Monte Carlo

STAT 251, Supplement 2

Overview

Potential difficulty in mathematical calculation

Numerical approximation

Monte Carlo Method

Monte Carlo Estimation of Percentiles

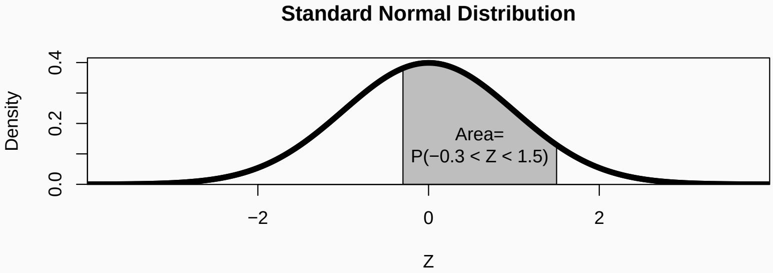
Potential difficulty in mathematical calculation

The problem

Sometimes it is very difficult (or at least very time-consuming) to derive a mathematical result using analytical techniques.

Consider this motivating example:

If $Z \sim N(0, 1^2)$, what is $Pr(-0.3 < Z < 1.5)$?



The problem

If $Z \sim N(0, 1^2)$, what is $Pr(-0.3 < Z < 1.5)$?

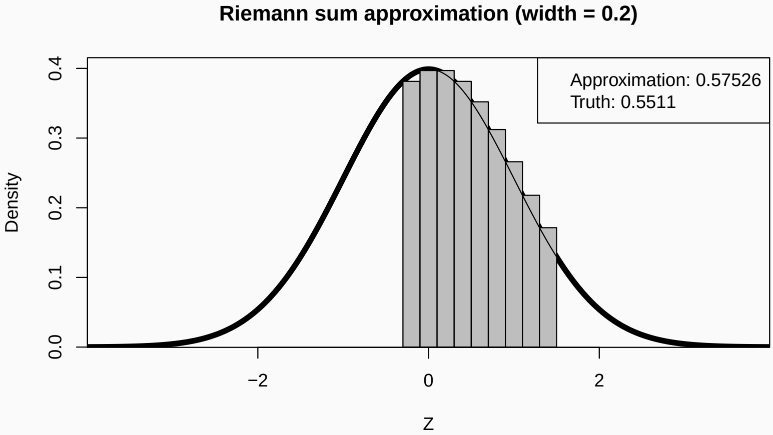
$$\int_{-0.3}^{1.5} f(z)dz \quad = \quad \int_{-0.3}^{1.5} \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}z^2 \right\} dz$$

This might not look so bad, but there is no closed-form expression for this integral.

