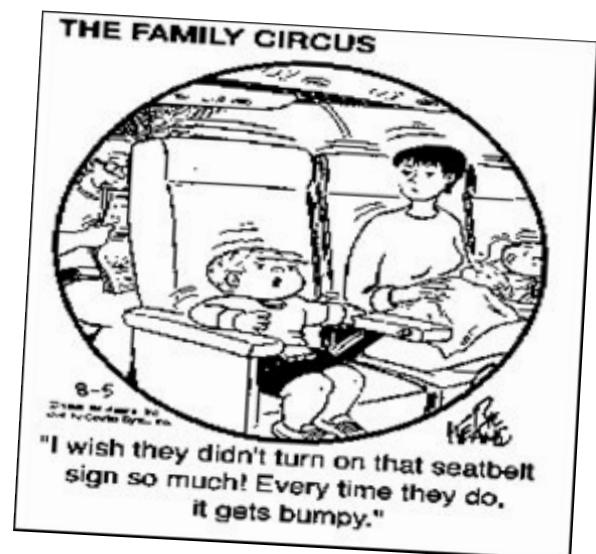




Quantitative Methods

Executive MBA

- Content
- References
- Material



"No matter how many instances of white swans we may have observed, this does not justify the conclusion that all swans are white."

Karl Popper

The Economist Espresso

Table of contents (Reader)

Cooper, D. and Schindler, P.S.: “Business Research Methods”	
Slides chapters 1-4, 6, 10-14, 17	1
Koop, G : “Analysis of Economic Data”	29
Chapter 1 Introduction	31
Chapter 3 Correlation (slides)	35
Chapter 4 An introduction to simple regression (slides)	39
Chapter 5 Statistical aspects of regression (slides)	46
Chapter 6 Multiple regression (slides)	50
Chapter 7 Regression with dummy variables (slides)	56
Chapter 9 Univariate time series analysis (slides)	61
Baye, M. “Managerial economics and business strategy“” task “College Town”	68

(Sources used: see below)



Selected **References**

Cooper, Donald and **Schindler**, Pamela: "Business Research Methods"

Koop, Gary: "Analysis of Economic Data"

Black, Thomas: "Understanding Social Science Research"

Baye, Michael: “Managerial economics and business strategy”

Blaxter, Loraine; **Hughes**, Christian and **Tight**, Malcolm: "HOW TO Research"

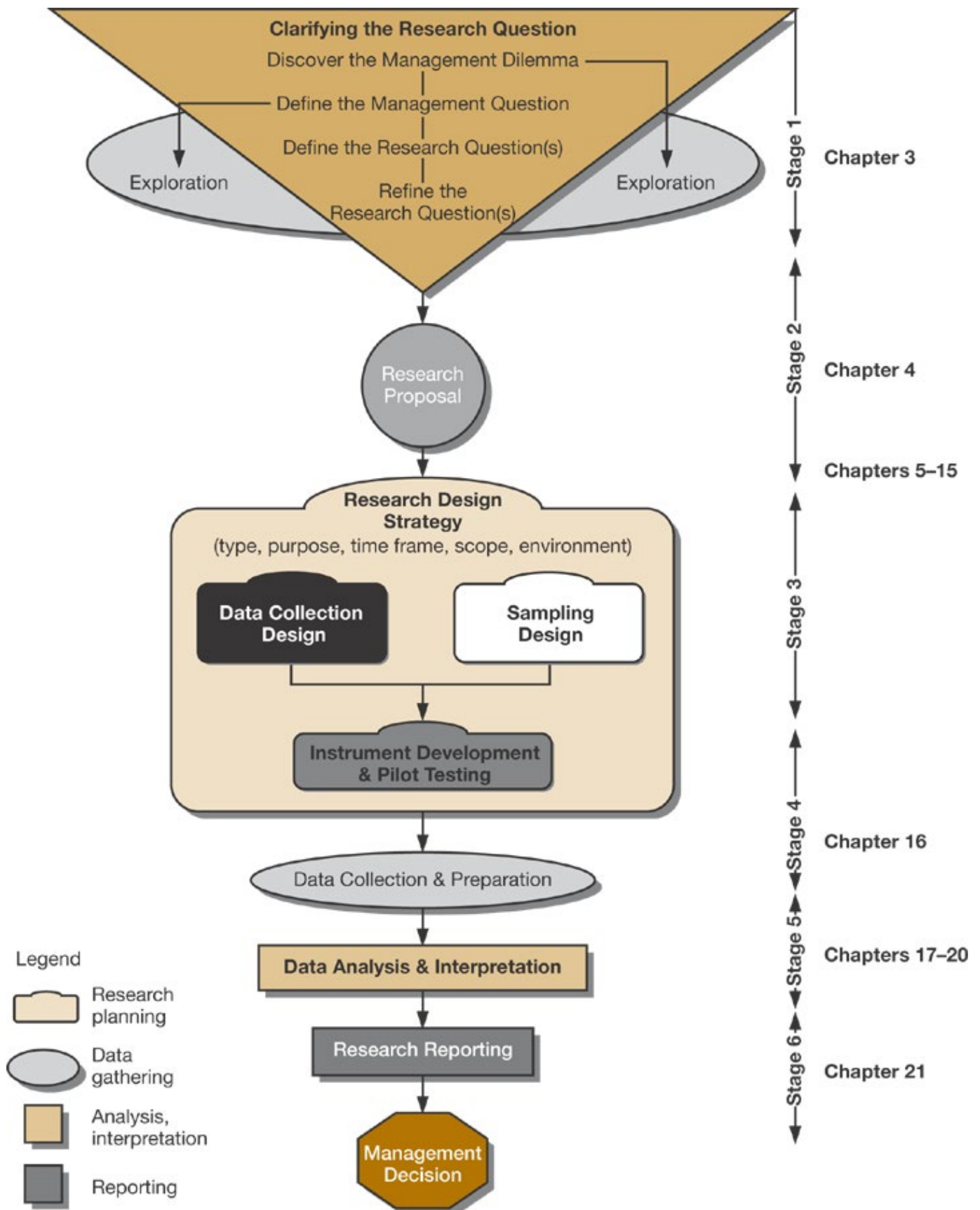
Bryman, Alan and **Cramer**, Duncan: “Quantitative data analysis with SPSS”

Kinnear, Paul and Gray, Colin: “SPSS made simple”

Oakshott, Les: "Essential Quantitative Methods for Business"

Studenmund, A. H.: “Using econometrics : a practical guide”

The Business Research Process



Source: Cooper, D. and Schindler, P.S.: "Business Research Methods"

Chapter 1

RESEARCH IN BUSINESS



McGraw-Hill/Irwin

Copyright © 2014 by The McGraw-Hill Companies, Inc. All rights reserved.

Learning Objectives

Understand . . .

- What business research is and how it differs from business decision support systems and business intelligence systems.
- Trends affecting business research and the emerging hierarchy of business decision makers.
- The distinction between good business research and research that falls short of professional quality.
- The nature of the research process.

Pull Quote

“Forward-thinking executives recognize that analytics may be the only true source of sustainable advantage since it empowers employees at all levels of an organization with information to help them make smarter decisions.”

Wayne Eckerson,
director of research, business applications and
architecture group,
TechTarget

Why Study Business Research?

Business research provides information to guide business decisions

Good research is like a parachute.
Without it, you could come to the wrong conclusion.

With research from JRP, you'll reach the right decision. For more than 40 years, we've worked with all agencies and corporate clients as partners, designing and fielding projects of all types. See why our seasoned staff of project directors, interviewers, coders and analysts have led so many companies to come to the same conclusion: JRP. Call Paul Frattolillo toll free at 877-JRP-2095 and ask about our full range of services.

JRP
MARKET RESEARCH
ANALYTICS & CONSULTING

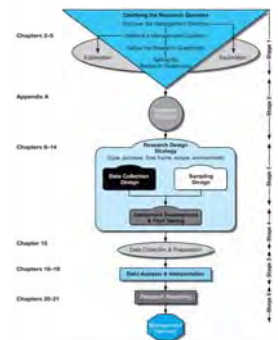
100 EAST 10TH STREET, SUITE 200, NEW YORK, NY 10003-3675
TEL: 212-691-2543 FAX: 212-691-2544
WWW.JRP-RESEARCH.COM

Business Research

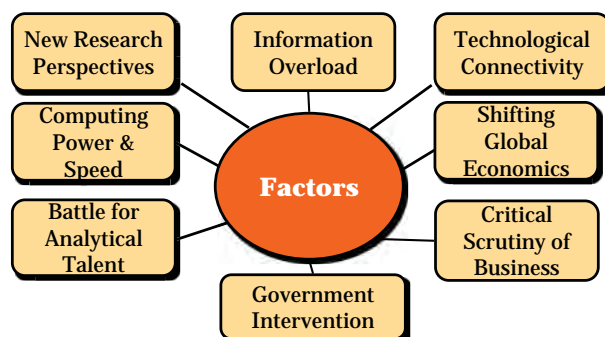
- A process of determining, acquiring, analyzing, synthesizing, and disseminating relevant business data, information, and insights to decision makers in ways that mobilize the organization to take appropriate business actions that, in turn, maximize business performance

The Research Process

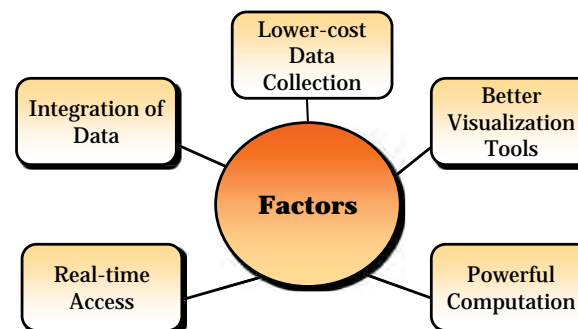
- Stage 1:** Clarifying the Research question
- Stage 2:** Proposing Research
- Stage 3:** Designing the Research
- Stage 4:** Data Collection & Preparation
- Stage 5:** Data Analysis & Interpretation
- Stage 6:** Reporting the Results



What's Changing in Business that Influences Research



Computing Power and Speed



Information Sources

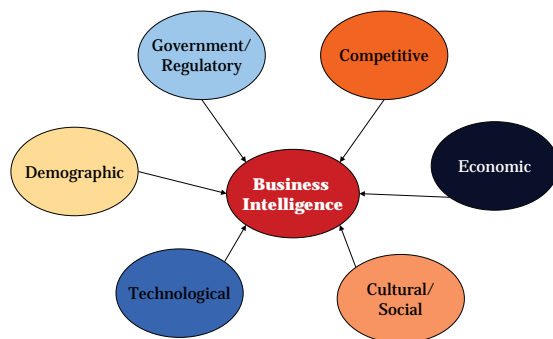
Decision Support Systems

- Numerous elements of data organized for retrieval and use in business decision making
- Stored and retrieved via
 - Intranets
 - Extranets

Business Intelligence Systems

- Ongoing information collection
- Focused on events, trends in micro and macro-environments

Sources of Business Intelligence



10

Hierarchy of Business Decision Makers



11

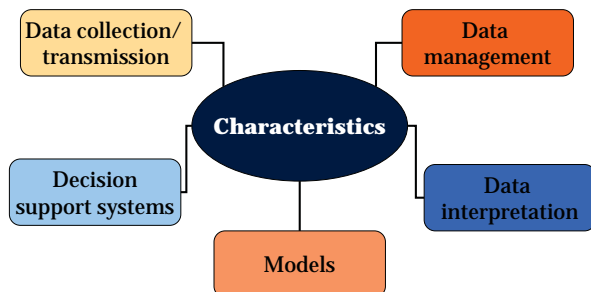
Research May Not Be Necessary

Can It Pass These Tests?

- Can information be applied to a critical decision?
- Will the information improve managerial decision making?
- Are sufficient resources available?

12

Information Value Chain



13

Chapter 2

ETHICS IN BUSINESS RESEARCH



McGraw-Hill/Irwin

Copyright © 2014 by The McGraw-Hill Companies, Inc. All rights reserved.

Ethical Treatment of Participants



Do no harm

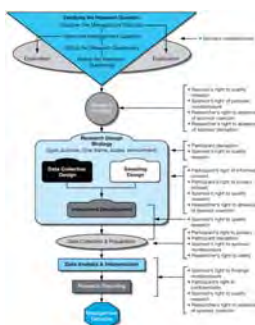
Explain study benefits

Explain participant rights and protections

Obtain informed consent

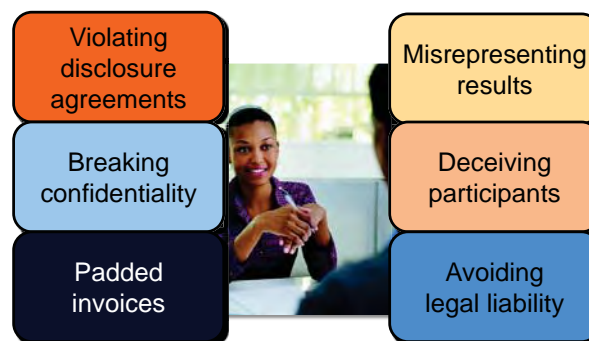
15

Ethical Issues and the Research Process



16

Types of Ethical Violations



17

Unethical Behavior to Avoid

Violating participant confidentiality

Changing data presentation

Changing data

Creating false data

Changing data interpretations

Injecting bias in interpretations

Omitting sections of data

Making recommendations beyond scope of data



18

Chapter 3

THINKING LIKE A RESEARCHER



McGraw-Hill/Irwin

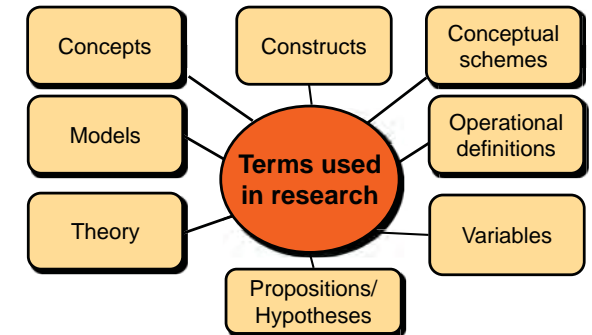
Copyright © 2014 by The McGraw-Hill Companies, Inc. All rights reserved.

Learning Objectives

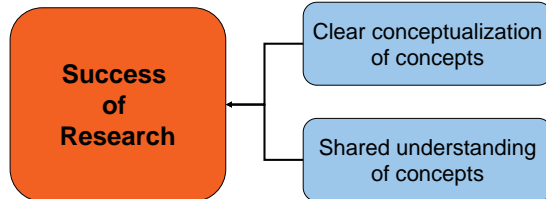
Understand . . .

- The terminology used by professional researchers employing scientific thinking.
- What you need to formulate a solid research hypothesis.
- The need for sound reasoning to enhance research results.

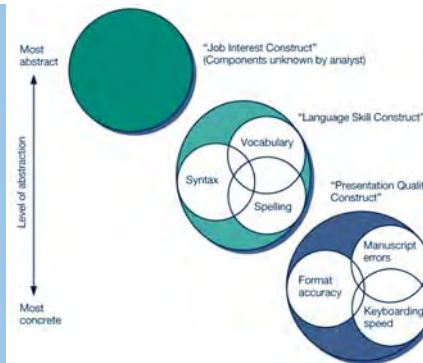
Language of Research



Language of Research



Job Redesign Constructs and Concepts

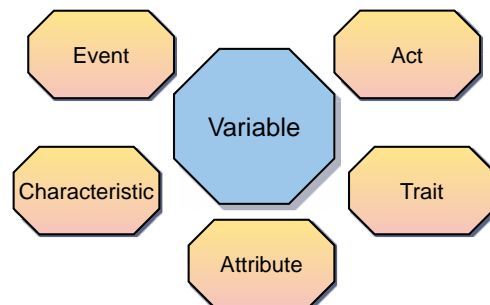


Operational Definitions

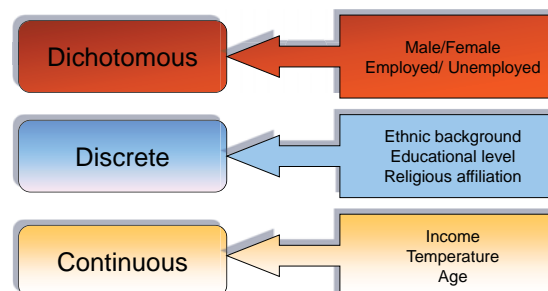
How can we define the variable "class level of students"?

- | | |
|-------------|----------------------|
| • Freshman | • < 30 credit hours |
| • Sophomore | • 30-50 credit hours |
| • Junior | • 60-89 credit hours |
| • Senior | • > 90 credit hours |

A Variable: Property Being Studied



Types of Variables



Independent and Dependent Variable Synonyms

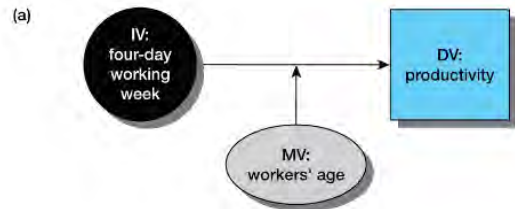
Independent Variable (IV)

- Predictor
- Presumed cause
- Stimulus
- Predicted from...
- Antecedent
- Manipulated

Dependent Variable (DV)

- Criterion
- Presumed effect
- Response
- Predicted to....
- Consequence
- Measured outcome

Relationships Among Variable Types



Relationships Among Variable Types

Moderating Variables (MV)

- The introduction of a four-day week (IV) will lead to higher productivity (DV), especially among younger workers (MV).
- The switch to commission from a salary compensation system (IV) will lead to increased sales (DV) per worker, especially more experienced workers (MV).
- The loss of mining jobs (IV) leads to acceptance of higher-risk behaviors to earn a family-supporting income (DV) – particularly among those with a limited education (MV).

Extraneous Variables (EV)

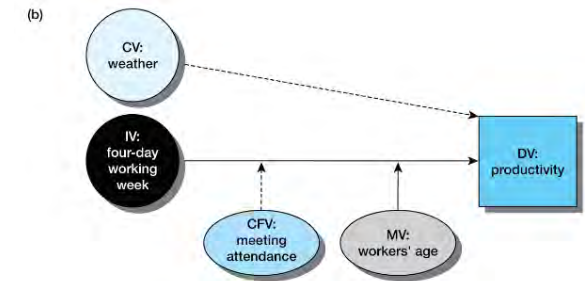
With new customers (EV-control), a switch to commission from a salary compensation system (IV) will lead to increased sales productivity (DV) per worker, especially among younger workers (MV).

Among residents with less than a high school education (EV-control), the loss of jobs (IV) leads to high-risk behaviors (DV), especially due to the proximity of the firing range (MV).

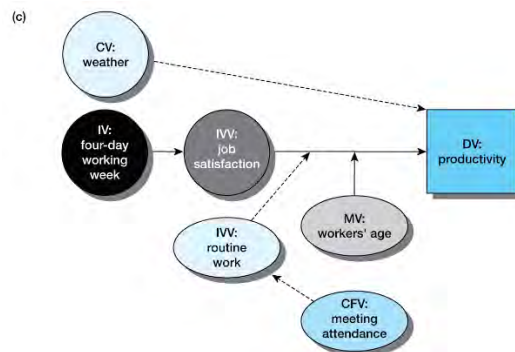
Intervening Variables (IVV)

- The switch to a commission compensation system (IV) will lead to higher sales (DV) by increasing overall compensation (IVV).
- A promotion campaign (IV) will increase savings activity (DV), especially when free prizes are offered (MV), but chiefly among smaller savers (EV-control). The results come from enhancing the motivation to save (IVV).

Relationships Among Variable Types



Relationships Among Variable Types



Descriptive Hypothesis Formats

Descriptive Hypothesis

- In Detroit, our potato chip market share stands at 13.7%.
- American cities are experiencing budget difficulties.

Research Question

- What is the market share for our potato chips in Detroit?
- Are American cities experiencing budget difficulties?

Relational Hypotheses Formats

• Correlational

- Young women (under 35) purchase fewer units of our product than women who are older than 35.
- The number of suits sold varies directly with the level of the business cycle.

• Causal

- An increase in family income leads to an increase in the percentage of income saved.
- Loyalty to a grocery store increases the probability of purchasing that store's private brand products.

