Tutorial 2

Rui Qiu

2018-02-28

(**2.1**) Given that $X_1, X_2, \ldots, X_n$ is a random sample from $U[0, \theta]$, find the p.d.f. of $X_{(n)}$, the largest of X_i.

Show that $2\bar{X}$ and $(n+1)X_{(n)}/n$ are both cosistent estimators of θ and compare their variances.

Solution:

Suppose $Y_1, Y_2, \ldots, Y_n$ are the order statistics of $X_1, X_2, \ldots, X_n$ with $Y_1 \leq Y_2 \leq \cdots \leq Y_n$, i.e. $Y_n = X_{(n)}$. The distribution of Y_i is an upper tail of a binomial distribution. Consider $X \leq y$ as an success and $F(y) = P(X \leq y)$ is the probability of success, then the drawing of each sample item is just a Bernoulli trial.

Hence the distribution of Y_i can be expressed as:

$$F_{Y_i(y)} = P(Y_i \leq y) = \sum_{k=i}^{n} \binom{n}{k} F(y)^k (1 - F(y))^{n-k}$$

We can take derivative to solve for the p.d.f. of Y_i. But this can also be generated through the following reasoning:

- The density $f_{Y_i}(y)$ is the probability that ith order statistics Y_i is right around y, with $i-1$ samples less than y and $n-i$ samples greater than y.
- To represent with statistical expression, we have

$$F(y)^{i-1} f(y) (1 - F(y))^{n-i}$$

- $F(y)^{i-1}$ is the probability that $i-1$ samples less than y.
- $f(y)$ is the probability that one sample is y.
- $(1 - F(y))^{n-i}$ is the probability that $n-i$ samples greater than y.
- This is just one possible combination, we should multiply it by multinomial coefficient.

Therefore:

$$f(Y_i(y)) = \frac{n!}{(i-1)!1!(n-i)!} F(y)^{i-1} f(y) (1 - F(y))^{n-i}$$

We plug in $i = 1$ and $Y_i = X_{(n)}$, then

$$f_{X_{(n)}}(y) = n \cdot F(y)^{n-1} f(y)$$

Recall that for a uniform distribution, the c.d.f. and p.d.f. are as following:

$$F(y) = \begin{cases} 0 & y < 0 \\ \frac{y}{\theta} & 0 \leq y < \theta \\ 1 & y \geq \theta \end{cases}$$

$$f(y) = \begin{cases} \frac{1}{\theta} & 0 < y < \theta \\ 0 & \text{otherwise} \end{cases}$$

Hence,

$$f_{X_{(n)}}(y) = n \cdot \left(\frac{y}{\theta}\right)^{n-1} \frac{1}{\theta} = \frac{ny^{n-1}}{\theta^n}$$

To show that $2\bar{X}$ and $\frac{(n+1)X_{(n)}}{n}$ are both consistent, it is equivalent to show the sufficient condition that both the bias and the variance of these estimators approach zeros when $n \to \infty$. Therefore, we do the following calculations:

$$\begin{aligned}
2\bar{X} &\approx X_{(n)} \\
E(2\bar{X}) &= \int_0^\theta x \cdot \frac{nx^{n-1}}{\theta^n} dx \\
&= \int_0^\theta \frac{n}{\theta^n} x^n dx \\
&= \frac{n}{\theta^n} \frac{1}{n+1} x^{n+1} \Big|_0^\theta \\
&= \frac{n}{\theta^n} \frac{1}{n+1} (\theta^{n+1} - 1) \\
&\approx \frac{\theta^{n+1}}{\theta^n} \\
&= \theta \\
V(2\bar{X}) &= E(X^2) - E(X)^2 \\
&= \int_0^\theta \frac{nx^{n+1}}{\theta^n} dx - \theta^2 \\
&= \frac{n}{\theta^n} \frac{1}{n+2} x^{n+2} \Big|_0^\theta - \theta^2 \\
&= \frac{n}{\theta^n (n+2)} (\theta^{n+2}) - \theta^2 \\
&\approx \frac{\theta^{n+2}}{\theta^n} - \theta^2 \\
&= 0
\end{aligned}$$

$$
\begin{aligned}
E(\frac{n+1}{n}X_{(n)}) &= \int_0^\theta \frac{n+1}{n}x \cdot \frac{nx^{n-1}}{\theta^n}dx \\
&= \frac{n+1}{\theta^n}\frac{1}{n+1}x^{n+1}\Big|_0^\theta \\
&= \frac{\theta^{n+1}}{\theta^n} \\
&= \theta \\
V(\frac{n+1}{n}X_{(n)}) &= E\left(\frac{(n+1)^2}{n^2}x^2\right) - \theta^2 \\
&= \int_0^\theta \left(\frac{n+1}{n}\right)^2 \cdot n\frac{x^{n+1}}{\theta^n}dx - \theta^2 \\
&= \frac{(n+1)^2}{n(n+2)}x^{n+2} \cdot \frac{1}{\theta^n}\Big|_0^\theta - \theta^2 \\
&= \frac{n^2+2n+1}{n^2+2n} \cdot \theta^2 - \theta^2 \\
&\approx 0
\end{aligned}
$$

Hence, both are consistent estimators.

