

## 6. Order Statistics

### Definitions

Suppose again that we have a basic [random experiment](#), and that  $X$  is a real-valued [random variable](#) for the experiment with [distribution function](#)  $F$  and probability density function  $f$ .

We perform  $n$  independent replications of the basic experiment to generate a random sample  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  of size  $n$  from the distribution of  $X$ . Recall that this is a sequence of [independent](#) random variables, each with the distribution of  $X$ .

Let  $X_{n,k}(\mathbf{X})$  denote the  $k^{\text{th}}$  smallest element of the sample  $\mathbf{X}$ . This statistics is called the **order statistic** of order  $k$ . Often the first step in a statistical study is to order the data; thus order statistics occur naturally. Our goal in this section is to study the distribution of the order statistics in terms of the sampling distribution. Note in particular that the **extreme order statistics** are the minimum and maximum values:

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

**1.** In the [order statistic experiment](#), use the default settings and run the experiment a few times. Note the following:

- a. The table on the left shows the values of the order statistics.
- b. The graph on the left shows the density function of the sampling distribution in blue and the sample values in red.
- c. The graph on the right shows the density function of the selected order statistic in blue and the empirical density function in red. The mean/standard deviation bar of the distribution is shown in blue while the empirical mean/standard deviation bar is shown in red.
- d. The table on the right gives numerical values of the density function and moments and the empirical density function and moments.

### Distributions

#### The Distribution of the $k^{\text{th}}$ Order Statistic

Let  $G_{n,k}$  denote the distribution function of  $X_{n,k}$ . Define

$$N_{n,y} = \#\{i \in \{1, 2, \dots, n\} : X_i \leq y\}, \quad y \in \mathbb{R}$$

**2.** Show that  $N_{n,y}$  has the [binomial distribution](#) with parameters  $n$  and  $F(y)$  for each  $y \in \mathbb{R}$ .

3. Show that  $X_{n,k} \leq y$  if and only if  $N_{n,y} \geq k$  for  $y \in \mathbb{R}$  and  $k \in \{1, 2, \dots, n\}$ .

4. Use the results of Exercises 2 and 3 to show that

$$G_{n,k}(y) = \sum_{j=k}^n \binom{n}{j} F(y)^j (1 - F(y))^{n-j}, \quad y \in \mathbb{R}$$

5. In particular, show that  $G_{n,1}(y) = 1 - (1 - F(y))^n$ ,  $y \in \mathbb{R}$ .

6. In particular, show that  $G_{n,n}(y) = F(y)^n$ ,  $y \in \mathbb{R}$ .

7. Suppose now that  $X$  has a **continuous distribution**. Show that  $X_{n,k}$  has a continuous distribution with probability density function

$$g_{n,k}(y) = \binom{n}{k-1, 1, n-k} F(y)^{k-1} (1 - F(y))^{n-k} f(y), \quad y \in \mathbb{R}$$

*Hint:* Differentiate the expression in [Exercise 4](#) with respect to  $y$ .

8. In the **order statistic experiment**, select the uniform distribution on  $[0, 1]$  and  $n = 5$ . Vary  $k$  from 1 to 5 and note the shape of the density function of  $X_{n,k}$ . For each value of  $k$ , run the simulation 1000 times with and update frequency of 10. Note the apparent convergence of the empirical density function to the true density function.

There is a simple heuristic argument for the result in [Exercise 7](#). First,  $g_{n,k}(y) dy$  is the probability that  $X_{n,k}$  is in an infinitesimal interval of size  $dy$  about  $y$ . On the other hand, this event means that one of sample variables is in the infinitesimal interval,  $k - 1$  sample variables are less than  $y$ , and  $n - k$  sample variables are greater than  $y$ . The number of ways of choosing these variables is the multinomial coefficient

$$\binom{n}{k-1, 1, n-k} = \frac{n!}{(k-1)! 1! (n-k)!}$$

By independence, the probability that the chosen variables are in the specified intervals is

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

9. Consider a random sample of size  $n$  from the **exponential distribution** with rate parameter  $r$ . Compute the probability density function of the  $k^{\text{th}}$  order statistic  $X_{n,k}$ . In particular, note that the minimum of the variables  $X_{n,1}$  has the exponential distribution with rate parameter  $nr$ .



10. In the **order statistic experiment**, select the exponential (1) distribution and  $n = 5$ . Vary  $k$  from 1 to 5 and note the shape of the probability density function of  $X_{n,k}$ . For each value of  $k$ , run the simulation 1000 times with and update frequency of 10. Note the apparent convergence of the empirical density function to the true density function.

11. Consider a random sample of size  $n$  from the **uniform distribution** on the interval  $[0, 1]$ .

- Show that  $X_{n,k}$  has **beta distribution** with parameters  $k$  and  $n - k + 1$ .
- Given the mean and variance of  $X_{n,k}$ .



