MATH 321: Mathematical Statistics

Lecture 2: Order Statistics

Chapter 6: Point Estimation (6.3)

Order statistics

Introduction

- Sample values such as the smallest, largest, or middle observation from a random sample can provide additional summary information. For example, the median price of houses sold during the previous month might be useful for estimating the cost of living.

Definition

- The **order statistics** of a random sample $X_1, \ldots, X_n$ are the sample values placed in ascending order. They are denoted by $X_{(1)}, \ldots, X_{(n)}$.

  The order statistics are random variables that satisfy $X_{(1)} \leq \cdots \leq X_{(n)}$. In particular

$$\begin{aligned} X_{(1)} &= \min_{1 \leq i \leq n} X_i, \\ X_{(2)} &= \text{second smallest } X_i \\ &\vdots \\ X_{(n)} &= \max_{1 \leq i \leq n} X_i. \end{aligned}$$

- The formulas for the pdfs of the order statistics for a random sample from a continuous population will be the main topic in this section.
- Notes:
  - The distribution of $X_{(j)}$ is not the same as the distribution of X_j
  - The range / support of is always the same as the random variable you are sampling from.

Bivariate case, min and max of two random variables

- Before we generalize order statistics to n random variables, we will look at the bivariate case.

  This means we are studying the min and max of two random variables.

- Derivation of the distributions of the functions $min(X_1, X_2)$ and $max(X_1, X_2)$.

  Setup: Let $X_i \overset{iid}{\sim} f(x)$ for $i = 1, 2$. We also have $F_X(x)$ and $S_X(x) = 1 - F_X(x)$.

  - Minimum: Using the above notation with $n = 2$, let $X_{(1)} = min(X_1, X_2)$.
  - We need to set up a probability statement that will make finding the distribution of $min(X_1, X_2)$ easier.

  - Now we can find the distribution.

  - Maximum: Let $X_{(2)} = max(X_1, X_2)$.

- Examples:

  1. Let $X_i \overset{\perp\!\!\!\perp}{\sim} \text{Exp}(\lambda_i)$. Find the distribution of $min(X_1, X_2)$.

     - Adding context: Suppose X_1 and X_2 are independent waiting times for accidents in two towns where $X_i \overset{iid}{\sim} \text{Exp}(\lambda = 1)$. Then $Min = min(X_1, X_2) \sim$
     - This can be interpreted in a natural way. In each of two separate towns we are waiting for the first accident in a process where the average number of accidents is 1 per month. When we study the accidents of both towns we are waiting for the first accident in the process where the average number of accidents is a total of 2 per month.

  2. Let $X_i \overset{iid}{\sim} \text{Uniform}(0, 10)$ for $i = 1, 2$.
     (a) Find the distribution of $min(X_1, X_2)$ and $max(X_1, X_2)$.

(b) Find the expected value of $min(X_1, X_2)$ and $max(X_1, X_2)$.

Generalizing order statistics

- Example: Let X be a random variable with pdf $f(x) = 2x, \quad 0 < x < 1$ and let $X_1, \ldots, X_5$ be a random sample from X.

  (a) Find the pdf of $X_{(1)}$, the first order statistic.

  Note: One strategy is to find cdf first, and then take derivative to find pdf.

  $f_{X_{(j)}}(x) = F'_{X_{(j)}}(x)$.

  Also we are going to frame everything using cdfs rather than survival functions like with the bivariate case.

(b) Find the pdf of $X_{(4)}$, the fourth order statistic.

(c) Find $P(X < 1/2)$, $P(X_{(1)} < 1/2)$, and $P(X_{(4)} < 1/2)$.

Order statistics distribution theorem

- Theorem: Let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics of a random sample, $X_1, \ldots, X_n$, from a continuous population with cdf $F_X(x)$ and pdf $f_X(x)$. Then the **cdf of $X_{(j)}$** is

$$F_{X_{(j)}}(x) = \sum_{k=j}^{n} \binom{n}{k} [F_X(x)]^k [1 - F_X(x)]^{n-k}$$

and the **pdf of $X_{(j)}$** is

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} [F_X(x)]^{j-1} f_X(x) [1 - F_X(x)]^{n-j}$$

- Walk through for proof of theorem:
  - Obtaining the pdf for the jth order statistic is the main goal. To do this, we first find the cdf for $X_{(j)}$ and then differentiate it to get the pdf.

