SOCIOECONOMIC DATA ANALYSIS TRAINING WORKSHOP

Instructors: Matt Gorstein (NOAA/NOS) and Supin Wongbusarakum (JIMAR, NOAA/PIFSC/ESD/CREP)

Location: Palau International Coral Reef Center, Koror, Palau

Dates: September 12-17, 2016
Introduction

This document compiles presentation slides that were used in the socioeconomic data analysis training workshop in Koror, Palau, from September 12-17, 2016. The workshop was funded by NOAA’s Coral Reef Conservation Program, with financial and logistical support from many partners—including the Micronesia Islands Nature Alliance, NOAA’s Pacific Islands Regional Office, Pacific Islands Managed and Protected Areas Community, Micronesia Conservation Trust, Palau International Coral Reef Center, and several other organizations and agencies involved in marine conservation and resource management in Micronesia. Participants attended from Guam, Commonwealth of the Northern Mariana Islands, Federated States of Micronesia (Kosrae, Pohnpei, and Yap), Palau, Republic of the Marshall Islands, and Hawai‘i.

The training used IBM SPSS Statistics version 24 and Excel. The example data set used in the training was a part of a survey conducted in the Merizo community of Manell-Geus in Guam. It is not included here. For questions related to the data set, please contact Matt Gorstein matt.gorstein@noaa.gov or Supin Wongbusarakum supin.wongbusarakum@noaa.gov
<table>
<thead>
<tr>
<th>Day, Date, Topics and Suggested Length of Session</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DAY 1: Monday, September 12, 2016</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Morning: Introduction to Data Analysis</strong></td>
<td></td>
</tr>
<tr>
<td>1. Introductions, agenda, goals, objectives (25 mins)</td>
<td>2</td>
</tr>
<tr>
<td>2. Different levels of data (nominal, ordinal, interval, ratio) (30 min)</td>
<td>5</td>
</tr>
<tr>
<td>3. Quiz 1 (15 mins)</td>
<td>11</td>
</tr>
<tr>
<td>4. Intro to Excel and data entry (75 min)</td>
<td>15</td>
</tr>
<tr>
<td><strong>Afternoon: Data Entry and Organization</strong></td>
<td></td>
</tr>
<tr>
<td>4. (continued) Data entry, codebook creation, documenting workflow (80 min)</td>
<td>30</td>
</tr>
<tr>
<td>5. Import data from Excel into SPSS and SPSS intro (60 min)</td>
<td>44</td>
</tr>
<tr>
<td>6. Introduction to qualitative data (20 min)</td>
<td>61</td>
</tr>
<tr>
<td>Address differences between qualitative and quantitative data (15 mins)</td>
<td></td>
</tr>
<tr>
<td>7. Quiz 2 (15 mins)</td>
<td>69</td>
</tr>
<tr>
<td><strong>DAY 2: Tuesday, September 13, 2016</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Morning: Qualitative Data</strong></td>
<td></td>
</tr>
<tr>
<td>1. Coding open text and qualitative data analysis (60 mins) Interviews and focus groups; what to do with large blocks of text?</td>
<td>74</td>
</tr>
<tr>
<td>2. Analyze qualitative data (45 mins)</td>
<td>89</td>
</tr>
<tr>
<td>3. Data Visualization for Qualitative Data (45 min)</td>
<td>90</td>
</tr>
<tr>
<td>4. Quiz 3 (15 mins)</td>
<td>94</td>
</tr>
<tr>
<td><strong>Afternoon: Descriptive Statistics</strong></td>
<td></td>
</tr>
<tr>
<td>5. Overview of descriptive statistics</td>
<td>97</td>
</tr>
<tr>
<td>(Central tendency, normal distributions, frequencies) (45 mins)</td>
<td></td>
</tr>
<tr>
<td>6. Frequency and summary stats in SPSS (60 mins)</td>
<td>105</td>
</tr>
<tr>
<td>7. Data visualization for descriptive stats (60 min)</td>
<td>118</td>
</tr>
<tr>
<td>8. Quiz 4 (15 mins)</td>
<td>136</td>
</tr>
<tr>
<td>Day, Date, Topics and Suggested Length of Session</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td><strong>DAY 3: Wednesday, September 14, 2016</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Morning: Inferential Statistics</strong></td>
<td></td>
</tr>
<tr>
<td>1. Overview of inferential statistics</td>
<td>140</td>
</tr>
<tr>
<td>(Confidence intervals, hypothesis testing, p-values, t-tests)</td>
<td></td>
</tr>
<tr>
<td>2. Variable transformations</td>
<td>162</td>
</tr>
<tr>
<td>(Dummies, index creation, normalization, treating not sures, etc.)</td>
<td>5</td>
</tr>
<tr>
<td>3. Quiz 5 (15 mins)</td>
<td>175</td>
</tr>
<tr>
<td>4. Proposing stats questions and hypothesis (What issue do you want to address with your data? What questions are you trying to answer?) (60 mins)</td>
<td>178</td>
</tr>
<tr>
<td><strong>Afternoon: Stats Questions and Inferential Stats in SPSS</strong></td>
<td></td>
</tr>
<tr>
<td>5. Inferential Stats using SPSS</td>
<td>182</td>
</tr>
<tr>
<td>6. Quiz 6 (15 mins)</td>
<td>198</td>
</tr>
<tr>
<td><strong>DAY 4: Thursday, September 15, 2016</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Morning: Chi-square, T-Test and ANOVA</strong></td>
<td></td>
</tr>
<tr>
<td>1. Creating contingency tables and doing Chi-square tests</td>
<td>202</td>
</tr>
<tr>
<td>(90 mins)</td>
<td></td>
</tr>
<tr>
<td>2. Various T-tests and ANOVA</td>
<td>217</td>
</tr>
<tr>
<td>(90 mins)</td>
<td></td>
</tr>
<tr>
<td>3. Quiz 7 (15 mins)</td>
<td>235</td>
</tr>
<tr>
<td><strong>DAY 5: Friday, September 16, 2016</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Morning: Correlation and Regression</strong></td>
<td></td>
</tr>
<tr>
<td>1. Correlation analysis</td>
<td>239</td>
</tr>
<tr>
<td>(45 min)</td>
<td></td>
</tr>
<tr>
<td>2. Simple linear regression</td>
<td>248</td>
</tr>
<tr>
<td>(60 mins)</td>
<td></td>
</tr>
<tr>
<td>3. Multiple linear regression and model validity (75 mins)</td>
<td>256</td>
</tr>
<tr>
<td>4. Quiz 8 (15 mins)</td>
<td>270</td>
</tr>
<tr>
<td><strong>Afternoon: Data Visualization</strong></td>
<td></td>
</tr>
<tr>
<td>5. Data visualization for inferential stats: contingency tables, t-tests, correlations, regressions (120 mins)</td>
<td>273</td>
</tr>
<tr>
<td>6. Quiz 9 (15 mins)</td>
<td>288</td>
</tr>
<tr>
<td><strong>DAY 6: Saturday, September 17, 2016</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Morning: Multiple Response, Non-Parametric Tests, Recap Qualitative VS Quantitative, and Best Practices</strong></td>
<td></td>
</tr>
<tr>
<td>1. Multiple Response Analysis</td>
<td>292</td>
</tr>
<tr>
<td>(30 mins)</td>
<td></td>
</tr>
<tr>
<td>2. Normality and Non-Parametric Tests</td>
<td>295</td>
</tr>
<tr>
<td>(60 mins)</td>
<td></td>
</tr>
<tr>
<td>3. Best practices, ethics of data analysis and data management</td>
<td>303</td>
</tr>
<tr>
<td>(45 mins)</td>
<td></td>
</tr>
<tr>
<td>4. When to use quantitative or qualitative data</td>
<td>308</td>
</tr>
<tr>
<td>(30 mins)</td>
<td></td>
</tr>
<tr>
<td><strong>Afternoon: Q&amp;A, Training Workshop feedback, and Closing</strong></td>
<td></td>
</tr>
<tr>
<td>Answer Key for Quizzes</td>
<td>318</td>
</tr>
</tbody>
</table>
Day 1

- Introduction to Data Analysis

- Data Entry and Organization
Introduction, Workshop Goals, and Agenda

Day 1: September 12, 2016

Participant Introduction

- What is your name?
- What is your organization and type of work?
- What is your experience in data analysis?

HOPE: What are the most important things you would like to learn from this workshop?

CONCERN: Anything you are concerned or worried about?
Goal of Data Analysis

- Data analysis is the process of making sense out of the data.
- Data do not speak for themselves; there is always an interpreter, or a translator (Ratcliffe 1983)

Agenda

**Day 1**: Intros, database, data coding and entering, intro to qualitative data
**Day 2**: Qualitative data analysis, descriptive data analysis
**Day 3**: Inferential Statistics, hypothesis testing, variable transformations
**Day 4**: Hands on SPSS exercises (contingency tables, chi square, t tests)
**Day 5**: Hands on SPSS exercises (correlation, regression), data visualization
**Day 6**: Data management and best practices
Workshop Objectives

- To understand basic statistics
- To understand principles of qualitative and quantitative data analysis
- To understand how to properly code and document your data
- To be able to use SPSS to run descriptive data analysis and test hypotheses
- To better communicate results of data analysis and effectively communicate data visually

Sharing Projects and Data Collected

- What type of projects are you working on?
- What are you hoping to find from these projects?
- What types of data are being collected?
- What is the goal of your research?
  - Inform management, public, or a private entity?
  - Inform planning?
  - What else?
- Is there something specific that you would like to learn here that can be used to address an issue with your project?
Different Types of Data

Day 1: September 12, 2016

Data Types

- Different “types” of data describe the characteristics of your data

- Choosing what types of analysis to run depends on your type of data
  - Certain statistical tests may be inappropriate (or invalid) for certain types of data
Discrete and Continuous Variables

- A variable is a characteristic that can vary in value among subjects in a sample or population.

- A variable is **discrete** if its possible values form a set of separate numbers, such as 0, 1, 2, 3, ....

- It is **continuous** if it can take an infinite continuum of possible real number values.

Types of Measurement Scales

**Categorical (Discrete) Data**

- **Nominal**
  - *Qualitative data (eg. label, type, yes/no)*
  - *Cannot be placed in meaningful order*

- **Ordinal**
  - *Qualitative and quantitative data (eg. rating, likert scale)*
  - *Can rank order from lowest to highest but we cannot claim anything about the space between the values*
    - (eg: We can’t say for sure that a 5 star hotel is 5 times better than 1 star hotel)
Types of Measurement Scales

**Continuous Data**

- **Interval**
  - Quantitative data
  - Can add and subtract when comparing values (example: Celsius scale)
  - The distance between attributes DOES have meaning (unlike ordinal)

- **Ratio**
  - Quantitative data with true 0 point (eg. age, weight, # people, %, $)
  - All math operations can be utilized

The distinction is important because it affects which statistical techniques we can use in data analysis

---

**Data Type Examples from Manell-Geus Questionnaire**

- **Nominal Variable**

  - From the file “Manell_Geus_codebook.xlsx”
  - Question #13 is coded as an nominal variable because all of the responses are mutually exclusive and none of them have any numerical significance
  - In this case, the codes are basically just labels, there is no relationship between the numbers themselves
### Ordinal Variable

#### Question #7

Coral reefs protect Guam from coastal/shoreline erosion and natural disasters like typhoons and tsunamis.

- **Variable Name**: agreement_protect
- **Answer Options**: Strongly Disagree
- **Code**: 1

- **Question #7 is coded as an ordinal variable because it the order of the responses is significant (higher numbers indicate more agreement), but the differences between each one is not quantifiable.**

### Interval Variable

#### Question #38

What is your annual household income?

<table>
<thead>
<tr>
<th>Question Number</th>
<th>Question</th>
<th>Variable Name</th>
<th>Answer Options</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
<td>What is your annual household income?</td>
<td>income</td>
<td>Under $30,000</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$30,000 to $39,999</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$40,000 to $49,999</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$50,000 to $59,999</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$60,000 to $69,999</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$70,000 to $79,999</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$80,000 to $89,999</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$90,000 to $99,999</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$100,000 or More</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Refused to answer</td>
<td>99</td>
</tr>
</tbody>
</table>

- **Question #38 is coded as an interval variable because it the order of the responses is significant (higher numbers indicate more income), and the differences between each category is not quantifiable (i.e. $60,000 is twice as much as $30,000).**
Data Type Examples from Manell-Geus Questionnaire

- **Ratio Variable**

<table>
<thead>
<tr>
<th>Question Number</th>
<th>Question</th>
<th>Variable Name</th>
<th>Answer Options</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Of the seafood that you and your family eat, how much of it comes from</td>
<td>Percent_Merizo</td>
<td>Continuous (percentage from 0-100)</td>
<td>0-100</td>
</tr>
<tr>
<td>21</td>
<td>Merizo?</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Question #3 is coded as a ratio variable because it asks the respondent to specify the percentage of their family’s seafood that comes from Merizo on a continuous scale with a true zero point.

Accuracy vs Precision

- **Accuracy (Validity) = how close the values are to the true value**
  - Are we asking the right question?
  - Are we using the relevant method?
  - Are we making valid conclusions?

- **Precision (Reliability) = how consistent the values are with repeated measurement**
  - Are we using an unbiased sample?
  - Is the sample size large enough?

Bias can lead to measures that are precise but not accurate.
Practice!

- Let’s open the “Manell_Geus_codebook” and look through some of the variables
  - *Identify some that are nominal, ordinal, interval, and ratio*
Quiz #1
Day 1: September 12, 2016

1.1 Which of the following are discrete (categorical) data?

A. Nominal data  
B. Ratio data  
C. Interval data  
D. Ordinal data
1.2 True or False: Some statistical tests may be invalid based on your type of data

A. True  
B. False

1.3 What is the definition of “variable”?

A. A variable is data that can take an infinite continuum of possible real number values  
B. A variable is a characteristic that can vary in value among subjects in a sample or population  
C. A variable is a type of statistical test  
D. A variable represents the main research question in a project
1.4 True or false: Interval data contains a true “zero point”
A. True
B. False

1.5 In which type of data does the order of responses matter but the differences between each choice is not quantifiable?
A. Nominal
B. Ordinal
C. Interval
D. Ratio
1.6 True or false: If the data are precise, then they are accurate.

A. True
B. False
Why Use Excel?

- Standard Microsoft program
- Variety of functions (sums, averages, counts, If/then, math operations)
- Flexible for many types of data
- Integrated graphics
  - Can create graphs, tables, charts
- Compatibility
- Can perform statistical tests
- A lot of support available online
- Widely used across the world
Excel Basics

- Multiple worksheet tabs within one file
- Cells can contain numbers, text, or functions
- Basic Word text editing and formatting
- Can sort, rearrange, cut/copy/paste, drag

Getting Started with Spreadsheets

- Open the file “Manell_Geus_Data_GettingStarted.xlsx”
- Data table = grid of rows and columns for organizing data
- Every ROW contains a unique record (case)
  - ROWS are identified by numbers (1, 2, 3, etc.)
- Every COLUMN contains a different attribute (variable) that relates to the records
  - COLUMNS are identified by letters (A, B, C, etc.)
- Cell: The intersection of a row and column, containing the data
  - Cells are identified by a letter/number combination (A1, A2, B3, C56, etc.)
- Header Row: the text name of each attribute, typically the first row of the table
- Row ID: a unique identifier (usually a number or text code) assigned to each row, to aid in data management
Ribbons – Home and Insert

Ribbons – Data, Formulas, and View
Data Entry Process

- **Best Practices:**
  - Name each sheet in the workbook with a name that represents what is contained in the sheet
  - Start with row IDs and column names
  - Enter data record-by-record (row-by-row) going from left to right
  - Keep an entry log of what is complete, with any additional comments
  - Backup and save after every session

- Enter data manually in Excel cell by cell

- Correct data entry will set you up for success in analysis later
  - Quality control is continuous
    - Checking ranges and value lists
    - Finding data in large tables
    - Narrowing to a subset
    - Checking for typos

Setting Up the Fields

- In first row create a field name for every attribute (piece of information or variable), column by column
  - Short, descriptive, unique
  - Avoid spaces, symbols, and numbers
  - Some data formats may restrict to 8 characters

- Decide what type of data it is
  - Nominal - categorical
  - Ordinal - categorical
  - Interval - continuous
  - Ratio - continuous

- Set up coding scheme to assign numbers for each response item

- Create a codebook or reference sheet with question number, question text, data type, and response codes
  (We will go into more detail about this step later)
Database design principles

- All data in a row should refer to a single record (case)
- Only one piece of information (‘attribute’ or variable) per column

Few things to help in reading data

- Text wrapping
- Merging
- Formatting as decimal, percent, etc.
  - Format painter (similar to copy and paste)
- Highlighting/Shading
- Insert/Delete
- Hide/Unhide
- Freeze Panes
Navigation Tips

- Arrows move through cells one at a time; Tab between columns; Enter between rows
- CTRL+arrow jumps to end of row or column
- SHIFT+arrow selects the current and adjacent cell
- CTRL+SHIFT+arrow selects to the end of the row or column
- CTRL+Z undoes last move
- CTRL+mouse click copies non-connected cells
- Clicking and dragging will select multiple cells
- Dragging (clicking the bottom right corner of a cell) copies a cell across rows and columns
- CTRL+F brings up the “find and replace” dialogue box
  - This function lets you find any text or numbers in a worksheet

Widely Used Formulas

- Formulas can be typed directly into a cell, into the formula bar, or they can be done by clicking the “function” key
- ALL FORMULAS START WITH AN EQUAL SIGN (=)
  - =SUM/SUMIF
  - =AVERAGE
  - =COUNT/COUNTIF
  - =MAX
  - =MIN
  - =MEDIAN
  - =IF, =OR, =AND
  - =VLOOKUP
Find and Replace Example

- Now, let’s open the file “Manell_Geus_Data_FunctionPractice.xlsx”
- Find the variable “condition_ocean”
- Click on CTRL+F
- Type “condition_ocean” under the “Find” tab

Find and Replace Example

- Now we have “found” the “condition_ocean” variable
- What if we want to replace all responses of “not sure” with a missing value (.)
- Highlight all of column L
- Click CTRL+F again
- Go to the “replace tab”
- Type 8 into “find what”
- Type a period into “replace with”
- Under “options,” we can click “match entire cell contents”
COUNT Example

- The **COUNT** function counts the number of cells in a range that contain a number.
- So we can see how many respondents answered a particular (numerically coded) question.
- Examine “agreement_mangroves”.
- In cell AJ310, type “=count(AJ2:AJ307)”.
- A figure of 304 is now in the cell.
  - Therefore, 304 people answered this question in the survey.

COUNTIF Example

- The **COUNTIF** function counts the number of cells in a range that contain a SPECIFIC number.
- We can see how many respondents agreed or strongly agreed that “mangroves are not important for protecting the coast from erosion”.
- In cell AJ311, type “=countif(AJ2:AJ307,4)”.
- In cell AJ312, type “=countif(AJ2:AJ307,5)”.
- 90 people agree, 44 people strongly agree.
SUM Example

- The SUM function adds all numeric cells in the range that you specify
- We want to investigate how many people were neutral, disagree, or strongly disagree with “mangroves are not important for protecting the coast from erosion”
- We can use the COUNTIF function, then the SUM function
- In cell AJ313, type “=countif(AJ2:AJ307,1)”
- In cell AJ314, type “=countif(AJ2:AJ307,2)”
- In cell AJ315, type “=countif(AJ2:AJ307,3)”
- In cell AJ316, type “=sum(AJ313:AJ315)”
- 157 people disagreed, or were neutral concerning the mangrove statement
  - \(\frac{157}{306} = 51\%\)
  - 306 is our sample size

SUMIF Example

- The SUMIF function adds specified numeric cells in the range that you specify
- “Add up this range of cells, ONLY if it contains this certain value”
- For instance, we can find how many respondents perceive ocean acidification to be a top threat to coral reefs
- In cell AP310, type “=sumif(AP2:AP307,1)”
- 29 respondents think ocean acidification is a threat
  - \(\frac{29}{306} = 9\%\)
- Note: The SUM function “sums” cells, while the COUNT function merely “counts” each cell as 1, regardless of cell contents
AVERAGE Example

- What is the average age of our sample?
- In cell IL310, type “=average(IL2:IL307)”
- Average age = 40.38 years old

AVERAGEIF Example

- We want to know the average age of homeowners in our sample
- “The AVERAGE age IF the respondent owns a home”
- In cell IL311, type “=AVERAGEIF(IV2:IV307,1,IL2:IL307)”
- The average age of homeowners in our sample is 44.52 years old
MIN and MAX Example

- These functions give us an idea of the range of values within a variable
- In our sample, what is the most amount of years that a respondent has lived in Merizo? What is the least amount of years that a respondent has lived in Merizo?
- In cell IT310, type “=max(IT2:IT307)”
- In cell IT311, type “=min(IT2:IT307)”
- Sample respondents have lived in Merizo anywhere from 0.25-68 years (Range = 67.75)

IF, OR, AND Example

- These functions represent logic statements
- “If cell __ is __ , then return a value of 1, if not, then return a value of 0”
- “If cell__ is __ OR __ , then return a value of 1, if not, then return a value of 0”
- “If cell__ is __ AND cell __ is __ , then return a value of 1, if not, then return a value of 0”
IF, OR, AND Example

- We want to make a variable that represents those who “strongly agree” with “regulating commercial fishing”
- Insert a column to the right of “mng_commfish”
- In cell GS2, type “=IF(GR2=5,1,0)”
- We need to keep missing values consistent, so insert another column to the right of column GS and in cell GT2, type “=IF(GR2=".",".",GS2)”
- Copy both formulas down the rest of the spreadsheet
- Column GT can now be named “mng_commfish_SA”
- Copy GT, paste as values, delete column GS

<table>
<thead>
<tr>
<th>OR</th>
<th>GS</th>
<th>GT</th>
</tr>
</thead>
<tbody>
<tr>
<td>mng_commfish</td>
<td>mng_commfish_SA</td>
<td>mng_commfish_A or SA</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

IF, OR, AND Example

- We want to make a variable that represents those who “strongly agree” OR “agree” with “regulating commercial fishing”
- Insert 2 columns to the right of “mng_commfish_SA”
- In cell GT2, type “=IF(OR(GR2=4,GR2=5),1,0)”
- In cell GU2, type “=IF(GR2=".",".",GT2)”
- Copy both formulas down the rest of the spreadsheet
- Column GU can now be named “mng_commfish_A or SA”
- Copy GU, paste as values, delete column GT

<table>
<thead>
<tr>
<th>OR</th>
<th>GS</th>
<th>GT</th>
</tr>
</thead>
<tbody>
<tr>
<td>mng_commfish</td>
<td>mng_commfish_SA</td>
<td>mng_commfish_A or SA</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
IF, OR, AND Example

- We want to make a variable that represents those who fish for “fun” AND also fish to “sell”
- Insert 2 columns to right of column H
- In cell I2, type “=IF(AND(G2>1,E2>1), 1,0)”
- In cell J2, type “=IF(OR(G2=".", E2=".", ".", I2))”
- Copy both formulas down the rest of the spreadsheet
- Column J can now be named “fish_funANDsell”
- Copy J, paste as values, delete column I

VLOOKUP Example

- This function is great for survey data and coding
- If we want to know the text that the numeric codes represent
- Let’s examine “consume_fish” ; what do each of the codes mean?
- Insert a new column to right of column J
- In a new sheet type out the “lookup array”
- In cell K2, type “=VLOOKUP(J2,Sheet1!$A$1:$B$8,2,FALSE)”
  - Dollar signs lock your cells in your table array
- New variable can be named “consume_fish_text”
- Copy, then paste as values
The formula bar is just above the column names (A, B, C, etc.)
Formulas can be directly typed into here

If you don’t want to type directly into the formula bar, you can click the Function button:

After you click, the window to the right will open so that you can search and choose the type of function you need.
Formula Bar

- After you chose which function to use, the function window will show each component that the function needs to operate
  - In the case of “=SUMIF”, you need a range, a criteria, and a sum range
  - Each of these is described if you click in the field
  - You can type the cell range you want, or you can click on the cells themselves and drag

Formula Tips

- “$” in front of a column letter or a row number will “lock” the formula to that particular row, column, or cell
  - Highlight the cell or cell text in the formula bar and hit F4
- If you want to copy and paste, but not paste the formula, you can “paste values”
DATA ENTRY AND CODEBOOK CREATION

Day 1: Monday September 12, 2016

Data Entry and Codebook Creation

Day 1: Monday September 12, 2016
Coding Data – Creating a Codebook

- A codebook is necessary for any data set
  - Codebooks provide information concerning each variable in the data set and what the codes for each variable represent
- Continuous number - enter as is, with standardized units (i.e. percent, dollars, etc.)
- Categorical – create numerical codes
  - Binary – enter as 0 or 1
  - Nominal – assign a number code that will be easy to remember, such as the order listed on the survey instrument
  - Ordinal – assign a number code that matches the order of the data
- Open ended – directly transcribe open-ended text
  - You may create codes for these later, we will cover this later in the workshop
- For all data, use 0 to indicate true zeros
- Code missing values as a “.” (period) or as a “#null!”
- Include comment fields where needed (usually in final column)

Continuous Variable

Q3. Of the seafood that you and your family eat, how much of it comes from Merizo? % [INTERVIEWER: TRY TO GET A PERCENTAGE OR WRITE DON'T KNOW]

<table>
<thead>
<tr>
<th>Question Number</th>
<th>Question</th>
<th>Variable Name</th>
<th>Answer Options</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>3</td>
<td>Of the seafood that you and your family eat, how much of it comes from Merizo?</td>
<td>Percent_Merizo</td>
<td>Continuous (percentage from 0-100)</td>
</tr>
</tbody>
</table>

- Open the file “Manell_Geus_codebook.xlsx”
- Question #3 is coded as a continuous variable because it asks the respondent to specify the percentage of their family’s seafood that comes from Merizo and has a true zero point.
Ordinal Variable

<table>
<thead>
<tr>
<th>Question Number</th>
<th>Question</th>
<th>Variable Name</th>
<th>Answer Options</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>7.1 Coral reefs protect Guam from coastal/shoreline erosion and natural disasters like typhoons and tsunamis</td>
<td>agreement_protect</td>
<td>Strongly Disagree</td>
<td>1</td>
</tr>
<tr>
<td>51</td>
<td>7.2 Diving and snorkeling are not harmful to coral reefs.</td>
<td>agreement_divesnoor</td>
<td>Disagree</td>
<td>2</td>
</tr>
<tr>
<td>52</td>
<td>7.3 Coral reefs provide sustainable resources that support the development of our Micronesian communities.</td>
<td>agreement_resources</td>
<td>Neither agree nor disagree</td>
<td>3</td>
</tr>
<tr>
<td>53</td>
<td>7.4 Coral reefs have an important role in our culture</td>
<td>agreement_culture</td>
<td>Agree</td>
<td>4</td>
</tr>
<tr>
<td>54</td>
<td>7.5 Coral reefs are important to my family’s way of life</td>
<td>agreement_life</td>
<td>Strongly Agree</td>
<td>5</td>
</tr>
<tr>
<td>55</td>
<td>7.6 Effects from climate change can severely affect coral reefs.</td>
<td>agreement_climate</td>
<td>Not Sure</td>
<td>8</td>
</tr>
</tbody>
</table>

- Question #7 is coded as an ordinal variable because the order of the responses is significant (higher numbers indicate more agreement), but the differences between each one is not quantifiable.

