1 Central Limit Theorem

Suppose you have a sample of size n, $X_1, X_2, ..., X_n$ each independent with common CDF, F with corresponding **finite** mean and variance– μ, σ^2. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, denote the sample mean. Then the expected value of the sample mean is

$$\begin{aligned} E(\overline{X}_n) &= E\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^n E(X_i) \\ &= \frac{1}{n}\sum_{i=1}^n \mu \\ &= \frac{1}{n} n\mu = \mu \end{aligned}$$

and the variance is

$$\begin{aligned} \text{var}(\overline{X}_n) &= \text{var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^n \text{var}(X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^n \sigma^2 \\ &= \frac{1}{n^2} n\sigma^2 = \sigma^2/n \end{aligned}$$

So the sample mean has expected value μ, and variance σ^2/n. Let $F_{\overline{X}_n}$ denote the CDF of $\overline{X}_n$. The Central Limit Theorem (CLT) says that

$$\lim_{n\to\infty} F_{\overline{X}_n} = \lim_{n\to\infty} P(\overline{X}_n \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2/n}} e^{-n(y-\mu)^2/2\sigma^2} dy = \Phi\left(\frac{(x-\mu)\sqrt{n}}{\sigma}\right) / \sqrt{\sigma^2/n}$$

where Φ denotes the standard normal CDF. This theorem basically says that as the sample size increases, the CDF of $\overline{X}_n$ becomes more and more like the CDF of a $N(\mu, \sigma^2/n)$; this gives a basis for hypothesis testing and confidence intervals that do not depend on the distribution of the data. The rate at which this convergence occurs depends primarily on how similar F is to the normal CDF. To demonstrate this we look at the distribution of $\overline{X}_n$ for sample sizes $n = 5, 12, 25$ for three population structures:

1. **flat:** $X \sim \text{Uniform}(0, 1)$; $E(X) = 1/2$, $\text{var}(X) = 1/12$.
2. **skewed:** $X \sim \text{Exponential}(1/2)$; $E(X) = 2$, $\text{var}(X) = 4$.

3. **long-tailed:** $X \sim t(3)$; $E(X) = 0$, $\text{var}(X) = 3$.

To look at how closely the distribution of the sample mean is to the normal distribution we generate 1000 sample means from each distribution, for each sample size, and plot the histogram with the corresponding normal curve overlaying it. A normal qq plot for each parent distribution is included to give perspective on what a "bad" qq plot looks like when you are looking to invoke the CLT.

```
# function to plot all histograms and a qqplot on the same layout
plot4hist <- function(Xbar5, Xbar12, Xbar25, M, V, Y, k)
{
   par(mfrow=c(2,2))
   s <- seq(M - k*sqrt(V), M + k*sqrt(V), by=.025)
   r <- hist(Xbar5, freq=FALSE, breaks=20, col=4, xlim=c(M - 2*sqrt(V), M + 2*sqrt(V))
   lines(s, dnorm(s, mean=M, sd=sqrt(V/N[1])), col=2)

   r <- hist(Xbar12, freq=FALSE, breaks=20, col=4, xlim=c(M - 2*sqrt(V)/1.5, M + 2*sqr
   lines(s, dnorm(s, mean=M, sd=sqrt(V/N[2])), col=2)

   r <- hist(Xbar25, freq=FALSE, breaks=20, col=4, xlim=c(M - 2*sqrt(V)/2, M + 2*sqrt(
   lines(s, dnorm(s, mean=M, sd=sqrt(V/N[3])), col=2)

   qqnorm(Y)
   qqline(Y)
}

# sample sizes
N <- c(5, 12, 25)

# generate 1000 sample of uniforms for each sample size
X_5 <- matrix( runif(1000*N[1]), 1000, N[1])
X_12 <- matrix( runif(1000*N[2]), 1000, N[2])
X_25 <- matrix( runif(1000*N[3]), 1000, N[3])

# get the sample means for each of the 10000 samples
Xbar_5 <- apply(X_5, 1, mean)
Xbar_12 <- apply(X_12, 1, mean)
Xbar_25 <- apply(X_25, 1, mean)

# plot histograms
plot4hist(Xbar_5, Xbar_12, Xbar_25, .5, 1/12, runif(1000),2)

# generate 1000 sample of exp(.5) for each sample size
X_5 <- matrix( rexp(1000*N[1],rate=.5), 1000, N[1])
X_12 <- matrix( rexp(1000*N[2],rate=.5), 1000, N[2])
X_25 <- matrix( rexp(1000*N[3],rate=.5), 1000, N[3])

# get the sample means for each of the 10000 samples
```

```
Xbar_5 <- apply(X_5, 1, mean)
Xbar_12 <- apply(X_12, 1, mean)
Xbar_25 <- apply(X_25, 1, mean)

# plot histograms
plot4hist(Xbar_5, Xbar_12, Xbar_25, 2, 4, rexp(1000),2)

# generate 1000 sample of t(3) for each sample size
X_5 <- matrix( rt(1000*N[1],df=3), 1000, N[1])
X_12 <- matrix( rt(1000*N[2],df=3), 1000, N[2])
X_25 <- matrix( rt(1000*N[3],df=3), 1000, N[3])

# get the sample means for each of the 10000 samples
Xbar_5 <- apply(X_5, 1, mean)
Xbar_12 <- apply(X_12, 1, mean)
Xbar_25 <- apply(X_25, 1, mean)

# plot histograms
plot4hist(Xbar_5, Xbar_12, Xbar_25, 0, 3, rt(1000,df=3),8)
```

