DATA130004: Homework 6

董晴园 14300680173

2017.12.2

1. Verify the counterexample in Midterm Exam: let X and W be independent variables, where X follows a standard normal distribution, and W is a Rademacher random variable, i.e., $W = 1$ with probability 0.5 and $W = -1$ with probability 0.5. Define $Y = WX$. Show that

   (a) X and Y are uncorrelated;

   $$\because E(X) = 0, E(W) = 0, and X, W are independent$$
   $$\therefore Cov(X, Y) = E(XY) - E(X)E(Y) = E(WX^2) - 0 = E(W)E(X^2) = 0$$

   (b) both X and Y have the same normal distribution;

   $$\begin{aligned} F_Y(y) = P(Y \le y) &= P(X \le y|W = 1)P(W = 1) + P(-X \le y|W = -1)P(W = -1) \\ &= 0.5F_X(y) + 0.5[1 - F_X(-y)] \\ &= 0.5F_X(y) + 0.5F_X(y) = F_X(y) \end{aligned}$$

   (c) and X and Y are not independent.

   If X and Y are independent, then $P(XY) = P(X)P(Y)$ is always true.
   Assume that $X = 0.1, Y = 0.05$, then $P(X = 0.1, WX = 0.05) = 0$, while $P(X = 0.1)$ and $P(WX = 0.05)$ are both positive, which means X and Y are not independent.

2. Exercise 6.6: Estimate the 0.025, 0.05, 0.95, and 0.975 quantiles of the skewness $b1$ under normality by a Monte Carlo experiment. Compute the standard error of the estimates using the normal approximation for the density (with exact variance formula $Var(hat(x)_q) = \frac{q(1-q)}{nf(x_q)^2}$). Compare the estimated quantiles with the quantiles of the large sample approximation $b1 \sim N(0, 6/n)$.

```
n <- 500 #sample sizes
sk <- function(x) {
  #computes the sample skewness coeff.
  xbar <- mean(x)
  m3 <- mean((x - xbar)^3)
  m2 <- mean((x - xbar)^2)
  return( m3 / m2^1.5 )
}

m <- 10000      #num. repl. each sim.
sk.sample <- numeric(m)
for (j in 1:m) {
    x <- rnorm(n)
    sk.sample[j] <- sk(x)
```

```
}
hist(sk.sample)
sk.sort <- sort(sk.sample)
q <- c(0.025,0.05,0.95,0.975)
q.est <- numeric(length(q))
q.norm <- numeric(length(q))
q.se <- numeric(length(q))
for(i in 1:length(q)){
  q.est[i] <- sk.sort[m*q[i]]
  q.norm[i] <- qnorm(q[i], 0, sqrt(6/n))
  q.se[i] <- (q[i]*(1-q[i]))/(n*(dnorm(q.norm[i]))^2)
}
q.est
[1] -0.2113199 -0.1762131  0.1820819  0.2154592
q.norm
[1] -0.2147033 -0.1801847  0.1801847  0.2147033
q.se
[1] 0.0003207557 0.0006166000 0.0006166000 0.0003207557
```

3. Exercise 6.9: Let X be a non-negative random variable with $\mu = E[X] < \infty$. For a random sample $x_1, ..., x_n$ from the distribution of X, the Gini ratio is defined by

$$G = \frac{1}{2n^2\mu} \sum_{j=1}^{n} \sum_{i=1}^{n} |x_i - x_j|.$$

The Gini ratio is applied in economics to measure inequality in income distribution. Note that G can be written in terms of the order statistics $x_{(i)}$ as

$$G = \frac{1}{n^2\mu} \sum_{i=1}^{n} (2i - n - 1)x_{(i)}.$$

If the mean is unknown, let $\hat{G}$ be the statistic G with μ replaced by $\overline{x}$. Estimate by simulation the mean, median and deciles of $\hat{G}$ if X is standard lognormal. Repeat the procedure for the uniform distribution and Bernoulli(0.1). Also construct density histograms of the replicates in each case.

