The Law of Large Numbers Can Fail: When Not To Trust An Average

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

Law of large numbers

Let $\tilde{x}_1$, $\tilde{x}_2$, ..., be random variables with mean μ and variance σ^2

Sample mean

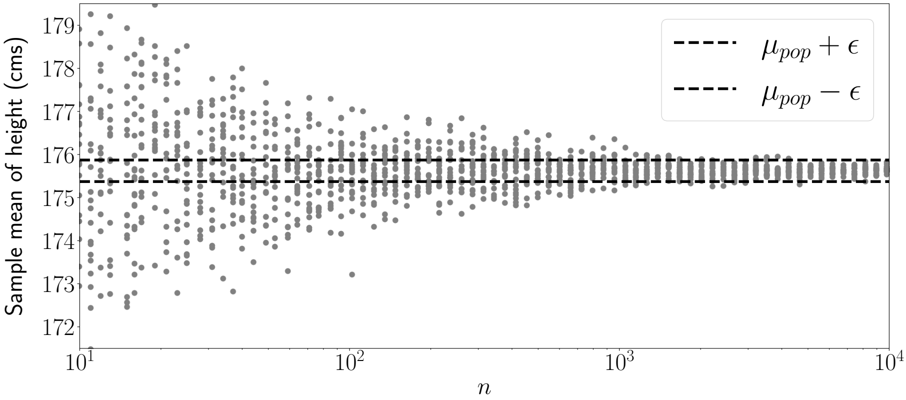
$$\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i$$

converges to μ in mean square and probability

Height data

$\mu_{\text{pop}} := 175.6$ cm, $\sigma_{\text{pop}} = 6.85$ cm

Total population $N := 4{,}082$



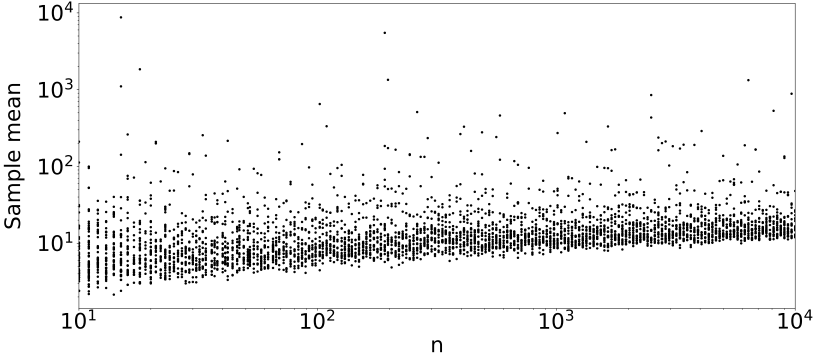
Bet

We flip a fair coin until it lands on heads

We receive a prize of 2^k (k: number of flips)

Mean of winnings?

Sample mean



Bet

Distribution of number of flips?

Geometric with parameter $\frac{1}{2}$

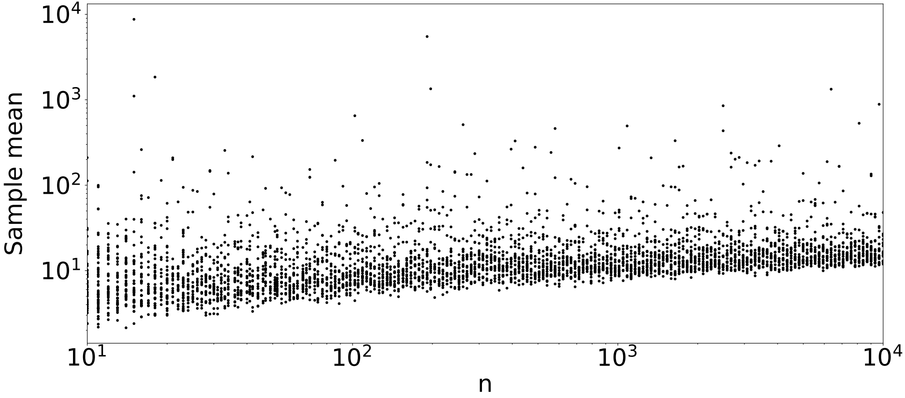
$$
\begin{aligned}
\mathrm{E}\left[\tilde{w}\right] &= \mathrm{E}\left[2^{\tilde{k}}\right] \\
&= \sum_{k=1}^{\infty} 2^k p_{\tilde{k}}\left(k\right) \\
&= \sum_{k=1}^{\infty} 2^k \cdot \frac{1}{2^k} \\
&= \infty
\end{aligned}
$$

