Sums of Random Variables

- We are often interested in the behavior of a sum of random variables $S_n = X_1 + X_2 + \cdots + X_n$.

- For instance, the sample mean $\hat{\mu}_x$ or M_n

$$\boxed{\hat{\mu}_x = M_n = \frac{1}{n}\sum_{i=1}^{n} X_i}$$

is frequently to estimate the mean from data.

- How many samples n are needed to obtain a good estimate?

- Ideally, we could answer this question precisely using the PMF (or PDF) of a sum of random variables.

- Unfortunately, calculating the exact PMF (or PDF) is difficult, even in simple scenarios.

- As a starting point, we will see how to calculate the mean and variance of a sum of random variables $S_n = X_1 + X_2 + \cdots + X_n$.

$$\mathbb{E}[S_n] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i]$$ using Linearity of Expectation

$$\begin{aligned}
\mathrm{Var}[S_n] &= \mathbb{E}\left[(S_n - \mathbb{E}[S_n])^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}[X_i]\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} (X_i - \mathbb{E}[X_i])\right)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} (X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}\left[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])\right] \\
&= \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}[X_i, X_j]}_{\text{Only requires pairwise second-order statistics.}}
\end{aligned}$$

$$\left(\sum_{i=1}^{n} a_i\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j$$

Linearity of Expectation

- To further simplify our calculations, we sometimes make additional assumptions on $X_1, X_2, \ldots, X_n$.

- The random variables $X_1, X_2, \ldots, X_n$ are said to be independent and identically distributed (i.i.d.) if they are independent and have the same underlying marginal PMF, $P_{X_i}(x_i) = P_X(x_i)$, or PDF, $f_{X_i}(x_i) = f_X(x_i)$ for $i = 1, 2, \ldots, n$,

Discrete Case: $P_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = P_X(x_1)P_X(x_2)\cdots P_X(x_n) = \prod_{i=1}^{n} P_X(x_i)$

Continuous Case: $f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f_X(x_1)f_X(x_2)\cdots f_X(x_n) = \prod_{i=1}^{n} f_X(x_i)$

- Ex: $X_1, X_2, \ldots, X_n$ are i.i.d. Bernoulli(p), $P_X(x) = p^x(1-p)^{1-x}$

$$P_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

- Ex: $X_1, X_2, \ldots, X_n$ are i.i.d. Gaussian(μ, σ^2), $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

- The mean and variance of a sum $S_n = \sum_{i=1}^{n} X_i$ of i.i.d. random variables $X_1, \dots, X_n$ are

$$\boxed{\mathbb{E}[S_n] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = n\,\underbrace{\mathbb{E}[X]} \qquad \mathrm{Var}[S_n] = \mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = n\,\underbrace{\mathrm{Var}[X]}}$$

Compute using the marginal PMF $P_X(x)$ or PDF $f_X(x)$.

Why? $$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] \overset{\text{Identically Distributed}}{=} \sum_{i=1}^{n} \mathbb{E}[X] = n\,\mathbb{E}[X]$$

$$\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}[X_i, X_j] = \sum_{i=1}^{n}\Big(\mathrm{Cov}[X_i, X_i] + \sum_{j \neq i} \underbrace{\cancel{\mathrm{Cov}[X_i, X_j]}}_{0 \text{ independent}}\Big)$$

$\mathrm{Cov}[X,X] = \mathrm{Var}[X]$

$$= \sum_{i=1}^{n} \mathrm{Var}[X_i] = n\,\mathrm{Var}[X]$$

Identically Distributed

- The mean and variance of the sample mean $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ of i.i.d. random variables $X_1, \dots, X_n$ are

$$\boxed{\mathbb{E}[M_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \mathbb{E}[X] \qquad \mathrm{Var}[M_n] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \underbrace{\frac{1}{n}\mathrm{Var}[X]}}$$

Goes to 0 as n increases.

Why? $$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\cdot n\,\mathbb{E}[X]$$

$$\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\cdot n\,\mathrm{Var}[X]$$

$\mathrm{Var}[aX] = a^2\,\mathrm{Var}[X]$

- The sample variance $\hat{\sigma}_X^2$ or V_n is frequently used to estimate the variance from data

$$\boxed{\hat{\sigma}_X^2 = V_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n)^2 \quad \text{where} \quad M_n = \frac{1}{n}\sum_{i=1}^{n} X_i}$$

- Why do we multiply by $\frac{1}{n-1}$ instead of $\frac{1}{n}$?

→ Consider $U_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - M_n)^2 \underset{\uparrow}{=} \frac{1}{n}\sum_{i=1}^{n} X_i^2 - M_n^2$.

can be shown

→ Calculate $\mathbb{E}[U_n]$. Is it $\mathrm{Var}[X]$?

$$\mathbb{E}[U_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - M_n^2\right]$$

Linearity of Expectation

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \mathbb{E}[M_n^2]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{Var}[X] + (\mathbb{E}[X])^2\right) - \mathrm{Var}[M_n] - (\mathbb{E}[M_n])^2$$

$$= \mathrm{Var}[X] + \cancel{(\mathbb{E}[X])^2} - \frac{1}{n}\mathrm{Var}[X] - \cancel{(\mathbb{E}[X])^2}$$

$$= \frac{n-1}{n}\mathrm{Var}[X]$$

This is a biased estimator.

→ Using a $\frac{1}{n-1}$ factor corrects this issue.