[1]

Probability Concepts and Random Variables

Graeme Warren
Leavey School of Business
gwarren@scu.edu

> Probability is expectation founded upon partial knowledge. A perfect acquaintance with all the circumstances affecting the occurrence of an event would change expectation into certainty, and leave nether room nor demand for a theory of probabilities.
>
> - George Boole

Contents

Learning Objectives

After completing this module you should be able to:

1. Explain how probabilities are interpreted, assigned, and mathematically characterized.
2. Explain and apply the concepts of mutual exclusivity, collective exhaustivity, and independence.
3. Apply the complement rule, addition rule, multiplication law, and definition of conditional probability to compute marginal, joint, conditional, and union probabilities.
4. Compute the cumulative distribution function, expected value, and variance of a discrete random variable.
5. Execute calculations using the cumulative distribution function and inverse cumulative distribution functions of the normal distribution using Microsoft Excel.
6. Compute the expected value and variance of linear combinations of independent random variables.

Probability

Random variables and their associated probability distributions provide analytic models of random behavior that can be used to support business decision making in the presence of risk and variability. Probability has many other applications. One of particular interest to us - the calculation of the expected value and variance of a linear combination of random variables - will reappear in our investigation of sampling distributions in a future module. Sampling distributions underpin the statistical techniques of interval estimation and significance testing, which in turn are foundations of the scientific method. See (Diez et al, 2015) section 2.1 for additional reading on this topic.

Assignment and Interpretation of Probability

A *random process* or *experiment* gives rise to two or more outcomes. Random processes that may be of interest to business analysts or researchers could include sales of a particular product in different stores, the number of dependents claimed on tax returns, social media posts on a particular topic, national capacity utilization over time, percentage changes in security prices of firms in a market sector, etc. Two critically important random processes in business are those of controlled experiments (in which variables of interest are systematically manipulated to discern their effect) and surveys.

The set of possible outcomes of these processes are quantitative (e.g., integer-valued or real-valued) or qualitative (e.g., preferences expressed on an “Excellent, Good, Average, Poor, Terrible” scale). For example, the set of outcomes for the number of dependents claimed on a tax return is $\{0, 1, 2, 3, \ldots\}$, and the national capacity utilization takes on values in the set $[0\%, 100\%]$. Other common examples of random processes include tossing of a single coin (for which the outcomes are $\{head, tail\}$), rolling of a single die (for which the outcomes are $\{1, 2, 3, 4, 5, 6\}$), or drawing of a card from a deck of playing cards (for which there are 52 possible outcomes).

An *event* is a collection of one or more outcomes of an experiment of interest to an analyst or researcher. For example, an analyst may be interested in the event that the national capacity utilization is considered high (e.g., in the range $[85\%, 100\%]$) or the event that a tax return claims more than four independents. She may be interested in the event that the stock market is in a bear market phase or the event of obtaining a face card (jack, queen, or king) when drawing a card from a deck of playing cards.

Two events are said to be *mutually exclusive* if they cannot occur simultaneously. For example, the event that we obtain an odd-numbered outcome and the event that we obtain an even-numbered outcome on a single die roll are mutually exclusive. Similarly, a stock market cannot be in both a bull market and bear market phase. These phases are therefore mutually exclusive. An individual’s credit rating cannot be both excellent and under 700. The credit events “excellent” and “under 700” are therefore mutually exclusive. See (Diez et al, 2015) section 2.1.2 for additional reading on this topic.

A set of outcomes are said to be *collectively exhaustive* if they fully describe all possible outcomes that can occur in a random process or experiment. For example, {Strong Buy, Buy, Hold, Underperform, Sell} is the set of collectively exhaustive forecast recommendations for a security at nasdaq.com. Similarly, $\{1, 2, 3, 4, 5, 6\}$ is the set of collectively exhaustive outcomes of a single die roll.

A *probability distribution* is a mathematical function that conveys the probability of occurrence of all the possible different outcomes of an experiment. We will use the notation $P(x)$ to denote the probability of occurrence of outcome x. For example, the probability distribution of outcomes of a single toss of a fair coin can be written as $P(Head) = P(Tail) = 0.5$. A probability distribution is a list of the possible outcomes with corresponding probabilities that satisfies three rules:

1. The outcomes listed must be disjoint and collectively exhaustive.

2. The probability of each outcome must be between 0 and 1.

3. The sum of the probabilities of all outcomes must total 1.

Three candidate distributions that reference events based upon a roll of a fair die appear in Tables 1, 2 and 3. "Even" is event denoting an outcome in the set

$\{2, 4, 6\}$. "Odd" is an event denoting an outcome in the set $\{1, 3, 5\}$. "Prime" is an event denoting an outcome in the set $\{2, 3, 5\}$. "> 4 is an event denoting an outcome in the set $\{5, 6\}$. The "even"/"prime" distribution in Table 1 is defective as the events are not disjoint (2 is both even and prime). The "even"/"> 4" distribution in Table 2 is defective as the probabilities do not sum to 1. Finally, the distribution in Table 3 satisfies the three probability rules.

These rules can be characterized mathematically as follows. Letting $S = \{O_1, O_2, \ldots, O_k\}$ denote the list of outcomes $O_1, O_2, \ldots, O_k$ of an experiment (S is the so-called *sample space* of the experiment), the rules require that:

1. $O_1, O_2, \ldots, O_k$ are mutually exclusive and collectively exhaustive,

2. $0 \leq P(O_i) \leq 1$ for all i, and

3. $\sum_{i=1}^{k} P(O_i) = 1$.

These requirements are a special case of Kolmogorov's probability axioms. See (Diez et al, 2015) section 2.1.4 for additional reading.

