AUA CS 108, Statistics, Fall 2019

Lecture 10

Michael Poghosyan
YSU, AUA
michael@ysu.am, mpoghosyan@aua.am

16 Sep 2019

Contents

- Sample Quantiles
- Theoretical Quantiles
- Q-Q Plots

Quiz

Quiz 01: For Density Histogram, **divide relative frequencies by the length of the corresponding interval!**

Last Lecture ReCap

- What is a BoxPlot?

Last Lecture ReCap

- What is a BoxPlot?
- What is an Outlier?

Last Lecture ReCap

- What is a BoxPlot?
- What is an Outlier?
- What is it for?

Last Lecture ReCap

- What is a BoxPlot?
- What is an Outlier?
- What is it for?
- What are the Quantiles?

Last Lecture ReCap

- What is a BoxPlot?
- What is an Outlier?
- What is it for?
- What are the Quantiles?
- What is the difference between Quartiles and Quantiles?

Sample Quantiles, ReCap

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, ReCap

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.

Sample Quantiles, ReCap

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.

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

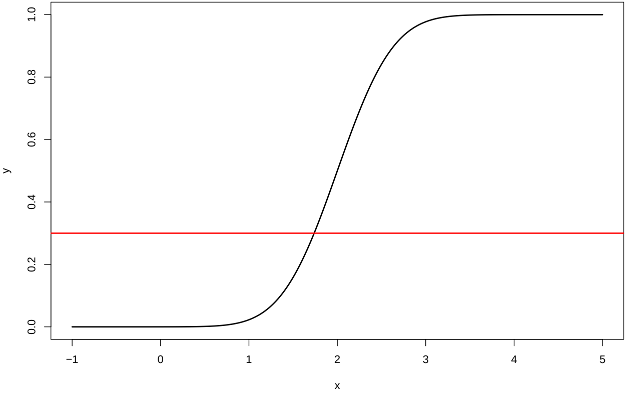
First we draw the CDF $y = F(x)$ graph, then draw the line $y = \alpha$.

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

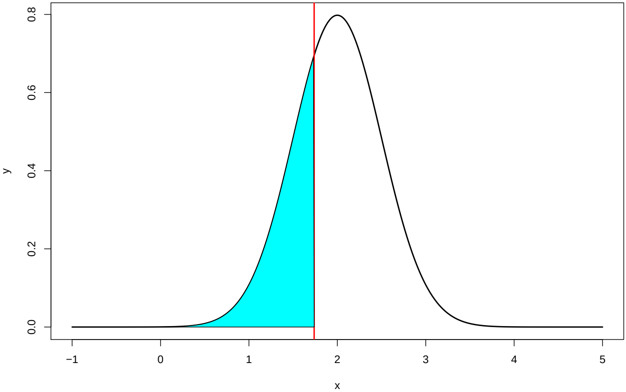
Now, assume our Distribution is continuous. We plot the graph of the PDF $y = f(x)$.

