Introduction to Probability

SECOND EDITION

Dimitri P. Bertsekas and John N. Tsitsiklis

Massachusetts Institute of Technology

WWW site for book information and orders

http://www.athenasc.com

Athena Scientific, Belmont, Massachusetts

1

Sample Space and Probability

Excerpts from Introduction to Probability: Second Edition

by Dimitri P. Bertsekas and John N. Tsitsiklis ©

Massachusetts Institute of Technology

Contents

1.1 SETS

1.2 PROBABILISTIC MODELS

Elements of a Probabilistic Model

- The **sample space** Ω, which is the set of all possible **outcomes** of an experiment.
- The **probability law**, which assigns to a set A of possible outcomes (also called an **event**) a nonnegative number $\mathbf{P}(A)$ (called the **probability** of A) that encodes our knowledge or belief about the collective "likelihood" of the elements of A. The probability law must satisfy certain properties to be introduced shortly.

Probability Axioms

1. **(Nonnegativity)** $\mathbf{P}(A) \geq 0$, for every event A.
2. **(Additivity)** If A and B are two disjoint events, then the probability of their union satisfies

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B).$$

More generally, if the sample space has an infinite number of elements and $A_1, A_2, \ldots$ is a sequence of disjoint events, then the probability of their union satisfies

$$\mathbf{P}(A_1 \cup A_2 \cup \cdots) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \cdots.$$

3. **(Normalization)** The probability of the entire sample space Ω is equal to 1, that is, $\mathbf{P}(\Omega) = 1$.

Discrete Probability Law

If the sample space consists of a finite number of possible outcomes, then the probability law is specified by the probabilities of the events that consist of a single element. In particular, the probability of any event $\{s_1, s_2, \ldots, s_n\}$ is the sum of the probabilities of its elements:

$$\mathbf{P}\big(\{s_1, s_2, \ldots, s_n\}\big) = \mathbf{P}(s_1) + \mathbf{P}(s_2) + \cdots + \mathbf{P}(s_n).$$

Discrete Uniform Probability Law

If the sample space consists of n possible outcomes which are equally likely (i.e., all single-element events have the same probability), then the probability of any event A is given by

$$\mathbf{P}(A) = \frac{\text{number of elements of } A}{n}.$$

Some Properties of Probability Laws

Consider a probability law, and let A, B, and C be events.

(a) If $A \subset B$, then $\mathbf{P}(A) \leq \mathbf{P}(B)$.

(b) $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$.

(c) $\mathbf{P}(A \cup B) \leq \mathbf{P}(A) + \mathbf{P}(B)$.

(d) $\mathbf{P}(A \cup B \cup C) = \mathbf{P}(A) + \mathbf{P}(A^c \cap B) + \mathbf{P}(A^c \cap B^c \cap C)$.

1.3 CONDITIONAL PROBABILITY

Properties of Conditional Probability

- The conditional probability of an event A, given an event B with $\mathbf{P}(B) > 0$, is defined by

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)},$$

  and specifies a new (conditional) probability law on the same sample space Ω. In particular, all properties of probability laws remain valid for conditional probability laws.
- Conditional probabilities can also be viewed as a probability law on a new universe B, because all of the conditional probability is concentrated on B.
- If the possible outcomes are finitely many and equally likely, then

$$\mathbf{P}(A \mid B) = \frac{\text{number of elements of } A \cap B}{\text{number of elements of } B}.$$

1.4 TOTAL PROBABILITY THEOREM AND BAYES' RULE

Total Probability Theorem

Let $A_1, \ldots, A_n$ be disjoint events that form a partition of the sample space (each possible outcome is included in exactly one of the events $A_1, \ldots, A_n$) and assume that $\mathbf{P}(A_i) > 0$, for all i. Then, for any event B, we have

$$\begin{aligned}\mathbf{P}(B) &= \mathbf{P}(A_1 \cap B) + \cdots + \mathbf{P}(A_n \cap B) \\ &= \mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \cdots + \mathbf{P}(A_n)\mathbf{P}(B \mid A_n).\end{aligned}$$

1.5 INDEPENDENCE

Independence

- Two events A and B are said to be **independent** if

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B).$$

  If in addition, $\mathbf{P}(B) > 0$, independence is equivalent to the condition

$$\mathbf{P}(A \,|\, B) = \mathbf{P}(A).$$

- If A and B are independent, so are A and B^c.
- Two events A and B are said to be **conditionally independent**, given another event C with $\mathbf{P}(C) > 0$, if

$$\mathbf{P}(A \cap B \,|\, C) = \mathbf{P}(A \,|\, C)\mathbf{P}(B \,|\, C).$$

  If in addition, $\mathbf{P}(B \cap C) > 0$, conditional independence is equivalent to the condition

$$\mathbf{P}(A \,|\, B \cap C) = \mathbf{P}(A \,|\, C).$$

- Independence does not imply conditional independence, and vice versa.

Definition of Independence of Several Events

We say that the events $A_1, A_2, \ldots, A_n$ are **independent** if

$$\mathbf{P}\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} \mathbf{P}(A_i), \qquad \text{for every subset } S \text{ of } \{1, 2, \ldots, n\}.$$

1.6 COUNTING

The Counting Principle

Consider a process that consists of r stages. Suppose that:

(a) There are n_1 possible results at the first stage.

(b) For every possible result at the first stage, there are n_2 possible results at the second stage.

(c) More generally, for any sequence of possible results at the first $i-1$ stages, there are n_i possible results at the ith stage.

Then, the total number of possible results of the r-stage process is

$$n_1 n_2 \cdots n_r.$$

Summary of Counting Results

- **Permutations** of n objects: $n!$.
- **k-permutations** of n objects: $n!/(n-k)!$.
- **Combinations** of k out of n objects: $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$.
- **Partitions** of n objects into r groups, with the ith group having n_i objects:

$$\binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1!\,n_2!\cdots n_r!}.$$

1.7 SUMMARY AND DISCUSSION

2

Discrete Random Variables

Excerpts from Introduction to Probability: Second Edition

by Dimitri P. Bertsekas and John N. Tsitsiklis ©

Massachusetts Institute of Technology

Contents

2.1 BASIC CONCEPTS

Main Concepts Related to Random Variables

Starting with a probabilistic model of an experiment:

- A **random variable** is a real-valued function of the outcome of the experiment.
- A **function of a random variable** defines another random variable.
- We can associate with each random variable certain "averages" of interest, such as the **mean** and the **variance**.
- A random variable can be **conditioned** on an event or on another random variable.
- There is a notion of **independence** of a random variable from an event or from another random variable.

Concepts Related to Discrete Random Variables

Starting with a probabilistic model of an experiment:

- A **discrete random variable** is a real-valued function of the outcome of the experiment that can take a finite or countably infinite number of values.
- A discrete random variable has an associated **probability mass function (PMF)**, which gives the probability of each numerical value that the random variable can take.
- A **function of a discrete random variable** defines another discrete random variable, whose PMF can be obtained from the PMF of the original random variable.

2.2 PROBABILITY MASS FUNCTIONS

Calculation of the PMF of a Random Variable X

For each possible value x of X:

1. Collect all the possible outcomes that give rise to the event $\{X = x\}$.
2. Add their probabilities to obtain $p_X(x)$.

2.3 FUNCTIONS OF RANDOM VARIABLES

2.4 EXPECTATION, MEAN, AND VARIANCE

Expectation

We define the **expected value** (also called the **expectation** or the **mean**) of a random variable X, with PMF p_X, by

$$\mathbf{E}[X] = \sum_x x p_X(x).$$

Expected Value Rule for Functions of Random Variables

Let X be a random variable with PMF p_X, and let $g(X)$ be a function of X. Then, the expected value of the random variable $g(X)$ is given by

$$\mathbf{E}\big[g(X)\big] = \sum_x g(x) p_X(x).$$

Variance

The variance $\mathrm{var}(X)$ of a random variable X is defined by

$$\mathrm{var}(X) = \mathbf{E}\left[\big(X - \mathbf{E}[X]\big)^2\right],$$

and can be calculated as

$$\mathrm{var}(X) = \sum_x \big(x - \mathbf{E}[X]\big)^2 p_X(x).$$

It is always nonnegative. Its square root is denoted by σ_X and is called the **standard deviation**.

Mean and Variance of a Linear Function of a Random Variable

Let X be a random variable and let

$$Y = aX + b,$$

where a and b are given scalars. Then,

$$\mathbf{E}[Y] = a\mathbf{E}[X] + b, \qquad \mathrm{var}(Y) = a^2\,\mathrm{var}(X).$$

Variance in Terms of Moments Expression

$$\mathrm{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2.$$

2.5 JOINT PMFS OF MULTIPLE RANDOM VARIABLES

Summary of Facts About Joint PMFs

Let X and Y be random variables associated with the same experiment.

- The **joint PMF** $p_{X,Y}$ of X and Y is defined by

$$p_{X,Y}(x,y) = \mathbf{P}(X = x, Y = y).$$

- The **marginal PMF**s of X and Y can be obtained from the joint PMF, using the formulas

$$p_X(x) = \sum_y p_{X,Y}(x,y), \qquad p_Y(y) = \sum_x p_{X,Y}(x,y).$$

- A function $g(X,Y)$ of X and Y defines another random variable, and

$$\mathbf{E}\big[g(X,Y)\big] = \sum_x \sum_y g(x,y)p_{X,Y}(x,y).$$

  If g is linear, of the form $aX + bY + c$, we have

$$\mathbf{E}[aX + bY + c] = a\mathbf{E}[X] + b\mathbf{E}[Y] + c.$$

- The above have natural extensions to the case where more than two random variables are involved.

2.6 CONDITIONING

Summary of Facts About Conditional PMFs

Let X and Y be random variables associated with the same experiment.

- Conditional PMFs are similar to ordinary PMFs, but pertain to a universe where the conditioning event is known to have occurred.
- The conditional PMF of X given an event A with $\mathbf{P}(A) > 0$, is defined by
$$p_{X|A}(x) = \mathbf{P}(X = x \,|\, A)$$
and satisfies
$$\sum_x p_{X|A}(x) = 1.$$
- If $A_1, \ldots, A_n$ are disjoint events that form a partition of the sample space, with $\mathbf{P}(A_i) > 0$ for all i, then
$$p_X(x) = \sum_{i=1}^{n} \mathbf{P}(A_i) p_{X|A_i}(x).$$
(This is a special case of the total probability theorem.) Furthermore, for any event B, with $\mathbf{P}(A_i \cap B) > 0$ for all i, we have
$$p_{X|B}(x) = \sum_{i=1}^{n} \mathbf{P}(A_i \,|\, B) p_{X|A_i \cap B}(x).$$
- The conditional PMF of X given $Y = y$ is related to the joint PMF by
$$p_{X,Y}(x, y) = p_Y(y) p_{X|Y}(x \,|\, y).$$
- The conditional PMF of X given Y can be used to calculate the marginal PMF of X through the formula
$$p_X(x) = \sum_y p_Y(y) p_{X|Y}(x \,|\, y).$$
- There are natural extensions of the above involving more than two random variables.

Summary of Facts About Conditional Expectations

Let X and Y be random variables associated with the same experiment.

- The conditional expectation of X given an event A with $\mathbf{P}(A) > 0$, is defined by
$$\mathbf{E}[X \mid A] = \sum_x x p_{X|A}(x).$$
For a function $g(X)$, we have
$$\mathbf{E}\big[g(X) \mid A\big] = \sum_x g(x) p_{X|A}(x).$$
- The conditional expectation of X given a value y of Y is defined by
$$\mathbf{E}[X \mid Y = y] = \sum_x x p_{X|Y}(x \mid y).$$
- If $A_1, \ldots, A_n$ be disjoint events that form a partition of the sample space, with $\mathbf{P}(A_i) > 0$ for all i, then
$$\mathbf{E}[X] = \sum_{i=1}^{n} \mathbf{P}(A_i)\mathbf{E}[X \mid A_i].$$
Furthermore, for any event B with $\mathbf{P}(A_i \cap B) > 0$ for all i, we have
$$\mathbf{E}[X \mid B] = \sum_{i=1}^{n} \mathbf{P}(A_i \mid B)\mathbf{E}[X \mid A_i \cap B].$$
- We have
$$\mathbf{E}[X] = \sum_y p_Y(y)\mathbf{E}[X \mid Y = y].$$

2.7 INDEPENDENCE

Summary of Facts About Independent Random Variables

Let A be an event, with $\mathbf{P}(A) > 0$, and let X and Y be random variables associated with the same experiment.

- X is independent of the event A if

$$p_{X|A}(x) = p_X(x), \qquad \text{for all } x,$$

  that is, if for all x, the events $\{X = x\}$ and A are independent.

- X and Y are independent if for all pairs (x, y), the events $\{X = x\}$ and $\{Y = y\}$ are independent, or equivalently

$$p_{X,Y}(x, y) = p_X(x)p_Y(y), \qquad \text{for all } x, y.$$

- If X and Y are independent random variables, then

$$\mathbf{E}[XY] = \mathbf{E}[X]\,\mathbf{E}[Y].$$

  Furthermore, for any functions g and h, the random variables $g(X)$ and $h(Y)$ are independent, and we have

$$\mathbf{E}\big[g(X)h(Y)\big] = \mathbf{E}\big[g(X)\big]\,\mathbf{E}\big[h(Y)\big].$$

- If X and Y are independent, then

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y).$$

2.8 SUMMARY AND DISCUSSION

Summary of Results for Special Random Variables

Discrete Uniform over $[a,b]$:

$$p_X(k) = \begin{cases} \dfrac{1}{b-a+1}, & \text{if } k = a, a+1, \ldots, b, \\ 0, & \text{otherwise,} \end{cases}$$

$$\mathbf{E}[X] = \frac{a+b}{2}, \qquad \text{var}(X) = \frac{(b-a)(b-a+2)}{12}.$$

Bernoulli with Parameter p: (Describes the success or failure in a single trial.)

$$p_X(k) = \begin{cases} p, & \text{if } k = 1, \\ 1-p, & \text{if } k = 0, \end{cases}$$

$$\mathbf{E}[X] = p, \qquad \text{var}(X) = p(1-p).$$

Binomial with Parameters p and n: (Describes the number of successes in n independent Bernoulli trials.)

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n,$$

$$\mathbf{E}[X] = np, \qquad \text{var}(X) = np(1-p).$$

Geometric with Parameter p: (Describes the number of trials until the first success, in a sequence of independent Bernoulli trials.)

$$p_X(k) = (1-p)^{k-1} p, \qquad k = 1, 2, \ldots,$$

$$\mathbf{E}[X] = \frac{1}{p}, \qquad \text{var}(X) = \frac{1-p}{p^2}.$$

Poisson with Parameter λ: (Approximates the binomial PMF when n is large, p is small, and $\lambda = np$.)

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \qquad k = 0, 1, \ldots,$$

$$\mathbf{E}[X] = \lambda, \qquad \text{var}(X) = \lambda.$$

3

General Random Variables

Excerpts from Introduction to Probability: Second Edition

by Dimitri P. Bertsekas and John N. Tsitsiklis ©

Massachusetts Institute of Technology

Contents

3.1 CONTINUOUS RANDOM VARIABLES AND PDFS

Summary of PDF Properties

Let X be a continuous random variable with PDF f_X.

- $f_X(x) \geq 0$ for all x.
- $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$.
- If δ is very small, then $\mathbf{P}\big([x, x+\delta]\big) \approx f_X(x) \cdot \delta$.
- For any subset B of the real line,

$$\mathbf{P}(X \in B) = \int_B f_X(x)\,dx.$$

Expectation of a Continuous Random Variable and its Properties

Let X be a continuous random variable with PDF f_X.

- The expectation of X is defined by

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx.$$

- The expected value rule for a function $g(X)$ has the form

$$\mathbf{E}\big[g(X)\big] = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx.$$

- The variance of X is defined by

$$\text{var}(X) = \mathbf{E}\Big[\big(X - \mathbf{E}[X]\big)^2\Big] = \int_{-\infty}^{\infty} \big(x - \mathbf{E}[X]\big)^2 f_X(x)\,dx.$$

- We have

$$0 \leq \text{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2.$$

- If $Y = aX + b$, where a and b are given scalars, then

$$\mathbf{E}[Y] = a\mathbf{E}[X] + b, \qquad \text{var}(Y) = a^2 \text{var}(X).$$

3.2 CUMULATIVE DISTRIBUTION FUNCTIONS

Properties of a CDF

The CDF F_X of a random variable X is defined by

$$F_X(x) = \mathbf{P}(X \le x), \qquad \text{for all } x,$$

and has the following properties.

- F_X is monotonically nondecreasing:

$$\text{if } x \le y, \text{ then } F_X(x) \le F_X(y).$$

- $F_X(x)$ tends to 0 as $x \to -\infty$, and to 1 as $x \to \infty$.
- If X is discrete, then $F_X(x)$ is a piecewise constant function of x.
- If X is continuous, then $F_X(x)$ is a continuous function of x.
- If X is discrete and takes integer values, the PMF and the CDF can be obtained from each other by summing or differencing:

$$F_X(k) = \sum_{i=-\infty}^{k} p_X(i),$$

$$p_X(k) = \mathbf{P}(X \le k) - \mathbf{P}(X \le k-1) = F_X(k) - F_X(k-1),$$

  for all integers k.
- If X is continuous, the PDF and the CDF can be obtained from each other by integration or differentiation:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt, \qquad f_X(x) = \frac{dF_X}{dx}(x).$$

  (The second equality is valid for those x at which the PDF is continuous.)

3.3 NORMAL RANDOM VARIABLES

Normality is Preserved by Linear Transformations

If X is a normal random variable with mean μ and variance σ^2, and if $a \neq 0$, b are scalars, then the random variable

$$Y = aX + b$$

is also normal, with mean and variance

$$\mathbf{E}[Y] = a\mu + b, \qquad \mathrm{var}(Y) = a^2\sigma^2.$$

CDF Calculation for a Normal Random Variable

For a normal random variable X with mean μ and variance σ^2, we use a two-step procedure.

(a) "Standardize" X, i.e., subtract μ and divide by σ to obtain a standard normal random variable Y.

(b) Read the CDF value from the standard normal table:

$$\mathbf{P}(X \leq x) = \mathbf{P}\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = \mathbf{P}\left(Y \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

The standard normal table. The entries in this table provide the numerical values of $\Phi(y) = \mathbf{P}(Y \leq y)$, where Y is a standard normal random variable, for y between 0 and 3.49. For example, to find $\Phi(1.71)$, we look at the row corresponding to 1.7 and the column corresponding to 0.01, so that $\Phi(1.71) = .9564$. When y is negative, the value of $\Phi(y)$ can be found using the formula $\Phi(y) = 1 - \Phi(-y)$.

3.4 JOINT PDFS OF MULTIPLE RANDOM VARIABLES

Summary of Facts about Joint PDFs

Let X and Y be jointly continuous random variables with joint PDF $f_{X,Y}$.

- The **joint PDF** is used to calculate probabilities:

$$\mathbf{P}\big((X,Y) \in B\big) = \int\int_{(x,y)\in B} f_{X,Y}(x,y)\,dx\,dy.$$

- The **marginal PDF**s of X and Y can be obtained from the joint PDF, using the formulas

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy, \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx.$$

- The **joint CDF** is defined by $F_{X,Y}(x,y) = \mathbf{P}(X \leq x,\, Y \leq y)$, and determines the joint PDF through the formula

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}}{\partial x \partial y}(x,y),$$

  for every (x,y) at which the joint PDF is continuous.

- A function $g(X,Y)$ of X and Y defines a new random variable, and

$$\mathbf{E}\big[g(X,Y)\big] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dx\,dy.$$

  If g is linear, of the form $aX + bY + c$, we have

$$\mathbf{E}[aX + bY + c] = a\mathbf{E}[X] + b\mathbf{E}[Y] + c.$$

- The above have natural extensions to the case where more than two random variables are involved.

3.5 CONDITIONING

Conditional PDF Given an Event

- The conditional PDF $f_{X|A}$ of a continuous random variable X, given an event A with $\mathbf{P}(A) > 0$, satisfies

$$\mathbf{P}(X \in B \,|\, A) = \int_B f_{X|A}(x)\,dx.$$

- If A is a subset of the real line with $\mathbf{P}(X \in A) > 0$, then

$$f_{X|\{X \in A\}}(x) = \begin{cases} \dfrac{f_X(x)}{\mathbf{P}(X \in A)}, & \text{if } x \in A, \\ 0, & \text{otherwise.} \end{cases}$$

- Let $A_1, A_2, \ldots, A_n$ be disjoint events that form a partition of the sample space, and assume that $\mathbf{P}(A_i) > 0$ for all i. Then,

$$f_X(x) = \sum_{i=1}^{n} \mathbf{P}(A_i) f_{X|A_i}(x)$$

(a version of the total probability theorem).

Conditional PDF Given a Random Variable

Let X and Y be jointly continuous random variables with joint PDF $f_{X,Y}$.

- The joint, marginal, and conditional PDFs are related to each other by the formulas

$$f_{X,Y}(x, y) = f_Y(y) f_{X|Y}(x \,|\, y),$$

$$f_X(x) = \int_{-\infty}^{\infty} f_Y(y) f_{X|Y}(x \,|\, y)\,dy.$$

The conditional PDF $f_{X|Y}(x \,|\, y)$ is defined only for those y for which $f_Y(y) > 0$.

- We have

$$\mathbf{P}(X \in A \,|\, Y = y) = \int_A f_{X|Y}(x \,|\, y)\,dx.$$

Summary of Facts About Conditional Expectations

Let X snd Y be jointly continuous random variables, and let A be an event with $\mathbf{P}(A) > 0$.

- **Definitions:** The conditional expectation of X given the event A is defined by
$$\mathbf{E}[X \mid A] = \int_{-\infty}^{\infty} x f_{X|A}(x)\,dx.$$
The conditional expectation of X given that $Y = y$ is defined by
$$\mathbf{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\,dx.$$

- **The expected value rule:** For a function $g(X)$, we have
$$\mathbf{E}\big[g(X) \mid A\big] = \int_{-\infty}^{\infty} g(x) f_{X|A}(x)\,dx,$$
and
$$\mathbf{E}\big[g(X) \mid Y = y\big] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x \mid y)\,dx.$$

- **Total expectation theorem:** Let $A_1, A_2, \ldots, A_n$ be disjoint events that form a partition of the sample space, and assume that $\mathbf{P}(A_i) > 0$ for all i. Then,
$$\mathbf{E}[X] = \sum_{i=1}^{n} \mathbf{P}(A_i)\mathbf{E}[X \mid A_i].$$
Similarly,
$$\mathbf{E}[X] = \int_{-\infty}^{\infty} \mathbf{E}[X \mid Y = y] f_Y(y)\,dy.$$

- There are natural analogs for the case of functions of several random variables. For example,
$$\mathbf{E}\big[g(X,Y) \mid Y = y\big] = \int g(x,y) f_{X|Y}(x \mid y)\,dx,$$
and
$$\mathbf{E}\big[g(X,Y)\big] = \int \mathbf{E}\big[g(X,Y) \mid Y = y\big] f_Y(y)\,dy.$$

Independence of Continuous Random Variables

Let X and Y be jointly continuous random variables.

- X and Y are **independent** if

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \qquad \text{for all } x, y.$$

- If X and Y are independent, then

$$\mathbf{E}[XY] = \mathbf{E}[X]\,\mathbf{E}[Y].$$

  Furthermore, for any functions g and h, the random variables $g(X)$ and $h(Y)$ are independent, and we have

$$\mathbf{E}\big[g(X)h(Y)\big] = \mathbf{E}\big[g(X)\big]\,\mathbf{E}\big[h(Y)\big].$$

- If X and Y are independent, then

$$\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y).$$

3.6 BAYES' RULE AND APPLICATIONS IN INFERENCE

Bayes' Rule Relations for Random Variables

Let X and Y be two random variables.

- If X and Y are discrete, we have for all x, y with $p_X(x) \neq 0$, $p_Y(y) \neq 0$,

$$p_X(x)p_{Y|X}(y\,|\,x) = p_Y(y)p_{X|Y}(x\,|\,y),$$

  and the terms on the two sides in this relation are both equal to

$$p_{X,Y}(x,y).$$

- If X is discrete and Y is continuous, we have for all x, y with $p_X(x) \neq 0$, $f_Y(y) \neq 0$,

$$p_X(x)f_{Y|X}(y\,|\,x) = f_Y(y)p_{X|Y}(x\,|\,y),$$

  and the terms on the two sides in this relation are both equal to

$$\lim_{\delta\to 0} \frac{\mathbf{P}(X = x,\, y \leq Y \leq y + \delta)}{\delta}.$$

- If X and Y are continuous, we have for all x, y with $f_X(x) \neq 0$, $f_Y(y) \neq 0$,

$$f_X(x)f_{Y|X}(y\,|\,x) = f_Y(y)f_{X|Y}(x\,|\,y),$$

  and the terms on the two sides in this relation are both equal to

$$\lim_{\delta\to 0} \frac{\mathbf{P}(x \leq X \leq x + \delta,\, y \leq Y \leq y + \delta)}{\delta^2}.$$

3.7 SUMMARY AND DISCUSSION

Summary of Results for Special Random Variables

Continuous Uniform Over $[a, b]$:

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & \text{if } a \le x \le b, \\ 0, & \text{otherwise,} \end{cases}$$

$$\mathbf{E}[X] = \frac{a+b}{2}, \qquad \text{var}(X) = \frac{(b-a)^2}{12}.$$

Exponential with Parameter λ:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \ge 0, \\ 0, & \text{otherwise,} \end{cases} \qquad F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{if } x \ge 0, \\ 0, & \text{otherwise,} \end{cases}$$

$$\mathbf{E}[X] = \frac{1}{\lambda}, \qquad \text{var}(X) = \frac{1}{\lambda^2}.$$

Normal with Parameters μ and $\sigma^2 > 0$:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/2\sigma^2},$$

$$\mathbf{E}[X] = \mu, \qquad \text{var}(X) = \sigma^2.$$

4

Further Topics on Random Variables

Excerpts from Introduction to Probability: Second Edition

by Dimitri P. Bertsekas and John N. Tsitsiklis ©

Massachusetts Institute of Technology

Contents

4.1 DERIVED DISTRIBUTIONS

Calculation of the PDF of a Function $Y = g(X)$ of a Continuous Random Variable X

1. Calculate the CDF F_Y of Y using the formula

$$F_Y(y) = \mathbf{P}\big(g(X) \le y\big) = \int_{\{x \mid g(x) \le y\}} f_X(x)\,dx.$$

2. Differentiate to obtain the PDF of Y:

$$f_Y(y) = \frac{dF_Y}{dy}(y).$$

The PDF of a Linear Function of a Random Variable

Let X be a continuous random variable with PDF f_X, and let

$$Y = aX + b,$$

where a and b are scalars, with $a \neq 0$. Then,

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

PDF Formula for a Strictly Monotonic Function of a Continuous Random Variable

Suppose that g is strictly monotonic and that for some function h and all x in the range of X we have

$$y = g(x) \qquad \text{if and only if} \qquad x = h(y).$$

Assume that h is differentiable. Then, the PDF of Y in the region where $f_Y(y) > 0$ is given by

$$f_Y(y) = f_X\big(h(y)\big)\left|\frac{dh}{dy}(y)\right|.$$

4.2 COVARIANCE AND CORRELATION

Covariance and Correlation

- The **covariance** of X and Y is given by

$$\text{cov}(X,Y) = \mathbf{E}\Big[\big(X - \mathbf{E}[X]\big)\big(Y - \mathbf{E}[Y]\big)\Big] = \mathbf{E}[XY] - \mathbf{E}[X]\,\mathbf{E}[Y].$$

- If $\text{cov}(X,Y) = 0$, we say that X and Y are **uncorrelated**.
- If X and Y are independent, they are uncorrelated. The converse is not always true.
- We have

$$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X,Y).$$

- The **correlation coefficient** $\rho(X,Y)$ of two random variables X and Y with positive variances is defined by

$$\rho(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}},$$

  and satisfies

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

4.3 CONDITIONAL EXPECTATION AND VARIANCE REVISITED

Law of Iterated Expectations: $\mathbf{E}\big[\mathbf{E}[X \mid Y]\big] = \mathbf{E}[X]$.

Law of Total Variance: $\text{var}(X) = \mathbf{E}\big[\text{var}(X \mid Y)\big] + \text{var}\big(\mathbf{E}[X \mid Y]\big)$.

Properties of the Conditional Expectation and Variance

- $\mathbf{E}[X \mid Y = y]$ is a number whose value depends on y.
- $\mathbf{E}[X \mid Y]$ is a function of the random variable Y, hence a random variable. Its value is $\mathbf{E}[X \mid Y = y]$ whenever the value of Y is y.
- $\mathbf{E}\big[\mathbf{E}[X \mid Y]\big] = \mathbf{E}[X]$ (**law of iterated expectations**).
- $\mathbf{E}[X \mid Y = y]$ may be viewed as an estimate of X given $Y = y$. The corresponding error $\mathbf{E}[X \mid Y] - X$ is a zero mean random variable that is uncorrelated with $\mathbf{E}[X \mid Y]$.
- $\text{var}(X \mid Y)$ is a random variable whose value is $\text{var}(X \mid Y = y)$ whenever the value of Y is y.
- $\text{var}(X) = \mathbf{E}\big[\text{var}(X \mid Y)\big] + \text{var}\big(\mathbf{E}[X \mid Y]\big)$ (**law of total variance**).

4.4 TRANSFORMS

Summary of Transforms and their Properties

- The transform associated with a random variable X is given by

$$M_X(s) = \mathbf{E}[e^{sX}] = \begin{cases} \displaystyle\sum_x e^{sx} p_X(x), & X \text{ discrete,} \\ \displaystyle\int_{-\infty}^{\infty} e^{sx} f_X(x)\,dx, & X \text{ continuous.} \end{cases}$$

- The distribution of a random variable is completely determined by the corresponding transform.
- Moment generating properties:

$$M_X(0) = 1, \qquad \frac{d}{ds} M_X(s)\bigg|_{s=0} = \mathbf{E}[X], \qquad \frac{d^n}{ds^n} M_X(s)\bigg|_{s=0} = \mathbf{E}[X^n].$$

- If $Y = aX + b$, then $M_Y(s) = e^{sb} M_X(as)$.
- If X and Y are independent, then $M_{X+Y}(s) = M_X(s) M_Y(s)$.

Transforms for Common Discrete Random Variables

Bernoulli(p) $(k = 0, 1)$

$$p_X(k) = \begin{cases} p, & \text{if } k = 1, \\ 1-p, & \text{if } k = 0, \end{cases} \qquad M_X(s) = 1 - p + pe^s.$$

Binomial(n, p) $(k = 0, 1, \ldots, n)$

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad M_X(s) = (1 - p + pe^s)^n.$$

Geometric(p) $(k = 1, 2, \ldots)$

$$p_X(k) = p(1-p)^{k-1}, \qquad M_X(s) = \frac{pe^s}{1 - (1-p)e^s}.$$

Poisson(λ) $(k = 0, 1, \ldots)$

$$p_X(k) = \frac{e^{-\lambda}\lambda^k}{k!}, \qquad M_X(s) = e^{\lambda(e^s - 1)}.$$

Uniform(a, b) $(k = a, a+1, \ldots, b)$

$$p_X(k) = \frac{1}{b - a + 1}, \qquad M_X(s) = \frac{e^{sa}\left(e^{s(b-a+1)} - 1\right)}{(b-a+1)(e^s - 1)}.$$

Transforms for Common Continuous Random Variables

Uniform(a, b) $(a \le x \le b)$

$$f_X(x) = \frac{1}{b - a}, \qquad M_X(s) = \frac{e^{sb} - e^{sa}}{s(b-a)}.$$

Exponential(λ) $(x \ge 0)$

$$f_X(x) = \lambda e^{-\lambda x}, \qquad M_X(s) = \frac{\lambda}{\lambda - s}, \qquad (s < \lambda).$$

Normal(μ, σ^2) $(-\infty < x < \infty)$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/2\sigma^2}, \qquad M_X(s) = e^{(\sigma^2 s^2/2) + \mu s}.$$

4.5 SUM OF A RANDOM NUMBER OF INDEPENDENT RANDOM VARIABLES

Properties of the Sum of a Random Number of Independent Random Variables

Let $X_1, X_2, \ldots$ be identically distributed random variables with mean $\mathbf{E}[X]$ and variance $\mathrm{var}(X)$. Let N be a random variable that takes nonnegative integer values. We assume that all of these random variables are independent, and we consider the sum

$$Y = X_1 + \cdots + X_N.$$

Then:

- $\mathbf{E}[Y] = \mathbf{E}[N]\,\mathbf{E}[X]$.
- $\mathrm{var}(Y) = \mathbf{E}[N]\,\mathrm{var}(X) + \big(\mathbf{E}[X]\big)^2 \mathrm{var}(N)$.
- We have
$$M_Y(s) = M_N\big(\log M_X(s)\big).$$
  Equivalently, the transform $M_Y(s)$ is found by starting with the transform $M_N(s)$ and replacing each occurrence of e^s with $M_X(s)$.

4.6 SUMMARY AND DISCUSSION

5

Limit Theorems

Excerpts from Introduction to Probability: Second Edition

by Dimitri P. Bertsekas and John N. Tsitsiklis ©

Massachusetts Institute of Technology

Contents

5.1 MARKOV AND CHEBYSHEV INEQUALITIES

Markov Inequality

If a random variable X can only take nonnegative values, then

$$\mathbf{P}(X \geq a) \leq \frac{\mathbf{E}[X]}{a}, \qquad \text{for all } a > 0.$$

Chebyshev Inequality

If X is a random variable with mean μ and variance σ^2, then

$$\mathbf{P}\big(|X - \mu| \geq c\big) \leq \frac{\sigma^2}{c^2}, \qquad \text{for all } c > 0.$$

5.2 THE WEAK LAW OF LARGE NUMBERS

The Weak Law of Large Numbers

Let $X_1, X_2, \ldots$ be independent identically distributed random variables with mean μ. For every $\epsilon > 0$, we have

$$\mathbf{P}\big(|M_n - \mu| \geq \epsilon\big) = \mathbf{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right) \to 0, \qquad \text{as } n \to \infty.$$

5.3 CONVERGENCE IN PROBABILITY

Convergence of a Deterministic Sequence

Let $a_1, a_2, \ldots$ be a sequence of real numbers, and let a be another real number. We say that the sequence a_n converges to a, or $\lim_{n\to\infty} a_n = a$, if for every $\epsilon > 0$ there exists some n_0 such that

$$|a_n - a| \le \epsilon, \qquad \text{for all } n \ge n_0.$$

Convergence in Probability

Let $Y_1, Y_2, \ldots$ be a sequence of random variables (not necessarily independent), and let a be a real number. We say that the sequence Y_n **converges to** a **in probability**, if for every $\epsilon > 0$, we have

$$\lim_{n\to\infty} \mathbf{P}\big(|Y_n - a| \ge \epsilon\big) = 0.$$

5.4 THE CENTRAL LIMIT THEOREM

The Central Limit Theorem

Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with common mean μ and variance σ^2, and define

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}.$$

Then, the CDF of Z_n converges to the standard normal CDF

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\, dx,$$

in the sense that

$$\lim_{n\to\infty} \mathbf{P}(Z_n \leq z) = \Phi(z), \qquad \text{for every } z.$$

Normal Approximation Based on the Central Limit Theorem

Let $S_n = X_1 + \cdots + X_n$, where the X_i are independent identically distributed random variables with mean μ and variance σ^2. If n is large, the probability $\mathbf{P}(S_n \leq c)$ can be approximated by treating S_n as if it were normal, according to the following procedure.

1. Calculate the mean $n\mu$ and the variance $n\sigma^2$ of S_n.
2. Calculate the normalized value $z = (c - n\mu)/\sigma\sqrt{n}$.
3. Use the approximation

$$\mathbf{P}(S_n \leq c) \approx \Phi(z),$$

where $\Phi(z)$ is available from standard normal CDF tables.

De Moivre-Laplace Approximation to the Binomial

If S_n is a binomial random variable with parameters n and p, n is large, and k, l are nonnegative integers, then

$$\mathbf{P}(k \le S_n \le l) \approx \Phi\left(\frac{l + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right).$$

5.5 THE STRONG LAW OF LARGE NUMBERS

The Strong Law of Large Numbers

Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with mean μ. Then, the sequence of sample means $M_n = (X_1 + \cdots + X_n)/n$ converges to μ, **with probability 1**, in the sense that

$$\mathbf{P}\left(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu\right) = 1.$$

Convergence with Probability 1

Let $Y_1, Y_2, \ldots$ be a sequence of random variables (not necessarily independent). Let c be a real number. We say that Y_n converges to c **with probability 1** (or **almost surely**) if

$$\mathbf{P}\left(\lim_{n\to\infty} Y_n = c\right) = 1.$$

5.6 SUMMARY AND DISCUSSION

6

The Bernoulli and Poisson Processes

Excerpts from Introduction to Probability: Second Edition

by Dimitri P. Bertsekas and John N. Tsitsiklis ©

Massachusetts Institute of Technology

Contents

6.1 THE BERNOULLI PROCESS

Some Random Variables Associated with the Bernoulli Process and their Properties

- **The binomial with parameters p and n.** This is the number S of successes in n independent trials. Its PMF, mean, and variance are

$$p_S(k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n,$$

$$\mathbf{E}[S] = np, \qquad \mathrm{var}(S) = np(1-p).$$

- **The geometric with parameter p.** This is the number T of trials up to (and including) the first success. Its PMF, mean, and variance are

$$p_T(t) = (1-p)^{t-1} p, \qquad t = 1, 2, \ldots,$$

$$\mathbf{E}[T] = \frac{1}{p}, \qquad \mathrm{var}(T) = \frac{1-p}{p^2}.$$

Independence Properties of the Bernoulli Process

- For any given time n, the sequence of random variables $X_{n+1}, X_{n+2}, \ldots$ (the future of the process) is also a Bernoulli process, and is independent from $X_1, \ldots, X_n$ (the past of the process).
- Let n be a given time and let $\overline{T}$ be the time of the first success after time n. Then, $\overline{T} - n$ has a geometric distribution with parameter p, and is independent of the random variables $X_1, \ldots, X_n$.

Alternative Description of the Bernoulli Process

1. Start with a sequence of independent geometric random variables T_1, $T_2, \ldots$, with common parameter p, and let these stand for the interarrival times.
2. Record a success (or arrival) at times T_1, $T_1 + T_2$, $T_1 + T_2 + T_3$, etc.

Properties of the kth Arrival Time

- The kth arrival time is equal to the sum of the first k interarrival times

$$Y_k = T_1 + T_2 + \cdots + T_k,$$

and the latter are independent geometric random variables with common parameter p.

- The mean and variance of Y_k are given by

$$\mathbf{E}[Y_k] = \mathbf{E}[T_1] + \cdots + \mathbf{E}[T_k] = \frac{k}{p},$$

$$\mathrm{var}(Y_k) = \mathrm{var}(T_1) + \cdots + \mathrm{var}(T_k) = \frac{k(1-p)}{p^2}.$$

- The PMF of Y_k is given by

$$p_{Y_k}(t) = \binom{t-1}{k-1} p^k (1-p)^{t-k}, \qquad t = k, k+1, \ldots,$$

and is known as the **Pascal PMF of order** k.

Poisson Approximation to the Binomial

- A Poisson random variable Z with parameter λ takes nonnegative integer values and is described by the PMF

$$p_Z(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots.$$

  Its mean and variance are given by

$$\mathbf{E}[Z] = \lambda, \qquad \mathrm{var}(Z) = \lambda.$$

- For any fixed nonnegative integer k, the binomial probability

$$p_S(k) = \frac{n!}{(n-k)!\,k!} \cdot p^k (1-p)^{n-k}$$

  converges to $p_Z(k)$, when we take the limit as $n \to \infty$ and $p = \lambda/n$, while keeping λ constant.

- In general, the Poisson PMF is a good approximation to the binomial as long as $\lambda = np$, n is very large, and p is very small.

6.2 THE POISSON PROCESS

Definition of the Poisson Process

An arrival process is called a Poisson process with rate λ if it has the following properties:

(a) **(Time-homogeneity)** The probability $P(k, \tau)$ of k arrivals is the same for all intervals of the same length τ.

(b) **(Independence)** The number of arrivals during a particular interval is independent of the history of arrivals outside this interval.

(c) **(Small interval probabilities)** The probabilities $P(k, \tau)$ satisfy

$$\begin{aligned} P(0,\tau) &= 1 - \lambda\tau + o(\tau), \\ P(1,\tau) &= \lambda\tau + o_1(\tau), \\ P(k,\tau) &= o_k(\tau), \qquad \text{for } k = 2, 3, \ldots \end{aligned}$$

Here, $o(\tau)$ and $o_k(\tau)$ are functions of τ that satisfy

$$\lim_{\tau \to 0} \frac{o(\tau)}{\tau} = 0, \qquad \lim_{\tau \to 0} \frac{o_k(\tau)}{\tau} = 0.$$

Random Variables Associated with the Poisson Process and their Properties

- **The Poisson with parameter $\lambda\tau$.** This is the number N_τ of arrivals in a Poisson process with rate λ, over an interval of length τ. Its PMF, mean, and variance are

$$p_{N_\tau}(k) = P(k,\tau) = e^{-\lambda\tau}\frac{(\lambda\tau)^k}{k!}, \quad k = 0, 1, \ldots,$$

$$\mathbf{E}[N_\tau] = \lambda\tau, \qquad \mathrm{var}(N_\tau) = \lambda\tau.$$

- **The exponential with parameter λ.** This is the time T until the first arrival. Its PDF, mean, and variance are

$$f_T(t) = \lambda e^{-\lambda t}, \quad t \geq 0, \qquad \mathbf{E}[T] = \frac{1}{\lambda}, \qquad \mathrm{var}(T) = \frac{1}{\lambda^2}.$$

Independence Properties of the Poisson Process

- For any given time $t > 0$, the history of the process after time t is also a Poisson process, and is independent from the history of the process until time t.
- Let t be a given time and let $\overline{T}$ be the time of the first arrival after time t. Then, $\overline{T} - t$ has an exponential distribution with parameter λ, and is independent of the history of the process until time t.

Alternative Description of the Poisson Process

1. Start with a sequence of independent exponential random variables $T_1, T_2, \ldots$, with common parameter λ, and let these represent the interarrival times.
2. Record an arrival at times T_1, $T_1 + T_2$, $T_1 + T_2 + T_3$, etc.

Properties of the kth Arrival Time

- The kth arrival time is equal to the sum of the first k interarrival times

$$Y_k = T_1 + T_2 + \cdots + T_k,$$

  and the latter are independent exponential random variables with common parameter λ.
- The mean and variance of Y_k are given by

$$\mathbf{E}[Y_k] = \mathbf{E}[T_1] + \cdots + \mathbf{E}[T_k] = \frac{k}{\lambda},$$

$$\mathrm{var}(Y_k) = \mathrm{var}(T_1) + \cdots + \mathrm{var}(T_k) = \frac{k}{\lambda^2}.$$

- The PDF of Y_k is given by

$$f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}, \qquad y \geq 0,$$

  and is known as the **Erlang PDF of order** k.

Properties of Sums of a Random Number of Random Variables

Let $N, X_1, X_2, \ldots$ be independent random variables, where N takes nonnegative integer values. Let $Y = X_1 + \cdots + X_N$ for positive values of N, and let $Y = 0$ when $N = 0$.

- If X_i is Bernoulli with parameter p, and N is binomial with parameters m and q, then Y is binomial with parameters m and pq.
- If X_i is Bernoulli with parameter p, and N is Poisson with parameter λ, then Y is Poisson with parameter λp.
- If X_i is geometric with parameter p, and N is geometric with parameter q, then Y is geometric with parameter pq.
- If X_i is exponential with parameter λ, and N is geometric with parameter q, then Y is exponential with parameter λq.

6.3 SUMMARY AND DISCUSSION

7

Markov Chains

Excerpts from Introduction to Probability: Second Edition

by Dimitri P. Bertsekas and John N. Tsitsiklis ©

Massachusetts Institute of Technology

Contents

7.1 DISCRETE-TIME MARKOV CHAINS

Specification of Markov Models

- A Markov chain model is specified by identifying:
  (a) the set of states $\S = \{1, \ldots, m\}$,
  (b) the set of possible transitions, namely, those pairs (i, j) for which $p_{ij} > 0$, and,
  (c) the numerical values of those p_{ij} that are positive.
- The Markov chain specified by this model is a sequence of random variables $X_0, X_1, X_2, \ldots$, that take values in $\S$, and which satisfy

$$\mathbf{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = p_{ij},$$

  for all times n, all states $i, j \in \S$, and all possible sequences $i_0, \ldots, i_{n-1}$ of earlier states.

Chapman-Kolmogorov Equation for the n-Step Transition Probabilities

The n-step transition probabilities can be generated by the recursive formula

$$r_{ij}(n) = \sum_{k=1}^{m} r_{ik}(n-1)p_{kj}, \qquad \text{for } n > 1, \quad \text{and all } i,\ j,$$

starting with

$$r_{ij}(1) = p_{ij}.$$

7.2 CLASSIFICATION OF STATES

Markov Chain Decomposition

- A Markov chain can be decomposed into one or more recurrent classes, plus possibly some transient states.
- A recurrent state is accessible from all states in its class, but is not accessible from recurrent states in other classes.
- A transient state is not accessible from any recurrent state.
- At least one, possibly more, recurrent states are accessible from a given transient state.

Periodicity

Consider a recurrent class R.

- The class is called **periodic** if its states can be grouped in $d > 1$ disjoint subsets $S_1, \ldots, S_d$, so that all transitions from S_k lead to S_{k+1} (or to S_1 if $k = d$).
- The class is **aperiodic** (not periodic) if and only if there exists a time n such that $r_{ij}(n) > 0$, for all $i, j \in R$.

7.3 STEADY-STATE BEHAVIOR

Steady-State Convergence Theorem

Consider a Markov chain with a single recurrent class, which is aperiodic. Then, the states j are associated with steady-state probabilities π_j that have the following properties.

(a) For each j, we have

$$\lim_{n\to\infty} r_{ij}(n) = \pi_j, \qquad \text{for all } i.$$

(b) The π_j are the unique solution to the system of equations below:

$$\pi_j = \sum_{k=1}^{m} \pi_k p_{kj}, \qquad j = 1, \ldots, m,$$
$$1 = \sum_{k=1}^{m} \pi_k.$$

(c) We have

$$\pi_j = 0, \qquad \text{for all transient states } j,$$
$$\pi_j > 0, \qquad \text{for all recurrent states } j.$$

Steady-State Probabilities as Expected State Frequencies

For a Markov chain with a single class which is aperiodic, the steady-state probabilities π_j satisfy

$$\pi_j = \lim_{n\to\infty} \frac{v_{ij}(n)}{n},$$

where $v_{ij}(n)$ is the expected value of the number of visits to state j within the first n transitions, starting from state i.

Expected Frequency of a Particular Transition

Consider n transitions of a Markov chain with a single class which is aperiodic, starting from a given initial state. Let $q_{jk}(n)$ be the expected number of such transitions that take the state from j to k. Then, regardless of the initial state, we have

$$\lim_{n\to\infty} \frac{q_{jk}(n)}{n} = \pi_j p_{jk}.$$

7.4 ABSORPTION PROBABILITIES AND EXPECTED TIME TO ABSORPTION

Absorption Probability Equations

Consider a Markov chain where each state is either transient or absorbing, and fix a particular absorbing state s. Then, the probabilities a_i of eventually reaching state s, starting from i, are the unique solution to the equations

$$\begin{aligned} a_s &= 1, & & \\ a_i &= 0, & & \text{for all absorbing } i \neq s, \\ a_i &= \sum_{j=1}^{m} p_{ij} a_j, & & \text{for all transient } i. \end{aligned}$$

Equations for the Expected Times to Absorption

Consider a Markov chain where all states are transient, except for a single absorbing state. The expected times to absorption, $\mu_1, \ldots, \mu_m$, are the unique solution to the equations

$$\mu_i = 0, \qquad \text{if } i \text{ is the absorbing state,}$$

$$\mu_i = 1 + \sum_{j=1}^{m} p_{ij}\mu_j, \qquad \text{if } i \text{ is transient.}$$

Equations for Mean First Passage and Recurrence Times

Consider a Markov chain with a single recurrent class, and let s be a particular recurrent state.

- The mean first passage times μ_i to reach state s starting from i, are the unique solution to the system of equations

$$\mu_s = 0, \qquad \mu_i = 1 + \sum_{j=1}^{m} p_{ij}\mu_j, \qquad \text{for all } i \neq s.$$

- The mean recurrence time μ_s^* of state s is given by

$$\mu_s^* = 1 + \sum_{j=1}^{m} p_{sj}\mu_j.$$

7.5 CONTINUOUS-TIME MARKOV CHAINS

Continuous-Time Markov Chain Assumptions

- If the current state is i, the time until the next transition is exponentially distributed with a given parameter ν_i, independent of the past history of the process and of the next state.
- If the current state is i, the next state will be j with a given probability p_{ij}, independent of the past history of the process and of the time until the next transition.

Alternative Description of a Continuous-Time Markov Chain

Given the current state i of a continuous-time Markov chain, and for any $j \neq i$, the state δ time units later is equal to j with probability

$$q_{ij}\delta + o(\delta),$$

independent of the past history of the process.

Steady-State Convergence Theorem

Consider a continuous-time Markov chain with a single recurrent class. Then, the states j are associated with steady-state probabilities π_j that have the following properties.

(a) For each j, we have

$$\lim_{t\to\infty} \mathbf{P}\big(X(t) = j \,|\, X(0) = i\big) = \pi_j, \qquad \text{for all } i.$$

(b) The π_j are the unique solution to the system of equations below:

$$\pi_j \sum_{k\neq j} q_{jk} = \sum_{k\neq j} \pi_k q_{kj}, \qquad j = 1, \ldots, m,$$

$$1 = \sum_{k=1}^{m} \pi_k.$$

(c) We have

$$\pi_j = 0, \qquad \text{for all transient states } j,$$
$$\pi_j > 0, \qquad \text{for all recurrent states } j.$$

7.6 SUMMARY AND DISCUSSION

8

Bayesian Statistical Inference

Excerpts from Introduction to Probability: Second Edition

by Dimitri P. Bertsekas and John N. Tsitsiklis ©

Massachusetts Institute of Technology

Contents

Major Terms, Problems, and Methods in this Chapter

- **Bayesian statistics** treats unknown parameters as random variables with known prior distributions.
- In **parameter estimation**, we want to generate estimates that are close to the true values of the parameters in some probabilistic sense.
- In **hypothesis testing**, the unknown parameter takes one of a finite number of values, corresponding to competing hypotheses; we want to choose one of the hypotheses, aiming to achieve a small probability of error.
- Principal Bayesian inference methods:
  (a) **Maximum a posteriori probability** (MAP) rule: Out of the possible parameter values/hypotheses, select one with maximum conditional/posterior probability given the data (Section 8.2).
  (b) **Least mean squares** (LMS) estimation: Select an estimator/function of the data that minimizes the mean squared error between the parameter and its estimate (Section 8.3).
  (c) **Linear least mean squares** estimation: Select an estimator which is a linear function of the data and minimizes the mean squared error between the parameter and its estimate (Section 8.4). This may result in higher mean squared error, but requires simple calculations, based only on the means, variances, and covariances of the random variables involved.

8.1 BAYESIAN INFERENCE AND THE POSTERIOR DISTRIBUTION

Summary of Bayesian Inference

- We start with a prior distribution p_Θ or f_Θ for the unknown random variable Θ.
- We have a model $p_{X|\Theta}$ or $f_{X|\Theta}$ of the observation vector X.
- After observing the value x of X, we form the posterior distribution of Θ, using the appropriate version of Bayes' rule.

The Four Versions of Bayes' Rule

- Θ discrete, X discrete:

$$p_{\Theta|X}(\theta\,|\,x) = \frac{p_\Theta(\theta)p_{X|\Theta}(x\,|\,\theta)}{\displaystyle\sum_{\theta'} p_\Theta(\theta')p_{X|\Theta}(x\,|\,\theta')}.$$

- Θ discrete, X continuous:

$$p_{\Theta|X}(\theta\,|\,x) = \frac{p_\Theta(\theta)f_{X|\Theta}(x\,|\,\theta)}{\displaystyle\sum_{\theta'} p_\Theta(\theta')f_{X|\Theta}(x\,|\,\theta')}.$$

- Θ continuous, X discrete:

$$f_{\Theta|X}(\theta\,|\,x) = \frac{f_\Theta(\theta)p_{X|\Theta}(x\,|\,\theta)}{\displaystyle\int f_\Theta(\theta')p_{X|\Theta}(x\,|\,\theta')\,d\theta'}.$$

- Θ continuous, X continuous:

$$f_{\Theta|X}(\theta\,|\,x) = \frac{f_\Theta(\theta)f_{X|\Theta}(x\,|\,\theta)}{\displaystyle\int f_\Theta(\theta')f_{X|\Theta}(x\,|\,\theta')\,d\theta'}.$$

8.2 POINT ESTIMATION, HYPOTHESIS TESTING, AND THE MAP RULE

The Maximum a Posteriori Probability (MAP) Rule

- Given the observation value x, the MAP rule selects a value $\hat{\theta}$ that maximizes over θ the posterior distribution $p_{\Theta|X}(\theta\,|\,x)$ (if Θ is discrete) or $f_{\Theta|X}(\theta\,|\,x)$ (if Θ is continuous).
- Equivalently, it selects $\hat{\theta}$ that maximizes over θ:

$$
\begin{aligned}
&p_\Theta(\theta)p_{X|\Theta}(x\,|\,\theta) && \text{(if } \Theta \text{ and } X \text{ are discrete)},\\
&p_\Theta(\theta)f_{X|\Theta}(x\,|\,\theta) && \text{(if } \Theta \text{ is discrete and } X \text{ is continuous)},\\
&f_\Theta(\theta)p_{X|\Theta}(x\,|\,\theta) && \text{(if } \Theta \text{ is continuous and } X \text{ is discrete)},\\
&f_\Theta(\theta)f_{X|\Theta}(x\,|\,\theta) && \text{(if } \Theta \text{ and } X \text{ are continuous)}.
\end{aligned}
$$

- If Θ takes only a finite number of values, the MAP rule minimizes (over all decision rules) the probability of selecting an incorrect hypothesis. This is true for both the unconditional probability of error and the conditional one, given any observation value x.

Point Estimates

- An **estimator** is a random variable of the form $\hat{\Theta} = g(X)$, for some function g. Different choices of g correspond to different estimators.
- An **estimate** is the value $\hat{\theta}$ of an estimator, as determined by the realized value x of the observation X.
- Once the value x of X is observed, the **Maximum a Posteriori Probability (MAP)** estimator, sets the estimate $\hat{\theta}$ to a value that maximizes the posterior distribution over all possible values of θ.
- Once the value x of X is observed, the **Conditional Expectation (LMS)** estimator sets the estimate $\hat{\theta}$ to $\mathbf{E}[\Theta \mid X = x]$.

The MAP Rule for Hypothesis Testing

- Given the observation value x, the MAP rule selects a hypothesis H_i for which the value of the posterior probability $\mathbf{P}(\Theta = \theta_i \mid X = x)$ is largest.
- Equivalently, it selects a hypothesis H_i for which $p_\Theta(\theta_i)p_{X|\Theta}(x \mid \theta_i)$ (if X is discrete) or $p_\Theta(\theta_i)f_{X|\Theta}(x \mid \theta_i)$ (if X is continuous) is largest.
- The MAP rule minimizes the probability of selecting an incorrect hypothesis for any observation value x, as well as the probability of error over all decision rules.

8.3 BAYESIAN LEAST MEAN SQUARES ESTIMATION

Key Facts About Least Mean Squares Estimation

- In the absence of any observations, $\mathbf{E}\big[(\Theta - \hat{\theta})^2\big]$ is minimized when $\hat{\theta} = \mathbf{E}[\Theta]$:

$$\mathbf{E}\Big[\big(\Theta - \mathbf{E}[\Theta]\big)^2\Big] \leq \mathbf{E}\big[(\Theta - \hat{\theta})^2\big], \qquad \text{for all } \hat{\theta}.$$

- For any given value x of X, $\mathbf{E}\big[(\Theta - \hat{\theta})^2 \,|\, X = x\big]$ is minimized when $\hat{\theta} = \mathbf{E}[\Theta \,|\, X = x]$:

$$\mathbf{E}\Big[\big(\Theta - \mathbf{E}[\Theta \,|\, X = x]\big)^2 \;\Big|\; X = x\Big] \leq \mathbf{E}\big[(\Theta - \hat{\theta})^2 \,|\, X = x\big], \quad \text{for all } \hat{\theta}.$$

- Out of all estimators $g(X)$ of Θ based on X, the mean squared estimation error $\mathbf{E}\Big[\big(\Theta - g(X)\big)^2\Big]$ is minimized when $g(X) = \mathbf{E}[\Theta \,|\, X]$:

$$\mathbf{E}\Big[\big(\Theta - \mathbf{E}[\Theta \,|\, X]\big)^2\Big] \leq \mathbf{E}\Big[\big(\Theta - g(X)\big)^2\Big], \qquad \text{for all estimators } g(X).$$

Properties of the Estimation Error

- The estimation error $\tilde{\Theta}$ is **unbiased**, i.e., it has zero unconditional and conditional mean:

$$\mathbf{E}[\tilde{\Theta}] = 0, \qquad \mathbf{E}[\tilde{\Theta} \mid X = x] = 0, \quad \text{for all } x.$$

- The estimation error $\tilde{\Theta}$ is uncorrelated with the estimate $\hat{\Theta}$:

$$\text{cov}(\hat{\Theta}, \tilde{\Theta}) = 0.$$

- The variance of Θ can be decomposed as

$$\text{var}(\Theta) = \text{var}(\hat{\Theta}) + \text{var}(\tilde{\Theta}).$$

8.4 BAYESIAN LINEAR LEAST MEAN SQUARES ESTIMATION

Linear LMS Estimation Formulas

- The linear LMS estimator $\hat{\Theta}$ of Θ based on X is

$$\hat{\Theta} = \mathbf{E}[\Theta] + \frac{\text{cov}(\Theta, X)}{\text{var}(X)}\big(X - \mathbf{E}[X]\big) = \mathbf{E}[\Theta] + \rho\frac{\sigma_\Theta}{\sigma_X}\big(X - \mathbf{E}[X]\big),$$

where

$$\rho = \frac{\text{cov}(\Theta, X)}{\sigma_\Theta \sigma_X}$$

is the correlation coefficient.

- The resulting mean squared estimation error is equal to

$$(1 - \rho^2)\sigma_\Theta^2.$$

8.5 SUMMARY AND DISCUSSION

9

Classical Statistical Inference

Excerpts from Introduction to Probability: Second Edition

by Dimitri P. Bertsekas and John N. Tsitsiklis ©

Massachusetts Institute of Technology

Contents

Major Terms, Problems, and Methods in this Chapter

- **Classical statistics** treats unknown parameters as constants to be determined. A separate probabilistic model is assumed for each possible value of the unknown parameter.
- In **parameter estimation**, we want to generate estimates that are nearly correct under any possible value of the unknown parameter.
- In **hypothesis testing**, the unknown parameter takes a finite number m of values ($m \geq 2$), corresponding to competing hypotheses; we want to choose one of the hypotheses, aiming to achieve a small probability of error under any of the possible hypotheses.
- In **significance testing**, we want to accept or reject a single hypothesis, while keeping the probability of false rejection suitably small.
- Principal classical inference methods in this chapter:
  (a) **Maximum likelihood (ML) estimation**: Select the parameter that makes the observed data "most likely," i.e., maximizes the probability of obtaining the data at hand (Section 9.1).
  (b) **Linear regression**: Find the linear relation that matches best a set of data pairs, in the sense that it minimizes the sum of the squares of the discrepancies between the model and the data (Section 9.2).
  (c) **Likelihood ratio test**: Given two hypotheses, select one based on the ratio of their "likelihoods," so that certain error probabilities are suitably small (Section 9.3).
  (d) **Significance testing**: Given a hypothesis, reject it if and only if the observed data falls within a certain rejection region. This region is specially designed to keep the probability of false rejection below some threshold (Section 9.4).

9.1 CLASSICAL PARAMETER ESTIMATION

Terminology Regarding Estimators

Let $\hat{\Theta}_n$ be an **estimator** of an unknown parameter θ, that is, a function of n observations $X_1, \ldots, X_n$ whose distribution depends on θ.

- The **estimation error**, denoted by $\tilde{\Theta}_n$, is defined by $\tilde{\Theta}_n = \hat{\Theta}_n - \theta$.
- The **bias** of the estimator, denoted by $\mathrm{b}_\theta(\hat{\Theta}_n)$, is the expected value of the estimation error:

$$\mathrm{b}_\theta(\hat{\Theta}_n) = \mathbf{E}_\theta[\hat{\Theta}_n] - \theta.$$

- The expected value, the variance, and the bias of $\hat{\Theta}_n$ depend on θ, while the estimation error depends in addition on the observations $X_1, \ldots, X_n$.
- We call $\hat{\Theta}_n$ **unbiased** if $\mathbf{E}_\theta[\hat{\Theta}_n] = \theta$, for every possible value of θ.
- We call $\hat{\Theta}_n$ **asymptotically unbiased** if $\lim_{n\to\infty} \mathbf{E}_\theta[\hat{\Theta}_n] = \theta$, for every possible value of θ.
- We call $\hat{\Theta}_n$ **consistent** if the sequence $\hat{\Theta}_n$ converges to the true value of the parameter θ, in probability, for every possible value of θ.

Maximum Likelihood Estimation

- We are given the realization $x = (x_1, \ldots, x_n)$ of a random vector $X = (X_1, \ldots, X_n)$, distributed according to a PMF $p_X(x; \theta)$ or PDF $f_X(x; \theta)$.
- The maximum likelihood (ML) estimate is a value of θ that maximizes the likelihood function, $p_X(x; \theta)$ or $f_X(x; \theta)$, over all θ.
- The ML estimate of a one-to-one function $h(\theta)$ of θ is $h(\hat{\theta}_n)$, where $\hat{\theta}_n$ is the ML estimate of θ (the invariance principle).
- When the random variables X_i are i.i.d., and under some mild additional assumptions, each component of the ML estimator is consistent and asymptotically normal.

Estimates of the Mean and Variance of a Random Variable

Let the observations $X_1, \ldots, X_n$ be i.i.d., with mean θ and variance v that are unknown.

- The sample mean

$$M_n = \frac{X_1 + \cdots + X_n}{n}$$

is an unbiased estimator of θ, and its mean squared error is v/n.

- Two variance estimators are

$$\overline{S}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - M_n)^2, \qquad \hat{S}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - M_n)^2.$$

- The estimator $\overline{S}_n^2$ coincides with the ML estimator if the X_i are normal. It is biased but asymptotically unbiased. The estimator $\hat{S}_n^2$ is unbiased. For large n, the two variance estimators essentially coincide.

Confidence Intervals

- A **confidence interval** for a scalar unknown parameter θ is an interval whose endpoints $\hat{\Theta}_n^-$ and $\hat{\Theta}_n^+$ bracket θ with a given high probability.
- $\hat{\Theta}_n^-$ and $\hat{\Theta}_n^+$ are random variables that depend on the observations $X_1, \ldots, X_n$.
- A $1 - \alpha$ confidence interval is one that satisfies

$$\mathbf{P}_\theta\big(\hat{\Theta}_n^- \leq \theta \leq \hat{\Theta}_n^+\big) \geq 1 - \alpha,$$

for all possible values of θ.

9.2 LINEAR REGRESSION

Linear Regression

Given n data pairs (x_i, y_i), the estimates that minimize the sum of the squared residuals are given by

$$\hat{\theta}_1 = \frac{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2}, \qquad \hat{\theta}_0 = \overline{y} - \hat{\theta}_1 \overline{x},$$

where

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Bayesian Linear Regression

- **Model:**
  (a) We assume a linear relation $Y_i = \Theta_0 + \Theta_1 x_i + W_i$.
  (b) The x_i are modeled as known constants.
  (c) The random variables $\Theta_0, \Theta_1, W_1, \ldots, W_n$ are normal and independent.
  (d) The random variables Θ_0 and Θ_1 have mean zero and variances σ_0^2, σ_1^2, respectively.
  (e) The random variables W_i have mean zero and variance σ^2.
- **Estimation Formulas:**
  Given the data pairs (x_i, y_i), the MAP estimates of Θ_0 and Θ_1 are

$$\hat{\theta}_1 = \frac{\sigma_1^2}{\sigma^2 + \sigma_1^2 \sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sum_{i=1}^{n}(x_i - \overline{x})\big(y_i - \overline{y}\big),$$

$$\hat{\theta}_0 = \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}\,(\overline{y} - \hat{\theta}_1\overline{x}),$$

where

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

9.3 BINARY HYPOTHESIS TESTING

Likelihood Ratio Test (LRT)

- Start with a target value α for the false rejection probability.
- Choose a value for ξ such that the false rejection probability is equal to α:

$$\mathbf{P}\big(L(X) > \xi; H_0\big) = \alpha.$$

- Once the value x of X is observed, reject H_0 if $L(x) > \xi$.

Neyman-Pearson Lemma

Consider a particular choice of ξ in the LRT, which results in error probabilities

$$\mathbf{P}\big(L(X) > \xi; H_0\big) = \alpha, \qquad \mathbf{P}\big(L(X) \leq \xi; H_1\big) = \beta.$$

Suppose that some other test, with rejection region R, achieves a smaller or equal false rejection probability:

$$\mathbf{P}(X \in R; H_0) \leq \alpha.$$

Then,

$$\mathbf{P}(X \notin R; H_1) \geq \beta,$$

with strict inequality $\mathbf{P}(X \notin R; H_1) > \beta$ when $\mathbf{P}(X \in R; H_0) < \alpha$.

9.4 SIGNIFICANCE TESTING

Significance Testing Methodology

A statistical test of a hypothesis H_0 is to be performed, based on the observations $X_1, \ldots, X_n$.

- The following steps are carried out before the data are observed.
  - (a) Choose a **statistic** S, that is, a scalar random variable that will summarize the data to be obtained. Mathematically, this involves the choice of a function $h : \Re^n \to \Re$, resulting in the statistic $S = h(X_1 \ldots, X_n)$.
  - (b) Determine the **shape of the rejection region** by specifying the set of values of S for which H_0 will be rejected as a function of a yet undetermined critical value ξ.
  - (c) Choose the **significance level,** i.e., the desired probability α of a false rejection of H_0.
  - (d) Choose the **critical value** ξ so that the probability of false rejection is equal (or approximately equal) to α. At this point, the rejection region is completely determined.
- Once the values $x_1, \ldots, x_n$ of $X_1, \ldots, X_n$ are observed:
  - (i) Calculate the value $s = h(x_1, \ldots, x_n)$ of the statistic S.
  - (ii) Reject the hypothesis H_0 if s belongs to the rejection region.

The Chi-Square Test:

- Use the statistic

$$S = \sum_{k=1}^{m} N_k \log\left(\frac{N_k}{n\theta_k^*}\right)$$

  (or possibly the related statistic T) and a rejection region of the form

$$\text{reject } H_0 \text{ if } 2S > \gamma$$

  (or $T > \gamma$, respectively).

- The critical value γ is determined from the CDF tables for the χ^2 distribution with $m-1$ degrees of freedom so that

$$\mathbf{P}(2S > \gamma; H_0) = \alpha,$$

  where α is a given significance level.

9.5 SUMMARY AND DISCUSSION