Homework 6-STAT:5400

Dev Narayan Baiju and Yu Chao Huang

Due: Oct 11, 2024 9:30 AM

Problems

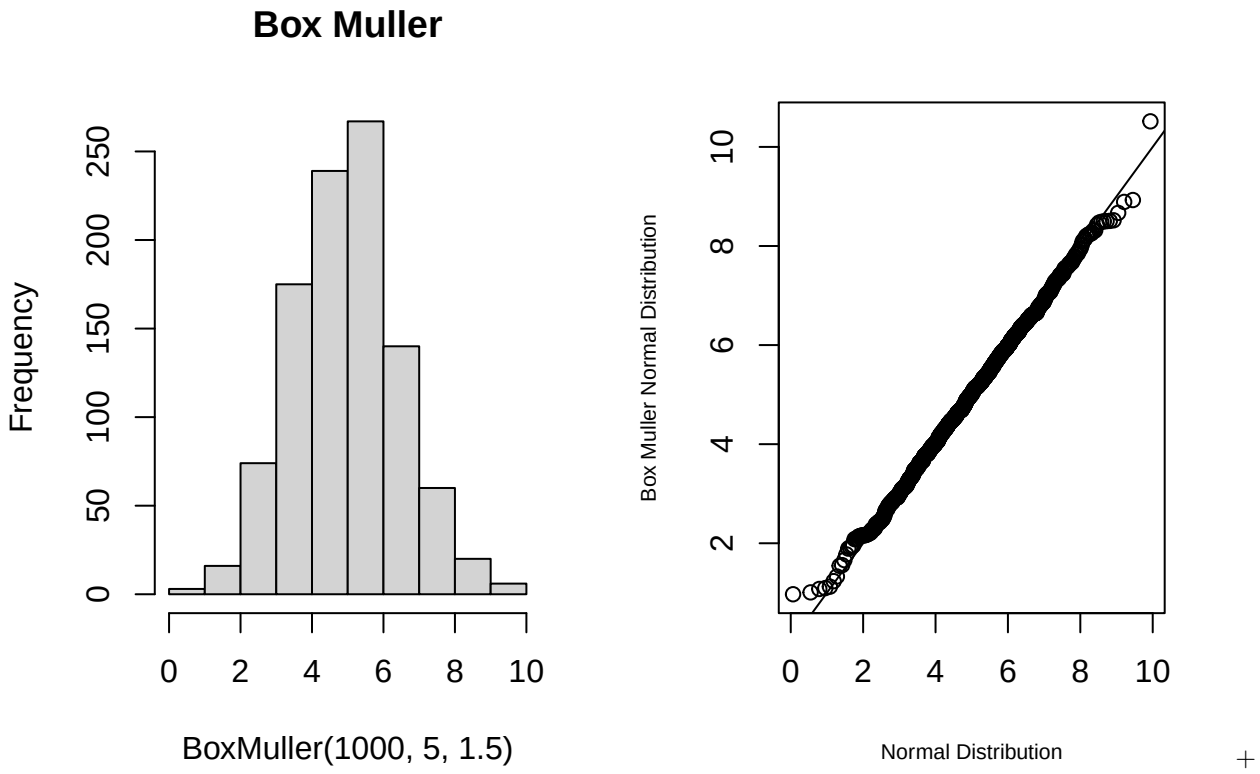
Submit your solutions as an .Rmd file and accompanying .pdf file. Include all the **relevant** R code and output. Always interpret your result whenever it is necessary.

Problems

1. Generators of normal distributions.

- Write an R function `BoxMuller` that generates a sample of n values from $N(\mu, \sigma^2)$ using the Box-Muller method. This function should have three arguments:
  - `n`: number of observations,
  - `mu`: the value of mean μ,
  - `sigma`: the value of standard deviation σ. This function should return a vector of n normal variates.

```
set.seed(5200)
BoxMuller <- function(n, mu, sigma){
  U1 <- runif(n/2, 0, 1)
  U2 <- runif(n/2, 0, 1)
  Z1 <- sqrt(-2*log(U1))*cos(2*pi*U2)
  Z2 <- sqrt(-2*log(U1))*sin(2*pi*U2)
  return(mu+sigma*c(Z1,Z2))
}
par(mfrow = c(1,2))
hist(BoxMuller(1000, 5, 1.5), main = "Box Muller")
plot(qnorm(ppoints(1000), 5, 1.5),
     sort(BoxMuller(1000, 5, 1.5)),
     xlab = "Normal Distribution",
     ylab = "Box Muller Normal Distribution",
     cex.lab = 0.7)
abline(0,1)
```

