Module 5

Point estimation

Topics: Simple random sample, Estimation of mean and variance, Estimation of parameters, Introduction to regression analysis

Simple random sample

Given a random experiment and a probability measure for it, probability theory enables us to determine probability of an event, calculate conditional probabilities, check independence of different events, etc. If the outcomes of the experiment are real numbers or are converted into real numbers, then probability theory allows us to obtain mean and variance of the outcomes and other interesting quantities. For vector-valued outcomes or such conversions, probability theory helps us obtain conditional mean and variance, correlation between components, and other quantities of interest. In short, probability theory enables us to anticipate behavior of a random experiment when we have a probability measure for it.

If you look back at all the examples that we discussed, all the problems that you solved, and all the questions that I asked in the exams, the probability measure was classical in most of the cases. In a few cases where the classical measure was not applicable, probability values were explicitly specified. In real-life problems, where classical measure is not applicable or we are unsure about it, we need a way to construct the probability measure. We already know the way – it's the frequentist's measure, if we can perform the experiment several times, else it's the subjective measure with (Bayesian) updating. In this course, we restrict ourselves to the frequentist's measure, i.e., we assume that we can perform the experiment several times in order to construct the probability measure.

In most real-life situations, we are interested only in some specific aspect of the probability measure. For example, consider the portfolio optimization problem where we want to invest some amount of money in some risky assets so that the expected return is maximized while the risk (measured in terms of variance) stays within a threshold. Let us consider the money available for investment to be 1 unit. With respect to this unit, let X_i denote the random return from 1 unit of investment in the i-th asset, for $i = 1,2,\ldots,n$. Let $E[X_i] = \mu_i$, $Var(X_i) = \sigma_i^2$, and $Cov(X_i, X_j) = \sigma_{ij}$ for $i < j$. Let $w_1, w_2, \ldots, w_n$ denote the amounts of investments in the assets (decision variables). Then the portfolio return is: $\sum_{i=1}^{n} w_i X_i$ with mean $\sum_{i=1}^{n} w_i \mu_i$ and variance $\sum_{i=1}^{n} w_i^2 \sigma_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_i w_j \sigma_{ij}$. Then the problem is:

$$\underset{w_1, w_2, \ldots, w_n}{\text{Maximize}} \sum_{i=1}^{n} w_i \mu_i$$

$$\text{such that } \sum_{i=1}^{n} w_i^2 \sigma_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_i w_j \sigma_{ij} \le \tau \text{ (threshold)}$$

$$w_1 + w_2 + \cdots + w_n \le 1 \text{ (budget constraint)}$$

$$\text{and } w_1, w_2, \ldots, w_n \ge 0$$

We have optimization methods to solve the above problem, provided we know the numerical values of $\mu_i, \sigma_i^2, \sigma_{ij}$. If we know the joint distribution of $(X_1, X_2, \dots, X_n)$, we can obtain these quantities. Of course, we don't know the distribution. We can look into the past data (which is same as repeating the underlying experiment several times) and construct the distribution, and then obtain the quantities. Alternately, we can directly estimate the quantities of interest from the data. The second approach is simpler.

In statistics, our aim is to estimate some aspects of the distribution of a random variable or a random vector from the data, which is essentially observed values of several repetitions of the underlying random experiment. In most cases, the repetitions are independent and identical. Then the data is referred to as simple random sample. For random variable X, the simple random sample of size n can be represented as $X_1, X_2, \dots, X_n$, a set of *iid* random variables having the distribution of X. For random vector (X, Y), the simple random sample can be represented by *iid* random vectors $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$. If outcomes of the random experiment are not numerical valued, we can convert them into numbers, and thereby, restrict our study of statistics to random variables and vectors. Estimating something from data is essentially constructing a function $g(X_1, X_2, \dots, X_n)$ or $g\big((X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\big)$ that must pass some goodness criteria. In the remaining classes, we will learn to construct these functions and study their goodness. Once the function is constructed and the data is observed, the rest is numerical calculations. Note that the data before its observation is random and so is the estimation function, which is the subject of our study.

Data in the form of simple random sample, i.e., independence and identicalness of the parts, is desirable. Sometimes, due to limitations in data collection procedure, this does not happen. For example, consider the delay of a morning train at Kanpur Station. It's a random variable and we want to estimate its mean. Starting on 1st November, we note down the delay for the next two months from the website of Indian Railways. Let $X_1, X_2, \dots, X_{61}$ denote this data. This is not a simple random sample, because presence/absence of fog influences the delay of morning trains in Kanpur. In November, there is no fog, and in later-half of December, most days are foggy. Thus, X_1 and X_{61} and more such combinations are not identical. Here, we can have a simple random sample if we pick the dates randomly throughout a year.

Consider another example where we want to estimate expected duration of stay for tourists in a small island. We picked one of the hotels and obtained length of stay for n tourists from the hotel's registrar. Let $X_1, X_2, \dots, X_n$ denote the data. This may not be a simple random sample because boarders of the same hotel are from similar socio-economic background, and thus, have spent similar amount of time in the island. Then $X_1, X_2, \dots, X_n$ may not be independent. Here, we can have a simple random sample if we pick data points from different hotels. In this course, we restrict ourselves to simple random samples. Estimation functions that we develop here cannot be applied directly if the data is not a simple random sample.

In certain situations, the quantity of interest is not associated with any random experiment. For example, consider the average annual income of all residents of Uttar Pradesh in 2021. This quantity is not associated with any random experiment. If we manage to reach out to all the residents, then we can obtain the exact value of the average annual income. This is not the case with the mean delay of the train at Kanpur. No amount of data can lead us to the exact value of the mean delay with certainty. Now, in the income example, it is nearly impossible to reach out to all the residents. So, we reach out to a randomly selected subset of them, and estimate the average annual income based on their responses. This random selection of the sample leads to randomness in the estimation. In fact, the sample can be viewed as a simple random sample of a random variable that describes the income of a randomly selected person from the population, as explained below. Then all estimation functions that we develop here can be applied to random samples of a population.