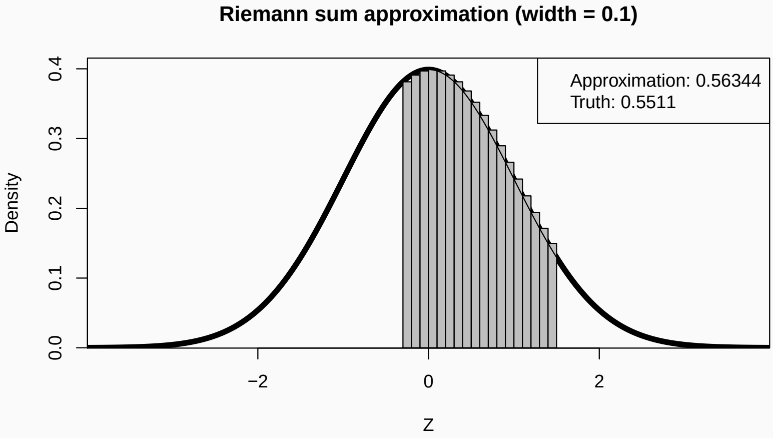
Numerical approximation

Riemann sum approximations

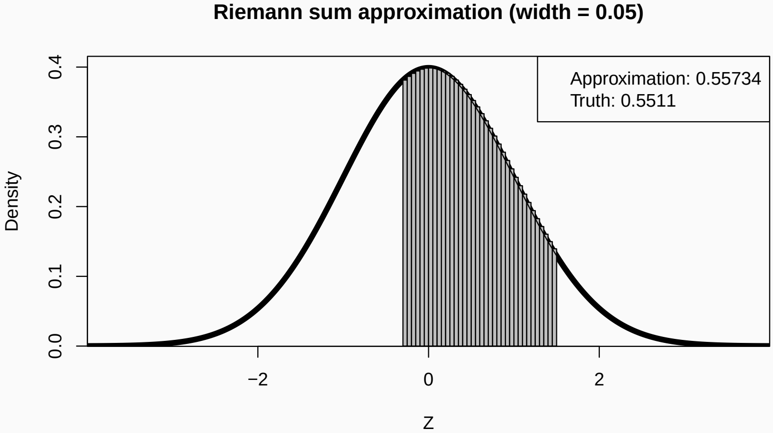
An alternative to exactly deriving the definite integral is to approximate it using Riemann sums. Consider the Riemann sums as the width of each rectangle shrinks. The area of the shaded rectangles is an estimate of the probability.



Riemann sum approximations (cont.)



Riemann sum approximations (cont.)



The estimate improves as the rectangles become narrower.

We won't be using Riemann sums in this class, but you can learn more at `https://en.wikipedia.org/wiki/Riemann_sum`

Monte Carlo Method

Monte Carlo

Another alternative to exactly deriving the integral is to estimate it using simulation (randomly generated values) in a principled manner.

This is the Monte Carlo method.
`https://en.wikipedia.org/wiki/Monte_Carlo_method`

Basic idea: Repeatedly simulate the random variable from its distribution. Then use the simulated data values to estimate the desired quantity.

Monte Carlo Details

Suppose we are interested in the expected value of $g(X)$ and that X has pdf $f(x)$.

Then by definition, $E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$.

Examples:

- If $g(\cdot)$ is the identity function, $g(x) = x$, then $E(g(X)) = E(X) = \mu = \int_{-\infty}^{\infty} xf(x)dx$.
- If $g(\cdot)$ is the function $g(x) = (x - \mu)^2$, then $E(g(X)) = E(X - \mu)^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx$.
- If $g(x) = \mathbb{1}_{(a<x<b)}$, then $E(g(X)) = \int_{-\infty}^{\infty} \mathbb{1}_{(a<x<b)} f(x)dx = \int_a^b f(x)dx = P(a < X < b)$.

Monte Carlo Details

If we have a large collection of values, $X_{(1)}, X_{(2)}, \ldots, X_{(J)}$, that are sampled from the distribution $f(x)$, then

$$
\begin{aligned}
E(g(X)) &= \int_{-\infty}^{\infty} g(x) f(x) dx \\
&\approx \frac{1}{J} \sum_{j=1}^{J} g(X_{(j)})
\end{aligned}
$$

Explanation: We are interested in $E(g(X))$, so we apply the $g(\cdot)$ function to each simulated value for the random variable X and then average these quantities.

Goal: Estimate $P(-0.3 < Z < 1.5)$, where $Z \sim N(0, 1^2)$

- underlying distribution, $f(z)$, is the standard normal pdf
- $g(z) = \mathbb{1}_{(-0.3<z<1.5)}$

$$\begin{aligned} E(g(Z)) &= \int_{-\infty}^{\infty} g(z)f(z)dz = \int_{-\infty}^{\infty} \mathbb{1}_{(-0.3<z<1.5)}f(z)dz \\ &= \int_{-0.3}^{1.5} f(z)dz \quad = \quad P(-0.3 < Z < 1.5). \end{aligned}$$

