ with probability 0.95, due to sampling variation.

### 2.1. Inverting a Likelihood Ratio test

Let $X_1, \ldots, X_n$ be i.i.d from $f(x|\lambda)$, where $f(x|\lambda) = \lambda e^{-\lambda x}$ for $x \geq 0$. This is the Exponential distribution with parameter $\lambda$. We want to derive an interval estimator for $\lambda$.

Consider the Likelihood Ratio Test of $H_0 : \lambda = \lambda_0$ versus $H_1 : \lambda \neq \lambda_0$.

The likelihood function is $L(\lambda|x_1,\ldots,x_n) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$. The Maximum Likelihood estimator is obtained via first order condition as $1/\bar{X}$.

Given $n$ realized random sample $x_1,\ldots,x_n$ from the population, the Likelihood Ratio Test Statistic is:

$$\frac{\lambda_0^n e^{-\lambda_0 \sum_{i=1}^n x_i}}{\bar{x}^{-n} e^{-n}} = (\bar{x}\lambda_0)^n e^n e^{-\lambda_0 n \bar{x}}$$

Suppose the critical value $c$ is such that the test has size 0.05, i.e:

$$P((\bar{X}\lambda_0)^n e^n e^{-\lambda_0 n \bar{X}} \leq c \,|\, \lambda = \lambda_0) = 0.05$$

Using asymptotic approximation, we know that $-2\log\left((\bar{X}\lambda_0)^n e^n e^{-\lambda_0 n \bar{X}}\right) \to \chi_1^2$ under the null hypothesis as $n \to \infty$. Therefore, we can find the critical value $c^*$ as $P(\chi_1^2 \geq -2\log c^*) = 0.05$ or $P(\chi_1^2 \leq -2\log c^*) = 0.95$. The inverse cdf of the chi-squared distribution at 0.95 is 3.8414, which can be found using the Mathematica command `InverseCDF[ChiSquareDistribution[1], 0.95]`. Solving for $c^*$, we get $c^* = 0.1465$.

Now we have:

$$P((\bar{X}\lambda_0)^n e^n e^{-\lambda_0 n \bar{X}} > 0.1465 \,|\, \lambda = \lambda_0) = 0.95$$
$$P((\bar{X}\lambda)^n e^n e^{-\lambda n \bar{X}} > 0.1465) = 0.95$$

In the last line above, we replace $\lambda_0$ with $\lambda$ (because the above holds true for any arbitrary $\lambda_0$, in particular, it holds true for the true unknown $\lambda$).

Therefore, the interval estimator $\{\lambda : (\bar{X}\lambda)^n e^n e^{-\lambda n \bar{X}} > 0.1465\}$ has a 0.95 coverage probability.

How does this interval estimator look like with real data? Suppose our dataset is such that the sample mean is 1.25, and the sample size is $n = 20$. We can use Mathematica's Reduce command to solve for $(\bar{X}\lambda)^n e^n e^{-\lambda n \bar{X}} > 0.1465$ in terms of $\lambda$, which gives us an interval estimate of $0.498628 < \lambda < 1.20359$. Now if our observed sample mean is 2.0, and the sample size is $n = 100$, then we get $0.408297 < \lambda < 0.604503$, a narrower interval around $\frac{1}{2}$.

We can further visualize that the function $(\bar{X}\lambda)^n e^n e^{-\lambda n \bar{X}}$ is unimodal in $\lambda$, and so the interval estimate takes the form of a compact connected set. Knowing that the interval estimate takes the form of a connected interval, it is numerically faster to solve for the root of the equation $(\bar{X}\lambda)^n e^n e^{-\lambda n \bar{X}} = 0.1465$ in terms of $\lambda$.

In addition, there are other ways to solve for the critical value $P((\bar{X}\lambda_0)^n e^n e^{-\lambda_0 n\bar{X}} \leq c \,|\, \lambda = \lambda_0) = 0.05$. For exponential variables, $\sum X_i \sim \text{Gamma}(n, \lambda_0)$, therefore $(\bar{X}\lambda_0)^n e^n e^{-\lambda_0 n\bar{X}}$ is a transformation of the Gamma distribution. We can also simulation to find the density of $(\bar{X}\lambda_0)^n e^n e^{-\lambda_0 n\bar{X}}$.

### 2.2. Another example

Consider the example from the last lecture.

$X_1, \ldots, X_n \sim$ *i.i.d.* Bernoulli with probability $p$. Test $H_0: \; p = p_0$ vs. $H_1: \; p \neq p_0$.

