Columbia Business School

Introduction to Probability and Statistics

1. Average

Suppose the sales at a grocery store for 3 days are \$150, \$100, and \$80. What "typical" daily sales, if repeated 3 times, would give the same total sales? The typical value, denoted $\bar{x}$, satisfies $3\bar{x} = 150 + 100 + 80$. The typical, or average, value for daily sales is $\bar{x} = 330/3 = 110$. Three days of average sales would give the same total sales at the grocery store.

In general, the *average* of the n numbers $x_1, x_2, \ldots, x_n$, denoted $\bar{x}$, is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Notes: The terms average and mean are synonymous. The previous definition of $\bar{x}$ is an *arithmetic* average, as compared to the geometric average defined below. Also, $\bar{x}$ is a *sample* average, since it is derived from a sample of data. In the Excel spreadsheet program, an average of numbers can be calculated using the function =AVERAGE. For example, if the numbers 150, 100, and 80 are placed in a spreadsheet in cells A1, A2, and A3, then =AVERAGE(A1:A3) = 110.

2. Geometric Average

Suppose the returns from an investment for 3 months are 2%, 4%, and −1%. This means that a \$1 investment grew to \$1.02 after 1 month, then \$1.02 grew to $1.02 \times 1.04 = \$1.0608$ after two months, which then grew to $1.0608 \times 0.99 = 1.050192$ at the end of three months. The total return for the three month period is 5.0192% (*not* 5% = 2% + 4% − 1%). What "typical" monthly return, if repeated 3 times, would give the same total return? The typical value, denoted r, satisfies $(1 + r)^3 = (1.02)(1.04)(0.99) = 1.050192$. This gives $1 + r = (1.050192)^{1/3}$, or $r = 0.016458$. Three monthly returns of r would give the same total return on the investment. The quantity $1 + r$ is called the *geometric* average of 1.02, 1.04, and 0.99.

In general, the *geometric average* of the n numbers $x_1, x_2, \ldots, x_n$ is the n^{th}-root of their product, i.e.,

$$\sqrt[n]{\prod_{i=1}^{n} x_i}.$$

Notes: Although the geometric average is more appropriate for averaging series of returns, the simpler arithmetic average is often used instead. The differences between the two averages are often small. (Alternatively, the arithmetic average of the logarithms of the numbers can be used. This is equivalent to the geometric average.)

3. Sample Variance and Sample Standard Deviation

Variance and standard deviation are measures of how widely a sample of data points are scattered. Variance is the average of the squared deviations from the mean. Standard deviation is the square root of variance. More precisely, let the n data points be denoted $x_1, x_2, \ldots, x_n$. The *sample variance*, denoted S^2, is given by

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}.$$

The *sample standard deviation*, denoted S, is the square root of the sample variance, or

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

The following formula for sample variance is equivalent to the previous formula, but is computationally simpler.

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2}{n-1} - \frac{(\sum_{i=1}^{n} x_i)^2}{n(n-1)}.$$

The derivation of this result follows. It uses the definition $n\bar{x} = \sum_{i=1}^{n} x_i$.

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}(x_i^2 - 2\bar{x}x_i + \bar{x}^2) = \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2\\ &= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}.\end{aligned}$$

Notes: If the data is measured in units of dollars, then sample variance is in units of \$ × \$, which is not very intuitive. The units of sample standard deviation are the same as the original data, i.e., \$. The $n-1$ in the denominator of the expressions for S^2 and S is a minor technical point: using $n-1$ in place of n makes S^2 an unbiased estimator of the true variance of the random variable that generated the data. For any reasonable sample sizes, dividing by n or $n-1$ is indistinguishable. The sample variance, S^2, and the sample standard deviation, S, can be easily calculated in a spreadsheet using the functions =VAR and =STDEV, respectively.

