AUA CS108, Statistics, Fall 2020

Lecture 12

Michael Poghosyan

23 Sep 2020

Contents

Sample Quantiles

Now we want to generalize the idea of the Median and Quartiles.

Sample Quantiles

Now we want to generalize the idea of the Median and Quartiles. Recall that:

Sample Quantiles

Now we want to generalize the idea of the Median and Quartiles. Recall that:

- 50% of Datapoints are to the left of the Median, and 50% are to the right, so Median divides the (sorted) Dataset in the (approximate) proportion 50% - 50%

Sample Quantiles

Now we want to generalize the idea of the Median and Quartiles. Recall that:

- 50% of Datapoints are to the left of the Median, and 50% are to the right, so Median divides the (sorted) Dataset in the (approximate) proportion 50% - 50%
- 25% of Datapoints are to the left of the Lower Quartile Q_1, and 75% are to the right, so Q_1 divides the (sorted) Dataset in the (approximate) proportion 25%-75%

Sample Quantiles

Now we want to generalize the idea of the Median and Quartiles. Recall that:

- 50% of Datapoints are to the left of the Median, and 50% are to the right, so Median divides the (sorted) Dataset in the (approximate) proportion 50% - 50%
- 25% of Datapoints are to the left of the Lower Quartile Q_1, and 75% are to the right, so Q_1 divides the (sorted) Dataset in the (approximate) proportion 25%-75%
- 75% of Datapoints are to the left of the Upper Quartile Q_3, and 25% are to the right, so Q_3 divides the (sorted) Dataset in the (approximate) proportion 75%-25%

Sample Quantiles

Now we want to generalize the idea of the Median and Quartiles. Recall that:

- 50% of Datapoints are to the left of the Median, and 50% are to the right, so Median divides the (sorted) Dataset in the (approximate) proportion 50% - 50%
- 25% of Datapoints are to the left of the Lower Quartile Q_1, and 75% are to the right, so Q_1 divides the (sorted) Dataset in the (approximate) proportion 25%-75%
- 75% of Datapoints are to the left of the Upper Quartile Q_3, and 25% are to the right, so Q_3 divides the (sorted) Dataset in the (approximate) proportion 75%-25%

Now, let $\alpha \in (0, 1)$.

Sample Quantiles

Now we want to generalize the idea of the Median and Quartiles. Recall that:

- 50% of Datapoints are to the left of the Median, and 50% are to the right, so Median divides the (sorted) Dataset in the (approximate) proportion 50% - 50%
- 25% of Datapoints are to the left of the Lower Quartile Q_1, and 75% are to the right, so Q_1 divides the (sorted) Dataset in the (approximate) proportion 25%-75%
- 75% of Datapoints are to the left of the Upper Quartile Q_3, and 25% are to the right, so Q_3 divides the (sorted) Dataset in the (approximate) proportion 75%-25%

Now, let $\alpha \in (0, 1)$. We want to find a real number q_α dividing our (sorted) Dataset into the proportion $100\alpha\% - 100(1-\alpha)\%$, i.e., q_α is a point such that the α-portion of the Datapoints are to the left to q_α, and others are to the right.

Sample Quantiles

Let $x: \; x_1, x_2, ..., x_n$ be our 1D numerical Dataset. Assume also that $\alpha \in (0, 1)$.

Definition: The Quantile of order α (or $100\alpha\%$ order, the α-Quantile) of x is defined by

$$q_\alpha = q_\alpha^x = x_{([\alpha \cdot n])}.$$

Sample Quantiles

Let $x: \ x_1, x_2, ..., x_n$ be our 1D numerical Dataset. Assume also that $\alpha \in (0, 1)$.

Definition: The Quantile of order α (or $100\alpha\%$ order, the α-Quantile) of x is defined by

$$q_\alpha = q_\alpha^x = x_{([\alpha \cdot n])}.$$

Note: $[\alpha \cdot n]$ is the integer part of $\alpha \cdot n$, and $x_{([\alpha \cdot n])}$ is the $[\alpha \cdot n]$-th Order Statistics of x.