Nominal Variable

<table>
<thead>
<tr>
<th>Question Number</th>
<th>Question</th>
<th>Variable Name</th>
<th>Answer Options</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>106</td>
<td>Please answer yes or no if you or your family do these activities, and whether you do them in A-hang or in both A-hang and the Cocos lagoon areas</td>
<td>activity_gather</td>
<td>no</td>
<td>1</td>
</tr>
<tr>
<td>107</td>
<td>Gathering of animals from the reef (ex. trochus (aillinging), clams (hima), sea cucumbers (balate), octopus (gaman))</td>
<td>activity_fish</td>
<td>yes, in Cocos Lagoon</td>
<td>2</td>
</tr>
<tr>
<td>108</td>
<td>Fishing (ask the following fishing methods only if they fish)</td>
<td>activity_spear</td>
<td>yes, in A-hang preserve</td>
<td>3</td>
</tr>
<tr>
<td>109</td>
<td>Spear fishing</td>
<td>activity_castnet</td>
<td>yes, in both places</td>
<td>4</td>
</tr>
<tr>
<td>110</td>
<td>Gillnet, surround net and drag net-fishing (tekken, chenchulu)</td>
<td>activity_gillnet</td>
<td>not sure</td>
<td>8</td>
</tr>
</tbody>
</table>

- Question #13 is coded as a nominal variable because all of the responses are mutually exclusive and none of them have any numerical significance.
- In this case, the codes are basically just labels, there is no relationship between the numbers themselves.
Binary Variable

Q15. Does your household benefit from the Achang Marine Preserve?

1. Yes 2. No

- Question #15 is coded as a binary variable because there are only 2 choices
  - Yes, they receive benefit
  - No, they do not receive benefit

Open Ended Text Variable

- Question #23.8 is coded as an open text variable because it asks the respondent to specify if their household participates in a pro-environmental activity (if any) that is not listed in the survey
Coding missing data

- DO NOT leave blanks
- Letters can create data type conflicts
- Symbols (- * .) may be read as functions
- Excel, SPSS, and SAS will interpret the period (".") to be a missing value and will not take it into account when performing calculations
- Be consistent
- Be prepared to remove, replace, or exclude codes later

Special Cases in Coding: Multi-Part Question

Q5. In your opinion, how is each of the following natural resources currently doing in Merito? We'll use the following scale: 1 very bad, 2 bad, 3 neither bad nor good, 4 good, or 5 very good.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Ocean water quality (clean and clear)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.2</td>
<td>Amount of coral</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.3</td>
<td>Size of Fish</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.4</td>
<td>Number of fish</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.5</td>
<td>Number of fish that eat seaweed or algae</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.6</td>
<td>Number of turtles</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.7</td>
<td>Beach/shoreline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.8</td>
<td>Stream water quality (clean and clear)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.9</td>
<td>Forest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- In a multi-part question such as this, we need to create variable codes for each part of the question
  - i.e. ocean water quality gets its own column, amount of coral gets its own column, etc.
- However, part of the variable name should correspond to this “question group”
- This is an ordinal variable, so the coding should reflect that
Special Cases in Coding: Multi-Part Question

The leading word of “condition” is used to show that these individual questions are grouped together.

The 1-5 scale goes in order from bad to good, indicating its ordinal nature.

Special Cases in Coding: “Rank Top 3”

In a multiple response ranking question such as this, we need to create variable codes for each part of the question.

- i.e. coral bleaching gets its own column, ocean acidification gets its own column, etc.

Again, part of the variable name should correspond to this “question group”.

This is a binary variable (i.e. the respondent can only say “Yes, this is a top threat” or “No, this is not a top threat”), so the coding should reflect that.
Special Cases in Coding: “Rank Top 3”

- The leading word of “threat” is used to show that these individual questions are grouped together
- The 0-1 scale makes calculations of averages easier

![Image of a table showing threats and their codes]

Special Cases in Coding: “Other – please list”

- In this group of ordinal-type questions, there is an open ended questions at the end
- For this, two separate variables must be created:
  - One for the “success rating” of “other”
  - One to specify what the respondent put as “other”
- Again, part of the variable name should correspond to this “question group”
Special Cases in Coding: “Other – please list”

- The leading word of “success” is used to show that these individual questions are grouped together
- Success_other is rated on the 1-5 ordinal scale
- Success_other_specify is an open ended text response

Data Cleaning

- Things to look for:
  - Missing or duplicate records
  - Missing data
  - Values out of range
  - Varying format
  - Inconsistencies in text fields
  - Data in the wrong field
  - Numbers or text in cells below the last row of data
    - This can affect calculations when a data set is imported from Excel to SPSS
Sorting and Filters

- Both are under the “Data” tab of the Excel Ribbon
- Sorting:
  - Order your data based on a column or set of successive columns
  - Sort numerically (high-to-low and low-to-high) or alphabetically (A-to-Z and Z-to-A)
- Filters:
  - Quick viewing and sorting of data
  - View subsets of data
  - Perform simple queries with one or more variables
  - Avoid data error and loss

Sorting Example

- Open “Manell_Geus_SortFilter.xlsx”
- For if you want your data set sorted based on a certain variable
- We want to order our data based on tenure in Merizo from high to low
- Under the data tab, click “Sort”
- In “sort by,” pick “tenure”
- In “order,” pick “largest to smallest”
Sorting Example

- Multiple levels can be added to the sort
- We want to sort by “income,” from low to high, then sort by “tenure” from low to high
- Under the data tab, click “Sort”
- Click “add level”
- In the first “sort by,” pick “income”
- In the second “sort by,” pick “tenure”
- In “order,” pick “smallest to largest” for both

Filter example

- Under the data tab, click “filter”
  - This will enable the filter function for all columns
- This function is used if you only want to observe certain data
- For example, if we want to only look at respondents who fish or harvest for marine resources
- Go to “fish_harvest” in column C and click on the filter button in that column
- Select that you only want to view data with a code of 1 for “fish_harvest”
Filtered Data

<table>
<thead>
<tr>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>fish_harvest</td>
<td>fish_myself</td>
<td>fish_sell</td>
<td>fish_give</td>
<td>fish_fish</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>.</td>
<td>4</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Data Management

- **Best practices:**
  - *Unique records*
  - *Backups*
  - *Metadata (data about your data)*
  - *Keep data file, codebook file, and work log in the same place*
    - Can be the same workbook, same folder, same hard drive, etc.
  - *Analysis log (keep track of your workflow and all decisions made)*
Documenting Workflow

- Data Entry and Analysis log
  - *You WILL forget what you did – write it down!*
- As you’re entering and coding data, if an analysis idea comes to mind, write/type it down
- All data decisions need to be kept track of
  - *Keep track of any issues or inconsistencies with the data entry*

Example Data Entry Error Decision – The Skip Pattern

*Example: There are a few data entry errors related to a “skip pattern”*

---

**Q12. Are you familiar with the Achang Reef Flat Marine Preserve?**

[Interviewer: If respondents are somewhat familiar, choose “YES”]

1. Yes
2. No

[Interviewer: If the respondent answers NO, please ask only Q13 and Q14, then skip to Q18. If YES, please proceed with all questions.

Provide all respondents with Map of Achang/Cocos Lagoon.]

**Q15. Does your household benefit from the Achang Marine Preserve?**

1. Yes
2. No

*If yes, what are the three most important benefits? [Interviewer: Do not read. Check only 3 answers based on the responses. Prompt if necessary.]*

1. Increasing the amount of fish for your household consumption or sale
2. Improving marine-related job opportunities
3. Protecting cultural heritage of your family
4. Improving recreational use by your family
5. Other: Please list
Example Data Entry Error Decision – The Skip Pattern

- This is a hypothetical data set
- Respondent 4 and Respondent 7 both answered that they are not familiar with the Achang Preserve
- Therefore, these respondents should not have been asked if their household benefits form the Preserve
  - Respondent 6 and Respondent 8 were coded correctly
  - There should not be responses for the “benefit” variables if there is a “0” under “familiar_Achang”
    - These “non responses” should be denoted with the period (“.”)
- A data coding decision must be made and logged

The Data Decision

Since the skip pattern was not followed correctly, and the survey instrument specifies that respondents who were not familiar with the Achang Preserve to NOT be asked the questions about their household receiving benefits from Achang Preserve, we must re-code the data according to how the survey instrument is meant to be employed
Documenting the Decision

<table>
<thead>
<tr>
<th>Date</th>
<th>Issue</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>9-12-16</td>
<td>Skip pattern was not followed correctly from Question 12; some respondents were allowed to answer Question 15 even if they chose “no” for Q12</td>
<td>Since data should not have been entered for Q15, all data values were changed to missing values to comply with survey instrument</td>
</tr>
</tbody>
</table>

• Keep a running log of all decisions like this and be as detailed as possible
• So that you will remember your logic and motivation for making the decisions
• And so you can explain yourself to colleagues and others
Introduction to SPSS

Day 1: September 12, 2016

Basic Overview

- SPSS can be used to analyze data and form statistical conclusions
- User-friendly drop down menu format
- Can perform more types of analyses when compared to Excel
- Can perform many of the same analyses as Excel, but faster
Opening SPSS

- Double click on the SPSS desktop icon to open the software
- This is the first step before entering/importing data

Blank SPSS Data Set
Data and Output Windows

■ When you click to open an SPSS data set, two separate windows are opened:

- **Data Editor Window** displays your data set and the variables inside your data set
- **Output Viewer Window** displays any resulting output from using SPSS functions in analyzing your data

*Analysis functions can be used in both windows*
Data Editor Window

- **Data View** – Shows your data set with each variable sitting in its own column and the data underneath the variable heading.

- **Variable view** – Shows each variable in your data set in rows with the characteristics (variable classification, number of decimal places, etc.) of each variable in the columns.

Data View
Manually Entering Data

*Open the file “Manell_Geus_ManualEnter.xlsx”

Inserting a Variable

*Untitled2 [DataSet] - IBM SPSS Statistics Data Editor
Naming the Variable

Classifying the Variable
Classifying the Variable

![Image of a software interface with columns for Name, Type, Width, Decimals, Label, Values, Missing, Columns, Align, Measure, and Role. The interface shows columns for Respondent_ID, fish_hornet, fish_self, and fish_sue, each with numeric types and specified attributes.]
A Manually Entered Data Set

**Data View:**

![Data View Image]

**Variable View:**

![Variable View Image]

Importing a Dataset

![Importing a Dataset Image]
Opening an SPSS data file

Open the "Manell_Geus_Getting Started.xlsx" file

Opening an Excel Data File

Open the "Manell_Geus_Getting Started.xlsx" file
Creating a Codebook in SPSS

- The first step is to format the “Variable View” window correctly
  - Classifying your variables as numeric or text strings
  - Classifying your variables as nominal, ordinal, or scale
  - Input values and codes (i.e. 1 = “yes”; 0 = “no”)

- Initially, all variables have been entered as “numeric” variables
  - We established in our codebook that some variables must be coded as “open ended” text responses
  - We must change these to “string” variables as they should not be analyzed numerically

- Initially, all variables have been entered as a “nominal” type of measure
  - Some of our variables are “ordinal” and some are “scale” (interval or ratio); therefore they must be coded as such

Numeric or String?

![Image of SPSS Variable View window with variables and types]
### Type of Measure?

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Width</th>
<th>Decimals</th>
<th>Label</th>
<th>Missing</th>
<th>Values</th>
<th>Column</th>
<th>Align</th>
<th>Measure</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>String</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Left</td>
<td>Nominal</td>
<td>Input</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Numeric</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Right</td>
<td>Nominal</td>
<td>Input</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Numeric</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Right</td>
<td>Ordinal</td>
<td>Input</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Numeric</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Right</td>
<td>Ordinal</td>
<td>Input</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Numeric</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Right</td>
<td>Ordinal</td>
<td>Input</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Numeric</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Right</td>
<td>Ordinal</td>
<td>Input</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Numeric</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Right</td>
<td>Ordinal</td>
<td>Input</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Numeric</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Right</td>
<td>Ordinal</td>
<td>Input</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Numeric</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Right</td>
<td>Ordinal</td>
<td>Input</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Numeric</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Right</td>
<td>Ordinal</td>
<td>Input</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Numeric</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Right</td>
<td>Ordinal</td>
<td>Input</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Numeric</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Right</td>
<td>Ordinal</td>
<td>Input</td>
<td></td>
</tr>
</tbody>
</table>

### Type of Measure?

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Width</th>
<th>Decimals</th>
<th>Label</th>
<th>Missing</th>
<th>Values</th>
<th>Column</th>
<th>Align</th>
<th>Measure</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>String</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Left</td>
<td>Nominal</td>
<td>Input</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Numeric</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Right</td>
<td>Nominal</td>
<td>Input</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Numeric</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Right</td>
<td>Ordinal</td>
<td>Input</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Numeric</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Right</td>
<td>Ordinal</td>
<td>Input</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Numeric</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Right</td>
<td>Ordinal</td>
<td>Input</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Numeric</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Right</td>
<td>Ordinal</td>
<td>Input</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Numeric</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Right</td>
<td>Ordinal</td>
<td>Input</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Numeric</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Right</td>
<td>Ordinal</td>
<td>Input</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Numeric</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Right</td>
<td>Ordinal</td>
<td>Input</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Numeric</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Right</td>
<td>Ordinal</td>
<td>Input</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Numeric</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Right</td>
<td>Ordinal</td>
<td>Input</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Numeric</td>
<td>12</td>
<td>0</td>
<td>None</td>
<td>None</td>
<td>12</td>
<td>Right</td>
<td>Ordinal</td>
<td>Input</td>
<td></td>
</tr>
</tbody>
</table>
Illustrating Variable Codes
Illustrating Variable Codes

Copy/Pasting Variable Codes
Copy/Pasting Variable Codes

Open the file "Manell_Geus_datafile1.sav"

All Variables Classified and Coded

Open the file “Manell_Geus_datafile1.sav”
Creating a Codebook in SPSS

1. Open SPSS and load your dataset.
2. Go to the 'Analyze' menu and select 'Descriptive Statistics' then 'Descriptives'.
3. Click on the variable(s) you want to include in the codebook and move them to the right-hand window.
4. Click 'Options' and select the variables you want to include in the codebook.
5. Click 'Continue' and then 'OK' to generate the codebook.

The codebook will display the following information for each variable:
- Name
- Variable type
- Value labels
- Mean
- Standard deviation
- Minimum
- Maximum

You can also export the codebook as a PDF file for future reference.
Creating a Codebook in SPSS

- The Output window in SPSS will then display variable-by-variable information and will also provide frequencies of each code and number of missing values if data is available.

- Note that code values were not imputed for open ended responses (text or number).
What is Qualitative Data?

- Data that is observed, describe to approximate or characterize but does not measure
  - We seek to understand and interpret peoples’ responses
- Arranged into categories that are not numerical
  - These categories can be physical traits, gender, colors or anything that does not have a number associated to it
- The product is richly descriptive; words not numbers
Sources of Qualitative Data

- Answers to open-ended questions
- Quotations from interviews or focus groups
- Observations of activities or behaviors
- Document excerpts, quotations, or passages

Who Should Collect the Data?

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Insider</strong></td>
<td><strong>Cons</strong></td>
</tr>
<tr>
<td>o In-depth context knowledge</td>
<td>o Differences in power can affect data</td>
</tr>
<tr>
<td>o Personal connections &amp; trust</td>
<td>o Can’t guarantee anonymity</td>
</tr>
<tr>
<td>o Aware of power dynamics</td>
<td>o Might miss things that are “too normal” to notice</td>
</tr>
<tr>
<td>o Logistically easier (finding informants etc.)</td>
<td></td>
</tr>
</tbody>
</table>

| **Outsider**              |                           |
| o Can ask the “silly” questions | o Differences in power can affect data |
| o Can see beyond the local context | o Prone to making mistakes |
| o Can provide anonymity    | o Longer to create trust       |
| o Can be perceived as “neutral” |                         |
Ethical Concerns

■ **Research fatigue** - people have other things to do!

■ **Anonymity** - from the interview to the handling of the data

■ **Coercion** – FPIC! Never force or guilt people into participation

■ **Respect for the community & individuals** - don’t just extract data and run

How Do You Collect Qualitative Data?

■ **In depth interviews**
  - Include both one-on-one interviews as well as “group” interviews (e.g. focus groups)
  - The data can be recorded in many ways including not taking, audio recording, video recording

■ **Direct observations**
  - Differ from interviewing in that the observer does not actively interact with the respondent
  - Ranging from field research where one lives in another context or culture for a period of time to photographs that illustrate some aspect of the phenomenon
  - Data can be recorded in the same ways as interviews (written notes, audio, video) and through pictures, photos or drawings

■ **Written documents**
  - Usually this refers to existing documents (as opposed to transcripts of interviews conducted for the research)
  - It can include newspapers, magazines, books, websites, memos, transcripts of conversations, annual reports, etc.
  - Usually written documents are analyzed with some form of **content analysis**
Content Analysis

- A research methodology that examines words or phrases within a wide range of texts
- A research technique used to make replicable and valid inferences by interpreting and coding textual material
- By systematically evaluating texts (e.g., documents, oral communication, and graphics), qualitative data can be converted into quantitative data
  - *It is an important bridge between purely quantitative and purely qualitative research methods*
- Through content analysis, we can examine the prevailing themes in a qualitative data set, code them accordingly, and analyze the results

Rigor and Validity

- **Rigor**: derives from the researcher’s presence, the nature of the interaction between researcher and participants, the triangulation of data, the interpretation of perception, and thick description

- **Validity**: whether the conclusions being drawn from the data are credible, defensible, warranted, and able to withstand alternative explanations.
How to Promote Rigor and Validity

- Triangulation
- Respondent validation
- Adequate engagement in data collection
  (saturation of data, time, discrepant analysis)
- Reflexivity
- Peer review
- Audit trail

Qualitative or Quantitative?

- **QUANTITATIVE**
  - Σ
  - ≠
  - X²
  - %
  - ≈
  - ±

- **QUALITATIVE**
  - what?
  - context
  - why?
  - sticky questions
  - thick description
  - how?
Qualitative or Quantitative?

It depends on your question!

Different Approaches

**Quantitative**
- Good for: Generalizable results
- Purpose: Predicting & Testing
- Questions: Structured
- Sampling: Large & Random
- Collecting Method: Experiments & Surveys
- Data: Numbers

**Qualitative**
- Purpose: In-depth understanding
- Exploring in Context
- Questions: Semi-structured & Open
- Sampling: Small & Purposeful
- Collecting Method: Interviews, Focus Groups, Ethnography & Observation
- Data: Words
Different Approaches: Analysis

**Quantitative**
- Data Analysis starts after data collection is completed
- Researcher can be distant, removed from the research site
- Collecting methods include experiments, surveys

**Qualitative**
- Data analysis and collection take place simultaneously
- Researcher is the data collecting instrument, often in the field
- Collecting methods include interviewing, facilitated focus group, ethnography (participating observation, key informant interviews)

Mixed Methods: Qualitative and Quantitative

- Mixed methods research offers you the best of both worlds:
  - *The in-depth, contextualized, and natural but more time-consuming insights of qualitative research*
  - *The more-efficient but less rich or compelling predictive power of quantitative research*
- The purpose of this form of research is that both qualitative and quantitative research, in combination, provide a better understanding of a research problem or issue than either research approach alone
- The Manell-Geus Data set is a good example of mixed methods research in that we are collecting qualitative and quantitative data through the same survey instrument to be analyzed together
Mixed methods

Mixed Methods are good for
Triangulation of results
Problem based approaches
Collaborative interdisciplinary approaches

But watch out for
Loss of rigor & validity
Lack of disciplinary interaction
2.1 Where can a formula be typed in Excel?

A. Into the formula bar
B. Into an individual cell
C. By using the Function box
D. All of the above
2.2 True or False: A missing data point should be coded with a zero

A. True
B. False

2.3 Which of the following is qualitative data?

A. The percentage of people who participate in beach clean ups
B. The number of times a household has been affected by a flood
C. Respondents’ opinions concerning the success of coral reef management
D. Respondents' Age
2.4 Which of the following is not sources for qualitative data?
A. Interviewing
B. Observation
C. Close-ended questions in a survey
D. Excerpt from a document

2.5 True or False: Quantitative methods can not be used with qualitative methods in the same research project
A. True
B. False
2.6 True or False: You can open an Excel data set in SPSS

A. True
B. False
Day 2

- Qualitative data

- Descriptive statistics
Coding Open Text Qualitative Data
Day 2: September 13, 2016

Qualitative Data Analysis Process

Making sense out of the data is a process of consolidating, reducing, and interpreting what people have said and what the researchers have observed, read, and taken notes about.

“It involves going back and forth between concrete bits of data and abstract concepts, between inductive and deductive reasoning, between description and interpretation.”

(Merriam, 2009)
Deductive analysis

Inductive analysis

Qualitative Data Analysis Process

1. Collect & Transcribe
2. Code (Descriptive)
3. Code (Analytical/Interpretive)
4. Sort & Refine Categories
5. Synthesize, Interpret & Write
Analysis during data collection

“To wait until all data are collected is to lose the opportunity to gather more reliable and valid data.”
(Merriam, 2009)

- Qualitative data analysis happens simultaneously with data collection
- It is ongoing
- Emerging insights direct to the next phase of data collection

Steps for data collection

1. Develop questions as you go
2. Findings from previous data guide next sessions
3. All records, notes & transcribing done ASAP!
4. Document your learning process
5. Create an inventory system of the data set
6. Discuss findings with key informants to advance analysis and fill gaps
7. Enhance understanding with related literature
8. Clarify understanding with metaphors, analogies, concepts, & visuals

(Merriam 2009)
When to stop collecting data?

- No more time and/or money
- Diminishing return of new information

Examples of Questions

**Purpose:** to understand impacts of MPAs on people who live in the area.

**Questions:**

- *How do people adjust their way of living when the area became an MPA?*
- *What do they think about the benefits of the MPA to their household?*
- *How does MPA influence the way they think about marine and coastal resources?*
Steps for Descriptive Coding

① Summarize & extract meaning from text
② Establish categories, themes, or patterns in responses
③ Use shorthand devices to label, separate, compile, manage and organize data
④ Use shorthand devices to summarize, synthesize and sort observations made of the data
⑤ Use shorthand devices to easily retrieve specific pieces of data

Question: “If you could choose an occupation for your children, what would it be?”

Examples of Descriptive Coding

<table>
<thead>
<tr>
<th>Coding</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Past infrastructure</td>
<td>See, When I was young, the school on our island only has 4 grades and it was enough to learn to read. I wanted to go to the main island and study more, but I was the eldest son and my dad wanted me to help with sea cucumber collecting and trap fishing. There were lots of sea cucumber back then. My mother and sisters had to help cleaning and drying them for the market. It was lot of work. Now it is hard to find sea cucumbers and fishing is not as good. I want my children study as much as they want. (laugh..) I don’t know I can pay...Maybe they can find a job in the city. Our island is small and you can only live on fishing. In the city there are many jobs. Here boys can go fish but no job for girls. Here we only have enough to eat but never money to buy things, ...to save. Working in the city is better. You get cash. You can see a doctor when you’re sick.</td>
</tr>
<tr>
<td>- Past harvesting (sea cucumbers), past resource condition, past gender roles</td>
<td></td>
</tr>
<tr>
<td>- Change of marine resource conditions</td>
<td></td>
</tr>
<tr>
<td>- Perception of alternative livelihood</td>
<td></td>
</tr>
<tr>
<td>- Alternative livelihood</td>
<td></td>
</tr>
<tr>
<td>- Present infrastructure</td>
<td></td>
</tr>
</tbody>
</table>
Steps for Analytic Coding

1. Interpret and reflect on meaning
2. Group categories from first set of data
3. Keep a running list of groupings, either attached to the transcript or on a separate document
4. Repeat process on next set of data
5. Check for regularities and emerging categories between the list of the first and second data sets.
6. Merge groupings into one list, but keep the original lists
7. Move onto the next set of data, continue
8. Categories should emerge more and more clearly as they recur across your data

Steps to Sort Categories & Data

1. Creating codes for categories
2. Sort your data for each category into relevant codes as evidence for each category.
3. If needed revise and flesh out categories to make them more robust by searching through the data for more and/or better data units.
Criteria for Categories

- Responsive to research purpose and questions
- Exhaustive (enough categories to encompass all relevant data)
- Conceptually congruent (all categories are at the same conceptual level)

Steps to Synthesize, Interpret & Write

① Keep research purpose and questions in mind
② Write the answers to the research questions based on the categories, sub-categories, and data
③ Explain interrelationship of different (sub-)categories and how to make sense of the data
④ Use quotes and anecdotes that allow readers to quickly understand what you are trying to communicate
⑤ State analysis as simply as possible
⑥ Support quotes with interpretation (don’t make them speak for themselves)
Using Content Analysis

- Taking purely qualitative data and turning into quantitative to perform analysis by systematically evaluating texts

- Through content analysis, we can examine the prevailing themes in a qualitative data set, code them accordingly, and analyze the results
Coding Open Text Data

- Coding is a preliminary component of analysis
- Summarizes and condenses data, according to themes
- Not a precise science – it is primarily an interpretive act
- Generally requires refinement and multiple iterations
- Helps to get perspectives of multiple individuals to verify consistency of interpretation

What to Code for?

- Pre-assigned themes
  - Relevant to a certain research or policy need (e.g. fishermen interviews – support or lack of support for instituting MPAs?)
  - Relevant to research theory (supporting or in contrast)
- “Serendipity” – themes that emerge while analyzing for themes
  - (e.g. unexpected reasons for resource use, such as religious, cultural...)
Coding Methods

Multiple methods:

- Colored highlighters and post-it note flags on paper
- **Create “code” columns in excel**
  - *Useful for simple coding exercises, short open-ended questions*
- Qualitative analysis software (the high tech way...)
  - *Useful for more complicated coding exercises, lengthy text files*
  - *Can link themes, more easily create sub-themes...*

Background Example

- In NOAA’s National Coral Reef Monitoring Program (NCRMP), we survey the people in the US coral jurisdictions to understand human use, resource management, and peoples’ knowledge, attitudes and perceptions toward reefs
- In the Hawaii survey, we asked people to “define their community”

16. How involved is your local community in protecting and managing coral reefs?
   a. Not at all involved
   b. Somewhat involved
   c. Moderately involved
   d. Involved
   e. Very involved
   f. Not sure

17. In thinking about the previous question, how would you briefly define “your local community”? *[open ended]*
Background Example

- 421 people responded to this question
- Here is an excerpt of the first 20 responses

<table>
<thead>
<tr>
<th>Example: How do you define “community”?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>in relation to marine resource management</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category 1: spatial, social, activity</th>
<th>Category 2: community involvement, activism, and character</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definitions of community</td>
<td>Qualities related to interest and concern for environmental issues</td>
</tr>
<tr>
<td>Theme 1: GE (spatial)</td>
<td>Theme 1: EI (interest and involvement)</td>
</tr>
<tr>
<td>Theme 2: SO (social)</td>
<td>Theme 2: QC (general community quality or characteristic)</td>
</tr>
<tr>
<td>Theme 3: AL (activities and livelihood)</td>
<td>Theme 3: AA (activism and apathy)</td>
</tr>
</tbody>
</table>
Background Example

- 2 categories, broken down into 3 themes each
  - Themes are also broken down into “sub-themes”

### Category 1: Spatial (GE) – Sub-themes:

<table>
<thead>
<tr>
<th>Code/name</th>
<th>description</th>
<th>Second-order theme</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>GE 1</td>
<td>Political</td>
<td></td>
<td>n/a</td>
</tr>
<tr>
<td>GE 2</td>
<td>Landscape</td>
<td>2.1 Landscape (general)</td>
<td>County of Honolulu, Kailua community, Volcano, Hawaii</td>
</tr>
<tr>
<td>GE 3</td>
<td>Island</td>
<td>3.1 Island</td>
<td>east side of big island of Hawaii; Maui</td>
</tr>
<tr>
<td>GE 4</td>
<td>Area and Neighbors</td>
<td>4.1 Area (general)</td>
<td>Rural, city, small town, large village “within 50 miles from Hilo” “country” rural “neighbors included here because relationship is spatially defined “neighbors gather and talk”</td>
</tr>
</tbody>
</table>

Findings: Defining “Community”

When respondents were asked the question “How would you define your local community?” answers fell into two categories (see figure 1). In the first category, respondents (38% of the total sample) provided a definition for what they considered to be their local community. In the second category (76% of the total sample), respondents described community involvement (or lack thereof) in marine resource management. Many respondents (26%) provided both a definition of community as well as a description of how their community was or was not engaged. Twelve percent of respondents stated that they could not define their local community.