The likelihood function is $L(p|x_1, \ldots, x_n) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}$.

$$\lambda(\vec{X}) = \left(\frac{p_0}{\bar{X}}\right)^{n\bar{X}} \left(\frac{1-p_0}{1-\bar{X}}\right)^{n-n\bar{X}}$$

Using the same asymptotic approxmation:

$$\begin{aligned}
0.05 &= P(\lambda(\vec{X}) \leq 0.1465 \,\big|\, p = p_0) \\
0.95 &= P\left(\left(\frac{p_0}{\bar{X}}\right)^{n\bar{X}} \left(\frac{1-p_0}{1-\bar{X}}\right)^{n-n\bar{X}} > 0.1465 \,\Big|\, p = p_0)\right) \\
0.95 &= P\left(\left(\frac{p}{\bar{X}}\right)^{n\bar{X}} \left(\frac{1-p}{1-\bar{X}}\right)^{n-n\bar{X}} > 0.1465)\right)
\end{aligned}$$

Therefore, the interval estimator for $p$ that has a coverage probability of 0.95 is:

$$\left\{0 \leq p \leq 1 : \left(\frac{p}{\bar{X}}\right)^{n\bar{X}} \left(\frac{1-p}{1-\bar{X}}\right)^{n-n\bar{X}} > 0.1465\right\}$$

To see what this confidence interval looks like, plug in some numbers. Say $\bar{x} = 0.4$ and $n = 10$, then $0.145 < p < 0.700$. With $\bar{x} = 0.1$, $n = 10$, we get a narrower interval: $0.00595 < p < 0.372$. Similarly with $\bar{x} = 0.4$ and $n = 100$, we get $0.307 < p < 0.497$, but with $\bar{x} = 0.01$ and $n = 100$, we get $0.000569 < p < 0.0433$.

```
Reduce[ReplaceAll[((1 - p)/(1 - x))^(n - n x) (p/x)^(n x) >
0.1465, {x -> 0.4, n -> 100}] && p > 0 && 1 > p, p]
```

### 2.3. Bayesian intervals

(*Optional reading)

Confidence interval is defined as the probability that an interval *covers* the parameter, not the probability that the parameter *lies within* the interval. This is to emphasize that the random quantity is the interval, not the parameter.

However, in the Bayesian setup, we have the posterior distribution of the parameter $f(\theta|\boldsymbol{x})$, which allows us to more naturally say what the probability that $\theta$ lies within an interval.

Let $C$ be a 95% credible set for $\theta$, and let $f(\theta|\boldsymbol{x})$ be the posterior distribution, then:

$$\int_C f(\theta|\boldsymbol{x})\, d\theta = 0.95$$

For example, recall $X_1, \ldots, X_n$ are iid $\sim \mathcal{N}(\theta, \sigma^2)$, and suppose that the prior distribution is $\pi(\theta) = \mathcal{N}(\mu, \tau^2)$, assuming that $\tau, \mu, \sigma$ are known. Then the posterior distribution $\theta|\boldsymbol{x} \sim \mathcal{N}\left(\frac{\frac{n}{\sigma^2}\bar{x}+\frac{1}{\tau^2}\mu}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}, \left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\right)$.

In general, because there are many possible intervals $C$ such that $\int_C f(\theta|\boldsymbol{x})\, d\theta = 1-\alpha$, there are many possible $1-\alpha$ credible sets. The simplest one would be symmetric around the posterior mean.

One common way to select among the possible credible sets is to select the shortest credible *intervals*. That is, the shortest length interval $C$ such that $\int_C f(\theta|\boldsymbol{x})\, d\theta = 1-\alpha$. This is special region is called the Highest Posterior Density (HPD) region.

If the posterior density is unimodal, then this is a straightforward task. The $1-\alpha$ HPD region for $\theta$ is $\{\theta : f(\theta|\boldsymbol{x}) \geq k\}$ such that:

$$\int_{\{\theta: f(\theta|\boldsymbol{x}) \geq k\}} f(\theta|\boldsymbol{x})\, d\theta = 1-\alpha$$

## 3. Monte Carlo method (simulation-based methods)

(*Optional reading)

### 3.1. Monte Carlo sampling

Let $X \sim f(x)$. We can approximate $P(a \leq X \leq b)$ using simulations. For example, this arises in hypothesis testing when we wish to compute the power functions.

Draw $x_1, \ldots, x_S$ from the pdf $f(x)$. We are generating artificial sample, the sample size can be as large as our computers permit. Drawing from a density can be done using Inverse Probability Transform, which consists of generating $U[0,1]$ and plugging into the inverse cdf $F^{-1}$. Later, we show another more powerful method (Markov Chain Monte Carlo) to generate draws from any pdf, including multivariate ones.

$$P(a \leq X \leq b) \approx \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}(a \leq x_s \leq b) \tag{1}$$

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx \approx \frac{1}{S} \sum_{s=1}^{S} g(x_s) \tag{2}$$

In general, given the data-generating model $X_1, \ldots, X_n \sim f(x_1, \ldots, x_n)$, and given the statistic or estimator $T(X_1, \ldots, X_n)$. We can approximate the sampling distribution of $T(X_1, \ldots, X_n)$ using Monte Carlo sampling.

