# The idea

Again, the goal is to estimate a parameter based on observations.
The path to a "good" estimate will be optimization (across some set of parameters) of an objective function (which depends on the observed data).
We choose only one function of the observations, namely, the likelihood function. Of course, there are others.

- This time, we allow for the observations to come from *different* distributions, and that the observations themselves are **not** individual. We introduce:
- $X_1, X_2, \ldots, X_n$ ... independent random variables whose realizations we assume are our observations;
- $A_1, \ldots, A_n$ ... sets within which the $n$ observations fall, e.g., single points (singletons $\{x_j\}$ - individual unomdified observations), intervals ($(\alpha, \beta]$ - grouping, truncation, censoring), ...
- So, we observe the event:

$$\{X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n\}$$

- Finally (and this is what makes our method work!), $X_j$ are all assumed to depend on the same (vector) parameter $\theta$

# The likelihood function

- **"Definition:"**
  The **likelihood function** is given by

$$L(\theta) = \prod_{j=1}^{n} \mathbb{P}[X_j \in A_j \,;\, \theta]$$

  for $\theta$ from a certain set of admissible parameters.
  If there exists an admissible parameter $\theta^*$ which maximizes the likelihood function, it is called the **maximum likelihood** estimate of $\theta$.
- Note that the likelihood function **depends** on the data, even though this is not explicit from the notation for it.

# The likelihood function - clarification

- The notation $\mathbb{P}[X_j \in A_j\,;\,\theta]$ should be interpreted liberally; in case that $A_j$ is a singleton $\{x_j\}$ and $X_j$ is a continuous random variable, we have

$$\mathbb{P}[X_j \in A_j\,;\,\theta] = 0$$

  and this would "kill" the likelihood function
- What they really want to have is something like

$$\mathbb{P}[X_j \in (x_j - \varepsilon, x_j + \varepsilon)\,;\,\theta]$$

  for infinitesimally small $\varepsilon$; this is what we approximate by

$$f_{X_j}(x_j\,;\,\theta)$$

# The loglikelihood function

Since the logarithmic function is strictly increasing, and has some other nice properties, we usually maximize the natural logarithm of the likelihood function:

$$l(\theta) = \ln[L(\theta)]$$

# Special case: Complete, individual data

When there is no truncation, no censoring and all observations are singletons:

$$x_1, x_2, \ldots, x_n,$$

then the likelihood and the loglikelihood functions have the forms:

$$L(\theta) = \prod_{j=1}^{n} f_{X_j}(x_j\,;\,\theta)$$

$$l(\theta) = \sum_{j=1}^{n} \ln[f_{X_j}(x_j\,;\,\theta)]$$

# Special case: Complete, grouped data

When there is no truncation, no censoring but the data are grouped into "bins" defined by the partition:

$$c_0 < c_1 < \cdots < c_k$$

and there are $n_j$ observations in the interval $(c_{j-1}, c_j]$, then the likelihood and the loglikelihood functions have the forms:

$$L(\theta) = \prod_{j=1}^{k} [F_{X_j}(c_j\,;\,\theta) - F_{X_j}(c_{j-1}\,;\,\theta)]^{n_j}$$

$$l(\theta) = \sum_{j=1}^{k} n_j \ln[F_{X_j}(c_j\,;\,\theta) - F_{X_j}(c_{j-1}\,;\,\theta)]$$

# Special case: Truncated or censored data

The easy part:
If the $j^{th}$ data-point is said to be censored at a value $u_j$, then we set

$$A_j = [u_j, \infty)$$

The more complicated part:
If the $j^{th}$ data point is truncated at $d_j$, we need to take one of the following two approaches:

- **shifting:** with $x_j$ being the uncensored values

$$L(\theta) = L(\theta \,|\, (x_1 - d_1)_+, \ldots, (x_n - d_n)_+)$$

- **conditioning:** omit $x_j < d_j$, for the rest write conditional expectations in the likelihood function, e.g., for "individual" data

$$L(\theta) = \prod_{j=1}^{n} \frac{f_{X_j}(x_j\,;\,\theta)}{1 - F_{X_j}(d_j\,;\,\theta)}$$

and proceed similarly for other types of data (e.g., when observations are intervals)