Chapter 4. Probability: The Study of Randomness

4.1. Randomness

Random: A phenomenon- individual outcomes are uncertain but there is nonetheless a regular distribution of outcomes in a large number of repetitions.
Probability: long-term relative frequency.

I. Probability Theory (section 4.2 and 4.5)

The term probability refers to the study of randomness and uncertainty. The theory provides methods for qualifying the chances associated with the various outcomes. For many years a simple relative frequency definition of probability was all that was known and was all that many felt was necessary. This definition proceeds roughly as follows. Suppose that an experiment is to be performed; thus there are several possible outcomes which can occur when the experiment is performed. If an event $A$ occurs with $m$ of these outcomes, then the probability of $A$ occurring is the ratio $m/n$ where $n$ is the total number of outcomes possible.

There are many problem for which this definition is appropriate. The mathematical advances in probability theory were relatively limited and difficult to establish on a firm basis until the Russian mathematician A.N. Kolmogorov gave a simple set of three axioms which probability are assumed to obey.

When building a probability model for an experiment, we are concerned with specifying: (1) what the total collection of outcomes could be and (2) the relative frequency of occurrence of these outcomes, based on an analysis of the experiment. The probability model then consists of the assumed collection of possible outcomes and the assigned relative frequencies or probabilities of these outcomes.

Equally Likely Outcomes
If a random phenomenon has $k$ possible outcomes, all equally likely, then each individual outcome has probability $1/k$. The Probability of any event $A$ is

$$P(A) = \frac{\text{count of outcomes in event } A}{\text{count of outcomes in } S} = \frac{\text{count of outcomes in event } A}{k}$$

Properties of Probability
1. Any probability is a number between 0 and 1.
2. All possible outcomes together must have probability 1.
3. The probability that an event does not occur is 1 minus the probability that the event does not occur.
4. If the two events have no outcomes in common, the probability that one or the other occurs is the sum of their individual probabilities.

Definition:
Experiment : An action or process that generates observations.
Sample Space, $S$: A set of all possible outcomes of a random experiment(phenomenon).
Event : an outcome or a set of the sample space.

Example :
We roll a die one time. A sample space for this experiment could be
$$S = \{1,2,3,4,5,6\}$$
Then each of the sets
\[ A = \{1\}, \quad B = \{1,3,5\}, \quad C = \{2,4,6\}, \quad D = \{4,5,6\}, \quad E = \{1,2,3,4\}, \quad F = \{2,4,5\} \]

is an event.

**Some Relations from Set Theory:**

1. **Union** of two events \( A \) and \( B \), denoted by \( A \cup B \) is an event consisting of all outcomes that are either in \( A \) or in \( B \) or in both events.

2. **Intersection** of two events \( A \) and \( B \), denoted by \( A \cap B \) is an event consisting of outcomes in both \( A \) and \( B \).

3. **Complement** of \( A \), denoted by \( A^c \) is the set of all outcomes in \( S \) that are not contained in \( A \).

**Definition: Disjoint (or Mutually exclusive)**

When \( A \) and \( B \) have no outcomes in common, they are said to be disjoint events, that is,
\[ A \cap B = \emptyset \]

**Example**: \( A \cap C, \quad A \cap D, \quad A \cap F, \quad B \cap C, \quad \) and \( A \cap A^c \) are disjoint.

Given an experiment and a sample space \( S \), the objective of probability is to assign to each event \( A \) a number \( P(A) \), called the probability of the event \( A \), which will give a precise measure of the chance that \( A \) will occur. All assignments should satisfy the following axioms.

