STA238 Tutorial 2

Luis Ledesma

2023-02-01

1 Announcements

- You can upload your work on Crowdmark from the end of the tutorial session to 5pm Friday of that week.
- All questions must be solved using RStudio.

2 Recall: Last tutorial

We did a brief overview of R and RStudio. Again, **please make sure that you have properly set up R and RStudio in your computer**. See the base R cheatsheet for a brief recap. In addition, we also looked at ways at simulating random numbers from a specific probability distribution, along with other related functions.

We also looked at the central limit theorem: For X_i iid, with finite mean and variance:

$$\frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \to_d N(0,1)$$

Where convergence is in distribution. The tutorial activity consisted in showing that the above holds for large sample sizes with a $\text{Bin}(t,p)$ distribution.

3 Some notes on order statistics

Let $Y_1, Y_2, \dots, Y_n$ denote random variables taken from a sample. Define $Y_{(1)}$ as the minimum of the random variables, and $Y_{(n)}$ as their maximum. In more generality, one can define the order statistic $Y_{(i)}$, where:

$$Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$$

$Y_{(i)}$ can be understood as the random variable that is in position i, for $i \in \{1, \dots, n\}$.

3.1 Maximum of random variables

Assuming the same set-up as above, the distribution function of $Y_{(n)}$ will be given by (where $f(y)$ is the density function of Y_i and $F(y)$ the distribution function of Y_i):

$$F_{Y_{(n)}}(y) = [F(y)]^n$$

And the density function will be:

$$f_{Y_{(n)}}(y) = nf(y)[F(y)]^{n-1}$$

3.2 Minimum of random variables

Correspondingly, the distribution function of $Y_{(1)}$ will be given by:

$$F_{Y_{(1)}}(y) = 1 - [1 - F(y)]^n$$

And the density function will be:

$$f_{Y_{(1)}}(y) = nf(y)[1 - F(y)]^{n-1}$$

4 Unbiased estimators and variance

In the tutorial activity, we have a sample of size 3 taken from an exponential distribution with parameter θ (this is the scale parameter parametrization):

$$f(y) = \begin{cases} \frac{1}{\theta}e^{-y/\theta} & \text{y} > 0 \\ 0 & \text{elsewhere} \end{cases}$$

In other words, $Y \sim \text{Exp}(1/\theta)$, which means that Y has an exponential distribution with parameter $1theta$.

Theorem: If $X \sim \text{Exp}(\lambda)$, then $E(X) = \frac{1}{\lambda}$ and $\text{var}(Y) = \frac{1}{\lambda^2}$.

To show this, recall that for a continuous random variable X, $E(X) = \int_{-\infty}^{\infty} xf(x)dx$ where f is the density of X. Moreover, $\text{var}(X) = E(X^2) - E(X)^2$.

Thus, $E(Y) = \theta$ and $\text{var}(Y) = \theta^2$.

Now, suppose that we have defined the estimators:

1. $\hat{\theta}_1 = Y_1$
2. $\hat{\theta}_2 = \frac{Y_1+Y_2}{2}$
3. $\hat{\theta}_3 = \frac{Y_1+2Y_2}{3}$
4. $\hat{\theta}_4 = Y_{(1)}$
5. $\hat{\theta}_5 = \hat{Y}$

We will compute the mean of $\hat{\theta}_1$ and $\hat{\theta}_2$. The mean of $\hat{\theta}_3$ and $\hat{\theta}_5$ should follow easily from these calculations.

$$E(\hat{\theta}_1) = E(Y_1) = \theta$$

$$E(\hat{\theta}_2) = E\left(\frac{Y_1 + Y_2}{2}\right) = \frac{1}{2}E(Y_1 + Y_2) = \frac{1}{2}[E(Y_1) + E(Y_2)] = \theta$$

Question: Why can we split the expectation of the sum?

Now:

$$E(\hat{\theta}_4) = E(Y_{(1)})$$

From our previous formula, we find that:

$$f_{Y_{(1)}}(y) = 3\frac{1}{\theta}e^{-y/\theta}[1 - F(y)]^2$$

$F(y)$ can be found by integrating the density function:

$$F(y) = \int_0^y \frac{1}{\theta}e^{-y/\theta}ds = -e^{-s/\theta}|_0^y = 1 - e^{-y/\theta}$$

Theorem: If X is a continuous random variable, then under enough regularity conditions, the distribution function of X will be $F(x) = \int_{-\infty}^{x} f(s)ds$ where f is the density function of X. Similarly, $f(x) = F'(x)$.

Using this, one gets for $y > 0$:

$$f_{Y_{(1)}}(y) = 3\frac{1}{\theta}e^{-y/\theta}[e^{-y/\theta}]^2 = \frac{3}{\theta}e^{-3y/\theta}$$

And 0 otherwise, as the density is 0 for $y \leq 0$. We recognize this as an exponential distribution with parameter $3/\theta$, so by using our previous theorem, $E(\hat{\theta}_4) = E(Y_{(1)}) = \theta/3$.

What does it mean for an estimator to be unbiased? Using this, which estimators would be unbiased?

4.1 Computing the variance of unbiased estimators

It is clear that $\text{var}(\hat{\theta}_1) = \theta^2$. Now:

$$\text{var}(\hat{\theta}_2) = \text{var}\left(\frac{Y_1 + Y_2}{2}\right) = \frac{1}{4}\text{var}(Y_1 + Y_2) = \frac{1}{4}[\text{var}(Y_1) + \text{var}(Y_2)] = \frac{\theta^2}{2}$$

Question: Why can we split the variance of the sum?

Using a similar procedure as the above:

$$\text{var}(\hat{\theta}_3) = \text{var}\left(\frac{Y_1 + 2Y_2}{3}\right) = \frac{1}{9}\text{var}(Y_1 + 2Y_2) = \frac{1}{9}[\text{var}(Y_1) + 4\text{var}(Y_2)] = \frac{5}{9}\theta^2$$

An identical procedure can be used to find $\text{var}(\hat{\theta}_5)$. Now, between the unbiased estimators, which one would have the smallest variance?

Challenge: Compute the variance for $\hat{\theta}_4$.