Sensitivity of the Mean to Extreme Values

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

Discrete random variable

The mean of a discrete random variable $\tilde{a}$ with range A is

$$\mathrm{E}\left[\tilde{a}\right] := \sum_{a \in A} a \, p_{\tilde{a}}\left(a\right)$$

if the sum converges

Continuous random variable

The mean of a continuous random variable $\tilde{a}$ is

$$\mathrm{E}\left[\tilde{a}\right] := \int_{a=-\infty}^{\infty} a f_{\tilde{a}}\left(a\right) \, \mathrm{d}a$$

if the integral converges

Sample mean

The sample mean of $X := \{x_1, x_2, \ldots, x_n\}$ is

$$m(X) := \frac{\sum_{i=1}^{n} x_i}{n}$$

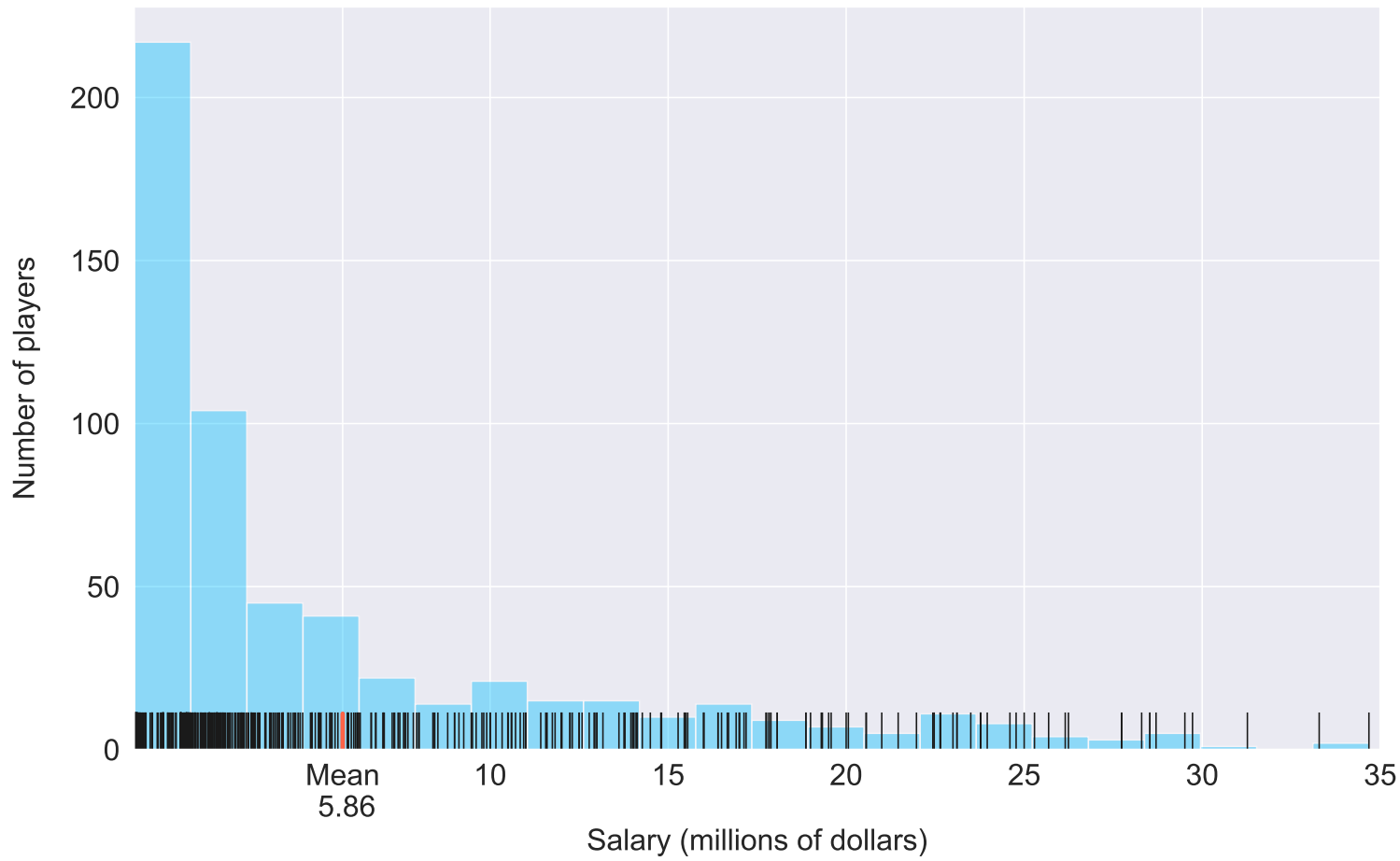
Question

Is the mean a typical value?