**Axiom 1)** For any event \( A \), \[ P(A) \geq 0 \]

**Axiom 2)** \[ P(S) = 1 \]

**Axiom 3)** If \( A_1, A_2, ..., A_n \) is a finite collection of disjoint events, then \[ P(A_1 \cup A_2 \cup ... \cup A_n) = \sum_{i=1}^{n} P(A_i) \]

**Properties of Probabilities**

1. For any event \( A \),
\[ P(A) = 1 - P(A^c) \]

2. If \( A \) and \( B \) are disjoint, then
\[ P(A \cap B) = 0 \]

3. For any two events \( A \) and \( B \),
\[ P(A \cup B) = P(A) + P(B) - P(A \cap B) \]

**Conditional Probability**

For any two events \( A \) and \( B \) with \( P(B) \), the conditional probability of \( A \) given that \( B \) has occurred is defined by
\[ P(A|B) = \frac{P(A \cap B)}{P(B)} \]

**Independence**

Two events \( A \) and \( B \) are independent if knowing that one occurs does not change the probability that the other occurs. If \( A \) and \( B \) are independent,
\[ P(A \text{ and } B) = P(A)P(B) \]

Equivalently, two events \( A \) and \( B \) are independent if \( P(A|B) = P(A) \).
Independence of More than Two Events

Events $A_1, \ldots, A_n$ are mutually independent if for every $k(k=2,3,\ldots,n)$ and every subset of indices $i_1, i_2, \ldots, i_k$, $P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = P(A_{i_1}) p(A_{i_2}) \ldots P(A_{i_k})$.

To paraphrase the definition, the events are mutually independent if the probability of intersection of any subset of the $n$ events is equal to the product of the individual probabilities.

Example: U.S. postal deliveries are summarized in the following table.

<table>
<thead>
<tr>
<th></th>
<th>Number of letters mailed</th>
<th>Number of arriving on time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Los Angeles</td>
<td>500</td>
<td>425</td>
</tr>
<tr>
<td>New York</td>
<td>500</td>
<td>415</td>
</tr>
<tr>
<td>Washington, D.C.</td>
<td>500</td>
<td>405</td>
</tr>
<tr>
<td>Nationwide</td>
<td>6000</td>
<td>5220</td>
</tr>
</tbody>
</table>

(a) $P$(on time delivery in Los Angeles) = 425/500
(b) $P$(late delivery in Washington, D.C.) = 1 - 405/500 = 0.19
(c) $P$(two letters mailed in New York are both delivered on time) = (415/500)(415/500) = 0.6889
(d) $P$(on time delivery nationwide) = 5220/6000 = 0.87

Two-Way Table for Categorical variables

Often two categorical variables are displayed in a table, one being the row variable, the other being the column. The table contains the counts or probabilities for each combination. The row, column and grand totals are in the marginals.

Example: A large computer software company received 500 applications for a single position. The applications are summarized in the table below.

<table>
<thead>
<tr>
<th>Degree</th>
<th>Sex</th>
<th>Computer science</th>
<th>Computer Engineering</th>
<th>Business</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>120</td>
<td>100</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>60</td>
<td>20</td>
<td>120</td>
<td></td>
</tr>
</tbody>
</table>
a) Determine the probability of applicants that had degrees in computer science or computer engineering.

b) Determine the probability of applicants that were female and had a business degree.

c) Assuming that two applicants are selected independently of each other, what is the probability that both were female applicants with computer engineering degrees.

II. Random variable, its Mean and Variance (section 4.3 and 4.4)

Whether an experiment yields qualitative or quantitative outcomes, methods of statistical analysis require that we focus on certain numerical aspects of the the data (such as a sample proportion $\frac{x}{n}$, mean $\overline{x}$, or standard deviation $s$). The concepts of a random variable allow us to pass from the experimental outcomes themselves to a numerical functions of the outcomes. There are two fundamental different types of random variables - discrete random variables and continuous random variables.

Definitions:

Random variable (r.v.): 1. assigns a real number to each element in sample space.
2. A variable whose value is a numerical outcome of a random experiment (phenomenon).

Discrete random variable: 1. has a finite number of possible values.
2. Possible values of discrete random variable are isolated points along the number line.