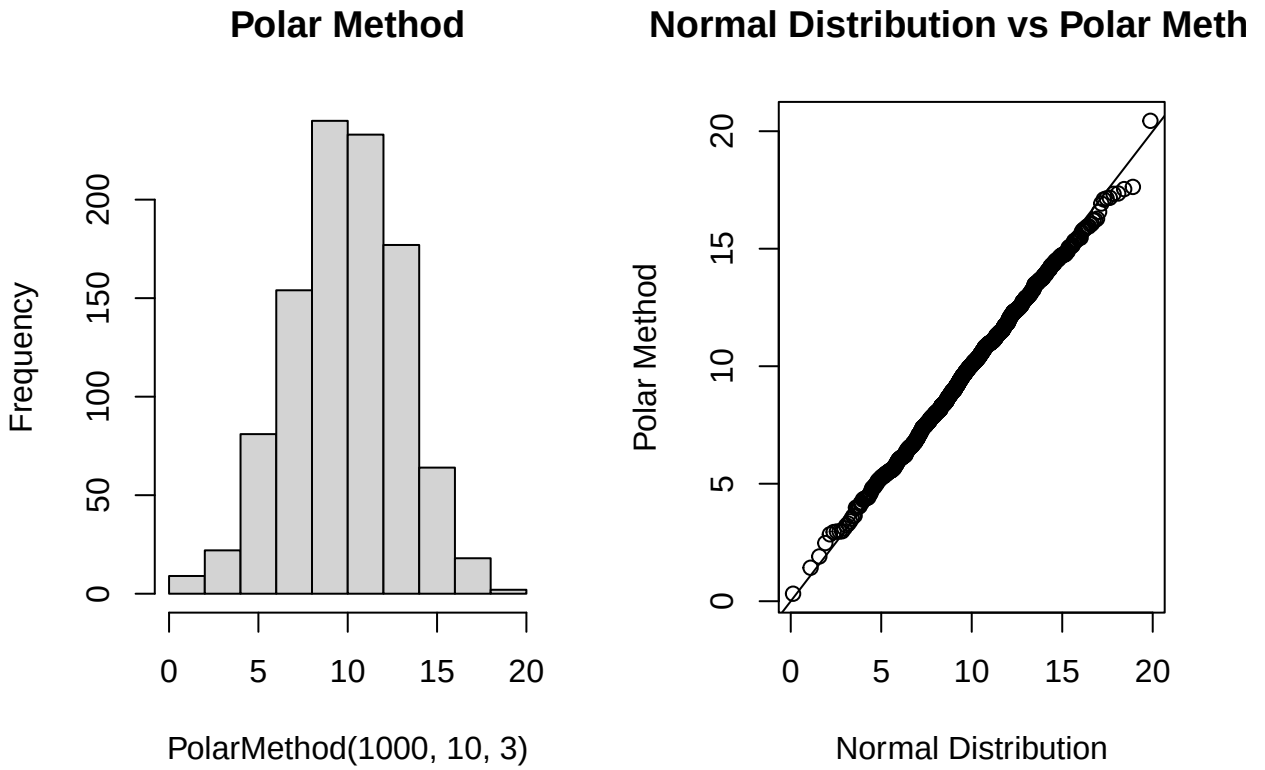


Write an R function `PolarMethod` that generates a sample of n values from $N(\mu, \sigma^2)$ using the polar method. This function should have the same arguments and return values as the `BoxMuller` function. When the polar method is used, it might be tricky if you want to control `n` when the accept-reject sampling method is used to generate points in the unit disk. You need a different implementation from the one on the slides.

```
PolarMethod <- function(n, mu, sigma){
  m <- 100*n
  V1 <- runif(m,-1,1)
  V2 <- runif(m,-1,1)
  V12 <- V1^2 + V2^2
  V1_new <- numeric(0)
  V2_new <- numeric(0)
  Rsq <- numeric(0) #using rep function is better but Rsq will have n elements
  attempts = 0
  while (length(Rsq) < n/2 && attempts < m){
    attempts = attempts + 1
    for (i in 1:m){
      if (V12[i] <= 1){
        Rsq <- c(Rsq, V12[i])
        V1_new <- c(V1_new, V1[i])
        V2_new <- c(V2_new, V2[i])
        if (length(Rsq) >= n / 2){
          break
        }
      }
    }
  }
  Rsq <- Rsq + .Machine$double.xmin
  m <- sqrt(-(2 * log(Rsq)) / (Rsq))
```

```
  X1 <- m * V1_new
  X2 <- m * V2_new
  PM <- mu + sigma * c(X1,X2)
  return(PM)
}

par(mfrow = c(1,2))
hist(PolarMethod(1000,10,3), main = "Polar Method")
plot(qnorm(ppoints(1000), 10, 3),
     sort(PolarMethod(1000,10,3)),
     xlab = "Normal Distribution",
     ylab = "Polar Method",
     main = "Normal Distribution vs Polar Method")
abline(0,1)
```