    The equation for the cdf of the jth order statistic is closely related to the cdf of the binomial distribution.

  - $X_{(j)}$ represents the jth smallest value. So the cdf is

    $$F_{X_{(j)}}(x) = P(X_{(j)} \leq x)$$

    * Interpretation: This is the probability that at least j of X_is are less than or equal to x.

      So, we are essentially just counting something, specifically the number of random variables in our random sample less than x.

    * Example: Let $X_1, \ldots, X_5$ be a random sample from $f(x)$. We are interested finding in $P(X_{(3)} \leq 4)$.

      Note: If $X_{(j)} \leq x$, then $X_{(j-1)} \leq x$ must be true. But $X_{(j+1)} \leq x$ can also be true.

Data	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$
(a)	2	3	4	5	6
(b)	2	3	4	4	5
(c)	2	3	4	4	4
(d)	2	4	4	5	6
(e)	2	5	4	5	6

  - Thus, we can use the binomial distribution to find the cdf of $X_{(j)}$.

    We can define the event of success as $\{X_j \leq x\}$, because we are counting how many of the original sample $X_1, \ldots, X_n$ are less than x.

  - Let Y be a random variable that counts the number of $X_1, \ldots, X_n$ less than or equal to x.

    Then, we see that $Y \sim$

- Then $f_{X_{(j)}}(x) = \frac{d}{dx} F_{X_{(j)}}(x)$.

  This derivation is not straightforward, but it can be intuitively understood.

- Concept: The pdf assigns our n random variables to three groups of sizes:

  Recall a **partition** of n objects into k groups of sizes $n_1, \ldots, n_k$ equals $\frac{n!}{n_1! \cdot \ldots \cdot n_k!}$.

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!\,1!\,(n-j)!} \left[P(X \leq x)\right]^{j-1} f_X(x) \left[P(X > x)\right]^{n-j}$$

- It is worth noting the special cases for the extreme order statistics (then show for bivariate case):

  Smallest: $f_{X_{(1)}}(x) = nf(x)[1 - F(x)]^{n-1} \rightarrow$

  Largest: $f_{X_{(n)}}(x) = n[F(x)]^{n-1} f(x) \rightarrow$

Specific order statistics and functions of order statistics

- Several very common statistics are actually order statistics.

  The importance of order statistics has increased because of more frequent use of non-parametric inferences and robust procedures.

- Sample median:

  - The sample median, which we will denote by M, is a number such that approximately one-half of the observations less than M and one-half are greater.
  - In terms of the order statistics, M is defined by

$$M = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \left[X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right]/2 & \text{if } n \text{ is even} \end{cases}$$

    So it is the sole middle observation or the average of the two middle observations.

  - The median is a measure of location that might be considered as alternative to the sample mean.
  - Sample mean vs sample median.
    * Sample mean: Can be **more efficient** (i.e. more accurate in some sense because it's using all of the data ($\sum X_i$)).

      But **less robust** because it can be affected by outliers when using all of the data.
    * Sample median: **Less efficient** because it only uses the first half of the data (e.g. if $\boldsymbol{x} = \{1, 2, 3, 4, 5\}$, it is only using 1, 2 and 3 to find the median (starts from left and stops when it gets to the median)).

      But **more robust** because it is only using half the data.

- Sample range, $R = X_{(n)} - X_{(1)} = max(X_1, \ldots, X_n) - min(X_1, \ldots, X_n)$.

  This is a measure of spread which gives the distance spanned by the entire sample.

- $IQR = Q_3 - Q_1$.

  In terms of order statistics, given an even $2m$ or odd $2m + 1$ random variables:

  $$\begin{array}{rcl} Q_1 & = & \text{median of the smallest } m \text{ values} \\ Q_3 & = & \text{median of the largest } m \text{ values} \end{array}$$

  This is a measure of spread that might be considered as alternative to the standard deviation. It is better for skewed data or when there is outliers.

- Midrange $= \dfrac{X_{(1)} + X_{(n)}}{2}$.

  This is a measure of location like the sample mean or median. It is found by averaging (or taking the midpoint) of the min and max of the random sample.

