Basic Probability Concepts, Random Variables and Sampling Distribution
Chapters 6, 7, and 8 (Siegel)

Rationale

For practical reasons, variables are observed to collect data. The sampled data is then analyzed to elicit information for decision making in business and indeed in all human endeavors. However, sampled information is incomplete and not free from sampling error. Its use in decision-making processes introduces an element of chance. Therefore, it is important for a decision-maker to know the amount of chance associated with a statistical decision of it being wrong. To quantify the amount of chance due to sampling error, basic probability concepts are indispensable via modeling sampled populations and testing of research hypotheses.

Objectives

Understand

- Different interpretations of probability of an event
- Rules for assigning probabilities to events
- Conditional probability and statistical independence
- Probability trees
- Binomial and normal distributions
- Sampling distribution of the sample mean.

Key words

- Mutually exclusive events
- Conditional probability, independence of events and probability trees
- Random variable, probability distribution
- Normal distribution, binomial distribution
- Sampling distribution of the sample mean
1. Basic Probability Concepts

Consider the following chance statements for events within (brackets).

- Chance of (rain today) is 10%.
- Chance of a (correct medical diagnosis) is 80%.
- Chance of (an adult Americans not being able to read and write adequately) is 50%.
- Chance of (the DJIA to hit 10,000 by the end of 1999) is 70%.
- Chance of (the face with head up in a toss of a dime) is 50%.
- Chance of (winning a lottery) is 0.000001%

Each statement gives odds in favor of an event as a percentage. Technically, an event is a subset of possible outcomes of an experiment. For example, tossing of a six-faced die has six possible outcomes (1, 2, 3, 4, 5, 6). A subset (2, 4, 6) is an event which in words states the outcome is a face with even number on it. Capital letters A, B, C, etc. are used to denote events. The chance of an event A is expressed as a percentage but probability, denoted by P(A), is a number between 0 and 1.

\[ P(A) \times 100 = \text{chance of } A \]

Examples in Section 6.3.

Laws of Probability

1. For an impossible event A, P(A) = 0, and P(A) = 1 if the occurrence of A is a certainty.

2. Complement of A, denoted as “Not A”, is an event which occurs whenever A does not. If an experimental outcome is in A as well as in B, then the event "A and B" is said to have occurred. Two events A and B are mutually exclusive events if P(A and B) = 0. A compound
event denoted as "AorB" means that at least one of them (either A or B or both A and B) occurred. The law of addition of probability is:

3. \( P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \)

If A and B are mutually exclusive events, then \( P(A \text{ or } B) = P(A) + P(B) \).

**Conditional Probability**

The probability of an event A given the information that an event B has occurred is denoted by \( P(A/B) \). It is called the conditional probability. The multiplication law of probability is:

4. \[ P(A/B) = \frac{P(A \text{ and } B)}{P(B)} \]

Equivalently the above can rewritten as:

\( P(A \text{ and } B) = P(A/B) \cdot P(B) \) or \( P(A \text{ and } B) = P(B/A) \cdot P(A) \).

**Independence**

Two events A and B are statistically independent if the following equivalent statements hold.

i) \( P(A) = P(A/B) \),  
ii) \( P(B) = P(B/A) \),  
iii) \( P(A \text{ and } B) = P(A) \cdot P(B) \)

To prove independence of two events, check any one of the three equivalent statements.

*Examples in Section 6.4*

**Probability Trees**

Construction of probability trees is an easy tool for computing and understanding probabilities.

*Examples in Section 6.6.*
2. Random Variables and Probability Models

A random variable is a numerical measurement, which quantifies the same aspect for each unit in a ‘population’. Consider the following statements.

- Amount of weight loss from following an exercise plan for a population of humans
- Rate of return from investment for a population of mutual funds
- Amount of money spent yearly on research and development by utility companies
- Number of bathrooms per home in a township
- Preference score of consumers for a product

The numerical quantity in each example is likely to vary from unit to unit. That is why, it is a ‘variable’. Capital letters X, Y etc. are used to denote variables. The totality of all scores of a variable X is considered its population. Since not all scores of a variable are equally likely, a model describing its population is called its probability distribution.

Consider consumers' preference score X for a product. For simplicity, suppose X is -1, 0, or 1 if a consumer does not like, has no preference, or likes the product respectively. Suppose 20% consumers do not like, 50% have no preference and the remaining 30% like the product. Then the population of X has as many X scores as the number of consumers consisting of numbers -1, 0 and 1 occurring with frequency 20%, 50% and 30% respectively. In statistical terminology, this is called a probability distribution of X, and written as:

\[
\begin{array}{ccc}
X: & -1 & 0 & 1 \\
P(X): & .20 & .50 & .30 \\
\end{array}
\]

P(-1), for example, denotes the probability of X = -1

Graphically the above probability distribution of X can be presented as:
The mean (expected value), and the standard deviation, S.D., of the distribution of a variable X are given by

\[ \mu = \sum X \cdot P(X), \text{ and} \]
\[ \sigma = \sqrt{\sum (X - \mu)^2 \cdot P(X)}, \text{ respectively.} \]

For the distribution of consumer preference score X given above,

\[ \mu = \sum X \cdot P(X) = -1(.2) + 0(.5) + 1(.3) = 0.1, \text{ and} \]
\[ \sigma = \sqrt{\sum (X - \mu)^2 \cdot P(X)} = \sqrt{(-1 - .1)^2(.2) + (0 -.1)^2(.5) + (1 - .1)^2(.3)}} \]
\[ = \sqrt{.242 + .005 + .243} = \sqrt{.49} = 0.7 \]

*Examples in Section 7.1*

**The Binomial Distribution:**

Consider the following statements. In each, interest is in the number of times a specified event occurs when an experiment is observed a fixed number of times.

