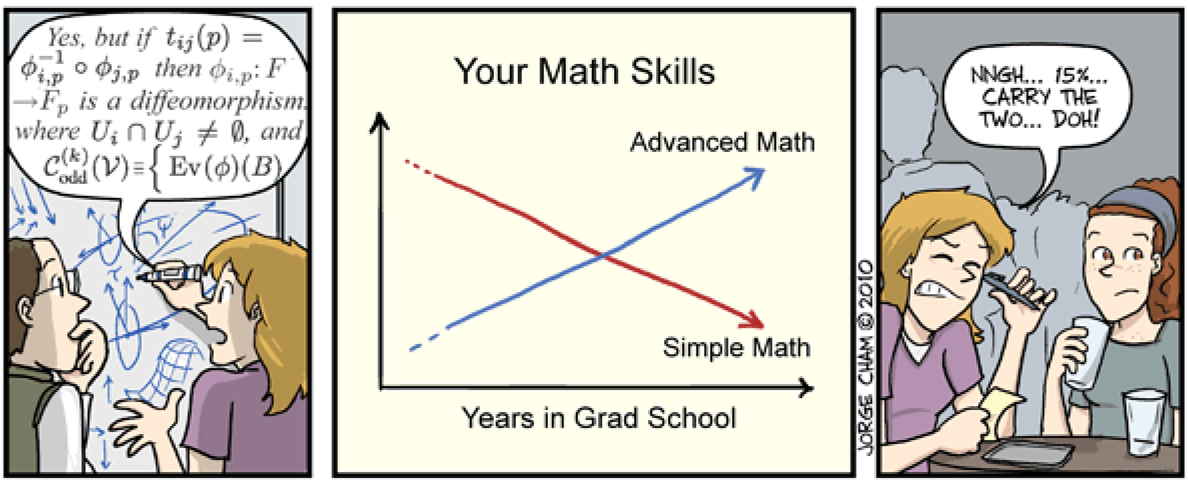
Foundations of Econometrics
Using SAS Simulations and Examples

ECNS 561 Course Notes



Anton Bekkerman

Any errors in this document are the responsibility of the author. Corrections and comments regarding any material in this text are welcomed and appreciated. The material in this document is intended to be supplementary (not as a substitute) to attending lectures regularly.

Direct correspondence to: anton.bekkerman@montana.edu

This page intentionally left blank.

Contents

How to Use These Notes

Students of applied econometrics can often struggle to "connect the dots" when introduced to the more rigorous statistical and mathematical properties underlying various concepts of econometrics. Often, when asked to provide a detailed explanation of even fundamental concepts, students are unable to explain the steps involved in getting from a question to an answer. This is, of course, normal, and most students begin feeling comfortable with the numerous aspects of econometrics only after working on and completing several research projects, which offer opportunities to rigorously work with data, modeling, and analysis inference. Providing such experiential opportunities earlier in a student's learning process can lay a stronger foundation for understanding econometrics and help "connect the dots" between theory and application. This is a major objective of this text.

A useful approach to providing opportunities to better understand theoretical concepts and their connections to applied work is through data-driven examples. In this text, all such examples will be in the form of simulations. Simulations offer the important advantage of controlling data and modeling structures, allowing for the most flexibility in illustrating important concepts through data analysis. This text provides numerous examples of such illustrations, but students are encouraged to use these examples as a foundation for further exploratory analysis.

Throughout the text, many concepts important to econometrics will be explained using both statistical theory and simulation-based examples. All code for executing simulation-based examples is written for use with the SAS software and was coded using SAS version 9.3. Most examples use either the matrix algebra-based `IML` procedure or the `DATA` step, with a multitude of other SAS procedures used to illustrate important concepts. All code used to generate simulations and examples is presented throughout the text and can be downloaded from the course website.

An overview of SAS basics is provided in Appendix 2 of this text and students are encouraged to reference this material when reading and using the SAS-based simulations and examples.

Chapter 1: Introduction to Econometrics

Econometrics is the empirical side of economics. An economist seeks to ask relevant and timely questions; econometrics provides tools that can be used to quantify answers to these questions. We can loosely generalize econometrics to be a tool for the following:

- Quantifying economic relationships
  - How much does library use increase as patrons become more affluent?
  - How do the prices of corn affect planting decisions for soybeans?
  - How does the size of a cattle's rib-eye area affect the cattle's sale price?
- Test competing hypothesis
  - Does one economic model explain the world more accurately than another economic model?
- Answer questions about policies
  - How does a tax cut affect unemployment?
  - Can a policy increase social welfare?
  - Did the No Child Left Behind policy positively or adversely affect children's education levels?
- Forecast
  - What will the price of corn be in two months?
  - How might unemployment rates behave in the next three quarters?
  - How do inflation rates behave in good and bad economic states?

1.1 Economic vs. Econometric Models

While economic models provide us with a theoretical representation of a phenomenon, econometric models offer an opportunity to empirically measure the phenomenon.

1.1.1 Economic models

Economic models are simplified representations of reality, which are often described through mathematical equations. Let's consider several economic models:

- Single-equation models: the behavior of a dependent variable Y is described by one or more independent variables X. That is:

  $$Y = f(X_1, X_2, \ldots, X_n)$$

  For example, we may wish to characterize the demand for beef at the retail level as a function of variables such as consumer income, the price of beef, and the price of a close substitute. That is, $D_{\text{beef}} = f(\text{Income},\ P_{\text{beef}},\ P_{\text{pork}})$.

- Multi-equation models: the behavior of a dependent variable Y is described by several independent variables X, which can are themselves explained by other equations. That is:

  $$Y = f(X_1, X_2, \ldots, X_n)$$
  $$X_1 = f(X_2, Z_1, Z_2, \ldots, Z_n)$$

  For example, while the demand for retail level beef may be characterized as above, we can also hypothesize that an individual's income is a function of the individual's education level and age. That is,

  $$D_{\text{beef}} = f(\text{Income},\ P_{\text{beef}},\ P_{\text{pork}})$$
  $$\text{Income} = f(\text{Education, Age})$$

- Time-series models: the behavior of a dependent variable Y is described by past values of the dependent variable Y_{t-1} as well as independent variables X. That is:

  $$Y_t = f(Y_{t-1}, Y_{t-2}, X_1, X_2, \ldots)$$

  For example, it is reasonable to assume that an individual's demand for beef in month t is partially explained by their demand for beef in the previous month, $t-1$, as well as their income, the price of beef, and the price of substitutes. That is, $D_{\text{beef},t} = f(D_{\text{beef},t-1},\ \text{Income}_t,\ P_{\text{beef},t},\ P_{\text{pork},t})$.

When we construct economic models, we conceptualize a conductible experiment. For example, consider the following questions and the approach that we might take to quantify answers:

1. What is the effect of advertisement on attendance of farmers' markets?

   Hypothesis: increasing advertisement will increase the number of people visiting farmers' markets.

   Experiment: choose two identical locations; use a higher amount of advertisement in one location; compare the attendance at each farmer's market.

   Although it would be almost impossible to find two identical locations (each may have differences in education levels, per capita income, ability to attend farmer's market, etc.) in order to conduct the experiment, it is useful to conceptualize how the experiment might be conducted.

2. Is the adoption of new technology positively related to education?

   Hypothesis: more educated individuals will adopt new technology faster than less educated individuals.

   Experiment: randomly survey individuals to determine their education level and how quickly they acquire a new technological tool after it is introduced.

In both cases, data from either the "experiment" in (1) or the survey results in (2) can be used to estimate an *econometric model*.

1.1.2 Econometric models

When we construct economic models, we assume that the relationship between the dependent variable Y and the independent (or lagged dependent) variables is *deterministic*. That is, given a value of X, we can exactly predict a value of Y. For example, consider a linear model:

$$Y = \beta_0 + \beta_1 X_1$$

Suppose that $X_1 = 0$. Then the exact value of the dependent variable is $Y = \beta_0$. If $X_1 = 1$, then $Y = \beta_0 + \beta_1$.

Life would be simple if the deterministic economic models depicted the truth. However, there is a lot of randomness that exists in life. For example, if you roll a die twice, you may get two different outcomes. Or, by increasing X_1 from 0 to 1 may not necessarily increases the value of Y by β_1 units (e.g., increasing fertilizer will increase yield in year 1, but there is no effect in year 2 because of different weather conditions).

In an econometric model, we account for this randomness. That is, an econometric model indicates that the behavior of Y may not fully be determined by certain independent (or lagged dependent) variables. Rather, there is always a chance that increasing X_1 from 0 to 1 will change Y by more or less than β_1 units.

What causes random behavior in Y?

- *Observable differences* – for example, differences in age, experience, and other characteristics.
- *Unobservable differences* – for example, ability levels, motivation.
- *Unpredictable events* – for example, weather events, space invasions.

Because we are unable to characterize the real world as being purely deterministic, econometrics is used to quantify the randomness. Consider how the economic model is re-written in order to account for randomness:

$$Y = f(X_1, X_2, X_3, \ldots, X_n, \varepsilon)$$

where

$$f(X_1, X_2, X_3, \ldots, X_n) \equiv \text{deterministic, observable components}$$
$$\varepsilon \equiv \text{random or unobservable component}$$

Because introducing the random component implies that any particular realization of Y can also be random, we treat Y as a random variable. This entails making assumptions about the true distribution from which each observed value of Y was drawn.

1.2 Data types and issues

Data are the backbone to any empirical economic study because it allows us to measure relationships about variables, test hypotheses, and forecast values. There are three basic structures of data:

1. Cross-sectional data: information about entities (e.g. individuals, firms, households, farms) at a particular point in time. That is, these data are a "snapshot" of an economic situation. An important assumption is that each observation in a cross-sectional data set is *independent* of all other observations in that data set.

   Example: Bull sale prices at a 2009 auction.

Bull ID	Sale price	Birth weight	365-day weight	Age	Daily gain
1	2,500	76	1,122	381	2.96
2	3,000	72	1,224	381	3.34
3	3,000	76	1,171	370	3.05
⋮	⋮	⋮	⋮	⋮	⋮

2. Time-series data: information on the same entity over a period of time. Often, each observation is dependent on preceding observations.

   Example: Wheat planting and yield data for a county in Montana.

Commodity	Year	Planted Acres	Yield
WHEAT	1998	173,100	30.60
WHEAT	1999	175,400	29.90
WHEAT	2000	158,900	29.90
⋮	⋮	⋮	⋮

3. Panel data: combination of cross-sectional and time-series types. Provides information about a variety of entities over a time period. This data type is becoming more readily available as technological advances have allowed for better data collection.

   Example: Information on public library operation and circulation statistics over several years.