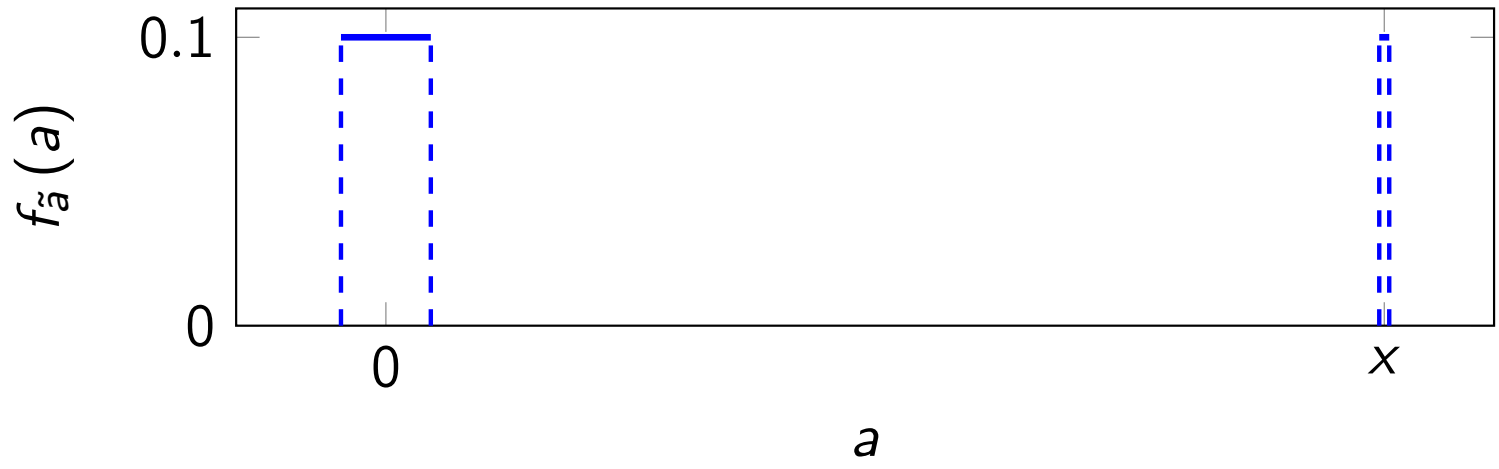
NBA salaries

How many earn more than mean?

Less than 1/3 of players (32.1%)



Extreme values



Random variable $\tilde{a}$ with uniform density in $[-4.5, 4.5]$ and $[x - 0.5, x + 0.5]$

