HA1 for Monte Carlo and Empirical Methods

Jingmo Bai Zuoyi Yu

February 7, 2023

1 Random number generation

In this problem, X is a random variable on $\mathbb{R}$. The probability density f_x, invertible distribution function F_x, and the inverse are assumed to be known. I = (a, b) is an interval such that $\mathbb{P}(X \in I{>}0)$.

1.1

Our task is to find the conditional distribution function $F_{X|X\in I}(x) = P(X \leq x|X \in I)$ and density $f_{X|X\in I}(x)$ of X given that $X \in I$. We know that when $a < \ x < \ b$, $P(X \in I){>}0$, so $F_{X|X\in I}(x) = 0$ when $x \leq a$, and $F_{X|X\in I}(x) = 1$ when $x \geq b$. We can obtain the conditional distribution function given that $X \in I$ by

$$F_{X|X\in I}(x) = \mathbb{P}(X \leq x|a < \ x < \ b) = \frac{\mathbb{P}(X \leq x, a < \ X < \ b)}{\mathbb{P}(a < \ X < \ b)} = \frac{\mathbb{P}(a < \ X < \ x)}{\mathbb{P}(a < \ X < \ b)} = \frac{F_X(x) - F_X(a)}{F_X(b) - F_X(a)}$$

The density function $f_{X|X\in I}(x)$ can be obtained by the derivative of distribution function $F_{X|X\in I}(x)$. We know that when $x \leq a$ and $x \geq b$, $f_{X|X\in I}(x) = 0$. When $a < \ x < \ b$,

$$f_{X|X\in I}(x) = \frac{d}{dx}(F_{X|X\in I}(x)) = \frac{F_X(x)}{F_X(b) - F_X(a)}$$

1.2

Our task is to find the inverse $F^{-1}_{X|X\in I}$ given that $X \in I$.
By the definition of inverse function: $F_X(F_X^{-1}(x)) = x$

$$F_{X|X\in I}(x) = \frac{F_X\ (x) - F_X(a)}{F_X\ (b) - F_X(a)}$$

Let $x = F^{-1}_{X|X\in I}(y)$

$$\begin{aligned}
F_{X|X\in I}(F^{-1}_{X|X\in I}\ (y)) = y &= \frac{F_X\left(F^{-1}_{X|X\in I}\ (y)\right) - F_X\ (a)}{F_X\ (b) - F_X\ (a)} \\
F_X\left(F^{-1}_{X|X\in I}\ (y)\right) &= y\left\{F_X\ (b) - F_X\ (a)\right\} + F_X\ (a) \\
{F_X}^{-1}\left\{F_X\left(F^{-1}_{X|X\in I}(y)\right)\ \right\} &= \ {F_X}^{-1}\left\{y\left\{F_X\ (b) - F_X\ (a)\right\}\ +\ F_X\ (a)\right\} \\
F^{-1}_{X|X\in I}(y) &= \ {F_X}^{-1}\left\{y\left\{F_X\ (b) - F_X\ (a)\right\}\ +\ F_X\ (a)\right\} \\
F^{-1}_{X|X\in I}(x) &= \ {F_X}^{-1}\left\{x\left\{F_X\ (b) - F_X\ (a)\right\}\ +\ F_X\ (a)\right\}
\end{aligned}$$

We can use the inverse transform method to sample the random values of for a random variable X conditionally on the interval $X \in I$, with u is uniform random variable on (0,1), and the distribution function F is known and invertible. The random sample x of X can be obtained by

$$x = \ F^{-1}(u)$$

2 Power production of a wind turbine)

The Two-sided confidence interval can be calculated through the following equation:

$$I_\alpha = \tau_N \pm \lambda_{\alpha/2} \frac{\sigma(\phi)}{\sqrt{N}} \tag{1}$$

In this assignment, all the questions require 95% confidence interval, so according to the confidence interval table, we could know all $I_\alpha = 1.96$. And we choose the number size 10,000 for all problems. Therefore, when we calculate the confidence interval, we just need find $\sigma(\phi)$, which is the standard deviation of objective function, and τ_N.