Should we invest our life savings on this?

Half the time, winnings = 1 dollar!

Infinite mean

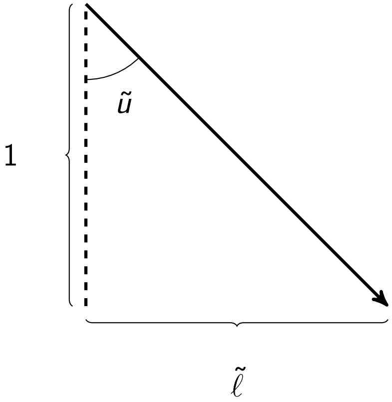
Sample mean of i.i.d. random variables



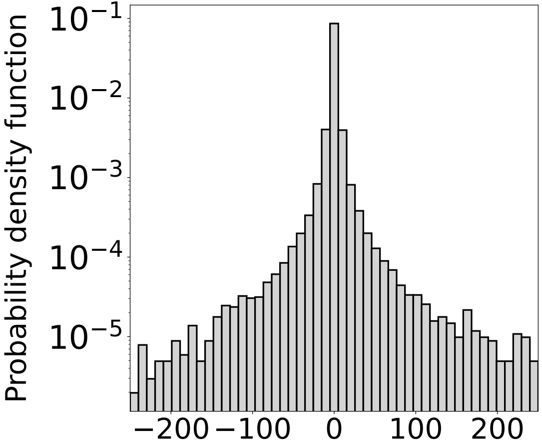
Physics experiment

$\tilde{u}$: Uniformly distributed in $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$

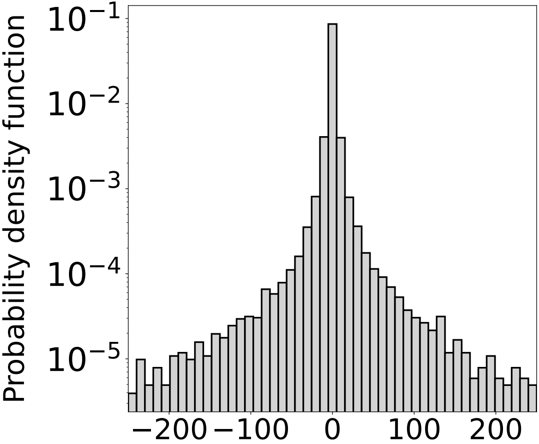
Mean of $\tilde{\ell}$?