Continuous random variable: 1. takes all values in an interval of numbers.
2. Possible values form an interval along the number line.

The mean of $X$: $\mu$

The variance of $X$: $\sigma^2$

The standard deviation of $X$: $\sigma$

1. Population Distribution for a Discrete Random Variable, $X$

The probability distribution of $X$ lists the values of $X$ and their probabilities.

<table>
<thead>
<tr>
<th>Value of $X$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>...</th>
<th>$x_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>$p_1$</td>
<td>$p_2$</td>
<td>$p_3$</td>
<td>...</td>
<td>$p_k$</td>
</tr>
</tbody>
</table>

Probabilities $p_i$ must satisfy two requirements:
1. Every probability $p_i$ is a number between 0 and 1.
2. $p_1 + p_2 + ... + p_k = 1$.

Example 1: When a student attempts to log on to computer system, either all ports could be busy (F), in which case the student will fail to obtain access, or else there will be at least one port free (S), in which case the student will be successful in accessing the system. With $S = \{S, F\}$, define an r.v. $X$ by $X(S) = 1$, $X(F) = 0$. The r.v. $X$ indicates whether (1) or not (0) the student can log on.
An Infinite Set of Possible \( X \) Values:
We consider an experiment in which batteries coming off an assembly line are examined until a good one (\( S \)) is obtained. The sample space is \( S=\{S,FS,FFS,FFFS,\ldots\} \). Define an r.v. \( X \) by \( X \) = the number of batteries examined before the experiment terminates. Then \( X(S)=1, X(FS)=2, X(FFS)=3, \ldots \). Any positive integer is a possible value of \( X \), so the set of possible value is infinite.

Probability Mass Function (pmf):
The probability mass function of a discrete r.v. is defined for every number \( x \) by
\[
P(X=x) \overset{\text{def}}{=} P(\text{all } s \in S : X(s)=x)
\]
\( P(X=x) \) is the sum of the probabilities for all events for which \( X=x \).

Example 2: An automobile service facility specializing in engine tune-ups knows that 45% of all tune-ups are done on four-cylinder automobiles, 40% on six-cylinder automobiles, and 15% on eight-cylinder automobiles. Let \( X = \) number of cylinders of the next car to be tuned. What is the pmf of \( X \)?

Cumulative Distribution Function (cdf):
The cdf \( F(x) \) of a discrete r.v. \( X \) with pmf \( P(X=x) \) is defined for every number \( x \) by
\[
F(x) \overset{\text{def}}{=} P(X \leq x) = \sum_{y:y \leq x} P(y)
\]

Example 3: What is the cdf of \( X \) in example 2?

Expected Value of \( X \):
Let \( X \) be a discrete r.v. with set of possible values \( x_1, x_2, x_3, \ldots \), occurring with probabilities \( P(X=x_1), P(X=x_2), P(X=x_3), \ldots \), then the expected value of \( X \) or mean of \( X \), denoted by \( E(X) \) or \( \mu_X \), is
\[
E(X) = \mu_X = \sum_{i=1}^{\infty} x_i \cdot P(X=x_i)
\]

Variance and Standard Deviation of \( X \):
Variance of \( X = Var(X) = \sigma^2 = \sum_{i=1}^{\infty} (x_i - \mu_X)^2 \cdot P(X=x_i) \)
Standard deviation of \( X = \sqrt{Var(X)} = \sigma = \sqrt{\sum (x_i - \mu_X)^2 \cdot P(X=x_i)} \)

Example 4: If \( X \) has the following pmf, \( E(X) \)? \( Var(X) \)?

