BoxPlot

BoxPlot (or Box and Whiskers Plot) is another very common method of visualization. To draw the BoxPlot, we calculate the following:

- The Quartiles $Q_1, Q_2 = Median, Q_3$
- the Lower and Upper Fences
  $W_1 = \min\{x_i : x_i \geq Q_1 - 1.5 \cdot IQR\}$ and
  $W_2 = \max\{x_i : x_i \leq Q_3 + 1.5 \cdot IQR\}$,i.e., the first and last observations lying in

$$\left[Q_1 - \frac{3}{2}IQR,\ Q_3 + \frac{3}{2}IQR\right];$$

the lines joining that fences to corresponding quartiles are the *Whiskers*;

- the set of all Outliers

$$O = \left\{x_i : x_i \notin \left[Q_1 - \frac{3}{2}IQR, Q_3 + \frac{3}{2}IQR\right]\right\}$$

Then we draw the points W_1, Q_1, Q_2, Q_3, W_2 on the real line and add all outliers, and make a box over $[Q_1, Q_3]$.

Example: Draw the Boxplot of

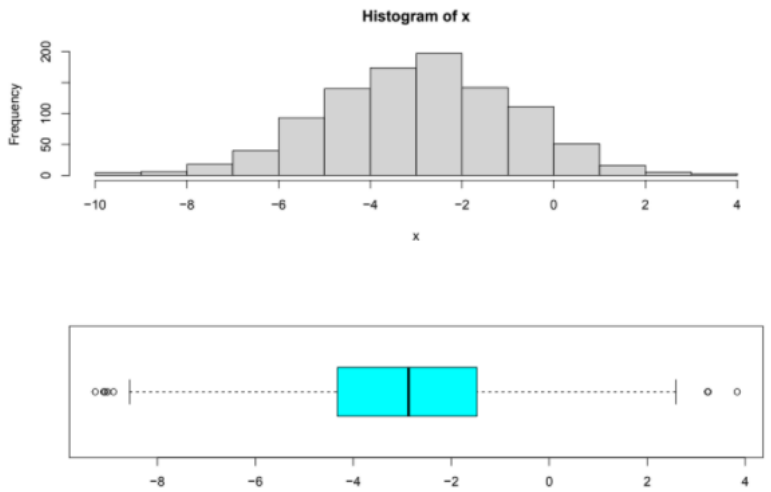
$$x: \ 0, -2, 2, 1, 5, 6, 4, 1, 2, 1, 12$$

Now, using **R**:

```
x <- c(0, -2, 2, 1, 5, 6, 4, 1, 2, 1, 12)
boxplot(x)
```

Here are some Datasets' Histograms along with the BoxPlots:

```
x <- rnorm(1000, mean = -3, sd = 2)
par(mfrow=c(2,1)); hist(x)
boxplot(x, horizontal = T, col = "cyan")
```



Note

Recall that an **Outlier** in the BoxPlot sense is a Datapoint x_k with

$$x_k \notin \left[Q_1 - \frac{3}{2} IQR,\ Q_3 + \frac{3}{2} IQR \right].$$

Another way to define an **Outlier:** Datapoint x_k is an Outlier, if

$$|x_k - \bar{x}| \geq 3 \cdot sd(x).$$

Note: Where the coefficient $\frac{3}{2}$ in front of the IQR comes from? This comes from the Normal Distribution: if our r.v. X is Normally Distributed, then (with theoretical Quartiles)

$$\mathbb{P}(X \in [Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR]) \approx 0.993,$$

so the chances that an Observation will be outside of this interval are very small. So if we see that kind of Observation, we think that this number is an Outlier.

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.

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.

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.

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: Find the 20% and 60% quantiles of

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

Theoretical Quantiles

Now assume X is a r.v. with CDF $F(x)$ or, in the case X is continuous, with 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