In the case of standard lognormal:

```
G.hat <- function(x){
  n <- 1000
  x <- rlnorm(n)
  x.order <- sort(x)
  st <- numeric(n)
  for(i in 1:n){
    st[i] <- (2*i-n-1)*x[i]
  }
  return(sum(st)/(n^2*mean(x)))
}

m <- 10000
G.es <- numeric(m)
for (i in 1:m){
  G.es[i] <- G.hat(x)
}
hist(G.es)
mean(G.es)
[1] -0.0002665302
```

```
median(sort(G.es))
[1] -0.0003551756
quantile(sort(G.es), 0.1)
        10%
-0.03118238
```

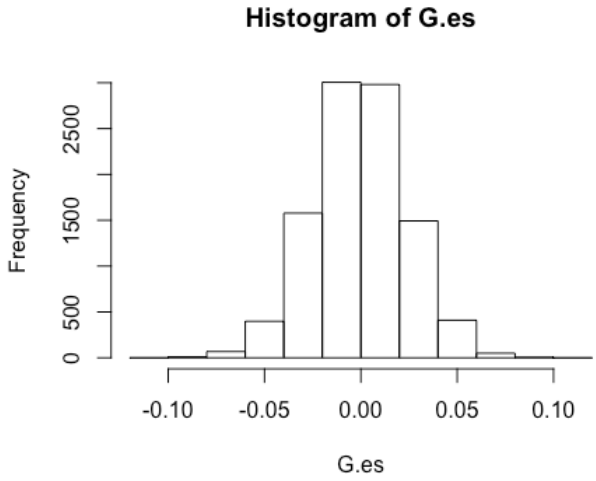

Figure 1: histogram of estimated G with standard lognormal distribution

Repeat the procedure for Unif(0,1):

```
> mean(G.es)
[1] 9.665334e-05
> median(sort(G.es))
[1] 9.1984e-05
> quantile(sort(G.es), 0.1)
        10%
-0.01357987
```

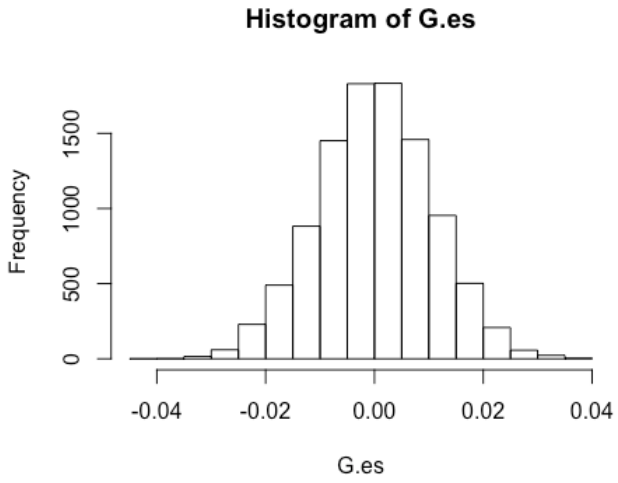

Figure 2: histogram of estimated G with N(0,1) distribution

Repeat again the procedure for Bernoulli(0.1):

```
> mean(G.es)
[1] -0.0006969219
> median(sort(G.es))
[1] -0.0007657143
> quantile(sort(G.es), 0.1)
        10%
-0.07109354
```