Library ID	Year	Hrs. Open	# Visits	Adult circulation	Child circulation
AK0001	2004	1,300	4,500	6,657	1,039
AK0001	2005	1,300	4,600	6,407	1,218
AK0002	2004	11,280	963,000	1,011,765	487,481
AK0002	2005	11,352	966,154	988,804	450,083
AK0003	2004	486	3,260	3,992	1,199
AK0003	2005	490	2,978	2,453	875
⋮	⋮	⋮	⋮	⋮	⋮

1.2.1 Data sources

Data are typically collected using original surveys and from existing data sets. Because there may be significant costs in collecting your own data, it is typically the case that much research uses data that has already been collected. Examples of data sources include:

- Bureau of Labor Statistics
- U.S. Census
- Current Population Survey (CPS)
- Agricultural Census
- USDA Risk Management Agency (RMA)
- USDA National Agricultural Statistics Service (NASS)

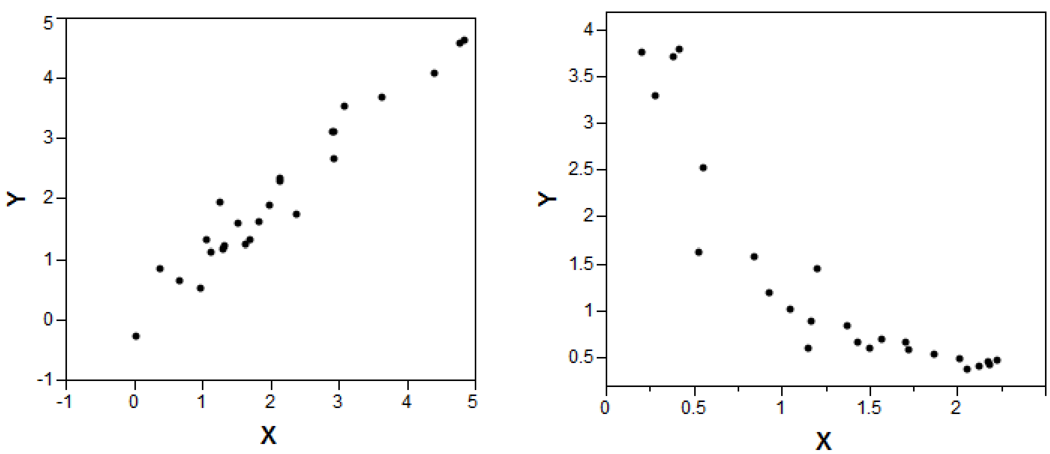
It is important to realize that *all* data must be handled with care. You must spend time inspecting and cleaning your data – being careless can lead to unexpected and often incorrect results. For example, suppose that you examine prices of shoes over the past 50 years. However, you forget to adjust for inflation. As a result, you forecast that the price of shoes in 2015 will be $500/pair.

1.2.2 Functional form issues

When specifying an econometric model, an important realization is that there may be numerous relationships between a dependent variable Y and a particular independent variable X. For example, a simple assumption is to assume that Y is a linear function of X; that is, $Y = \beta_0 + \beta_1 X_1$. This implies that for each additional unit of X_1, Y increases at the same rate. Alternatively, the dependence of Y on X may be non-linear. Such differences in the data are shown in figure 1.1.

The choice of a functional form is important because it represents your hypothesis about the relationship between Y and each X. Often, you can get an idea of these relationships by examining the data (e.g., plotting the relationship between the dependent and independent variables). However, imposing a wrong functional relationship can lead to poor inferences.

Figure 1.1: An Example of Linear and Non-linear Relationships Between Variables



1.3 Conducting an econometric project

1. *Pick a topic / pose a question* – when completing this step, you should remember to ask yourself why you are choosing a particular topic. That is, why does answering this question matter? Who cares that this question be answered? Is it possible to answer this question? Are there available data for answering the question?

2. *Construct a theoretical economic model* to answer the question.

3. *Specify an econometric model.*

4. *Gather data.*
5. *Estimate the econometric model.* The estimation should be such that the results are precise and there is a low margin of error.
6. *Interpret the results* and *perform statistical tests.*
7. *Evaluate whether econometric results are congruent with the theoretical model.* That is, do your outcomes make economic sense?

1.4 Additional resources

If you are struggling to understand a particular topic, wish to obtain a more in-depth knowledge about the topic, or are simply looking for an alternative explanation that can your understanding, you are encouraged to reference other sources. The Econometric References and Resources offers a good starting point for such references.

Chapter 2: Matrix algebra

Matrix algebra is a powerful tool for analyzing statistics. Most of the statistical analyses performed by statistical analysis software are performed using matrices. When we begin to learn about probability theory, mathematical statistics, and linear regressions, matrices will be used extensively. This chapter is intended as an overview and reference guide for using matrices.

2.1 Basic Definitions

- *Vector*: is a row or column of entities (e.g., values, variables). A column vector of dimension $(n \times 1)$ and a row vector of dimension $(1 \times m)$ are as follows:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{(n\times 1)} \qquad \boldsymbol{y} = [y_1 \; y_2 \; \ldots \; y_m]_{(1\times m)}$$

- *Matrix*: a rectangular array. A matrix of dimension $(n \times m)$ is as follows:

$$\boldsymbol{y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ \vdots & \ddots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nm} \end{bmatrix}_{(n\times m)}$$

  Notation: we will specify the number of rows in a matrix/vector by n, and the number of columns in a matrix/vector by m. **You should always write out the dimensions of matrices/vectors.**

  Example: what are the dimensions of the following matrix?

$$\boldsymbol{y} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{bmatrix}$$

There are two rows and three columns. Thus, the dimensions are (2×3).

- *Scalar*: a matrix of dimension (1×1). A scalar is a constant.

$$\boldsymbol{y} = \begin{bmatrix} y_{11} \end{bmatrix}_{(1\times1)}$$

- *Special matrices*: there are a number of matrices that can be identified directly by their name.
  - *Square matrix*: a matrix in which the number of rows is the same as the number of columns. That is, it is a matrix of dimension $(n \times n)$.
  - *Diagonal matrix*: a square matrix in which only the diagonal elements are non-zero. All off-diagonal terms are zero. An $(n \times n)$ diagonal matrix is as follows:

$$\boldsymbol{y} = \begin{bmatrix} y_{11} & 0 & \dots & 0 \\ 0 & y_{22} & \dots & 0 \\ \vdots & \ddots & \vdots & \\ 0 & 0 & \dots & y_{nn} \end{bmatrix}_{(n\times n)}$$

  - *Identity matrix*: a diagonal matrix with ones on the diagonal and zeros on the off-diagonal. An $(n \times n)$ identity matrix is as follows:

$$\boldsymbol{I} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \ddots & \vdots & \\ 0 & 0 & \dots & 1 \end{bmatrix}_{(n\times n)}$$

2.2 Matrix Properties and Manipulations

- *Matrix dimensional equality*: two matrices are dimensionally equal if each matrix has the same number of rows and columns as the other matrix. That is, a matrix of dimensions $(n \times m)$ is dimensionally equivalent to any other matrix of dimensions $(n \times m)$. Two dimensionally equal matrices are as follows:

$$\begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{bmatrix}_{(2\times3)} \qquad == \qquad \begin{bmatrix} 4 & 14 & 32 \\ 22 & 44 & 64 \end{bmatrix}_{(2\times3)}$$

- *Matrix addition/subtraction*: elementwise addition/subtraction of dimensionally equal matrices. The resulting matrix is of the same dimensions as the two original matrices. An example of matrix addition is as follows:

$$\begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{bmatrix}_{(2\times3)} \quad + \quad \begin{bmatrix} 4 & 14 & 32 \\ 22 & 44 & 64 \end{bmatrix}_{(2\times3)} \quad = \quad \begin{bmatrix} 5 & 16 & 35 \\ 25 & 48 & 69 \end{bmatrix}_{(2\times3)}$$

- *Scalar multiplication/division*: elementwise multiplication of a matrix by a constant. An example of scalar multiplication is as follows:

$$2 \cdot \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{bmatrix}_{(2\times3)} \quad = \quad \begin{bmatrix} 2 & 4 & 6 \\ 6 & 8 & 10 \end{bmatrix}_{(2\times3)}$$

- *Matrix multiplication*: requires that two matrices are of *appropriate* dimensions. That is, for two matrices $\boldsymbol{y}$ and $\boldsymbol{z}$, the product $\boldsymbol{yz}$ requires that the number of columns in the matrix $\boldsymbol{y}$ is equal to the number of rows in matrix $\boldsymbol{z}$. For example, the following matrices can be multiplied:

$$\boldsymbol{y} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{bmatrix}_{(2\times3)} \qquad \boldsymbol{z} = \begin{bmatrix} 1 & 3 \\ 2 & 1 \\ 3 & 0 \end{bmatrix}_{(3\times2)}$$

  The first matrix is of dimension (2×3) and the second is of dimension (3×2). It is always easy to know whether you can multiply two matrices by placing the dimensions next to each, and seeing if the "inside" numbers match. That is, the dimensions (2×3) (3×2) can be multiplied because the "inside" numbers 3 match. The resulting matrix will have the dimensions of the "outside" numbers. That is, the resulting matrix product $\boldsymbol{yz}$ will be of dimensions (2×2).

  To multiply matrices, the method to remember is "row 1 times column 1 plus row 2 times column 2 plus ...." That is, a new element in the product matrix $\boldsymbol{yz}$ is obtained by multiplying the element in the i^{th} row in $\boldsymbol{y}$ by the element in the j^{th} column of $\boldsymbol{z}$, and adding these products together. Consider how to multiply the following matrices:

$$\begin{bmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \end{bmatrix}_{(2\times3)} \cdot \begin{bmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \\ z_{31} & z_{32} \end{bmatrix}_{(3\times2)} =$$

$$\begin{bmatrix} (y_{11} \cdot z_{11} + y_{12} \cdot z_{21} + y_{13} \cdot z_{31}) & (y_{11} \cdot z_{12} + y_{12} \cdot z_{22} + y_{13} \cdot z_{32}) \\ (y_{21} \cdot z_{11} + y_{22} \cdot z_{21} + y_{23} \cdot z_{31}) & (y_{21} \cdot z_{12} + y_{22} \cdot z_{22} + y_{23} \cdot z_{32}) \end{bmatrix}_{(2\times2)}$$

Example: Suppose you have matrices $\boldsymbol{y}$ and $\boldsymbol{z}$. Find the matrix product $\boldsymbol{yz}$.

$$\boldsymbol{y} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{bmatrix}_{(2\times3)} \qquad \boldsymbol{z} = \begin{bmatrix} 1 & 3 \\ 2 & 1 \\ 3 & 0 \end{bmatrix}_{(3\times2)}$$

$$\boldsymbol{yz} = \begin{bmatrix} (1\cdot1+2\cdot2+3\cdot3) & (1\cdot3+2\cdot1+3\cdot0) \\ (3\cdot1+4\cdot2+5\cdot3) & (3\cdot3+4\cdot1+5\cdot0) \end{bmatrix}_{(2\times2)} = \begin{bmatrix} 14 & 5 \\ 26 & 13 \end{bmatrix}_{(2\times2)}$$

Properties of Matrix Addition and Multiplication

Assume that α and β represent any scalar, and $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$ are matrices.

