4. Order Statistics

Basic Theory

Random Variables

Suppose that the objects in our population are numbered from 1 to m, so that $D = \{1, 2, ..., m\}$. For example, the population might consist of manufactured items, and the labels might correspond to serial numbers. As in the basic sampling model we select n objects at random, without replacement from D:

$$\boldsymbol{X} = (X_1, X_2, ..., X_n)$$

where $X_i \in D$ is the i^{th} object chosen. Recall that $\boldsymbol{X}$ is uniformly distributed over the set of permutations of size n chosen from D. Recall also that

$$\boldsymbol{W} = \{X_1, X_2, ..., X_n\}$$

is the unordered sample, which is uniformly distributed on the set of combinations of size n chosen from D.

For $i \in \{1, 2, ..., n\}$ let

$$X_{n,i} = i^{\text{th}} \text{ smallest element of } \{X_1, X_2, ..., X_n\}$$

The random variable $X_{n,i}$ is known as the **order statistic** of order i for the sample $\boldsymbol{X}$. Note that in particular, the extreme order statistics are

$$X_{n,1} = \min\{X_1, X_2, ..., X_n\}$$
$$X_{n,n} = \max\{X_1, X_2, ..., X_n\}$$

1. Show that $X_{n,i}$ takes values in $\{i, i+1, ..., m-n+1\}$.

We will denote the vector of order statistics by

$$\boldsymbol{Y} = \left(X_{n,1}, X_{n,2}, ..., X_{n,n}\right)$$

Note that $\boldsymbol{Y}$ takes values in $L = \{(x_1, x_2, ..., x_n) \in D^n : x_1 < x_2 < \cdots < x_n\}$.

2. Run the order statistic experiment. Note that you can vary the population size m and the sample size n. The order statistics are recorded on each update.

Distributions

3. Show that L has $\binom{m}{n}$ elements and that $\boldsymbol{Y}$ is uniformly distributed on L. *Hint*: $\boldsymbol{Y} = (x_1, x_2, ..., x_n)$ if and only if $\boldsymbol{W} = \{x_1, x_2, ..., x_n\}$ if and only if $\boldsymbol{X}$ is one of the $n!$ permutations of $(x_1, x_2, ..., x_n)$.

4. Use a combinatorial argument to show that the probability density function of $X_{n,i}$ is

$$\mathbb{P}\left(X_{n,i} = k\right) = \frac{\binom{k-1}{i-1}\binom{m-k}{n-i}}{\binom{m}{n}}, \quad k \in \{i, i+1, ..., m-n+i\}$$

5. In the order statistic experiment, vary the parameters and note the shape and location of the probability density function. For selected values of the parameters, run the experiment 1000 times, updating very 10 runs. Note the apparent convergence of the relative frequency function to the probability density function.

Moments

The probability density function in Exercise 4 can be used to obtain an interesting identity involving the binomial coefficients. This identity, in turn, can be used to find the mean and variance of $X_{n,i}$.

6. Show that for $1 \le i \le n \le m$,

$$\sum_{k=i}^{m-n+i} \binom{k-1}{i-1}\binom{m-k}{n-i} = \binom{m}{n}$$

7. Use the identity in the Exercise 6 to show that

$$\mathbb{E}\left(X_{n,i}\right) = i\,\frac{m+1}{n+1}$$

8. Use the identity in Exercise 6 to show that

$$\operatorname{var}\left(X_{n,i}\right) = i\,(n-i+1)\,\frac{(m+1)\,(m-n)}{(n+1)^2\,(n+2)}$$

9. In the order statistic experiment, vary the parameters and note the size and location of the mean/standard deviation bar. For selected values of the parameters, run the experiment 1000 times, updating every 10 runs. Note the apparent convergence of the sample mean and standard deviation to the distribution mean and standard deviation.

Estimators

10. Use the result of Exercise 7 to show that for $i \in \{1, 2, ..., n\}$, the following statistic is an unbiased estimator of m:

$$U_{n,i} = \frac{n+1}{i}\,X_{n,i} - 1$$

Since $U_{n,i}$ is unbiased, its variance is the **mean square error**, a measure of the quality of the estimator.

11. Use the result of Exercise 8 to show that

$$\operatorname{var}(U_{n,i}) = \frac{(m+1)(m-n)(n-i+1)}{i(n+2)}$$

12. Show that for fixed m and n, $\operatorname{var}(U_{n,i})$ decreases as i increases. Thus, the estimators improve as i increases; in particular, $U_{n,n}$ is the best and $U_{n,1}$ the worst.

13. Verify the following ratio, known as the **relative efficiency** of $U_{n,j}$ with respect to $U_{n,i}$:

$$\frac{\operatorname{var}(U_{n,i})}{\operatorname{var}(U_{n,j})} = \frac{j(n-i+1)}{i(n-j+1)}$$

.