(**2.2**) Suppose that $\hat{\theta}_1$ and $\hat{\theta}_2$ are independent unbiased estimators of a parameter θ, with variances σ_1^2, σ_2^2, respectively, and $\tilde{\theta} = k_1\hat{\theta}_1 + k_2\hat{\theta}_2$, where k_1, k_2 are constants. Find the values of k_1, k_2 for which $\tilde{\theta}$ is unbiased and has the smallest possible variance.

Solution:

$$
\begin{aligned}
E(\hat{\theta}_1) = E(\hat{\theta}_2) &= \theta \\
V(\hat{\theta}_1) = \sigma_1^2 &= E(\hat{\theta}_1^2) - E(\hat{\theta}_1)^2 \\
V(\hat{\theta}_2) = \sigma_2^2 &= E(\hat{\theta}_2^2) - E(\hat{\theta}_2)^2 \\
\tilde{\theta} &= k_1\hat{\theta}_1 + k_2\hat{\theta}_2
\end{aligned}
$$

Since $E(\tilde{\theta}) = E(k_1\hat{\theta}_1 + k_2\hat{\theta}_2) = \theta$, $\hat{\theta}_1$ and $\hat{\theta}_2$ are independent, then we can see $k_1\theta + k_2\theta = \theta$, therefore $k_1 + k_2 = 1$.

We want $V(\tilde{\theta})$ to be as small as possible:

$$
\begin{aligned}
V(\tilde{\theta}) &= E(\tilde{\theta}^2) - E(\tilde{\theta})^2 \\
&= E(k_1^2\hat{\theta}_1^2 + 2k_1k_2\hat{\theta}_1\hat{\theta}_2 + k_2^2\hat{\theta}_2) - \theta^2 \\
&= k_1^2E(\hat{\theta}_1^2) + 2k_1k_2E(\hat{\theta}_1\hat{\theta}_2) + k_2^2E(\hat{\theta}_2) - \theta^2 \\
&= k_1^2(\sigma_1^2 + \theta^2) + k_2^2(\sigma_2^2 + \theta^2) + 2k_1k_2\theta^2 - \theta^2 \\
&= \theta^2(k_1^2 + k_2^2 + 2k_1k_2) - \theta^2 + k_1\sigma_1^2 + k_2\sigma_2^2 \\
&= \theta^2(k_1 + k_2)^2 - \theta^2 + k_1^2\sigma_1^2 + k_2^2\sigma_2^2 \\
&= k_1^2\sigma_1^2 + k_2^2\sigma_2^2
\end{aligned}
$$

Let $k_2 = 1 - k_1$ then,

$$
\begin{aligned}
V(\tilde{\theta}) &= k_1^2\sigma_1^2 + k_2^2\sigma_2^2 \\
&= k_1^2\sigma_1^2 + (1-k_1)^2\sigma_2^2 \\
&= (\sigma_1^2+\sigma_2^2)k_1^2 - 2\sigma_2^2 k_1 + \sigma_2^2
\end{aligned}
$$

Solve $2(\sigma_1^2+\sigma_2^2)k_1 - 2\sigma_2^2 = 0$, then the minimizer is

$$
k_1 = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}
$$

(**2.4**) Suppose $\hat{\theta}$ is an estimator for θ with probability function $Pr[\hat{\theta}=\theta] = (n-1)/n$ and $Pr[\hat{\theta}=\theta+n] = 1/n$ (and no other values of $\hat{\theta}$ are possible); show that $\hat{\theta}$ is consistent but that bias $(\hat{\theta}) \not\to$ as $n \to \infty$.

Solution:

By definition, $Pr\left(|\hat{\theta}-\theta| > \epsilon\right) \to 0$ as $n \to \infty$ then such estimator should be consistent. For our case, when n is close to positive infinity, the probabilty that $\theta = \hat{\theta}$ is close to 1. Thus, $\hat{\theta}$ should be a consistent estimator. But our usual sufficient condition is not necessary, due to the following reasoning:

$$
\begin{aligned}
\text{bias}(\hat{\theta}) &= E(\hat{\theta}) - \theta \\
&= \frac{n-1}{n}\theta + \frac{1}{n}(\theta+n) - \theta \\
&= \theta - \frac{1}{n}\theta + \frac{1}{n}\theta + 1 - \theta \\
&= 1 \neq 0
\end{aligned}
$$

