AFRE 835: Introductory Econometrics

Appendix 1A: Background Material

Spring 2017

Outline

Appendix A: Basic Mathematical Tools

- This appendix reviews some basic mathematical material regarding:
  - The properties of the summation operator;
  - Some basic descriptive statistics using the summation operators;
  - Linear Functions;
  - Proportions and Percentages;
  - Quadratic, Logarithmic and Exponential Functions;
  - Some Basic Calculus
- You should be sure that you are comfortable with this material.

Appendix B: Fundamentals of Probability

- This appendix covers key concepts from basic probability, beginning with the notion of random variables and their probability distributions.
- Wooldridge uses the convention of letting capital letters (X) denote random variables and lower case letters (x) denote their realizations.
- Random variables are typically divided into two categories:
  1. *Discrete random variables*: take on only a finite or countably infinite number of values ($x_j, j = 1, \ldots, k$).
     - *Bernoulli (or binary) random variables* are a special case in which X takes on only two possible values $X = 1$ ("success") or $X = 0$ ("failure").
  2. *Continuous random variables* take on any real value with zero probability.
- Combined discrete/continuous random variables do exist, but we will ignore this complication for most of this class.

Discrete Random Variables

- The *probability density function (pdf)* of X summarizes the information concerning the possible outcomes of X and the corresponding probabilities.

$$P(X = x) = f(x) = \begin{cases} p_j & x \in \{x_1, \ldots, x_k\} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

- If we have more than one random variable (say X and Y), the corresponding pdf's are denoted by f_X and f_Y.

Continuous Random Variables

- With a continuous random variable, the probability of any given outcome has zero probability.
- Instead, we characterize the probability that the random variable falls within a given set of possible outcomes, typically a range of values; e.g., $P(a \leq X \leq b)$ for $a < b$.
- We characterize probabilities for a continuous random variable using the *cumulative distribution function (cdf)* $F(x)$ such that:

$$F(x) = P(X \leq x) \tag{2}$$

- Two important properities of cdf's are

$$P(X > c) = 1 - F(c) \quad \forall c \tag{3}$$

$$P(a \leq X \leq b) = F(b) - F(a) \quad \forall a < b \tag{4}$$

- There is a corresponding *probability density function* (pdf) $f(x)$ where

$$F(x) = \int_{-\infty}^{x} f(s)ds \tag{5}$$

Joint and Conditional Distributions and Independence

- We are often interested in how two or more random variables are related; e.g.,
  - snow fall amounts and highway fatalities;
  - the wage an individual earns and their education level;
  - the amount of electricity a household uses and the appliances they own;
  - the amount of fertilizer a farmer applies and algae levels in local waterways.
- The outcomes for multiple random variables are characterized in term of their *joint probability distributions*.
- For two discrete random variables, we have the *joint pdf*:

$$f_{X,Y}(x,y) = P(X = x, Y = y) \tag{6}$$

- Two random variables are *independent* if, and only if, their joint pdf is the product of their marginal pdf's; i.e., $f_{X,Y}(x,y) = f_X(x)f_Y(y)$
- The *conditional distribution* of Y given X is given by $f_{Y|X} = f_{X,Y}(x,y)/f_X(x)$

Features of Probability Distributions

- Central Tendency: Expected Value
  - A key attribute of a random variable is its expected value $E(X)$.
    - For a discrete random variable

$$E(X) = x_1 P(X = x_1) + \cdots + x_k P(X = x_k) = \sum_{j=1}^{k} x_j P(X = x_j) = \sum_{j=1}^{k} x_j f(x_j) \tag{7}$$

    - For a continuous random variable

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx \tag{8}$$

  - For functions of random variables (rv's) (e.g., $g(x)$), we have

$$E[g(x)] = \begin{cases} \sum_{j=1}^{k} g(x_j) f(x_j) & \text{for discrete rv's} \\ \int_{-\infty}^{\infty} g(x) f(x) dx & \text{for continuous rv's} \end{cases} \tag{9}$$

Features of Probability Distributions (cont'd)

- Properties of Expected Values
  1. For any constant c, $E(c) = c$;
  2. For any constants a and b, $E(aX + b) = aE(X) + b$
  3. More generally, for any constants a_j and rv's X_j ($j = 1, \ldots, k$), $E\left(\sum_{j=1}^{k} a_j X_j\right) = \sum_{j=1}^{k} a_j E(X_j)$
- Measures of Variability: Variance and Standard Deviation
  - Variance: $Var(X) \equiv E[(X - \mu_X)^2] = E(X^2) - \mu_X^2$, where $\mu_X = E(X)$
  - Standard Deviation: $sd(X) \equiv +\sqrt{Var(X)}$
  - Section B.3 lists a number of properties of variances and standard deviations.
- Measures of Association: Covariance and Correlation
  - Covariance: $Cov(X, Y) \equiv E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X \mu_Y$
  - Correlation Coefficient: $\frac{Cov(X,Y)}{Var(X)Var(Y)} \in [-1, 1]$

Features of Probability Distributions (cont'd)