Outcome	$P(Outcome)$
Even	0.5
Prime	0.5
Total	1

Table 1. *"Even"/"prime."*

Outcome	$P(Outcome)$
Even	0.5
> 4	$0.\dot{3}$
Total	$0.\dot{8}$

Table 2. *"Even"/"> 4."*

Outcome	$P(Outcome)$
Even	0.5
Odd	0.5
Total	1

Table 3. *"Even"/"odd."*

We consider three ways of assigning or interpreting probabilities: the *classical*, *Bayesian*, and *frequentist* approaches. Probabilities must satisfy the three rules irrespective of how they are assigned. The key assumption in the classical assignment of probability is that all outcomes of a random process or experiment are equally likely. Thus, if there are n outcomes, classical probability assigns a likelihood of $1/n$ to each outcome. For example, the probability of rolling any number in the set $\{1, 2, 3, 4, 5, 6\}$ on an unbiased die is $1/6$. Classical probability assignments may be useful in Vegas, but tend to have limited application in business because the equally-likely-outcome assumption is typically associated only with special or artificial decision-making situations in which more refined knowledge about the likelihood of outcomes is unavailable. There are therefore two main ways of interpreting probability that are of interest to business professionals - the Bayesian and frequentist approaches. The objectives and tools used in the two approaches differ. There are pros and cons associated with both.

The Bayesian approach regards probability as a measure of belief that is updated as new information becomes available. Bayesian statistics has essentially one tool - Bayes' theorem - which is used to successively update an initial probability distribution, as new information becomes available, to obtain a revised probability distribution. This revised distribution then becomes the initial distribution during the next update. The Bayesian approach is often seen to involve subjective probability estimates as the original initial probability distribution is subjectively chosen.

In this course we focus solely on the frequentist approach. The frequentist approach (also known as the empirical-probability approach or relative-frequency approach), in contrast, views probabilities as the frequency with which an outcome or event has occurred or will occur over many replications of a random process. The frequentist approach draws on historical or simulation data. The probability of an outcome using the frequentist approach is the fraction of the process replications it has occurred in the past. The frequentist estimate of the likelihood of an outcome occurring converges by the law of large numbers to the true probability as the number of trials used to figure the estimate increases. The law of large numbers holds that the fraction of occurrences of a particular outcome converges to the probability of that outcome as the number of process replications increases. An app that shows this convergence in the case of coin flips can be viewed here.

The frequentist probability of an outcome is the proportion of times it would occur if the process was observed or replicated an infinite number of times. Calculations based upon finite data are therefore estimates of frequentist probabilities. Structural changes (such as changes in the "causal mechanism") over time of a random process challenge the use of frequentist probability estimates.

For example, suppose that in the last 1,000 trading days the S&P 500 index has been up at the close 542 times and down 458 times. The frequentist estimates of the up and down probabilities are, respectively, $P(Up) = 0.542$ and $P(Down) = 0.458$. Up and $Down$ are obviously mutually exclusive and assumed to be collectively exhaustive (i.e., we assume that there were no days on which the index ended flat). Confirm that the rules are satisfied.

Constructs and Rules

We need to be able to perform consistent calculations (i.e., calculations that ensure compliance with the rules) with probabilities once they are assigned. We consider probability constructs and rules for this purpose. The four probability constructs we consider are:

1. *Marginal probability*, which deals with the probability of a single event. The marginal probability of event A is written $P(A)$.

2. *Joint probability*, which deals with the probability of two events. A joint probability is the probability of the compound event comprised of the two component events. The joint probability of two events A and B is written $P(A \cap B)$ or $P(AB)$. $A \cap B$ is the compound event that both A and B occur.

3. *Conditional probability*, which deals with the probability that an event A will occur *given* that some other event B has occurred. The probability of event A given that event B has occurred is written $P(A|B)$.

4. The union probability of events A or B (or both) occurring is written $P(A \cup B)$. $A \cup B$ is the event that either A or B or both occur.

To illustrate these constructs, consider the crosstab in Table 4 showing the housing situation by generation in the San Francisco Bay area extracted from a 2016 survey of the San Francisco Foundation.

	Housing Situation		
Generation	Own	Rent	Total
18-44	125	211	336
45-64	169	95	264
65+	166	34	200
Total	460	340	800

Table 4. *Crosstab of housing situation by generation in the SF Bay area based on survey of 800 respondents.*

	Housing Situation		
Generation	Own	Rent	Total
18-44	0.16	0.26	0.42
45-64	0.21	0.12	0.33
65+	0.21	0.04	0.25
Total	0.58	0.42	1

Table 5. *Related table showing marginal (shown in cyan) and joint probabilities (shown in red) of housing situation by generation.*

To calculate marginal probabilities we divide by the total number of respondents (800) in each case. For example, to find the marginal $P(Own)$, divide the number of respondents who own (460) by the total of respondents (800). We get

$$P(Own) = 460/800 \approx 0.58$$

This is a frequentist estimate of the probability that a randomly chosen person in the San Francisco Bay area owns a home. Check your knowledge on the following questions:

- What is[2] $P(65+)$?
- What is[3] $P(Rent)$?
- What is[4] $P(18-44)$?

Notice that the marginal probabilities appear with a cyan background in the bottom row and rightmost column of Table 5.