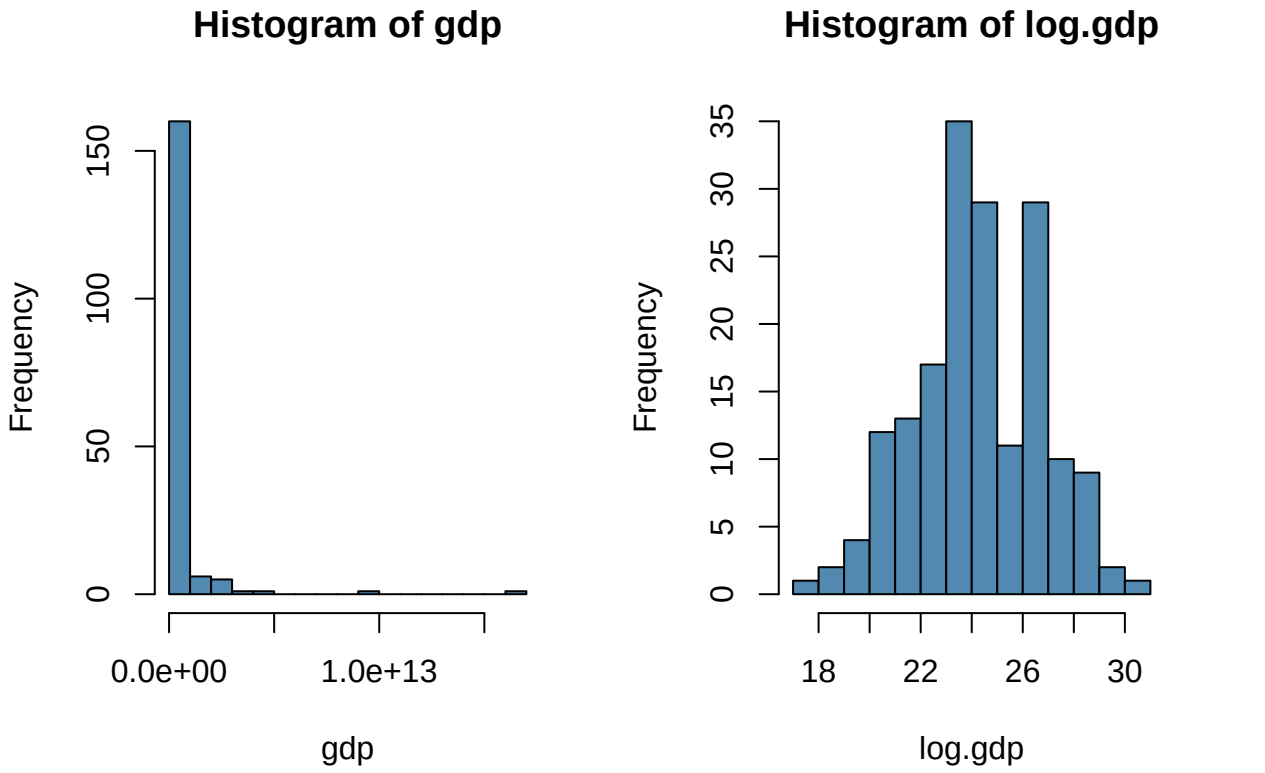
Thus, $\hat{\theta}$ is biased though.

(a)

The original GDP data is rather ugly as we plot the density and find out it is heavily right-skewed, which is somehow reflecting the reality of this world, that the development of this world is unbalanced. But after taking natural logarithm, the central tendency of data is shifted right a little bit. Though not perfect, not even unimodal, it looks better than before, thus further approaches can be applied to the transformed data.

Note that, for all countries with unavailable data (`NA`), we simply ignored them when processing the data.

```
data <- read.table('gdp2013.txt', header = TRUE, sep =' ')
data <- na.omit(data)
gdp <- data$Y2013
log.gdp <- log(data$Y2013)
par(mfrow=c(1,2))
hist(gdp,breaks = 12,col='#5289B1')
hist(log.gdp,breaks = 12,col='#5289B1')
```



(b) The 6-number summary is shown below:

```
summary(log.gdp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.46   22.73   24.15   24.23   26.13   30.45
```

(c)

There is only one country, *Tuvalu* with a GDP below the critical value of 1st Quantile - 1.5 IQR, which is identified as an outlier.

```
IQR <- 26.13-22.73
lbound <- 22.73-1.5*IQR
ubound <- 26.13+1.5*IQR
data[log(data$Y2013)>ubound,]
```

```
## [1] Country.Name Country.Code Y2013
## <0 rows> (or 0-length row.names)
```

```
data[log(data$Y2013)<lbound,]
```

```
##     Country.Name Country.Code    Y2013
## 199       Tuvalu          TUV 38134775
```

(d)

The best guesses for T_1 and T_2 are the sample mean and sample variance from this data set.

```
mean(log.gdp)
```

```
## [1] 24.22638
```

```
var(log.gdp)
```

```
## [1] 6.093419
```

```
(length(log.gdp)-1)/length(log.gdp)*var(log.gdp)
```

```
## [1] 6.058599
```

The expected sample mean is 24.22638 and expected sample variance is 6.058599.

In fact, if we run a Monte Carlo simulation wtih 10000 iterations, we will have a very similar result.

```
set.seed(8027)
skip <- 10

sample.mean <- c(); sample.var <- c()

for (sim in 1:10000) {
    hide.index <- sample(1:length(log.gdp), skip)
    new.gdp <- log.gdp[-hide.index]
    sample.mean[sim] <- mean(new.gdp)
    sample.var[sim] <- var(new.gdp)
}

mean(sample.mean)
```

```
## [1] 24.22584
```

```
mean(sample.var)*(length(log.gdp)-skip-1)/(length(log.gdp)-skip)
```

```
## [1] 6.056588
```