- We are often interesting in understanding how two random variables depend upon each other.
- One way to summarize this relationship is in terms of the *conditional expectation* of, say, Y given X.
- In the case of discrete random variables, this conditional mean takes the form

$$E(Y|X = x) = \sum_{j=1}^{k} y_j f_{Y|X}(y_j|x) \tag{10}$$

- Section B.3 lists several properties of conditional means
- Two particularly useful ones are:
  1. If X and Y are independent, then $E(Y|X) = E(Y)$.
  2. The *Law of Iterated Expectations*: $E_X[E(Y|X)] = E(Y)$

The Normal and Related Distributions

- The normal distribution is particularly useful.
- In general, if $X \sim Normal(\mu, \sigma^2)$, then $E(X) = \mu$ and $Var(X) = \sigma^2$.
- The distribution is symmetric, with pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-(x-\mu)^2/2\sigma^2\right] \quad -\infty < x < \infty \tag{11}$$

- The standard normal corresponds to

$$Z = \frac{X-\mu}{\sigma} \sim Normal(0,1) \tag{12}$$

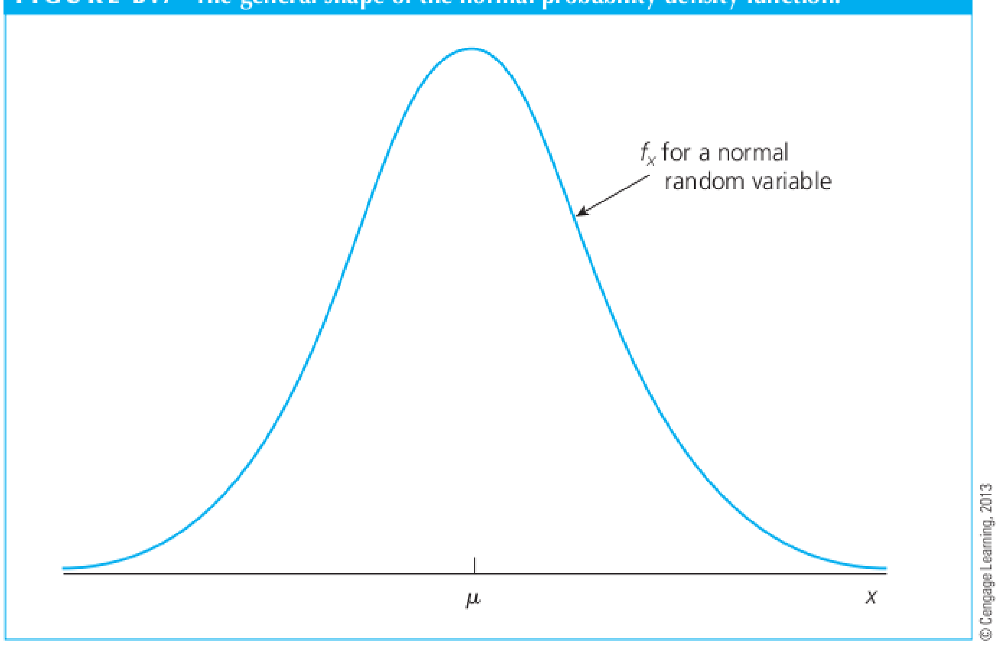
with pdf $\phi(z)$, where

$$\phi(z) = \frac{1}{\sqrt{2\pi}} exp\left[-z^2/2\right] \quad -\infty < z < \infty \tag{13}$$

- The corresponding cdf is denoted by $\Phi(z)$,

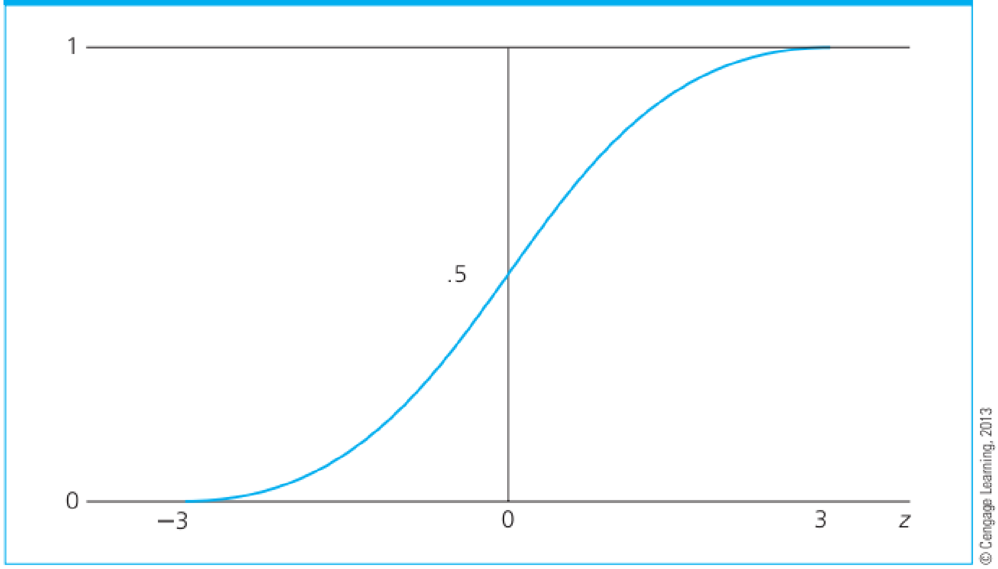
Normal pdf

FIGURE B.7 The general shape of the normal probability density function.



