The Cumulative Distribution Function

Probability and Statistics for Data Science

Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

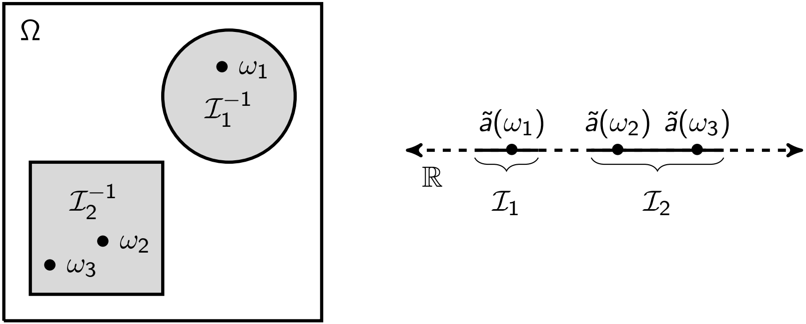
Plan

Define the cumulative distribution function

Define the quantiles of a distribution

Describe how to estimate them from data

Continuous random variables



Continuous random variables

We describe continuous random variables in terms of the probability that they belong to any interval

How do we encode this information?

Cumulative distribution function

The cumulative distribution function (cdf) of a random variable $\tilde{a}$ is

$$F_{\tilde{a}}(a) := \mathrm{P}(\tilde{a} \leq a)$$

Probability that $\tilde{a}$ is less than or equal to a, for all $a \in \mathbb{R}$

Properties

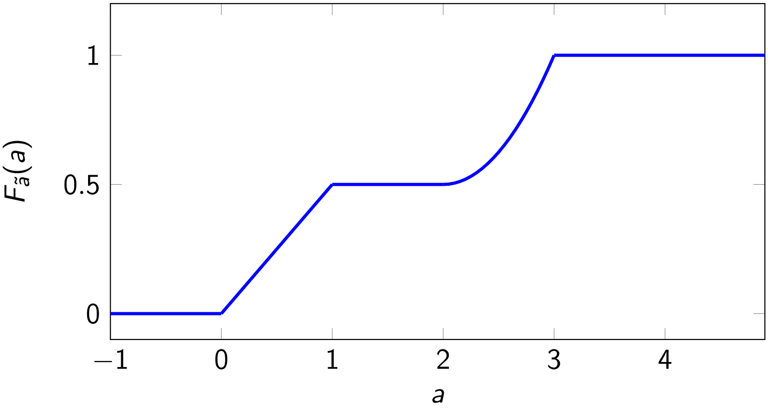
$$\lim_{a \to \infty} F_{\tilde{a}}(a) = \mathrm{P}(\tilde{a} \in \mathbb{R}) = 1$$

$$\lim_{a \to -\infty} F_{\tilde{a}}(a) = 1 - \mathrm{P}(\tilde{a} \in \mathbb{R}) = 0$$

Can $F_{\tilde{a}}(b) < F_{\tilde{a}}(a)$ if $b > a$?

No, $\{\tilde{a} \leq b\} = \{\tilde{a} \leq a\} \cup \{a < \tilde{a} \leq b\}$

Cumulative distribution function



Probability of an interval

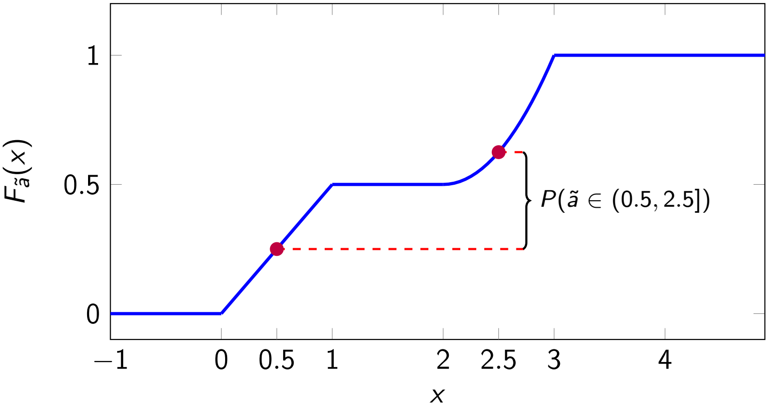
For any a, $b \in \mathbb{R}$, $\mathrm{P}(a < \tilde{a} \leq b)$?

$$\begin{aligned}\mathrm{P}(\tilde{a} \leq b) &= \mathrm{P}(\tilde{a} \in (-\infty, a] \cup (a, b]) \\ &= \mathrm{P}(a < \tilde{a} \leq b) + \mathrm{P}(\tilde{a} \leq a)\end{aligned}$$

