Lecture 3, Jan 15, 2024

Sum of Random Variables

- Let $S_n = \sum_{i=1}^{n} X_i$
- We can show that $E[S_n] = E[X_1 + \cdots + X_n] = \sum_{i=1}^{n} E[X_i]$
  - Note that although $E[S_n]$ is on the joint PDF of all the X random variables, $E[X_i]$ is on the marginal only, i.e. f_{X_i}
  - The expected value of a sum is always the sum of the expected values in all cases
- For variance: $\text{Var}[S_n] = E[(S_n - E[S_n])^2]$

$$\begin{aligned}
&= E\left[\left(\sum_{i=1}^{n} (X_i - m_{X_i})\right)^2\right] \\
&= E\left[\left(\sum_{i=1}^{n}(X_i - m_{X_i})\right)\left(\sum_{j=1}^{n}(X_j - m_{X_j})\right)\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \text{Cov}(X_i, X_j) \\
&= \sum_{i=1}^{n} \text{Var}[X_i] + \sum_{i \neq j}\sum_{j} \text{Cov}(X_i, X_j)
\end{aligned}$$

  - If all X_i, X_j are pairwise uncorrelated, then $\text{Var}[S_n] = \sum_{i=1}^{n} \text{Var}[X_i]$
  - But in general, the variance of a sum of RVs is not the sum of the variances
- Suppose that the Xs are *independent and identically distributed* (IID)
  - This means $f_{\boldsymbol{X}}(x_1, \dots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n) = f_X(x_1) \cdots f_X(x_n) = \prod_{i=1}^{n} f_X(x_i)$
  - It follows that all the Xs will have the same mean m and variance σ^2
  - Therefore $E[S_n] = nm, \text{Var}[S_n] = n\sigma^2$
- Let the *sample mean* be $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n} S_n$
  - Assuming IID:
    * $E[M_n] = E\left[\frac{1}{n} S_n\right] = \frac{1}{n} E[S_n] = m$
    * $\text{Var}[M_n] = \text{Var}\left[\frac{1}{n} S_n\right] = \frac{1}{n^2}\text{Var}[S_n] = \frac{\sigma^2}{n}$
  - With increasing n, the expected value is unchanged but the variance decreases; this means to estimate $E[X]$, we can repeat the same experiment and take the mean to get a smaller variance in our results
- To formalize this, we can apply Chebyshev's inequality to the mean
  - $P[|X - m_X| \geq \epsilon] \leq \frac{\sigma_X^2}{\epsilon^2}$
  - Applied to the sample mean: $P[|M_n - E[X]| \geq \epsilon] \leq \frac{\text{Var}[M_n]}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} = 1 - \delta$
  - Given any error tolerance ϵ and probability $1 - \delta$, we can always select n such that the probability of M_n being within the tolerance of the true mean is $1 - \delta$ or greater
  - This is also known as *convergence in probability*

Theorem

Chebyshev's Inequality:

$$P[|X - m_X| \geq \epsilon] \leq \frac{\sigma_X^2}{\epsilon^2}$$

Alternatively stated as

$$P[|X - m_X| \geq k\sigma] \leq \frac{1}{k^2}$$

Theorem

Weak Law of Large Numbers:

$$\lim_{n\to\infty} P[|M_n - E[X]| < \epsilon] = 1$$

That is, as the sample size N increases, the probability of the sample mean being within ϵ of the true mean approaches 1, where ϵ is any arbitrarily small positive number.

Strong Law of Large Numbers: Given IID X_i with finite mean,

$$P\left[\lim_{n\to\infty} |M_n - E[X]| < \epsilon\right] = 1$$

- SLLN asserts a much stronger form of convergence to $E[X]$
  - Notice that for SLLN the limit is outside the probability
  - The weak law states that for a certain value of n, most of the observed values of M_n will be close to $E[X]$
    - WLLN does not address what happens to the sample mean for a specific sequence of random variables
  - The strong law states that every sequence of sample mean calculations will eventually approach and stay close to $E[X]$
- Consider an event A and suppose we want to find $p = P[A]$
  - Let the *indicator function* for A be $I = \begin{cases} 1 & s \in A \\ 0 & s \notin A \end{cases}$
    - Note that $E[I] = 1 \cdot P[A] + 0 \cdot (1 - P[A]) = P[A] = p$
    - $\text{Var}[I] = E[(I - E[I])^2] = E[(I - p)^2] = (1 - p)^2 p + (-p)^2(1 - p) = p(1 - p)$
  - Repeat the experiment n times so we have $S_n = I_1 + I_2 + \cdots + I_n$ equal to the number of times that A occurred
  - The relative frequency of A is $f_n = \frac{S_n}{n}$, so $E[f_n] = \frac{E[S_n]}{n} = p$
  - $\text{Var}[f_n] = \frac{\sigma^2}{n} = \frac{p(1-p)}{n} \leq \frac{1}{4n}$
    - But we don't know p, so instead we note $p(1 - p)$ is bounded by $1/4$
    - Therefore $\text{Var}[f_n] \leq \frac{1}{4n}$
  - This gives us a way to estimate p while bounding the variance on our estimate
    - e.g. we want to be within $\frac{1}{10}$ of the true probability 90% of the time
      - Chebyshev: $P[|f_n - p| > \underbrace{0.1}_{\epsilon}] \leq \underbrace{0.1}_{\delta}$, then $0.1 = \frac{p(1-p)}{n_0 \left(\frac{1}{10}\right)^2} \leq \frac{1}{4n_0 \left(\frac{1}{10}\right)^2}$
      - Solve to get $n_0 > 250$

Introduction to Parameter Estimation

- Given an IID sequence of random variables, we want to estimate a parameter θ of the distribution X
  - The distribution depends on θ; it can be e.g. for Bernoulli it is $\theta = P[X = 1]$; for a Gaussian

    $\theta = (m_X, \sigma^2)$
  - $\hat{\Theta}_n$ is an estimator of the unknown parameter
- Note that the estimator is a function of the RVs, $\hat{\Theta}_n(\boldsymbol{X})$
- Estimators have the following properties:
  - The *error* is $\hat{\Theta}_n(\boldsymbol{X}) - \theta$
    - This is how much the estimate is off by from the true value
  - The *bias* is $E[\hat{\Theta}_n(\boldsymbol{X})] - \theta$
    - This is whether we get the correct estimate on average
  - An estimator is *unbiased* if the expected value of the error is zero, i.e. the bias is zero
    - i.e. on average, our estimate will be correct
  - An estimator is *asymptotically unbiased* if $\lim_{n \to \infty} E[\Theta_n(\boldsymbol{X})] = \theta$, even if it's not unbiased
  - An estimator is *consistent* if as $n \to \infty$, the distribution of $\hat{\Theta}_n$ converges to θ (weak law)
    - i.e. as the sample size increases, the estimates become more and more concentrated around θ
    - Consistency implies asymptotic unbiasedness (if the estimator has finite variance) but the reverse is not true