Standard Normal cdf, $\Phi(z)$

FIGURE B.8 The standard normal cumulative distribution function.



Properties of the Normal Distribution

- If $X \sim Normal(\mu, \sigma^2)$, then $Y = aX + b \sim Normal(a\mu + b, a^2\sigma^2)$
- If X and Y are jointly normally distributed, then they are independent if, and only if, $Cov(X, Y) = 0$.
- Any linear combination of independent, identically distribution (*iid*) normal random variables has a normal distribution.

Related Distributions

- *Chi-Square Distribution*: Let $Z_i, i = 1, \ldots, n$ be *iid* standard normal random variables. Then
$$X = \sum_{i=1}^{n} Z_i^2 \sim \chi_n^2 \tag{14}$$
is distributed chi-squared with n degress of freedom.
- *t-Distribution*: If $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi_n^2$ and Z and X are independent, then
$$T = \frac{Z}{\sqrt{X/n}} \tag{15}$$
has a t-distribution with n degrees of freedom.
- *F Distribution*: If $X_1 \sim \chi_{k_1}^2$ and $X_2 \sim \chi_{k_2}^2$, and the two random variables are independent, then
$$F = \frac{X_1/k_1}{X_2/k_2} \tag{16}$$
has an F distribution with (k_1, k_2) degrees of freedom.

Statistical Inference

- *Statistical Inference* involves learning something about a population given the availability of a sample from that population.
  - a sample is needed because it is usually too costly to obtain information about the entire population.
- There are several key issues in this definition:
  1. What is our population of interest?
  2. How will we obtain a sample?
     - A *random* sample is often convenient.
- Definition: If $Y_1, Y_2, \ldots, Y_n$ are independent random variables with a common probability density function $f(y;\theta)$, then $\{Y_1, Y_2, \ldots, Y_n\}$ is said to be a random sample from $f(y;\theta)$.
- The realization of a random sample would be denotes by $\{y_1, y_2, \ldots, y_n\}$

Estimators and Estimates

- We often know, or are willing to assume, a specific type of distribution for a random variable X, say $f(x;\theta)$, but do not know the specific values of the parameters of that distribution (i.e., θ).
- *Estimation* is the procedure used to learn about these unknown values from an available sample.
- More specifically, an *estimator* of, say, θ is a rule that assigns to each possible outcome of the sample a value of θ.
  - It is key to understand that the rule is specified before any sampling is carried out; i.e., it is *not* dependent upon the sample.
- Example: One estimator of the population mean, μ, is the sample average, given by

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{17}$$

- The corresponding *estimate* is $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Properties of Estimators

- In general, since an *estimator* is a function random variables, it is also a random variable;

$$W = h(Y_1, \ldots, Y_n). \tag{18}$$

- The distribution of W is known as its sampling distribution.
- We are often interested in some key properties of an estimator.
- An estimator, W of θ, is *unbiased* if $E(W) = \theta$.
- More generally, the *bias* of an estimator of θ is given by $Bias(W) = E(W) - \theta$.
- It is easy to show that the sample mean $\bar{X}$ is an unbiased estimator of the population mean $E(X) = \mu_X$.
- One can also show that the sample variance

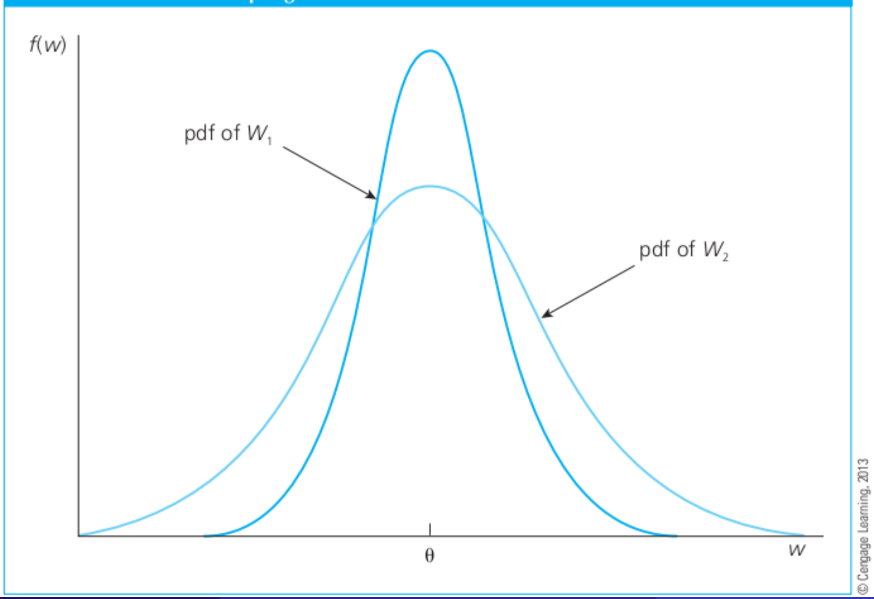
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{19}$$

is an unbiased estimator of the population variance $Var(X) = \sigma^2$.

Additional Properties of Estimators - Efficiency

- *Relative Efficiency*: If W_1 and W_2 are two unbiased estimators of θ, then W_1 is efficient relative to W_2 if $Var(W_1) \leq Var(W_2)$ for all θ, with strict inequality for at least one θ.