2.1

We begin with finding confidence interval using standard Monte Carlo method. For standard MC, we need a large amount number of independent random numbers that follow Weibull distribution, which could be drawn with *wblrnd* function automatically in MATLAB. Then, by the law of large numbers, when N is a infinite number,

$$\tau_N = \frac{1}{N} \sum \phi(X_i) \to \mathbb{E}(\phi(X)) \tag{2}$$

Therefore, we can get τ_u through calculating the expectation of independent random generated numbers. The results are shown in Table 1.

Table 1: 95% confidence intervals and widths for the Standard Monte Carlo Method

Month	Lower Bound	Upper Bound	Width
Jan	4582921.0614	4726576.096	143655.0346
Feb	4064575.6228	4204531.8085	139956.1856
Mar	3753857.6383	3889453.4021	135595.7638
Apr	2937319.8372	3062786.1785	125466.3413
May	2839704.3252	2963136.9218	123432.5967
Jun	2991527.3377	3118322.4509	126795.1132
Jul	2813536.737	2936355.9415	122819.2045
Aug	3051737.1741	3179466.435	127729.261
Sep	3644665.5663	3779222.6107	134557.0445
Oct	4106885.2264	4249193.9653	142308.7389
Nov	4562427.4232	4705032.6726	142605.2494
Dec	4626735.2607	4770426.9585	143691.6978

For the truncated version, what we can learn from Problem 1 is that when the random variable X is set between an interval, we cannot just use *wblrnd* to generate independent random numbers but need some calculation. Therefore, from the result we get in 1.2,

$$x = F^{-1}(F(a) + u(F(b) - F(a)) \tag{3}$$

where F^{-1} could be found through *wblinv* in MATLAB, a and b are the lower and upper limit of wind speed(3.5 and 25 in this case), F(x) is the CDF which could be calculated by *wblcdf* in MATLAB.
After generating independent random variable X as above, we can calculate the confidence interval as for standard Monte Carlo method. The results for truncated version are shown in Table 2.
From Table 1 and Table 2, we can find the widths for the truncated version are smaller than the widths for standard Monte Carlo method, since standard Monte Carlo will generate random numbers out of the speed interval which will generate zero power. Therefore, it can say that when we are given an input interval, truncated version will have a better performance than the standard Monte Carlo method.

2.2

When using control variate to decrease the variance, we assume that we have another random variable Y, which we know $\mathbb{E}(Y) = m$ and Y has the same complexity as $\phi(X)$. Then we set some $\beta \in mathbbR$ and Z that follow the equation:

$$Z = \phi(X) + \beta(Y - m) \tag{4}$$

where X is the random numbers follow the Weibull distribution, m is given by $\mathbb{E}[V^m] = \Gamma(1 + m/k)\lambda^m$

Table 2: 95% confidence intervals and widths for the Truncated Version

Month	Lower Bound	Upper Bound	Width
Jan	4624182.0262	4746619.7559	122437.7296
Feb	4086875.4486	4204339.0752	117463.6266
Mar	3761755.2615	3875240.0968	113484.8353
Apr	2959134.5464	3060430.3388	101295.7923
May	2876975.7207	2975886.4196	98910.6989
Jun	3050924.4984	3153031.6951	102107.1968
Jul	2834559.8949	2932576.9679	98017.073
Aug	2952823.0697	3054822.0426	101998.9729
Sep	3750571.7572	3863831.4495	113259.6923
Oct	4147505.2476	4266328.9767	118823.7291
Nov	4570363.2594	4692627.5213	122264.2619
Dec	4550642.7667	4673016.4873	122373.7206

Month	Lower Bound	Upper Bound	Width
Nov	4624862.1392	4689614.2994	64752.1602
Dec	4638451.211	4704125.4754	65674.26446

