STAT 3610: Supplemental Notes

Finding the Median and Other Quantiles

Keywords: *median, quartiles, quantiles, order statistics, lower forth, upper forth, forth spread, interquartile range*

One way to organize and study a collection of observations from a sample of size n, is to order the values from smallest to largest. The ordered data values are called the **order statistics**. We use these order statistics to explore the center of the data, as well as, the spread. For example, a sample average you often see cited, in addition to the sample mean, is the sample median, $\tilde{x}$, which is in the middle of the sorted data, i.e., if n is large, approximately 50% of the sample values are below the median and approximately 50% are above the median. The **quartiles**, Q_1, Q_2 and Q_3, break the sample distribution into quarters (25%, 50% and 75%). For large samples, approximately 25% of the sample values are below the first quartile, Q_1, the second quartile is the median, and approximately 75% of the sample values are below the third quartile, Q_3. The median and the quartiles are examples of **quantiles**. As you may note, quantiles are statistics that depend on ranked data values.

Order Statistics: Let $x_1, \ldots, x_n$ be an observed sample of size n from some population. Sort the data from smallest to largest. The **order statistics** are the values $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ such that, $x_{(1)} = \min\{x_1, \ldots, x_n\}$, $x_{(2)}$ is the second smallest observed value, $x_{(3)}$ is the third smallest, ..., $x_{(n)} = \max\{x_1, \ldots, x_n\}$

Median: Suppose $x_1, \ldots, x_n$ is an observed sample of size n from some population. Let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ represent the ordered values from the sample (order statistics). To calculate the sample median, because a sample is a discrete set of points, we need to consider whether n is odd or even.

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{if n is odd} \\ \\ \frac{x_{(n/2)} + x_{(n/2+1)}}{2} & \text{if n is even} \end{cases}$$

Example 1: Suppose we took a sample of size $n = 10$ and got $x_1 = 5.8, x_2 = 8.1, x_3 = 6.5, x_4 = 7.8, x_5 = 3.1, x_6 = 10.1, x_7 = 2.3, x_8 = 11.0, x_9 = 12.1, x_{10} = 4.5$. Easy to see that
$x_{(1)} = \min\{5.8, 8.1, 6.3, 7.4, 3.1, 10.1, 2.3, 11.0, 12.1, 4.5\} = 2.3$
and $x_{(10)} = \max\{5.8, 8.1, 6.3, 7.4, 3.1, 10.1, 2.3, 11.0, 12.1, 4.5\} = 12.1$

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
Original sample	5.8	8.1	6.3	7.4	3.1	10.1	2.3	11.0	12.1	4.5

Ordered values	2.3	3.1	4.5	5.8	6.3	7.4	8.1	10.1	11.0	12.1
Rank	1	2	3	4	5	6	7	8	9	10
Order statistics	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$

Since $n = 10$ is even, we know that the sample median is the average/midpoint of the

$10/2 = 5th$ and $10/2 + 1 = 6th$ ordered values or

$$\tilde{x} = \frac{x_{(5)} + x_{(6)}}{2} = \frac{6.3 + 7.4}{2} = 6.85$$

the first quartile is located at the third smallest value, since there are 5 data values in the lower half of the sample. So, $Q_1 = x_{(3)} = 4.5$ is the first sample quartile and $Q_3 = x_{(8)} = 10.1$.

In the above example, I used an approach similar to the approach given on page 39 of the textbook. There, they refer to the first quartile (the 25th percentile) as the **lower forth** of the data and the third quartile (the 75th percentile) as the **upper forth**. These are used in creating a **boxplot** and are sufficient way to compute the quartiles. Later we will discuss measures of spread. One such measure, is the **interquartile range** or what the book refers to as the **box width** of the box plot, denoted as f_s = upper forth - lower forth $\approx Q_3$ - Q_1. Sometimes, the computed quartiles from statistical packages, such as, MINITAB, SAS and R, may slightly differ from each other and your hand calculations. These slight differences reflect the different computation methods .

Example 2: Suppose the ordered values (order statistics) of a sample of size $n = 9$ are

Ordered Data:	9.69	13.16	17.09	18.12	23.70	24.07	24.29	26.43	30.75
Rank:	1	2	3	4	5	6	7	8	9

Since the sample size $n = 9$ is odd, the median is exactly the $(n+1)/2 = 10/2 = 5$th ordered value, so the median is $\tilde{x} = x_5 = 23.70$. The quartiles will be the medians of the lower and upper halves of the data (including the median in each). So the median of the first five values, 9.69, 13.16, 17.09, 18.12, 23.70 is $Q_1 = 17.09$. The median of the upper half (top 5 values), 23.70, 24.07, 24.29, 26.43, and 30.75 is $Q_3 = 24.29$.