Written by:
-Aditya Chilwal
-Paul Bakshi
-Muzammil Arshad

To estimate the size of a country on Earth, we will be using three algorithms involving Monte Carlo sampling and integration**: uniform random sampling**, **importance sampling on a rectangle using a Beta distribution, and importance sampling on a sphere using the von Mises-Fisher distribution.**

We will lay down a few assumptions while tackling the problem:

- The sphere we will be sampling on is the unit sphere of radius 1, and the randomly sampled xyz coordinates on the unit sphere will be transformed into Latitude and Longitudinal coordinates which can then be used to indicate whether a point lies within a country or not.

- Following is that we will assume that the Earth is a perfect sphere to ease calculations even though in actuality it is an irregularly shaped ellipsoid.

- For reference purposes, to get the "true" size of a country's surface area we will query **rworldmap's countryExData** dataset.(note that this data is itself not perfect because they are themselves estimating the true surface areas of countries)

- When scaling the estimates and variance we will be using earth's total surface area (510,000,000 km^2) divided by 4 pi as the scaling constant.

- Some of our algorithms require that we know where the center of a country is, thus we will be using **rgeos' gCentroid** function to acquire the center of a given country.

- The rescaled beta on rectangle and the Von Mises Fisher importance use size categorization as "large", (larger than 1 mill km^2), "medium" (in between 300k km^2 and 1 million km^2) and "small" (between 300k km^2 and 100k km^2) or "very small"(below 100k km^2). These categorizations are assumed to be known, the case where the categorization is unknown will be dealt with in the final method.

- Our methods rely on the importance sampling, so whatever the measure of the domain of the probability distribution we are using is will be used as the denominator and the surface area of the Earth will be the numerator for the scaling factor when

rescaling the estimate. (For the unit sphere this will be 5 million/4pi, for the [-90,90] x[-180,180] rectangle it will be 5 million/64800))

Our first solution to the problem will also serve as the baseline for improvement, the uniform random sampling method. This method utilizes the **rsphere.normal()** function in class to generate randomly uniform distributed points on the unit sphere which are then converted into latitude and longitudinal coordinates. The estimates calculated are roughly close to their "true" value, but the problem arises with the resulting variance and confidence interval. The confidence intervals range from being roughly which leaves a lot of room for improvement and the goal of the next two methods are to reduce the variance. This method is still good for the larger countries because it is able to sample more evenly from the entire country, and its higher variance more accurately represents the uncertainty inherent in measuring the surface area of large countries.

The second solution to try is importance sampling using a Beta (a,a) distribution for some $a > 0$. Since the beta(a,a) has the same alpha and beta parameter, its mean is centered at $a/(a+b)=a/(a+a) =½$. So we use the country's center latitude as c2 and longitude as c1 and rescale the beta with lower bound as l_lat=c2-90 and upper bound as u_lat=c2+90. Then we rescale the other beta lower bound as l_long=c1-180 and upper bound is u_long=c1+180. This will make he center of rescaled distribution match with the center of the country. Additionally since the variance of the beta distribution is $(ab)/((a+b)^2)(a+b+1)) = (a^2)/((2a)^2)(2a+1)) =1/((4)(2a+1))$, meaning that as the a increases the variance decreases. Therefore, the idea is to center the rescaled beta distribution at the center of the country and increase the a the smaller the country is. We created a custom function to compute the Beta weights for the aforementioned rescaled Beta random variables. The value of the particular distribution is based on the l_lat,u_lat and l_long,u_long bounds, however, the function which checks if a point is in a country cannot take input that is outside of [-90,90] x[-180,180], but we just made it wrap around the -90 to 90 and -180 to 180 to rectify this.

While we do get a significant improvement for the variance our estimates however are starting to deviate a lot more from the "true" value compared to the uniform estimate. This can be explained by the fact that we are sampling from a rectangle which is a "squished" version of the surface area of the sphere which leads to unavoidable distortion of the sphere. The estimates are worse for large countries as the rectangle greatly distorts their actual size but it is the opposite for small countries as this distortion does not seem to have as significant of an effect on their size. The von Mises method is one solution that we propose that can have both a good estimate and reasonable variance.

The third solution and the best out of the previous proposed solutions is monte carlo importance sampling using the von Mises Fisher distribution. This uses a similar idea as the beta distribution with centering the mean of the distribution at the center of the country by setting the mean parameter as the xyz position which the country would be in. The concentration parameter for this distribution more or less functions as the “a” parameter of the beta distribution in the previous method for our purposes because it decreases the variance. If we wish to estimate the size of a larger country, the concentration parameter should be smaller.