Sample Quantiles

Let $x: \ x_1, x_2, ..., x_n$ be our 1D numerical Dataset. Assume also that $\alpha \in (0,1)$.

Definition: The Quantile of order α (or $100\alpha\%$ order, the α-Quantile) of x is defined by

$$q_\alpha = q_\alpha^x = x_{([\alpha \cdot n])}.$$

Note: $[\alpha \cdot n]$ is the integer part of $\alpha \cdot n$, and $x_{([\alpha \cdot n])}$ is the $[\alpha \cdot n]$-th Order Statistics of x.

Note: There are different definitions of the α-quantile in the literature and in software implementations. Say, **R** has 9 methods to calculate quantiles.

Sample Quantiles

Definition: The Quantile of order α (or $100\alpha\%$ order, the α-Quantile) of x is defined by

$$q_\alpha = q_\alpha^x = x_{([\alpha \cdot n])}.$$

Sample Quantiles

Definition: The Quantile of order α (or $100\alpha\%$ order, the α-Quantile) of x is defined by

$$q_\alpha = q_\alpha^x = x_{([\alpha \cdot n])}.$$

Note: In the case when $[\alpha \cdot n] = 0$, we take $x_{(0)} = x_{(1)}$.

Sample Quantiles

Definition: The Quantile of order α (or $100\alpha\%$ order, the α-Quantile) of x is defined by

$$q_\alpha = q_\alpha^x = x_{([\alpha \cdot n])}.$$

Note: In the case when $[\alpha \cdot n] = 0$, we take $x_{(0)} = x_{(1)}$.

Note: Quartiles are not always Quantiles (in the sense of our definitions). Say, Q_1 is not always the 25% Quantile (despite their idea is to split the Dataset into the proportion 25%-75%).

Sample Quantiles

Definition: The Quantile of order α (or $100\alpha\%$ order, the α-Quantile) of x is defined by

$$q_\alpha = q_\alpha^x = x_{([\alpha \cdot n])}.$$

Note: In the case when $[\alpha \cdot n] = 0$, we take $x_{(0)} = x_{(1)}$.

Note: Quartiles are not always Quantiles (in the sense of our definitions). Say, Q_1 is not always the 25% Quantile (despite their idea is to split the Dataset into the proportion 25%-75%). By our definition, *Quantile is a Datapoint*, but a Quartile is not necessarily a Datapoint.

Sample Quantiles

Definition: The Quantile of order α (or $100\alpha\%$ order, the α-Quantile) of x is defined by

$$q_\alpha = q_\alpha^x = x_{([\alpha \cdot n])}.$$

Note: In the case when $[\alpha \cdot n] = 0$, we take $x_{(0)} = x_{(1)}$.

Note: Quartiles are not always Quantiles (in the sense of our definitions). Say, Q_1 is not always the 25% Quantile (despite their idea is to split the Dataset into the proportion 25%-75%). By our definition, *Quantile is a Datapoint*, but a Quartile is not necessarily a Datapoint.

Note: Sometimes Quantiles are called Percentiles.

Example

Example: Find the 20% and 60% quantiles of

$$x : -2, 3, 5, 7, 8, -3, 4, 5, 2$$

Solution: OTB

Example

Now, let us calculate Quantiles in **R**:

```r
x <- 1:15
quantile(x,0.21)
```

```
##  21%
## 3.94
```

```r
quantile(x, c(0.1,0.3,0.7))
```

```
##  10%  30%  70%
##  2.4  5.2 10.8
```

Theoretical Quantiles

Now assume X is a r.v. with CDF $F(x)$ and PDF $f(x)$.

