Nonparametric Models

Last modified: 2026-09-29 00:05:17 (PDT)

Empirical CDF

Definition 0.1 (Empirical CDF)**.** For observed values $x_1, \dots, x_n$, the **empirical cumulative distribution function** (**empirical CDF**) is the function

$$\hat{F}(t) \overset{\text{def}}{=} \frac{1}{n} \sum_{j=1}^{n} I(x_j \le t), \quad t \in \mathbb{R},$$

where $I(A)$ is the indicator function: $I(A) = 1$ if A holds, and $I(A) = 0$ otherwise.

Example 0.1 (Numerical example: empirical CDF)**.** Suppose the observed values are $4, 1, 7, 3$, so $n = 4$. Two of the four values (1 and 3) are at most 3.5, so:

$$\begin{aligned}
\hat{F}(3.5) &= \frac{1}{4}(I(4 \le 3.5) + I(1 \le 3.5) + I(7 \le 3.5) + I(3 \le 3.5)) && \text{(definition of } \hat{F}\text{)} \\
&= \frac{1}{4}(0 + 1 + 0 + 1) && \text{(evaluate each indicator)} \\
&= 0.5 && \text{(arithmetic)}
\end{aligned}$$

Similarly, $\hat{F}(0) = 0$, because no value is at most 0, and $\hat{F}(7) = 1$, because every value is at most 7.

Order statistics

Definition 0.2 (Order statistics)**.** For observed values $x_1, \dots, x_n$, the **order statistics** are the same values sorted in increasing order, written

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

So $x_{(1)}$ is the smallest value and $x_{(n)}$ is the largest.

Example 0.2 (Numerical example: order statistics)**.** If the observed values are $4, 1, 7, 3$, then the order statistics are $x_{(1)} = 1$, $x_{(2)} = 3$, $x_{(3)} = 4$, and $x_{(4)} = 7$.

Sample quantiles

Definition 0.3 (Sample quantile function)**.** Given an empirical CDF $\hat{F}$, the **sample quantile function** (also called the **empirical quantile function**) is

$$\hat{Q}(p) \stackrel{\text{def}}{=} \inf\left\{t : \hat{F}(t) \geq p\right\}, \quad 0 < p \leq 1.$$

For a given p, the value $\hat{Q}(p)$ is the **sample p quantile**, also called the **sample $100p$th percentile**.

The CDF F and the quantile function Q describe a probability distribution, while $\hat{F}$ and $\hat{Q}$ describe a sample, and serve as estimates of F and Q.

Theorem 0.1 (Order-statistics form of the sample quantile)**.** *If $x_{(1)} \leq \cdots \leq x_{(n)}$ are the order statistics and $i \in \{1, \ldots, n\}$, then $\hat{Q}(p) = x_{(i)}$ for every p with $(i-1)/n < p \leq i/n$. In particular, $\hat{Q}(i/n) = x_{(i)}$.*

Proof

Proof. Fix p with $(i-1)/n < p \leq i/n$.
For any $t < x_{(i)}$, the $n - i + 1$ values $x_{(i)}, \ldots, x_{(n)}$ all exceed t, so at most $i - 1$ values are at most t, and:

$$\begin{aligned}\hat{F}(t) &\leq \frac{i-1}{n} && \text{(at most } i-1 \text{ indicators equal 1)}\\ &< p && \text{(choice of } p\text{)}\end{aligned}$$

So no $t < x_{(i)}$ belongs to the set $\left\{t : \hat{F}(t) \geq p\right\}$.
At $t = x_{(i)}$, the i values $x_{(1)}, \ldots, x_{(i)}$ are all at most $x_{(i)}$, so:

$$\begin{aligned}\hat{F}\left(x_{(i)}\right) &\geq \frac{i}{n} \quad \text{(at least } i \text{ indicators equal 1)} \\ &\geq p \quad \text{(choice of } p\text{)}\end{aligned}$$

So $x_{(i)}$ belongs to the set, and it is the smallest member of the set, because no smaller t belongs to it. Therefore $\hat{Q}(p) = x_{(i)}$. $\square$

Example 0.3 (Numerical example: sample quantiles)**.** Using the same data $4, 1, 7, 3$, with order statistics $1, 3, 4, 7$, Theorem 0.1 says the sample quantile function is a step function:

- $\hat{Q}(p) = 1$ for $0 < p \leq 1/4$;
- $\hat{Q}(p) = 3$ for $1/4 < p \leq 1/2$;
- $\hat{Q}(p) = 4$ for $1/2 < p \leq 3/4$;
- $\hat{Q}(p) = 7$ for $3/4 < p \leq 1$.

