**Module 5: Covariance and Portfolios**

# Dependence and Independence

In many situations, we are interested in the joint outcomes of two or more random variables.

For a randomly chosen person, *X* = height and *Y* = weight.

For a randomly selected store selling a given product, *X* = price and *Y* = sales.

For next January, *X* = return on Disney and *Y* = return on McDonald’s.

We can think of each of these pairs as a draw from a population of (*x*, *y*) values.

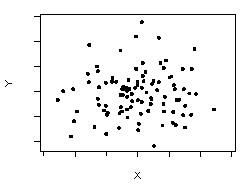
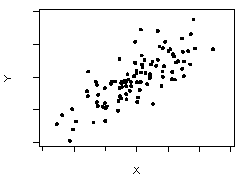
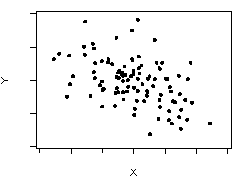
Recall from Module 4 (pg 4-12) that if *X* and *Y* are *independent*, then P(*Y*=*y*) = P(*Y*=*y*|*X*=*x*),

which says that the probability distribution of *Y* is unaffected by observing the outcome of *X* (or vice-versa). If this is not the case, then *X* and *Y* are said to be *dependent*.

Are the above pairs of r.v.s dependent? How would you describe the dependence?

# Linear Association

Suppose that we repeat the process that attaches values to *X* and *Y*, and then plot the joint outcomes. The plot might resemble one of the following:



Note how *X* and *Y* vary together along the diagonals in the first two plots.

Which of these would you associate with the previous (*X*, *Y*) pairs?

The tendency to cluster along the diagonals (seen in the first two plots) is called *linear association*.

Does linear association imply dependence?

Does dependence imply linear association?

# Covariance and Correlation

For *n* pairs of observations (*x*1,*y*1),…,(*xn*,*yn*) covariance and correlation measure the linear association between the two sets of measurements.

Examples:

For *n* individuals, *xi* = height and *yi* = weight.

For *n* days, *xi* = return on Disney and *yi* = return on McDonald’s.

# Covariance

The (sample) covariance between these two sets of measurements is defined as the average cross product[[1]](#footnote-1)



In the product, pairs with both values above (or below) the mean contribute positive summands. Pairs with one value above the mean and another below contribute negative summands.

Interpretation is difficult: covariance depends on the measurement units of *x* and *y*.

# Correlation

The (sample) correlation between *X* and *Y* is defined as



Good news: *rxy* is unit-free. It measures linear association.

*rxy* has the same sign as *sxy*  but satisfies -1 ≤ *rxy* ≤ 1

As *rxy* goes from -1 to 1, linear association goes from negative to positive.[[2]](#footnote-2)

*rxy* = 1 (or −1) indicates perfect positive (or negative) linear association.

What are the signs of *rxy* for the three plots on page 5-2? The signs of *sxy*?

# Population Parameters

We have previously seen sample means and population means, sample variances and population variances. Similarly there are also population[[3]](#footnote-3) covariances and correlations. They are obtained as follows.

Let *μx* , *μy* , *σx* and *σy* denote the population means and standard deviations of *X* and *Y*, respectively. Then, analogous to the sample value (which is the average product)

Intuition: (When you see “*E*,” think “on average.”)

Similarly, the population correlation is



# Correlation and Independence

*Cov* (*X*,*Y*) ≠ 0 and *ρxy* ≠ 0 imply that *X* and *Y* are dependent.

*ρxy* ≠ 0 implies Dependence

*X* and *Y* independent implies *Cov* (*X*,*Y*) = 0 and *ρxy* = 0

Independence implies *ρxy* = 0

# Careful!

Neither *Cov* (*X*,*Y*) = 0 nor *ρxy* = 0 independence. Why?

Think of a plot of data that has *ρxy* = 0 but is not independent. Think: dependence means knowing the position along the x-axis helps you to predict the position along the *y*-axis.

# Weighted Sums of Random Variables[[4]](#footnote-4)

A weighted sum of random variables *X* and *Y* is

*a* *X* + *b* *Y*

where *a* and *b* stand for fixed numbers that are not random.