+ Generate two vectors of 30 standard normal variates, by both `BoxMuller` and `PolarMethod`. Run a test to check if the two vectors are from the same distribution. One solution is to use `ks.test` in R to perform a Kolmogorov-Smirnov test.

```
BM <- BoxMuller(30,0,1)
PM <- PolarMethod(30,0,1)
ks.test(BM,PM)
```

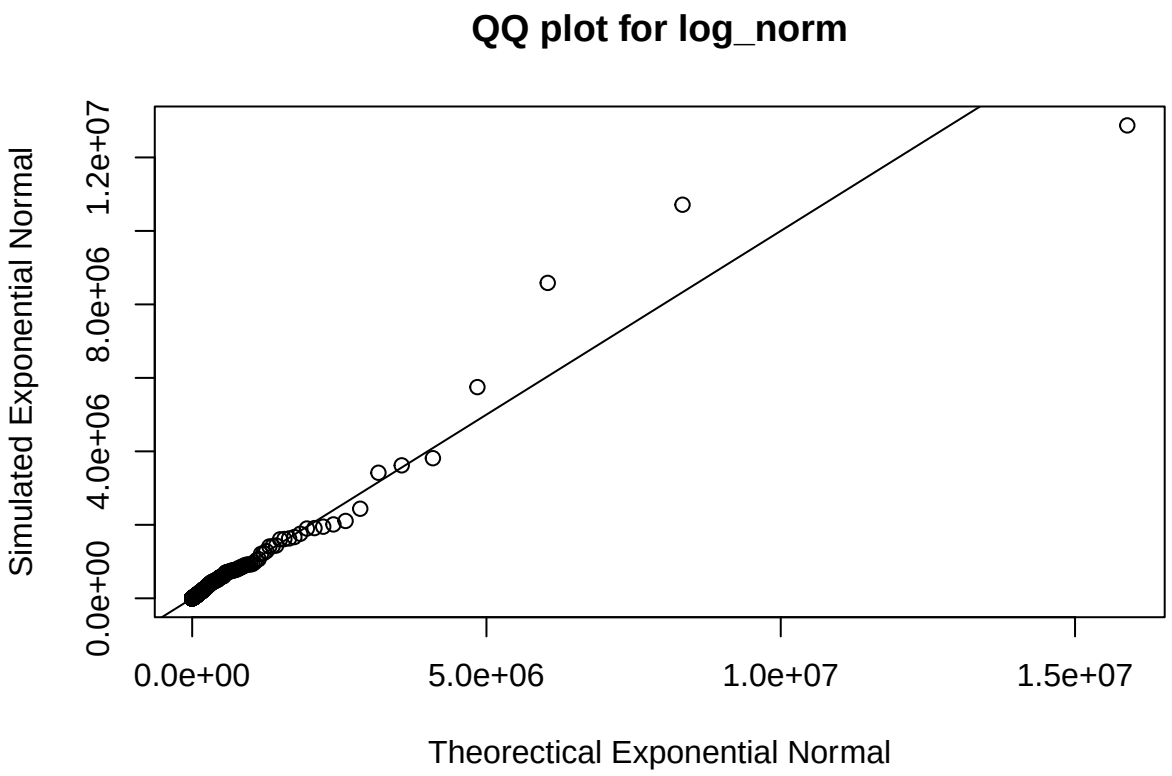
```
##
##  Exact two-sample Kolmogorov-Smirnov test
##
## data:  BM and PM
## D = 0.13333, p-value = 0.9578
## alternative hypothesis: two-sided
```