For $s = 1, \ldots, S$, draw $\boldsymbol{x}_s = (x_{1s}, \ldots, x_{ns})$ from the joint pdf $f(x_1, \ldots, x_n)$. That is, for the $s$-th experiment/simulation, we draw $(x_1, \ldots, x_n)$ from the joint pdf $f(x_1, \ldots, x_n)$, and label the simulated dataset as $\boldsymbol{x}_s = (x_{1s}, \ldots, x_{ns})$.

$$P(a \leq T(X_1, \ldots, X_n) \leq b) \approx \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}(a \leq T(x_{1s}, \ldots, x_{ns}) \leq b) \tag{3}$$

$$\mathbb{E}[g(T(X_1, \ldots, X_n))] \approx \frac{1}{S} \sum_{s=1}^{S} g(T(x_{1s}, \ldots, x_{ns})) \tag{4}$$

Consider again the example: $X_1, \ldots, X_n \sim$ *i.i.d.* Bernoulli with probability $p$. Test $H_0 : \ p = p_0$ vs. $H_1 : \ p \neq p_0$.

$$\lambda(\vec{X}) = \frac{(p_0)^{n\bar{X}} (1 - p_0)^{n - n\bar{X}}}{\left(\bar{X}\right)^{y_n} \left(1 - \bar{X}\right)^{n - n\bar{X}}}.$$

We can use simulation to determine the exact critical value such that $P(\lambda(\vec{X}) \leq c^*|p = p_0) = 0.05$. We can use simulation to verify that $-2\log\lambda(\vec{X}) \to \chi^2_1$ under the null hypothesis. Given the asymptotic critical value of $c^* = 0.1465$, we can compute the power function $P(\lambda(\vec{X}) \leq c^*)$ as a function of $p$.

## 4. Importance sampling

(*Optional reading)

Importance sampling is a more efficient form of Monte carlo sampling. For example, we want to calculate $P(X > 3)$, where $X \sim \mathcal{N}(0, 1)$. Let $H = \mathbb{1}(X > 3)$, then $P(X > 3) = \mathbb{E}[H] \approx \frac{1}{S}\sum_{s=1}^{S} \mathbb{1}(x_s > 3)$.

Suppose we draw 100 random samples from $\mathcal{N}(0, 1)$, how many are above 3? None! We are "wasting" a lot of draws by not drawing from important regions.

Rather than sampling from $f$, consider sampling from a different probability density function, $g$, as the proposal distribution.

Let $X \sim f(x)$, we have $\mathbb{E}[h(X)] = \int h(x)f(x)\,dx$. Consider some other arbitrary pdf $g(x)$ (integrates to one under the same support as $f(x)$).

$$\mathbb{E}[h(X)] = \int h(x)f(x)\,dx \tag{5}$$

$$= \int h(x)\frac{f(x)}{g(x)}g(x)\,dx \tag{6}$$

$$= \mathbb{E}_g\left[h(x)\frac{f(x)}{g(x)}\right] \tag{7}$$

$\frac{f(x)}{g(x)}$ is the importance weight. In this example, even though we want to compute $P(X > 3)$ for $X \sim \mathcal{N}(0, 1)$, it is more efficient to sample from $\mathcal{N}(3, 2)$ and apply the importance sampling weights. Therefore, draw $x_1, \ldots, x_S$ from the $\mathcal{N}(3, 2)$.

$$P(X > 3) \approx \frac{1}{S}\sum_{s=1}^{S} h(x_s)\frac{f(x_s)}{g(x_s)} \tag{8}$$

Where $f$ is the density of $\mathcal{N}(0, 1)$ and $g$ is the density of $\mathcal{N}(3, 2)$, and $h(x_s) = \mathbb{1}(x_s > 3)$. Compare the accuracy of monte carlo integration with and without importance weight, and compare them to the ground truth.

