Sample Quartiles

We have seen that the **sample median** of a data set \( \{x_1, x_2, \ldots, x_n\} \), *sorted* in increasing order, is a value that divides it in such a way, that exactly half (i.e., 50%) of the sample observations fall below the median, and exactly half (50%) are above it.

- If the sample size \( n \) is odd, then precisely **one** of the data values will lie at the exact center; this value is located at position \((n + 1)/2\) in the data set, and corresponds to the median.

- If the sample size \( n \) is even however, then the exact middle of the data set will fall *between* two values, located at positions \(n/2\) and \(n/2 + 1\). In this case, it is customary to define the median as the average of the two values, which lies midway between them.

**Sample quartiles** are defined similarly: they divide the data set into quarters (i.e., 25%). The first quartile, designated, \( Q_1 \), marks the cutoff below which lies the lowest 25% of the sample. Likewise, the second quartile \( Q_2 \) marks the cutoff between the second lowest 25% and second highest 25% of the sample; *note that this coincides with the sample median!* Finally, the third quartile \( Q_3 \) marks the cutoff above which lies the highest 25% of the sample.

This procedure of ranking data is not limited to quartiles. For example, if we wanted to divide a sample into ten intervals of 10% each, the cutoff points would be known as **sample deciles**. In general, the cutoff values that divide a data set into any given proportion(s) are known as **sample quantiles** or **sample percentiles**. For example, receiving an exam score in the “90th percentile” means that 90% of the scores are below it, and 10% are above.

For technical reasons, the strict definitions of quartiles and other percentiles follow rigorous mathematical formulas; however, these formulas can differ slightly from one reference to another. As a consequence, different statistical computing packages frequently output slightly different values from one another. On the other hand, these differences are usually very minor, especially for very large data sets.

**Exercise 1 (not required):** Using the R code given below, generate and view a random sample of \( n = 40 \) positive values, and find the quartiles via the so-called “**five number summary**” that is output.

```r
# Create and view a sorted sample, rounded to 3 decimal places.
x = round(sort(rchisq(40, 1)), 3)
print(x)
y = rep(0, 40)

# Plot it along the real number line.
plot.new()
plot(x, y, pch = 19, cex = .5, xlim = range(0, 1 + max(x)), ylim = range(0, 0.01), ylab = "", axes = F)
axis(1)

# Identify the quartiles.
summary(x)

# Plot the median Q_2 (with a filled red circle).
Q2 = summary(x)[3]
points(Q2, 0, col = "red", pch = 19)

# Plot the first quartile Q_1 (with a filled blue circle).
Q1 = summary(x)[2]
points(Q1, 0, col = "blue", pch = 19)

# Plot the third quartile Q_3 (with a filled green circle).
Q3 = summary(x)[5]
points(Q3, 0, col = "green", pch = 19)
```
Exercise 2 (not required): Using the same sample, sketch and interpret

\texttt{boxplot(x, pch = 19)}

identifying all relevant features. From the results of these two exercises, what can you conclude about the general “shape” of this distribution?

\textbf{NOTE:} Finding the approximate quartiles (or other percentiles) of \texttt{grouped data} can be a little more challenging. Refer to the Lecture Notes, pages 2.3-4 to 2.3-6, \textit{and especially} 2.3-11.

\textit{CONTINUED...}
To illustrate the idea of estimating quartiles from the **density histogram** of grouped data, let us consider a previous, posted exam problem (Fall 2013).

First, we find the **median** $Q_2$, i.e., the value on the $X$-axis that divides the total area of 1 into .50 area on either side of it. By inspection, the cumulative area below the left endpoint 4 is equal to $.10 + .20 = .30$, too small. Likewise, the cumulative area below the right endpoint 12 is $.10 + .20 + .30 = .60$, too big. Therefore, in order to have .50 area both below and above it, $Q_2$ must lie in the third interval (4, 12), in such a way that its corresponding rectangle of .30 area is split into left and right sub-areas of .20 + .10, respectively.
Now just focus on this particular rectangle…

\[
A = 0.20 \quad B = 0.10
\]

\[
a = 4 \quad Q_2 \quad b = 12
\]

… and use any of the three boxed formulas on page 2.3-5 of the Lecture Notes, with the quantities labeled above. For example, the third formula (which I think is easiest) yields

\[
Q_2 = \frac{Ab + Ba}{A + B} = \frac{(0.2)(12) + (0.1)(4)}{0.2 + 0.1} = \frac{2.8}{0.3} = 9.333.
\]

- The other quartiles are computed similarly. For example, the first quartile \(Q_1\) is the cutoff for the lowest 25% of the area. By the same logic, this value must lie in the second interval [2, 4), and split its corresponding rectangle of 0.20 area into left and right sub-areas of 0.15 + 0.05, respectively:

\[
A = 0.15 \quad B = 0.05
\]

\[
\text{sum} = 0.25
\]

Therefore,

\[
Q_1 = \frac{Ab + Ba}{A + B} = \frac{(0.15)(4) + (0.05)(2)}{0.15 + 0.05} = \frac{0.7}{0.2} = 3.5.
\]
Likewise, the third quartile $Q_3$ is the cutoff for the highest 25% of the area. By the same logic as before, this value must lie in the fourth interval $[12, 22)$, and split its corresponding rectangle of .25 area into left and right sub-areas of .15 + .10, respectively:

$$Q_3 = \frac{Ab + Ba}{A + B} = \frac{(.15)(22) + (.10)(12)}{.15 + .10} = \frac{4.5}{.25} = 18.$$

Estimating a sample proportion between two known quantile values is done pretty much the same way, except in reverse, using the formulas on the bottom of the same page, 2.3-5. For example, the same problem asks to estimate the sample proportion in the interval $[9, 30)$. This interval consists of the disjoint union of the subintervals $[9, 12)$, $[12, 22)$, and $[22, 30)$.

- The first subinterval $[9, 12)$ splits the corresponding rectangle of area .30 over the class interval $[4, 12)$ into unknown left and right subareas $A$ and $B$, respectively, as shown below. Since it is the right subarea $B$ we want, we use the formula $B = (b - Q) \times \text{Density} = (12 - 9) \times .0375 = .1125$.
- The next subinterval $[12, 22)$ contains the entire corresponding rectangular area of .25.
- The last subinterval $[22, 30)$ splits the corresponding rectangle of area .15 over the class interval $[22, 34)$ into unknown left and right subareas $A$ and $B$, respectively, as shown below. In this case, it is the left subarea $A$ that we want, so we use $A = (Q - a) \times \text{Density} = (30 - 22) \times .0125 = .10$.

Adding these three areas together yields our answer, $$.1125 + .25 + .10 = .4625.$$

Page 2.3-6 gives a way to calculate quartiles, etc., from the cumulative distribution function (cdf) table, without using the density histogram.