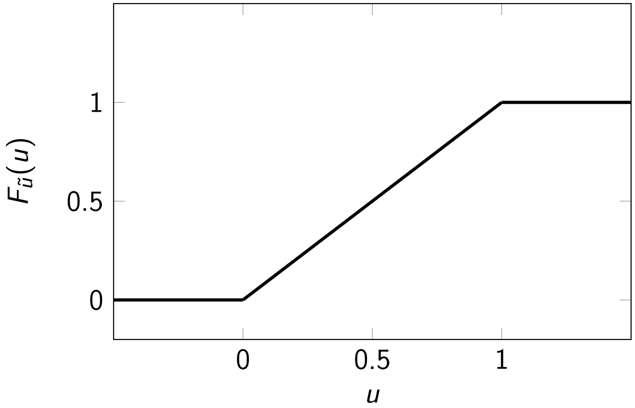
so

$$\mathrm{P}(a < \tilde{a} \leq b) = F_{\tilde{a}}(b) - F_{\tilde{a}}(a)$$

Probability of an interval



Linear cdf



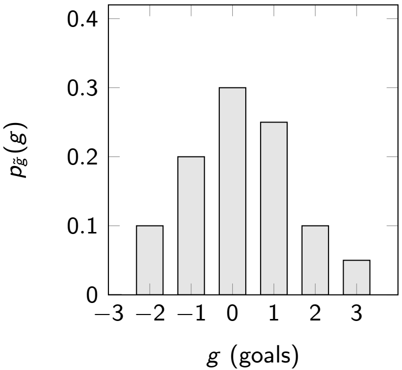
Linear cdf

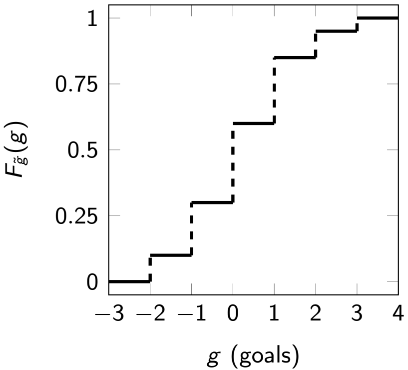
$$F_{\tilde{u}}(u) := \begin{cases} 0 & \text{for } u < 0 \\ u & \text{for } 0 \leq u \leq 1 \\ 1 & \text{for } u > 1 \end{cases}$$

$$\begin{aligned} \mathrm{P}(a < \tilde{u} \leq b) &= F_{\tilde{u}}(b) - F_{\tilde{u}}(a) \\ &= b - a \end{aligned}$$

Probability is proportional to the length of the interval

Cdf of a discrete random variable



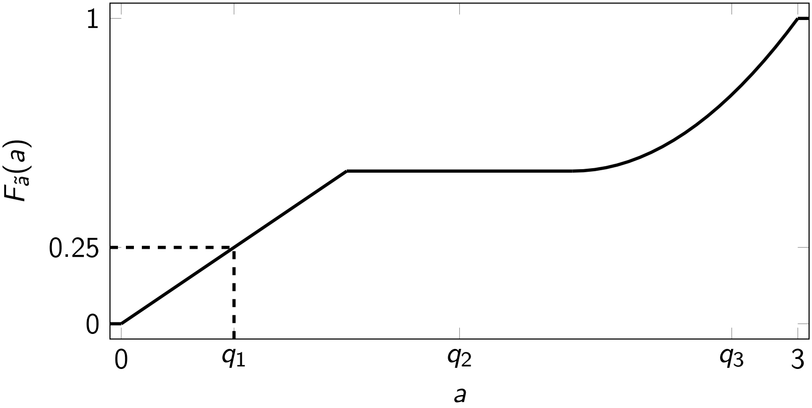


Continuous random variable

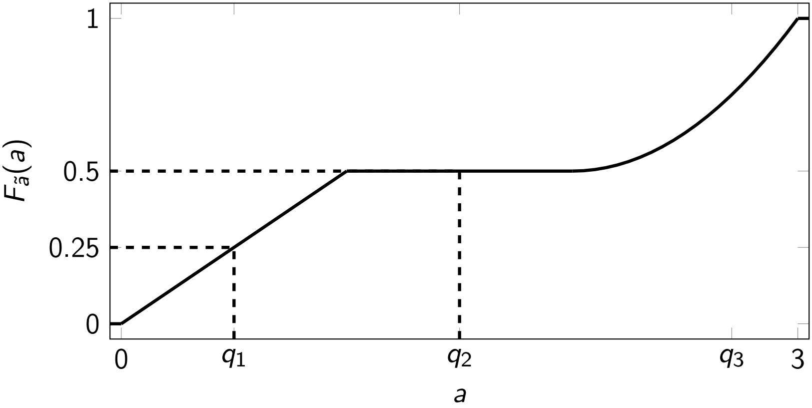
A random variable is continuous if its cdf is continuous

$$\begin{aligned}\mathrm{P}\left(\tilde{a}=a\right) &= F_{\tilde{a}}\left(a\right)-\lim_{\epsilon\to 0} F_{\tilde{a}}\left(a-\epsilon\right) \\ &= 0\end{aligned}$$

