1. In previous lecture, we generated pseudo populations based on parametric model $N(\mu, \sigma^2)$. Now we show a technique called bootstrap where no parametric assumptions are made. See Efron and Tibshirani (1993) An introduction to the bootstrap for detail. Given a random sample $x = (x_1, \ldots, x_n)$, The CDF of true population, $F$ was estimated as the proportion of the total sample points that is equal or less than $x$, i.e.

$$\hat{F}(x) = \sum_{i=1}^{n} I_{(-\infty, x_i]}(x).$$

Note that $\hat{F}(x(i)) = i/n$ and $\hat{F}(x(i+1)) = (i+1)/n$. So we are treating each sample point to occur with equal probability $1/n$. Based on this idea, we resample with replacement from $x$ and we will denote the new sample as $x^*$. If we resample $m$ times, we denote them by $x^1, \ldots, x^m$.

2. Suppose $X = (X_1, \ldots, X_n) \sim F$. Suppose we are estimating $\theta$ by $\hat{\theta}(X)$. The distribution of $\hat{\theta}(X)$ would be an interest for statistical inference but it might be difficult to compute analytically. In this situation, we generate $m$ bootstrap data $X^*_{i1}, \ldots, X^*_im$ and obtain the bootstrap replications of $\theta$,

$$\hat{\theta}^* = \hat{\theta}(X^i).$$

These bootstrap resamples provide us with an estimate of the distribution of $\theta$. In particular, we can estimate $E \hat{\theta}(X)$ and $\text{Var} \hat{\theta}(X)$ by

$$E \hat{\theta}(X) \approx \bar{\theta} = \frac{m}{m} \sum_{i=1}^{m} \hat{\theta}(X^i)/m$$

$$\text{Var} \hat{\theta}(X) \approx \sum_{i=1}^{m} \left( \hat{\theta}(X^i) - \bar{\theta} \right)^2 / (m - 1)$$

From strength.data, let us estimate the population mean and variance based on 200 bootstrap replications. For population mean, we use $\hat{\theta}(X) = \bar{X}$. For population variance, we use $\hat{\theta}(X) = \sqrt{n\bar{X}}$.