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")
```



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 $X \sim \mathit{Bernoulli}(0.2)$, 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**.

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**. 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.

Theoretical Quantiles, again

Later we will use a lot quantiles. When constructing Confidence Intervals or Hypothesis Testing, we will use Quantiles of the Normal Distribution, t-Distribution, χ^2-Distribution.

Theoretical Quantiles, again

Later we will use a lot quantiles. When constructing Confidence Intervals or Hypothesis Testing, we will use Quantiles of the Normal Distribution, t-Distribution, χ^2-Distribution.

Say, later, by z_α we will denote the α-quantile of the Standard Normal Distribution, $\mathcal{N}(0, 1)$.

Theoretical Quantiles, again

Later we will use a lot quantiles. When constructing Confidence Intervals or Hypothesis Testing, we will use Quantiles of the Normal Distribution, t-Distribution, χ^2-Distribution.

Say, later, by z_α we will denote the α-quantile of the Standard Normal Distribution, $\mathcal{N}(0,1)$.

Say, we will take $\alpha \in (0,1)$ and find two points $a, b \in \mathbb{R}$ such that for $X \sim \mathcal{N}(0,1)$

$$\mathbb{P}(X \leq a) = \mathbb{P}(X \geq b) = \frac{\alpha}{2}.$$

Theoretical Quantiles, again

Later we will use a lot quantiles. When constructing Confidence Intervals or Hypothesis Testing, we will use Quantiles of the Normal Distribution, t-Distribution, χ^2-Distribution.

Say, later, by z_α we will denote the α-quantile of the Standard Normal Distribution, $\mathcal{N}(0,1)$.

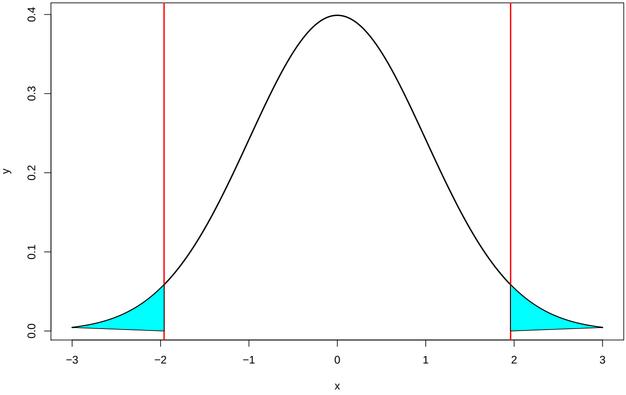
Say, we will take $\alpha \in (0,1)$ and find two points $a, b \in \mathbb{R}$ such that for $X \sim \mathcal{N}(0,1)$

$$\mathbb{P}(X \le a) = \mathbb{P}(X \ge b) = \frac{\alpha}{2}.$$

The idea is to find a symmetric (in fact, the smallest length) interval $[a, b]$ such that for a Standard Normal r.v. X, the chances of $X \notin [a, b]$ are small, are exactly α.

Graphically