Theoretical Quantiles

Now assume X is a r.v. with CDF $F(x)$ and PDF $f(x)$. For $\alpha \in (0, 1)$, we define the α-quantile q_α to be the real number satisfying:

$$q_\alpha = q_\alpha^F = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}.$$

Theoretical Quantiles

Now assume X is a r.v. with CDF $F(x)$ and PDF $f(x)$. For $\alpha \in (0,1)$, we define the α-quantile q_α to be the real number satisfying:

$$q_\alpha = q_\alpha^F = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}.$$

If F is stricly increasing and continuous, then we can define

$$F(q_\alpha) = \alpha, \qquad i.e., \qquad q_\alpha = F^{-1}(\alpha).$$

Theoretical Quantiles

Now assume X is a r.v. with CDF $F(x)$ and PDF $f(x)$. For $\alpha \in (0,1)$, we define the α-quantile q_α to be the real number satisfying:

$$q_\alpha = q_\alpha^F = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}.$$

If F is stricly increasing and continuous, then we can define

$$F(q_\alpha) = \alpha, \qquad i.e., \qquad q_\alpha = F^{-1}(\alpha).$$

If F has a Density, $f(x)$, then q_α can be calculated from

$$\int_{-\infty}^{q_\alpha} f(x)dx = \alpha.$$

Theoretical Quantiles, Geometrically, by CDF

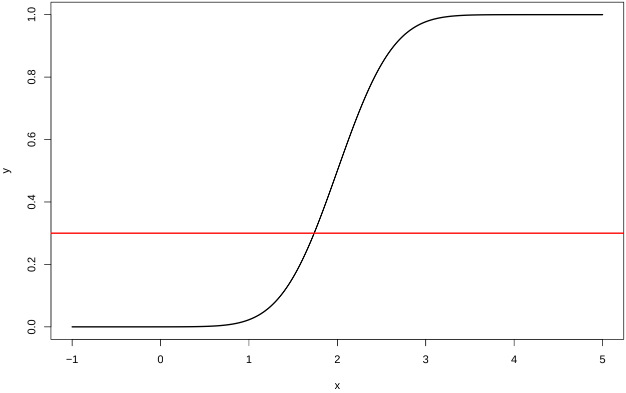
First we draw the CDF $y = F(x)$ graph, then draw the line $y = \alpha$.

Theoretical Quantiles, Geometrically, by CDF

First we draw the CDF $y = F(x)$ graph, then draw the line $y = \alpha$. Now, we keep the portion of the graph of $y = F(x)$ above (or on) the line $y = \alpha$. Then we take the leftmost point of the remaining part, and the x-coordinate of that point will be q_α.

Theoretical Quantiles, Geometrically, by CDF

```
alpha <- 0.3
x <- seq(-1,5, by = 0.01)
y <- pnorm(x, mean = 2, sd = 0.5)
plot(x,y, type = "l", xlim = c(-1,5), lwd = 2)
abline(h = alpha, lwd = 2, col = "red")
```



Theoretical Quantiles, Geometrically, by PDF

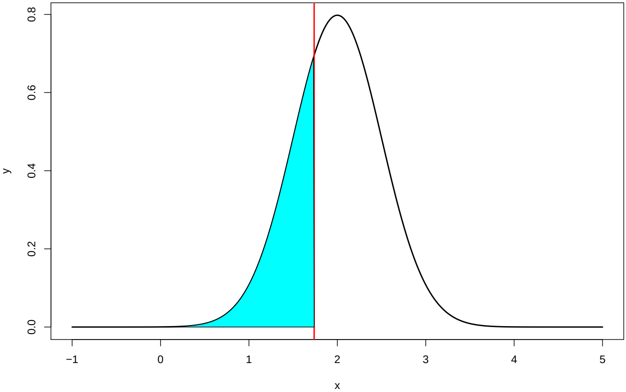
Now, assume our Distribution is continuous. We plot the graph of the PDF $y = f(x)$.