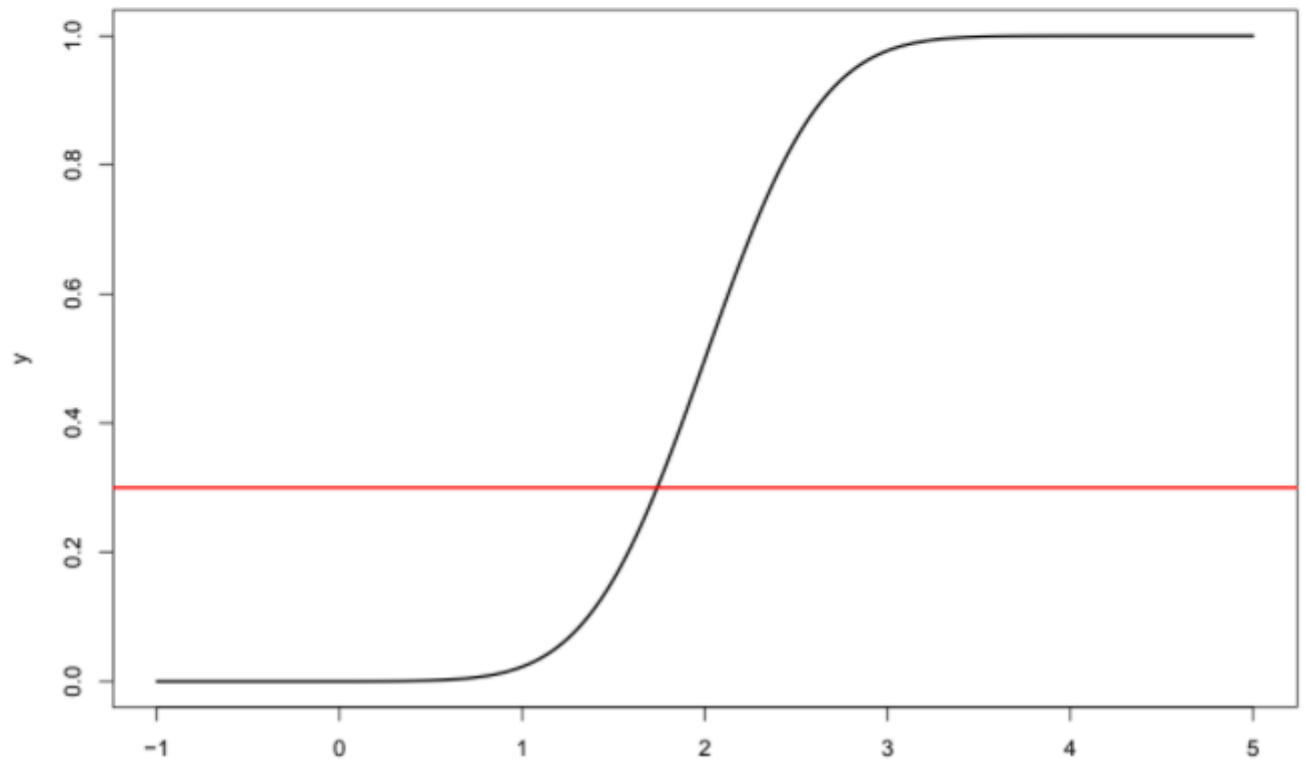
$$\int_{-\infty}^{q_\alpha} f(x)dx = \alpha.$$

Theoretical Quantiles, Geometrically, by CDF

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

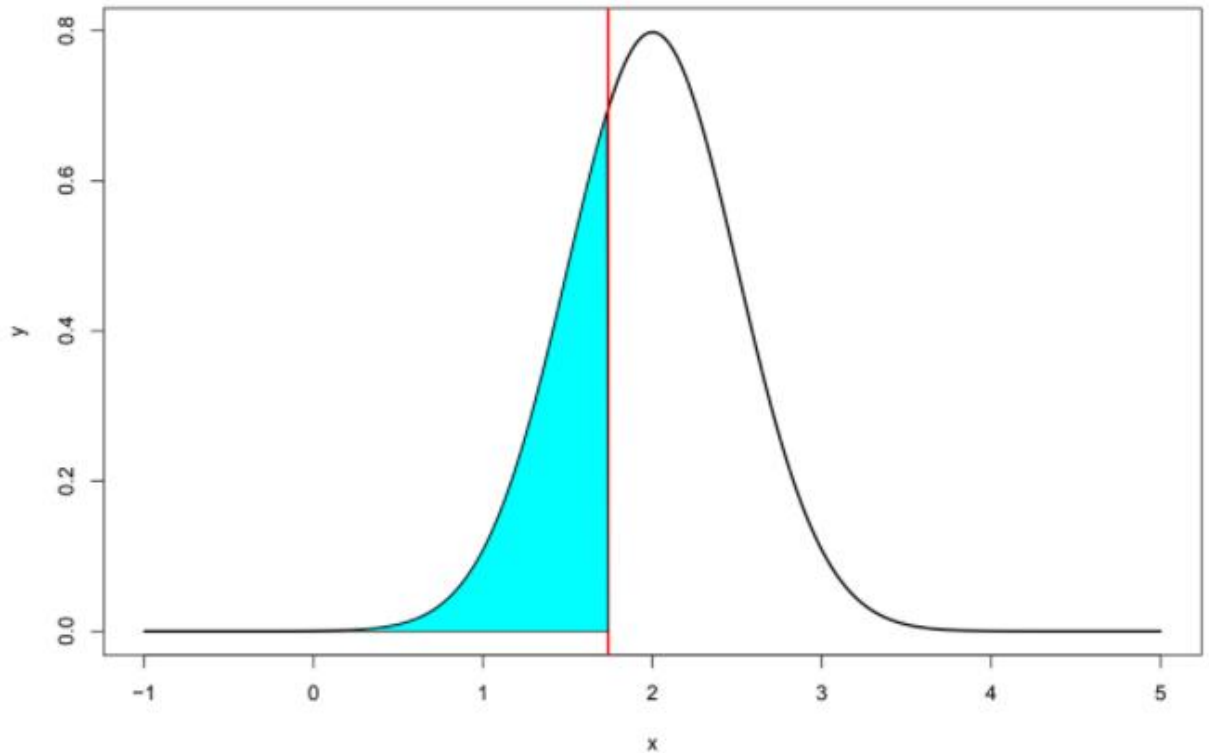


Now, assume our Distribution is continuous. We plot the graph of the PDF $y = f(x)$.

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")
```



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

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