1. $(\alpha+\beta)\boldsymbol{y} = \alpha\boldsymbol{y} + \beta\boldsymbol{y}$
2. $\alpha(\boldsymbol{y}+\boldsymbol{z}) = \alpha\boldsymbol{y} + \alpha\boldsymbol{z}$
3. $(\alpha\beta)\boldsymbol{y} = \alpha(\beta\boldsymbol{y})$
4. $\alpha(\boldsymbol{yz}) = (\alpha\boldsymbol{y})\boldsymbol{z}$
5. $\boldsymbol{y}+\boldsymbol{z} = \boldsymbol{z}+\boldsymbol{y}$
6. $(\boldsymbol{yz})\boldsymbol{x} = \boldsymbol{y}(\boldsymbol{zx})$
7. $\boldsymbol{y}(\boldsymbol{z}+\boldsymbol{x}) = \boldsymbol{yz}+\boldsymbol{yz}$
8. $\boldsymbol{Iy} = \boldsymbol{yI} = \boldsymbol{y}$
9. Generally, $\boldsymbol{yz} \neq \boldsymbol{zy}$, even if $\boldsymbol{y}$ and $\boldsymbol{z}$ are both square matrices.

Carefully note the property (9). This is an important property that you need to remember.

Example: Recall from our previous example that $\boldsymbol{yz} = \begin{bmatrix} 14 & 5 \\ 26 & 13 \end{bmatrix}_{(2\times2)}$. Now, calculate $\boldsymbol{zy}$ and show that $\boldsymbol{yz} \neq \boldsymbol{zy}$.

$$\boldsymbol{zy} = \begin{bmatrix} 10 & 14 & 18 \\ 5 & 8 & 11 \\ 3 & 6 & 9 \end{bmatrix}_{(3\times3)} \neq \begin{bmatrix} 14 & 5 \\ 26 & 13 \end{bmatrix}_{(2\times2)}$$

- *Transpose*: the interchanging of rows and columns in a matrix. You can also think of a transpose as flipping the matrix along the diagonal. The transpose is denoted by either a tilde $'$ or by a superscript ^{T}.

$$\boldsymbol{y} = \begin{bmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \end{bmatrix}_{(2\times3)} \qquad \boldsymbol{y'} = \boldsymbol{y^T} = \begin{bmatrix} y_{11} & y_{21} \\ y_{12} & y_{22} \\ y_{13} & y_{23} \end{bmatrix}_{(3\times2)}$$

As you can see, the first row of $\boldsymbol{y}$ became the first column of $\boldsymbol{y'}$. The second row of $\boldsymbol{y}$ became the second column of $\boldsymbol{y'}$.

Example: Find the transpose of the matrix $\boldsymbol{y} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{bmatrix}_{(2\times3)}$

$$\boldsymbol{y} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{bmatrix}_{(2\times3)} \qquad \boldsymbol{y'} = \begin{bmatrix} 1 & 3 \\ 2 & 4 \\ 3 & 5 \end{bmatrix}_{(3\times2)}$$

Properties of Matrix Transposes

1. $(\boldsymbol{y}')' = \boldsymbol{y}$
2. $(\boldsymbol{yz})' = \boldsymbol{z}'\boldsymbol{y}'$
3. $(\boldsymbol{y} + \boldsymbol{z})' = \boldsymbol{y}' + \boldsymbol{z}'$
4. If $\boldsymbol{x}$ is an $(n \times 1)$ vector, then $\boldsymbol{x}'\boldsymbol{x} = \sum_{i=1}^{n} x_i^2$
5. If a matrix is square and $\boldsymbol{y}' = \boldsymbol{y}$, then $\boldsymbol{y}$ is a symmetric matrix.

- *Inverse*: only square matrices can be inverted. The inverse if an $(n \times n)$ matrix $\boldsymbol{y}$ is identified as $\boldsymbol{y}^{-1}$. The inverse of a matrix has the same dimensions as the original matrix.

Example: suppose you have a 2× 2 matrix $\boldsymbol{y} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$. Find the inverse of $\boldsymbol{y}$.

For a 2× 2 matrix, the inverse is as follows:

$$\boldsymbol{z}^{-1} = \frac{1}{\det(\boldsymbol{z})} \cdot \begin{bmatrix} z_{22} & -z_{12} \\ -z_{21} & z_{11} \end{bmatrix}$$

where $\det(\boldsymbol{z}) = (z_{11}z_{22} - z_{12}z_{21})$ is the determinate of the matrix $\boldsymbol{y}$. Therefore, for the provided matrix, the inverse can be solved for as follows:

$$\boldsymbol{y}^{-1} = \frac{1}{(1 \cdot 4 - 2 \cdot 3)} \cdot \begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ 1.5 & -0.5 \end{bmatrix}$$

Properties of Matrix Inverses

1. For a matrix $\boldsymbol{y}$ of any dimension $(n \times m)$, you can take the inverse of:
   (a) $\boldsymbol{y}'\boldsymbol{y}$
   (b) $\boldsymbol{y}\boldsymbol{y}'$
2. $(\boldsymbol{yz})^{-1} = \boldsymbol{z}^{-1}\boldsymbol{y}^{-1}$
3. $\boldsymbol{y}^{-1}\boldsymbol{y} = \boldsymbol{y}\boldsymbol{y}^{-1} = I$
4. $(\boldsymbol{y}^{-1})' = (\boldsymbol{y}')^{-1}$

- *Idempotent matrix*: any matrix for which $\boldsymbol{yy} = \boldsymbol{y}$ is true. This concept will be important when we discuss projection matrices.

2.3 Linear independence

Linear independence is a crucial issue in econometrics. We will discuss the reasons later, but for now, let's explore how linear independence (and lack thereof) may affect matrix manipulations.

Consider a set of $(n \times 1)$ vectors, $\{\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots \boldsymbol{y_k}\}$, and all scalars in the set $\{c_1, c_2, \ldots, c_k\}$ are zero. The vectors are *linearly independent* if and only if:

$$c_1\boldsymbol{y_1} + c_2\boldsymbol{y_2} + c_3\boldsymbol{y_3} + \ldots c_k\boldsymbol{y_k} = \mathbf{0}$$

If this equation holds when any of the scalars are non-zero, then the set of vectors are *linearly dependent*. Another way to think about linear dependence is asking whether any $\boldsymbol{y_k}$ is a linear combination of any other two vectors. That is, $\boldsymbol{y_k}$ is a linear combination of $\boldsymbol{y_1}$ and $\boldsymbol{y_2}$ if:

$$\boldsymbol{y_k} = c_1\boldsymbol{y_1} + c_2\boldsymbol{y_2}$$

2.3.1 Matrix rank

The *rank* of a matrix reveals the maximum number of linearly independent columns are in a matrix. If the matrix $\boldsymbol{y}$ has the dimensions $(n \times m)$ and the $rank(\boldsymbol{y}) = m$, then the matrix $\boldsymbol{y}$ is *full rank*.

Example: What is the rank of the following two matrices?

$$\boldsymbol{y} = \begin{bmatrix} 1 & 2 & 13 \\ 3 & 4 & 2 \end{bmatrix}_{(2\times 3)} \qquad \boldsymbol{z} = \begin{bmatrix} 1 & 3 & 2.5 \\ 2 & 4 & 4 \\ 3 & 5 & 5.5 \end{bmatrix}_{(3\times 2)}$$

The rank of $\boldsymbol{y}$ is $rank(\boldsymbol{y}) = 3$ because there is no column that is a linear combination of other columns. However, the rank of $\boldsymbol{z}$ is $rank(\boldsymbol{z}) = 2$ because the third column is a linear combination of columns 1 and 2 ($col_1 + 0.5 \cdot col_2 = col_3$). Thus, $\boldsymbol{y}$ is full rank, but $\boldsymbol{z}$ is not full rank.

Note that if a square matrix is *not full rank*, then it is not invertible (because essentially, it is not a square matrix).

SAS Code: Example of Matrix Operations

```
/* Entering into IML (interactive matrix language) mode */
proc iml;

/* Set option to display row and column numbers */
reset autoname;

/*Constructing a vector:
    columns separated by space, rows separated by "," */
v1 = {1 2 3 4};
print v1;

/*Constructing a matrix:
    columns separated by space, rows separated by "," */
m1 = {1 2, 3 4};
print m1;

/* Change a value in a vector or a matrix */
/* Example: Change the value of the second column in the vector v1 */
```

```
v1[,2] = 5;
print v1;

/* Example: Change the value of the second row,
            first column in matrix m1 */
m1[2,1]        = 5;
print m1;

/* Construct an identity matrix; I(num columns) */
i1 = i(2);
print i1;

/* Construct a matrix of all ones; J(rows, cols, value) */
ones1 = j(2,2,1);
print ones1;

/* Transpose a vector or matrix:
    t(vector name) or (vector name)` */
v1t = t(v1);
m1t = m1`;
print v1t, m1t;

/* Inverse of square matrix; inv(matrix name) */
m1inv = inv(m1);
print m1inv;

/* Adding or subtracting a scalar */
v2 = v1 - 1;
m2 = m1 - 1;
print v2, m2;

/* Multiplying/dividing by scalars */
v3 = v1 # 2;
m3 = m1 # 2;
print v3, m3;

/* Raising to a power */
v4 = v1 ## 2;
m4 = m1 ## (0.5);
print v4, m4;
```

```
/* Vector / matrix addition/substraction */
v5 = v1 + v1;
m5 = m1 + m1;
print v5, m5;

/* Vector/matrix multiplication */
v6 = v1 * v1‘;
m6 = m1 * m1‘;
print v6, m6;

/* Determine the rank of a matrix */
m1rank=round(trace(ginv(m1)*m1));
print m1rank;
```

Chapter 3: Foundation of Probability Theory

Most empirical economics is non-experimental. That is, researchers do not control treatments. Rather, treatments, explanatory variables, and associated outcomes are simply observed. However, it is possible to come up with a framework that can relate the observed treatments and explanatory variables to outcomes.

What's the catch? There is almost never 100% *a priori* certainty that a particular treatment and/or explanatory variable will lead to a particular outcome. That is:

$$Outcome = f(explanatory\ variables,\ treatments) + random\ error$$

Because of the error term, repeating an experiment using the identical explanatory variables and treatments may not yield the same outcome as a preceding iteration of the experiment. In other words, the *Outcome* variable is a random variable.

3.1 Random Variables

Random variable: a variable whose values are unknown prior to carrying out an experiment. Another way to think about is that a random variable can have different outcomes depending on a "state of nature."

There are two types of random variables: *discrete* and *continuous*.

3.1.1 Discrete random variables

Discrete random variables are ones that take on a limited number of distinct values. That is, the outcomes are countable. Examples include: yes/no replies; number of cattle on a ranch; number of children.

Recall the question we asked earlier: What is the effect of advertisement attendance of farmers' markets? We can set up an advertisement campaign and *count* the number of additional people attending the farmer's market with the advertisement campaign relative to the farmer's market without the promotion.

3.1.2 Continuous random variables

Continuous random variables are ones that can take on an infinite number of outcomes. For example, a continuous random variable can be any value on the interval [0,1]. Examples include: farm yields, prices, birth weights.

Suppose that we ask the following question: How does attendance of a farmer's market affect the price of tomatoes sold at the market? Again, we can measure the number of people attending the farmer's market and the prices of tomatoes.

What is the outcome variable? Why is the outcome variable a *random variable*?

3.2 Depicting Random Variables

Because random variables cannot be predicted with certainty, it is useful to have a way to "describe" these variables. We do this by assigning *probabilities* to certain outcomes. That is, although no outcome is certain to occur every single time, some outcomes may be more probable than others.

Example 1: If you have a six-sided fair die, what is the probability of rolling a "one"? What about rolling a "two"?