Let $1,2,\dots,N$ denote all the residents and $\{a_1,a_2,\dots,a_m\}$ denote the set of their incomes. Let k_i number of persons have income a_i, for $i=1,2,\dots,m$. Note that $k_1+k_2+\dots+k_m=N$. The average income that we want to obtain is $\sum_{i=1}^{m} a_i k_i/N$. Let X denote the income of a randomly chosen person from the population. It is a discrete random variable with $p_X(a_i)=k_i/N$ for $i=1,2,\dots,m$ and $E[X]=\sum_{i=1}^{m} a_i p_X(a_i)=\sum_{i=1}^{m} a_i k_i/N$, same as the quantity of interest. Let $X_1,X_2,\dots,X_n$ denote the sample, i.e., incomes of the randomly selected persons. In order to show that $X_1,X_2,\dots,X_n$ is a simple random sample of X, we need to show that the random variables have the mass function of X and are independent.

Let us consider that a sequential selection scheme with replacement is adopted for sampling, i.e., one person is randomly selected from the population, his/her response is recorded, and then returned to the population, and this process is repeated n times to obtain $X_1,X_2,\dots,X_n$. Since the population returns to the original state before every random selection, $X_1,X_2,\dots,X_n$ have the mass function of X. Also, the random selections do not influence with one another, and therefore, $X_1,X_2,\dots,X_n$, are independent. If the 'replacement part' is removed from the selection scheme, we still have identical $X_1,X_2,\dots,X_n$ (both for sequential and simultaneous selection schemes), but they are no more independent.

Estimation of mean and variance

Let us begin our study of statistical inference with the estimation of mean and variance of a random variable. Let X denote the random variable of interest, whose probability distribution is unknown to us, and $X_1,X_2,\dots,X_n$ denote its simple random sample. The natural estimators of its mean μ and variance σ^2 are: $\hat{\mu}=\bar{X}=\sum_{i=1}^{n} X_i/n$ and $\hat{v}=\sum_{i=1}^{n}(X_i-\bar{X})^2/n$. We refer to these as natural estimators because of their resemblance with the definitions of mean and variance of a random variable.

The natural estimators are not the only estimators. For example, $\widehat{\mu_1}=5$, $\widehat{\mu_2}=X_1$, $\widehat{\mu_3}=\beta_0+\sum_{i=1}^{n}\beta_i X_i$ for constants $\beta_0,\beta_1,\dots,\beta_n$ are valid estimators of μ. Likewise, one can construct many more estimators. We need to have a way of comparing different estimators. We call an estimator good if it is ‘close’ to the underlying quantity of interest. The closeness is measured primarily through bias and mean squared error.

Consider θ to be the quantity of interest associated with random variable X. Let $X_1, X_2, \dots, X_n$ denote its random sample and $\hat{\theta} = g(X_1, X_2, \dots, X_n)$ denote an estimator of θ. Depending on the observed values of $X_1, X_2, \dots, X_n$, $\hat{\theta}$ can be less than θ, more than θ, or even same as θ. If we repeat the estimation process several times, then we have several values of $\hat{\theta}$. Average of these values should ideally converge to θ as the number of repetitions increases. This is same as saying that $E[\hat{\theta}] = \theta$. Note that $\hat{\theta} = g(X_1, X_2, \dots, X_n)$ is a random variable. An estimator having this feature is known as unbiased estimator. It’s a desirable property. An estimator can be biased, which is measured as: $B[\hat{\theta}] = E[\hat{\theta}] - \theta$.

Let us check if the estimators we mentioned for μ are unbiased or not. $E[\hat{\mu}] = E[\bar{X}] = \mu$. So, the natural estimator of mean is unbiased. $E[\widehat{\mu_1}] = E[5] = 5$. It may or may not be biased; we cannot tell for sure. $E[\widehat{\mu_2}] = E[X_1] = \mu$. It is unbiased, like the natural estimator. $E[\widehat{\mu_3}] = E[\beta_0 + \sum_{i=1}^{n} \beta_i X_i] = \beta_0 + \sum_{i=1}^{n} \beta_i E[X_i] = \beta_0 + \mu \sum_{i=1}^{n} \beta_i$. It is unbiased if $\beta_0 + \mu \sum_{i=1}^{n} \beta_i = \mu \equiv \beta_0 + \mu(\sum_{i=1}^{n} \beta_i - 1) = 0$. Without knowing μ, the only way to ensure unbiasedness is to set $\beta_0 = 0$ and choose $\beta_1, \beta_2, \dots, \beta_n$ such that $\sum_{i=1}^{n} \beta_i = 1$. One such choice is: $\beta_1 = \beta_2 = \cdots = \beta_n = 1/n$, which gives us the natural estimator.

Let us compare $\hat{\mu} = \bar{X}$ and $\widehat{\mu_2} = X_1$. Both are unbiased estimators of mean, but $Var(\hat{\mu}) = Var(\bar{X}) = \sigma^2/n$, whereas $Var(\widehat{\mu_2}) = Var(X_1) = \sigma^2$. Since the natural estimator has less variance, it is better than $\widehat{\mu_2}$. If we include biased estimators in the comparison, then variance of the estimator can be misleading. This is because $Var(\hat{\theta})$ is measured with respect to $E[\hat{\theta}]$, which is not same as θ for biased estimators. For the correct comparison, we need to capture $E\left[(\hat{\theta} - \theta)^2\right]$, which is referred to as the mean squared error (MSE). It can be expressed in terms of bias and variance as follows:

$$MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right] = E\left[\left\{(\hat{\theta} - E[\hat{\theta}]) + (E[\hat{\theta}] - \theta)\right\}^2\right]$$
$$= E\left[(\hat{\theta} - E[\hat{\theta}])^2 + 2(\hat{\theta} - E[\hat{\theta}])B[\hat{\theta}] + B^2[\hat{\theta}]\right] = Var(\hat{\theta}) + B^2[\hat{\theta}]$$

