1. Explain how these estimators relate to your answers to the Exercise in the previous chapter (Ex-3.6.3).

$$E(X) = \frac{\alpha}{\alpha+1} = 1 - \frac{1}{\alpha+1} \Leftrightarrow 1 - E(X) = \frac{1}{\alpha+1} \Leftrightarrow \frac{1}{1-E(X)} = \alpha + 1 \Leftrightarrow \frac{1}{1-E(X)} - 1 = \alpha \Leftrightarrow \alpha = \frac{E(X)}{1-E(X)}$$

The expectation of a sample is the mean of that sample.

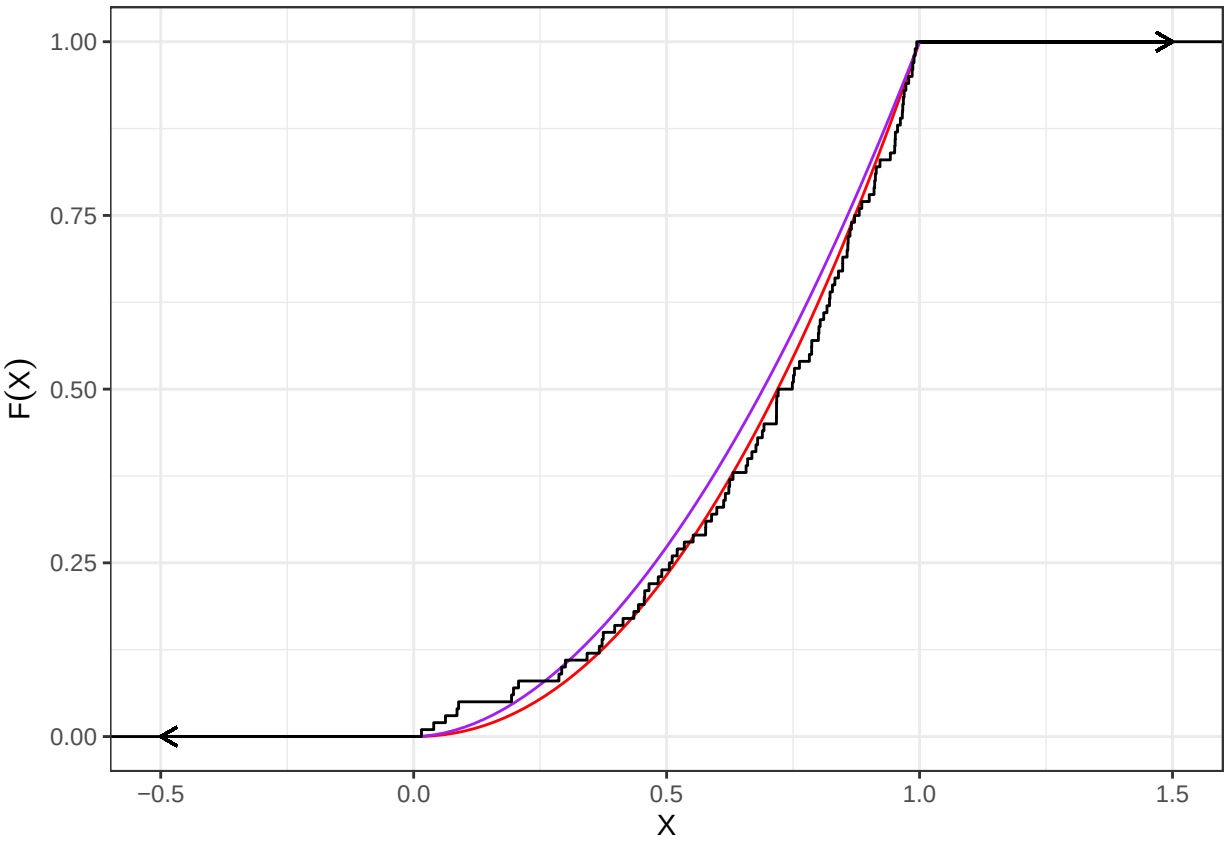
$$\hat{\alpha} = \frac{\bar{x}}{1-\bar{x}}$$

2: Download the dataset `ExBetaSim_1.csv` from the data folder, which contains a simulated sample from this distribution. Use both estimators to estimate α.

For the data set and estimator 1, the estimated $\hat{\alpha}$ is 2.1052162. For the data set and estimator 2, the estimated $\tilde{\alpha}$ is 1.8726.

3. Plot the cdf implied by your estimates, and also show the "empirical cumulative density function" of your data, which you can do in `ggplot2` using `stat_ecdf`.

The red line is the first estimator, and the purple line is the second.



4. (Simulation exercise) Fix $\alpha = 0.7$. Simulate some properties of these estimators for a sample size of $N = 30$. Are the estimators biased? Does one stand out as better than the other? Hint: You can simulate the distribution of X by transforming uniform random numbers. Specifically, if $U \sim \mathrm{U}[0,1]$, then: $X = U^{\frac{1}{\alpha}}$ will have the correct distribution.