3.2.1 Discrete random variables

Let's consider another example. Suppose you are measuring attendance of women at a farmer's market. A general way to indicate whether the next person at the farmer's market is a woman is:

$$P[Woman = Yes] = \phi$$
$$P[Woman = No] = 1 - \phi$$

where ϕ is some probability that a person at a farmer's market is a woman.

We can summarize this information using a *probability density function* (pdf). In the case of a discrete random variable, the pdf describes the probability that a random variable will take on a particular value. In notation:

$$\text{pdf} = f_Y(y) = P[Y = y]$$

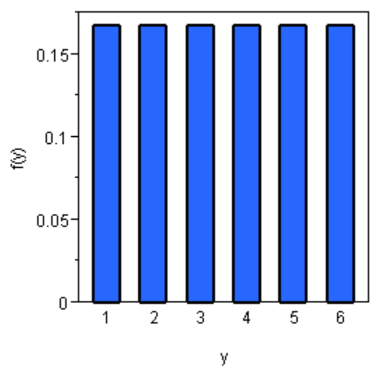
where Y is the random variable and y is the outcome that the random variable takes on. Two ways that a pdf can be described are with a table or a plot are presented in Figure 3.1.

Example: Calculate $P[Y \leq 3]$.

$$P[Y \leq 3] = P[Y = 1] + P[Y = 2] + P[Y = 3] = \frac{1}{2}$$

Figure 3.1: Discrete Probability Density Function

y	$f_Y(y)$
1	0.167
2	0.167
3	0.167
4	0.167
5	0.167
6	0.167



SAS Code: Simulate a Discrete Random Variable

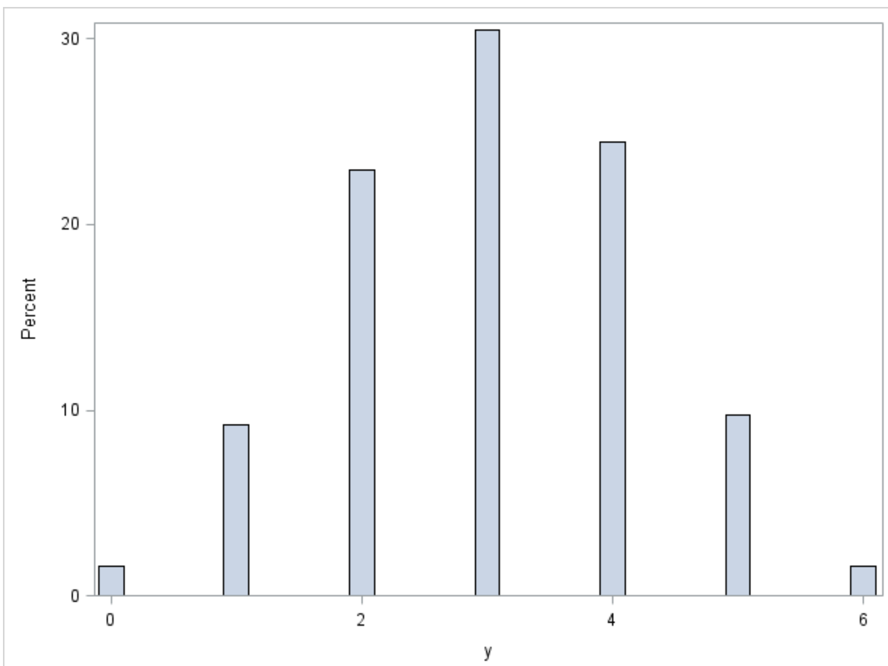
```
/* Generate data */
data random;
/* Set seed value to exactly replicate simulation */
call streaminit(12345);
do i = 1 to 10000;
    y = rand("Binomial",0.5,6);
    output;
end;
drop i;
run;

/* Summarize the random variable */
proc sgplot data=random;
histogram y;
run; quit;
```

The above SAS code generates random values from the discrete Binomial(0.5) distribution and plots a histogram of the outcomes. The `call streaminit` statement sets a simulation seed, which enables you to exactly replicate the simulated values in a data set each time that you execute the code. If no seed is set, then one is randomly generated by SAS, and the values in the data set "random" will be different in every execution. The `rand` function is used to specify the distribution name and parameters from which random draws are made.

The `SGPLOT` procedure is used to generate a histogram of the resulting simulated data set. Figure 3.2 shows the histogram, which demonstrates the probability of obtaining a particular outcome. The outcome $Y = 3$ is the most likely; that is, $P[Y = 3] = 31\%$.

Figure 3.2: Histogram of Values Simulated from the Binomial(0.5) Distribution



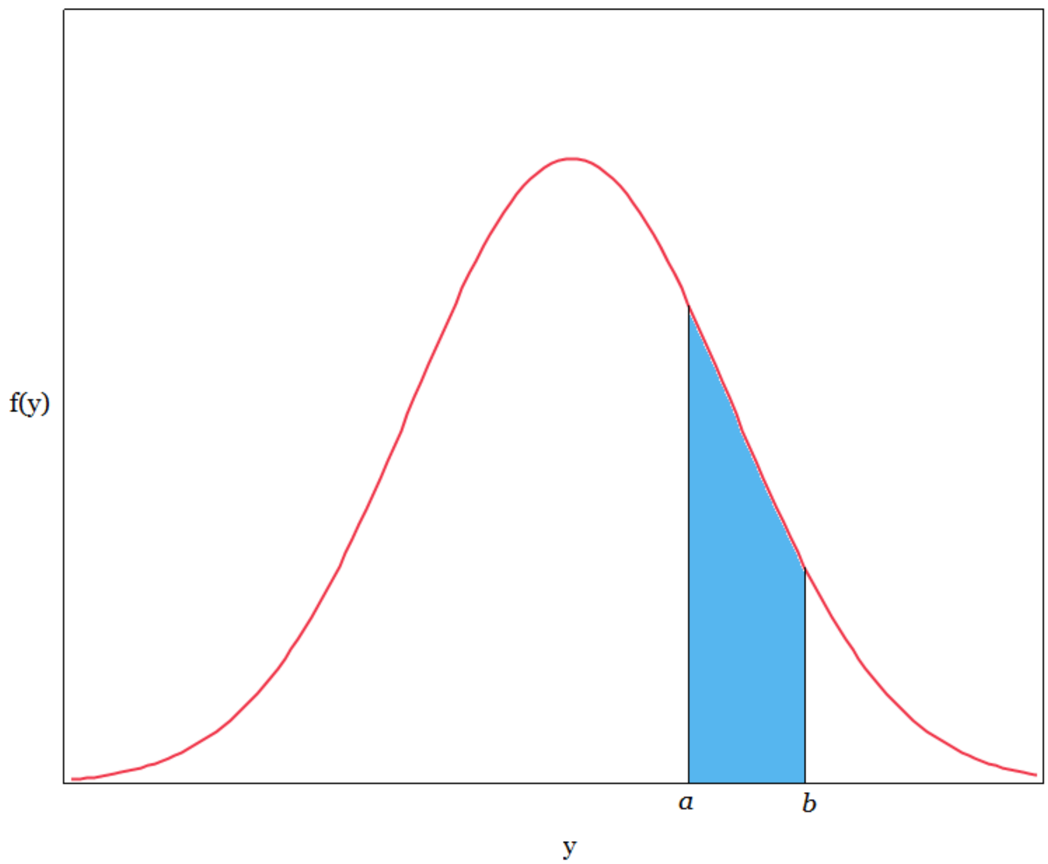
3.2.2 Continuous random variables

Summarizing continuous random variables is a bit less intuitive. Because a random variable Y can take on an unlimited set of values, the probability that Y takes on any unique value is zero. So, instead of attempting to identify the probability of Y attaining a particular value, we determine the probability that Y takes on a value within a range of values. In notation:

$$\text{pdf} = P[a \leq Y \leq b] = \int_a^b f_Y(y)\mathrm{d}y$$

The integration sign indicates that you are trying to find the area under the function $f_Y(y)$ between the points $y = a$ and $y = b$. This is shown in figure 3.3.

Figure 3.3: Continuous Probability Density Function



SAS Code: Simulate a Discrete Random Variable

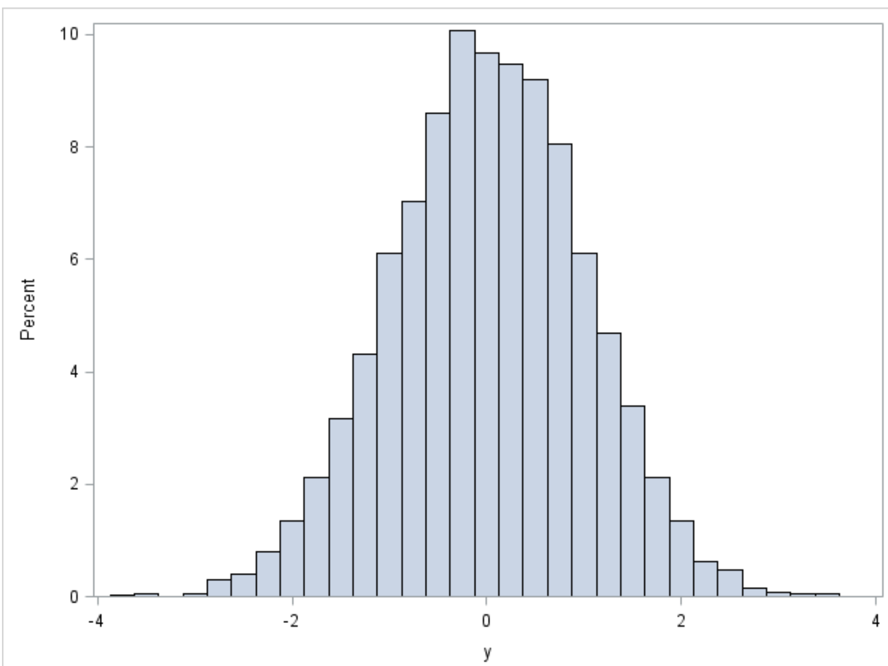
```
/* Generate random data from a Normal(0,1) distribution */
data random;
call streaminit(12345);
do i = 1 to 10000;
    y = rand("Normal",0,1);
    output;
```

```
end;
drop i;
run;

/* Summarize the random variable */
proc sgplot data=random;
histogram y;
run; quit;
```

The above SAS code generates random values from the continuous Normal distribution, $N(0, 1^2)$, and plots a histogram of the outcomes. The `SGPLOT` procedure is used to generate a histogram of the resulting simulated data set. Figure 3.4 shows the histogram, which demonstrates the probability of obtaining a particular outcome. The outcome $Y = 0$ is the most likely and corresponds to the mean of the Normal distribution from which values were drawn.