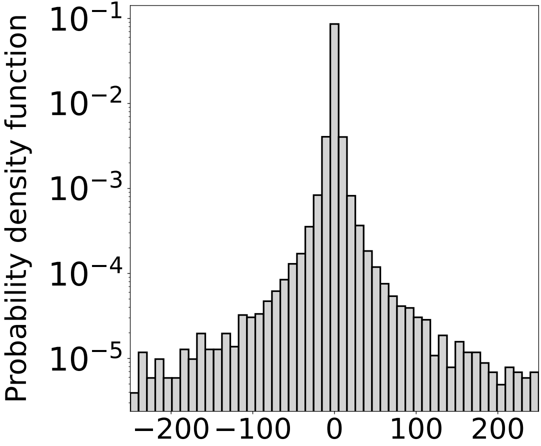
Sample mean $n = 100$



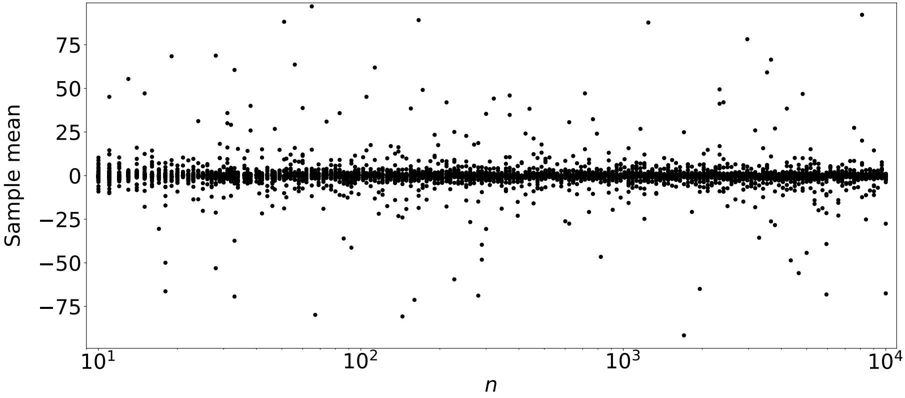
Sample mean n =1,000



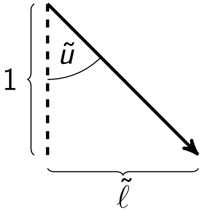
Sample mean $n = 10{,}000$



Sample mean



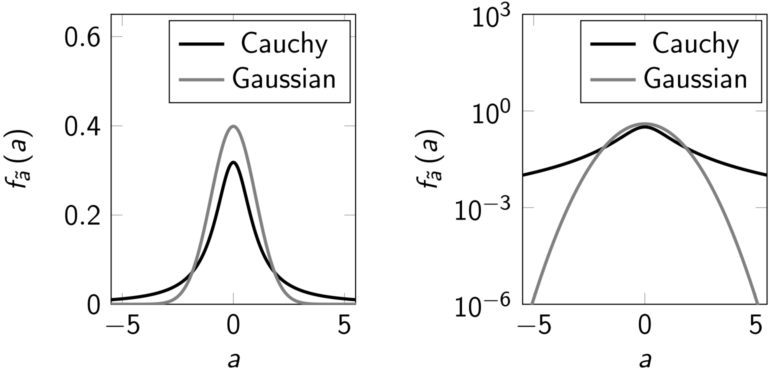
Physics experiment



$$
\begin{aligned}
F_{\tilde{\ell}}(\ell) &= \mathrm{P}(\tilde{\ell} \leq \ell) \\
&= \mathrm{P}(\tan \tilde{u} \leq \ell) \\
&= \mathrm{P}(\tilde{u} \leq \arctan \ell) \\
&= \frac{1}{\pi} \int_{-\pi/2}^{\arctan \ell} \mathrm{d}u = \frac{1}{2} + \frac{\arctan \ell}{\pi}
\end{aligned}
$$

$$
f_{\tilde{\ell}}(\ell) = \frac{1}{\pi(1 + \ell^2)}
$$

Cauchy random variable



Cauchy random variable

$$f_{\tilde{a}}(a) = \frac{1}{\pi(1+a^2)}$$

$$\mathrm{E}[\tilde{a}] = \int_{-\infty}^{\infty} \frac{a}{\pi(1+a^2)}\,\mathrm{d}x = \int_{0}^{\infty} \frac{a}{\pi(1+a^2)}\,\mathrm{d}a - \int_{0}^{\infty} \frac{a}{\pi(1+a^2)}\,\mathrm{d}a$$

