Properties of 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

Describe two important properties of the mean

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

Mean cost of a latte

Price per kg of coffee:

Random variable $\tilde{c}$ with mean 2.5

Price per gallon of milk:

Random variable $\tilde{m}$ with mean 3.5

$\tilde{c}$ and $\tilde{m}$ are *not* independent

A latte has 0.02 kg of coffee and 0.1 gallons of milk

Mean cost of a latte?

Mean cost of a latte

$$
\begin{aligned}
\mathrm{E}[\tilde{\ell}] &= \mathrm{E}(0.02\tilde{c} + 0.1\tilde{m}) \\
&= \int_{c\in\mathbb{R}} \int_{m\in\mathbb{R}} (0.02c + 0.1m) f_{\tilde{c},\tilde{m}}(c,m)\,\mathrm{d}c\,\mathrm{d}m \\
&= 0.02 \int_{c\in\mathbb{R}} \int_{m\in\mathbb{R}} c f_{\tilde{c},\tilde{m}}(c,m)\,\mathrm{d}c\,\mathrm{d}m \\
&\quad + 0.1 \int_{c\in\mathbb{R}} \int_{m\in\mathbb{R}} m f_{\tilde{c},\tilde{m}}(c,m)\,\mathrm{d}c\,\mathrm{d}m \\
&= 0.02 \int_{c\in\mathbb{R}} c f_{\tilde{c}}(c)\,\mathrm{d}c + 0.1 \int_{m\in\mathbb{R}} m f_{\tilde{m}}(m)\,\mathrm{d}m \\
&= 0.02\,\mathrm{E}[\tilde{c}] + 0.1\,\mathrm{E}[\tilde{m}] \\
&= 0.4 \qquad \text{(40 cents)}
\end{aligned}
$$

Linearity of expectation

For any constants $c_1, c_2 \in \mathbb{R}$, any functions $h_1, h_2 : \mathbb{R}^n \to \mathbb{R}$ and any continuous or discrete random variables $\tilde{a}$ and $\tilde{b}$

$$\mathrm{E}\left[c_1\, h_1(\tilde{a}, \tilde{b}) + c_2\, h_2(\tilde{a}, \tilde{b})\right] = c_1\, \mathrm{E}\left[h_1(\tilde{a}, \tilde{b})\right] + c_2\, \mathrm{E}\left[h_2(\tilde{a}, \tilde{b})\right]$$

Follows from linearity of sums and integrals

Binomial random variable

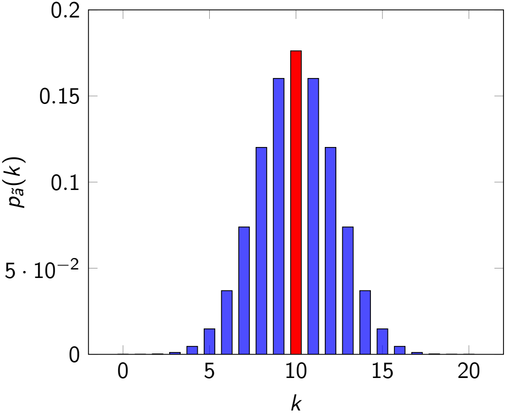
Mean of binomial random variable $\tilde{a}$ with parameters n and θ?

Sum of n independent Bernoulli random variables with parameter θ

$$\begin{aligned}
\mathrm{E}\left[\tilde{a}\right] &= \mathrm{E}\left[\sum_{k=1}^{n} \tilde{b}_k\right] \\
&= \sum_{k=1}^{n} \mathrm{E}(\tilde{b}_k) \\
&= n\theta
\end{aligned}$$

Do we need independence?

Binomial, $n := 20$ $\theta := 0.5$



Independent random variables

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

Same for discrete random variables

Restaurant

Goal: Estimate expected revenue

Mean number of customers: 50

Mean amount spent per customer: 40 dollars

Is mean revenue necessarily 2000? No!

Restaurant

Each night is busy or calm with probability $\frac{1}{2}$

Busy nights: 80 customers who spend 60 dollars each

Calm nights: 20 customers who spend 20 dollars each

Mean number of customers: 50

Mean amount spent per customer: 40 dollars

$$\begin{aligned} \mathrm{E}(\tilde{c}\tilde{a}) &= \frac{80 \cdot 60}{2} + \frac{20 \cdot 20}{2} \\ &= 2600 \neq 2000 \end{aligned}$$

What have we learned?

The mean is linear

The mean of the product of independent random variables is the product of their means