Then we can find τ_N through

$$\mathbb{E}(Z) = \mathbb{E}(\phi(X) + \beta(Y - m)) = \tau_N \tag{5}$$

In this case, $\phi(X)$ is given by P(X), m can be calculated directly, so all what we need to do is to find out β. To find the optimal β, we can calculate the variance of Z:

$$\mathbb{V}(Z) = \mathbb{V}(\phi(X) + \beta Y) = \mathbb{V}(\phi(X)) + 2\beta\mathbb{C}(\phi(X), Y) + \beta^2\mathbb{V}(Y) \tag{6}$$

We want $\mathbb{V}(Z)$ could be as close to $\mathbb{V}(\phi(X))$ as possible, so we need find a β that makes $2\beta\mathbb{C}(\phi(X), Y)+\beta^2\mathbb{V}(Y)$ as small as possible, which can be calculated by differentiating:

$$0 = 2\mathbb{C}(\phi(X), Y) + 2\beta\mathbb{V}(Y) \leftrightarrow \beta = \beta^* = -\frac{\mathbb{C}(\phi(X), Y)}{\mathbb{V}(Y)} \tag{7}$$

where $\beta^* is the optimal \beta we need.$ After finding out m and β^*, we calculate new variable Z according to Equation 4 and then get τ_N using control variate. The 95% confidence interval and widths can be found in Table 3.

Table 3: 95% confidence intervals and widths after using control variate to decrease variance

Month	Lower Bound	Upper Bound	Width
Jan	4624962.2045	4691764.9128	66802.7082
Feb	4132204.944	4181685.6904	49480.7463
Mar	3803718.5947	3850701.43	46982.8353
Apr	2974766.7896	3013960.9563	39194.1667
May	2847733.9568	2885895.3585	38161.4017
Jun	3055768.7132	3096407.0391	40638.3258
Jul	2845001.871	2882524.1392	37522.2682
Aug	3081144.5956	3119601.5463	38456.9506
Sep	3717342.8651	3763390.7636	46047.8986
Oct	46047.8986	4243801.5009	61600.266

As Table 3 shown, it is obviously that control variate method decrease the widths a lot.

2.3

In the importance sampling method, the most important part is to find an instrumental density g on X such that:

$$g(x) = 0 \rightarrow f(x) = 0 \tag{8}$$

Then we can rewrite the integral as

$$\tau_N = \mathbb{E}_f(\phi(X)) = \int_X \phi(x)f(x)dx = \int_{g(x)>0} \phi(x)\frac{f(x)}{g(x)}g(x)dx = \mathbb{E}_g(\phi(X)\frac{f(X)}{g(X)}) = \mathbb{E}_g(\phi(X)\omega(X)) \tag{9}$$

where

$$\omega : \{x \in X : g(x) > 0\} \ni x \to \frac{f(x)}{g(x)} \tag{10}$$

Then, we can estimate τ_N using standard MC as 2.1. Therefore, we need to find out the function g(x). Ideally, we wan to choose the function g(x) that makes $\phi(x)\frac{f(x)}{g(x)}$ is a constant, which means the variance of the approximation error of this equation is close to zero. To make $\phi(x)\frac{f(x)}{g(x)}$ a constant represents that g(x) should have similar shape as the function $\phi(x)f(x)$. Therefore, we plot the figure of $\phi(x)f(x)$ in Figure 1, where the figure looks just like a normal distribution. So we choose the g(x) as a normal distribution, with the mean same as joint function $\phi(x)f(x)$ and tune σ by hand to make two curves similar. Our final choice of parameters for g(x) is mu = 12, σ = 5.099.

