Statistical Computing and Simulation

統碩一 106354003 林健宏

Question 01. Find I by calculus techniques

$$I = \int_{-20}^{20} e^{-|x|} dx$$

< Sol >

$$\int_{-20}^{20} e^{-|x|} dx = 2\int_{0}^{20} e^{-x} dx = 2\left(-e^{-x}\Big|_0^{20}\right)$$

$$\because e^{-20} \approx 0, \therefore \; -e^{-x}\Big|_0^{20} = 1 - e^{-20} \approx 1$$

$$I = \int_{-20}^{20} e^{-|x|} dx = 2\int_{0}^{20} e^{-x} dx = 2\left(-e^{-x}\Big|_0^{20}\right) = 2\left(1 - e^{-20}\right) \approx 2$$

Question 02. Simulation Study

(1)

Find mean and variance of importance sampling estimates using proposal $\phi(x,0,1)$ and $\phi(x,0,5)$ or (more proposals). Compare it with that Monte Carlo.

(Which proposal gives a smaller variance? Give arguments to support your finding with using numbers and graphs.)

< Sol >

Derive of important sampling:

$$E_f[h(x)] = \int h(x)f(x)dx, where\ f\ is\ a\ pdf$$

$$E_f[h(x)] = \int h(x)f(x)dx = \int h(x)\frac{f(x)}{g(x)}g(x)dx = E_g[h(x)\frac{f(x)}{g(x)}]$$

$$g\ is\ another\ pdf\ and\ \frac{f(x)}{g(x)}$$

$$\therefore \mathrm{E_f}[h(x)] = E_g[h(x)\frac{f(x)}{g(x)}]$$

$$\mathrm{Var_f[h(x)]} = E_f[h(x)^2] - E_f[h(x)]^2$$

$$\mathrm{Var_g}\left[h(x)\frac{f(x)}{g(x)}\right] = E_g\left[\left(h(x)\frac{f(x)}{g(x)}\right)^2\right] - \mathrm{E_g}\left[h(x)\frac{f(x)}{g(x)}\right]^2$$

So if $\mathrm{I} = E_f[h(x)] = E_g[h(x)\frac{f(x)}{g(x)}]$, which is better method ? → The smaller variance is better.

$$\mathrm{Var_f[h(x)]} - \mathrm{Var_g}\left[h(x)\frac{f(x)}{g(x)}\right] =$$

$$E_f[h(x)^2] - E_f\,[h(x)]^2 - E_g\left[\left(h(x)\frac{f(x)}{g(x)}\right)^2\right] + \mathrm{E_g}\left[h(x)\frac{f(x)}{g(x)}\right]^2$$

$$\because \mathrm{E_f}[h(x)] = E_g\left[h(x)\frac{f(x)}{g(x)}\right] = \mathrm{I}$$

$$\therefore \mathrm{Var_f[h(x)]} - \mathrm{Var_g}\left[h(x)\frac{f(x)}{g(x)}\right] = E_f[h(x)^2] - E_g\left[\left(h(x)\frac{f(x)}{g(x)}\right)^2\right] = \mathrm{E_g[h(x)^2}\frac{f(x)}{g(x)}(1-\frac{f(x)}{g(x)})]$$

If $\frac{f(x)}{g(x)}(1-\frac{f(x)}{g(x)}) > 0 \rightarrow 1 > \frac{f(x)}{g(x)}$, $E_g[h(x)\frac{f(x)}{g(x)}]$ is better than $E_f[h(x)]$

So, it is important to choose g(x).

這個方法為 Important sampling，若是能找到一個適合的 Important function，便能達到縮減變異數的效果。

下表我們使用 Mnote Carlo、Important sampling N(0,1)以及 Important sampling N(0,5)去估計$I = \int_{-20}^{20} e^{-|x|}dx$

（使用 1000 個樣本去估計一個 I，並總共估計 5000 個 I，計算這 5000 個 估計量 I 的樣本平均與樣本變異數）

	Monte Carlo	Important sampling N(0,1)	Important sampling N(0,5)
Mean	1.99851	1.98926	1.99872
Variance	0.03619	0.14084	0.00865

