# The parametric and data-dependent distributions: Definitions

- **Definition:** A parametric distribution is a set of distribution functions where each of these distribution functions is fully specified through one or more (a **fixed and finite** number) parameters.
- A data-dependent distribution is at least as complex as the data or knowledge that produced it, and the number of "parameters" increases as the number of data points or the amount of knowledge increases.
- For example, the empirical distribution is data-dependent

# The empirical distribution (for complete, individual data)

- **Definition:**
  Let $x_1, x_2, \ldots, x_n$ be a data set.
  The empirical distribution function is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{[x_i \leq x]}$$

  for every $x \in \mathbb{R}$, where $\mathbb{I}$ denotes the indicator function.
  Less formally,

$$F_n(x) = \frac{\text{number of observations} \leq x}{n}$$

# The risk set and some notation

- The set of observed values is referred to as the risk set (the number of observations is also sometimes called the same thing)
- **Notation:**

  $n$ ... sample size

  $y_1 < y_2 < \cdots < y_k$ ... distinct observed values

  $s_j$ ... the number of times value $y_j$ was observed ($j = 1, 2, \ldots, k$)

  $r_j$ ... the number of observations greater than or equal to $y_j$ ($j = 1, 2, \ldots, k$), i.e.,

$$r_j = \sum_{i=j}^{n} s_i$$

- Note that $\sum_{i=1}^{k} s_i = n$.
- The empirical distribution function can now be written as

$$F_n(x) = \begin{cases} 0, & x < y_1 \\ 1 - \frac{r_j}{n}, & y_{j-1} \leq x < y_j,\ j = 2, \ldots, k \\ 1, & x \geq y_k. \end{cases}$$

# The cumulative hazard rate function

- **Definition:**
  The cumulative hazard rate function is defined as

$$H(x) = -\ln[S(x)].$$

- Note that

$$S(x) = e^{-H(x)}$$

  so that

$$F(x) = 1 - e^{-H(x)}$$

- **If** $S$ is differentiable, then

$$H'(x) = h(x), \text{ i.e., } H(x) = \int_{-\infty}^{x} h(y)\, dy$$

- All in all - it is worthwhile to find **empirical estimates** for $H(x)$

# The Nelson-Åalen estimate

- **Definition:**
  The Nelson-Åalen estimate of the cumulative hazard rate function is defined as

$$\hat{H}(x) = \begin{cases} 0, & x < y_1 \\ \sum_{i=1}^{j-1} \frac{s_i}{r_i}, & y_{j-1} \leq x < y_j,\ j = 2, \ldots, k \\ \sum_{i=1}^{k} \frac{s_i}{r_i}, & x \geq y_k. \end{cases}$$

- The set (or number) of observed values still greater or equal to some $y_j$ is referred to as the risk set at time $j$
- Let's look at a heuristic argument for the Nelson-Åalen estimate