Table 5 is obtained by dividing all entries of Table 4 by the number of respondents (800). It is called a joint probability table. To calculate the joint probability $P(Own \cap 18-44)$ we take the number of respondents who own **and** are in the 18-44 age group and divide by the total number of respondents. We get

$$P(Own \cap 18-44) = 125/800 \approx 0.16$$

[2] $P(65+) = 200/800 \approx 0.25$

[3] $P(Rent) = 340/800 \approx 0.42$

[4] $P(18-44) = 336/800 \approx 0.42$

This is a frequentist estimate of the probability that a randomly person in the San Francisco Bay area is aged 18 to 44 and owns a home. Check your knowledge on the following questions:

- What is[5] $P(Own \cap 65{+})$?
- What is[6] $P(65 + \cap Rent)$?
- What is [7] $P(Rent \cap 18 - 44)$?

Notice that the joint probabilities appear with a red background in the "center" of Table 5.

We will now apply the following rules, all of which can be shown to be direct consequences of the probability axioms:

- The complement rule,
- Conditional probability,
- The multiplication rule (also known as chain rule), and
- The addition rule.

Complement Rule. The complement rule gives us the probability of the complementary event. For event A, the complementary event "not A" is denoted by A^c and its probability is $P(A^c) = 1 - P(A)$. For example, using the data in Tables 1 and 2,

$$P(Own^c) = 1 - P(Own) \approx 1 - 0.58 = 0.42$$

This is a frequentist estimate of the probability that a randomly chosen person in the San Francisco Bay area does not own a home. See (Diez et al, 2015) section 2.1.5 for additional reading on this topic. Check your knowledge on the following questions using the data in tables 4 and 5,

- What is[8] $P(45 - 64^c)$?
- What is[9] $P(65{+}^c)$?

[5] $P(Own \cap 65{+}) = 166/800 \approx 0.21$

[6] $P(65 + \cap Rent) = 34/800 \approx 0.04$

[7] $P(Rent \cap 18 - 44) = 211/800 \approx 0.26$

[8] $P(45 - 64^c) = 1 - P(45 - 64) \approx 1 - 0.33 = 0.67$

[9] $P(65{+}^c) = 1 - P(65{+}) = 1 - 0.25 = 0.75$

Conditional Probability. The conditional probability of an event A given an event B is defined as

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

Notice how the definition of conditional probability establishes a relationship between the conditional probability $P(A|B)$, marginal probability $P(B)$, and the joint probability $P(A \cap B)$. See (Diez et al, 2015) section 2.2.2 for additional reading on this topic.

For example, using the data in tables 4 and 5

$$P(Own|18-44) = \frac{P(Own \cap 18-44)}{P(18-44)} \approx \frac{0.16}{0.42} \approx 0.38$$

This is a frequentist estimate of the probability that a randomly chosen person in the San Francisco Bay area owns a home **given** that she is 18 to 44 years old. Interestingly, the estimate of the probability that a randomly chosen person in the San Francisco Bay area is 18 to 44 years old **given** that she owns a home is

$$P(18-44|Own) = \frac{P(18-44 \cap Own)}{P(Own)} \approx \frac{0.16}{0.58} \approx 0.28$$

So $P(18-44|Own) \neq P(Own|18-44)$. In general $P(A|B) \neq P(B|A)$. Check your knowledge on the following questions:

- What is $P(65+|Rent)$[10]?
- What is $P(Rent|65+)$?[11]?

Multiplication Rule. The multiplication rule is a near equivalent of the conditional probability formula. It establishes the relationship

$$P(A \cap B) = P(A|B)P(B)$$

We could use it, for example, to calculate the probability that a randomly-chosen respondent rents and is 65+ as

$$P(Rent \cap 65+) = P(Rent|65+)P(65+) \approx 0.16 \times 0.25 = 0.04$$

Two events are said to be *independent* if the occurrence of either one of them has no effect on the probability of occurrence of the other. We can check this mathematically: events A and B are independent if and only if

$$P(A|B) = P(A)$$

[10] $P(65+|Rent) \approx P(65+\cap Rent)/P(Rent) \approx 0.04/0.42 \approx 0.1$

[11] $P(Rent|65+) \approx P(65+\cap Rent)/P(65+) \approx 0.04/0.25 = 0.16$

otherwise they are not independent. This is equivalent to showing that

$$P(A \cap B) = P(A)P(B)$$

Since $P(Own|18{-}44) \approx 0.38 \neq 0.58 \approx P(Own)$, we conclude that owning a home and membership of the 18-44 age group are not independent events in the San Francisco Bay area. See (Diez et al, 2015) section 2.1.6 and 2.2.4 for additional reading on this topic. Check your knowledge on the following questions:

- Are home ownership and membership of the 65+ age group in the San Francisco Bay area independent events[12]?
- Are renting and membership of 18-44 age group independent events[13]?

Addition Rule. The addition rule provides a relationship between $P(A \cup B)$ and marginal and joint probabilities:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

For example, we could estimate the probability that a randomly-chosen respondent in the San Francisco Bay area rents or is 65+ (or both):

$$P(\mathit{Rent} \cup 65{+}) = P(\mathit{Rent}) + P(65{+}) - P(\mathit{Rent} \cap 65{+}) \approx 0.42 + 0.25 - 0.04 = 0.63$$

If events A and B are mutually exclusive, then $P(A \cap B) = 0$ and the addition law simplifies to $P(A \cup B) = P(A) + P(B)$. See (Diez et al, 2015) section 2.1.2 for additional reading on this topic. Check your knowledge on the following questions:

- What is[14] $P(Own \cup 18 - 44)$?
- What is[15] $P(Own \cup \mathit{Rent})$?
- What is[16] $P(45 - 64 \cup \mathit{Rent})$?

As our second example, consider the survey data from 1,000 respondents shown in Table 6. *Legalize* denotes the event that a student favors the legalization of marijuana. *Share* denotes the event that a student shares the political views of her parents. The related joint probability table is shown in Table 7, obtained by dividing all entries in Table 6 by the number of respondents, 1,000.