Another example, you want to simulate the mean of a standard normal distribution, truncated to the unit interval [0,1]. That is, $\mathbb{E}[X|X \in [0,1]]$. The brute-force way is to sample from $\mathcal{N}(0,1)$ and throw away those samples outside of $[0,1]$, i.e.

$$\mathbb{E}[X|X \in [0,1]] \approx \frac{\sum_{s=1}^{S} x_s \mathbb{1}(0 < x_s < 1)}{\sum_{s=1}^{S} \mathbb{1}(0 < x_s < 1)} \tag{9}$$

Importance sampling: draw from $U[0,1]$, so that $g(x) = 1$ for $x \in [0,1]$. The sampling density of $X|X \in [0,1]$ is:

$$f(x) = \frac{\phi(x)\mathbb{1}(x \in [0,1])}{\int_0^1 \phi(x)\,dx} \tag{10}$$

For each draw, the importance weight is $w^s = f(x^s) = \frac{\phi(x^s)}{\int_0^1 \phi(x)\,dx} = \frac{\phi(x^s)}{0.34135}$. The simulated mean is $\frac{1}{S}\sum x_s w_s$.

Probabilities and expectation involving multivariate Normal is extremely difficult to compute, it involves multi-dimensional integration. Importance sampling is crucial here. In fact, it has been given a name – GHK simulator – enables us to efficiently draw from truncated multivariate normal distribution.

## 5. Bootstrap methods

This section is accompanied by the R Markdown "bootstrap.Rmd" or the Python Jupyter Notebook.

Given a random sample $X_1, \ldots, X_n$, consider a statistic $T(X_1, \ldots, X_n)$. Everything that we have done so far involved a statistic: (1) estimator of a population parameter, (2) test statistic of a hypothesis test, etc.

Statistical inference relies on knowing the sampling distribution of $T(X_1, \ldots, X_n)$. For example, to evaluate $\mathbb{E}[T(X_1, \ldots, X_n)]$ or $\text{Var}[T(X_1, \ldots, X_n)]$. For hypothesis testing, we also need to know the distribution of the test statistic in order to determine the rejection region.

There are several ways to determine the sampling distribution.

(i) Make assumption about the data-generating process. Assume that the data are generated from a family of distributions $f(x_1, \ldots, x_n|\theta)$. Estimate the unknown parameter $\theta$ (or in hypothesis testing, set it to the null value). Then, use Monte Carlo simulation to sample from $f(x_1, \ldots, x_n|\hat{\theta})$ to determine the sampling distribution of $T(X_1, \ldots, X_n)$.

(ii) Simple transformation of random variables. In some cases, $Y = T(X_1, \ldots, X_n)$ is a simple transformation (convolution) of random variables, for example, sum of independent exponentials is a gamma distribution, sum of independent normals is a normal distribution.

(iii) Asymptotic approximation. $T(X_1, \ldots, X_n)$ might have a known asymptotic distribution. If $T$ is the sample mean, then it is asymptotically normal. LR test statistic has a chi-squared distribution asymptotically, Maximum Likelihood estimator is asymptotically Normal with variance equals to the inverse of the Fisher information matrix, etc.

(iv) Bootstrapping. Also, called non-parametric bootstrapping, to emphasize that we do not need to make specific assumptions about the form of the data-generating process.

### 5.1. Bootstrap algorithm

(1) Given sample $x_1, \ldots, x_n$. Treat sample as if it is the population.

(2) Draw $n$ random samples *with replacement* from $x_1, \ldots, x_n$. Call this a bootstrapped sample $\boldsymbol{x}^* = (x_1^*, \ldots, x_n^*)$.

(3) Draw $B$ number of bootstrapped samples $\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots, \boldsymbol{x}_B^*$. Each bootstrapped sample $\boldsymbol{x}_b^*$ has $n$ observations.

(4) Compute the statistic $T$ using the bootstrapped samples, that is, $(T(\boldsymbol{x}_1^*), \ldots, T(\boldsymbol{x}_B^*))$.

(5) The empirical sampling distribution of the statistic $T$ is approximated by $(T(\boldsymbol{x}_1^*), \ldots, T(\boldsymbol{x}_B^*))$.

Example. Consider the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$. We want to know the distribution of the sample variance. We could assume that $X_i$ is Normally distributed, so that $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. We could rely on asymptotics: the sample variance is asymptotically normal with $\sqrt{n}(s^2 - \sigma^2) \xrightarrow{d} \mathcal{N}(0, \mu_4 - \sigma^4)$, where $\mu_4 = \mathbb{E}[(X - \mathbb{E}[X])^4]$, and $\sigma^4 = \mathrm{Var}(X)^2$.

However, if we are unwilling to assume a data-generating process for $X_i$, and if the sample size is not large, then we can try bootstrapping.

Bootstrapping is very useful in constructing hypothesis test and confidence intervals. We see from our simulation exercise that bootstrapping becomes more accurate when the sample size is large. When the sample size is small, bootstrapping can be very misleading. Now if the sample size is large, why don't we just use the asymptotic distribution instead? In this example, the statistic of interest is the sample variance,

so the asymptotic distribution is known – there are many cases where the asymptotic sampling distribution is not known.