FIGURE C.2 The sampling distributions of two unbiased estimators of θ.



Additional Properties of Estimators - MSE and Consistency

- Sometimes we are faced with a trade-off between bias and efficiency.
  - A typical metric in this case is *mean square error*:

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = var(\hat{\theta}) + [bias(\hat{\theta})]^2 \tag{20}$$

- Another desirable property of an estimator is *consistency*:
  - An estimator $\hat{\theta}$ is said to be consistent if it approaches the true value as the sample size increases; denote as $\hat{\theta} \xrightarrow{p} \theta$
  - More formally, $\hat{\theta}$ is said to be consistent if for every $\epsilon > 0$,

$$\lim_{n \to \infty} P(|\hat{\theta} - \theta| < \epsilon) = 1 \tag{21}$$

  - This is often written in shorthand as $\text{plim}(\hat{\theta}) = \theta$.

Some Useful Large Sample (Asymptotic) Properties of Estimators

- If $\text{plim}(\hat{\theta}) = \theta$ and $g(\theta)$ is a continuous function of θ, then $\text{plim}[g(\hat{\theta})] = g(\theta)$
- $\text{plim}(b) = b$ for any constant b.
- If $\text{plim}(\hat{\theta}_1) = \theta_1$ and $\text{plim}(\hat{\theta}_2) = \theta_2$, then
  - If $\text{plim}(\hat{\theta}_1 + \hat{\theta}_2) = \theta_1 + \theta_2$
  - If $\text{plim}(\hat{\theta}_1\hat{\theta}_2) = \theta_1\theta_2$
  - If $\text{plim}(\frac{\hat{\theta}_1}{\hat{\theta}_2}) = \frac{\theta_1}{\theta_2}$
- Other related asymptotic concepts include asymptotic efficiency and asymptotic normality.

Interval Estimation and Confidence Intervals

- While point estimates are useful, they say nothing directly about how certain we are that the estimate approximates the population parameter of interest.
- The sampling standard deviation of an estimator $(S_n = \sqrt{S_n^2})$ provides on measure of assessing the uncertainty of that estimator.
- Perhaps more useful is the corresponding *confidence interval*.
- Example #1: Suppose we know $X \sim \mathcal{N}(\mu, 1)$.
  - Then $\bar{X} \sim \mathcal{N}(\mu, \frac{1}{n})$
  - Normalizing, $\frac{\bar{X}-\mu}{1/\sqrt{n}} \sim \mathcal{N}(0,1)$
  - This implies that

$$\begin{aligned} 0.95 &= P\left(-1.96 < \frac{\bar{X}-\mu}{1/\sqrt{n}} < 1.96\right) \\ &= P\left(\bar{X} - 1.96/\sqrt{n} < \mu < \bar{X} + 1.96/\sqrt{n}\right) \end{aligned} \tag{22}$$

  - The corresponding *95%-confidence interval estimate* is given by $\left[\bar{x} - 1.96/\sqrt{n}, \bar{x} + 1.96/\sqrt{n}\right]$

Interpreting the Confidence Interval

- It is incorrect to say that "...there is a 95 percent probability that the true value of μ falls in the estimated confidence interval.
- What is random is the confidence interval estimator, not the confidence interval estimate nor the true value of μ.
- Rather, one wants to say that "...for 95% of all random samples, the constructed confidence intervals will contain μ."
- Wooldridge provides a nice example using 20 samples using a $X \sim \mathcal{N}(2, 1)$.

TABLE C.2 Simulated Confidence Intervals from a Normal(μ,1) Distribution with $\mu = 2$

Replication	$\bar{y}$	95% Interval	Contains μ?
1	1.98	(1.36,2.60)	Yes
2	1.43	(0.81,2.05)	Yes
3	1.65	(1.03,2.27)	Yes
4	1.88	(1.26,2.50)	Yes
5	2.34	(1.72,2.96)	Yes
6	2.58	(1.96,3.20)	Yes
7	1.58	(.96,2.20)	Yes
8	2.23	(1.61,2.85)	Yes
9	1.96	(1.34,2.58)	Yes
10	2.11	(1.49,2.73)	Yes
11	2.15	(1.53,2.77)	Yes
12	1.93	(1.31,2.55)	Yes
13	2.02	(1.40,2.64)	Yes
14	2.10	(1.48,2.72)	Yes
15	2.18	(1.56,2.80)	Yes
16	2.10	(1.48,2.72)	Yes
17	1.94	(1.32,2.56)	Yes
18	2.21	(1.59,2.83)	Yes
19	1.16	(.54,1.78)	No
20	1.75	(1.13,2.37)	Yes

Example #2: Confidence Intervals More Generally

- Suppose now that $X \sim \mathcal{N}(\mu, \sigma^2)$
- If σ is known, then $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$
- Working through the same logic, one can show that the appropriate confidence interval estimator becomes:

$$\left[\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n}\right] \tag{23}$$

- If, instead, σ is unknown, we must find an estimator for it.
- The sample standard deviation is a logical choice, where

$$S = \left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^{\frac{1}{2}} \tag{24}$$

