

Material related to the Foundations of Finance course

Most of you are taking the Foundations of Finance course this semester. There are certain statistical concepts that come up relatively early on in the semester in that course, and for that reason it is useful to highlight them here. We will cover most of this material in much more detail during the semester, but it is important for you to have some exposure to it now. Note that the “Data presentation and summary” handout also covers some of this material, and it should be read before this handout is read. Note also that all of the measures described here are *sample* versions, based on observed data; they each have corresponding *population* versions, and the sample versions are used to estimate those population versions. The concepts discussed in this handout will be illustrated using daily returns for the Korean KOSPI stock index and the Japanese Nikkei 225 stock index for January through October of 2003 (I am indebted to Jason Sohn for gathering these data).

The normal curve and “typical” variation

The output below summarizes the 2003 daily returns. We can see that both the levels of the return (as represented by the mean daily return) and the variabilities of the return (as represented by the standard deviation of the daily return) are similar in the two markets, although interestingly the Japanese market was both more profitable and less volatile during this time period.

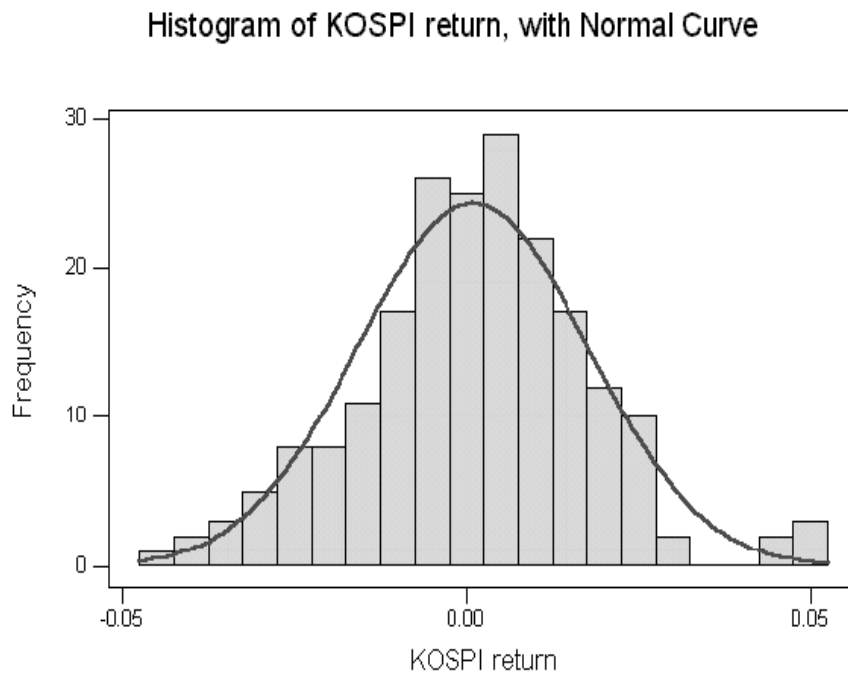
Descriptive Statistics: KOSPI return, Nikkei return

Variable	N	Mean	Median	TrMean	StDev
KOSPI re	203	0.00093	0.00167	0.00088	0.01665
Nikkei r	203	0.00104	0.00044	0.00140	0.01361

Variable	SE Mean	Minimum	Maximum	Q1	Q3
KOSPI re	0.00117	-0.04458	0.04998	-0.00864	0.01104
Nikkei r	0.00096	-0.05092	0.03381	-0.00731	0.01052

Many variables (although certainly not all) follow a symmetric, bell-shaped distribution. Such distributions often (although again, not always) can be approximated using

the **normal** or **Gaussian** distribution. We will talk more about this distribution later on in the semester, but it is helpful to know a little about it now. Here is a histogram of the KOSPI returns, with a normal curve that is consistent with the observed mean and standard deviation superimposed:

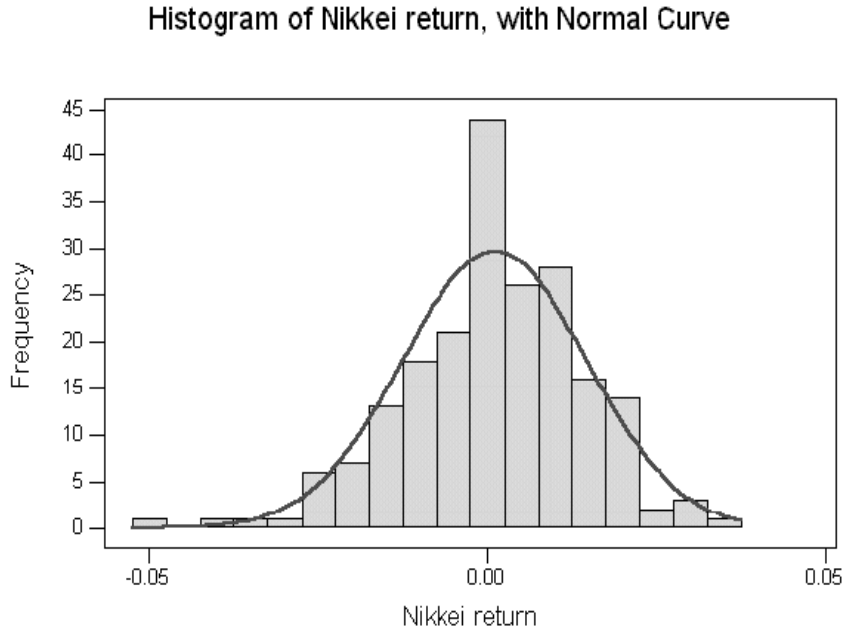


The curve doesn't follow the observed histogram perfectly, but we shouldn't expect that it would; even if the true underlying distribution of returns was normal, random fluctuation would result in some variation from the curve.

The normal distribution (as a representation of “well-behaved” data that are roughly symmetric without outliers) has the following useful property: for such data we can expect that about two-thirds of the observations will fall within one standard deviation of the mean; about 95% will fall within two standard deviations of the mean; and roughly 99.7% will fall within three standard deviations of the mean. So, for example, for the KOSPI returns, we would expect that on about two-thirds of the days the return would be within $0.00093 \pm .01665$, or $(-.01572, .01758)$, on about 95% of the days it would be within $0.00093 \pm (2)(.01665)$, or $(-.03237, .03423)$, and on virtually all days it would be within

$0.00093 \pm (3)(.01665)$, or $(-.04902, .05088)$. In fact, for these data, the observed proportions in these ranges are 70.0%, 94.6%, and 100%, respectively, illustrating a nice correspondence to the normal-based rule.

Consider now the Nikkei returns:

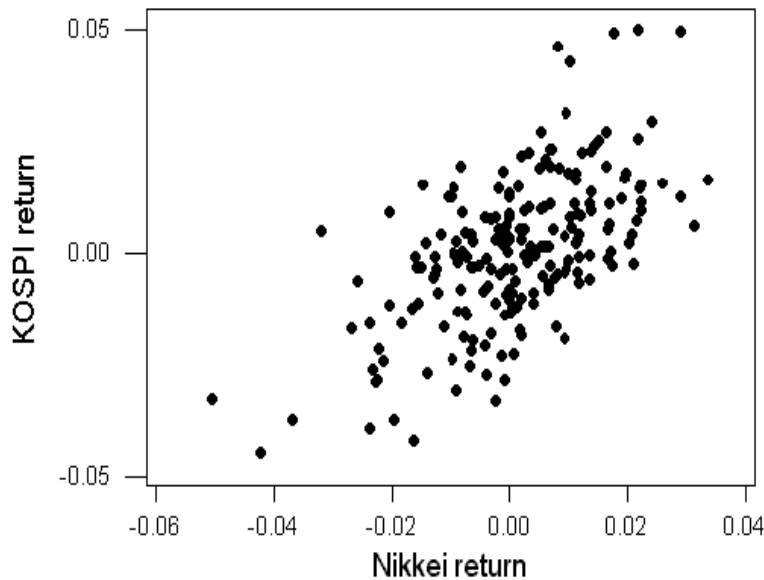


The agreement with the normal curve is noticeably worse, as the distribution is more “peaked” in the center than would be expected. Despite this, the informal rules given earlier are still reasonably effective, as the proportions of days with returns within one, two, and three standard deviations of the mean are 70.4%, 95.6%, and 99.0%, respectively.