12. In the **order statistic experiment**, select the uniform distribution on  $[0, 1]$  and  $n = 6$ . Vary  $k$  from 1 to 6 and note the size and location of the mean/standard deviation bar. For each value of  $k$ , run the simulation 1000 times with an update frequency of 10. Note the apparent convergence of the empirical moments to the distribution moments.

13. Four fair dice are rolled. Find the probability density function of each of the order statistics.



14. In the **dice experiment**, select the following order statistic and die distribution. Increase the number of dice from 1 to 20, noting the shape of the probability density function at each stage. Now with  $n = 4$ , run the simulation 1000 times, updating every 10 runs. Note the apparent convergence of the relative frequency function to the density function.

- Maximum score with fair dice.
- Minimum score with fair dice.
- Maximum score with ace-six flat dice.
- Minimum score with ace-six flat dice.

## Joint Distributions

Suppose again that  $X$  has a continuous distribution.

15. Suppose that  $j < k$ . Use an heuristic argument to show that the joint density of  $(X_{n,j}, X_{n,k})$  is

$$g_{n,j,k}(y, z) = \binom{n}{j-1, 1, k-j-1, 1, n-k} F(y)^{j-1} f(y) (F(z) - F(y))^{k-j-1} f(z) (1 - F(z))^{n-k}, \quad y < z$$

Similar arguments can be used to obtain the joint probability density function of any number of the order statistics. Of course, we are particularly interested in the joint probability density function of *all* of the order statistics; the following exercise gives this joint probability density function, which has a remarkably simple form.

16. Show that  $(X_{n,1}, X_{n,2}, \dots, X_{n,n})$  has joint probability density function given by

$$g_n(y_1, y_2, \dots, y_n) = n! f(y_1) f(y_2) \cdots f(y_n), \quad y_1 < y_2 < \cdots < y_n$$

- For each permutation  $\mathbf{i} = (i_1, i_2, \dots, i_n)$  of  $(1, 2, \dots, n)$ , let  $S_{\mathbf{i}} = \{\mathbf{x} \in \mathbb{R}^n : x_{i_1} < x_{i_2} < \cdots < x_{i_n}\}$ .
- On  $S_{\mathbf{i}}$ , the mapping  $(x_1, x_2, \dots, x_n) \mapsto (x_{i_1}, x_{i_2}, \dots, x_{i_n})$  is one-to-one, has continuous first partial derivatives, and has Jacobian 1.

- c. The sets  $S_i$  where  $i$  ranges over the  $n!$  permutations of  $(1, 2, \dots, n)$  are disjoint.
- d. The probability that  $(X_1, X_2, \dots, X_n)$  is not in one of these sets is 0.
- e. Now use the multivariate [change of variables formula](#).

Again, there is a simple heuristic argument for the formula in Exercise 16. For each  $\mathbf{y} \in \mathbb{R}^n$  with  $y_1 < y_2 < \dots < y_n$ , there are  $n!$  permutations of the coordinates of  $\mathbf{y}$ . The probability density of  $(X_1, X_2, \dots, X_n)$  at each of these points is  $f(y_1)f(y_2)\dots f(y_n)$ . Hence the probability density of  $(X_{n,1}, X_{n,2}, \dots, X_{n,n})$  at  $\mathbf{y}$  is  $n!$  times this product.

17. Consider a random sample of size  $n$  from the exponential distribution with rate parameter  $r$ . Compute the joint probability density function of the order statistics  $(X_{n,1}, X_{n,2}, \dots, X_{n,n})$ .



18. Suppose that  $(X_1, X_2, \dots, X_n)$  is a random sample of size  $n$  from the uniform distribution on the interval  $[a, b]$ , where  $a < b$ . Show that

- a.  $(X_1, X_2, \dots, X_n)$  is uniformly distributed on  $[a, b]^n$ .
- b.  $(X_{n,1}, X_{n,2}, \dots, X_{n,n})$  is uniformly distributed on  $\{\mathbf{x} \in [a, b]^n : a < x_1 < x_2 < \dots < x_n < b\}$ .

19. Four fair dice are rolled. Find the joint probability density function of the order statistics.



## Derived Statistics

We will study several important statistics that are based on order statistics.

### Sample Range

The **sample range** is the random variable

$$R = X_{n,n} - X_{n,1}$$

This statistic gives a simple measure of the dispersion of the sample. Note the distribution of the sample range can be obtained from the joint distribution of  $(X_{n,1}, X_{n,n})$  given earlier.

20. Consider a random sample of size  $n$  from the exponential distribution with rate parameter  $r$ . Show that the sample range  $R$  has the same distribution as the maximum of a random sample of size  $n - 1$  from this exponential distribution.

21. Consider a random sample of size  $n$  from the uniform distribution on  $[0, 1]$ .

- a. Show that  $R$  has the beta distribution with left parameter  $n - 1$  and right parameter 2.

- b. Give the mean and variance of  $R$ .  
 c. What happens as  $n \rightarrow \infty$ ?