Theoretical Quantiles, Geometrically, by PDF

Now, assume our Distribution is continuous. We plot the graph of the PDF $y = f(x)$. We take q_α to be the smallest point such that the area under the PDF curve **left to** q_α is exactly α.

Theoretical Quantiles, Geometrically, by PDF

```
alpha <- 0.3; q.alpha <- qnorm(alpha, mean = 2, sd = 0.5)
x <- seq(-1,5, by = 0.01)
y <- dnorm(x, mean = 2, sd = 0.5)
plot(x,y, type = "l", xlim = c(-1,5), lwd = 2)
abline(v = q.alpha, lwd = 2, col = "red")
polygon(c(x[x<=q.alpha], q.alpha),c(y[x<=q.alpha],0),col="cyan")
```



Examples

Example: Find the 30% quantile of $\mathit{Unif}\,[3, 10]$

Solution: OTB

Examples

Example: Find the 30% quantile of $Unif[3, 10]$

Solution: OTB

Example: Find the 70% quantile of the Distribution with the PDF

$$f(x) = \begin{cases} 3x^2, & x \in [0, 1] \\ 0, & otherwise \end{cases}$$

Solution: OTB

Theoretical Quantiles, again

Now, if q_α is the α-quantile of some Distribution, and X is a r.v. from that Distribution, then

$$\mathbb{P}(X \leq q_\alpha) \geq \alpha \qquad \text{and} \qquad \mathbb{P}(X \geq q_\alpha) \geq 1 - \alpha.$$

Theoretical Quantiles, again

Now, if q_α is the α-quantile of some Distribution, and X is a r.v. from that Distribution, then

$$\mathbb{P}(X \leq q_\alpha) \geq \alpha \qquad \text{and} \qquad \mathbb{P}(X \geq q_\alpha) \geq 1 - \alpha.$$

Note: Here we are taking inequalities, and not, say, $\mathbb{P}(X \leq q_\alpha) = \alpha$, since, in the Discrete r.v. case, we can have no q_α with exact equality. Say, if $X \sim \mathit{Bernoulli}(0.2)$, and $\alpha = 0.4$, then no q_α exists with $\mathbb{P}(X \leq q_\alpha) = \alpha$.

Theoretical Quantiles, again

Now, if q_α is the α-quantile of some Distribution, and X is a r.v. from that Distribution, then

$$\mathbb{P}(X \leq q_\alpha) \geq \alpha \qquad \text{and} \qquad \mathbb{P}(X \geq q_\alpha) \geq 1 - \alpha.$$

Note: Here we are taking inequalities, and not, say, $\mathbb{P}(X \leq q_\alpha) = \alpha$, since, in the Discrete r.v. case, we can have no q_α with exact equality. Say, if $X \sim \textit{Bernoulli}(0.2)$, and $\alpha = 0.4$, then no q_α exists with $\mathbb{P}(X \leq q_\alpha) = \alpha$.

Note: If $\alpha = 0.5$, we call $q_\alpha = q_{0.5}$ to be the **Median of the Distribution**.

Theoretical Quantiles, again

Now, if q_α is the α-quantile of some Distribution, and X is a r.v. from that Distribution, then

$$\mathbb{P}(X \leq q_\alpha) \geq \alpha \qquad \text{and} \qquad \mathbb{P}(X \geq q_\alpha) \geq 1 - \alpha.$$

Note: Here we are taking inequalities, and not, say, $\mathbb{P}(X \leq q_\alpha) = \alpha$, since, in the Discrete r.v. case, we can have no q_α with exact equality. Say, if $X \sim \mathit{Bernoulli}(0.2)$, and $\alpha = 0.4$, then no q_α exists with $\mathbb{P}(X \leq q_\alpha) = \alpha$.

Note: If $\alpha = 0.5$, we call $q_\alpha = q_{0.5}$ to be the **Median of the Distribution**. So if we consider a Continuous r.v. and draw the PDF of that r.v., then the Median is the (leftmost) point dividing the area under the PDF curve into 50%-50% portions.