Goal: Estimate $P(-0.3 < Z < 1.5)$, where $Z \sim N(0, 1^2)$

- underlying distribution, $f(z)$, is the standard normal pdf
- $g(z) = \mathbb{1}_{(-0.3<z<1.5)}$

$$
\begin{aligned}
E(g(Z)) = \int_{-\infty}^{\infty} g(z)f(z)dz &= \int_{-\infty}^{\infty} \mathbb{1}_{(-0.3<z<1.5)} f(z)dz \\
= \int_{-0.3}^{1.5} f(z)dz \quad &= \quad P(-0.3 < Z < 1.5).
\end{aligned}
$$

- To do: sample J=5 values from the pdf, f(z). Then calculate $g(Z_{(j)})$ for each of the five values. Then average them together.

```r
simvals <- rnorm(5, 0, 1)
round(simvals, 3)
```

```
## [1] -1.775 -0.468  0.097 -0.005 -0.874
```

Goal: Estimate $P(-0.3 < Z < 1.5)$, where $Z \sim N(0, 1^2)$

In R, if a logical condition evaluates to TRUE or FALSE, then the logical condition can be converted to a numeric value (where TRUE becomes 1 and FALSE becomes 0).

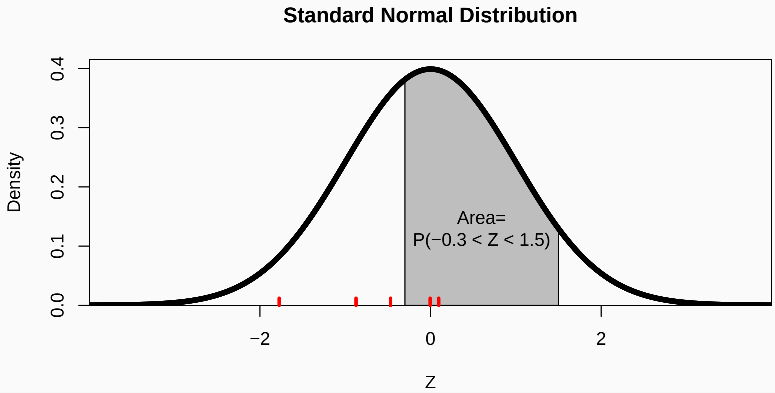
We need $g(z) = \mathbb{1}_{(-0.3<z<1.5)}$. Here is a user-defined g function that takes an input "z" and that returns either a 1 (if $-0.3 < z$ and $z < 1.5$) or a 0 (otherwise).

Goal: Estimate $P(-0.3 < Z < 1.5)$, where $Z \sim N(0, 1^2)$

In R, if a logical condition evaluates to TRUE or FALSE, then the logical condition can be converted to a numeric value (where TRUE becomes 1 and FALSE becomes 0).

We need $g(z) = \mathbb{1}_{(-0.3<z<1.5)}$. Here is a user-defined g function that takes an input "z" and that returns either a 1 (if $-0.3 < z$ and $z < 1.5$) or a 0 (otherwise).

```
round(simvals, 3)


## [1] -1.775 -0.468  0.097 -0.005 -0.874


g <- function(z){as.numeric(-0.3 < z & z < 1.5)}
g(simvals)


## [1] 0 0 1 1 0
```



```
mean(g(simvals))


## [1] 0.4
```

Sample proportion of simulated data in the (-0.3, 1.5) interval is the Monte Carlo estimate of the true probability, $P(-0.3 < Z < 1.5) = 0.5511$.

Monte Carlo estimation of a probability: alternate R code

To estimate the probability of event A using Monte Carlo, you simply need to calculate the proportion of times A occurred in the Monte Carlo-sampled values.

In R, if a condition evaluates to TRUE or FALSE, then this will be treated as TRUE=1 and FALSE=0 when used in an arithmetic expression.