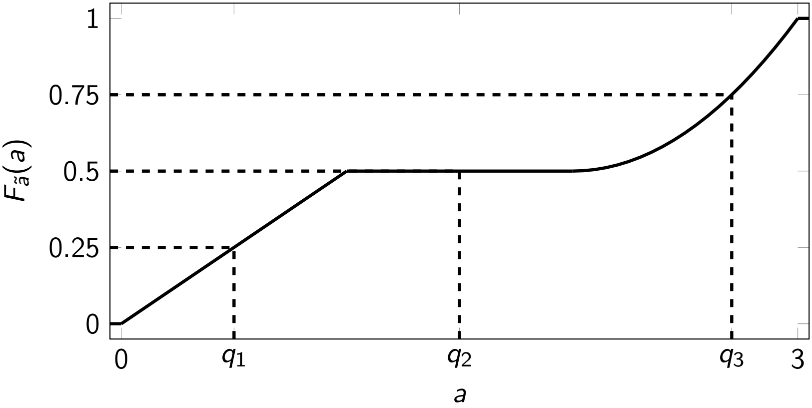
Quantiles



Quantiles



Quantiles



Quantiles

The n-quantiles of $\tilde{a}$ are $n-1$ points q_1, q_2, ..., q_n such that

$$\mathrm{P}(\tilde{a} \leq q_1) = \mathrm{P}(q_1 \leq \tilde{a} \leq q_2) = \cdots = \mathrm{P}(\tilde{a} \geq q_{n-1})$$

or equivalently

$$F_{\tilde{a}}(q_i) = \mathrm{P}(\tilde{a} \leq q_i) = \frac{i}{n} \qquad i = 1, 2, \ldots, n-1$$

4-quantiles are called quartiles: q_1, q_2, q_3

Median

The median q_2 of a continuous random variable $\tilde{a}$ satisfies

$$\mathrm{P}(\tilde{a} \leq q_2) = \mathrm{P}(\tilde{a} > q_2) = \frac{1}{2}$$

or equivalently

$$F_{\tilde{a}}(q_2) = \frac{1}{2}$$

Estimating the cdf from data

For any a $F_{\tilde{a}}(a) := \mathrm{P}(\tilde{a} \leq a)$ is a probability

Use empirical probability!

Empirical cdf

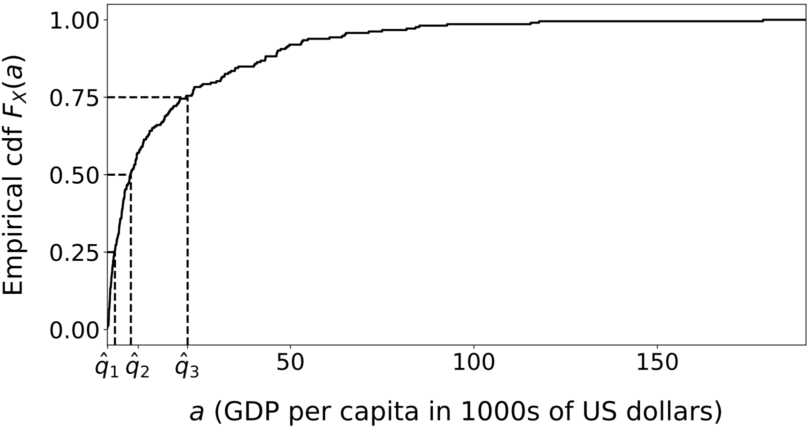
Dataset $X := \{x_1, x_2, \ldots, x_n\}$

The empirical cumulative distribution function $F_X : \mathbb{R} \to [0, 1]$ equals

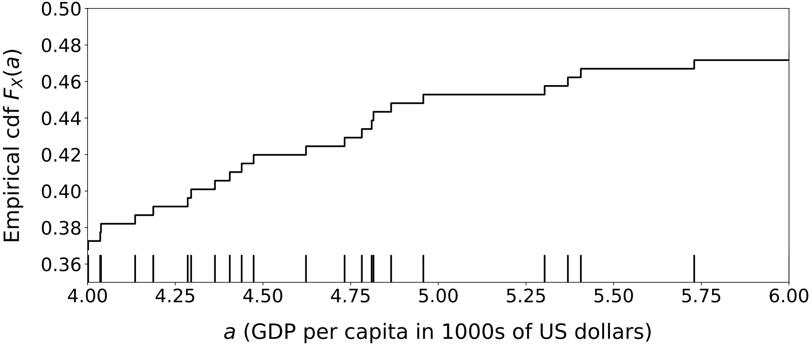
$$F_X(a) := \frac{1}{n} \sum_{i=1}^{n} 1_{x_i \leq a}$$