- Generate n values from the log-normal distributions. Note that if X has a normal distribution, then $Y = \exp(X)$ has a log-normal distribution. Thus you may generate log-normal realizations based on the normal variates that you have generated.

```
log_norm <- function(n, mu, sigma){
  Y <- exp(BoxMuller(n, mu, sigma))
  return(Y)
}
```

- Have a Q-Q plot to check whether the sample comforts with the log-normal distribution. You may use `qlnorm` function in R to compute the quantiles of log-normal distributions.

```
plot(qlnorm(ppoints(1000), 10, 2), sort(log_norm(1000, 10, 2)),
     xlab = "Theorectical Exponential Normal",
     ylab = "Simulated Exponential Normal",
     main = "QQ plot for log_norm")
abline(0,1)
```



2. Generators of multivariate normal distributions. + Write an R function called `mymvrnorm` that generates n random observations from $\mathrm{N}_p(\mu, \Sigma)$. Only calls to R's standard uniform generators `runif` are permitted. This function should have three arguments:

```
+ ‘‘n‘‘, the random sample size,
+ ‘‘mu‘‘, the mean vector with $p$ entries,
+ ‘‘Sigma‘‘, the variance-covariance matrix, which is symmetric and positive semi-definite.
```

- This function should return a matrix with n rows and p columns, where the ith row has the realization of Y_i. The `eigen` function should be called in your definition of `mymvrnorm`.

```r
mymvrnorm <- function(n, mu, sigma){
  p <- length(mu)
  eig <- eigen(sigma)
  e_values <- eig$values
  e_vectors <- eig$vectors
  D <- diag(sqrt(e_values))
  Z <- matrix(0, nrow = n, ncol = p)
  for(i in 1:n){
    u1 <- runif(p, 0, 1)
    u2 <- runif(p, 0, 1)
    Z[i,] <- sqrt(-2*log(u1)) * cos(2*pi*u2)
  }
  Sigma.sqrt <- e_vectors %*%
    tcrossprod(diag(sqrt(e_values)), e_vectors)

  MVnorm <- matrix(mu, n, p, byrow = TRUE) + Z %*% Sigma.sqrt
  return(MVnorm)
}
mu <- c(1, 2, 3)
Sigma1 <- matrix(c(1, 0.4, 0.5, 0.2, 1, 0.52, 0.8, 0.6, 1),
                 3, 3, byrow = TRUE)
n <- 10
print(mymvrnorm(n, mu, Sigma1))
```

```
##             [,1]       [,2]       [,3]
##  [1,]  0.6509283  1.2235302  3.1401981
##  [2,] -1.0506448  2.0809556 -0.8389636
##  [3,]  1.2773876  2.1646853  4.6759197
##  [4,]  0.1466507 -0.2222981  2.3490302
##  [5,] -0.5537332  3.2872511  1.0944934
##  [6,]  0.5067483  2.5244796  2.2658486
##  [7,]  0.8618609  2.1256978  2.4786832
##  [8,]  1.0748014  1.3140052  3.0951378
##  [9,] -0.1721972  1.5006774  0.5297280
## [10,] -0.3700318  1.8797768  1.5751654
```

- Generate 200 random observations from the 3-dimensional multivariate normal distribution having mean vector $\mu = (0, 1, 2)$ and covariance matrix