Note that the relative efficiency depends only on the orders i and j and the sample size n, but not on the population size m. In particular, the relative efficiency of $U_{n,n}$ with respect to $U_{n,1}$ is n^2.

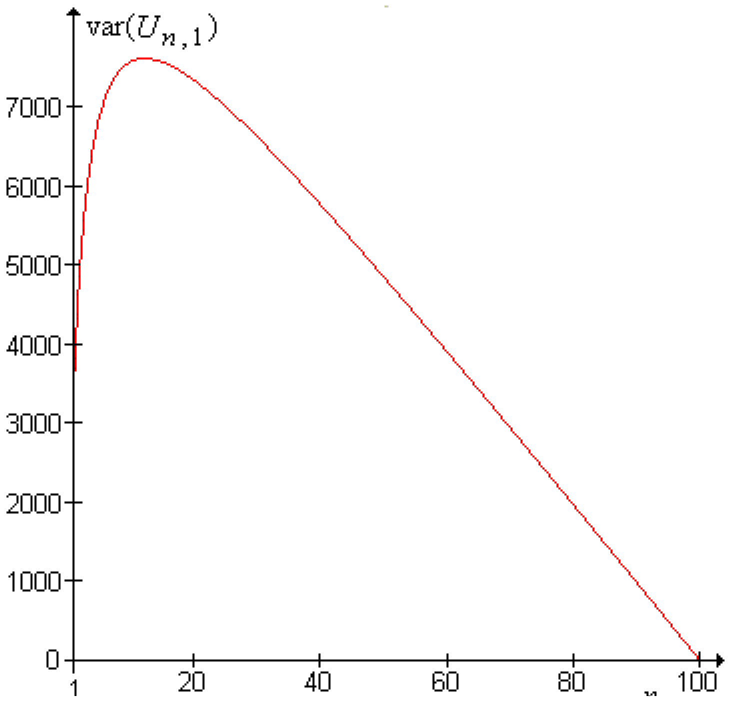
Usually, we hope that an estimator improves (in the sense of mean square error) as the sample size n increases (the more information we have, the better our estimate should be). This general idea is known as **consistency**.

14. Verify the following result. Thus, $\operatorname{var}(U_{n,n})$ decreases to 0 as n increases from 1 to m, and so $U_{n,n}$ is consistent:

$$\operatorname{var}(U_{n,n}) = \frac{(m+1)(m-n)}{n(n+2)}$$

15. Show that for fixed i, $\operatorname{var}(U_{n,i})$ at first increases and then decreases to 0 as n increases from i to m. Thus, $U_{n,i}$ is inconsistent.

The following graph shows $\operatorname{var}(U_{n,1})$ as a function of n for $m = 100$.



Sampling with Replacement

If the sampling is *with* replacement, then the sample $\boldsymbol{X} = (X_1, X_2, ..., X_n)$ is a sequence of independent and identically distributed random variables. The order statistics from such samples are studied in the chapter on Random Samples.

Examples and Applications

16. Suppose that in a lottery, tickets numbered from 1 to 25 are placed in a bowl. Five tickets are chosen at random and without replacement.

a. Find the probability density function of $X_{5,3}$.
b. Find $\mathbb{E}(X_{5,3})$.
c. Find $\text{var}(X_{5,3})$.

The German Tank Problem

The estimator $U_{n,n}$ was used by the Allies during World War II to estimate the number of German tanks m that had been produced. German tanks had serial numbers, and captured German tanks and records formed the sample data. The statistical estimates turned out to be much more accurate than intelligence estimates. Some of the data are given in the table below.

German Tank Data

Date	Statistical Estimate	Intelligence Estimate	German Records
June 1940	169	1000	122
June 1941	244	1550	271
August 1942	327	1550	342

One of the morals, evidently, is not to put serial numbers on your weapons!

17. Suppose that in a certain war, 100 enemy tanks have been captured. The largest serial number of the captured tanks is 1423. Estimate the total number of tanks that have been produced.

18. In the order statistic experiment, and set $m = 100$, $n = 10$. Run the experiment 50 times, updating after each run. For each run, compute the estimate of m based on each order statistic. For each estimator, compute the square root of the average of the squares of the errors over the 50 runs. Based on these empirical error estimates, rank the estimators of m in terms of quality.

19. Suppose that in a certain war, 100 enemy tanks have been captured. The smallest serial number of the captured tanks is 23. Estimate the total number of tanks that have been produced.

?

Virtual Laboratories > 12. Finite Sampling Models > 1 2 3 4 5 6 7 8 9
Contents | Applets | Data Sets | Biographies | External Resources | Keywords | Feedback | ©