The final method we propose combines the importance of the sphere and the Von Mises Fisher methods. It aims to eliminate the circular reasoning of using a size categorization to estimate size. This method utilizes the estimate of the uniform random method which as mentioned above gives decent estimates of the true size. This estimate then serves as a can determine whether a country’s surface area is large, medium, small, or very small to tune the concentration parameter of the von Mises method while the base von Mises method relies on the fact that the user has a general idea of what the country’s total surface area could be which could lead to the wrong concentration parameter. The final method is capable of making fairly accurate estimations, where the confidence interval is around 10,000 away from the estimated size, for the smallest countries.

Appendix:
The derivation of the rescaled beta distribution is below:

Rescaling Beta distribution from $[0,1]$ to $[l,u]$

If $X \sim \text{Beta}(\alpha=K, \beta=K)$ where $K>0$, and

$R = uX + l(1-X) = uX + l - lX = (u-l)X + l$, which is rescaled such that $R \in [l,u]$,

then $\boxed{f_R(r) = \frac{\Gamma(2K)}{(u-l)(\Gamma(K))^2}\left[\frac{(r-l)(u-r)}{(u-l)^2}\right]^{K-1}\left(1_{[l,u]}(r)\right)}$

$F_R(r) = P(R<r) = P((u-l)X + l < r) \underset{(u>l)}{=} P\left(X < \frac{r-l}{u-l}\right) = F_X\left(\frac{r-l}{u-l}\right)$

$\therefore f_R(r) = \frac{dF_R(r)}{dr} = \frac{dF_X\left(\frac{r-l}{u-l}\right)}{dr} = f_X\left(\frac{r-l}{u-l}\right)\left(\frac{1}{u-l}\right)$

$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left(\frac{r-l}{u-l}\right)^{\alpha-1}\left(1-\left(\frac{r-l}{u-l}\right)\right)^{\beta-1}\left(1_{[0,1]}\left(\frac{r-l}{u-l}\right)\right)\left(\frac{1}{u-l}\right)$

$\left(\overset{*}{\text{since } \alpha=\beta=K}\right) = \frac{\Gamma(2K)}{(\Gamma(K))^2}\left(\left(\frac{r-l}{u-l}\right)\left(1-\left(\frac{r-l}{u-l}\right)\right)\right)^{K-1}\left(1_{[l,u]}(r)\right)\left(\frac{1}{u-l}\right)$

$= \frac{\Gamma(2K)}{(\Gamma(K))^2}\left[\left(\frac{r-l}{u-l}\right)\left(\frac{u-r}{u-l}\right)\right]^{K-1}\left(1_{[l,u]}(r)\right)\left(\frac{1}{u-l}\right)$

$= \frac{\Gamma(2K)}{(u-l)(\Gamma(K))^2}\left(\frac{(r-l)(u-r)}{(u-l)^2}\right)^{K-1}\left(1_{[l,u]}(r)\right)$

Therefore,

1b) if $K=\frac{1}{2}$, $l=-1$, $u=1$

then $f_R(r) = \frac{\Gamma(2(\frac{1}{2}))}{(2)(\Gamma(\frac{1}{2}))^2}\left(\frac{(r-(-1))(1-r)}{(2)^2}\right)^{-\frac{1}{2}} 1_{[-1,1]}(r)$

$= \frac{1}{\pi\sqrt{(r+1)(1-r)}}\left(1_{[-1,1]}(r)\right) = \frac{1}{\pi\sqrt{-(r^2-1)}} 1_{[-1,1]}(r)$

1c) if $K=2$, $l=-1$, $u=1$

$f_R(r) = \frac{\Gamma(2(2))}{(2)(\Gamma(2))^2}\left(\frac{(r-(-1))(1-r)}{2^2}\right)^1 1_{[-1,1]}(r)$

$= \frac{3!}{2^3(1)}\left((r+1)(1-r)\right) 1_{[-1,1]}(r) = \frac{-3}{4}(r^2-1)\left(1_{[-1,1]}(r)\right)$

2b) if $K=\frac{1}{2}$, $l=0$, $u=\pi$,

$f_R(r) = \frac{\Gamma(2(\frac{1}{2}))}{\pi(\Gamma(\frac{1}{2}))^2}\left(\frac{(r-(0))(\pi-r)}{\pi^2}\right)^{-\frac{1}{2}} 1_{[0,\pi]}(r)$

$= \frac{1}{\pi\sqrt{\pi r - r^2}}\left(1_{[0,\pi]}(r)\right)$

2d) if $K=2$, $l=0$, $u=\pi$, $f_R(r) = \frac{\Gamma(2(2))}{\pi(\Gamma(2))^2}\left(\frac{(r-0)(\pi-r)}{\pi^2}\right)^1 \cdot 1_{[0,\pi]}(r) =$

$= \frac{3!}{\pi^3}(\pi r - r^2)\cdot\left(1_{[0,\pi]}(r)\right) = \frac{6(\pi r - r^2)}{\pi^3}\left(1_{[0,\pi]}(r)\right)$