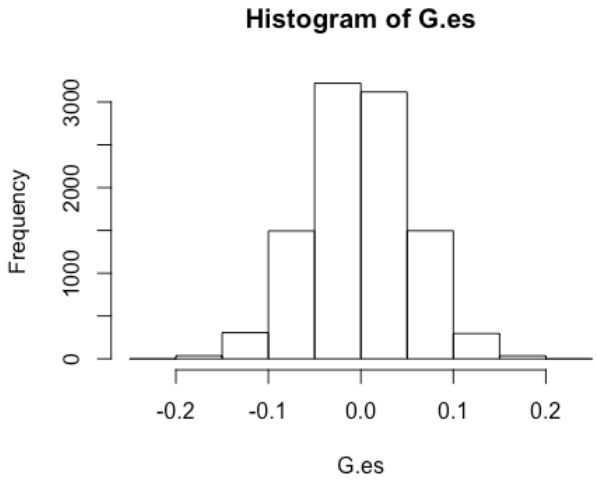

Figure 3: histogram of estimated G with Bernoulli(0.1) distribution

4. Exercise 6.10: Construct an approximate 95% confidence interval for the Gini ratio $\gamma = E[G]$ if X is lognormal with unknown parameters. Assess the coverage rate of the estimation procedure with a Monte Carlo experiment.

   95% confidence interval:

$$(\overline{\hat{G}} - 1.96\frac{s.e.(G)}{\sqrt{n}}, \overline{\hat{G}} + 1.96\frac{s.e.(G)}{\sqrt{n}})$$

   Assume the real coverage Gini ratio $\gamma = E[G] = 0$, then assess the coverage rate:

```
G.hat <- function(x){
  n <- 20
  x <- rlnorm(n)
  x.order <- sort(x)
  st <- numeric(n)
  for(i in 1:n){
    st[i] <- (2*i-n-1) * x[i]
  }
  return(sum(st)/(n^2 * mean(x)))
}

r <- 1000
y <- numeric(r)
for (i in 1:r){
  m <- 100
```

```
  G.es <- numeric(m)
  for (j in 1:m){
    G.es[j] <- G.hat(x)
  }
  U.CI <- mean(G.es)+1.96*sqrt(var(G.es)/m)
  L.CI <- mean(G.es)-1.96*sqrt(var(G.es)/m)
  y[i] <- ifelse(0 <= U.CI & 0 >= L.CI, 1, 0)
}
mean(y)
[1] 0.946
```

5. Project 6.D: Repeat Example 6.11 for multivariate tests of normality. Mardia defines multivariate kurtosis as

$$\beta_{2,d} = E\big[(X-\mu)^T\Sigma^{-1}(X-\mu)\big]^2.$$

For d-dimensional multivariate normal distributions the kurtosis coefficient is $\beta_{2,d} = d(d+2)$. The multivariate kurtosis statistic is

$$b_{2,d} = \frac{1}{n}\sum_{i=1}^{n}((X_i-\overline{X})^T\hat{\Sigma}^{-1}(X_i-\overline{X}))^2.$$

The large sample test of multivariate normality based on $b_{2,d}$ rejects the null hypothesis at significance level α if

$$\left|\frac{b_{2,d}-d(d+2)}{\sqrt{8d(d+2)/n}}\right| \geq \Phi^{-1}(1-\alpha/2).$$

However, $b_{2,d}$ converges very slowly to the normal limiting distribution. Compare the empirical power of Mardia's skewness and kurtosis tests of multivariate normality with the energy test of multivariate normality `mvnorm.etest (energy)`. Consider multivariate normal location mixture alternatives where the two samples are generated from `mlbench.twonorm` in the `mlbench` package.

We use `mardiaTest` in the `MVN` package to find the P value of Mardia's skewness and kurtosis tests, and compare their power with the energy test of multivariate normality under dimensions $2-8$.
First, we use a small sample ($n = 30$):

```
library(energy)
library(MVN)
library(mlbench)
alpha <- .1
n <- 30
m <- 500
test1 <- test2 <- test3 <- numeric(m)
sim <- matrix(0, 9, 4)

# estimate power
for (i in 2:10) {
  d <- i
  for (j in 1:m) {
    x <- mlbench.twonorm(n, d=d)[[1]]
    test1[j] <- as.integer(
      mardiaTest(x)@p.value.skew <= alpha)
    test2[j] <- as.integer(
      mardiaTest(x)@p.value.kurt <= alpha)
```

```
    test3[j] <- as.integer(
      mvnorm.etest(x, R=200)$p.value <= alpha)
  }
  print(c(d, mean(test1), mean(test2), mean(test3)))
  sim[i-1, ] <- c(d, mean(test1), mean(test2), mean(test3))
}

# plot the empirical estimates of power
plot(sim[,1], sim[,2], ylim = c(0, 1), type = "l",
     xlab = "dimension", ylab = "power")
lines(sim[,1], sim[,3], lty = 2, col=2)
lines(sim[,1], sim[,4], lty = 4, col=3)
abline(h = alpha, lty = 3)
legend("topright", 1, c("M skew", "M kurt", "energy"),
       lty = c(1,2,4), col=1:3, inset = .02)
```

