Statistics for Data Science - 2

Week 7 Notes

Statistics from samples and Limit theorems

1. **Empirical distribution:**
   Let $X_1, X_2, \ldots, X_n \sim X$ be i.i.d. samples. Let $\#(X_i = t)$ denote the number of times t occurs in the samples. The empirical distribution is the discrete distribution with PMF
   $$p(t) = \frac{\#(X_i = t)}{n}$$
   - The empirical distribution is random because it depends on the actual sample instances.
   - **Descriptive statistics:** Properties of empirical distribution. Examples :
     - Mean of the distribution
     - Variance of the distribution
     - Probability of an event
   - As number of samples increases, the properties of empirical distribution should become close to that of the original distribution.

2. **Sample mean:**
   Let $X_1, X_2, \ldots, X_n \sim X$ be i.i.d. samples. The sample mean, denoted $\overline{X}$, is defined to be the random variable
   $$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$
   - Given a sampling $x_1, \ldots, x_n$ the value taken by the sample mean $\overline{X}$ is $\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$. Often, $\overline{X}$ and $\overline{x}$ are both called sample mean.

3. **Expected value and variance of sample mean:**
   Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples whose distribution has a finite mean μ and variance σ^2. The sample mean $\overline{X}$ has expected value and variance given by
   $$E[\overline{X}] = \mu, \quad \text{Var}(\overline{X}) = \frac{\sigma^2}{n}$$
   - Expected value of sample mean equals the expected value or mean of the distribution.
   - Variance of sample mean decreases with n.

4. **Sample variance:**
   Let $X_1, X_2, \ldots, X_n \sim X$ be i.i.d. samples. The sample variance, denoted S^2, is defined to be the random variable

$$S^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}{n-1},$$

   where $\overline{X}$ is the sample mean.

5. **Expected value of sample variance:**
   Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples whose distribution has a finite variance σ^2. The sample variance $S^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}{n-1}$ has expected value given by

$$E[S^2] = \sigma^2$$

   - Values of sample variance, on average, give the variance of distribution.
   - Variance of sample variance will decrease with number of samples (in most cases).
   - As n increases, sample variance takes values close to distribution variance.

6. **Sample proportion:**
   The sample proportion of A, denoted $S(A)$, is defined as

$$S(A) = \frac{\text{number of } X_i \text{ for which } A \text{ is true}}{n}$$

   - As n increases, values of $S(A)$ will be close to $P(A)$.
   - Mean of $S(A)$ equals $P(A)$.
   - Variance of $S(A)$ tends to 0.

7. **Weak law of large numbers:**
   Let $X_1, X_2, \ldots, X_n \sim$ iid X with $E[X] = \mu, \text{Var}(X) = \sigma^2$.
   Define sample mean $\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$. Then,

$$P(|\overline{X} - \mu| > \delta) \leq \frac{\sigma^2}{n\delta^2}$$