By the change of variables $t = a^2$

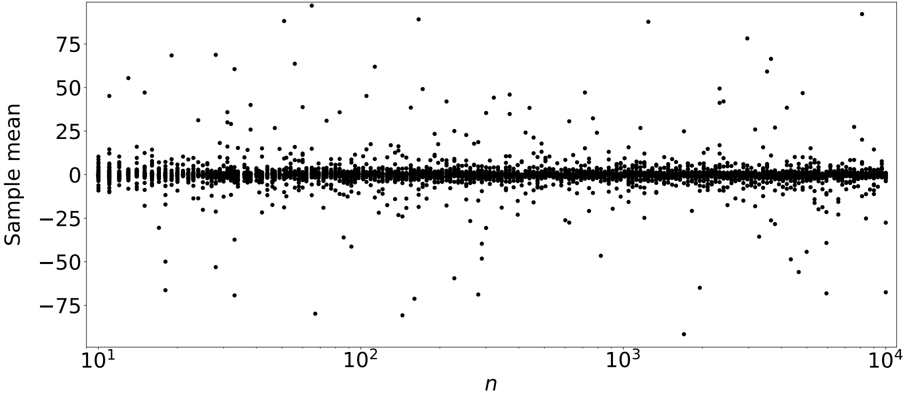
$$\int_{0}^{\infty} \frac{a}{\pi(1+a^2)}\,\mathrm{d}a = \int_{0}^{\infty} \frac{1}{2\pi(1+t)}\mathrm{d}t = \lim_{b\to\infty} \frac{\log(1+b)}{2\pi} = \infty$$

The mean does not exist!

Non-existent mean

Sounds like a mathematical curiosity

Important consequence: Sample mean does not converge!



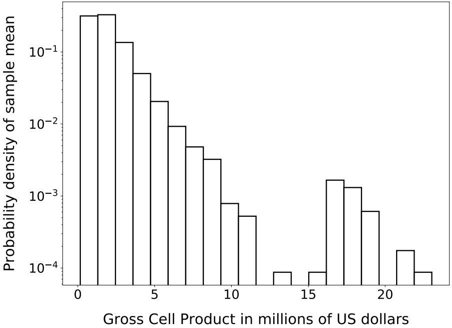
Local economic activity

Gross cell products of small regions

Population mean $\mu_{\text{pop}} = 2$ million dollars

Total population $N := 20{,}100$

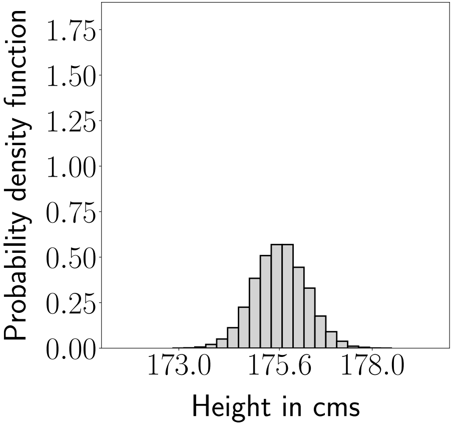
10^4 sample means of $n = 100$ random samples



Height data

Population mean $\mu_{\text{pop}} := 175.6$ cm

10^4 sample means of $n = 100$ random samples

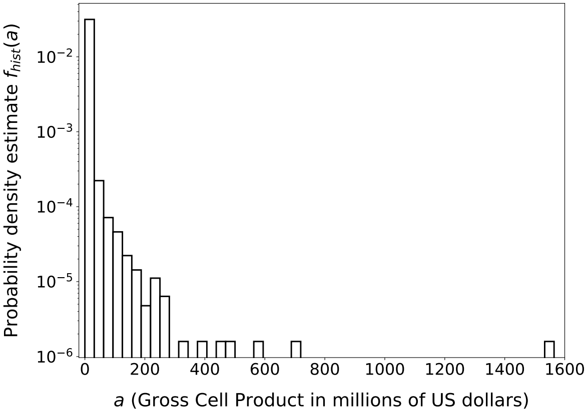


Population

Population mean $\mu_{\text{pop}} = 2$ million dollars

Extreme values shift the mean by $\frac{200}{n} = 2$ million

Population standard deviation $\sigma_{\text{op}} = 17.7$ million!



What have we learned

Law of large numbers does not always hold

Extreme values distort the sample mean