Example. The computation of S^2 is illustrated next. Suppose we want to compute the sample variance of the numbers 2, 9, 10, 5, and 4. Here $n = 5$, $\sum_{i=1}^{n} x_i = 30$, and $\sum_{i=1}^{n} x_i^2 = 226$. So $\bar{x} = 6$ and $S^2 = 226/4 - (30)^2/(5)(4) = 56.5 - 45 = 11.5$. Also $S = 3.391$. When using the computational formula for S^2 it is easy to add and delete data points and recompute all value. The recomputation involves updating n, $\sum_{i=1}^{n} x_i$, and $\sum_{i=1}^{n} x_i^2$, and then recomputing. Using the original definition, $\bar{x}$ would have to be recomputed, as well as all of the terms $(x_i - \bar{x})^2$. The calculation can be done in a spreadsheet by placing the numbers 2, 9, 10, 5, and 4 in the cells A1 through A5. Then =VAR(A1:A5) = 11.5 and =STDEV(A1:A5) = 3.391.

Statistics deals with actual data: estimating parameters (e.g., means and standard deviations), computing confidence intervals, hypothesis testing, etc. *Probability* deals with the underlying theoretical models that could generate the data. Probability provides the theoretical foundation for statistics.

The preceding definitions of average, sample variance, and sample standard deviation deal with actual data. The following concepts are similar, but deal with theoretical probabilities rather than actual data.

4. Expected Value

Suppose an investment has a 30% probability of a 10% return, a 40% probability of a 20% return, and a 30% probability of a 25% return. The *expected return* is given by weighting the outcomes by their respective probabilities. In this case, the expected return is $0.3(0.10) + 0.4(0.20) + 0.3(0.25) = 0.185$.

In general, if a random variable X takes on the n values $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$, respectively, then the expected value of X, denoted $E(X)$ or μ_X, is

$$E(X) = \mu_X = \sum_{i=1}^{n} p_i x_i.$$

Notes: For a random variable, the terms expected value, mean, and average are synonymous. The expected value is a "true" or theoretical value. By contrast, the average of a sample of data, $\bar{x}$, is only an estimate of the true expected value, μ_X. Expected value is only defined for a random variable; it does not makes sense to have a sample expected value.

The Greek letter μ is often used to denote expected value because of the mnemonic *m*u – *m*ean.

The expected value in the example above can be calculated in a spreadsheet by placing the probabilities 0.3, 0.4, and 0.3 in the cells A1, A2, and A3 and the returns 0.1, 0.2, 0.25 in the cells B1, B2, and B3. Then the expected return can be calculated by the =SUMPRODUCT function, i.e., =SUMPRODUCT(A1:A3, B1:B3) = 0.185.

The word expected in expected value does not have a literal English interpretation. For example, suppose X is a random variable that takes on the values 0 and 1, each with probability 1/2. Then the expected value of X is $E(X) = 0.5(0) + 0.5(1) = 0.5$. While 0.5 is the expected value of X, it is not really expected to happen; in this case the outcome 0.5 will *never* happen since X only takes on the values 0 and 1.

To compare the formulas for expected value and sample average, suppose a series of 10 returns is generated from the preceding random variable. The observed returns are 10%, 20%, 10%, 20%, 25%, 25%, 10%, 25%, 20%, and 20%. The average return is $(.1 + .2 + .1 + .2 + .25 + .25 + .1 + .25 + .2 + .2)/10$. Regrouping gives $.3(.1) + .4(.2) + .3(.25) = 0.185$. Notice that this is the same formula as before. If the observed frequency of outcomes exactly matches the probabilities of occurrence, then the formulas for average and expected value are the same. If the observed frequencies of a sample differ from the true probabilities of the random variable, then the sample average will differ from the true expected value.

5. Variance and Standard Deviation

The variance of a random variable is a measure of the dispersion of the outcomes around the mean. It is the expected value of the squared deviations from the mean. More precisely, if X is a random variable that takes on the n values $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$, respectively, then the variance of X, denoted $Var(X)$ or σ_X^2, is

$$\mathrm{Var}(X) = \sigma_X^2 = E([X - E(X)]^2) = \sum_{i=1}^{n} p_i (x_i - \mu_X)^2.$$

The standard deviation of a random variable X, denoted σ_X, is simply the square root of its variance:

$$\sigma_X = \sqrt{\sum_{i=1}^{n} p_i(x_i - \mu_X)^2}.$$

Notes: If the n values of the random variable X occur with equal probability, $1/n$, then the formula for the variance of X is the same as the sample variance (with μ_X in place of $\bar{x}$ and with n in place of $n-1$ in the denominator).

