STA511 Homework #5

Ezgi Karaesmen

11/30/15

1. Using simulation $H_0 : \mu = 0$ and $H_a : \mu \neq 0$ was tested on a sample of size of $n = 20$ with a Normal distribution of μ and $\sigma^2 = 1$, i.e. N(μ,1),

   (a) $Pr(Reject\, H_0|X_1, X_2, ..., X_{20} \sim N(0.5, 1))$ was estimated with a rejection criteria for H_0 determined as $|\frac{\bar{X}}{1\sqrt{n}}| > 1.96$. Probability of rejecting $H_0 : \mu = 0$ was estimated as `0.6076` by simulating random sampling with size n=20 from a N(0.5, 1) distribution, taking the sample mean and applying the rejection criteria. Then the number of the rejected simulations were divided by the total number of simulations.

   `R` code for the estimation is presented below:

```
sim <- 10000
n=20
rej <- NULL
xi <- rnorm(sim, 0.5, 1)

for(i in 1:sim){
  xsam <- sample(xi, n, replace=T)
  xbar <- mean(xsam)
  rej[i] <- (abs(xbar/(1/sqrt(n))) > 1.96)*1
}
sum(rej)/sim
[1] 0.6076
```

   (b) $Pr(Reject\, H_0|X_1, X_2, ..., X_{20} \sim N(0.5, 1))$ is the probability that 20 samples that were drawn from a N(0.5, 1) distribution does not have a population mean (i.e. μ) that is equal to 0. In terms of confidence intervals, when the paramater μ is unknown, we can simply compute the interval that μ falls into with 95% confidence by $Pr(-1.96 \leq |\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}| \geq 1.96)$. The rejection criteria used in the problem therefore contains a pivot function of $|\frac{\bar{X}-0}{1\sqrt{n}}|$, meaning that if the rejection criteria is met, with 95% confidence we can say that the data does not have a $\mu = 0$. The problem can be seen as a Monte Carlo integration problem as the testing requires the estimation of $Pr(|\frac{\bar{X}}{1\sqrt{n}}| > 1.96)$ which can also be defined as:

$$\begin{aligned} Pr\left(|\frac{\bar{X}}{1\sqrt{n}}| > 1.96\right) &= Pr\left(\frac{\bar{X}}{1\sqrt{n}} > 1.96 \mid \bar{X} \geq 0\right) \cup Pr\left(\frac{\bar{X}}{1\sqrt{n}} < -1.96 \mid \bar{X} < 0\right) \\ &= Pr\left(\bar{X} > \frac{1.96}{\sqrt{n}} \mid \bar{X} \geq 0\right) \cup Pr\left(\bar{X} < \frac{-1.96}{\sqrt{n}} \mid \bar{X} < 0\right) \end{aligned}$$

Since the distribution of X_i is N(0.5, 1), the distribution of $\bar{X}$ would be N(0.5, 1/n), where $n = 20$. Therefore it can be further defined as:

$$
\begin{aligned}
&= Pr\left(\bar{X} > \frac{1.96}{\sqrt{20}} \mid \bar{X} \geq 0\right) \; + \; Pr\left(\bar{X} < \frac{-1.96}{\sqrt{20}} \mid \bar{X} < 0\right) \\
&= \left(1 \, - \, Pr\left(\bar{X} < \frac{1.96}{\sqrt{20}} \mid \bar{X} \geq 0\right)\right) \; + \; Pr\left(\bar{X} < \frac{-1.96}{\sqrt{20}}\right) \\
&= \left(1 \, - \, F_{\bar{X}}\left(\frac{1.96}{\sqrt{20}} \mid \bar{X} \geq 0\right)\right) \; + \; F_{\bar{X}}\left(\frac{-1.96}{\sqrt{20}}\right), \quad where\, F_{\bar{X}} \, = CDF \sim N(0.5, (1/20)) \\
&= \left(1 \, - \int_0^{\frac{1.96}{\sqrt{20}}} \frac{1}{(1/20)\sqrt{2\pi}} e^{-(x-0.5)^2/2(1/20)^2} dx\right) \; + \; \int_{-\infty}^{\frac{-1.96}{\sqrt{20}}} \frac{1}{(1/20)\sqrt{2\pi}} e^{-(x-0.5)^2/2(1/20)^2} dx \\
&= \left(1 \, - \, \frac{2(1/20)}{\sqrt{2\pi}} \int_0^{\frac{1.96}{\sqrt{20}}} e^{\frac{-x^2+2x}{2(1/20)^2}} \, \frac{1}{2(1/20)^2} \, e^{\frac{-x}{2(1/20)^2}} \, dx\right) \; + \; \frac{2(1/20)}{\sqrt{2\pi}} \int_{-\infty}^{\frac{-1.96}{\sqrt{20}}} e^{\frac{-x^2+2x}{2(1/20)^2}} \, \frac{1}{2(1/20)^2} \, e^{\frac{-x}{2(1/20)^2}} \, dx \\
&= \left(1 \, - \, \frac{2(1/20)}{\sqrt{2\pi}} \int_0^{\frac{1.96}{\sqrt{20}}} g(x)\, f(x) dx\right) \; + \; \frac{2(1/20)}{\sqrt{2\pi}} \int_{-\infty}^{\frac{-1.96}{\sqrt{20}}} g(x)\, f(x) dx \\
&\quad where\, g(x) = e^{\frac{-x^2+2x}{2(1/20)^2}} \, and\, f(x) \sim \; Exponential(\theta = 2(1/20)^2), \quad where\, f_{Exp(\theta)}(x) = \frac{1}{\theta} e^{-x/\theta}
\end{aligned}
$$

To achieve this estimation via Monte Carlo integration, the sample of $X_1, X_2, ..., X_n$ can be generated from an Exponential($\theta = 2(1/20)^2$) distribution and can then be used to generate $\frac{\sum_{i=1}^{n} gx}{n}$. This cannot be achieved for the second part of the probability where $X_i < 0$ as the exponential distribution is only defined for $0 \leq x < \infty$. However as it can be seen in Figure 1, the area under the $\bar{X} \sim N(0, 1/20)$ pdf curve for $\bar{x} < -1.96/(\sqrt{20})$ becomes a very small number and approches to zero. Therefore it can be argued that it can be ignored.

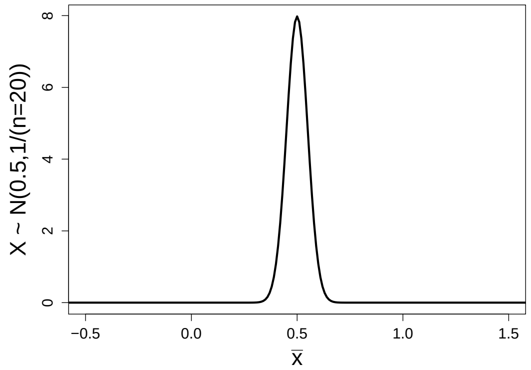

Figure 1: Distribution of the sample means of variables drawn from a N(0.5,1) distribution. Presented density function ~ N(0.5, 1/20)

2. $X_1, X_2, ; ..., X_n$ follow a Binomial distribution with size 10 and probability of success θ.

   (a) MLE for θ (i.e. $\hat{\theta}_{MLE}$) was computed as follows:

$$
\begin{aligned}
f_{Bin(10,\theta)}(x) &= \frac{10!}{x!(10-x)!}\theta^{x}(1-\theta)^{10-x} \\
L(\theta) &= \prod_{i=1}^{n} \frac{10!}{x_i!(10-x_i)!}\theta^{x_i}(1-\theta)^{10-x_i} \\
L(\theta) &= \left(\prod_{i=1}^{n}\left(\frac{10!}{x_i!(10-x_i)!}\right)\right)\theta^{\sum\limits_{i=1}^{n} x_i}(1-\theta)^{\sum\limits_{i=1}^{n}(n-x_i)} \\
\ell(\theta) &= ln\left(\prod_{i=1}^{n}\left(\frac{10!}{x_i!(10-x_i)!}\right)\right) \;+\; \sum_{i=1}^{n} x_i\, ln(\theta) \;+\; \sum_{i=1}^{n}\big(10-x_i\big)ln(1-\theta) \\
\frac{d\ell(\theta)}{d\theta} &= \frac{1}{\theta}\sum_{i=1}^{n} x_i \;-\; \left(\frac{1}{1-\theta}\right)\sum_{i=1}^{n}\big(10-x_i\big) \\
&= \frac{1}{\hat{\theta}}\sum_{i=1}^{n} x_i \;-\; \left(\frac{1}{1-\hat{\theta}}\right)\sum_{i=1}^{n}\big(10-x_i\big) \overset{\text{set}}{=} 0 \\
&= (1-\hat{\theta})\;\sum_{i=1}^{n} x_i \;-\; \hat{\theta}\left(10n - \sum_{i=1}^{n} x_i\right) \;=\; 0 \\
&= \sum_{i=1}^{n} x_i - \hat{\theta}\sum_{i=1}^{n} x_i - 10n\hat{\theta} \;+\; \hat{\theta}\sum_{i=1}^{n} x_i \;=\; 0 \\
&= \sum_{i=1}^{n} x_i \;=\; 10n\hat{\theta} \\
\hat{\theta}_{MLE} \;&=\; \frac{\sum\limits_{i=1}^{n} x_i}{10n}
\end{aligned}
$$

   (b) Moment of methods estimator was found for θ (i.e $\tilde{\theta}_{MOM}$) as follows:

$$
E(X_{Bin(n,\theta)}) = n\theta
$$

$$
\begin{aligned}
E(X_{data}) &= \;10\theta \;=\; \bar{X} \\
&= 10\theta \;=\; \frac{\sum\limits_{i=1}^{n} x_i}{n} \\
&= \tilde{\theta}_{MOM} \;=\; \frac{\sum\limits_{i=1}^{n} x_i}{10n}
\end{aligned}
$$

3. The data X_1, X_2, ..., X_n was generated by drawing Y_1, Y_2, ... , $Y_n \sim N(0,1)$ and setting $X_i = e^{Y_i}, i = 1, \ldots, n$ where $n = 25$. The skewness function was estimated by the Monte Carlo (MC) integration where the function for MC integration was determined as follows:

$$\begin{aligned}\theta F &= \int \frac{(x-\mu)^3}{\sigma^3}\, f(x)\, dx \\ &= \int g(x)\, f(x)\, dx \quad , where\; g(x) \;=\; \frac{(x-\mu)^3}{\sigma^3} \\ &= E[g(x)]\end{aligned}$$

$$\text{Estimation by MC integration,} \quad E[g(x)] = \frac{\sum_{i=1}^{n}(x_i-\mu)^3}{n\sigma^3} \;=\; \hat{S}$$

Using the plug-in estimators of $\bar{x}$ for μ and $\hat{\sigma}$ for σ $\hat{S}$ was estimated as `1.97`. Also 100 confidence intervals were generated through repetitions of the bootstrapping method. For bootstrapping, 25 random data points were sampled with replacement from the data (that was generated as described above). None of the intervals contained the true skewness value of `6.185`, meaning 0% of the times . This indicates that $g(x)$ is not a good estimator of skewness.

```
n <- 25
set.seed(938)
xi <- exp(rnorm(n,0,1))
mu<- mean(xi)
s <- sqrt(var(xi))
s_hat <- sum((xi - mu)^3/s^3)/n # estimated skewness value

nsim1 <- 1000
nsim2 <- 100
s_err <- NULL
Tboot <- NULL


for(i in 1:nsim2){ # repeating bootstrapping
  for(j in 1:nsim1){ # bootstrapping
    boot <- sample(xi, 25, replace=T)
    skew <- (boot - mu)^3/s^3
    Tboot[j] <- mean(skew)
  }
s_err[i] <- sqrt(var(Tboot))
}

ci1 <- s_hat - s_err*1.96
ci2 <- s_hat + s_err*1.96
CIs <- data.frame(ci1,ci2)
real_skew <- (exp(1)+2)*sqrt(exp(1)-1) # true skewness
Tval <- cbind( CIs[,1] <= real_skew & CIs[,2] >= real_skew )*1 # of CI with true value
sum(Tval)/length(Tval) * 100 #\% of the CI containing true value
```