**Category 1: Defined ‘Local Community’**

The majority of respondents who provided a definition for their local community provided a spatial or geographic definition (61%). Social characteristics were also mentioned (42%), as were references to activities and sources of livelihood (35%). While definitions of ‘local community’ varied tremendously, the types of definitions mirrored definitions within the academic literature, which has multiple definitions of community generally falling into 3 broad categories:

1) geographic expression, 2) local social system, and 3) relationship type (eg. identity). Most individuals are in fact members of multiple communities at any point in their life.

**Category 2: Described Community Involvement**

Rather than define their ‘local community,’ most respondents (76%) described how their community was involved in marine resource management. The majority of respondents who provided a definition of local community also provided additional detail describing their community (60%). This indicates that community characteristics and actions are a fundamental component of how people relate to and define their community. Indeed, members giving their time, effort, and devotion to the public good are considered to be sustaining forces for community.
Breakout Session

From:
American Samoa Community-based Fisheries Management Area Program village survey

Survey Questions Regarding Education and Outreach

Q. 43 = open-ended – needs post-coding

40. Has a VMIA outreach activity happened in your community?
   Yes, No, Not sure

41. If yes, were you able to attend?
   a. Yes, I attended
   b. No, I was not able to attend
   c. No, I did not want to attend

42. If yes, what topics were discussed? (circle all that apply)
   a. Coral Reefs
   b. Fishing Regulations
   c. Erosion
   d. Piggeries
   e. MPAs
   f. Other
   g. Not sure

43. Are there any outreach topics that you would like to have presented to your community?
Breakout Session

From:
American Samoa Community-based Fisheries Management Area Program village survey:

- Are there any outreach topics that you would like to have presented to your community?

*What are the common themes that you find in these answers?*

*Code and analyze responses*

---

Breakout Session

- Open “Open Text Coding Lesson Data.xlsx”
- Break into groups and go through process of coding open text data
- Each group will be responsible for a set of responses
  - 302 total responses
Breakout Session

■ Step 1:
  – See if any answers to Q43 would fit into themes of Q42 (in the file “Q40-43.png”)
    ■ Example: Row 3 – “Fishing advice” could fit into “fishing regulations” theme
    ■ Example: Row 37 – “Topics that will benefit the reefs” could fit into “coral reefs” theme

■ Step 2:
  – Some of these open text responses will not fit into pre-determined theme areas, so you must
develop your own general theme areas

■ Step 3:
  – Some open text responses may fall into multiple theme areas
    ■ This is ok, we can code these as such
    ■ Example: Row 51 – “MPAs, Streams: how to reduce pollution/destructive fishing” could fit
      into themes of “MPA importance,” “pollution,” and “destructive fishing”

■ Step 4:
  – Create a numeric code for each theme and document your coding

■ Step 5:
  – Determine more general “categories” that your themes can fit in

Breakout Session

■ Discuss coding results
  – Challenges
  – Lessons learned

■ We have our themes, now what are our categories?
  – Determine 2-4 main general categories that responses
    can fall under and place into spreadsheet

■ We will analyze after the break!
Analyzing Qualitative Data

Day 2: September 13, 2016

Time to Analyze!

- Based on previous discussion/breakout group in coding open text data
- Open “Open Text Coding Lesson Data.xlsx”
- Use codes that were generated to understand what types of outreach that American Samoan survey respondents want
  - We can examine percentages and overlaps in Excel
  - What % fall into each theme?
  - What % fall into each category?
Visualizing Qualitative Data

Timelines

(Brooks et al. 2008)
Visualizing Qualitative Data

Example: Gender differences in ecological mental models

(Levine et al. 2016)

Example: causes and effects of degradation of fisheries resources

A “flow chart”
Visualizing Qualitative Data

Venn Diagrams

This diagram illustrates the categories that the data are sorted into, as well as the overlap between the categories that may exist.

Open the file “Community Poster Draft.pdf”

The Process

- Once qualitative data is categorized, sorted, and coded according to category, we can then quantitatively examine the distribution of responses.

- Example: X% of respondents fell into category 1
  Y% of respondents fell into category 2
  Z% of respondents fell into category 3

- Example: A% of category 1 responses contained sub-element #1
  B% of category 1 responses contained sub-element #2
  C% of category 1 responses contained sub-element #3
Practice!

- Open “Open Text Coding Lesson Data.xlsx”
- Open a blank PowerPoint Presentation
- Using previous discussions and coding from breakout session, let’s make a venn diagram and table of responses similar to Levine et al. (2016)
3.1 When does qualitative data analysis take place?

A. Before data collection  
B. After data collection  
C. During data collection  
D. During AND after data collection
3.2 True or false: We can use quantitative data analysis methods with open-coded text responses

A. True
B. False

3.3 What is content analysis?

A. The process of taking qualitative data and turning it into quantitative data through the use of interpretation and systematic coding to perform analysis
B. The diminishing returns of new information as sample size increases
C. Keeping the initial research questions in mind throughout the data analysis process
D. Stating and communicating your analysis
3.4 Check each of the following that are criteria for sorting open text qualitative data into categories

A. Responsive to research purpose and questions
B. Exhaustive (enough categories to encompass all relevant data)
C. Accepted (the categories must be determined by the survey respondent)
D. Mutually exclusive (a relevant unit of data can be placed in only one category)
E. Conceptually congruent (all categories are at the same conceptual level)
F. Limited (only a certain number of data points can be placed in each category)

3.5 True or false: The amount of new qualitative information that you can obtain diminishes as sample size gets really big

A. True
B. False
Overview of Descriptive Statistics

Day 2: September 13, 2016

Types of statistics

- **Descriptive statistics** = statistics that describe or display data in a meaningful way
  - This is our focus today

- **Inferential statistics** = statistics that draw generalizable conclusions about a population based on a sample of that population
Purposes of Descriptive Statistics

- Summarize data
- Identify useful and interesting findings
- Reporting and communications
- “Getting to know” your data
  - *Exploratory analysis*

Things to consider

- What information is needed/what is the question or issue that your are addressing
- What subgroups are of interest
- Grouping/reclassifying
Measures of Central Tendency

- **Mean** – The average of set of values
- **Median** – The middle value in a list of numbers
  - 50% of the data points are greater than the value, and
  - 50% of the data points are less than the value
- **Mode** – The value that occurs most often in a set of values

Measures of Spread

- **Standard deviation (SD)** - Measures how far a set of numbers are spread out from their mean
  - Expressed in the same units as the data
  - Low SD: data points tend to be close to the mean
  - High SD: data points are spread out over a wider range of values
- **Variance** - The expectation of the squared deviation of a random variable from its mean
  - $SD^2$
- **Range** – The difference between the greatest value and the lowest value in a set of values
- **Quartiles** - Tell us about the spread of a data set by breaking the data set into quarters, just like the median breaks it in half
  - Interquartile Range (IQR) – 3\textsuperscript{rd} Quartile minus the 1\textsuperscript{st} Quartile
Examples of Central Tendency

- Open the file “Manell_Geus_Data_Freqs.xlsx”
- Let’s examine a continuous variable like “tenure”
- We can calculate measures of central tendency and spread in excel
- In cell IT310, type “=AVERAGE(IT2:IT307)”
  - 21.54 is your mean
- In cell IT311, type “=MEDIAN(IT2:IT307)”
  - 19 is your median
- In cell IT312, type “=MODE.SNGL(IT2:IT307)”
  - 20 is your mode
- While measures of central tendency are all different values, they are still close to one another

Examples of Spread

- In cell IT313, type “=MAX(IT2:IT307)-MIN(IT2:IT307)”
  - Our range is 67.75
- In cell IT314, type “=VAR.S(IT2:IT307)”
  - Our variance is 253.87
- In cell IT315, type “=STDEV.S(IT2:IT307)”
  - Our standard deviation is 15.93
- In cell IT316, type “=QUARTILE.INC(IT2:IT307,3)-QUARTILE.INC(IT2:IT307,1)”
  - Our IQR is 20
Outliers

- An outlier is an observation point that is distant from other observations.
- Means are more susceptible to outliers.
- Example:
  - Set of numbers \(1,2,3,4,5\)
    - Mean = 3; median = 3
  - Set of numbers \(1,2,3,4,99\)
    - Mean = 21.8; median = 3
- In many cases, a few outliers can distort statistical conclusions based on the mean.
  - Outliers are sometimes left absent from statistical analysis.

Frequency tables

- Frequency Tables are the easiest way to provide a snapshot of categorical data.
- They are meant to take a specific variable and break it down into its various possible values and examine the frequency of each.
- This can show:
  - What is the most frequent response?
  - What is the least frequent response?
  - Should some responses be grouped together?
  - Are the responses clustered around one value?
  - Are the responses spread out evenly throughout all values?
Frequency table example

- Let’s examine the Frequency Distribution of “education”
- Use COUNTIF functions
- Start in cell IM30 and IN310
- Copy frequency and “paste as values” on another sheet
- Widen the column in order to read all of the text
- Calculate percentage column
- Add a “total” column
- Add borders
- Add shading to top row

<table>
<thead>
<tr>
<th>Level of Education</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>8th grade or less</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>9th to 11th grade</td>
<td>16</td>
<td>5%</td>
</tr>
<tr>
<td>12th grade, high school grad, GED</td>
<td>138</td>
<td>47%</td>
</tr>
<tr>
<td>some community college or vocational training</td>
<td>77</td>
<td>26%</td>
</tr>
<tr>
<td>college graduate</td>
<td>56</td>
<td>19%</td>
</tr>
<tr>
<td>graduate school, law school, medical school</td>
<td>9</td>
<td>3%</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>296</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>

The Normal Distribution

- An arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.
- Most statistical tests assume a normal distribution
  - Height is one simple example of something that follows a normal distribution pattern:
    - Most people are of average height
    - The numbers of people that are taller and shorter than average are fairly equal
    - A very small number of people are either extremely tall or extremely short
How does standard deviation affect the Normal Distribution?

The larger the standard deviation, the further the individual cases are from the mean.

The smaller the standard deviation, the closer the individual scores are to the mean.

Histograms

- A histogram is a display of statistical information that uses rectangles to show the frequency of data items in successive numerical intervals of equal size.
Box Plots

- A box plot is a graphical rendition of statistical data based on the minimum, first quartile, median, third quartile, and maximum.
- The term "box plot" comes from the fact that the graph looks like a rectangle with lines extending from the top and bottom.

Other types of distributions

- Bi-modal (polarized opinions)
- Skewed left (score on an easy exam)
- Skewed right (annual income)
Descriptive Statistics in SPSS

Day 2: September 13, 2016

Overview

- SPSS is a great tool for descriptive analysis
- It can work faster than excel
- User-friendly drop down menu format

- Open “Manell_Geus_datafile.sav”
Frequency Tables

**Frequency Tables – Last10**

- Let's examine the frequencies of the “last10_” questions.
- How would you say the condition of each of the following has changed in Merizo over the last 10 years?
Frequency Tables – Last10

- **Frequency** – number of cases in the category
- **Percent** – percent of cases in the category
  - Missing values included in calculations
- **VALID percent** - percent of cases in the category
  - Missing values NOT included in calculations
- **Cumulative percent** – a “running total” of the percent of responses included in the current category, as well as all preceding categories too

Some Conclusions about “Last10_”

- In examining the frequency tables of the “last10_” questions, a few conclusions can be drawn:
  - **Most responses seem to be in the “somewhat better” category**
    - Except “beach,” “stream,” and “forest”
  - **People had the most negative perception about the change in the condition of beaches/shorelines (40.9% said “a lot worse” or “somewhat worse”)**
  - **People had the most positive perception about the change in the condition of the amount of coral and the number of fish (47.4% said “a lot better” or “somewhat better”)**
  - **People were the most not sure about the change in the condition of the number of fish that eat seaweed or algae and the number of turtles (8.9%)**
Frequency Tables – Fire

- Lets examine the frequencies of the “fire_” questions
- Please rate the severity of the fires on the Merizo community

### fire_public

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>very low</td>
<td>36</td>
<td>11.8</td>
<td>11.8</td>
<td>11.8</td>
</tr>
<tr>
<td>low</td>
<td>74</td>
<td>24.2</td>
<td>24.4</td>
<td>36.3</td>
</tr>
<tr>
<td>medium</td>
<td>131</td>
<td>42.8</td>
<td>43.2</td>
<td>75.5</td>
</tr>
<tr>
<td>high</td>
<td>36</td>
<td>11.8</td>
<td>11.9</td>
<td>91.4</td>
</tr>
<tr>
<td>very high</td>
<td>21</td>
<td>6.9</td>
<td>6.9</td>
<td>98.3</td>
</tr>
<tr>
<td>not sure</td>
<td>5</td>
<td>1.6</td>
<td>1.7</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>303</td>
<td>99.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

### fire_health

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>very low</td>
<td>36</td>
<td>11.8</td>
<td>11.9</td>
<td>11.9</td>
</tr>
<tr>
<td>low</td>
<td>74</td>
<td>24.2</td>
<td>24.4</td>
<td>36.3</td>
</tr>
<tr>
<td>medium</td>
<td>131</td>
<td>42.8</td>
<td>43.2</td>
<td>75.5</td>
</tr>
<tr>
<td>high</td>
<td>36</td>
<td>11.8</td>
<td>11.9</td>
<td>91.4</td>
</tr>
<tr>
<td>very high</td>
<td>21</td>
<td>6.9</td>
<td>6.9</td>
<td>98.3</td>
</tr>
<tr>
<td>not sure</td>
<td>5</td>
<td>1.6</td>
<td>1.7</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>303</td>
<td>99.0</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>3</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>306</td>
<td>100.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Some Conclusions about “Fire_”

- In examining the frequency tables of the “fire_” questions, a few conclusions can be drawn:
  - “Medium” is the most frequent response for all questions about rating the severity of fires
  - Respondents felt that fires impacted the community economy the least (40.9% said “very low” or “low” severity)
  - Respondents felt that fires impacted the public property and infrastructure the most (27.2% said “very high” or “high” severity)
  - Respondents were the most not sure about seawater quality (2.6%)

Summary Statistics

- While frequency tables are great for examining the distributions of nominal and ordinal variables, we want to use the “descriptives” function in SPSS for analyzing the distribution of continuous data
Summary Statistics

Let’s analyze the descriptive statistics of some continuous data:

- Percent_merizo
- Flood_impact
- Fire_impact
- Family
- Age
- Tenure

Summary Statistics

- Click “options”
- Click “range”
Some conclusions based on output:

- Some people get 100% of their seafood from Merizo
- On average, a third of peoples’ seafood comes from Merizo
- In the last 5 years, the average family has been affected by 2 floods (max 20) and 3 fires (max 60)
- The average family size is 6.35 persons (max 15)
- Average age = 40.38
- Average number of years living in Merizo = 21.55 (max 68)
Histograms in SPSS

- At this point, we have examined distributions through numbers and text, but what about graphics?
- A histogram can graphically display a variable’s distribution.

Then, double click the graph in the output viewer to pull up the “chart editor”.
Histograms in SPSS

- The histogram of the percentage of respondents’ food that comes from Merizo is depicted here to the right.
- The normal distribution curve is drawn for comparison.
- It looks as if the distribution of “percent_merizo” is not normal.
- It looks to be bi-modal.
  - We observe the highest frequencies in the low percentages and relatively high frequencies again in the high percentages.

Practice!

Generate a frequency table for “success_ecosystem” and “success_water”
- What is the mode for each?
- Which did respondents think was more of a success? (high or very high)
- Which were respondents more unsure about?
■ Mode of both = 3 = “medium”
■ Respondents felt that “Protecting the whole coral reef ecosystem” has been more of a success (27.2%)
■ Respondents were more unsure about “Protecting the whole coral reef ecosystem” (4%)

Practice!

Generate a frequency table for “mng_gillnet” and “mng_dregnnet”
- What is the mode for each?
- Which did respondents support more? (support or strongly support)
- Which were respondents more unsure about?
- Mode for both = 2 = “oppose”
- Respondents supported “Prohibit drag nets (chenchulu)” more (32%)
- Respondents were more unsure about “Prohibit gill nets (tekken)” (2.3%)

Practice!

- What is the mean of “familiar_achang”?
- What is the range of “flood_impact”?
- What is the interquartile range of “percent_merizo”?
  - Hint: use “frequencies” > ”statistics”
- What is the sample size of “agreement_protect”?
What is the mean of “familiar_achang”?  
- 56% are familiar with Achang Preserve

What is the range of “flood_impact”?  
- 20 (Respondents have been affected by a range of zero to 20 floods in the last five years)

What is the interquartile range of “percent_merizo”?  
- 1st quartile = 10; 3rd quartile = 50; IQR = 40

What is the sample size of “agreement_protect”?  
- 306

Practice!

Generate a Histograms with Normal curve comparison for “age” and “tenure”
“Age” appears to be Skewed right

- Evidenced by the long right tail
- Median is less than the mean

“Tenure” also appears to be Skewed right

- Evidenced by the long right tail
- Median is less than the mean

- Save your output as “Manell_Geus_Output_Descriptive.spv”
Data Visualization for Descriptive Statistics

Day 2: September 13, 2016

Communicating Data

- Visualizing your data and your analysis is how you can communicate your data to a wider audience.
- Not everyone is going to read all of the text in a lengthy report.
  - *While these reports are important for providing the entire context of a research project, the graphs/charts/figures are easier to understand and can reach a wider audience.*
Preparation steps

- Revisit the purpose of your study and select key messages to communicate.
- Decide who is your main audience and what communication tools and types of messages most appropriate to your audience.
- Ask what they will do with the results (what type of decision or reaction).
- Decide team tasks and schedule of products and their delivery/dissemination.

Design Objectives

- Eliminate clutter and distraction
- Group data into logical sections
- Highlight what’s most important
- Support meaningful comparisons
- Ensure readers can understand the message

Smith & Azzam (2010)
Telling stories with data

- **Facts** – TNC has protected 113,283,164 acres of land, which is greater than the entire state of California.

- **Comparisons** – TNC has protected more land in Meso-/South America than in the U.S. and Canada combined.

- **Trends and patterns** – The number of acres TNC has protected has increased by 13% since 2006.

- **Relationships** – Protected areas increase the value and desirability of surrounding lands (McConnell and Walls, 2005).

---

Graphs or tables?

<table>
<thead>
<tr>
<th>Tables</th>
<th>Graphs</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Look up individual values</td>
<td>• Message conveyed via shape</td>
</tr>
<tr>
<td>• Compare individual values</td>
<td>• Relationships among multiple values</td>
</tr>
<tr>
<td>• &gt; 1 quantitative unit of measure</td>
<td></td>
</tr>
</tbody>
</table>
Table design

- White space/light fill for delineation (avoid grids)
- Calculated columns to right
- Align numbers to right
- Align all non-number text to left
- Consistent row spacing
- Adjust columns to best fit
- Effective grouping

```
<table>
<thead>
<tr>
<th>Sales Rep</th>
<th>Quota</th>
<th>Variance to Quota</th>
<th>% of Quota</th>
<th>Forecast</th>
<th>Actual Bookings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albright, Gary</td>
<td>200,000</td>
<td>-16,062</td>
<td>92</td>
<td>205,000</td>
<td>183,938</td>
</tr>
<tr>
<td>Brown, Sheryl</td>
<td>150,000</td>
<td>84,983</td>
<td>157</td>
<td>260,000</td>
<td>234,983</td>
</tr>
<tr>
<td>Cartwright, Bonnie</td>
<td>100,000</td>
<td>-56,125</td>
<td>44</td>
<td>50,000</td>
<td>43,875</td>
</tr>
<tr>
<td>Caruthers, Michael</td>
<td>300,000</td>
<td>-25,125</td>
<td>92</td>
<td>324,000</td>
<td>274,875</td>
</tr>
<tr>
<td>Garibaldi, John</td>
<td>250,000</td>
<td>143,774</td>
<td>158</td>
<td>410,000</td>
<td>393,774</td>
</tr>
<tr>
<td>Girard, Jean</td>
<td>75,000</td>
<td>-48,117</td>
<td>36</td>
<td>50,000</td>
<td>26,883</td>
</tr>
<tr>
<td>Jone, Suzanne</td>
<td>140,000</td>
<td>-5,204</td>
<td>96</td>
<td>149,000</td>
<td>134,796</td>
</tr>
<tr>
<td>Larson, Tam</td>
<td>350,000</td>
<td>238,388</td>
<td>168</td>
<td>600,000</td>
<td>588,388</td>
</tr>
<tr>
<td>LeShan, George</td>
<td>200,000</td>
<td>-75,126</td>
<td>62</td>
<td>132,000</td>
<td>124,874</td>
</tr>
<tr>
<td>Levensen, Bernard</td>
<td>175,000</td>
<td>-9,267</td>
<td>95</td>
<td>193,000</td>
<td>165,733</td>
</tr>
<tr>
<td>Mulligan, Robert</td>
<td>225,000</td>
<td>34,383</td>
<td>115</td>
<td>259,000</td>
<td>259,383</td>
</tr>
<tr>
<td>Tetracell, Sheila</td>
<td>50,000</td>
<td>-1,263</td>
<td>97</td>
<td>50,000</td>
<td>48,737</td>
</tr>
<tr>
<td>Woytisek, Gillian</td>
<td>190,000</td>
<td>-3,648</td>
<td>98</td>
<td>210,000</td>
<td>186,352</td>
</tr>
</tbody>
</table>

```

NOT A Properly Formatted Table

- Alphabetical sorting makes it hard to compare
- No unit of measure
- Most useful measure furthest from names
- Grid too heavy – not needed
- Derived from Bookings – should not appear before Bookings column
- No total row. Can’t assess overall performance
A Properly Formatted Table

Quarter-to-Date Sales Rep Performance Summary
Quarter 2, 2003 as of March 15, 2003

<table>
<thead>
<tr>
<th>Sales Rep</th>
<th>Actual Bookings</th>
<th>% of Total Bookings</th>
<th>Forecasted Bookings</th>
<th>Quota</th>
<th>Bookings to Quota Variance</th>
<th>Bookings % of Quota</th>
</tr>
</thead>
<tbody>
<tr>
<td>Larson, Terri</td>
<td>588,388</td>
<td>22.1%</td>
<td>600,000</td>
<td>350,000</td>
<td>238,388</td>
<td>168%</td>
</tr>
<tr>
<td>Garibaldi, John</td>
<td>393,774</td>
<td>14.8%</td>
<td>410,000</td>
<td>250,000</td>
<td>143,774</td>
<td>158%</td>
</tr>
<tr>
<td>Caruthers, Michael</td>
<td>274,875</td>
<td>10.3%</td>
<td>324,000</td>
<td>300,000</td>
<td>-25,125</td>
<td>92%</td>
</tr>
<tr>
<td>Mulligan, Robert</td>
<td>259,383</td>
<td>9.7%</td>
<td>275,000</td>
<td>225,000</td>
<td>34,383</td>
<td>115%</td>
</tr>
<tr>
<td>Brown, Sheryll</td>
<td>234,983</td>
<td>8.8%</td>
<td>260,000</td>
<td>150,000</td>
<td>84,983</td>
<td>157%</td>
</tr>
<tr>
<td>Woytisek, Gillian</td>
<td>186,352</td>
<td>7.0%</td>
<td>210,000</td>
<td>190,000</td>
<td>-3,648</td>
<td>98%</td>
</tr>
<tr>
<td>Albright, Gary</td>
<td>183,938</td>
<td>6.9%</td>
<td>205,000</td>
<td>200,000</td>
<td>-16,062</td>
<td>92%</td>
</tr>
<tr>
<td>Levensen, Bernard</td>
<td>165,733</td>
<td>6.2%</td>
<td>193,000</td>
<td>175,000</td>
<td>-9,267</td>
<td>95%</td>
</tr>
<tr>
<td>Jone, Suzanne</td>
<td>134,796</td>
<td>5.1%</td>
<td>149,000</td>
<td>140,000</td>
<td>-5,204</td>
<td>96%</td>
</tr>
<tr>
<td>LeShan, George</td>
<td>124,874</td>
<td>4.7%</td>
<td>132,000</td>
<td>200,000</td>
<td>-75,126</td>
<td>62%</td>
</tr>
<tr>
<td>Tetracell, Sheila</td>
<td>48,737</td>
<td>1.8%</td>
<td>50,000</td>
<td>50,000</td>
<td>-1,263</td>
<td>97%</td>
</tr>
<tr>
<td>Cartwright, Bonnie</td>
<td>43,875</td>
<td>1.6%</td>
<td>50,000</td>
<td>100,000</td>
<td>-56,125</td>
<td>44%</td>
</tr>
<tr>
<td>Girard, Jean</td>
<td>26,883</td>
<td>1.0%</td>
<td>50,000</td>
<td>75,000</td>
<td>-48,117</td>
<td>36%</td>
</tr>
<tr>
<td>Total</td>
<td>$2,666,591</td>
<td>100.0%</td>
<td>$2,908,000</td>
<td>$2,405,000</td>
<td>$261,591</td>
<td>111%</td>
</tr>
</tbody>
</table>


Graph selection and design

<table>
<thead>
<tr>
<th>Type of data</th>
<th>Chart to use</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Time series or continuous data</td>
<td>• Line graph</td>
</tr>
<tr>
<td>• Discrete categories or few time points</td>
<td>• Bars/histograms</td>
</tr>
<tr>
<td>• Relative contribution mutually exclusive categories</td>
<td>• Stacked bar</td>
</tr>
<tr>
<td>• Proportions/Percentages</td>
<td>• Pie Chart</td>
</tr>
<tr>
<td>• Ranges, IQRs, Confidence Intervals</td>
<td>• Box Plot</td>
</tr>
<tr>
<td>• Correlation between 2 continuous variables</td>
<td>• Scatter plot</td>
</tr>
</tbody>
</table>
Avoid 3-D

Smith & Azzam (2010)
Radar graph vs. bar graph

Bars are better than circles for relative size
Making Graphs in Excel – Line Graph

Since this data is a time-series (i.e. illustrates change over time), a line graph is appropriate

Open the data set “USA Real GDP.xlsx”
Making Graphs in Excel – Bar Graph

- Open “Manell_Geus_data_Graphs.xlsx”
- Let’s examine “flood_economy”
- This question asks respondent to rate their opinion on the level of severity of floods as it pertains community economy
- Make a new sheet and name it “charts”
- Set your data up correctly on the “charts” sheet with the COUNTIF function
- Click on the “Insert” Tab and go to the “Charts” section
- Click the column/bar chart icon
Making Graphs in Excel – Bar Graph
Making Graphs in Excel – Bar Graph

Perceptions on Severity Level for Flooding's Effect on Community Economy

- Very low: 10%
- Low: 31%
- Medium: 37%
- High: 15%
- Very High: 4%
- Not Sure: 2%

Perceptions on Severity Level for Flooding's Effect on Community Economy

- Very low: 10%
- Low: 31%
- Medium: 37%
- High: 15%
- Very High: 4%
- Not Sure: 2%
Making Graphs in Excel – Histogram

- Histograms display continuous data broken into bins, and then illustrate the frequency of occurrence for each bin
- Let’s examine “tenure”
- In the “Charts” worksheet, Click on the “Insert” Tab and go to the “Charts” section
- Click the statistical charts icon
Making Graphs in Excel – Histogram

We can examine “flood_economy” again since our data is already set up.