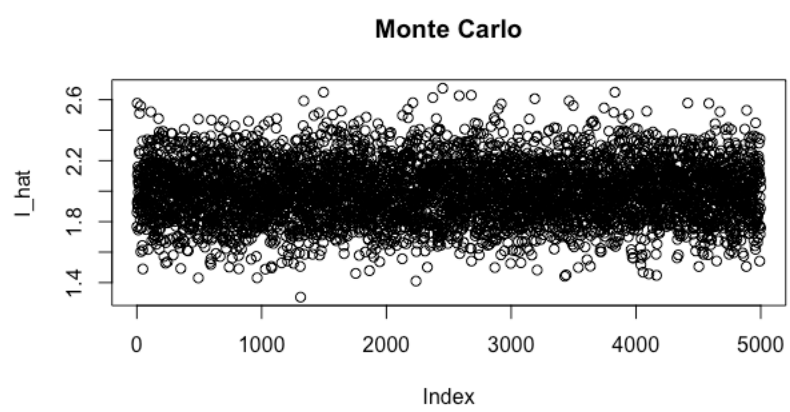
小結：

三種方法用於估計$I = \int_{-20}^{20} e^{-|x|}dx$都很接近我們使用微積分機算的數值 2，符合理論的推導，三種方法皆為不偏估計的方法。為了比較方法的好壞，我們比較三種方法的變異數。

由上表可以清楚看出三種方法變異數的差異，最大的為 Important sampling 選擇 N(0,1)當作 important function 時得到的變異數 0.14084 以及最小的為 Important sampling 選擇 N(0,5)當作 important function 時得到的變異數 0.00865。可見 Important sampling 不一定可以降低變異數，反而會造成變異數變大。

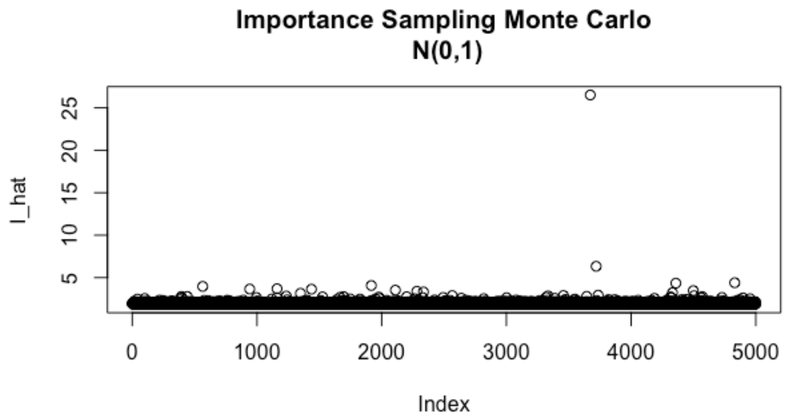
我們以圖表解釋為何差異這麼大，將 5000 個估計值，每一個估計值由 1000 個樣本構成，做成散佈圖，看三種方法的的估計值浮動範圍差異。

(a) Monte Carlo



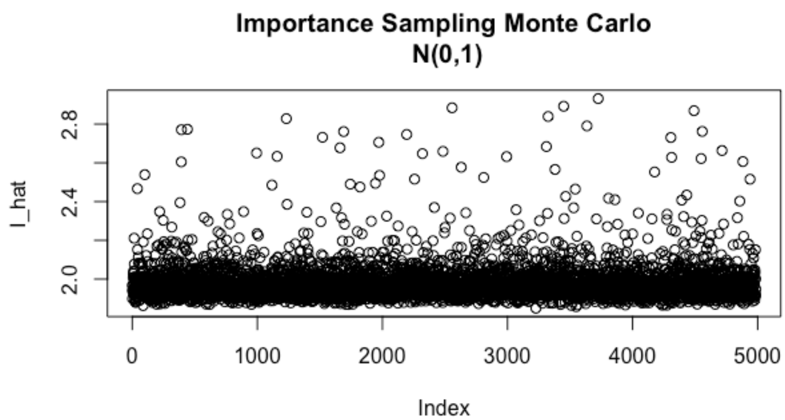
由上圖可見，使用 Monte Carlo 方法去估計$I = \int_{-20}^{20} e^{-|x|}dx$時，估計值的範圍大約為 (1.4, 2.6)，其中大部分都集中在(1.8, 2.3)。

(b) Important sampling with N(0,1)