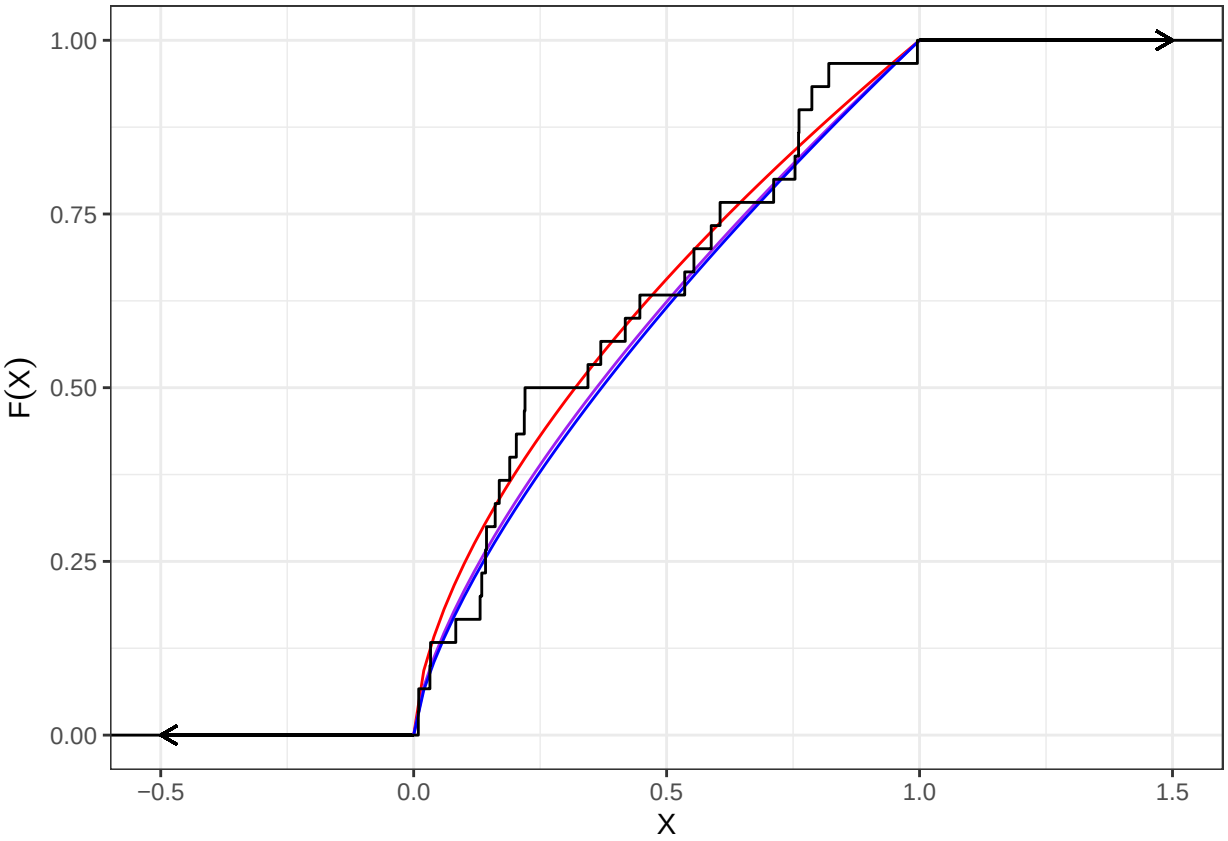
The random sample taken from F is:

0.1313068, 0.1691426, 0.2186458, 0.0091835, 0.1439862, 0.142047, 0.0333814, 0.7538026, 0.1345585, 0.3445668, 0.6056829, 0.5882005, 0.5357871, 0.8205557, 0.1611333, 0.009935, 0.5539809, 0.4471379, 0.0832208, 0.0321068, 0.1898836, 0.2200646, 0.7115699, 0.7616829, 0.7608469, 0.9957098, 0.3698259, 0.787154, 0.4182283, 0.2028776

The seed for generating this specific sample was 123.

This sample has an $\hat{\alpha}$ of 0.6073902 This sample has an $\tilde{\alpha}$ of 0.6812157

Here is a plot of the data (the red line is an estimate based on the sample, the blue line is the actual cdf):



I don't think either estimator is particularly biased. I have experimented with larger sample sizes - (which you can also do by changing the N variable in the Rmd file) - and the results speak for themselves. At 10000 values, the data was a perfect match. Is one more biased than the other at lower values? I don't think so. I could not find a problem with either for the seeds I tested. I was able to find a few seeds where the second was significantly better than the first. However, I had also found a few seeds where the first was closest.

In conclusion: They're both kinda good. If I was forced to pick, I'd go for the second.