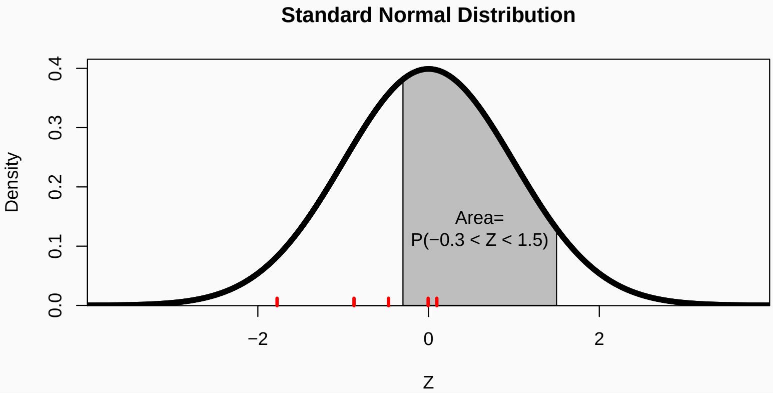
The proportion of times A occurred can be calculated as the average for these TRUE/FALSE values.

Pedagogical Implementation

Goal: Estimate $P(-0.3 < Z < 1.5)$, where $Z \sim N(0, 1^2)$.

- underlying distribution, $f(z)$, is the standard normal pdf
- $g(z) = \mathbb{1}_{(-0.3<z<1.5)}$
- To do: sample J=5 values from the pdf, f(z), then determine the proportion of these values in $(-0.3, 1.5)$.

```
simvals <- rnorm(5, 0, 1)
round(simvals, 3)


## [1] -1.775 -0.468  0.097 -0.005 -0.874


## proportion that are above -0.3 AND below 1.5
mean(simvals > -0.3 & simvals < 1.5)


## [1] 0.4
```

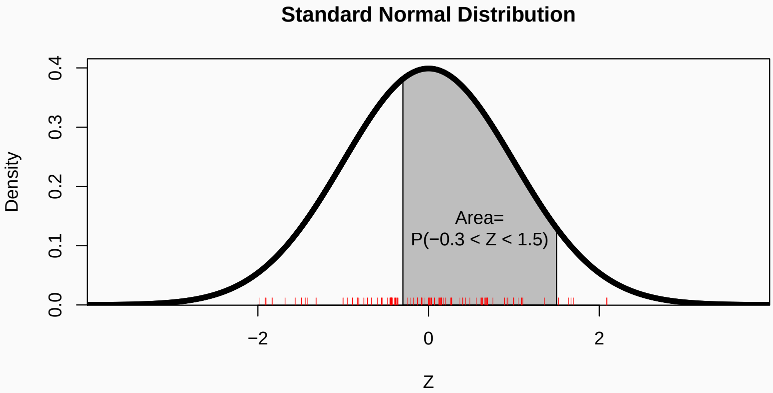


```r
mean(simvals > -0.3 & simvals < 1.5)

## [1] 0.4
```

Sample proportion of simulated data in the (-0.3, 1.5) interval is the Monte Carlo estimate of the true probability, $P(-0.3 < Z < 1.5) = 0.5511$.

Better Estimate

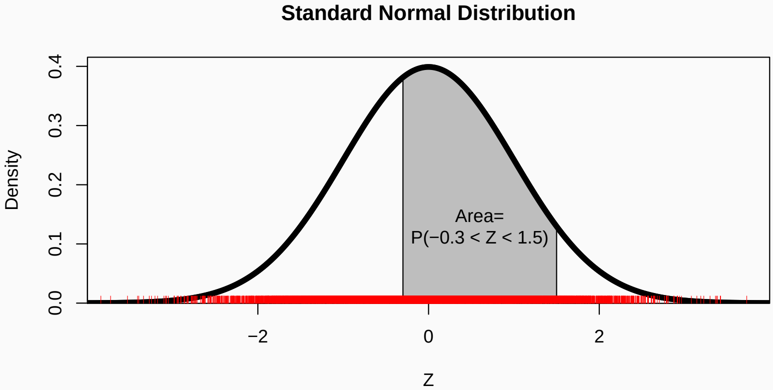
```r
simvals2 <- rnorm(100, 0, 1)
mean(simvals2 > -0.3 & simvals2 < 1.5)


## [1] 0.51
```