where $1_{x_i \leq a}$ equals one if $x_i \leq a$ and zero otherwise

GDP per capita



GDP per capita



Quantile estimation

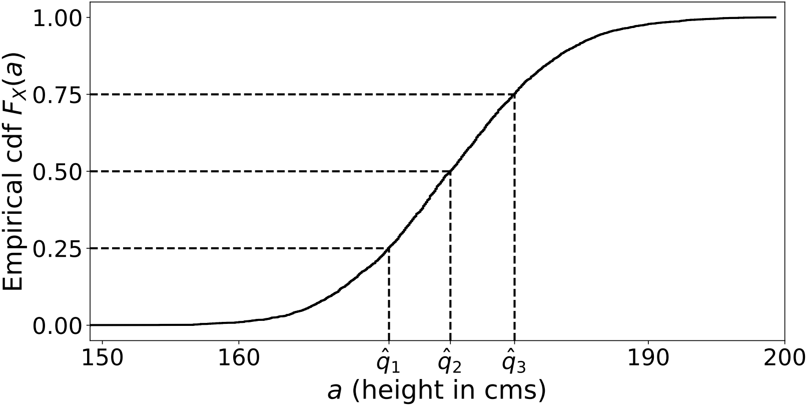
Dataset $X := \{x_1, x_2, \ldots, x_n\}$

The n-quantiles of the data are $n-1$ points $\hat{q}_1$, $\hat{q}_2$, ..., $\hat{q}_n$ such that

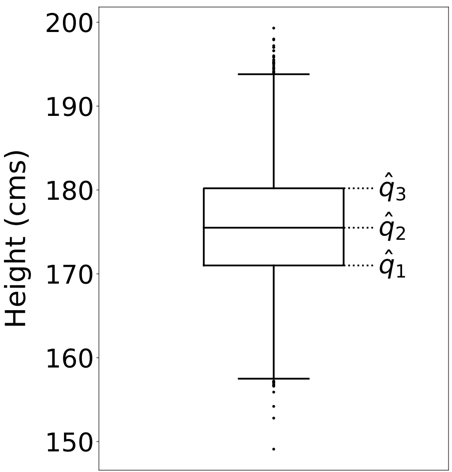
$$\mathrm{P}_X(\tilde{a} \leq \hat{q}_1) = \mathrm{P}_X(\hat{q}_1 \leq \tilde{a} \leq \hat{q}_2) = \cdots = \mathrm{P}_X(\tilde{a} \geq \hat{q}_{n-1})$$

where P_X is the empirical probability of the data

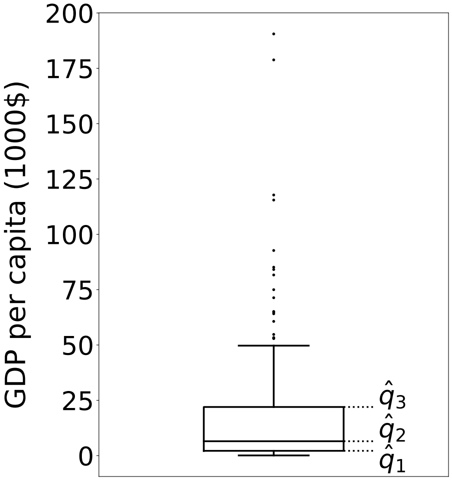
Height in US army



Box plot

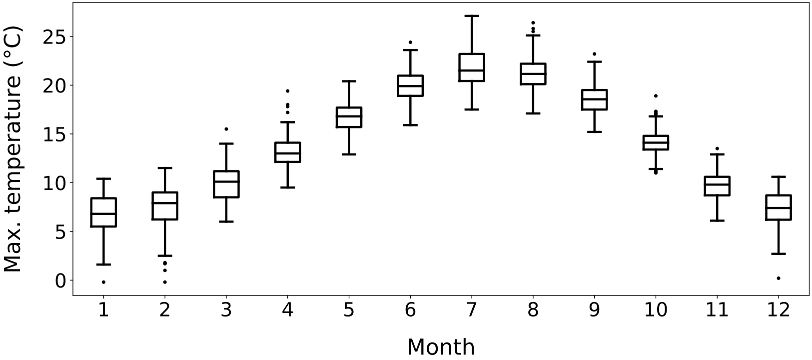


Box plot



GDP per capita (1000$)
200
175
150
125
100
75
50
25
0
$\hat{q}_3$
$\hat{q}_2$
$\hat{q}_1$

Weather in Oxford



What have we learned?

Definition of cumulative distribution function

Definition of quantiles

How to estimate them from data