- To find the distributions of functions of order statistics, e.g. involving more than than one statistic such as the sample range R or midrange, we have two options:

  a) Find the pdf of multiple ordered statistics (i.e. multivariate transformation).

  b) OR we can use simulation! (like we did in our sampling distribution R notes)

  - First we could simulate the sampling distribution of the statistic of interest.

    Then use those results to approximate any quantity we need!
  - For example, suppose we have 10,000 values for $\hat{R} = max(x_1, \ldots, x_n) - min(x_1, \ldots, x_n)$.

    $E(R) \approx$

    If we want to estimate $P(R > x)$. Let I be an indicator variable such that

    $$I = \begin{cases} 1 & = & \text{if } R > x \\ 0 & = & \text{if } R \leq x \end{cases}$$

    $P(R > x) \approx$

- Simulation is a very powerful tool that allows researchers to study things that don't have theoretical solutions.

- Examples:

  1. Continuing previous example:

     $X_1, \ldots, X_5 \overset{iid}{\sim} f(x) = 2x$ and $F(x) = x^2 \quad 0 < x < 1.$

     (a) Find the cdf of the sample median $X_{(3)}$.

     (b) Find the pdf of the sample median $X_{(3)}$.

2. Wind damage to insured homes are independent random variables with common pdf and cdf

$$f(x) = \frac{3}{x^4} \quad x > 1 \qquad \longrightarrow \qquad F(x) = 1 - \frac{1}{x^3} \quad x > 1$$

where x is in thousands of dollars. Find the expected value of the largest of three such claims.

Order statistics as estimators of population percentiles

- Expected value of the "position" of order statistics.

  Theorem: Let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics of a random sample of size n from a continuous population with cdf $F_X(x)$. Then

$$E[F_X(X_{(j)})] = \frac{j}{n+1}, \quad j = 1, \ldots, n$$

- Breaking down theorem:
  - For our population distribution, $F_X(x) = P(X \leq x) = p$ represents the cumulative probability up to and including x, or equivalently the area under $f(x)$ less than x.

    Recall that probability is a function, so here we are inputting a constant, particular x value and getting the corresponding probability p as a result (which is also a constant).

- Now if we input the jth order statistic $X_{(j)}$ (which is a random variable) into $F_X(x)$, the output is a random area ($\approx$ random variable p), which represents the probability X is less than or equal to $X_{(j)}$:

$$F_X(X_{(j)}) = P(X \leq X_{(j)})$$

- Because it is a random variable, we can find the expected value.

$$E[F_X(X_{(j)})] = \frac{j}{n+1}$$

Example: Let $n = 9$ and $j = 6$. Find $E[F_X(X_{(6)})]$.

- Using this theorem:
  - Recall for $0 \leq p \leq 1$ the **$100p^{th}$ percentile of X** is the number x_p defined by

$$P(X \leq x_p) = F(x_p) = p$$

  - Thus, we can use $X_{(j)}$ as an estimator of x_p, where $p = j/(n+1)$.

    Note that p is a function of j and $n \Longrightarrow$ We are figuring out which percentile, x_p, $X_{(j)}$ estimates.

$$F(x_p) = p \quad \longrightarrow \quad F(x_{j/(n+1)}) = \frac{j}{n+1}$$

q–q plots

- Extension of previous theorem:
  - Now let's consider the previous order statistic $X_{(j-1)}$ as well, which is of course another random variable.

    $F_X(X_{(j)}) - F_X(X_{(j-1)})$ represents the probability (area under curve) between two adjacent order statistics $X_{(j)}$ and $X_{(j-1)}$. The expected value of this random area is

$$E[F_X(X_{(j)}) - F_X(X_{(j-1)})] =$$

  - We could also show the area below the first order stat and the area above the last order stat:

$$E[F_X(X_{(1)})] = \frac{1}{n+1} \qquad \text{and} \qquad E[1 - F_X(X_{(n)})] = \frac{1}{n+1}$$

– This means the order statistics $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ partition the range of X into $n+1$ parts and thus create $n+1$ areas under $f(x)$ and above the x-axis.

On average, each of the $n+1$ areas equals $1/(n+1)$.

- So, we can use the relationships shown above to test whether a random variable X has a certain distribution by "matching up" the sample order statistics with the theoretical percentiles. This is the process to get the numbers used in a q–q plot.