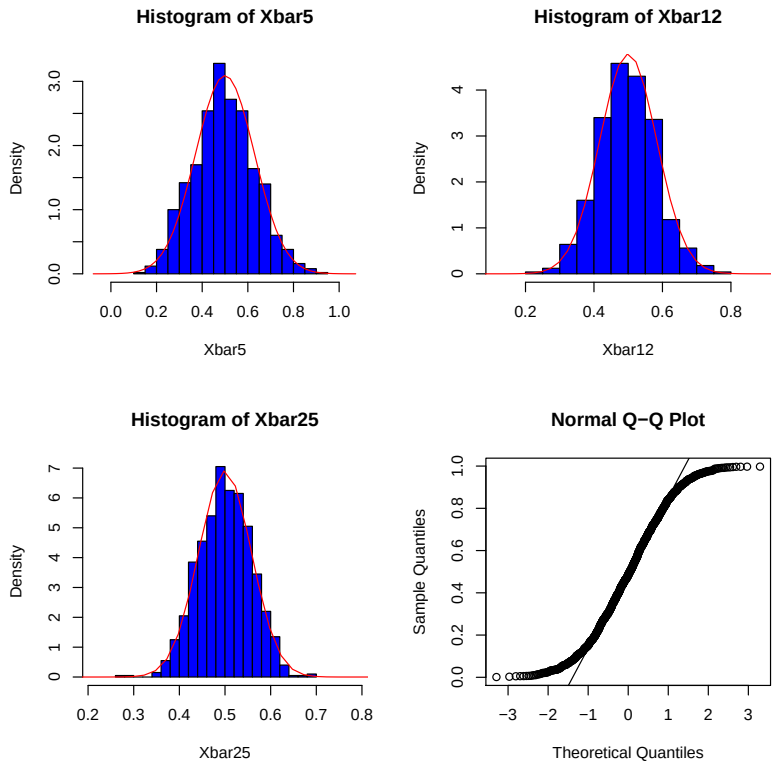

Figure 1: Approximate distribution of the sample mean for sample sizes 5,12,25 from the Uniform(0,1) population, with normal density overlayed. Normal QQ plot of sample of size 1000 from a uniformly distributed population in bottom right

when the parent population is flat (uniform), then the CLT appears to kick in almost immediately, so when the QQ plot looks like that of a uniform distribution, the CLT