The Role of Hypotheses

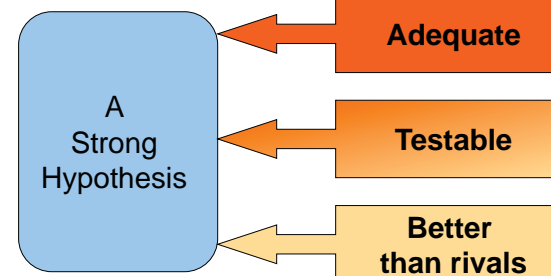
Guide the direction of the study

Identify relevant facts

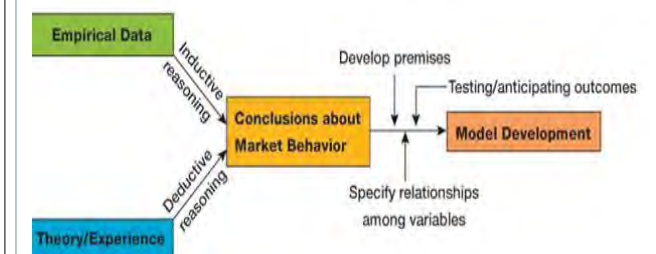
Suggest most appropriate research design

Provide framework for organizing resulting conclusions

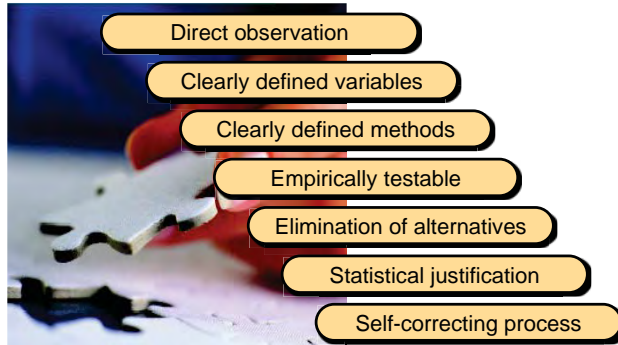
Characteristics of Strong Hypotheses



The Role of Reasoning



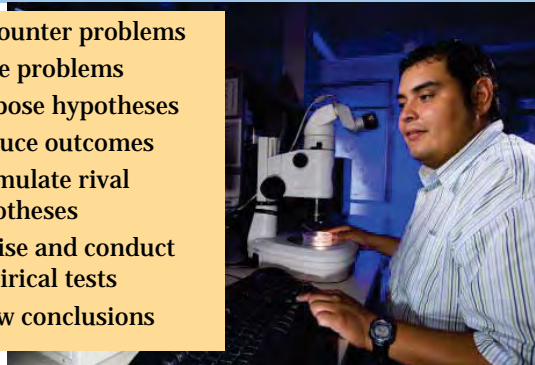
The Scientific Method



37

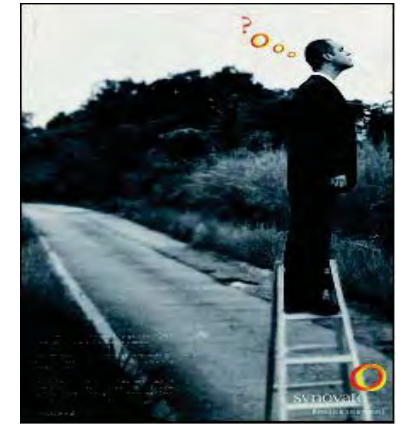
Researchers

- Encounter problems
- State problems
- Propose hypotheses
- Deduce outcomes
- Formulate rival hypotheses
- Devise and conduct empirical tests
- Draw conclusions



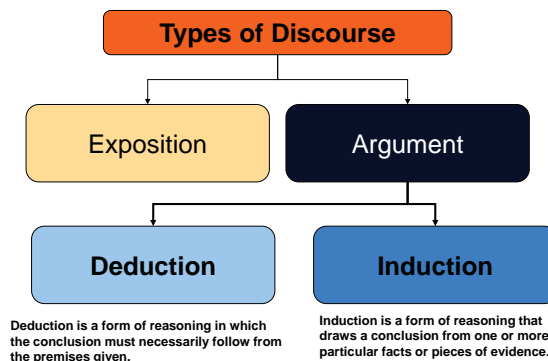
38

Why is curiosity important?



39

Sound Reasoning



40

Deductive Reasoning



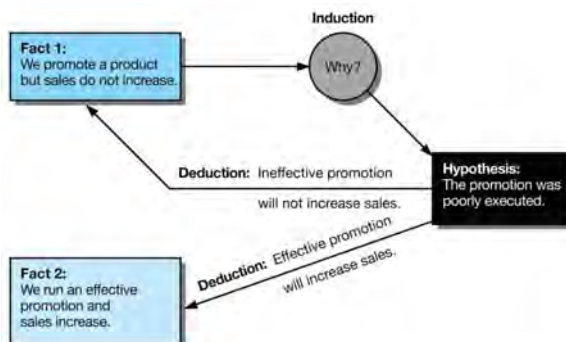
41

Inductive Reasoning

- **Why didn't sales increase during our promotional event?**
 - Regional retailers did not have sufficient stock to fill customer requests during the promotional period
 - A strike by employees prevented stock from arriving in time for promotion to be effective
 - A hurricane closed retail outlets in the region for 10 days during the promotion

42

Why Didn't Sales Increase?



43

Chapter 4

THE RESEARCH PROCESS: AN OVERVIEW



McGraw-Hill/Irwin

Copyright © 2014 by The McGraw-Hill Companies, Inc. All rights reserved.

Learning Objectives

Understand ...

- That research is decision- and dilemma-centered.
- That the clarified research question is the result of careful exploration and analysis and sets the direction for the research project.

45

Learning Objectives

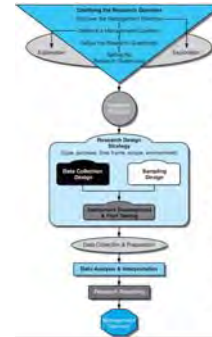
Understand . . .

- How value assessments and budgeting influence the process for proposing research, and ultimately, research design.
- What is included in research design, data collection, and data analysis.
- Research process problems to avoid.

46

The Research Process

- Stage 1:** Clarifying the Research question
- Stage 2:** Proposing Research
- Stage 3:** Designing the Research
- Stage 4:** Data Collection & Preparation
- Stage 5:** Data Analysis & Interpretation
- Stage 6:** Reporting the Results



47

Stage 1: Clarifying the Research Question

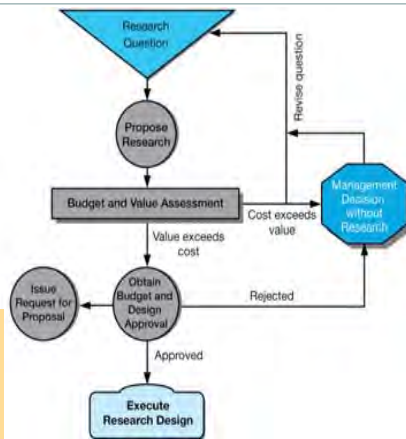


Management-research question hierarchy process begins by identifying the management dilemma

48

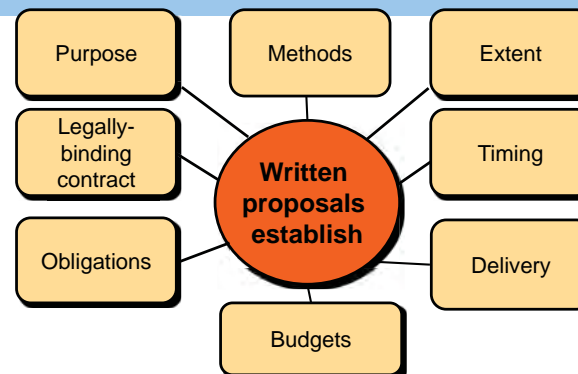
Stage 2: Proposing Research

- Budget Types**
 - Rule-of-thumb
 - Departmental
 - Task



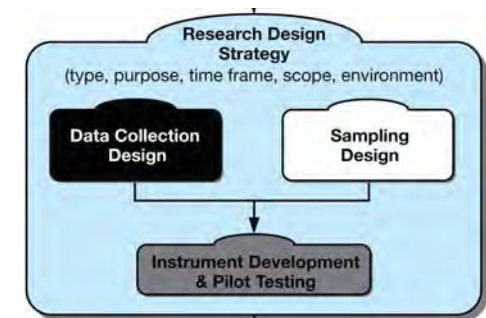
49

The Research Proposal



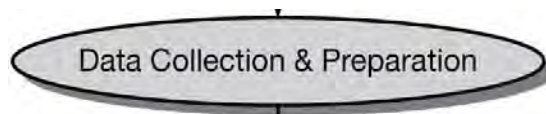
50

Stage 3: Designing the Research



51

Stage 4: Data Collection



52

Data Characteristics

- Abstractness
- Verifiability
- Elusiveness
- Closeness



53

Data Types

Primary

Secondary



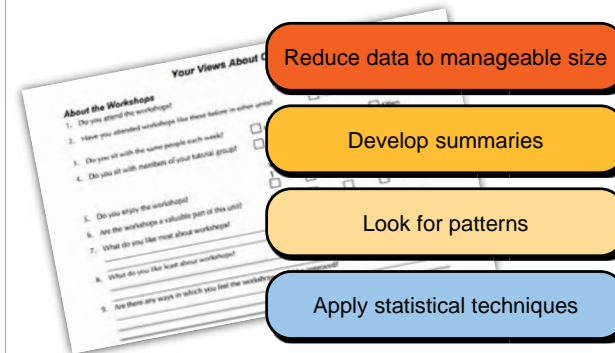
54

Stage 5: Data Analysis & Interpretation

Data Analysis & Interpretation

55

Steps in Data Analysis and Interpretation



56

Stage 6: Reporting the Results

Research Reporting

Management Decision

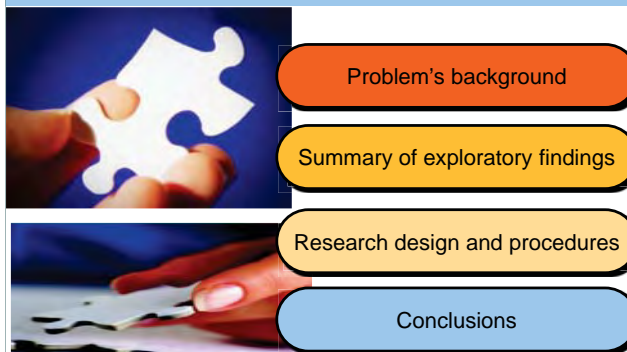
57

Parts of the Research Report



58

The Research Report Overview



59

Chapter 6

RESEARCH DESIGN: AN OVERVIEW



McGraw-Hill/Irwin

Copyright © 2014 by The McGraw-Hill Companies, Inc. All rights reserved.

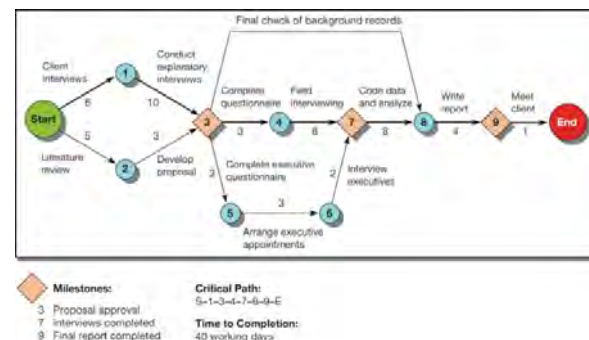
Learning Objectives

Understand . . .

- The basic stages of research design.
- The major descriptors of research design.
- The major types of research designs.
- The relationships that exist between variables in research design and the steps for evaluating those relationships.

61

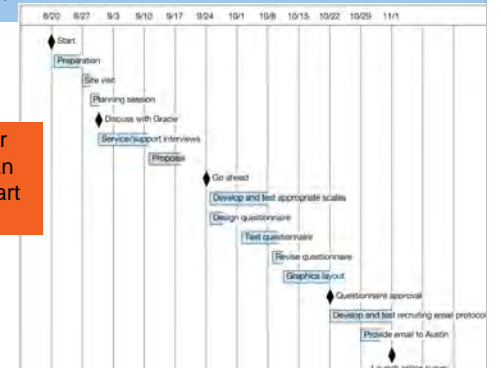
What Tools Are Used in Designing Research?



62

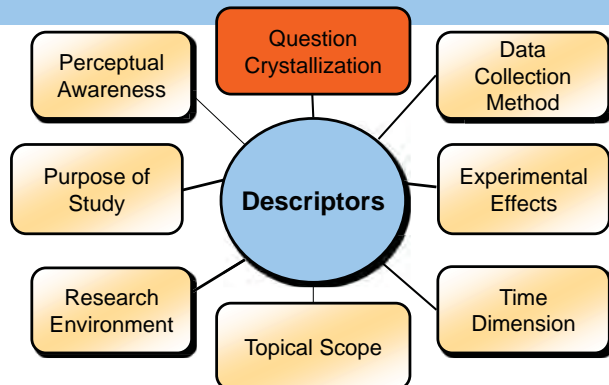
What Tools Are Used in Designing Research?

MindWriter Project Plan in Gantt chart format



63

Research Design Descriptors



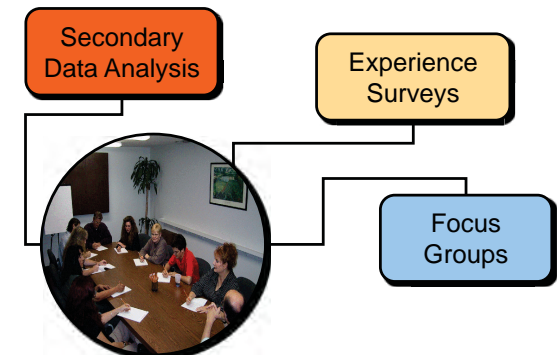
64

Degree of Question Crystallization

- **Exploratory Study**
 - Loose structure
 - Expand understanding
 - Provide insight
 - Develop hypotheses
- **Formal Study**
 - Precise procedures
 - Begins with hypotheses
 - Answers research questions

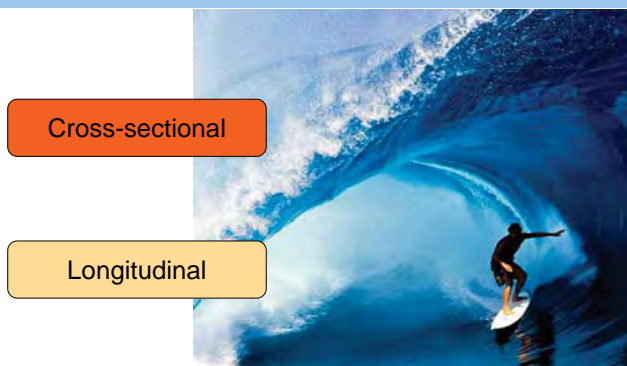
65

Commonly Used Exploratory Techniques



66

The Time Dimension



67

The Research Environment



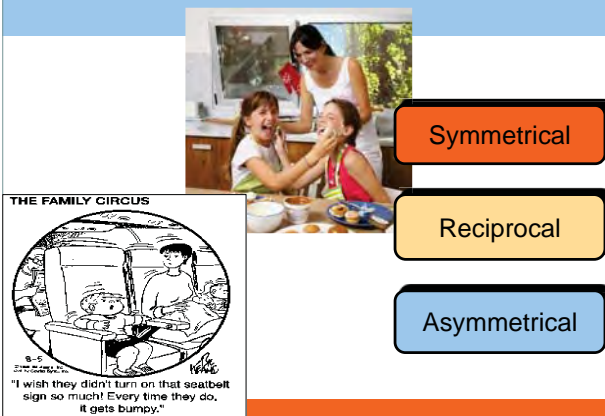
68

Descriptive Studies



69

Causal Studies



70

Understanding Casual Relationships



71

Chapter 10 SURVEYS



McGraw-Hill/Irwin

Copyright © 2014 by The McGraw-Hill Companies, Inc. All rights reserved.

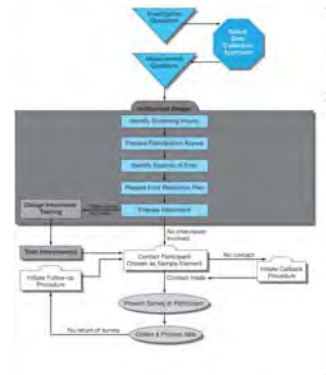
Learning Objectives

Understand . . .

- The process for selecting the appropriate and optimal communication approach.
- Factors affect participation in communication studies.
- Sources of error in communication studies and how to minimize them.
- Major advantages and disadvantages of the three communication approaches.
- Why an organization might outsource a communication study.

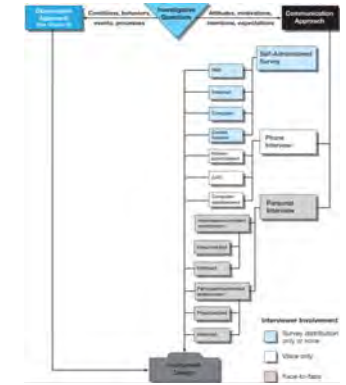
73

Data Collection Approach



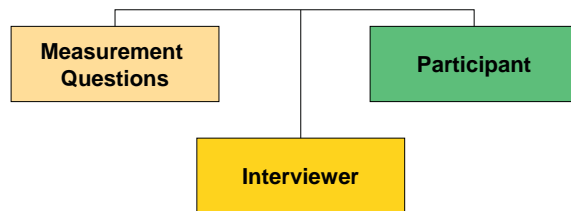
74

Selecting a Communication Data Collection Approach



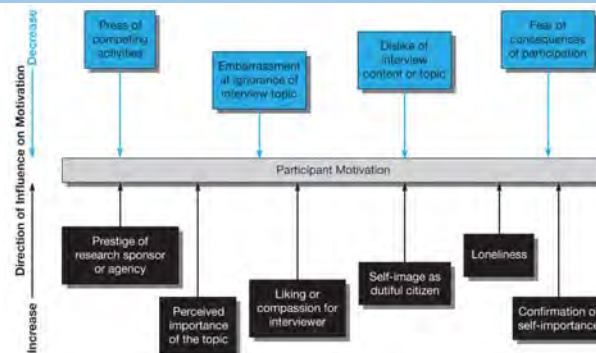
75

Sources of Error



76

Participant Motivation



77

Response Terms

Noncontact rate

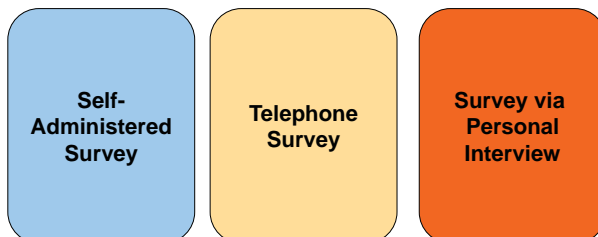
Refusal rate

Incidence rate



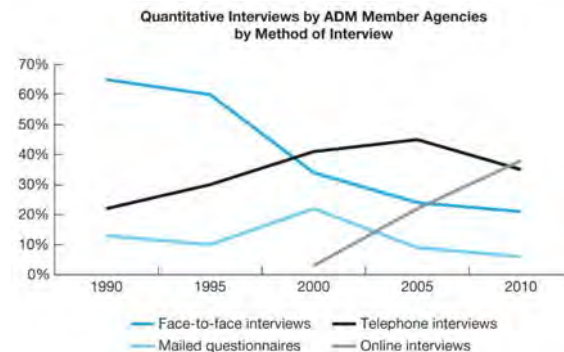
78

Communication Approaches



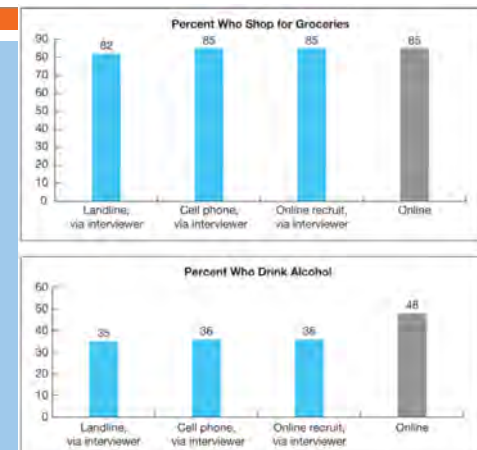
79

PicProfile: Methodology Trends



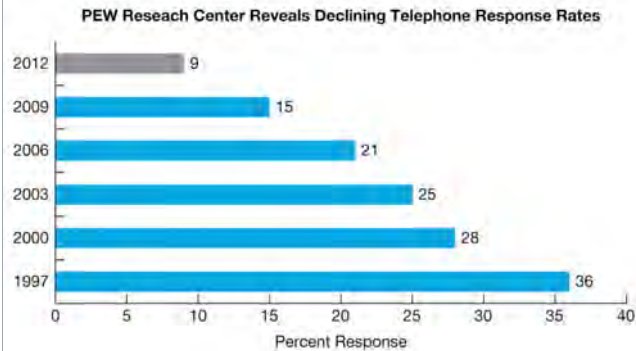
80

PicProfile: Mixed-mode Research



81

PicProfile: Declining Phone Response



82

Chapter 11

MEASUREMENT



McGraw-Hill/Irwin

Copyright © 2014 by The McGraw-Hill Companies, Inc. All rights reserved.

84

Learning Objectives

Understand . . .

- The distinction between measuring objects, properties, and indicants of properties.
- The similarities and differences between the four scale types used in measurement and when each is used.
- The four major sources of measurement error.
- The criteria for evaluating good measurement.

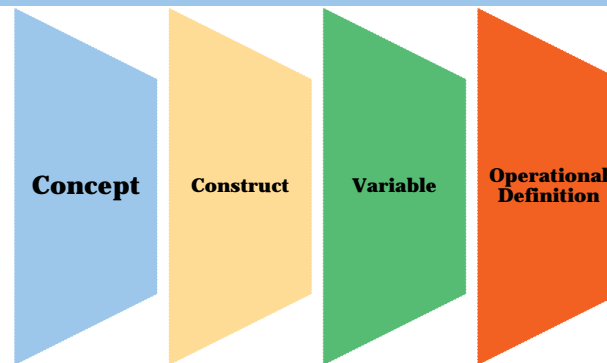
Pull Quote

“You’re trying too hard to find a correlation here. You don’t know these people, you don’t know what they intended. You try to compile statistics and correlate them to a result that amounts to nothing more than speculation.”