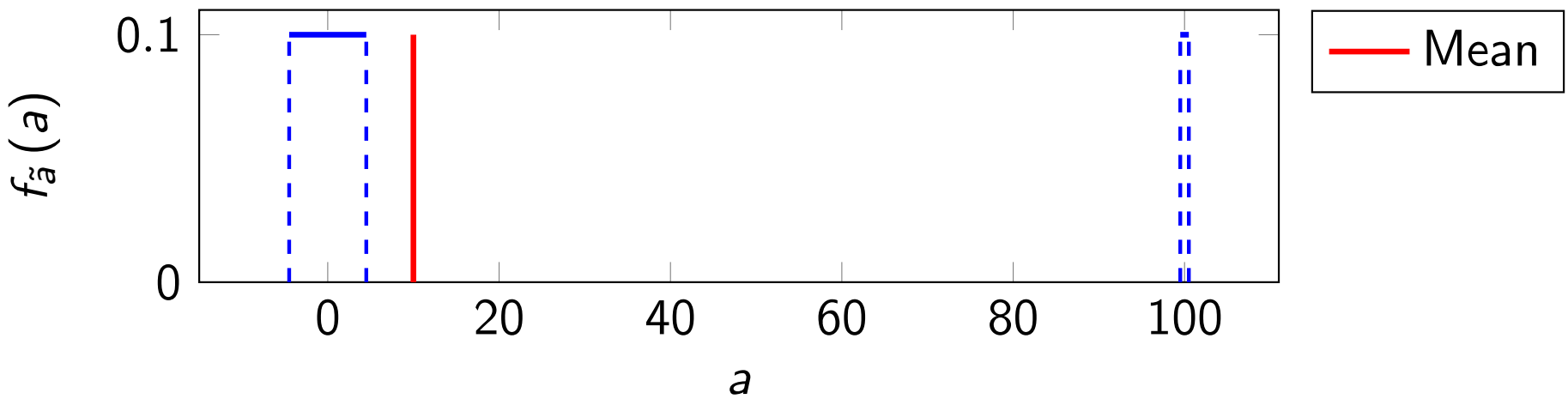
Extreme values

Random variable $\tilde{a}$ uniform in $[-4.5, 4.5]$ and $[x - 0.5, x + 0.5]$

Mean:

$$\begin{aligned}
\mathrm{E}[\tilde{a}] &= \int_{a=-4.5}^{4.5} a f_{\tilde{a}}(a)\, \mathrm{d}a + \int_{a=x-0.5}^{x+0.5} a f_{\tilde{a}}(a)\, \mathrm{d}a \\
&= \frac{1}{10} \frac{(4.5)^2 - (-4.5)^2}{2} + \frac{1}{10} \frac{(x+0.5)^2 - (x-0.5)^2}{2} \\
&= \frac{1}{10} \frac{x^2 + x + 0.25 - x^2 + x - 0.25}{2} \\
&= \frac{x}{10}
\end{aligned}$$

$x = 100$



Alternative characterization of typical value?

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

$$\mathrm{P}\left(\tilde{a} \leq q\right) = F_{\tilde{a}}\left(q\right) = \frac{1}{2}$$

Extreme values

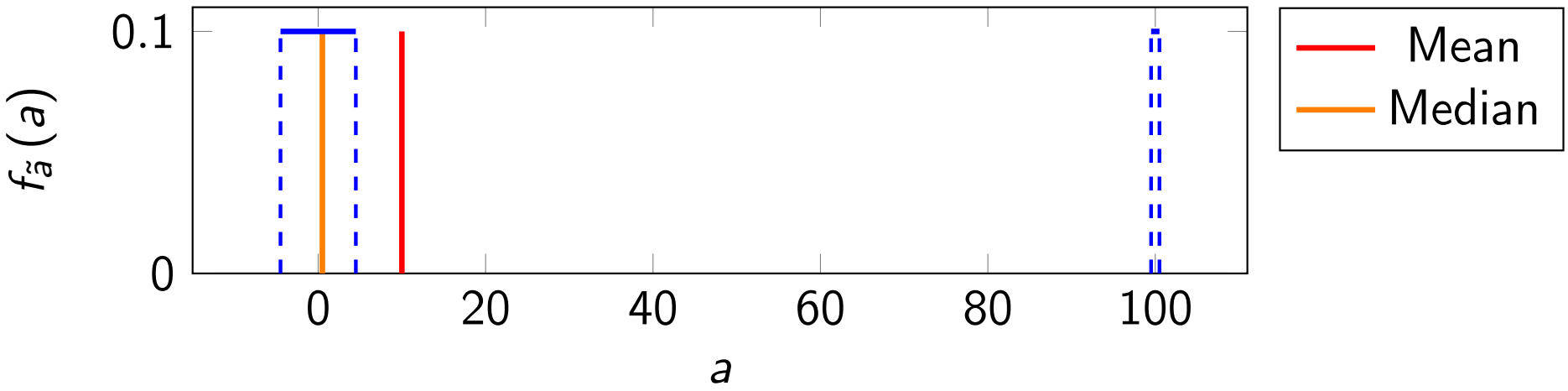
Random variable $\tilde{a}$ uniform in $[-4.5, 4.5]$ and $[x - 0.5, x + 0.5]$

$$F_{\tilde{a}}(q) = \int_{-4.5}^{q} f_{\tilde{a}}(a)\, \mathrm{d}a$$
$$= \frac{q + 4.5}{10} = \frac{1}{2}$$

