The Mean

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

Goals

Define an averaging operation for random variables

Motivation

Data: 3,4,3,4,6,3, . . .

Averaging is a reasonable way to compute *typical* value

$$\frac{3+4+3+4+\cdots}{n}$$

What is the average of a random variable?

Intuitive definition of probability

If we observe many samples from $\tilde{a}$

$$\mathrm{P}(\tilde{a} = a) = \frac{\text{number of data equal to } a}{\text{total}}$$

Discrete random variable

Data interpreted as samples from random variable $\tilde{a}$ with range A

$$\frac{3+4+3+4+\cdots}{n} = \sum_{a \in A} a \cdot \frac{\text{number of data equal to } a}{n}$$
$$\approx \sum_{a \in A} a \, p_{\tilde{a}}(a)$$

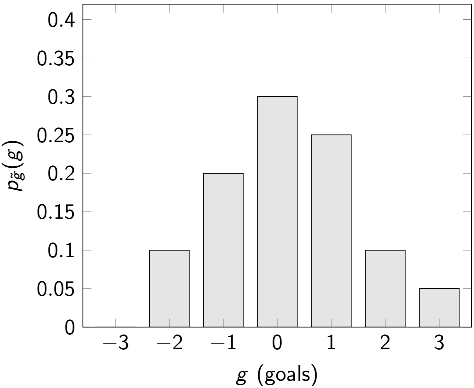
Mean of a discrete random variable

The mean, first moment or expected value 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

Goal difference



$$\begin{aligned}
\mathrm{E}[\tilde{g}] &= \sum_{g=-2}^{2} g\, p_{\tilde{g}}(g) \\
&= -2 \cdot 0.1 - 1 \cdot 0.2 + 0 \cdot 0.3 + 1 \cdot 0.25 + 2 \cdot 0.1 + 3 \cdot 0.05 \\
&= 0.2
\end{aligned}$$

Function of a random variable

Data: 3,4,3,4,6,3, ...

We are interested in a function of the data (e.g. their square)

Average of transformed values

$$\begin{aligned}\frac{h(3)+h(4)+h(3)+h(4)+\cdots}{n} &= \sum_{a\in A} h(a)\cdot\frac{\text{number of data equal to } a}{n} \\ &\approx \sum_{a\in A} h(a)\, p_{\tilde{a}}(a)\end{aligned}$$

Function of a random variable

The expected value of $h(\tilde{a})$, $h : \mathbb{R} \to \mathbb{R}$ is

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

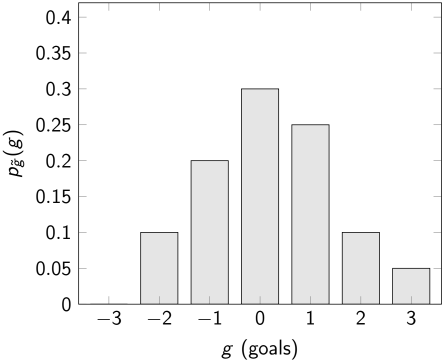
if $\tilde{a}$ is discrete and the sum converges

Converting goal difference to points

Points: $\tilde{x} := h(\tilde{g})$, where

$$h(g) := \begin{cases} 0 & \text{if } g < 0 \\ 1 & \text{if } g = 0 \\ 3 & \text{if } g > 0 \end{cases}$$

Goal difference



$$
\begin{aligned}
\mathrm{E}[\tilde{x}] &= \mathrm{E}\left[h(\tilde{g})\right] \\
&= \sum_{g=-2}^{2} h(g) p_{\tilde{g}}(g) \\
&= 0 \cdot 0.1 + 0 \cdot 0.2 + 1 \cdot 0.3 + 3 \cdot 0.25 + 3 \cdot 0.1 + 3 \cdot 0.05 \\
&= 1.5
\end{aligned}
$$

Multiple discrete random variables

Data: (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)

Function of the data $h(x, y)$

Average:

$$\frac{h(x_1, y_1) + h(x_2, y_2) + \cdots + h(x_n, y_n)}{n}$$
$$= \sum_{a \in A} \sum_{b \in B} h(a, b) \cdot \frac{\text{number of pairs } (x, y) \text{ for which } x = a \text{ and } y = b}{n}$$
$$\approx \sum_{a \in A} \sum_{b \in B} h(a, b) p_{\tilde{a}, \tilde{b}}(a, b)$$

Multiple discrete random variables

If $\tilde{a}$ (range: A) and $\tilde{b}$ (range: B) are discrete, the expected value of $h(\tilde{a}, \tilde{b})$ is

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

if the sum converges

Function of discrete random vector

If $\tilde{x}$ is a d-dimensional discrete random vector the expected value of $h(\tilde{x})$ of $\tilde{x}$ is

$$\mathrm{E}\left[h(\tilde{x})\right] := \sum_{x[1]\in X_1} \sum_{x[2]\in X_2} \cdots \sum_{x[d]\in X_d} h\left(x\right) p_{\tilde{x}}\left(x\right)$$

if the sum converges

Cats and dogs

		Cats			
		0	**1**	**2**	**3**
Dogs	**0**	0.35	0.15	0.1	0.05
	1	0.2	0.05	0.03	0
	2	0.05	0.02	0	0

$$\mathrm{E}[\tilde{c} + \tilde{d}]$$

$$= \sum_{c=0}^{3} \sum_{d=0}^{2} (c + d) p_{\tilde{c},\tilde{d}}(c, d)$$

$$= 0.15 + 2 \cdot 0.1 + 3 \cdot 0.05 + 0.2 + 2 \cdot 0.05 + 3 \cdot 0.03 + 2 \cdot 0.05 + 3 \cdot 0.02$$

$$= 1.05$$

Continuous quantity

Data: 3.67, 4.91, 3.02, 4.83, . . .

