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

PIMPAC

NOAA CORAL REEF CONSERVATION PROGRAM

MICRONESIA CONSERVATION TRUST

mina Micronesia Islands Nature Alliance

Micronesia Challenge

Palau International Coral Reef Center

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 OF CONTENTS

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 form 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

3	Question Number	Question	Variable Name	Answer Options	Code
106	**Please answer yes or no if you or your family do these activities, and whether you do them in Achang, or in both Achang and the Cocoos lagoon areas**				
107	13.1	Gathering of animals from the reef (ex. trochus (ailingling), clams (hima), sea cucumbers (balate), octopus (gamson))	activity_gather	no	1
108	13.2	Fishing (ask the following fishing methods only if they fish)	activity_fish	yes, in Cocos Lagoon	2
109	13.3	Spear fishing	activity_spear	yes, in Achang preserve	3
110	13.4	13.4 Cast net-fishing (talaya)	activity_castnet	yes, in both places	4
111	13.5	Gillnet, surround net and drag net-fishing (tekken, chenchulu)	activity_gillnet	not sure	8

- 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

Data Type Examples from Manell-Geus Questionnaire

- Ordinal Variable

3	Question Number	Question	Variable Name	Answer Options	Code
49	**To what extent do you agree with each of the following statements?**				
50	7.1	Coral reefs protect Guam from coastal/shoreline erosion and natural disasters like typhoons and tsunamis	agreement_protect	Strongly Disagree	1
51	7.2	Diving and snorkeling are not harmful to coral reefs.	agreement_divesnork	Disagree	2
52	7.3	Coral reefs provide sustainable resources that support the development of our Merizo communities.	agreement_resources	Neither agree nor disagree	3
53	7.4	Coral reefs have an important role in our culture	agreement_culture	Agree	4
54	7.5	Coral reefs are important to my family's way of life	agreement_life	Strongly Agree	5
55	7.6	Effects from climate change can severely affect coral reefs.	agreement_climate	Not Sure	8

- 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

Data Type Examples from Manell-Geus Questionnaire

- Interval Variable

Question Number	Question	Variable Name	Answer Options	Code
38	What is your annual household income?	income	Under $10,000	1
			$10,000 to $19,999	2
			$20,000 to $29,999	3
			$30,000 to $39,999	4
			$40,000 to $49,999	5
			$50,000 to $59,999	6
			$60,000 to $74,999	7
			$75,000 to $99,999	8
			$100,000 to $149,999	9
			$150,000 or More	10
			Refused to answer	99

- 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

3	Question Number	Question	Variable Name	Answer Options	Code
21	3	Of the seafood that you and your family eat, how much of it comes from Merizo?	Percent_Merizo	Continuous (percentage from 0-100)	0-100

- Question #3 is coded as a ratio variable because it asks the respondent to specify the percentage of their family's seafood that comes form 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?

Accuracy without precision

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

Precision without accuracy

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

Intro to Excel and Data Entry

Day 1: Monday September 12, 2016

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

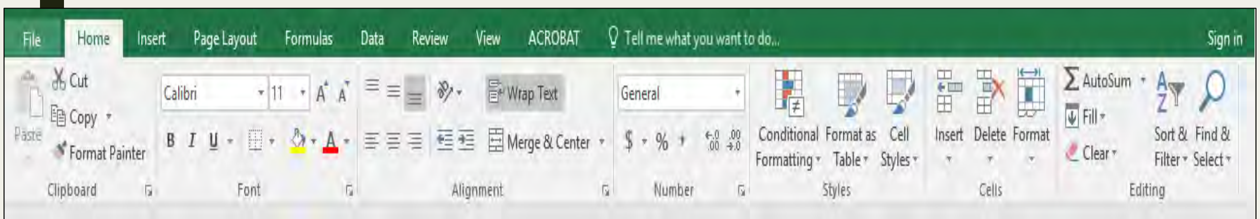
Excel Basics

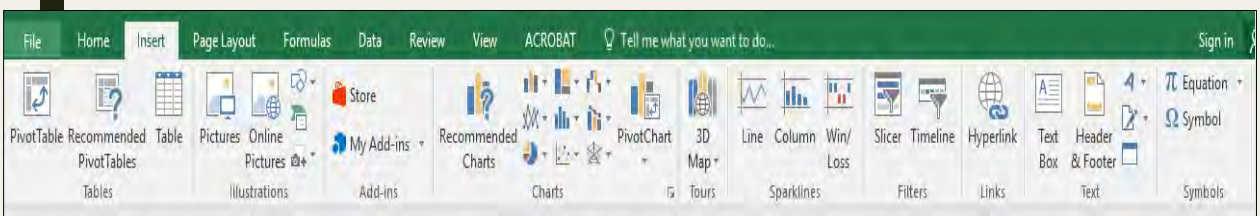
- Multiple worksheet tabs within one file
- Cells can contain numbers, text, or functions
- Basic Word text editing and formatting
- Can sort, rearrange, cut/copy/paste, drag

Getting Started with Spreadsheets

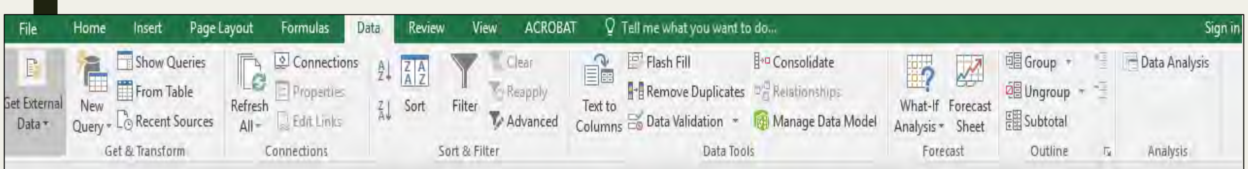
- Open the file “Manell_Geus_Data_GettingStarted.xlsx”
- Data table = grid of rows and columns for organizing data
- Every ROW contains a unique record (case)
  - *ROWS are identified by numbers (1, 2, 3, etc.)*
- Every COLUMN contains a different attribute (variable) that relates to the records
  - *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

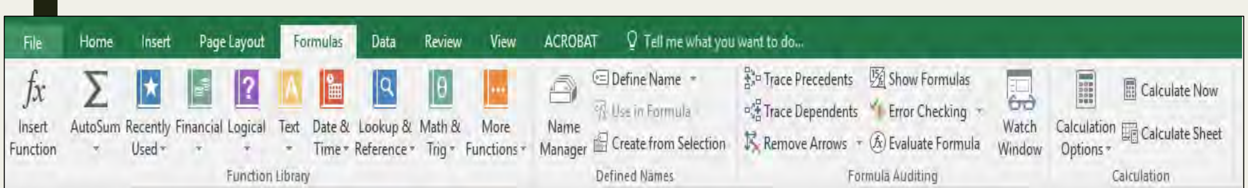
Ribbons – Home and Insert

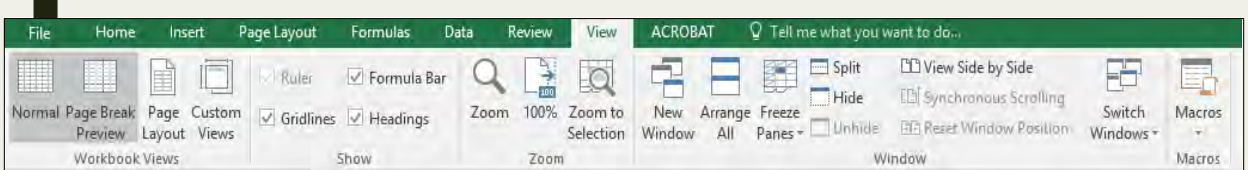




Ribbons – Data, Formulas, and View







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

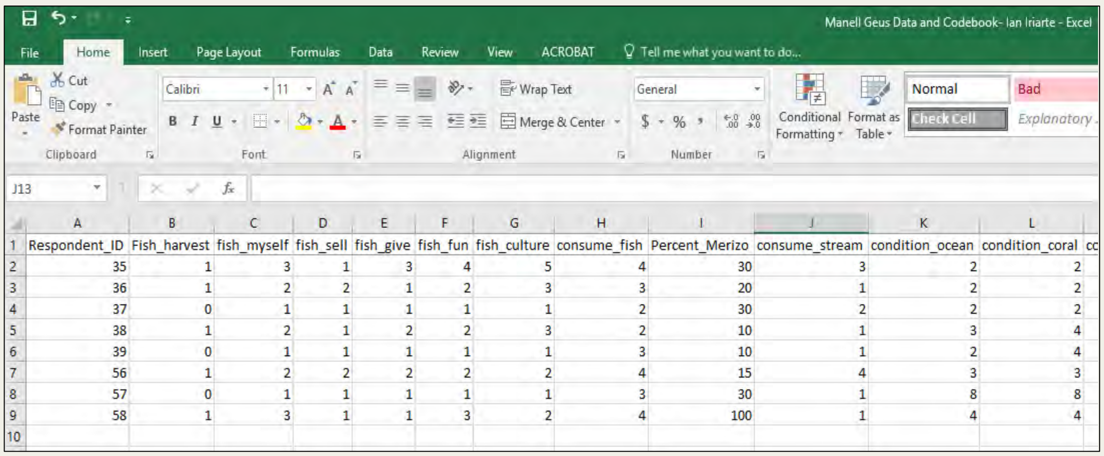
- Correct data entry will set you up for success in analysis later
  - *Quality control is continuous*
    - Checking ranges and value lists
    - Finding data in large tables
    - Narrowing to a subset
    - Checking for typos

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



Respondent_ID	Fish_harvest	fish_myself	fish_sell	fish_give	fish_fun	fish_culture	consume_fish	Percent_Merizo	consume_stream	condition_ocean	condition_coral
35	1	3	1	3	4	5	4	30	3	2	2
36	1	2	2	1	2	3	3	20	1	2	2
37	0	1	1	1	1	1	2	30	2	2	2
38	1	2	1	2	2	3	2	10	1	3	4
39	0	1	1	1	1	1	3	10	1	2	4
56	1	2	2	2	2	2	4	15	4	3	3
57	0	1	1	1	1	1	3	30	1	8	8
58	1	3	1	1	3	2	4	100	1	4	4

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

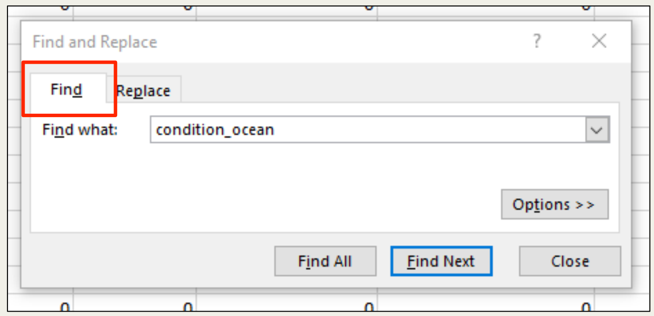
- Arrows move through cells one at a time;
  Tab between columns;
  Enter between rows
- CTRL+arrow jumps to end of row or column
- SHIFT+arrow selects the current and adjacent cell
- CTRL+SHIFT+arrow selects to the end of the row or column
- CTRL+Z undoes last move
- CTRL+mouse click copies non-connected 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 EQUSL SIGN (=)
- =SUM/SUMIF
- =AVERAGE
- =COUNT/COUNTIF
- =MAX
- =MIN
- =MEDIAN
- =IF, =OR, =AND
- =VLOOKUP

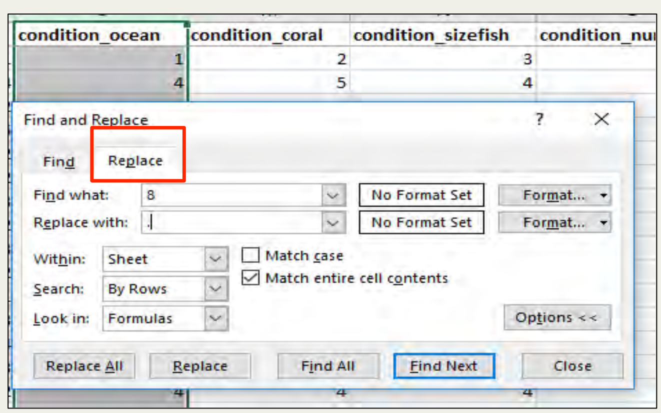
Find and Replace Example

- 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
  - *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
  - *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

GR	GS
mng_commfish	mng_commfish_SA
2	0
2	0
3	0
2	0
2	0
4	0
4	0
.	.
2	0
5	1
2	0
2	0
2	0
3	0
5	1
3	0
4	0
4	0
5	1
4	0
4	0
5	1

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

GR	GS	GT
mng_commfish	mng_commfish_SA	mng_commfish_A or SA
2	0	0
2	0	0
3	0	0
2	0	0
2	0	0
4	0	1
4	0	1
.	.	.
2	0	0
5	1	1
2	0	0
2	0	0
2	0	0
3	0	0
5	1	1
3	0	0
4	0	1
4	0	1
5	1	1
4	0	1
4	0	1
5	1	1
4	0	1

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

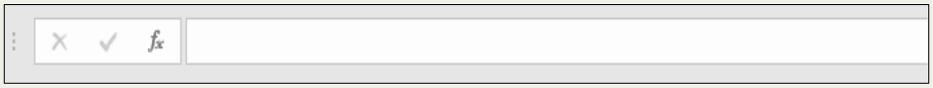
E	F	G	H	I
fish_sell	fish_give	fish_fun	fish_culture	fish_funANDsell
.	.	.	.	.
.	.	.	.	.
3	3	3	3	1
.	.	.	.	.
.	.	.	.	.
1	2	3	2	0
.	.	.	.	.
.	.	.	.	.
2	3	3	3	1
1	3	2	1	0
2	3	2	4	1
.	.	.	.	.
3	2	3	3	1
3	2	4	.	1
1	2	2	2	0
1	3	3	3	0
3	2	2	3	1
3	2	2	2	1
2	3	3	2	1
3	2	2	2	1
.	.	.	.	.

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

J	K
consume_fish	consume_fish_text
5	At least once a week
6	Almost daily
4	A few times a month
6	Almost daily
4	A few times a month
2	A few times a year
2	A few times a year
2	A few times a year
4	A few times a month
3	Once a month
2	A few times a year
1	Almost never/never
5	At least once a week
5	At least once a week
2	A few times a year
3	Once a month
4	A few times a month
3	Once a month
3	Once a month
3	Once a month
6	Almost daily
3	Once a month
6	Almost daily

Formula Bar

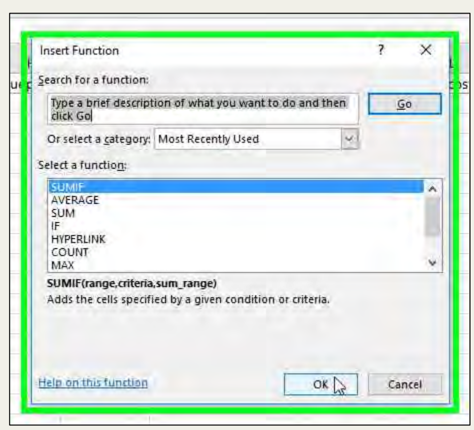


- The formula bar is just above the column names (A, B, C, etc.)
- Formulas can be directly typed into here

Formula Bar

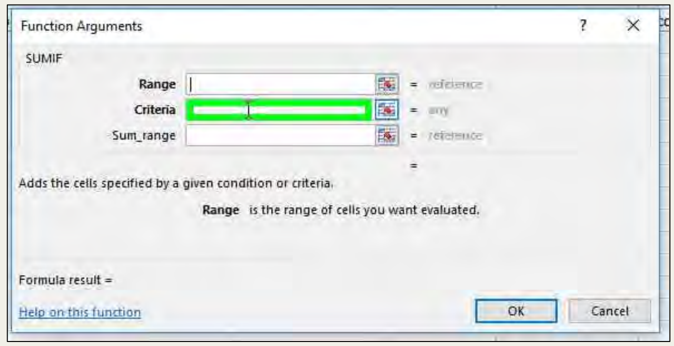
- If you don't want to type directly into the formula bar, you can click the Function button:

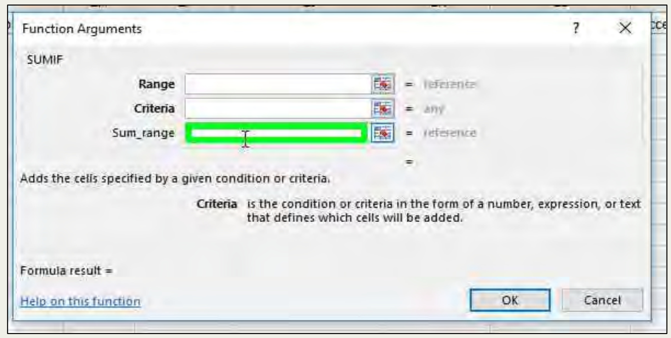
- After you click, the window to the right will open so that you can search and chose the type of function you need



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]

3	Question Number	Question	Variable Name	Answer Options	Code
21	3	Of the seafood that you and your family eat, how much of it comes from Merizo?	Percent_Merizo	Continuous (percentage from 0-100)	0-100

- 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 form Merizo and has a true zero point.

Ordinal Variable

3	Question Number	Question	Variable Name	Answer Options	Code
49	To what extent do you agree with each of the following statements?				
50	7.1	Coral reefs protect Guam from coastal/shoreline erosion and natural disasters like typhoons and tsunamis	agreement_protect	Strongly Disagree	1
51	7.2	Diving and snorkeling are not harmful to coral reefs.	agreement_divesnork	Disagree	2
52	7.3	Coral reefs provide sustainable resources that support the development of our Merizo communities.	agreement_resources	Neither agree nor disagree	3
53	7.4	Coral reefs have an important role in our culture	agreement_culture	Agree	4
54	7.5	Coral reefs are important to my family's way of life	agreement_life	Strongly Agree	5
55	7.6	Effects from climate change can severely affect coral reefs.	agreement_climate	Not Sure	8

- 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

Nominal Variable

3	Question Number	Question	Variable Name	Answer Options	Code
106	Please answer yes or no if you or your family do these activities, and whether you do them in Achang, or in both Achang and the Cocoos lagoon areas				
107	13.1	Gathering of animals from the reef (ex. trochus (ailingling), clams (hima), sea cucumbers (balate), octopus (gamson))	activity_gather	no	1
108	13.2	Fishing (ask the following fishing methods only if they fish)	activity_fish	yes, in Cocos Lagoon	2
109	13.3	Spear fishing	activity_spear	yes, in Achang preserve	3
110	13.4	13.4 Cast net-fishing (talaya)	activity_castnet	yes, in both places	4
111	13.5	Gillnet, surround net and drag net-fishing (tekken, chenchulu)	activity_gillnet	not sure	8

- 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

Binary Variable

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

1. Yes   2. No

3	Question Number	Question	Variable Name	Answer Options	Code
136				yes	1
137	15	Does your household benefit from the Achang Marine Preserve?	benefit	no	0
138				not sure	8

- 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

3	Question Number	Question	Variable Name	Answer Options	Code
226	**How often do members of your household do following activities to help protect the environment?**				
227	23.1	Pick up trash	enviro_trash	never	1
228	23.2	Community fire watch	enviro_fire	once a year	2
229	23.3	Report fires	enviro_report	a few times a year	3
230	23.4	Plant trees or native plants to prevent erosion	enviro_trees	Once a month	4
231	23.5	Attend local education/awareness initiatives	enviro_education	weekly	5
232	23.6	Remove abandoned fishing gear on the reef or beach	enviro_gear	not sure	8
233	23.7	Report marine preserve violations	enviro_violations		
234	23.8	Other (frequency)	enviro_other		
235	23.8	Other: Please list	enviro_other_specify	open ended	open ended

- 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 Merizo? We'll use the following scale: 1 very bad, 2 bad, 3 neither bad nor good, 4 good, or 5 very good.***

		1. Very bad	2. Bad	3. Neither bad nor good	4. Good	5. Very good	Not sure
5.1	Ocean water quality (clean and clear)						
5.2	Amount of coral						
5.3	Size of Fish						
5.4	Number of fish						
5.5	Number of fish that eat seaweed or algae						
5.6	Number of turtles						
5.7	Beach/shoreline						
5.8	Stream water quality (clean and clear)						
5.9	Forest						

- 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

29	In your opinion, how is each of the following natural resources currently doing in Merizo?				
30	5.1	Ocean water quality (clean and clear)	condition_ocean	Very bad	1
31	5.2	Amount of coral	condition_coral	Bad	2
32	5.3	Size of fish	condition_sizefish	Neither good nor bad	3
33	5.4	Number of fish	condition_numfish	Good	4
34	5.5	Number of fish that eat seaweed or algae	condition_numfish_that_eat	Very good	5
35	5.6	Number of turtles	condition_numturt	Not sure	8
36	5.7	Beach/shoreline	condition_beach		
37	5.8	Stream water quality (clean and clear)	condition_stream		
38	5.9	Forest	condition_forest		

K	L	M	N	O	P
condition_ocean	condition_coral	condition_sizefish	condition_numfish	condition_numfish_that_eat	condition_numturt
2	2	2	2	2	3
2	2	2	4	4	4
2	2	4	4	4	2
3	4	2	4	4	3
2	4	2	3	4	2
3	3	3	3	3	3
8	8	8	8	8	8
4	4	4	3	4	5

- 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"

Q8. ***What do you think are the 3 greatest threats to the reefs in Merizo?***
[INTERVIEWER: DO NOT READ. CHECK UP TO 3 ANSWERS BASED ON THE RESPONSES. PROMPT IF NECESSARY. IT IS OK IF THEY PROVIDE FEWER. WRITE IN 8.25 ,ANY THREAT NOT ON THE TABLE.]

		Check 3
8.1	Coral bleaching from sea surface temperature increase	□
8.2	Ocean acidification	□
8.3	Lack of vegetation in the mountains	□
8.4	Erosion in the mountains	□
8.5	Stream bank erosion	□
8.6	Fires in the mountains	□
8.7	Sedimentation caused by fire	□
8.8	Sedimentation caused by floods	□
8.9	Chemical runoff (pesticides, herbicides, fertilizers)	□
8.10	Sewage discharge	□
8.11	Typhoons	□
8.12	Storm water runoff	□
8.13	Shoreline erosion	□
8.14	Algal bloom or seaweed cover	□
8.15	Harmful Fishing practices	□
8.16	Illegal fishing	□
8.17	Overfishing	□
8.18	Overuse for recreation / tourism	□
8.19	Scuba divers	□
8.20	Ships and boats grounding on reefs	□
8.21	Off-roading	□
8.22	Increased development	□
8.23	Trash	□
8.24	Poor water quality	□
8.25	Other: Please list	□

- 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"

61	What do you think are the 3 greatest threats to the reefs in Merizo?				
62	8.1	Coral bleaching from sea surface temperature increase	threat_bleach	respondent chose as top 3	1
63	8.2	Ocean acidification	threat_acid	respondent did not choose as top 3	0

AN	AO	AP	AQ	AR	AS	AT	AU
threat_bleach	threat_acid	threat_veg	threat_mterosion	threat_streamerosion	threat_mtfire	threat_sedfire	threat_sedflood
0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	1
0	0	0	0	0	0	0	0
0	0	1	0	0	0	1	0
0	0	0	1	0	1	0	0
0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

84	8.23	Trash	threat_trash		
85	8.24	Poor water quality	threat_water		
86	8.25	Other: Please list	threat_other	open ended	open ended

- 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

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"

Q19. *Please rate the success of the following natural resource management activities in Merizo. On a scale from 1 to 5, where 1 is very low, 2 low, 3 medium, 4 high, and 5 very high.*

		1. Very low	2. Low	3. Medium	4. High	5. Very high	Not sure
19.1	These efforts increased availability of locally sourced marine and terrestrial foods						
19.2	Increasing household participation in natural resources management planning or decision making						
19.3	Increasing use of community input and scientific data in decision making of the Micronesia Challenge						
19.4	Protecting the whole coral reef ecosystem						
19.5	Improving the water quality of the area, including reducing contamination						
19.6	Reducing sedimentation						
19.7	Reducing algae or seaweeds that are harmful to the reefs						
19.8	Increasing the public environmental awareness						
19.9	Increasing of number of community driven management plans endorsed by stakeholders						
19.10	Increasing tourism						
19.11	Reduce violations and illegal activities related to fishing, harvesting, and use of natural resources						
19.12	Reducing user conflicts						
19.13	Protecting cultural artifacts or traditions						
19.14	Other: Please list						

Special Cases in Coding: “Other – please list”

179 **Please rate the success of the following natural resource management activities in Merizo.**

ER	ES	ET	EU	EV	EW
success_tourism	success_reduce_violations	success_conflict	success_culture	success_other	success_other_specify
3	3	4	4	.	.
5	3	2	5	.	.
4	3	5	3	.	.
3	3	4	6	.	.
1	1	2	2	1	reducing over-harvestation
3	3	3	3	.	.
8	8	8	8	.	.
4	4	6	8	.	.

192	19.13	Protecting cultural artifacts or traditions	success_culture		
193	19.14	Other (rating)	success_other		
194	19.14	Other: Please list	success_other_specify	open ended	open ended

- 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

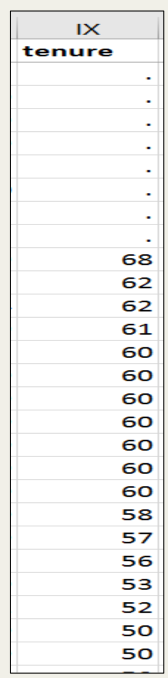
- Things to look for:
  - *Missing or duplicate records*
  - *Missing data*
  - *Values out of range*
  - *Varying format*
  - *Inconsistencies in text fields*
  - *Data in the wrong field*
  - *Numbers or text in cells below the last row of data*
    - This can affect calculations when a data set is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

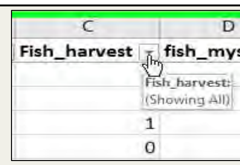
- Multiple levels can be added to the sort
- We want to sort by "income," from low to high, then sort by "tenure" from low to high
- Under the data tab, click "Sort"
- Click "add level"
- In the first "sort by," pick "income"
- In the second "sort by," pick "tenure"
- In "order," pick "smallest to largest"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

IW	IX
income	tenure
1	2
1	3
1	3
1	11
1	18
1	19
1	19
1	20
1	20
1	28
1	45
1	52
2	1
2	2
2	4
2	6
2	8
2	11
2	14
2	17
2	30
2	32
2	40
2	42
2	50
3	8
3	10

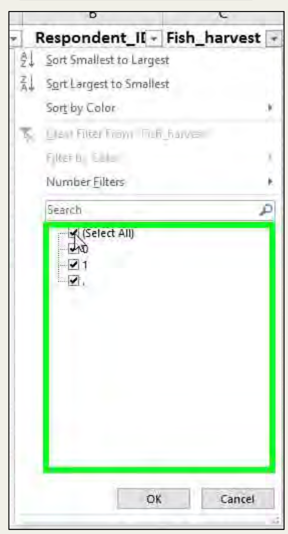
Filter example

- 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"

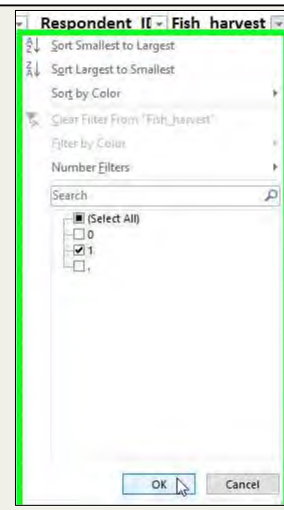
1



2



3



Filtered Data

C	D	E	F	G	H	
Fish_harvest	fish_myself	fish_sell	fish_give	fish_fun	fish_culture	f
1	3	3	2	3	3	
1	3	3	4	3	3	
1	2	2	3	2	3	
1	3	2	3	3	3	
1	3	2	2	4	4	
1	2	2	3	2	3	
1	2	3	2	4	.	
1	3	3	3	3	3	
1	4	1	4	4	4	
1	3	3	4	3	4	
1	2	2	3	2	3	
1	3	3	2	2	2	
1	4	.	4	3	4	
1	4	4	3	4	4	
1	4	.	4	4	4	
1	.	.	.	.	.	
1	3	2	3	3	3	
1	4	4	4	4	4	
1	2	3	2	3	3	

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

A	CB	DA	DB	DC	DD	DE
Respondent_ID	familiar_Achang	benefit	benefit_fish	benefit_jobs	benefit_culture	benefit_rec
1	1	1	1	0	1	1
2	1	0	.	.	.	.
3	1	0	.	.	.	.
4	0	1	0	1	1	1
5	1	1	1	1	0	0
6	0	.	.	.	.	.
7	0	1	1	0	1	1
8	0	.	.	.	.	.

- 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**

A	CB	DA	DB	DC	DD	DE
Respondent_ID	familiar_Achang	benefit	benefit_fish	benefit_jobs	benefit_culture	benefit_rec
1	1	1	1	0	1	1
2	1	0	.	.	.	.
3	1	0	.	.	.	.
4	0	.	.	.	.	.
5	1	1	1	1	0	0
6	0	.	.	.	.	.
7	0	.	.	.	.	.
8	0	.	.	.	.	.