The Greek letter σ is often used to denote standard deviation because of the mnemonic *s*igma – *s*tandard deviation.

The following formula for variance of a random variable is equivalent to the previous formula, but is computationally simpler.

$$\begin{aligned}\mathrm{Var}(X) &= \sum_{i=1}^{n} p_i x_i^2 - \mu_X^2 \\ &= E(X^2) - E(X)^2.\end{aligned}$$

The derivation of this result follows. It uses the fact that the probabilities must sum to one: $\sum_{i=1}^{n} p_i = 1$.

$$\begin{aligned}\sum_{i=1}^{n} p_i(x_i - \mu_X)^2 &= \sum_{i=1}^{n} p_i(x_i^2 - 2\mu_X x_i + \mu_X^2) = \sum_{i=1}^{n} p_i x_i^2 - 2\mu_X \sum_{i=1}^{n} p_i x_i + \mu_X^2 \sum_{i=1}^{n} p_i \\ &= \sum_{i=1}^{n} p_i x_i^2 - 2\mu_X^2 + \mu_X^2 = \sum_{i=1}^{n} p_i x_i^2 - \mu_X^2.\end{aligned}$$

6. Properties of Expected Value, Variance, and Std. Deviation

Suppose X is a random variable that represents the profit on a \$1 investment. Further, suppose that we have computed the expected profit and variance of the profit. The profit on a \$2 investment is also a random variable, and if the profit is proportional to the amount of the investment, then the random profit is $2X$. We might also be interested in the expected value and variance in the profit of the \$2 investment. Without doing any recalculations, the following properties give the answer.

Suppose X is a random variable and a and b are constants. Then the following properties hold:

$$\begin{aligned}E(aX + b) &= aE(X) + b \\ \mathrm{Var}(aX + b) &= a^2\mathrm{Var}(X) \\ \sigma(aX + b) &= a\sigma(X)\end{aligned}$$

Thus, if the expected profit on a \$1 investment is $E(X)$, then the expected profit on a \$2 investment is $E(2X) = 2E(X)$. The first line says that expected value is a *linear operator*; however, variance and standard deviations are *not* linear operators. The derivations are given next. As before, assume that X takes on the n values $x_1, x_2, \ldots, x_n$ with probabilities p_1, p_2, $\ldots, p_n$, respectively. The derivations use the fact that the probabilities sum to one.

$$\begin{aligned}
E(aX+b) &= \sum_{i=1}^{n} p_i(ax_i+b) = a\sum_{i=1}^{n} p_i x_i + b\sum_{i=1}^{n} p_i \\
&= aE(X)+b \\
\mathrm{Var}(aX+b) &= \sum_{i=1}^{n} p_i[ax_i+b-(aE(X)+b)]^2 = \sum_{i=1}^{n} p_i[ax_i-aE(X)]^2 \\
&= a^2\sum_{i=1}^{n} p_i[x_i-E(X)]^2 = a^2\mathrm{Var}(X)
\end{aligned}$$

7. Properties of Two Random Variables

If one models the returns on securities as random variables and an investor holds several securities in his portfolio, then the portfolio return is a new random variable that is a linear combination of the original random variables. Thus it is important to understand the properties of expected value and variance for linear combinations of two or more random variables.

As an illustration, suppose that the returns of two stocks occur with the probabilities given in Table 1.

Table 1.

Probability	Stock X Return	Stock Y Return
0.5	0.5	−0.1
0.5	0.1	0.7
Expected value	0.3	0.3
Variance	0.04	0.16
Standard deviation	0.2	0.4

It is straightforward to compute the expected returns, variances, and standard deviations of the return for the two stocks using the previous formulas. But now suppose that an investor is going to invest 60% of his funds in stock X and 40% of his funds in stock Y. Then the return on the portfolio is $0.6X + 0.4Y$. The expected value, variance, and standard deviation of the portfolio return can be calculated, and the results are given in Table 2.

