ONLINE MASTERS IN **DATA SCIENCE**

DSC 257R - UNSUPERVISED LEARNING

LOCATION & SCALE

SANJOY DASGUPTA, PROFESSOR

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

Mean

The *mean (expected value)* of random variable $\mathcal{X}$, or equivalently of its distribution, is

$$\mathbb{E}(\mathcal{X}) = \begin{cases} \sum_x x \Pr(\mathcal{X} = x) & \text{if } \mathcal{X} \text{ is discrete} \\ \int x\, p(x)\, dx & \text{if } \mathcal{X} \text{ is continuous with density } p(x) \end{cases}$$

The *empirical mean* of a set of data points $x_1, \ldots, x_n$:

$$\frac{1}{n} \sum_{i=1}^{n} x_i$$

What is the relationship between these two definitions?

Median

Two ways of summarizing a set of numbers by a single number.

- The **(empirical) mean**
- The **(empirical) median**: the number in the middle, if you sort them

Find the median of the following sets of numbers:

- $10, -20, 100, 20, 50$

- $50, 100, 60, 90, 20, 10$

How can we define the median of a random variable X?

Mean vs Median

In a certain neighborhood, there are 100 houses.

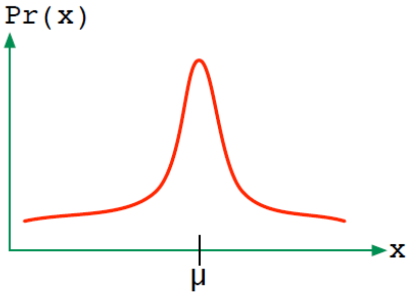
- 10 of the houses cost $100K
- 60 of the houses cost $200K
- 29 of the houses cost $300K
- one house costs $100M

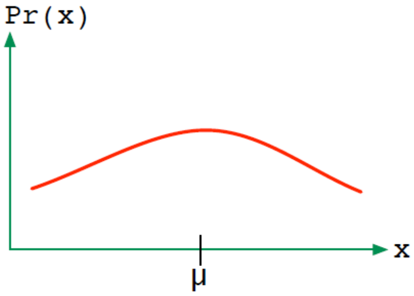
What is the mean house cost, roughly?

What is the median cost?

Variance & Standard Deviation

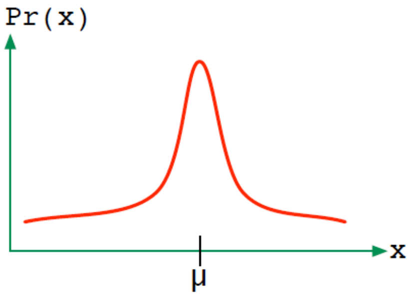
We can summarize a random variable X by its mean μ (or median). But this doesn't capture the ***spread*** of X.

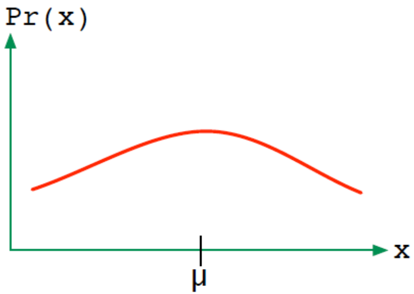




Variance & Standard Deviation

We can summarize a random variable $\mathcal{X}$ by its mean μ (or median). But this doesn't capture the ***spread*** of $\mathcal{X}$.





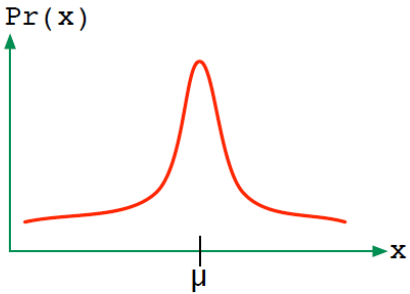
The **variance** of $\mathcal{X}$ is defined as

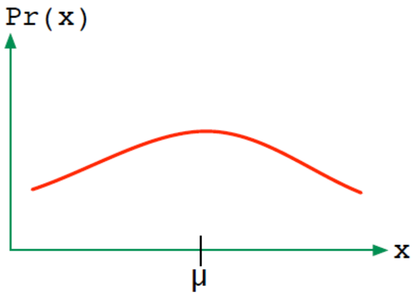
$$\mathrm{var}(\mathcal{X}) = \mathbb{E}(\mathcal{X} - \mu)^2 = \mathbb{E}(\mathcal{X}^2) - \mu^2,$$

where $\mu = \mathbb{E}(\mathcal{X})$. It is always ≥ 0.

Variance & Standard Deviation

We can summarize a random variable $\mathcal{X}$ by its mean μ (or median). But this doesn't capture the ***spread*** of $\mathcal{X}$.





The **variance** of $\mathcal{X}$ is defined as

$$\mathrm{var}(\mathcal{X}) = \mathbb{E}(\mathcal{X} - \mu)^2 = \mathbb{E}(\mathcal{X}^2) - \mu^2,$$

where $\mu = \mathbb{E}(\mathcal{X})$. It is always ≥ 0.

The **standard deviation** of $\mathcal{X}$ is $std(\mathcal{X}) = \sqrt{\mathrm{var}(\mathcal{X})}$. It is, *roughly*, the average amount by which $\mathcal{X}$ differs from its mean.