Essentials of Statistics

FIFTH EDITION

Mario F. Triola

ALWAYS LEARNING

PEARSON

Symbol Table

$\overline{A}$	complement of event A
H_0	null hypothesis
H_1	alternative hypothesis
α	alpha; probability of a type I error or the area of the critical region
β	beta; probability of a type II error
r	sample linear correlation coefficient
ρ	rho; population linear correlation coefficient
r^2	coefficient of determination
r_s	Spearman's rank correlation coefficient
b_1	point estimate of the slope of the regression line
b_0	point estimate of the y-intercept of the regression line
$\hat{y}$	predicted value of y
d	difference between two matched values
$\overline{d}$	mean of the differences d found from matched sample data
s_d	standard deviation of the differences d found from matched sample data
s_e	standard error of estimate
$\mu_{\overline{x}}$	mean of the population of all possible sample means $\overline{x}$
$\sigma_{\overline{x}}$	standard deviation of the population of all possible sample means $\overline{x}$
E	margin of error of the estimate of a population parameter, or expected value
Q_1, Q_2, Q_3	quartiles
$D_1, D_2, \ldots, D_9$	deciles
$P_1, P_2, \ldots, P_{99}$	percentiles
x	data value
f	frequency with which a value occurs
Σ	capital sigma; summation
Σx	sum of the values
Σx^2	sum of the squares of the values
$(\Sigma x)^2$	square of the sum of all values
Σxy	sum of the products of each x value multiplied by the corresponding y value
n	number of values in a sample
$n!$	n factorial
N	number of values in a finite population; also used as the size of all samples combined
k	number of samples or populations or categories
$\overline{x}$	mean of the values in a sample
μ	mu; mean of all values in a population
s	standard deviation of a set of sample values
σ	lowercase sigma; standard deviation of all values in a population
s^2	variance of a set of sample values
σ^2	variance of all values in a population
z	standard score
$z_{\alpha/2}$	critical value of z
t	t distribution
$t_{\alpha/2}$	critical value of t
df	number of degrees of freedom
F	F distribution
χ^2	chi-square distribution
χ^2_R	right-tailed critical value of chi-square
χ^2_L	left-tailed critical value of chi-square
p	probability of an event or the population proportion
q	probability or proportion equal to $1 - p$
$\hat{p}$	sample proportion
$\hat{q}$	sample proportion equal to $1 - \hat{p}$
$\overline{p}$	proportion obtained by pooling two samples
$\overline{q}$	proportion or probability equal to $1 - \overline{p}$
$P(A)$	probability of event A
$P(A \mid B)$	probability of event A, assuming event B has occurred
${}_nP_r$	number of permutations of n items selected r at a time
${}_nC_r$	number of combinations of n items selected r at a time

important parameters of mean, standard deviation, and variance for a probability distribution. Most importantly, we describe how to determine whether outcomes are *unlikely* to occur by chance. We begin with the related concepts of *random variable* and *probability distribution.*

Part 1: Basic Concepts of a Probability Distribution

DEFINITIONS

A **random variable** is a variable (typically represented by x) that has a single numerical value, determined by chance, for each outcome of a procedure.

A **probability distribution** is a description that gives the probability for each value of the random variable. It is often expressed in the format of a table, formula, or graph.

In Section 1-3 we made a distinction between discrete and continuous data. Random variables may also be discrete or continuous, and the following two definitions are consistent with those given in Section 1-3.

DEFINITIONS

A **discrete random variable** has a collection of values that is finite or countable. (If there are infinitely many values, the number of values is countable if it is possible to count them individually, such as the number of tosses of a coin before getting tails.)

A **continuous random variable** has infinitely many values, and the collection of values is not countable. (That is, it is impossible to count the individual items because at least some of them are on a continuous scale.)

This chapter deals exclusively with discrete random variables, but the subsequent chapters will deal with continuous random variables.

Every probability distribution must satisfy each of the following three requirements.

Probability Distribution: Requirements

1. There is a numerical random variable x and its values are associated with corresponding probabilities.
2. $\Sigma P(x) = 1$ where x assumes all possible values. (The sum of all probabilities must be 1, but sums such as 0.999 or 1.001 are acceptable because they result from rounding errors.)
3. $0 \leq P(x) \leq 1$ for every individual value of the random variable x. (That is, each probability value must be between 0 and 1 inclusive.)