Jargon - weighted sums such as these are often called linear combinations.

Key application:

Portfolios of investments are weighted sums of random variables each denoting the returns on one of the component investments.

Example: From Module 3, Pink = .5 Red + .5 White

We will now see how

*E*(*aX* + *bY*) can be simply expressed in terms of *E*(*X*) and *E*(*Y*)

*Var*(*aX* + *bY*) can be simply expressed in terms of *Var*(*X*), *Var*(*Y*) and *Cov* (*X*,*Y*)

# FACT #1

For any weighted sum of random variables[[5]](#footnote-5)

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

In Module 3 (p. 3-13), we saw a special case of this formula

*E*(Pink) = *E*(.5 Red + .5 White)

= .5 *E*(Red) + .5 *E*(White)

= .5 (1.71) + .5 (1.008) = 1.359

# FACT #2

For *any* weighted sum of random variables, such as real stocks that are dependent,[[6]](#footnote-6)

*Var*(*aX* + *bY*) = *a*2*Var*(*X*) + *b*2*Var*(*Y*) + 2*abCov*(*X*,*Y*)

In Module 3 (p. 3-13), we saw a special case of this formula

*Var*(Pink) = *Var*(.5 Red + .5 White)

= .52 *Var*(Red) + .52 *Var*(White) + 2(.5)(.5)*Cov*(Red,White)

= .52 (1.755) + .52 (.002) = .439

What happened to *Cov*(Red,White)?

Some useful special cases of these facts:

*E*(*aX*) = *aE*(*X*), *Var*(*aX*) = *a*2*Var*(*X*), *SD*(*aX*) = |*a|SD*(*X*)

For independent *X* and *Y*, the variance of a sum is the sum of the variances

*Var*(*X* + *Y*) = *Var*(*X*) + *Var*(*Y*), 

The same applies to differences of independent random variables[[7]](#footnote-7)

*Var*(*X* − *Y*) = *Var*(*X*) + *Var*(*Y*) , 

# The Mean and Variance of Some Real Portfolios

Let’s now move beyond the Pink portfolio and consider portfolios of real stocks. Two new aspects need to be considered:

1) Returns on the individual investments are typically not independent of each other.

2) The probability distributions of the returns on the individual investments are unknown.

Thus,

1) We cannot assume independence.

2) The unknown characteristics must be estimated from data.

# Stock Market Data Files

*StockReturns.JMP* contains the monthly (net) returns on the stocks of 35 companies from 1975-1999. *StockReturnsSummary.JMP* contains a summary of these returns. [[8]](#footnote-8)

The 35 companies were chosen as the dominant companies in 7 industries in December 1974. Only 22 firms survive the entire 300 month period. The following output summarizes the volatility-adjusted returns for all 35 stocks. The highlighted stocks lasted less than 200 months during 1975-1999.



What happened to the firms that dropped out of the sample?

Of the stocks that survived the whole 25 years (300 months), Disney had the highest volatility-adjusted return 0.0149. Exxon was a close second with adjusted return 0.0148.

Let’s combine these into the equally-weighted portfolio[[9]](#footnote-9) DisExx = .5 Disney + .5 Exxon.



From these summaries[[10]](#footnote-10) we obtain the volatility-adjusted returns[[11]](#footnote-11) as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Investment | Mean | Variance | Mean – Var/2 |
| Disney | .0190 | .0082 | .0149 |
| Exxon | .0159 | .0023 | .0148 |
| DisExx | .0174 | .0033 | .0158 |

Once again, a portfolio offers an improvement in long-term gains over investing 100% in either of the individual investments.

For the equally-weighted portfolio of Disney and Exxon stocks, the return on the portfolio is just the average of the returns on the two stocks

Avg(DisExx) = .0174 = .5(.0190) + .5(.0159)

but for the variance we find Var(DisExx) = .0033 ≠ .52(.0082) + .52(.0023)

which was the formula we used for Pink. What’s missing?

# Role of Covariance

It turns out that[[12]](#footnote-12) that returns on Disney and Exxon are dependent, though only weakly.



The sample correlation between Disney and Exxon is .2988.