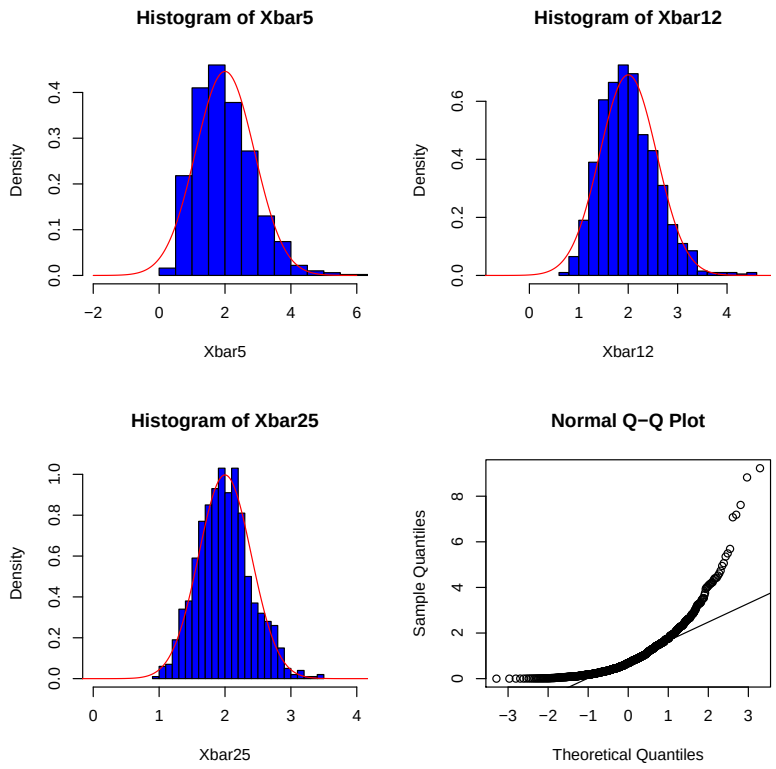

Figure 2: Approximate distribution of the sample mean for sample sizes 5,12,25 from the Exponential(2) population, with normal density overlayed. Normal QQ plot of sample of size 1000 from a exponential distributed population in bottom right

can be invoked even for minscule sample sizes. When the QQ plot indicates skewness, then this can be more of a concern for small sample sizes, since the distribution of the sample mean appears markedly different from that expecte under the CLT. If the QQ plot indicates long tailedness then certain realizations of $\overline{X}$ can be very far out in the tails of the normal distribution, indicating that the p-values calculated by invoking the CLT can be seriously underestimated. For each of these examples, since the parent population has finite mean and variance, the CLT will eventually kick in.

Exercise: On your own, investigate the sample size necessary for the CLT to be a reliable approximation for the sample mean from a $t(3)$ distribution. Can you generate from a distribution where the CLT does not appear to apply for any sample size?

2 Basic Monte Carlo Integration

2.1 Moments of functions of random variables

One common place where one may want to calculate an intractable integral is when you want to find *moments* of complicated functions of a random variable. Denote an arbitrary function by $h(X)$, where X is a random variable with density p. Then

$$E(h(X)) = \int h(x)p(x)dx$$

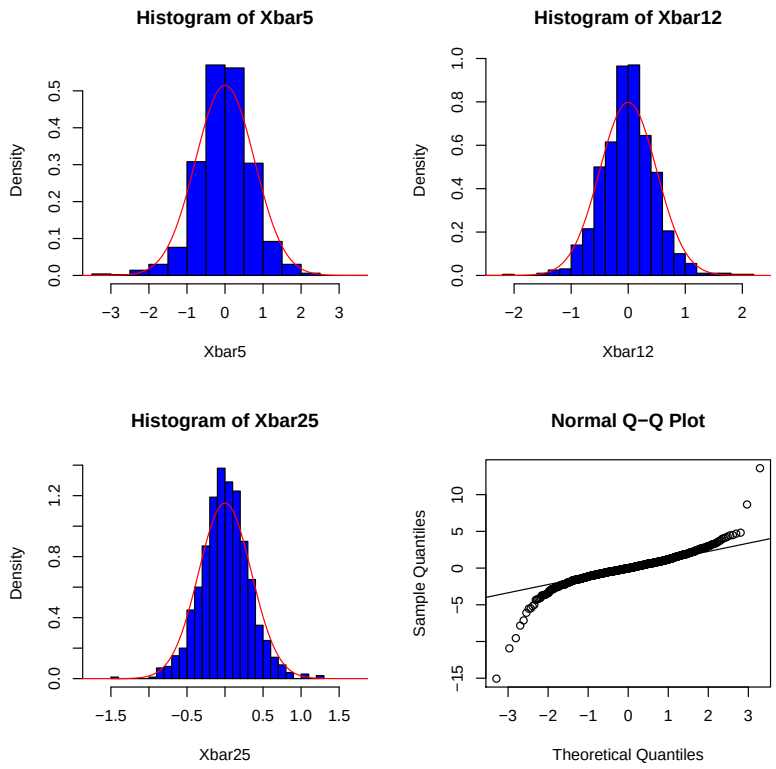

Figure 3: Approximate distribution of the sample mean for sample sizes 5,12,25 from the $t(3)$ population, with normal density overlayed. Normal QQ plot of sample of size 1000 from a $t(3)$ distributed population in bottom right