Adequate Estimate (True probability is 0.5511)

```
simvals3 <- rnorm(10000, 0, 1)
mean(simvals3 > -0.3 & simvals3 < 1.5)
```

```
## [1] 0.5518
```



Number of simulated values

When performing Monte Carlo estimation, the more simulated values we obtain, the better our estimate should be. For most of what we will do in this class, a sample size of 5000 is sufficiently large; don't use fewer than 1000.

In Monte Carlo, we are not making up data. There are no new *observations*. Rather, it is just a tool to avoid (potentially very difficult) exact mathematical calculations by sampling from the desired distribution. We are not "cheating" by getting larger Monte Carlo sample sizes, but rather are getting more precise estimates of a deterministic mathematical expression. The only reason to limit the Monte Carlo sample size is the computational cost. But for what we do in this class, it takes a negligible amount of time to get 5000 values, so always use AT LEAST 1000 in your Monte Carlo estimation.

Negligible run time for 1,000,000 simulated values

```
system.time({
             simvals4 <- rnorm(1000000, 0, 1)
             print(mean(simvals4 > -0.3 & simvals4 < 1.5))
            })
```

```
## [1] 0.550685
##    user  system elapsed
##   0.083   0.006   0.090
```

Well under a second!

Compare above estimate with the true probability:

```
pnorm(1.5, 0, 1 ) - pnorm(-0.3, 0, 1)
```

```
## [1] 0.5511042
```

Monte Carlo Estimation of μ

- What is the mean of the beta(4,8) distribution? That is, if $\theta \sim Beta(4, 8)$, what would $E(\theta)$ be?
- Easy enough to show that this is equal to $a/(a + b) = 4/(4 + 8) = 1/3$ mathematically, but we'll demonstrate with Monte Carlo methodology.

Monte Carlo Estimation of μ

- Monte Carlo steps to estimating μ from a given distribution, $f(x)$. **Note: we are assuming the distribution f(x) is completely specified and that we know how to simulate values from f(x).**
    1. Simulate a random sample of size J (for J very large) from $f(x)$;
    2. Calculate the sample mean of the simulated values;
    3. Use this as an estimate of the expected value.

Monte Carlo Estimate of $E(\theta)$ if $\theta \sim Beta(4, 8)$

```
set.seed(9703)  ## so that results are reproducible

### Sample 5000 values from the Beta(4,8) distribution
vals <- rbeta(5000,4,8)

### Calculate the sample average, our estimate of
### the expected value from this distribution
mean(vals)

## [1] 0.3310814
```

Actual $E(\theta) = 1/3$.

Stochastic nature of Monte Carlo Estimation

Because we are simulating *random* values from the distribution of interest, a Monte Carlo estimate is also random. Repeating the process (with a different seed) will yield a new (random) estimate. We call this Monte Carlo variation.

It is undesirable for such variation to be large. This is why we need large sample sizes in the Monte Carlo samples—so that the estimates will be more precise.

What if we repeat this process?

```
## Reset the seed
set.seed(97145)
## Get a new sample of 5000 and get the estimate
vals.alt <- rbeta(5000,4,8)
mean(vals.alt)

## [1] 0.332642

### Another repetition
vals2 <- rbeta(5000,4,8)
mean(vals2)

## [1] 0.3345063
```

Optional: Standard Error

The *standard error* is defined as the square root of an estimator's variance. If our estimator is of a population mean, i.e. if we are estimating μ, then compute $\frac{s}{\sqrt{J}}$, where s is the sample standard deviation of the simulated values. This would measure about how far a (random) Monte Carlo estimate is expected to be from the true mean.