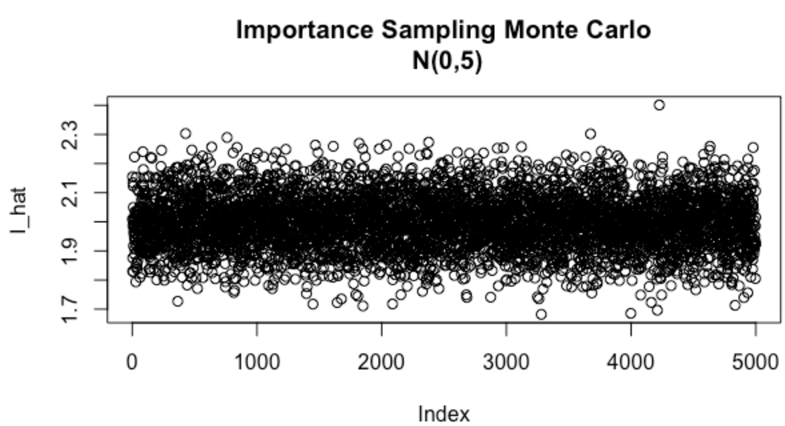
由上圖可見，使用 Important sampling with N(0,1) 方法去估計$I = \int_{-20}^{20} e^{-|x|}dx$時，估計值浮動範圍很大，但大部分都集中在 5 以下，偶而會出現極端值。

接著將極端值刪除（大於 3 的值），並重新繪製散佈圖



圖中顯示，雖然大部分的資料點都幾鐘在(1.8,2.2)之間，但仍有很多點出現在 (2.2, 3.0) 之間，甚至更大，浮動範圍比 Monte Carlo 的範圍(1.4, 2.6)來得大。也因此使用 Important sampling 非但沒有降低 Monte Carlo 方法的變異數，反正比 Monte Carlo 的變異數來得大。

(c) Important sampling with N(0,5)



這次 Important sampling 方法去估計$I = \int_{-20}^{20} e^{-|x|} dx$時，我們改變 Important function，由 N(0,1)改成 N(0,5)。從上圖中發現，這 5000 個估計值範圍為(1.7, 2.4)附近，大多都集中在(1.8,2.1)之間，比 Monte Carlo 的範圍(1.4, 2.6)來得小，因此變異數比 Monte Carlo 的變異數小，縮減量約為 76.1%。

結論：

選擇適當的 Important function 可以使得估計效果與原本使用的方法一樣，同時變異數達到縮減的效果，但選擇不好的 Important function 則會造成變異數變大，雖然估計值差不多。以本題的兩個 Important function $\phi(x,0,1)$ 與$\phi(x,0,5)$ ，$\phi(x,0,5)$可以讓整體變異量下降 76.1%，而$\phi(x,0,1)$造成整體變異量上升。

Question 03.

(a) Obtain the mean and variance of N(0,1) using importance sampling with g ~ t_3.

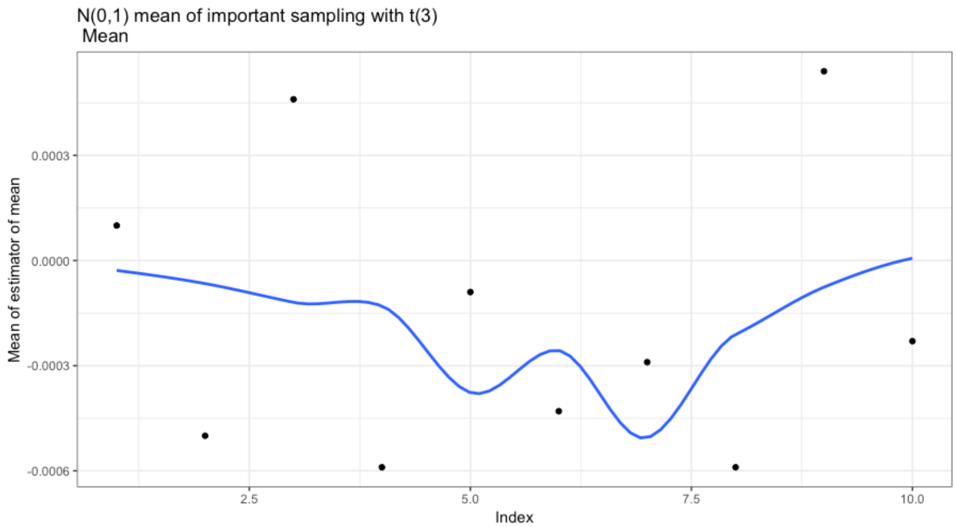
同樣取 5000 個估計量，每一個估計量分別取 100, 200, ⋯ ,1000 去探討，樣本數是否影響估計效果<