The expected return is the same for the portfolio as for the individual stocks, but the variance is now less than an investment in either stock. If the investor chose to put 50% of his funds in each stock, then the same process could be repeated to compute the expected value, variances, and standard deviations. However, a little thought shows that most of the calculations would be repeated, and a simpler approach can be used.

Suppose that X and Y are random variables that take on the values (x_i, y_i) with probability p_i for $i = 1, \ldots, n$. Also, let a and b be constants. Then the expected value of a linear

Table 2.

Probability	Portfolio Return
0.5	0.26 = 0.6(0.5) + 0.4(−0.1)
0.5	0.34 = 0.6(0.1) + 0.4(0.7)
Expected value	0.3
Variance	0.0016
Standard deviation	0.04

combination of two random variables is given by

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

The formula is easy to show by plugging through the definition of expected value:

$$\begin{aligned} E(aX + bY) &= \sum_{i=1}^{n} p_i(ax_i + by_i) = a\sum_{i=1}^{n} p_i x_i + b\sum_{i=1}^{n} p_i y_i \\ &= aE(X) + bE(Y). \end{aligned}$$

The variance of a linear combination of two random variables is derived as follows.

$$\begin{aligned} \mathrm{Var}(aX + bY) &= \sum_{i=1}^{n} p_i[ax_i + by_i - (a\mu_X + b\mu_Y)]^2 \\ &= \sum_{i=1}^{n} p_i[a^2(x_i - \mu_X)^2 + 2ab(x_i - \mu_X)(y_i - \mu_Y) + b^2(y_i - \mu_Y)^2] \\ &= a^2\sum_{i=1}^{n} p_i(x_i - \mu_X)^2 + 2ab\sum_{i=1}^{n} p_i(x_i - \mu_X)(y_i - \mu_Y) + b^2\sum_{i=1}^{n} p_i(y_i - \mu_Y)^2 \\ &= a^2\mathrm{Var}(X) + 2ab\mathrm{Cov}(X, Y) + b^2\mathrm{Var}(Y). \end{aligned}$$

The last equality is uses the definition of *covariance*, which is discussed next.

8. Covariance

The covariance between X and Y, denoted $\mathrm{Cov}(X, Y)$ or σ_{XY}, is defined to be

$$\mathrm{Cov}(X, Y) = \sigma_{XY} = E([X - E(X)][Y - E(Y)]) = \sum_{i=1}^{n} p_i(x_i - \mu_X)(y_i - \mu_Y).$$

The covariance measures the degree that the random variable X varies with the random variable Y. If X is larger than its mean when Y is larger than its mean, the term $(x_i - \mu_X)(y_i - \mu_Y)$ is positive. Similarly, if X is smaller than its mean when Y is smaller than its mean, the term $(x_i - \mu_X)(y_i - \mu_Y)$ is also positive. Thus, the covariance of X and Y is positive if X and Y move up and down together. Covariance measures the degree of linear association between X and Y.

Using the definition of covariance, it is straightforward to show the following results. In the results, X, Y, and Z are random variables, and a and b are constants.

$$\begin{aligned}
\mathrm{Cov}(aX, bY) &= ab\mathrm{Cov}(X, Y) \\
\mathrm{Cov}(X, a) &= 0 \\
\mathrm{Cov}(X, X) &= \mathrm{Var}(X) \\
\mathrm{Cov}(X, Y + Z) &= \mathrm{Cov}(X, Y) + \mathrm{Cov}(X, Z)
\end{aligned}$$

The following formula for the covariance of two random variables is equivalent to the earlier definition of covariance, but is computationally simpler.

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \sum_{i=1}^{n} p_i x_i y_i - \mu_X \mu_Y \\
&= E(XY) - E(X)E(Y).
\end{aligned}$$

The derivation of this result follows. It uses the fact that the probabilities must sum to one: $\sum_{i=1}^{n} p_i = 1$.