The second requirement comes from the simple fact that the random variable x represents all possible events in the entire sample space, so we are certain (with probability 1) that one of the events will occur. The third requirement comes from the basic principle that any probability value must be 0 or 1 or a value between 0 and 1.

Example 1 **Genetics**

Although the Chapter Problem involves 945 births, let's consider a simpler example that involves only two births with the following random variable:

$$x = \text{number of girls in two births}$$

The above x is a random variable because its numerical values depend on chance. With two births, the number of girls can be 0, 1, or 2, and Table 5-1 is a probability distribution because it gives the probability for each value of the random variable x and it satisfies the three requirements listed earlier:

1. The variable x is a numerical random variable and its values are associated with probabilities, as in Table 5-1.
2. $\Sigma P(x) = 0.25 + 0.50 + 0.25 = 1$
3. Each value of $P(x)$ is between 0 and 1. (Specifically, 0.25 and 0.50 and 0.25 are each between 0 and 1 inclusive.)

The random variable x in Table 5-1 is a *discrete* random variable, because it has three possible values (0, 1, 2), and 3 is a finite number, so this satisfies the requirement of being finite or countable.

Table 5-1 Probability Distribution for the Number of Girls in Two Births

Number of Girls x	$P(x)$
0	0.25
1	0.50
2	0.25

Notation

In tables such as Table 5-1 or the Binomial Probabilities in Table A-1 in Appendix A, we sometimes use 0+ to represent a probability value that is positive but very small, such as 0.000000123. (When rounding a probability value for inclusion in such a table, a rounded value of 0 would be misleading because it incorrectly suggests that the event is impossible.)

Probability Distribution: Graph

There are various ways to graph a probability distribution, but we will consider only the **probability histogram.** Figure 5-3 is a probability histogram corresponding to Table 5-1. Notice that it is similar to a relative frequency histogram (described in Section 2-3), but the vertical scale shows *probabilities* instead of relative frequencies based on actual sample results.

In Figure 5-3, we see that the values of 0, 1, 2 along the horizontal axis are located at the centers of the rectangles. This implies that the rectangles are each 1 unit wide, so the areas of the rectangles are 0.25, 0.50, and 0.25. The *areas*

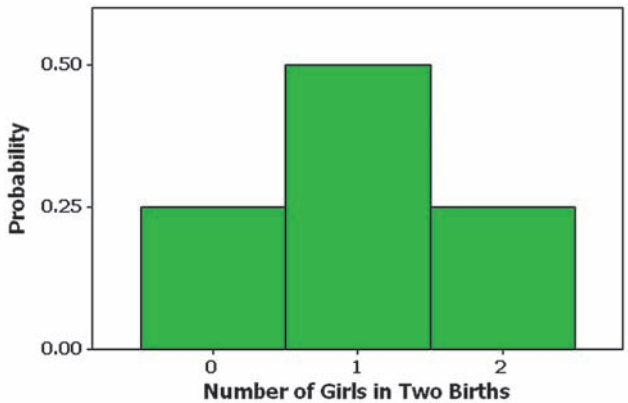

Figure 5-3 **Probability Histogram for Number of Girls in Two Births**

of these rectangles are the same as the *probabilities* in Table 5-1. We will see in Chapter 6 and future chapters that such a correspondence between area and probability is very useful.

Example 2 **Marijuana Survey**

In a Pew Research Center poll, subjects were asked if the use of marijuana should be made legal, and the results from that poll have been used to create Table 5-2. Does Table 5-2 describe a probability distribution?

Table 5-2 Should Marijuana Use Be Legal?

Response	$P(x)$
Yes	0.41
No	0.52
Don't know	0.07

Solution

Consider the three requirements listed earlier.

1. The responses (yes, no, don't know) are not numerical, so we do not have a numerical random variable. The first requirement is violated, so Table 5-2 does not describe a probability distribution.
2. The sum of the probabilities is 1.
3. Each probability is between 0 and 1 inclusive.

Because one of the three requirements is not met, Table 5-2 does not describe a probability distribution.