[12] No, because $P(Own|65{+}) \approx 0.21/0.25 = 0.84 \neq 0.58 \approx P(Own)$. To establish that these two events are not independent, we could alternatively have shown that $P(65 + |Own) \neq P(65{+})$ or $P(65 + \cap Own) \neq P(65{+})P(Own)$.

[13] No, because $P(\mathit{Rent}|18{-}44) \approx 0.26/0.42 \approx 0.62 \neq 0.42 \approx P(\mathit{Rent})$. To establish that these two events are not independent, we could alternatively have shown that $P(18 - 44|\mathit{Rent}) \neq P(18 - 44)$ or $P(18 - 44 \cap \mathit{Rent}) \neq P(18 - 44)P(\mathit{Rent})$.

[14] $P(Own \cup 18 - 44) = P(Own) + P(18 - 44) - P(Own \cap 18 - 44) \approx 0.58 + 0.42 - 0.16 = 0.84.$

[15] $P(Own \cup \mathit{Rent}) = P(Own) + P(\mathit{Rent}) - P(Own \cap \mathit{Rent}) \approx 0.58 + 0.42 - 0 = 1.$

[16] $P(45 - 64 \cup \mathit{Rent}) = P(45 - 64) + P(\mathit{Rent}) - P(45 - 64 \cap \mathit{Rent}) \approx 0.33 + 0.42 - 0.12 = 0.63.$

	Parents' position		
Marijuana	$Share$	$Share^c$	Total
$Legalize$	610	120	730
$Legalize^c$	70	200	270
Total	680	320	1,000

Table 6. *Marijuana survey data.*

	Parents' position		
Marijuana	$Share$	$Share^c$	Total
$Legalize$	0.61	0.12	0.73
$Legalize^c$	0.07	0.2	0.27
Total	0.68	0.32	1

Table 7. *Related joint probability table.*

Check the following probability calculations:

$$P(Share^c \cap Legalize^c) = 200/1,000 = 0.2$$

$$P(Share|Legalize^c) = \frac{P(Share \cap Legalize^c)}{P(Legalize^c)} = \frac{0.07}{0.27} \approx 0.26$$

$$P(Share \cup Legalize^c) = P(Share) + P(Legalize^c) - P(Share \cap Legalize^c) = 0.68 + 0.27 - 0.07 = 0.88$$

Check your knowledge on the following questions:

- What is[17] $P(Legalize)$?
- What is[18] $P(Share)$?
- What is[19] $P(Legalize^c)$?
- What is[20] $P(Share^c)$?
- What is[21] $P(Legalize \cap Share^c)$?
- What is[22] $P(Share \cap Legalize^c)$?
- What is[23] $P(Legalize|Share^c)$?
- What is[24] $P(Share^c|Legalize)$?
- What is[25] $P(Share^c \cup Legalize)$?
- What is[26] $P(Share \cup Legalize)$?

[17] $P(Legalize) = 730/1000 = 0.73$

[18] $P(Share) = 680/1000 = 0.68$

[19] $P(Legalize^c) = 0.27$

[20] $P(Share^c) = 0.32$

[21] $P(Legalize \cap Share^c) = 120/1,000 = 0.12$

[22] $P(Share \cap Legalize^c) = 70/1,000 = 0.07$

[23] $P(Legalize|Share^c) = P(Legalize \cap Share^c)/P(Share^c) = 0.12/0.32 = 0.375$

[24] $P(Share^c|Legalize) = (P(Share^c \cap Legalize))/P(Legalize) = 0.12/0.73 = 0.16$

[25] $P(Share^c \cup Legalize) = P(Share^c) + P(Legalize) - P(Share^c \cap Legalize) = 0.32 + 0.73 - 0.12 = 0.93$

[26] $P(Share \cup Legalize) = P(Share) + P(Legalize) - P(Share \cap Legalize) = 0.68 + 0.73 - 0.61 = 0.8$

For our third example, consider the data from 115 market evaluations by prominent marketing analysts shown in Table 8. *Pos* denotes the event that an evaluation was favorable. *Low*, *Medium*, and *High* denote the events that the market demand is, respectively, low, medium, and high. The related joint probability table is shown in Table 9, obtained by dividing all entries in Table 8 by the number of evaluations, 115.

	Market Condition			
Eval	Low	Medium	High	Total
Pos	0	12	70	82
Pos^c	20	11	2	33
Total	20	23	72	115

Table 8. *Market survey data.*

	Market Condition			
Eval	Low	Medium	High	Total
Pos	0	0.10	0.61	0.71
Pos^c	0.17	0.10	0.02	0.29
Total	0.17	0.20	0.63	1

Table 9. *Related joint probability table.*

We can compute the following probabilities:

$$P(Pos^c \cap Medium) = 11/115 \approx 0.10$$

$$P(Pos|Medium) = \frac{P(Pos \cap Medium)}{P(Medium)} \approx \frac{0.10}{0.20} \approx 0.5$$

$$P(Pos \cup Medium) = P(Pos) + P(Medium) - P(Pos \cap Medium) \approx 0.71 + 0.20 - 0.10 = 0.81$$

Check your knowledge on the following questions:

- What is[27] $P(High)$?
- What is[28] $P(Pos^c)$?
- What is[29] $P(High \cap Pos^c)$?
- What is[30] $P(High|Pos^c)$?
- What is[31] $P(Pos^c|High)$?
- What is[32] $P(Pos^c \cup High)$?
- What is[33] $P(Pos \cup Low)$?