Smaller the mean squared error, better the estimator. Let us study $\widehat{\mu_3} = \sum_{i=1}^{n} \beta_i X_i$ having $\sum_{i=1}^{n} \beta_i = 1$. It is the general linear unbiased estimator of μ; $\hat{\mu} = \sum_{i=1}^{n} X_i/n$ is a special case of $\widehat{\mu_3}$. $MSE(\widehat{\mu_3}) = Var(\sum_{i=1}^{n} \beta_i X_i) = \sum_{i=1}^{n} \beta_i^2 Var(X_i) = \sigma^2 \sum_{i=1}^{n} \beta_i^2$. Now,

$$\sum_{i=1}^{n} \beta_i^2 = \sum_{i=1}^{n} \left\{\left(\beta_i - \frac{1}{n}\right) + \frac{1}{n}\right\}^2 = \sum_{i=1}^{n} \left(\beta_i - \frac{1}{n}\right)^2 + \frac{2}{n} \sum_{i=1}^{n} \left(\beta_i - \frac{1}{n}\right) + \frac{n}{n^2}$$
$$= \sum_{i=1}^{n} \left(\beta_i - \frac{1}{n}\right)^2 + \frac{2}{n}\left(\sum_{i=1}^{n} \beta_i - 1\right) + \frac{1}{n} = \sum_{i=1}^{n} \left(\beta_i - \frac{1}{n}\right)^2 + \frac{1}{n} \geq \frac{1}{n}.$$

Then $MSE(\widehat{\mu_3}) = \sigma^2 \sum_{i=1}^{n} \beta_i^2 \geq \sigma^2/n = MSE(\bar{X})$. Evidently, $\bar{X}$ is the best linear unbiased estimator (BLUE) of μ. In addition to bias and mean squared error, we check for two more properties of theoretical importance – consistency and sufficiency. A consistent estimator is one that converges to the quantity of interest as the sample size $n \to \infty$, i.e., $\hat{\theta}$ is consistent if $\lim_{n\to\infty} P(|\hat{\theta} - \theta| > \epsilon) = 0$ for all $\epsilon > 0$. Consistency of an estimator can be verified from its MSE, as explained below. Sufficiency is about capturing all meaningful information about the quantity of interest from the sample. We will not discuss it in this course.

Chebyshev's Inequality states that $P(|X - \mu| > k\sigma) \leq 1/k^2 \ \forall k > 0$, where $\mu = E[X]$ and $\sigma^2 = Var(X)$. An alternate form of the inequality is: $P(|X - \mu| > \epsilon) \leq \sigma^2/\epsilon^2 \ \forall \epsilon > 0$. The generalized inequality considers deviation of X from an arbitrary point α. Let $Y = (X - \alpha)^2$. Since Y is non-negative, by Markov's Inequality, $P(Y > t) \leq E[Y]/t \ \forall t > 0$, where $E[Y] = E[(X - \alpha)^2] = E[X^2] - 2\alpha E[X] + \alpha^2 = E[X^2] - \mu^2 + \mu^2 - 2\alpha\mu + \alpha^2 = \sigma^2 + (\mu - \alpha)^2$. Let us take $t = \epsilon^2$. Then for every $\epsilon > 0$, there is a $t > 0$. Then $P((X - \alpha)^2 > \epsilon^2) \leq \{\sigma^2 + (\mu - \alpha)^2\}/\epsilon^2 \equiv P(|X - \alpha| > \epsilon) \leq \{\sigma^2 + (\mu - \alpha)^2\}/\epsilon^2 \ \forall \epsilon > 0$.

Let us apply the generalized Chebyshev's Inequality to an arbitrary estimator $\hat{\theta}$ of a quantity of interest θ. By setting $\alpha = \theta$, we obtain $P(|\hat{\theta} - \theta| > \epsilon) \leq \left\{Var(\hat{\theta}) + \left(E[\hat{\theta}] - \theta\right)^2\right\}/\epsilon^2 = MSE(\hat{\theta})/\epsilon^2 \ \forall \epsilon > 0$. From this inequality, it is clear that if $MSE(\hat{\theta}) \to 0$ as the sample size $n \to \infty$, then $\lim_{n\to\infty} P(|\hat{\theta} - \theta| > \epsilon) = 0 \ \forall \epsilon > 0$, implying consistency of $\hat{\theta}$. On the other hand, if $MSE(\hat{\theta}) \nrightarrow 0$ as $n \to \infty$, then $\hat{\theta}$ is not consistent. Since $MSE(\bar{X}) = \sigma^2/n \to 0$ as $n \to \infty$, $\bar{X}$ is a consistent estimator of μ.

Let us study the natural estimator of variance $\hat{v} = \sum_{i=1}^{n} (X_i - \bar{X})^2/n$.

$$\hat{v} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{2\bar{X}}{n}\sum_{i=1}^{n} X_i + \frac{n\bar{X}^2}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$$

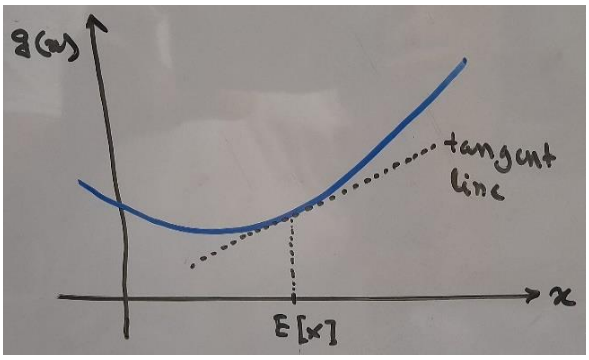
$$\Rightarrow E[\hat{v}] = \frac{1}{n}\sum_{i=1}^{n} E[X_i^2] - E[\bar{X}^2] = \frac{1}{n}\sum_{i=1}^{n} \{Var(X_i) + E^2[X_i]\} - \{Var(\bar{X}_i) + E^2[\bar{X}]\}$$

$$= \frac{1}{n}\sum_{i=1}^{n} (\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2\right) = \sigma^2\left(1 - \frac{1}{n}\right)$$

Observe that $\hat{v} = \sum_{i=1}^{n} (X_i - \bar{X})^2/n$ is not an unbiased estimator of variance σ^2. The above calculations suggest that the estimator $S^2 \coloneqq \sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1)$ is an unbiased estimator of σ^2. Division by $n - 1$ is more appropriate because one on the X_i values is redundant, i.e., one of the X_i values can be obtained from the remaining values and $\bar{X}$. MSE of S^2 involves higher order moments of X. We won't calculate it here. However, it vanishes as $n \to \infty$, implying consistency of S^2. Overall, S^2 is a very good estimator of σ^2.

Other than mean and variance, sometimes, we are interested in estimating standard deviation, quantiles, distribution function values, covariance, correlation, etc. We can construct natural estimators for these quantities following their definitions. For example, square root of $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ can be regarded as the natural estimator for standard deviation. This, however, is not an unbiased estimator of σ, as explained below.

In Module 2, while obtaining expectation of real-valued function of a random variable, we noted that expectation of a linear function is same as the function of expectation, but the same may not be the case for non-linear functions. Jensen's Inequality provides more information on this. It states that $E[g(X)] \geq g(E[X])$ whenever g is a convex function.



Let g be an arbitrary convex function. Let $a + bx$ denote the tangent line of g at $E[X]$, as shown in the diagram. Due to convexity, $g(x) \geq a + bx\ \forall x \in \mathbb{R}$. Then for continuous X, $\int_{-\infty}^{\infty} g(x) f_X(x)dx \geq \int_{-\infty}^{\infty}(a + bx) f_X(x)dx \Rightarrow E[g(X)] \geq a + bE[X] = g(E[X])$, as claimed. One can prove for the discrete case as well.

If g is a concave function, then $-g$ is convex, and by Jensen's Inequality, $E[-g(X)] \geq -g(E[X]) \equiv -E[g(X)] \geq -g(E[X]) \equiv E[g(X)] \leq g(E[X])$. The above proof also tells that strict inequality holds for strictly convex/concave functions.

Square root function is strictly concave. By Jensen's Inequality, $E[S] = E\left[\sqrt{S^2}\right] < \sqrt{E[S^2]} = \sqrt{\sigma^2} = \sigma$. So, S is biased. In fact, there is no single estimator of σ that is unbiased for all X. For practical purposes, S is a reasonably good estimator of σ. *Construct natural estimators of some other quantities of interest and try to study their properties.*

Estimation of parameters

So far, we considered no knowledge about the random variable X, and estimated quantities of interest about X following the definitions of those quantities. Sometimes, we have partial knowledge about the distribution of X. For example, consider estimation of p, the probability of getting head in a coin toss, when we have outcome data of n tosses. Here, we know that the underlying random variable X follows Bernoulli distribution, but we do not know the success probability p. We shall make use of this information about the distribution of X while estimating p. Maximum likelihood method provides a generic way of estimating distribution parameter(s) when the form of the distribution is known.

Let $X \sim Ber(p)$, where p is unknown. Let 0,1,1,0,1 denote the data, i.e., the realized value of a simple random sample of X, where 1 denotes success and 0 denotes failure. We want to estimate success probability p in this setup. Let us hypothetically assume that p is either 0.4 or 0.7 and we must pick one of these values as estimate for p. If we choose $p = 0.4$, then the probability of observing the data 0,1,1,0,1 is: $0.4^3 \times 0.6^2 = 0.02304$. On the other hand, if we choose $p = 0.7$, the probability of observing the given data is: $0.7^3 \times 0.3^2 = 0.03087$. The likelihood of observing the given data is more if we choose $p = 0.7$, and therefore, we shall pick 0.7 as our estimate for p. This is the core idea of the maximum likelihood method, i.e., choose that parameter value which maximizes the likelihood of observing the given data.

In the above example, we artificially restricted p to be either 0.4 or 0.7. Let us remove the restriction. Then $p \in (0,1)$ and we need to pick a value that maximizes the likelihood of observing the data 0,1,1,0,1. The likelihood for an arbitrary $p \in (0,1)$ is: $p^3(1-p)^2$. We need to maximize the likelihood function with respect to p.

Stationarity condition: $3p^2(1-p)^2 - 2p^3(1-p) = 0 \Rightarrow 3(1-p) = 2p \Rightarrow p = 0.6$
Second derivative: $6p(1-p)^2 - 6p^2(1-p) - 6p^2(1-p) + 2p^3$,
which is $6 \times 0.6 \times 0.4^2 - 12 \times 0.6^2 \times 0.4 + 2 \times 0.6^3 = -0.72$ at $p = 0.6$

So, $p = 0.6$ offers the maximum likelihood of observing the given data among all $p \in (0,1)$. Therefore, the maximum likelihood estimate of p is 0.6.

In the above example, we worked with a realized value of the simple random sample, which can be something else as well. So, we need to place ourselves before the realization takes place and obtain the maximum likelihood estimator (MLE) as a function of the sample $X_1, X_2, \ldots, X_n$. Let us find MLE of p for $Ber(p)$. The likelihood function is obtained as:

$$L_X(p) = P\big(\text{Observing } X_1, X_2, \ldots, X_n | p \in (0,1)\big) = \prod_{i=1}^{n} P\big(\text{Observing } X_i | p \in (0,1)\big)$$
$$= \prod_{i=1}^{n} p_{X|p}(X_i) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$$

Note that $p_{X|p}$ denotes the mass function of X with parameter value p. We need to maximize $L_X(p)$ with respect to p. Given the nature of $L_X(p)$, maximizing $l_X(p) = \ln\big(L_X(p)\big)$ seems easier. Since, the logarithm function is strictly increasing, p-value that maximizes $\ln\big(L_X(p)\big)$ would also maximize $L_X(p)$. The log-likelihood function is maximized next.

$$l_X(p) = \ln\big(L_X(p)\big) = \sum_{i=1}^{n} \ln\big(p^{X_i}(1-p)^{1-X_i}\big) = \sum_{i=1}^{n} \{X_i \ln(p) + (1-X_i)\ln(1-p)\}$$
$$\Rightarrow l_X'(p) = \sum_{i=1}^{n} \left(\frac{X_i}{p} - \frac{1-X_i}{1-p}\right) = \sum_{i=1}^{n} \frac{X_i - p}{p(1-p)}$$
$$\Rightarrow l_X''(p) = \sum_{i=1}^{n} \frac{-p(1-p) - (X_i - p)(1-2p)}{p^2(1-p)^2} = -\sum_{i=1}^{n} \frac{p^2 + X_i(1-2p)}{p^2(1-p)^2}$$
$$= -\sum_{i=1}^{n} \frac{(p-X_i)^2 + X_i(1-X_i)}{p^2(1-p)^2} < 0 \Rightarrow l_X(p) \text{ is concave.}$$

Maxima: $l_X'(\hat{p}) = 0 \Rightarrow \sum_{i=1}^{n}(X_i - \hat{p}) = 0 \Rightarrow \hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$

So, the MLE of p for $Ber(p)$ is: $\hat{p} = \bar{X} = \sum_{i=1}^{n} X_i/n$. We can study properties of this MLE. *Verify that $\hat{p}$ is unbiased, obtain its MSE, and establish its consistency.*

Let us obtain MLE of p for $Geo(p)$ from simple random sample $X_1, X_2, \dots, X_n$. Here, $X_i \geq 1$ for all i and it denotes number of trials required to get the first success.

$$L_X(p) = P\big(\text{Observing } X_1, X_2, \dots, X_n | p \in (0,1)\big) = \prod\nolimits_{i=1}^{n} P\big(\text{Observing } X_i | p \in (0,1)\big)$$
$$= \prod\nolimits_{i=1}^{n} p_{X|p}(X_i) = \prod\nolimits_{i=1}^{n} p(1-p)^{X_i - 1} = p^n (1-p)^{\sum_{i=1}^{n} X_i - n}$$
$$\Rightarrow l_X(p) = \ln\big(L_X(p)\big) = n\ln(p) + \left(\sum_{i=1}^{n} X_i - n\right)\ln(1-p)$$
$$\Rightarrow l_X'(p) = \frac{n}{p} - \frac{\sum_{i=1}^{n} X_i - n}{1-p} = \frac{n}{p} - \frac{n\bar{X} - n}{1-p}, \text{where } \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$\Rightarrow l_X''(p) = -\frac{n}{p^2} - \frac{n\bar{X} - n}{(1-p)^2} = -n\left\{\frac{1}{p^2} + \frac{\bar{X} - 1}{(1-p)^2}\right\} < 0 \Rightarrow l_X(p) \text{ is concave.}$$

Maxima: $l_X'(\hat{p}) = 0 \Rightarrow \frac{n}{\hat{p}} = \frac{n(\bar{X} - 1)}{1 - \hat{p}} \Rightarrow \hat{p} = \frac{1}{\bar{X}}$

So, the MLE of p for $Geo(p)$ is: $\hat{p} = 1/\bar{X}$. *Using Jensen's Inequality, show that $\hat{p}$ is biased.* It's not easy to study MSE and consistency of this estimator. Later, we will mention about a property of MLEs that will give us some idea about goodness of $\hat{p} = 1/\bar{X}$.

Following the above principle, show that MLE of μ for $Pois(\mu)$, where $\mu = \lambda t$, is: $\hat{\mu} = \bar{X}$. One can study its goodness properties easily. In a similar manner, one can establish that $\hat{p} = \bar{X}/N$ is the MLE of p for $Bin(N, p)$, where N is known, and study its properties. Now, let us turn our attention to continuous random variable. Let $\theta \in \Theta$ denote the unknown parameter and $f_{X|\theta}$ denote the density function of X with parameter value θ. Then

$$P(\text{Observing } X_1, X_2, \dots, X_n | \theta \in \Theta) = \prod\nolimits_{i=1}^{n} P(\text{Observing } X_i | \theta \in \Theta)$$
$$= \prod\nolimits_{i=1}^{n} \lim_{\delta \to 0} f_{X|\theta}(X_i)\delta = \left[\prod\nolimits_{i=1}^{n} f_{X|\theta}(X_i)\right] \lim_{\delta \to 0} \delta^n$$

Observe that maximizing the above likelihood is same as maximizing $\prod_{i=1}^{n} f_{X|\theta}(X_i)$, which is considered as the likelihood function for the continuous case. For the discrete case, likelihood function is: $\prod_{i=1}^{n} p_{X|\theta}(X_i)$. In a similar manner, joint mass/density function can be used to construct likelihood functions for random vectors.

Let us obtain MLE of a and b for $U(a, b)$ from simple random sample $X_1, X_2, \dots, X_n$.

$$L_X(a,b) = \prod_{i=1}^{n} f_{X|a,b}(X_i) = \begin{cases} 1/(b-a)^n & \text{if } a \le X_i \le b\ \forall i \\ 0 & \text{otherwise} \end{cases}$$

We must choose a, b such that $a \le X_i \le b\ \forall i$, otherwise $L_X(a,b) = 0$ which is its lowest possible value. Any $a \le \min\{X_1, X_2, \dots, X_n\}$ and $b \ge \max\{X_1, X_2, \dots, X_n\}$ would ensure that $L_X(a,b) = 1/(b-a)^n > 0$. Now, $1/(b-a)^n$ is maximized when $b-a$ is least, which is ensured by setting $a = \min\{X_1, X_2, \dots, X_n\}$ and $b = \max\{X_1, X_2, \dots, X_n\}$. Thus, MLE of a and b of $U(a,b)$ are: $\hat{a} = \min\{X_1, X_2, \dots, X_n\}$ and $\hat{b} = \max\{X_1, X_2, \dots, X_n\}$. *Using the distribution of minimum and maximum, obtain bias and MSE of $\hat{a}$ and $\hat{b}$. Are these consistent*?

Let us obtain MLE of μ and $\sigma^2 = v$ for $N(\mu, \sigma^2)$ from simple random sample $X_1, X_2, \dots, X_n$.

$$L_X(\mu, v) = \prod_{i=1}^{n} f_{X|\mu,v}(X_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi v}} e^{-\frac{(X_i-\mu)^2}{2v}} = \frac{1}{(2\pi v)^{n/2}} e^{-\frac{1}{2v}\sum_{i=1}^{n}(X_i-\mu)^2}$$

$$\Rightarrow l_X(\mu, v) = \ln\big(L_X(\mu, v)\big) = -\frac{n}{2}\ln(2\pi v) - \frac{1}{2v}\sum_{i=1}^{n}(X_i-\mu)^2$$

$$\Rightarrow \frac{\partial l_X}{\partial \mu} = \frac{1}{v}\sum_{i=1}^{n}(X_i-\mu) \text{ and } \frac{\partial l_X}{\partial v} = -\frac{n}{2v} + \frac{1}{2v^2}\sum_{i=1}^{n}(X_i-\mu)^2$$

$$\Rightarrow H(\mu, v) = \begin{bmatrix} -\frac{n}{v} & -\frac{1}{v^2}\sum_{i=1}^{n}(X_i-\mu) \\ -\frac{1}{v^2}\sum_{i=1}^{n}(X_i-\mu) & \frac{n}{2v^2} - \frac{1}{v^3}\sum_{i=1}^{n}(X_i-\mu)^2 \end{bmatrix}$$

$$\text{Now,} \frac{\partial l_X}{\partial \mu} = 0 \Rightarrow \frac{1}{\hat{v}}\sum_{i=1}^{n}(X_i-\hat{\mu}) = 0 \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

$$\text{and } \frac{\partial l_X}{\partial v} = 0 \Rightarrow \frac{n}{2\hat{v}} = \frac{1}{2\hat{v}^2}\sum_{i=1}^{n}(X_i-\hat{\mu})^2 \Rightarrow \hat{v} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

$$H(\hat{\mu}, \hat{v}) = \begin{bmatrix} -\frac{n^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2} & 0 \\ 0 & -\frac{n^3}{2\{\sum_{i=1}^{n}(X_i-\bar{X})^2\}^2} \end{bmatrix} \sim \text{negative definite}$$

Since the Hessian matrix is negative definite, $(\hat{\mu}, \hat{v})$ is the maxima. Therefore, $\hat{\mu} = \bar{X}$ and $\hat{v} = \sum_{i=1}^{n}(X_i - \bar{X})^2/n$ are the MLEs of mean and variance of $N(\mu, \sigma^2)$. We already have studies properties of these estimators in the previous section.

Following the above principle, show that MLE of λ for $Exp(\lambda)$ is: $\hat{\lambda} = 1/\bar{X}$. Let $\mu = 1/\lambda$. Then $f_X(x) = \lambda e^{-\lambda x} = (1/\mu)e^{-x/\mu}$ for $x > 0$, zero otherwise. So, we can consider μ as the parameter for exponential distribution, instead of λ. *Verify that $\hat{\mu} = \bar{X}$ is MLE for $Exp(\lambda)$.* Observe that $\hat{\mu} = 1/\hat{\lambda}$, i.e., MLE of the function $(1/\lambda)$ is function of the MLE $(1/\hat{\lambda})$. This is no coincidence. If $\hat{\theta}$ is MLE of θ, then $g(\hat{\theta})$ is MLE of $g(\theta)$, where g is an arbitrary real-valued function. This is known as the invariance property of MLE.

Let the distribution of X can be written in terms of θ or in terms of $\beta = g(\theta)$. Let $L_X(\theta) = \prod_{i=1}^{n} f_{X|\theta}(X_i) = h_1(\theta, X_1, \ldots, X_n)$ and $L_X(\beta) = \prod_{i=1}^{n} f_{X|\beta}(X_i) = h_2(\beta, X_1, \ldots, X_n)$ denote the likelihood functions in terms of θ and β respectively. Since they represent the same function, $h_2(\beta, X_1, \ldots, X_n) = h_2(g(\theta), X_1, X_2, \ldots, X_n) = h_1(\theta, X_1, \ldots, X_n)$ for all θ. If $\hat{\theta}$ is MLE, then $h_1(\hat{\theta}, X_1, \ldots, X_n) = h_2(g(\hat{\theta}), X_1, X_2, \ldots, X_n)$ is the highest value of the likelihood function. Therefore, $\hat{\beta} = g(\hat{\theta})$ must be MLE, as claimed.

Another important property of the MLE is its asymptotic normality, which says that MLE $\hat{\theta}$ of any parameter θ converges to $N(\theta, 1/n\mathbb{I}(\theta))$ as the sample size $n \to \infty$, where

$$\mathbb{I}(\theta) = Var\left(\frac{\partial}{\partial\theta}\ln\big(f(X|\theta)\big)\right) = E\left[\left(\frac{f'(X|\theta)}{f(X|\theta)}\right)^2\right]$$

denotes Fisher Information. It measures the amount of information that an observable random variable X carries about an unknown parameter θ of a distribution f that models X. We won’t discuss any further details on this, though we will study its implications. Note that f is to be treated as mass function if X is discrete. An immediate consequence of asymptotic normality of MLE is its consistency. Since $\mathbb{I}(\theta) > 0$ (unless θ has nothing to do with X), $1/n\mathbb{I}(\theta) \to 0$ as $n \to \infty$. So, $\hat{\theta}$ ‘converges’ to θ as $n \to \infty$, implying consistency of $\hat{\theta}$. One can also see that $E[\hat{\theta}] \approx \theta$ for large n, i.e., bias (if any) vanishes as sample size increases.

We talked about (large sample) unbiasedness and consistency of MLE. The next result talks about MSE. If $X(\Omega)$, also known as the support of the random variable X, does not depend on θ, then any estimator $\hat{\theta}$ of θ (MLE or anything else) satisfies

$$Var(\hat{\theta}) \geq \frac{\left(E'[\hat{\theta}]\right)^2}{n\mathbb{I}(\theta)}, \text{where } E'[\hat{\theta}] = \frac{\partial E[\hat{\theta}]}{\partial\theta}.$$

The above bound is known as Cramer-Rao Lower Bound. If $\hat{\theta}$ is unbiased, then $MSE(\hat{\theta}) = Var(\hat{\theta}) \geq 1/n\mathbb{I}(\theta)$. Observe the similarity with the asymptotic normality of MLE. $1/n\mathbb{I}(\theta)$ is regarded as the ideal value of MSE. Efficiency of estimators are measured with respect to this ideal value. Efficiency of $\hat{\theta}$ is: $MSE(\hat{\theta})/\{1/n\mathbb{I}(\theta)\}$.

Most of the MLEs are 100% efficient. For example, consider p of $Ber(p)$. We determined MLE of p as: $\hat{p} = \bar{X}$. Then $MSE(\hat{p}) = Var(\bar{X}) + B^2[\bar{X}] = Var(X)/n + 0^2 = p(1-p)/n.$

Now, $f(X|p) = p^X(1-p)^{1-X} \Rightarrow f'(X|p) = Xp^{X-1}(1-p)^{1-X} - p^X(1-X)(1-p)^{-X}$

$$\Rightarrow \frac{f'(X|\theta)}{f(X|\theta)} = \frac{X}{p} - \frac{1-X}{1-p} = \frac{X-p}{p(1-p)} \Rightarrow \mathbb{I}(\theta) = E\left[\left(\frac{f'(X|\theta)}{f(X|\theta)}\right)^2\right] = E\left[\left(\frac{X-E[X]}{p(1-p)}\right)^2\right]$$

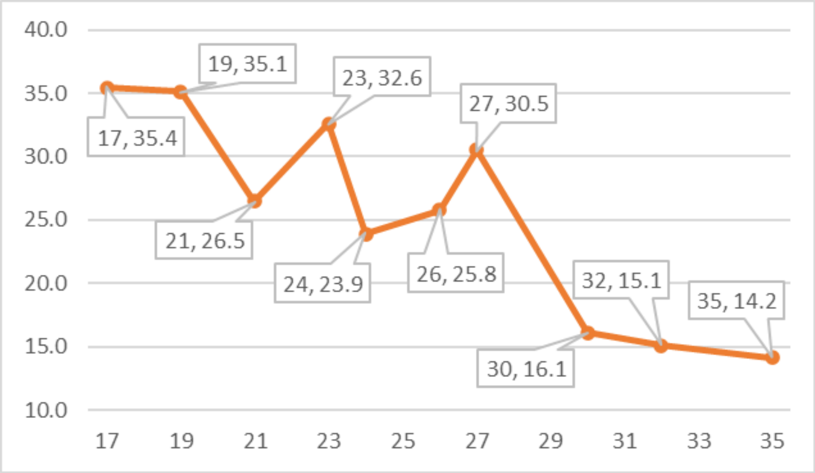
$$= \frac{Var(X)}{p^2(1-p)^2} = \frac{1}{p(1-p)} \Rightarrow \text{Efficiency} = \frac{MSE(\hat{p})}{1/n\mathbb{I}(\theta)} = \frac{p(1-p)/n}{p(1-p)/n} = 100\%.$$

Check efficiency of some of the other MLEs that we obtained.

Introduction to regression analysis

So far, we estimated some quantities associated with random variables. We did not consider influence of other factors on the quantity of interest. For example, consider yearly demand of smart phones in a state, denoted by Y, and we are interested in estimating its expected value $E[Y]$. Our knowledge of estimation tells that $\bar{Y} = \sum_{i=1}^{n} Y_i/n$ is a good estimator of $E[Y]$, where $Y_1, Y_2, \dots, Y_n$ denotes a simple random sample of Y. Let 14.2,15.1,30.5,16.1,25.8,23.9, 26.5,32.6,35.1,35.4 be the yearly demand of past ten years (in '000 units). Sample average is 25.5, which is our prediction for next (or any other) year's expected demand.

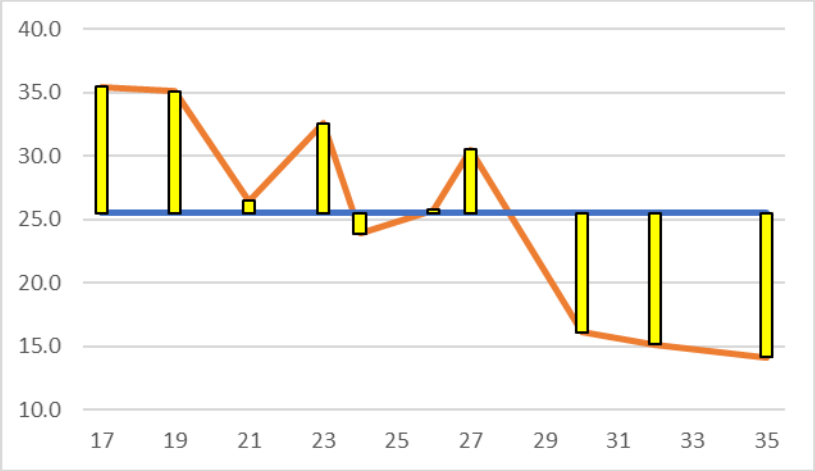
If we look closely at the data, we see an increasing trend. Out of 9 consecutive changes, 7 are of increasing type and 2 are of decreasing type. If variation in Y-values is entirely due to randomness, we would expect these changes to be more evenly matched. This suggests that something other than randomness may be influencing Y. When we look at the average prices of smartphones in these years (in '000 rupee), say 35,32,27,30,26,24,21,23,19,17, we notice that the prices have been generally declining. This may be contributing to the increasing trend in demand, as illustrated in the plot below. It does not show the data in chronological manner. Instead, it shows demand (in Y-axis) with respect to price (in X-axis).



With the above observation, it is appropriate that we use price information in our estimation of expected demand. For this, we need the relationship between price and demand. Economic theories on price and demand seem to suggest a linear relation between these two quantities. Let us consider the relation to be: $Y = a + bx + \epsilon$, where x denotes the price, a, b denote the determinant of the linear relation, and ϵ denote the randomness in Y that is not captured by the relation. This is a simple linear regression model. Here, the random sample is represented as collection of pairs $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$, where $Y_i = a + bx_i + \epsilon_i$ for $i = 1,2, \dots, n$.

Note that x is not random, and the randomness in Y is due to ϵ. It is reasonable to assume $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ to be *iid* random variables with zero expectation, which leads to $E[Y_i] = a + bx_i$ for all i and independence of $Y_1, Y_2, \dots, Y_n$. Note that $Y_1, Y_2, \dots, Y_n$ are not identical anymore. Here, we our objective is to estimate a and b, denoted by $\hat{a}$ and $\hat{b}$, from the data, and then our estimate for next year's expected demand is: $\hat{E}[Y_{n+1}] = \hat{a} + \hat{b}x_{n+1}$, where x_{n+1} denotes next year's average price (either known or predicted separately).

Now, $\hat{a}$ and $\hat{b}$ can assume any value leading to different amount of error, i.e., the deviation of expected model predictions $\hat{E}[Y_i] = \hat{a} + \hat{b}x_i$ from observations Y_i for $i = 1,2, \dots, n$. One such choice with $\hat{a} = \bar{Y}$ and $\hat{b} = 0$, leading to $\hat{E}[Y_i] = \bar{Y}$ for all i, is depicted below in blue. Actual observations and deviations are shown in orange and yellow respectively.



It is appropriate to choose $\hat{a}$ and $\hat{b}$ such that the total error is minimized. It is customary to consider squared distance as error. Then we have the following optimization problem:

$$\text{Maximize } E(\hat{a}, \hat{b}) = \sum_{i=1}^{n} \left(Y_i - \hat{E}[Y_i]\right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{a} - \hat{b}x_i\right)^2$$

$$\Rightarrow \frac{\partial E}{\partial \hat{a}} = \sum_{i=1}^{n} 2\left(Y_i - \hat{a} - \hat{b}x_i\right)(-1) = 2\hat{a}n - 2\sum_{i=1}^{n}\left(Y_i - \hat{b}x_i\right) = 2n\left[\hat{a} - \left(\bar{Y} - \hat{b}\bar{x}\right)\right]$$

$$\Rightarrow \frac{\partial E}{\partial \hat{b}} = \sum_{i=1}^{n} 2\left(Y_i - \hat{a} - \hat{b}x_i\right)(-x_i) = 2\left[\hat{b}\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n}(Y_i - \hat{a})x_i\right]$$

$$\Rightarrow H(\hat{a}, \hat{b}) = \begin{bmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2\sum_{i=1}^{n} x_i^2 \end{bmatrix} \sim \text{positive definite}$$

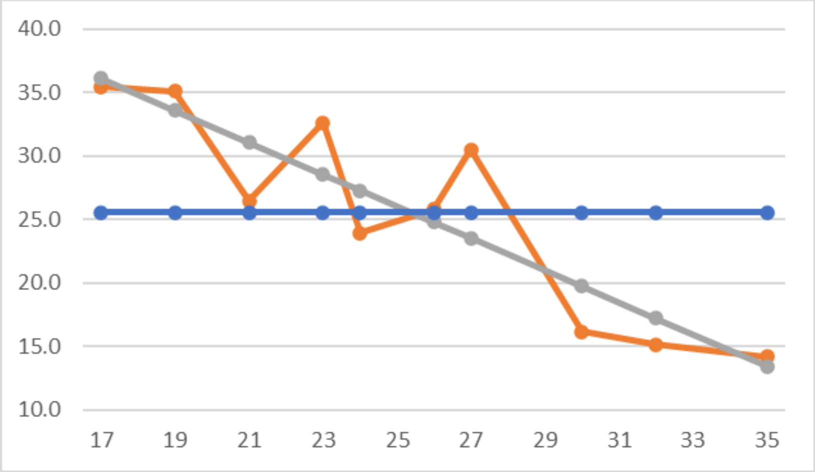
So, there is only one stationary point and that is the minima.

$$\frac{\partial E}{\partial \hat{a}} = 0 \Rightarrow \hat{a} = \bar{Y} - \hat{b}\bar{x} \;\; \left(\text{represented in terms of } \hat{b}\right)$$

$$\frac{\partial E}{\partial \hat{b}} = 0 \Rightarrow \hat{b}\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n}(Y_i - \hat{a})x_i = \sum_{i=1}^{n}\left(Y_i - \left(\bar{Y} - \hat{b}\bar{x}\right)\right)x_i = \sum_{i=1}^{n} Y_i x_i - \left(\bar{Y} - \hat{b}\bar{x}\right)\sum_{i=1}^{n} x_i$$

$$\Rightarrow \hat{b}\left(\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} Y_i x_i - \bar{Y}\sum_{i=1}^{n} x_i \Rightarrow \hat{b} = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

The above estimators are referred to as the least square estimators. With our data, $\hat{b} = -1.26$ and $\hat{a} = 57.6$. The corresponding regression line is shown below in grey.



It's evident that the least square estimators fit the data better than the blue line (corresponding to $\hat{a} = \bar{Y}$ and $\hat{b} = 0$). In fact, it can be shown that the least square estimators are best among all linear unbiased estimators of a and b. However, this goodness property is meaningful only within the regression model $Y = a + bx + \epsilon$. There can be other regression models, e.g., $Y = a + bx + cx^2 + \epsilon$, $\ln(Y) = a + b\ln(x) + \epsilon$, etc. and the associated estimations. We need to have a measure of model fitness. R^2-value provides one such measurement.

Without any regression model, our prediction for Y_i is $\bar{Y}$ for all i. This leads to a total squared error of $SS_{tot} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$, which we consider as benchmark. With a regression model, our prediction for Y_i is $\hat{E}[Y_i]$. This reduces the error, and the residual squared error is $SS_{res} = \sum_{i=1}^{n}\left(Y_i - \hat{E}[Y_i]\right)^2$. With these two error terms, coefficient of determination of the regression model is defined as: $R^2 = 1 - SS_{res}/SS_{tot}$. Note that $R^2 \in [0,1]$; higher the value of R^2, better is the model fit. In the above example, $SS_{tot} = 594.8$, $SS_{res} = 119.7$, and $R^2 = 0.799$ or 79.9%. There is lot more in regression analysis than this introduction. We have discussed only the core idea and the method of least squares.

Practice problems

Book-1: A Modern Introduction to Probability and Statistics by Dekking et al.

Bias and mean squared error
Book-1, Chapter-19, Exercise No. 1, 2, 5, 6

Book-1, Chapter-20, Exercise No. 3, 4, 5, 8, 10

Maximum likelihood method
Book-1, Chapter-21, Exercise No. 1, 3, 6, 7, 11, 14

Regression analysis
Book-1, Chapter-22, Exercise No. 3, 7, 10, 11