Figure 3.4: Histogram of Values Simulated from the Normal(0,1) Distribution



Properties of continuous probability density functions

1. The probability that an outcome Y is between some range of the distribution must be greater than zero. That is:

$$P[a \leq Y \leq b] = \int_a^b f_Y(y)\mathrm{d}y \geq 0$$

2. The total area under a pdf curve is equal to one. In other words, the probability that you will observe a value over the entire range of outcomes is 100%. That is:

$$P[-\infty \leq Y \leq \infty] = \int_{-\infty}^{\infty} f_Y(y)\mathrm{d}y = 1$$

3.3 Cumulative Density Function

When working with continuous random variables, we can use the cumulative density function (CDF) to describe the random variable. A CDF describes the probability that the random variable Y assumes a value that is less than or equal to y.

$$F_Y(y) \equiv P[Y \leq y]$$

3.3.1 Discrete random variables

In the case of discrete random variables, the CDF for some outcome y is the sum of all pdf values such that $\tilde{y} \leq y$. For example, consider the pdf table from above. The CDF for each value is as follows:

y	$f_Y(y)$	$F_Y(y) \equiv P[Y \leq y]$
1	0.167	0.167
2	0.167	0.334
3	0.167	0.501
4	0.167	0.668
5	0.167	0.835
6	0.167	1

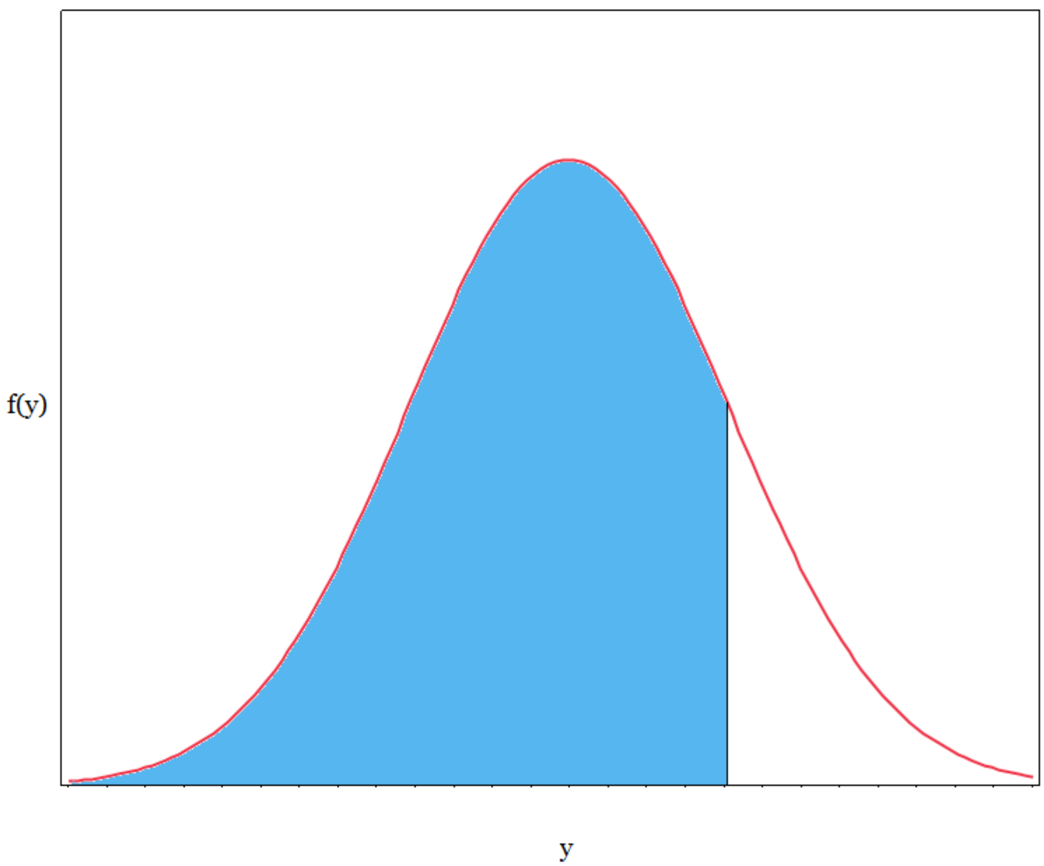
3.3.2 Continuous random variables

In the case of continuous random variables, the CDF is the area under the pdf to the left of the an outcome y. Consider the continuous pdf from above. The CDF can be calculated as follows:

$$\mathrm{CDF} = F_Y(y) = \int_{-\infty}^{y} f_Y(u)\mathrm{d}u$$

where $f_Y(u)$ is the pdf of Y. Graphically, this is shown in figure 3.5.

Figure 3.5: Continuous PDF



3.3.3 Properties of CDFs

1. $F_Y(y)$ is simply a probability. Thus, $0 \leq F_Y(y) \leq 1$.

2. The CDF is an increasing function of y. That is, if $y_1 \leq y_2$, then $F_Y(y_1) \leq F_Y(y_2) \equiv P[Y \leq y_1] \leq P[Y \leq y_2]$.

3. For any number c, it is equivalent to state that: $P[Y > c] = 1 - P[Y \leq c] \equiv 1 - F_Y(c)$.

4. For any numbers a and b such that $a \leq b$, the following holds:

$$P[a \leq Y \leq b] = F_Y(b) - F_Y(a)$$

3.4 Describing Random Variables

Probability and cumulative density functions provide a large amount of information about a random variable. However, at times we are simply interested in simpler descriptive statistics that can give us a feeling about what the random variable's distribution is without giving us the entire distribution. This is especially helpful when there are a large number of observations, which is typical of many data sets. There are three common descriptive statistics: central tendency, dispersion, and correlation/covariance.

3.4.1 Central tendency

The central tendency describes typical outcomes of a random variable. For example, if you look at the ages of 1,000 college students, you might find that the central tendency of age is 19.5 years of age. That is, if you asked a group of 1,000 college students their age, you will find that 19.5 years old is the answer that you will receive with highest probability.

The most common descriptor of central tendency is *expected value*. Expected value is a weighted average of all possible outcomes of a random variable Y. In other words, you weigh each possible outcome by the probability of that outcome occurring.

Discrete random variables

To calculate the central tendency for a discrete random variably Y, compute:

$$E(Y) = \sum_{n=1}^{N} (y_n \cdot P[Y = y_n]) \equiv \sum_{n=1}^{N} (y_n \cdot f_Y(y_n))$$

Example: Suppose that you have the grades from an undergraduate econometrics exam and the probability of each grade occurring. Compute the central tendency.

y = Score	$f_Y(y)$
100	0.2
95	0.2
90	0.2
88	0.2
76	0.2

What do we typically use as the expected value of a random variable? That is, if we know nothing about the probability of the occurrence of a certain outcome, how do we determine its central tendency?

Continuous random variables

The analogous calculation for determining expected values of a continuous variable involves an integration under the pdf:

$$E(Y) = \int_{-\infty}^{\infty} y \cdot f_Y(y)\mathrm{d}y$$

Example: Suppose the $Y \sim uniform(0,5)$. The pdf for a uniform distribution is: $f_Y(y) = \frac{1}{b-a}$ if $a < y < b$, and $f_Y(y) = 0$ otherwise. Calculate the expected value of Y.

$$E(Y) = \int_0^5 \left(y \cdot \frac{1}{5-0}\right) \mathrm{d}y = \frac{1}{5}\int_0^5 y \cdot \mathrm{d}y$$

$$\frac{1}{5} \cdot \left[\frac{y^2}{2}\right]_0^5 = \frac{1}{5} \cdot \frac{25}{2} = 2.5$$

Properties of expected values

1. The expected value of a constant c is the constant itself (e.g., if you get the same grade on every exam, what is your average across exams?)

$$E(c) = c$$

2. The expectation operator can be applied to each term in a linear function.

$$E(cY + d) = cE(Y) + d$$

More generally, if c_i are constants and Y_i are random variables, then:

$$E(c_1Y_1 + c_2Y_2 + \ldots + c_nY_n) = c_1E(Y_1) + c_2E(Y_2) + \ldots + c_nE(Y_n)$$

$$E\left(\sum_{i=1}^{n} c_iY_i\right) = \sum_{i=1}^{n} (c_iE(Y_i))$$

3. Jensen's inequality: for a non-linear function of a random variable $h(Y)$, you cannot assume that $E(h(Y)) = h(E(Y))$. That is, you cannot assume the expected value of the function of Y can be determined by the expected value of Y. Specifically, for a convex function of Y ($h''(Y) \geq 0$), Jensen's inequality holds:

$$E(h(Y)) \geq h(E(Y))$$

Alternatively, for a concave function, $E(h(Y)) \leq h(E(Y))$.

Example: consider the function of Y, $h = Y^2$. Suppose that you can have three outcomes (1, 2, 3), each with probability 0.33. Determine the expected value of h.

$$E(h(Y)) = E(Y^2) \neq (E(Y))^2$$

That is, you cannot approximate the expected value of Y^2 by squaring the expectation of Y.

$$E(Y^2) = \frac{1}{3} \cdot (1^2 + 2^2 + 3^2) = \frac{14}{3}$$

$$E(Y)^2 = \left(\frac{1}{3} \cdot (1 + 2 + 3)\right)^2 = 4$$

Clearly, $E(Y^2) \neq E(Y)^2$

3.5 Measures of Dispersion

The measure of dispersion for a random variable describes the *variance* (or the *spread*). Variance is a complement to the central tendency because knowing the central tendency and the dispersion, we can develop a good idea of the random variable's properties. Along with variance, we can use *standard deviation* as a descriptor of dispersion.

3.5.1 Variance

The variance of a distribution describes the expected measure of how far an outcome of Y is from the expected value of Y. In notation, variance is defined as:

$$\mathrm{Var}(Y) = \sigma_Y^2 = E([Y - E(Y)]^2)$$

In other words, variance is the weighted average of squared difference between outcomes of a random variable and the random variable's expected value. Squaring insures that positive differences do not eliminate negative differences. It should be noted that $Var(Y)$ is itself a random variable, because it depends on the random outcomes of Y.

Example: Show that $\sigma_Y^2 = E([Y - E(Y)]^2) = E(Y^2) - E(Y)^2$

$$\begin{array}{l} E([Y - E(Y)]^2) \\ E(Y^2 - 2YE(Y) + E(Y)^2) \\ E(Y^2) - 2E(YE(Y)) + E(Y)^2 \end{array}$$

Recall that $E(Y)$ is the expected value, which is a constant (typically the mean). Let's notate $E(Y) = \mu$.

$$\begin{array}{l} E(Y^2) - 2E(Y\mu) + \mu^2 \\ E(Y^2) - 2\mu E(Y) + \mu^2 \\ E(Y^2) - 2\mu \cdot \mu + \mu^2 \\ E(Y^2) - 2\mu^2 + \mu^2 \\ E(Y^2) - \mu^2 \\ E(Y^2) - E(Y)^2 \end{array}$$

Properties of variance

1. Variance is always nonnegative.
2. The variance of a constant is zero: $\text{Var}(c) = 0$. Remember, constants don't vary.
3. Adding a constant to a random variable does not change the variance. However, multiplying the random variable by a constant changes the dispersion, which also changes the variance of the random variable.

$$\text{Var}(cY + d) = c^2\text{Var}(Y)$$

Example: Show that $\text{Var}(cY + d) = c^2\text{Var}(Y)$.