並觀察估計是否準確。

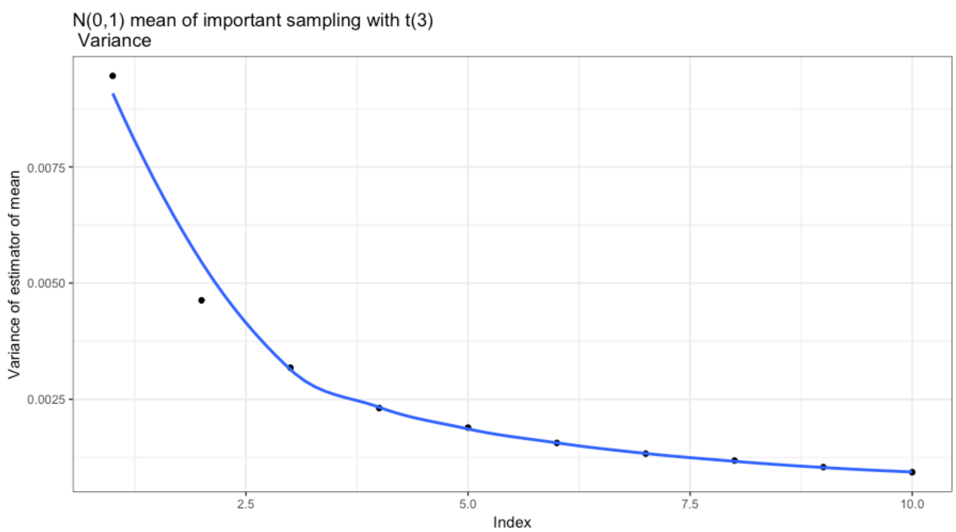
Important sampling t(3)去估計 Normal (0, 1)的平均與變異數

估計標準常態的 mean 的樣本平均與樣本變異數：

n	100	200	300	400	500
Mean of $\hat{\mu}$	0.0001	-0.0005	0.00046	-0.00059	-0.00009
Variance of $\hat{\mu}$	0.00946	0.00463	0.00318	0.00231	0.00189
n	600	700	800	900	1000
Mean of $\hat{\mu}$	-0.00043	-0.00029	-0.00059	0.00054	-0.00023
Variance of $\hat{\mu}$	0.00156	0.00133	0.00118	0.00104	0.00093