which, depending on the structure of h and p, can be a very difficult quantity to calculate. The Monte Carlo method appeals to the Law of Large Numbers and estimates $E(h(X))$ by the sample mean of $h(X)$:

$$\overline{h(X)} = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

Recall the LLN ensures that $\overline{h(X)} \to E(h(X))$ as $n \to \infty$. You can also make confidence intervals for $E(h(X))$ by appealing to the CLT. The sample variance of $h(X)$ is

$$\widehat{\text{var}(h(X))} = \frac{1}{n}\sum_{i=1}^{n}(h(X_i) - \overline{h(X)})^2$$

Then the CLT tells us that the approximate distribution of $\overline{h(X)}$ is $N\left(E(h(X)), \text{var}(h(X))/n\right)$. Therefore,

$$\overline{h(X)} \pm z_{\alpha/2}\sqrt{\frac{\widehat{\text{var}(h(X))}}{n}}$$

is an *approximate* $(1-\alpha)$ confidence interval for $E(h(X))$.

Example: In logistic regression the goal is to predict an individual's probability of "yes"

based on a number of predictors. A typical example is where the response is a disease of some kind, and the predictors are potential risk factors. As a convenience the probability being modeled is tranformed to the logistic scale, which maps probabilities to the real interval, $(-\infty, \infty)$. The function which defines this mapping, called the *logit* function, is

$$h(x) = \log(\frac{x}{1-x})$$

As x approaches 0, $h(x)$ approaches $-\infty$; as x approaches 1, $h(x)$ approaches ∞. Suppose the fitted logistic regression model is

$$\text{logit}(P(Y = 1|X_1, X_2)) = .7 + 2.1X_1 - 1.3X_2 + \varepsilon$$

where $\varepsilon \sim N(0, 1.43)$. Predict the probability that a subject with $X_1 = 1$ and $X_2 = 1$ responds with $Y = 1$, and give a 95 percent confidence interval for this probability based on 1000 monte carlo replications.

Solution: The fitte logit-transformed probability of responding with $Y = 1$ for this subject is $.7 + 2.1 - 1.3 = 1.5$. This means, for this subject, their logit-tranformed probability is $N(1.5, 1.43)$. So the expected probability that $Y = 1$ is given by the integral

$$\int_{-\infty}^{\infty} \frac{1}{1+e^{-x}} \frac{1}{\sqrt{2\pi \cdot 1.43}} e^{-(x-1.5)^2/2.86} dx$$

This is because the inverse of the logit function is given by $1/(1 + e^{-x})$, which needs to be integrated against the $N(1.5, 1.43)$ density. We can estimate this by generating $Z_1, Z_2, ..., Z_{1000}$ from $N(1.5, 1.43)$ and calculating

$$\widehat{p} = \frac{1}{1000} \sum_{i=1}^{1000} \frac{1}{1+e^{-Z_i}}$$

and calculating the corresponding variance as

$$\widehat{\text{var}(\widehat{p})} = \frac{1}{1000} \sum_{i=1}^{1000} (\hat{p} - \frac{1}{1+e^{-Z_i}})^2$$

and giving the corresponding confidence interval. The following R code does this:

```
# underlying normal variables
Z = rnorm(1000, mean=1.5, sd=sqrt(1.43))

# the transformed variables
P = 1/(1 + exp(-Z))

# sample mean
p = mean(P)

# sample variance
var_p = (n-1)*var(P)/n

# multiplier
```

```
z <- qnorm(1-(.05/2))

c(p - z*sqrt(var_p/1000), p + z*sqrt(var_p/1000))
[1] 0.7431204 0.7659213
```

So, for this subject, $P(Y = 1) \approx .75$.