Example #2: Confidence Intervals More Generally

- It now turns out that

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \tag{25}$$

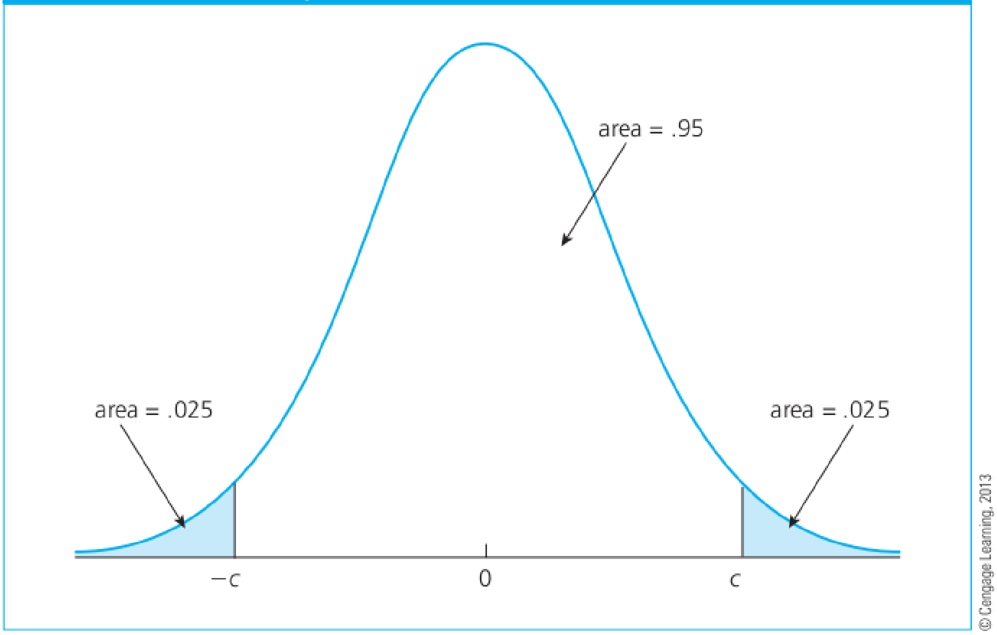
  where $S/\sqrt{n}$ is sometimes referred to as the standard error of $\bar{X}$.
- Similar to how we proceeded before, we have

$$\begin{aligned} 0.95 &= P\left(-c_{\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < c_{\alpha/2}\right) \\ &= P\left(\bar{X} - 1.96S/\sqrt{n} < \mu < \bar{X} + 1.96S/\sqrt{n}\right) \end{aligned} \tag{26}$$

  where c_α denotes the $100(1-\alpha)$ percentile in a t_{n-1} distribution.
- The corresponding *95%-confidence interval estimate* is given by $\left[\bar{x} - c_{\alpha/2}s/\sqrt{n}, \bar{x} + c_{\alpha/2}s/\sqrt{n}\right]$

Critical Values in a t-Distribution

FIGURE C.4 The 97.5[th] percentile, c, in a t distribution.



Numerical Example

- Suppose we have a random sample of MSU PhD program graduates, including data on their number of years to graduation (Y), with $y = \{4.5, 4, 7, 5, 5, 5.5, 5, 5, 3.5, 4, 4, 4.5, 6, 5.5, 4, 4.5,$ $6, 5, 5, 4.5, 4, 5, 5, 4.5, 4, 4, 4.5, 7.5, 3.5, 5\}$.
- Using this data

$$\begin{aligned} n &= 30. \\ \bar{y} &= 4.817. \\ s &= 0.924. \\ c_{\alpha/2} &= 2.045 \text{ for } \alpha = 0.025 \end{aligned}$$

- The corresponding confidence interval is given by: $\left[\bar{y} \pm c_{\alpha/2} s/\sqrt{n}\right] = [4.817 \pm 0.345] = [4.472, 5.162]$.

Hypothesis Testing

- We are often interesting in specific questions regarding a population; e.g.,
  - Does a job training program increase average worker wages?
  - Do stricter drunk driving laws reduce the number of drunk driving arrests?
  - Does an increase in the minimum wage increase the unemployment rates?
  - Is the average time to graduation for MSU PhD students the targeted 5 years?
- Consider the latter question. In the language of hypothesis testing:
  - Our *null hypothesis* is given by

$$H_0 : \mu_Y = 5 \tag{27}$$

    where Y is the number of years to graduation.
  - The *alternative hypothesis* is given by $H_1 : \mu_Y \neq 5$ (a two-sided alternative)

Two Types of Errors

- In hypothesis testing, we can make two types of mistakes:
  1. *Type I error*: Rejecting the null hypothesis when it is true.
     - The *significance level* of a test (α) is the probability of a Type I error.
     - Mathematically, $\alpha = P(RejectH_0|H_0)$.
  2. *Type II error*: Failing to rejecting the null hypothesis when it is false.

Hypothesis Testing Using a Test Statistic

- A *test statistic*, T, is some function of the random sample.
- The realized value of the test statistic for any given sample is given by t.
- Given the test statistic, we can define a rejection rule under which values of t that H_0 is rejected in favor of H_1.
- In Wooldridge, he focuses on rejection rules that compare the observed test statistic, t, to a critical level c.
- The values of t that result in rejection of H_0 are referred to as the *rejection region*.
- The rejection region depends on the alternative hypothesis:
  - $H_1 : \mu \neq \mu_0$
  - $H_1 : \mu > \mu_0$
  - $H_1 : \mu < \mu_0$

The Two-Sided Case

- With $H_1 : \mu \neq \mu_0$, we would want to reject the null hypothesis if the sample average differs *too* much from the hypothesized value μ_0 in either direction; i.e., if $|\bar{y} - \mu_0|$ is large.