[27] $P(High) = 72/115 = 0.63$

[28] $P(Pos^c) = 0.29$

[29] $P(High \cap Pos^c) = 2/115 \approx 0.02$

[30] $P(High|Pos^c) = P(High \cap Pos^c)/P(Pos^c)) \approx 0.02/0.29 \approx 0.07$

[31] $P(Pos^c|High) = P(Pos^c \cap High)/P(High) \approx 0.02/0.63 \approx 0.03$

[32] $P(Pos^c \cup High) = P(Pos^c) + P(High) - P(Pos^c \cap High) \approx 0.29 + 0.63 - 0.02 = 0.9$

[33] $P(Pos \cup Low) = P(Pos) + P(Low) - P(Pos \cap Low) \approx 0.71 + 0.17 - 0 = 0.88$

> A scientist worthy of a lab coat should be able to make original discoveries while wearing a clown suit, or give a lecture in a high squeaky voice from inhaling helium. It is written nowhere in the math of probability theory that one may have no fun.
>
> Eliezer Yudkowsky

Random Variables

A *random variable* assigns numerical representations to experimental outcomes. For example, instead of enumerating the outcomes of a coin-tossing experiment as $\{Head, Tail\}$, we could employ a numerical representation such as $\{1, 0\}$. The numerical labeling of events enables the calculation of statistics. See (Diez et al, 2015) section 2.4 for additional reading.

The random variable described above for the coin-tossing experiment is an example of a *discrete random variable*. Discrete random variables take on at most a countable number of values. Continuous random variables take on an uncountable number of values.

A *probability distribution* describes the values that a random variable can assume and their associated probabilities. A probability distribution is therefore a model of the likelihood of occurrence of outcomes of the underlying experiment.

Discrete Distributions

In the case of a discrete random variable the probability distribution $P(X = x)$ is called a probability mass function (pmf). The rules for the pmf of a discrete random variable are already familiar:

1. $0 \leq P(x) \leq 1$ for all x, and

2. $\sum_{all\ x} P(x) = 1$

An associated function, the cumulative distribution function (cdf), is

$$F(x) = P(X \leq x) = \sum_{i=-\infty}^{x} P(X = i)$$

For example, the pmf $P(X = x)$ and cdf $P(X \leq x)$ of a coin toss with a fair coin is shown in Table 10 and the pmf and cdf of a die toss with a fair die is shown in Table 11. As a third example, Table 12 shows the pmf and cdf of revenue X associated with four different business scenarios.

	pmf	cdf
x	$P(X = x)$	$P(X \leq x)$
0 (Head)	0.5	0.5
1 (Tail)	0.5	1

Table 10. *pmf and cdf of a fair coin toss.*

	pmf	cdf
x	$P(X = x)$	$P(X \leq x)$
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	1

Table 11. *pmf and cdf of a fair die roll.*

Expected Value. The *expected value*, denoted by $E(X)$ or μ, of a discrete random variable X, is the average over a large number of outcomes (of random variable X). It is calculated by the sum of each experimental outcomes weighted by the probability of their occurrence. That is:

$$E(X) = \mu = x_1 P(x_1) + x_2 P(x_2) + \ldots + x_n P(x_n)$$

See (Diez et al, 2015) section 2.4.1 for additional reading on this topic.

The expected value of a coin toss (using the random variable assignment shown in Table 7) is:

$$E(X) = \mu = 0(0.5) + 1(0.5) = 0.5$$

The expected value of a large number of die rolls is (see Table 8 for the associated pmf):

$$E(X) = 1.\frac{1}{6} + 2.\frac{1}{6} + 3.\frac{1}{6} + 4.\frac{1}{6} + 5.\frac{1}{6} + 6.\frac{1}{6} = 3.5$$

Finally, the expected value of the distribution shown in Table 9 is

$$E(X) = 1(0.2) + 2(0.5) + 5(0.1) + 10(0.2) = 3.7$$

Notice that the expected value is not, in all three examples, an experimental outcome (i.e., not a value that X can assume).

Let us revisit the issue of interpretation of $E(X)$ using the die-rolling example: $E(X) = \mu$ is

- Estimated by the simple (arithmetic) average value of X over a large number of trials of the underlying random experiment. See Table 13 for outcomes of repeated rolls of a fair die (the data) and associated interpretation of $E(X)$.

	pmf	cdf
Revenue (x)	$P(X = x)$	$P(X \leq x)$
1	0.2	0.2
2	0.5	0.7
5	0.1	0.8
10	0.2	1

Table 12. *pmf and cdf of business scenario revenues.*

- The *weighted* average of the possible values of X where the weights are the probabilities of different values of X in the probability distribution *model* of X. See Table 14 for the probability model of a roll of a fair die and associated interpretation of $E(X)$.

The following is an important insight. The population mean, μ, is the *simple (arithmetic)* average of the population values. A sample mean, $\bar{x}$, is the *simple arithmetic* average of the data values, and is an *estimate* of μ. See, for example, Table 13. $E(X)$ is the *weighted* average of the different values that X can assume in the probability model (weighted by the probabilities). See, for example, Table 14. The data and probability model are connected because $\mu = E(X)$.

Trial	Outcome (x)
1	5
2	6
3	1
4	1
⋮	⋮

Table 13. *Data. μ is (or is estimated by)* $\sum_{i=1}^{n} x_i/n = \frac{5+6+1+1+\ldots}{n}$.

	pmf
x	$P(X = x)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Table 14. *Probability model (distribution).* $E(X) = \sum_x P(X = x)x = 1.\frac{1}{6} + 2.\frac{1}{6} + 3.\frac{1}{6} + 4.\frac{1}{6} + 5.\frac{1}{6} + 6.\frac{1}{6} = 3.5.$

Variance. Let V(X) denote the variance over a large number of outcomes $x_1, x_2, \ldots, x_n$ of random variable X. It is calculated as follows:

$$V(X) = (x_1 - \mu)^2 P(x_1) + (x_2 - \mu)^2 P(x_2) + \ldots + (x_n - \mu)^2 P(x_n)$$

See (Diez et al, 2015) section 2.4.2 for additional reading on this topic.