2.2 Calculating arbitrary integrals

To calculate more general integrals, the trick is to re-write the integral as an expectation against some density, generate from that density, and look at the sample mean as before. For definite integrals over a finite interval, the uniform distribution will suffice (but may not be optimal in terms of estimation precision). To calculate the integral of an arbitrary function h over the interval (a, b):

$$\int_a^b h(x)dx$$

this can be rewritten as an integral against the Uniform(a, b) distribution by writing

$$(b-a)\int_a^b h(x) \cdot \frac{1}{b-a}dx$$

since the density of a Uniform(a, b) is just $\frac{1}{b-a}$. As an example consider integrating the function $h(x) = \sin(x\cos(x))$ over the interval $(0, 2\pi)$. Step 1 is to rewrite the integral as

$$2\pi\int_0^{2\pi} \frac{\sin(x\cos(x))}{2\pi}dx$$

Then we estimate this by generating $U_1, U_2, ..., U_N$ distribution from the uniform distribution on $(0, 2\pi)$ and calculating

$$\frac{2\pi}{N}\sum_{i=1}^{N} \sin(U_i\cos(U_i))$$

We can calculate this in R by the following code and use $N = 100000$ to minimize the effect of sampling variation:

```
# underlying uniforms
U <- runif(100000, 0, 2*pi)

# calculate h(U)
hU <- sin(U*cos(U))

# monte carlo approximation
2*pi*mean(hU)
[1] -1.048418

# R's numerical integration approximation
h <- function(u) sin(u*cos(u))
integrate(h, 0, 2*pi)$val
[1] -1.041727
```

This approach will not work when one or both of the bounds are not finite, since $(b-a)$ will be undefined in that case. In that case we can use any distribution defined on the entire real line and do a similar approximation. This time we rewrite the integral as

$$\int_{-\infty}^{\infty} \frac{h(x)}{p(x)} p(x)dx$$

where $p(x)$ is the density for a distribution defined over $\mathcal{R}$ and h is any function. You then generate X_i from $p(x)$ and look at the sample mean of $g(X_i)/p(X_i)$ for your approximation.

Consider again $h(x) = \sin(x\cos(x))$ except now we want to integrate over $(-\infty, \infty)$. It turns out h is an odd function– that is $h(x) = -h(-x)$ for any x, so the integral over $\mathcal{R}$ is 0. To verify this we will calculate this quantity by monte carlo. The normal distribution is supported over $\mathcal{R}$, so we can use the standard normal distribution for the monte carlo. The approach will be to generate $X_1, X_2, ..., X_N$ from the normal distribution and calculate

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\sin(X_i\cos(X_i))}{\phi(X_i)}$$

as the estimate, where $\phi(x)$ denotes the standard normal density. The following R code does this

```
X <- rnorm(100000)

Y <- sin(X*cos(X))/dnorm(X)
mean(Y)
[1] -0.3657281
```

In this case, the standard normal density appears in the denominator, so the integrand gets very big in the low mass areas of the normal distribution, so the integral estimate has substantial variance– estimated around 3 in my simulation.

Exercise: See you if you can find a better distribution than the normal for approximating this integral. A good choice will have a lower estimated variance than when the normal distribution in used, and should be supported on the entire interval $(-\infty, \infty)$.

3 Importance Sampling

The methods we've introduced so far generate arbitrary points from a distribution to approximate integrals– in some cases many of these points correspond to points where the function value is very close to 0, and therefore contributes very little to the approximation. In many cases the integral "comes with" a given density, such as integrals involving calculating an expectation. However, there will be cases where another distribution gives a better fit to integral you want to approximate, and results in a more accurate estimate; importance sampling is useful here. In other cases, such as when you want to evaluate $E(X)$ where you can't even generate from the distribution of X, importance sampling is necessary.

The logic underlying importance sampling lies in a simple rearrangement of terms in the target integral and multiplying by 1:

$$\int h(x)p(x)dx = \int h(x)\frac{p(x)}{g(x)}g(x)dx = \int h(x)w(x)g(x)dx$$

here $g(x)$ is another density function whose support is the same as that of $p(x)$. That is, the sample space corresponding to $p(x)$ is the same as the sample space corresponding to $g(x)$ (at least over the range of integration). $w(x)$ is called the importance function; a good importance function will be large when the integrand is large and small otherwise.