So $\hat{Q}(0.5) = 3$ and $\hat{Q}(0.9) = 7$. R's `quantile()` function computes this definition when called with `type = 1`:

```
quantile(c(4, 1, 7, 3), probs = c(0.5, 0.9), type = 1)
#> 50% 90%
#>   3   7
```

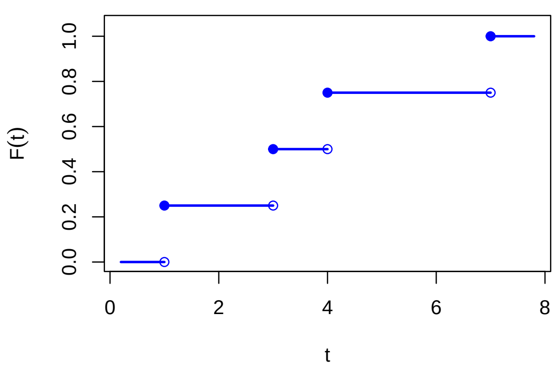
Other sources, and other software defaults, define sample quantiles differently, mostly by interpolating between order statistics. R's `quantile()` offers nine definitions through its `type` argument, and its default (`type = 7`) interpolates, so `quantile(c(4, 1, 7, 3), 0.5)` returns 3.5, not 3. The usual sample median is another interpolated quantile: for an even number of observations, it averages the two middle order statistics. These notes use Definition 0.3, which always returns one of the observed values.

The empirical CDF and quantile function as generalized inverses

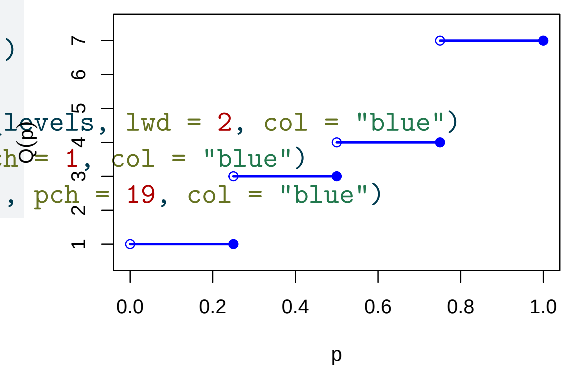
Both functions in Figure 1 are step functions, so neither is one-to-one. Each flat piece of one function corresponds to a jump of the other: for example, $\hat{F}(t) = 0.5$ for every $t \in [3, 4)$, and $\hat{Q}$ jumps from 3 to 4 at $p = 0.5$. So $\hat{Q}$ is a *generalized* inverse of $\hat{F}$, not an inverse in the usual sense: $\hat{Q}(\hat{F}(t)) \leq t$ for every $t \geq x_{(1)}$, with equality only at the observed values.

```r
x <- c(4, 1, 7, 3)
n <- length(x)
x_ord <- sort(x)
p_ord <- seq_len(n) / n
ecdf_x_padding <- 0.8
x_min <- min(x_ord) - ecdf_x_padding
x_max <- max(x_ord) + ecdf_x_padding

x_left <- c(x_min, x_ord)
x_right <- c(x_ord, x_max)
f_levels <- c(0, p_ord)

plot(
  0,
  0,
  type = "n",
  xlim = c(x_min, x_max),
  ylim = c(0, 1.05),
  xlab = "t",
  ylab = expression(hat(plain("F"))(t))
)
segments(x_left, f_levels, x_right, f_levels, lwd = 2, col = "blue")
points(x_ord, f_levels[seq_len(n)], pch = 1, col = "blue")
points(x_ord, f_levels[seq_len(n) + 1], pch = 19, col = "blue")
```

```r
eqf_y_padding <- 0.5
q_left <- c(0, p_ord[-n])
q_right <- p_ord
plot(
  0,
  0,
  type = "n",
  xlim = c(0, 1),
  ylim = c(min(x_ord) - eqf_y_padding, max(x_ord) + ec
  xlab = "p",
  ylab = expression(hat(Q)(p))
)
segments(q_left, x_ord, q_right, x_ord, lwd = 2, col =
points(q_left, x_ord, pch = 1, col = "blue")
points(q_right, x_ord, pch = 19, col = "blue")
```



(a) Empirical CDF, horizontal pieces only. Closed circles mark included endpoints; open circles mark excluded endpoints.



(b) Sample quantile function, horizontal pieces only. Open circles mark excluded left endpoints; closed circles mark included right endpoints.

Figure 1: The empirical CDF and the sample quantile function for the data 4, 1, 7, 3 (order statistics 1, 3, 4, 7).