Marc Racicot, former governor of Montana and chairman of the Republican Party

85

Review of Terms



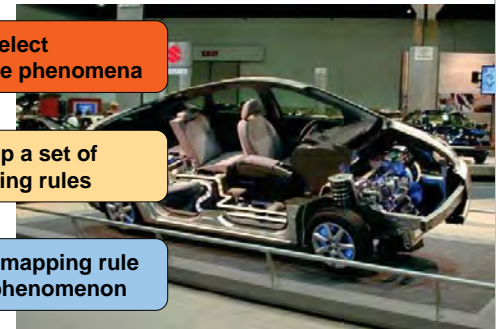
86

Measurement

Select measurable phenomena

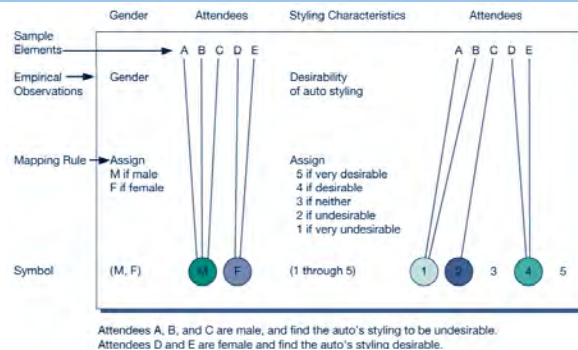
Develop a set of mapping rules

Apply the mapping rule to each phenomenon



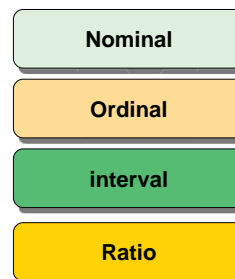
87

Characteristics of Measurement



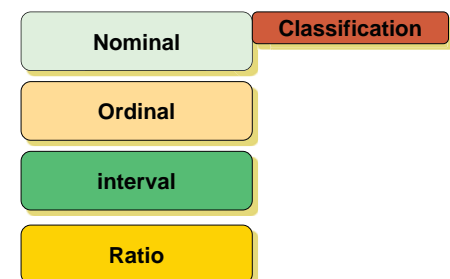
88

Types of Scales



89

Levels of Measurement



90

Nominal Scales

Mutually Exclusive

Collectively Exhaustive Categories

Classification Only



91

Levels of Measurement

Nominal

Classification

Ordinal

Classification

Order

Interval

Ratio

92

Ordinal Scales



Nominal Scale Characteristics

Order

Implies greater than or less than

93

Levels of Measurement

Nominal

Classification

Ordinal

Classification

Order

Interval

Classification

Order

Distance

Ratio

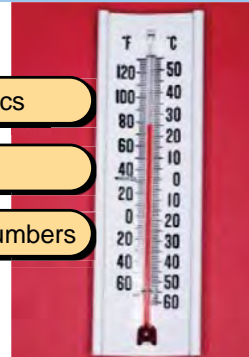
94

Interval Scales

Ordinal Scale Characteristics

Equality of interval

Equality of distance between numbers



95

Levels of Measurement

Nominal

Classification

Ordinal

Classification

Order

Interval

Classification

Order

Distance

Ratio

Classification

Order

Distance

Natural Origin

96

Ratio Scales

Interval Scale Characteristics

Absolute Zero



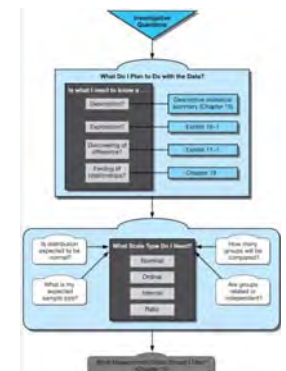
97

Examples of Data Scales

Type of Scale	Example
Nominal	Gender (male, female)
Ordinal	Doneness of meat (well, medium well, medium rare, rare)
Interval	Temperature in degrees
Ratio	Age in years

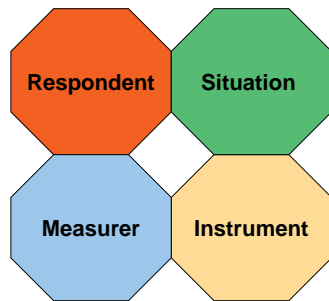
98

From Investigative to Measurement Questions



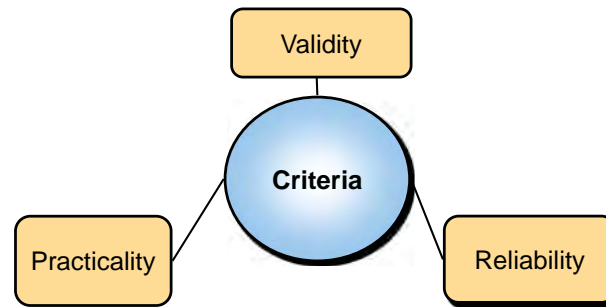
99

Sources of Error



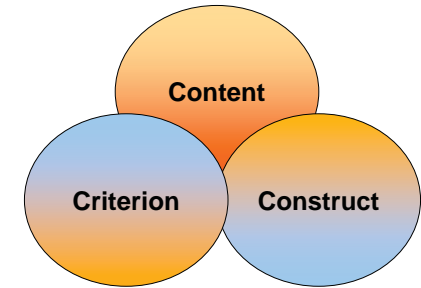
100

Evaluating Measurement Tools



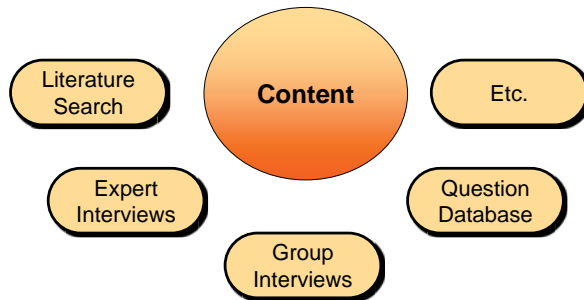
101

Validity Determinants



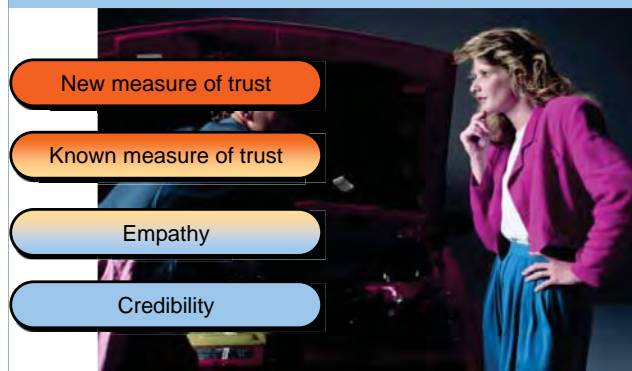
102

Increasing Content Validity



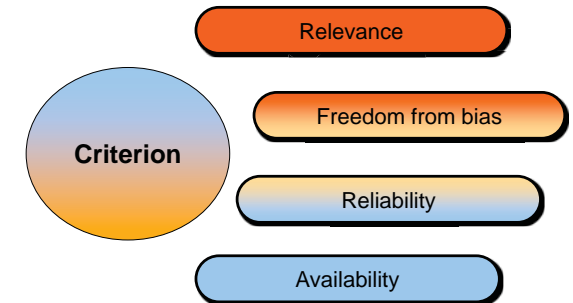
103

Increasing Construct Validity



104

Judging Criterion Validity



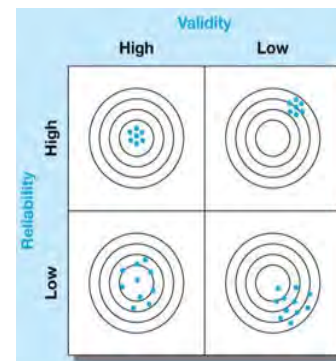
105

Summary of Validity Estimates

Types	What Is Measured	Methods
Content	Degree to which the content of the items adequately represents the universe of all relevant items under study.	<ul style="list-style-type: none"> • Judgmental • Panel evaluation with content validity ratio
Criterion-Related	Degree to which the predictor is adequate in capturing the relevant aspects of the criterion.	<ul style="list-style-type: none"> • Correlation
Concurrent	Description of the present; criterion data are available at the same time as predictor scores.	<ul style="list-style-type: none"> • Correlation
Predictive	Prediction of the future; criterion data are measured after the passage of time.	<ul style="list-style-type: none"> • Correlation
Construct	Answers the question, "What accounts for the variance in the measure?", attempts to identify the underlying construct(s) being measured and determine how well the test represents it (them).	<ul style="list-style-type: none"> • Judgmental • Correlation of proposed test with established one • Convergent-discriminant techniques • Factor analysis • Multitrait-multimethod analysis

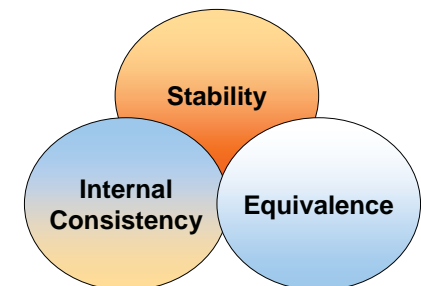
106

Understanding Validity and Reliability



107

Reliability Estimates



108

Practicality

Economy

Convenience

Interpretability

109

Chapter 12

MEASUREMENT SCALES



McGraw-Hill/Irwin

Copyright © 2014 by The McGraw-Hill Companies, Inc. All rights reserved.

Learning Objectives

Understand...

- The nature of attitudes and their relationship to behavior.
- The critical decisions involved in selecting an appropriate measurement scale.
- The characteristics and use of rating, ranking, sorting, and other preference scales.

111

Pull Quote

“No man learns to know his inmost nature by introspection, for he rates himself sometimes too low, and often too high, by his own measurement. Man knows himself only by comparing himself with other men; it is life that touches his genuine worth.”

Johann Wolfgang von Goethe
German writer, artist, politician
(1749–1832)

112

Nature of Attitudes

Cognitive

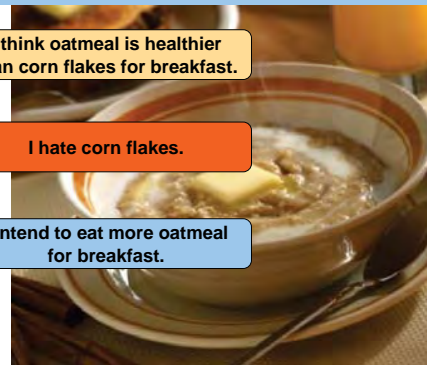
I think oatmeal is healthier than corn flakes for breakfast.

Affective

I hate corn flakes.

Behavioral

I intend to eat more oatmeal for breakfast.



113

Selecting a Measurement Scale

Research objectives

Response types

Data properties

Number of dimensions

Balanced or unbalanced

Forced or unforced choices

Number of scale points

Rater errors

114

Response Types

Rating scale

Ranking scale

Categorization

Sorting

115

Number of Dimensions

Unidimensional

Multi-dimensional



116

Balanced or Unbalanced

How good an actress is Jennifer Lawrence?

Very bad
Bad
Neither good nor bad
Good
Very good

Poor
Fair
Good
Very good
Excellent

117

Forced or Unforced Choices

How good an actress is Jennifer Lawrence?

Very bad
Bad
Neither good nor bad
Good
Very good

Very bad
Bad
Neither good nor bad
Good
Very good
No opinion
Don't know

118

Number of Scale Points

How good an actress is Jennifer Lawrence?

Very bad
Bad
Neither good nor bad
Good
Very good

Very bad
Somewhat bad
A little bad
Neither good nor bad
A little good
Somewhat good
Very good

119

Rater Errors

Error of
central tendency
Error of leniency

- Adjust strength of descriptive adjectives
- Space intermediate descriptive phrases farther apart
- Provide smaller differences in meaning between terms near the ends of the scale
- Use more scale points

120

Rater Errors

Primacy Effect
Recency Effect

Reverse order of
alternatives periodically
or randomly

121

Rater Errors

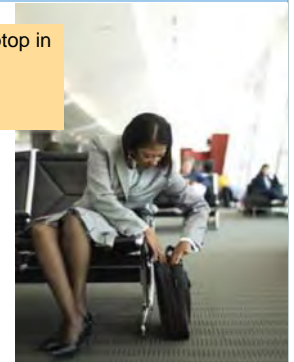
Halo Effect

- Rate one trait at a time
- Reveal one trait per page
- Reverse anchors periodically

122

Simple Category Scale

I plan to purchase a MindWriter laptop in the next 12 months.
☐ Yes
☐ No

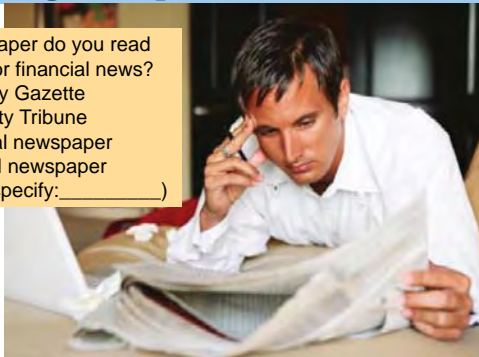


123

Multiple-Choice, Single-Response Scale

What newspaper do you read most often for financial news?

- ☐ East City Gazette
- ☐ West City Tribune
- ☐ Regional newspaper
- ☐ National newspaper
- ☐ Other (specify: _____)



124

Multiple-Choice, Multiple-Response Scale

Check any of the sources you consulted when designing your new home.

- ☐ Online planning services
- ☐ Magazines
- ☐ Independent contractor/builder
- ☐ Designer
- ☐ Architect
- ☐ Other (specify: _____)

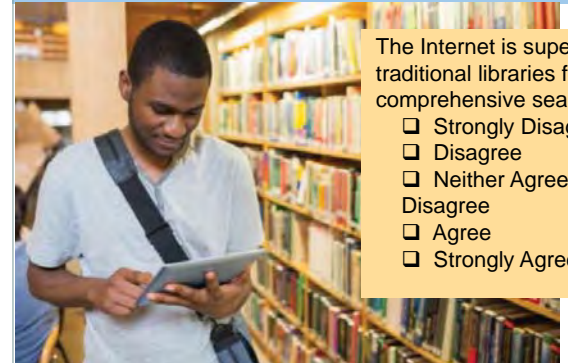


125

Likert Scale

The Internet is superior to traditional libraries for comprehensive searches.

- ☐ Strongly Disagree
- ☐ Disagree
- ☐ Neither Agree nor Disagree
- ☐ Agree
- ☐ Strongly Agree



126

Semantic Differential



Lands' End Catalog
FAST _____ SLOW
HIGH QUALITY _____ LOW QUALITY

127

Adapting SD Scales

Convenience of Reaching the Store from Your Location	
Nearby _____	Distant _____
Short time required to reach store _____	Long time required to reach store _____
Difficult drive _____	Easy Drive _____
Difficult to find parking place _____	Easy to find parking place _____
Convenient to other stores I shop _____	Inconvenient to other stores I shop _____
Products offered	
Wide selection of different kinds of products _____	Limited selection of different kinds of products _____
Fully stocked _____	Understocked _____
Undependable products _____	Dependable products _____
High quality _____	Low quality _____
Numerous brands _____	Few brands _____
Unknown brands _____	Well-known brands _____

128

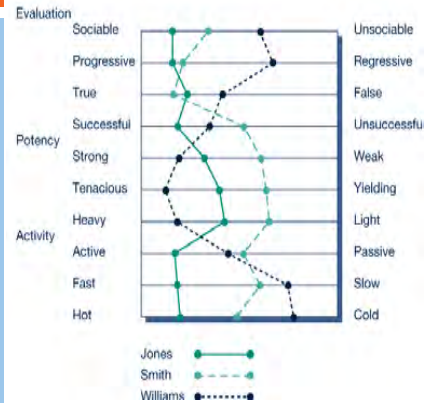
SD Scale for Analyzing Actor Candidates

Analyze (candidate) for current position:

(E) Sociable	(7): _____	(1) Unsociable
(P) Weak	(1): _____	(7) Strong
(A) Active	(7): _____	(1) Passive
(E) Progressive	(7): _____	(1) Regressive
(P) Yielding	(1): _____	(7) Tenacious
(A) Slow	(1): _____	(7) Fast
(E) True	(7): _____	(1) False
(P) Heavy	(7): _____	(1) Light
(A) Hot	(7): _____	(1) Cold
(E) Unsuccessful	(1): _____	(7) Successful

129

Graphic of SD Analysis



130

Numerical Scale

EXTREMELY FAVORABLE	5	4	3	2	1	EXTREMELY UNFAVORABLE
Employee's cooperation in teams _____						
Employee's knowledge of task _____						
Employee's planning effectiveness _____						

131

Multiple Rating List Scales

"Please indicate how important or unimportant each service characteristic is:"

	IMPORTANT					UNIMPORTANT				
Fast, reliable repair	7	6	5	4	3	2	1			
Service at my location	7	6	5	4	3	2	1			
Maintenance by manufacturer	7	6	5	4	3	2	1			
Knowledgeable technicians	7	6	5	4	3	2	1			
Notification of upgrades	7	6	5	4	3	2	1			
Service contract after warranty	7	6	5	4	3	2	1			



132

Stapel Scales

(Company Name)		
+5	+5	+5
+4	+4	+4
+3	+3	+3
+2	+2	+2
+1	+1	+1
Technology Leader	Exciting Products	World-Class Reputation
-1	-1	-1
-2	-2	-2
-3	-3	-3
-4	-4	-4
-5	-5	-5

133

Constant-Sum Scales

"Taking all the supplier characteristics we've just discussed and now considering cost, what is their relative importance to you (dividing 100 units between)":

Being one of the lowest-cost suppliers	<input type="text"/>
All other aspects of supplier performance	<input type="text"/>
Sum	100

134

Graphic Rating Scales

MindWriter

"How likely are you to recommend CompleteCare to others?" (Place an X at the position along the line that best reflects your judgment.)

VERY LIKELY VERY UNLIKELY

(alternative with graphic)

135

Ranking Scales



Paired-comparison scale

Forced ranking scale

Comparative scale

136

Paired-Comparison Scale



"For each pair of two-seat sports cars listed, place a check beside the one you would most prefer if you had to choose between the two."

<input type="checkbox"/> BMW Z4 M Coupe	<input type="checkbox"/> Chevrolet Corvette Z06
<input type="checkbox"/> Porsche Cayman S	<input type="checkbox"/> Porsche Cayman S
<input type="checkbox"/> Chevrolet Corvette Z06	<input type="checkbox"/> Dodge Viper SRT10
<input type="checkbox"/> BMW Z4 M Coupe	<input type="checkbox"/> Dodge Viper SRT10
<input type="checkbox"/> Chevrolet Corvette Z06	<input type="checkbox"/> Dodge Viper SRT10
<input type="checkbox"/> Dodge Viper SRT10	<input type="checkbox"/> BMW Z4 M Coupe

137

Forced Ranking Scale



"Rank the radar detection features in your order of preference. Place the number 1 next to the most preferred, 2 by the second choice, and so forth."

- ☐ User programming
- ☐ Cordless capability
- ☐ Small size
- ☐ Long-range warning
- ☐ Minimal false alarms

138

Comparative Scale

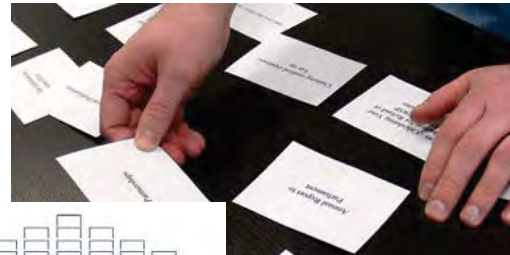


"Compared to your previous hair dryer's performance, the new one is":

SUPERIOR	ABOUT THE SAME	INFERIOR
1	2	3

139

Sorting



Example of a Q-Sort

140

CloseUp: MindWriter Scaling



Likert Scale

The problem that prompted service/repair was resolved

Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
1	2	3	4	5

Numerical Scale (MindWriter's Favorite)

To what extent are you satisfied that the problem that prompted service/repair was resolved?

Very Dissatisfied				Very Satisfied
1	2	3	4	5

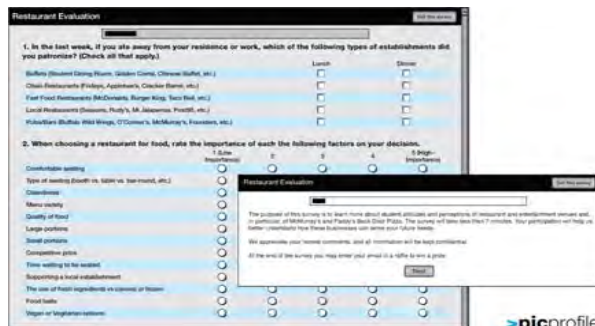
Hybrid Expectation Scale

Resolution of the problem that prompted service/repair.

Met Few Expectations	Met Some Expectations	Met Most Expectations	Met All Expectations	Exceeded Expectations
1	2	3	4	5

141

PicProfile: Online Surveys



142

Chapter 13

QUESTIONNAIRES AND INSTRUMENTS



McGraw-Hill/Irwin

Copyright © 2014 by The McGraw-Hill Companies, Inc. All rights reserved.

Learning Objectives

Understand...

- The link forged between the management dilemma and the communication instrument by the management-research question hierarchy.
- The influence of the communication method on instrument design.
- The three general classes of information and what each contributes to the instrument.

143

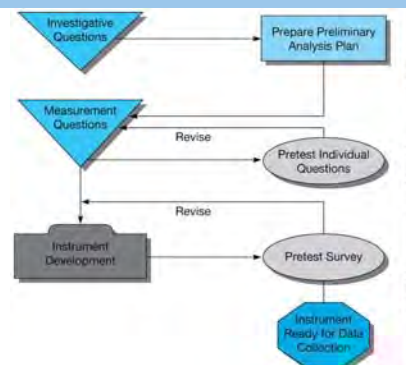
Learning Objectives

Understand . . .

- The influence of question content, question wording, response strategy, and preliminary analysis planning on question construction.
- Each of the numerous question design issues influencing instrument quality, reliability, and validity.
- The sources for measurement questions
- The importance of pretesting questions and instruments.

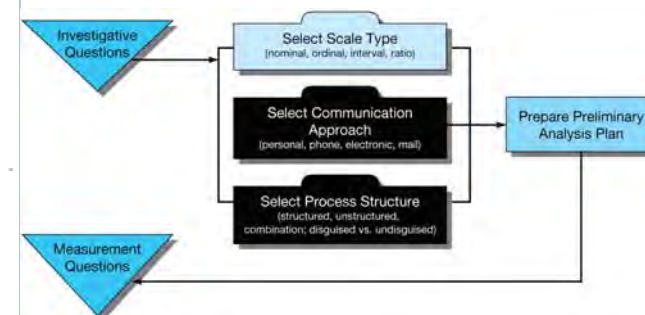
145

Overall Flowchart for Instrument Design



146

Flowchart for Instrument Design Phase 1



147

Strategic Concerns in Instrument Design

What type of scale is needed?

What communication approach will be used?

Should the questions be structured?

Should the questioning be disguised?