Example 1: As a first example we will look at a case where importance sampling provides a reduction in the variance of an integral approximation. Consider the function $h(x) = 10\exp(-2|x-5|)$. Suppose that we want to calculate $E(h(X))$, where $X \sim$ Uniform$(0,1)$. That is, we want to calculate the integral

$$\int_0^{10} \exp(-2|x-5|)\,dx$$

The true value for this integral is about 1. The simple way to do this is to use the approach from section 2 and generate X_i from the uniform(0,10) density and look at the sample mean of $10 \cdot h(X_i)$ (notice this is equivalent to importance sampling with importance function $w(x) = p(x)$):

```
X <- runif(100000,0,10)
Y <- 10*exp(-2*abs(X-5))
c( mean(Y), var(Y) )
[1] 0.9919611 3.9529963
```

The function h in this case is peaked at 5, and decays quickly elsewhere, therefore, under the uniform distribution, many of the points are contributing very little to this expectation. Something more like a gaussian function (ce^{-x^2}) with a peak at 5 and small variance, say, 1, would provide greater precision. We can re-write the integral as

$$\int_0^{10} 10\exp(-2|x-5|)\,\frac{1/10}{\frac{1}{\sqrt{2\pi}}e^{-(x-5)^2/2}}\frac{1}{\sqrt{2\pi}}e^{-(x-5)^2/2}dx$$

That is, $E(h(X)w(X))$, where $X \sim N(5,1)$, and $w(x) = \frac{\sqrt{2\pi}e^{(x-5)^2/2}}{10}$ is the importance function in this case. This integral can be more compactly written as

$$\int_0^{10} \exp(-2|x-5|)\,\sqrt{2\pi}e^{(x-5)^2/2} \cdot \frac{1}{\sqrt{2\pi}}e^{-(x-5)^2/2}dx$$

where the part on the left is the quantity whose expectation is being calculated, and the part on the right is the density being integrated against ($N(5,1)$). We implement this second approach in R by

```
X=rnorm(1e5,mean=5,sd=1)
Y=10*f(X)*dunif(X,0,10)/dnorm(X,mean=5,sd=1)
c( mean(Y), var(Y) )
[1] 0.9999271 0.3577506
```

Notice that the integral calculation is still correct, but with a variance this is approximately 1/10 of the simple monte carlo integral approximation. This is one case where

importance sampling provided a substantial increase in precision. A plot of the integrand from solution 1:

$$10\exp\left(-2|x-5|\right),$$

along with the density it is being integrated against:

$$p(x) = 1/10,$$

and a second plot of the integrand from solution 2:

$$\exp\left(-2|x-5|\right)\sqrt{2\pi}e^{(x-5)^2/2},$$

along with the density it is being integrated against:

$$p(x) = \frac{1}{\sqrt{2\pi}}e^{-(x-5)^2/2},$$

gives some intuition about why solution 2 was so much more efficient. Notice the integrands are on much larger scale than the densities (since densities must integrate to 1), so the integrands are normalized to make the plots comparable.

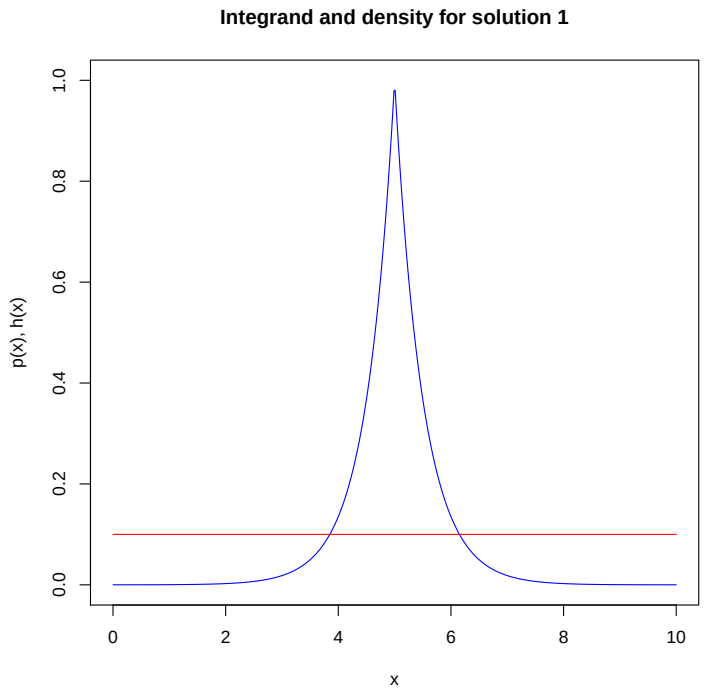

Figure 4: Normalized integrand (blue) and the density being integrated against (red) for approach 1