Covariance and correlation

Understanding the relationships between and among variables is typically the primary goal of any investigation of data. Consider the following question: are the returns of the KOSPI index related to those of the Nikkei 225? Given the status of these two countries in the Asian financial world, we would certainly expect that this would be the case. The following is a plot of the daily returns for the KOSPI index versus the daily returns for

the Nikkei index.



As we expected, there is a clear direct (positive) relationship between the two returns. That is, on days when the Nikkei index goes up, the KOSPI index also tends to go up (note that these are returns for the same day, so it would not be possible to use the Nikkei variable to predict KOSPI return; we would expect to see a far weaker relationship between returns on different days, if any at all).

We would like to be able to quantify the relationship in this scatter plot. There are two related measures that are commonly used to do this. Consider two variables X and Y , with observed values $\{x_1, \dots, x_n\}$ and $\{y_1, \dots, y_n\}$. That is, the data form n pairs of numbers (KOSPI and Nikkei returns for each day, in the current example). The **sample covariance** between the two variables has the following form:

$$COV(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y}),$$

where \bar{X} is the sample mean of the X values, and \bar{Y} is the sample mean of the Y values. If two variables are directly related (as are the KOSPI and Nikkei returns), values above the

mean for one will tend to occur with values above the mean for the other, and values below the mean for one will tend to occur with values below the mean for the other, implying that most of the terms in the covariance will be positive. That is, a positive covariance corresponds to a direct relationship between the two variables. Similarly, if two variables are inversely related, values above the mean for one will tend to occur with values below the mean for the other, and values below the mean for one will tend to occur with values above the mean for the other, implying that most of the terms in the covariance will be negative. Thus, a negative covariance corresponds to an inverse relationship between the two variables.

The most important use of the covariance is in calculating the variance of a sum of two variables. Specifically, if a and b are two constants, the variance of the weighted sum of two variables X and Y satisfies

$$V(aX + bY) = a^2V(X) + b^2V(Y) + 2ab \text{COV}(X, Y).$$

This is useful when considering the volatility of a portfolio. Consider a portfolio that reflects an investment weighted 40% on the KOSPI and 60% on the Nikkei. The return of this hypothetical investment would be $Z = .4X + .6Y$, where X is the KOSPI return and Y is the Nikkei return. The average return of the portfolio is the weighted average of the returns,

$$\bar{Z} = .4\bar{X} + .6\bar{Y},$$

while the formula given earlier tells us what the variability of the portfolio's return would be. Here are the variances and covariance of the KOSPI and Nikkei returns as given by Minitab:

Covariances: KOSPI return, Nikkei return

	KOSPI re	Nikkei r
KOSPI re	0.00027731	
Nikkei r	0.00013671	0.00018527

The variance of our hypothetical portfolio would thus be

$$V(Z) = (.4^2)(.00027731) + (.6^2)(.00018527) + (2)(.4)(.6)(.00013671) = .0001766876,$$

with square root $SD(Z) = .013292$. Sure enough, here is corresponding output from Minitab for the portfolio:

Descriptive Statistics: Portfolio return

Variable	N	Mean	Median	TrMean	StDev
Portfoli	203	0.00099	0.00123	0.00129	0.01329

Variable	SE Mean	Minimum	Maximum	Q1	Q3
Portfoli	0.00093	-0.04345	0.03755	-0.00597	0.01048