Follows similar procedure to that of a regular bar graph.

Move percentages directly to the right of the labels (very low, low, etc.).

Making Graphs in Excel – Stacked Bar Graph

- We can examine “flood_economy” again since our data is already set up.
- Follows similar procedure to that of a regular bar graph.
- Move percentages directly to the right of the labels (very low, low, etc.).
Making Graphs in Excel – Stacked Bar Graph

1

2

3

Making Graphs in Excel – Stacked Bar Graph
Making Graphs in Excel – Stacked Bar Graph

Perceptions on Severity Level for Flooding’s Effect on Community Economy

- Very High: 2%
- High: 4%
- Medium: 15%
- Low: 37%
- Very low: 31%
- Not Sure: 10%

Making Graphs in Excel – Pie Chart

- Good for showing proportions
- Let’s examine “activity_fish”
- We can analyze what proportion of respondents fish in Cocos Lagoon, in Achang Preserve, in neither, or in both
- First we set the data up with a COUNTIF on our “Charts” sheet

<table>
<thead>
<tr>
<th>Where do Respondents Fish?</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Do not Fish</td>
<td>42%</td>
<td>70</td>
</tr>
<tr>
<td>In Cocos Lagoon</td>
<td>23%</td>
<td>38</td>
</tr>
<tr>
<td>In Achang preserve</td>
<td>13%</td>
<td>22</td>
</tr>
<tr>
<td>In both places</td>
<td>21%</td>
<td>34</td>
</tr>
<tr>
<td>not sure</td>
<td>1%</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>165</td>
</tr>
</tbody>
</table>
Making Graphs in Excel – Pie Chart

- Instructions on how to create a pie chart in Excel.
- A screenshot showing the pie chart creation process.
- A pie chart illustrating the percentage distribution of respondents.

Making Graphs in Excel – Pie Chart

- Description of a pie chart with categories and corresponding percentages.
- Example data used to create the pie chart.
Making Graphs in SPSS – Scatter Plot

- We’ll do Scatter Plots in SPSS, since we are now examining a bivariate relationship
- Open “manell_geus_datafile.sav”

Making Graphs in SPSS – Scatter Plot

- Choose variables: Respondent_ID and Fish harvest
- Create a scatter plot

Example scatter plot:

- X-axis: age
- Y-axis: income
- Points represent individual data points

Example scatter plot:

- X-axis: age
- Y-axis: tenure
- Points represent individual data points

Example scatter plot:

- X-axis: income
- Y-axis: tenure
- Points represent individual data points

Example scatter plot:

- X-axis: income
- Y-axis: age
- Points represent individual data points

Example scatter plot:

- X-axis: tenure
- Y-axis: age
- Points represent individual data points
Making Graphs in SPSS – Scatter Plot with a Fit Line

Save your output as “Manell_Geus_Outout_Scatter.spv”

Exporting Graphs and Tables from Excel

■ Open a blank Word document
■ Select data tables – copy and paste
  - *Insert as table (default)*
  - *Link data*
  - *Picture*
  - *Text*
■ Select figures – copy and paste
  - *Embed data (default)*
  - *Link data*
  - *Picture*
  - *To PDF a figure: select the object, “save as…” to .pdf*
4.1 Which of the following are measures of central tendency?

A. Range  
B. Median  
C. Mode  
D. Variance  
E. Maximum  
F. Mean
4.2 What is a statistical outlier?

A. Another way of saying “average”
B. An observation point that is distant from other observations
C. A non-normal distribution
D. Any point outside the interquartile range

4.3 True or false: A normally distributed data set with a smaller standard deviation will have a narrower, taller bell shape

A. True
B. False
4.4 What is this distribution?

A. Normal
B. Right (positive) skewed
C. Left (negative) skewed
D. Bi-modal

4.5 3-D graphs are a great way to compare data across time points

A. True
B. False
Day 3

- Inferential Statistics

- Stats Questions and Inferential Stats in SPSS
Overview of Inferential Statistics

Day 3: September 14, 2016

Types of statistics

- **Descriptive statistics** = statistics that describe or display data in a meaningful way

- **Inferential statistics** = statistics that draw generalizable conclusions about a population based on a sample of that population
  
  - This is our focus today
Purposes of Inferential Statistics

- We distinguish between summaries of samples (statistics) and summaries of populations (parameters).
- Inferential statistics uses a sample to make inferences about a population with a certain level of confidence.
- Since it is time-consuming and nearly impossible for an average researcher to collect data on an entire population, a sample of the population is studied instead in order to form conclusions about the population of interest.
  - Inferential statistics is the method by which we form these conclusions.

Variables

- To perform inferential statistics, you need:
  - A dependent variable (often the “Y” variable) – a variable whose value depends on that of another.
  - An independent variable (often the “X” variable) - a variable whose variation does not depend on that of another.
  - We use X or (multiple Xs) to try to predict Y.
  - The dependent variable is “dependent” upon the independent variable.
Things to Consider

■ Sample size
  - Larger sample sizes mean more statistical robustness
    ■ More Robust = more resistant to errors
  - However, beyond a certain point, larger samples have diminishing returns on precision

■ Sampling bias
  - Sampling bias is a bias in which a sample is collected in such a way that some members of the intended population are less likely to be included than others

Types of Sampling Bias

■ Non response bias - Results when respondents differ in meaningful ways from non-respondents.
  - Often problem with mail surveys, where the response rate can be very low, and the people that take the time to respond are usually a certain type of person

■ Response bias - results from problems in the measurement process
  - Leading questions - The wording of the question may be loaded in some way to unduly favor one response over another
  - Social desirability - Most people like to present themselves in a favorable light, so they will be reluctant to admit to unsavory attitudes or illegal activities in a survey

■ Undercoverage - Undercoverage occurs when some members of the population are inadequately represented in the sample

■ Voluntary response bias - occurs when sample members are self-selected volunteers
Sampling

- Random sampling - a procedure for sampling from a population in which:
  - The selection of a sample unit is based on chance
  - Every element of the population has a known, non-zero probability of being selected

- All inferential stats should be based on a random sample of data

Sampling Methods

- **Simple random sample**
  - The population consists of N objects.
  - The sample consists of n objects.
  - All possible samples of n objects are equally likely to occur

- **Stratified sampling**
  - Population is divided into groups, based on some characteristic (ex. county/municipality). Then, within each group, a probability sample (often a simple random sample) is selected
  - In stratified sampling, the groups are called strata
Sampling Methods

■ Cluster sampling
  - Every member of the population is assigned to one, and only one, group
  - Each group is called a cluster
  - A sample of clusters is chosen, using a probability method (often simple random sampling)
  - Only individuals within sampled clusters are surveyed

■ Multistage sampling
  - select a sample by using combinations of different sampling methods
  - For example, in Stage 1, we might use cluster sampling to choose clusters from a population. Then, in Stage 2, we might use simple random sampling to select a subset of elements from each chosen cluster for the final sample.

Probability

■ The chances of something happening
■ A number between 0 (never happens) and 1 (always happens)
■ When something happens half the time, it has a probability of 0.5 or 50%
■ Probability in Inferential stats
  - Often called the p-value
Distribution

- Any sample of a population will have a distribution of values with variability and a central tendency.
- Variance/standard deviation or range are descriptions of the spread of the distribution.
- Central tendency is usually described by the sample mean or median.

Confidence Level

- Confidence level - how often you expect to get similar results.
- If you have a 95% confidence level, it means that if you conducted the same survey 100 times, 95 times the confidence interval will include the true value.
- At 95% confidence level, we allow 5% mistakes due to chance.
Confidence Interval

- We use a confidence interval to express the degree of uncertainty associated with a sample statistic
  - Since we are using a sample instead of the population, there is inherently going to be some margin of error involved

- A confidence interval is an interval estimate combined with a probability statement
  - Most common = 95% confidence interval
    - Also common: 90% (smaller interval) and 99% (larger interval)
  - “We are 95% sure that the population parameter lies within the values of X and Y”

- Confidence intervals are preferred to individual numbers because they indicate:
  - The precision of the estimate
  - The uncertainty of the estimate

- What it means is that you have a high confidence that the true value is within a certain range

Confidence Interval

- Formula
  \[
  X \pm t \frac{s}{\sqrt{n}}
  \]

- Where \(X\) = sample mean
  - \(t\) = t statistic from t table based on sample size and alpha level
  - \(s\) = standard deviation of sample
  - \(n\) = sample size

- In the end, you “are 95% (or 90%, 99%) confident that the true population mean lies within the interval of \((X_{i1} \text{ and } X_{i2})\)” ; Where \(X_{i2} > X_{i1}\)

- In the lesson later today, we will calculate confidence intervals with SPSS
Hypothesis Testing

- Hypothesis testing is an inferential procedure that uses sample data to evaluate the credibility of a hypothesis about a population

- The logic:
  - **State the Hypothesis**: We state a hypothesis (guess) about a population. Usually the hypothesis concerns the value of a population parameter
  - **Define the Decision Method**: We define a method to make a decision about the hypothesis. The method involves sample data
  - **Gather Data**: We obtain a random sample from the population
  - **Make a Decision**: We compare the sample data with the hypothesis about the population. Usually we compare the value of a statistic computed from the sample data with the hypothesized value of the population parameter

Hypothesis Testing

- When you want to know whether:
  - A value is different from expected
  - There are differences between groups
  - There is a relationship between two variables
Hypothesis Testing

- Null hypothesis - usually refers to a general statement or default position that there is no relationship between two measured phenomena, or no association among groups (i.e. any difference is due to randomness)
  - Denoted by \( H_0 \)
  - Example: “There is no difference between the perceptions of coral reef condition amongst college educated and non college educated respondents”

- Alternative hypothesis – The opposite of the null hypothesis
  - Denoted by \( H_A \)
  - The hypothesis that sample observations are influenced by some non-random cause
  - Example: “College educated respondents have a more negative perception concerning coral reef condition when compared to non-college educated respondents”

P-values

- Statistically significant = The patterns we observe are unlikely to happen by chance
- The significance level (\( \alpha \); alpha) is the threshold for statistical significance
- Alpha levels are used in hypothesis tests
  - Usually, these tests are run with an alpha level of .05 (5%), but other levels commonly used are .01 and .10.
  - We “reject” the null hypothesis if our resulting p-value is less than our chosen alpha value
  - We “fail to reject” the null hypothesis if our resulting p-value is greater than our chosen alpha value
Statistical Significance

- Hypothesis tests form the basis for statistical conclusions
  - Example conclusion:
    - “We are 95% confident that college educated respondents have a more negative perception concerning coral reef condition when compared to non-college educated respondents”
    - This conclusion was reached because the p-value of the hypothesis test was less than the alpha level threshold of 0.05 (or 5%)
  - We will learn how to perform hypothesis tests and interpret results in a lesson later today

Statistical Tests

- Certain types of data call for different statistical tests
Bivariate Analysis

- Bivariate analysis involves the analysis of two variables (often denoted as $X$, $Y$), for the purpose of determining the relationship between them
  - Null hypothesis = no relationship
  - Alternative hypothesis = there is some relationship
  - The purpose is to understand the relationship between the two variables:
    - How does variable $X$ relate to variable $Y$?
    - How strong is that relationship?
- Bivariate analysis can be helpful in testing simple hypotheses of association
- Bivariate analysis can help determine to what extent it becomes easier to know and predict a value for one variable (dependent variable) if we know the value of the other variable (independent variable)

Contingency Tables

- Contingency tables are mostly used with categorical data
- Called “cross-tabulation” because it crosses and tabulates each of the categories of one variable with each of the categories of a second variable
- A “statistical relationship” between two variables indicates a recognizable pattern of changes in one variable as the other changes
- Dependent variable goes in the rows of the table, and independent variable goes in the columns
Contingency Tables

- **Steps:**
  - *First, we test determine if a relationship exists between the variables*
    - This shows whether or not the observed relationship is significant
  - *Next, we describe the relationship between the variables using “measures of association”*
    - This shows the strength of the association, and (for ordinal variables) the direction of the relationship

**Basic 2x2 Contingency Table**

<table>
<thead>
<tr>
<th></th>
<th>IV Attribute Lo</th>
<th>IV Attribute Hi</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>DV attribute Lo</td>
<td>A</td>
<td>B</td>
<td>A + B</td>
</tr>
<tr>
<td>DV attribute Hi</td>
<td>C</td>
<td>D</td>
<td>C + D</td>
</tr>
<tr>
<td>Total</td>
<td>A + C</td>
<td>B + D</td>
<td>N = A+B+C+D</td>
</tr>
</tbody>
</table>

- The totals around the perimeter = “marginals”
- The cross-classifications = “cells”
- N = “grand total”
Creating Contingency Tables

- The dependent variable goes in the rows
  - Variable attributes should be ordered lowest to highest (least to most, etc.) from top to bottom
- The independent variable goes in the columns
  - Variable attributes should be ordered lowest to highest (least to most, etc.) from left to right
- Start by filling in cell frequencies
- Convert the frequencies to column percentages
  - Divide the cell frequency by the total for the column and multiply by 100

Chi Square

- The Chi-Square test is used to determine if a statistical relationship exists between two categorical variables
  - This is the test to use for contingency tables/cross tabulations
- The null hypothesis of every chi-square test is that “no statistically significant relationship exists” between the 2 variables
Chi Square

- Chi-square determines the extent to which the observed distribution differs from what it is expected to be if no relationship exists
  - If there is no relationship, the “expected” frequencies for each column would be the same, and the column percentages would be the same as the aggregate totals
  - The greater the difference between the observed and expected frequencies, the greater the relationship between the variables

Chi-square

- Three assumptions:
  - The variables must be categorical
  - The observations must be independent
  - All cells must have at least 5 expected observations
    - If not, we can use the Fisher’s Exact Test to test for a statistical relationship
- Chi-square is always a positive number (it doesn’t have “direction”)
  - If there is no relationship, chi-square = 0
  - To determine direction, “measures of association” are used
- Fisher’s Exact Test
  - The equivalent of a chi-square test, but for smaller sample sizes in a 2x2 table (i.e. some cells in the contingency table have less than 5 expected observations)
Measures of Association

- Once you’ve determined that a relationship exists by using the Chi-square test, you want to further describe the nature of that relationship
  - How strong is the association between the variables (nominal and ordinal)?
  - What direction (positive or negative) is the association? (for ordinal variables) (for ordinal)

- Measures of association are used to answer the above 2 questions

- Characteristics of measures of association:
  - If the relationship between the variables is perfect, the measure equals 1 or -1
  - If there is no relationship between the variables, the measure equals 0
  - For ordered variable measures, the sign of the measure indicates the direction of the relationship

Measures of Association - Nominal

- Cramer’s V used if your DV and IV are both nominal
  - Ranges from 0 (no relationship) to 1
  - Cannot be negative - the relationships between nominal variables do not have direction

<table>
<thead>
<tr>
<th>Table 10.2</th>
<th>How to Interpret Measures of Association</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measure of Association ($X$)</td>
<td>Qualitative Interpretation</td>
</tr>
<tr>
<td>$0 \leq X &lt; 0.10$</td>
<td>Very Weak</td>
</tr>
<tr>
<td>$0.10 \leq X &lt; 0.20$</td>
<td>Weak</td>
</tr>
<tr>
<td>$0.20 \leq X &lt; 0.30$</td>
<td>Moderate</td>
</tr>
<tr>
<td>$X \geq 0.30$</td>
<td>Strong</td>
</tr>
</tbody>
</table>
Measures of Association - Ordinal

- Kendall’s tau-b is for “square” tables, with equal numbers of rows & columns
- Kendall’s (Stuart’s) tau-c is for “rectangular” tables, with different numbers of rows & columns
- Both of these can be negative depending on the direction of the relationship

T-test

- Are two means significantly different?
- When to use:
  - When you have continuous data that is normally distributed
  - Some other special cases (we will discuss)
- Types of T-tests
  - One sample
  - Two sample - paired
  - Two sample – independent
- In two sample t-tests, the difference of the means of the separate groups are calculated and confidence intervals are created around the difference
  - If the confidence interval does not contain zero, then there is a statistically significant difference
One Sample t-test

- Used to examine the mean difference between the sample and the known value of the population mean
- In one sample t-test, we know the population mean (or we assume it to be some value)
- The basic idea of the test is a comparison of the average of the sample (observed average) and the population (expected average)
  - Compare the sample mean with the population mean and make a statistical decision as to whether or not the sample mean is different from the population mean
- Null hypothesis: assumes that there are no significance differences between the population mean and the sample mean
- Alternative hypothesis: assumes that there is a significant difference between the population mean and the sample mean

Two Sample t-test: Paired Samples

- Paired sample t-tests are used in ‘before-after’ studies, or when the samples are the matched pairs, or when it is a case-control study
- For example, we give weight loss treatment to a group of voluntary participants
  - The paired t-test can be used to test for a statistically significant reduction in weight from before the treatment to after the treatment
    - Null hypothesis: there is no difference in weight from before to after the treatment
    - Alternative hypothesis: there is a significant decrease in weight after the treatment
- In paired t-tests, each “group” (i.e. the before and after) is dependent upon each other
  - Measurements are taken from the same group of respondents “before” and “after”
Two sample t-test – Independent Samples

- Helps you compare whether two groups have statistically significant different mean values
  - For example, whether men and women have different mean heights
- Null hypothesis: there is no significant difference between the groups
- Alternative hypothesis: There is a significant difference between the groups
- There are 2 ways this test can be performed:
  - Assuming Equal Variances
  - Assuming Unequal Variances
  - The F-test is used to determine if variances are equal or not
    - Null hypothesis if F-test = variances are assumed equal

Interpreting Results from a t-test

- P(T<=t) = the probability of observing an equal or greater t-statistic by chance
- To be considered significant, P < Alpha
- t critical: the cutoff value of t where P = Alpha
- Two-tailed: use when you don’t know which will be greater before you run the test
  - Null: means are equal
  - Alternative: means are not equal
- One-tailed: use when you expect one group mean will be greater or less than the other
  - Null: means are equal
  - Alternative: Mean #1 > Mean #2 (or vice versa)
Correlations

- Correlation is a statistical technique that is used to measure and describe the strength and direction of the linear relationship between two variables
  - Correlation coefficient ranges from -1 to 1
  - Zero = no linear relationship
  - Closer to zero = weaker linear relationship
  - Closer to 1 = strong positive linear relationship
  - Closer to -1 = strong negative linear relationship
- Can only be performed with 2 continuous or binary variables
- A positive correlation means that if one variable gets bigger, the other variable tends to get bigger
- A negative correlation means that if one variable gets bigger, the other variable tends to get smaller
- Correlation DOES NOT mean Causation

Multivariate Analysis

- Statistical procedure for analysis of data involving more than one type of measurement or observation
- It may also mean solving problems where more than one independent variable is analyzed simultaneously with other variables
- Basically, we use multivariate analysis when we are interested in looking at more than 2 variables at once
One way ANOVA

- Essentially an Independent t-test with >2 groups
- Tests for statistically significant differences in the means of 2 or more groups
  - For example, when testing for differences in median household income by region (north, south, east, and west)
    - Null hypothesis: there is no significant difference between the groups
    - Alternative hypothesis: There is a significant difference between the groups
- In SPSS, the “dependent list” corresponds to the variables you want to take the mean of (income), and the “factor” represents the grouping variable (i.e. the region)
- One-way ANOVA is an omnibus F-test statistic and cannot tell you which specific groups were significantly different from each other (i.e. p-value<0.05), only that at least two groups were
  - To determine which specific groups differed from each other, you need to use a post hoc test

One way ANOVA Post-Hoc Test

- The most common post-hoc test is the Tukey’s HSD test
- The post-hoc test will tell you which groups’ means are significantly different
- The difference of the means of the separate groups are calculated and confidence intervals are created around the difference
  - If the confidence interval does not contain zero, then there is a statistically significant difference
Simple Linear Regression

- One dependent variable (Y) and one independent variable (X)
- At the center of the regression analysis is the task of fitting a single line through a scatter plot.
- A technique used to determine the linear relationship between two variables, and in turn, attempt to predict changes in Y due to changes in X.
  - Defined by the formula \( Y = c + b \times X \)
  - \( c \) = constant (intercept)
  - \( b \) = regression coefficient (slope)
  - \( X \) = independent variable
- As \( X \) changes by (1), we expect \( Y \) to change by (b)
- When \( X=0 \), \( Y = c \)
- \( b \) is tested for statistical significance (i.e. does \( X \) have a significant effect on \( Y \)?)
  - Using p-values; is p-value less than alpha?

Multiple Linear Regression

- Same as simple linear regression, but with multiple independent (X) variables
  - \( X_{11}, X_{12}, X_{13}, X_{1n}, \) etc.
- Incorporates multiple “predictor” variables to “predict” the value of \( Y \)
- Strength of a regression model given by \( R^2 \)
  - \( R^2 \) ranges from 0-1, with stronger (better predictive) models having an \( R^2 \) value closer to 1
  - Interpretation: if \( R^2 = 0.50 \), then “50% of the variation in \( Y \) is explained by \( X_{11}, X_{12}, \text{ and } X_{13} \)”
Multiple Linear Regression

- Model given by:
  - \( Y = c + b_1 + b_2 + ... + b_k \)
- Each coefficient \( b \) is tested for statistical significance using p-values
- Interpretation:
  - Each predictor \( X \) variable is interpreted on its own (i.e. “all else held equal”)
  - Example: if \( b \) is significant, that “all else held equal, \( b \) significantly effects \( Y \)”
  - If \( b \) changes by 1, we expect \( Y \) to change by
  - If changes by 1, we expect \( Y \) to change by
  - If \( b_1 = b_2 = b_3 = b_4 = b_5 = 0 \), then we expect \( Y = c \)
What is a Variable Transformation?

- The replacement of a variable by a function of that variable
- In some cases, variable transformations (change in coding) are necessary to perform certain types of analysis
  - Need continuous variables for correlations and regressions
    - Categorical (qualitative) data can be transformed into continuous (quantitative) data through the use of variable transformations
- When a variable is transformed, you keep the “old” version of the variable in your data set, AND the “new” version as well
  - You ALWAYS document what you did (keep track of work flow)
  - You ALWAYS add newly transformed variables to your codebook
Preparation for Transformation

- Open “Manell_Geus_codebook.xlsx”
- In Row 419 of your codebook, type “Variable Transformation Stipulations”
  - Highlight the cell
  - Below this line, we will type in any decisions made concerning variable transformations and what the corresponding coding looks like

How to Treat “Not Sure”

- Since a response of “not sure” is not quantifiable on a scale of perception (condition questions), agreement (management option questions), use frequency (activity participation questions), etc.:
  - They must be coded as “missing” (.) for most analysis
  - Since in many cases, a “not sure” is coded as “8,” this can mess up calculations if not addressed
- Open “Manell_Geus_datafile_trans.sav”
- Let’s remove “not sures” from “condition_ocean”
Treating “Not Sure” as Missing

[Diagram of data analysis software interface]

1. Recode into Different Variables
2. Select conditions for reclassification
3. Apply changes and review results

Treating “Not Sure” as Missing
Treating “Not Sure” as Missing

- Now, the “not sure” responses will NOT be included in analysis calculations and our responses are all on the correct scale without the interference of “not sure” responses
  - “Condition” questions are only measured from negative perception to positive perception
  - “Agreement” questions are only measured from less agreement to more agreement
  - “Success” questions are only measured from low success to high success

- Add to your Variable Transformation Stipulations
- *NOTE: You can transform multiple variables at once using this method

419 VARIABLE TRANSFORMATION STIPULATIONS
420 Any variable name with "_NS" after it represents the variable coded with "not sure" responses as MISSING
The Dummy Variable

- A dummy variable is a variable that is binary (only 2 possible responses)
- Coded as zero (0) or one (1)
- Usually:
  - 1 = Exhibits a certain attribute
  - 0 = Does NOT exhibit a certain attribute

- For example, to analyze “consume_fish” (How often does your family eat fish/seafood?), we can change the coding from:
  - Almost never/never = 1
  - A few times a year = 2
  - Once a month = 3
  - A few times a month = 4
  - At least once a week = 5
  - Almost daily = 6
  - Not Sure = 8
  - Almost never/never = 0
  - A few times a year = 0
  - Once a month = 0
  - A few times a month = 0
  - At least once a week = 1
  - Almost daily = 1
  - Not Sure = missing (.)

- This transformation allows us to analyze this ordinal variable as a continuous variable
- We can now calculate the percentage of people that consume seafood AT LEAST once per week and we can try to see how this variable relates to other variables in the data through quantitative analysis
- We can name it “consume_once_week” and enter this into our “Variable Transformation Stipulations” part of our codebook
Dummy Transformation in SPSS

- Now, the “not sure” responses will NOT be included in analysis calculations and our responses are all on the correct scale without the interference of “not sure” responses
- Additionally, “consume_once_week” can be analyzed as a continuous variable since it is now binary on a 0-1 scale
- Add to your Variable Transformation Stipulations
- *NOTE: You can transform multiple variables at once using this method

<table>
<thead>
<tr>
<th>VARIABLE TRANSFORMATION STIPULATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>consume_once_week</td>
</tr>
</tbody>
</table>

The Index Variable

- Combines the values of other variables into a single indicator or score (usually by adding or taking an average)
- The variables that comprise the indicator can be weighted differently (i.e. some variables are more important than other), or they can all be equally weighted
- This method turns a series of categorical (qualitative) variables into a single continuous (quantitative) variable
The Index Variable: Additive Index

- Usually this type of variable transformation is suited for questions that are related in some way
  - Example: The “last 10 years” questions in the Manell-Geus survey
    - We can use an additive index technique to create a variable that represents each respondent’s overall perception concerning the change in condition of marine resources over the last 10 years
    - This technique takes a series of ordinal variables and turns them into a single continuous variable

The Index Variable: Additive Index

- Last 10 years index
  - Since each of the individual questions that comprise this index are coded ordinally from 1-5 (from more negative opinion to more positive opinion), this index variable will have “direction”

  - Direction = as the index increases, positive perception concerning the change in the condition of marine resources over the last 10 years increases
The Index Variable: Additive Index

- Rules for the “Last 10 Years” Index:
  - Respondents have to answer EVERY QUESTION that is in the index to be assigned an index value
  - If a respondent left one or more questions in the index unanswered, they will be assigned a missing value for the “Last 10 Years” Index
  - If a respondent answered “not sure” to one or more questions in the index, they will be assigned a missing value for the “Last 10 Years” Index
    - A response of “Not sure” does not indicate direction (i.e. It isn’t a negative or a positive perception concerning marine resource condition)
    - Therefore “not sure” responses will be coded as missing for the purposes of creating the index

The Index Variable: Last 10 Years Index

- Working out of “Manell_Geus_Datafile_trans.sav”
- We are focused on the “Last10_NS” variables
The Index Variable: Last 10 Years Index

- This variable is now continuous and can be used in correlations, regressions, etc.
- Add to your Variable Transformation Stipulations

- Let’s run some summary statistics on this new index variable and see what we find…..
- The range is 9-45
  - Make sense since the index is made of 9 questions that range individually from 1-5
  - Minimum: 9*1 = 9; Maximum 9*5 = 45
- Mean = 27.7
- Midpoint of index: 9 + (45-9)/2 = 27
- Perception concerning the change in marine resource condition is about neutral
The Normalized Variable

- In some cases, we may want to scale variables in a simpler way
  - *i.e. instead of the “Last 10 Years” Index ranging from 9-45, we want it to range from 0-100*
    - This can make interpretation easier as some may misinterpret a value of 9 to be greater than some other value when it is, in fact, the minimum
- The variable can be normalized with the “min-max” scaling method
  - $X_{\text{norm}} = \frac{X - \text{min}\downarrow x}{\text{max}\downarrow x - \text{min}\downarrow x}$
  - *This equation transforms any variable to a 0-1 scale*
    - When multiplied by 100, we can go to a 0-100 scale

The Normalized Variable – Last 10 Years Index

- Another variable transformation in SPSS
The Normalized Variable – Last 10 Years Index

Practice!