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \sum_{i=1}^{n} p_i(x_i - \mu_X)(y_i - \mu_Y) = \sum_{i=1}^{n} p_i(x_i y_i - \mu_X y_i - x_i \mu_Y + \mu_X \mu_Y) \\
&= \sum_{i=1}^{n} p_i x_i y_i - \mu_X \sum_{i=1}^{n} p_i y_i - \mu_Y \sum_{i=1}^{n} p_i x_i + \sum_{i=1}^{n} p_i \mu_X \mu_Y \\
&= \sum_{i=1}^{n} p_i x_i y_i - \mu_X \mu_Y - \mu_Y \mu_X + \mu_X \mu_Y = \sum_{i=1}^{n} p_i x_i y_i - \mu_X \mu_Y \\
&= E(XY) - E(X)E(Y).
\end{aligned}$$

The expected value and variance of a linear combination of two random variables can be summarized in the following two formulas.

$$\begin{aligned}
E(aX + bY) &= aE(X) + bE(Y) \\
\mathrm{Var}(aX + bY) &= a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X, Y).
\end{aligned}$$

These results can be used to simplify the calculations of expected values and variances for two random variables. Returning to the example, the covariance of X and Y is -0.08 which can be computed using the earlier formula. The covariance is negative, because the returns of the stocks move in opposite directions – when one stock's return is above its mean the other's return is usually below its mean. The variance of the return of a portfolio that has 60% of the funds invested in stock X and 40% of the funds in stock Y is $0.6^2(0.04) + 0.4^2(0.16) + 2(0.6)(0.4)(-0.08) = 0.0016$. This gives the same result as the direct calculation before.

9. Properties of Several Random Variables

The previous formulas can be extended to linear combinations of several random variables. Suppose there are m random variables, denoted $X_1, \ldots, X_m$. Let $a_1, \ldots, a_m$ be constants. Then the following formulas hold.

$$E(\sum_{j=1}^{m} a_j X_j) = \sum_{j=1}^{m} a_j E(X_j)$$

$$\text{Var}(\sum_{j=1}^{m} a_j X_j) = \sum_{j=1}^{m} a_j^2 \text{Var}(X_j) + 2 \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j>i}}^{m} a_i a_j \text{Cov}(X_i, X_j)$$

For the special case of two random variables, the formulas reduce to the ones given earlier. The derivations of these formulas follow the same lines as the previous ones, with the algebra being a little messier. The derivation of the variance result uses the fact that covariance is *symmetric*, i.e., $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$.

10. Independent Random Variables

Intuitively speaking, two random variables are *independent* if the outcome of one random variable gives no information about the outcome of the other random variable. Table 3 gives an example.

Table 3. Probability table of independent random variables

		Y			
		y_1	y_2	y_3	Row total
X	x_1	0.56	0.16	0.08	0.80
	x_2	0.14	0.04	0.02	0.20
Column total		0.70	0.20	0.10	1.00

X takes on the values x_1 and x_2 and Y takes on the values y_1, y_2, and y_3. Table 3 shows that x_1 is four times as likely to occur as x_2, since $P(X = x_1) = 0.80$ and $P(X = x_2) = 0.20$. If $Y = y_1$, then x_1 is still four times as likely to occur as x_2, since $P(X = x_1 \text{ and } Y = y_1) = 0.56$ and $P(X = x_2 \text{ and } Y = y_1) = 0.14$. No matter what is the outcome of Y, x_1 is always four times as likely to occur as x_2. Similarly, y_1 is seven times as likely to happen as y_3 for any outcome of X. This means X and Y are independent.

Independence is formally defined as follows. Suppose that X takes on the values x_1, ..., x_n with probabilities p_1, ..., p_n and Y takes on the values y_1, ..., y_m with probabilities q_1, ..., q_m. Then X and Y are independent if

$$P(X = x_i \text{ and } Y = y_j) = P(X = x_i)P(Y = y_j) = p_i q_j,$$

for all i and j. If X and Y are not independent, they are called *dependent*.

If X and Y are independent then it is easy to show that $\text{Cov}(X,Y) = 0$. The converse is *not* true. That is, if $\text{Cov}(X,Y) = 0$ then X and Y might not be independent. Roughly speaking, this is because covariance measures the degree of linear association between two random variables. Two random variables can be dependent, but not linearly related.