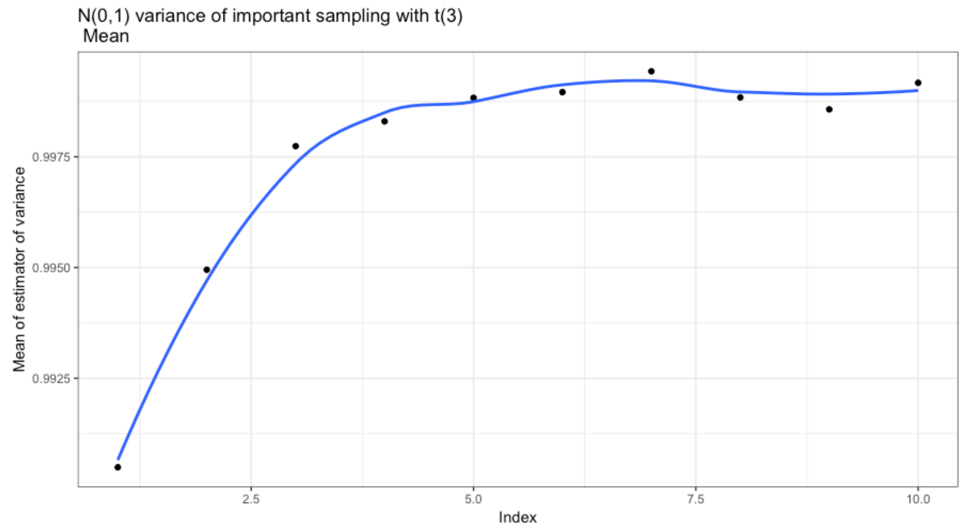
（在不同樣本數下的樣本平均）



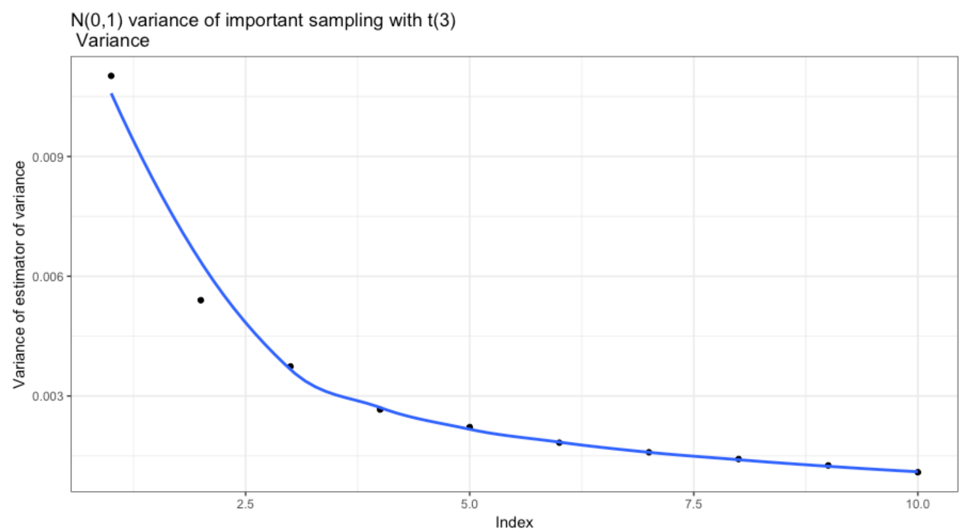
（在不同樣本數下的樣本變異數）

估計標準常態的 Variance 的樣本平均與變異數：

n	100	200	300	400	500
mean of $\widehat{\sigma^2}$	0.99049	0.99495	0.99774	0.9983	0.99883
Variance of $\widehat{\sigma^2}$	0.01102	0.0054	0.00374	0.00266	0.00222
n	600	700	800	900	1000
Mean of $\widehat{\sigma^2}$	0.99896	0.99943	0.99884	0.99857	0.99917
Variance of $\widehat{\sigma^2}$	0.00183	0.00159	0.00142	0.00126	0.00109



（在不同樣本數下的樣本平均）



（在不同樣本數下的樣本變異數）

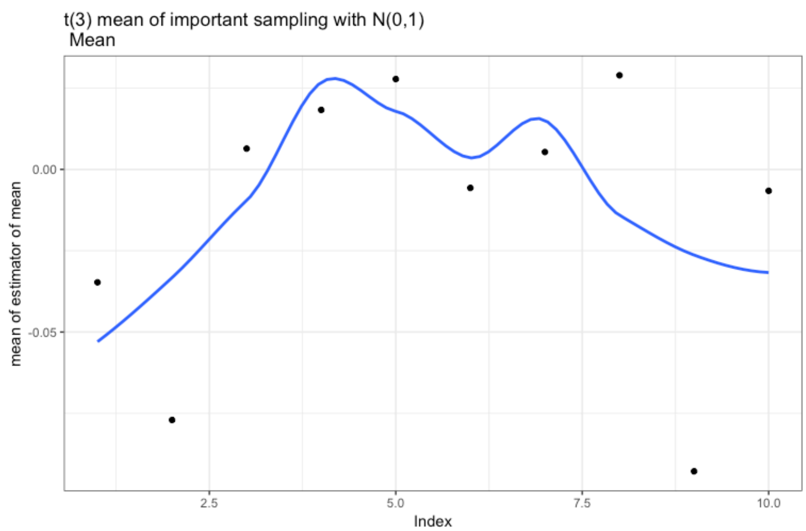
小結：

由 Important sampling with t(3)去估計標準常態的 mean 與 Variance，在不同的抽樣樣本數下都蠻接近的標準常態的 meann 0 以及 Variance 1，特別是當抽樣樣本數逐漸變多時，估計值的變異數都逐漸下降，可知樣本數會影響準確度。