For example, the variance of a fair coin toss (see Table 7) is:

$$V(X) = (0 - 0.5)^2(0.5) + (1 - 0.5)^2(0.5) = 0.25$$

The variance of a fair die roll (see Table 8) is:

$$V(X) = (1-3.5)^2\frac{1}{6}+(2-3.5)^2\frac{1}{6}+(3-3.5)^2\frac{1}{6}+(4-3.5)^2\frac{1}{6}+(5-3.5)^2\frac{1}{6}+(6-3.5)^2\frac{1}{6} \approx 2.92$$

The variance of our final example, the pmf shown in Table 9, is:

$$V(X) = (1-3.7)^2(0.2) + (2-3.7)^2(0.5) + (5-3.7)^2(0.1) + (10-3.7)^2(0.2) \approx 11.01$$

Check your knowledge on the following questions using the pmf in Table 15:

- What is[34] $P(X \leq 30)$?
- What is[35] $E(X)$?
- What is[36] $V(X)$?

x	pmf $P(X = x)$
10	0.1
20	0.11
30	0.12
50	0.67

Table 15. *pmf.*

Continuous Distributions

In continuous distributions the counterpart of the probability mass function (of discrete distributions) is the so-called probability density function (pdf), which we will denote by $f(x)$. The rules for a probability density function are:

1. $f(x) \geq 0$ for all x, and
2. $\int_{-\infty}^{+\infty} f(x)dx = 1$

The mathematical notation $\int_{-\infty}^{+\infty} f(x)dx$ is the integral of the function $f(x)$, and symbolizes, for our purposes, the "area" under the function $f(x)$. The cumulative distribution function (cdf), denoted by $F(x) = P(X \leq x)$, is, for our purposes, the "area" under f on the domain $[-\infty, x]$, i.e., $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$. It does not make sense to talk about $P(X = x)$ in a continuous distribution – because the

[34] 0.33
[35] 40.3
[36] 212.91

"area" under $f(x)$ at a single point x is zero. Instead, we compute the probability that the random variable will assume a value in an interval of interest, e.g.,

$$P(x_1 \leq X \leq x_2)$$

This can be computed as follows:

$$P(x_1 \leq X \leq x_2) = P(X \leq x_2) - P(X \leq x_1) = F(x_2) - F(x_1)$$

See Figure 1. See section 2.5 for additional reading on this topic.

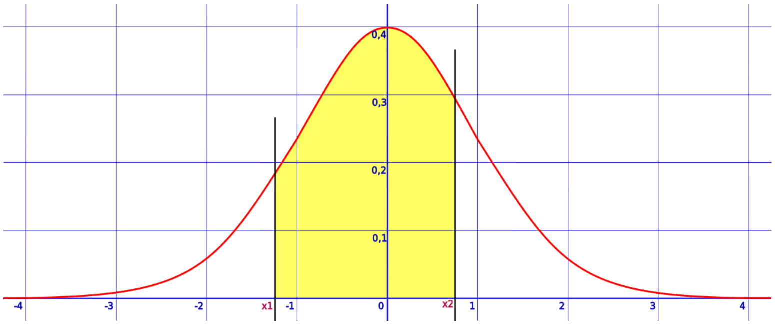

Figure 1. $P(x_1 \leq X \leq x_2)$.

We now consider a special case of a continuous distribution, namely the normal distribution. The normal distribution has direct application in statistical inference.

Normal Distribution. The *normal distribution* is commonly (but informally) referred to as the "bell curve" distribution. It has two parameters: μ and σ. μ is the mean and σ is the standard deviation of the distribution. The random variable Z is commonly used to denote the standard normal random variable. The standard normal distribution is a special case of the normal distribution with a mean of $\mu = 0$ and standard deviation of $\sigma = 1$. The pdf of the normal distribution is

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

The pdf of the standard normal distribution is shown in Figure 2. The related cdf is shown in Figure 3.

The pdf of the normal distribution is difficult to work with algebraically. The standard workaround for this issue in the past involved a transformation

$$z = \frac{x - \mu}{\sigma}$$

of a normal distribution problem to an equivalent problem involving the standard normal distribution and the use of tables (see Appendix B in (Diez et al, 2015)) to find associated cumulative probabilities or z values. See section 3.1.2 – 3.1.4 of (Diez et al 2015) for more details.

Microsoft Excel offers formulas that allow us to avoid the use of tables:

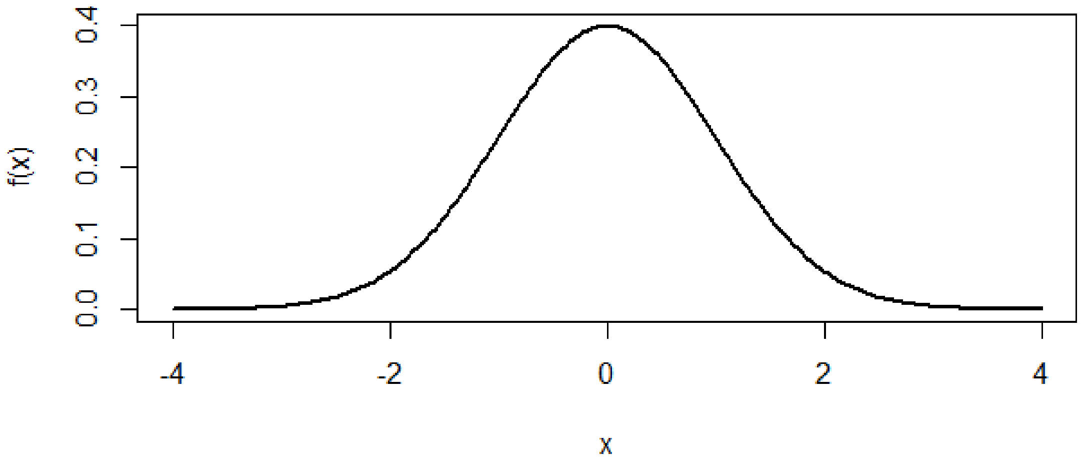

Figure 2. Pdf of normal distribution with $\mu = 0$ and $\sigma = 1$.

- In Excel, $NORM.DIST(x, \mu, \sigma, 1)$ returns the probability $F(x) = P(X \leq x)$ if X is normally distributed with mean μ and standard deviation σ. Notice that $NORM.DIST(x, \mu, \sigma, 1)$ returns the *left* (lower) tail probability of the distribution.

- In Excel, $NORM.INV(p, \mu, \sigma)$ returns the value of x for which $P(X \leq x) = p$ if X is normally distributed with mean μ and standard deviation σ. Notice that p is a *left* (upper) tail probability.

Consider the following example. Suppose a product is advertised as containing 16 fluid ounces of beer. The process used to manufacture the product produces a variable quantity of beer in the product. The beer in each unit is normally distributed with a mean of 16.05 fluid ounces and a standard deviation of 0.2 fluid ounces. What is the probability that a unit is under filled (i.e., contains less than 16 fluid ounces)? The answer is $NORM.DIST(16, 16.05, 0.2, 1) \approx 0.40$.

Let's flip this question and ask what the advertised product quantity should be for a process that has a mean and standard deviation of 16.05 fluid ounces and 0.2 fluid ounces, respectively, if the required maximum underfill probability is 1%. The answer is $NORM.INV(0.01, 16.05, 0.2) \approx 15.58$.

Consider the following new example. Suppose that bank accounts for which complaints have been received in the last quarter have a mean value of \$12,784 and a standard deviation of \$3,100. What fraction of accounts for which complaints were received have a value in excess of \$10,000? The answer is $1 - NORM.DIST(10000, 12784, 3100, 1) \approx 0.82$. Here we need $1 - NORM.DIST(\ldots)$ in order to compute the *right* tail probability $F(X \geq 10000) = 1 - F(X \leq 10000)$.

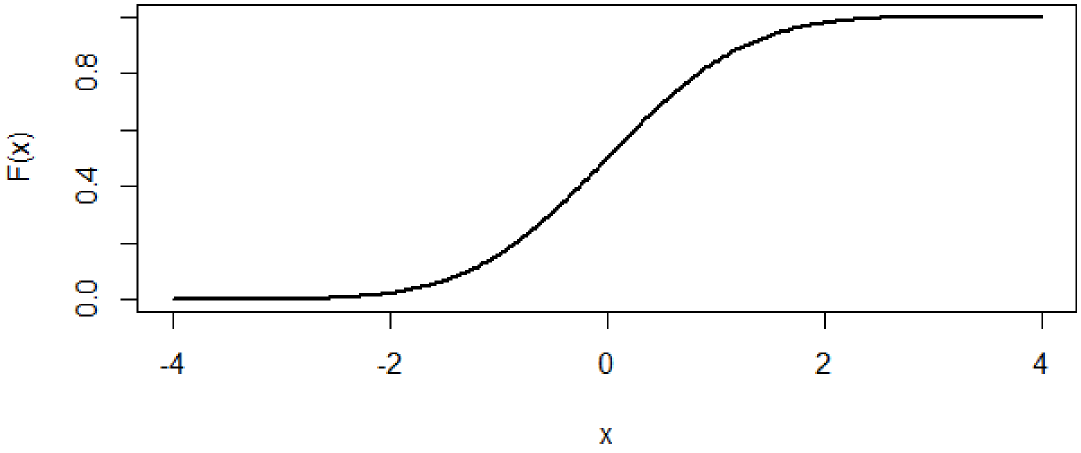

Figure 3. Cdf of normal distribution with $\mu = 0$ and $\sigma = 1$.

Suppose that individual human intelligence is scored on a scale with a mean of 100 and a standard deviation of 15. What is $P(85 \leq IQ \leq 115)$? The answer is $P(85 \leq IQ \leq 115) = P(IQ \leq 115) - P(IQ \leq 85) = NORM.DIST(115, 100, 15, 1) - NORM.DIST(85, 100, 15, 1) \approx 0.683$.

A few special-case computations using the standard normal distribution are worth recording for future use. Let z_α denote the value of Z such that $P(Z \geq z_\alpha) = \alpha$. Then we can calculate z_α (using $NORM.S.INV(1 - \alpha)$) in Excel for various values of α shown in Table 16.

α	z_α
.01	2.33
.025	1.96
.05	1.645

Table 16. z_α *for selected values of* α.

Check your knowledge on the following questions:

- Suppose that X is normal with a mean of 100 and a standard deviation of 15. What is[37] $P(X \leq 130)$?

- Suppose that X is normal with a mean of 100 and a standard deviation of 15.

[37] $NORM.DIST(130, 100, 15, 1) \approx 0.98$. Here we use $NORM.DIST(\ldots)$ to calculate a *left* (lower) tail probability.