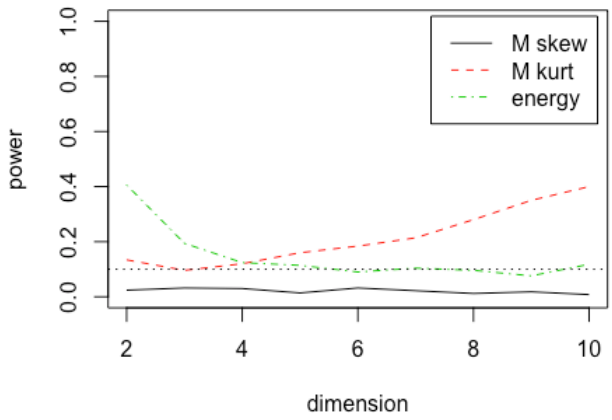

Figure 4: Empirical Power of tests of multivariate normality ($n = 30$)

Then we repeat the procedure with larger samples ($n = 100$ and $n = 500$), and the result is presented in the next page (Figure 5 & Figure 6).

As can be seen from the figures, the result is very interesting. No matter in which case, the empirical power of Mardia's skewness is very low, so it's not a good test we will use to test multivariate normality. Then compare Mardia's kurtosis test and energy test. In small samples ($n = 30$ and $n = 100$), energy test has higher power in low dimensions while Maedia's kurtosis has higher power in high dimensions; and in large samples ($n = 500$), energy test always has a higher power than Mardia's kurtosis.

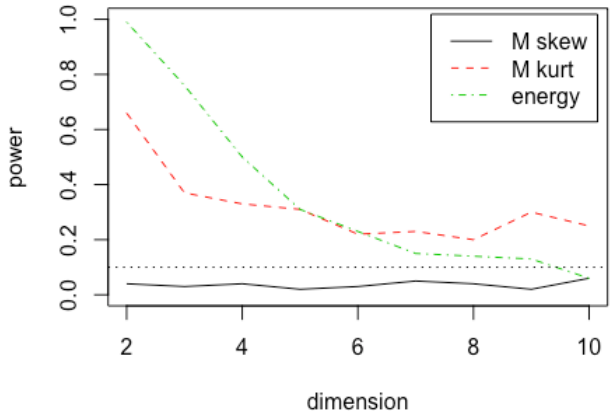

Figure 5: Empirical Power of tests of multivariate normality ($n = 100$)

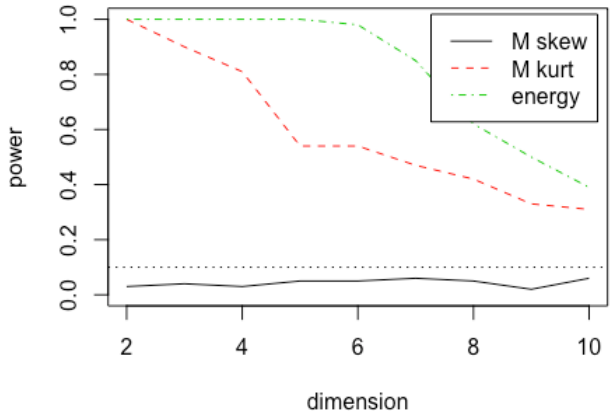

Figure 6: Empirical Power of tests of multivariate normality ($n = 500$)