The covariance has a notable flaw as a measure of the strength of the (straight-line) relationship between two variables, since it is in units that correspond to the product of the two variables. A standardized version of the covariance, the **correlation coefficient**, is therefore more commonly used to measure the strength of association between two variables. This statistic, usually called r , has the form

$$r = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{(n-1)s_X s_Y},$$

where s_X is the sample standard deviation of the X values and s_Y is the sample standard deviation of the Y values. Note that

$$r = \frac{COV(X, Y)}{s_X s_Y},$$

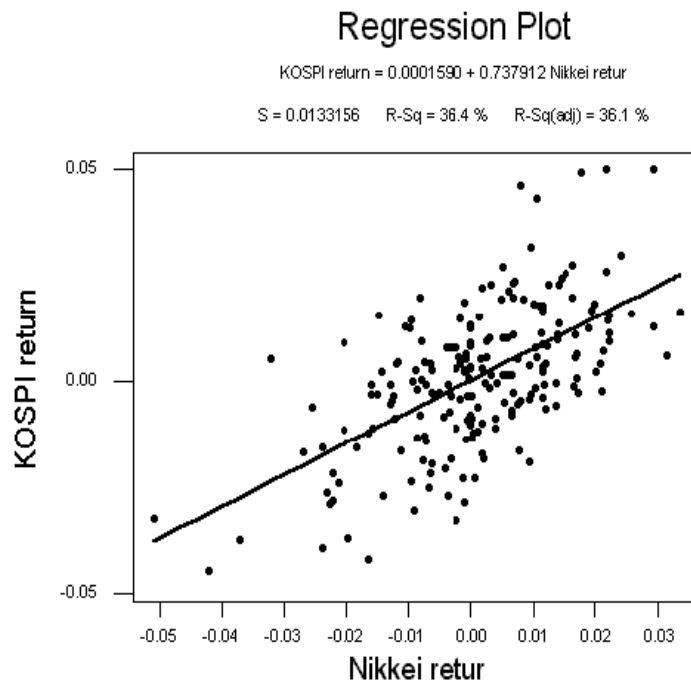
making this a “pure” number (that is, it has no units). It takes on values between -1 and 1 , with $r = 1$ representing a perfect direct linear relationship, $r = -1$ representing a perfect inverse linear relationship, and $r = 0$ representing no linear relationship. The KOSPI and Nikkei returns are reasonably strongly correlated ($r \approx .6$), as the following Minitab output shows:

Correlations: KOSPI return, Nikkei return

Pearson correlation of KOSPI return and Nikkei return = 0.603

Simple linear regression

The correlation coefficient is inherently symmetric; its view of the association between X and Y treats the two variables identically. Sometimes there is an inherent difference in the two variables, in that we would like to model or predict one as a function of the other. This is the essence of a **regression** problem. We will spend a great deal of time talking about regression during the semester, but for now it is useful to recognize the most basic aspects of a regression line. The following graph is obtainable in **Minitab** as a *fitted line plot*:



This is, of course, just the same scatter plot we saw before, with a line relating the two variables superimposed. This line is the **least squares regression line**, and it provides a way to model the KOSPI return as a function of the Nikkei return. The equation of the superimposed line is given in the plot:

The regression equation is

$$\text{KOSPI return} = 0.0001590 + 0.737912 \text{ Nikkei return}$$

There are many aspects to the way this equation is obtained, and its full interpretation, that we will discuss later on in the semester, but at this point you should understand the meaning of the two coefficients, the intercept and the slope. The intercept of the equation (0.001590) represents an estimate of the expected KOSPI return when the Nikkei return is zero. The slope (0.737912) is interpreted as follows: a one unit increase in the Nikkei return (that is, a one percentage point increase in Nikkei return) is estimated to be associated with an expected .738 unit increase in the KOSPI return (that is, a .738 percentage point increase). Note that a negative coefficient would be defined similarly: a slope coefficient of -1.5 , for example, implies that a one unit increase in the predictor is estimated to be associated with an expected 1.5 unit decrease in the response variable.

Minitab commands

To obtain the covariance between two variables, or among a set of variables, click on **Stat** → **Basic Statistics** → **Covariance**. Enter the variable names under **Variables**. Entries along the diagonal of the output matrix are the variances of each variable, while entries along the off-diagonal are the covariances between the pairs of variables.