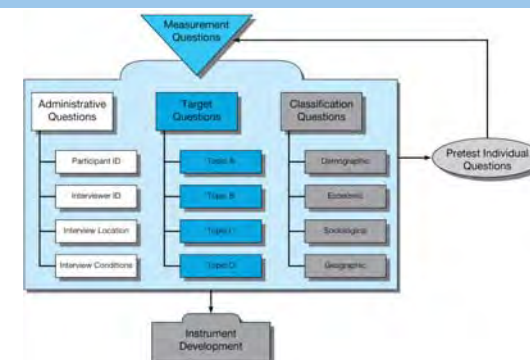
148

Factors Affecting Respondent Honesty

Syndrome	Description	Example
Peacock	Desire to be perceived as smarter, wealthier, happier, or better than others.	Respondent who claims to shop Harrods in London (twice as many as those that do).
Pleaser	Desire to help by providing answers they think the researchers want to hear, to please or avoid offending or being socially stigmatized.	Respondent gives a politically correct or assumed correct answer about degree to which they revere their elders, respect their spouse, etc.
Gamer	Adoption of answers to play the system.	Participants who take membership to a specific demographic to participate in high remuneration study; that they drive an expensive car when they don't or that they have cancer when they don't.
Disengager	Don't want to think deeply about a subject.	Falsely ad recall or purchase behavior (didn't recall or didn't buy) when they actually did.
Self-delusionist	Participants who lie to themselves.	Respondent who falsifies behavior, like the level they recycle.
Unconscious Decision Maker	Participants who are dominated by irrational decision making.	Respondent who cannot predict with any certainty his future behavior.
Ignoramus	Participant who never knew or doesn't remember an answer and makes up a lie.	Respondent who can't identify on a map where they live or remember what they ate for supper the previous evening.

149

Flowchart for Instrument Design Phase 2



150

Question Categories and Structure



Administrative

Target

Classification

151

Engagement = Convenience

"Participants are becoming more and more aware of the value of their time. The key to maintaining a quality dialog with them is to make it really convenient for them to engage, whenever and wherever they want."

Tom Anderson
managing partner
Anderson Analytics

152

Question Content

Should this question be asked?

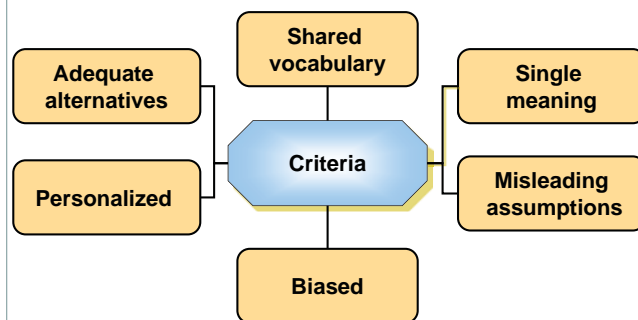
Is the question of proper scope and coverage?

Can the participant adequately answer this question as asked?

Will the participant willingly answer this question as asked?

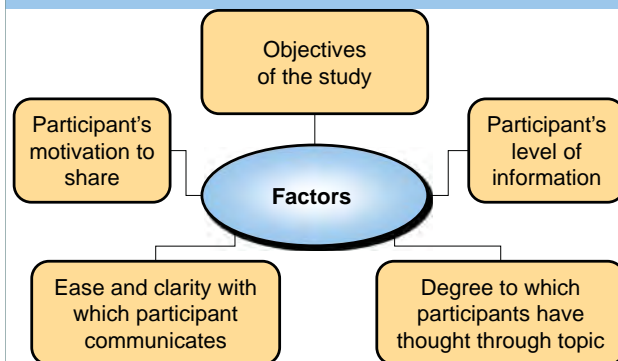
153

Question Wording



154

Response Strategy



155

Free-Response Strategy



What factors influenced your enrollment in Metro U?

156

Dichotomous Response Strategy

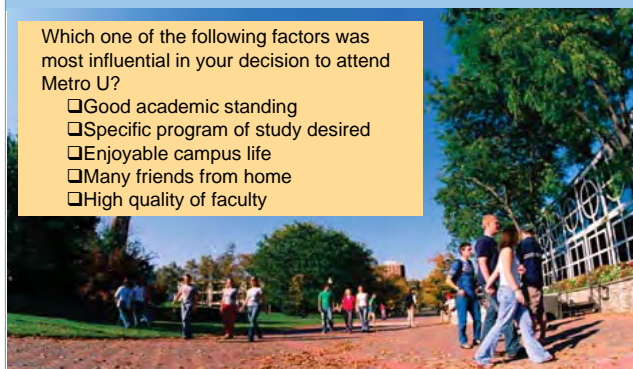


Did you attend the "A Day at College" program at Metro U?

- ☐ Yes
☐ No

157

Multiple Choice Response Strategy

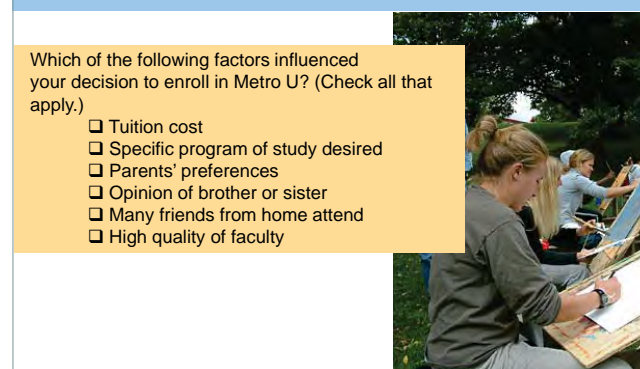


Which one of the following factors was most influential in your decision to attend Metro U?

- ☐ Good academic standing
☐ Specific program of study desired
☐ Enjoyable campus life
☐ Many friends from home
☐ High quality of faculty

158

Checklist Response Strategy



Which of the following factors influenced your decision to enroll in Metro U? (Check all that apply.)

- ☐ Tuition cost
☐ Specific program of study desired
☐ Parents' preferences
☐ Opinion of brother or sister
☐ Many friends from home attend
☐ High quality of faculty

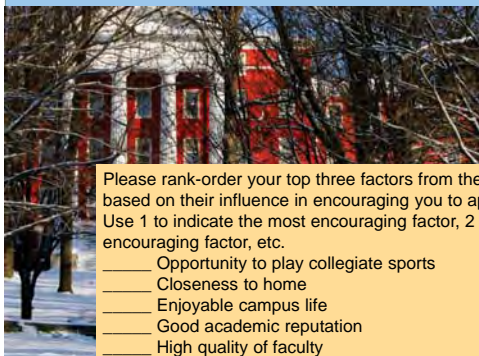
159

Rating Response Strategy

	Strongly influential	Somewhat influential	Not at all influential
Good academic reputation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Enjoyable campus life	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Many friends	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
High quality faculty	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Semester calendar	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

160

Ranking



Please rank-order your top three factors from the following list based on their influence in encouraging you to apply to Metro U. Use 1 to indicate the most encouraging factor, 2 the next most encouraging factor, etc.

____ Opportunity to play collegiate sports
____ Closeness to home
____ Enjoyable campus life
____ Good academic reputation
____ High quality of faculty

161

Summary of Scale Types

Type	Restrictions	Scale Items	Data Type
Rating Scales			
Simple Category Scale	• Needs mutually exclusive choices	One or more	Nominal
Multiple Choice Single-Response Scale	• Needs mutually exclusive choices • May use exhaustive list or "other"	Many	Nominal
Multiple Choice Multiple-Response Scale (checklist)	• Needs mutually exclusive choices • Needs exhaustive list or "other"	Many	Nominal
Likert Scale	• Needs definitive positive or negative statements with which to agree/disagree	One or more	Ordinal
Likert-type Scale	• Needs definitive positive or negative statements with which to agree/disagree	One or more	Ordinal

162

Summary of Scale Types

Type	Restrictions	Scale Items	Data Type
Rating Scales			
Numerical Scale	•Needs concepts with standardized meanings; •Needs number anchors of the scale or end-points •Score is a measurement of graphical space	One or many	Ordinal or Interval
Multiple Rating List Scale	•Needs words that are opposites to anchor the end-points on the verbal scale	Up to 10	Ordinal
Fixed Sum Scale	•Participant needs ability to calculate total to some fixed number, often 100.	Two or more	Interval or Ratio

163

Summary of Scale Types

Type	Restrictions	Scale Items	Data Type
Rating Scales			
Stapel Scale	•Needs verbal labels that are operationally defined or standard.	One or more	Ordinal or Interval
Graphic Rating Scale	•Needs visual images that can be interpreted as positive or negative anchors •Score is a measurement of graphical space from one anchor.	One or more	Ordinal (Interval, or Ratio)

164

Summary of Scale Types

Type	Restrictions	Scale Items	Data Type
Ranking Scales			
Paired Comparison Scale	• Number is controlled by participant's stamina and interest.	Up to 10	Ordinal
Forced Ranking Scale	• Needs mutually exclusive choices.	Up to 10	Ordinal or Interval
Comparative Scale	• Can use verbal or graphical scale.	Up to 10	Ordinal

165

Internet Survey Scale Options

What ONE magazine do you read most often for computing news?

Please select your answer

PC Magazine
Wired
Computing Magazine
Computing World
PC Computing
Laptop

Multiple Choice, Single Response using pull-down box

Which of the following computing magazines did you look at in the last 30 days?

PC Magazine
Wired
Computing Magazine
Computing World
PC Computing
Laptop

Checklist using checkbox (may also use radio buttons)

166

Internet Survey Scale Options

Where have you seen advertising for Mini/Write laptop computers?

Free Response/Open Question using textbox

I plan to purchase a Mini/Write laptop in the next 3 months:

Yes
No

Dichotomous Question using radio buttons (may also use pull-down box)

My next laptop computer will have:

More memory
More processing speed.

Paired Comparison using radio buttons (may also use pull-down box)

What ONE magazine do you read most often for computing news?

PC Magazine
Wired
Computing Magazine
Computing World
PC Computing
Laptop

Multiple Choice, Single Response using radio buttons (may also use pull-down box or checkbox)

167

Internet Survey Scale Options

Please indicate the importance of each of the characteristics in choosing your next laptop. (Select one answer in each row. Scroll to see the complete list of options.)

	Very Important	Neutral Important/Not Important	Not at all Important
Fast reliable repair service	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Service at my location	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Maintenance by the manufacturer	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Knowledgeable technicians	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Notification of upgrades	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Rating Grid (may also use checkboxes) Requires a single response per line. The longer the list, the more likely the participant must scroll.

From the list below, please choose the three most important service options when choosing your next laptop.

Fast reliable repair service
Service at my location
Maintenance by the manufacturer
Knowledgeable technicians
Notification of upgrades

Ranking Question using pull-down box (may also use textboxes, in which ranks are entered) [This question asks for a limited ranking of only three of the listed elements.]

1
2
3

168

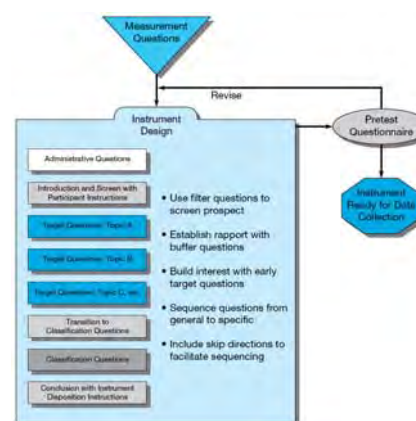
Sources of Questions

- Handbook of Marketing Scales
- The Gallup Poll Cumulative Index
- Measures of Personality and Social-Psychological Attitudes
- Measures of Political Attitudes

- Index to International Public Opinion
- Sourcebook of Harris National Surveys
- Marketing Scales Handbook
- American Social Attitudes Data Sourcebook

169

Flowchart for Instrument Design: Phase 3



170

Guidelines for Question Sequencing

- Interesting topics early
- Simple topics early
- Sensitive questions later
- Classification questions later
- Transition between topics
- Reference changes limited

171

PicProfile: Branching Question

2. Which of the following attributes do you like about the automobile you just saw? (Select all that apply)

- ☒ Overall appeal
- ☒ Headroom
- ☐ Design
- ☐ Color
- ☒ Height from the ground
- ☐ Other
- ☐ None of the above

Next Question

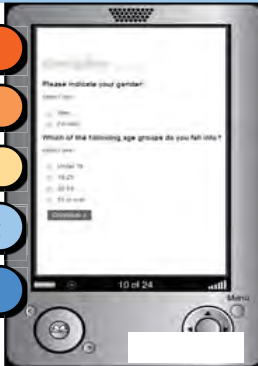
3. For those items that you selected, how important is each? (Provide one answer for each attribute)

	Extremely Important	Very Important	Neither Important nor not important	Not at all Important	Don't know
a) Overall appeal	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
b) Height from the ground	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
c) Headroom	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

172

Snapshot: Mobile Questionnaires

- 10 or fewer questions
- Simple question modes
- Minimize scrolling
- Minimize non-essential content
- Minimize distraction



173

Research Thought Leader

“Research that asks consumers what they did and why is incredibly helpful. Research that asks consumers what they are going to do can often be taken with a grain of salt.”

Al Ries
author, co-founder, and chairman
Ries & Ries.

174

Chapter 14

SAMPLING



McGraw-Hill/Irwin

Copyright © 2014 by The McGraw-Hill Companies, Inc. All rights reserved.

175

Learning Objectives

Understand . . .

- The two premises on which sampling theory is based.
- The accuracy and precision for measuring sample validity.
- The five questions that must be answered to develop a sampling plan.

Learning Objectives

Understand . . .

- The two categories of sampling techniques and the variety of sampling techniques within each category.
- The various sampling techniques and when each is used.

176

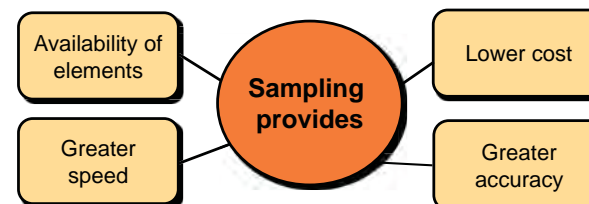
The Nature of Sampling

- Population
- Population Element
- Sampling Frame
- Census
- Sample



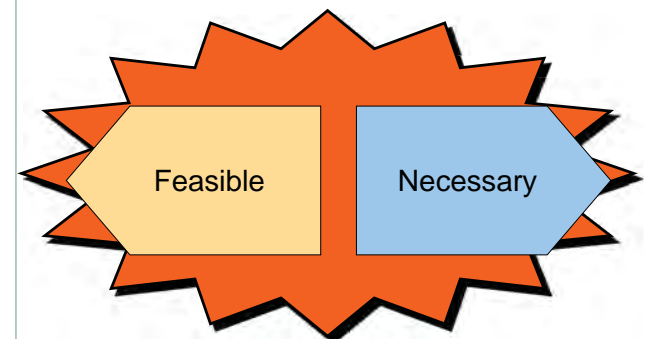
178

Why Sample?



179

When Is a Census Appropriate?



180

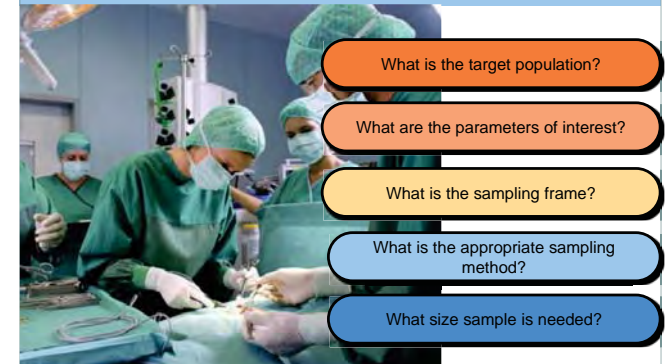
What Is a Valid Sample?



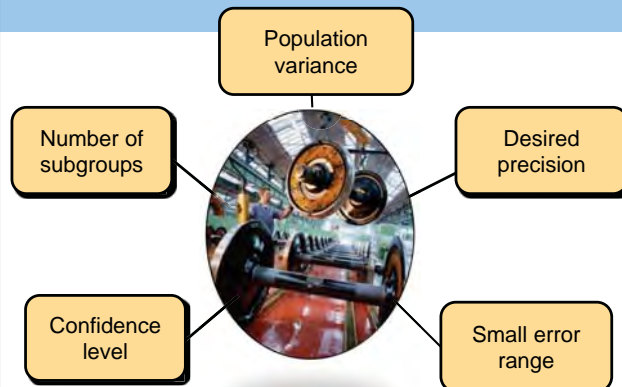
Types of Sampling Designs

Element Selection	Probability	Nonprobability
• Unrestricted	• Simple random	• Convenience
• Restricted	• Complex random	• Purposive
	• Systematic	• Judgment
	• Cluster	• Quota
	• Stratified	• Snowball
	• Double	

Steps in Sampling Design



When to Use Larger Sample?



Simple Random

Advantages

- Easy to implement with random dialing

Disadvantages

- Requires list of population elements
- Time consuming
- Larger sample needed
- Produces larger errors
- High cost

Systematic

Advantages

- Simple to design
- Easier than simple random
- Easy to determine sampling distribution of mean or proportion

Disadvantages

- Periodicity within population may skew sample and results
- Trends in list may bias results
- Moderate cost

Stratified

Advantages

- Control of sample size in strata
- Increased statistical efficiency
- Provides data to represent and analyze subgroups
- Enables use of different methods in strata

Disadvantages

- Increased error if subgroups are selected at different rates
- Especially expensive if strata on population must be created
- High cost

Cluster

Advantages

- Provides an unbiased estimate of population parameters if properly done
- Economically more efficient than simple random
- Lowest cost per sample
- Easy to do without list

Disadvantages

- Often lower statistical efficiency due to subgroups being homogeneous rather than heterogeneous
- Moderate cost

Stratified and Cluster Sampling

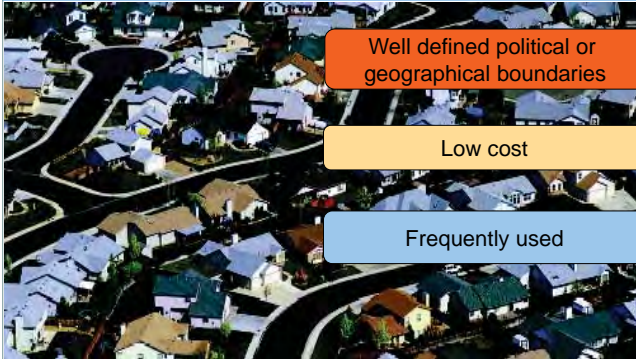
Stratified

- Population divided into few subgroups
- Homogeneity within subgroups
- Heterogeneity between subgroups
- Choice of elements from within each subgroup

Cluster

- Population divided into many subgroups
- Heterogeneity within subgroups
- Homogeneity between subgroups
- Random choice of subgroups

Area Sampling



190

Double Sampling

Advantages

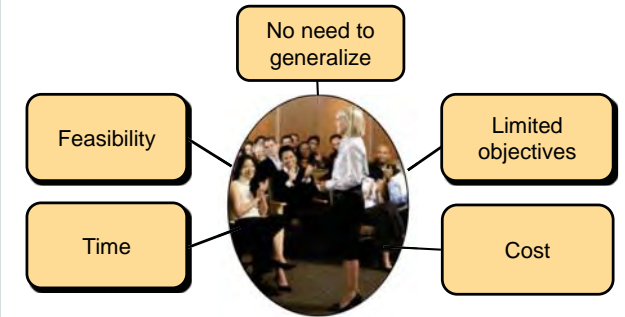
- May reduce costs if first stage results in enough data to stratify or cluster the population

Disadvantages

- Increased costs if discriminately used

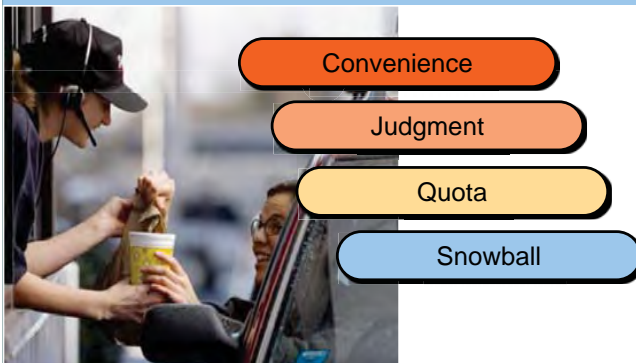
191

Nonprobability Samples



192

Nonprobability Sampling Methods



193

Appendix 15a

Describing Data Statistically



15

194

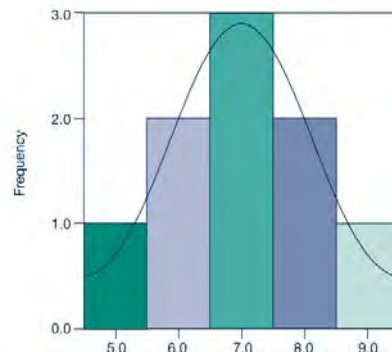
Frequencies

	Unit Sales Increase (%)	Frequency	Percentage	Cumulative Percentage
	5	1	11.1	11.1
	6	2	22.2	33.3
	7	3	33.3	66.7
	8	2	22.2	88.9

B	Unit Sales Increase (%)	Frequency	Percentage	Cumulative Percentage
Origin, foreign (1)	6	1	11.1	11.1
	7	2	22.2	33.3
	8	2	22.2	55.5
Origin, foreign (2)	5	1	11.1	66.6
	6	1	11.1	77.7
	7	1	11.1	88.8
	9	1	11.1	100.0
Total		9	100.0	

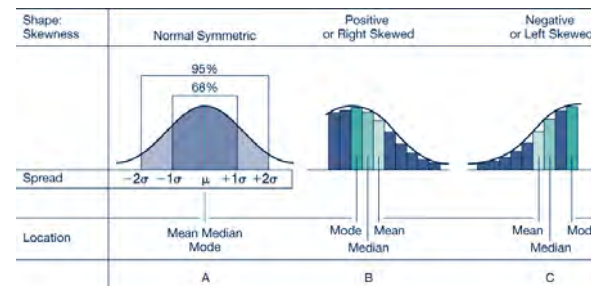
App 15a-195

Distributions



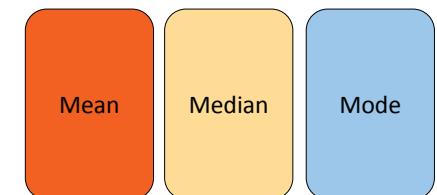
App 15a-196

Characteristics of Distributions



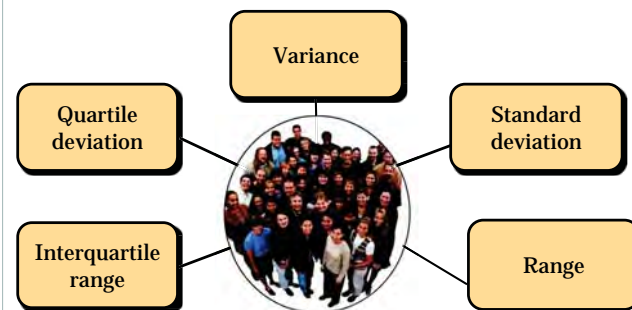
App 15a-197

Measures of Central Tendency



App 15a-198

Measures of Variability



App 14a-199

Summarizing Distribution Shape



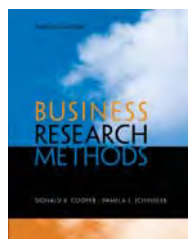
App 14a-200

Symbols

Variable	Population	Sample
Mean	μ	\bar{x}
Proportion	Π	p
Variance	σ^2	s^2
Standard deviation	σ	s
Size	N	n
Standard error of the mean	$\sigma_{\bar{x}}$	$S_{\bar{x}}$
Standard error of the proportion	σ_p	S_p