```
alpha <- 0.05; z.alpha <- qnorm(alpha/2, mean = 0, sd = 1)
x <- seq(-3,3, by = 0.01)
y <- dnorm(x, mean = 0, sd = 1)
plot(x,y, type = "l", xlim = c(-3,3), lwd = 2)
abline(v = z.alpha, lwd = 2, col = "red")
abline(v = -z.alpha, lwd = 2, col = "red")
polygon(c(x[x<=z.alpha], z.alpha),c(y[x<=z.alpha],0),col="cyan")
polygon(c(x[x>=-z.alpha], -z.alpha),c(y[x>=-z.alpha],0),col="cyan")
```



Theoretical Quantiles, again

Then, it is easy to see, if $\alpha \in (0, 0.5)$ because of the symmetry, that $b = -a$, and

$$a = z_{\alpha/2}.$$

Theoretical Quantiles, again

Then, it is easy to see, if $\alpha \in (0, 0.5)$ because of the symmetry, that $b = -a$, and

$$a = z_{\alpha/2}.$$

So

$$b = -z_{\alpha/2} = z_{1-\alpha/2}$$

Theoretical Quantiles, again

Then, it is easy to see, if $\alpha \in (0, 0.5)$ because of the symmetry, that $b = -a$, and

$$a = z_{\alpha/2}.$$

So

$$b = -z_{\alpha/2} = z_{1-\alpha/2}$$

Note: Please be careful when using Normal Tables. Usually, there is a picture above the table, on which you can find the explanation of the process. Just search "Normal tables" in Google Images.

Q-Q Plots

Next, we consider three important statistical problems: Check visually if

Q-Q Plots

Next, we consider three important statistical problems: Check visually if

- two given Datasets (possibly, of different sizes) are from the same Distribution;

Q-Q Plots

Next, we consider three important statistical problems: Check visually if

- two given Datasets (possibly, of different sizes) are from the same Distribution;
- a given Dataset comes from a given Distribution;

Q-Q Plots

Next, we consider three important statistical problems: Check visually if

- two given Datasets (possibly, of different sizes) are from the same Distribution;
- a given Dataset comes from a given Distribution;
- given two theoretical Distributions, check if one of them is a shifted-scaled version of the other one, or check if one has *fatter tails* than the other one

Q-Q Plots, Data vs Data

Now, assume we have two Datasets, not necessarily of the same size:

$$x: \ x_1, x_2, ..., x_n \qquad \text{and} \qquad y: \ y_1, y_2, ..., y_m$$

Q-Q Plots, Data vs Data

Now, assume we have two Datasets, not necessarily of the same size:

$$x: \; x_1, x_2, ..., x_n \qquad \text{and} \qquad y: \; y_1, y_2, ..., y_m$$

Question: Are x and y coming from the same Distribution?

Q-Q Plots, Data vs Data

Now, assume we have two Datasets, not necessarily of the same size:

$$x : \; x_1, x_2, ..., x_n \qquad \text{and} \qquad y : \; y_1, y_2, ..., y_m$$

Question: Are x and y coming from the same Distribution?

Q-Q Plot helps to answer to this question visually.

Q-Q Plots, Data vs Data

Now, assume we have two Datasets, not necessarily of the same size:

$$x: \ x_1, x_2, ..., x_n \qquad \text{and} \qquad y: \ y_1, y_2, ..., y_m$$

Question: Are x and y coming from the same Distribution?

Q-Q Plot helps to answer to this question visually. To draw the Q-Q Plot for Datasets, we take some levels of quantiles, say, for some n,

$$\alpha = \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}$$

and then draw the points (q_α^x, q_α^y).

Q-Q Plots, Data vs Data

Now, assume we have two Datasets, not necessarily of the same size:

$$x: \; x_1, x_2, ..., x_n \qquad \text{and} \qquad y: \; y_1, y_2, ..., y_m$$

Question: Are x and y coming from the same Distribution?

Q-Q Plot helps to answer to this question visually. To draw the Q-Q Plot for Datasets, we take some levels of quantiles, say, for some n,

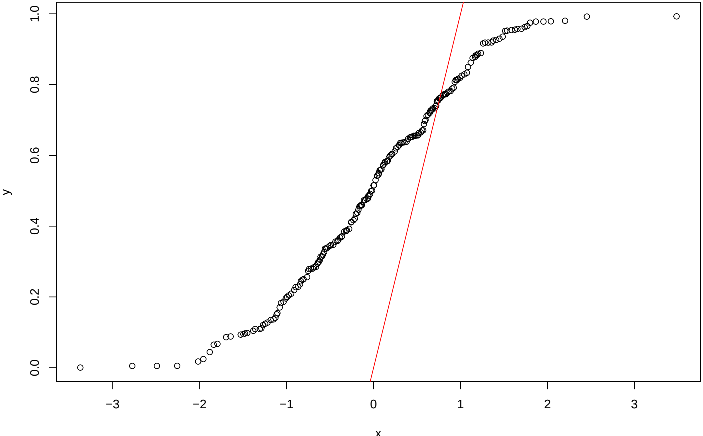
$$\alpha = \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}$$

and then draw the points (q_α^x, q_α^y).

Idea: If x and y are coming from the same Distribution, then the Quantiles of x and y need to be approximately the same, $q_\alpha^x \approx q_\alpha^y$, so geometrically, the points (q_α^x, q_α^y) need to be close to the bisector line.

Example, Q-Q Plots, Data vs Data

```
x <- rnorm(1000)
y <- runif(200)
qqplot(x,y)
abline(0,1, col="red")
```



Example, Q-Q Plots, Data vs Data

```
x <- rnorm(1000)
y <- rnorm(500)
qqplot(x,y)
abline(0,1, col="red")
```