Documenting the Decision

Date	Issue	Decision
9-12-16	Skip pattern was not followed correctly from Question 12; some respondents were allowed to answer Question 15 even if they chose "no" for Q12	Since data should not have been entered for Q15, all data values were changed to missing values to comply with survey instrument

- Keep a running log of all decisions like this and be as detailed as possible
  - So that you will remember you 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

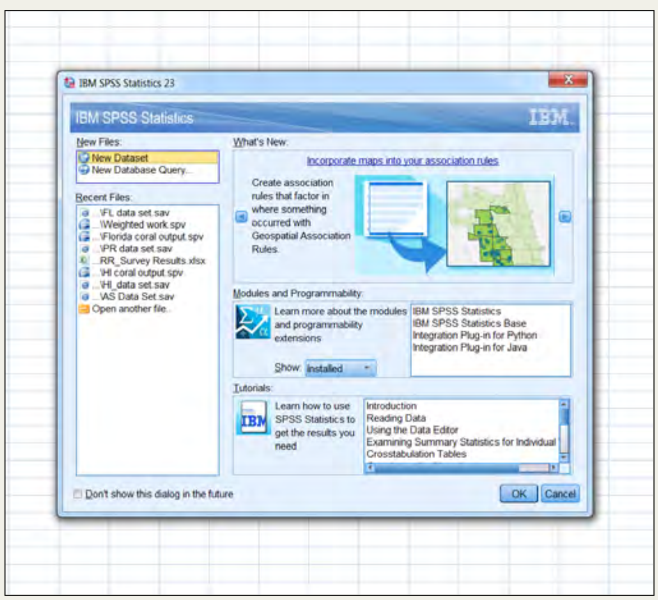
Basic Overview

- 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

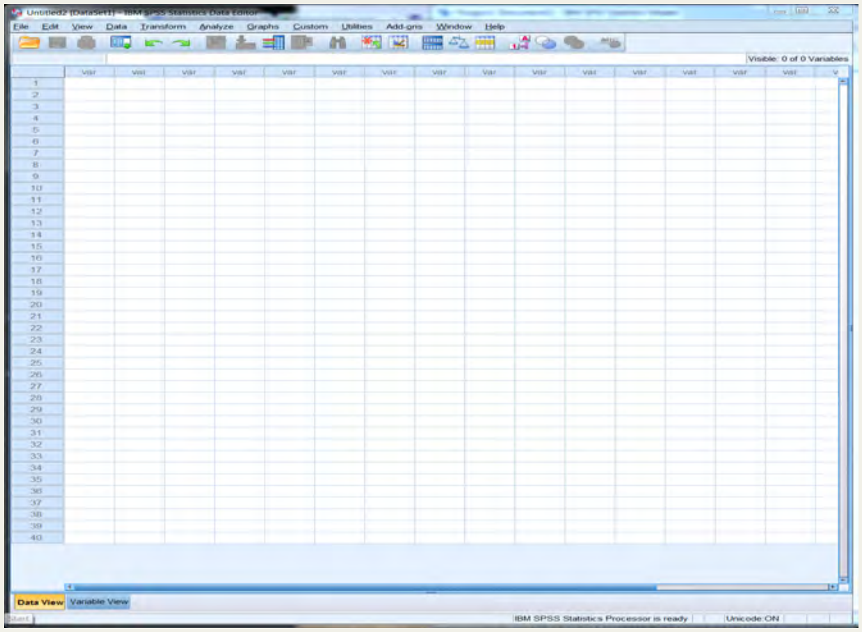
Opening SPSS

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

Blank SPSS Data Set



Blank SPSS Data Set



Data and Output Windows

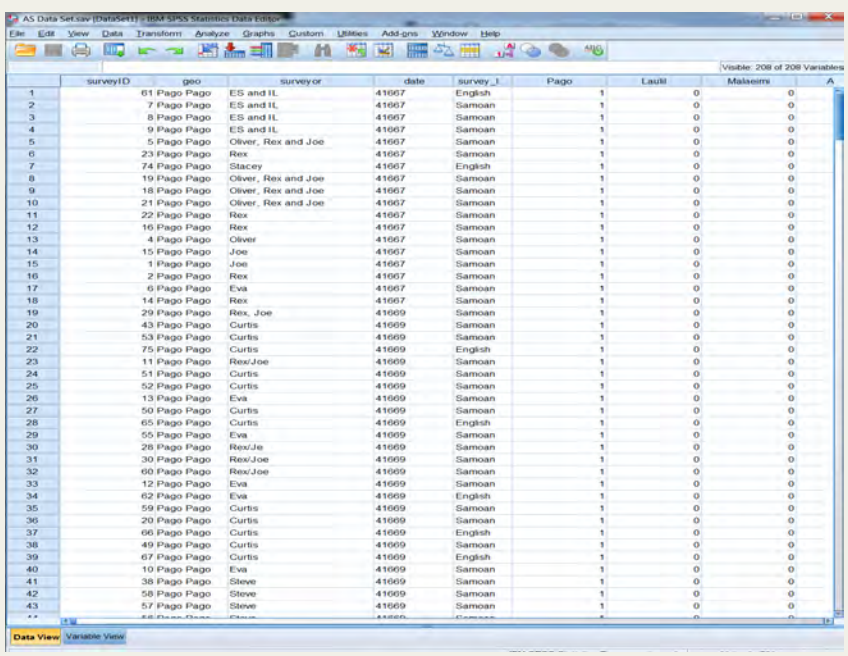
- When you click to open an SPSS data set, two separate windows are opened:
  - ***Data Editor Window*** *displays your data set and the variables inside your data set*
  - *Output Viewer Window displays any resulting output from using SPSS functions in analyzing your data*

Analysis functions can be used in both windows

Data Editor Window

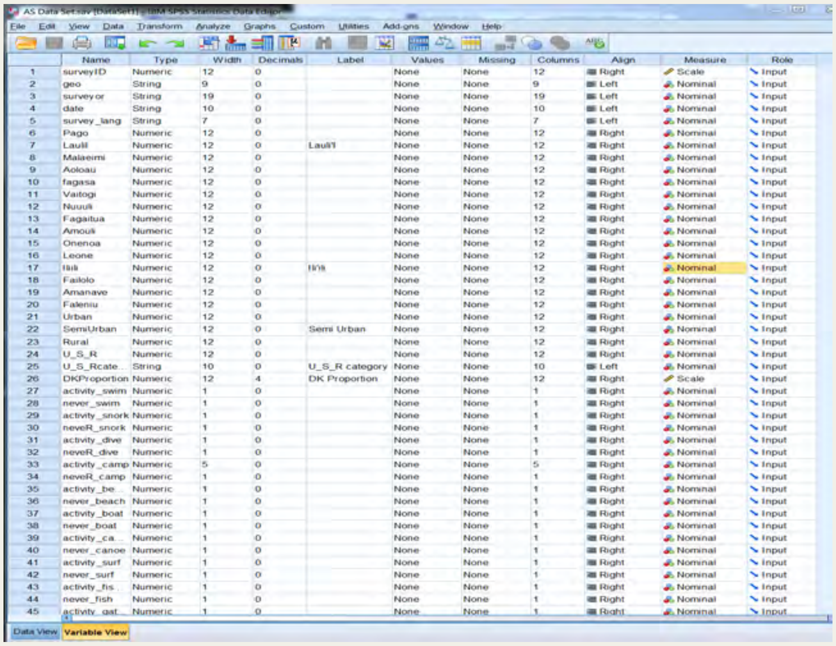
- Data View – Shows your data set with each variable sitting in its own column and the data underneath the variable heading
- Variable view – Shows each variable in your data set in rows with the characteristics (variable classification, number of decimal places, etc.) of each variable in the columns

Data View



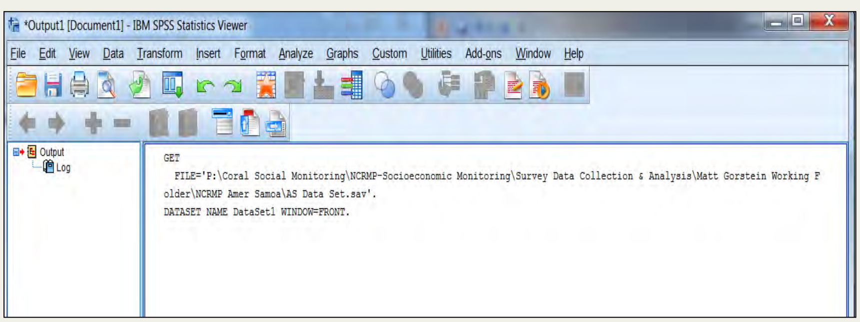
	surveyID	geo	surveyor	date	survey_l	Pago	Laulii	Malaeimi	A
1	61	Pago Pago	ES and IL	41667	English	1	0	0	
2	7	Pago Pago	ES and IL	41667	Samoan	1	0	0	
3	8	Pago Pago	ES and IL	41667	Samoan	1	0	0	
4	9	Pago Pago	ES and IL	41667	Samoan	1	0	0	
5	5	Pago Pago	Oliver, Rex and Joe	41667	Samoan	1	0	0	
6	23	Pago Pago	Rex	41667	Samoan	1	0	0	
7	74	Pago Pago	Stacey	41667	English	1	0	0	
8	19	Pago Pago	Oliver, Rex and Joe	41667	Samoan	1	0	0	
9	18	Pago Pago	Oliver, Rex and Joe	41667	Samoan	1	0	0	
10	21	Pago Pago	Oliver, Rex and Joe	41667	Samoan	1	0	0	
11	22	Pago Pago	Rex	41667	Samoan	1	0	0	
12	16	Pago Pago	Rex	41667	Samoan	1	0	0	
13	4	Pago Pago	Oliver	41667	Samoan	1	0	0	
14	15	Pago Pago	Joe	41667	Samoan	1	0	0	
15	1	Pago Pago	Joe	41667	Samoan	1	0	0	
16	2	Pago Pago	Rex	41667	Samoan	1	0	0	
17	6	Pago Pago	Eva	41667	Samoan	1	0	0	
18	14	Pago Pago	Rex	41667	Samoan	1	0	0	
19	29	Pago Pago	Rex, Joe	41669	Samoan	1	0	0	
20	43	Pago Pago	Curtis	41669	Samoan	1	0	0	
21	53	Pago Pago	Curtis	41669	Samoan	1	0	0	
22	75	Pago Pago	Curtis	41669	English	1	0	0	
23	11	Pago Pago	Rex/Joe	41669	Samoan	1	0	0	
24	51	Pago Pago	Curtis	41669	Samoan	1	0	0	
25	52	Pago Pago	Curtis	41669	Samoan	1	0	0	
26	13	Pago Pago	Eva	41669	Samoan	1	0	0	
27	50	Pago Pago	Curtis	41669	Samoan	1	0	0	
28	65	Pago Pago	Curtis	41669	English	1	0	0	
29	55	Pago Pago	Eva	41669	Samoan	1	0	0	
30	28	Pago Pago	Rex/Je	41669	Samoan	1	0	0	
31	30	Pago Pago	Rex/Joe	41669	Samoan	1	0	0	
32	60	Pago Pago	Rex/Joe	41669	Samoan	1	0	0	
33	12	Pago Pago	Eva	41669	Samoan	1	0	0	
34	62	Pago Pago	Eva	41669	English	1	0	0	
35	59	Pago Pago	Curtis	41669	Samoan	1	0	0	
36	20	Pago Pago	Curtis	41669	Samoan	1	0	0	
37	66	Pago Pago	Curtis	41669	English	1	0	0	
38	49	Pago Pago	Curtis	41669	Samoan	1	0	0	
39	67	Pago Pago	Curtis	41669	English	1	0	0	
40	10	Pago Pago	Eva	41669	Samoan	1	0	0	
41	38	Pago Pago	Steve	41669	Samoan	1	0	0	
42	58	Pago Pago	Steve	41669	Samoan	1	0	0	
43	57	Pago Pago	Steve	41669	Samoan	1	0	0	

Variable View



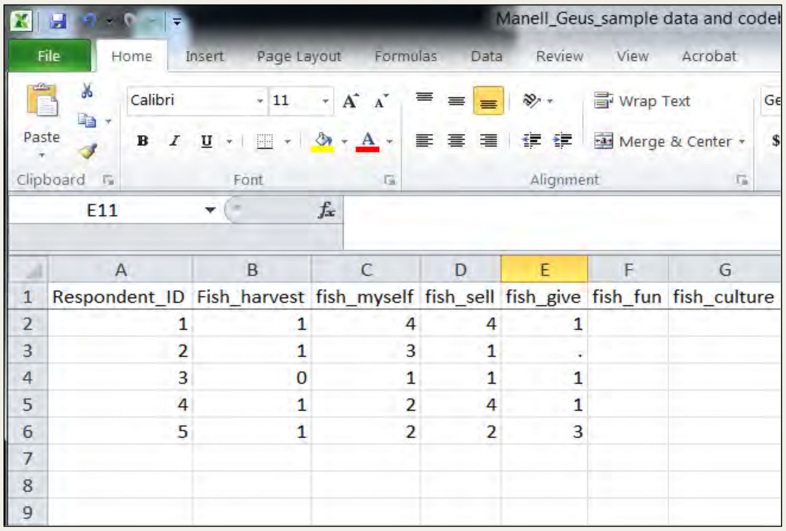
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	surveyID	Numeric	12	0		None	None	12	Right	Scale	Input
2	geo	String	9	0		None	None	9	Left	Nominal	Input
3	surveyor	String	19	0		None	None	19	Left	Nominal	Input
4	date	String	10	0		None	None	10	Left	Nominal	Input
5	survey_lang	String	7	0		None	None	7	Left	Nominal	Input
6	Pago	Numeric	12	0		None	None	12	Right	Nominal	Input
7	Laulil	Numeric	12	0	Lauli'I	None	None	12	Right	Nominal	Input
8	Malaeimi	Numeric	12	0		None	None	12	Right	Nominal	Input
9	Aoloau	Numeric	12	0		None	None	12	Right	Nominal	Input
10	fagasa	Numeric	12	0		None	None	12	Right	Nominal	Input
11	Vaitogi	Numeric	12	0		None	None	12	Right	Nominal	Input
12	Nuuuli	Numeric	12	0		None	None	12	Right	Nominal	Input
13	Fagaitua	Numeric	12	0		None	None	12	Right	Nominal	Input
14	Amouli	Numeric	12	0		None	None	12	Right	Nominal	Input
15	Onenoa	Numeric	12	0		None	None	12	Right	Nominal	Input
16	Leone	Numeric	12	0		None	None	12	Right	Nominal	Input
17	Ilili	Numeric	12	0	Ili'ili	None	None	12	Right	Nominal	Input
18	Failolo	Numeric	12	0		None	None	12	Right	Nominal	Input
19	Amanave	Numeric	12	0		None	None	12	Right	Nominal	Input
20	Faleniu	Numeric	12	0		None	None	12	Right	Nominal	Input
21	Urban	Numeric	12	0		None	None	12	Right	Nominal	Input
22	SemiUrban	Numeric	12	0	Semi Urban	None	None	12	Right	Nominal	Input
23	Rural	Numeric	12	0		None	None	12	Right	Nominal	Input
24	U_S_R	Numeric	12	0		None	None	12	Right	Nominal	Input
25	U_S_Rcate...	String	10	0	U_S_R category	None	None	10	Left	Nominal	Input
26	DKProportion	Numeric	12	4	DK Proportion	None	None	12	Right	Scale	Input
27	activity_swim	Numeric	1	0		None	None	1	Right	Nominal	Input
28	never_swim	Numeric	1	0		None	None	1	Right	Nominal	Input
29	activity_snork	Numeric	1	0		None	None	1	Right	Nominal	Input
30	neveR_snork	Numeric	1	0		None	None	1	Right	Nominal	Input
31	activity_dive	Numeric	1	0		None	None	1	Right	Nominal	Input
32	neveR_dive	Numeric	1	0		None	None	1	Right	Nominal	Input
33	activity_camp	Numeric	5	0		None	None	5	Right	Nominal	Input
34	neveR_camp	Numeric	1	0		None	None	1	Right	Nominal	Input
35	activity_be...	Numeric	1	0		None	None	1	Right	Nominal	Input
36	never_beach	Numeric	1	0		None	None	1	Right	Nominal	Input
37	activity_boat	Numeric	1	0		None	None	1	Right	Nominal	Input
38	never_boat	Numeric	1	0		None	None	1	Right	Nominal	Input
39	activity_ca...	Numeric	1	0		None	None	1	Right	Nominal	Input
40	never_canoe	Numeric	1	0		None	None	1	Right	Nominal	Input
41	activity_surf	Numeric	1	0		None	None	1	Right	Nominal	Input
42	never_surf	Numeric	1	0		None	None	1	Right	Nominal	Input
43	activity_fis...	Numeric	1	0		None	None	1	Right	Nominal	Input
44	never_fish	Numeric	1	0		None	None	1	Right	Nominal	Input
45	activity_gat...	Numeric	1	0		None	None	1	Right	Nominal	Input

Output Window



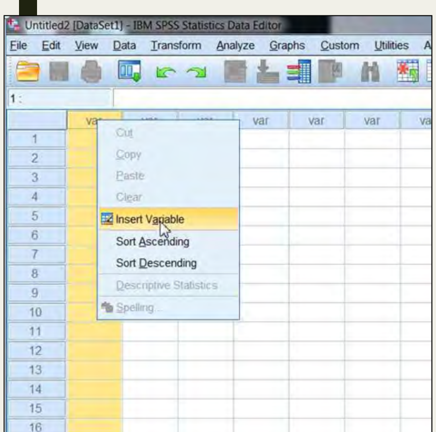
Manually Entering Data

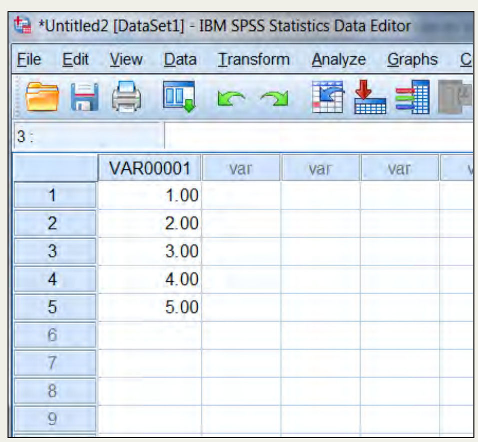
*Open the file “Manell_Geus_ManualEnter.xlsx”



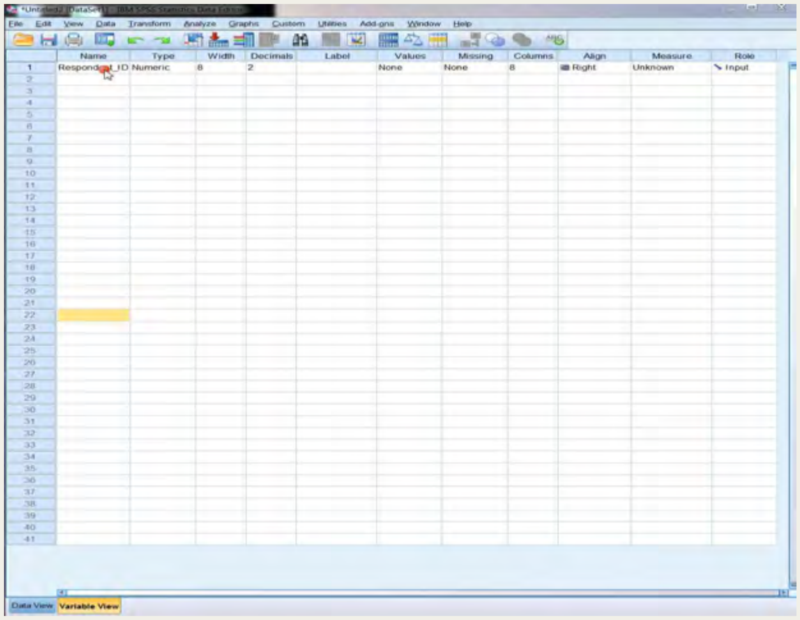
	A	B	C	D	E	F	G
1	Respondent_ID	Fish_harvest	fish_myself	fish_sell	fish_give	fish_fun	fish_culture
2	1	1	4	4	1		
3	2	1	3	1	.		
4	3	0	1	1	1		
5	4	1	2	4	1		
6	5	1	2	2	3		
7							
8							
9							

Inserting a Variable

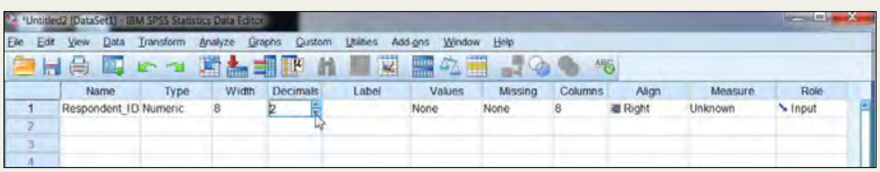


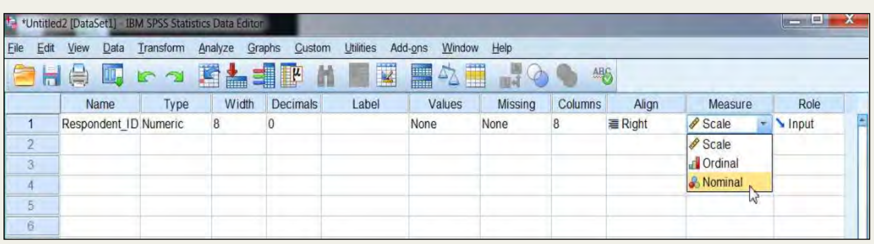


Naming the Variable

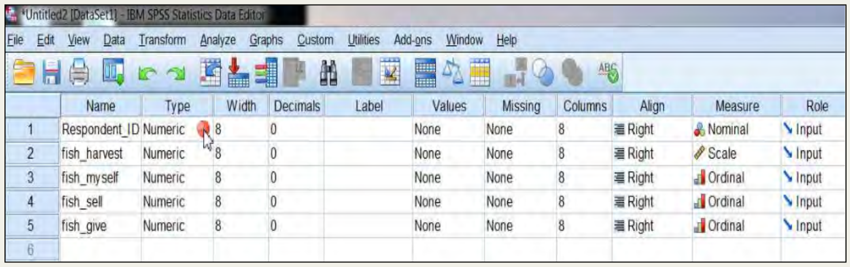


Classifying the Variable

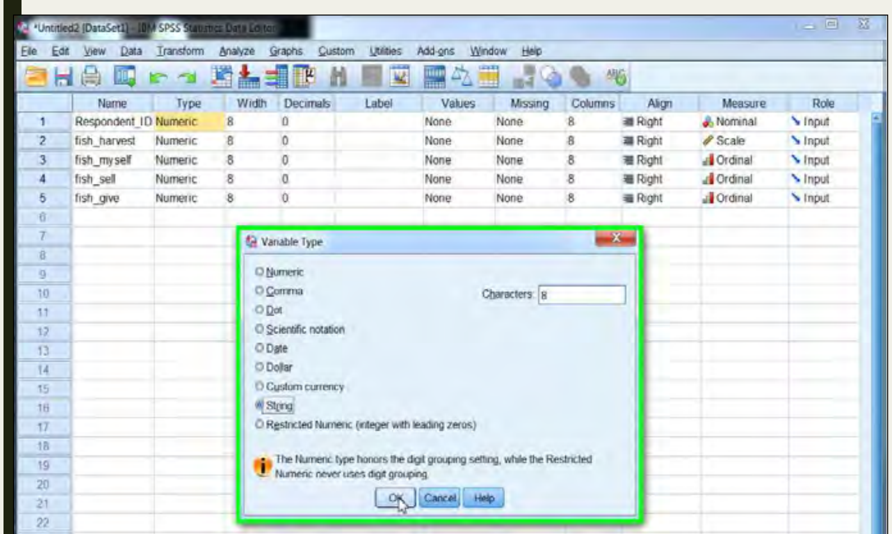




Classifying the Variable

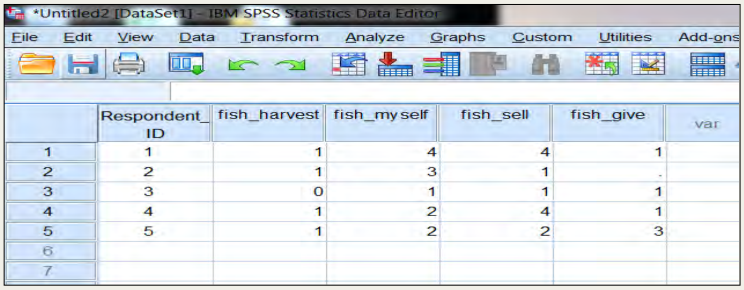


Classifying the Variable



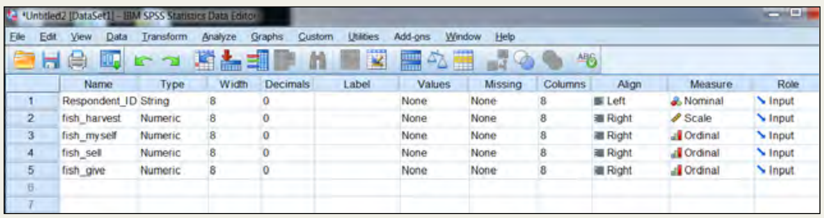
A Manually Entered Data Set

Data View:



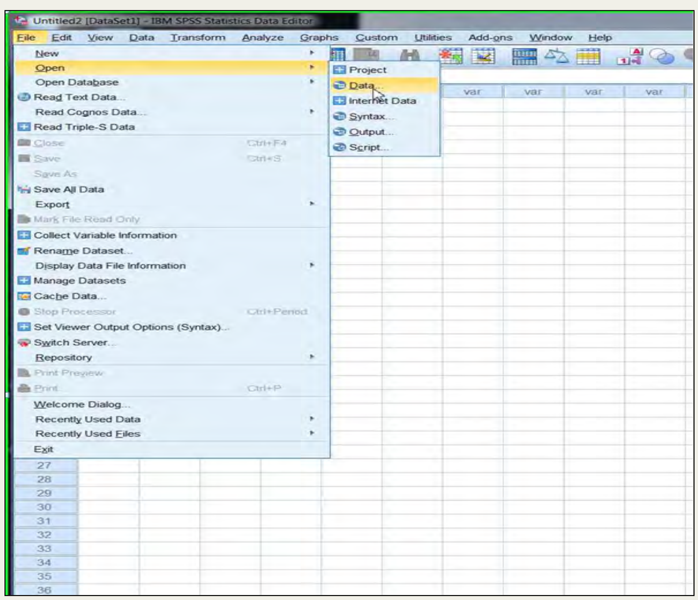
	Respondent_ID	fish_harvest	fish_myself	fish_sell	fish_give	var
1	1	1	4	4	1	
2	2	1	3	1	.	
3	3	0	1	1	1	
4	4	1	2	4	1	
5	5	1	2	2	3	
6						
7						

Variable View:

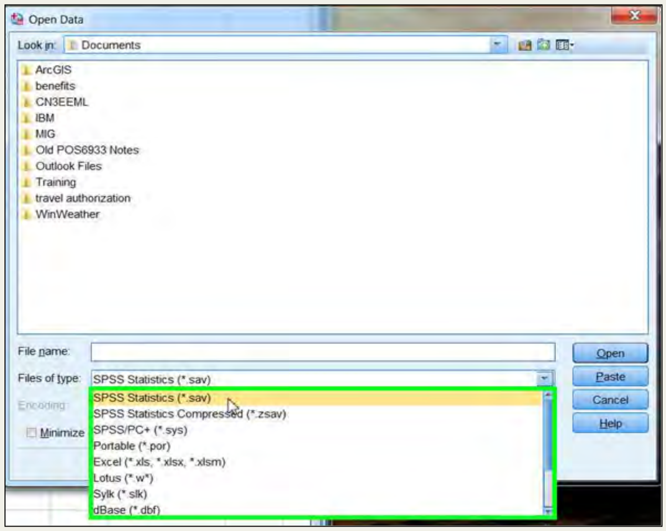


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Respondent_ID	String	8	0		None	None	8	Left	Nominal	Input
2	fish_harvest	Numeric	8	0		None	None	8	Right	Scale	Input
3	fish_myself	Numeric	8	0		None	None	8	Right	Ordinal	Input
4	fish_sell	Numeric	8	0		None	None	8	Right	Ordinal	Input
5	fish_give	Numeric	8	0		None	None	8	Right	Ordinal	Input
6											
7											

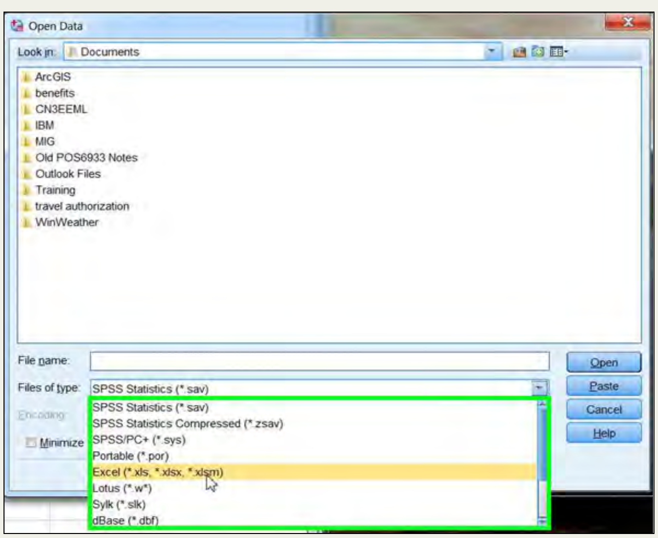
Importing a Dataset



Opening an SPSS data file



Opening an Excel Data File

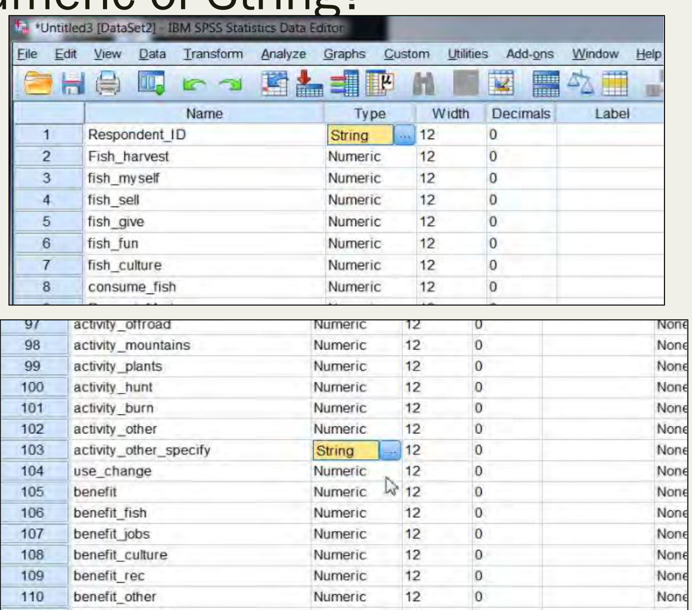


Open the "Manell_Geus_Getting Started.xlsx" file

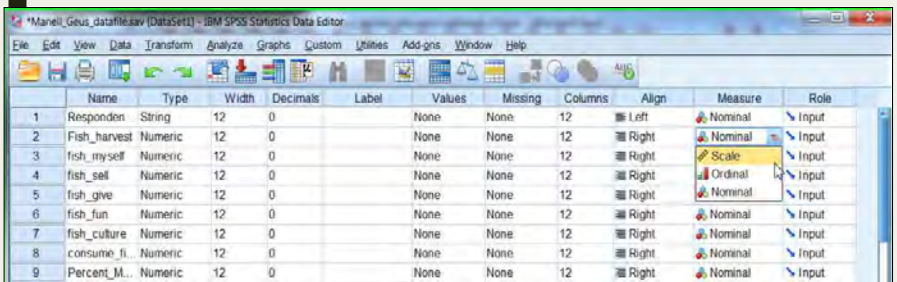
Creating a Codebook in SPSS

- 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*

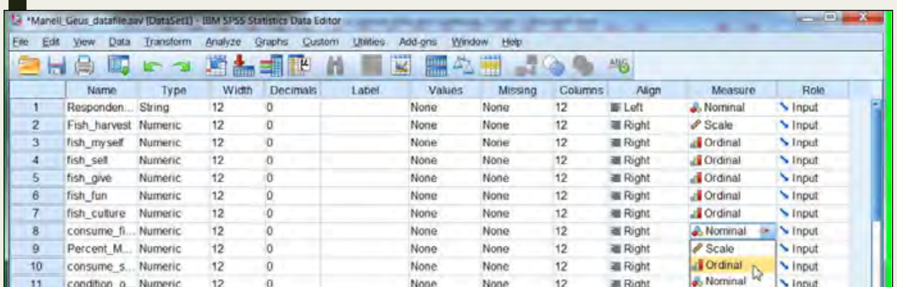
Numeric or String?



Type of Measure?