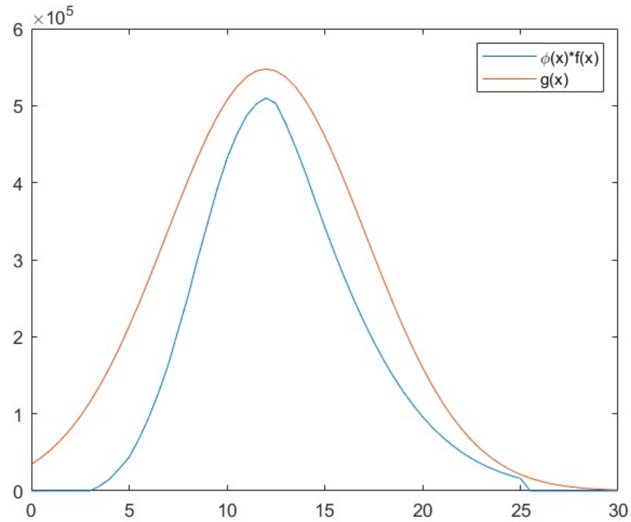

Figure 1: Plot of $\phi(x) * f(x)$ and chosen g(x)

After finding out the optimal g(x), since we assume that independent random variable X follows the instrumental density g(x), which is a normal distribution function, we can generate random variable through using function *randn* multiply with σ and then plus mu. Then we can calculate τ_N according to Equation 9, the results are shown in Table 4.

Table 4: 95% confidence intervals using importance sampling based on normal distribution

Month	Lower Bound	Upper Bound	Width
Jan	4634638.1521	4702580.6947	67942.5426
Feb	4115259.225	4174735.7664	59476.5414
Mar	3785737.7003	3840312.4666	54574.7662
Apr	2997683.8954	3028652.8421	30968.9467
May	2865173.9505	2889888.5747	24714.6242
Jun	3058156.5719	3090678.7485	32522.1766
Jul	2861527.9375	2887209.62	25681.6824
Aug	3057366.8372	3089602.4853	32235.648
Sep	3732070.357	3785557.6856	53487.3286
Oct	4198679.7933	4248305.6127	49625.8194
Nov	4630439.1265	4698792.7997	68353.6732
Dec	4610994.5794	4679362.1298	68367.5504

From Table 4, we can find that IS method has a obvious narrower confidence interval than standard MC and truncated version. And IS also has the general same confidence interval as control variate method.

2.4

In antithetic sampling, we assume another two variables V and $\tilde{V}$. For V, we set V = $\phi(X)$ to make $\tau = \mathbb{E}(V)$. For $\tilde{V}$, we set $\tilde{V}$ has the same complexity as V, $\mathbb{E}(\tilde{V}) = \tau$ and $\mathbb{V}(\tilde{V}) = \mathbb{V}(V) = \sigma^2(\phi)$. Then define

$$W = \frac{V + \tilde{V}}{2} \tag{11}$$

where $\mathbb{E}(W) = \tau$ and $\mathbb{V}(W) = \mathbb{V}(\frac{V+\tilde{V}}{2}) = \frac{1}{2}(\mathbb{V}(V) + \mathbb{C}(V, \tilde{V}))$ According to the slides in Page 19 in L4, we want to find $\mathbb{V}$ such that V and $\mathbb{V}$ are negatively correlated, which could be done using the application of the theorem in Page 20: Let F be a distribution function and ϕ a monotone function. Then, we set $U \sim \mho(0, 1), T(u) = 1 - u, \varphi(u) = \phi(F^{-1}(u))$, then we can get

$$V = \phi(F^{-1}(U)), \tilde{V} = \phi(F^{-1}(1 - U)) \tag{12}$$

After getting V and $\tilde{V}$, we can get W and then get τ_N. The final results are shown in Table 5.