<table>
<thead>
<tr>
<th>( X=x )</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(X=x) )</td>
<td>0.3</td>
<td>0.4</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Probabilistic Properties of a Discrete Random Variable, \( X \):
1) \( P(X > a) = 1 - P(X \leq a) \), where \( a \) is a constant value.
2) \( P(X \geq a) = 1 - P(X < a) \)
3) \( P(X \leq a) + P(X > a) = 1 \)
4) \( P(a < X < b) = P(X < b) - P(X \leq a) \), where \( a \) and \( b \) is constant values.
5) \( P(a \leq X \leq b) = P(X \leq b) - P(X < a) \)

**Example 5:** A pizza shop sells pizzas in four different sizes. The 1000 most recent orders for a single pizza gave the following proportions for the various sizes.

<table>
<thead>
<tr>
<th>size</th>
<th>12&quot;</th>
<th>14&quot;</th>
<th>16&quot;</th>
<th>18&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>proportion</td>
<td>0.2</td>
<td>0.25</td>
<td>0.5</td>
<td>0.05</td>
</tr>
</tbody>
</table>

With \( X \) denoting the size of a pizza in a single-pizza order, the above table is an approximation to the population distribution of \( X \)

a) Construct a histogram to represent the approximate distribution of this variable.
b) Approximate \( P(X < 16) \)
c) Approximate \( P(X \leq 16) \)
d) Approximate \( P(14 < X \leq 16) \)
e) Approximate \( P(14 \leq X \leq 16) \)
f) Show that the mean value of \( X \) is approximately 14.8".
g) What is the approximate probability that \( X \) is within 2" of this mean value?
h) What is the variance of \( X \)?

**Binomial random variable and its distribution (section 5.1)**

There are many experiments that conform the following list of requirements

1. The experiment consists of a sequence of \( n \) trials, where \( n \) is fixed in advance of the experiment.
2. The trials are identical, and each trial can result in one of the same two possible outcomes, which we denote by success(S) and failure(F).
3. The trials are independent, so that the outcome on any particular trial does not influence the outcome on any other outcomes.
4. The probability of success is constant from trial to trial; we denote this probability by \( p \).

An experiment for which conditions 1-4 are satisfied is called a **binomial experiment**.

Given a binomial experiment consisting of \( n \) trials, the **binomial random variable** \( X \) associated with this experiment is defined as \( X = \) the number of S's among the \( n \) trials.

\[ X \sim Bin(n, p) \]

pmf of \( X \): \( P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \) for \( x = 0,1,2,3,... \)

cdf of \( X \): \( F(x) = P(X \leq x) = \sum_{t=0}^{x} P(X = t) = \sum_{t=0}^{x} \frac{n!}{t!(n-t)!} p^t (1-p)^{n-t} \)

\[ \mu_X = np \quad Var(X) = \sigma_X^2 = np(1-p) \quad \text{and} \quad \sigma_X = \sqrt{np(1-p)} \]

2. Population Distribution for a Continuous Random Variable, \( X \)
A continuous probability distribution is a smooth curve, that serves as a model for the population distribution of a continuous random variable.

The probability distribution of X is described by a density curve. The probability of any event is the area under the density curve and above the values of X that make up the event.

**Probability Density Function (pdf):**
Let X be a continuous r.v., then a pdf of X is a function \( f(x) \) such that for any two numbers \( a \) and \( b \) with \( a < b \),
\[
P(a \leq X \leq b) = \int_a^b f(x) \, dx.
\]
That is, the probability that X takes on a value in the interval \([a,b]\) is the area under the graph of the density function.

**Example 6:** Suppose I take a bus to work, and that every 5 minutes a bus arrives at my stop. Because of variation in the time that I leave my house, I don't always arrive at the bus stop at the same time, so my waiting time \( X \) for the next bus is a continuous r.v.. The set of possible values of \( X \) is the interval \([0,5]\).

One possible pdf for \( X \) is
\[
f(x) = \frac{1}{5}, \text{ if } 0 \leq x \leq 5.
\]
Thus the probability that I wait between 1 and 3 minutes is
\[
P(1 \leq X \leq 3) = \frac{2}{5}.
\]
Because whenever \( 0 \leq a \leq b \leq 5 \), \( P(a \leq X \leq b) \) depends only on the length \( b-a \) of the interval, X is said to have a **uniform distribution**.