  ...but we want to take into account how varied Y is to begin with.
  - We can do this by using the standardized test statistic:

$$T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \tag{28}$$

  under the null hypothesis.
  - The corresponding realized test statistic is given by

$$t = \frac{\bar{y} - \mu_0}{se(\bar{y})} \tag{29}$$

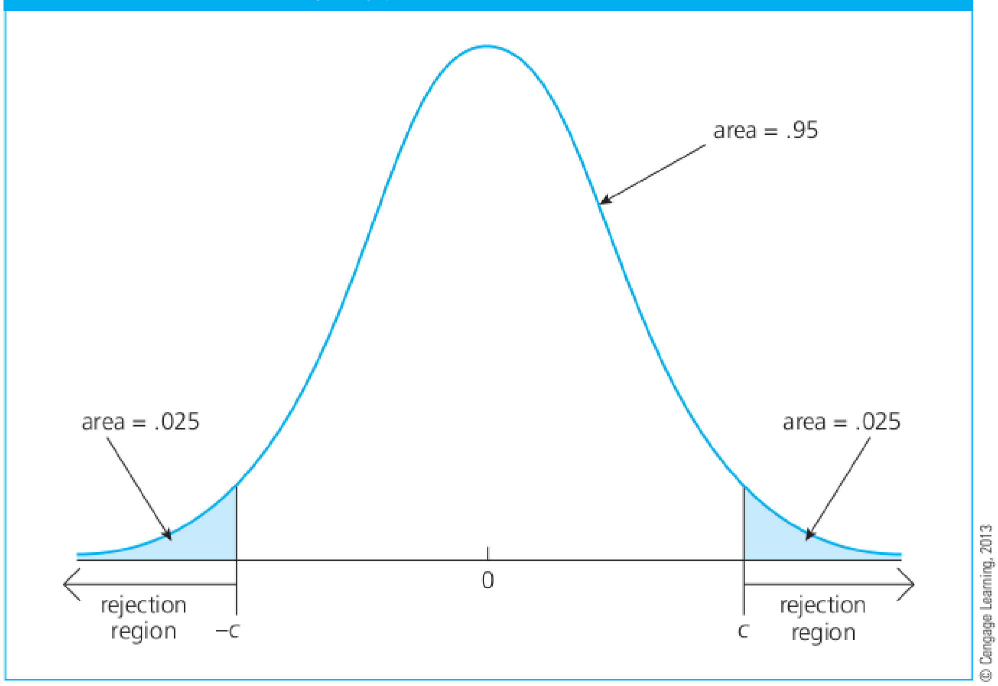
  where $se(\bar{y}) = S/\sqrt{n}$
- We want to reject the null hypothesis if $|t|$ is too large.

The Two-Sided Case (cont'd)

- The size of the rejection region will depend upon how confident we want to be regarding our rejection of the null;
- ...or to put it another way, how small we want the probability of a Type I error (significance level) to be.
- If we want a significance level to be $100 \cdot \alpha$, then the critical level becomes $c_{\alpha/2}$
- This splits the rejection region evenly between $\bar{y}$ being too big and $\bar{y}$ being too small.

Critical Region in a Two-Tailed Test

FIGURE C.6 Rejection region for a 5% significance level test against the two-sided alternative H_1: $\mu \neq \mu_0$.