  (1) Compute the percentiles $x_{\frac{1}{n+1}}, \ldots, x_{\frac{n}{n+1}}$ of the population distribution we are testing.

  (2) Compare (1) to the observed sample order statistics $x_{(1)}, \ldots, x_{(n)}$.

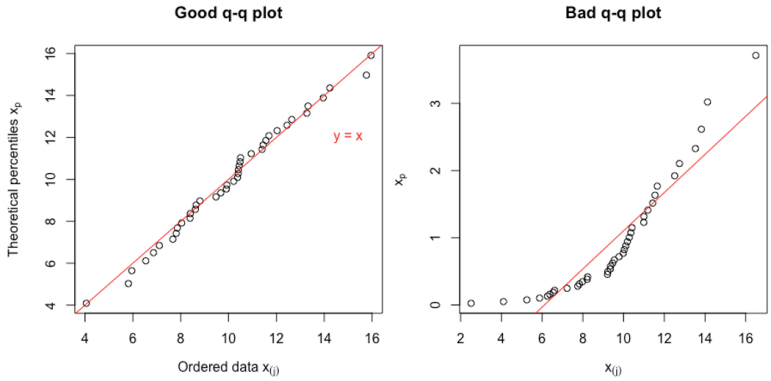
  If the theoretical distribution is a good model for the observations, then we should see

$$x_{(1)} \approx x_{\frac{1}{n+1}} \quad , \quad \cdots \quad , \quad x_{(n)} \approx x_{\frac{n}{n+1}}$$

- Definition: Let X be a random variable, $x_{(1)}, \ldots, x_{(n)}$ be the observed sample order statistics of a random sample of size n, and $x_{\frac{1}{n+1}}, \ldots, x_{\frac{n}{n+1}}$ be the percentiles from some particular distribution. A plot of the points

$$(x_{(1)}, x_{\frac{1}{n+1}}) \ , \ \cdots \ , \ (x_{(n)}, x_{\frac{n}{n+1}})$$

is known as a **quantile–quantile plot**, or more simply a **q–q plot**.



- Interpretation of a q–q plot.
  - If we picked a good model (i.e. X has the particular distribution), then $x_{(j)} \approx x_{\frac{j}{n+1}}$ and the q–q plot should be nearly a straight line through the origin with slope = 1 (i.e. diagonal line).
  - Conversely, a strong deviation from this line is evidence that the distribution did not produce the data.
  - Sidenote: It's called a quantile–quantile plot because the sample order statistics $x_{(1)}, \ldots, x_{(n)}$ associated with the sample $x_1, \ldots, x_n$ are called the **sample quantiles of order $j/(n+1)$** and the percentile x_p of a theoretical distribution is the **quantile of order p**, and we are using $p = j/(n+1)$ to match them up.

- Using q–q plots.
  - Usually we are not trying to see if the data come from a particular distribution, but rather from a parametric family of distributions (such as the normal, uniform, or exponential, etc.).

    We are usually forced into this situation because we don't know the parameters. So typically, the next step, after the q–q plot, may be to estimate the parameters, which we will learn how to do later.

- q–q plots for the normal distribution.
  - q–q plots are often used to test whether a random sample is from a normal distribution.
  - When creating the plot, we of course need to calculate the theoretical percentiles. This requires specifying μ and σ^2 when using invNorm() or qnorm(); but as mentioned these are usually unknown.

    So we have two strategies:

    1. We could use the sample statistics as best guess of the population parameters ($\bar{X} \to \mu$ and $S^2 \to \sigma^2$), as we know these are unbiased estimators.

       If we do this, the q–q plot should follow the diagonal line.

    2. If we don't want to make this assumption. We can make use of the relationship to the standard normal distribution:

Thus if we vary p and plot (x_p, z_p), we get a straight line with slope $1/\sigma$.

This means we can still test if a random sample came from a normal distribution without having to know / guess the mean and standard deviation.

So if we plot $(x_{(1)}, z_{\frac{1}{n+1}})$, ... , $(x_{(n)}, z_{\frac{n}{n+1}})$, which has our ordered sample data on the x-axis now as estimates of the population percentiles, we should see an approximately straight line. If so, then $\frac{1}{\text{slope}}$ is an approximation of σ.