- Let’s practice some variable transformations
- Code “not sures” as missing for all “condition_” questions
- Create a Condition index and normalize it
- Run some Descriptive Stats to make sure means are correct
Descriptive Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>condition_ocean_NS</td>
<td>299</td>
<td>1</td>
<td>5</td>
<td>3.08</td>
<td>1.097</td>
</tr>
<tr>
<td>condition_total_NS</td>
<td>289</td>
<td>1</td>
<td>5</td>
<td>3.21</td>
<td>1.093</td>
</tr>
<tr>
<td>condition_sizefish_NS</td>
<td>293</td>
<td>1</td>
<td>5</td>
<td>3.26</td>
<td>1.174</td>
</tr>
<tr>
<td>condition_numfish_NS</td>
<td>285</td>
<td>1</td>
<td>5</td>
<td>3.36</td>
<td>1.126</td>
</tr>
<tr>
<td>condition_numfish_underwater_NS</td>
<td>276</td>
<td>1</td>
<td>5</td>
<td>3.34</td>
<td>1.095</td>
</tr>
<tr>
<td>condition_numturt_NS</td>
<td>276</td>
<td>1</td>
<td>5</td>
<td>3.30</td>
<td>1.166</td>
</tr>
<tr>
<td>condition_beach_NS</td>
<td>298</td>
<td>1</td>
<td>5</td>
<td>2.90</td>
<td>1.107</td>
</tr>
<tr>
<td>condition_stream_NS</td>
<td>284</td>
<td>1</td>
<td>5</td>
<td>2.90</td>
<td>1.076</td>
</tr>
<tr>
<td>condition_forest_NS</td>
<td>282</td>
<td>1</td>
<td>5</td>
<td>3.05</td>
<td>1.070</td>
</tr>
<tr>
<td>Condition_index_norm</td>
<td>246</td>
<td>0</td>
<td>100</td>
<td>53.97</td>
<td>22.026</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>246</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

New Files

- Our new full data set that includes all newly transformed variables in addition to all initial variables is the file “Manell_Geus_Data_Transformed.xlsx”
- Our new full codebook that includes all newly transformed variables in addition to all initial variables is the file “Manell_Geus_Codebook_Transformed.xlsx”
- Our new full SPSS data set that includes all newly transformed variables in addition to all initial variables is the file “Manell_Geus_transformed_datafile.sav”
- We will mostly be relying on these files for the rest of the workshop
5.1 What is the definition of a variable transformation?

A. The replacement of a variable by a function of that variable
B. Changing a categorical variable to a continuous variable
C. When you leave responses of “not sure” out of analysis
D. When data is filtered in Excel
5.2 How does non-response bias happen?

A. From problems in the measurement process
B. Data entry errors
C. Some members of the population are inadequately represented in the sample
D. When respondents differ in meaningful ways from non-respondents

5.3 When is it ok to accept the alternative hypothesis in a hypothesis test?

A. When you reject the null hypothesis
B. When you fail to reject the null hypothesis
C. You never “accept” a hypothesis in a hypothesis test
D. If you have a significant p-value
5.4 What statistical test should be used to determine if a statistical relationship exists between two categorical variables?

A. Two sample (paired) t-test  
B. Chi square test  
C. Correlation  
D. Regression

5.5 What is the formula for normalizing a variable on a 0-1 scale?

A. \( \text{X_{norm}} = \frac{X - \text{max}_{x}}{\text{max}_{x} - \text{min}_{x}} \)  
B. \( \text{X_{norm}} = \frac{X - \text{min}_{x}}{\text{max}_{x} - \text{min}_{x}} \)  
C. \( \text{X_{norm}} = \frac{\text{min}_{x} - X}{\text{max}_{x} - \text{min}_{x}} \)
Proposing Questions and Hypotheses

Day 3: September 14, 2016

Discussion

- When examining the Manell-Geus questionnaire, what possible analyses come to mind?
  - Look at the questionnaire pdf, the codebook and the data itself
    - What questions would you like to answer with this data set?
    - How will you answer the question?
      - With what statistical test?
    - What do you expect the results to look like? (i.e. what is your “hypothesis”?)
    - How will you communicate the results?
      - With a table? A specific type of graph?
What Questions Can We Answer?

Some questions from initial participant feedback:

1. Are people that rate floods and fires as low severity less inclined to participate in reef protection activities?
2. Are people that support the Achang Marine Preserve more likely to participate in reef protection activities?
3. What types of people are unsure or “neither agree or disagree” about “every resident should be responsible to help take care of reefs”
4. How to fill knowledge gap (maybe where they get their info from, could introduce easy-to-understand infographics for outreach purposes)
Best Way to Answer Each Question?

Hypothesis for Each Question
Best Way to Communicate Each Result?
Exploratory Analysis

- Exploratory Data Analysis is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to
  - maximize insight into a data set
  - uncover underlying structure
  - extract important variables
  - detect outliers and anomalies
  - test underlying assumptions
- “Playing with” the data
- Seeing what the distribution of a variable(s) looks like
- Seeing what variables look like in relation to one another
Scatter Plots

- Open “Manell_Geus_transformed_datafile.sav”
- A quick way to graphically examine 2 continuous variables together
- Let’s look at “percent_merizo” and “tenure”
- Graphs > Chart builder > scatter/dot > then drag variables
- Click OK > Double click chart in output window > add fit line

Scatter Plots – Percent_Merizo and Tenure

- Judging by this graph, there doesn’t appear to be a strong relationship between the amount of years people have lived in Merizo and the percentage of seafood that they eat that comes from Merizo
Dual Bar Graphs

- A quick way to graphically display frequencies or means for different groups
- Let’s examine “fishing index_norm” and “college”
  - To see if there is a difference between the average fishing index for those who completed college and those who did not complete college
- Graphs > Chart builder > bar > then drag variables

- We can adjust the statistic to what we want to compare
- Here, we are examining means
- But, we can also look at frequencies, median, min, max, etc.
Dual Bar Graphs

• Judging by this graph, there appears to be a difference between in fishing frequency between college educated and non-college educated respondents
• Those that did not complete college tend to fish more
• *NOTE: we are NOT saying they are statistically different, just that they are different

Contingency Tables

■ A quick way to graphically examine 2 categorical variables together
  - A “cross-tabulation” of the variables
■ Divides variables into categories and displays frequencies for each category
■ Let’s examine “consume_stream_anyfreq” and “boat”
### Contingency Tables

- “consume_stream_anyfreq” and “boat”
- Analyzing these two variables will tell us if there is some sort of relationship between owning a boat and consuming fish/seafood from a stream
- Dependent variable = “consume_stream_anyfreq”
  - We want to know if consuming fish/seafood from the stream is “dependent” upon owning a boat
  - “boat” is our “predictor”, there it is the independent variable
- In contingency tables, the DV is always in the rows and the IV is always in the columns

### Basic 2x2 Contingency Table

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Independent Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IV Attribute Lo</td>
</tr>
<tr>
<td>DV attribute Lo</td>
<td>A</td>
</tr>
<tr>
<td>DV attribute Hi</td>
<td>C</td>
</tr>
<tr>
<td>Total</td>
<td>A + C</td>
</tr>
</tbody>
</table>
We want “column percentages” because we are concerned with how the proportion of those who consume fish/seafood from the stream varies by boat ownership.

Displaying the column percentages breaks the “boat” variable into its 2 groups (own a boat or don’t own a boat).

Contingency Tables

- 56.4% of boat owners consume fish/seafood from the stream.
- 53.3% of those who do not own a boat consume fish/seafood from the stream.
- We see a difference here in our contingency table.
- *NOTE: we are NOT saying they are statistically different, just that they are different.
Inferential Statistics

- After doing some exploratory analysis to “get to know” your data better, we can now move on to inferential stats.
- With these exercises, we will be able to form statistical conclusions about our data:
  - Performing hypothesis tests
  - Checking for significant differences
  - Communicating results based on correct methods

Generating Confidence Intervals

- We want to generate a 95% confidence interval for the mean of “fish_harvest”.
- Our test value is zero in this case because we are not trying to see if the mean of “fish_harvest” is different from some value, we only want the confidence interval around the mean.
Generating Confidence Intervals

- Our 95% confidence interval for the mean of “fish_harvest” is (0.45, 0.56)
- The sample mean is 0.51
  - 51% of our sample fishes or harvests for marine resources
- Conclusion:
  - We are 95% confident that the true population mean for the percentage of people that fish and harvest for marine resources in Merizo is between 45% and 56%

Let’s examine the confidence interval for some of the “threat_” questions
Generating Confidence Intervals

5% of our sample believes that coral bleaching is a top 3 threat to coral reefs
- We are 95% confident that the true population mean for the percentage of people who believe coral bleaching is a top 3 threat to coral reefs is between 2% and 7%.

10% of our sample believes that ocean acidification is a top 3 threat to coral reefs
- We are 95% confident that the true population mean for the percentage of people who believe ocean acidification is a top 3 threat to coral reefs is between 6% and 13%.

8% of our sample believes that lack of mountain vegetation is a top 3 threat to coral reefs
- We are 95% confident that the true population mean for the percentage of people who believe lack of mountain vegetation is a top 3 threat to coral reefs is between 5% and 11%.

Testing Hypothesis

- Building off of confidence intervals, we can also test hypothesis in SPSS.
- Here we will specify a “test value” different from zero to make a conclusion about our sample mean in relation to a hypothesis.
- Let’s examine “fish_harvest” again.
- Imagine you have read in past literature that 50% of Merizo residents fish or harvest for marine resources.
  - 50% is our “test value”
Testing Hypothesis

- Null Hypothesis: The percentage of people that fish or harvest for marine resources in Merizo is **NOT** significantly different from 50%
- Alternative Hypothesis: The percentage of people that fish or harvest for marine resources in Merizo is significantly different from 50%

• Our p-value is under “Sig. (2-tailed)”
• We reject the null hypothesis if our p-value is <0.05
• Since our p-value = 0.774 > 0.05,
  • “we fail to reject the null hypothesis at the 95% confidence level”
  • The population mean is not statistically different from 50%
  • Our results are consistent with (hypothetical) past findings
• Notice that the 95% confidence interval contains zero, meaning that difference between the sample mean (51%) and the test value (50%) is **NOT STATISTICALLY SIGNIFICANT**
Testing Hypothesis

- Let’s now examine “percent_merizo”
- Imagine you have read in past literature that for an average Merizo resident, 40% of the seafood that they consume comes from Merizo — 40% is our “test value”
- Null hypothesis: The percentage of seafood that comes from Merizo that is consumed by an average Merizo resident is NOT different from 40%
- Alternative hypothesis: The percentage of seafood that comes from Merizo that is consumed by an average Merizo resident is different from 40%

Since our p-value = 0.000 < 0.05,
- “we reject the null hypothesis at the 95% confidence level”
- We are 95% confident that the population mean is statistically different from 40%
- Our results are NOT consistent with (hypothetical) past findings
- Notice that the 95% confidence interval DOES NOT contains zero, meaning that difference between the sample mean (17.51%) and the test value (40%) is STATISTICALLY SIGNIFICANT
Practice!

- What is the confidence interval for the percentage of people that use the newspaper of a coral reef information source?
  - What is our conclusion?

  - Confidence interval
    - (0.45, 0.56)
    - Sample mean = 50%
  
  - We are 95% confident that the true population mean for the percentage of people who use newspaper as a top 5 choice of information source about coral reefs is between 45% and 56%
Practice!

- What is the confidence interval for the percentage of people that have lived in Merizo for 15 or more years?
  - What is our conclusion?

Practice!

- Confidence interval
  - (.56 , .67)
  - Sample mean = 62%
- We are 95% confident that the true population mean for the percentage of people who have lived in Merizo for 15 or more years is between 45% and 56%
Is the average age of Merizo’s population statistically different from 35?
- What are your null and alternative hypotheses?
- What is your p-value?
- What is your conclusion?

Null hypothesis: The average age of Merizo’s population is NOT different from 35 years old
Alternative hypothesis: The average age of Merizo’s population is different from 35 years old
P-value = 0.000 < 0.05

- Conclusion
  - We reject the null hypothesis at the 95% confidence level
  - We are 95% confident that the population mean age is greater than 35 years old
  - We can say “greater” because our 95% CI for the difference is all positive
Practice!

- Is the average amount of times that a household has been impacted by a flood statistically different from 5?
  - What are your null and alternative hypotheses?
  - What is your p-value?
  - What is your conclusion?

Practice!

- Null hypothesis: The average amount of times an average household in Merizo has been impacted by a flood is NOT different from 5 instances
- Alternative hypothesis: The average amount of times an average household in Merizo has been impacted by a flood is different from 5 instances
- P-value = 0.000 < 0.05

- Conclusion
  - We reject the null hypothesis at the 95% confidence level
  - We are 95% confident that the population mean number of times impacted by a flood is **less** than 5 instances
  - We can say “less” because our 95% CI for the **difference** is all negative
Practice!

- Is the proportion of Merizo residents that are familiar with Achang Preserve statistically different from 55%?
  - What are your null and alternative hypotheses?
  - What is your p-value?
  - What is your conclusion?

Practice!

- Null hypothesis: The proportion of Merizo residents that are familiar with Achang Preserve is NOT different from 55%
- Alternative hypothesis: The proportion of Merizo residents that are familiar with Achang Preserve is different from 55%
- P-value = 0.796 > 0.05
- Conclusion
  - We fail reject the null hypothesis at the 95% confidence level
  - The population proportion of Merizo residents that are familiar with Achang Preserve is not statistically different from 55%
  - Our 95% CI for the difference contains zero

*Save your output as “Manell_Geus_Output_inferential.spv”*
6.1 True or false: A dummy variable **can not** be analyzed as continuous data

A. True
B. False
6.2 What does a 95% confidence interval tell us?

A. “We are 95% sure that the population parameter lies within the values of X and Y”
B. “We reject the null hypothesis”
C. “There is 95% sampling bias in my data”
D. “My data is 95% normal”

6.3 What is a scatter plot useful for?

A. Calculating means
B. Visually displaying means
C. As an initial step before using bar graphs
D. Exploring the relationship between 2 variables
6.4 Why do we do some data analysis with “not sure” coded as missing?

A. Because we don’t care if respondents answered “not sure”
B. A response of “not sure” is not quantifiable on an ordinal or continuous scale
C. To decrease our sample size
D. To increase our sample size

6.5 True or false: A large p-value means we reject the null hypothesis

A. True
B. False
Day 4

- Chi-square, T-Test and ANOVA

- Correlation and Regression
Contingency Tables, Chi-Square, and Measures of Association

Day 4: September 15, 2016

Recap

- Contingency tables are mostly used with categorical data
- Called “cross-tabulation” because it crosses and tabulates each of the categories of one variable with each of the categories of a second variable
- A “statistical relationship” between two variables indicates a recognizable pattern of changes in one variable as the other changes
- Dependent variable goes in the rows of the table, and independent variable goes in the columns
Recap

- The dependent variable goes in the rows
  - Variable attributes should be ordered lowest to highest (least to most, etc.) from top to bottom
- The independent variable goes in the columns
  - Variable attributes should be ordered lowest to highest (least to most, etc.) from left to right
- Start by filling in cell frequencies
- Convert the frequencies to column percentages
  - Divide the cell frequency by the total for the column and multiply by 100
- Typically display the final table in percentage form

Basic 2x2 Contingency Table

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>IV Attribute Lo</th>
<th>IV Attribute Hi</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>DV attribute Lo</td>
<td>A</td>
<td>B</td>
<td>A+B</td>
</tr>
<tr>
<td>DV attribute Hi</td>
<td>C</td>
<td>D</td>
<td>C+D</td>
</tr>
<tr>
<td>Total</td>
<td>A+C</td>
<td>B+D</td>
<td>N=A+B+C+D</td>
</tr>
</tbody>
</table>

The totals around the perimeter = “marginals”

The cross-classifications = “cells”

The totals around the perimeter = “marginals”

N = “grand total”
Recap

- The Chi-Square test is used to determine if a statistical relationship exists between two categorical variables
  - *This is the test to use for contingency tables/cross tabulations*

- The null hypothesis of every chi-square test is that “no statistically significant relationship exists” between the 2 variables

---

Recap

- **Chi Square**
  - *Three assumptions:*
    - The variables must be categorical
    - The observations must be independent
    - All cells must have at least 5 expected observations
      - If not, we can use the Fisher's Exact Test to test for a statistical relationship
    - *Chi-square is always a positive number (it doesn’t have “direction”)*
      - If there is no relationship, chi-square = 0
      - To determine direction, “measures of association” are used

- **Fisher’s Exact Test**
  - *The equivalent of a chi-square test, but for smaller sample sizes in a 2x2 table (i.e. some cells in the contingency table have less than 5 expected observations)*
Recap

- After constructing a contingency table:

- Steps:
  - First, we test determine if a relationship exists between the variables
    - This shows whether or not the observed relationship is significant
  - Next, we describe the relationship between the variables using "measures of association"
    - This shows the strength of the association, and (for ordinal variables) the direction of the relationship

Measures of Association

- Once you’ve determined that a relationship exists by using the Chi-square test, you want to further describe the nature of that relationship
  - How strong is the association between the variables?
  - What direction (positive or negative) is the association? (for ordinal variables)

- Measures of association are used to answer the above 2 questions

- Characteristics of measures of association:
  - If the relationship between the variables is perfect, the measure equals 1 or -1
  - If there is no relationship between the variables, the measure equals 0
  - For ordered variable measures, the sign of the measure indicates the direction of the relationship
Measures of Association

- **Nominal**
  - *Cramer’s V used if your DV and IV are both nominal*
    - Ranges from 0 (no relationship) to 1
    - Cannot be negative - the relationships between nominal variables do not have direction

- **Ordinal**
  - *Kendall’s tau-b is for “square” tables, with equal numbers of rows & columns*
  - *Kendall’s (Stuart’s) tau-c is for “rectangular” tables, with different numbers of rows & columns*
  - *Both of these can be negative depending on the direction of the relationship*

---

**Table 10.2: How to Interpret Measures of Association**

<table>
<thead>
<tr>
<th>Measure of Association ($X$)</th>
<th>Qualitative Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 ≤ $X$ &lt; 0.10</td>
<td>Very Weak</td>
</tr>
<tr>
<td>0.10 ≤ $X$ &lt; 0.20</td>
<td>Weak</td>
</tr>
<tr>
<td>0.20 ≤ $X$ &lt; 0.30</td>
<td>Moderate</td>
</tr>
<tr>
<td>$X$ ≥ 0.30</td>
<td>Strong</td>
</tr>
</tbody>
</table>
Contingency Tables in SPSS

- Open “Manell_Geus_transformed_datafile.sav”
- We want to determine if someone’s gender affects their belief that coral bleaching is a top threat to coral reefs
- Let’s cross-tabulate “male gender” and “threat_bleach”
- Dependent variable = “threat_bleach” in rows
- Independent variable = “male gender” in columns
Contingency Tables in SPSS

• Click on “cells”
• Make sure “column” percentages is checked
• Run the analysis

Approximately 4% of males and 4% of females chose coral bleaching as a top threat to coral reefs.

Gender does not seem to have an effect on choosing whether coral bleaching is a top threat to coral reefs.

However, we must perform the chi-square test to be certain.
Chi Square and Measures of Association in SPSS

- In our “crosstabs” window, click on “Statistics”
- Check the box for Chi Square
  - To test if there is a relationship
- Check the box for Tau-B
  - To test the strength of the relationship
  - Since the variables are ordinal and the table is square

Null hypothesis = no relationship
Alternative hypothesis = there is a relationship
Chi square is a valid test since all cells have expected counts greater than 5
Chi square statistic value: 0.007
  - P-value = 0.934
  - Tau-B statistic = 0.005
    - P-value = 0.934
  - “very weak” relationship
Since 0.934 > 0.05, there is no significant relationship between gender and the belief that coral bleaching is a top threat to coral reefs
- We fail to reject the null hypothesis

Test Statistic | Significance/P-value
---|---
Chi-Square Tests
  - Pearson Chi-Square: 0.007
    - Asymptotic Significance (2-sided): 0.934
  - Continuity Correction: 0.000
    - Exact Sig. (2-sided): 1.000
  - Likelihood Ratio: 0.007
    - Exact Sig. (2-sided): 0.934
  - Fisher’s Exact Test: 0.007
    - Linear-by-Linear Association: 1.000
    - Asymptotic Significance (1-sided): 0.576

Symmetric Measures
  - Kendall’s tau-b: 0.005
    - Asymptotic Standard Error: 0.948
    - Asymptotic Approx. T: 0.093
    - Approximate Significance: 0.934

N of Valid Cases: 302
Contingency Tables in SPSS

- We want to determine if peoples’ perceptions concerning whether sedimentation caused by fire is a top threat to coral reefs has an effect on whether they would participate in volunteering to practice better fire safety to prevent wildfires.
- Let’s cross-tabulate “threat_sedfire” and “volunteer_safety_NS”.
- Dependent variable = “threat_sedfire” in rows.
- Independent variable = “fire_economy_NS” in columns.
Contingency Tables in SPSS

- Click on "cells"
  - Make sure "column" percentages is checked
- Click on "statistics"
  - Make sure Chi-Square and Tau-c are checked (ordinal data and rectangular table)
- Run the analysis

3x2 table
- Similar column percentages at all levels of volunteer participation
  - 6% and 5%
  - 40% and 38%
  - 55% and 57%
- The belief that sedimentation caused by fires is a top threat to coral reefs does not seem to have an effect on whether someone would volunteer to practice better fire safety
  - However, we must perform the chi-square test to be certain
Chi Square and Measures of Association in SPSS

- For tables larger than 2x2, a minimum expected count of 1 is permissible as long as no more than about 20% of the cells have expected values below 5 (Cochran, 1954)
- Null hypothesis = no relationship
- Alternative hypothesis = there is a relationship
- Chi square statistic value: 0.114
  - P-value = 0.945
- Tau-C statistic = 0.014
  - P-value = 0.736
  - “very weak” relationship
- Since 0.945 > 0.05, there is no significant relationship between the belief that sedimentation caused by fires is a top threat to coral reefs and whether someone would volunteer to practice better fire safety
  - We fail to reject the null hypothesis

Contingency Tables in SPSS

- We want to determine if boat ownership affects where people go in Merizo to do hook-and-line fishing
- Let’s cross-tabulate “boat” and “activity_hookline”
- Dependent variable = “activity_hookline” in rows
- Independent variable = “boat” in columns
Contingency Tables in SPSS

- Click on “cells”
  - Make sure “column” percentages is checked
- Click on “statistics”
  - Make sure Chi-Square and Cramer’s V are checked (nominal data)
- Run the analysis
Contingency Tables in SPSS

- 4x2 table
- Differing column percentages at each fishing location
  - 22% and 15%
  - 39% and 52%
  - 15% and 2%
  - 23% and 30%
- Boat ownership may have an effect on where people go to do hook-and-line fishing
  - However, we must perform the chi-square test to be certain

Chi Square and Measures of Association in SPSS

- Null hypothesis = no relationship
- Alternative hypothesis = there is a relationship
- Chi square statistic value: 8.36
  - P-value = 0.039
- Cramer’s V statistic = 0.184
  - P-value = 0.039
  - “weak” relationship, but significant
- Since 0.039 < 0.05, there is a significant relationship between boat ownership and hook-and-line fishing location
  - We reject the null hypothesis
Practice!

- Create a contingency table based on “fish_harvest” and “condition_numfish_NS”
  - We want to know if the participation in the fishing/harvesting of marine resources has an effect on someone’s perception concerning the number of fish
- What is your dependent variable?
- What is your independent variable?
- What is the chi-square test telling you?
- Which measure of association will you use?
  - What does the measure of association tell you?

Practice!

- DV = perception of number of fish
- IV = participation in fishing/harvesting of marine resources
- Use Tau-C for ordinal data and rectangular table (5x2)
- Chi square = 15.272
  - p-value = 0.004 < 0.05
- Tau-C = 0.163
  - P-value = 0.012
- We reject the null hypothesis
  - There is a significant positive relationship between fishing/harvesting and perception of the number of fish
  - Those who fish/harvest have a "more positive" perception concerning the number of fish
  - The relationship is statistically significant, but it is relatively weak
Practice!

Create a contingency table based on “enviro_education_NS” and “resident_responsible.”
- We want to know if there is a relationship between those who believe that “every resident is responsible for taking care of the reefs” and the frequency at which they attend local education/awareness initiatives.

What is your dependent variable?
What is your independent variable?
What is the chi-square test telling you?
Which measure of association will you use?
- What does the measure of association tell you?

DV = Frequency of attendance of local education/awareness initiatives
IV = Belief that “every resident is responsible for taking care of the reefs”
Use Tau-C for ordinal data and rectangular table (5x2)
Chi square = 7.484
- p-value = 0.112 > 0.05
Tau-C = -0.033
- P-value = 0.552

We fail to reject the null hypothesis.
- There is no significant relationship between the belief that all resident are responsible for reef protection and frequency of attendance at education/awareness initiatives.

Save your output as “Manell_Geus_Output_Contingency.spv”
T-tests and ANOVA

Day 4: September 15, 2016

T-tests

- Widely used, very important
- When you want to know if there is a “statistically significant difference,” some form of t-test is used
- Example: Monitoring data
  - 50% of people participate in pro-environmental behavior in 2005
  - 60% of people participate in pro-environmental behavior in 2015
  - Is this a statistically significant increase?
    - It depends on sample sizes and standard deviations......
- Null hypothesis = no significant difference
Types of T-tests

- One sample
  - Compare a sample mean to a point estimate
- Two sample - paired
  - “Before and after”
- Two sample – independent
  - Difference of means across groups

In two sample t-tests, the difference of the means of the separate groups are calculated and confidence intervals are created around the difference.