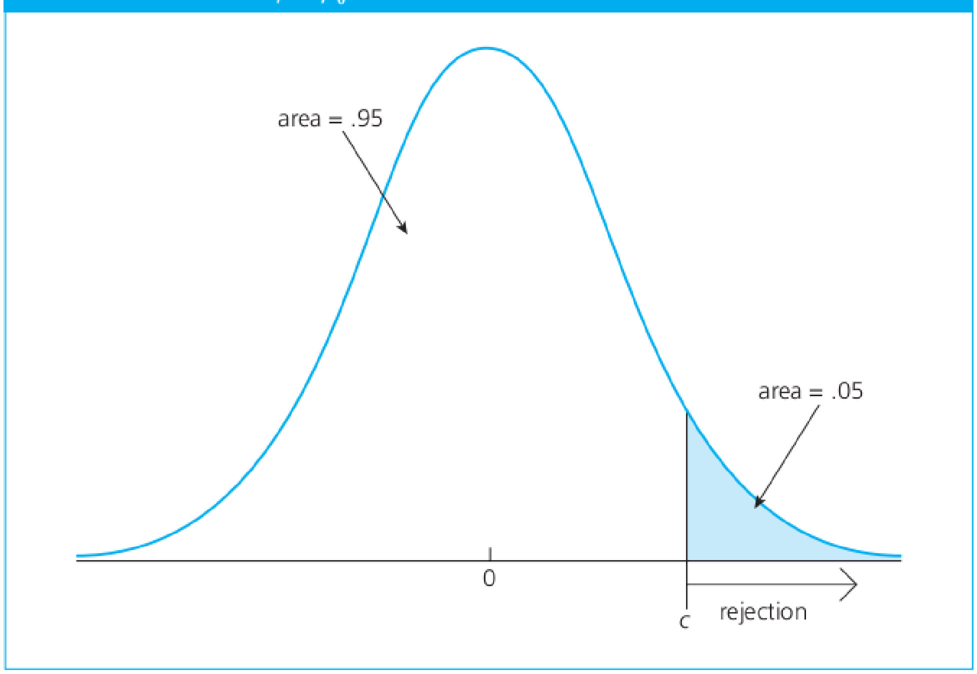


Graduate Student Example

- For our example, $t = \frac{\bar{y}-\mu_0}{se(\bar{y})} = \frac{4.817-5}{0.169} = -1.085$
- Using a significance level of 5%, the corresponding critical level becomes $c_{0.025} = 2.045$.
- We would not reject the null hypothesis in this case, since $|t| < c_{0.025}$.
- If instead our alternative hypothesis was one-sided, we would define the rejection region as one-sided as well.
- For example, if $H_1 : \mu > 5$, then the corresponding critical value would be $c_{0.05} = 1.699$.
- We would reject the null if $t > 1.699$, which is clearly not the case in our application.

Critical Region in a One-Tailed Test

FIGURE C.5 Rejection region for a 5% significance level test against the one-sided alternative $\mu > \mu_0$.



Computing and Using p-values

- An alternative approach to hypothesis testing is to compute the corresponding *p-value* for a test statistic.

  ..., where the p-value is the probability of obtaining a result equal to or "more extreme" than what was actually observed, when the null hypothesis is true; i.e.,

$$p - value = P(|T| > |t||H_0) \tag{30}$$

  - If this probability is small, it provides evidence against the null hypothesis.

- In our MSU graduate student example, we have

$$\begin{aligned} p\text{-}value &= P(|T_{n-1}| > |t||H_0) \\ &= 2 * P(T_{n-1} > 1.084|H_0) \\ &= 2 * (0.1434) = 0.2868 \end{aligned} \tag{31}$$

- In this case, we would not want to reject H_0, since there is nearly a 30 percent chance of observing a t-statistic as big or bigger than the one found in our current sample.

A Summary of Matrix Algebra

- Matrix algebra is often a convenient tool in econometrics, simplifying notation.
- Throughout the course, I will rewrite results in terms of matrices.
- Understanding matrix notation is useful, but not essential to the course.

Basic Matrix Notation

- A matrix is a rectangular array of numbers or elements arranged in rows and columns.
- An $M \times N$ (M rows and N columns) matrix $\boldsymbol{A}$ can be expressed as:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1N} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & a_{M3} & \dots & a_{MN} \end{bmatrix} \tag{32}$$

  where a_{ij} represents the element in i^{th} row and j^{th} column.
- For example a_{23} stands for the element in the 2^{nd} row and 3^{rd} column. $[a_{ij}]$ is a shorthand expression for the matrix $\boldsymbol{A}$.
- $\boldsymbol{A} = \begin{bmatrix} 2 & 4 & -3 \\ 8 & 1 & 12 \end{bmatrix}$ is a 2×3 matrix with $a_{23} = 12$.

Column and Row Vectors

- An $M \times 1$ matrix (M rows, one column) is called a *column vector*.
- Letting bold lowercase letters denote vectors, an example of 3×1 column vector can be written as:

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \tag{33}$$

- A $1 \times N$ matrix (1 row, N columns) is called a *row vector*.
- An example of 1×4 row vector can be written as:

$$\boldsymbol{y} = \begin{bmatrix} y_1 & y_2 & y_3 & y_4 \end{bmatrix} \tag{34}$$

Matrix Tranposition

- Let $\boldsymbol{A} = [a_{ij}]$ be an $M \times N$ matrix.
  - The transpose of $\boldsymbol{A}$, denoted $\boldsymbol{A}'$, is an $N \times M$ matrix obtained by interchanging the rows and columns of $\boldsymbol{A}$.
  - In short, we can write $\boldsymbol{A}'$ in short form as $\boldsymbol{A}' = [a_{ji}]$.
- For example, if $\boldsymbol{A} = \begin{bmatrix} 2 & 4 & -3 \\ 8 & 1 & 12 \end{bmatrix}$, then $\boldsymbol{A}' = \begin{bmatrix} 2 & 8 \\ 4 & 1 \\ -3 & 12 \end{bmatrix}$.
- Transposition of a row vector is a column vector, and the transpose of a column vector is a row vector.
- For example, if $\boldsymbol{x} = \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix}$, then $\boldsymbol{x}' = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$.

Types of Matrices

- *Submatrix*: Given an $M \times N$ matrix, if all but r rows and s columns of $\boldsymbol{A}$ are deleted, the remaining matrix is called a submatrix of $\boldsymbol{A}$.
- *Square matrix*: A matrix that has the same number of rows as columns.
- *Diagonal matrix*: A matrix that has all of its off-diagonal elements equal to zero and at least one nonzero element on the diagonal.
- *Identity matrix* is a diagonal matrix whose diagonal elements are all 1.
- *Symmetric matrix* is a square matrix whose elements above the main diagonal are mirror images of the elements below the main diagonal.
  - Equivalently, $\boldsymbol{A}$ is a symmetric matrix if $\boldsymbol{A} = \boldsymbol{A}'$.
  - All the identity matrices are symmetric matrices.
- *Null Matrix* is a matrix whose elements are all zero, denoted by $\mathbf{0}$.
- *Null vector* is a row or column vector whose elements are all zero, also denoted by $\mathbf{0}$.