Table 5: 95% confidence intervals using antithetic sampling to decrease variance

Month	Lower Bound	Upper Bound	Width
Jan	4651215.08	4669322.6178	18107.5378
Feb	4124556.3324	4149815.3737	25259.0413
Mar	3798210.0935	3829293.4022	31083.3087
Apr	2981179.7145	3025266.0974	44086.3829
May	2824546.5282	2869456.6989	44910.1707
Jun	3087993.6251	3131507.6017	43513.9765
Jul	2840955.801	2886435.9499	45480.1488
Aug	3074976.741	3118522.3127	43545.5718
Sep	3749925.0107	3782174.1453	32249.1346
Oct	4196008.0552	4221946.4893	25938.4341
Nov	4643668.804	4664182.9556	20514.1516
Dec	4644612.8408	4664066.28398	19453.443

As we can see in Table 5, antithetic sampling performs better than standard Monte Carlo and truncated version very much. Compared with controlling variate and IS, it has wider confidence interval in the several middle months but on average antithetic sampling has the smallest interval. So we can say antithetic sampling has the best performance through these methods in this case.

2.5

As we are estimating the probability tat the turbine delivers power, we need generate independent random numbers using *wblrnd* function in MATLAB first, then calculate corresponding power according to the function P. What we want to know is $\mathbb{P}(P(V) > 0$, so we can find out the number of $P(V) \neq 0$ divided by the total number of power, which is just the number of generated random numbers. The results are shown in Table 6.

Table 6: The probability that the turbine delivers power

Month	Probability
Jan	0.8962
Feb	0.8727
Mar	0.8633
Apr	0.8133
May	0.8067
Jun	0.8205
Jul	0.7984
Aug	0.8249
Sep	0.8643
Oct	0.8717
Nov	0.8933
Dec	0.8926

2.6

In this problem we want to estimate

$$\frac{\mathbb{E}P(V)}{\mathbb{E}P_{tot}(V)} \tag{13}$$

where P_{tot} could be calculated directly through given equations: $P_{tot}(v) = \frac{1}{2}\rho\pi\frac{d^2}{4}v^3$ and d = 164, $\rho = 1.225kg/m^3, v^3 = \Gamma(1+3/k)\lambda^3$ Then, $\mathbb{E}P(V)$ is the mean power we calculate before, where V is the independent random variable generated with *wblrnd* in MATLAB. The 95% confidence intervals of the average ration are presented in Table 7.

Table 7: 95% confidence intervals for the average ratio of actual wind turbine output to the total power

Month	Lower Bound	Upper Bound	Width
Jan	0.22195	0.22895	0.007001
Feb	0.25535	0.26421	0.0088595
Mar	0.28594	0.29615	0.010207
Apr	0.31612	0.32958	0.013458
May	0.3247	0.33895	0.014251
Jun	0.31291	0.32605	0.013137
Jul	0.31893	0.33295	0.014021
Aug	0.30874	0.32176	0.013021
Sep	0.28594	0.29636	0.010415
Oct	0.23452	0.2426	0.0080803
Nov	0.22575	0.23277	0.0070166
Dec	0.22169	0.2287	0.0070111

2.7

In the last problem, we need to calculate two factors. For the *capacity factor*, it can be calculated through $\frac{\mathbb{E}P(V)}{9.5MW}$, where $\mathbb{E}P(V)$ is the mean of actual power we calculate before. For the *availability factor*, it is just the mean ratio of the result we calculate in 2.e. The results of these two factors are shown in Table 8.

Table 8: Capacity Factor and Availability Factor

capacity factor	availability factor
0.3936	0.8516

We can see although capacity factor is among 20%-40%, the availability is less than 90%. So this seems not to be a good site to build a wind turbine.