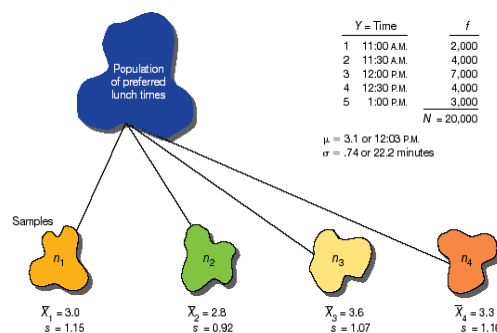
App 14a-201

Appendix 14a Determining Sample Size



202

Random Samples



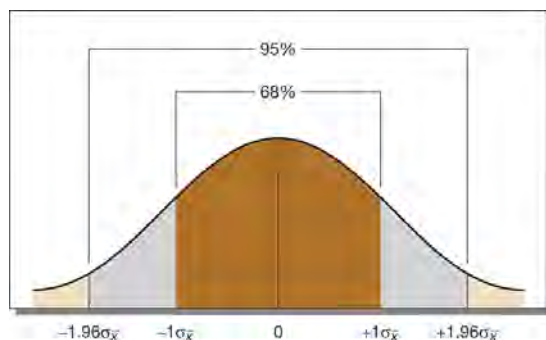
203

Increasing Precision

Reducing the Standard Deviation by 50%	Quadrupling the Sample
$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$ $\sigma_{\bar{x}} = \frac{0.74}{\sqrt{10}} = 0.234$ $\sigma_{\bar{x}} = \frac{0.37}{\sqrt{10}} = 0.117$	$\sigma_{\bar{x}} = \frac{0.8}{\sqrt{25}} = 0.16$ $\sigma_{\bar{x}} = \frac{0.8}{\sqrt{100}} = 0.08$
<p>where:</p> <p>$\sigma_{\bar{x}}$ = standard error of the mean</p> <p>s = standard deviation of the sample</p> <p>n = sample size</p> <p>Note: A 400 percent increase in sample size (from 25 to 100) would yield only a 200 percent increase in precision (from 0.16 to 0.08). Researchers are often asked to increase precision, but the question should be, at what cost? Each of those additional sample elements adds both time and cost to the study.</p>	

App 14a-204

Confidence Levels & the Normal Curve



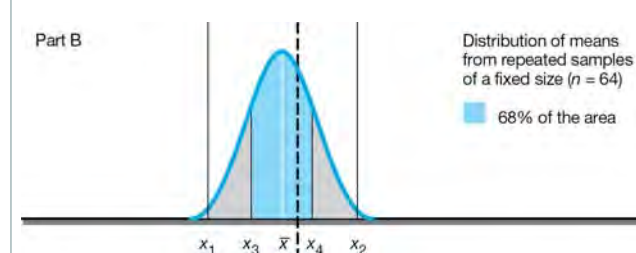
App 14a-205

Standard Errors

Standard Error (Z score)	% of Area	Approximate Degree of Confidence
1.00	68.27	68%
1.65	90.10	90%
1.96	95.00	95%
3.00	99.73	99%

App 14a-206

Central Limit Theorem



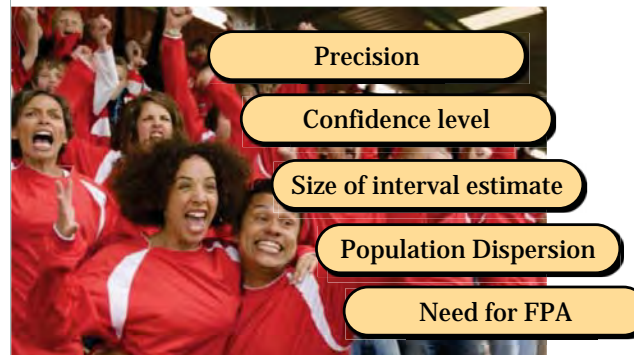
App 14a-207

Estimates of Dining Visits

Confidence	Z score	% of Area	Interval Range (visits per month)
68%	1.00	68.27	9.48-10.52
90%	1.65	90.10	9.14-10.86
95%	1.96	95.00	8.98-11.02
99%	3.00	99.73	8.44-11.56

App 14a-208

Calculating Sample Size for Questions Involving Means



App 14a-209

Metro U Sample Size for Means

Steps	Information
Desired confidence level	95% ($z = 1.96$)
Size of the interval estimate	$\pm .5$ meals per month
Expected range in population	0 to 30 meals
Sample mean	10
Standard deviation	4.1
Need for finite population adjustment	No
Standard error of the mean	$.5/1.96 = .255$
Sample size	$(4.1)^2 / (.255)^2 = 259$

App 14a-210

Proxies of the Population Dispersion

Previous Research

Pilot or Pretest

Rule-of-thumb: 1/6 of range



App 14a-211

Metro U Sample Size for Proportions

Steps	Information
Desired confidence level	95% ($z = 1.96$)
Size of the interval estimate	$\pm .10$ (10%)
Expected range in population	0 to 100%
Sample proportion with given attribute	30%
Sample dispersion	$Pq = .30(1-.30) = .21$
Finite population adjustment	No
Standard error of the proportion	$.10/1.96 = .051$
Sample size	$.21 / (.051)^2 = 81$

App 14a-212

Chapter 17

HYPOTHESIS TESTING



McGraw-Hill/Irwin

Copyright © 2014 by The McGraw-Hill Companies, Inc. All rights reserved.

Learning Objectives

Understand . . .

- The nature and logic of hypothesis testing.
- A statistically significant difference
- The six-step hypothesis testing procedure.

214

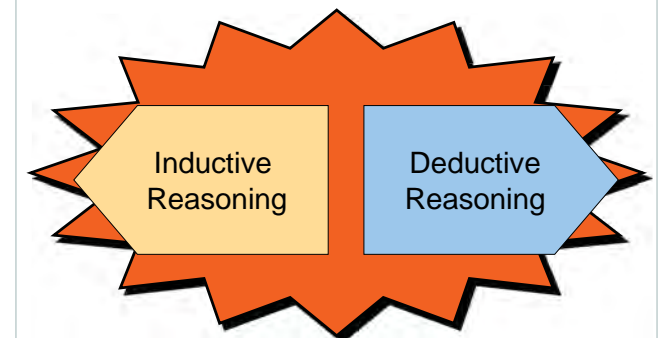
Learning Objectives

Understand . . .

- The differences between parametric and nonparametric tests and when to use each.
- The factors that influence the selection of an appropriate test of statistical significance.
- How to interpret the various test statistics

215

Hypothesis Testing



216

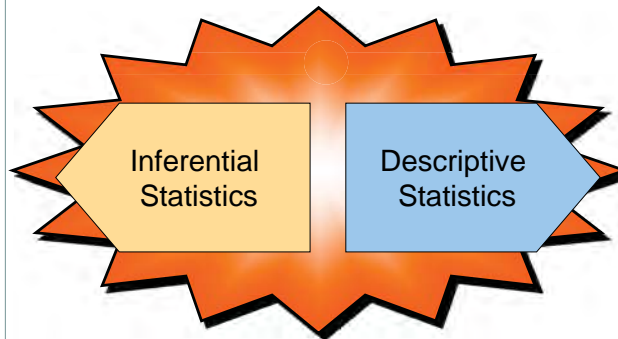
Hypothesis Testing Finds Truth

"One finds the truth by making a hypothesis and comparing the truth to the hypothesis."

David Douglass
physicist
University of Rochester

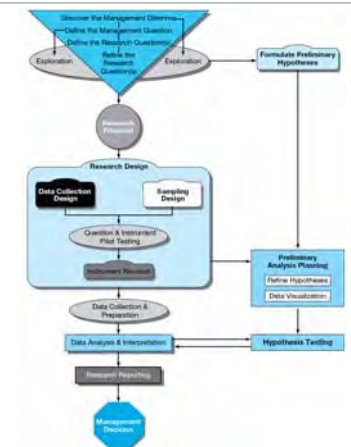
217

Statistical Procedures



218

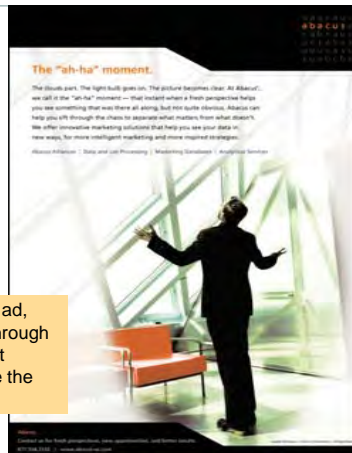
Hypothesis Testing and the Research Process



219

When Data Present a Clear Picture

As Abacus states in this ad, when researchers 'sift through the chaos' and 'find what matters' they experience the "ah ha!" moment.



220

Approaches to Hypothesis Testing

Classical statistics

- Objective view of probability
- Established hypothesis is rejected or fails to be rejected
- Analysis based on sample data

Bayesian statistics

- Extension of classical approach
- Analysis based on sample data
- Also considers established subjective probability estimates

221

Statistical Significance



222

Types of Hypotheses

Null

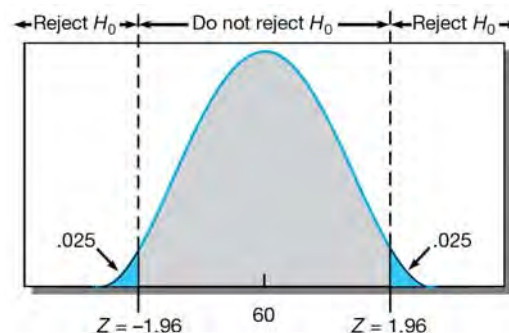
- $H_0: \mu = 50$ mpg
- $H_0: \mu \leq 50$ mpg
- $H_0: \mu \geq 50$ mpg

Alternate

- $H_A: \mu \neq 50$ mpg
- $H_A: \mu > 50$ mpg
- $H_A: \mu < 50$ mpg

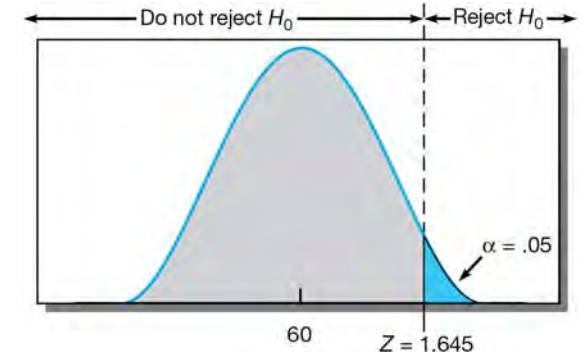
223

Two-Tailed Test of Significance



224

One-Tailed Test of Significance



225

Decision Rule

Take no corrective action if the analysis shows that one **cannot reject** the null hypothesis.

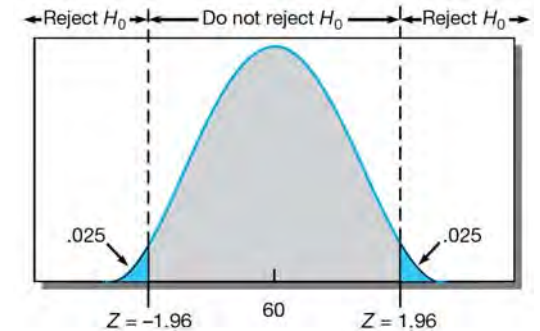
226

Statistical Decisions

		State of Nature	
		H_0 is true	H_A is true
Decision: Accept H_0	Correct decision Power of test Probability = $1 - \alpha$	Innocent of crime Found not guilty	Guilty of crime Unjustly acquitted
	Type II error Power of test Probability = β	Innocent Unjustly convicted	Guilty Justly convicted
Decision: Accept H_A	Type I error Significance level Probability = α	Innocent of crime Found not guilty	Guilty of crime Unjustly acquitted
	Correct decision Power of test Probability = $1 - \beta$	Innocent Unjustly convicted	Guilty Justly convicted

227

Probability of Making a Type I Error



228

Critical Values

$Z = 1.96$ (significance level = .05)
 \bar{X}_c = the critical value of the sample mean
 μ = the population value stated in $H_0 = 50$
 $\sigma_{\bar{X}}$ = the standard error of a distribution of means of samples of 25

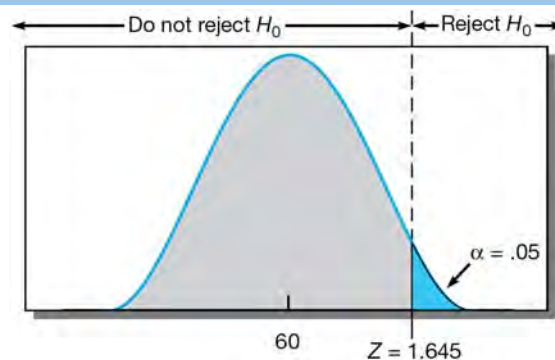
$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

$$-1.96 = \frac{\bar{X}_c - 50}{2} \quad 1.96 = \frac{\bar{X}_c - 50}{2}$$

$$\bar{X}_c = 46.08 \quad \bar{X}_c = 53.92$$

229

Probability of Making A Type I Error



230

Factors Affecting Probability of Committing a β Error

True value of parameter

Alpha level selected

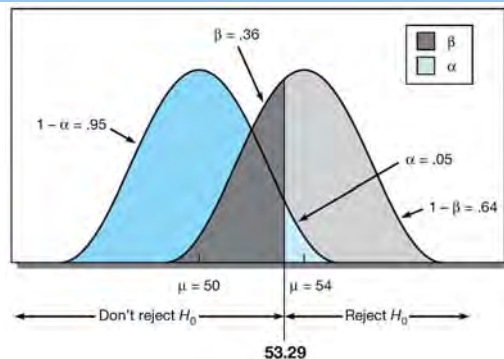
One or two-tailed test used

Sample standard deviation

Sample size

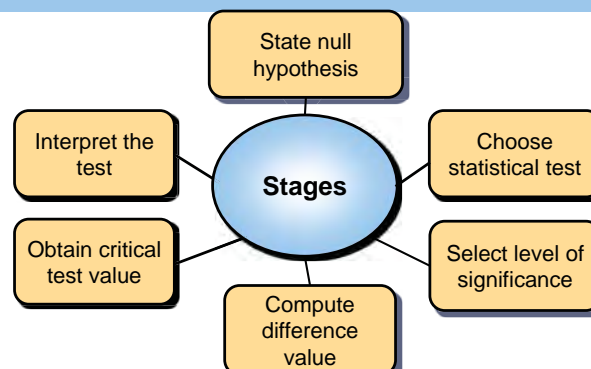
231

Probability of Making A Type II Error



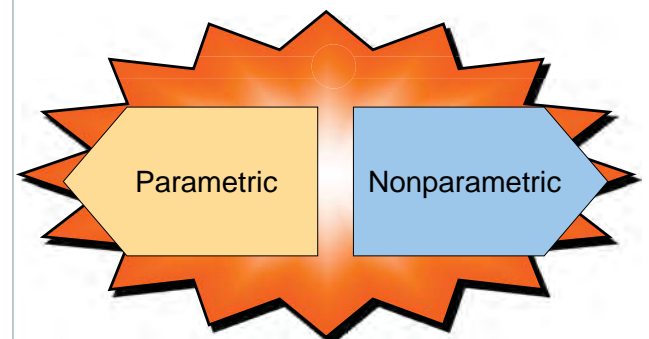
232

Statistical Testing Procedures



233

Tests of Significance



234

Assumptions for Using Parametric Tests

Independent observations

Normal distribution

Equal variances

Interval or ratio scales

215

Advantages of Nonparametric Tests

Easy to understand and use

Usable with nominal data

Appropriate for ordinal data

Appropriate for non-normal population distributions

216

How to Select a Test

How many samples are involved?

If two or more samples:
are the individual cases independent or related?

Is the measurement scale
nominal, ordinal, interval, or ratio?

217

Recommended Statistical Techniques

Measureme nt Scale	One-Sample Case	Two-Sample Tests		k-Sample Tests	
		Related Samples	Independent Samples	Related Samples	Independent Samples
Nominal	• Binomial • χ^2 one-sample test	• McNemar	• Fisher exact test • χ^2 two- samples test	• Cochran Q	• χ^2 for k samples
Ordinal	• Kolmogorov- Smirnov one- sample test • Runs test	• Sign test • Wilcoxon matched- pairs test	• Median test • Mann- Whitney U • Kolmogorov- Smirnov • Wald- Wolfowitz	• Friedman two-way ANOVA	• Median extension • Kruskal- Wallis one- way ANOVA
Interval and Ratio	• t-test • Z test	• t-test for paired samples	• t-test • Z test	• Repeated- measures ANOVA	• One-way ANOVA • n-way ANOVA

218

Questions Answered by One-Sample Tests

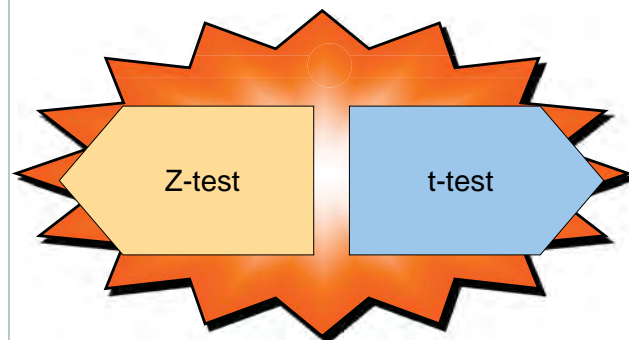
Is there a difference between observed frequencies and the frequencies we would expect?

Is there a difference between observed and expected proportions?

Is there a significant difference between some measures of central tendency and the population parameter?

219

Parametric Tests



240

One-Sample t-Test Example

Null	$H_0: = 50 \text{ mpg}$
Statistical test	t-test
Significance level	.05, $n=100$
Calculated value	1.786
Critical test value	1.66 (from Appendix C, Exhibit C-2)

241

One Sample Chi-Square Test Example

Living Arrangement	Intend to Join	Number Interviewed	Percent (no. interviewed/200)	Expected Frequencies (percent x 60)
Dorm/fraternity	16	90	45	27
Apartment/rooming house, nearby	13	40	20	12
Apartment/rooming house, distant	16	40	20	12
Live at home	15	30	15	9
Total	60	200	100	60

242

Two-Sample Parametric Tests

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)0}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

243

Two-Sample t-Test Example

	A Group	B Group
Average hourly sales	$X_1 = \$1,500$	$X_2 = \$1,300$
Standard deviation	$s_1 = 225$	$s_2 = 251$

244

Two-Sample t-Test Example

Null	$H_0: A \text{ sales} = B \text{ sales}$
Statistical test	t-test
Significance level	.05 (one-tailed)
Calculated value	1.97, d.f. = 20
Critical test value	1.725 (from Appendix C, Exhibit C-2)

245

Two-Sample Nonparametric Tests: Chi-Square

		On-the-Job-Accident		
	Cell Designation Count Expected Values	Yes	No	Row Total
Smoker	Heavy Smoker	1,1	1,2	16
		12, 8.24	4 7.75	
		2,1	2,2	
	Moderate	9 7.73	6 7.27	15
		3,1	3,2	
		13 18.03	22 16.97	
	Nonsmoker			35
	Column Total	34	32	66

246

Two-Sample Chi-Square Example

Null	There is no difference in distribution channel for age categories.
Statistical test	Chi-square
Significance level	.05
Calculated value	6.86, d.f. = 2
Critical test value	5.99 (from Appendix C, Exhibit C-3)

247

SPSS Cross-Tabulation Procedure

INCOME BY POSSESSION OF MBA				
INCOME	MBA		Row Total	
	Yes	No		
High	30	30	60	
Low	30	30	60	
Column Total	60	60	120	
Chi-Square	Value	D.F.	Significance	
Pearson	6.25000	1	.01242	
Continuity Correction	5.25174	1	.02142	
Likelihood Ratio	6.43786	1	.01117	
Nagelkerke	6.18750	1	.01287	
Minimum Expected Frequency: 16.000				

248

k-Independent-Samples Tests: ANOVA

Tests the null hypothesis that the means of three or more populations are equal.

One-way: Uses a single-factor, fixed-effects model to compare the effects of a treatment or factor on a continuous dependent variable.