What is[38] the probability of observing values of the random variable in excess of six standard deviations from the mean?

- Suppose that X is normal with a mean of 100 and a standard deviation of 15. What is[39] the probability of observing values of the random variable within one standard deviation of the mean?
- Suppose that X is normal with a mean of 100 and a standard deviation of 15. Below what value[40] x do 25% of the observations of X lie?
- Suppose that X is normal with a mean of 100 and a standard deviation of 15. Above what value[41] x do 10% of the observations of X lie?

Linear Combinations of Random Variables

A bivariate distribution is the joint distribution of two variables. The probability rules for a bivariate distribution, $P(x, y) = P(X = x, Y = y) = P(X = x \cap Y = y)$, are:

1. $0 \leq P(x, y) \leq 1$ for all x and y, and
2. $\sum_{all\ x} \sum_{all\ y} P(x, y) = 1$.

The marginal probabilities can be obtained by summing over values of the opposing variable, e.g. $P(x) = \sum_{all\ x} P(x, y)$. The *covariance* of two discrete random variables X and Y is:

$$COV(X,Y) = \sigma_{xy} = \sum_{all\ x} \sum_{all\ y} (x - \mu_x)(y - \mu_y)P(x, y) = \sum_{all\ x} \sum_{all\ y} xyP(x, y) - \mu_x\mu_y$$

where $\mu_x = E(X)$ and $\mu_y = E(Y)$. The *correlation coefficient* is

$$\rho = \frac{\sigma_{xy}}{\sigma_y\sigma_y}$$

The linear combination of two random variables X and Y is $aX + bY$ where a and b are constants. The expected value of a linear combination of two random variables is:

$$E(aX + bY) = aE(X) + bE(Y)$$

[38] $1 - NORM.DIST(190, 100, 15, 1) = 9.86588 \times 10^{-10}$. Note that we use $1 - NORM.DIST(\ldots)$ to calculate a *right* (upper) tail probability.

[39] $NORM.DIST(115, 100, 15, 1) - NORM.DIST(85, 100, 15, 1) \approx 0.68$. Here we use the idea that $P(\mu - \sigma \leq X \leq \mu + \sigma) = P(X \leq \mu + \sigma) - P(X \leq \mu - \sigma) = F(\mu + \sigma) - F(\mu - \sigma) = F(115) - F(85)$.

[40] $NORM.INV(0.25, 100, 15) = 89.88$

[41] $NORM.INV(0.9, 100, 15) \approx 119.22$. Note that we use $p = 0.9$ here because the problem deals with a *right* (upper) tail of 10%.

The variance of a linear combination of two random variables is:

$$V(aX + bY) = a^2V(X) + b^2V(Y) + 2abCov(X, Y)$$

If X and Y are independent then $Cov(X, Y) = 0$ and we get

$$V(aX + bY) = a^2V(X) + b^2V(Y)$$

More generally, suppose we have n random variables $X_1, \ldots, X_n$. The expectation of $a_1X_1 + a_2X_2 + \ldots + a_nX_n$ is

$$E(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_1E(X_1) + a_2E(X_2) + \ldots + a_nE(X_n)$$

If these random variables are independent then the covariance terms are equal to zero and we get

$$V(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_1^2V(X_1) + a_2^2V(X_2) + \ldots + a_n^2V(X_n)$$

See (Diez et al, 2015) section 2.4.3 for additional reading on this topic.

For example, suppose that the daily movement of an equity index has an expected value of $E(X) = 5$ points. Then the expected movement of a perfectly-tracking 3X leveraged exchange-traded fund has an expected return that is three times that of the equity index, or $E(3X) = 3E(X) = 3(5) = 15$ points.

Consider the following new example. Suppose that the daily movement of an equity index has a variance of $V(X) = 10$ points. Then the daily variance of a perfectly-tracking 3X ETF based upon the equity index is $V(3X) = 3^2V(X) = 9(10) = 90$ points.

Consider the following more elaborate example. Suppose that a portfolio is composed of two securities with expected returns and variance of returns as shown in Table 17. Suppose that the covariance of the returns of the two securities is 0.

Security	Percentage of Portfolio Value	Expected Return	Variance of Return
1	40%	.1	.03
2	60%	.16	.09

Table 17. *Portfolio composition and return information.*

The expected return of the portfolio is:

$$E(Portfolio\ Return) = E(.4Return_1 + .6Return_2) =$$

$$.4E(Return_1) + .6E(Return_2) = (.4 \times .1) + (.6 \times .16) = .136$$

and

$$V(Portfolio\ Return) = V(.4Return_1 + .6Return_2) =$$

$$.4^2 V(Return_1) + .6^2 V(Return_2) = (.16 \times .03) + (.36 \times .09) = .0372$$

Check your knowledge on the following questions. Suppose that X and Y are independent random variables.

- Suppose $E(X) = 10$, $E(Y) = 2$. What is[42] $E(5X - 2Y)$?
- Suppose $E(X) = 100$, $E(Y) = 100$. What is[43] $E(X - Y)$?
- Suppose $V(X) = 100, V(Y) = 100$. What is[44] $V(X - Y)$?
- Suppose $V(X) = 100, V(Y) = 100$. What is[45] $V(2X - 6Y)$?

[42] $E(5X - 2Y) = 5(10) - 2(2) = 46$

[43] $E(X - Y) = E(X) - E(Y) = 100 - 100 = 0$

[44] $V(X - Y) = V(X) + V(Y) = 100 + 100 = 200$

[45] $V(2X - 6Y) = 4V(X) + V(Y) = 4(100) + 36(100) = 4,000$