Median: $q = 0.5$

No dependence on x

$x = 100$



Sample median

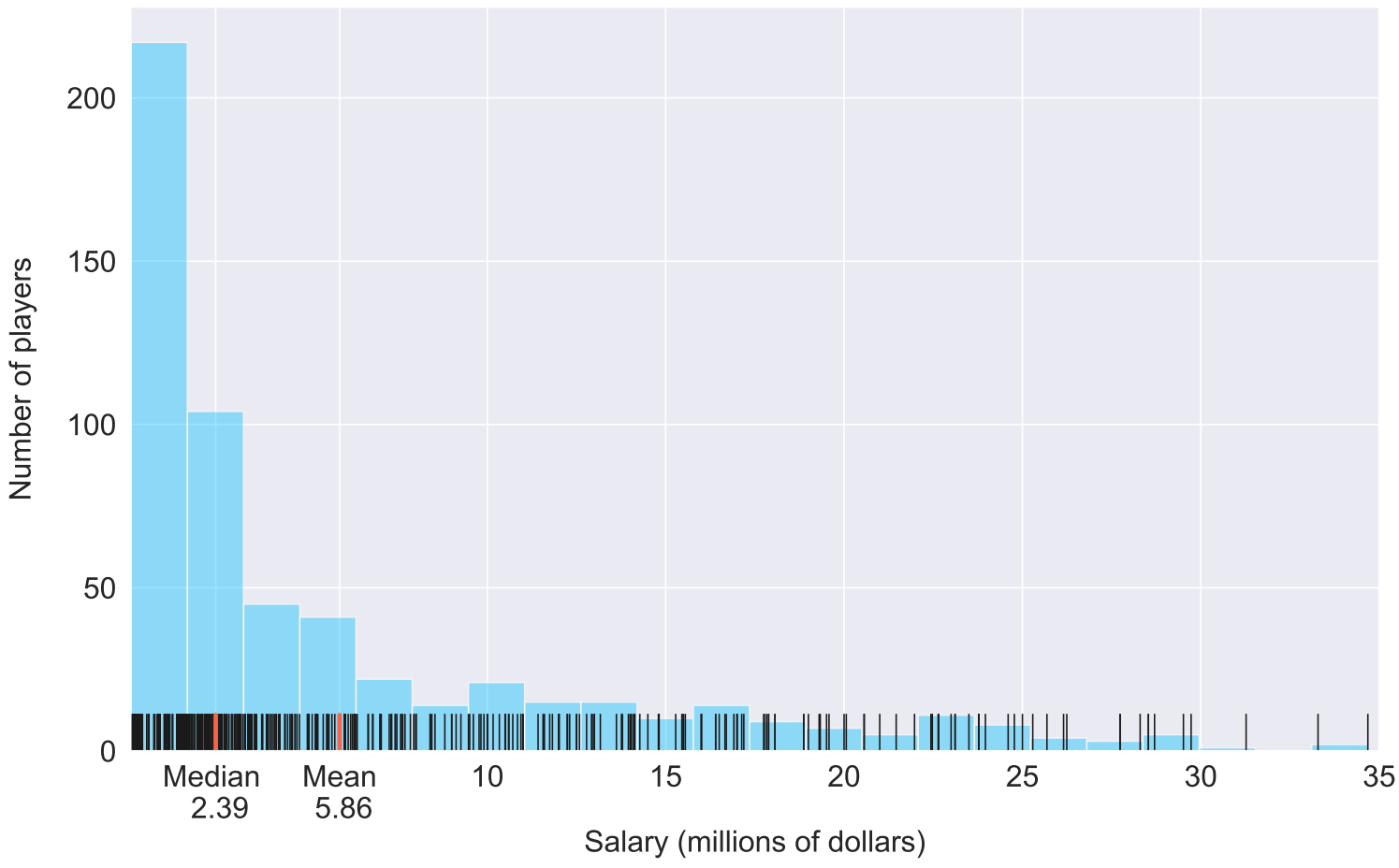
Dataset $X := \{x_1, x_2, \ldots, x_n\}$

The sample median $\hat{q}$ satisfies

$$\mathrm{P}_X(\tilde{a} \leq \hat{q}) = \frac{1}{2}$$

where P_X is the empirical probability of the data

NBA salaries



What have we learned?

Sensitivity of the mean to extreme values