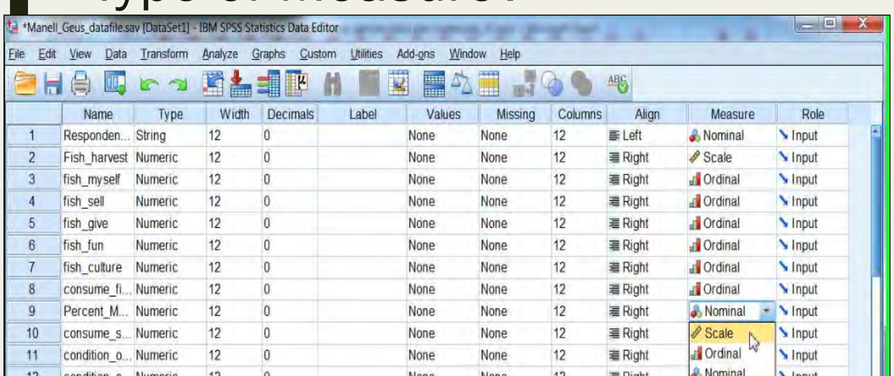


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Responden	String	12	0		None	None	12	Left	Nominal	Input
2	Fish_harvest	Numeric	12	0		None	None	12	Right	Nominal	Input
3	fish_myself	Numeric	12	0		None	None	12	Right	Scale	Input
4	fish_sell	Numeric	12	0		None	None	12	Right	Ordinal	Input
5	fish_give	Numeric	12	0		None	None	12	Right	Nominal	Input
6	fish_fun	Numeric	12	0		None	None	12	Right	Nominal	Input
7	fish_culture	Numeric	12	0		None	None	12	Right	Nominal	Input
8	consume_fi...	Numeric	12	0		None	None	12	Right	Nominal	Input
9	Percent_M...	Numeric	12	0		None	None	12	Right	Nominal	Input



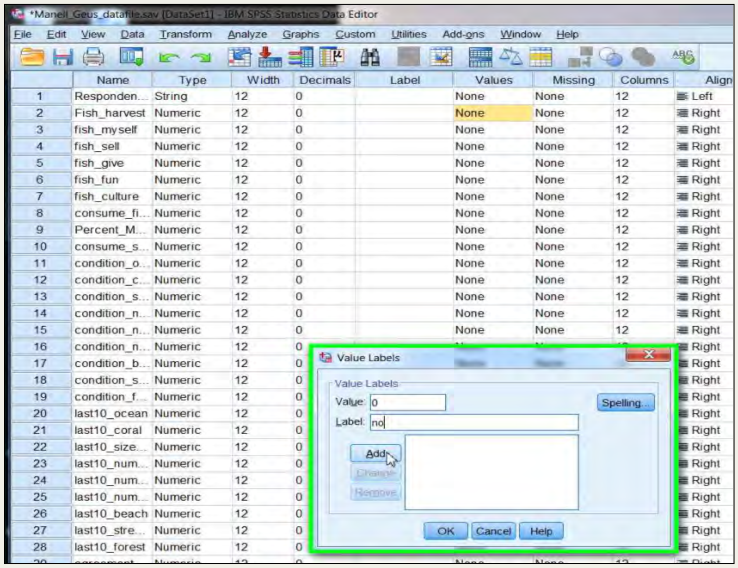
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Responden...	String	12	0		None	None	12	Left	Nominal	Input
2	Fish_harvest	Numeric	12	0		None	None	12	Right	Scale	Input
3	fish_myself	Numeric	12	0		None	None	12	Right	Ordinal	Input
4	fish_sell	Numeric	12	0		None	None	12	Right	Ordinal	Input
5	fish_give	Numeric	12	0		None	None	12	Right	Ordinal	Input
6	fish_fun	Numeric	12	0		None	None	12	Right	Ordinal	Input
7	fish_culture	Numeric	12	0		None	None	12	Right	Ordinal	Input
8	consume_fi...	Numeric	12	0		None	None	12	Right	Nominal	Input
9	Percent_M...	Numeric	12	0		None	None	12	Right	Scale	Input
10	consume_s...	Numeric	12	0		None	None	12	Right	Ordinal	Input
11	condition_o...	Numeric	12	0		None	None	12	Right	Nominal	Input

Type of Measure?

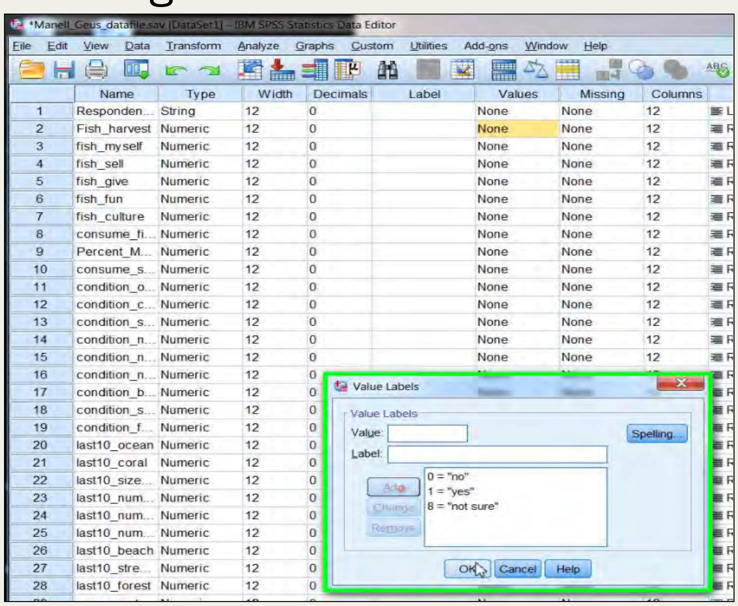


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Responden...	String	12	0		None	None	12	Left	Nominal	Input
2	Fish_harvest	Numeric	12	0		None	None	12	Right	Scale	Input
3	fish_myself	Numeric	12	0		None	None	12	Right	Ordinal	Input
4	fish_sell	Numeric	12	0		None	None	12	Right	Ordinal	Input
5	fish_give	Numeric	12	0		None	None	12	Right	Ordinal	Input
6	fish_fun	Numeric	12	0		None	None	12	Right	Ordinal	Input
7	fish_culture	Numeric	12	0		None	None	12	Right	Ordinal	Input
8	consume_fi...	Numeric	12	0		None	None	12	Right	Ordinal	Input
9	Percent_M...	Numeric	12	0		None	None	12	Right	Nominal	Input
10	consume_s...	Numeric	12	0		None	None	12	Right	Scale	Input
11	condition_o...	Numeric	12	0		None	None	12	Right	Ordinal	Input
12	condition_c...	Numeric	12	0		None	None	12	Right	Nominal	Input

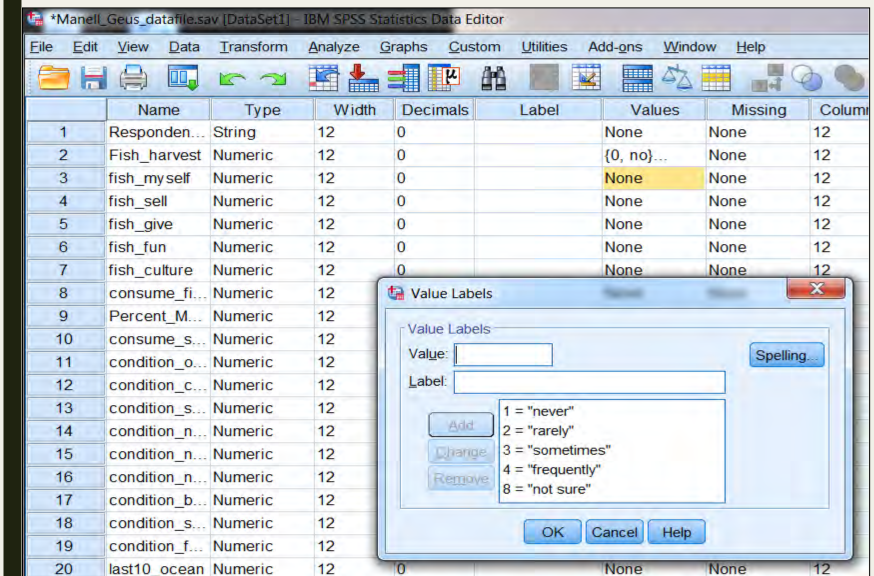
Illustrating Variable Codes



Illustrating Variable Codes

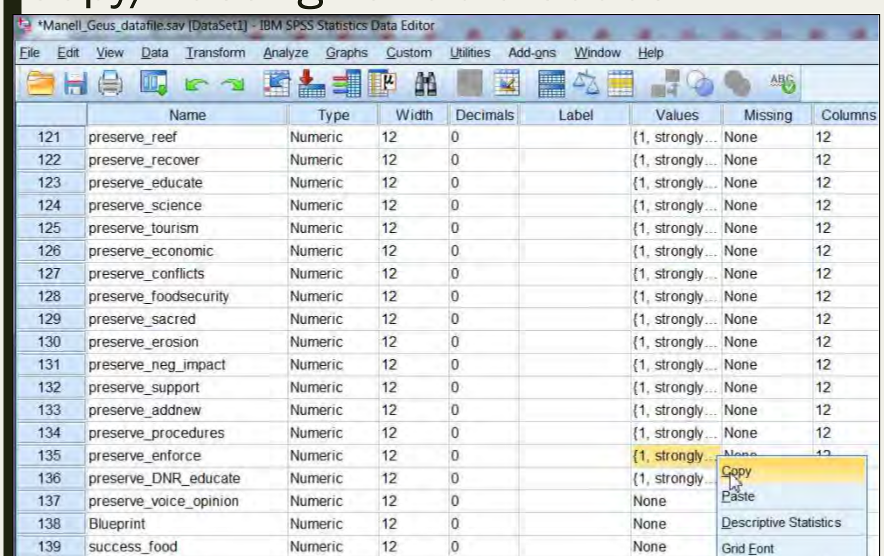


Illustrating Variable Codes



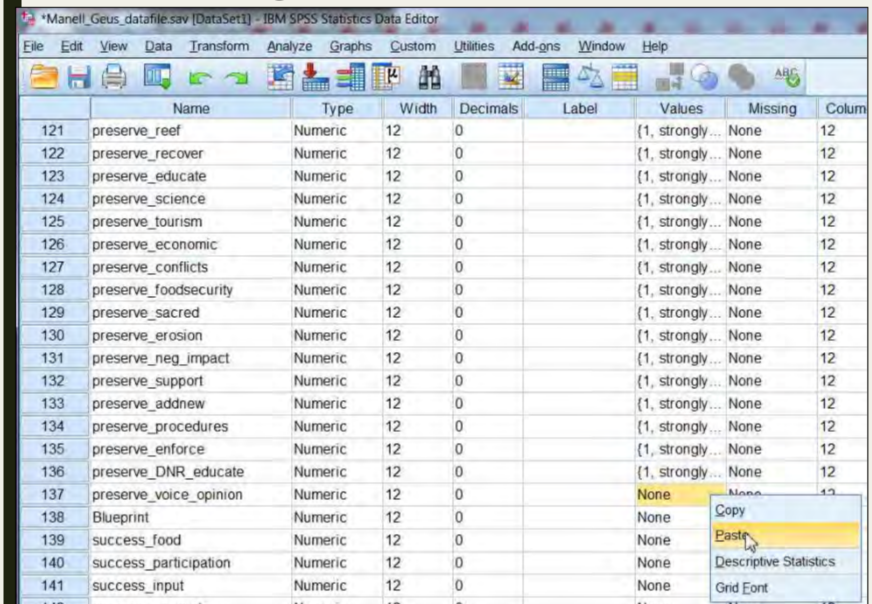
	Name	Type	Width	Decimals	Label	Values	Missing	Column
1	Responden...	String	12	0		None	None	12
2	Fish_harvest	Numeric	12	0		{0, no}...	None	12
3	fish_myself	Numeric	12	0		None	None	12
4	fish_sell	Numeric	12	0		None	None	12
5	fish_give	Numeric	12	0		None	None	12
6	fish_fun	Numeric	12	0		None	None	12
7	fish_culture	Numeric	12	0		None	None	12
8	consume_fi...	Numeric	12					
9	Percent_M...	Numeric	12					
10	consume_s...	Numeric	12					
11	condition_o...	Numeric	12					
12	condition_c...	Numeric	12					
13	condition_s...	Numeric	12					
14	condition_n...	Numeric	12					
15	condition_n...	Numeric	12					
16	condition_n...	Numeric	12					
17	condition_b...	Numeric	12					
18	condition_s...	Numeric	12					
19	condition_f...	Numeric	12					
20	last10_ocean	Numeric	12	0		None	None	12

Copy/Pasting Variable Codes



	Name	Type	Width	Decimals	Label	Values	Missing	Columns
121	preserve_reef	Numeric	12	0		{1, strongly...	None	12
122	preserve_recover	Numeric	12	0		{1, strongly...	None	12
123	preserve_educate	Numeric	12	0		{1, strongly...	None	12
124	preserve_science	Numeric	12	0		{1, strongly...	None	12
125	preserve_tourism	Numeric	12	0		{1, strongly...	None	12
126	preserve_economic	Numeric	12	0		{1, strongly...	None	12
127	preserve_conflicts	Numeric	12	0		{1, strongly...	None	12
128	preserve_foodsecurity	Numeric	12	0		{1, strongly...	None	12
129	preserve_sacred	Numeric	12	0		{1, strongly...	None	12
130	preserve_erosion	Numeric	12	0		{1, strongly...	None	12
131	preserve_neg_impact	Numeric	12	0		{1, strongly...	None	12
132	preserve_support	Numeric	12	0		{1, strongly...	None	12
133	preserve_addnew	Numeric	12	0		{1, strongly...	None	12
134	preserve_procedures	Numeric	12	0		{1, strongly...	None	12
135	preserve_enforce	Numeric	12	0		{1, strongly...		
136	preserve_DNR_educate	Numeric	12	0		{1, strongly...		
137	preserve_voice_opinion	Numeric	12	0		None		
138	Blueprint	Numeric	12	0		None		
139	success_food	Numeric	12	0		None		

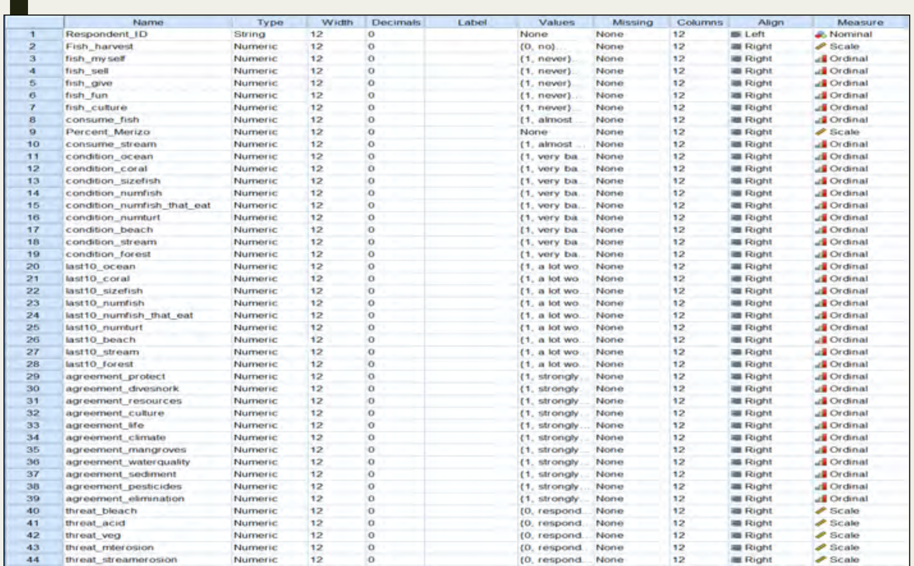
Copy/Pasting Variable Codes



	Name	Type	Width	Decimals	Label	Values	Missing	Colum
121	preserve_reef	Numeric	12	0		{1, strongly...	None	12
122	preserve_recover	Numeric	12	0		{1, strongly...	None	12
123	preserve_educate	Numeric	12	0		{1, strongly...	None	12
124	preserve_science	Numeric	12	0		{1, strongly...	None	12
125	preserve_tourism	Numeric	12	0		{1, strongly...	None	12
126	preserve_economic	Numeric	12	0		{1, strongly...	None	12
127	preserve_conflicts	Numeric	12	0		{1, strongly...	None	12
128	preserve_foodsecurity	Numeric	12	0		{1, strongly...	None	12
129	preserve_sacred	Numeric	12	0		{1, strongly...	None	12
130	preserve_erosion	Numeric	12	0		{1, strongly...	None	12
131	preserve_neg_impact	Numeric	12	0		{1, strongly...	None	12
132	preserve_support	Numeric	12	0		{1, strongly...	None	12
133	preserve_addnew	Numeric	12	0		{1, strongly...	None	12
134	preserve_procedures	Numeric	12	0		{1, strongly...	None	12
135	preserve_enforce	Numeric	12	0		{1, strongly...	None	12
136	preserve_DNR_educate	Numeric	12	0		{1, strongly...	None	12
137	preserve_voice_opinion	Numeric	12	0		None		
138	Blueprint	Numeric	12	0		None		
139	success_food	Numeric	12	0		None		
140	success_participation	Numeric	12	0		None		
141	success_input	Numeric	12	0		None		

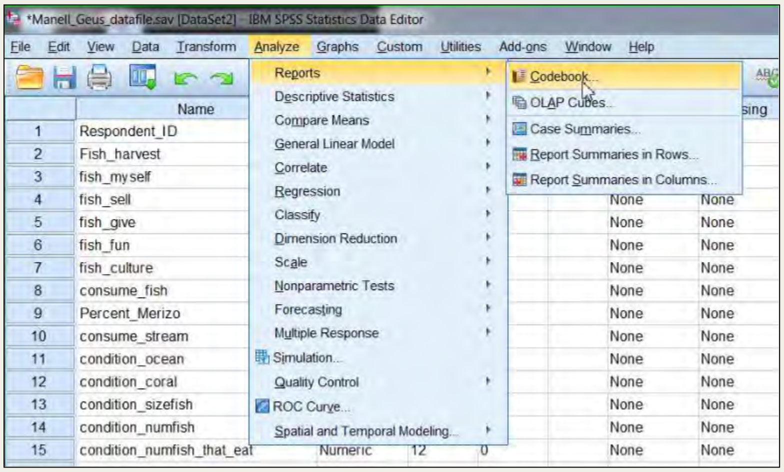
All Variables Classified and Coded

Open the file "Manell_Geus_datafile1.sav"

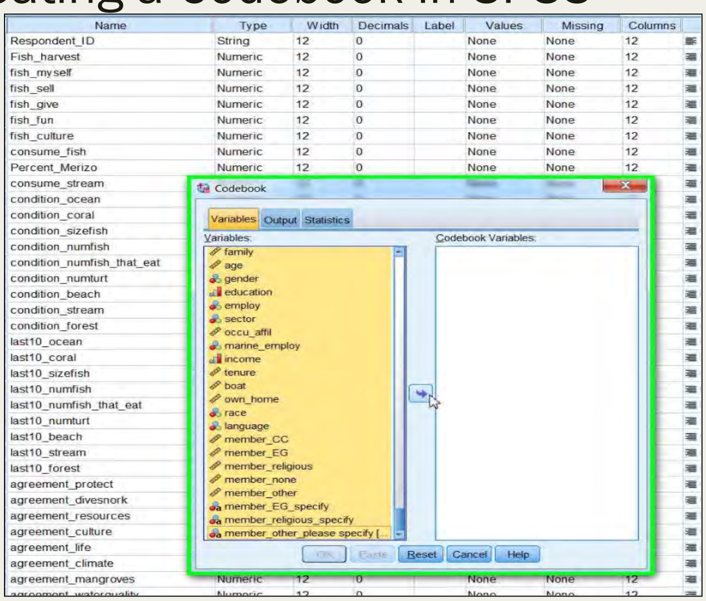


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	Respondent_ID	String	12	0		None	None	12	Left	Nominal
2	Fish_harvest	Numeric	12	0		{0, no}...	None	12	Right	Scale
3	fish_my self	Numeric	12	0		{1, never}...	None	12	Right	Ordinal
4	fish_sell	Numeric	12	0		{1, never}...	None	12	Right	Ordinal
5	fish_give	Numeric	12	0		{1, never}...	None	12	Right	Ordinal
6	fish_fun	Numeric	12	0		{1, never}...	None	12	Right	Ordinal
7	fish_culture	Numeric	12	0		{1, never}...	None	12	Right	Ordinal
8	consume_fish	Numeric	12	0		{1, almost ...	None	12	Right	Ordinal
9	Percent_Merizo	Numeric	12	0		None	None	12	Right	Scale
10	consume_stream	Numeric	12	0		{1, almost ...	None	12	Right	Ordinal
11	condition_ocean	Numeric	12	0		{1, very ba...	None	12	Right	Ordinal
12	condition_coral	Numeric	12	0		{1, very ba...	None	12	Right	Ordinal
13	condition_sizefish	Numeric	12	0		{1, very ba...	None	12	Right	Ordinal
14	condition_numfish	Numeric	12	0		{1, very ba...	None	12	Right	Ordinal
15	condition_numfish_that_eat	Numeric	12	0		{1, very ba...	None	12	Right	Ordinal
16	condition_numturt	Numeric	12	0		{1, very ba...	None	12	Right	Ordinal
17	condition_beach	Numeric	12	0		{1, very ba...	None	12	Right	Ordinal
18	condition_stream	Numeric	12	0		{1, very ba...	None	12	Right	Ordinal
19	condition_forest	Numeric	12	0		{1, very ba...	None	12	Right	Ordinal
20	last10_ocean	Numeric	12	0		{1, a lot wo...	None	12	Right	Ordinal
21	last10_coral	Numeric	12	0		{1, a lot wo...	None	12	Right	Ordinal
22	last10_sizefish	Numeric	12	0		{1, a lot wo...	None	12	Right	Ordinal
23	last10_numfish	Numeric	12	0		{1, a lot wo...	None	12	Right	Ordinal
24	last10_numfish_that_eat	Numeric	12	0		{1, a lot wo...	None	12	Right	Ordinal
25	last10_numturt	Numeric	12	0		{1, a lot wo...	None	12	Right	Ordinal
26	last10_beach	Numeric	12	0		{1, a lot wo...	None	12	Right	Ordinal
27	last10_stream	Numeric	12	0		{1, a lot wo...	None	12	Right	Ordinal
28	last10_forest	Numeric	12	0		{1, a lot wo...	None	12	Right	Ordinal
29	agreement_protect	Numeric	12	0		{1, strongly...	None	12	Right	Ordinal
30	agreement_divesnork	Numeric	12	0		{1, strongly...	None	12	Right	Ordinal
31	agreement_resources	Numeric	12	0		{1, strongly...	None	12	Right	Ordinal
32	agreement_culture	Numeric	12	0		{1, strongly...	None	12	Right	Ordinal
33	agreement_life	Numeric	12	0		{1, strongly...	None	12	Right	Ordinal
34	agreement_climate	Numeric	12	0		{1, strongly...	None	12	Right	Ordinal
35	agreement_mangroves	Numeric	12	0		{1, strongly...	None	12	Right	Ordinal
36	agreement_waterquality	Numeric	12	0		{1, strongly...	None	12	Right	Ordinal
37	agreement_sediment	Numeric	12	0		{1, strongly...	None	12	Right	Ordinal
38	agreement_pesticides	Numeric	12	0		{1, strongly...	None	12	Right	Ordinal
39	agreement_elimination	Numeric	12	0		{1, strongly...	None	12	Right	Ordinal
40	threat_bleach	Numeric	12	0		{0, respond...	None	12	Right	Scale
41	threat_acid	Numeric	12	0		{0, respond...	None	12	Right	Scale
42	threat_veg	Numeric	12	0		{0, respond...	None	12	Right	Scale
43	threat_mterosion	Numeric	12	0		{0, respond...	None	12	Right	Scale
44	threat_streamerosion	Numeric	12	0		{0, respond...	None	12	Right	Scale
45	threat_mtfire	Numeric	12	0		{0, respond...	None	12	Right	Scale

Creating a Codebook in SPSS



Creating a Codebook in SPSS



Creating a Codebook in SPSS

fish_harvest

		Value	Count	Percent
Standard Attributes	Position	2		
	Label	<none>		
	Type	Numeric		
	Format	F12		
	Measurement	Scale		
	Role	Input		
N	Valid	0		
	Missing	414		
Central Tendency and Dispersion	Mean	.		
	Standard Deviation	.		
	Percentile 25			
	Percentile 50			
	Percentile 75			
Labeled Values	0	no	0	0.0%
	1	yes	0	0.0%
	9	not sure	0	0.0%

fish_myself

		Value	Count	Percent
Standard Attributes	Position	3		
	Label	<none>		
	Type	Numeric		
	Format	F12		
	Measurement	Ordinal		
	Role	Input		
Valid Values	1	never	0	0.0%
	2	rarely	0	0.0%
	3	sometimes	0	0.0%
	4	frequently	0	0.0%
	9	not sure	0	0.0%
Missing Values	System		414	100.0%

fish_sell

		Value	Count	Percent
Standard Attributes	Position	4		
	Label	<none>		
	Type	Numeric		
	Format	F12		
	Measurement	Ordinal		
	Role	Input		
Valid Values	1	never	0	0.0%
	2	rarely	0	0.0%
	3	sometimes	0	0.0%
	4	frequently	0	0.0%
	9	not sure	0	0.0%
Missing Values	System		414	100.0%

- 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)

Introduction To Qualitative Data

Day 1: September 12, 2016

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?

	Pros	**Cons**
Insider	o In-depth context knowledge o Personal connections & trust o Aware of power dynamics o Logistically easier (finding informants etc.)	o Differences in power can affect data o Can't guarantee anonymity o Might miss things that are "too normal" to notice
Outsider	o Can ask the "silly" questions o Can see beyond the local context o Can provide anonymity o Can be perceived as "neutral"	o Differences in power can affect data o Prone to making mistakes o Longer to create trust

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
  - *Differs 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

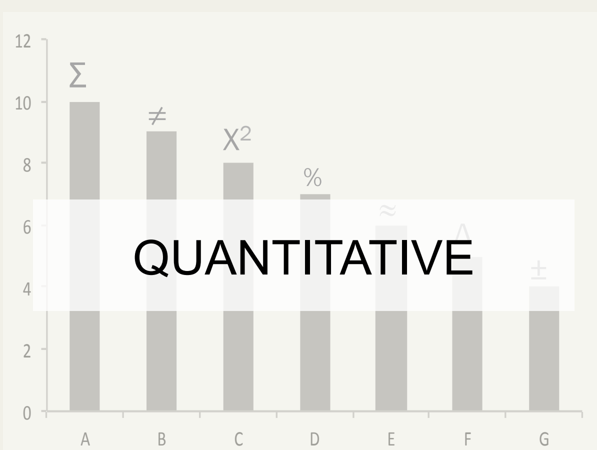
Rigor and Validity

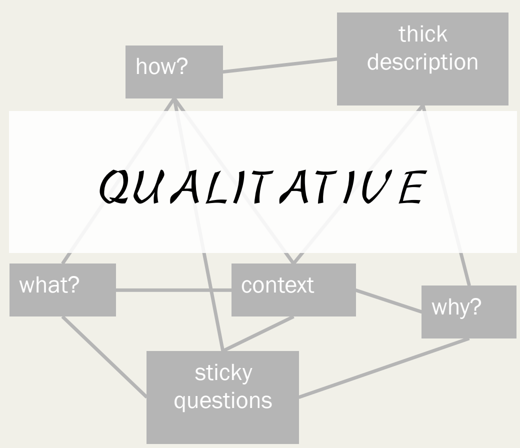
- **Rigor:** derives from the researcher's presence, the nature of the interaction between researcher and participants, the triangulation of data, the interpretation of perception, and thick description
- **Validity:** whether the conclusions being drawn from the data are credible, defensible, warranted, 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?





Qualitative or Quantitative?

It depends on your question!

Different Approaches

	Quantitative	Qualitative
Good for	Generalizable results	In-depth understanding
Purpose	Predicting & Testing	Exploring in Context
Questions	Structured	Semi-structured & Open
Sampling	Large & Random	Small & Purposeful
Collecting Method	Experiments & Surveys	Interviews, Focus Groups, Ethnography & Observation
Data	Numbers	Words

Different Approaches: Analysis

Quantitative	**Qualitative**
Data Analysis starts after data collection is completed	Data analysis and collection take place simultaneously
Researcher can be distant, removed from the research site	Researcher is the data collecting instrument, often in the field
Collecting methods include experiments, surveys	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

Quiz #2

Day 1: September 12, 2016

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

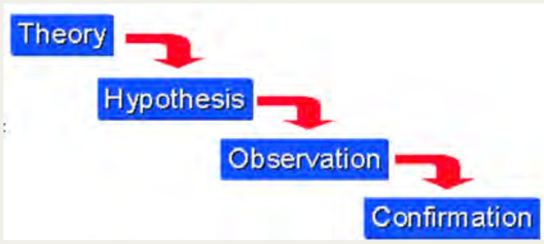
Coding Open Text Qualitative Data

Day 2: September 13, 2016

Qualitative Data Analysis Process

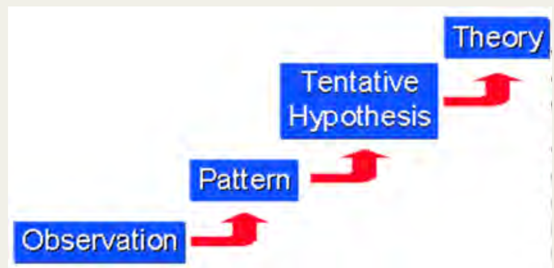
Making sense out of the data is a process of *consolidating*, *reducing*, and *interpreting* what people have said and what the researchers have observed, read, and taken notes about.

"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)

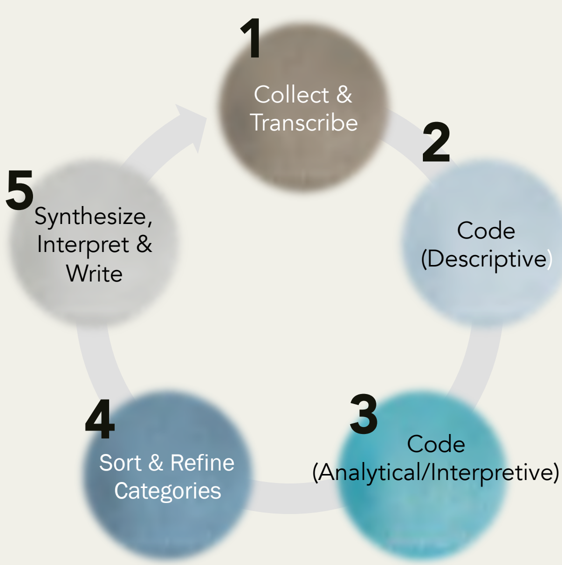


Deductive analysis

Inductive analysis



Qualitative Data Analysis Process



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

1 Collect & Transcribe

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

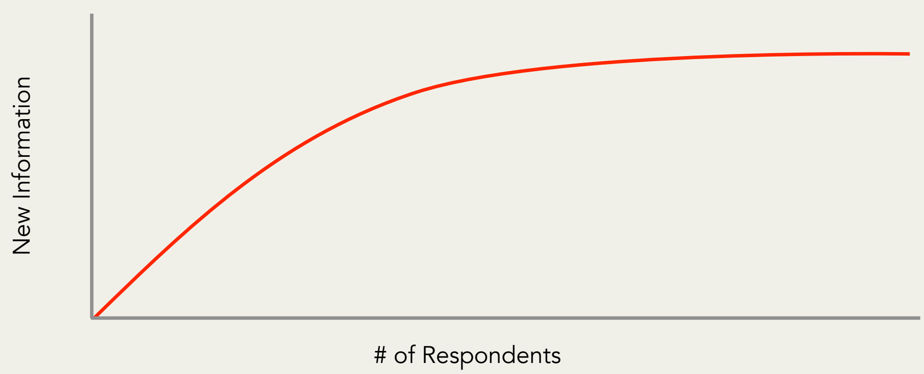
⑧ Clarify understanding with metaphors, analogies, concepts, & visuals

(Merriam 2009)

1 Collect & Transcribe

When to stop collecting data?

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



1 Collect & Transcribe

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?*

1 Collect & Transcribe

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

2
Code (Descriptive)

Examples of Descriptive Coding

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

Coding	Text
• Past infrastructure • Past harvesting (sea cucumbers), past resource condition, past gender roles • Change of marine resource conditions • Perception of alternative livelihood • Alternative livelihood • Present infrastructure	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.

2
Code (Descriptive)

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

3 Code (Analytical)

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.