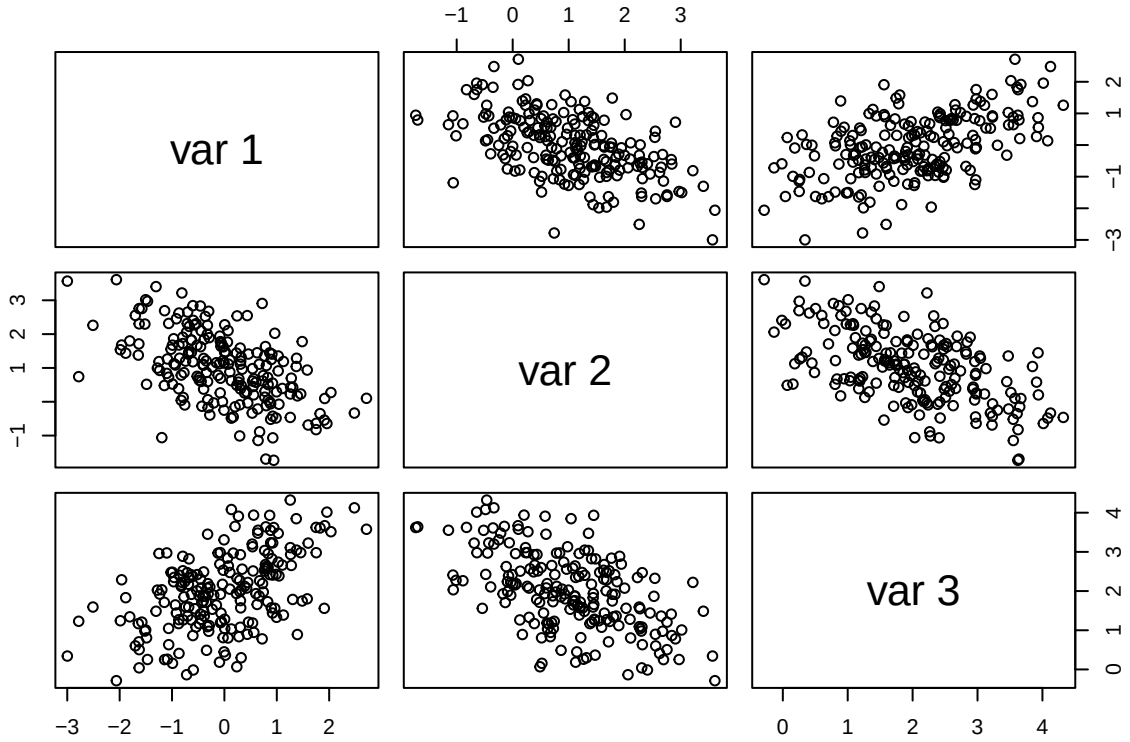
$$\Sigma = \begin{bmatrix} 1.0 & -0.5 & 0.5 \\ -0.5 & 1.0 & -0.5 \\ 0.5 & -0.5 & 1.0 \end{bmatrix}$$

  using your `mymvrnorm` function.

```r
mu <- c(0, 1, 2)
n <- 200
Sigma1 <- matrix(c(1,-0.5,0.5,-0.5,1,-0.5,0.5,-0.5,1),3,3,byrow = TRUE)
```

- Use the R `pairs` plot to graph an array of scatter plots for each pair of variables. For each pair of variables, (visually) check that the location and correlation approximately agree with the theoretical parameters of the corresponding bivariate normal distribution.

```
pairs(mymvrnorm(n, mu, Sigma1))
```



3. Truncated normal Here we consider sampling the $\mathrm{N}(0,1)$ distribution truncated to the interval $[a,\infty)$ where $a>0$. Inverting the truncated CDF leads to the formula $X = g_1(U) \equiv \Phi^{-1}(\Phi(a) + (1-\Phi(a))U)$ where $U \sim \mathbf{U}(0,1)$.

- Now find the smallest integer $a \geqslant 1$ for which $X = g_1(U)$ fails to work. For concreteness we can define failure as delivering at least one NaN or $\pm\infty$ when tested with $U \in \{(i-1/2)/1000 \mid, i = 1,2,\ldots,1000\}$.

```
trunc8 = function(a, U){
  X <- qnorm(pnorm(a) + (1 - pnorm(a)) * U)
  return(X)
}
find_a <- function(){
  for(a in 1:100){
    U <- (seq(1, 1000) - 0.5)/1000 #same as ppoints(1000)
    results <- sapply(U, function(u) trunc8(a,U))
    if(any(is.nan(results)) || any(is.infinite(results)))
      return(a)
  }
  return(NA)
}
find_a()
```

```
## [1] 8
```

- Consider $X = g_2(U) \equiv -\Phi^{-1}(\Phi(-a)(1-U))$ that also generates X from the same distribution. Find the smallest integer $a \geqslant 1$ at which g_2 fails using the same criterion as for g_1.

```r
trunc8_2 = function(a, U){
  X <- -qnorm(pnorm(-a) * (1 - U))
  return(X)
}
find_a_2 <- function(){
  for(a in 1:100){
    U <- (seq(1, 1000) - 0.5)/1000 #same as ppoints(1000)
    results <- sapply(U, function(u) trunc8_2(a,U))
    if(any(is.nan(results)) || any(is.infinite(results)))
      return(a)
  }
  return(NA)
}
find_a_2()
```

```
## [1] 38
```