11. Identically Distributed Random Variables

Two random variables are *identically distributed* if they take on the same values with the same probabilities. More precisely, X and Y are identically distributed if both take on the values z_1, …, z_n with probabilities p_1, …, p_n.

Identically distributed random variables might or might not be independent. For example, suppose that X and Y take on the values 0 and 1 each with probability 1/2. It could be that half of the time both X and Y take on the value 0 and half of the time they both take on the value 1. In this case, X and Y are dependent, e.g., because $P(X = 0, Y = 0) = 1/2 \neq P(X = 0)P(Y = 0)$.

If X and Y are identically distributed, then $E(X) = E(Y)$ and $Var(X) = Var(Y)$.

12. Properties of I.I.D. Random Variables

Suppose there are m independent and identically distributed (i.i.d.) random variables, denoted $X_1, \ldots, X_m$. Let $a_1, \ldots, a_m$ be constants. Denote their common mean by $E(X)$ and common variance by $Var(X)$. Then the following formulas hold.

$$E(\sum_{j=1}^{m} a_j X_j) = E(X) \sum_{j=1}^{m} a_j$$

$$\text{Var}(\sum_{j=1}^{m} a_j X_j) = \text{Var}(X) \sum_{j=1}^{m} a_j^2$$

The first formula does not require independence. The second formula uses independence (which implies that the covariance terms, $\text{Cov}(X_i, X_j)$, are zero).

13. Correlation

The covariance of two random variables is *scale dependent*. That is, suppose $\text{Cov}(X,Y) = \sigma_{XY}$ is the covariance between X and Y and suppose that X is measured in dollars. Suppose that $\hat{X}$ is X measured in units of hundreds of dollars, i.e., $\hat{X} = 0.01X$. Then the covariance of $\hat{X}$ and Y is $\text{Cov}(\hat{X},Y) = \text{Cov}(0.01X,Y) = 0.01\text{Cov}(X,Y)$, which is one hundredth of the covariance of X and Y. In order to measure the strength of the relation between two random variables, it is useful to have a measure that does not depend on units, i.e., a measure that is scale independent. This can be accomplished by defining the *correlation* between X and Y, denoted ρ_{XY}, to be

$$\rho_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}.$$

The correlation between X and Y is the same as the correlation between aX and bY:

$$\begin{aligned}\rho(aX,bY) &= \frac{\text{Cov}(aX,bY)}{\sigma(aX)\sigma(bY)} = \frac{ab\text{Cov}(X,Y)}{a\sigma(X)b\sigma(Y)} \\ &= \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \rho_{XY}.\end{aligned}$$

It can be shown that ρ_{XY} always lies between -1 and $+1$:

$$-1 \le \rho_{XY} \le 1.$$

Corresponding to the previous definitions of covariance and correlation for random variables, there are similar concepts of sample covariance and sample correlation for samples of data.

14. Sample Covariance

Sample covariance is a measure of how two sets of data vary with each other. The two data sets could be the returns of two securities over time. If both sets of data tend to go up and down together, then the covariance will be positive. If own goes up when the other tends to go down, then their covariance will be negative. Suppose the two sets of data points are denoted (x_i, y_i) for $i = 1, \dots, n$. The *sample covariance*, denoted S_{XY}, is given by

$$S_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}.$$

The following formula for sample covariance is equivalent to the previous one, but is computationally simpler.

$$S_{XY} = \frac{\sum_{i=1}^{n} x_i y_i}{n-1} - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n(n-1)}.$$

The derivation of this result follows. It uses the definitions $n\bar{x} = \sum_{i=1}^{n} x_i$ and $n\bar{y} = \sum_{i=1}^{n} y_i$.