4 Sort & Refine Categories

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)

4 Sort & Refine Categories

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)

5 Synthesize, Interpret & Write

EXAMPLES AND PRACTICE

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

	A
1	How would you define your "community"?
2	I would say it is environmentally minded and engaged. I am convinced that it is progressive on evironmental matters.
3	My local community is involved in resource management. I believe that the local people do not overfish. They take what they need.
4	It is trustworthy. It is a nice community. I live in a paradise.
5	I'm not sure how to answer that. There are a lot of fishermen out here, and they all have invested interest in the resources and the management of the resources in our environment.
6	They are involved in protecting the coral reefs.
7	It is very good. Everybody works together. There are always a couple people of who overfih. If the police just stick to what they do and just catch the people and not just give them warnings, then everything should be good.
8	I don't know what they do, so I don't know how to define them.
9	They love the ocean, the reef and everything involving wildlife.
10	There are several community organizations that are devoted or are involved in the ocean and beach preservation, like my county government, fish and wildlife resources and agency, and other smaller community groups.
11	It is just fine because everybody gets along.
12	I don't have any knowledge about our local community.
13	It is very rural. We are near the beach, but other things are very important. Right now, the community is more involved in GMO.
14	It is diverse. There are a lot of people who wants different things and have different opinions. There are a lot of activities.
15	They are involved in environment saving activities.
16	A lot of people here care and are trying to educate people. There are some who just litter and just do not seem to care at all.
17	It is a community of divers.
18	It is conscious of the environment. We have some big businesses that are taking over the land and are putting pesticides, like planting GMO or Genetically Modified Organism. They are damaging the environment. Our community is fighting for it, We don't agree with it, but they are big companies so they won. They say it's not damaging the environment, but it is actually destroying it.
19	It is passive.
20	I do not know.
21	They are a local Hawaiian community.

Background Example

- When coding, a few prevailing themes emerged

Example: How do you define "community"?
(in relation to marine resource management)

Category 1: spatial, social, activity	**Category 2: community involvement, activism, and character**
Definitions of community	Qualities related to interest and concern for environmental issues
Theme 1: GE (spatial)	Theme 1: EI: (interest and involvement)
Theme 2: SO (social)	Theme 2: QC (general community quality or characteristic)
Theme 3: AL (activities and livelihood)	Theme 3: AA (activism and apathy)

Background Example

- 2 categories, broken down into 3 themes each
  - *Themes are also broken down into "sub-themes"*

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

Code/name	description	Second-order theme	Examples
GE 1 Political	Political boundary: town, county, district, state	n/a	County of Honolulu, Kailua community, Volcano, Hawaii
GE 2 Landscape	Physical geography landmark: near/far from mountains, beach	2.1 Landscape (general) 2.2 Beach, shoreline 2.3 Mountains 2.4 Volcano 2.5 Ocean, water, coral reefs 2.6 Land	Beach, Mountains
GE 3 Island	Refers to "this" island, name of island or region of island	3.1 Island 3.2 District, side of island	east side of big island of Hawaii; Maui
GE 4 Area and Neighbors	Description of the built/ human environment or distance	4.1 Area (general) 4.2 Rural, up-country 4.3 City 4.4 Small, small town, village 4.5 institutions: military base, university 4.6 condos, complex, subdivision, resort 4.7 neighbors, neighborhood	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"

Background Example Results

- "Themes" are then rolled up into more general "categories"
- While some responses can contain multiple themes, each response should only fit into ONE category

Levine *et al.* 2016

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 who 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]: 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 there life.[2]

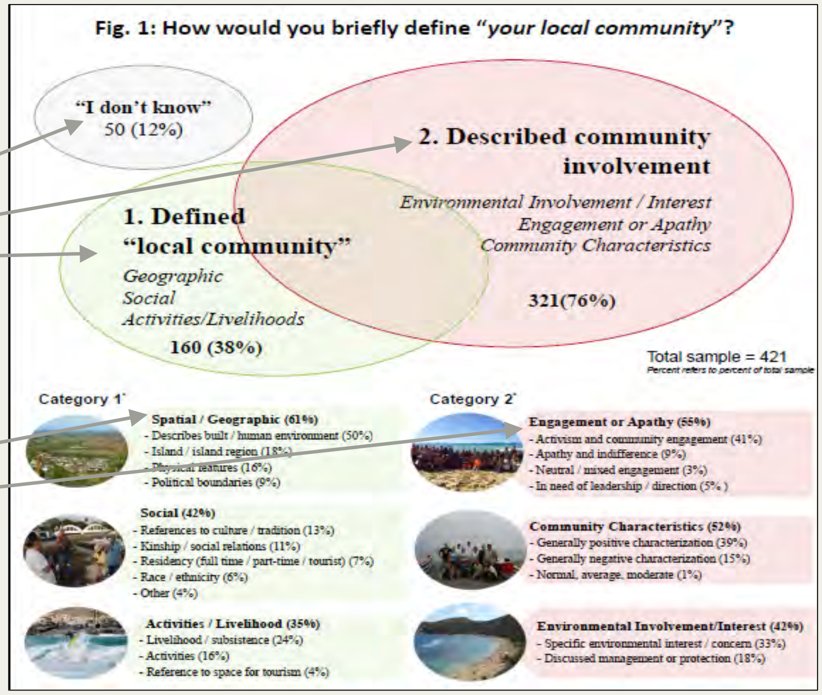
Category 2: Described Community Involvement

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 (69%). 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.[3]

Background Example Results

Categories

Themes



Levine *et al.* 2016

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 VMPA 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?*

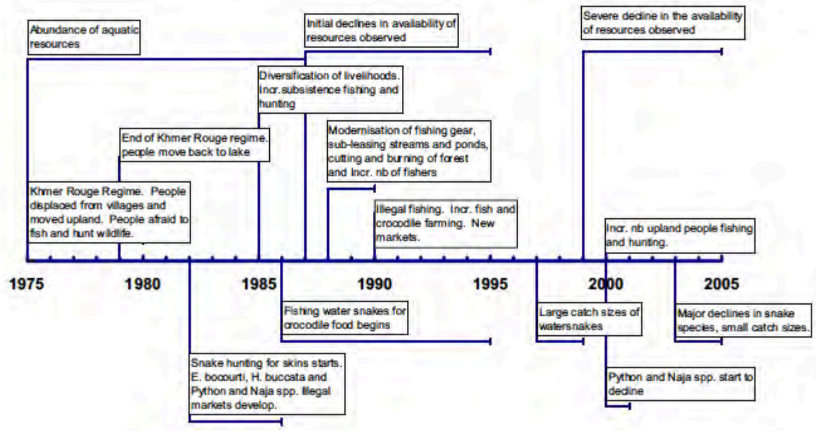
Data Visualization for Qualitative Data

Day 2: September 13, 2016

Visualizing Qualitative Data

Timelines

Fig. 2 Time line of events and changes perceived by members of the focus group discussion meetings. Generic events and changes are above the time line and events and changes regarding the snakes are shown underneath. Variation between groups in the timing of events and changes represented by the duration *bar* beneath the text

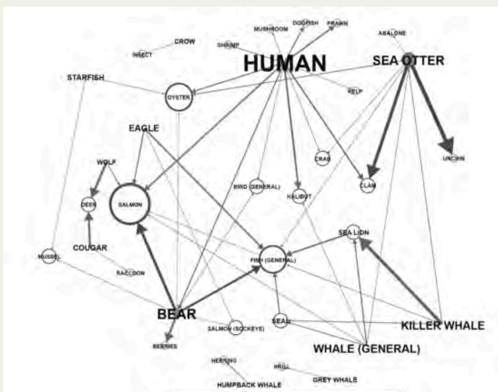


(Brooks et al. 2008)

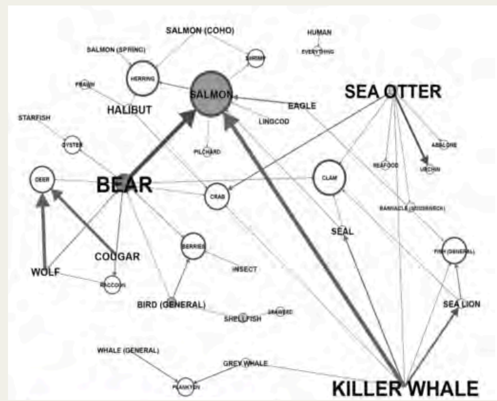
Visualizing Qualitative Data

Example: Gender differences in ecological mental models

women



men



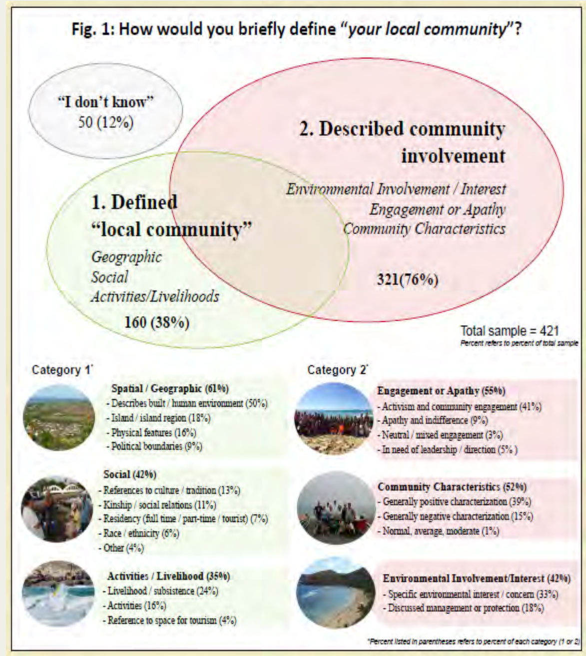
(Levine et al. 2016)

Visualizing Qualitative Data

Example: causes and effects of degradation of fisheries resources

A “flow chart”

Visualizing Qualitative Data



(Levine et al. 2016)

Venn Diagrams

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

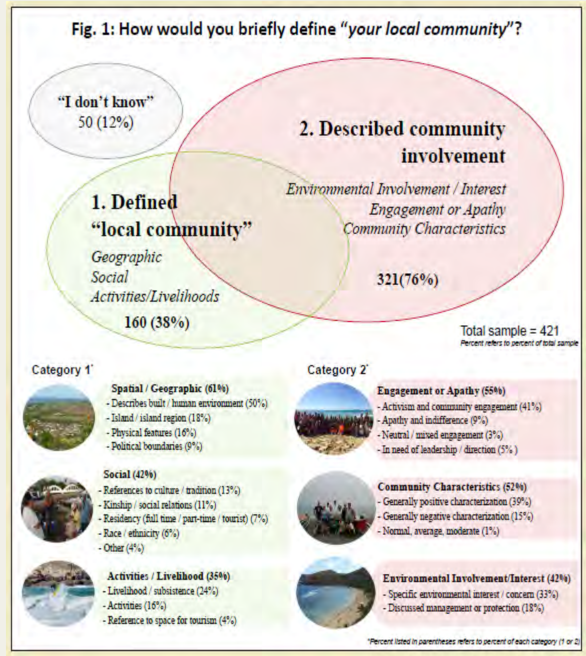
Open the file "Community Poster Draft.pdf"

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)



Quiz #3

Day 2: September 13, 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) – 3rd Quartile minus the 1st Quartile*

Examples of Central Tendency

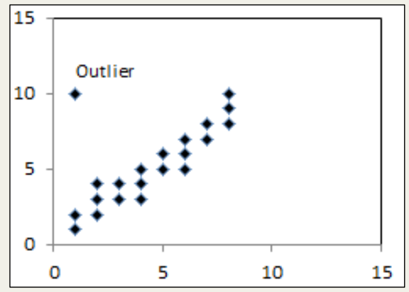
- Open the file “Manell_Geus_Data_Freqs.xlsx”
- Let’s examine a continuous variable like “tenure”
- We can calculate measures of central tendency and spread in excel
- In cell IT310, type “=AVERAGE(IT2:IT307)”
  - *21.54 is your mean*
- In cell IT311, type “=MEDIAN(IT2:IT307)”
  - *19 is your median*
- 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 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

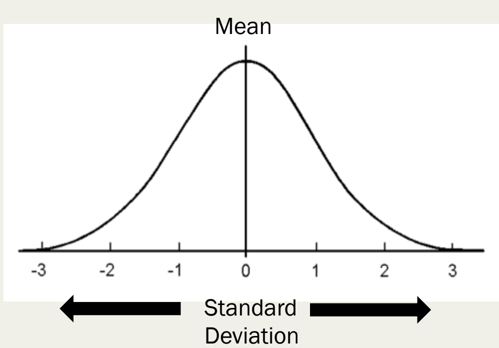
- Frequency Tables are the easiest way to provide a snapshot of categorical data
- They are meant to take a specific variable and break it down into its various possible values and examine the frequency of each
- This can show:
  - *What is the most frequent response?*
  - *What is the least frequent 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 IM310 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

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

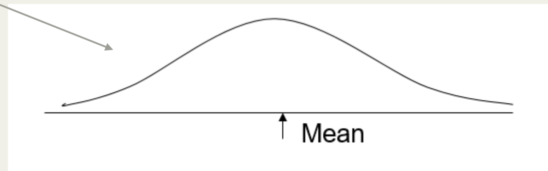
The Normal Distribution



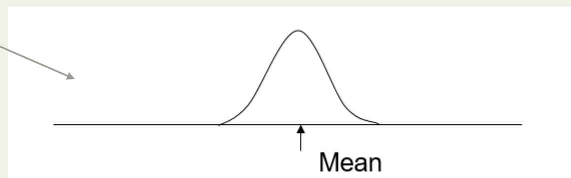
- An arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.
- Most statistical tests assume a normal distribution
  - Height is one simple example of something that follows a normal distribution pattern:
    - Most 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

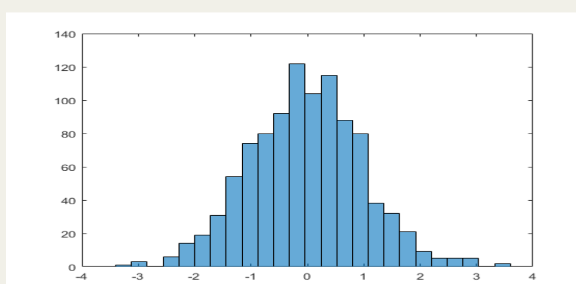


The smaller the standard deviation, the closer the individual scores are to the mean.



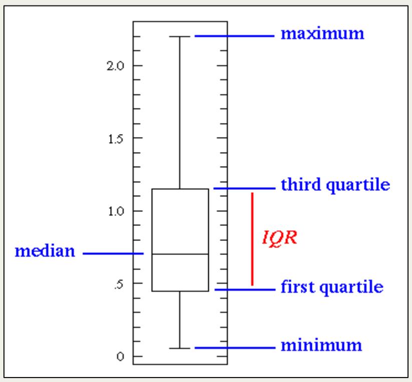
Histograms

- 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



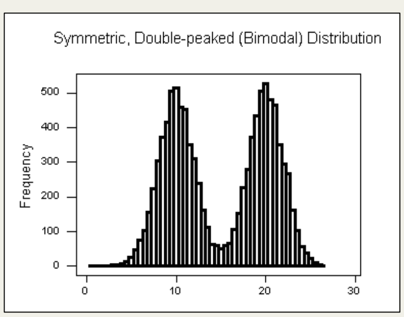
Box Plots

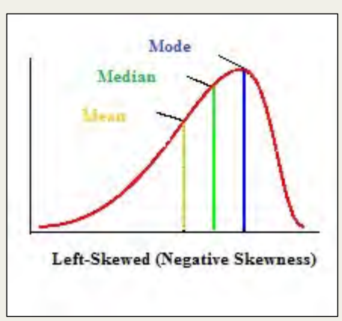
- A box plot is a graphical rendition of statistical data based on the minimum, first quartile, median, third quartile, and maximum
- The term "box plot" comes from the fact that the graph looks like a rectangle with lines extending from the top and bottom.

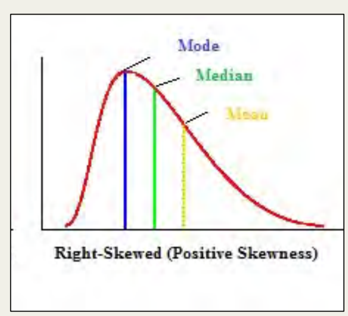


Other types of distributions

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







Descriptive Statistics in SPSS

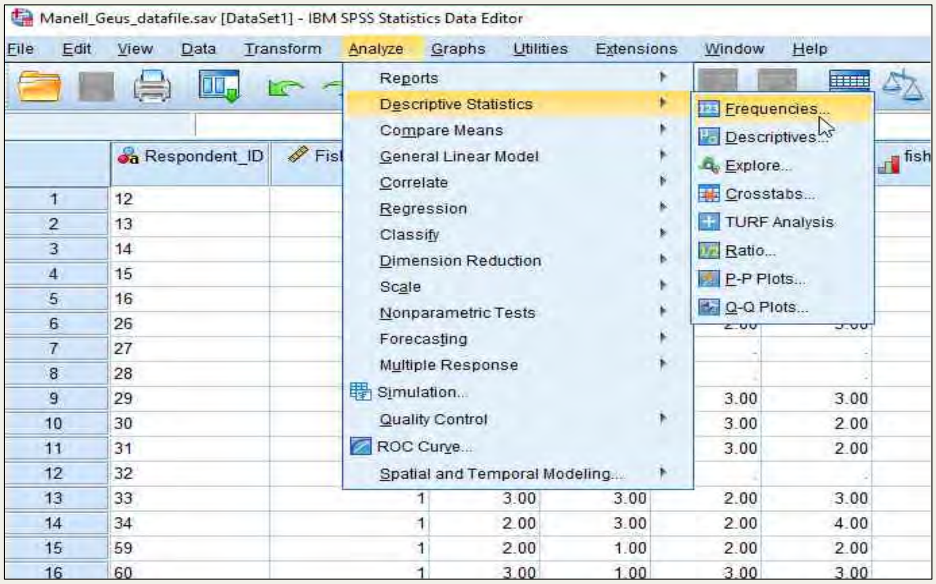
Day 2: September 13, 2016

Overview

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

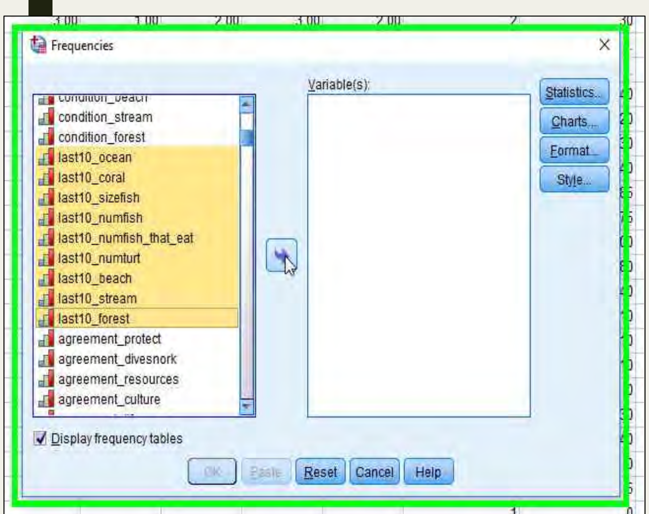
- Open "Manell_Geus_datafile.sav"

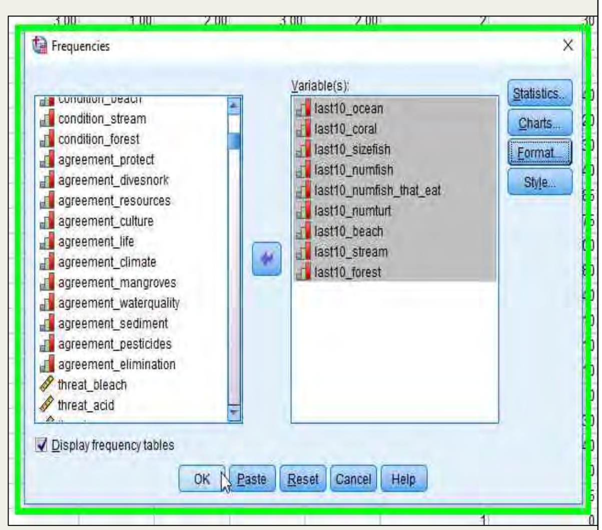
Frequency Tables



Frequency Tables – Last10

- Lets 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

last10_ocean

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	a lot worse	26	8.5	8.5	8.5
	somewhat worse	89	29.1	29.2	37.7
	no change	58	19.0	19.0	56.7
	somewhat better	101	33.0	33.1	89.8
	a lot better	22	7.2	7.2	97.0
	not sure	9	2.9	3.0	100.0
	Total	305	99.7	100.0	
Missing	System	1	.3		
Total		306	100.0		

last10_coral

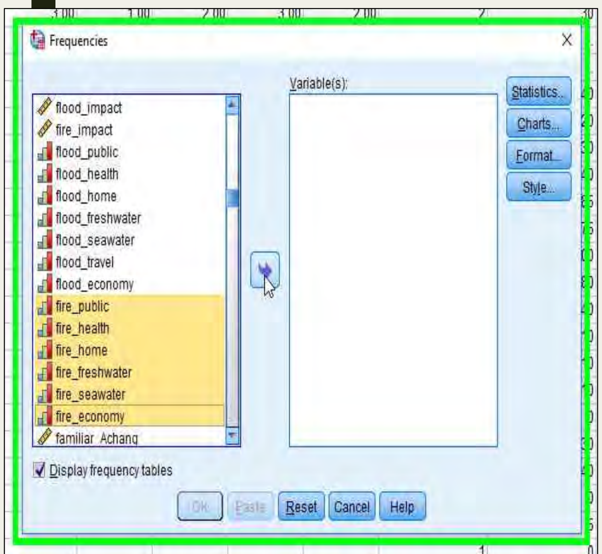
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	a lot worse	29	9.5	9.5	9.5
	somewhat worse	62	20.3	20.4	29.9
	no change	49	16.0	16.1	46.1
	somewhat better	111	36.3	36.5	82.6
	a lot better	33	10.8	10.9	93.4
	not sure	20	6.5	6.6	100.0
	Total	304	99.3	100.0	
Missing	System	2	.7		
Total		306	100.0		

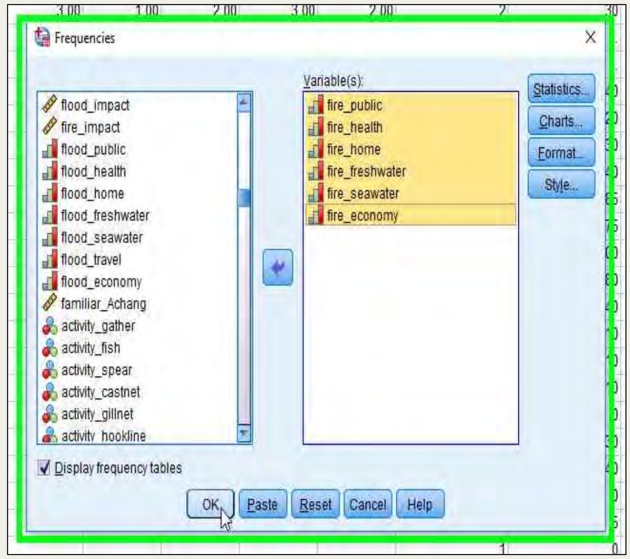
Some Conclusions about "Last10_"

- 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





Frequency Tables – Fire

fire_public

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	very low	36	11.8	11.8	11.8
	low	71	23.2	23.3	35.1
	medium	112	36.6	36.7	71.8
	high	65	21.2	21.3	93.1
	very high	18	5.9	5.9	99.0
	not sure	3	1.0	1.0	100.0
	Total	305	99.7	100.0	
Missing	System	1	.3		
Total		306	100.0		

fire_health

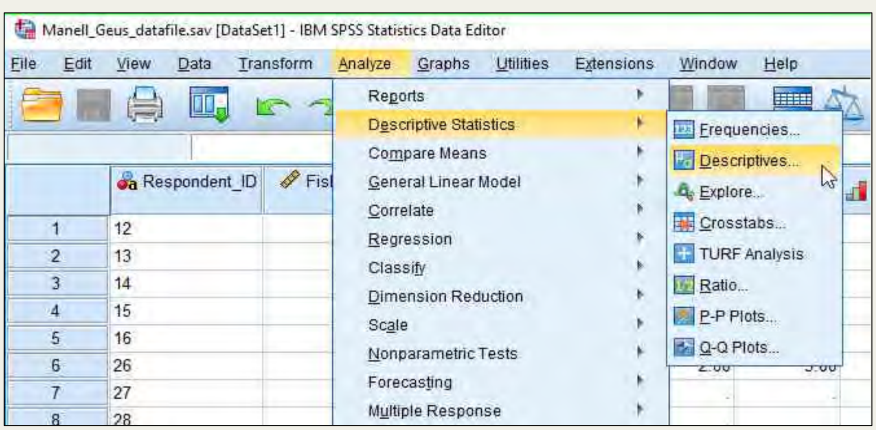
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	very low	36	11.8	11.9	11.9
	low	74	24.2	24.4	36.3
	medium	131	42.8	43.2	79.5
	high	36	11.8	11.9	91.4
	very high	21	6.9	6.9	98.3
	not sure	5	1.6	1.7	100.0
	Total	303	99.0	100.0	
Missing	System	3	1.0		
Total		306	100.0		

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

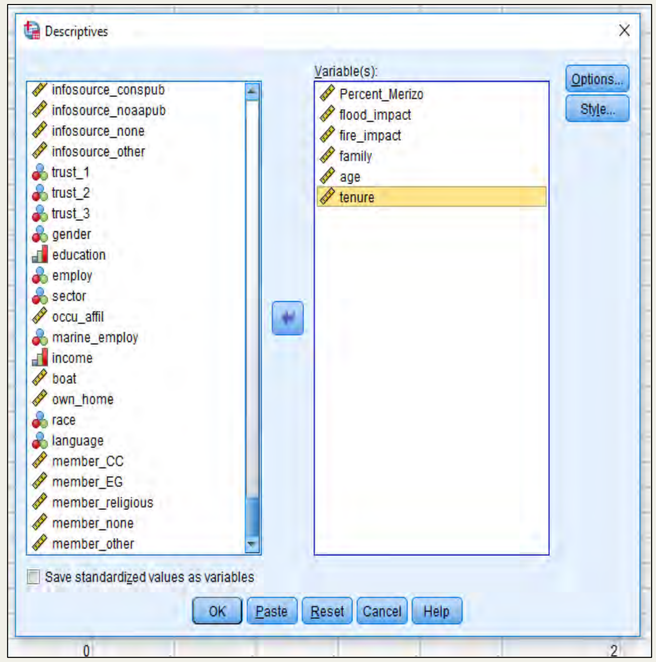
- While frequency tables are great for examining the distributions of nominal and ordinal variables, we want to use the "descriptives" function in SPSS for analyzing the distribution of continuous data



Summary Statistics

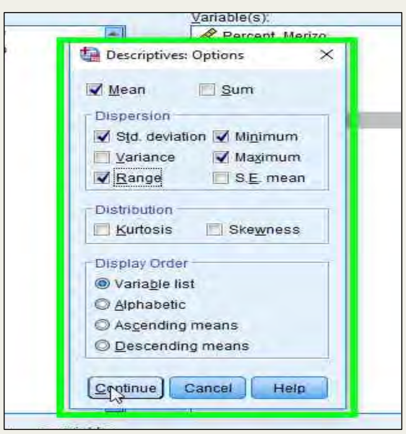
Let's analyze the descriptive statistics of some continuous data:

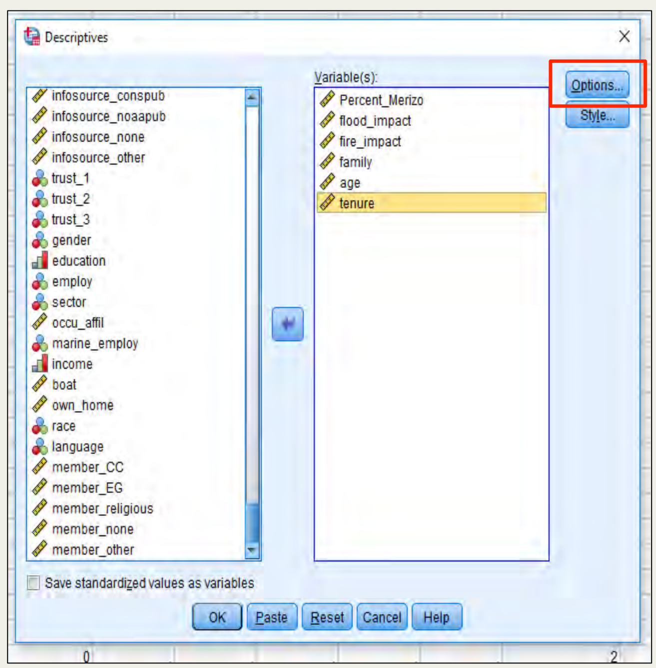
- *Percent_merizo*
- *Flood_impact*
- *Fire_impact*
- *Family*
- *Age*
- *Tenure*



Summary Statistics

- Click "options"
- Click "range"





Summary Statistics

Descriptives

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation
Percent_Merizo	283	100	0	100	33.35	32.037
flood_impact	243	20	0	20	2.27	3.462
fire_impact	232	60	0	60	3.16	8.166
family	303	14	1	15	6.35	2.733
age	304	62	18	80	40.38	15.040
tenure	298	68	0	68	21.55	15.933
Valid N (listwise)	202					

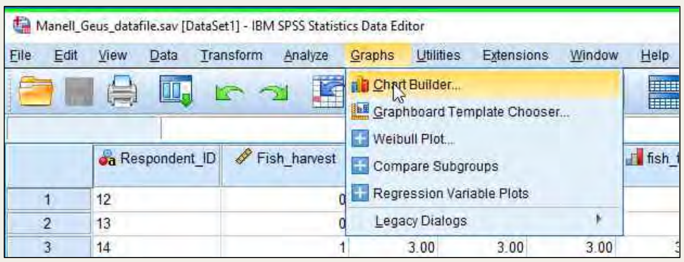
Summary Statistics

- 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)*

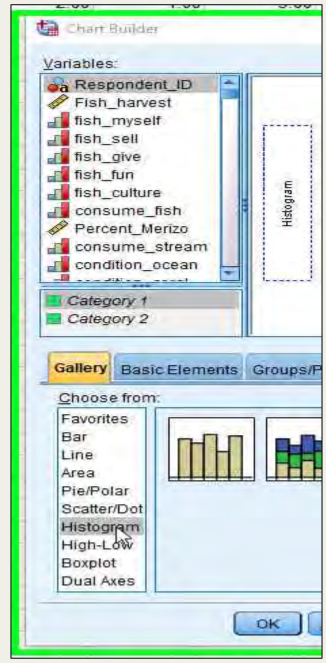
Histograms in SPSS

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

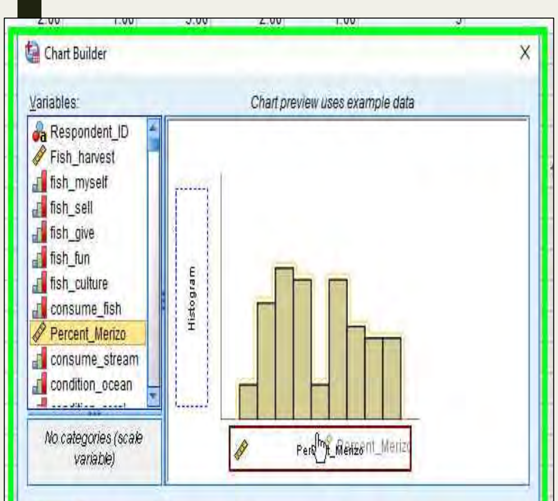
1

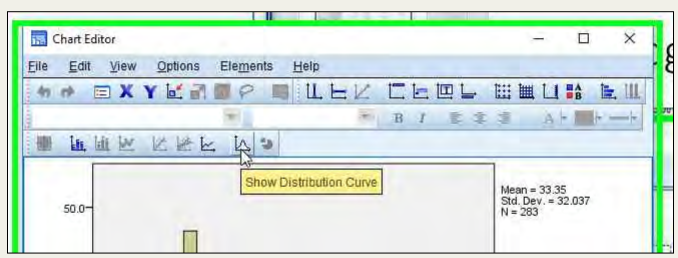


2



Histograms in SPSS

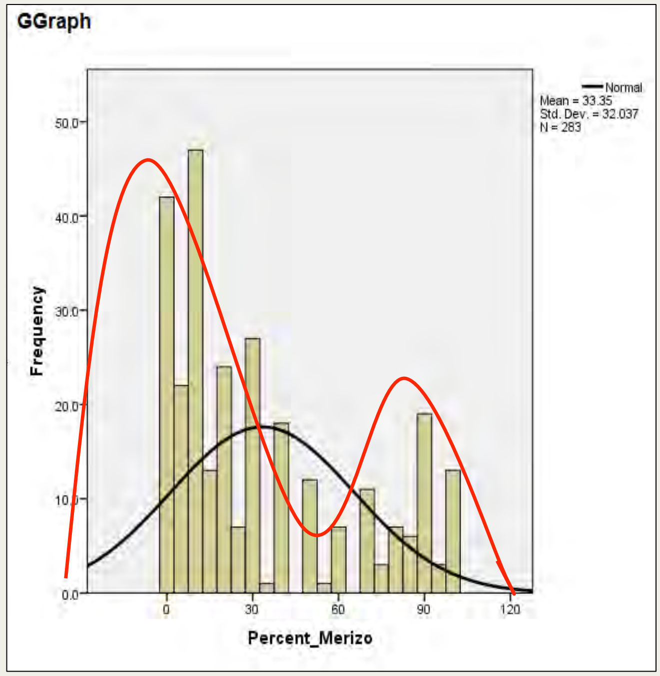




Then, double click the graph in the output viewer to pull up the "chart editor"

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 from 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%)

Frequency Table

success_ecosystem

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	very low	29	9.5	9.6	9.6
	low	68	22.2	22.5	32.1
	medium	111	36.3	36.8	68.9
	high	64	20.9	21.2	90.1
	very high	18	5.9	6.0	96.0
	not sure	12	3.9	4.0	100.0
	Total	302	98.7	100.0	
Missing	System	4	1.3		
Total		306	100.0		

success_water

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	very low	40	13.1	13.3	13.3
	low	70	22.9	23.3	36.5
	medium	111	36.3	36.9	73.4
	high	55	18.0	18.3	91.7
	very high	15	4.9	5.0	96.7
	not sure	10	3.3	3.3	100.0
	Total	301	98.4	100.0	
Missing	System	5	1.6		
Total		306	100.0		

Practice!