(a) First, let's show that adding the constant does not change the variance. Let's specify a new random variable $Z = (Y + d)$.

$$\begin{aligned}
\text{Var}(Y + d) &= \text{Var}(Z) \\
\text{Var}(Z) &= E(Z^2) - E(Z)^2 \\
&= E((Y + d)^2) - [E(Y + d)]^2 \\
&= E(Y^2 + 2dY + d^2) - [E(Y) + d]^2 \\
&= E(Y^2) + 2dE(Y) + d^2 - E(Y)^2 - 2E(Y)d - d^2 \\
E(Y^2) - E(Y)^2 &= \text{Var}(Y)
\end{aligned}$$

(b) Next, let's show that multiplying by a constant changes the variance. Again, let's specify $Z = cY$.

$$\begin{aligned}
\text{Var}(cY) &= \text{Var}(Z) \\
\text{Var}(Z) &= E(Z^2) - E(Z)^2 \\
&= E(c^2Y^2) - [E(cY)]^2 \\
&= c^2E(Y^2) - [cE(Y)]^2 \\
&= c^2E(Y^2) - c^2E(Y)^2 \\
&= c^2[E(Y^2) - E(Y)^2] \\
&= c^2\text{Var}(Y)
\end{aligned}$$

3.5.2 Standard deviation

The second measure of dispersion is *standard deviation*. Standard deviation is simply the square root of the variance. We take the square root because it allows us to have a more intuitive measure of an outcome's average distance from the expected value. That is, deviations are expressed in the same units as the data, rather than squared units.

$$\mathrm{SD}(Y) = \sigma_Y = \sqrt{\sigma_Y^2}$$

Properties of standard deviation

1. The standard deviation of a constant is zero: $\mathrm{SD}(c) = 0$.
2. Adding a constant to a random variable does not change the standard deviation. However, multiplying the random variable by a constant changes the dispersion, which also changes the standard deviation of the random variable.

   $$\mathrm{SD}(cY + d) = |c| \cdot \mathrm{SD}(Y)$$

   When c is nonnegative, $\mathrm{SD}(cY) = c \cdot \mathrm{SD}(Y)$.

3.6 The Normal Distribution

Many random variables in empirical economics are continuous. There are numerous probability density functions that exist for describing these random variables. The one that is most often used is the *normal distribution*. A normally distributed random variable is one that is continuous and can take on any value. The normal distribution can also be called a *Gaussian distribution*, after the statistician C.F. Gauss. The distribution has a bell shape, as shown in figure 3.6.

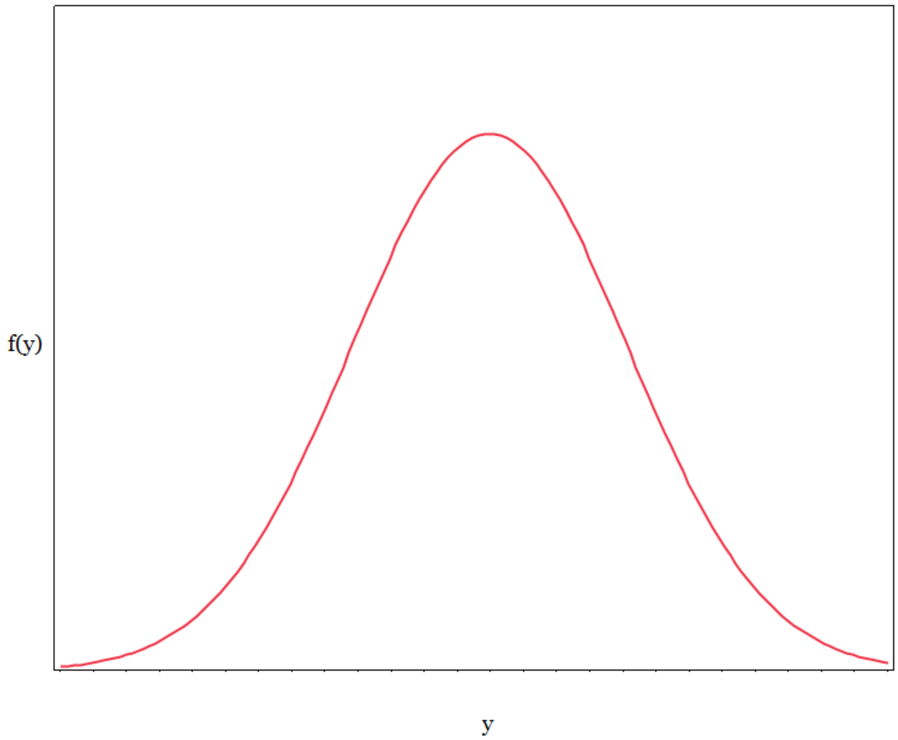
The pdf of a random variable Y is defined as follows:

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\left\{\frac{-(y-\mu_Y)^2}{2\sigma_Y^2}\right\}$$

To specify that a random variable Y is distributed normally, we use the notation: $Y \sim N(\mu_Y, \sigma_Y^2)$, where μ_Y is the mean and σ_Y^2 is the variance.

When a random variable is not normally distributed, it may be possible to transform the variable such that it's distribution becomes normal. For example, there are many cases in which taking the natural log of outcomes for a particular random variable will cause the logged distribution to be normal. This is called a *log-normal distribution*.

Figure 3.6: Normal Distribution



Properties of normal distributions

1. If $Y \sim N(\mu_Y, \sigma_Y^2)$, then $aY + c \sim N(a\mu_Y + c, a^2\sigma_Y^2)$
2. Any linear combination of independent, identically distributed (i.i.d.) normal random variables has a normal distribution.
3. The normal distribution is symmetric about its mean. That is, mean = median = mode.

SAS Code: Properties of Normal Distributions

```
/* Generate random data from a Normal(0,1) distribution */
data random;
call streaminit(12345);
do i = 1 to 1000;
    y = rand("Normal",0,1);
    y_t = 2*y + 4;
```

```
    output;
end;
drop i;
run;

/* Summarize the Normal random variable */
proc means data=random mean median var;
var y y_t;
run;quit;
```

The above SAS code generates random values from the Normal distribution, $N(0,1^2)$, which are used to create a transformed variable, $Y_T = 2Y + 4$. The `MEANS` procedure is used to genereate summary statistics for the original and transformed random variables. Table 3.1 presents these summary statistics, include the mean, median, and variance of both Y and Y_T. The pdf of both variables is symmetric about the mean, and the mean and variance of Y_T are 4 times larger than the mean and variance of Y.

Table 3.1: Summary Statistics of the Original and Transformed Random Variables

	$N(0,1)$	$N(2\mu_Y + 4, 2^2\sigma_Y^2)$
Mean	-0.011	3.979
Median	0.029	4.059
Var	1.102	4.075

3.6.1 Standard normal distribution

The standard normal distribution is a special case of the normal. If a variable is distributed with a standard normal, it is notated as: $Y \sim N(0,1)$. This implies that the mean of the random variable is zero and the variance (and standard deviation) is one. The pdf of a standard normal distribution typically denoted by ϕ and is as follows:

$$f_Y(y) = \phi_Y(y) = \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{-y^2}{2}\right\}$$

An important property associated with standard normal distributions is that for $Y \sim N(\mu_Y, \sigma_Y^2)$, then $\frac{(Y-\mu_Y)}{\sigma_Y} \sim N(0,1)$. That is, we can turn any normal distribution into a standard normal distribution by standardizing the random variable (i.e., subtracting its mean and dividing through by the standard deviation).

3.7 Joint Distributions

So far, we have looked only at the properties of a single random variable. Assuming that we can explain economic relationships with a single variable is similar to saying that incomes of economics professors are not related to any other variables. A more realistic and more interesting scenario would be to examine the relationship between two or more random variables.

For example, suppose that you wanted to examine exam scores for an undergraduate econometrics class. You have a pdf of the grades, and you also have the pdf of student ages. Suppose that you would like to determine the *joint probability* that a certain grade occurs and that the grade is earned by a student of a certain age.

For two discrete random variables Y and Z, the joint probability density function of (Y, Z) is as follows:

$$f_{Y,Z}(y, z) = P[Y = y, Z = z]$$

To determine the joint probability of an outcome, a simplifying assumption that we will make is that the random variable Y is *independent* from the random variable Z. That is, the outcome of Y does not change the probability of an outcome of Z, and vice versa. For example, consider rolling a fair die. There is always a 1/6 chance that you roll a six. Now consider that you want to determine the joint probability that out of 100 undergraduate and graduate college students, a graduate student is chosen and she rolls a six. The chances of choosing a graduate student does not change the probability that a six is rolled. That probability is still 1/6.

Using this assumption, we can define the joint probability density of (Y, Z) to be:

$$f_{Y,Z}(y, z) = f_Y(y)f_Z(z) = P[Y = y] \cdot P[Z = z]$$

Note that each individual pdf, $f_Y(y)$, is known as the *marginal distribution*.

Example: Suppose that you have the following information about die outcomes and students:

y = Die outcome	$f_Y(y)$	**z = Student**	$f_Z(z)$
1	0.167	Undergraduate	0.6
2	0.167	Masters	0.3
3	0.167	Ph.D.	0.1
4	0.167		
5	0.167		
6	0.167		

1. Determine the joint probability that a masters student rolls a 3.

$$
\begin{aligned}
f_{Y,Z}(3, M) &= f_Y(3) \cdot f_Z(M) = P[Y = 3] \cdot P[Z = M] \\
f_{Y,Z}(3, M) &= (0.167) \cdot (0.3) = 0.0501
\end{aligned}
$$

2. Construct a table of the joint pdf, $f_{Y,Z}(y, z)$.

   To construct the table, simply find the other joint probabilities:

				Z	
		z	**Undgrd**	**Masters**	**PhD**
		$f_Z(z)$	0.6	0.3	0.1
	y	$f_Y(y)$			
	1	0.167	0.1002	0.0501	0.0167
	2	0.167	0.1002	0.0501	0.0167
Y	3	0.167	0.1002	0.0501	0.0167
	4	0.167	0.1002	0.0501	0.0167
	5	0.167	0.1002	0.0501	0.0167
	6	0.167	0.1002	0.0501	0.0167

Note that adding across a particular row or column of joint probabilities yields the marginal probability for that row or column.

Additionally, the assumption of independence among random variables allows us to come up with a general notation for many random variables. That is, for a series of random variables $\{Y_1, Y_2, \ldots, Y_n\}$, the joint pdf under the independence assumption can be written as follows:

$$f_{Y_1,Y_2,\ldots,Y_n}(y_1, y_2, \ldots, y_n) = f_{Y_1}(y_1) f_{Y_2}(y_2) \cdots f_{Y_n}(y_n)$$

SAS Code: Joint Probability Density Simulation

```
/* Generate random data for two variables from Normal distributions */
data random;
call streaminit(12345);
do i = 1 to 1000;
    y = rand("Normal",0,1);
    z = rand("Normal",5,2);
    output;
end;
```

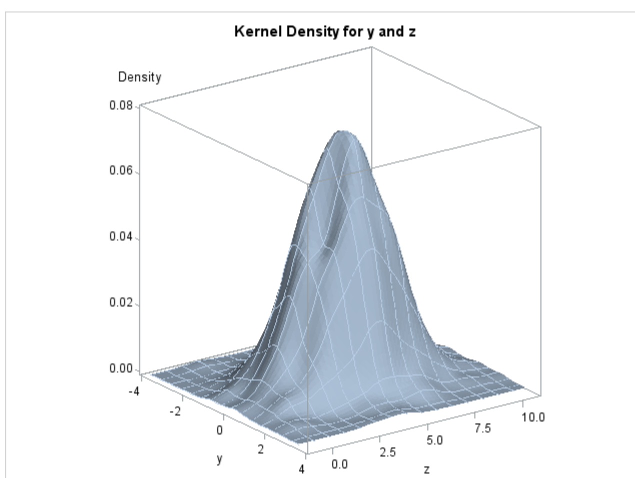
```
drop i;
run;

/* Calculate the joint probability of each outcome and plot the
   bivariate density */
proc kde data=random;
bivar y z / out=bivar plots=(contour surface);
run;quit;


/* Check that the highest joint probability; it should occur at the
   means of each marginal distribution */
proc means data=bivar max;
var density;
run;
```

The above SAS code generates data for two random variables, $y \sim N(0, 1^2)$ and $z \sim N(5, 2^2)$, and examines the joint probability properties of the variables. The `KDE` procedure is used to plot the joint probability density function, which is shown in Figure 3.7. The `out=bivar` statement creates a new SAS dataset, `bivar`, which contains the numerical representation of the joint pdf. The `MEANS` procedure is then used to output the value of the most likely joint outcome of the random variables y and z, which is 0.075.