3 Combined power production of two wind turbines

3.1

To prove the joint expectation is n one dimensional problem is equal to prove the equation:

$$\mathbb{E}(P(V_1) + P(V_2)) = \mathbb{E}(P(V_1)) + \mathbb{E}(P(V_2)) \tag{14}$$

In this case, two wind turbines are placed in the same area and exposed to similar winds V_1 and V_2 follow the same Weibull distribution. Therefore, the expectation of the power of wind speed should be equal:

$$\mathbb{E}(P(V_1)) = \mathbb{E}(P(V_2)) \tag{15}$$

Then,

$$\mathbb{E}(P(V_1) + P(V_2)) = \mathbb{E}(P(V_1) + P(V_1)) = \mathbb{E}(2P(V_1)) = 2\mathbb{E}(P(V_1)) = \mathbb{E}(P(V_1)) + \mathbb{E}(P(V_2)) \tag{16}$$

Therefore, this joint expectation could reduce to a one dimensional problem, and to estimate the joint expectation is only to estimate $\mathbb{E}(P(V_1))$ or $\mathbb{E}(P(V_2))$ using important sampling as what we do in 2.c. The curves of the joint function $\phi(x) * f(x)$ and instrumental density function g(x) is shown in Figure 2, and our final choice for g(x) is $mu = 11, \sigma = 4.4$.

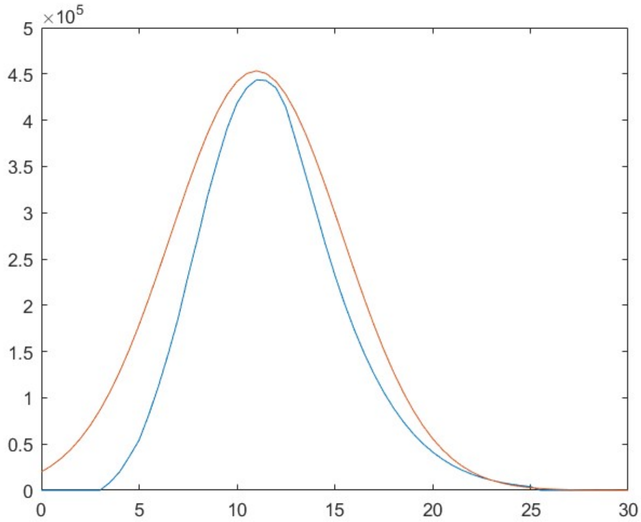

Figure 2: Plot of $\phi(x) * f(x)$ and chosen g(x)

Then use the IS method with detail above to calculate the expectation. The final joint expecatation is:

$$\mathbb{E}(P(V_1) + P(V_2)) = 7.5MW \tag{17}$$

3.2

According to the definition of the covariance:

$$\mathbb{C}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \tag{18}$$

which is equal to the following equation in our case:

$$\mathbb{C}(P(V_1), P(V_2)) = \mathbb{E}(P(V_1)P(V_2)) - \mathbb{E}(P(V_1))\mathbb{E}(P(V_2)) \tag{19}$$

Since we have found $\mathbb{E}(P(V_1))$ and $\mathbb{E}(P(V_2))$, we only need calculate $\mathbb{E}(P(V_1)P(V_2))$, which implies a joint expectation. Therefore our goal is to find g(x,y) that makes $\phi(x) * \phi(y) * f(x, y)$ over g(x, y) is a constant. Even though this is a two-dimensional problem, the method for finding IS parameters is still the same as before. Our final choice for this joint function is mu = 12, $\sigma = 41$. Then we can use *mvnrnd* in MATLAB to generate multivariate normal random numbers X and Y with given mu and σ. Then use *mvnpdf* to generate multivariate normal probability density function g(x,y). The expectation of the joint function in this case is:

$$P(V_1)P(V_2) = P(X)P(Y)\frac{f_xy(X, Y)}{g(X, Y)} \tag{20}$$

The final covariance of $\mathbb{C}(P(V_1), P(V_2))$ is:

$$\mathbb{C}(P(V_1), P(V_2)) = 6.732e12 \tag{21}$$

3.3

According to the definition of variability

$$\mathbb{V}(P(V_1) + P(V_2)) = \mathbb{V}(P(V_1)) + \mathbb{V}(P(V_2)) + 2\mathbb{C}(P(V_1), P(V_2)) \tag{22}$$

As V_1 and V_2 are similar wind speed,

$$\mathbb{V}(P(V_1)) = \mathbb{V}(P(V_2)) \tag{23}$$

where V_1 and V_2 are both followed Weibull distribution. Therefore, we can generate random numbers through *wblrnd* in MATLAB and apply to P function, then calculate the variance for V. Using the covariance we get in the last problem, we can find out variability according to Equation 22. Then the standard deviation is just the root of the variability. The final result is shown in Table 9.

