A Probability Primer

Johar M. Ashfaque

I. EXPECTATION

A. Expectation of a Random Variable

The expectation or the mean of a random variable X is the average value of X. The formal definition is as follows

Definition I.1 *The expected value or the mean or the first moment of X is defined to be*

$$\mathbb{E}(X) = \int x dF(x) = \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int_x x f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

assuming that the sum (or integral) is well-defined. We use the following notation to denoted the expected value of X:

$$\mathbb{E}(X) = \mathbb{E}X = \int x dF(x) = \mu = \mu_X.$$

The expectation if a one-number summary of the distribution. From now on whenever we discuss expectations, we assume that they exist. Let $Y = r(X)$. How do we compute $\mathbb{E}(Y)$?

Theorem I.2 *Let $Y = r(X)$. Then*

$$\mathbb{E}(Y) = \mathbb{E}(r(X)) = \int r(x) dF_X(x).$$

This result makes intuitive sense. Think of playing a game where we draw X at random and then I pay you $Y = r(X)$. Your average income is $r(x)$ times the chance that $X = x$ summed or integrated over all values of x. The k^{th} moment of X is defined to be $\mathbb{E}(X^k)$ assuming that $\mathbb{E}(X^k) < \infty$. We shall rarely make much use of moments beyond $k = 2$.

B. Properties of Expectation

Theorem I.3 *If $X_1, ..., X_n$ are random variables and $a_1, ..., a_n$ are constants then*

$$\mathbb{E}\Big(\sum_i a_i X_i\Big) = \sum_i a_i \mathbb{E}(X_i).$$

Theorem I.4 *Let $X_1, ..., X_n$ be independent random variables. Then*

$$\mathbb{E}\Big(\prod_i^n X_i\Big) = \prod_i \mathbb{E}(X_i).$$

Notice that the summation rule does not require independence but the multiplication does.

C. Variance and Covariance

The variance measures the spread of the distribution.

Definition I.5 *Let X be a random variable with mean μ. The variance of X denoted σ^2 or σ_X^2 or $\mathbb{V}(X)$ is defined by*

$$\sigma^2 = \mathbb{E}(X-\mu)^2 = \int (x-\mu)^2 dF(x)$$

assuming this expectation exists. The standard deviation is $sd(X) = \sigma$.

Theorem I.6 *Assuming the variance is well-defined, it has the following properties:*

1. $\mathbb{V}(X) = \mathbb{E}(X^2) - E(X)^2 = \mathbb{E}(X^2) - \mu^2$.
2. *If a and b are constants then $\mathbb{V}(aX+b) = a^2\mathbb{V}(X)$.*
3. *If $X_1, ..., X_n$ are independent and $a_1, ..., a_n$ are constants then*

$$\mathbb{V}\Big(\sum_{i=1}^n a_i X_i\Big) = \sum_{i=1}^n a_i^2 \mathbb{V}(X_i).$$

If $X_1, ..., X_n$ are random variables then we define the sample mean to be

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$$

and the sample variance to be

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2.$$

Theorem I.7 *Let $X_1, ..., X_n$ be IID and let $\mu = \mathbb{E}(X_i)$, $\sigma^2 = \mathbb{V}(X_i)$. Then*

$$\mathbb{E}(\overline{X}_n) = \mu, \qquad \mathbb{V}(\overline{X}_n) = \frac{\sigma^2}{n}, \qquad \mathbb{E}(S_n^2) = \sigma^2.$$

If X and Y are random variables then the covariance and correlation between X and Y measure how strong the linear relationship is between X and Y.

Definition I.8 *Let X and Y be random variables with means μ_X and μ_Y and standard deviations σ_X and σ_Y. Define the covariance between X and Y by*

$$Cov(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$$

and the correlation by

$$\rho = \rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}.$$

Theorem I.9 *The covariance satisfies*

$$Cov(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

The correlation satisfies

$$-1 \le \rho(X,Y) \le 1.$$

If $Y = a + bX$ for some constants a and b then $\rho(X, Y) = 1$ if $b > 0$ and $\rho(X, Y) = -1$ if $b < 0$. If X and Y are independent, then $\text{Cov}(X, Y) = \rho(X, Y) = 0$. The converse is not true in general.

Distribution	Mean	Variance
Point mass at a	a	0
Bernoulli (p)	p	$p(1-p)$
Binomial (n, p)	p	$np(1-p)$
Geometric (p)	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson (λ)	λ	λ
Uniform (a, b)	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal (μ, σ^2)	μ	σ^2
Exponential (β)	β	β^2
Gamma (α, β)	$\alpha\beta$	$\alpha\beta^2$
Beta (α, β)	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
t_ν	0 $(\nu > 1)$	$\frac{\nu}{\nu-2}$ $(\nu > 2)$
χ^2_p	p	$2p$

II. LINEAR REGRESSION

The term "regression" is due to Sir Francis Galton. Regression is a method for studying the relationship between a response variable Y and a covariate X. The covariate is also called a predictor variable or a feature. There can be one or more covariates. The data are of the form

$$(Y_1, X_1), ..., (Y_n, X_n).$$

One way to summarize the relationship between X and Y is through the regression function

$$r(x) = \mathbb{E}(Y|X = x) = \int y f(y|x) dy.$$

A. Simple Linear Regression

The simplest version of regression is when X_i is simple (a scalar not a vector) and $r(x)$ is assumed to be linear

$$r(x) = \beta_0 + \beta_1 x.$$

This model is called the simple linear regression model. Let $\epsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$. Then

$$\mathbb{E}(\epsilon_i|X_i) = 0.$$

Let $\sigma^2(x) = \mathbb{V}(\epsilon_i|X = x)$. We will make the further simplifying assumption that $\sigma^2(x) = \sigma^2$ does not depend on x. We can thus write the linear regression model as follows.

Definition II.1 *(The Linear Regression Model)*

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where

$$\mathbb{E}(\epsilon_i|X_i) = 0$$

and

$$\sigma^2(x) = \mathbb{V}(\epsilon_i|X_i).$$

The unknown parameters in the model are the intercept β_0 and the slope β_1 ad the variance σ^2 Let $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the estimates of β_0 and β_1 respectively. The fitted line is defined to be

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1(x).$$

The predicted values or fitted values are $\hat{Y}_i = \hat{r}(X_i)$ and the residuals are defined to be

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i).$$

The residual sums of squares or RSS is defined by

$$\text{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2.$$

The quantity RSS measures how well the fitted line fits the data.

Definition II.2 *The least squares estimates are the values $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize*

$$RSS = \sum_{i=1}^{n} \hat{\epsilon}_i^2.$$

Theorem II.3 *The least squares estimates are given by*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}$$
$$\hat{\beta}_0 = \overline{Y}_n - \hat{\beta}_1 \overline{X}_n.$$

An unbiased estimate of σ^2 is

$$\hat{\sigma}^2 = \left(\frac{1}{n-2}\right) \sum_{i=1}^{n} \hat{\epsilon}_i^2.$$

III. GRAPHICAL MODELS

Graphical models are a class of multivariate statistical models that useful for representing independence relations. Graphical models often require fewer parameters and may lead to estimators with smaller risk. There are two main types of graphical models: undirected and directed.

An undirected graph $\mathcal{G} = (V, E)$ has a finite set V of vertices and a set E of edges consisting of a pair of edges. The vertices correspond to random variables X, Y, Z,... and edges are written as unordered pairs. For example, $(X, Y) \in E$ means that X and Y are joined by an edge. Two vertices are adjacent denoted $X \sim Y$ if there is an edge between them. A graph is complete if there is an edge between every pair of vertices. A subset $U \subset V$ of vertices together with their edges is called a subgraph. If A, B and C are three distinct subsets of V, we say that C separates A and B if every path from a variable in A to a variable in B intersects a variable in C.

Directed graphs are similar to undirected graphs except that there are arrows between vertices instead of edges. Like undirected graphs, directed graphs can be used to represent independence relations. A directed graph $\mathcal{G}$ consists of a set of vertices V and an edge set E of ordered pairs of variables. If $(X, Y) \in E$ then there is an arrow pointing from X to Y. If an arrow connects two variables X and Y in either direction, we say that X and Y are adjacent. If there is an arrow from X to Y then X is a parent of Y and Y is the child of X. The set of all parents of X is denoted $\pi(X)$. A directed path from X to Y is a set of vertices beginning with X and ending with Y such that each pair is connected by an arrow and all the arrows point in the same direction. A sequence of adjacent vertices starting with X and ending with Y but ignoring the direction of the arrows is called an undirected path. X is an ancestor of Y if there is a directed path from X to Y. We also say that Y is a descendent of X. A directed path that starts and ends at the same variable is called a cycle. A directed graph is acyclic if it has no cycles. In thie case we say that the graph is a directed acyclic graph (DAG).

IV. LOG-LINEAR MODELS

Log-linear models are useful for modelling multivariate discrete data. There is a strong connection between log-linear models and undirected graphs.

A. The Log-Linear Model

Let $X = (X_1, ..., X_m)$ be a random vector with probability function

$$f(x) = \mathbb{P}(X = x) = \mathbb{P}(X_1 = x_1, ..., X_m = x_m)$$

where $x = (x_1, ..., x_m)$. Let r_j be the number of values that X_j takes. Without loss of generality, we can assume that $X_j \in \{0, 1, ..., r_j - 1\}$. Suppose now that we have n such random vectors. We can think of the data as a sample from a multinomial with $N = r_1 \times r_2 \times ... \times r_m$ categories. The data can be represented as counts in a $r_1 \times r_2 \times ... \times r_m$ table. Let $p = (p_1, ..., p_N)$ denote the multinomial parameter. Let $S = \{1, ..., m\}$. Given a vector $x = (x_1, ..., x_m)$ and a subset $A \subset S$, let $x_A = (x_j : j \in A)$. For example, if $A = \{1, 3\}$ then $x_A = (x_1, x_3)$.

Theorem IV.1 *The joint probability function $f(x)$ of a single random vector $X = (X_1, ..., X_m)$ can be written as*

$$\log f(x) = \sum_{A \subset S} \psi_A(x)$$

where the sum is over all subsets A of $S = \{1, ..., m\}$ and the ψ's satisfy the following conditions

- *$\psi_\emptyset(x)$ is a constant*
- *For every $A \subset S$, $\psi_A(x)$ is only a function of x_A and not the rest of the x_j's*
- *If $i \in A$ and $x_i = 0$ then $\psi_A(x) = 0$*

The formula

$$\log f(x) = \sum_{A \subset S} \psi_A(x)$$

is called the log-linear expansion of f. Note that this is the probability function for a single draw. Each $\psi_A(x)$ will depend on some unknown parameters β_A. Let $\beta = (\beta_A : A \subset S)$ be the set of all these parameters. We will write $f(x) = f(x; \beta)$ when we want to estimate the dependence on the unknown parameters β.

In terms of the multinomial, the parameter space is

$$\mathcal{P} = \left\{ p = (p_1, ..., p_N) : p_j \geq 0, \sum_{j=1}^{N} p_j = 1 \right\}.$$

This is the $N - 1$ dimensional space. In the log-linear representation, the parameter space is

$$\Theta = \left\{ \beta = (\beta_1, ..., \beta_N) : \beta = \beta(p), p \in \mathcal{P} \right\}$$

where $\beta(p)$ is the set of β values associated with p. The set Θ is a $N - 1$ dimensional surface in $\mathcal{R}^n$.

V. INTRODUCTION TO STOCHASTIC PROCESSES

We will consider sequences of dependent random variables. For example, daily temperatures will form a sequence of time-ordered random variables and clearly the temperature on day one is not independent of the temperature on day two.

A stochastic process $\{X_t : t \in T\}$ is a collection of random variables.The variables X_t take values in some set $\mathcal{X}$ called the state space. The set T is called the index set and for our purposes can be thought of as time. The index set can either be discrete or continuous.

Example 1. Let $\mathcal{X} = \{\text{sunny}, \text{cloudy}\}$. A typical sequence might be

$$\text{sunny}, \text{sunny}, \text{cloudy}, \text{sunny}, \text{cloudy}, \cdots.$$

This process has a discrete state space and a discrete index set.

Example 2. A sequence of IID random variables can be written as $\{X_t : t \in T\}$ where $T = \{1, 2, 3, ...\}$. Hence a sequence of IID random variables is an example of a stochastic process.

If $X_1, ..., X_n$ are random variables then we can write the joint density as

$$f(x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i|\text{past}_i)$$

where past_i refers to all the variables before X_i.

VI. MARKOV CHAINS

The simplest stochastic process is a Markov chain in which the distribution of X_t depends only on X_{t-1}. We will assume that the state space is discrete $\mathcal{X} = \{1, ..., N\}$ and that the index set is $T = \{0, 1, 2, ...\}$.

Definition VI.1 *The process $\{X_n : n \in T\}$ is a Markov chain if*

$$\mathbb{P}(X_n = x|X_0, ..., X_{n-1}) = \mathbb{P}(X_n = x|X_{n-1})$$

for all n and for all $x \in \mathcal{X}$.

For the Markov chain, the joint probability is

$$f(x_1, ..., x_n) = f(x_1)f(x_2|x_1)...f(x_n|x_{n-1}).$$

A. Transition Probabilities

The key quantities of a Markov chain are the probabilities of jumping from one state into another.

Definition VI.2 *We call*

$$\mathbb{P}(X_{n+1} = j|X_n = i)$$

the transition probabilities. If the transition probabilities do not change with time, we call the chain homogeneous. In this case, we define

$$p_{ij} = \mathbb{P}(X_{n+1} = j|X_n = i).$$

The matrix $\mathbf{P}$ whose (i, j) element is p_{ij} called the transition matrix.

We will only be interested in homogeneous Markov chains. Notice that $\mathbf{P}$ has two properties

- $p_{ij} \geq 0$
- $\sum_i p_{ij} = 1$

Each row is a probability mass function. A matrix with these properties is called a stochastic matrix. Let

$$p_{ij}(n) = \mathbb{P}(X_{m+n} = j | X_m = i)$$

be the probability of going from state i to state j in n steps. Let $\mathbf{P}_n$ be the matrix whose (i, j) element is $p_{ij}(n)$. These are called the n-step transition probabilities.

Theorem VI.3 *(The Chapman-Kolmogorov Equations) The n-step transition probabilities satisfy*

$$p_{ij}(m + n) = \sum_k p_{ik}(m) p_{kj}(n).$$

This statement of the theorem is nothing more than the equation for matrix multiplication. Hence

$$\mathbb{P}_{m+n} = \mathbb{P}_m \mathbb{P}_n.$$

1) Transition Matrix: $\mathbb{P}(i, j) = \mathbb{P}(X_{n+1} = j \| X_n = i)$.
2) n-step Matrix, $\mathbb{P}_n(i, j) = \mathbb{P}(X_{m+n} = j \| X_m = i)$.

TABLE I. Summary

B. States

The states of a Markov chain can be classified according to various properties.

Definition VI.4 *We say that i reaches j or j is accessible from I if $p_{ij}(n) > 0$ for some n and we write $i \to j$. If $i \to j$ and $j \to i$ then we write $i \leftrightarrow j$ and we say that i and j communicate.*

Theorem VI.5 *The communication relation satisfies the following properties:*

1. *$i \leftrightarrow i$.*
2. *If $i \leftrightarrow j$ then $j \leftrightarrow i$.*
3. *If $i \leftrightarrow j$ and $j \leftrightarrow k$ then $i \leftrightarrow k$.*
4. *The set of states $\mathcal{X}$ can be written as a disjoint union of classes $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \cdots$ where two states i and j communicate with each other if and only if they are in the same class.*

If all states communicate with each other then the chain is called irreducible. A set of states is closed if once you enter that set of states you never leave. A closed set consisting of a single state is called an absorbing state. Let $\mathcal{X} = \{1, 2, 3, 4\}$ and

$$\mathbf{P} = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The classes are $\{1, 2\}$, $\{3\}$ and $\{4\}$. State 4 is an absorbing state.

Theorem VI.6 *A state i is recurrent if and only if*

$$\sum_n p_{ii}(n) = \infty.$$

A state i is transient if and only if

$$\sum_n p_{ii}(n) < \infty.$$

Theorem VI.7 *Some facts.*

- *If state i is recurrent and $i \leftrightarrow j$ then j is recurrent.*
- *If state i is transient and $i \leftrightarrow j$ then j is transient.*
- *A finite Markov chain must have at least one recurrent state.*
- *The states of a finite, irreducible Markov chain are all recurrent.*

VII. POISSON PROCESSES

One of the most studied and useful stochastic processes is the Poisson process. It arises when we count occurrences of events over time. For example, traffic accidents, radioactive decay etc. As the name suggests the Poisson process is related to the Poisson distribution.

Definition VII.1 *A Poisson process is a stochastic process*

$$\{X_t : t \in [0, \infty)\}$$

with state space

$$\mathcal{X} = \{0, 1, 2, ...\}$$

such that

- $X(0) = 0$.
- *For any $0 = t_0 < t_1 < t_2 < \cdots < t_n$, the increments*

  $$X(t_1) - X(t_0), X(t_2) - X(t_1), \cdots, X(t_n) - X(t_{n-1})$$

  are independent.
- *There is a function $\lambda(t)$ called an intensity function. This is to say the number of events in any interval of length t is a Poisson random variable with parameter (or mean) λt.*

Definition VII.2 *A Poisson process with intensity function $\lambda(t) \equiv \lambda$ for some $\lambda > 0$ is called a homogeneous Poisson process with rate λ. In this case*

$$X(t) \sim Poisson(\lambda t).$$