

Extra Material for Chapter 12: Percentiles and Quantiles

The material of this section is NOT in the text and will be covered only very briefly in lecture. (It is not included in the lecture notes for the course.)

This material is needed for some homework and practice exercises and it might be on the final exam. You may ask about this material at discussion, at the review for the final or during office hours.

The mean and standard deviation are arithmetic summaries of data; i.e. we get them by performing arithmetic operations: adding, dividing, and so on. Percentiles and quantiles are not arithmetic summaries; rather, they focus on positions in the sorted list of data.

We learned about the median in lecture. The median is also the 50th percentile and the 0.50 quantile. Quantiles are the decimal version of a percentile. For example, the 63rd percentile is the same as the 0.63 quantile.

To make this abstract, but I hope not confusing, let π be a number strictly between 0 and 1; i.e. $0 < \pi < 1$. The π quantile is the same number as the 100π percentile. My example of the previous paragraph illustrates this notion with $\pi = 0.63$; the 0.63 quantile is the same number as the $100(0.63) = 63$ rd percentile.

By convention, there are 99 percentiles—corresponding to the integers 1 thru 99—allowing π to be any of the values: 0.01, 0.02, 0.03, ... 0.99.

As stated earlier, the median is the 0.50 quantile and the 50th percentile. We will explore this idea further now. Consider a sample of three distinct numbers sorted, for example: 4, 9 and 20. The median is ‘the number in the center position,’ which is 9. For a sample of four distinct numbers sorted, for example: 4, 7, 9 and 20. There are now two center positions, which contain the numbers 7 and 9. By convention, the median is taken to be the arithmetic average of these numbers, 8, which has the property that half the data are smaller than it and half are larger than it. So what is the median? Well, there is no one simple answer. But there is one not-so-simple answer, which we have not seen yet.

It turns out that for math purposes, the idea ‘half the data larger, half the data smaller’ is closer to what we want than the idea ‘in the center position.’ Thus, we define the median as follows.

The median is any number with the following two properties.

- *At least* one-half of the data are *less than or equal to* the median.
- *At least* one-half of the data are *greater than or equal to* the median.

For any odd sample size, there is a unique number that satisfies this definition, namely the number in the center position. For our example data of 4, 9 and 20, the number in the center position, 9, is the unique number that satisfies this definition, as argued below.

- Two-thirds of the data are less than or equal to 9.
- Two-thirds of the data are greater than or equal to 9.

Thus, 9 satisfies our definition. But is it unique? Yes. (For any candidate smaller than 9, the first item fails; for any candidate larger than 9, the second item fails.)

Next, consider an even sample size. This is trickier, so I will use our earlier data of 4, 7, 9 and 20. You can check that our median 8 satisfies both conditions. But it is not unique. Actually, any number between 7 and 9 inclusive will satisfy our definition of the median. Also, any number smaller than 7 or larger than 9 will fail our definition. Thus, in a strictly literal math sense, for an even sample size there can be an entire interval of numbers that satisfy the definition of median. Most statisticians and mathematicians agree that if there is an interval of medians, we call the midpoint of the interval *the median*. And we will always do this in this course.

You might well wonder what was the point of all of the above. Before you read this we all agreed on what the median was, and now I have just made it more complicated. Well, we need the above for percentiles and quantiles. I will illustrate.

First, let us be specific. Suppose we want to find the 35th percentile, which we will denote by $P_{35} = Q_{0.35}$, the 0.35 quantile. We define the 35th percentile to be any number that has the following two properties.

- *At least 35% of the data are less than or equal to it.*
- *At least 65% of the data are greater than or equal to it.*

If an entire interval of numbers satisfy both properties, we take the 35th percentile to be the midpoint of the interval.

Next, I give you the algorithm for calculating the 35th percentile.

1. Calculate $k = 0.35n$.
2. If k is an integer, then the 35th percentile equals

$$[x_{(k)} + x_{(k+1)}]/2.$$

3. If k is not an integer, round it up to the next integer and call it k' . The 35th percentile equals $x_{(k')}$.

For example, suppose $n = 100$. Then,

$$0.35n = 0.35(100) = 35$$

is an integer and the 35th percentile equals

$$[x_{(35)} + x_{(36)}]/2.$$

Let's check that this works.

First, suppose that $x_{(35)}$ and $x_{(36)}$ are different numbers. Then exactly 35% of the data are less than the percentile and exactly 65% of the data are greater than the percentile. If $x_{(35)}$ and $x_{(36)}$ are the same number then at least 35% of the data are less than or equal to the percentile and at least 65% of the data are greater than or equal to the percentile.

As another example, suppose that $n = 150$. Then

$$k = 0.35n = 0.35(150) = 52.5$$

is not an integer, so we round it up to $k' = 53$ and the 35th percentile equals $x_{(53)}$. Let's check that it works.

First, clearly at least 53 observations are less than or equal to $x_{(53)}$, and $53/150 = 0.353$ is at least 35%. Second, at least $150 - 52 = 98$ observations are greater than or equal to $x_{(53)}$, and $98/150 = 0.653$ is at least 65%.

Now that we understand the 35th percentile, we will consider any arbitrary percentile. First, the definition. We define the 100π percentile to be any number that has the following two properties.

- At least $100\pi\%$ of the data are *less than or equal to* it.
- At least $100(1 - \pi)\%$ of the data are *greater than or equal to* it.

If an entire interval of numbers satisfies both properties, we take the 100π percentile to be the midpoint of the interval.

Next, I give you the algorithm for calculating the 100π percentile.

1. Calculate $k = \pi n$.
2. If k is an integer, then the 100π percentile equals

$$[x_{(k)} + x_{(k+1)}]/2.$$

3. If k is not an integer, round it up to the next integer and call it k' . The 100π percentile equals $x_{(k')}$.

Final comment. The 25th percentile is called the first quartile; the 50th percentile—in addition to being called the median—is called the second quartile; and the 75th percentile is called the third quartile. Note the word is quartile, not quantile. To add to the confusion, the quartiles are denoted Q_1 , Q_2 and Q_3 . Thus, a Q with a subscript can be a quantile or a quartile. If the subscript is 1, 2 or 3, it's a quartile; if the subscript is smaller than 1, it is a quantile.

The text presents a method different than the one in this section for computing quartiles. Actually, the two methods give the same answers *unless* the sample size n is equal to $4m + 1$, where m is a positive integer. In other words, the two methods give the same answer unless

$$n = 5, 9, 13, \dots$$

The final will *not* ask you to calculate quartiles for any $n = 4m + 1$, for m an integer. To summarize: you don't need to learn the book's method for calculating quartiles, but if you do, it won't effect your performance on the final.