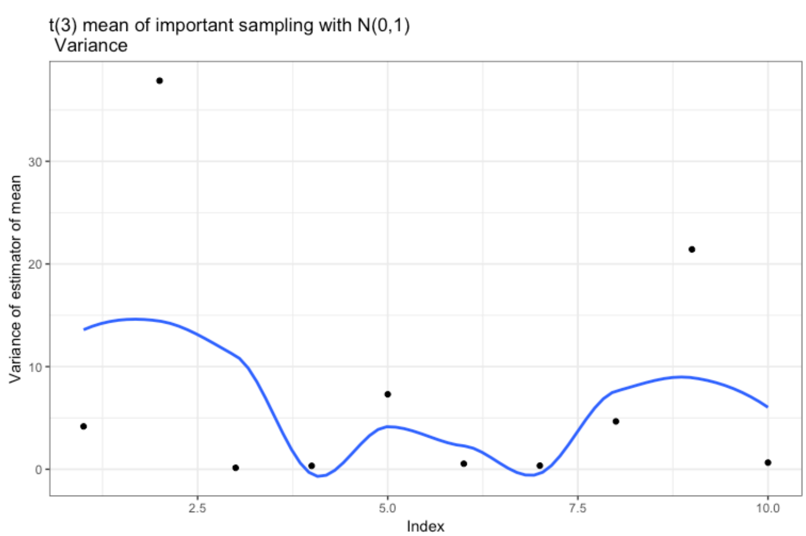
(b) Obtain the mean and variance of t_3 using importance sampling with $g \sim N(0,1)$.

估計 t(3)的 mean 的樣本平均與樣本變異數：

n	100	200	300	400	500
Mean of $\hat{\mu}$	-0.03472	-0.07704	0.0064	0.01828	0.02777
Variance of $\hat{\mu}$	4.17662	37.85109	0.14083	0.33347	7.30422
n	600	700	800	900	1000
Mean of $\hat{\mu}$	-0.00569	0.00535	0.02893	-0.09281	-0.00659
Variance of $\hat{\mu}$	0.54173	0.35007	4.66431	21.42051	0.64627



