INFO 2950   2/12/2019

DISTRIBUTIONS AND SAMPLES

A probability distribution is a function that maps events to probabilities.

ex: coin=H, die=6, points=10

The set of events for which a distribution has P(event) > 0 is its support.

ex: {H,T}, [1,2,...,6], [0,1,2,...]

From the distribution $P(X)$, we can:

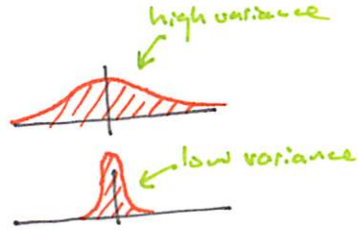
1. calculate $\mathbb{E}[X]$ the "mean"

$$\mu = \sum_x x P(x)$$

sum over full support

2. calculate variance(X)

$$\sigma^2 = \sum_x (x-\mu)^2 P(x)$$



We can also generate samples.

i'th element of sample $X_i \sim P(X)$ "distributed as"

X_i is x[i]
(math) (Python)

The size of the sample array is N.

From a sample array we can calculate:

!! I find that I refer to the array $[X_1, X_2, X_3, \ldots, X_N]$ and an element of that array as a "sample". !! Hopefully it will be clear from context which I mean.

1. sample mean

"X bar" $\bar{X} = \frac{1}{N}\sum_i X_i$

sum is over each sampled ~~value~~ value, not over full support, may contain duplicates!

2. sample variance

$$S^2 = \frac{1}{N-1}\sum_i (X_i - \bar{X})^2$$

As N increases, the empirical distribution, $\frac{\#\text{ of times } X_i = \text{event}}{N}$ becomes close to $P(X)$.

As N increases, $\bar{X}$ gets very close to μ (usually, not for all P(x)!)

" " " $\frac{1}{N}\sum(X_i - \mu)^2$ gets very close to σ^2

BUT As N increases, $\frac{1}{N}\sum(X_i - \bar{X})^2$ is always a tiny bit too small.

thats why we use $\frac{1}{N-1}$! it accounts for the bias using $\bar{X}$ instead of μ, which we usually don't know.