Example 3 **When to Discuss Salary**

Senior executives were asked when job applicants should discuss salary, and Table 5-3 is based on their responses (based on data from an Accountemps survey). Does Table 5-3 describe a probability distribution?

Table 5-3 When to Discuss Salary

Number of Interviews x	$P(x)$
1	0.30
2	0.26
3	0.10

Solution

To be a probability distribution, we must have a numerical random variable x such that $P(x)$ must satisfy the preceding three requirements.

1. The variable x is a numerical random variable and its values are associated with probabilities, as in Table 5-3.
2. $\Sigma P(x) = P(1) + P(2) + P(3)$
   $= 0.30 + 0.26 + 0.10$
   $= 0.66$ [showing that $\Sigma P(x) \neq 1$]
3. Each value of $P(x)$ is between 0 and 1 inclusive.

Because the second requirement is not satisfied, we conclude that Table 5-3 does *not* describe a probability distribution.

Example 4

Does $P(x) = \frac{x}{3}$ (where x can be 0, 1, or 2) determine a probability distribution?

Solution

To be a probability distribution, we must have a numerical random variable x such that $P(x)$ must satisfy the preceding three requirements. From the given formula we find that $P(0) = 0/3$ and $P(1) = 1/3$ and $P(2) = 2/3$.

continued

Life Data Analysis

Life data analysis deals with the longevity and failure rates of manufactured products. In one application, it is known that Dell computers have an "infant mortality" rate, whereby the failure rate is highest immediately after the computers are produced. Dell therefore tests or "burns-in" the computers before they are shipped. Dell can optimize profits by using an optimal burn-in time that identifies failures without wasting valuable testing time. Other products, such as cars, have failure rates that increase over time as parts wear out. If General Motors or Dell or any other company were to ignore the use of statistics and life data analysis, it would run the serious risk of going out of business because of factors such as excessive warranty repair costs or the loss of customers who experience unacceptable failure rates.

The *Weibull distribution* is a probability distribution commonly used in life data analysis applications. That distribution is beyond the scope of this book.

1. The variable x is a numerical random variable and its values (0, 1, 2) are associated with probabilities, as determined by the given formula.
2. $\Sigma P(x) = P(0) + P(1) + P(2) = \frac{0}{3} + \frac{1}{3} + \frac{2}{3} = 1$
3. Each value of $P(x)$ is between 0 and 1 inclusive.

Because the three requirements are satisfied, we conclude that the given formula does describe a probability distribution.

Parameters of a Probability Distribution: Mean, Variance, and Standard Deviation

Remember that with a probability distribution, we have a description of a *population* instead of a sample, so the values of the mean, standard deviation, and variance are *parameters* instead of statistics. The mean is the central or "average" value of the random variable. The variance and standard deviation measure the variation of the random variable. These parameters can be found with the following formulas:

Formula 5-1 $\mu = \Sigma [x \cdot P(x)]$

Mean for a probability distribution

Formula 5-2 $\sigma^2 = \Sigma [(x - \mu)^2 \cdot P(x)]$

Variance for a probability distribution (This format is easier to understand.)

Formula 5-3 $\sigma^2 = \Sigma [x^2 \cdot P(x)] - \mu^2$

Variance for a probability distribution (This format is easier for manual computations.)

Formula 5-4 $\sigma = \sqrt{\Sigma [x^2 \cdot P(x)] - \mu^2}$

Standard deviation for a probability distribution

When applying Formulas 5-1 through 5-4, use the following rule for rounding results.

Round-off Rule for μ, σ, and σ^2 from a Probability Distribution

Round results by carrying one more decimal place than the number of decimal places used for the random variable x. If the values of x are integers, round μ, σ, and σ^2 to one decimal place.

It is sometimes necessary to use a different rounding rule because of special circumstances, such as results that require more decimal places to be meaningful. For example, with four-engine jets the mean number of jet engines working successfully throughout a flight is 3.999714286, which becomes 4.0 when rounded to one more decimal place than the original data. Here, 4.0 would be misleading because

it suggests that all jet engines always work successfully. We need more precision to correctly reflect the true mean, such as the precision in the number 3.999714.