From the relationship *rxy* = *sxy*/ *sxsy* on p. 5-4, we can use this output to obtain the covariance between Disney and Exxon

*sxy* = .0013 = (.2988) (.0903)(.0484)

Using the covariance, we can see that .0033, the variance of DissExx on p. 5-14, is obtained using the full formula[[13]](#footnote-13) for the variance of a weighted sum (Fact #2, p. 5-9),

.0033 = .52(.0082) + .52(.0023) + 2(.5)(.5)(.0013)

Would a smaller or a larger covariance value have been preferable?

Let’s now consider pairing Disney with McDonald’s, which had the third largest volatility adjusted return, at .0142.



Compared to Exxon, McDonald’s has a larger average return .0164, but McDonald’s also has a larger variance (.0044) and a higher correlation (.557) with Disney. The plot confirms the stronger positive relationship (notice the more flattened elliptical shape).

For the equally-weighted portfolio of Disney and McDonald’s, we obtain

|  |  |  |  |
| --- | --- | --- | --- |
| Investment | Average | Variance | Avg – Var/2 |
| Disney | .0190 | .0082 | .0149 |
| McDonald’s | .0164 | .0044 | .0142 |
| DisMcD | .0177 | .0048 | .0153 |

Although the portfolio DisMcD has a higher average return than DisExx, it does not outperform DisExx. Why?

In looking to form a portfolio to get higher returns, might it make sense to

use other weights?

use other companies?

use more than two companies?

Whatever the choice, we should still ask whether our means, variances, and covariances will continue to describe future returns accurately.

# Take-Away Review

Portfolios can be represented as weighted sums of random variables, allowing us to study methods for optimizing the performance of a portfolio.

Finding the variance of such a portfolio requires that we know the covariances among the different investments.

Covariance and correlation measure the strength of linear association between pairs of random variables.

*Cov*(*X*,*Y*) ≠ 0 implies dependence, but

*Cov*(*X*,*Y*) = 0 *does not* imply independence.

Covariance is more useful for working with portfolios, but correlation will become more useful in Stat 621.

# Next Module

Samples and populations are key ideas in inferential statistics, and we next take a closer look at these concepts.

# Appendix: Constructing an “Optimal” Portfolio

The pairwise portfolios that we considered put equal weight on the two stocks in the portfolio. There’s no need to divide the investment equally.

More generally, for two investments with returns *X* and *Y*, we could consider portfolios of the form

*Z = pX + (1 − p)Y*

where *p* is the proportion[[14]](#footnote-14) of the total invested in *X*.

Given the means, variances, and covariance between *X* and *Y*, we can use our previous facts to obtain

*E*(*Z*)= *p E*(*X*) + (1 − *p*) *E*(*Y*)

*Var*(*Z*) = *p*2*Var*(*X*) + (1 − *p*)2*Var*(*Y*) + 2*p*(1 − *p*) *Cov*(*X*,*Y*)

An “optimal” portfolio can then be obtained by choosing the value of *p* that maximizes the volatility adjusted return

This problem can be solved with calculus if you want a formula or the Solver in Excel to solve the optimal weights in particular situation.[[15]](#footnote-15)

For example, the optimal Disney-Exxon portfolio is .52 Disney + .48 Exxon yielding a volatility adjusted return of .0158 (only a small improvement).

The optimal Disney-McDonald’s portfolio is .62 Disney + .38 McDonald’s yielding a volatility adjusted return of .0154 (again only a small improvement).

# Larger Portfolios

More importantly, why restrict portfolios to only two investments?

More generally, for *K* investments *X*1,…,*XK* and weights *p*1,…,*pK* (that sum to 1), the general form of a portfolio is given by (another linear combination, just bigger)



The mean and variance of *Z* are given by the formulas





Remain calm – these formulas are just straightforward generalizations of the two variable formulas.[[16]](#footnote-16)

As before, an “optimal” portfolio might then be obtained by choosing the values of *p*1,…,*pK* which maximize the volatility adjusted return *E*(*Z*) – *Var*(*Z*)/2.

Another strategy, sometimes considered, fixes the “risk” of the portfolio (its variance) and then chooses the values mixture weights *p*1,…,*pK* that maximize the expected return.