22. Four fair dice are rolled. Find the probability density function of the sample range.



## The Sample Median

If  $n$  is odd, the **sample median** is the middle of the ordered observations, namely

$$X_{n,k} \text{ where } k = \frac{n+1}{2}$$

If  $n$  is even, there is not a single middle observation, but rather two middle observations. Thus, the **median interval** is

$$[X_{n,k}, X_{n,k+1}] \text{ where } k = \frac{n}{2}$$

In this case, the **sample median** is defined to be the midpoint of the median interval

$$\frac{1}{2} (X_{n,k} + X_{n,k+1}) \text{ where } k = \frac{n}{2}$$

In a sense, this definition is a bit arbitrary because there is no compelling reason to prefer one point in the median interval over another. For more on this issue, see the discussion of [error functions](#) in the section on [Variance](#). In any event, sample median is a natural statistic that is analogous to the [median of the distribution](#). Moreover, the distribution of the sample median can be obtained from our results on order statistics.

## Sample Quantiles

We can generalize the sample median discussed above to other sample quantiles. Suppose that  $p \in (0, 1)$ . Let  $k = \lfloor (n+1)p \rfloor$ , the integer part of  $(n+1)p$ , and let  $q = (n+1)p - k$ , the fractional part of  $(n+1)p$ . Using [linear interpolation](#), we define the **sample quantile** of order  $p$  to be

$$X_{n,k} + q (X_{n,k+1} - X_{n,k}) = (1-q)X_{n,k} + qX_{n,k+1}$$

Once again, the sample quantile of order  $p$  is a natural statistic that is analogous to the distribution quantile of order  $p$ . Moreover, the distribution of a sample quantile can be obtained from our results on order statistics.

The sample quantile of order  $\frac{1}{4}$  is known as the **first sample quantile** and is frequently denoted  $Q_1$ . The sample quantile of order  $\frac{3}{4}$  is known as the **third sample quantile** and is frequently denoted  $Q_3$ . Note that

the sample median is the quartile of order  $\frac{1}{2}$  and is sometimes denoted  $Q_2$ . The **interquartile range** is defined to be

$$\text{IQR} = Q_3 - Q_1$$

The IQR is a statistic that measures the spread of the distribution about the median, but of course this number gives less information than the *interval*  $[Q_1, Q_3]$ .

### Exploratory Data Analysis

The five statistics  $(X_{n,1}, Q_1, Q_2, Q_3, X_{n,n})$  are often referred to as the **five-number summary**. Together, these statistics give a great deal of information about the distribution in terms of the center, spread, and skewness. Graphically, the five numbers are often displayed as a **boxplot**, which consists of a line extending from the minimum  $X_{n,1}$  to the maximum  $X_{n,n}$ , with a rectangular box from the first quartile  $Q_1$  to the third quartile  $Q_3$  and tick marks at the minimum, the median  $Q_2$ , and the maximum.

23. In the **interactive histogram**, select boxplot. Construct a frequency distribution with at least 6 classes and at least 10 values. Compute the statistics in the five-number summary by hand and verify that you get the same results as the applet.

24. In the **interactive histogram**, select boxplot. Set the class width to 0.1 and construct a distribution with at least 30 values of each of the types indicated below. Then increase the class width to each of the other four values. As you perform these operations, note the shape of the boxplot and the relative positions of the statistics in the five-number summary:

- A uniform distribution.
- A symmetric, unimodal distribution.
- A unimodal distribution that is skewed right.
- A unimodal distribution that is skewed left.
- A symmetric bimodal distribution.
- A  $u$ -shaped distribution.

25. In the **interactive histogram**, select boxplot. Start with a distribution and add additional points as follows. Note the effect on the boxplot:

- Add a point below  $X_{n,1}$ .
- Add a point between  $X_{n,1}$  and  $Q_1$ .
- Add a point between  $Q_1$  and  $Q_2$ .
- Add a point between  $Q_2$  and  $Q_3$ .
- Add a point between  $Q_3$  and  $X_{n,n}$ .
- Add a point above  $X_{n,n}$ .

In the last problem, you may have noticed that when you add an additional point to the distribution, one or more of the five statistics does not change. In general, quantiles can be relatively insensitive to changes in the data.

26. Compute the five number summary and sketch the boxplot for the velocity of light variable in **Michelson's data**. Compare the median with the “true value” of the velocity of light.



27. Compute the five number summary and sketch the boxplot for the density of the earth variable in **Cavendish's data**. Compare the median with the “true value” of the density of the earth.



28. Compute the five number summary and sketch the boxplot for the net weight variable in the **M&M data**.



29. Compute the five number summary for the sepal length variable in **Fisher's iris data**, using the cases indicated below. Plot the boxplots on parallel axes, so you can compare.

- All cases
- Type Setosa only
- Type Verginica only
- Type Versicolor only



[Virtual Laboratories](#) > [6. Random Samples](#) > [1](#) [2](#) [3](#) [4](#) [5](#) **[6](#)** [7](#)

[Contents](#) | [Applets](#) | [Data Sets](#) | [Biographies](#) | [External Resources](#) | [Key words](#) | [Feedback](#) | ©