Generate a frequency table for “mng_gillnet” and “mng_dragnet”

- *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%)

Frequency Table

mng_gillnet

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	strongly oppose	63	20.6	20.7	20.7
	oppose	80	26.1	26.3	47.0
	neither support nor oppose	63	20.6	20.7	67.8
	support	63	20.6	20.7	88.5
	strongly support	28	9.2	9.2	97.7
	not sure	7	2.3	2.3	100.0
	Total	304	99.3	100.0	
Missing	System	2	.7		
Total		306	100.0		

mng_dragnet

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	strongly oppose	68	22.2	22.6	22.6
	oppose	79	25.8	26.2	48.8
	neither support nor oppose	53	17.3	17.6	66.4
	support	59	19.3	19.6	86.0
	strongly support	39	12.7	13.0	99.0
	not sure	3	1.0	1.0	100.0
	Total	301	98.4	100.0	
Missing	System	5	1.6		
Total		306	100.0		

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”?

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation
familiar_Achang	305	1	0	1	.56	.498
flood_impact	243	20	0	20	2.27	3.462
Percent_Merizo	283	100	0	100	33.35	32.037
agreement_protect	306	7	1	8	3.64	1.226
Valid N (listwise)	231					

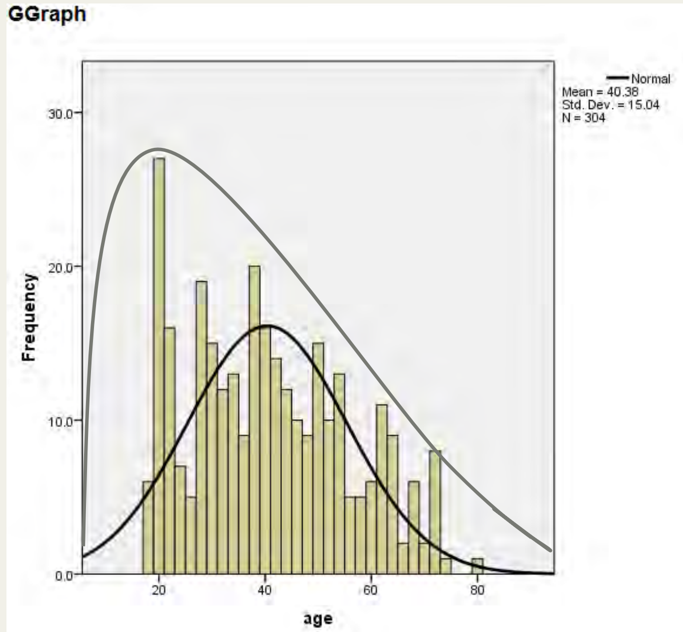
- What is the mean of "familiar_achang"?
  - *56% are familiar with Achang Preserve*
- What is the range of "flood_impact"?
  - *20 (Respondents have been affected by a range of zero to 20 floods in the last five years)*
- What is the interquartile range of "percent_merizo"?
  - *1st quartile = 10; 3rd 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

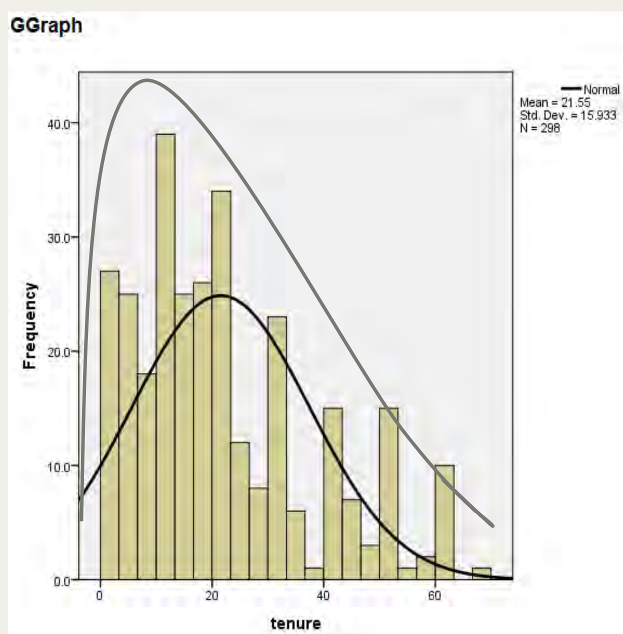
- Evidenced by the long right tail
- Median is less than the mean



“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

3

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)

4

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?

Tables	Graphs
• Look up individual values	• Message conveyed via shape
• Compare individual values	• Relationships among multiple values
• > 1 quantitative unit of measure	

6

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

7

Alphabetical sorting makes it hard to compare

NOT A Properly Formatted Table

No unit of measure

Most useful measure furthest from names

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

Sales Rep	Quota	Variance to Quota	% of Quota	Forecast	Actual Bookings
Albright, Gary	200,000	-16,062	92	205,000	183,938
Brown, Sheryll	150,000	84,983	157	260,000	234,983
Cartwright, Bonnie	100,000	-56,125	44	50,000	43,875
Caruthers, Michael	300,000	-25,125	92	324,000	274,875
Garibaldi, John	250,000	143,774	158	410,000	393,774
Girard, Jean	75,000	-48,117	36	50,000	26,883
Jone, Suzanne	140,000	-5,204	96	149,000	134,796
Larson, Terri	350,000	238,388	168	600,000	588,388
LeShan, George	200,000	-75,126	62	132,000	124,874
Levensen, Bernard	175,000	-9,267	95	193,000	165,733
Mulligan, Robert	225,000	34,383	115	275,000	259,383
Tetracelli, Sheila	50,000	-1,263	97	50,000	48,737
Woytisek, Gillian	190,000	-3,648	98	210,000	186,352

Grid too heavy – not needed

Derived from *Bookings* – should not appear before *Bookings* column

No total row. Can't assess overall performance

Source: Few, S. (2004)

A Properly Formatted Table

Sorted by *Bookings*

Grid eliminated - Simple rule lines

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

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

Total row added

Bookings moved and % of total added

Source: Few, S. (2004)

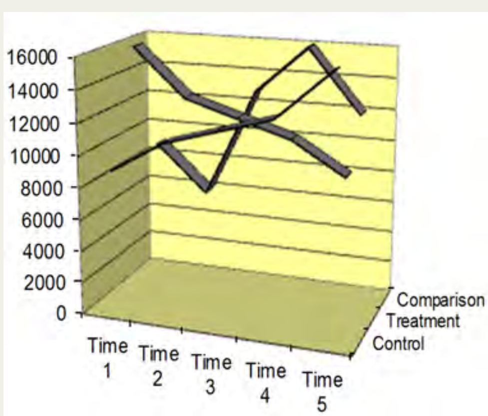
Graph selection and design

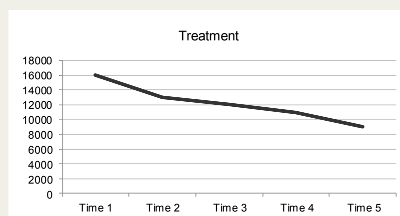
Type of data	Chart to use
• Time series or continuous data	• Line graph
• Discrete categories or few time points	• Bars/histograms
• Relative contribution mutually exclusive categories	• Stacked bar
• Proportions/Percentages	• Pie Chart
• Ranges, IQRs, Confidence Intervals	• Box Plot
• Correlation between 2 continuous variables	• Scatter plot

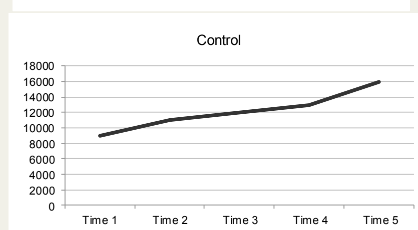
House of Commons library: Chart Format Guide

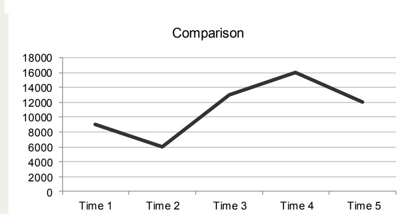
	Positive	Negative
Column	Simple, clear, recognisable, works for categories and time series	Trends less clear for very long time series, small space for long category names, inflexible
Bar	Simple, clear, recognisable, works for categories including those with long names, good for very large number of categories	Not appropriate for time series, less recognisable than column chart
Line	Simple, clear, recognisable, works for time series and index charts	Data markers can be clunky, not appropriate for category charts, interpolation of gaps, stacked charts not clear
Area	Recognisable, works for time series and stacked charts	Not category charts, less flexible and much more data ink than a line chart
Pie	Simple, recognisable	Difficulty in perceiving values especially with more than a few slices, needs labels, inflexible, cannot be combined with other types, no time series charts, never looks right in Excel
Scatter	The only choice for comparing two variables, correct interpretation of date values	Limited other uses

Avoid 3-D





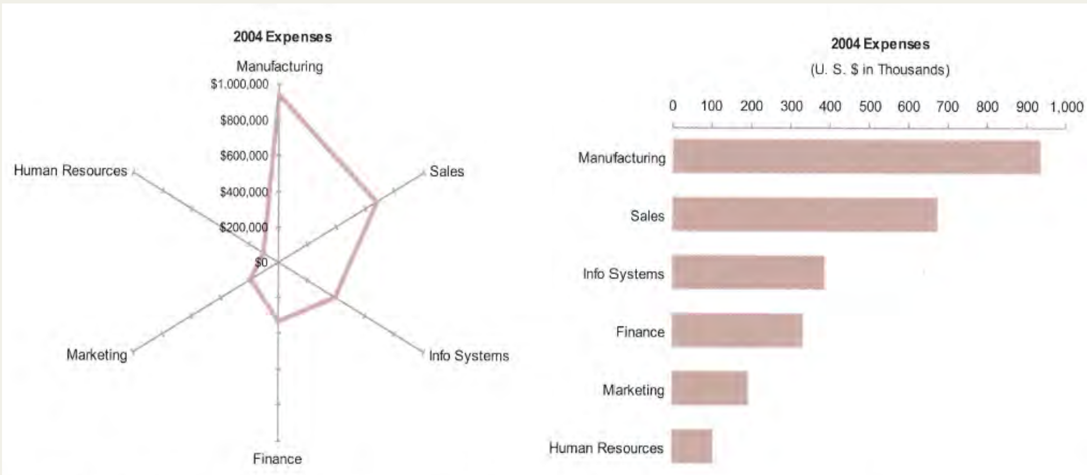




Smith & Azzam (2010)

12

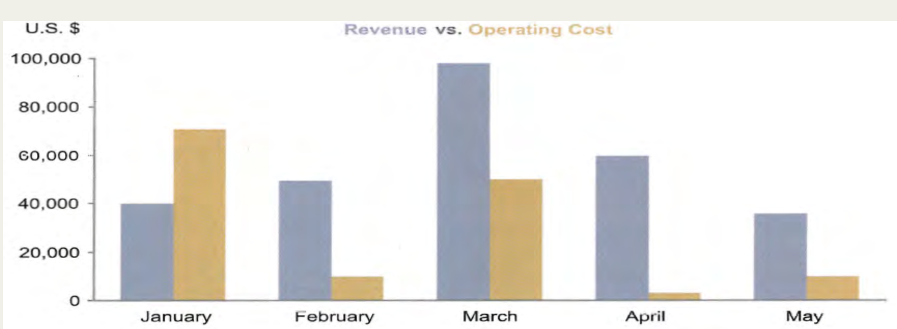
Radar graph vs. bar graph



Source: Few, S. (2004)

13

Bars are better than circles for relative size



Source: Few, S. (2004)

14

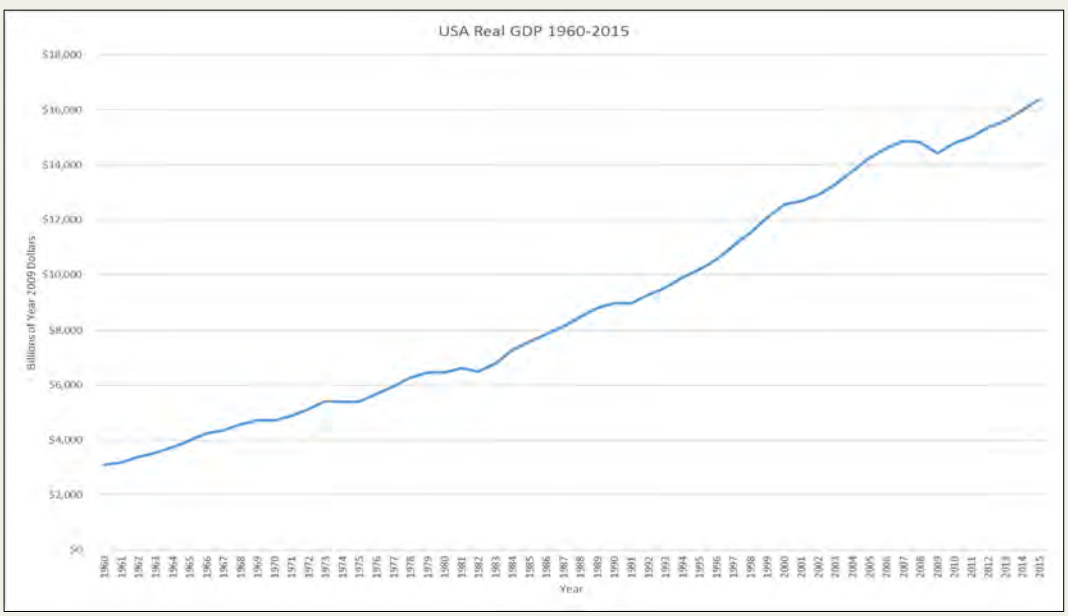
Making Graphs in Excel – Line Graph

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Table 1.1.6. Real Gross Domestic Product, Chained Dollars**											
2	[Billions of chained (2009) dollars]											
3	Bureau of Economic Analysis											
4	Last Revised on: July 29, 2016 - Next Release Date August 26, 2016											
5												
6		**1960**	**1961**	**1962**	**1963**	**1964**	**1965**	**1966**	**1967**	**1968**	**1969**	**1970**
7	**Gross domestic product**	3108.7	3188.1	3383.1	3530.4	3734	3976.7	4238.9	4355.2	4569	4712.5	4722
8												

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”

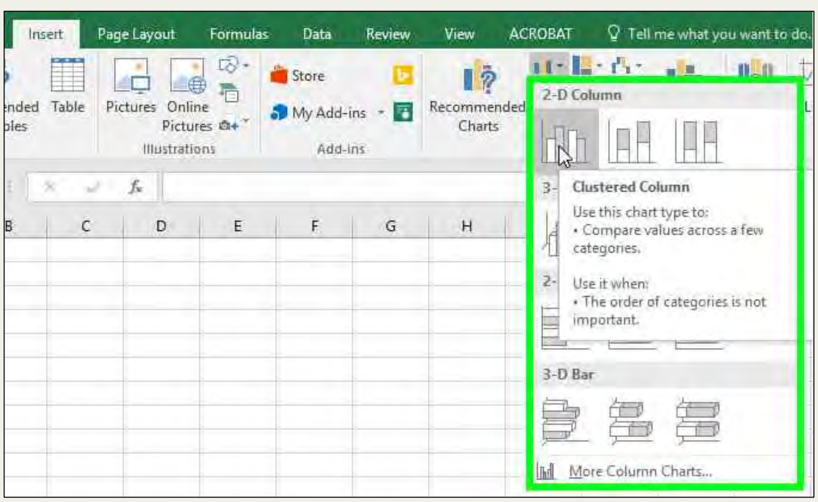
Making Graphs in Excel – Line Graph



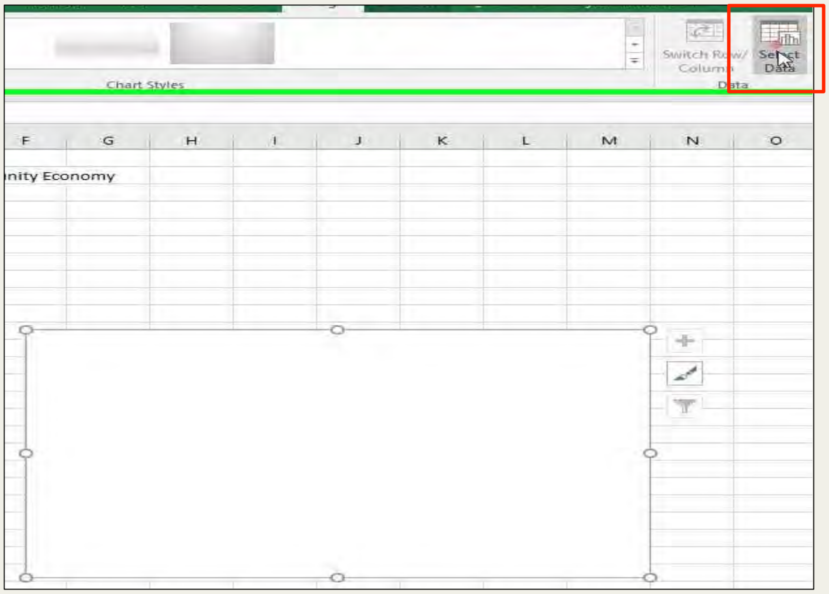
Making Graphs in Excel – Bar Graph

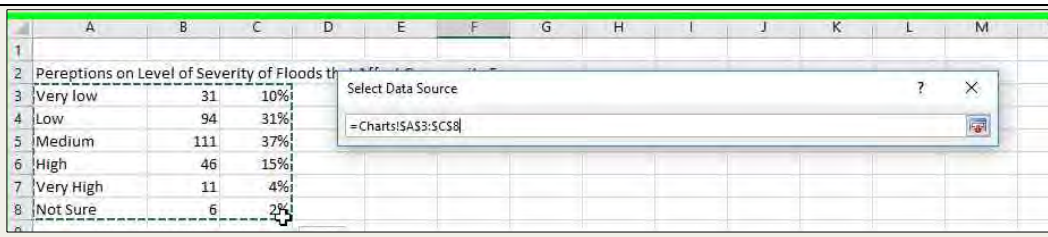
- Open "Manell_Geus_data_Graphs.xlsx"
- Let's examine "flood_economy"
- This question asks respondent to rate their opinion on the level of severity of floods as it pertains community economy
- Make a new sheet and name it "charts"
- Set your data up correctly on the "charts" sheet with the COUNTIF function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

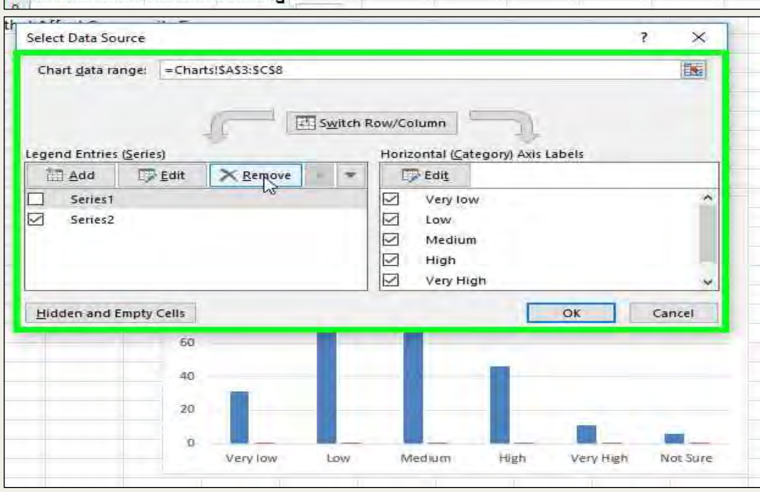
Making Graphs in Excel – Bar Graph



Making Graphs in Excel – Bar Graph

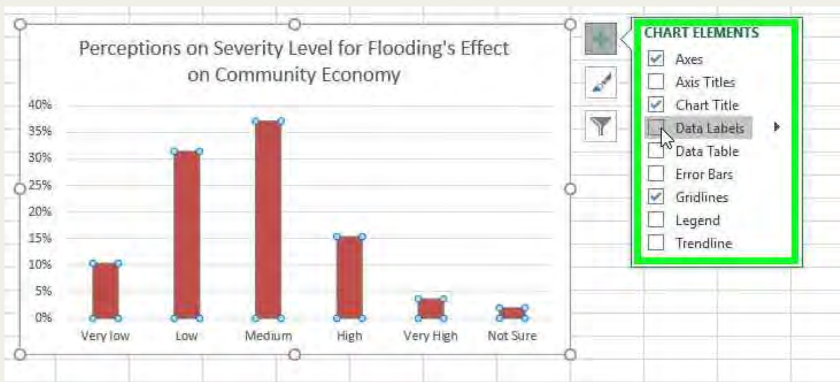




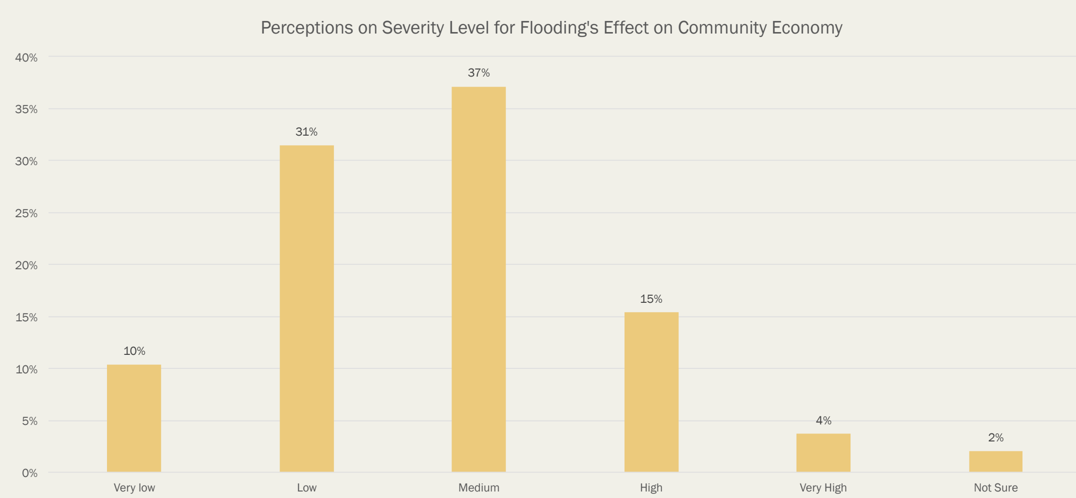


Making Graphs in Excel – Bar Graph

Making Graphs in Excel – Bar Graph



Making Graphs in Excel – Bar Graph

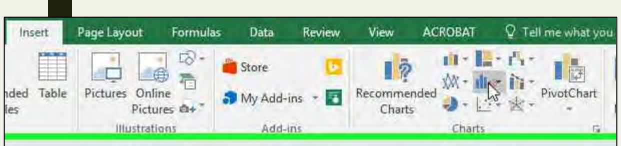


Making Graphs in Excel – Histogram

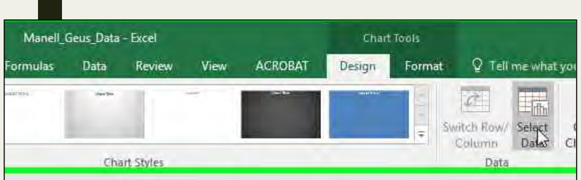
- Histograms display continuous data broken into bins, and then illustrate the frequency of occurrence for each bin
- Let's examine "tenure"
- In the "Charts" worksheet, Click on the "Insert" Tab and go to the "Charts" section
- Click the statistical charts icon

Making Graphs in Excel – Histogram

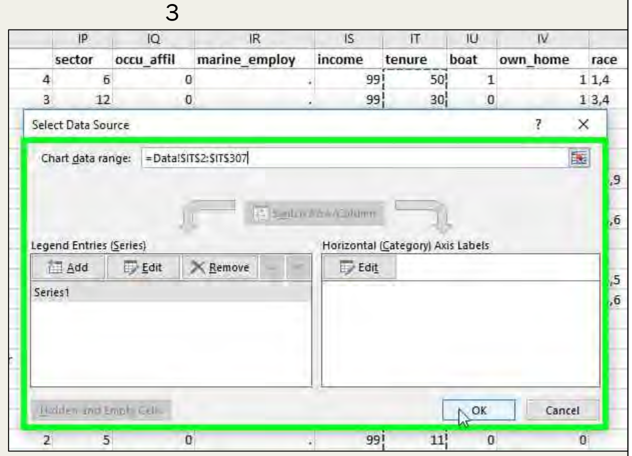
1



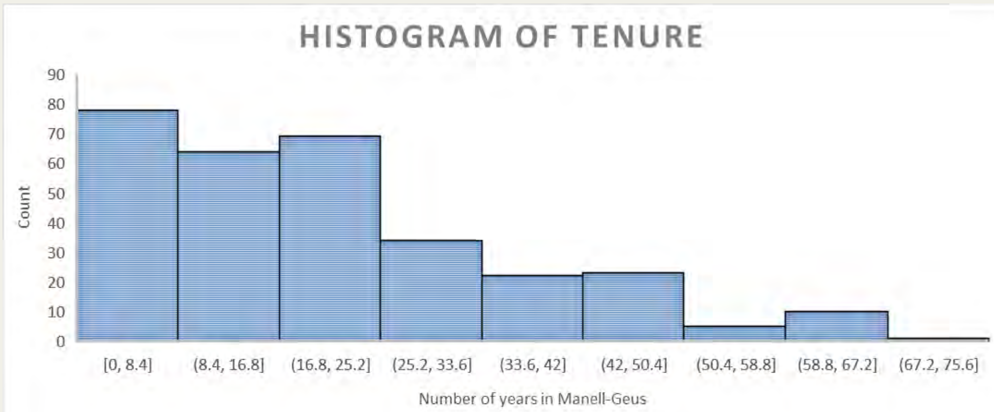
2



3

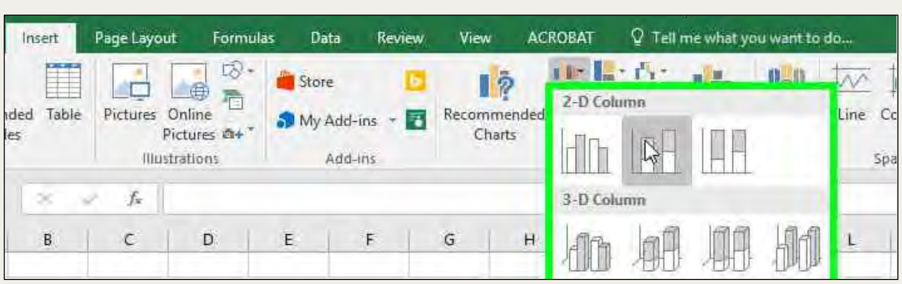


Making Graphs in Excel – Histogram



Making Graphs in Excel – Stacked Bar Graph

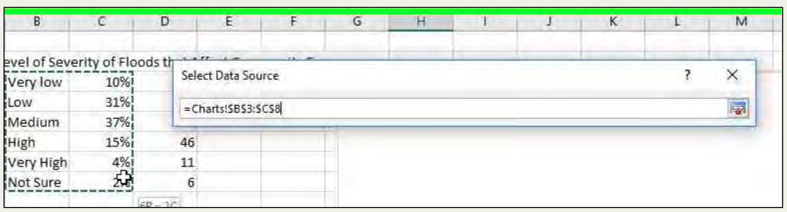
- We can examine “flood_economy” again since our data is already set up
- Follows similar procedure to that of a regular bar graph
- Move percentages directly to the right of the labels (very low, low, etc.)