For example, using the Solver in Excel we find that for a portfolio of a subset of our 35 stocks

|  |  |  |
| --- | --- | --- |
| Var(Z) | Maximum E(Z) | E(Z) - Var(Z)/2 |
| 0.00135 | 0.0139 | 0.013225 |
| 0.00140 | 0.0149 | 0.014200 |
| 0.00150 | 0.0161 | 0.015350 |
| 0.00200 | 0.0191 | 0.018100 |
| 0.00300 | 0.0225 | 0.021000 |
| 0.00400 | 0.0250 | 0.023000 |
| 0.00500 | 0.0271 | 0.024600 |

The increased volatility allows for increased expected returns.[[17]](#footnote-17)

For our previous two stock portfolios we found

|  |  |  |  |
| --- | --- | --- | --- |
| Investment | Average | Variance | Avg – Var/2 |
| DisMcD | .0177 | .0048 | .0153 |
| DisExx | .0174 | .0033 | .0158 |

How do the larger portfolios compare to these?

# Caveat!

The means, variances, and covariances always may change in the future.

1. Section 10.4 of SF, page 225, discussed covariance and correlation (which comes later in this module). [↑](#footnote-ref-1)
2. See SF page 113 for an illustration. [↑](#footnote-ref-2)
3. Although we refer to these as population characteristics, they can also be thought of characteristics of pairs of random variables that represent draws from the population. [↑](#footnote-ref-3)
4. Properties of sums of random variables are in Chapter 10 of SF, particularly section 10.3 (page 224) and section 10.6 (page 232). [↑](#footnote-ref-4)
5. This fact becomes a lot more natural if you think about it like this; how is the average of a sum related to the sum of the averages? They are the same, right? This fact just takes that notion a little farther. [↑](#footnote-ref-5)
6. SF bundles this expression and the previous equation for the expected value as the “Addition Rule for Weighted Sums” (page 232). [↑](#footnote-ref-6)
7. It seems natural that the variance of a sum might be the sum of the variances. But why should the variance of the difference also be the sum of the variances? It might help to think of Var(X-Y) as Var(X+[–Y]) = Var(X) + Var(–Y). Changing the sign of Y does not change its variance, Var(Y) = Var(–Y). For further discussion, see the caution on variances of differences on page 232-233 of SF. [↑](#footnote-ref-7)
8. These data were obtained from the Center for Research in Security Prices (CRSP), which is available to you on Wharton Research Data Services (WRDS). [↑](#footnote-ref-8)
9. As in the dice simulation, this portfolio rebalances after each month so that half of the value of the portfolio is kept in Disney and the other half in Exxon. [↑](#footnote-ref-9)
10. We can get the variance by simply squaring the SD that appears in the output. You can also ask JMP to give you the variance, as well. You do this by adding “more moments” to the output of the description command. [↑](#footnote-ref-10)
11. The volatility adjusted return formula for gross returns from Module 3 (p. 3-11) also holds for net returns. [↑](#footnote-ref-11)
12. This output is obtained using Analyze > Fit Y by X with Exxon as Y and Disney as X, clicking on the red triangle on Bivariate Fit title bar, and selecting Density Ellipse .95. The more concentrated the football shape of the ellipse along a line, the more linearly dependent the pair of measurements. [↑](#footnote-ref-12)
13. Fact 1 on p. 5-8 and Fact 2 on p. 5-9 also hold for sample means, variances and covariances. [↑](#footnote-ref-13)
14. Ordinarily, *p* is chosen between 0 and 1. However, negative values could be used to represent short selling. [↑](#footnote-ref-14)
15. The Excel files *portfolio1.XLS* and *portfolio2.XLS* are set up to do the Solver optimization for the Disney-McDonald’s, and the Disney-Exxon portfolios. Just select Solver from the Tools menu. [↑](#footnote-ref-15)
16. At least remain calm for now. You’ll see them again in finance. They *really* like these formulas! [↑](#footnote-ref-16)
17. At some point, however, accepting more risk will not lead to higher volatility adjusted returns. You’ll see more of this concept when you study the “efficient frontier” in finance. [↑](#footnote-ref-17)