249

ANOVA Example

Model Summary					
Source	d.f.	Sum of Squares	Mean Square	F Value	p Value
Model (airline)	2	11644.033	5822.017	28.304	0.0001
Residual (error)	57	11724.550	205.694		
Total	59	23368.583			

Means Table				
	Count	Mean	Std. Dev.	Std. Error
Lufthansa	20	38.950	14.006	3.132
Malaysia Airlines	20	58.900	15.089	3.374
Cathay Pacific	20	72.900	13.902	3.108

All data are hypothetical

250

ANOVA Example Continued

Null	$\mu A1 = \mu A2 = \mu A3$
Statistical test	ANOVA and F ratio
Significance level	.05
Calculated value	28.304, d.f. = 2, 57
Critical test value	3.16 (from Appendix C, Exhibit C-9)

251

Contents

Preface to the first edition	xi
Preface to the second edition	xiii
Chapter 1 Introduction	1
Organization of the book	3
Useful background	4
Appendix 1.1: Mathematical concepts used in this book	4
Endnote	7
Chapter 2 Basic data handling	9
Types of economic data	9
Obtaining data	14
Working with data: graphical methods	18
Working with data: descriptive statistics	23
Chapter summary	25
Appendix 2.1: Index numbers	25
Appendix 2.2: Advanced descriptive statistics	31
Endnotes	32
Chapter 3 Correlation	35
Understanding correlation	35
Understanding correlation through verbal reasoning	36
Understanding why variables are correlated	39
Understanding correlation through XY -plots	42
Correlation between several variables	45

	Chapter summary	46
	Appendix 3.1: Mathematical details	46
	Endnotes	47
Chapter 4	An introduction to simple regression	49
	Regression as a best fitting line	50
	Interpreting OLS estimates	54
	Fitted values and R^2 : measuring the fit of a regression model	57
	Nonlinearity in regression	61
	Chapter summary	65
	Appendix 4.1: Mathematical details	65
	Endnotes	67
Chapter 5	Statistical aspects of regression	69
	Which factors affect the accuracy of the estimate $\hat{\beta}$?	70
	Calculating a confidence interval for β	73
	Testing whether $\beta = 0$	79
	Hypothesis testing involving R^2 : the F -statistic	84
	Chapter summary	86
	Appendix 5.1: Using statistical tables for testing whether $\beta = 0$	87
	Endnotes	88
Chapter 6	Multiple regression	91
	Regression as a best fitting line	92
	Ordinary least squares estimation of the multiple regression model	93
	Statistical aspects of multiple regression	93
	Interpreting OLS estimates	94
	Pitfalls of using simple regression in a multiple regression context	97
	Omitted variables bias	99
	Multicollinearity	100
	Chapter summary	106
	Appendix 6.1: Mathematical interpretation of regression coefficients	106
	Endnotes	107
Chapter 7	Regression with dummy variables	109
	Simple regression with a dummy variable	111
	Multiple regression with dummy variables	112
	Multiple regression with both dummy and non-dummy explanatory variables	115

	Interacting dummy and non-dummy variables	117
	What if the dependent variable is a dummy?	119
	Chapter summary	120
	Endnote	120
Chapter 8	Regression with time lags: distributed lag models	121
	Aside on lagged variables	123
	Aside on notation	125
	Selection of lag order	128
	Chapter summary	131
	Appendix 8.1: Other distributed lag models	131
	Endnotes	133
Chapter 9	Univariate time series analysis	135
	The autocorrelation function	138
	The autoregressive model for univariate time series	142
	Nonstationary versus stationary time series	145
	Extensions of the AR(1) model	146
	Testing in the AR(p) with deterministic trend model	151
	Chapter summary	155
	Appendix 9.1: Mathematical intuition for the AR(1) model	156
	Endnotes	157
Chapter 10	Regression with time series variables	159
	Time series regression when X and Y are stationary	160
	Time series regression when Y and X have unit roots: spurious regression	164
	Time series regression when Y and X have unit roots: cointegration	165
	Time series regression when Y and X are cointegrated: the error correction model	171
	Time series regression when Y and X have unit roots but are not cointegrated	175
	Chapter summary	176
	Endnotes	177
Chapter 11	Applications of time series methods in macroeconomics and finance	179
	Volatility in asset prices	179
	Granger causality	186
	Vector autoregressions	193
	Chapter summary	205

	Appendix 11.1: Hypothesis tests involving more than one coefficient	205
	Endnotes	209
Chapter 12	Limitations and extensions	211
	Problems that occur when the dependent variable has particular forms	212
	Problems that occur when the errors have particular forms	213
	Problems that call for the use of multiple equation models	216
	Chapter summary	220
	Endnotes	220
Appendix A	Writing an empirical project	223
	Description of a typical empirical project	223
	General considerations	225
	Project topics	226
Appendix B	Data directory	229
Index		233

CHAPTER 1

Introduction

There are several types of professional economists working in the world today. Academic economists in universities often derive and test theoretical models of various aspects of the economy. Economists in the civil service often study the merits and demerits of policies under consideration by government. Economists employed by a central bank often give advice on whether or not interest rates should be raised, while in the private sector, economists often predict future variables such as exchange rate movements and their effect on company exports.

For all of these economists, the ability to work with data is an important skill. To decide between competing theories, to predict the effect of policy changes, or to forecast what may happen in the future, it is necessary to appeal to facts. In economics, we are fortunate in having at our disposal an enormous amount of facts (in the form of “data”) that we can analyze in various ways to shed light on many economic issues.

The purpose of this book is to present the basics of data analysis in a simple, non-mathematical way, emphasizing graphical and verbal intuition. It focusses on the tools that economists apply in practice (primarily regression) and develops computer skills that are necessary in virtually any career path that the economics student may choose to follow.

To explain further what this book does, it is perhaps useful to begin by discussing what it does **not** do. **Econometrics** is the name given to the study of quantitative tools for analyzing economic data. The field of econometrics is based on probability and statistical theory; it is a fairly mathematical field. This book does not attempt to teach probability and statistical theory. Neither does it contain much mathematical content. In both these respects, it represents a clear departure from traditional econometrics textbooks. Yet, it aims to teach most of the practical tools used by applied econometricians today.

Books that merely teach the student which buttons to press on a computer without providing an understanding of what the computer is doing, are commonly referred to as “cookbooks”. The present book is **not** a cookbook. Some econometricians may interject at this point: “But how can a book teach the student to use the tools of econometrics, without teaching the basics of probability and statistics?” My answer is that much of what the econometrician does in practice can be understood intuitively, without resorting to probability and statistical theory. Indeed, it is a contention of this book that most of the tools econometricians use can be mastered simply through a thorough understanding of the concept of correlation, and its generalization, regression. If a student understands correlation and regression well, then he/she can understand most of what econometricians do. In the vast majority of cases, it can be argued that regression will reveal most of the information in a data set. Furthermore, correlation and regression are fairly simple concepts that can be understood through verbal intuition or graphical methods. They provide the basis of explanation for more difficult concepts, and can be used to analyze many types of economic data.

This book focusses on the **analysis** of economic data; **it is not a book about collecting economic data**. With some exceptions, it treats the data as given, and does not explain how the data is collected or constructed. For instance, it does not explain how national accounts are created or how labor surveys are designed. It simply teaches the reader to make sense out of the data that has been gathered.

Statistical theory usually proceeds from the formal definition of general concepts, followed by a discussion of how these concepts are relevant to particular examples. The present book attempts to do the opposite. That is, **it attempts to motivate general concepts through particular examples**. In some cases formal definitions are not even provided. For instance, P-values and confidence intervals are important statistical concepts, providing measures relating to the accuracy of a fitted regression line (see Chapter 5). The chapter uses examples, graphs and verbal intuition to demonstrate how they might be used in practice. But no formal definition of a P-value nor derivation of a confidence interval is ever given. This would require the introduction of probability and statistical theory, which is not necessary for using these techniques sensibly in practice. For the reader wishing to learn more about the statistical theory underlying the techniques, many books are available; for instance *Introductory Statistics for Business and Economics* by Thomas Wonnacott and Ronald Wonnacott (Fourth edition, John Wiley & Sons, 1990). For those interested in how statistical theory is applied in econometric modeling, *Undergraduate Econometrics* by R. Carter Hill, William E. Griffiths and George G. Judge (Second edition, John Wiley & Sons, 2000) provides a useful introduction.

This book reflects my belief that the use of concrete examples is the best way to teach data analysis. Appropriately, each chapter presents several examples as a means of illustrating key concepts. One risk with such a strategy is that some students might

interpret the presence of so many examples to mean that myriad concepts must be mastered before they can ever hope to become adept at the practice of econometrics. This is not the case. At the heart of this book are only a few basic concepts, and they appear repeatedly in a variety of different problems and data sets. The best approach for teaching introductory econometrics, in other words, is to illustrate its specific concepts over and over again in a variety of contexts.

Organization of the book

In organizing the book, I have attempted to adhere to the general philosophy outlined above. Each chapter covers a topic and includes a general discussion. However, most of the chapter is devoted to empirical examples that illustrate and, in some cases, introduce important concepts. Exercises, which further illustrate these concepts, are included in the text. Data required to work through the empirical examples and exercises can be found in the website which accompanies this book <http://www.wileyeurope.com/go/koopdata2ed>. By including real-world data, it is hoped that students will not only replicate the examples, but will feel comfortable extending and/or experimenting with the data in a variety of ways. Exposure to real-world data sets is essential if students are to master the conceptual material and apply the techniques covered in this book.

The empirical examples in this book are designed for use in conjunction with the computer package Excel. The website associated with this book contains Excel files. Excel is a simple and common software package. It is also one that students are likely to use in their economic careers. However, the data can be analyzed using many other computer software packages, not just Excel. Many of these packages recognize Excel files and the data sets can be imported directly into them. Alternatively, the website also contains all of the data files in ASCII text form. Appendix B at the end of the book provides more detail about the data.

Mathematical material has been kept to a minimum throughout this book. In some cases, a little bit of mathematics will provide additional intuition. For students familiar with mathematical techniques, appendices have been included at the end of some chapters. However, students can choose to omit this material without any detriment to their understanding of the basic concepts.

The content of the book breaks logically into two parts. Chapters 1–7 cover all the basic material relating to graphing, correlation and regression. A very short course would cover only this material. Chapters 8–12 emphasize time series topics and analyze some of the more sophisticated econometric models in use today. The focus on the underlying intuition behind regression means that this material should be easily accessible to students. Nevertheless, students will likely find that these latter chapters are more difficult than Chapters 1–7.

Useful background

As mentioned, this book assumes very little mathematical background beyond the pre-university level. Of particular relevance are:

1. Knowledge of simple equations. For instance, the equation of a straight line is used repeatedly in this book.
2. Knowledge of simple graphical techniques. For instance, this book is full of graphs that plot one variable against another (i.e. standard XY -graphs).
3. Familiarity with the summation operator is useful occasionally.
4. In a few cases, logarithms are used.

For the reader unfamiliar with these topics, the appendix at the end of this chapter provides a short introduction. In addition, these topics are discussed elsewhere, in many introductory mathematical textbooks.

This book also has a large computer component, and much of the computer material is explained in the text. There are myriad computer packages that could be used to implement the procedures described in this book. In the places where I talk directly about computer programs, I will use the language of spreadsheets and, particularly, that most common of spreadsheets, Excel. I do this largely because the average student is more likely to have knowledge of and access to a spreadsheet rather than a specialized statistics or econometrics package such as E-views, Stata or MicroFit.¹ I assume that the student knows the basics of Excel (or whatever computer software package he/she is using). In other words, students should understand the basics of spreadsheet terminology, be able to open data sets, cut, copy and paste data, etc. If this material is unfamiliar to the student, simple instructions can be found in Excel's on-line documentation. For computer novices (and those who simply want to learn more about the computing side of data analysis) *Computing Skills for Economists* by Guy Judge (John Wiley & Sons, 2000) is an excellent place to start.

Appendix 1.1: Mathematical concepts used in this book

This book uses very little mathematics, relying instead on intuition and graphs to develop an understanding of key concepts (including understanding how to interpret the numbers produced by computer programs such as Excel). For most students, previous study of mathematics at the pre-university level should give you all the background knowledge you need. However, here is a list of the concepts used in this book along with a brief description of each.

The equation of a straight line

Economists are often interested in the relationship between two (or more) variables. Examples of variables include house prices, gross domestic product (GDP), interest rates, etc. In our context a variable is something the economist is interested in and can collect data on. I use capital letters (e.g. Y or X) to denote variables. A very general way of denoting a relationship is through the concept of a function. A common mathematical notation for a function of X is $f(X)$. So, for instance, if the economist is interested in the factors that explain why some houses are worth more than others, he/she may think that the price of a house depends on the size of the house. In mathematical terms, he/she would then let Y denote the variable “price of the house” and X denote the variable “size of the house” and the fact that Y depends on X is written using the notation:

$$Y = f(X)$$

This notation should be read “ Y is a function of X ” and captures the idea that the value for Y depends on the value of X . There are many functions that one could use, but in this book I will usually focus on linear functions. Hence, I will not use this general “ $f(X)$ ” notation in this book.

The equation of a straight line (what was called a “linear function” above) is used throughout this book. Any straight line can be written in terms of an equation:

$$Y = \alpha + \beta X$$

where α and β are **coefficients**, which determine a particular line. So, for instance, setting $\alpha = 1$ and $\beta = 2$ defines one particular line while $\alpha = 4$ and $\beta = -5$ defines a different line.

It is probably easiest to understand straight lines by using a graph (and it might be worthwhile for you to sketch one at this stage). In terms of an XY graph (i.e. one which measures Y on the vertical axis and X on the horizontal axis) any line can be defined by its intercept and slope. In terms of the equation of a straight line, α is the intercept and β the slope. The intercept is the value of Y when $X = 0$ (i.e. point at which the line cuts the Y -axis). The slope is a measure of how much Y changes when X is changed. Formally, it is the amount Y changes when X changes by one unit. For the student with a knowledge of calculus, the slope is the first derivative, $\frac{dY}{dX}$.

Summation notation

At several points in this book, subscripts are used to denote different observations from a variable. For instance, a labor economist might be interested in the wage of every one of 100 people in a certain industry. If the economist uses Y to denote this variable, then he/she will have a value of Y for the first individual, a value of Y for the second individual, etc. A compact notation for this is to use subscripts so that Y_1

is the wage of the first individual, Y_2 the wage of the second individual, etc. In some contexts, it is useful to speak of a generic individual and refer to this individual as the i -th. We can then write, Y_i for $i = 1, \dots, 100$ to denote the set of wages for all individuals.

With the subscript notation established, summation notation can now be introduced. In many cases we want to add up observations (e.g. when calculating an average you add up all the observations and divide by the number of observations). The Greek symbol, Σ , is the summation (or “adding up”) operator and superscripts and subscripts on Σ indicate the observations that are being added up. So, for instance,

$$\sum_{i=1}^{100} Y_i = Y_1 + Y_2 + \dots + Y_{100}$$

adds up the wages for all of the 100 individuals. As other examples,

$$\sum_{i=1}^3 Y_i$$

adds up the wages for the first 3 individuals and

$$\sum_{i=47}^{48} Y_i$$

adds up the wages for the 47th and 48th individuals.

Sometimes, where it is obvious from the context (usually when summing over all individuals), the subscript and superscript will be dropped and I will simply write:

$$\sum Y_i.$$

Logarithms

For various reasons (which are explained later on), in some cases the researcher does not work directly with a variable but with a transformed version of this variable. Many such transformations are straightforward. For instance, in comparing the incomes of different countries the variable GDP per capita is used. This is a transformed version of the variable GDP. It is obtained by dividing GDP by population.

One particularly common transformation is the logarithmic one. The logarithm (to the base B) of a number, \mathcal{A} , is the power to which B must be raised to give \mathcal{A} . The notation for this is: $\log_B(\mathcal{A})$. So, for instance, if $B = 10$ and $\mathcal{A} = 100$ then the logarithm is 2 and we write $\log_{10}(100) = 2$. This follows since $10^2 = 100$. In economics, it is common to work with the so-called natural logarithm which has $B = e$ where $e \approx 2.71828$. We will not explain where e comes from or why this rather unusual-looking base is chosen. The natural logarithm operator is denoted by \ln ; i.e. $\ln(\mathcal{A}) = \log_e(\mathcal{A})$.

In this book, you do not really have to understand the material in the previous paragraph. The key thing to note is that the natural logarithmic operator is a common one (for reasons explained later on) and it is denoted by $\ln(\mathcal{A})$. In practice, it can be easily calculated in a spreadsheet such as Excel (or on a calculator).

Endnote

1. I expect that most readers of this book will have access to Excel (or a similar spreadsheet or statistics software package) through their university computing labs or on their home computers (note, however, that some of the methods in this book require the Excel Analysis ToolPak add-in which is not included in some basic installations of Microsoft Works). However, computer software can be expensive and, for the student who does not have access to Excel and is financially constrained, there is an increasing number of free statistics packages designed using open source software. R. Zelig, which is available at <http://gking.harvard.edu/zelig/>, is a good example of such a package.

ANALYSIS OF ECONOMIC DATA

by Gary Koop

Chapter 3: Correlation

Correlation measures numerically the relationship between two variables X and Y (e.g. population density and deforestation).

Correlation between X and Y is symbolised by r or r_{XY} .

$$r = \frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum (Y_i - \bar{Y})^2} \sqrt{\sum (X_i - \bar{X})^2}}$$

Understanding Correlation

Example: Y = deforestation, X = population density we obtain $r = .66$.

What does the fact r is positive mean?

- There is a positive relationship between population density and deforestation.
- Countries with high population densities also tend to have high deforestation rates.
- Countries with low population densities tend to have low deforestation rates.
- Deforestation and population densities both vary across countries. The high/low variation in deforestation rates tends to match up with the high/low variation in population densities.

Properties of Correlation

- r lies between -1 and +1.
- Positive values of r indicate positive correlation between X and Y, negative values indicate negative correlation, $r = 0$ implies X and Y are uncorrelated.
- Larger positive values of r indicate stronger positive correlation. $r = 1$ indicates perfect positive correlation.
- More negative values of r indicate stronger negative correlation. $r = -1$ indicates perfect negative correlation.
- The correlation between Y and X is the same as the correlation between X and Y.
- The correlation between any variable and itself is 1.
- Correlation only measures linear relations

Understanding Correlation (cont.)

What does the magnitude of r mean?

r^2 measures the proportion of the variance in deforestation that matches up with (or is explained by) the variance in population density.

$$r^2 = .66^2 = .44.$$

44% of the cross-country variation in deforestation can be explained by the cross-country variation in population density.

Why are Variables Correlated?

Correlation does not necessarily imply causality.

Example:

The correlation between workers' education levels and wages is strongly positive.

Does this mean education "causes" higher wages? We can't know for sure.

- **Possibility 1: Education improves skills and more skilled workers get better paying jobs.**

Education causes wages to increase.

- **Possibility 2: Individuals are born with quality A which is relevant for success in education and on the job (e.g. intelligence or talent or determination, etc.).**

Quality A (NOT education) causes wages to increase.

Why are Variables Correlated? (cont.)

Example:

Data on N=546 houses sold in Windsor, Canada

Y = sales price of a house

X = the size of the lot the house is on

$$r_{XY} = .54$$

- Houses with large lots tend to be worth more than houses with small lots.
- Economic theory tells us that the price of a house should depend on its characteristics.
- Economic theory suggests X is causing Y.
- Here economic theory suggests that correlation does imply causality.

Why are Variables Correlated? (cont.)

Example:

Assume:

- Cigarette smoking causes cancer.
- Drinking alcohol does not cause cancer.
- Smokers tend to drink more alcohol.

Suppose we collected data on many elderly people on how much they smoked (X), whether they had cancer (Y) and how much they drank (Z). What correlations would we find?

$r_{XY} > 0$ Direct causality.

$r_{YZ} > 0$ This does NOT reflect causality.

Why are Variables Correlated? (cont.)

Example:

High rural population density (X) causes farmers to clear new land in forested areas (Z) which in turn causes deforestation (Y).

Here we would find $r_{XY} > 0$ and $r_{ZY} > 0$.

X (population density) is an indirect (or proximate) cause of Y (deforestation).

Z (agricultural clearance) is a direct (or immediate) cause of Y (deforestation).

Why are Variables Correlated? Summary

- Correlations can be very suggestive, but cannot on their own establish causality.
- Correlation + a sensible theory suggests (but does not prove) causality

Understanding Correlation through XY-plots

Figure 3.1: House price versus lot size

Think of drawing a straight line that best fits the points in the XY-plot. It will have positive slope.

Positive slope=positive relationship=positive correlation.

Figure 3.2: XY plot with $r_{XY} = 1$

All points fit on a straight upward sloping line.

Understanding Correlation through XY-plots (continued)

Figure 3.3: XY plot with $r_{XY} = .51$

Points still exhibit an upward sloping pattern, but much more scattered.

Figure 3.4: XY plot with $r_{XY} = 0$

Completely random scattering of points.

Figure 3.5: XY plot with $r_{XY} = -.51$

Think of drawing a straight line that best fits the points in the XY-plot. It will have negative slope.

Negative slope=negative relationship=negative correlation.

Correlation Among Several Variables

Correlation relates precisely two variables.

What to do with three or more? Usually use regression (next chapter).

Or you can calculate the correlation between every possible pair of variables.

Example: Three variables: X, Y and Z.

Can calculate three correlations: r_{xy} , r_{xz} and r_{yz} .

Four variables: X, Y, Z and W.

Can calculate six correlation: r_{xy} , r_{xz} , r_{xw} , r_{yz} , r_{yw} and r_{zw} .

M variables: $M \times (M-1)/2$ correlations.

A Correlation Matrix

Example:

Column 1 = X, Column 2 = Y, Column 3 = Z

	<i>Column 1</i>	<i>Column 2</i>	<i>Column 3</i>
Column 1	1		
Column 2	0.3182369	1	
Column 3	-0.1309744	0.0969959	1

$$r_{xy} = 0.3182369$$

$$r_{xz} = -0.1309744$$

$$r_{yz} = 0.0969959$$

ANALYSIS OF ECONOMIC DATA

by Gary Koop

Chapter 4: An Introduction to Simple Regression

- Regression is the most common tool of the applied economist.
- Used to help understand the relationships between many variables.
- We begin with simple regression to understand the relationship between two variables, X and Y.

Copyright © 2009 John Wiley & Sons, Ltd

1

Regression as a Best Fitting Line

Example:

See Figure 2.3: XY-plot of deforestation versus population density.

Regression fits a line through the points in the XY-plot that best captures the relationship between deforestation and population density.

Question: What do we mean by “best fitting” line?

Copyright © 2009 John Wiley & Sons, Ltd

2

Simple Regression: Theory

Assume a linear relationship exists between Y and X:

$$Y = \alpha + \beta X$$

α = intercept of line

β = slope of line

Example:

Y = output of a good, X = labour input

$$Y = .8 \times X$$

$$\alpha = 0$$

$\beta = .8$ = marginal product of labour.

Copyright © 2009 John Wiley & Sons, Ltd

3

Simple Regression: Theory (cont.)