Expected Value

The mean of a discrete random variable x is the theoretical mean outcome for infinitely many trials. We can think of that mean as the *expected value* in the sense that it is the average value that we would expect to get if the trials could continue indefinitely. The uses of expected value (also called *expectation,* or *mathematical expectation*) are extensive and varied, and they play an important role in *decision theory*. (See Example 8 in Part 2 of this section.)

DEFINITION The **expected value** of a discrete random variable x is denoted by E, and it is the mean value of the outcomes, so $E = \mu$ and E can also be found by evaluating $\Sigma[x \cdot P(x)]$, as in Formula 5-1.

CAUTION An expected value need not be a whole number, even if the different possible values of x might all be whole numbers. We say that the expected number of girls in five births is 2.5, even though five specific births can never result in 2.5 girls. If we were to survey many couples with five children, we *expect* that the mean number of girls will be 2.5.

Example 5 Finding the Mean, Variance, and Standard Deviation

Table 5-1 describes the probability distribution for the number of girls in two births. Find the mean, variance, and standard deviation for the probability distribution described in Table 5-1 from Example 1.

Solution

In Table 5-4, the two columns at the left describe the probability distribution given earlier in Table 5-1; we create the three columns at the right for the purposes of the calculations required.

Using Formulas 5-1 and 5-2 and the table results, we get

Mean: $\mu = \Sigma[x \cdot P(x)] = 1.0$

Variance: $\sigma^2 = \Sigma[(x - \mu)^2 \cdot P(x)] = 0.5$

The standard deviation is the square root of the variance, so

Standard deviation: $\sigma = \sqrt{0.5} = 0.707107 = 0.7$ (rounded)

Table 5-4 Calculating μ and σ for a Probability Distribution

x	$P(x)$	$x \cdot P(x)$	$(x - \mu)^2 \cdot P(x)$
0	0.25	$0 \cdot 0.25 = 0.00$	$(0 - 1)^2 \cdot 0.25 = 0.25$
1	0.50	$1 \cdot 0.50 = 0.50$	$(1 - 1)^2 \cdot 0.50 = 0.00$
2	0.25	$2 \cdot 0.25 = 0.50$	$(2 - 1)^2 \cdot 0.25 = 0.25$
Total		1.00 ↑ $\mu = \Sigma[x \cdot P(x)]$	0.50 ↑ $\sigma^2 = \Sigma[(x - \mu)^2 \cdot P(x)]$

continued

Interpretation

The mean number of girls in two births is 1.0 girl, the variance is 0.50 "girl squared," and the standard deviation is 0.7 girl. Also, the expected value for the number of girls in two births is 1.0 girl, which is the same value as the mean. If we were to collect data on a large number of couples with two children, we expect to get a mean of 1.0 girl.

Making Sense of Results: Identifying Unusual Values

We present the following two different approaches for determining whether a value of a random variable x is unusually low or unusually high.

Identifying Unusual Results with the Range Rule of Thumb

The range rule of thumb (introduced in Section 3-3) may be helpful in interpreting the value of a standard deviation. According to the range rule of thumb, the vast majority of values should lie within 2 standard deviations of the mean, so we can consider a value to be unusual if it is more than 2 standard deviations away from the mean. (The use of 2 standard deviations is not an absolutely rigid value, and other values such as 3 could be used instead.) We can therefore identify "unusual" values by determining that they lie outside of these limits:

Range Rule of Thumb

$$\textbf{maximum usual value} = \mu + 2\sigma$$
$$\textbf{minimum usual value} = \mu - 2\sigma$$

CAUTION Know that the use of the number 2 in the range rule of thumb is somewhat arbitrary, and this rule is a guideline, not an absolutely rigid rule.

Example 6 **Identifying Unusual Results with the Range Rule of Thumb**

In Example 5 we found that for families with two children, the mean number of girls is 1.0 and the standard deviation is 0.7 girl. Use those results and the range rule of thumb to find the maximum and minimum usual values for the number of girls. Based on the results, if a couple has two children, is 2 girls an unusually high number of girls?

Solution

Using the range rule of thumb, we can find the maximum and minimum usual values as follows:

maximum usual value: $\mu + 2\sigma = 1.0 + 2(0.7) = 2.4$

minimum usual value: $\mu - 2\sigma = 1.0 - 2(0.7) = -0.4$