Averaging is still a reasonable way to compute *typical* value

$$\frac{3.67 + 4.91 + 3.02 + \cdots}{n}$$

What is the average of a continuous random variable?

Continuous random variables

Grid with step size ϵ

$a_m := m\epsilon$ where $m \in \mathbb{Z}$

As $\epsilon \to 0$ for large n

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} x_i &\approx \sum_{m\in\mathbb{Z}} \frac{a_m \cdot \text{number of data between } a_m - \epsilon \text{ and } a_m}{n} \\
&\approx \sum_{m\in\mathbb{Z}} a_m \mathrm{P}(a_m - \epsilon \leq \tilde{a} \leq a_m) \\
&\approx \sum_{m\in\mathbb{Z}} a_m f_{\tilde{a}}(a_m)\epsilon \\
&= \int_{a\in\mathbb{R}} a f_{\tilde{a}}(a)\,\mathrm{d}a \qquad \text{when } \epsilon \to 0
\end{aligned}
$$

Continuous random variable

The mean, first moment or expected value of a continuous random variable $\tilde{a}$ is

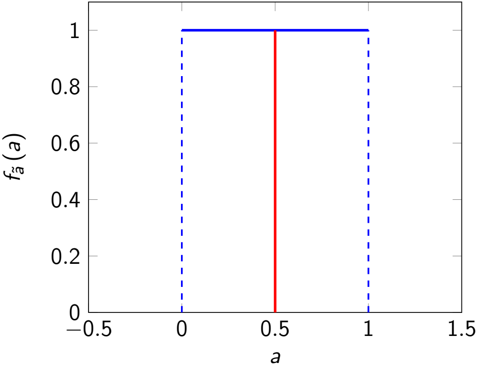
$$\mathrm{E}\left[\tilde{a}\right] := \int_{a=-\infty}^{\infty} a f_{\tilde{a}}\left(a\right)\, \mathrm{d}a$$

if the integral converges

Uniform random variable in $[a, b]$

$$
\begin{aligned}
\mathrm{E}\left[\tilde{u}\right] &= \int_{u=-\infty}^{\infty} u f_{\tilde{a}}\left(u\right)\,\mathrm{d}u \\
&= \int_{u=a}^{b} \frac{u}{b-a}\,\mathrm{d}u \\
&= \frac{b^2 - a^2}{2\left(b-a\right)} \\
&= \frac{a+b}{2}
\end{aligned}
$$

Uniform random variable in $[0, 1]$



Function of a random variable

The mean of $h(\tilde{a})$, $h : \mathbb{R} \to \mathbb{R}$ is

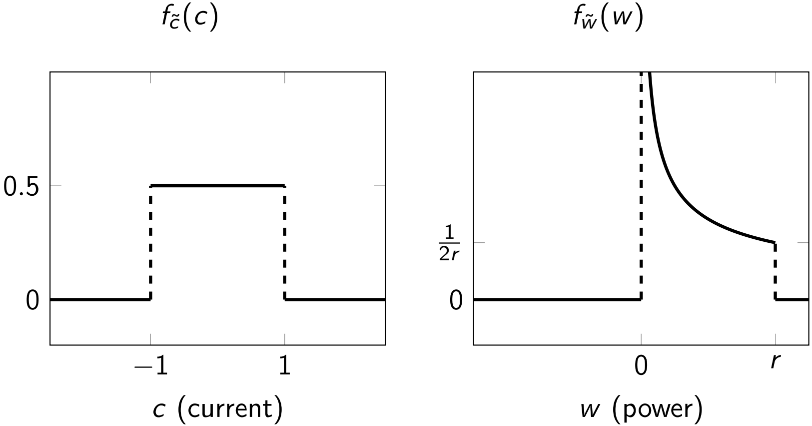
$$\mathrm{E}\left[h(\tilde{a})\right] := \int_{a=-\infty}^{\infty} h(a)\, f_{\tilde{a}}(a)\, \mathrm{d}a$$

if $\tilde{a}$ is continuous and the integral converges

Circuit

Current $\tilde{c}$ with pdf $f_{\tilde{c}}$

Power $\tilde{w} = r\tilde{c}^2$



Circuit

$$
\begin{aligned}
\mathrm{E}[\tilde{w}] &= \mathrm{E}\left[r\tilde{c}^2\right] \\
&= \int_{c=-1}^{1} \frac{rc^2}{2}\,\mathrm{d}c \\
&= \frac{r}{3}
\end{aligned}
$$