Figure 3.7: Joint Probability Density Function Surface Plot for Variables y and z



3.7.1 Covariance

Covariance is an indicator of how, on average, a random variable varies with variation in another random variable. That is, covariance provides an indicator of the direction that a random variable's outcomes move as the outcomes of another random variable move. It is important to note that covariance *does not* reveal the magnitude of co-movements. This is because covariances depend on the units of each random variables. So scaling the units would also change the magnitude of the covariance, but would not change the direction.

Covariance between two random variables Y and Z is defined as follows:

$$\text{Cov}(Y,Z) = \sigma_{Y,Z} = E[(Y - E(Y))(Z - E(Z))]$$

$\sigma_{Y,Z} > 0$ implies that when $Y > E(Y)$, then Z tends be greater than $E(Z)$.

$\sigma_{Y,Z} < 0$ implies that when $Y < E(Y)$, then Z tends be greater than $E(Z)$.

Example: Show that $\sigma_{Y,Z} = E[(Y - E(Y))(Z - E(Z))] = E(YZ) - E(Y)E(Z)$.

$$\begin{aligned}
\sigma_{Y,Z} &= E[(Y - E(Y))(Z - E(Z))] \\
&= E[YZ - YE(Z) - ZE(Y) + E(Y)E(Z)] \\
&= E(YZ) - E(Y)E(E(Z)) - E(Z)E(E(Y)) + E(E(Y)E(Z))
\end{aligned}$$

Recall that we can specify $E(Y) = \mu_Y$ and $E(Z) = \mu_Z$ because these are constant means.

$$\begin{aligned}
&= E(YZ) - E(Y)E(E(Z)) - E(Z)E(E(Y)) + E(E(Y)E(Z)) \\
&= E(YZ) - \mu_Y\mu_Z - \mu_Z\mu_Y + E(\mu_Y\mu_Z) \\
&= E(YZ) - \mu_Y\mu_Z - \mu_Z\mu_Y + \mu_Y\mu_Z \\
&= E(YZ) - \mu_Y\mu_Z \\
&= E(YZ) - E(Y)E(Z)
\end{aligned}$$

Properties of covariances

1. Multiplying a random variable by a constant will change the covariance.

$$\text{Cov}(aY + c, bZ + d) = ab\text{Cov}(Y,Z)$$

2. Variance of a sum of random variables is as follows:

$$\text{Var}(aY + bZ) = a^2\sigma_Y^2 + b^2\sigma_Z^2 + 2ab\sigma_{Y,Z}$$

   It is important to note that if Y and Z are independent, then $\text{Cov}(Y,Z) = 0$. However, the converse is not true. That is, if $\text{Cov}(Y,Z) = 0$, then the independence of Y and Z should not be inferred.

3.7.2 Correlation

An important downside to measuring directional relationships between random variables using covariance is that the covariance is dependent on the units of each random variable. Measuring the relationship between planted area and yields depends on whether acres are measured in acres or hectares.

A unit-less measure used to examine the relationship between variables is *correlation*. Correlation can reveal the strength of the linear relationship between random variables, because it is unit-less. The correlation coefficient is defined as follows:

$$\text{Corr}(Y,Z) = \rho_{Y,Z} = \frac{\text{Cov}(Y,Z)}{\text{SD}(Y)\text{SD}(Z)} = \frac{\sigma_{Y,Z}}{\sigma_Y \sigma_Z}$$

Properties of correlation

1. $\rho_{Y,Z} \in [-1, 1]$
2. If $\rho_{Y,Z} = 0$, then the two random variables are not correlated. (Note that a correlation of zero *does not* imply independence).
3. For any constants a, b, c and d:

$$\begin{aligned} \text{Corr}(aY + c, bZ + d) &= \text{Corr}(Y,Z), \text{ if } a \cdot b > 0 \\ \text{Corr}(aY + c, bZ + d) &= -\text{Corr}(Y,Z), \text{ if } a \cdot b < 0 \end{aligned}$$

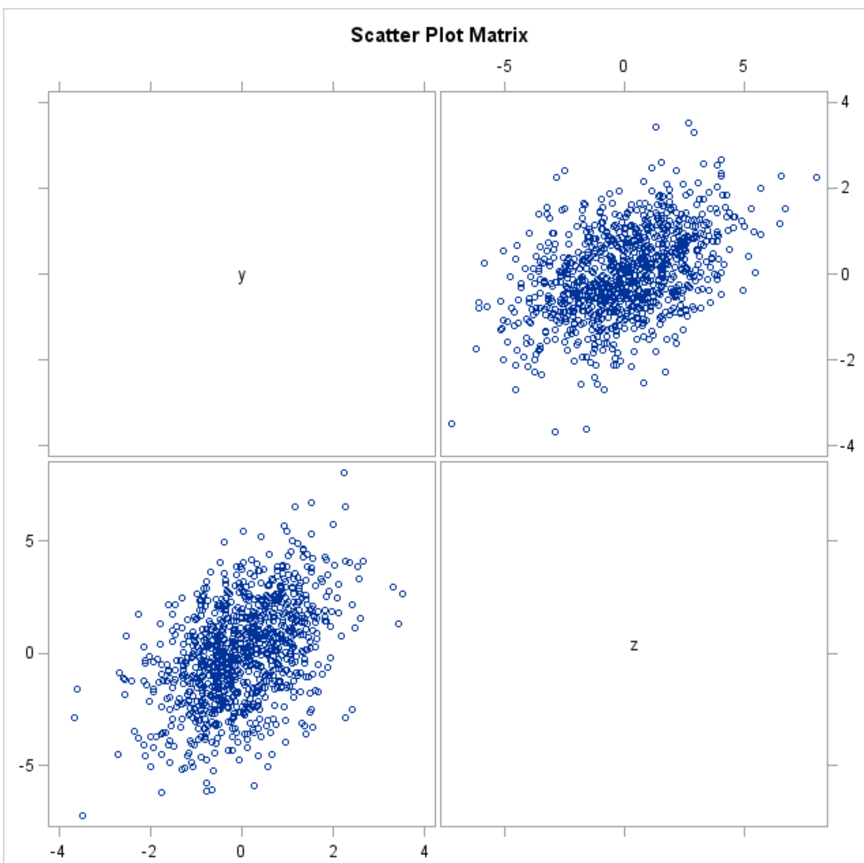
SAS Code: Correlation Simulation

```
/* Generate data for two correlated variables */
data random;
call streaminit(12345);
do i = 1 to 1000;
    y = rand("Normal",0,1);
    z = y + 2*rand("Normal",0,1);
    output;
end;
drop i;
run;
```

```
/* Calculate correlation and scatter plots; option "cov"
   produces covariance matrix */
proc corr data=random cov plots=matrix;
run;
```

The above SAS code illustrates the data generating process of two correlated variables, y and z. Specifically, the random variable z is related positively to variable y, because the value of y enters into the determination of the value of z. The `CORR` procedure calculates and displays the Pearson correlation matrix of the two variables. The `cov` option requests that the covariance matrix is also presented and the `plots=matrix` option requests that a matrix of correlation plots among all variable combinations is also displayed. Figure 3.8 shows the correlation plots.

Figure 3.8: Matrix Correlation Plots Between Variables y and z



SAS Code: Illustrating Changes in the Correlation Structure

```
/* Generate data for two correlated variables */
data random;
call streaminit(12345);
do i = 1 to 1000;
    y = rand("Normal",0,1);
    z = y + 2*rand("Normal",0,1);
    output;
end;
drop i;
run;

proc iml;

/* A function for calculating correlation */
start Correl(A);
   n = nrow(A);
   C = A - A[:,];
   cov = (C` * C) / (n-1);
   print cov;
   stdCol = sqrt(C[##,] / (n-1));
   print stdcol;
   stdC = C / stdCol;
   return( (stdC` * stdC) / (n-1) );
finish;

/* Variable names to be displayed*/
varNames = {"y" "z"};

use random;
read all into rand;
/* Call the correlation function and output result into "corr1" */
corr1 = Correl(rand);
print corr1[c=varnames r=varnames];

/* (1) Variables move in opposite directions */
yz1 = {-5 5};
rand = rand // yz1;
corr2 = Correl(rand);
print corr2[c=varnames r=varnames];
```

```
/* Numerator (Cov) is going down and denomenator (SD) is
   going up --> corr goes down */

/* (2) Variables move in opposite directions */
yz2 = {-10 10};
rand = rand // yz2;
corr3 = Correl(rand);
print corr3[c=varnames r=varnames];
/* Numerator (Cov) is going down and denomenator (SD) is
   going up --> corr goes down */

/* (3) Variables start moving in same direction */
yz3 = {10 10};
rand = rand // yz3;
corr4 = Correl(rand);
print corr4[c=varnames r=varnames];
/* Numerator (Cov) is going up and denomenator (SD) is going up,
   but not as fast as the covariance is increasing --> corr goes up */

quit;
```

The above SAS code demonstrates the manner in which correlations among variables can change under varying conditions. The `Correl` function is defined for calculating correlations between two variables. First, a correlation is calculated for the generated data, indicating that the variables are positively correlated. Next, an additional observation is added to the dataset, which represents a scenario in which the variables move in opposite directions (i.e., the value of y is positive but the value of z is negative). A re-calculaton of the correlation coefficient indicates that its value is dampened, because the covariance has decreased while the standard deviation of each variable increased.

In the second case, another set of observations is added to the dataset, with the values being even futher apart in opposite directions. The outcome is a further decrease of the correlation coefficient. The last set of additional observations are values that are positively related (i.e., both y and z increased), but by a large amount. This creates two opposing effects. Although the magnitude of the covariance between y and z is now larger and positive, the standard deviations of both variables is also larger, so the impact on the correlation coefficient is ambiguous. However, the correlation coefficient has increased, implying that the covariance increase was larger than the increase in the standard deviations.

Table 3.1 presents the values of the covariance, standard deviation, and correlation in each scenario.