Example 2: As another example we will consider estimating the moments of a distribution we are unable to sample from. Let

$$p(x) = \frac{1}{2}e^{-|x|}$$

which is called the double exponential density. The CDF is

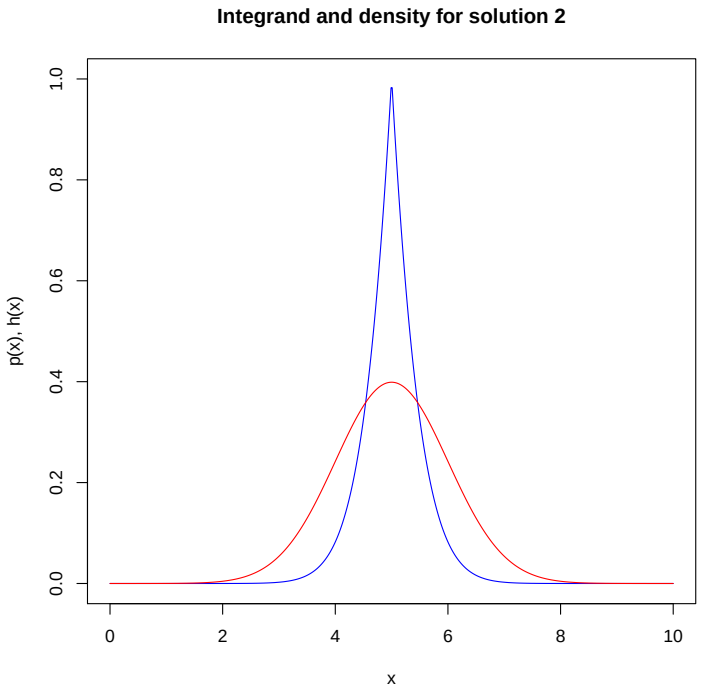

Figure 5: Normalized integrand (blue) and the density being integrated against (red) for approach 2

$$F(x) = \frac{1}{2}e^{x}\mathcal{I}(x \leq 0) + (1 - e^{-x}/2)\mathcal{I}(x > 0)$$

which is a piecewise function and difficult to invert (it is possible to generate from this distribution but lets pretend it is not). Suppose you want to estimate $E(X^2)$ for this distribution, which is support on $\mathcal{R}$. That is, we want to calculate the integral

$$\int_{-\infty}^{\infty} x^2 \frac{1}{2}e^{-|x|}dx$$

We cannot estimate this empirically without generating from p. However, we can re-write this as

$$\int_{-\infty}^{\infty} x^2 \frac{\frac{1}{2}e^{-|x|}}{\frac{1}{\sqrt{8\pi}}e^{-x^2/8}} \frac{1}{\sqrt{8\pi}}e^{-x^2/8}dx$$

Notice that $\frac{1}{\sqrt{8\pi}}e^{-x^2/8}$ is the $N(0, 4)$ density. Now this amounts to generating $X_1, X_2, ..., X_N$ from $N(0, 4)$ and estimating

$$E\left(X^2 \frac{\frac{1}{2}e^{-|X|}}{\frac{1}{\sqrt{8\pi}}e^{-X^2/8}}\right)$$

by the sample mean of this quantity. The following R code does this:

```
X <- rnorm(1e5, sd=2)
Y <- (X^2) * .5 * exp(-abs(X))/dnorm(X, sd=2)
mean(Y)
```

```
[1] 1.998898
```

The true value for this integral is 2, so importance sampling has done the job here. Other moments, $E(X), E(X^3), E(X^4), ...$ for $X \sim p$ can be estimated analogously, although it should be clear that all odd moments are 0.