Multiple random variables

If $\tilde{a}$, and $\tilde{b}$ are continuous, the expected value of $h(\tilde{a}, \tilde{b})$ is

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

if the integral converges

Function of random vector

If $\tilde{x}$ is a d-dimensional continuous random vector the expected value of $h(\tilde{x})$ is

$$\mathrm{E}\left[h(\tilde{x})\right] := \int_{x \in \mathbb{R}^d} h\left(x\right) f_{\tilde{x}}\left(x\right) \, \mathrm{d}x$$

if the integral converges

Sugar

You grab an amount of sugar uniformly distributed between 0 and 1 kg

You spill an amount that is uniformly distributed between 0 and the quantity that you grabbed

Expected amount of spilled sugar?

Sugar

Distribution of sugar $\tilde{g}$ you grab? Uniform in $[0, 1]$

Distribution of sugar $\tilde{s}$ you spill? Uniform in $[0, g]$ given $\tilde{g} = g$

Mean of $\tilde{s}$?

Sugar

$$
\begin{aligned}
\mathrm{E}[\tilde{s}] &= \int_g \int_s s\, f_{\tilde{g},\tilde{s}}(g, s)\, \mathrm{d}g\, \mathrm{d}s \\
&= \int_g \int_s s\, f_{\tilde{g}}(g) f_{\tilde{s}\,|\,\tilde{g}}(s \,|\, g)\, \mathrm{d}g\, \mathrm{d}s \\
&= \int_{g=0}^{1} \int_{s=0}^{g} \frac{s}{g}\, \mathrm{d}g\, \mathrm{d}s \\
&= \int_{g=0}^{1} \frac{g}{2}\, \mathrm{d}g \\
&= \frac{1}{4}
\end{aligned}
$$

Discrete and continuous quantities

If $\tilde{c}$ is continuous and $\tilde{d}$ is discrete with range D, the mean of $h\left(\tilde{c}, \tilde{d}\right)$ is

$$\begin{aligned}
\mathrm{E}\left[h(\tilde{c}, \tilde{d})\right] &:= \int_{c=-\infty}^{\infty} \sum_{d \in D} h(c, d) f_{\tilde{c}}(c)\, p_{\tilde{d}\,|\,\tilde{c}}(d \,|\, c)\, \mathrm{d}c \\
&= \sum_{d \in D} \int_{c=-\infty}^{\infty} h(c, d) p_{\tilde{d}}(d)\, f_{\tilde{c}\,|\,\tilde{d}}(c \,|\, d)\, \mathrm{d}c,
\end{aligned}$$

if the sum and integral converge

Bayesian coin flip

We flip a coin but don't know the probability of heads $\tilde{\theta}$

We assume $\tilde{\theta}$ is uniform in [0,1]

Mean of the coin flip (heads = 1, tails = 0)?

$$\begin{aligned}
\mathrm{E}\left[\tilde{a}\right] &= \int_{c=-\infty}^{\infty} \sum_{a=0}^{1} a f_{\tilde{\theta}}\left(\theta\right) p_{\tilde{a}\,|\,\tilde{\theta}}\left(a \,|\, \theta\right) \,\mathrm{d}\theta \\
&= \int_{0}^{1} \theta \,\mathrm{d}\theta \\
&= \frac{1}{2}
\end{aligned}$$

How do we estimate the mean from data?

We average

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

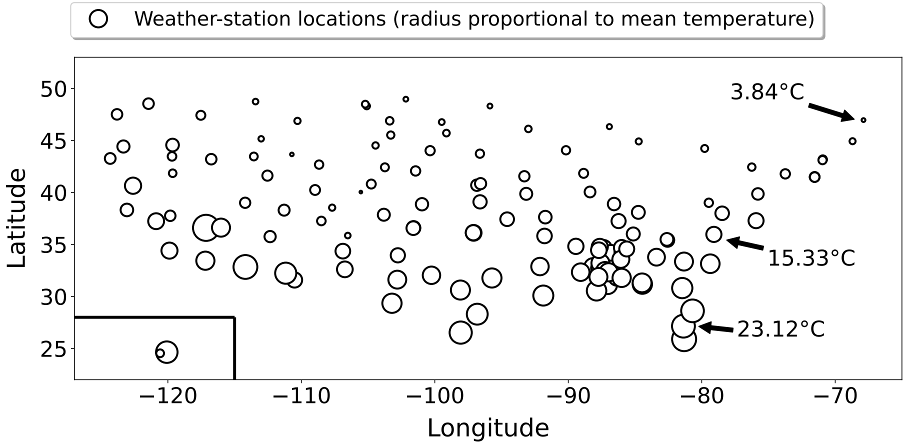
$$m(X) := \frac{\sum_{i=1}^{n} x_i}{n}$$

Same for discrete and continuous variables

If data are i.i.d. samples from distribution with finite variance, sample mean converges to the mean as $n \to \infty$ (law of large numbers)

Temperature dataset

Hourly temperatures at 134 weather stations in the US



What have we learned?

Definition of the mean, as an averaging operation for random variables