4. Negative binomial distribution The negative binomial distribution with parameters $r \in \{1, 2, \ldots\}$ and $p \in (0, 1)$, denoted Negbin(r, p), has probability mass function

$$p_k = \mathbb{P}(X = k) = \begin{pmatrix} k+r-1 \\ r-1 \end{pmatrix} p^r (1-p)^k, \quad k = 0, 1, \ldots$$

It describes the number of failures before the r' th success in a sequence of independent Bernoulli trials with success probability p. For $r = 1$, it reduces to the geometric distribution. The negative binomial distribution has the following compound representation: $X \sim \text{Poi}(\lambda)$ for $\lambda \sim \text{Gam}(r) \times (1-p)/p$. We don't need r to be an integer. Writing

$$\begin{pmatrix} k+r-1 \\ r-1 \end{pmatrix} = \frac{(k+r-1)!}{(r-1)!k!} = \frac{\Gamma(k+r)}{\Gamma(r)\Gamma(k+1)}$$

yields a valid distribution

$$p_k = \mathbb{P}(X = k) = \frac{\Gamma(k+r)}{\Gamma(r)\Gamma(k+1)} p^r (1-p)^k$$

for $k \in \{0, 1, 2, \ldots\}$ for any real $r > 0$. The Poisson-gamma mixture representation also holds for $r > 0$. Using the compound representation we find that $\mathbb{E}(X) = r(1-p)/p$ and $\text{Var}(X) = r(1-p)/p^2$.

Generate 1000 variables following negative binomial distribution through the compound representation. Show that the sample mean and sample variance are close to the population mean and variance.

```r
neg_binomial <- function(n, r, p){
  gamma_sample <- rgamma(n, shape = r, rate = (1-p)/p)
  neg_bin_sample <- rpois(n, lambda = gamma_sample)
  mean_negbin <- mean(neg_bin_sample)
  var_negbin <- var(neg_bin_sample)
  return(c(mean_negbin, var_negbin))
}
n <- 1000
r <- 7
p <- 0.5
print(neg_binomial(n, r, p))
```

```
## [1]  7.00800 14.28422
```

```r
theoretical_mean <- r*(1-p)/p
theoretical_var <- r*(1 - p)/(p*p)
print(c(theoretical_mean, theoretical_var))
```

```
## [1]  7 14
```

5. Rshiny The dataset in counties.rds contains the name of each county in the United States, the total population of the county and the percent of residents in the county who are White, Black, Hispanic, or Asian. During the presentation, we use the percent of residents in the county. In the homework, we hope that you could establish the census app by filling the missing code.

Link of counties.rds: https://shiny.rstudio.com/tutorial/written-tutorial/lesson5/census-app/data/counties.rds

Link of helper.R: https://shiny.rstudio.com/tutorial/written-tutorial/lesson5/census-app/helpers.R

RShiny Codes:

```r
# Load packages
library(shiny)
library(maps)
library(mapproj)
# Load data
counties <- readRDS("/Users/devanandabaiju/Downloads/counties.rds")
# Source helper functions
source("/Users/devanandabaiju/R/helpers.R")
# User interface
ui <- fluidPage(
  titlePanel("censusVis"),
  sidebarLayout(
    sidebarPanel(
      helpText("Create demographic maps with
information from the 2010 US Census."),
      selectInput("var",
                  label = "Choose a variable to display",
                  choices = c('Percent White','Percent Black',
                              'Percent Hispanic','Percent Asian'),
                  selected = "Percent White"),
      sliderInput("range",
                  label = "Range of interest:",
                  min = 0, max = 100, value = c(0, 100))
    ),
    mainPanel(plotOutput("map"))
  )
)
# Server logic
server <- function(input, output) {
  output$map <- renderPlot({
    data <- switch(input$var,
                   'Percent White' = counties$white,
                   'Percent Black' = counties$black,
                   'Percent Hispanic' = counties$hispanic,
                   'Percent Asian' = counties$asian)
    color <- switch(input$var,
```

```
                    'Percent White' = 'darkgreen',
                    'Percent Black' = 'darkorange',
                    'Percent Hispanic' = 'darkred',
                    'Percent Asian' = 'darkblue')
    legend <- switch(input$var,
                     'Percent White' = 'Percent White',
                     'Percent Black' = 'Percent Black',
                     'Percent Hispanic' = 'Percent Hispanic',
                     'Percent Asian' = 'Percent Asian')
    percent_map(data, color, legend, input$range[1], input$range[2])
  })
}
# Run app
shinyApp(ui, server)
```

Output: