Probability and Statistics Refresher

1. Introduction to random variables

A. What is a random variable?

> *A random variable can be viewed as the name of an experiment with a probabilistic outcome. It's value is the outcome of that experiment.*
>
> *Tom Mitchell, 1997*

- A *discrete random variable* can assume only a countable number of values.

  Example: Coin flip (heads/tails), Let's make a deal (Door 1, 2 or 3), Die (1, 2, 3, 4, 5 or 6).

- A *real (or continuous) random variable* can assume a range of values, $x \in [a, b]$.

Example: most sensor readings.

[Note: We'll be dealing a lot with both types of random variables.]

B. Some synonyms for "random"

- *stochastic*
- *nondeterministic*

2. Discrete random variables

A. Notation

- Let $A, B, C, \ldots$ and $A_1, A_2, A_3, \ldots$ denote discrete random variables.
- Let $a, b, c, \ldots$ and $a_1, a_2, a_3, \ldots$ denote possible values for corresponding discrete random variables $A, B, C, \ldots$ and $A_1, A_2, A_3, \ldots$, respectively. Hence, $P(A = a)$ denotes the *probability* that the random variable A assumes the values a.
- In the special case when the random variable A can assume only two values, *(T)rue* or *(F)alse*, denote $P(A)$ as the probability that the random variable A is true, and denote $\neg A$ or $\bar{A}$ as the probability that the random variable A is false.
- Let $P(A \wedge B) = P(A \cap B) = P(A, B) = P(A \text{ and } B)$.
- Let $P(A \vee B) = P(A \cup B) = P(A \text{ or } B)$.
- Let $P(A = a|B = b)$ denote the conditional probability that $A = a$ given that $B = b$. Often we use the shorthand $P(A|B)$ when A and B are true/false random variables.

B. Axioms of probability

$$0 \le P(A) \le 1 \qquad (1)$$

$$P(True) = 1 \qquad (2)$$

$$P(False) = 0 \qquad (3)$$

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B) \qquad (4)$$

C. Conditional probability

$$P(A|B) \equiv \frac{P(A \wedge B)}{P(B)},\ P(B) > 0 \tag{5}$$

From (5),

$$P(A \wedge B) = P(A|B)P(B) = P(B|A)P(A) \text{ (Chain rule)} \tag{6}$$

Equation (6) leads to *Bayes Theorem*:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{7}$$

Assume that precisely one of $A_1, A_2, \ldots, A_n$ mutually exclusive events can occur; that is,

$$P(A_i \wedge A_j) = 0,\ \forall i \neq j \text{, and,} \tag{8}$$

$$P(A_1 \vee A_2 \vee \ldots \vee A_n) = \sum_{i=1}^{n} P(A_i) = 1 \tag{9}$$

For such events A_i, the general form of Bayes Theorem is given by,

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} \tag{10}$$

where,

$$P(B) = \sum_{j=1}^{n} P(B|A_j)P(A_j) \tag{11}$$

D. Other laws of probability

- $P(A) = 1 - P(\neg A)$

  Easy to prove. From (4),

$$P(A \vee \neg A) = P(A) + P(\neg A) - P(A \wedge \neg A) \tag{12}$$

$$P(A \wedge \neg A) = 0 \text{ (something can't be true and false at the same time).} \tag{13}$$

$$P(A \vee \neg A) = 1 \text{ (binary random variable has to be either true or false).} \tag{14}$$

$$1 = P(A) + P(\neg A) \tag{15}$$

$$P(A) = 1 - P(\neg A) \tag{16}$$

- $P(A) = P(A \wedge \neg B) + P(A \wedge B)$
- $P(A \wedge B \wedge C) = P(A|B \wedge C)P(B|C)P(C)$

E. Independence

A and B are independent if and only if $P(A = a) = P(A = a|B = b)$, $\forall a, b$. From this definition, we can easily derive the following properties of *independent events*:

$$P(B) = P(B|A) \tag{17}$$

$$P(A \wedge B) = P(A)P(B) \tag{18}$$

Example: coin toss

F. Conditional independence

A and B are <u>*conditionally independent*</u> given C if and only if

$$P(A = a|B = b \wedge C = c) = P(A = a|C = c), \forall a, b, c \tag{19}$$

Example: Traveling salesman

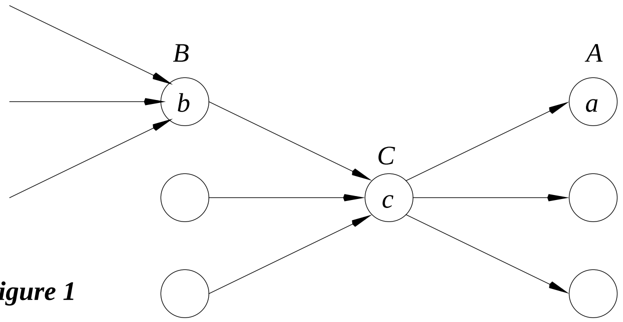


Figure 1

3. Probability distributions

A. Discrete random variables

For a given discrete random variable X, the probability distribution for X assigns a probability for each possible value or outcome of X,

$$P(X = x) = p(x), \forall x \tag{20}$$

For n random variables, $X_1, X_2, \ldots, X_n$, the joint probability distribution assigns a probability for all possible combinations of values,

$$P(X_1 = x_1 \wedge X_2 = x_2 \wedge \ldots \wedge X_n = x_n) = p(x_1, x_2, \ldots, x_n), \forall x_1, x_2, \ldots, x_n \tag{21}$$

Example: If each random variable can assume one of k different values, then the joint probability distribution for n different random variables is fully specified by k^n values.

Given two discrete random variables X and Y, the univariate probability distribution $p(x)$ is related to the joint probability distribution $p(x, y)$ by,

$$p(x) = \sum_y p(x, y) \tag{22}$$

B. Real random variables

A real-valued random variable X is characterized by a continuous probability distribution function,

$$F(x) = P(X \le x), -\infty < x < \infty \tag{23}$$

$$F(x_1) \ge F(x_2) \text{ if } x_1 > x_2 \tag{24}$$

$$F(-\infty) = 0, F(\infty) = 1 \tag{25}$$

$$P(X = x) = 0 \tag{26}$$

Often, however, the distribution is expressed as a *probability density function* (*pdf*) $p(x)$,

$$p(x) = \frac{dF(x)}{dx}, \tag{27}$$

$$p(x) \ge 0, \forall x \tag{28}$$

$$\int_{-\infty}^{\infty} p(x)dx = 1 \tag{29}$$

From (27),

$$F(x) = \int_{-\infty}^{x} p(t)dt \tag{30}$$

For n real-valued random variables, $X_1, X_2, \ldots, X_n$, the joint probability density function is given by,

$$p(x_1, x_2, \ldots, x_n) \geq 0 \text{ , } \forall x_1, x_2, \ldots, x_n \tag{31}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} p(x_1, x_2, \ldots, x_n)dx_1, dx_2, \ldots, dx_n = 1 \tag{32}$$

Note: $p(x)$ and $p(x_1, x_2, \ldots, x_n)$ are often referred to as the *likelihood* of x and $x_1, x_2, \ldots, x_n$, respectively.

Given two continuous random variables X and Y, the univariate probability distribution $p(x)$ is related to the joint probability distribution $p(x, y)$ by,

$$p(x) = \int_{-\infty}^{\infty} p(x, y)dy \tag{33}$$

C. Normal (Gaussian) distribution

A very important continuous probability distribution is the *Normal*, or *Gaussian*, distribution. It is defined by the following probability density function (pdf),

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] = N[x, \mu, \sigma^2] \tag{34}$$

for specific values of the scalar parameters μ and σ. In multiple dimensions,

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right] = N[\mathbf{x}, \mu, \Sigma] \tag{35}$$

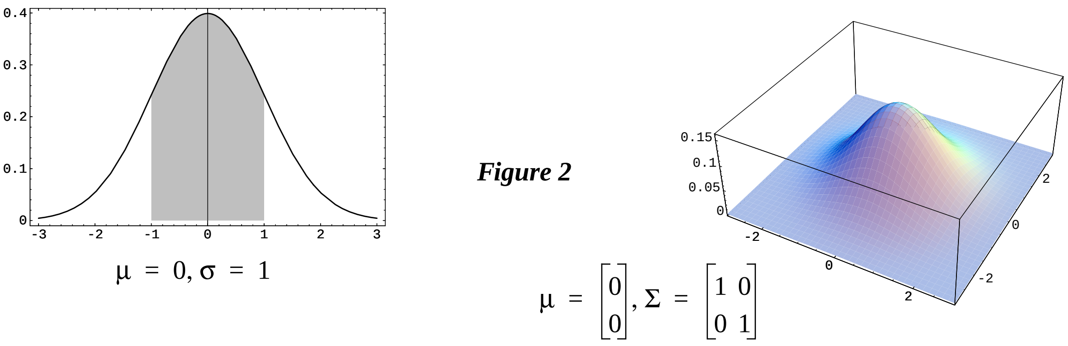
where μ and $\mathbf{x}$ are d-dimensional vectors, and Σ is a $d \times d$ positive-definite, symmetric matrix.

[Note: A square matrix $\mathbf{M}$ is positive definite if and only if,

$$\mathbf{v}^T\mathbf{M}\mathbf{v} > 0 \quad \forall\mathbf{v} \neq 0 \text{ .} \tag{36}$$

Equation (36) also implies that all the eigenvalues $\lambda_i > 0$ and that $\mathbf{M}$ is invertible.]

Example: In Figure 2, the normal distribution is plotted in one and two dimensions for unity variances.

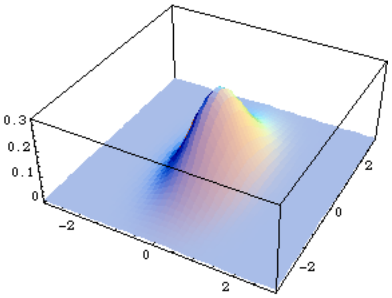


$\mu = 0, \sigma = 1$

Figure 2

$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

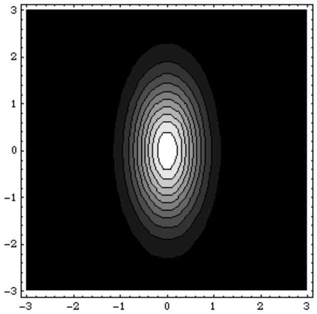
Example: In Figure 3, the normal distribution is plotted in two dimensions for a nonuniform diagonal covariance matrix.



Gaussian pdf in two dimensions

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1/4 & 0 \\ 0 & 1 \end{bmatrix}$$

Figure 3



Contour plot of Gaussian pdf

4. Expected value, variance, standard deviation

A. Expected value for a discrete random variable

The *expected value* for a discrete random variable is given by,

$$E[X] = \sum_{x} xp(x) \tag{37}$$

The *mean* $\bar{x}_n = \mu$ for a dataset $\mathbf{X} = \{x_j\}$, $j \in \{1, 2, \ldots, n\}$, is given by,

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{38}$$

where x_i denotes the *i*th measurement. The mean and expected value for a random variable are related by,

$$E[X] = \lim_{n \to \infty} \bar{x}_n \tag{39}$$

Example: Let A denote a random variable which denotes the outcome of a die roll. Find $E[A]$.

Assume a fair die. then $P(A = i) = p(i) = 1/6$, $\forall i$.

$$E[A] = 1 \cdot 1/6 + 2 \cdot 1/6 + 3 \cdot 1/6 + 4 \cdot 1/6 + 5 \cdot 1/6 + 6 \cdot 1/6 \tag{40}$$

$$E[A] = \frac{1}{6}\sum_{i=1}^{6} i = 3.5 \tag{41}$$

[Note: The term “expected value” can be misleading, since the expected value of a random variable may not be a possible value for the random variable itself.]

Example: You shake two 6-sided dice. What is the expected value of the lower of the two numbers (A)?

Define two random variables D_1 and D_2, which define the outcome of the die roll of dice 1 and 2.

$$P(A = 6) = P(D_1 = 6)P(D_2 = 6) = 1/36 \tag{42}$$

$$P(A = 5) = P(D_1 = 5)P(D_2 = 5) + P(D_1 = 5)P(D_2 = 6) + P(D_1 = 6)P(D_2 = 5) \tag{43}$$

$$P(A = 5) = 3/36 \tag{44}$$

Similarly,

$$P(A = i) = [(7 - i) + (6 - i)]/36 = (13 - 2i)/36 . \tag{45}$$

Therefore,

$$E[A] = \sum_{i=1}^{6} \frac{i(13-2i)}{36} = \frac{91}{36} = 2.52778 \tag{46}$$

B. Expected value for a continuous random variable

The *expected value* for a continuous random variable is given by,

$$E[X] = \int_{-\infty}^{\infty} xp(x)dx \tag{47}$$

Example: $E[X]$ for the normal distribution is μ.

C. Variance and standard deviation

The *variance* $V[X]$ of a random variable X is defined by,

$$V[X] = E[(X-\mu)^2] \tag{48}$$

where $\mu = E[X]$. The *standard deviation* σ of a random variable X is defined by,

$$\sigma = \sqrt{V[X]} \tag{49}$$

Because of (49), we often denote $V[X] = \sigma^2$.

Example: The variance of the normal distribution is σ^2.

D. Covariance of two random variables

The *covariance* of two random variables X and Y is defined by,

$$Cov[X, Y] = E[(X-\mu_x)(Y-\mu_y)] \tag{50}$$

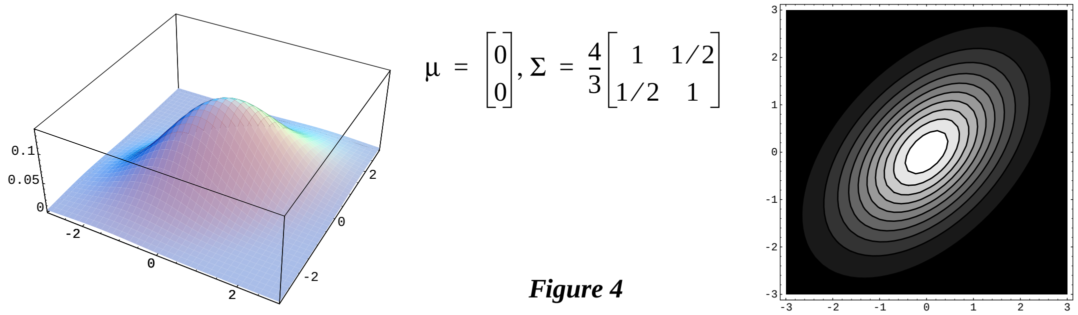
where $\mu_x = E[X]$ and $\mu_y = E[Y]$. If X and Y are independent, then,

$$Cov[X, Y] = 0 \tag{51}$$

Example: The matrix Σ in the multivariate normal distribution is the *covariance matrix*. The (i, j) element $\Sigma_{(i,j)}$ of the covariance matrix is given by,

$$\Sigma_{(i,j)} = E[(X_i-\mu_i)(X_j-\mu_j)] \tag{52}$$

where μ_i is the *i*th element of the μ vector, and X_i is the *i*th random variable in the joint normal distribution. Consider, for example, the Gaussian pdf with a full covariance matrix Σ shown in Figure 4 below.



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \frac{4}{3}\begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}$$

Gaussian pdf in two dimensions

Figure 4

Contour plot of Gaussian pdf

E. Properties of the expected value operator

- The expected value of a function $g(X)$ is given by,

$$E[g(X)] = \sum_x g(x)p(x) \text{ (discrete random variable)} \tag{53}$$

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)p(x)dx \text{ (continuous random variable)} \tag{54}$$

- *Linearity*: Let a, b be constants and let X, Y be random variables. Then,

$$E[af(X) + bg(Y)] = aE[f(X)] + bE[g(Y)] \tag{55}$$

5. Some simple problems

This section below lists some simple problems. Section 6 gives the solution to each of these problems.

A. Problem #1

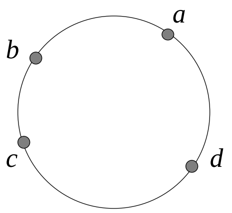
Suppose that you toss an unbiased coin 20 consecutive times, with two possible outcomes for each trial, heads (H) or tails, (T). What can you say about the relative probability of the following two experimental sequences O_1 and O_2?

$$O_1 = \{H, H, H, H, H, H, H, H, H, H, H, H, H, H, H, H, H, H, H, H\} \tag{56}$$

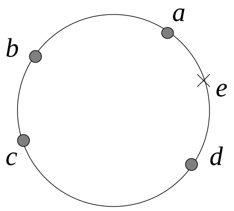
$$O_2 = \{H, T, H, H, T, T, H, T, T, T, H, T, H, H, T, H, H, T, H, T\} \tag{57}$$

B. Problem #2

(i) Consider a circle with circumference equal to 1 (radius $r = 1/(2\pi)$). Suppose that you place four points $\{a, b, c, d\}$ along the circumference of this circle at random (see picture below). What is the expected value of the arc length between adjacent points?



(ii) Now suppose that you place a fifth point e along the circumference of the circle at random. What is the expected value of the arc length between e and the two nearest points? For example, in the diagram below, what is the expected value of the arc length from a to e and from e to d?



C. Problem #3

Consider a family with exactly two children.

(i) You are told that the younger of two children is a girl. Compute the probability that one child is a boy.

(ii) You are told that at least one of the children is a girl. Compute the probability that one child is a boy.

(iii) Do your answers agree or differ for parts (i) and (ii)?

D. Problem #4

You are a contestant on *Let's Make A Deal*. Monty Hall directs your attention to three doors. He tells you that there is a prize of $1,000 behind one of the three doors, but that there is nothing behind the other two doors. You will win the prize of $1,000 if you pick the correct door.

The game proceeds as follows. First, Monty asks you to pick one of the three doors. After you have made your selection, Monty reveals that there is nothing behind one of the two doors *which you didn't pick*. He then gives you a choice: Stay with your original selection (i.e. the door that you picked originally) or switch to the other unrevealed door.

(i) What are your expected winnings if you *do not* switch from your original selection?

(ii) What are your expected winnings if you *do* switch from your original selection?

(iii) Which policy [(i) or (ii)] maximizes your expected winnings? Does it matter which policy you choose?

E. Problem #5

Consider the one-dimensional, continuous probability density function $p(x)$,

$$p(x) = \begin{cases} kx^{\lambda} & 0 \le x \le 1, \lambda > -1 \\ 0 & \text{elsewhere} \end{cases} \tag{58}$$

(i) What value of k ensures that $p(x)$ is indeed a probability density function?

(ii) Compute the expected value of x, $E[x]$ in terms of λ.

6. Simple problem solutions

This section gives the solution to each of the problems in the previous section.

A. Problem #1 solution

$P(O_1) = P(O_2) = (1/2)^{20}$, assuming independent trials and an unbiased coin.

B. Problem #2 solution

(i) Let,

$$d_1 = \text{distance between } a \text{ and } b; \tag{59}$$

$$d_2 = \text{distance between } b \text{ and } c; \tag{60}$$

$$d_3 = \text{distance between } c \text{ and } d; \tag{61}$$

$$d_4 = \text{distance between } d \text{ and } a; \tag{62}$$

Then,

$$d_1 + d_2 + d_3 + d_4 = 1 \tag{63}$$

$$E[d_1 + d_2 + d_3 + d_4] = E[1] \tag{64}$$

$$E[d_1] + E[d_2] + E[d_3] + E[d_4] = 1 \text{ (linearity)} \tag{65}$$

By symmetry,

$$E[d_1] = E[d_2] = E[d_3] = E[d_4] = E[d] \tag{66}$$

$$4E[d] = 1 \tag{67}$$

$$E[d] = 1/4 \tag{68}$$

(ii) Using a similar argument as above,

$$E[d_{ae}] = E[d_{ed}] = 1/5 \tag{69}$$

C. Problem #3 solution

(i) There are two possible cases. Let X_1 denote the younger child, and X_2 the older child; then,

$$P(X_1 = G, X_2 = B) = 1/2 \tag{70}$$

$$P(X_1 = G, X_2 = G) = 1/2 \tag{71}$$

Therefore, the probability that one child is a boy is $1/2$.

(ii) Now, there are three possible cases:

$$P(X_1 = G, X_2 = B) = 1/3 \tag{72}$$

$$P(X_1 = B, X_2 = G) = 1/3 \tag{73}$$

$$P(X_1 = G, X_2 = G) = 1/3 \tag{74}$$

Therefore, the probability that one child is a boy is now $2/3$.

(iii) No. The different results stem from the fact that the information provided in each case is different.

D. Problem #4 solution

(i) Let's label the doors A, B and C. Let X denote a discrete random variable, which denotes the door behind which the prize of $1,000 resides. Let Y denote a discrete random variable, which denotes the door that you choose. If you do not switch from your original door selection, there are nine equally likely combinations of $\{X, Y\}$, only three of which result in you winning the prize. The table below enumerates all possible outcomes, where $P(X, Y)$ denotes the probability of X and Y, and $R(X, Y)$ denotes the prize (or reward) given X and Y.

$\{X, Y\}$	$P(X, Y)$	$R(X, Y)$	$\{X, Y\}$	$P(X, Y)$	$R(X, Y)$	$\{X, Y\}$	$P(X, Y)$	$R(X, Y)$
$\{A, A\}$	$1/9$	$1,000	$\{B, A\}$	$1/9$	$0	$\{C, A\}$	$1/9$	$0
$\{A, B\}$	$1/9$	$0	$\{B, B\}$	$1/9$	$1,000	$\{C, B\}$	$1/9$	$0
$\{A, C\}$	$1/9$	$0	$\{B, C\}$	$1/9$	$0	$\{C, C\}$	$1/9$	$1,000

Therefore, our expected reward if we don't switch is given by,

$$R = \sum_{X, Y} P(X, Y)R(X, Y) \tag{75}$$

$$R = \left(\frac{1}{3}\right)(1000) = 333.33 . \tag{76}$$

(ii) If we do switch, the rewards in the above table change to $R'(X, Y)$ where,

$$R'(X, Y) = 1000 - R(X, Y) \tag{77}$$

To see that equation (77) is correct, consider the following three cases: $\{X, Y\} = \{A, A\}$, $\{X, Y\} = \{A, B\}$ and $\{X, Y\} = \{A, C\}$. The analysis is then trivially extended for $X \in \{B, C\}$.

In all three cases, we assume that the prize is behind door A. Now if we originally select door A, then Monty can reveal either of the other two doors. In either case, switching will result in 0 reward for us. Now suppose that we originally select door B. In that case, Monty is forced to reveal door C, since he can't show us the prize behind door A. By switching, we are guaranteed to win $1,000. The same is true if we originally select door C. In that case, Monty is forced to reveal door B, since he can't show us the prize behind door A. Again, by switching we are guaranteed to win $1,000.

Our expected reward if we do switch is therefore given by,

$$R = \sum_{X, Y} P(X, Y)R'(X, Y) \tag{78}$$

$$R = \left(\frac{2}{3}\right)(1000) = 666.67 \tag{79}$$

(iii) Policy (ii) (switching) maximizes our expected reward. Although this result may seem counterintuitive, you win $2/3$ of the time by switching, and only $1/3$ of the time by not switching away from your original selection. If you are still not convinced, I got a card game, I'd like to play with you...

E. Problem #5 solution

(i) In order for $p(x)$ to be a probability density function, we require that,

$$\int_0^1 p(x)dx = 1 \tag{80}$$

Expanding equation (80),

$$\int_0^1 p(x)dx = \int_0^1 kx^{\lambda}dx = \left.\frac{kx^{\lambda+1}}{\lambda+1}\right|_{x=0}^{x=1} = \frac{k}{\lambda+1} = 1 \tag{81}$$

$$k = \lambda + 1\,. \tag{82}$$

(ii) By definition of expected value,

$$E[x] = \int_{-\infty}^{\infty} xp(x)dx \tag{83}$$

$$\begin{aligned} E[x] &= \int_{-\infty}^{\infty} x(\lambda+1)x^{\lambda}dx \\ &= (\lambda+1)\int_{-\infty}^{\infty} x^{\lambda+1}dx \\ &= (\lambda+1)\left.\frac{x^{\lambda+2}}{\lambda+2}\right|_{x=0}^{x=1} \\ E[x] &= \frac{(\lambda+1)}{(\lambda+2)} \end{aligned} \tag{84}$$