Optional: Standard Error

The *standard error* is defined as the square root of an estimator's variance. If our estimator is of a population mean, i.e. if we are estimating μ, then compute $\frac{s}{\sqrt{J}}$, where s is the sample standard deviation of the simulated values. This would measure about how far a (random) Monte Carlo estimate is expected to be from the true mean.

```
# mean(vals) was 0.3310814.
### standard error of the estimate:
sd(vals)/sqrt(5000)


## [1] 0.001822705


### Because of how small this number is, we don't
### expect our Monte Carlo estimate to be far off
```

Monte Carlo Estimate of $Var(\theta)$ if $\theta \sim Beta(4,8)$

```r
set.seed(9703)  #for reproducibility of simulated vals

### Sample 8000 values from the Beta(4,8) distribution
sim.vars <- rbeta(8000,4,8)

### Calculate the sample variance, our estimate of
### the population variance from this distribution
var(sim.vars)
```

```
## [1] 0.01686282
```

Note that the exact variance of a random variable with the $Beta(4,8)$ distribution is $2/117 = 0.0171$.

Final Comments

- Can the same Monte Carlo ideas be used if $f(x)$ is a pmf, not a pdf? Yes!

```
# Estimate mu if X~Binomial(200, 0.4)
bin.vals <- rbinom(10000, 200, 0.4)
mean(bin.vals)

## [1] 80.1354

# True mean is (200*0.4)=80
```

Monte Carlo Estimation of Percentiles

Monte Carlo Estimation of Percentiles

Can use Monte Carlo to estimate expected values of a function of a random variable (e.g., μ, σ^2, $P(c < X < d)$; see Supplement 2)

Can also use Monte Carlo to estimate percentiles. Suppose $X \sim N(200, 18^2)$.

Actual 80th percentile: *qnorm(.80, 200, 18)* = 215.1492

Monte Carlo Estimation of Percentiles

Can use Monte Carlo to estimate expected values of a function of a random variable (e.g., μ, σ^2, $P(c < X < d)$; see Supplement 2)

Can also use Monte Carlo to estimate percentiles. Suppose $X \sim N(200, 18^2)$.

Actual 80th percentile: *qnorm(.80, 200, 18)* $= 215.1492$

Monte Carlo estimate of 80th percentile using *quantile* function:

```
set.seed(21397) # reproducibility
vals <- rnorm(15000, 200, 18)
quantile(vals, .80)


##      80%
## 215.4916
```

Credible Interval

A (posterior) 95% credible interval (CI) is an interval with 95% posterior probability of containing θ. That is, there is a 95% posterior probability that θ is in the interval.

One possibile construction of a 95% credible interval of θ: $(\{\theta|y\}_{0.025}, \{\theta|y\}_{0.975})$—the 2.5th and 97.5th percentiles from the posterior distribution.

There are other possibilities, such as $(\{\theta|y\}_{0.01}, \{\theta|y\}_{0.96})$ or $(\{\theta|y\}_{0.02}, \{\theta|y\}_{0.97})$.

We'll restrict attention to the *central credible interval*, which contains the *middle 95%* of the posterior's mass.

Fictional Example

Suppose that there are nine study participants in a study for an investigational medication.

Among the study objectives is to estimate the probability that someone taking the drug would experience nausea as a side effect.

Let $\theta \equiv$ the probability that a person taking the investigational medication would experience nausea as a side effect.

Suppose the prior distribution for θ is the Beta(1,1) distribution.

Suppose also that five of the nine study participants experience nausea as a side effect.

What is the posterior distribution of θ?

Fictional Example

Suppose that there are nine study participants in a study for an investigational medication.

Among the study objectives is to estimate the probability that someone taking the drug would experience nausea as a side effect.

Let $\theta \equiv$ the probability that a person taking the investigational medication would experience nausea as a side effect.

Suppose the prior distribution for θ is the Beta(1,1) distribution.

Suppose also that five of the nine study participants experience nausea as a side effect.