- If the confidence interval does not contain zero, then there is a statistically significant difference

One Sample T-test

- Good news!
  - We already know how to perform one sample T-tests because we practiced them in our hypothesis tests lesson
- We are seeking to determine if a sample mean is “statistically significantly” different from a “test value”
- Open the file “manell_geus_transformed_datafile.sav”
- Let’s examine “infosource_television”
  - Let’s imagine that a past socioeconomic monitoring study found that 30% of the population used television as a source for coral reef information
  - 30% will be our test value
One Sample T-test

- Null Hypothesis: The percentage of people that use the tv as a source of coral reef information in Merizo is **NOT** significantly different from 30%
- Alternative Hypothesis: The percentage of people that use the tv as a source of coral reef information in Merizo is significantly different from 30%

Since our p-value = 0.007 < 0.05,
- “we reject the null hypothesis at the 95% confidence level”
- We are 95% confident that the population mean is statistically different from 30%
- **Since the last round of monitoring surveys, the proportion of people in Merizo that use the TV as a source of coral reef information has increased**
- Notice that the 95% confidence interval **DOES NOT** contains zero, meaning that difference between the sample mean (38%) and the test value (30%) is **STATISTICALLY SIGNIFICANT**
Two Sample t-test: Paired Samples

- Paired sample t-tests are used in ‘before-after’ studies, or when the samples are the matched pairs, or when it is a case-control study.
- For example, we give weight loss treatment to a group of voluntary participants.
  - The paired t-test can be used to test for a statistically significant reduction in weight from before the treatment to after the treatment.
    - Null hypothesis: there is no difference in weight from before to after the treatment.
    - Alternative hypothesis: there is a significant decrease in weight after the treatment.
- In paired t-tests, each “group” (i.e. the before and after) is dependent upon each other.
  - Measurements are taken from the same group of respondents “before” and “after”.
- We test the difference between “before” and “after” for statistical significance.

Two Sample T-test: Paired Samples

- Since we don’t have truly “before and after” data in the Manell-Geus questionnaire, we will use a hypothetical data set for this lesson.
- Open the file “Paired T-test Example.sav” (SPSS data set).
  - What do we have here?
    - Weight loss treatment study.
Two Sample T-test: Paired Samples

- Let’s investigate if the weight loss treatment significantly decreased the participants’ weight.
- Null hypothesis: There has NOT been a statistically significant reduction in weight since the treatment.
- Alternative hypothesis: There has been a statistically significant reduction in weight since the treatment.

Sample Means

- Shows how related the 2 variables are
- Mean DIFFERENCE
- P-value
Since our p-value = 0.000 < 0.05
- "we reject the null hypothesis at the 95% confidence level"
- We are 95% confident that the weight loss treatment has led to a reduction in weight
- There has been a statistically significant weight reduction

Let’s investigate if the weight loss treatment significantly decreased the participants’ cholesterol

Null hypothesis: There has NOT been a statistically significant reduction in cholesterol since the treatment

Alternative hypothesis: There has been a statistically significant reduction in cholesterol since the treatment
Two Sample T-test: Paired Samples

- Since our p-value = 0.000 < 0.05
  - “we reject the null hypothesis at the 95% confidence level”
- We are 95% confident that the weight loss treatment has led to a reduction in cholesterol
- There has been a statistically significant cholesterol reduction

<table>
<thead>
<tr>
<th>Paired Samples Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>---------------------------</td>
</tr>
<tr>
<td>Paired Samples Correlations</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>---------------------------</td>
</tr>
<tr>
<td>Paired Samples Test</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Paired Differences</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>---------------------------</td>
</tr>
</tbody>
</table>

Likert Data

- Before moving forward with the Independent Samples T-test, we should discuss Likert Data
- Likert Data = Measures attitudes by asking people to respond to a series of statements about a topic, in terms of the extent to which they agree with them
  - Most widely used approach to scaling responses in survey research
  - Usually a five point scale
  - Examples from Manell-Geus Questionnaire:
    - “Please indicate your level of agreement with each of the following statements”
      - Strongly disagree, disagree, neither agree nor disagree, agree, strongly agree
    - “In your opinion, how is each of the following natural resources currently doing in Merizo?”
      - Very bad, bad, neither good no bad, good, very good
Likert Data

- Likert Data is ordinal, and with ordinal data, you (usually) cannot calculate means or perform the same statistical tests that you would on continuous data.
- *However, a contingent of scholars in the social sciences agree that Likert Data can be analyzed as continuous data*
  - i.e. we can calculate means and perform T-tests and ANOVA on Likert Data
  - "Generally speaking, the choice between the two analyses (parametric and non-parametric) is tie. If you need to compare two groups of five-point Likert data, it usually doesn’t matter which analysis you use. Both tests almost always provide the same protection against false negatives and always provide the same protection against false positives.”
  - (de Winter, J.C.F. and D. Dodou 2010)

Two Sample t-test: Independent Samples

- Helps you compare whether two groups have statistically significant different mean values
  - For example, whether men and women have different mean heights
- Null hypothesis: there is no significant difference between the groups
- Alternative hypothesis: There is a significant difference between the groups
- There are 2 ways this test can be performed:
  - Assuming Equal Variances
  - Assuming Unequal Variances
  - The F-test is used to determine if variances are equal or not
    - Null hypothesis if F-test = variances are assumed equal
Two Sample t-test: Independent Samples

- Open the file “Manell_Geus_transformed_datafile.sav”
- Let’s examine if college completion has any effect on peoples’ perceptions concerning the condition of the beach shoreline
- We need “college” and “condition_beach_NS”
- “College” is our independent (grouping) variable
- “condition_beach_NS” is our dependent variable
- We must “define groups”

Null hypothesis = college completion has NO effect on peoples’ perceptions of beach/shoreline quality
Alternative hypothesis = college completion does have an effect on peoples’ perceptions of beach/shoreline quality
Two Sample t-test: Independent Samples

Let’s breakdown this output:

Means of “condition_beach_NS” delineated by college completion
• 3.00 and 2.89 are fairly close
• College completion doesn’t “seem to have” an effect on peoples’ perceptions of beach/shoreline quality, but we must interpret the t-test to be sure

Significance of F-test tells us if we “assume equal variances” or not
• Since 0.213 > 0.05, we “fail to reject the null hypothesis of equal variances”
  • Therefore we assume equal variances and continue
  • The T statistics and associated p-values will be different depending on how you interpret your F-test, so we want the p-value that corresponds to “equal variances assumed”
    • T = -0.701   p-value = 0.484 (this is greater than 0.05, so.....)
  • We fail to reject the null hypothesis at the 95% confidence level – college completion DOES NOT have an effect on peoples’ perception of beach/shoreline quality
Two Sample t-test: Independent Samples

- Now, let’s examine if fishing “to sell” has any effect on peoples’ level of support for “creating areas for only traditional fishing”
- We need “fish_sell” and “mng_tradfish_NS”
  - $DV = mng\_tradfish\_NS$ $IV = “fish\_sell”$ (grouping variable)
- Null hypothesis: If someone fishes to sell, this has NO effect on their level of support for creating areas for only traditional fishing
- Alternative hypothesis: If someone fishes to sell, this DOES have an effect on their level of support for creating areas for only traditional fishing

Two Sample t-test: Independent Samples

- When we “define groups,” we use a “cut point”
- The cutpoint will always make the groups be “greater than or equal to the cutpoint” and “less than the cutpoint”
  - i.e., a cutpoint of 2 will result in a group of “greater than or equal to 2” and “less than 2”
  - For “fish\_myself,” using a cutpoint of 2 will result in 2 groups
    - Those that fish to sell at any frequency
    - Those that never fish to sell
Two Sample t-test: Independent Samples

- Significance of F test = 0.474 < 0.05
  - We fail to reject equality of variances (i.e. variances are assumed to be equal)
- Significance of t-test = 0.354 < 0.05
  - We fail to reject the null hypothesis of no relationship at the 95% confidence level
  - Fishing to sell DOES NOT have an effect on the level of support for creating areas for only traditional fishing

One-Way ANOVA

- Essentially an Independent t-test with >2 groups
- Tests for statistically significant differences in the means of 2 or more groups
- One-way ANOVA is an omnibus F-test statistic and cannot tell you which specific groups were significantly different from each other (i.e. p-value<0.05), only that at least two groups were
  - To determine which specific groups differed from each other, you need to use a post hoc test
One way ANOVA Post-Hoc Test

- The most common post-hoc test is the Tukey’s HSD test
- The post-hoc test will tell you which groups’ means are significantly different
- The difference of the means of the separate groups are calculated and confidence intervals are created around the difference
  - If the confidence interval does not contain zero, then there is a statistically significant difference
- Null hypothesis = no significant difference

One-Way ANOVA

- Let’s examine if where people participate in beach recreation has an effect on their overall perception concerning the change in condition of marine resources over the last 10 years
- We need “activity_beach” and “last10index_norm”
- In SPSS, the “dependent list” corresponds to the variables you want to take the mean of (last10index_norm), and the “factor” represents the grouping variable (activity_beach)
One-Way ANOVA

- The Tukey post-hoc test tells us which groups are significantly different from each other.
- The “last 10 index” increases as positive perception concerning the change in condition of marine resources increases.
- With 95% confidence, we conclude that those that participate in beach recreation at Achang Preserve AND Cocos Lagoon are more likely to have a more positive perception concerning the change in condition of marine resources when compared to those that participate in beach recreation only in Cocos Lagoon.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Last 10 index norm</th>
<th>N</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>10</td>
<td>50</td>
<td>56.44</td>
</tr>
<tr>
<td>Yes, in Cocos Lagoon</td>
<td>86</td>
<td>46.96</td>
<td></td>
</tr>
<tr>
<td>Yes, in Achang Preserve</td>
<td>10</td>
<td>66.51</td>
<td></td>
</tr>
<tr>
<td>Yes, in both places</td>
<td>55</td>
<td>62.85</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>291</td>
<td></td>
<td>52.41</td>
</tr>
</tbody>
</table>

Multiple Comparisons

<table>
<thead>
<tr>
<th>Activity</th>
<th>Last 10 index norm</th>
<th>Mean Difference</th>
<th>Std. Error</th>
<th>Sig.</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>10</td>
<td>10.357</td>
<td>4.157</td>
<td>.014</td>
<td>.2113</td>
<td>-17.59</td>
</tr>
<tr>
<td>Yes, in Cocos Lagoon</td>
<td>86</td>
<td>-10.357</td>
<td>4.157</td>
<td>.014</td>
<td>.2113</td>
<td>-17.59</td>
</tr>
<tr>
<td>Yes, in Achang Preserve</td>
<td>10</td>
<td>-8.742</td>
<td>1.000</td>
<td>.000</td>
<td>-17.53</td>
<td>.1401</td>
</tr>
<tr>
<td>Yes, in both places</td>
<td>55</td>
<td>-2.071</td>
<td>0.998</td>
<td>.037</td>
<td>-4.01</td>
<td>0.97</td>
</tr>
</tbody>
</table>

*The mean difference is significant at the 0.05 level.*

One-Way ANOVA

- Let’s examine if where people participate in hunting has an effect on their overall support for management options.
- We need “activity_hunt” and “ManagementAgreeIndex_norm”.
- In SPSS, the “dependent list” corresponds to the variables you want to take the mean of (ManagementAgreeIndex_norm), and the “factor” represents the grouping variable (activity_hunt).
One-Way ANOVA

• The “Management Agree index” increases as support for management options increases
• With 95% confidence, we conclude that the location in which people hunt has no effect on their overall support for management options
• There are no significant p-values when comparing the groups

Practice!

- Open the file “Paired T-test Example.sav”
- Let’s investigate if the weight loss treatment significantly decreased the participants’ body fat percentage
- Null hypothesis?
- Alternative hypothesis?
- What is your conclusion?
Practice!

- Null hypothesis: There has NOT been a statistically significant reduction in body fat percentage since the treatment
- Alternative hypothesis: There has been a statistically significant reduction in body fat percentage since the treatment
- Conclusion: we reject the null hypothesis at the 95 % confidence level
  - *There has been a statistically significant decrease in body fat percentage since the weight loss treatment*

```
<table>
<thead>
<tr>
<th>Paired Differences</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Paired Differences</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Std. Deviation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Std. Error</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>95% Confidence Interval of the Difference</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upper</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>df</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part 1 Body fat percentage before: Body fat percentage after</td>
<td>2.2960%</td>
<td>2.6814%</td>
<td>6.38016%</td>
<td>1.95194%</td>
<td>3.04369%</td>
</tr>
</tbody>
</table>
```

Practice!

- Open the file “Manell_Geus_transformed_datafile.sav”
- Let’s examine if receiving benefit from Achang Preserve (“benefit”) has any effect on their perception of overall management success (“ManagementSuccessIndex_norm”)
- Null hypothesis?
- Alternative hypothesis?
- What is your conclusion?
- *Don’t forget to “define groups”*
Practice!

- Null hypothesis: Receiving benefit form Achang Preserve has NO effect on peoples’ overall opinion of management success
- Alternative hypothesis: Receiving benefit form Achang Preserve DOES have an effect on peoples’ overall opinion of management success
- Conclusion: we fail to reject the null hypothesis at the 95% confidence level
  - Receiving benefit form Achang Preserve has NO effect on peoples’ overall opinion of management success

Practice!

- Let’s examine if participation in volunteering at least once a year to help protect reefs (’volunteer_protect_NS’) has any effect on peoples’ overall perception concerning the condition of marine resources (’ConditionIndex_norm’)
- What is your conclusion?
- *Don’t forget the Post-Hoc test*
Practice!

- Those that “would” volunteer at least once a year to help protect the reefs had a more positive perception concerning the condition of marine resources when compared to those who would “consider” volunteering at least once a year to help protect the reefs.

Save your output as “Manell_Geus_Output_Ttest and anova.spv”
7.1 Which T-test is suitable for “before and after” studies?

A. One sample t-test
B. Paired samples t-test
C. Independent samples t-test
D. One way ANOVA
7.2 What does an ANOVA post-hoc test do?

A. Tells us if our ANOVA model is significant
B. Tells us which t-test test is the proper one to use
C. Determines the overall “fit” of the model
D. Determines statistical significant differences between groups in ANOVA analysis

7.3 True or False: Likert data can be analyzed as continuous in some cases

A. True
B. False
7.4 What is the null hypothesis of the F test in an Independent samples T-test?
A. Assume equal variances
B. Assume unequal variances
C. There is no significant difference between the groups
D. There is significant difference between the groups

7.5 What happens when a p-value is less than 0.05?
A. We accept the null hypothesis
B. We fail to reject the null hypothesis
C. We reject the null hypothesis
D. We reject the alternative hypothesis
Day 5

- Correlation and Regression

- Data Visualization
Correlation Analysis

Day 5: September 16, 2016

Recap

- Correlation is a statistical technique that is used to measure and describe the STRENGTH and DIRECTION of the LINEAR relationship between two variables
  - Correlation coefficient ranges from -1 to 1
  - Zero = no linear relationship
  - Closer to zero = weaker linear relationship
  - Closer to 1 = strong positive linear relationship
  - Closer to -1 = strong negative linear relationship
- Can only be performed with 2 continuous or binary variables
- A positive correlation means that if one variable gets bigger, the other variable tends to get bigger
- A negative correlation means that if one variable gets bigger, the other variable tends to get smaller
- Correlation DOES NOT mean Causation
Correlation Analysis

- Correlation Analysis is good for creating “demographic profiles” of certain respondent types
  - Demographics – population characteristics such as gender, age, income, education, etc.
  - For example, if we want to answer what type of person fishes or harvests for marine resource?
- We can run a correlation analysis with “fish_harvest,” age, gender, etc. to create a “demographic profile” of those who fish or harvest for marine resources

Correlations in SPSS

- Open the file “Manell_Geus_transformed_datafile.sav”
- Let’s create a demographic profile of those who fish or harvest for marine resources
- “fish_harvest”
- “age”
- “tenure”
- “high school”
- “college”
- “male gender”
- “family”
- “own_home”
- “income_midpoint”
  - A transformed version of the income variable in which the midpoint of each category is used to make the variable continuous (continuous data is necessary for correlation analysis)
Correlations in SPSS

Correlation Coefficient = represents relationship between the 2 variables

Significance of Correlation Coefficient = p-value = represents statistical significance of the relationship between the 2 variables

Sample size of respondents that answered both questions

Correlation Coefficient =

Significance of Correlation Coefficient = p-value =

Sample size of respondents that answered both questions

241
Correlations in SPSS

For the purposes of creating a demographic profile of those who fish or harvest for marine resources, we focus on the correlations that correspond to "fish_harvest"

- Look at p-values
  - Are any less than 0.05?
    - No
  - Are any less than 0.10?
    - Yes, "male gender"

Male gender and fish_harvest are significantly positively correlated at the 90% confidence level

We are 90% confident that males are more likely to fish or harvest for marine resources when compared to females

Correlations in SPSS

What are the other correlations telling us?

- Age and income are significantly positively correlated
- Older people are more likely to have more money
- College completion and income are significantly positively correlated
- Those who have completed college are more likely to have more money
- Some other ones??
Correlations in SPSS

- Let's examine our normalized indices and see which ones correlate with each other.

Correlations in SPSS

- What are correlations telling us?
  - Condition index and Last 10 Index are significantly positively correlated.
    - Those with a more positive perception concerning the condition of marine resources have a more positive perception concerning the change in the condition of marine resources as well.
- Some other ones??
Correlations in SPSS

- Other results from our Index Correlation Analysis
  - Those with a more positive perception concerning the condition of marine resources have a more positive perception of management efficiency
  - Those that fish for marine resources more frequently have a more positive opinion of Achang Preserve
  - Those with a more positive perception concerning the change in the condition of marine resources are more likely to participate in environmental behavior more frequently

Practice!

- Let’s investigate the correlation between the number of times a family has been impacted by a flood (“flood_impact”) and their overall support for management (ManagementAgreeIndex_norm”)
- What is the correlation coefficient?
- Is the correlation significant?
- What is our conclusion?
Practice!

■ What is the correlation coefficient?
  - 0.010

■ Is the correlation significant?
  - P-value = 0.890 > 0.05; no significance

■ What is our conclusion?
  - There is not a statistically significant correlation between the number of times a family has been impacted by a flood and their overall support for management

Practice!

■ Let’s investigate the correlation between the percentage of a family’s seafood that they get from Merizo (“percent_Merizo”) and their overall opinion concerning marine resource condition (“ConditionIndex_norm”)

■ What is the correlation coefficient?

■ Is the correlation significant?

■ What is our conclusion?
Practice!

- What is the correlation coefficient?  
  - 0.307

- Is the correlation significant? 
  - P-value = 0.000 < 0.05; Yes, this is significant and positive

- What is our conclusion? 
  - As people obtain a higher percentage of their seafood from Merizo, their overall perception concerning the condition of marine resources becomes more positive

Practice!

- Let’s investigate the correlation between the belief that sedimentation caused by fires is a top threat to coral reefs (“threat_sedfire”) and their overall belief concerning the severity of fires (FireSeverityIndex_norm”)

- What is the correlation coefficient?

- Is the correlation significant?

- What is our conclusion?
Practice!

- What is the correlation coefficient?
  - -0.029

- Is the correlation significant?
  - $P$-value = 0.632 > 0.05; No significance

- What is our conclusion?
  - *There is no statistically significant correlation between the belief that sedimentation caused by fires is a top threat to coral reefs and their overall belief concerning the severity of fires*

Save your output as “Manell_Geus_Output_correlations.spv”
Simple Linear Regression

Day 5: September 16, 2016

Recap – Simple Linear Regression

- One dependent variable (Y) and one independent variable (X)
  - *A regression model is a model that describes how a variable X influences the value of another variable Y*

- At the center of the regression analysis is the task of fitting a single line through a scatter plot.
The Regression Line

- A regression line is fit through a scatter plot to minimize the deviations from the line
  - The line is meant to predict the outcome of Y, given a certain X value
  - Expected values are along the line
  - Observed values are as close as possible to the line
  - By minimizing the deviations from the line (minimizing the error between observed and expected outcomes), we increase our predicative power

Recap – Simple Linear Regression

- A technique used to determine the linear relationship between two variables, and in turn, attempt to predict changes in Y due to changes in X
  - Defined by the formula $Y = c + b \times X$
    where $Y = \text{estimated dependent variable}$
    $c = \text{constant (intercept)}$
    $b = \text{regression coefficient (slope)}$
    $X = \text{independent variable}$

- As $X$ changes by (1), we expect $Y$ to change by (b)
- When $X=0$, $Y = c$
- $b$ is tested for statistical significance (i.e. does X have a significant effect on Y?)
  - Using $p$-values; is $p$-value less than alpha?
Assumptions of Linear Regression

- Linearity: The existing relation between X and Y is linear
- Homogeneity: The mean value of the error is zero
- Homoscedasticity: The variance of the errors is constant
- Independence: The observations are independent
- Normality: The errors follow a normal distribution

If all of these assumptions are not satisfied, then your regression model is not valid

Linear Regression in SPSS

- Open the file “Manell_Geus_transformed_datafile.sav”
- Let’s investigate if the perception concerning the condition of marine resources (“ConditionIndex_norm”) is a good (or bad) predictor of the perception concerning the change in the condition of marine resources (“Last10Index_norm”)
Linear Regression in SPSS

- Adjusted R square = ___% of the variation in Y is explained by X
  - 68.5% of the variation in last10index_norm is explained by conditionindex_norm
  - Higher adj R square values = more predictive power

Conclusions?
- Regression line
  - Last 10 Index_norm = 4.687 + (0.881)*Condition Index_norm
  - Significance of X’s predictive power of Y = 0.000 < 0.05
    - X is a statistically significant predictor of Y
    - Condition Index is a statistically significant predictor of Last 10 index
  - As someone’s normalized condition index increases by 1, we expect their normalized last 10 years index to increase by 0.881
  - Peoples’ perceptions concerning the condition of marine resources is linearly positively related to their perceptions concerning the change in the condition of marine resources over the last 10 years
  - As positive perception concerning the condition of marine resources increases, positive perception concerning the change in condition of marine resources increases
Linear Regression in SPSS

- Let’s investigate if owning a home ("own_home") is a good (or bad) predictor of whether or not someone owns a boat ("boat")

- Conclusions?
  - Adjusted R square
    - 5.4% of the variation in boat ownership is explained by home ownership
  - Regression line
    - \( \text{boat} = 0.074 + (0.19) \times \text{own_home} \)
    - Significance of X’s predictive power of Y = 0.000 < 0.05
      - X is a statistically significant predictor of Y
      - Home ownership is a statistically significant predictor of boat ownership
  - If someone owns a home, we expect their chances of owning a boat to increase by 19%
  - Home ownership is linearly positively related to boat ownership
Let’s investigate if the perception concerning the change in the condition of marine resources (“Last10Index_norm”) is a good (or bad) predictor of overall support for management (“ManagementAgreeIndex_norm”).

Conclusions?
- Adjusted R square
  - 9.7% of the variation in overall management support is explained by perceptions concerning the change in the condition of marine resources over the last 10 years
- Regression line
  - ManagementAgreeIndex_norm = 36.401 + (0.279)*last10index_norm
- Significance of X's predictive power of Y = 0.000 < 0.05
  - X is a statistically significant predictor of Y
  - Perceptions concerning the change in condition of marine resources is a statistically significant predictor of overall management support
- As someone’s normalized last10 index increases by 1, we expect their normalized management agreement index to increase by 0.279
- Peoples’ perceptions concerning the change in condition of marine resources is linearly positively related to overall management support
  - As positive perception concerning the change in condition of marine resources increases, overall support for management statistically significantly increases as well
Practice!

- Let’s investigate if the amount of years one has lived in Merizo (tenure”) is a good (or bad) predictor of frequency of participation in environmental behavior (“EnvBehaviorFrequencyIndex_norm”)
- What is the Adjusted R square telling us?
- What is the regression line equation?
- Is X a significant predictor of Y?
- What is our overall conclusion?

Practice!

- What is the Adjusted R square telling us?
  - 1.2% of the variation in environmental behavior participation is explained by tenure in Merizo
- What is the regression line equation?
  - EnvBehaviorIndex_norm = 56.353 – 0.167*(tenure)
- Is X a significant predictor of Y?
  - Yes, p-value = 0.049 < 0.05
- What is our overall conclusion?
  - The amount of years one has lived in Merizo is linearly negatively related to their frequency of participation in environmental behavior
  - As tenure increases, the frequency of participation in environmental behavior decreases
  - As tenure increases by 1, the EnvBehaviorIndex_norm decreases by 0.167
Practice!

- Let's investigate if the belief that every resident is responsible for protecting coral reefs ("resident_responsible") is a good (or bad) predictor of their overall consideration of volunteering to protect coral reefs ("VolunteerIndex_norm")

- What is the Adjusted R square telling us?

- What is the regression line equation?

- Is X a significant predictor of Y?

- What is our overall conclusion?

Practice!

- What is the Adjusted R square telling us?
  - The belief that all residents are responsible for protecting coral reefs does not explain any variation in one's consideration of volunteering to protect coral reefs

- What is the regression line equation?
  - VolunteerIndex_norm = 73.039 - 2.001*(resident_responsible)

- Is X a significant predictor of Y?
  - No, p-value = 0.512 > 0.05

- What is our overall conclusion?
  - There is no statistically significant relationship between the belief that all residents are responsible for protecting coral reefs and one's consideration of volunteering to protect coral reefs

Save your output as “Manell_Geus_Output_simple regression.spv”
Multiple Linear Regression

Day 5: September 16, 2016

Recap

- Same as simple linear regression, but with multiple independent (X) variables
  - $X_1, X_2, X_3, X_n$, etc.
- Incorporates multiple “predictor” variables to “predict” the value of Y
- Strength of a regression model given by $R^2$
  - $R^2$ ranges from 0-1, with stronger (better predictive) models having an $R^2$ value closer to 1
  - Interpretation: if $R^2 = 0.50$, then “50% of the variation in $Y$ is explained by $X_1$, $X_2$, and $X_3$.”
Recap

- Model given by:
  \[ Y = c + b_1 + b_2 + \ldots + b_n \]
- Each coefficient (b) is tested for statistical significance using p-values
- Interpretation:
  - Each predictor (X) variable is interpreted on its own (i.e. “all else held equal”)
  - Example: if \( b_1 \) is significant, that “all else held equal, significantly effects Y”
  - If \( b_1 \) changes by 1, we expect Y to change by
  - If \( b_1 \) changes by 1, we expect Y to change by
  - If \( b_1 = b_2 = b_3 = b_4 = 0 \), then we expect \( Y = c \)

Multiple Regression in SPSS

- How can we predict overall support for management (“ManagementAgreeIndex_norm”)?
  - Let’s use frequency of environmental behavior participation, perception concerning the change in condition of marine resources, and amount of years lived in Merizo
  - EnvBehaviorIndex_norm, Last10Index_norm, tenure

Open the file “Maneli_Geus_datafile_transformed.sav”
Multiple Regression in SPSS

• Conclusions?
  • Adjusted R square
    • 21.3% of the variation in overall management support is explained by environmental behavior frequency, perception concerning the change in condition of marine resources, and the amount of years lived in Merizo

• Regression line
  • ManagementAgreeIndex_norm = 17.658 + (0.261)*EnvBehaviorIndex-norm + (0.317)*Last10Index_norm + (0.040)*tenure

• Significance of X's predictive power of Y
  • EnvBehaviorIndex p-value = 0.004 < 0.05
  • Last10Index p-value = 0.000 < 0.05
  • Tenure p-value = 0.705 > 0.05

• Frequency of environmental behavior participation and perception concerning the change in the condition of marine resources are significant predictors of overall management support, whereas tenure is not
• All else held equal, if EnvBehaviorIndex_norm increases by 1, we expect ManagementAgreeIndex_norm to increase by 0.261
• All else held equal, if Last10Index_norm increases by 1, we expect ManagementAgreeIndex_norm to increase by 0.317

Multiple Regression in SPSS

• How can we predict if people will consider volunteering to protect coral reefs (“VolunteerIndex_norm”)?
  • Let’s use amount of years lived in Merizo, college completion, Perception concerning management success, and perception concerning the change in condition of marine resources
  • Tenure, College, ManagementSuccessIndex_norm, Last10Index_norm
Multiple Regression in SPSS

- Conclusions?
  - Adjusted R square
    - 3.2% of the variation in peoples' consideration of volunteering to help protect coral reefs is explained by amount of years lived in Merizo, college completion, perception concerning management success, and perception concerning the change in condition of marine resources.