Table 9: Variability and Standard Deviation

variability	standard deviation
3.76e13	6.13e6

3.4

In this question, our task is to find 95% confidence interval for the probability $\mathbb{P}(P(V_1)+P(V_2) > \ 9.5MW)$ and $\mathbb{P}(P(V_1)+P(V_2) < \ 9.5MW)$ which are the probability of the total power being greater and smaller than half of the installed capacity respectively. We use importance sampling method as variance reduction technique. We denote two target function ϕ_1 and ϕ_2, for those two conditions. To find the probability that the total power are greater or less than 9.5 MW, we just take the expectation of our target function times the importance weight function

$$\begin{aligned}\mathbb{P}[P(X)+P(Y) > \ 9.5MW] = E[\phi_1(X,Y)\frac{f(X,Y)}{g(X,Y)}] \\ \mathbb{P}[P(X)+P(Y) < \ 9.5MW] = E[\phi_2(X,Y)\frac{f(X,Y)}{g(X,Y)}]\end{aligned} \tag{24}$$

For the probability of power less than 935MW, we still use multivariate normal distribution as g(X, Y) as previous. For the probability greater than 9.5MW, we still plot the function of $\phi \times f$ and g. The mean is still the value which obtain the maximum value of $\phi \times f$. However, as the region for greater ones is infinite, we want to find the g function decays slower than $\phi \times f$. Our final choice is $mu_2 = 12, \sigma_2 = 50$, and the plot of these two functions are shown in Figure 3.

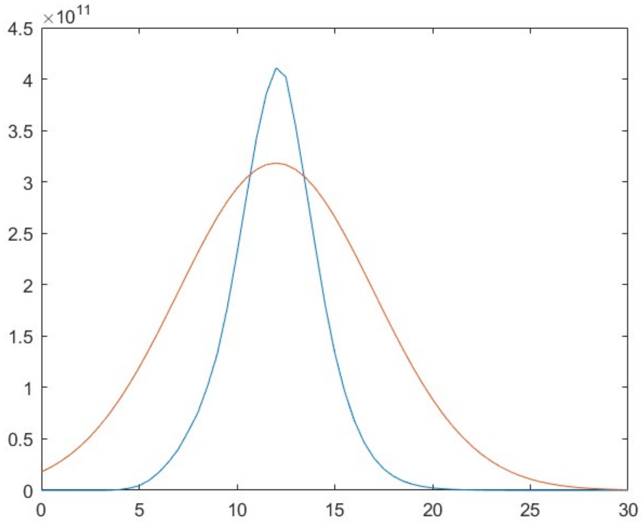

Figure 3: Plot of $\phi(x) * f(x)$ and chosen g(x)

Then we can get the expectations for both conditions which are just the probability we want and calculate their 95% confidence interval respectively. The final results are shown in Table 10

Table 10: The probability for two conditions and their 95% confidence intervals

Condition	Upper Bound	Lower Bound	width	Probability
$P < \ 9.5MW$	0.6209	0.6041	0.0168	0.6125
$P > \ 9.5MW$	0.3780	0.3589	0.0191	0.3684

As we can see from the above table, the sum of these two probabilities is not equal to 1. The main reason of this problem is the way we estimate the parameters, there must be some bias so we cannot get the perfect answer. What's more, there should be some probability that the the generated power is exact equal to 9.5MW.

4 References

Magnus Wiktorsson, Lecture slides, FMSN50, Monte Carlo and Empirical Methods for Stochastic Inference, Lund University