- Number of females children born among 30 births.
- Number of defective items in a shipment of 200 items.
- Number of yes responses in a poll of 500 citizens.

When an experiment is performed, suppose either an event of interest occurs with probability \( \pi \) or fails to occur with probability \( 1 - \pi \). Further suppose the experiment is performed \( n \) times independently. Let \( X \) be the number of times out of \( n \), the experiment resulting in the event of interest. The possible values of \( X \) are: 0, 1, 2, ......., \( n \), and it said
to have the binomial probability distribution with parameters \( n \) and \( \pi \). The binomial distribution is a function, which assigns a probability to each value of \( X \). It has been tabulated extensively for various combinations of \( n \) and \( \pi \) for calculating probabilities of binomial events. Also, statistical software can be used to calculate binomial probabilities.

The mean and standard deviation of a binomial distribution with parameters \( n \) and \( \pi \) are \( \text{E}(X) = \mu = n \pi \) and \( \text{S.D.} = \sigma = \sqrt{n\pi(1 - \pi)} \) respectively. For \( n = 25 \), and \( \pi = 0.25 \), \( \mu = 25(0.25) = 6.25 \), \( \text{SD} = \sqrt{25(0.25)(0.75)} = 2.165 \), and the binomial distribution is displayed below.

![Binomial distribution, \( n = 25, \ \pi = 0.45 \)]

Use Excel for examples in Section 7.2

Above probability distributions are examples of models for discrete variables. There are also variables that are recorded on continuous scales. Their probability model is seen by a curve (function). The total area under a probability curve is 100%. The probability of observing a continuous variable, say in an interval \((a \leq x \leq b)\), is given by the area under the curve spanned over the interval.

**The Normal Distribution:**

The population of a variable \( X \) is normally distributed if it can be described by a bell-shape curve shown below. Suppose the annual performance \( X \) of mutual funds is normally distributed with mean annual performance \( \mu = 12\% \) and S.D. \( \sigma = 2\% \), then the normal curve has maximum height at \( x = 12 \) and is symmetric around it. Also, the curve essentially ends at \( 12 \pm 3(2) \). The curve below is a sketch of this normal distribution. Probabilities of various events for a normal distribution can be calculated using Excel or other software. For example, The probability of the fund’s performance being in between 8\% and 16\%, may be calculated as:
P(8 < X < 16) = P(X < 16) - P(X < 8) = Normal CDF 16 - Normal CDF 8
= 0.9772 - 0.0228 = 0.9544

The Standard Normal Distribution:

A variable X with mean \( \mu \) and S.D. \( \sigma \) can be mapped into a variable Z with mean 0 and S.D. 1 by the following transformation.

\[
Z = \frac{(X - \mu)}{\sigma}
\]

If X scores are normally distributed, then so are the Z-scores. The distribution of the Z scores is called the standard normal Distribution. One can also use the tables of the standard normal distribution for calculating probabilities of the normal distribution. For example, the event (8 < X < 16) on the X-scale can be transformed to (- 2 < Z < 2) on the Z-scale. That is,

\[
P(8 < X < 16) = P\left[\frac{(8 - 12)}{2} < Z < \frac{(16 - 12)}{2}\right]
= P(-2 < Z < 2) = P(Z < 2) - P(Z < -2).
= .9772 - .0228 = .9544
\]

Examples in Section 7.4

3. Random Sampling

For applying statistical methods, data must be collected to have a representative random sample. There is a vast literature dealing with methods for collecting data scientifically to study variables of interest and their relationships to each other. For random sampling,
experimental devices are used to collect observations independently from the population of a variable following its probability distribution. In practice n units are randomly selected from a well-defined frame so that each unit of the frame has equal chance of being chosen. This can be accomplished with the help of a table of Random Numbers (see text) or a statistical software (Excel, for example) for generating random numbers.

*Examples in Section 8.2.*

**Sampling Distribution of the Sample Mean**

It is crucial to recognize that there are many different possible samples, each of size n, even though an experimenter would observe only one of them randomly to elicit information. Thus, any sample quantity (statistic) will have as many numerical values associated with it as the number of different samples. A statistic, when computed from an observed sample for use in decision-making, introduces an element of chance. To quantify this chance, which is inherent in all statistical decisions, statisticians study probability distributions (sampling distributions) of statistics.

One frequently used statistic is the sample mean. For large samples (n large), the sampling distribution of $\bar{x}$ is normal with mean equal to that of the population $\mu$, and the standard deviation equal to the standard deviation $\sigma$ of the population divided by the square root of n, i.e., $\sigma(\bar{x}) = \sigma / \sqrt{n}$. When $\sigma$ is replaced by the sample standard deviation $S$, that is, $S / \sqrt{n}$, it is called the standard error of the mean $\bar{x}$ (s.e. of $\bar{x}$). The s.e. of $\bar{x}$ quantifies the amount of distance one can typically expect in its observed value from its population mean $\mu$.

*Examples in Section 8.3 and 8.4.*