- Regression line
  - VolunteerIndex_norm = 63.219 + (0.139)^Tenure + (2.646)^College - (0.044)^ManagementSuccessIndex_norm + (0.153)^Last10Index_norm

- Significance of X’s predictive power of Y
  - Tenure p-value = 0.123 > 0.05
  - College p-value = 0.423 > 0.05
  - Management Success p-value = 0.567 > 0.05
  - Last10Index p-value = 0.009 < 0.05

- Perception concerning the change in the condition of marine resources is a significant predictor of peoples' consideration of volunteering to help protect coral reefs, whereas tenure, college completion, and perceptions of management success are not.
- All else held equal, if Last10Index_norm increases by 1, we expect VolunteerIndex_norm to increase by 0.153

Multiple Regression in SPSS

- How can we predict if people have heard of the coral reef conservation program? (“Program_CRCP”)?
  - Let’s use age, college completion, overall support for management, and whether their household receives benefit from Achang Preserve.

  - Age, college, ManagementAgreeIndex_norm, and benefit
Multiple Regression in SPSS

- **Conclusions?**
  - Adjusted R square
    - 5.5% of the variation in peoples’ knowledge of CRCP is explained by age, college completion, overall support for management, and whether their household receives benefit from Achang Preserve
  - Regression line
    - Program_CRCP = 0.465 - (0.002)*Age - (0.245)*College + (0.005)*ManagementAgreeIndex_norm - (0.139)*Benefit
  - Significance of X’s predictive power of Y
    - Age p-value = 0.574 > 0.05
    - College p-value = 0.029 < 0.05
    - Management Agree p-value = 0.027 < 0.05
    - Benefit p-value = 0.197 > 0.05
  - College completion and overall support for management options are significant predictors of peoples’ knowledge of CRCP, whereas age and whether a household receives benefit form Achang Preserve are not
  - All else held equal, if someone completed college, we expect their probability of knowing about CRCP to decrease by 24.5%
  - All else held equal, if ManagementAgreeIndex_norm increases by 1, we expect their probability of knowing about CRCP to increase by 0.5%

---

**Practice!**

- How can we predict peoples perceptions of the Achang Preserve? (“AchangPreserveIndex_norm”)
- Let’s use college completion (“college”), amount of years lived in Merizo (“tenure”), knowledge of the Micronesia challenge (Program_MC”), and perceptions of marine resource condition (“ConditionIndex_norm”)
- What is the adjusted R square telling us?
- What is the regression line equation?
- Are the X’s significant predictors of Y?
- What is our overall conclusion?
Practice!

- What is the adjusted R square telling us?
  - 16.8% of the variation in people's perceptions concerning Achang Preserve is explained by college completion, amount of years lived in Merizo, knowledge of the Micronesia Challenge, and perception of marine resource condition.

- What is the regression line equation?
  - AchangPreserveIndex_norm = 38.132 - (12.646)*college + (0.186)*tenure + (11.548)*Program_MC + (0.249)*ConditionIndex_norm

- Are the X's significant predictors of Y?
  - College p-value = 0.025 < 0.05
  - Tenure p-value = 0.175 > 0.05
  - Program_MC p-value = 0.018 < 0.05
  - Condition Index p-value = 0.015 < 0.05

Practice!

- What is our overall conclusion?
  - College completion, knowledge of the Micronesia Challenge, and perception of marine resource condition are significant predictors of peoples' perceptions of Achang Preserve, whereas the amount of years lived in Merizo is not.
  - All else held equal, if someone completed college, we expect their perception of Achang Preserve to be more negative (completion of college leads to a 12.6 unit decrease in the index).
  - All else held equal, if someone has heard of the Micronesia Challenge, we expect their perception of Achang Preserve to be more positive (knowledge of the MC leads to a 11.5 unit increase in the index).
  - All else held equal, as perception of marine resource condition is more positive, the perception of the Achang Preserve is more positive as well (a 1 unit increase in the Condition Index leads to a 0.249 unit increase in the Achang Preserve Index).
Practice!

■ How can we predict if someone is familiar with Achang Preserve? (“familiar_Achang”)
■ Let’s use overall support for management (“ManagementAgreeIndex_norm”), amount of years lived in Merizo (“tenure”), and the percentage of a household’s seafood that comes from Merizo (“Percent_merizo”)
■ What is the adjusted R square telling us?
■ What is the regression line equation?
■ Are the X’s significant predictors of Y?
■ What is our overall conclusion?

Practice!

■ What is the adjusted R square telling us?
  - 21% of the variation in peoples’ familiarity with Achang Preserve is explained by overall support for management, amount of years lived in Merizo, and the percentage of a household’s seafood that comes from Merizo
■ What is the regression line equation?
  - “familiar_Achang” = 0.443 – (0.003)*ManagementAgreeIndex_norm + (0.003)*tenure + (0.007)*percent_merizo
■ Are the X’s significant predictors of Y?
  - Management Agree p-value = 0.023 < 0.05
  - Tenure p-value = 0.193 > 0.05
  - Percent_merizo p-value = 0.000 < 0.05
Practice!

- What is our overall conclusion?
  - Overall support for management options and the percentage of a household’s seafood that comes from Merizo are significant predictors of peoples’ familiarity with Achang Preserve, whereas the amount of years lived in Merizo is not.
  - All else held equal, more support for management leads to less familiarity with Achang Preserve (A 1 unit increase in the Management Agreement Index decreases the probability of being familiar with Achang preserve by 0.3%)
  - All else held equal, as the percentage of a household’s seafood that comes from Merizo increase, their familiarity with Achang Preserve increases (A 1% increase in the percentage of a household’s seafood that comes from Merizo increase the probability of being familiar with Achang Preserve by 0.7%)

(IF WE HAVE TIME)

VALIDITY OF REGRESSION MODELS

Day 5: September 16, 2016
Checking the Validity of Your Regression Model

- Multicollinearity
  - A phenomenon in which two or more predictor (X) variables in a multiple regression model are highly correlated

- Heteroscedasticity
  - A phenomenon in which the variance of the errors is varies across values of your predictor variable(s)
  - Recall: the “error” is the distance between an observed value of Y (given a value of X) and the associated expected value of Y (given the regression equation)

- Autocorrelation
  - A phenomenon in which a variable’s value is a function of its past values
  - Usually only a problem with time series data

If your model exhibits ANY of the above characteristics, it is INVALID

Tests to Determine Validity

- Multicollinearity
  - Collinearity diagnostics (under “Statistics” in our regression box)
  - Use the VIF (variance inflation factor)

- Heteroscedasticity
  - Breusch-Pagan test
    - Not as easy to do in SPSS, but there is a way

- Autocorrelation
  - Durbin-Watson test
  - However, since we are not dealing with time series data (we are dealing with survey data), we will only focus on diagnosing \textbf{multicollinearity and heteroscedasticity}
Testing for Multicollinearity in SPSS

- Open the file "Manell_Geus_transformed_datafile_validity.sav"
- Let's go back to our regression model
  - ManagementAgreeIndex_norm = 17.658 + (0.261)*EnvBehaviorIndex-norm + (0.317)*Last10Index_norm + (0.040)*tenure
- In the "statistics" box, select "collinearity diagnostics"

To test if our model has multicollinearity, we are concerned with the variance inflation factors.

As a standard rule of thumb:
- If any VIFs are greater than 10, your model is exhibiting multicollinearity (Hair, Anderson, Tatham, and Black 1995; Kennedy 1992; Marquardt 1970; Neter, Wasserman, and Kutner 1989)
- Since all of our VIFs are <10, are model does not exhibit multicollinearity (YAY!)
  - But does the model exhibit heteroscedasticity?........
Testing for Heteroscedasticity in SPSS

To perform the Breusch-Pagan test, we need to “save” a new “residuals variable”
- In the “save” box, check the box for “unstandardized residuals”

Now, a new variable was created in our data set
- \( RES_1 = \) unstandardized residuals

From here, we must compute the squared residuals
Testing for Heteroscedasticity in SPSS

- Once square residuals are computed, we must run a new regression model, replacing our previous dependent variable with the squared residuals variable.
- Uncheck the box for “unstandardized residuals.”

This is our regression output with our squared residuals as the dependent variable.

To determine if the model exhibits Heteroscedasticity, we must examine the F test and its associated significance.

- Null hypothesis = no Heteroscedasticity (i.e. “homoscedasticity”)
- Alternative hypothesis = there is Heteroscedasticity

Since our p value = 0.185, we fail to reject the null hypothesis of homoscedasticity and conclude that our model DOES NOT exhibit heteroscedasticity (YAY!!)

- No multicollinearity and no heteroscedasticity = a valid regression model
Testing for multicollinearity and heteroscedasticity in SPSS: Practice!

- Let’s examine another of our previously ran regression models
  - \( \text{AchangPreserveIndex\_norm} = 38.132 - (12.646)\text{college} + (0.186)\text{tenure} + (11.548)\text{Program\_MC} + (0.249)\text{ConditionIndex\_norm} \)

<table>
<thead>
<tr>
<th>Coefficients( ^{a} )</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>Beta</th>
<th>t</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>4.43</td>
<td>.991</td>
<td>4.469</td>
<td>.667</td>
<td></td>
</tr>
<tr>
<td>Management agree index_norm</td>
<td>-0.003</td>
<td>.991</td>
<td>-2.367</td>
<td>.023</td>
<td></td>
</tr>
<tr>
<td>tenure</td>
<td>.003</td>
<td>.992</td>
<td>.078</td>
<td>1.065</td>
<td>.193</td>
</tr>
<tr>
<td>Percent freeze</td>
<td>.007</td>
<td>.991</td>
<td>.464</td>
<td>7.646</td>
<td>.000</td>
</tr>
</tbody>
</table>

\( ^{a} \text{Dependent Variable: famili\_Achang} \)

Testing for multicollinearity in SPSS: Practice!

- All VIFs<10, our model does not exhibit multicollinearity
Testing for heteroscedasticity in SPSS: Practice!

F test p-value = 0.418 > 0.05
Failed to reject null hypothesis of homoscedasticity
Our model does not exhibit multicollinearity or heteroscedasticity, therefore it is a valid regression model.

Save your output as Manell_Geus_Output_regression validity
Quiz #8
Day 5: September 16, 2016

8.1 What does a regression line do?

A. Provide the “best fit” through a scatter plot
B. Use X to predict Y
C. Express the linear relationship between an independent variable and a dependent variable
D. All of the above
8.2 What does “b” represent in the equation:
\[ Y = c + b \cdot x \]
A. Slope  
B. Y-intercept  
C. X-intercept  
D. The independent variable

8.3 What does “Y” represent in the equation:
\[ Y = c + b \cdot x \]
A. Slope  
B. Y-intercept  
C. Dependent variable  
D. Independent variable
8.4 What does the adjusted R square tell us?

A. If X is a significant predictor of Y
B. How much of Y’s variance is explained by X
C. If Y is a significant predictor of X
D. How much of X’s variance is explained by Y

8.5 True or False: Regression Analysis is used to predict observed values

A. True
B. False
Data Visualization for Inferential Stats

Day 5: September 16, 2016

SPSS Output

- While SPSS is great for generating statistical analysis output, the output is not in an ideal format for reports and presentations
- It is up to the researcher to convey statistical output in a more understandable fashion to communicate with stakeholders and the general public
- We CANNOT just copy and paste SPSS output into a report and call it final
  - This is where data visualization of inferential stats comes in
  - Microsoft Excel and Microsoft Word can help us
Visualizing SPSS Results

- Open the file “Data Visualization.xlsx”
- We will use this file to generate “report ready” results based off of previous SPSS output
- SPSS output usually reports more than what is needed for “report ready” results
  - *It is up to us to extract the key information that we must communicate to our audience*

Visualizing Contingency Table/Chi Square Results

- Open the file “Manell_Geus_Output_Contingency.spv” from Day 4
- This is the output from our first Contingency Table analysis
- Pasting this directly into a report would be confusing to the reader
- We must extract the necessary information
  - The column percentages
  - The total frequencies
  - Refine our labels
Visualizing Contingency Table/Chi Square Results

- This is the corresponding Chi-square output for the Contingency Table analysis on the previous slide.
- Similarly, pasting this directly into a report would be confusing to the reader.
- We must extract the necessary information again:
  - The chi-square statistic
  - The chi-square p-value
  - The measure of association

<table>
<thead>
<tr>
<th></th>
<th>Values</th>
<th>df</th>
<th>Asymp. Signif. (2-sided)</th>
<th>Exact Sig. (2-sided)</th>
<th>Exact Sig. (1-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Chi-Square</td>
<td>0.007</td>
<td>1</td>
<td>0.234</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continuity Correction</td>
<td>0.000</td>
<td>1</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>0.007</td>
<td>1</td>
<td>0.234</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fisher's Exact Test</td>
<td></td>
<td></td>
<td>1.000</td>
<td>0.576</td>
<td></td>
</tr>
<tr>
<td>Linear-by-Linear Association</td>
<td>0.007</td>
<td>1</td>
<td>0.234</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>302</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 5.65.
- b. Computed only for a 2x2 table.

Symmetric Measures

<table>
<thead>
<tr>
<th>Value</th>
<th>Asymp. Standard Error</th>
<th>Approximate</th>
<th>Approximate Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ordinal by Ordinal Kendall's tau b</td>
<td>0.05</td>
<td>0.058</td>
<td>0.093</td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>302</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Does Gender have an effect on the belief that coral bleaching is a top threat to coral reefs?

Does Gender have an effect on the belief that coral bleaching is a top threat to coral reefs?
- Yes, it does
- The p-value of the Chi-Square test is greater than 0.05, which indicates that there is no statistical relationship between gender and the belief that coral bleaching is a top threat to coral reefs.

**Pasted this table into this presentation with copy > paste “keep source formatting”**
Visualizing Contingency Table/Chi Square Results

*Let's create another “report ready” contingency table based on these results*

<table>
<thead>
<tr>
<th>boat</th>
<th>no</th>
<th>yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>45</td>
<td>7</td>
<td>52</td>
</tr>
<tr>
<td>% within boat</td>
<td>22.4%</td>
<td>13.2%</td>
<td>21.1%</td>
</tr>
<tr>
<td>Does not use a hookline</td>
<td>22.4%</td>
<td>13.2%</td>
<td>21.1%</td>
</tr>
<tr>
<td>Yes, in Cocos Lagoon</td>
<td>78</td>
<td>24</td>
<td>102</td>
</tr>
<tr>
<td>% within boat</td>
<td>38.8%</td>
<td>52.2%</td>
<td>41.3%</td>
</tr>
<tr>
<td>Yes, in Achang Preserve</td>
<td>31</td>
<td>1</td>
<td>32</td>
</tr>
<tr>
<td>% within boat</td>
<td>15.4%</td>
<td>2.2%</td>
<td>13.0%</td>
</tr>
<tr>
<td>Yes, in both places</td>
<td>47</td>
<td>14</td>
<td>61</td>
</tr>
<tr>
<td>% within boat</td>
<td>23.4%</td>
<td>10.4%</td>
<td>24.7%</td>
</tr>
<tr>
<td>Total</td>
<td>201</td>
<td>46</td>
<td>247</td>
</tr>
<tr>
<td>% within boat</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Chi-Square Tests

<table>
<thead>
<tr>
<th>Test</th>
<th>Value</th>
<th>Asymptotic Significance (2-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Chi-Square</td>
<td>8.360*</td>
<td>.039</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>10.467</td>
<td>.015</td>
</tr>
<tr>
<td>Linear-by-Linear Association</td>
<td>2.07</td>
<td>.149</td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>247</td>
<td></td>
</tr>
</tbody>
</table>

Symmetric Measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
<th>Approximate Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal by Nominal: Phi</td>
<td>.184</td>
<td>.039</td>
</tr>
<tr>
<td>Cramer's V</td>
<td>.184</td>
<td>.039</td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>247</td>
<td></td>
</tr>
</tbody>
</table>

Does owning a boat affect where someone uses a hookline?

<table>
<thead>
<tr>
<th>Does not own a boat</th>
<th>Owns a boat</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does not use a hookline</td>
<td>22.4%</td>
<td>15.2%</td>
</tr>
<tr>
<td>Uses a hookline in Cocos Lagoon</td>
<td>38.8%</td>
<td>52.2%</td>
</tr>
<tr>
<td>Uses a hookline in Achang Preserve</td>
<td>15.4%</td>
<td>2.2%</td>
</tr>
<tr>
<td>Uses a hookline in Cocos Lagoon and Achang Preserve</td>
<td>23.4%</td>
<td>30.4%</td>
</tr>
<tr>
<td>Total</td>
<td>201</td>
<td>46</td>
</tr>
</tbody>
</table>

Chi-Square statistic | 8.360 |
Chi-square p-value | 0.039 |
Cramer's V association measure | 0.184 |

- Does owning a boat affect where someone uses a hookline?
  - Yes, it does
  - The p-value of the Chi-Square test is less than 0.05, which indicates that there is a statistical relationship between owning a boat and where someone uses a hookline
  - The Cramer's V statistic of 0.184 indicates that this a relatively “weak,” yet statistically significant, relationship
Making “Report Ready” Contingency Tables: Practice

- Make a report ready table based on this output.

### Does owning fishing or harvesting marine resources affect peoples’ perceptions concerning the condition of the number of fish?

- Yes, it does.
- The p-value of the Chi-Square test is less than 0.05, which indicates that there is a statistical relationship between fishing/harvesting for marine resources and perception of the condition of the number of fish.
- The Tau-b statistic of 0.163 indicates that this a relatively “weak,” yet statistically significant, and positive relationship (i.e. as someone fishes/harvests, they’re more likely to have a more positive perception).

### Making “Report Ready” Contingency Tables: Practice

<table>
<thead>
<tr>
<th></th>
<th>Does not fish or harvest for marine resources</th>
<th>Does fish or harvest for marine resources</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Believes that the condition of the number of fish is very bad</td>
<td>9.6%</td>
<td>5.4%</td>
<td>21</td>
</tr>
<tr>
<td>Believes that the condition of the number of fish is bad</td>
<td>23.0%</td>
<td>8.8%</td>
<td>44</td>
</tr>
<tr>
<td>Believes that the condition of the number of fish is neither good nor bad</td>
<td>18.5%</td>
<td>28.4%</td>
<td>67</td>
</tr>
<tr>
<td>Believes that the condition of the number of fish is good</td>
<td>37.8%</td>
<td>41.2%</td>
<td>112</td>
</tr>
<tr>
<td>Believes that the condition of the number of fish is very good</td>
<td>11.1%</td>
<td>16.2%</td>
<td>39</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>135</td>
<td>148</td>
<td>283</td>
</tr>
</tbody>
</table>

Chi-Square statistic: 15.272
Chi-square p-value: 0.004
Tau-b association measure: 0.163

### Symmetric Measures

<table>
<thead>
<tr>
<th>Value</th>
<th>Asymptotic Standard Error</th>
<th>Approximate Phi</th>
<th>Approximate Phi Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Pearson’s Chi-Square

- None of the 0 cells (0.0%) have expected count less than 5. The minimum expected count is 13.02.

- Total: 283

- **Chi-Square:** 15.272
- **Chi-square p-value:** 0.004
- **Tau-b:** 0.163
Visualizing Paired Samples T-test Results

- This is the output from our first Paired t-test analysis.
- Pasting this directly into a report would be confusing to the reader.
- We must extract the necessary information:
  - The means
  - The mean difference
  - The sample sizes
  - The T statistic (shows direction)
  - The p-value (shows significance)
- Refine our labels.

Did the weight loss treatment lead to a reduction in weight?

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Average weight before weight loss treatment</th>
<th>Average weight after weight loss treatment</th>
<th>Mean Difference</th>
<th>T statistic</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>200.56</td>
<td>187.4</td>
<td>-13.16</td>
<td>5.383</td>
<td>0.000</td>
</tr>
</tbody>
</table>

- Did the weight loss treatment lead to a reduction in weight?
  - Yes, it did.
  - The p-value of the paired t-test is less than 0.05, which indicates that there is a statistically significant difference in the mean weights before and after the treatment.
  - Since the test was “before minus after,” the positive t-statistic of 5.383 indicates that the mean weight before the treatment was significantly greater than the mean weight after the treatment.
Visualizing Paired Samples T-test Results: Practice

- Create a report ready table based off of this output

![Paired Samples Statistics Table]

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std Deviation</th>
<th>Std Error</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Differences</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cholesterol Before</td>
<td>193.44</td>
<td>44.913</td>
<td>5.309</td>
<td>50</td>
</tr>
<tr>
<td>Cholesterol After</td>
<td>170.64</td>
<td>36.811</td>
<td>5.470</td>
<td>50</td>
</tr>
</tbody>
</table>

![Paired Samples Correlations Table]

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Correlation</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cholesterol Before &amp; Cholesterol After</td>
<td>50</td>
<td>.025</td>
<td>.021</td>
</tr>
</tbody>
</table>

![Paired Samples Test Table]

<table>
<thead>
<tr>
<th>Differences</th>
<th>Mean</th>
<th>Std Deviation</th>
<th>Std Error</th>
<th>Lower</th>
<th>Upper</th>
<th>df</th>
<th>Sig (2 tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cholesterol Before &amp; Cholesterol After</td>
<td>12.8</td>
<td>7.091</td>
<td>2.416</td>
<td>-7.44</td>
<td>17.05</td>
<td>5.297</td>
<td>.000</td>
</tr>
</tbody>
</table>

**Did the weight loss treatment lead to a reduction in cholesterol?**

<table>
<thead>
<tr>
<th></th>
<th>Sample Size</th>
<th>Average cholesterol before weight loss treatment</th>
<th>Average cholesterol after weight loss treatment</th>
<th>Mean Difference</th>
<th>T statistic</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
<td>183.44</td>
<td>170.64</td>
<td>-12.8</td>
<td>5.297</td>
<td>0.000</td>
</tr>
</tbody>
</table>

- Did the weight loss treatment lead to a reduction in cholesterol?
  - Yes, it did
  - The p-value of the paired t-test is less than 0.05, which indicates that there is a statistically significant difference in the mean cholesterol levels before and after the treatment
  - Since the test was “before minus after,” the positive t-statistic of 5.297 indicates that the mean cholesterol level before the treatment was significantly greater than the mean cholesterol level after the treatment
Visualizing Independent Samples T-test Results

- This is the output from our first independent samples t-test analysis
- We must extract the necessary information
  - The means
  - The mean difference
  - Refine our labels
  - The sample sizes
  - The CORRECT T statistic (shows direction)
  - The CORRECT p-value (shows significance)

Does college completion have an effect on peoples’ perceptions concerning the condition of the beach/shoreline?

<table>
<thead>
<tr>
<th>Did not complete college</th>
<th>Completed college</th>
<th>Mean difference</th>
<th>T statistic</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition of beach/shoreline</td>
<td>227</td>
<td>2.89</td>
<td>64</td>
<td>3.00</td>
</tr>
</tbody>
</table>

- Does college completion have an effect on peoples’ perceptions concerning the condition of the beach/shoreline?
  - No, it does not
  - The p-value of the independent samples t-test is greater than 0.05, which indicates that there is not a statistically significant difference in the mean perception of beach/shoreline amongst college graduates and the mean perception of beach/shoreline amongst those who have not completed college
Does fishing to sell have an effect on peoples’ agreement with creating areas for traditional fishing only?

<table>
<thead>
<tr>
<th></th>
<th>Does not fish to sell</th>
<th>Does fish to sell</th>
<th>Mean difference</th>
<th>T statistic</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreement with</td>
<td>Sample size</td>
<td>Mean</td>
<td>Sample size</td>
<td>Mean</td>
<td></td>
</tr>
<tr>
<td>creating areas for only</td>
<td>106</td>
<td>3.57</td>
<td>28</td>
<td>3.32</td>
<td>0.245</td>
</tr>
<tr>
<td>traditional fishing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.93</td>
</tr>
</tbody>
</table>

Does fishing to sell have an effect on peoples’ agreement with creating areas for traditional fishing only?

- No, it does not

The p-value of the independent samples t-test is greater than 0.05, which indicates that there is not a statistically significant difference in the mean agreement levels for creating areas for traditional fishing only when comparing those who fish to sell and those who do not fish to sell.
Visualizing one way ANOVA results

- This is the output from our first one way ANOVA analysis
- Pasting this directly into a report would be confusing to the reader
- We must extract the necessary information
  - The means
  - The sample sizes
  - Which groups are significant
  - Refine our labels
  - The p-values

Does where people participate in beach recreation have an effect on their overall perception concerning the change in the condition of marine resources over the last 10 years?

<table>
<thead>
<tr>
<th>(1) Does not participate in beach recreation</th>
<th>(2) Participates in beach recreation in Cocos Lagoon</th>
<th>(3) Participates in beach recreation in Achang Preserve</th>
<th>(4) Participates in beach recreation in both places</th>
<th>Significance Between Groups</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>Mean</td>
<td>Sample Size</td>
<td>Mean</td>
<td>Sample Size</td>
<td>Mean</td>
</tr>
<tr>
<td>Last 10 Years Index</td>
<td>50</td>
<td>56.44</td>
<td>88</td>
<td>46.09</td>
<td>16</td>
</tr>
</tbody>
</table>

** = significant at the 5% level

- Does where people participate in beach recreation have an effect on their overall perception concerning the change in the condition of marine resources over the last 10 years?
  - Yes, it does
  - Those that participate in beach recreation in both places have a more positive perception concerning the change in condition of marine resources when compared to those who participate in beach recreation only in Cocos Lagoon
Visualizing one way ANOVA results: Practice

- Create a report ready table based off of this output

<table>
<thead>
<tr>
<th>Condition index norm</th>
<th>N</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>would not do</td>
<td>4</td>
<td>35.42</td>
</tr>
<tr>
<td>would consider</td>
<td>82</td>
<td>47.59</td>
</tr>
<tr>
<td>would do</td>
<td>152</td>
<td>58.11</td>
</tr>
<tr>
<td>Total</td>
<td>236</td>
<td>54.11</td>
</tr>
</tbody>
</table>

### Multiple Comparisons

<table>
<thead>
<tr>
<th></th>
<th>(1) volunteer Protect</th>
<th>(2) volunteer Protect</th>
<th>(3) volunteer Protect</th>
<th>95% Confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>12.178</td>
<td>11.083</td>
<td>10.844</td>
<td></td>
</tr>
<tr>
<td>Std. Error</td>
<td>3.11</td>
<td>3.11</td>
<td>3.11</td>
<td></td>
</tr>
<tr>
<td>Sig.</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>Lower Bound</td>
<td>-3.82</td>
<td>-3.68</td>
<td>-3.76</td>
<td></td>
</tr>
<tr>
<td>Upper Bound</td>
<td>19.17</td>
<td>18.32</td>
<td>18.72</td>
<td></td>
</tr>
</tbody>
</table>

* The mean difference is significant at the 0.05 level.

---

Does one’s consideration of volunteering at least once a year to help protect coral reefs have an effect on their overall perception of marine resource condition?

- Yes, it does
- Those that “would” volunteer at least once a year to help protect the reefs have a more positive perception concerning the condition of marine resources when compared to those who would “consider” volunteering at least once a year to help protect the reefs
Correlation Matrix – Visualizing Correlations

- This is the output from our first correlation analysis.
- Pasting this directly into a report would be confusing to the reader.
- We must extract the necessary information.
  - The correlation coefficient.
  - Which coefficients are significant.
  - Refine our labels.
  - Get rid of half the values (redundancy).