**Probabilistic Properties of a Continuous Random Variables, \( X \):**

1) The total area under the curve is 1.
2) \( P(X=c) = 0 \) for any number \( c \). That is, all continuous probability distributions assign probability 0 to every individual outcome.
3) \( P(X < a) = P(X \leq a) \), where \( a \) is any constant.
\[
P(X > a) = 1 - P(X \leq a)
\]
\( P(a \leq X \leq b) = P(a < X < b) \), where \( b \) is any constant.
\[
P(X > a) = 1 - P(X \leq a)
\]
\( P(X < b) = P(X < b) - P(X \leq a) \)

**Normal Distribution (section 1.3 and 4.3)**

Normal distribution is a continuous probability distribution with mean \( \mu \) and variance \( \sigma^2 \). It can be written as \( N(\mu, \sigma^2) \). It is bell-shaped and therefore symmetric. The standard normal distribution, \( N(0,1) \) is the normal distribution with mean 0 and variance 1. It is known as Z curve and corresponding probabilities can be determined using the Standard Normal table (z table).
\[
X \sim N(\mu, \sigma^2) = \frac{1}{\sqrt{2 \pi \sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty
\]
\[
P(Z > z) = 1 - P(Z \leq z)
\]
\[ P(Z \leq z) + P(Z > z) = 1 \]
\[ P(z_1 < Z < z_2) = P(z_1 \leq Z < z_2) = p(z_1 < Z \leq z_2) = P(z_1 \leq Z \leq z_2) \]
\[ P(Z < z_2) - P(Z \leq z_1) = P(Z < z_2) - P(Z < z_1) = P(Z \leq z_2) - P(Z \leq z_1) = P(Z \leq z_2) - P(Z < z_1) \]

If \( X \) is a normally distributed variable with mean \( \mu \) and variance \( \sigma^2 \), then
\[ P(X > a) = P\left( \frac{X - \mu}{\sigma} > \frac{a - \mu}{\sigma} \right) = p\left( Z > \frac{a - \mu}{\sigma} \right) \]

since \( Z = \frac{X - \mu}{\sigma} \), that is, \( X = \mu + \sigma Z \).

**Example 8**: Compute the following probabilities.

a) \( P(Z \leq 1.25) \)

b) \( P(Z \geq 1.25) \)

c) \( P(Z \leq -1.25) \)

d) \( P(-0.38 \leq Z \leq 1.25) \)

\( z_{\alpha} \) notation:
\( z_{\alpha} \) will denote the value on the measurement axis for which \( \alpha \) of the area under the \( z \) curve lies to the right of \( z_{\alpha} \). The \( z_{\alpha} \)'s are usually referred to as \( z \) critical values.

**Empirical Rule**: If the population distribution of a variable is (approximately) normal, then
1. Roughly 68\% of the values are within 1SD of the mean.
2. Roughly 95\% of the values are within 2SDs of the mean.
3. Roughly 99.7\% of the values are within 3SDs of the mean.

**Percentiles of an Arbitrary Normal Distribution**: The \((100p)\)th percentile of a normal distribution with mean \( \mu \) and standard deviation \( \sigma \) is easily related to the \((100p)\)th percentile of the standard normal distribution. \((100p)\)th percentile for \( N(\mu, \sigma) \) = \( \mu + [(100p)\text{th percentile in Standard Normal distribution}] \times \sigma \).

**Checking Normality**: To see the population distribution is normal, we use normal probability plot.

(Note) Boxplots and histograms summarize distributions, hiding individual data points to simplify overall patterns. Normal quantile plot, in contrast, include points for every case in the distribution. It is harder to read than boxplot and histogram, but conveys more detailed information. The linearity of the plot supports the assumption that a distribution from which observations were drawn is normal.