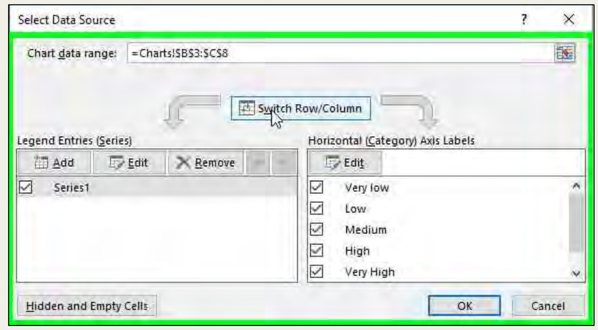
Making Graphs in Excel – Stacked Bar Graph

1

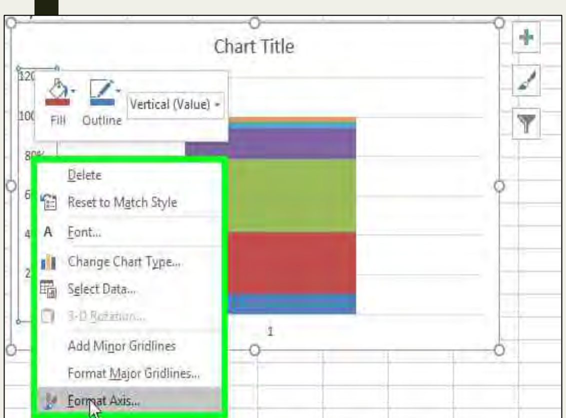
2

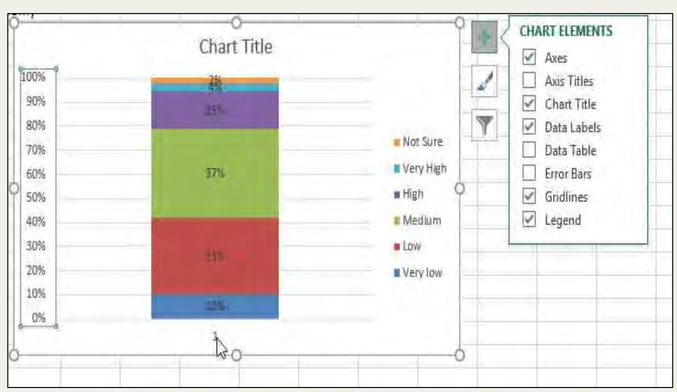


3

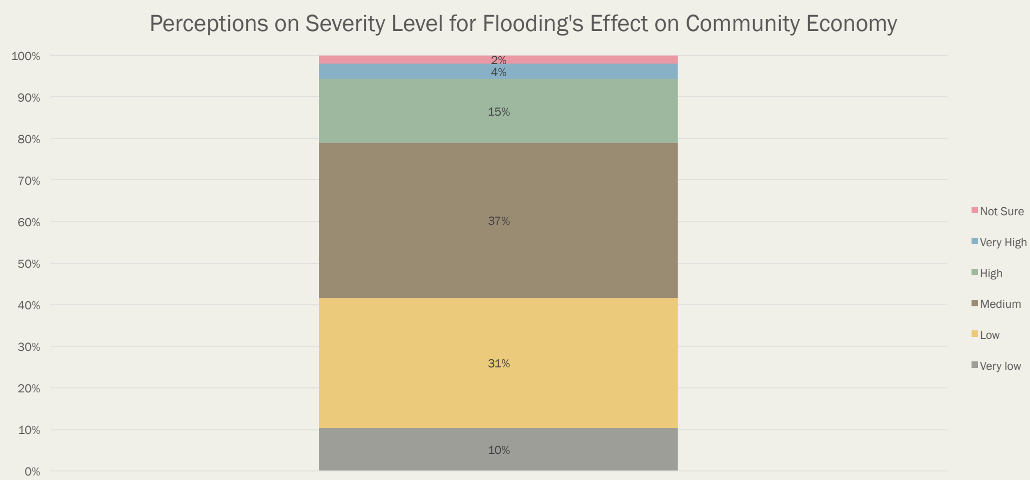


Making Graphs in Excel – Stacked Bar Graph





Making Graphs in Excel – Stacked Bar Graph

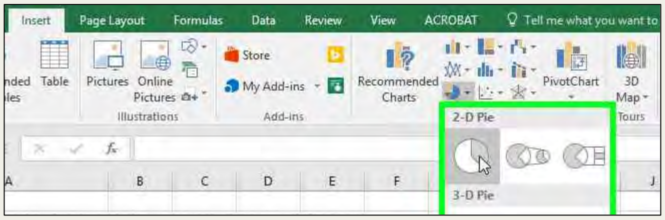


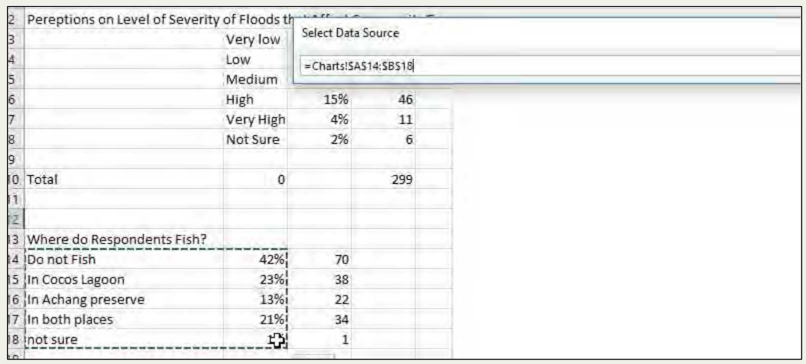
Making Graphs in Excel – Pie Chart

- 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

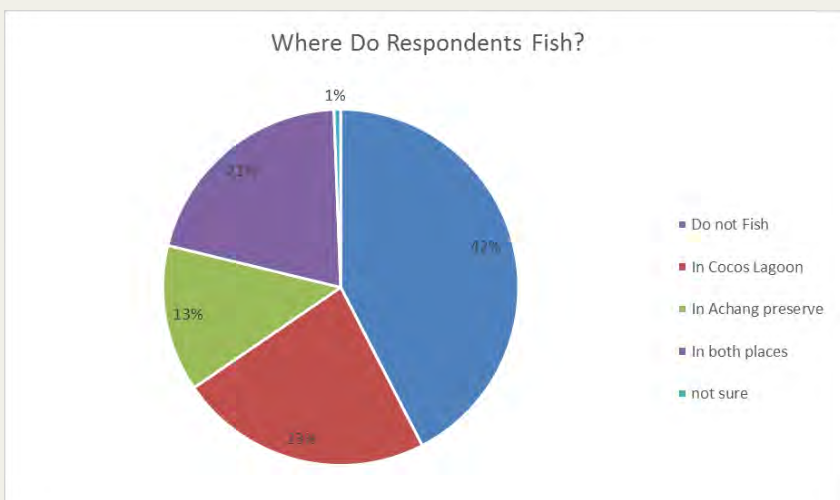
13	Where do Respondents Fish?		
14	Do not Fish	42%	70
15	In Cocos Lagoon	23%	38
16	In Achang preserve	13%	22
17	In both places	21%	34
18	not sure	1%	1
19			
20			165

Making Graphs in Excel – Pie Chart



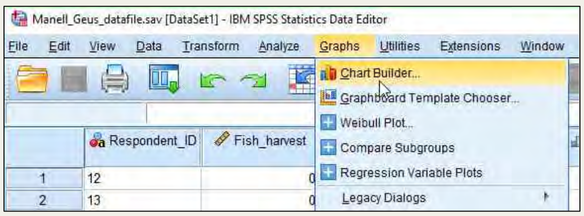


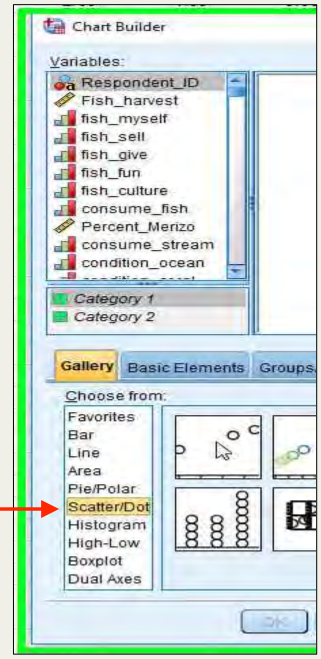
Making Graphs in Excel – Pie Chart



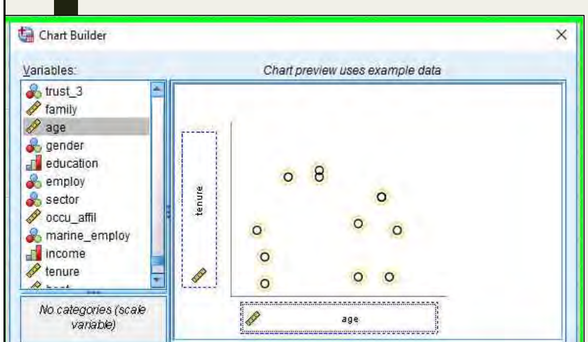
Making Graphs in SPSS – Scatter Plot

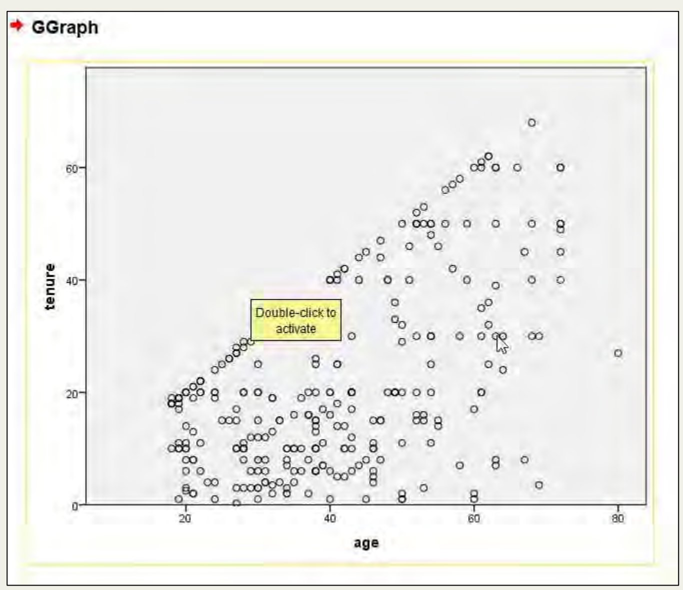
- We'll do Scatter Plots in SPSS, since we are now examining a bivariate relationship
- Open "manell_geus_datafile.sav"



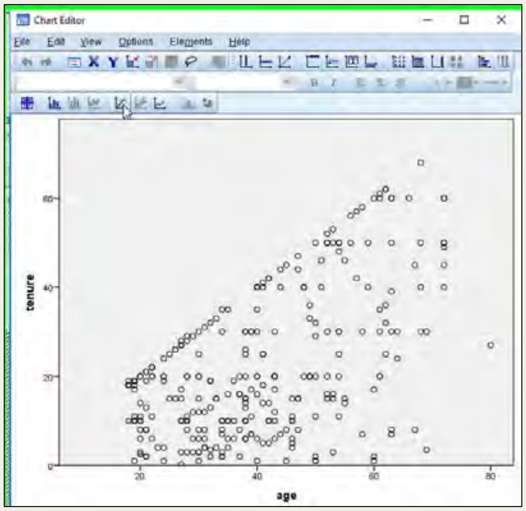


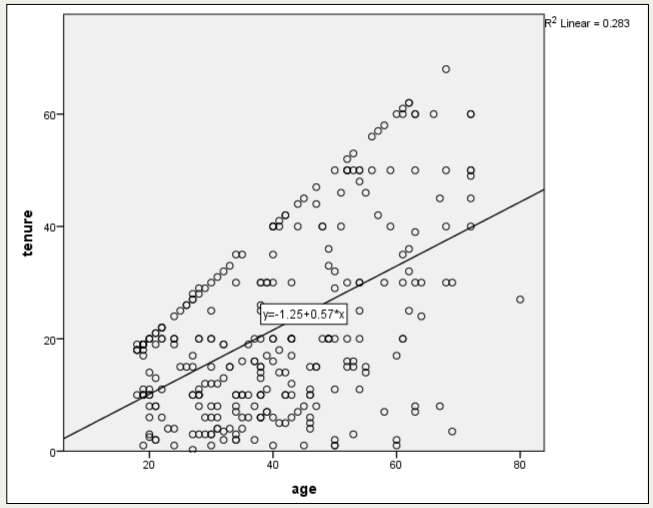
Making Graphs in SPSS – Scatter Plot





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*

Quiz #4

Day 2: September 13, 2016

4.1 Which of the following are measures of central tendency?

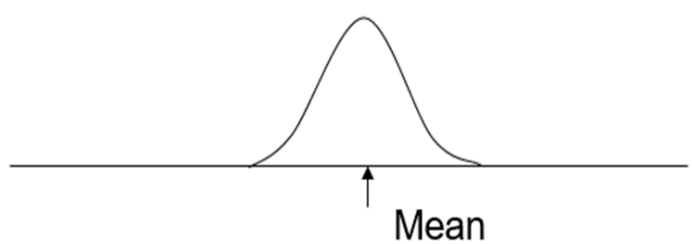
A. Range
B. Median
C. Mode
D. Variance
E. Maximum
F. Mean

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

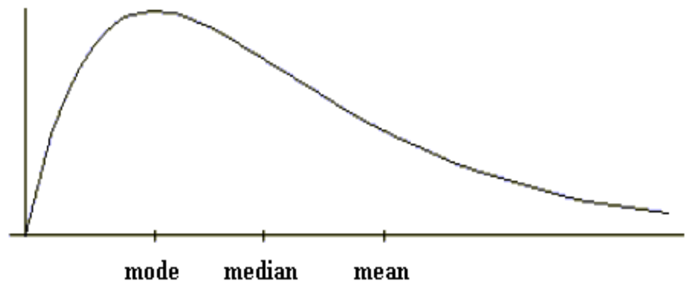
4.3 True or false: A normally distributed data set with a smaller standard deviation will have a narrower, taller bell shape

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

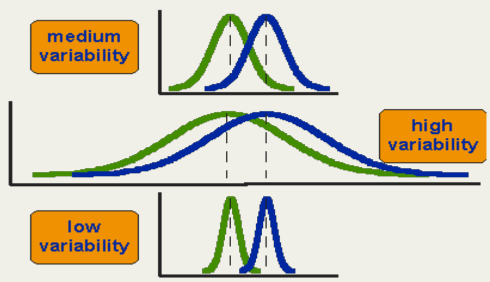
- Cluster sampling
  - *Every member of the population is assigned to one, and only one, group*
  - *Each group is called a cluster*
  - *A sample of clusters is chosen, using a probability method (often simple random sampling)*
  - *Only individuals within sampled clusters are surveyed*
- Multistage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

$$\overline{X} \pm t \frac{s}{\sqrt{n}}$$

- 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 $(X_1 \text{ and } X_2)$" ; Where $X_2 > X_1$
- 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 (α; 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

		Independent Variable		
		Nominal	Ordinal	Interval/Ratio
Dependent Variable	Nominal	*Crosstabs* Chi-square Lambda	*Crosstabs* Chi-square Lambda	
	Ordinal	*Crosstabs* Chi-square Lambda	*Crosstabs* Chi-square Lambda Gamma Kendall's tau Sommers' *d*	
	Interval/Ratio	*Means* *t*-test ANOVA	*Means* *t*-test ANOVA	*Correlate* Pearson *r* Regression (*R*)

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

The totals around the perimeter = "marginals"

Dependent Variable	Independent Variable: IV Attribute Lo	Independent Variable: IV Attribute Hi	Total
DV attribute Lo	A	B	A + B
DV attribute Hi	C	D	C + D
Total	A + C	B + D	N = A+B+C+D

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 10.2 How to Interpret Measures of Association

Measure of Association (X)	*Qualitative Interpretation*
$0 \leq X < 0.10$	Very Weak
$0.10 \leq X < 0.20$	Weak
$0.20 \leq X < 0.30$	Moderate
$X \geq 0.30$	Strong

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*X*
    where Y = estimated dependent variable
    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_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*"*

Multiple Linear Regression

- Model given by:
  – *Y = c + + + + … +*
- 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 is significant, that "all else held equal, significantly effects Y"*
  – *If changes by 1, we expect Y to change by*
  – *If changes by 1, we expect Y to change by*
  – *If = = = = 0, then we expect Y = c*

Variable Transformations

Day 3: September 14, 2016

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*

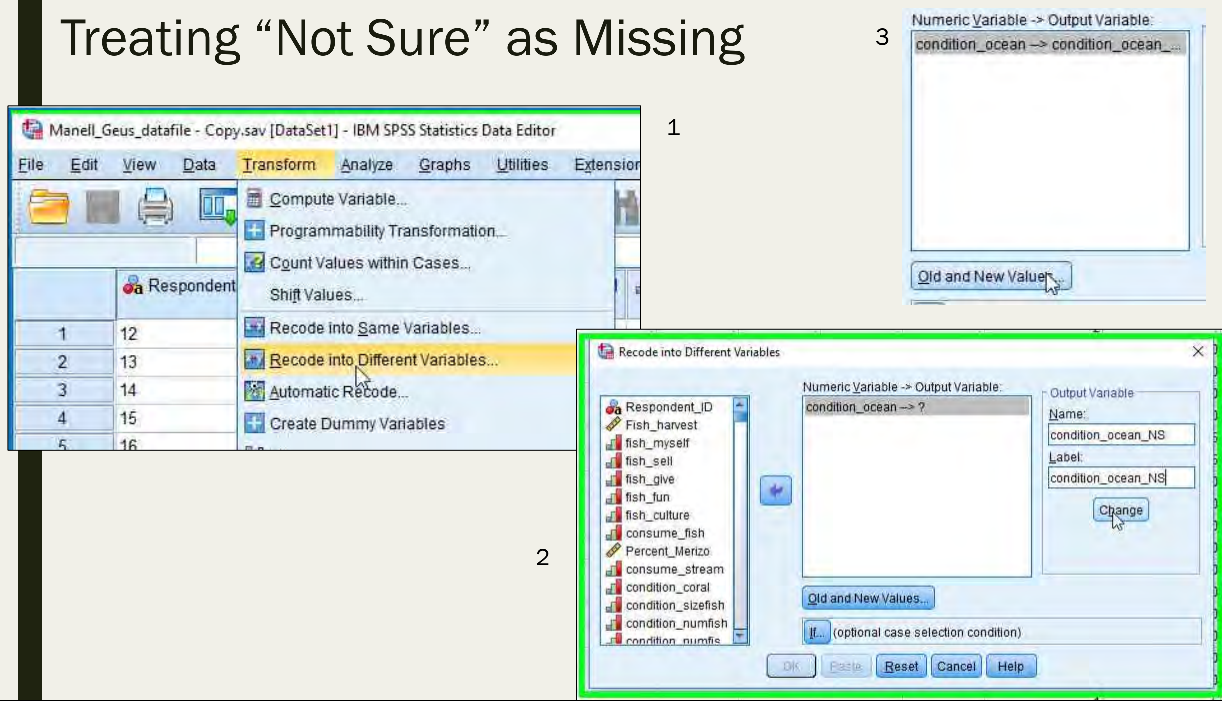
Preparation for Transformation

- 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*

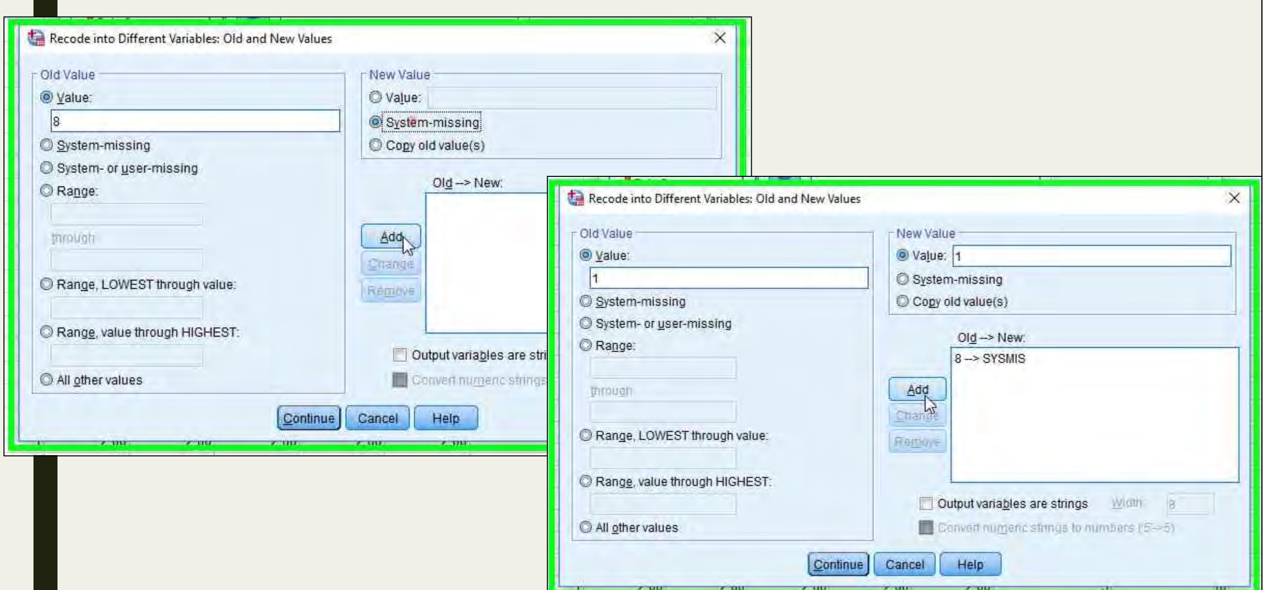
How to Treat “Not Sure”

- 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”

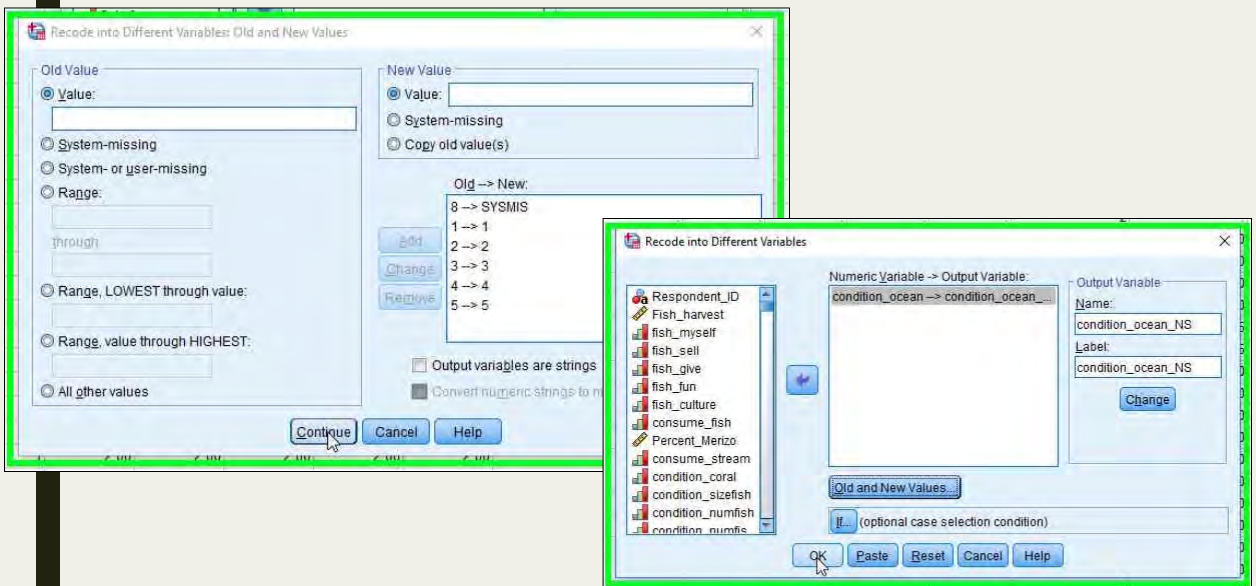
Treating “Not Sure” as Missing



Treating “Not Sure” as Missing



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

condition_ocean_NS
1.00
4.00
2.00
.
1.00
1.00
1.00
2.00
2.00
3.00
4.00
4.00
3.00
3.00
4.00
4.00
4.00
4.00
4.00
4.00
3.00
2.00
4.00
2.00
2.00
2.00
4.00
4.00

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*

The Dummy Variable

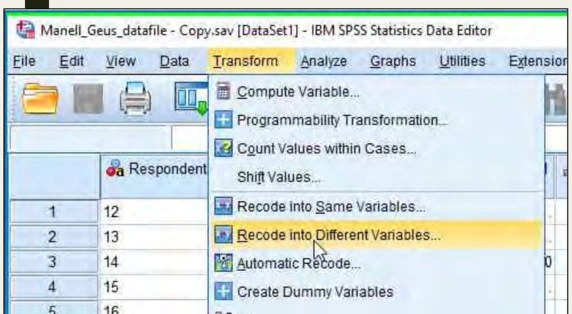
- *For example, to analyze "consume_fish" (How often does your family eat fish/seafood?), we can change the coding from:*
  - Almost never/never = 1
  - A few times a year = 2
  - Once a month = 3
  - A few times a month = 4
  - At lest once a week = 5
  - Almost daily = 6
  - Not Sure = 8

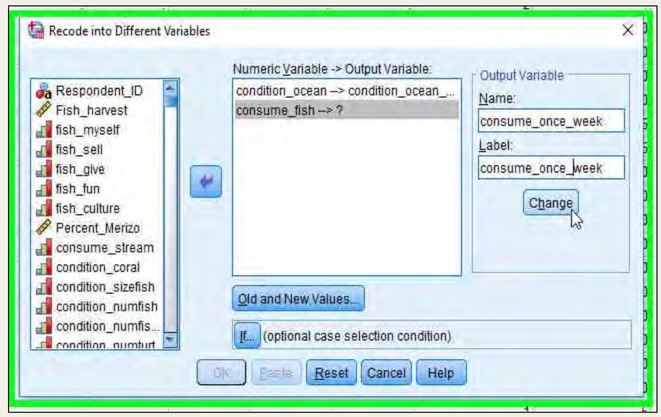
→

- 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 (.)

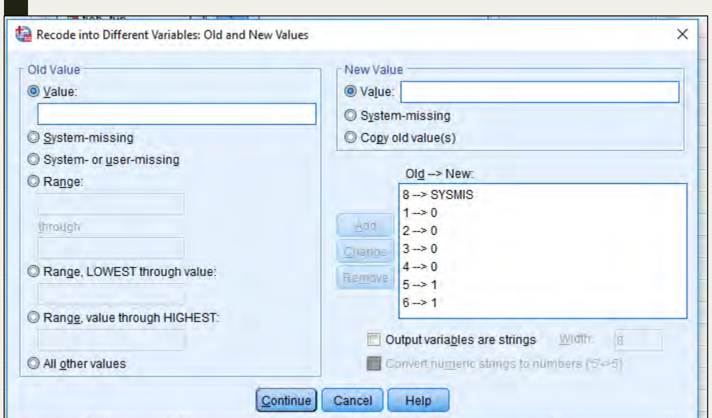
- *This transformation allows us to analyze this ordinal variable as a continuous variable*
- *We can now calculate the percentage of people that consume seafood AT LEAST once per week and we can try to see how this variable relates to other variables in the data through quantitative analysis*
- *We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

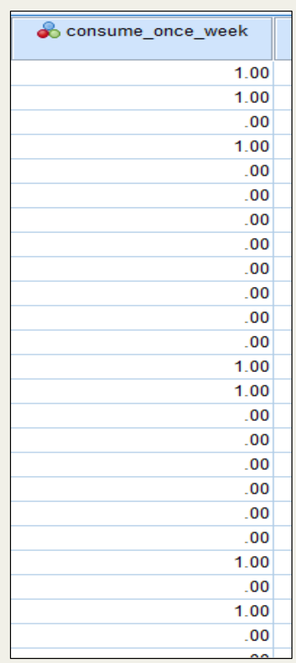
Dummy Transformation in SPSS





Dummy Transformation in SPSS





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

419	VARIABLE TRANSFORMATION STIPULATIONS	
423	consume_once week	Dummy variable representing if respondent eats seafood at least once per week

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

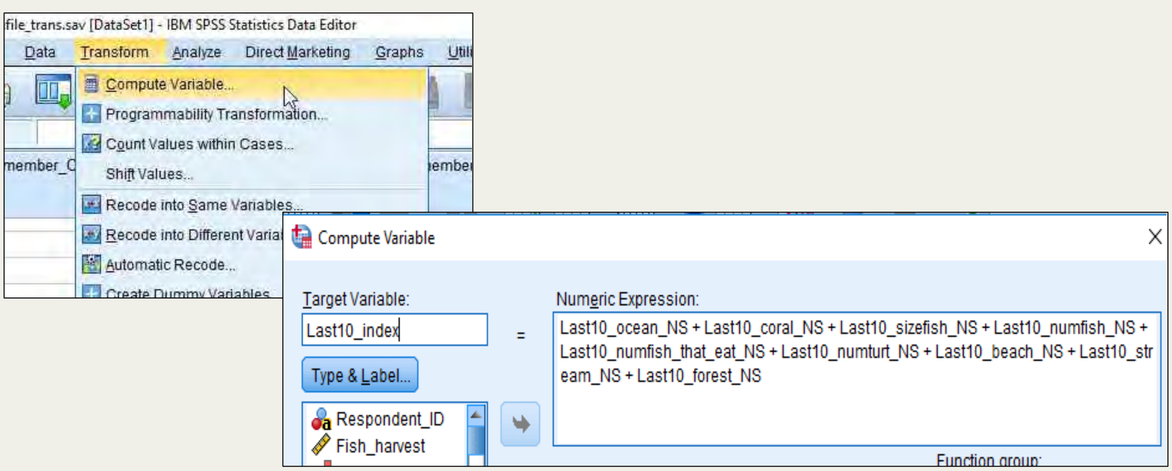
- Last 10 years index
  - *Since each of the individual questions that comprise this index are coded ordinally from 1-5 (from more negative opinion to more positive opinion), this index variable will have "direction"*
  - *Direction = as the index increases, positive perception concerning the change 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

last10_index
.
.
21.00
.
17.00
20.00
.
16.00
.
24.00
25.00
29.00
21.00
18.00
31.00
43.00
37.00
30.00
31.00
30.00
32.00
29.00
45.00
16.00
17.00
27.00
27.00

419	VARIABLE TRANSFORMATION STIPULATIONS		
426	Last 10 Index	Additive index of "last10_" questions	higher index=more positive opinion

The Index Variable: Last 10 Years Index

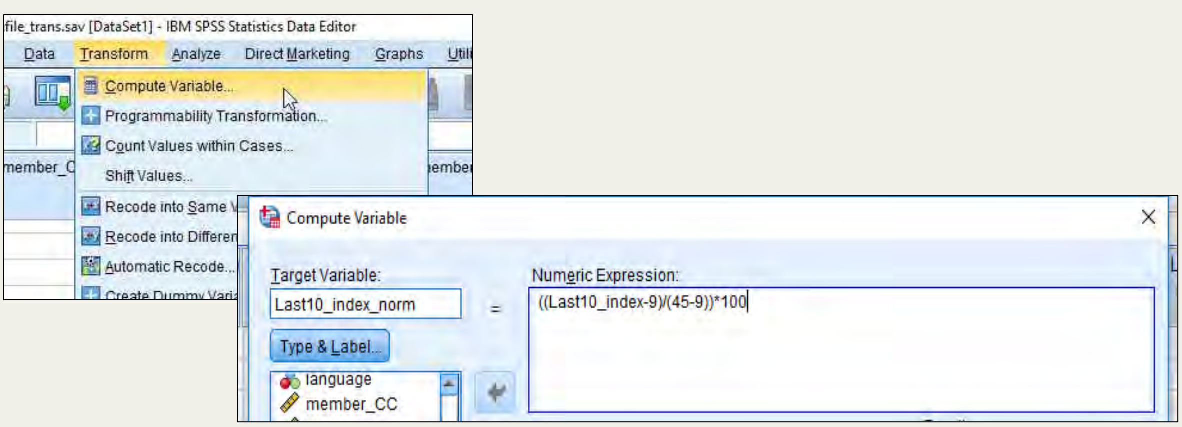
- Let's run some summary statistics on this new index variable and see what we find.....
- The range is 9-45
  - *Makes sense since the index is made of 9 questions that range individually from 1-5*
  - *Minimum: 9*1 = 9; Maximum 9*5 = 45*
- Mean = 27.7
- Midpoint of index: 9 + (45-9)/2 = 27
- Perception 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
  - $X_{norm} = \frac{X - min_x}{max_x - min_x}$
  - *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

last10_index	last10_index_norm
.	.
.	.
21.00	33.33
.	.
17.00	22.22
20.00	30.56
.	.
16.00	19.44
.	.
24.00	41.67
25.00	44.44
29.00	55.56
21.00	33.33
18.00	25.00
31.00	61.11
43.00	94.44
37.00	77.78

419	VARIABLE TRANSFORMATION STIPULATIONS
421	Any index with "_norm" after it represents the index normalized on a 0-100 scale

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

Practice!

*Save your output as "Manell_Geus_Output_trans"

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
condition_ocean_NS	299	1	5	3.08	1.097
condition_coral_NS	289	1	5	3.21	1.093
condition_sizefish_NS	293	1	5	3.26	1.174
condition_numfish_NS	286	1	5	3.36	1.126
condition_numfish_that_eat_NS	276	1	5	3.34	1.095
condition_numturt_NS	276	1	5	3.30	1.166
condition_beach_NS	298	1	5	2.90	1.107
condition_stream_NS	294	1	5	2.90	1.076
condition_forest_NS	292	1	5	3.05	1.070
Condition index_norm	246	0	100	53.97	22.026
Valid N (listwise)	246				

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

Quiz #5

Day 3: September 14, 2016

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. $X_{norm} = X - max_x / max_x - min_x$
B. $X_{norm} = X - min_x / max_x - min_x$
C. $X_{norm} = min_x - X / max_x - min_x$

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 infogrpahics for outreach purposes)

What Questions Can We Answer?

5.

Best Way to Answer Each Question?

Hypothesis for Each Question

Best Way to Communicate Each Result?