Matrix Operations

- Matrix addition: Two matrices $\boldsymbol{A} = [a_{ij}]$ and $\boldsymbol{B} = [b_{ij}]$, each having the same dimension $M \times N$, can be added element by element to form a matrix: $\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{B} = [a_{ij} + b_{ij}]$.
- $\boldsymbol{A}$ and $\boldsymbol{B}$ must have the same dimension in order to be conformable for addition.
- Let $\boldsymbol{A} = \begin{bmatrix} 1 & 3 & 7 \\ 2 & 5 & 3 \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} 2 & 1 & 5 \\ 3 & 4 & 6 \end{bmatrix}$, then $\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{B} = \begin{bmatrix} 3 & 4 & 12 \\ 5 & 9 & 9 \end{bmatrix}$.
- Matrix subtraction: matrix subtraction follows exactly the same rules as for addition; i.e., $\boldsymbol{C} = \boldsymbol{A} - \boldsymbol{B} = [a_{ij} - b_{ij}]$.

Matrix Operations (cont'd)

- Scalar Multiplication: To multiply a matrix $\boldsymbol{A}$ by any real number λ (also called scalar), we simply multiply each element of $\boldsymbol{A}$ by λ.
- Matrix Multiplication: To multiply matrix $\boldsymbol{A}$ by matrix $\boldsymbol{B}$ to form a new matrix $\boldsymbol{C}$, the column dimension of $\boldsymbol{A}$ must be equal the row dimension of $\boldsymbol{B}$. Let $\boldsymbol{A}$ be $M \times N$ and $\boldsymbol{B}$ be $N \times P$. Then each element of matrix $\boldsymbol{C}$ is obtained as

$$c_{ij} = \sum_{k=1}^{N} a_{ik} b_{kj} \quad i = 1, \ldots, M; j = 1, \ldots, P \tag{35}$$

Some Useful Properties of Matrix Multiplication

- $\alpha(\boldsymbol{AB}) = (\alpha\boldsymbol{A})\boldsymbol{B}$
- $(\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC})$
- $\boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C}) = \boldsymbol{AB}+\boldsymbol{AC}$
- $(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{C} = \boldsymbol{AC}+\boldsymbol{BC}$
- $I\boldsymbol{A} = \boldsymbol{A}I = \boldsymbol{A}$, where I is an identity matrix
- $\boldsymbol{AB}$ need not equal $\boldsymbol{BA}$, even when both products are defined. $\boldsymbol{AB} = \boldsymbol{BA}$ only under special circumstances.
- A row vector post-multiplied by a column vector is a scalar.
- A column vector post-multiplied by a row vector is a matrix.
- A matrix post-multiplied by a column vector is a column vector.
- A row vector post-multiplied by a matrix is a row vector

Matrix Inversion

- An inverse of an $N \times N$ square matrix $\boldsymbol{A}$, denoted by $\boldsymbol{A}^{-1}$, exists if

$$\boldsymbol{AA}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I} \tag{36}$$

  where $\boldsymbol{I}$ is an identity matrix.
- If a square matrix has an inverse, we say the matrix is *invertible* or *nonsingular*. Otherwise, it is said to be *noninvertible* or *singular*.
- Properties of an inverse
  1. If an inverse exists, it is unique.
  2. $(\alpha\boldsymbol{A})^{-1} = \frac{1}{\alpha}\boldsymbol{A}^{-1}$
  3. $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$ if $\boldsymbol{A}$ and $\boldsymbol{B}$ are both $N \times N$ and invertible.
  4. $(\boldsymbol{A}^{-1})' = (\boldsymbol{A}')^{-1}$.

Linear Independence

- Linear dependence: Let $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k\}$ be a set of $N \times 1$ vectors. We say they are linearly independent vectors if, and only if,

$$\alpha_1 \boldsymbol{x}_1 + \alpha_2 \boldsymbol{x}_2 + \cdots + \alpha_k \boldsymbol{x}_k = 0 \tag{37}$$

  implies that $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$.
- If there is at least one of the α's is not equal to zero, then $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k\}$ is linearly dependent.
- In other words, at least one vector in this set can be written as a linear combination of the others.

Appendix D: Summary of Matrix Algebra

The Rank of a Matrix

- The rank of the $N \times M$ matrix $\boldsymbol{A}$, denoted rank($\boldsymbol{A}$), is the maximum number of linearly independent columns of $\boldsymbol{A}$.
- If rank($\boldsymbol{A}$)$= M$, the number of columns of $\boldsymbol{A}$, then $\boldsymbol{A}$ is said to have a full column rank.
- If an $N \times M$ matrix $\boldsymbol{A}$ has full column rank, then its columns are linearly independent, and $\boldsymbol{A}'\boldsymbol{A}$ is nonsingular.
- If an $N \times N$ square matrix $\boldsymbol{A}$ has full column rank, then $\boldsymbol{A}$ is nonsingular.