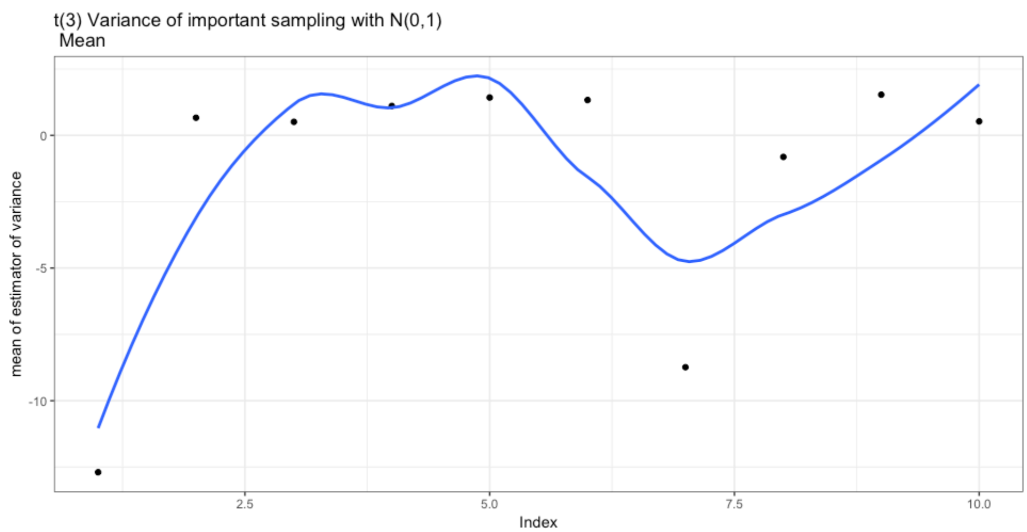
（在不同樣本數下的樣本平均）



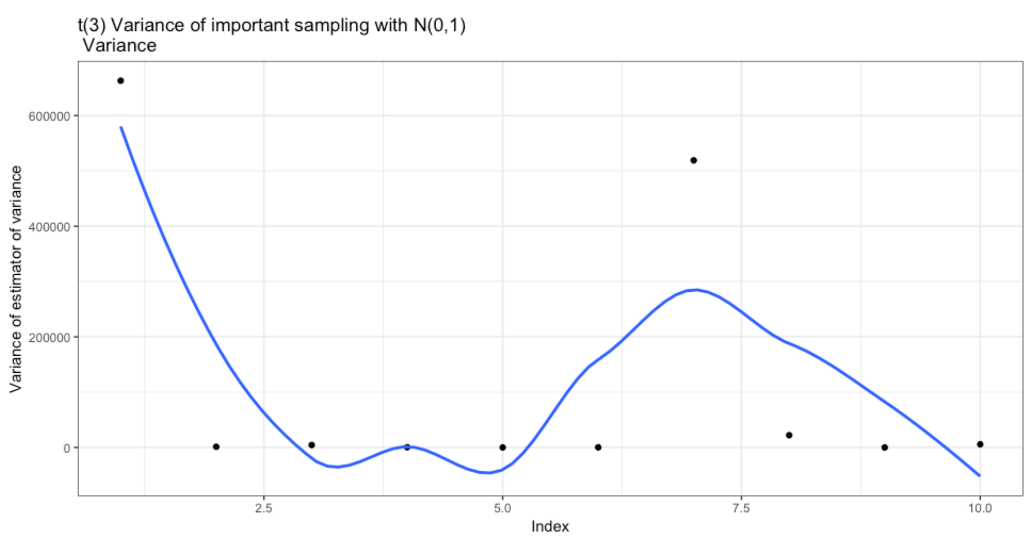
（在不同樣本數下的樣本變異數）

估計 t(3)的 variance 的樣本平均與樣本變異數：

n	100	200	300	400	500
mean of $\widehat{\sigma^2}$	-12.69518	0.66387	0.50941	1.10809	1.42652
Variance of $\widehat{\sigma^2}$	663024.834	1238.09275	4433.13618	387.9221	16.91081
n	600	700	800	900	1000
Mean of $\widehat{\sigma^2}$	1.33154	-8.73801	-0.81247	1.53318	0.52701
Variance of $\widehat{\sigma^2}$	272.9102	518980.652	22146.2786	19.25389	5697.17336



（在不同樣本數下的樣本平均）



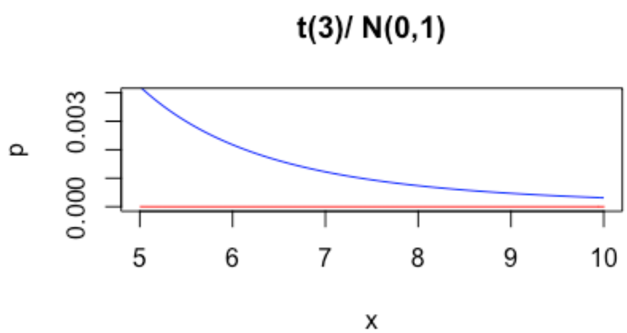
（在不同樣本數下的樣本變異數）

小結：

由 important sampling with N(0,1)去估計 t(3)的平均與變異數，在不同的抽樣樣本數下，產生極大的差異，估計平均值的部分相對於估計變異數來得準確，估計平均值大多落在 0 的上下，與 t(3)實際的母體平均 0 蠻接近的。再來看估計變異數的部分，跳動範圍蠻大的，甚至出現變異數小於 0 的部分，與 t(3)實際的母體變異數相差甚大，同時估計變異數的變異數跳動範圍也是很大。因此我認為用標準常態去估計 t(3)並不是可行的方法。

(c) Comment on the results.

比較(a)與(b)的部分，由 t(3)去估計 N(0,1)與 N(0,1)去估計 t(3)兩個差異很大，最主要差異性來自於兩個分配的尾巴的部分，t(3)尾巴較厚而 N(0,1)尾巴較薄（下圖），而 important sampling 要估計的好時，$\frac{f(x)}{g(x)}$要小於 1(f(X)為原分配，g(x)為 important function))，所以當 t(3)/N(0,1)，尾巴的部分會因為 N(0,1)的值太小，而讓比值過大，因而讓 Variance 變過大。因此，用 important sampling with t(3)去估計 N(0,1)是一個好的方法，但相反用 important sampling with N(0,1)去估計 t(3)則不是一個好的方法。



(上圖只截取一段尾巴，藍線為 t(3)分配，紅線為 N(0,1)分配)

Appendix

Code: http://rpubs.com/YaPi/388887