Inferential Stats in SPSS

Day 3: September 14, 2016

Exploratory Analysis

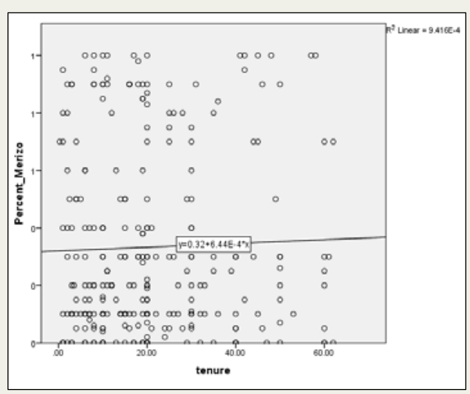
- Exploratory Data Analysis is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to
  - *maximize insight into a data set*
  - *uncover underlying structure*
  - *extract important variables*
  - *detect outliers and anomalies*
  - *test underlying assumptions*
- "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
- Lets’ 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

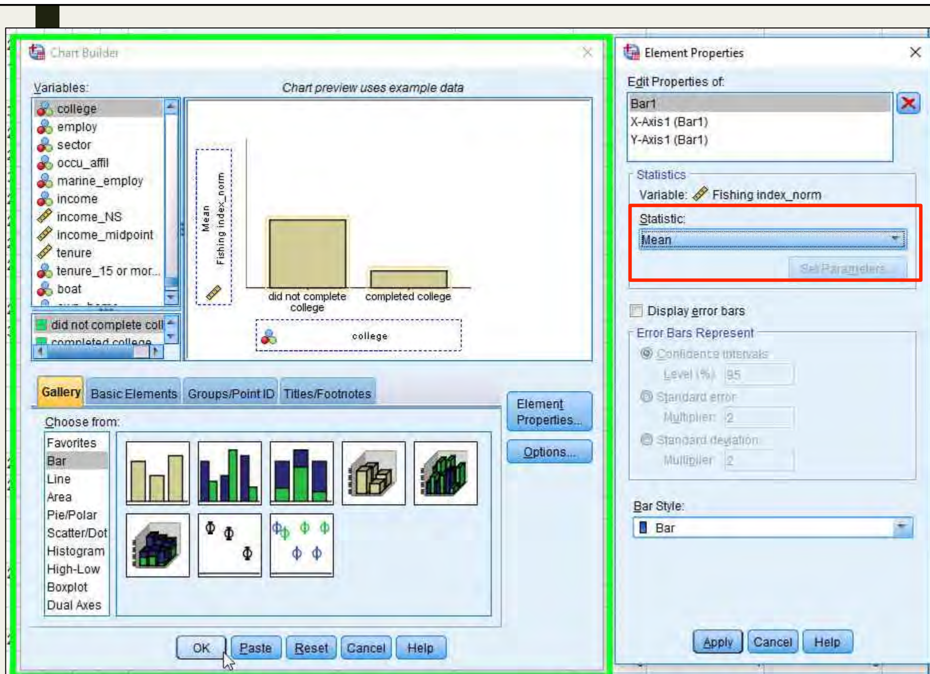
Scatter Plots – Percent_Merizo and Tenure

- 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

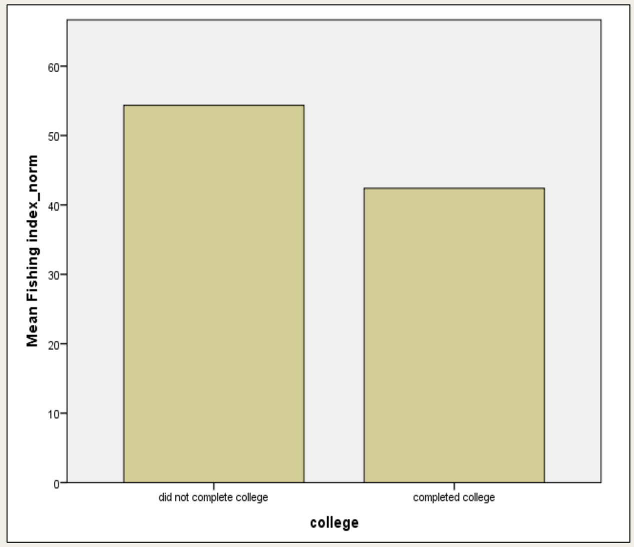
- A quick way to graphically display frequencies or means for different groups
- Let’s examine “fishing index_norm” and “college”
  - *To see if there is a difference between the average fishing index for those who completed college and those who did not complete college*
- Graphs > Chart builder > bar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

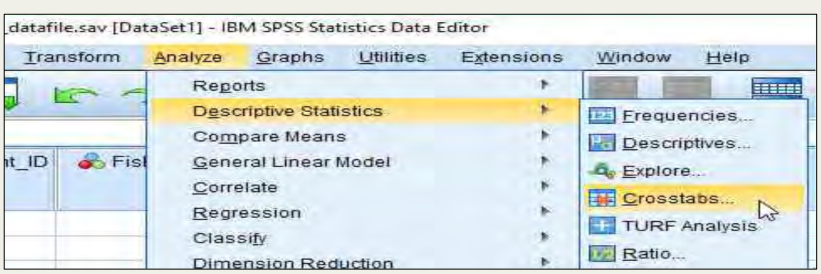
Dual Bar Graphs

- 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"



Basic 2x2 Contingency Table

Dependent Variable	Independent Variable		
	IV Attribute Lo	IV Attribute Hi	Total
DV attribute Lo	A	B	A + B
DV attribute Hi	C	D	C + D
Total	A + C	B + D	N = A+B+C+D

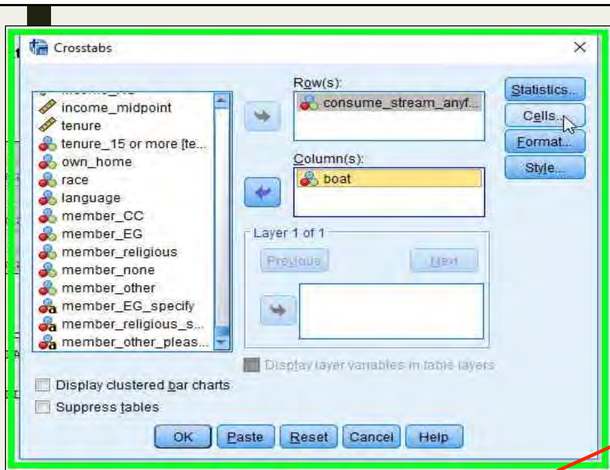
The totals around the perimeter = “marginals”

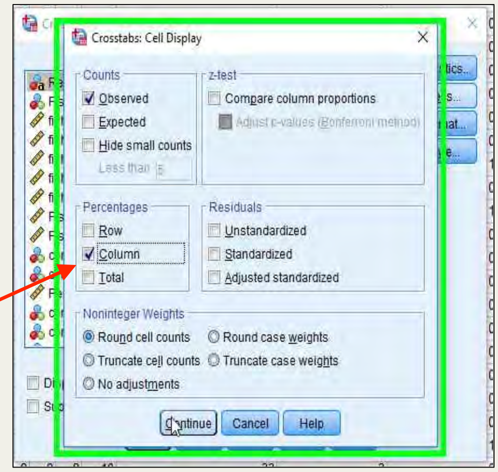
The cross-classifications = “cells”

N = “grand total”

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





- *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

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
consume_stream_anyfreq * boat	295	96.4%	11	3.6%	306	100.0%

consume_stream_anyfreq * boat Crosstabulation

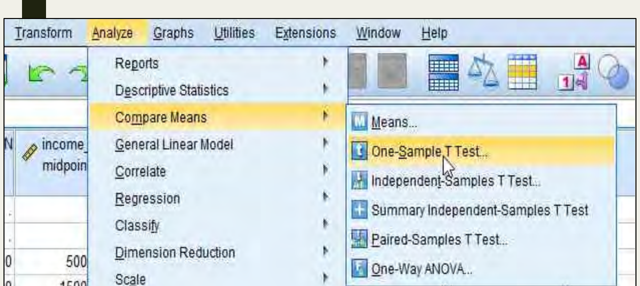
			boat		
			no	yes	Total
consume_stream_anyfreq	Does not consume fish from the stream	Count	112	24	136
		% within boat	46.7%	43.6%	46.1%
	Does consume fish from the stream	Count	128	31	159
		% within boat	53.3%	56.4%	53.9%
Total		Count	240	55	295
		% within boat	100.0%	100.0%	100.0%

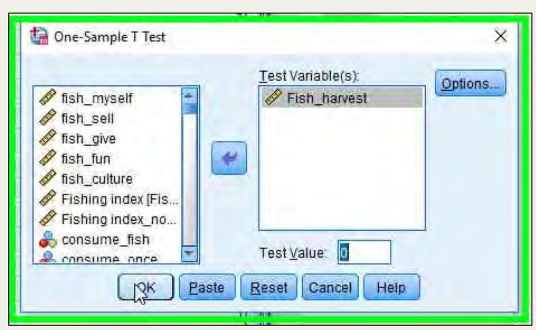
Inferential Statistics

- After doing some exploratory analysis to "get to know" your data better, we can now move on to inferential stats
- With theses 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%

T-Test

One-Sample Statistics

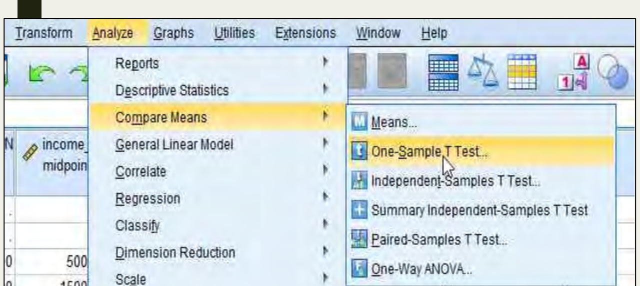
	N	Mean	Std. Deviation	Std. Error Mean
Fish_harvest	303	.51	.501	.029

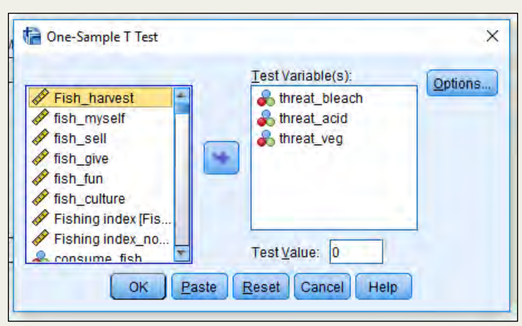
One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Fish_harvest	17.667	302	.000	.508	.45	.56

Generating Confidence Intervals

- 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%*

T-Test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
threat_bleach	305	.05	.210	.012
threat_acid	305	.10	.294	.017
threat_veg	305	.08	.264	.015

One-Sample Test

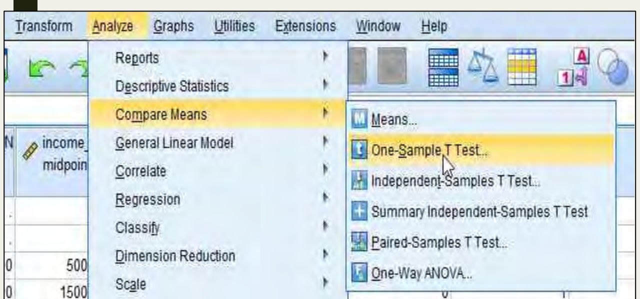
	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
threat_bleach	3.824	304	.000	.046	.02	.07
threat_acid	5.652	304	.000	.095	.06	.13
threat_veg	4.979	304	.000	.075	.05	.11

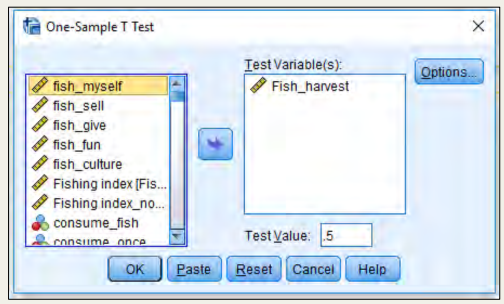
Testing Hypothesis

- 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%





Testing Hypothesis

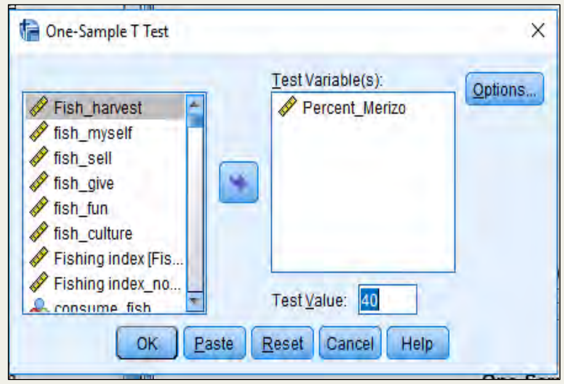
One-Sample Test

	Test Value = .5					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Fish_harvest	.287	302	.774	.008	-.05	.06

- 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%



Testing Hypothesis

One-Sample Test

	Test Value = 40					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Percent_Merizo	-2082.906	282	.000	-39.667	-39.70	-39.63

- 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?*

Practice!

- Confidence interval
  - (.45 , .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%

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
infosource_newspaper	304	.50	.501	.029

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
infosource_newspaper	17.522	303	.000	.503	.45	.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%

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
tenure_15 or more	298	.62	.487	.028

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
tenure_15 or more	21.894	297	.000	.617	.56	.67

Practice!

- 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?*

Practice!

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
age	304	40.38	15.040	.863

One-Sample Test

	Test Value = 35					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
age	6.235	303	.000	5.378	3.68	7.08

- 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

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
flood_impact	243	2.27	3.462	.222

One-Sample Test

	Test Value = 5					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
flood_impact	-12.286	242	.000	-2.728	-3.17	-2.29

- 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

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
familiar_Achang	305	.56	.498	.028

One-Sample Test

	Test Value = .55					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
familiar_Achang	.259	304	.796	.007	-.05	.06

*Save your output as "Manell_Geus_Output_inferential.spv"

- 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

Quiz #6

Day 3: September 14, 2016

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

The totals around the perimeter = "marginals"

	Independent Variable		
Dependent Variable	IV Attribute Lo	IV Attribute Hi	Total
DV attribute Lo	A	B	A + B
DV attribute Hi	C	D	C + D
Total	A + C	B + D	N = A+B+C+D

The totals around the perimeter = "marginals"

The cross-classifications = "cells"

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*

Measures of Association

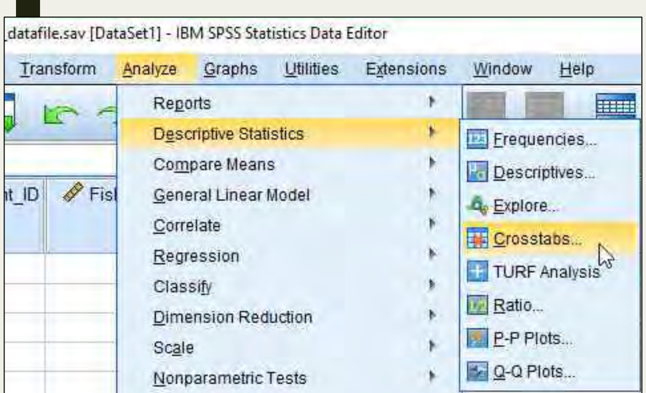
Table 10.2 How to Interpret Measures of Association

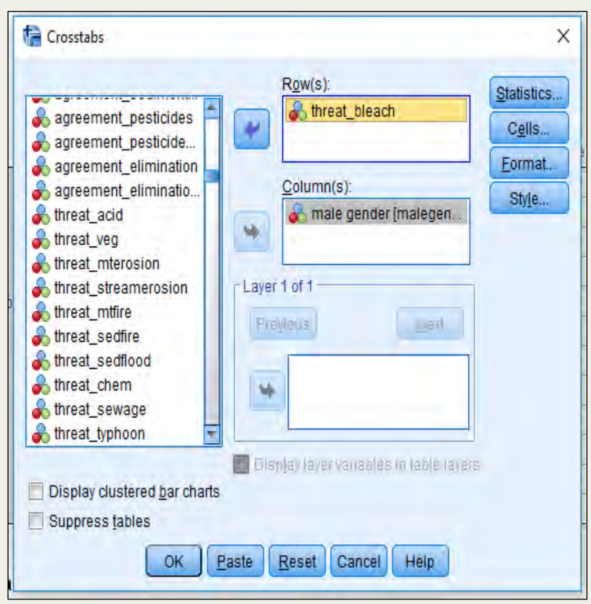
Measure of Association (X)	*Qualitative Interpretation*
$0 \leq X < 0.10$	Very Weak
$0.10 \leq X < 0.20$	Weak
$0.20 \leq X < 0.30$	Moderate
$X \geq 0.30$	Strong

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

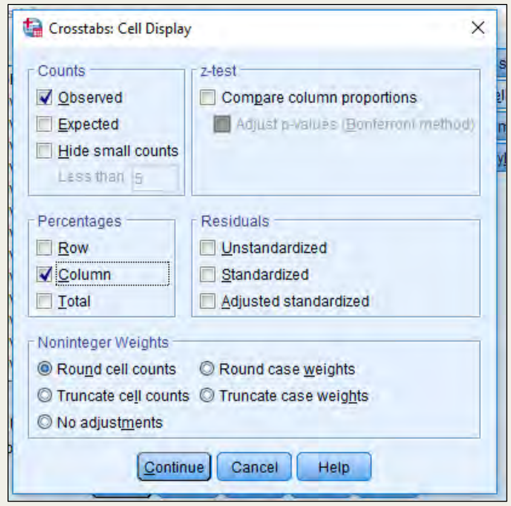
Contingency Tables in SPSS





Contingency Tables in SPSS

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



Contingency Tables in SPSS

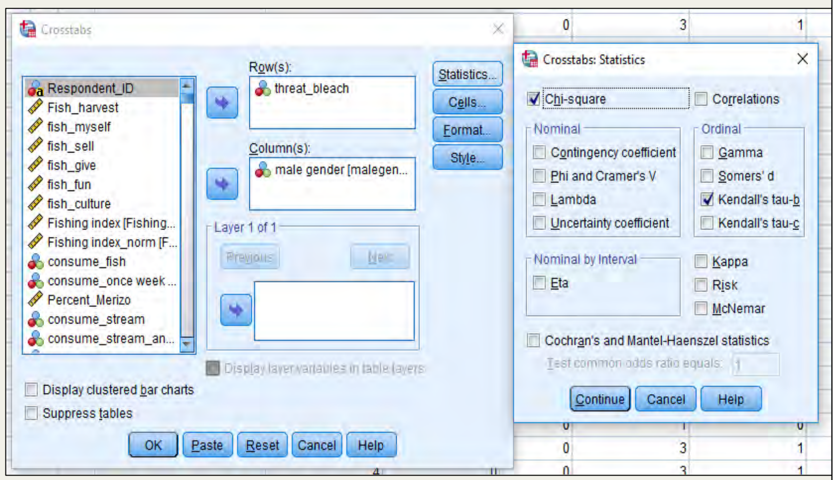
- Approximately 4% of males and 4% of females chose coral bleaching as a top threat to coral reefs
- Gender does not seem to have an effect on choosing whether coral bleaching is a top threat to coral reefs
  - However, we must perform the chi-square test to be certain

threat_bleach * male gender Crosstabulation

			male gender		
			female	male	Total
threat_bleach	respondent did not chose as top 3	Count	159	130	289
		% within male gender	95.8%	95.6%	95.7%
	respondent chose as top 3	Count	7	6	13
		% within male gender	4.2%	4.4%	4.3%
Total		Count	166	136	302
		% within male gender	100.0%	100.0%	100.0%

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*



Chi Square and Measures of Association in SPSS

- 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

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.007[a]	1	.934		
Continuity Correction[b]	.000	1	1.000		
Likelihood Ratio	.007	1	.934		
Fisher's Exact Test				1.000	.576
Linear-by-Linear Association	.007	1	.934		
N of Valid Cases	302				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 5.85.

b. Computed only for a 2x2 table

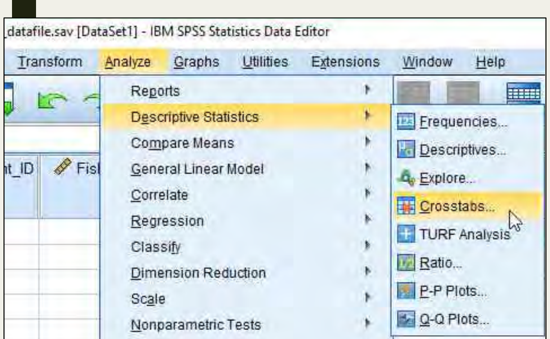
Symmetric Measures

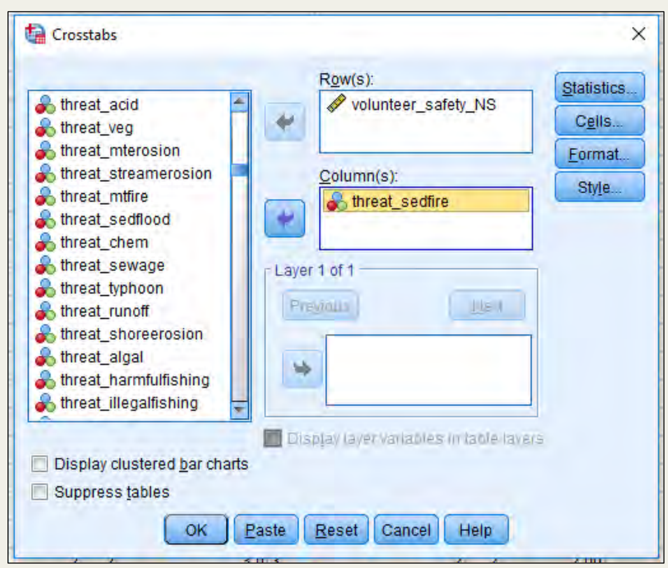
		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Kendall's tau-b	.005	.058	.083	.934
N of Valid Cases		302			

Contingency Tables in SPSS

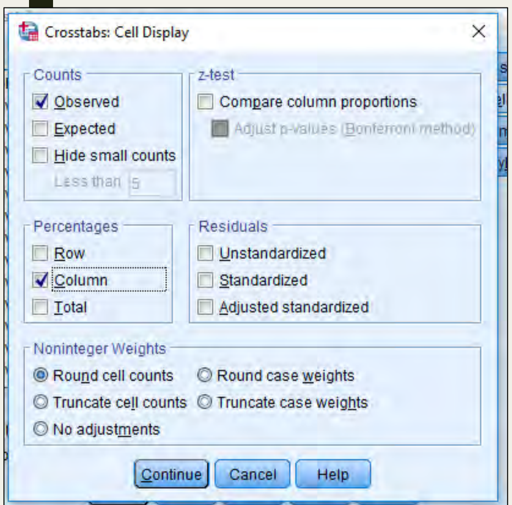
- We want to determine if peoples' perceptions concerning whether sedimentation caused by fire is a top threat to coral reefs has an effect on whether they would participate in volunteering to practice better fire safety to prevent wildfires
- Let's cross-tabulate "threat_sedfire" and "volunteer_safety_NS"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

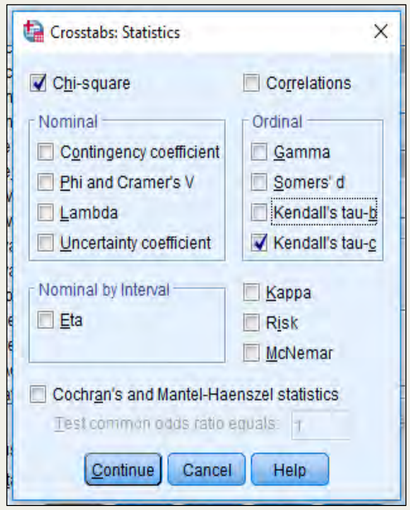
Contingency Tables in SPSS





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

Contingency Tables in SPSS

- 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

volunteer_safety_NS * threat_sedfire Crosstabulation

			threat_sedfire: respondent did not chose as top 3	threat_sedfire: respondent chose as top 3	Total
volunteer_safety_NS	would not do	Count	14	2	16
		% within threat_sedfire	5.5%	4.8%	5.4%
	would consider	Count	102	16	118
		% within threat_sedfire	40.0%	38.1%	39.7%
	would do	Count	139	24	163
		% within threat_sedfire	54.5%	57.1%	54.9%
Total		Count	255	42	297
		% within threat_sedfire	100.0%	100.0%	100.0%

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*

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	.114[a]	2	.945
Likelihood Ratio	.115	2	.944
Linear-by-Linear Association	.114	1	.736
N of Valid Cases	297		

a. 1 cells (16.7%) have expected count less than 5. The minimum expected count is 2.26.

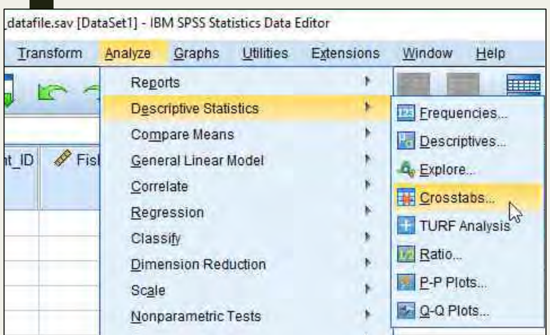
Symmetric Measures

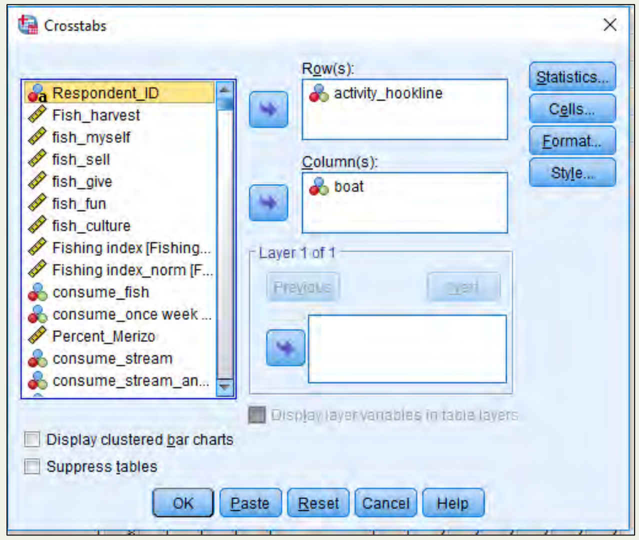
		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Kendall's tau-c	.014	.041	.337	.736
N of Valid Cases		297			

Contingency Tables in SPSS

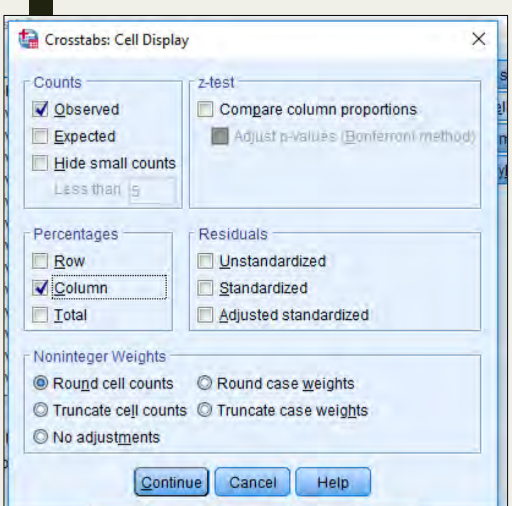
- We want to determine if boat ownership affects where people go in Merizo to do hook-and-line fishing
- Let's cross-tabulate "boat" and "activity_hookline"
- Dependent variable = "activity_hookline" in rows
- Independent variable = "boat" in columns

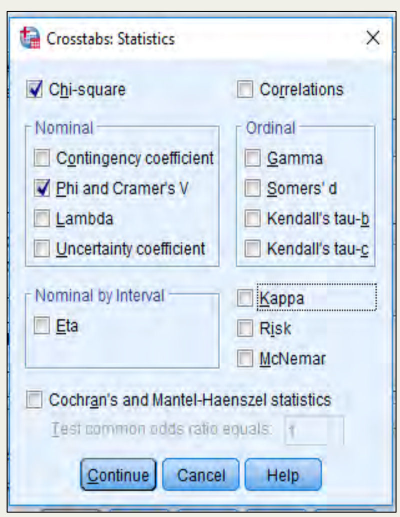
Contingency Tables in SPSS





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

activity_hookline * boat Crosstabulation

			boat		
			no	yes	Total
activity_hookline	no	Count	45	7	52
		% within boat	22.4%	15.2%	21.1%
	yes, in Cocos Lagoon	Count	78	24	102
		% within boat	38.8%	52.2%	41.3%
	yes, in Achang Preserve	Count	31	1	32
		% within boat	15.4%	2.2%	13.0%
	Yes, in both places	Count	47	14	61
		% within boat	23.4%	30.4%	24.7%
Total		Count	201	46	247
		% within boat	100.0%	100.0%	100.0%

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*

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	8.360[a]	3	.039
Likelihood Ratio	10.467	3	.015
Linear-by-Linear Association	.207	1	.649
N of Valid Cases	247		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 5.96.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.184	.039
	Cramer's V	.184	.039
N of Valid Cases		247	

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*

condition_numfish_NS * Fish_harvest Crosstabulation

			Fish_harvest		
			no	yes	Total
condition_numfish_NS	very bad	Count	13	8	21
		% within Fish_harvest	9.6%	5.4%	7.4%
	bad	Count	31	13	44
		% within Fish_harvest	23.0%	8.8%	15.5%
	neither good nor bad	Count	25	42	67
		% within Fish_harvest	18.5%	28.4%	23.7%
	good	Count	51	61	112
		% within Fish_harvest	37.8%	41.2%	39.6%
	very good	Count	15	24	39
		% within Fish_harvest	11.1%	16.2%	13.8%
Total		Count	135	148	283
		% within Fish_harvest	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	15.272[a]	4	.004
Likelihood Ratio	15.540	4	.004
Linear-by-Linear Association	7.323	1	.007
N of Valid Cases	283		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 10.02.

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Kendall's tau-c	.163	.065	2.503	.012
N of Valid Cases		283			

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?*

Practice!

- 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"

enviro_education_NS * resident_responsible Crosstabulation

			resident_responsible no	resident_responsible yes	Total
enviro_education_NS	never	Count	5	32	37
		% within resident_responsible	7.8%	15.5%	13.7%
	once a year	Count	15	30	45
		% within resident_responsible	23.4%	14.5%	16.6%
	a few times a year	Count	16	68	84
		% within resident_responsible	25.0%	32.9%	31.0%
	once a month	Count	25	61	86
		% within resident_responsible	39.1%	29.5%	31.7%
	weekly	Count	3	16	19
		% within resident_responsible	4.7%	7.7%	7.0%
Total		Count	64	207	271
		% within resident_responsible	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	7.484[a]	4	.112
Likelihood Ratio	7.620	4	.107
Linear-by-Linear Association	.361	1	.548
N of Valid Cases	271		

a. 1 cells (10.0%) have expected count less than 5. The minimum expected count is 4.49.