1. Even if straight line relationship were true, we would never get all points on an XY-plot lying precisely on it due to measurement error.
2. True relationship probably more complicated, straight line may just be an approximation.
3. Important variables which affect Y may be omitted.

Due to 1., 2. and 3. we add an error.

Copyright © 2009 John Wiley & Sons, Ltd

4

The Simple Regression Model

$$Y = \alpha + \beta X + e$$

where e is an error.

What we know: X and Y .

What we do not know: α , β or e .

Regression analysis uses data (X and Y) to make a guess or estimate of what α and β are.

Notation: $\hat{\alpha}$ and $\hat{\beta}$ are the estimates of α and β .

Distinction Between Errors and Residuals

True Regression Line:

$$Y = \alpha + \beta X + e$$

$$e = Y - \alpha - \beta X$$

e = error

Estimated Regression Line:

$$Y = \hat{\alpha} + \hat{\beta} X + u$$

$$u = Y - \hat{\alpha} - \hat{\beta} X$$

u = residual

How do we choose $\hat{\alpha}$ and $\hat{\beta}$?

Consider the following reasoning:

1. We want to find a best fitting line through the XY-plot.
2. With more than two points it is not possible to find a line that fits perfectly through all points. (See Figure 4.1)
3. Hence, find “best fitting” line which makes the residuals as small as possible.
4. What do we mean by “as small as possible”?
The one that minimizes the sum of squared residuals.
5. Hence, we obtain the “ordinary least squares” or OLS estimator.

Derivation of OLS Estimator

- We have data on $i=1,...,N$ individuals which we call Y_i and X_i .
- Any line we fit/choice of $\hat{\alpha}$ and $\hat{\beta}$ will yield residuals u_i .
- Sum of squared residuals = $SSR = \sum u_i^2$.
- OLS estimator chooses $\hat{\alpha}$ and $\hat{\beta}$ to minimise SSR.

Solution:

$$\hat{\beta} = \frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$$

and

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \times \bar{X}$$

Jargon of Regression

- Y = dependent variable.
- X = explanatory (or independent) variable.
- α and β are coefficients.
- $\hat{\alpha}$ and $\hat{\beta}$ are OLS estimates of coefficients
- “Run a regression of Y on X ”

Regression and Causality

How to decide which is dependent variable

- Ideally, explanatory variable should be the one which causes/influences the dependent variable.
- Ideally, X causes Y .

If you can, build models where causality assumptions make sense

Examples:

- Increases in X = population density cause Y = deforestation to increase (not vice versa).
- Increasing X = the lot size of a house causes Y = its value to increase (not vice versa).
- Increasing X = advertising expenditures causes Y = company sales to increase (not vice versa).

Regression and Causality (cont.)

In practice, great care must be taken in interpreting regression results as reflecting causality. Why?

- In some cases, your assumption that X causes Y may be wrong.
- In some cases, you may not know whether X causes Y .
- In some cases, X may cause Y but Y may also cause X (e.g. exchange rates and interest rates).
- In some cases, the whole concept of causality may be inappropriate.

Formally, the question regression addresses is: “How much of the variability in Y can be explained by X ?”

Interpreting OLS Estimates

$$Y = \hat{\alpha} + \hat{\beta}X + u$$

Ignore u for the time being (focus on regression line)

Interpretation of $\hat{\alpha}$

- Estimated value of Y if $X = 0$.
- This is often not of interest.

Example:

X = lot size, Y = house price

$\hat{\alpha}$ = estimated value of a house with lot size = 0

Interpreting OLS Estimates (cont.)

$$Y = \hat{\alpha} + \hat{\beta}X + u$$

Interpretation of $\hat{\beta}$

1. $\hat{\beta}$ is estimate of the marginal effect of X on Y

2. Using regression model:

$$\frac{dY}{dX} = \hat{\beta}$$

3. A measure of how much Y tends to change when you change X.

4. “If X changes by 1 unit then Y tends to change by $\hat{\beta}$ units”, where “units” refers to what the variables are measured in (e.g. \$, £, %, hectares, metres, etc.).

Example: Deforestation and Population Density

Development economists have theories which imply that increasing population density should increase deforestation.

Thus:

Y = deforestation (annual % change) = dependent variable

X = population density (people per thousand hectares) = explanatory variable

Using data on N = 70 tropical countries we find:

$$\hat{\beta} = 0.000842$$

Example: Deforestation and Population Density (continued)

Interpretation of $\hat{\beta}$

a) “If population density increases by 1 person per 1,000 hectares, then deforestation will tend to increase by .000842% per year”

b) “If population density increases by 100 people per 1,000 hectares, then deforestation will tend to increase by .0842% per year”

Note: if deforestation increased by .0842% per year an additional 5% of the forest will be lost over 50 years.

Example: Cost of Production in the Electric Utility Industry

Data on N = 123 electric utility companies in the U.S.

Y = cost of production (millions of \$)

X = output (thousands of kilowatt hours, Kwh)

$$\hat{\beta} = .005$$

“Increasing output by 1,000 Kwh tends to increase costs by \$5,000”

Note: $.005 \times 1,000,000 = 5,000$

“Decreasing output by 1,000 Kwh tends to decrease costs by \$5,000”

Example: The Effect of Advertising on Sales

Data on N = 84 companies in the U.S.

Y = sales (millions of \$)

X = advertising expenditure (millions of \$)

$$\hat{\beta} = .218$$

“Increases in advertising of \$1,000,000 are associated with increases in sales of \$218,000.”

Residual Analysis

“How well does the regression line fit through the points on the XY-plot?”

“Are there any outliers which do not fit the general pattern?”

Concepts:

1. Fitted value of dependent variable:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} \times X_i$$

- Y_i does not lie on the regression line, \hat{Y}_i does lie on line.
- What would you do if you did not know Y_i , but only X_i and wanted to predict what Y_i was? Answer \hat{Y}_i

2. Residual, u_i , is: $Y_i - \hat{Y}_i$

Residual Analysis (cont.)

- Good fitting models have small residuals (i.e. small SSR)
- If residual is big for one observation (relative to other observations) then it is an outlier.
- Looking at fitted values and residuals can be very informative.

R²: A Measure of Fit

Intuition:

“Variability” = (e.g.) how deforestation rates vary across countries

Total variability in dependent variable Y =

Variability explained by the explanatory variable (X) in the regression

+

Variability that cannot be explained and is left as an error.

R²: A Measure of Fit (cont.)

In mathematical terms,

$$TSS = RSS + SSR$$

where TSS = Total sum of squares

$$TSS = \sum (Y_i - \bar{Y})^2$$

Note similarity to formula for variance.

RSS = Regression sum of squares

$$RSS = \sum (\hat{Y}_i - \bar{Y})^2$$

SSR = sum of squared residuals

Properties of R²

- $0 \leq R^2 \leq 1$
- $R^2 = 1$ means perfect fit. All data points exactly on regression line (i.e. SSR=0).
- $R^2 = 0$ means X does not have any explanatory power for Y whatsoever (i.e. X has no influence on Y).
- Bigger values of R^2 imply X has more explanatory power for Y.
- R^2 is equal to the correlation between X and Y squared (i.e. $R^2 = r_{xy}^2$)

R²: A Measure of Fit (cont.)

$$R^2 = 1 - \frac{SSR}{TSS}$$

or (equivalently):

$$R^2 = \frac{RSS}{TSS}$$

R^2 is a measure of fit (i.e. how well does the regression line fit the data points)

Properties of R² (cont.)

R^2 measures the proportion of the variability in Y that can be explained by X.

Example:

In regression of Y = deforestation on X = population density we obtain $R^2 = 0.44$

“44% of the cross-country variation in deforestation rates can be explained by the cross-country variation in population density”

Nonlinearity

So far regression of Y on X:

$$Y = \alpha + \beta X + e$$

Could do regression of Y (or $\ln(Y)$ or Y^2) on X^2 (or $1/X$ or $\ln(X)$ or X^3 , etc.) and same techniques and results would hold.

E.g.

$$Y = \alpha + \beta X^2 + e$$

How might you know if relationship is nonlinear?

Answer: Careful examination of XY-plots or residual plots.

Figures 4.2, 4.3 and 4.4.

Chapter 5: Statistical Aspects of Regression

$\hat{\alpha}$ and $\hat{\beta}$ are only estimates of α and β

Key question: How accurate are these estimates?

Statistical procedures allow us to formally address this question.

What Factors Affect Accuracy of OLS Estimates?

Graphical Intuition:

- Figure 5.1 (small number of data points)
- Figure 5.2 (large number of data points but very scattered)
- Figure 5.3 (large number of data points but not very scattered)
- Figure 5.4 (large number of data points, but clustered near one value for X)

What Factors Affect Accuracy of OLS Estimates?

Consider fitting a line through the XY-plots in Figures 5.1-5.4.

You would be most confident in the line you fit in Figure 5.3

Larger number of data points + less scattering (i.e. less variability in errors) + more variability in X = more accurate estimates.

Note: Figures 5.1, 5.2, 5.3 and 5.4 all contain artificially generated data with $\alpha=0$, $\beta=1$.

A Confidence Interval for β

- Uncertainty about accuracy of the estimate $\hat{\beta}$ can be summarised in a “confidence interval”
- 95% confidence interval for β is given by:

$$[\hat{\beta} - t_{\alpha/2} s_b, \hat{\beta} + t_{\alpha/2} s_b]$$

- $t_{\alpha/2}$ is a “critical value” from the “Student t-distribution” --- calculated automatically in Excel
- s_b = standard error of $\hat{\beta}$ is a measure of the accuracy of $\hat{\beta}$

$$s_b = \sqrt{\frac{SSR}{(N-2) \times \sum (X_i - \bar{X})^2}}$$

A Confidence Interval for $\hat{\beta}$ (cont.)

- t_b controls the confidence level (e.g. t_b is bigger for 95% confidence than 90%).
- s_b varies directly with SSR (i.e. how variable the residuals are)
- s_b varies inversely with N, the number of data points
- s_b varies inversely with $\sum(X_i - \bar{X})^2$, which is related to the variance/variability of X.

Note: Excel calculates the confidence interval for you and labels bounds of confidence interval as “Lower 95%” and “Upper 95%”

Intuition of Confidence Interval

- Useful (but formally incorrect) intuition: “There is a 95% probability that the true value of β lies in the confidence interval”.
- Correct intuition: “If you repeatedly use the above formula for calculating a confidence interval, 95% of the intervals you construct will contain the true value for β ”.
- Can choose any level of confidence you want (e.g. 90%, 99%).

Example: Confidence Intervals for the Data sets in Figures 5.1-5.4

Data Set	$\hat{\beta}$	90% Confid. Interval	95% Confid. Interval	99% Confid. Interval
Figure 5.1	.91	[-.92,2.75]	[-1.57,3.39]	[-3.64,5.47]
Figure 5.2	1.04	[.75,1.32]	[.70,1.38]	[.59,1.49]
Figure 5.3	1.00	[.99,1.01]	[.99,1.02]	[.98,1.03]
Figure 5.4	1.52	[-1.33,4.36]	[-1.88,4.91]	[-2.98,6.02]

Example: The Regression of Deforestation on Population Density

Y = deforestation

X = population density

$$\hat{\beta} = .000842$$

95% Confidence interval: [.00061,.001075]

Example: The Regression of Lot Size on House Price

OLS results:

$$Y = 34,136 + 6.59X,$$

- The OLS estimate of the marginal effect of X on Y is 6.59.
- “Increasing lot size by an extra square foot is associated with a \$6.59 increase in house price.”
- The 95% confidence interval for β is [5.72,7.47].
- “We are 95% confident that the effect of lot size on house is at least \$5.72 and at most \$7.47.”

Hypothesis Testing

- Test whether $\beta=0$ (i.e. whether X has any explanatory power)
- One way of doing it: look at confidence interval, check whether it contains zero. If no, then you are confident $\beta \neq 0$.
- Alternative (equivalent) way is to use “t-statistic” (often called “t-ratio”)

$$t = \frac{\hat{\beta}}{s_b}$$

- “Big” values for t indicate $\beta \neq 0$.
- “Small” values for t indicate it $\beta=0$.
(more concretely: β might be 0)

Hypothesis Testing (cont.)

Q: What do we mean by “big” and “small”?

A: Look at P-value.

- If P-value $\leq .05$ then t is “big” and conclude $\beta \neq 0$.
- If P-value $> .05$ then t is “small” and conclude $\beta=0$.
- Useful (but formally incorrect) intuition:
P-value measures the probability that $\beta = 0$.
- .05 = 5% = level of significance
- Other levels of significance (e.g. 1% or 10%) occasionally used

Example: The Regression of Deforestation on Population Density (cont.)

95% Confidence interval: [.00061,.001075]

Confidence interval does not contain zero, so conclude that $\beta \neq 0$.

Alternatively:

t-ratio is 7.227937. Is this big?

Yes, the P-value is 5.5×10^{-10} which is much less than .05.

Hence, we conclude again that $\beta \neq 0$.

Jargon

- “The coefficient on population density is significantly different from zero.”
- “Population density has statistically significant explanatory power for deforestation.”
- “The hypothesis that $\beta = 0$ can be rejected at the 5% significance level.”

Testing on R^2 : The F-statistic

- Test whether $R^2=0$ (i.e. whether X has any explanatory power)
- Note: In simple regression testing $R^2=0$ and $\beta=0$ are the same, but in multiple regression they will be different.
- F-statistic is a test statistic analogous to t-statistic (e.g. small values of it indicate $R^2=0$).

$$F = \frac{(N-2)R^2}{(1-R^2)}$$

Testing on R^2 : The F-statistic (cont.)

- For test with 5% level of significance:
- If P-value is $> .05$ conclude $R^2=0$.
- If P-value is $\leq .05$ conclude $R^2 \neq 0$.
- Excel calls the P-value for this test “Significance F”

Example: The Regression of Deforestation on Population Density (cont.)

- P-value = Significance F = 5.5×10^{-10} .
- Since P-value $< .05$ conclude $R^2 \neq 0$.
- Population density does have explanatory power for Y.

Chapter 6: Multiple Regression

- Multiple regression same as simple regression except many explanatory variables: X_1, X_2, \dots, X_k
- Intuition and derivation of multiple and simple regression very similar.
- We will emphasise only the few differences between simple and multiple regression.

Example: Explaining House Prices

Data on N=546 houses sold in Windsor, Canada

Dependent variable:

Y = sales price of house

Explanatory variables:

X_1 = lot size of property (in square feet)

X_2 = number of bedrooms

X_3 = number of bathrooms

X_4 = number of storeys (excluding basement)

OLS Estimation

- Multiple regression model:

$$Y_i = \alpha + \beta_1 X_{1i} + \dots + \beta_k X_{ki} + e_i$$

- OLS estimates: $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$
- Minimise Sum of Squared Residuals:

$$SSR = \sum (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \dots - \hat{\beta}_k X_{ki})^2$$

- Solution to minimisation problem: A mess
- Excel will calculate OLS estimates

Statistical Aspects of Multiple Regression

- Largely the same as for multiple regression.
- Formulae of Chapter 5 have only minor modifications.
- R^2 still a measure of fit with same interpretation (although now it is no longer simply the correlation between Y and X squared).

Statistical Aspects of Multiple Regression (cont.)

- Can test $R^2=0$ in same manner as for simple regression.
- If you find $R^2 \neq 0$ then you conclude that the explanatory variables *together* provide significant explanatory power (Note: this does not necessarily mean each individual explanatory variable is significant).
- Confidence intervals can be calculated for each *individual* coefficient as before.
- Can test $\beta_j=0$ for each *individual* coefficient ($j=1,2,...,k$) as before.

Interpreting OLS Estimates in Multiple Regression Model

Mathematical Intuition

- Total vs. partial derivative
- Simple regression: $\frac{dY}{dX} = \beta$
- Multiple Regression: $\frac{\partial Y}{\partial X_j} = \beta_j$

Interpreting OLS Estimates in Multiple Regression Model (cont.)

Verbal intuition

- β_j is the marginal effect of X_j on Y , *ceteris paribus*
- β_j is the effect of a small change in the j th explanatory variable on the dependent variable, *holding all the other explanatory variables constant*.

Example: Explaining House Prices (cont.)

	Coeff.	St.Err	t-Stat	P-val.	Lower 95%	Upper 95%
Interc.	-4010	3603	-1.113	0.266	-11087	3068
Size	5.429	0.369	14.703	2.E-41	4.704	6.155
Bed.	2825	1215	2.325	0.020	438.3	5211
Bath.	17105	1734	9.862	3.E-21	13698	20512
Storeys	7635	1008	7.574	1.E-13	5655	9615

- $R^2=.54$ and the P-value for testing $R^2=0$ (which is labelled “Significance F” by Excel) is 1.18E-88.
- Fitted regression line:

$$\hat{Y} = -4010 + 5.429X_1 + 2825X_2 + 17105X_3 + 7635X_4$$

Example: Explaining House Prices (cont.)

Since $\hat{\beta}_1 = 5.43$:

- An extra square foot of lot size will tend to add \$5.43 onto the price of a house, *ceteris paribus*.
- For houses with the same number of bedrooms, bathrooms and storeys, an extra square foot of lots size will tend to add \$5.43 onto the price of a house.
- If we compare houses with the same number of bedrooms, bathrooms and storeys, those with larger lots tend to be worth more. In particular, an extra square foot in lot size is associated with an increased price of \$5.43.

Example: Explaining House Prices (cont.)

Since $\hat{\beta}_2 = 2,824.61$:

- Adding one bedroom to your house will tend to increase its value by \$2,824.61, *ceteris paribus*.
- If we consider houses with comparable lot sizes and numbers of bathrooms and storeys, then those with an extra bedroom tend to be worth \$2,824.61 more.

Pitfalls of Using Simple Regression in a Multiple Regression Context

- In multiple regression above, coefficient on number of bedrooms was 2,824.61.
- A simple regression of Y = house price on X = number of bedrooms yields a coefficient estimate of 13,269.98.
- Why are these two coefficients on the same explanatory variable so different? i.e. 13,269.98 >>> 2,824.61.

Answer 1: They just come from two different regressions which control for different explanatory variables (different *ceteris paribus* conditions).

Pitfalls of Using Simple Regression in a Multiple Regression Context (cont.)

Answer 2:

- Imagine a friend asked: “I have 2 bedrooms and I am thinking of building a third, how much will it raise the price of my house?”
- Simple regression: “Houses with 3 bedrooms tend to cost \$13,269.98 more than houses with 2 bedrooms”
- Does this mean adding a 3rd bedroom will tend to raise price of house by \$13,269.98? Not necessarily, other factors influence house prices.
- Houses with three bedrooms also tend to be desirable in other ways (e.g. bigger, with larger lots, more bathrooms, more storeys, etc.). Call these “good houses”.
- Simple regression notes “good houses” tend to be worth more than others.

Pitfalls of Using Simple Regression in a Multiple Regression Context (cont.)

- Number of bedrooms is acting as a proxy for all these “good house” characteristics and hence its coefficient becomes very big (13,269.98) in simple regression.
- Multiple regression can estimate separate effects due to lot size, number of bedroom, bathrooms and storeys.
- Tell your friend: “Adding a third bedroom will tend to raise your house price by \$2,824.61”.
- Multiple regressions which contains all (or most) of house characteristics will tend to be more reliable than simple regression which only uses one characteristic.

Pitfalls of Using Simple Regression in a Multiple Regression Context (cont.)

Statistical evidence:

Correlation matrix:

	Sale Price	Lot size	#bed	#bath	#storey
Sale price	1				
Lot size	0.5358	1			
#bed	0.3664	0.1519	1		
#bath	0.5167	0.1938	0.3738	1	
#storeys	0.4212	0.0837	0.4080	0.3241	1

- Positive correlations between explanatory variables indicate that houses with more bedrooms also tend to have larger lot size, more bathrooms and more storeys.

Omitted Variable Bias

“Omitted variable bias” is a statistical term for these issues.

IF

1. We exclude explanatory variables that should be present in the regression,

AND

2. these omitted variables are correlated with the included explanatory variables,

THEN

3. the OLS estimates of the coefficients on the included explanatory variables will be biased.

Omitted Variable Bias (cont.)

Example:

- Simple regression used Y = house prices and X = number of bedrooms.
- Many important determinants of house prices omitted.
- Omitted variables were correlated with number of bedrooms. Hence, the OLS estimate from the simple regression $\hat{\beta} = 13,269.98$ was biased.

Practical Advice for Selecting Explanatory Variables

- Include (insofar as possible) all explanatory variables which you think might possibly explain your dependent variable. This will reduce the risk of omitted variable bias.
- However, including irrelevant explanatory variables reduces accuracy of estimation and increases confidence intervals. So do t-tests to decide whether variables are significant. Run a new regression omitting the explanatory variables which are not significant.

KOOP, chapter 6 Multiple Regression

17

Multicollinearity

- **Intuition:** If explanatory variables are highly correlated with one another then regression model has trouble telling which individual variable is explaining Y.
- **Symptom:** Individual coefficients may look insignificant, but regression as a whole may look significant (e.g. R^2 big, F-stat big).
- Looking at a correlation matrix for explanatory variables can often be helpful in revealing extent and source of multicollinearity problem.

KOOP, chapter 6 Multiple Regression

18

Multicollinearity (cont.)

Example:

Y = exchange rate

Explanatory variable(s) = interest rate

X_1 = bank prime rate
 X_2 = Treasury bill rate

Using both X_1 and X_2 will probably cause multicollinearity problem

Solution: Include either X_1 or X_2 but not both.

In some cases this “solution” will be unsatisfactory if it causes you to drop out explanatory variables which economic theory says should be there.

KOOP, chapter 6 Multiple Regression

19

Example: Multicollinearity Illustrated using Artificial Data

True Model:

$$Y = .5X_1 + 2X_2 + e$$

Correlation between X_1 and X_2 = .98

	Coeff.	St. Error	t-Stat	P-val.	Lower 95%	Upper 95%
Inter.	.1662	.1025	1.579	.1211	-.0456	.3780
X1	2.084	.9529	2.187	.0338	.1667	4.001
X2	.1478	.9658	.1530	.8790	-1.795	2.091

- $R^2 = .76$
- P-value for testing $R^2 = 0$ is 1.87E-15.
- We want coefficient estimates of roughly .5 and 2 – this is not what we are getting. Plus X_2 is not significant.