What is the posterior distribution of θ? $\theta | y \sim Beta(6, 5)$.

Fictional Example (cont.)

$\theta | y \sim \mathit{Beta}(6, 5)$.

What is an *exact* 95% credible interval for the probability of experiencing nausea as a side effect?

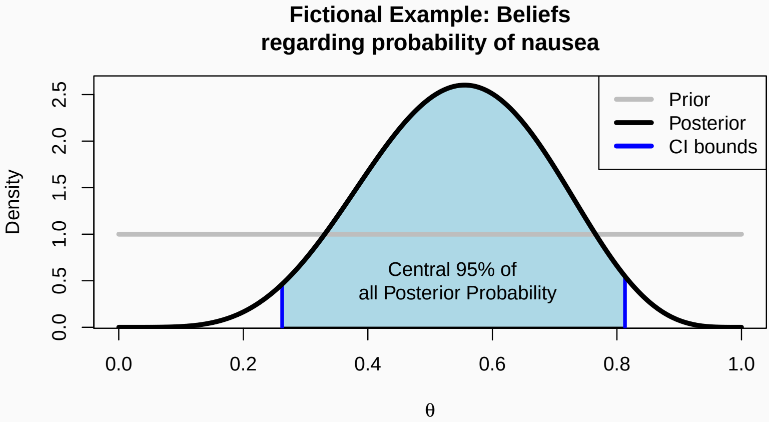
Fictional Example (cont.)

$\theta | y \sim Beta(6, 5)$.

What is an *exact* 95% credible interval for the probability of experiencing nausea as a side effect?

```
qbeta(c(.025, .975), 6, 5)


## [1] 0.2623781 0.8129140
```

We are 95% sure that the probability of experiencing nausea as a side effect of the drug is between 26% and 81%.



```r
qbeta(c(.025, .975), 6, 5)
```

```
## [1] 0.2623781 0.8129140
```

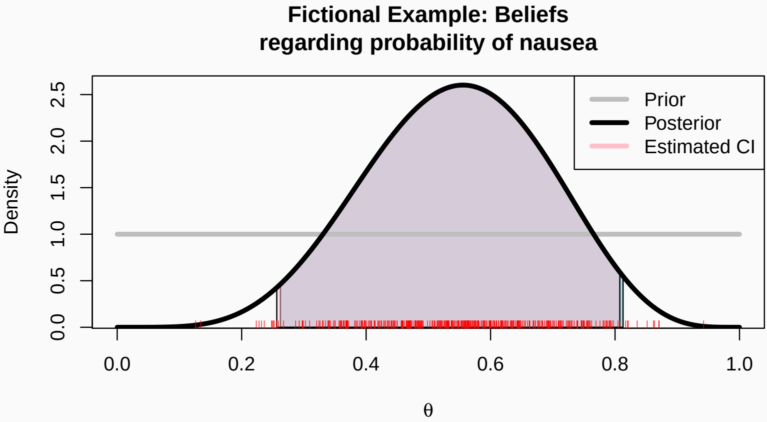
We are 95% sure that the probability of experiencing nausea as a side effect of the drug is between 26% and 81%.

Monte-Carlo estimated CIs

What is an *estimated* 95% credible interval for the probability of experiencing nausea as a side effect?

```r
set.seed(5547085)
post.thetas <- rbeta(400, 6, 5)
quantile(post.thetas, c(.025, .975))
```

```
##      2.5%     97.5%
## 0.2564087 0.8076524
```



```
quantile(post.thetas, c(.025, .975))


##       2.5%     97.5%
## 0.2564087 0.8076524
```

We are *approximately* 95% sure that the probability of experiencing nausea as a side effect of the drug is between 26% and 81%.

Final Comments

- When should we use Monte Carlo?
  - If the mathematical calculations are too tedious or difficult
  - To check the correctness of exact calculations
  - Get quick sanity check
- Your turn. Use Monte Carlo methods to obtain results from NFL home field example described in previous lecture.