### Correlation Matrix – Visualizing Correlations

- Copy and paste SPSS output into Excel and work from there.

---

**Note:** Correlation is significant at the 0.01 level (2-tailed).

**.** Correlation is significant at the 0.05 level (2-tailed).

---

### Example Table

<table>
<thead>
<tr>
<th></th>
<th>Fish or Harvests for marine resources</th>
<th>Age</th>
<th>Number of years lived in Merizo</th>
<th>Completed high school</th>
<th>Completed college</th>
<th>Male gender</th>
<th>Number of family members</th>
<th>Owns a home</th>
<th>Annual household income</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fish or Harvests</td>
<td>1.00</td>
<td></td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harvests for marine</td>
<td></td>
<td>-0.003</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>resources</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.019</td>
<td>0.532**</td>
<td></td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of years</td>
<td></td>
<td>-0.062</td>
<td>-0.060</td>
<td>0.154**</td>
<td>0.135**</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lived in Merizo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Completed high school</td>
<td></td>
<td>0.062</td>
<td>-0.040</td>
<td>0.029</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Completed college</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male gender</td>
<td>0.096</td>
<td>0.036</td>
<td>0.072</td>
<td>0.065</td>
<td>0.056</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of family</td>
<td>-0.060</td>
<td>-0.151**</td>
<td>-0.043</td>
<td>-0.038</td>
<td>-0.059</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>members</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Owns a home</td>
<td>-0.056</td>
<td>0.324**</td>
<td>0.397**</td>
<td>0.090</td>
<td>0.064</td>
<td>0.027</td>
<td>-0.117**</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Annual household</td>
<td>-0.128</td>
<td>0.337**</td>
<td>-0.122</td>
<td>0.200</td>
<td>0.698**</td>
<td>-0.027</td>
<td>-0.316**</td>
<td>0.060</td>
<td>1.000</td>
</tr>
<tr>
<td>income</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Correlation Matrix – Visualizing Correlations: Practice

**Make a report ready correlation matrix based off of this output**

### Correlation Matrix – Visualizing Correlations

<table>
<thead>
<tr>
<th></th>
<th>Fishing index</th>
<th>Condition index</th>
<th>Last 10 index</th>
<th>Flood severity index</th>
<th>Flood severity index</th>
<th>Achang Preserve Index</th>
<th>Management efficiency index</th>
<th>Management Success Index</th>
<th>Volunteer index</th>
<th>Environmental Behavior Frequency Index</th>
<th>Management agreement index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fishing index</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition index</td>
<td>0.078</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Last 10 index</td>
<td>0.026</td>
<td>0.828**</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flood severity index</td>
<td>0.401**</td>
<td>0.180**</td>
<td>0.087</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flood severity index</td>
<td>0.257**</td>
<td>0.152</td>
<td>0.022</td>
<td>0.713**</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Achang Preserve Index</td>
<td>0.400**</td>
<td>0.294**</td>
<td>0.216</td>
<td>0.426**</td>
<td>0.282**</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Management efficiency index</td>
<td>0.254**</td>
<td>0.399**</td>
<td>0.391**</td>
<td>0.204*</td>
<td>0.122</td>
<td>0.757**</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Management Success Index</td>
<td>0.208*</td>
<td>0.161*</td>
<td>0.162</td>
<td>0.552**</td>
<td>0.498**</td>
<td>0.412**</td>
<td>0.258**</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Volunteer index</td>
<td>0.186</td>
<td>0.134</td>
<td>0.202**</td>
<td>0.004</td>
<td>0.076</td>
<td>0.432**</td>
<td>0.444**</td>
<td>-0.059</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Environmental Behavior Frequency Index</td>
<td>0.277**</td>
<td>0.272**</td>
<td>0.368**</td>
<td>0.139*</td>
<td>0.130</td>
<td>0.234*</td>
<td>0.256**</td>
<td>0.079</td>
<td>0.384**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Management agreement index</td>
<td>0.191**</td>
<td>0.382**</td>
<td>0.311**</td>
<td>0.257**</td>
<td>0.209**</td>
<td>0.452**</td>
<td>0.381**</td>
<td>0.166</td>
<td>0.319**</td>
<td>0.371**</td>
<td>1.000</td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level

* Correlation is significant at the 0.05 level
Visualizing Regression Results

• This is the output from one of our multiple regression analyses
• Pasting this directly into a report would be confusing to the reader
• We must extract the necessary information
  • The adj R square
  • The beta coefficients
  • The T statistics (for direction)
  • The p-values (for significance)
  • Refine our labels

```
Visualizing Regression Results

<table>
<thead>
<tr>
<th>Dependent Variable: Management Agreement Index</th>
<th>Rsquare adj = 0.213</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent Variable</td>
<td>Coefficient</td>
</tr>
<tr>
<td>Constant</td>
<td>17.658</td>
</tr>
<tr>
<td>Environmental Behavior Frequency Index</td>
<td>0.261***</td>
</tr>
<tr>
<td>Last 10 Index</td>
<td>0.317***</td>
</tr>
<tr>
<td>Amount of years lived in Merizo</td>
<td>0.040</td>
</tr>
</tbody>
</table>

*** = significant at the 1% level
```

- All else held equal, as people participate in environmental behavior more frequently, they are more supportive of management options
- All else held equal, as people have a more positive perception concerning the change in the condition of marine resources, they are more supportive of management
Visualizing Regression Results: Practice

- Create a report ready regression results table based off of this output

### Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.384</td>
<td>.055</td>
<td>.055</td>
<td>.085</td>
</tr>
</tbody>
</table>

*a Predictors: (Constant), benefit, Management agreement index, norm, college, age*

### Coefficients

<table>
<thead>
<tr>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(Constant)</strong>*</td>
<td>0.465</td>
<td>3.062</td>
<td>.003</td>
</tr>
<tr>
<td>Age</td>
<td>-0.002</td>
<td>-0.564</td>
<td>.574</td>
</tr>
<tr>
<td>College</td>
<td>-0.245***</td>
<td>-2.212</td>
<td>.029</td>
</tr>
<tr>
<td>Management agreement index</td>
<td>0.005**</td>
<td>0.055</td>
<td></td>
</tr>
<tr>
<td>Receives benefit from Achang Preserve</td>
<td>-0.139</td>
<td>-1.296</td>
<td>.197</td>
</tr>
</tbody>
</table>

*a Dependent Variable: program CRCP*

### Visualizing Regression Results: Practice

**Dependent Variable: Familiarity with CRCP**  
\[ R^2 = 0.055 \]

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Coefficient</th>
<th>T Statistic</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.465</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.002</td>
<td>-0.564</td>
<td>0.574</td>
</tr>
<tr>
<td>Completed College</td>
<td>-0.245**</td>
<td>-2.212</td>
<td>0.029</td>
</tr>
<tr>
<td>Management Agreement Index</td>
<td>0.005**</td>
<td>2.239</td>
<td>0.027</td>
</tr>
<tr>
<td>Receives benefit from Achang Preserve</td>
<td>-0.139</td>
<td>-1.296</td>
<td>0.197</td>
</tr>
</tbody>
</table>

\* ** = significant at the 5% level

- All else held equal, those who completed college were less likely to be familiar with CRCP
- All else held equal, those who are more supportive of management were more likely to be familiar with CRCP
9.1 True or False: It is acceptable to simply copy and paste SPSS output into a report

A. True
B. False
9.2 What should we be reporting in a “report ready” contingency table?

A. The column percentages
B. The total frequencies
C. The chi-square statistic
D. The chi-square p-value
E. The measure of association
F. All of the above

9.3 What is wrong with this correlation matrix?

<table>
<thead>
<tr>
<th></th>
<th>Fish or Harvests for marine resources</th>
<th>Age</th>
<th>Number of years lived in Merizo</th>
<th>Completed high school</th>
<th>Completed college</th>
<th>Male gender</th>
<th>Number of family members</th>
<th>Owns a home</th>
<th>Annual household income</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fish or Harvests for marine resources</td>
<td>1</td>
<td>-0.003</td>
<td>-0.019</td>
<td>0.002</td>
<td>-0.000</td>
<td>0.095</td>
<td>-0.056</td>
<td>-0.128</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.003</td>
<td>1</td>
<td>0.532**</td>
<td>-0.040</td>
<td>0.014**</td>
<td>0.038</td>
<td>-0.151**</td>
<td>0.324*</td>
<td>0.337**</td>
</tr>
<tr>
<td>Number of years lived in Merizo</td>
<td>-0.019</td>
<td>-0.532**</td>
<td>1</td>
<td>0.029</td>
<td>-0.038</td>
<td>0.072</td>
<td>-0.015</td>
<td>0.397**</td>
<td>-0.122</td>
</tr>
<tr>
<td>Completed high school</td>
<td>0.002</td>
<td>-0.154</td>
<td>0.029</td>
<td>1</td>
<td>0.035</td>
<td>-0.043</td>
<td>0.090</td>
<td>0.200</td>
<td></td>
</tr>
<tr>
<td>Completed college</td>
<td>-0.090</td>
<td>-0.035</td>
<td>-0.038</td>
<td>1</td>
<td>0.095</td>
<td>-0.030</td>
<td>0.064</td>
<td>0.697</td>
<td></td>
</tr>
<tr>
<td>Male gender</td>
<td>0.006</td>
<td>0.036</td>
<td>0.072</td>
<td>0.065</td>
<td>0.056</td>
<td>1</td>
<td>-0.059</td>
<td>0.027*</td>
<td>-0.027*</td>
</tr>
<tr>
<td>Number of family members</td>
<td>0.003</td>
<td>-0.151</td>
<td>-0.015</td>
<td>-0.043</td>
<td>-0.030</td>
<td>-0.059</td>
<td>1</td>
<td>-0.317*</td>
<td>-0.316*</td>
</tr>
<tr>
<td>Owns a home</td>
<td>-0.006</td>
<td>-0.324*</td>
<td>-0.367*</td>
<td>0.090</td>
<td>0.064</td>
<td>0.027</td>
<td>-0.171</td>
<td>1</td>
<td>0.065*</td>
</tr>
<tr>
<td>Annual household income</td>
<td>-0.128</td>
<td>-0.337**</td>
<td>-0.122</td>
<td>0.200</td>
<td>-0.027</td>
<td>-0.316**</td>
<td>0.060</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

** Correlation is significant at the 0.01 level
* Correlation is significant at the 0.05 level
9.4 What is missing from this regression table to make it “report ready”?

<table>
<thead>
<tr>
<th>Dependent Variable: Familiarity with CRCP</th>
<th>Rsquare adj = 0.055</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent Variable</td>
<td>Coefficient</td>
</tr>
<tr>
<td>Constant</td>
<td>0.465</td>
</tr>
<tr>
<td>Age</td>
<td>-0.002</td>
</tr>
<tr>
<td>Completed College</td>
<td>-0.245**</td>
</tr>
<tr>
<td>Management Agreement Index</td>
<td>0.005**</td>
</tr>
<tr>
<td>Receives benefit from Achang Preserve</td>
<td>-0.139</td>
</tr>
</tbody>
</table>

** = significant at the 5% level

9.5 When is it suitable to use nominal data in a regression model?

A. When the dependent variable is nominal
B. When the independent variable is nominal
C. When both the dependent and independent variables are nominal
D. Never
Day 6

- Multiple Response
- Non-Parametric Tests
- Recap Qualitative vs. Quantitative
- Best Practices
Multiple Response Analysis

Day 6: September 17, 2016

For “ranking” questions
In our Manell-Geus questionnaire, respondents were asked to pick the top 3 threats to coral reefs as they perceive them
- Since it was not specified to “rank,” only to “choose 3,” we coded each threat as a different variable and coded as a yes or no denoting whether the respondent chose this threat as a “top” threat
What if respondents are asked to rank?

- If respondents are asked to rank their choices, the order matters
- Open the file “Data for multiple responses.sav”

4. From a **job perspective**, what are your top 2 **main expectations** from the training course?
   1. 
   2. 

Defining Variable Sets
Frequencies of Multiple Responses

Output

- The "percent" column represents the percentage of people that identified that particular expectation as a top one in terms of how many expectations were identified (rank 1 and rank 2)
  - 69 respondents identified 129 expectations
  - Ex: 2 people identified "improve aquaculture" as a top expectation (2 "improve aquaculture" identifications divided by 129 total identifications yields 1.6%)
- The "percent of cases" column represents the percentage of respondents that identified that particular expectation as a top one (regardless of rank)
  - Ex: 2 people identified "improve aquaculture" as a top expectation (2 people divided by 69 total people yields 2.9%)
Normality and Non Parametric Tests

Day 6: September 17, 2016

The Central Limit Theorem

- Given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined (finite) expected value and finite variance, will be approximately normally distributed, regardless of the underlying distribution.

- In English:
  - If our sample size is less than 30, we must test for normality
    - If normal, we can do parametric stats (t test, ANOVA, regression, etc.)
    - If not normal, we use non parametric stats (coming up...)
  - If our sample size is greater than 30, we can assume the data to be normal
Assessing Normality in SPSS

Testing for Normality in SPSS

- Transfer the variable that needs to be tested for normality into the “Dependent” Box
  - Let’s examine “tenure”
- Click on “Plots” and check the box for normality tests
**Interpreting Normality Test Output**

- **WE ONLY NEED TO RUN THESE TESTS IF N<30**
- This is merely an example to show you the output and how to interpret
- The Shapiro-Wilk Test and the Kolmogorov-Smirnov Test are both normality tests
- Null hypothesis of both = normal
- If p-value > 0.05, our data is normal
- If p-value < 0.05, our data is not normal
- However, our sample size is >30, so we don’t need to worry about this

<table>
<thead>
<tr>
<th>Tests of Normality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kolmogorov-Smirnov*</td>
</tr>
<tr>
<td>Statistic</td>
</tr>
<tr>
<td>tenure</td>
</tr>
</tbody>
</table>

*a. Lilliefors Significance Correction*

---

**What to do if data is not normal?**

- Use non-parametric tests
  - Non parametric tests do not assume normality
- Types of non-parametric tests:
  - Wilcoxon signed-rank test (non parametric equivalent of paired samples t test)
  - Mann-Whitney test (non parametric equivalent of independent samples t test)
  - Kruskall Wallis test (non parametric equivalent of one way ANOVA)
  - All of these can be performed in SPSS
Wilcoxon signed-rank test in SPSS

Open “Paired T-test Example.sav”
Wilcoxon signed-rank test in SPSS

- P-value = 0.000 < 0.05
- The median “Cholesterol After” is significantly less than the median “Cholesterol before”
- There has been a statistically significant reduction in cholesterol since the weight loss treatment

Mann-Whitney test

Let’s test to see if “gender” has an effect on overall perception concerning the condition of marine resources “Condition Index_norm”
Mann-Whitney test

- P-value = 0.381 > 0.05
- The medians are not statistically different
- There is no statistical relationship between gender and overall perception of condition of marine resources

Kruskall Wallis Test

Let’s test to see if education level (“education”) has an effect on overall agreement with management options (“Management Agree Index_norm”)
Kruskall Wallis Test

Kruskall Wallis Test Output

- The output tells us that there are differences in median overall management agreement amongst differing levels of education.....but which groups specifically??
- Let’s double click on our “hypothesis test summary” table to open “model viewer”
  - We want to examine the “pairwise comparisons” under “view” near the bottom left
Kruskall Wallis Test Output

- Each node (circle) represents the median management agree index score for each education level.
- If 2 nodes are connected by a yellow line, there is a significant difference (95% confidence level).
- Some community college median = 60.91
- 9th-11th grade median = 42.73
- Those who have completed some community college or vocational training exhibit statistically significantly more agreement with management options when compared to those who have only completed 9th-11th grade.
Best Practices and Ethics in Data Analysis and Management

Day 6: September 17, 2016

Why best practices?

- Research with people has ethical and moral consequences.
- Results are used in decision making that impacts people. This means we need accurate, unbiased evidence for decision making.
- No good reasons not to use best practices.
Ethical principles in conducting studies with human subjects

- Respect local culture and rights – “Free, prior and informed consent” (FPIC)
- No harm may be done to the participants – anonymity and confidentiality
- Justice--subjects should not be selected based on a compromised or manipulated position

More....

- Ensuring full, effective stakeholder participation wherever possible and appropriate
- Supporting transparency and accountability, disclosing and sharing information with stakeholders in a locally appropriate manner.
Confidentiality and anonymity

- Respect the privacy of the interviewees and make sure that information they give to you as a researcher does not cause them harm in any way.
- Have the data under good control. Do not leave transcripts, pictures, videotapes or whatever you are working with lying about in public.
- Remove personal identification as early as possible. If you don’t need these data, don’t collect them.
- Do not make unnecessary copies and keep good track of the location of all copies (in both electronic and other formats).
- Do not hand your material to anyone without going over the handling procedures.

Best practices in quantitative data analysis

- Develop and follow code book strictly.
  - Be consistent with coding
  - Update codebook regularly
- Use . (dot) for missing data
- Use accurate and clean data for analysis
- Know the level of your data
  - Begin with more detailed level of data
  - Keep continuous data when possible, category them later
- In descriptive stats, check on range, SD, mode and median
- Use inferential stats with representative samples from random sampling
- Apply $p$-value $\leq .05$, confidence level and confidence interval in inferential stats
Best practices in qualitative data analysis

- Start analyzing as data collecting is taking place
- Determine sample size by the diminishing rate of return
- Record and transcribe notes when possible
- Use coding to bring together the similar ideas, concepts, or themes that have been discovered.
  - Start with descriptive coding,
  - Then move on to analytical coding, constructing themes and categories,
  - revisit them on an on-going basis
- Go back and forth between deductive and inductive reasoning
- selection of quotes to support the presentation of the findings.
- Triangulate to ensure thrustworthyness of data
- Use trained qualitative data analyst. Good common sense is not enough!

Condensation of qualitative data

Qualitative data analysis is a process of condensation in which a vast amount of data has to be condensed in a meaningful way both theoretically and generally. Three areas to watch out for:

- **Drifting**, which means that the results are poorly rooted in the original data.
- **Dumping**, which means that the results are simply not based on the data and at best present an oversimplified picture.
- **Data drowning**, which means that too much data has been collected and the researcher fails to get any meaningful grip on the data.

Source: http://www.lse.ac.uk/media@lse/research/EUKidsOnline/BestPracticeGuide/FAQ31.aspx
Best practices in data analysis (continued)

- Well-designed assessment (appropriate methods, level of data, mixed complementing methods)

- Do NOT impose the ideas of quantitative analysis on qualitative data. If generalizability is what you want, use quantitative methods.

Resources for Ethnical Principles

University of Hawaii’s Policy and Guidance on Human Studies
https://www.hawaii.edu/researchcompliance/policies-guidance

Qualitative Vs. Quantitative: When to Use?

Day 6: September 17, 2016

Many data collecting methods for socioeconomic data

- Secondary data
- Surveys
- Interviews
- Focus groups
- Participatory/Rapid rural appraisal
- Visualization techniques
- (Participation) observation
Which methods to use depends on....

- Your research questions
- Your subjects
- Your resources and capacity
- Your intended use and audiences

What is a survey?

Way to collect data, usually from a relatively large group of people who are randomly selected to be included in the sample.

A questionnaire is used, with highly structured, mostly close-ended questions.
When to use a survey

- Researcher has clearly specified research questions
- When quantitative data are required
- When topics are relatively simple/straightforward
- Need to understand perception, attitude, opinion and knowledge of a large group of people.
- To generate data that are statistically representative of the larger population
- When you want comparisons between groups and examine correlations between variables.

Advantages of surveys

- Can cover large population in short time
- Researchers have control
- Precision through standardized questions and interview process
- Statistical significance
- Generate short answers that can be coded and analyzed quickly and easily by statistical software programs
Disadvantages of Survey

- **Inflexible**, does not allow for possibility to further explore a topic that may come up
- Does **not uncover important** information that researchers are not aware of or important **for the interviewee**
- Does **not provide deeper context** of an issue, not appropriate for complex topics
- Can be perceived as being ‘artificial’
- Unless the research is longitudinal, the respondents will be **interviewed only once**.

What is interview?

Loosely structured conversation with people who have specialized knowledge about the topic you wish to understand.

Key informants can provide useful information regarding a larger population or group.
When to conduct interviews?

- Explore a subject in depth
- Uncover and understand issues that may not be clear to the researchers, explore topics for further research
- Emphasis on interviewee’s perspectives and understanding the topics in their own terms
- Gather information in socio-cultural setting where a survey or focus groups are inappropriate
- Clarify findings from quantitative research. Explain the “why” and “how”

Advantages of Interviews

- Flexible responding to interviewee, exploring significant issues that emerge during interview
- Provide local context in local terms
- Can be done early in the process and help gain information to improve research design
- Greater depth and detailed information
- Can be conducted with same person more than once
Disadvantages of Interviews

- Interviewer-induced bias, different interviewers may get different results
- Can give misleading or biased information
- Require high interviewing skills
- Does not provide statistical validity
- Time consuming transcription of interviews
- Not easy to analyze qualitative data—needs good synthesizing and summarizing skills

What Does the Literature Say?

Revisiting the Quantitative-Qualitative Debate: Implications for Mixed-Methods Research

JOANNA E. M. SALE
Institute for Work & Health; Health Research Methodology Program, Department of Clinical Epidemiology & Biostatistics, McMaster University

LYNNE H. LOHFELD
St. Joseph’s Hospital and Home; Department of Clinical Epidemiology & Biostatistics, McMaster University

KEVIN BRAZIL
St. Joseph’s Health Care System Research Network, St. Joseph’s Community Health Centre; Department of Clinical Epidemiology & Biostatistics, McMaster University

Abstract. Health care research includes many studies that combine quantitative and qualitative methods. In this paper, we revisit the quantitative-qualitative debate and review the arguments for and against using mixed-methods. In addition, we discuss the implications stemming from our view that the paradigms upon which the methods are based have a different view of reality and therefore a different view of the phenomenon under study. Because the two paradigms do not study the same phenomena, quantitative and qualitative methods cannot be combined for cross-validation or triangulation purposes. However, they can be combined for complementary purposes. Future standards for mixed-methods research should clearly reflect this recommendation.
What Does the Literature Say?

- The quantitative paradigm is based on positivism
  - All phenomena can be reduced to empirical indicators which represent the truth
  - The position of the quantitative paradigm is that there is only one truth, an objective reality that exists independent of human perception
  - NUMBERS

- The qualitative paradigm is based on interpretivism and constructivism
  - There are multiple realities or multiple truths based on one’s construction of reality
  - Reality is socially constructed and so is constantly changing
  - DESCRIPTIONS

(Sale et al. 2002)

What Does the Literature Say?

- The underlying assumptions of the quantitative and qualitative paradigms result in differences which extend beyond philosophical and methodological debates

- The two paradigms have given rise to different journals, different sources of funding, different expertise, and different methods

- There are even differences in scientific language used to describe them

(Sale et al. 2002)
What Does the Literature Say? – Mixed Methods

Maps, Numbers, Text, and Context: Mixing Methods in Feminist Political Ecology*

Dianne Rocheleau
Clark University

Feminist post-structuralist theory, feminist empiricism, and field practice can all contribute to insights on the value of quantitative and qualitative methods in feminist geographical research. A political ecology study of gendered interests in a social forestry program in the Dominican Republic illustrates the methodological dilemmas and potentials of feminist research on environmental change. The study combined qualitative and quantitative data collection and analytical techniques. Examples from the case study address three methodological questions in feminist geography: (1) Should identity or affinity be the basis for situating ourselves and the subjects of our research? (2) How can we reconcile multiple subjectivities and quantitative methods in the quest for objectivity? and (3) Can we combine traditional positivist methods with participatory mapping and oral histories? The paper draws on theoretical literature as well as field experience to answer these questions. Key Words: feminist, gender, qualitative methods, political ecology.
What Does the Literature Say? – Mixed Methods

- The work of interpretative scholars and the "turn toward discourse" (Peet and Watts 1993) have opened a new space for the combination of traditional positivist methods such as resource mapping from remotely sensed data and questionnaire surveys about resource use and management with personal life histories, oral histories, text analysis, landscape interpretation, and participatory mapping methods.
- Simply put, the intersection of qualitative and quantitative data has become more prevalent in modern times.
- Today's researchers must be well-versed in both as many social science projects utilize both techniques.

Summary

- Qualitative
  - If we are investigating an issue that elicits a wide range of opinions and deep knowledge
  - Usually a semi-structured discussion guide to ensure that all topics under consideration are covered and that the discussion stays relevant
    - Questioning is open and participants are encouraged to explore the reasons for their responses
- Quantitative
  - If we are investigating an issue that has measurable units
  - Usually a structured questionnaire with mostly closed questions (i.e. the respondents select their answers from given lists of possible responses)
  - Because of its statistical nature, sample size is important for quantitative research
    - 30 is generally held to be the minimum number of responses for any area of interest although a larger sample size will produce more reliable data.
Summary – Mixed Methods

Mixed Methods

- Allows for the use of both qualitative and quantitative practices
- When is mixed methods suitable?
  - When qualitative research or quantitative research is insufficient to fully understand the problem
  - When we need different, multiple perspectives, or a more complete understanding
  - Example: Need to evaluate the success of a program by using a needs assessment AND a test of the success of the program

(Creswell 2013)

Number of Dissertations and Theses with “Mixed Methods” in the Title

<table>
<thead>
<tr>
<th>Year Range</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005-2009</td>
<td>2524</td>
</tr>
<tr>
<td>2000-2004</td>
<td>532</td>
</tr>
<tr>
<td>1995-1999</td>
<td>100</td>
</tr>
<tr>
<td>1990-1994</td>
<td>26</td>
</tr>
<tr>
<td>1985-1989</td>
<td>17</td>
</tr>
<tr>
<td>1980-1984</td>
<td>3</td>
</tr>
</tbody>
</table>

(Haines 2011)
Answer Key for Quizzes
Socioeconomic Data Analysis Workshop, Palau
September 12-17, 2016

DAY 1
Quiz 1
Question 1.1: A, D
Question 1.2: A
Question 1.3: B
Question 1.4: B
Question 1.6: B
Question 1.7: B

Quiz 2:
Question 2.1: D
Question 2.2: B
Question 2.3: C
Question 2.4: C
Question 2.5: B
Question 2.6: A

DAY 2
Quiz 3
Question 3.1: D
Question 3.2: B
Question 3.3: A
Question 3.4: A, B, E
Question 3.5: A

Quiz 4
Question 4.1: B, C, F
Question 4.2: B
Question 4.3: A
Question 4.4: B
Question 4.5: B

DAY 3
Quiz 5
Question 5.1: A
Question 5.2: D
Question 5.3: C
Question 5.4: B
Question 5.5: B

DAY 3
Quiz 6
Question 6.1: B
Question 6.2: A
Question 6.3: D
Question 6.4: B
Question 6.5: B

DAY 4
Quiz 7
Question 7.1: B
Question 7.2: D
Question 7.3: A
Question 7.4: A
Question 7.5: C

DAY 5
Quiz 8
Question 8.1: D
Question 8.2: A
Question 8.3: C
Question 8.4: B
Question 8.5: B

Quiz 9
Question 9.1: B
Question 9.2: F
Question 9.3: Redundancy in reporting; don’t need values on each side of the “1s”, only need values on one side of the “1s”
Question 9.4: the p-values for the independent variables
Question 9.5: D