KOOP, chapter 6 Multiple Regression

20

Example: Multicollinearity Illustrated using Artificial Data (cont.)

Drop X_2 and re-run the regression:

	Coeff.	St. Error	t-Stat	P-val.	Lower 95%	Upper 95%
Inter.	.1667	.1041	1.601	.1160	-.0427	.3761
X1	2.227	.1788	12.454	1.E-16	1.867	2.586

- $R^2=.76$
- P-value for testing $R^2=0$ is 1.2E-16.
- Coefficient on X1 is significant, but is nowhere near .5!

Chapter 7: Regression with Dummy Variables

- Dummy variable is either 0 or 1.
- Use to turn qualitative (Yes/No) data into 1/0.

Example: Explaining House Prices

Dependent variable: Price of house

Explanatory variables:

- $D_1 = 1$ if the house has a driveway (=0 if it does not).
- $D_2 = 1$ if the house has a recreation room (=0 if not).
- $D_3 = 1$ if the house has a basement (=0 if not).
- $D_4 = 1$ if the house has gas central heating (=0 if not).
- $D_5 = 1$ if the house has air conditioning (=0 if not).

Analysis of Variance

- Analysis of Variance = ANOVA
- Statistical technique commonly used in many fields (but not much used in economics).
- ANOVA is just a special case of regression with dummy variables.
- Regression with dummy variables is much more powerful and general tool.

Simple Regression with a Dummy Variable

$$Y = \alpha + \beta D + e$$

- OLS estimation, confidence intervals, testing, etc. carried out in standard way.
- Interpretation a little different.

Simple Regression with a Dummy Variable

- Fitted value for i^{th} observation (point on regression line):

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} D_i$$

- Since $D_i = 0$ or 1 either

$$\hat{Y}_i = \hat{\alpha}$$

or

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}$$

Example: Explaining house prices (continued)

- Regress Y = house price on D = dummy for air conditioning (=1 if house has air conditioning, = 0 otherwise).
- Result:

$$\begin{aligned}\hat{\alpha} &= 59,885 \\ \hat{\beta} &= 25,996 \\ \hat{\alpha} + \hat{\beta} &= 85,881\end{aligned}$$

- Average price of house with air conditioning is \$85,881
- Average price of house without air conditioning is \$59,885

Multiple Regression with Dummy Variables

$$Y = \alpha + \beta_1 D_1 + \dots + \beta_k D_k + e$$

Example: Explaining house prices (continued)

Regress Y = house price on D_1 = driveway dummy and D_2 = rec room dummy

Four types of houses:

1. Houses with a driveway and a rec room ($D_1=1$, $D_2=1$)
2. Houses with a driveway but no rec room ($D_1=1$, $D_2=0$)
3. Houses with a rec room but no driveway ($D_1=0$, $D_2=1$)
4. Houses with no driveway and no rec room ($D_1=0$, $D_2=0$)

Example: Explaining house prices (continued)

	Coeff.	St. Error	t Stat	P-value	Lower 95%	Upper 95%
Inter.	47099.1	2837.6	16.60	2.E-50	41525	52673
D1	21159.9	3062.4	6.91	1.E-11	15144	27176
D2	16023.7	2788.6	5.75	1.E-08	10546	21502

1. If $D_1=1$ and $D_2=1$, then

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 + \hat{\beta}_2 = 47,099 + 21,160 + 16,024 = 84,283$$

“The average price of houses with a driveway and rec room is \$84,283”.

Example: Explaining house prices (continued)

2. If $D_1=1$ and $D_2=0$, then

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 = 47,099 + 21,160 = 68,259$$

“The average price of houses with a driveway but no rec room is \$68,259”.

3. If $D_1=0$ and $D_2=1$, then

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_2 = 47,099 + 16,024 = 63,123$$

“The average price of houses with a rec room but no driveway is \$63,123”.

4. If $D_1=0$ and $D_2=0$, then

$$\hat{Y} = \hat{\alpha} = 47,099$$

“The average price of houses with no driveway and no rec room is \$47,099”.

Multiple Regression with Dummy and non-Dummy Explanatory Variables

$$Y = \alpha + \beta_1 D + \beta_2 X + e$$

Example: Explaining house prices (continued)

Regress Y = house price on D = air conditioning dummy and X = lot size.

OLS estimates:

$$\begin{aligned}\hat{\alpha} &= 32,693 \\ \hat{\beta}_1 &= 20,175 \\ \hat{\beta}_2 &= 5.64\end{aligned}$$

Example: Explaining house prices (continued)

- For houses with an air conditioner D = 1 and

$$\hat{Y}_i = 52,868 + 5.64 X_i$$

- For houses without an air conditioner D=0 and

$$\hat{Y}_i = 32,693 + 5.64 X_i$$

- Two different regression lines depending on whether the house has an air conditioner or not.
- Two lines have different intercepts but same slope (i.e. same marginal effect)

Example: Explaining house prices (continued)

Verbal ways of expressing OLS results:

- “An extra square foot of lot size will tend to add \$5.64 onto the price of a house” (Note: no *ceteris paribus* qualifications to statement since marginal effect is same for houses with and without air conditioners)
- “Houses with air conditioners tend to be worth \$20,175 more than houses with no air conditioners, *ceteris paribus*” (Note: Here we do have *ceteris paribus* qualification)
- “If we consider houses with similar lot sizes, those with air conditioners tend to be worth an extra \$20,175”

Another House Price Regression

$$Y = \alpha + \beta_1 D_1 + \beta_2 D_2 + \beta_3 X_1 + \beta_4 X_2 + e.$$

- Regress Y = house price on D₁ = dummy variable for driveway, D₂ = dummy variable for rec room, X₁ = lot size and X₂ = number of bedrooms
- OLS estimates:

$$\begin{aligned}\hat{\alpha} &= -2736 \\ \hat{\beta}_1 &= 12,598 \\ \hat{\beta}_2 &= 10,969 \\ \hat{\beta}_3 &= 5.197 \\ \hat{\beta}_4 &= 10,562\end{aligned}$$

Another House Price Regression (cont.)

1. If $D_1=1$ and $D_2=1$, then

$$\hat{Y} = 20,831 + 5.197X_1 + 10,562X_2.$$

This is the regression line for houses with a driveway and rec room.

2. If $D_1=1$ and $D_2=0$, then

$$\hat{Y} = 9,862 + 5.197 \times X_1 + 10,562 \times X_2.$$

This is the regression line for houses with a driveway but no rec room.

Another House Price Regression (cont.)

3. If $D_1=0$ and $D_2=1$, then

$$\hat{Y} = 8,233 + 5.197X_1 + 10,562X_2.$$

This is the regression line for houses with a rec room but no driveway.

4. If $D_1=0$ and $D_2=0$, then

$$\hat{Y} = -2,736 + 5.197X_1 + 10,562X_2.$$

This is the regression line for houses with no driveway and no rec room.

Another House Price Regression (cont.)

Examples of verbal statements:

- “Houses with driveways tend to be worth \$12,598 more than similar houses with no driveway.”
- “If we consider houses with the same number of bedrooms, then adding an extra square foot of lot size will tend to increase the price of a house by \$5.197.”
- “An extra bedroom will tend to add \$10,562 to the value of a house, *ceteris paribus*”

Interacting Dummy and Non-Dummy Variables

$$Y = \alpha + \beta_1 D + \beta_2 X + \beta_3 Z + e.$$

where

$$Z = D \times X.$$

- Z is either 0 (for observations with $D=0$) or X (for observations with $D=1$)
- If $D=1$ then $\hat{Y} = (\hat{\alpha} + \hat{\beta}_1) + (\hat{\beta}_2 + \hat{\beta}_3)X$.
- If $D=0$, then $\hat{Y} = \hat{\alpha} + \hat{\beta}_2 X$.
- Two different regression lines corresponding to $D=0$ and $D=1$ exist and have different intercepts and slopes.
- The marginal effect of X on Y is different for $D=0$ and $D=1$.

Yet Another House Price Example

- Regress Y = house price on D = air conditioner dummy, X = lot size and $Z = D \times X$

- OLS estimates:

$$\hat{\alpha} = 35,684$$

$$\hat{\beta}_1 = 7,613$$

$$\hat{\beta}_2 = 5.02$$

$$\hat{\beta}_3 = 2.25.$$

- The marginal effect of lot size on housing is 7.27 for houses with air conditioners and only \$5.02 for houses without.
- Increasing lot size will tend to add more to the value of a house if it has an air conditioner than if it does not.

Working with Dummy Dependent Variables

Example:

Dependent variable is a transport choice.

1 = “Yes I take my car to work”

0 = “No I do not take my car to work”

- We will not discuss this case.
- Note only the following points:
 1. There are some problems with OLS estimation. But OLS estimation might be adequate in many cases.
 2. Better estimation methods are “Logit” and “Probit”. Excel cannot easily do these, so you will need a different software package.

Chapter 9: Univariate Time Series Analysis

- In previous chapter we discussed distributed lag models. These can be misleading if:
 1. The dependent variable Y_t depends on *lags of the dependent variable* as well, possibly, as $X_t, X_{t-1}, \dots, X_{t-q}$.
 2. The variables are nonstationary.
- In this chapter and the next, we develop tools for dealing with both issues and define what we mean by “nonstationary”.
- To simplify the analysis, focus solely on one time series, Y . Hence, *univariate time series analysis*.
- It is important to understand the properties of each individual series before proceeding to regression modelling involving several series.

Copyright © 2009 John Wiley & Sons, Ltd

1

Example: log of US personal income from 1954Q1 through 1994Q4.

See Figure 9.1

Aside on logs

- It is common to take the natural logarithm of time series which are growing over time (i.e. work with $\ln(Y)$ instead of Y). Why?
- A time series graph of $\ln(Y)$ will often approximate a straight line.
- In regressions with logged variables coefficients can be interpreted as elasticities.
- $\ln(Y_t) - \ln(Y_{t-1})$ is (approximately) the percentage change in Y between period $t-1$ and t .

Copyright © 2009 John Wiley & Sons, Ltd

2

Example: US personal income (cont.)

- Note trend behaviour of personal income series.
- Many macroeconomic time series exhibit such trends.

Differencing

$$\Delta Y_t = Y_t - Y_{t-1}$$

- ΔY_t measures the change (or growth) in Y between periods $t-1$ and t .
- If Y_t is the log of a variable, then ΔY_t is the percentage change.
- ΔY_t is the difference of Y (or first difference).
- ΔY_t is often called “delta Y ”.

Copyright © 2009 John Wiley & Sons, Ltd

3

Copyright © 2009 John Wiley & Sons, Ltd

4

Example: US personal income (cont.)

See Figure 9.2.

- ΔY = % change in personal income
- Not trending, very erratic.
- The differences of many macroeconomic time series have such properties.

The Autocorrelation Function

- Correlation between Y and lags of itself shed important light of the properties of Y .
- Relates to the idea of a trend (discussed above) and nonstationarity (not discussed yet).

Example: Y = US personal income

- Correlation between Y_t and Y_{t-1} is .999716!
- Correlation between ΔY_t and ΔY_{t-1} is -.00235.
- These are *autocorrelations* (i.e. correlations between a variable and lags of itself).

The Autocorrelation Function: Notation

- r_1 = correlation between Y and Y lagged one period.
- r_p = correlation between Y and Y lagged p periods.
- *Autocorrelation function* treats r_p as a function of p .

Example: US personal income (cont.)

Autocorrelation functions of Y and ΔY

Lag length (p)	Personal Income	Change in Pers. Income
1	.9997	-.0100
2	.9993	.0121
3	.9990	.1341
4	.9986	.0082
5	.9983	-.1562
6	.9980	.0611
7	.9978	-.0350
8	.9975	-.0655
9	.9974	.0745
10	.9972	.1488
11	.9969	.0330
12	.9966	.0363

- Y is highly correlated with lags of itself, but the change in Y is not.
- Information could also be presented on bar charts. See Figures 9.3 and 9.4.

Autocorrelation: Intuition

- **Y is highly correlated over time. ΔY does not exhibit this property.**
- **If you knew past values of personal income, you could make a very good estimate of what personal income was this quarter. However, knowing past values of the change in personal income will not help you predict the change in personal income this quarter.**
- **Y “remembers the past”. ΔY does not.**
- **Y is a nonstationary series while ΔY is stationary. (Note: These words not formally defined yet.)**

The Autoregressive Model

- **Previous discussion has focussed on graphs and correlations, now we go on to regression.**

- **Autoregressive model of order 1 is written as AR(1) and given by:**

$$Y_t = \alpha + \phi Y_{t-1} + e_t$$

- **Figures 9.5, 9.6 and 9.7 indicate the types of behaviour that this model can generate.**
- **$\phi = 1$ generates trending behaviour typical of macro time series.**
- **$\phi = 0$ looks more like change in macro time series.**

Nonstationary vs. Stationary Time Series

- **Formal definitions require difficult statistical theory. Some intuition will have to suffice.**
- **“Nonstationary” means “anything which is not stationary”.**
- **Focus on a case of great empirical relevance: unit root nonstationarity.**

Ways of Thinking about Whether Y is Stationary or has a Unit Root

1. **If $\phi = 1$, then Y has a unit root. If $|\phi| < 1$ then Y is stationary.**
2. **If Y has a unit root then its autocorrelations will be near one and will not drop much as lag length increases.**
3. **If Y has a unit root, then it will have a long memory. Stationary time series do not have long memory.**
4. **If Y has a unit root then the series will exhibit trend behaviour.**
5. **If Y has a unit root, then ΔY will be stationary. Hence, series with unit roots are often referred to as *difference stationary*.**

More on the AR(1) Model

$$Y_t = \alpha + \phi Y_{t-1} + e_t$$

- Can rewrite as:

$$\Delta Y_t = \alpha + \rho Y_{t-1} + e_t$$

$$\text{where } \rho = \phi - 1$$

- If $\phi = 1$ (unit root) then $\rho = 0$ and:

$$\Delta Y_t = \alpha + e_t$$

- Intuition: if Y has a unit root, can work with differenced data --- differences are stationary.

More on the AR(1) Model

- Test if $\rho = 0$ to see if a unit root is present.
- $-1 < \phi < 1$ is equivalent to $-2 < \rho < 0$. This is called the *stationarity condition*.

Aside: The Random Walk Model:

$$Y_t = Y_{t-1} + e_t$$

- This is thought to hold for many financial variables such as stock prices, exchange rates.
- Intuition: Changes in Y are unpredictable, so no arbitrage opportunities for investors.

Extensions of the AR(1) Model

- AR(p) model:

$$Y_t = \alpha + \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + e_t,$$

- Properties similar to the AR(1) model.

- Alternative way of writing AR(p) model:

$$\Delta Y_t = \alpha + \rho Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \dots + \gamma_{p-1} \Delta Y_{t-p+1} + e_t.$$

- Coefficients in this alternative regression ($\rho, \gamma_1, \dots, \gamma_{p-1}$) are simple functions of ϕ_1, \dots, ϕ_p .

The AR(p) Model

- AR(p) is in the form of a regression model.
- $\rho=0$ implies that the time series Y contains a unit root (and $-2 < \rho < 0$ indicates stationarity).
- If a time series contains a unit root then a regression model involving only ΔY is appropriate (i.e. if $\rho = 0$ then the term Y_{t-1} will drop out of the equation).
- “If a unit root is present, then you can difference the data to induce stationarity.”

More Extensions: Adding a Deterministic Trend

- Consider the following model:

$$Y_t = \alpha + \delta t + e_t.$$

- The term δt is a *deterministic trend* since it is an exact (i.e. deterministic) function of time.
- Unit root series contain a so-called *stochastic trend*.
- Combine with the AR(1) model to obtain:

$$Y_t = \alpha + \phi Y_{t-1} + \delta t + e_t.$$

- Can generate behaviour that looks similar to unit root behaviour even if $|\phi| < 1$. (i.e. even if they are stationary).
- See Figure 9.8.

Summary

- The nonstationary time series variables on which we focus are those containing a unit root. These series contain a stochastic trend. If we difference these time series, the resulting time series will be stationary. For this reason, they are also called difference stationary.
- The stationary time series on which we focus have $-2 < \rho < 0$. But these series may exhibit trend behaviour through the incorporation of a deterministic trend. If this occurs, they are also called trend stationary.

AR(p) with Deterministic Trend Model

- Most general model we use:

$$\Delta Y_t = \alpha + \rho Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \dots + \gamma_p \Delta Y_{t-p+1} + \delta t + e_t.$$

- Why work with this form of the model?

- A unit root is present if $\rho = 0$. Easy to test.
- The specification is less likely to run into multicollinearity problems. Remember: in macroeconomics we often find Y is highly correlated with lags of itself but ΔY is not.

Estimation of the AR(p) with Deterministic Trend Model

- OLS can be done in usual way.

Example: Y = US personal income

- ΔY is the dependent variable in the regression below.

AR(4) with Deterministic Trend Model

	Coeff.	St. Error	t-Stat	P-val	Lower 95%	Upper 95%
Inter.	.138	.108	1.279	.203	-.075	.351
Y_{t-1}	-.018	.015	-1.190	.236	-.049	.012
ΔY_{t-1}	-.017	.081	-.217	.829	-.177	.142
ΔY_{t-2}	.014	.081	.172	.863	-.145	.173
ΔY_{t-3}	.130	.080	1.627	.106	-.028	.288
time	.00012	.00012	.955	.341	-.00013	.00037

Testing in AR(p) with Deterministic Trend Model

- For everything except ρ , testing can be done in usual way using t-statistics and P-values.
- Hence, can use standard tests to decide whether to include deterministic trend.

Lag length selection

- A common practice: begin with an AR(p) model and look to see if the last coefficient, γ_p is significant. If not, estimate an AR(p-1) model and see if γ_{p-1} is significant. If not, estimate an AR(p-2), etc.

Example: Y = US personal income

- Sequential testing strategy leads us to drop the deterministic trend and go all the way back to a model with one lag, an AR(1).
- ΔY is the dependent variable in the regression.

	Coeff.	St. Error	t-Stat	P-val	Lower 95%	Upper 95%
Inter.	.039	.014	2.682	.008	.010	.067
Y_{t-1}	-.004	.002	-2.130	.035	-.0077	-.0003

Testing for a Unit Root

- You might think you can test $\rho = 0$ in the same way (i.e. look at P-value and, if it is less than .05, reject the unit root hypothesis, if not accept the unit root).
- **THIS IS INCORRECT!**
- Justification: Difficult statistics.
- Essentially: Excel gets the t-statistic correct, but the P-value is wrong.
- A correct test is the Dickey-Fuller Test, which uses the t-statistic and compares it to a critical value.

Practical Advice on Unit Root Testing

- Learn more econometrics using another textbook.
- Use a computer software package which is suitable for time series.
- Use the following rough rule of thumb which should be okay if sample size is moderately large (e.g. $T > 50$).

Testing for a Unit Root: An Approximate Strategy

1. Use the sequential testing strategy outlined above to estimate the AR(p) with deterministic trend model. Record the t-stat corresponding to ρ (i.e. the coefficient on Y_{t-1}).
2. If the final version of your model includes a deterministic trend, the Dickey-Fuller critical value is approximately -3.45 . If the t-stat on ρ is more negative than -3.45 , reject the unit root hypothesis and conclude that the series is stationary. Otherwise, conclude that the series has a unit root.
3. If the final version of your model does not include a deterministic trend, the Dickey-Fuller critical value is approximately -2.89 . If the t-stat on ρ is more negative than this, reject the unit root hypothesis and conclude that the series is stationary. Otherwise, conclude that the series has a unit root.

Example: Y = US personal income

- Sequential testing strategy leads from:

$$\Delta Y_t = \alpha + \rho Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \dots + \gamma_p \Delta Y_{t-p+1} + \delta_t + e_t.$$

to

$$\Delta Y_t = \alpha + \rho Y_{t-1} + e_t$$

	Coeff.	St. Error	t-Stat	P-val	Lower 95%	Upper 95%
Inter.	.039	.014	2.682	.008	.010	.067
Y_{t-1}	-.004	.002	-2.130	.035	-.0077	-.0003

- t-stat on ρ is -2.13 , which is not more negative than -2.89 . Hence, we can accept the hypothesis that personal income does contain a unit root.

College Town

FCI owns 10 apartment buildings in a college town, which it rents exclusively to students. Each apartment building contains 100 rental units, but the owner is having cash flow problems due to an average vacancy rate of nearly 50 percent. The apartments in each building have comparable floor plans, but some buildings are closer to campus than others. The owner of FCI has data from last year on the number of apartments rented, the rental price (in dollars), and the amount spent on advertising (in hundreds of dollars) at each of the 10 apartments. These data, along with the distance (in miles) from each apartment building to campus, are presented in rows 1 through 11 of Table 3–9. The owner regressed the quantity demanded of apartments on price, advertising, and distance. The results of the regression are reported in rows 16 through 22 of Table 3–9. What is the estimated demand function for FCI's rental units? If FCI raised rents at one complex by \$100, what would you expect to happen to the number of units rented? If FCI raised rents at an average apartment building, what would happen to FCI's total revenues? What inferences should be drawn from this analysis?

Table 3–9 Input and Output from a Multiple Regression

	A	B	C	D	E	F	G
1	Observation	Quantity	Price	Advertising	Distance		
2	1	28	250	11	12		
3	2	69	400	24	6		
4	3	43	450	15	5		
5	4	32	550	31	7		
6	5	42	575	34	4		
7	6	72	375	22	2		
8	7	66	375	12	5		
9	8	49	450	24	7		
10	9	70	400	22	4		
11	10	60	375	10	5		
12	Average	53.10	420.00	20.50	5.70		
13							
14							
15							
16	Regression Statistics						
17							
18	Multiple R	0.89					
19	R-Square	0.79					
20	Adjusted R-Square	0.69					
21	Standard Error	9.18					
22	Observations	10.00					
23							

Quelle: Baye, Michael: Managerial Economics and Business Strategy, Mc Graw 2003, p 101

Analyse the data in Excel:

- **Describe the relation between the quantity and the independent variables in bivariate correlations and regressions. What do we learn?**
- Continue the multiple regression (already started above).
 - Which are the numerical results?
 - What do we learn? Interpret the results (for example answering the questions above); what is your recommendation for FCI?