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^{n}(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y}) \\ &= \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + \sum_{i=1}^{n}\bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i - \bar{y}n\bar{x} - \bar{x}n\bar{y} + n\bar{x}\bar{y} \\ &= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}.\end{aligned}$$

Notes: The $n - 1$ in the denominator makes S_{XY} an unbiased estimator of the true covariance of the random variables X and Y that generated the data. The sample covariance can be easily calculated in a spreadsheet using the formula given above. For example, suppose that x_1 to x_n are placed in the spreadsheet in the range A1:A50 and y_1 to y_n are in the range B1:B50. Then the spreadsheet formula

```
=SUMPRODUCT(A1:A50, B1:B50)/(COUNT(A1:A50) − 1)
  − SUM(A1:A50) ∗ SUM(B1:B50) / (COUNT(A1:A50) ∗ (COUNT(A1:A50) − 1))
```

calculates the sample covariance. Alternatively, it can be calculated from the output of a regression, which is also easily done in a spreadsheet; see *Introduction to Regression* for an example.

15. Sample Correlation

As before, suppose there a two sets of data points, denoted (x_i, y_i) for $i = 1, \ldots, n$. The *sample correlation*, denoted r_{XY}, is given by

$$r_{XY} = \frac{S_{XY}}{S_X S_Y},$$

where S_{XY} is the sample covariance of the data and S_X and S_Y are the sample standard deviations of the x and y data points, respectively.

Notes: The sample correlation can be easily calculated in a spreadsheet using the formula for S_{XY} and then dividing by S_X and S_Y using the =STDEV function. Alternatively, the sample correlation can be computed in a spreadsheet from the output of a regression; see *Introduction to Regression* for an example.

16. Homework Problems

1. Suppose Y_0 is the current known (i.e., not random) yield on a bond. Also suppose that the yield one day later is $Y_1 = Y_0 + \Delta Y_1$, where ΔY_1 is a random variable with mean 0 and standard deviation σ. Thus, σ is a measure of the daily volatility of the yield change of the bond. Similarly, suppose that the yield of the bond on day j (for $j = 1, \ldots, n$) is $Y_j = Y_{j-1} + \Delta Y_j$, where the ΔY_j are independent and identically distributed random variables with mean 0 and standard deviation σ. What is the volatility, i.e., standard deviation, of the bond's yield on day n? If the daily volatility of yield changes is 0.05%, then under this model, what would you expect to be the monthly volatility of yield changes? annual volatility of yield changes?

2. Suppose that the return of the market is a random variable R_M with mean μ_M and standard deviation σ_M. Suppose that there are two securities X and Y whose random returns are given by $R_X = \beta_X R_M + \epsilon_X$ and $R_Y = \beta_Y R_M + \epsilon_Y$. The parameters β_X and β_Y are constant. The random variable ϵ_X has mean 0 and standard deviation σ_X and the random variable ϵ_Y has mean 0 and standard deviation σ_Y. Furthermore, ϵ_X and ϵ_Y are independent of each other and independent of R_M. What are $\mathrm{Var}(R_X)$, $\mathrm{Var}(R_Y)$, and $\mathrm{Cov}(R_X, R_Y)$ (in terms of the parameters μ_M, σ_M, β_X, β_Y, σ_X, and σ_Y)?

3. Suppose that there are n securities with random returns denoted R_1, R_2, ..., R_n. The expected returns and standard deviations are μ_1, ..., μ_n, and σ_1, ..., σ_n, respectively. The correlation of R_i and R_j is denoted ρ_{ij}. The return of portfolio X is given by $R_X = \sum_{i=1}^{n} x_i R_i$, where the x_i are the weights on the n securities. Similarly, the return of portfolio Y is given by $R_Y = \sum_{i=1}^{n} y_i R_i$, where the y_i are the weights on the n securities. What is the correlation of the returns of the two portfolios? That is, what is a formula for $\rho(R_X, R_Y)$? Give a numerical answer for the parameter values given next: $n = 3$, $\mu_1 = 0.05$, $\mu_2 = 0.10$, $\mu_3 = 0.15$, $\sigma_1 = 0.04$, $\sigma_2 = 0.15$, $\sigma_3 = 0.20$, $\rho_{12} = 0.5$, $\rho_{13} = 0.3$, $\rho_{23} = 0.6$, $x_1 = 0.4$, $x_2 = 0.3$, $x_3 = 0.3$, $y_1 = 0.2$, $y_2 = 0.4$, and $y_3 = 0.4$.

   Hint: The problem is more easily solved using matrix notation. The calculations can then be done in MATLAB.