Table 3.1: Changes in the Covariance, Standard Deviation and Correlations

	$Cov_{1,2}$	$Cov_{2,1}$	σ_y	σ_z	$\rho_{y,z}$
Original Data	0.99	4.85	1.00	2.20	0.45
Perturbation 1	1.02	4.87	1.00	2.20	0.44
Perturbation 2	1.12	4.97	1.06	2.23	0.37
Perturbation 3	1.21	5.06	1.10	2.25	0.39

3.8 Practice with Computing Summary Statistics

Recall the example from above in we were interested in the effect of advertisements on attendance at a farmers' markets. Suppose that you observe the number of posters/fliers in area advertising the farmer's market as well as the attendance. Additionally, you are provided with the information about associated probabilities.

y = Attendance	$f_Y(y)$	**z = Posters**	$f_Z(z)$
50	0.025	20	0.45
60	0.175	35	0.2
45	0.2	10	0.1
75	0.35	25	0.25
80	0.2		
120	0.05		

Calculate the following:

1. Expected values of Y and Z.
2. Variances of Y and Z.
3. Correlation between Y and Z.

3.9 Conditional Probabilities

Correlation and covariances explain the co-movement relationships between two random variables. However, we are often interested in observing the behavior of one random variable *conditional* on an outcome of one or more other random variables. That is, we would like to know the probability distribution of Y conditional on the knowledge that Z has attained a particular outcome. For a discrete random variance, this is denoted as follows:

$$f_{Y|Z}(y|z) = P[Y = y|Z = z] = \frac{f_{Y,Z}(y,z)}{g_Z(z)}$$

The second equality tells us that a conditional distribution can be interpreted as the ratio of the joint probability between Y and Z and the marginal distribution of Z.

Specifying continuous conditional distributions is similar:

$$f_{Y|Z}(y|z) = \frac{f_{Y,Z}(y,z)}{g_Z(z)} = \frac{\int\int f_{Y,Z}(y,z)dydz}{\int f_Z(z)}$$

Example: suppose that you are interested in determining the probability of higher attendance at farmers' markets given that there is advertisement of a market. That is, we would like to know the distribution of attendance outcomes conditional on the number of advertisement fliers posted around town.

$$f_{At|Ad}(at|ad) = P[At = at|Ad = ad]$$

Consider the following pdf that describes the number of extra people attending farmers' markets and number of posted advertisements:

At = Attendance	$f_{At}(at)$	**Ad = Ads**	$f_{Ad}(ad)$
50	0.025	20	0.45
60	0.175	35	0.2
45	0.2	10	0.1
75	0.35	25	0.25
80	0.2		
120	0.05		

We can use our knowledge of joint probabilities to calculate the joint pdf for these two random variables. Let's suppose that we are interested in the probability that there will be 60 extra people conditional on 20 additional fliers. Assuming independence:

$$P[At = 60, Ad = 20] = P[At = 60] \cdot P[Ad = 20] = (0.175 \times 0.45) = 0.0785$$

Now we can calculate the conditional probability by dividing the joint probability by the probability that the number of posted fliers will is 20. That is:

$$f_{At|Ad}(at|ad) = \frac{P[At = 60, Ad = 20]}{P[Ad = 20]} = \frac{0.0785}{0.45} = 0.175$$

Note that the conditional probability of 60 additional people is the same as the unconditional probability of this higher attendance. This only holds if two random variables are independent.

3.9.1 Conditional expectations

As with unconditional random variables, we would like to find a way to describe conditional random variables. That is, we may be interested in finding the expected values of a random variable. For example, suppose that we want to know the central tendency of yields given a certain level of fertilizer. Or, we would like to know the most likely number of wins for a baseball team conditional on the total player payroll of that team.

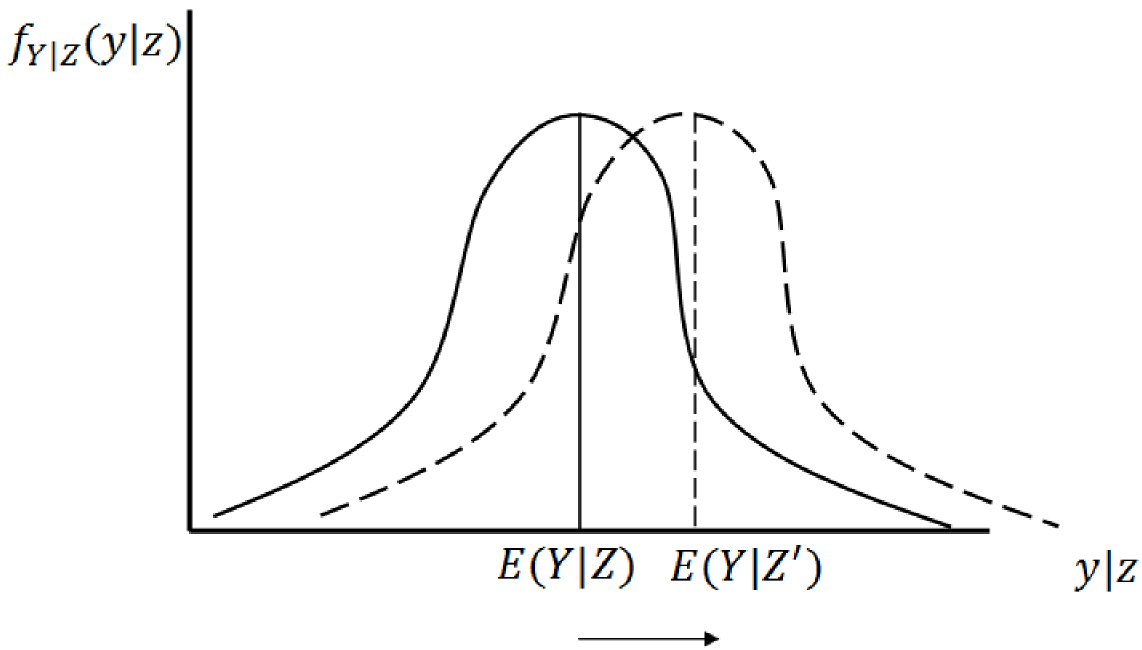
This *conditional expectation* is also often known as the *conditional mean*, and an integral concept for discussing least squares regressions. Another intuitive way of thinking about the conditional mean asking the question: how does the mean level of Y change with changes in Z? Figure 3.9 illustrates a *location* shift the conditional mean of Y given a change in Z.

For discrete random variables, we can formalize the conditional mean to be as follows:

$$E(Y|Z = z) = \sum_{i=1}^{n} y_i f_{Y|Z}(y_i|z) = \sum_{i=1}^{n} y_i \frac{f_{Y,Z}(y_i, z)}{f_Z(z)}$$

This shows that the expected value of Y is still a weighted average (as the case for unconditional expectations), but now the weights depend on the outcome of the random variable Z.

Figure 3.9: Location Shift of the Conditional Distribution



Similarly, the continuous case is as follows:

$$E(Y|Z=z) = \int_{-\infty}^{\infty} y \cdot f_{Y|Z}(y|z)dy = \int_{-\infty}^{\infty} y \cdot \left(\frac{f_{Y,Z}(y_i, z)}{f_Z(z)}\right) dy$$

Example: suppose that you find that attendance at farmers' markets and advertisements are not independent (for example, organizers of the farmers market may put up more fliers if less new people come). After carefully searching information sources, you find the following joint probability function of attendance and advertisement:

		ad	10	20	30
		$f_{Ad}(ad)$	0.45	0.35	0.2
at	$f_{At}(at)$				
50	0.65		0.2925	0.2275	0.13
60	0.25		0.1125	0.0875	0.05
75	0.1		0.045	0.035	0.02

You're interested in determining the expected value of additional attendance given that there are 20 posted advertisements. That is, $E(At|Ad = 20)$. This is as follows:

$$
\begin{aligned}
E(At|Ad = 20) &= \textstyle\sum_{i=1}^{n} at_i \left(\frac{f_{At,Ad}(at_i, ad=1)}{f_{Ad}(ad=1)}\right) \\
&= \frac{(50\times 0.2275)+(60\times 0.0875)+(75\times 0.02)}{0.35} \\
&= \frac{17.25}{0.35} \\
&\approx 52
\end{aligned}
$$

That is, if there are 20 fliers posted for advertisements, you are expected to attract approximately 52 additional farmers' market patrons.

3.9.2 Properties of conditional expectations

1. The conditional expectation of a function of Y is just the function of Y:

$$E(h(Y)|Y) = h(Y)$$

2. For a linear function with several random variables, applying the conditional expectation operator of Y only affects random variables other than Y:

$$E([h(Y)Z + k(Y)]|Y) = h(Y)E(Z|Y) + k(Y)$$

3. If two random variables are independent, then the conditional expectation of one random variable given another is the same as the unconditional expectation:

$$E(Y|Z) = E(Y)$$

4. Law of Iterated Expectations – the unconditional expected value of Y is the expectation of its conditional expectations:

$$E[E(Y|Z)|Z] = E(Y)$$

Proof:

$$
\begin{aligned}
E[E(Y|Z)|Z] &= \int_{-\infty}^{\infty} E(Y|Z) f_Z(z)dz \\
&= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} y \cdot f_{Y|Z}(y|z)dy \right\} f_Z(z)dz \\
&= \int\int y \cdot f_{Y|Z}(y|z) f_Z(z)dydz
\end{aligned}
$$

Note that in the last step, the term $f_Z(z)$ was carried inside of the dy integral because it is independent of dy. Next, note that $f_{Y|Z}(y|z)f_Z(z) = f_{Y,Z}(y,z)$ because recall that by definition, $f_{Y|Z}(y|z) = f(y,z)/f_Z(z)$. To continue:

$$
\begin{aligned}
&= \int\int y \cdot f_{Y,Z}(y,z)dydz \\
&= \int y \left(\int f_{Y,Z}(y,z)dz\right) dy \\
&= \int y \cdot f_Y(y)dy \\
&= E[Y]
\end{aligned}
$$

5. Law of Iterated Expectations – expanded form:

$$E[E(Y|Z,W)|Z] = E(Y|Z)$$

Example: suppose we want to determine corn yields based on observable characteristics Z (e.g., amount of fertilizer, farmer's experience and education, etc) and unobserved characteristics W such as ability of the farmer. We can't observe the farmer's ability level, but we can retrieve its expected value provided observable variables. That is:

$$
\begin{aligned}
E(Y|Z) &= E[E(Y|Z,W)|Z] \\
&= E[(\beta_1 z_1 + \beta_2 z_2 \ldots + \omega W)|Z] \\
&= \beta_1 z_1 + \beta_2 z_2 \ldots + \omega E(W|Z) \\
&= \beta_1 z_1 + \beta_2 z_2 \ldots + \omega(\phi_1 z_1 + \phi_2 z_2 \ldots) \\
&= \beta_1 z_1 + \beta_2 z_2 \ldots + \omega\phi_1 z_1 + \omega\phi_2 z_2 \ldots \\
&= (\beta_1 + \omega\phi_1)z_1 + (\beta_2 + \omega\phi_2)z_2 + \ldots \\
&= \eta_1 z_1 + \eta_2 z_2 \ldots \eta_n z_n
\end{aligned}
$$

Thank you for your interest in my notes. For record keeping and intellectual property purposes, I have elected to provide the full text to individuals only by request. If you would like to obtain a copy of these notes, I would be glad to share them, but I ask that you send an email request to anton.bekkerman@montana.edu.