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Kendall's tau-c	-.033	.056	-.594	.552
N of Valid Cases		271			

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

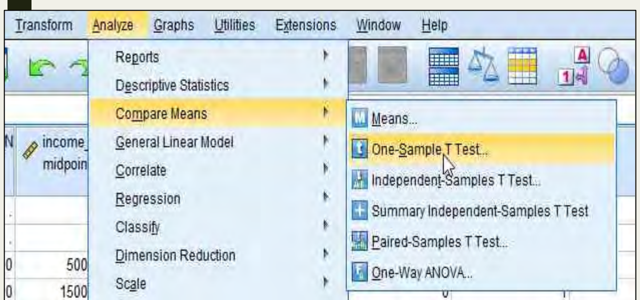
- One sample
  - *Compare a sample mean to a point estimate*
- Two sample - paired
  - *"Before and after"*
- Two sample – independent
  - *Difference of means across groups*
- In two sample t-tests, the difference of the means of the separate groups are calculated and confidence intervals are created around the **difference**
- If the confidence interval **does not contain zero**, then there is a statistically significant difference

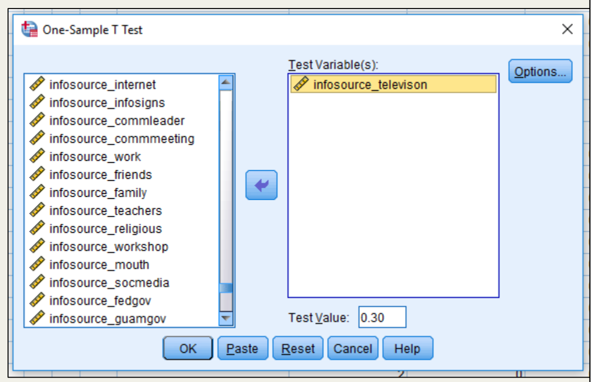
One Sample T-test

- 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%





One Sample T-test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
infosource_televison	304	.38	.485	.028

One-Sample Test

	Test Value = 0.30					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
infosource_televison	2.697	303	.007	.075	.02	.13

- 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

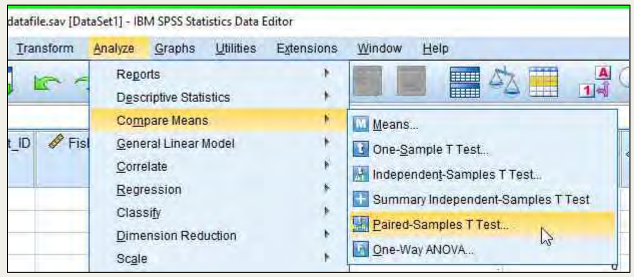
- Paired sample t-tests are used in ‘before-after’ studies, or when the samples are the matched pairs, or when it is a case-control study
- For example, we give weight loss treatment to a group of voluntary participants
  - *The paired t-test can be used to test for a statistically significant reduction 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

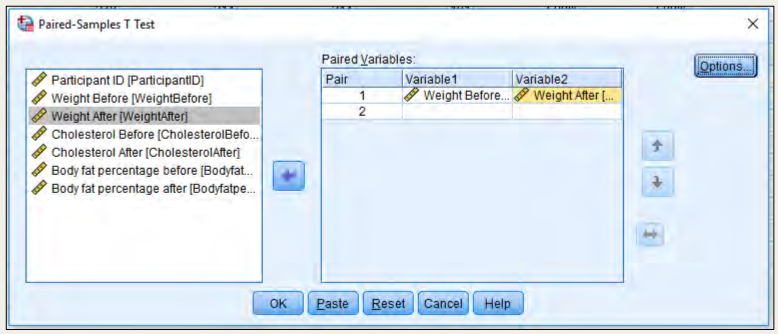
Two Sample T-test: Paired Samples

- 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





Two Sample T-test: Paired Samples

Sample Means

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Weight Before	200.56	50	58.208	8.232
	Weight After	187.40	50	52.594	7.438

Shows how related the 2 variables are

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Weight Before & Weight After	50	.956	.000

Mean DIFFERENCE

P-value

Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	Weight Before - Weight After	13.160	17.286	2.445	8.248	18.072	5.383	49	.000

Two Sample T-test: Paired Samples

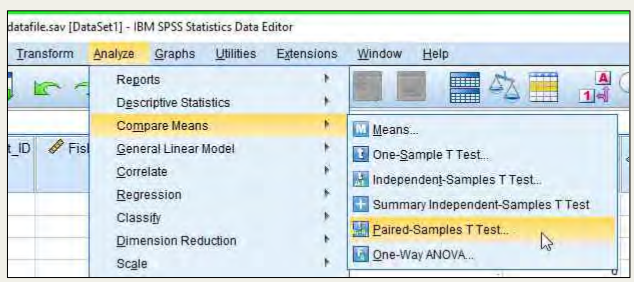
Paired Samples Test

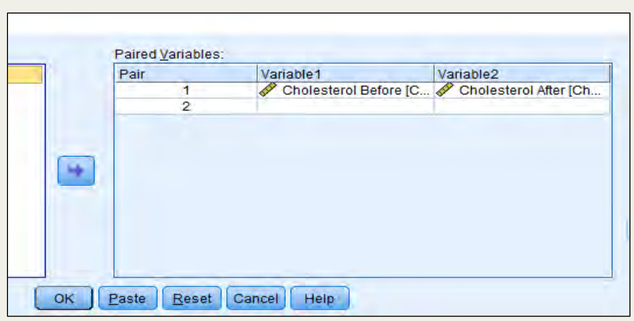
		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Weight Before - Weight After	13.160	17.286	2.445	8.248	18.072	5.383	49	.000

- 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***

Two Sample T-test: Paired Samples

- 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

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Cholesterol Before	183.44	50	44.613	6.309
	Cholesterol After	170.64	50	38.681	5.470

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Cholesterol Before & Cholesterol After	50	.926	.000

Paired Samples Test

		Paired Differences							
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	Cholesterol Before - Cholesterol After	12.800	17.087	2.416	7.944	17.656	5.297	49	.000

- 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**

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

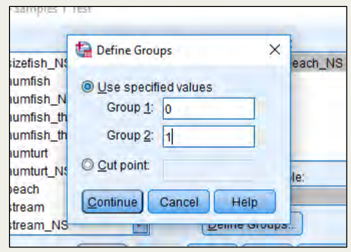
- Likert Data is ordinal, and with ordinal data, you (*usually) cannot calculate means or perform the same statistical tests that you would on continuous data
- *However, a contingent of scholars in the social sciences agree that **Likert Data can be analyzed as continuous data**
  - *i.e. we can calculate 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)

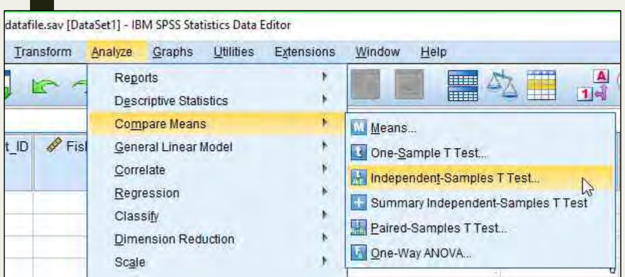
Two Sample t-test: Independent Samples

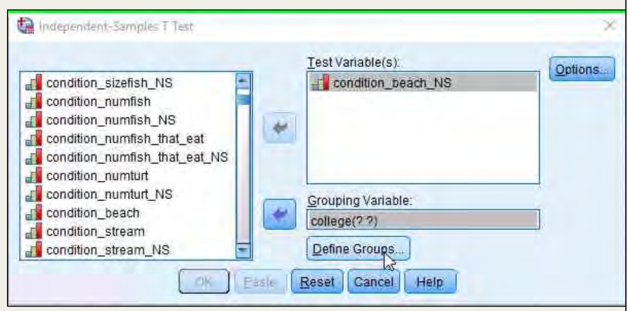
- Helps you compare whether two groups have statistically significant different mean values
  - *For example, whether men and women have different mean heights*
- Null hypothesis: there is no significant difference between the groups
- Alternative hypothesis: There is a significant difference between 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”







Two Sample t-test: Independent Samples

Group Statistics

	college	N	Mean	Std. Deviation	Std. Error Mean
condition_beach_NS	did not complete college	227	2.89	1.073	.071
	completed college	64	3.00	1.234	.154

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means					95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
condition_beach_NS	Equal variances assumed	1.560	.213	-.701	289	.484	-.110	.157	-.419	.199
	Equal variances not assumed			-.648	91.547	.519	-.110	.170	-.448	.227

- 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:

Group Statistics

	college	N	Mean	Std. Deviation	Std. Error Mean
condition_beach_NS	did not complete college	227	2.89	1.073	.071
	completed college	64	3.00	1.234	.154

- 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

Two Sample t-test: Independent Samples

Independent Samples Test

		Levene's Test for Equality of Variances		t-		
		F	Sig.	t	df	Sig. (2-tailed)
condition_beach_NS	Equal variances assumed	1.560	.213	-.701	289	.484
	Equal variances not assumed			-.648	91.547	.519

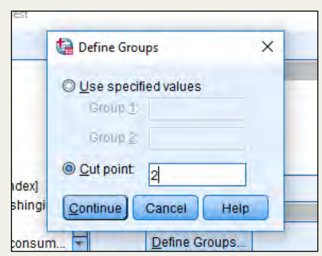
- Significance of F-test tells us if we "assume equal variances" or not
- Since 0.213 > 0.05, we "fail to reject the null hypothesis of equal variances"
  - Therefore we assume equal variances and continue
  - The T statistics and associated p-values will be different depending on how you interpret your 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

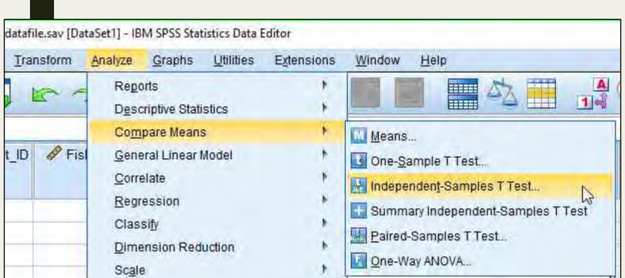
Two Sample t-test: Independent Samples

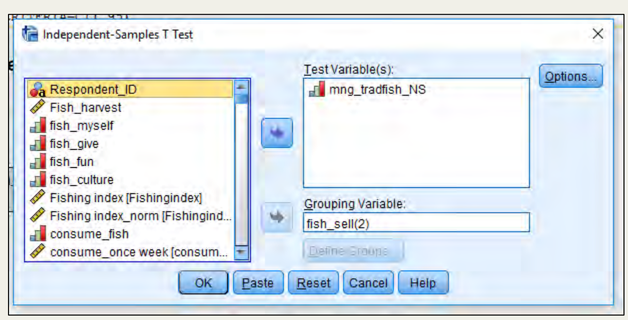
- Now, let’s examine if fishing “to sell” has any effect on peoples’ level of support for “creating areas for only traditional fishing”
- We need “fish_sell” and “mng_tradfish_NS”
  - *DV = mng_tradfish_NS   IV = “fish_sell” (grouping variable)*
- Null hypothesis: If someone fishes to sell, this has 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

Group Statistics

	fish_sell	N	Mean	Std. Deviation	Std. Error Mean
mng_tradfish_NS	>= 2	106	3.57	1.235	.120
	< 2	28	3.32	1.249	.236

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality o			
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference
mng_tradfish_NS	Equal variances assumed	.515	.474	.930	132	.354	.245
	Equal variances not assumed			.924	42.020	.361	.245

- 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

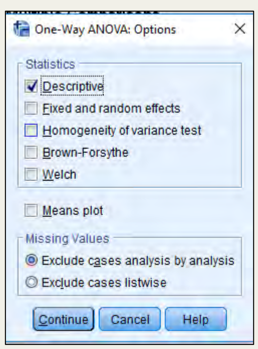
- Essentially an Independent t-test with >2 groups
- Tests for statistically significant differences in the means of 2 or more groups
- One-way ANOVA is an omnibus F-test statistic and cannot tell you which specific groups were significantly different from each other (i.e. p-value<0.05), only that at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

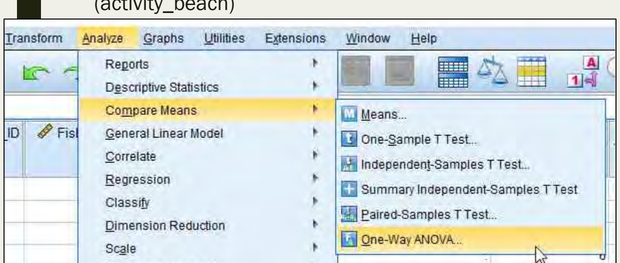
One way ANOVA Post-Hoc Test

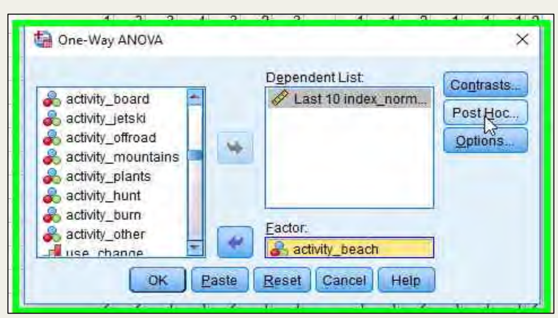
- The most common post-hoc test is the Tukey's HSD test
- The post-hoc test will tell you *which groups' means* are significantly different
- The difference of the means of the separate groups are calculated and confidence intervals are created around the **difference**
  - *If the confidence interval 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

Last 10 index_norm

	N	Mean
no	50	56.44
yes, in Cocos Lagoon	88	46.09
yes, in Achang Preserve	16	56.51
Yes, in both places	55	57.65
Total	209	52.41

Multiple Comparisons

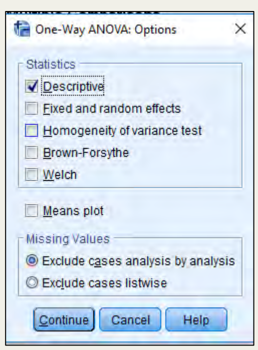
Dependent Variable: Last 10 index_norm
Tukey HSD

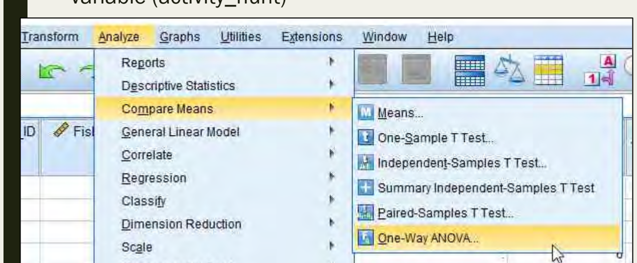
(I) activity_beach	(J) activity_beach	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
no	yes, in Cocos Lagoon	10.359	4.157	.064	-.41	21.13
	yes, in Achang Preserve	-.066	6.742	1.000	-17.53	17.40
	Yes, in both places	-1.207	4.586	.994	-13.09	10.67
yes, in Cocos Lagoon	no	-10.359	4.157	.064	-21.13	.41
	yes, in Achang Preserve	-10.425	6.379	.362	-26.95	6.10
	Yes, in both places	-11.566*	4.034	.024	-22.02	-1.12
yes, in Achang Preserve	no	.066	6.742	1.000	-17.40	17.53
	yes, in Cocos Lagoon	10.425	6.379	.362	-6.10	26.95
	Yes, in both places	-1.141	6.667	.998	-18.41	16.13
Yes, in both places	no	1.207	4.586	.994	-10.67	13.09
	yes, in Cocos Lagoon	11.566*	4.034	.024	1.12	22.02
	yes, in Achang Preserve	1.141	6.667	.998	-16.13	18.41

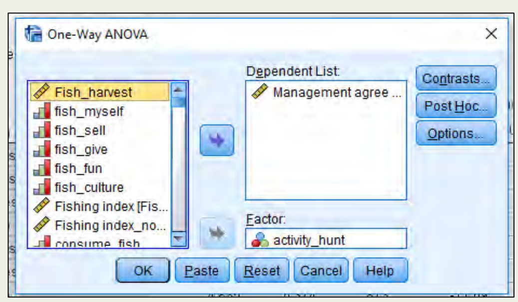
*. The mean difference is significant at the 0.05 level.

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

Management agree index_norm

	N	Mean
no	99	48.35
yes, in Cocos Lagoon	37	52.75
yes, in Achang Preserve	12	53.18
Yes, in both places	48	49.85
Total	196	49.84

Multiple Comparisons

Dependent Variable: Management agree index_norm

Tukey HSD

(I) activity_hunt	(J) activity_hunt	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
no	yes, in Cocos Lagoon	-4.405	4.018	.692	-14.82	6.01
	yes, in Achang Preserve	-4.835	6.374	.873	-21.35	11.69
	Yes, in both places	-1.501	3.668	.977	-11.01	8.00
yes, in Cocos Lagoon	no	4.405	4.018	.692	-6.01	14.82
	yes, in Achang Preserve	-.430	6.928	1.000	-18.38	17.52
	Yes, in both places	2.903	4.562	.920	-8.92	14.73
yes, in Achang Preserve	no	4.835	6.374	.873	-11.69	21.35
	yes, in Cocos Lagoon	.430	6.928	1.000	-17.52	18.38
	Yes, in both places	3.333	6.730	.960	-14.11	20.78
Yes, in both places	no	1.501	3.668	.977	-8.00	11.01
	yes, in Cocos Lagoon	-2.903	4.562	.920	-14.73	8.92
	yes, in Achang Preserve	-3.333	6.730	.960	-20.78	14.11

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*

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Body fat percentage before - Body fat percentage after	2.28000%	2.68814%	0.38016%	1.51604%	3.04396%	5.997	49	.000

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

Group Statistics

	benefit	N	Mean	Std. Deviation	Std. Error Mean
Management Success Index_norm	no	107	52.84	17.556	1.697
	yes	22	55.83	21.472	4.578

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality (			
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference
Management Success Index_norm	Equal variances assumed	.140	.709	-.698	127	.486	-2.986
	Equal variances not assumed			-.612	27.068	.546	-2.986

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"

Condition index_norm

	N	Mean
would not do	4	35.42
would consider	82	47.59
would do	152	58.11
Total	238	54.11

Multiple Comparisons

Dependent Variable: Condition index_norm
Tukey HSD

(I) volunteer_protect_NS	(J) volunteer_protect_NS	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
would not do	would consider	-12.178	11.063	.515	-38.27	13.92
	would do	-22.697	10.944	.097	-48.51	3.12
would consider	would not do	12.178	11.063	.515	-13.92	38.27
	would do	-10.519*	2.960	.001	-17.50	-3.54
would do	would not do	22.697	10.944	.097	-3.12	48.51
	would consider	10.519*	2.960	.001	3.54	17.50

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

Quiz #7

Day 4: September 15, 2016

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

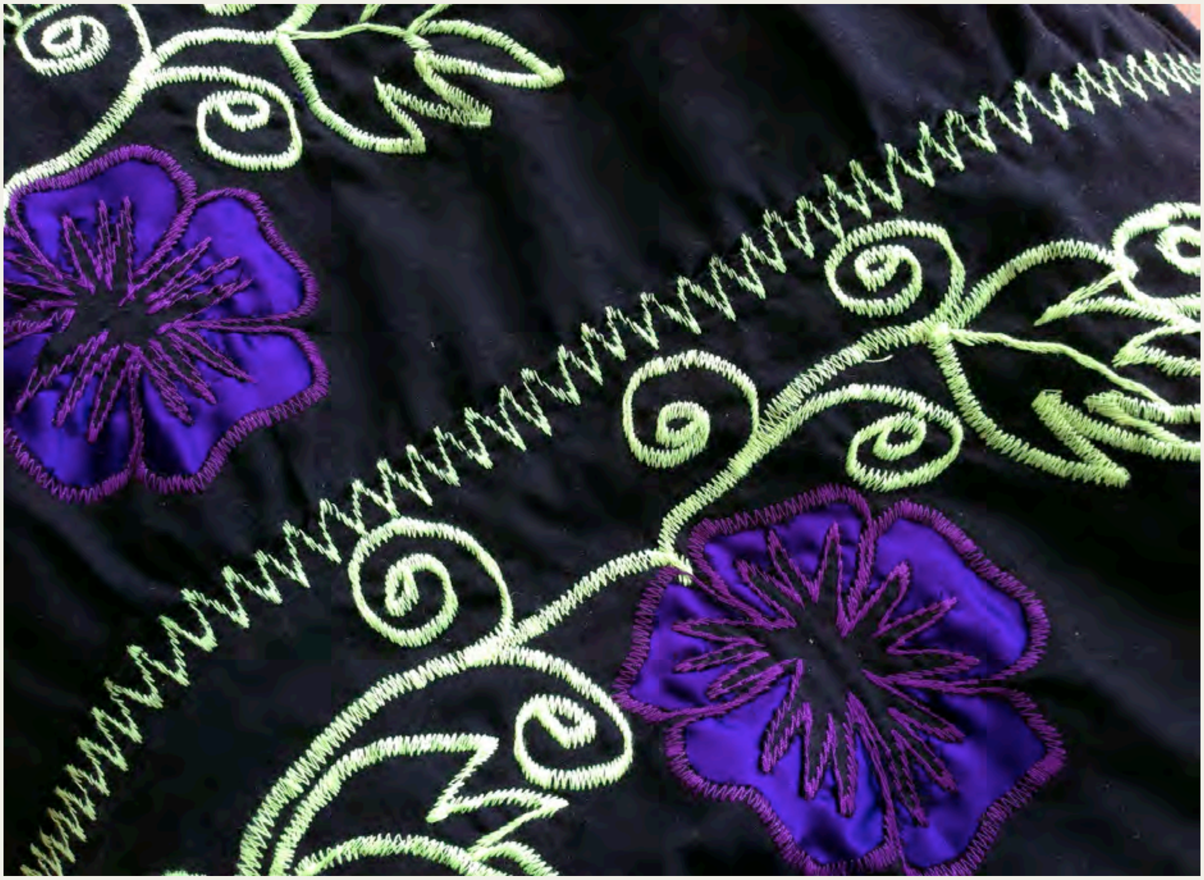
7.5 What happens when a p-value is less than 0.05?

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**

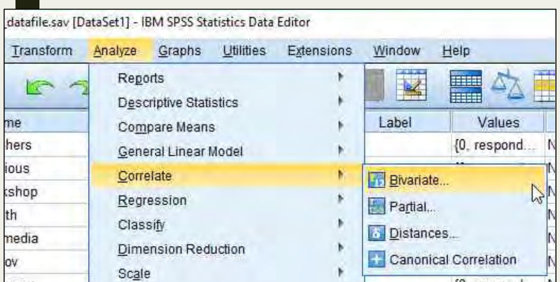
Correlation Analysis

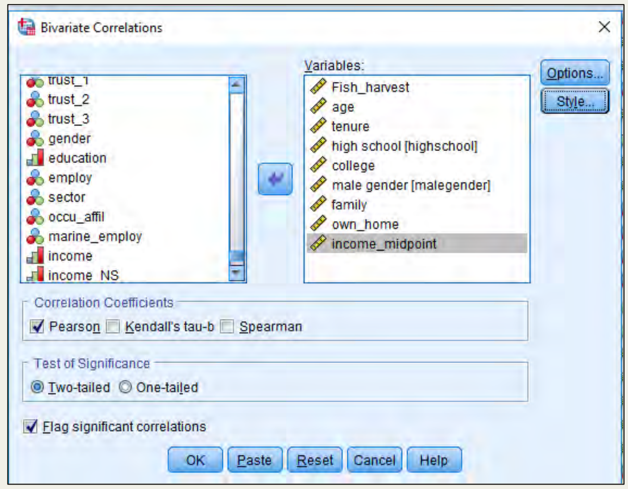
- Correlation Analysis is good for creating "demographic profiles" of certain respondent types
  - *Demographics – population characteristics such as gender, age, income, education, etc.*
  - *For example, if we want to answer what type of person fishes or harvests for marine resource?*
    - We can 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





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

Correlations

		Fish_harvest	age	tenure	high school	college	male gender	family	own_home	income_midpoint
Fish_harvest	Pearson Correlation	1	-.003	-.019	.062	-.060	.096	.060	-.056	-.128
	Sig. (2-tailed)		.961	.741	.290	.304	.096	.301	.333	.273
	N	303	301	295	296	296	300	300	297	75
age	Pearson Correlation	-.003	1	.532**	-.040	.154**	.036	-.151**	.324**	.337**
	Sig. (2-tailed)	.961		.000	.493	.008	.537	.008	.000	.003
	N	301	304	297	297	297	302	303	299	74
tenure	Pearson Correlation	-.019	.532**	1	.029	-.038	.072	-.015	.397**	-.122
	Sig. (2-tailed)	.741	.000		.620	.523	.217	.801	.000	.299
	N	295	297	298	291	291	296	296	295	74
high school	Pearson Correlation	.062	-.040	.029	1	.135*	.065	-.043	.090	.200
	Sig. (2-tailed)	.290	.493	.620		.020	.265	.458	.124	.088
	N	296	297	291	299	299	297	296	293	74
college	Pearson Correlation	-.060	.154**	-.038	.135*	1	.056	-.030	.064	.698**
	Sig. (2-tailed)	.304	.008	.523	.020		.335	.609	.273	.000
	N	296	297	291	299	299	297	296	293	74
male gender	Pearson Correlation	.096	.036	.072	.065	.056	1	-.059	.027	-.027
	Sig. (2-tailed)	.096	.537	.217	.265	.335		.308	.643	.819
	N	300	302	296	297	297	303	301	298	74
family	Pearson Correlation	.060	-.151**	-.015	-.043	-.030	-.059	1	-.117*	-.316**
	Sig. (2-tailed)	.301	.008	.801	.458	.609	.308		.044	.006
	N	300	303	296	296	296	301	303	298	74
own_home	Pearson Correlation	-.056	.324**	.397**	.090	.064	.027	-.117*	1	.060
	Sig. (2-tailed)	.333	.000	.000	.124	.273	.643	.044		.613
	N	297	299	295	293	293	298	298	300	74
income_midpoint	Pearson Correlation	-.128	.337**	-.122	.200	.698**	-.027	-.316**	.060	1
	Sig. (2-tailed)	.273	.003	.299	.088	.000	.819	.006	.613	
	N	75	74	74	74	74	74	74	74	75

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

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

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**

		Fish_harvest
Fish_harvest	Pearson Correlation	1
	Sig. (2-tailed)	
	N	303
age	Pearson Correlation	-.003
	Sig. (2-tailed)	.961
	N	301
tenure	Pearson Correlation	-.019
	Sig. (2-tailed)	.741
	N	295
high school	Pearson Correlation	.062
	Sig. (2-tailed)	.290
	N	296
college	Pearson Correlation	-.060
	Sig. (2-tailed)	.304
	N	296
male gender	Pearson Correlation	.096
	Sig. (2-tailed)	.096
	N	300
family	Pearson Correlation	.060
	Sig. (2-tailed)	.301
	N	300
own_home	Pearson Correlation	-.056
	Sig. (2-tailed)	.333
	N	297
income_midpoint	Pearson Correlation	-.128
	Sig. (2-tailed)	.273
	N	75

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

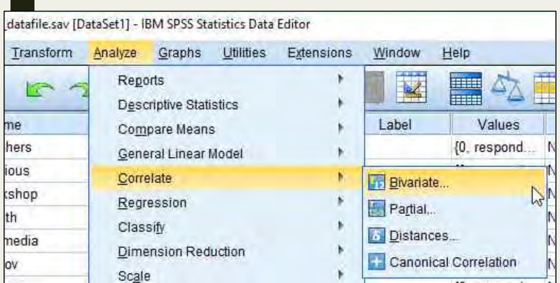
		Fish_harvest	age	tenure	high school	college	male gender	family	own_home	income_midpoint
Fish_harvest	Pearson Correlation	1	-.003	-.019	.062	-.060	.096	.060	-.056	-.128
	Sig. (2-tailed)		.961	.741	.290	.304	.096	.301	.333	.273
	N	303	301	295	296	296	300	300	297	75
age	Pearson Correlation	-.003	1	.532**	-.040	.154**	.036	-.151**	.324**	.337**
	Sig. (2-tailed)	.961		.000	.493	.008	.537	.008	.000	.003
	N	301	304	297	297	297	302	303	299	74
tenure	Pearson Correlation	-.019	.532**	1	.029	-.038	.072	-.015	.397**	-.122
	Sig. (2-tailed)	.741	.000		.620	.523	.217	.801	.000	.299
	N	295	297	298	291	291	296	296	295	74
high school	Pearson Correlation	.062	-.040	.029	1	.135*	.065	-.043	.090	.200
	Sig. (2-tailed)	.290	.493	.620		.020	.265	.458	.124	.088
	N	296	297	291	299	299	297	296	293	74
college	Pearson Correlation	-.060	.154**	-.038	.135*	1	.056	-.030	.064	.698**
	Sig. (2-tailed)	.304	.008	.523	.020		.335	.609	.273	.000
	N	296	297	291	299	299	297	296	293	74
male gender	Pearson Correlation	.096	.036	.072	.065	.056	1	-.059	.027	-.027
	Sig. (2-tailed)	.096	.537	.217	.265	.335		.308	.643	.819
	N	300	302	296	297	297	303	301	298	74
family	Pearson Correlation	.060	-.151**	-.015	-.043	-.030	-.059	1	-.117*	-.316**
	Sig. (2-tailed)	.301	.008	.801	.458	.609	.308		.044	.006
	N	300	303	296	296	296	301	303	298	74
own_home	Pearson Correlation	-.056	.324**	.397**	.090	.064	.027	-.117*	1	.060
	Sig. (2-tailed)	.333	.000	.000	.124	.273	.643	.044		.613
	N	297	299	295	293	293	298	298	300	74
income_midpoint	Pearson Correlation	-.128	.337**	-.122	.200	.698**	-.027	-.316**	.060	1
	Sig. (2-tailed)	.273	.003	.299	.088	.000	.819	.006	.613	
	N	75	74	74	74	74	74	74	74	75

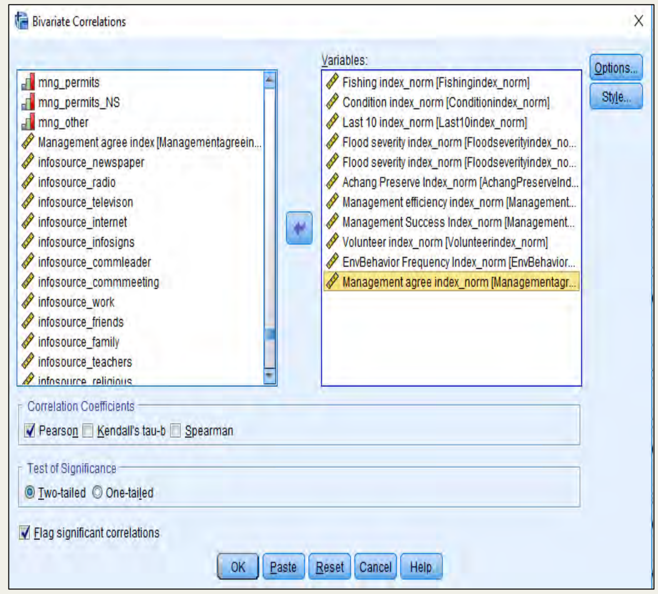
**. Correlation is significant at the 0.01 level (2-tailed).

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

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

		Fishing index_norm	Condition index_norm	Last 10 index_norm	Flood severity index_norm	Flood severity index_norm	Achang Preserve index_norm	Management efficiency index_norm	Management Success index_norm	Volunteer index_norm	EnvBehavior Frequency index_norm	Management agree index_norm
Fishing index_norm	Pearson Correlation	1	.078	.026	.401**	.257**	.400**	.254*	.208*	.186	.277**	.191*
	Sig. (2-tailed)		.400	.775	.000	.003	.000	.014	.027	.055	.003	.043
	N	133	120	121	128	129	78	93	113	107	113	113
Condition index_norm	Pearson Correlation	.078	1	.828**	.180**	.152*	.294**	.399**	.161*	.134	.272**	.382**
	Sig. (2-tailed)	.400		.000	.007	.021	.003	.000	.019	.060	.000	.000
	N	120	246	230	226	230	99	124	213	199	209	204
Last 10 index_norm	Pearson Correlation	.026	.828**	1	.087	.022	.216*	.391**	.162*	.202**	.368**	.311**
	Sig. (2-tailed)	.775	.000		.192	.745	.028	.000	.019	.005	.000	.000
	N	121	230	241	224	226	104	128	210	195	205	200
Flood severity index_norm	Pearson Correlation	.401**	.180**	.087	1	.713**	.426**	.204*	.552**	.004	.139*	.257**
	Sig. (2-tailed)	.000	.007	.192		.000	.000	.016	.000	.959	.036	.000
	N	128	226	224	272	257	112	138	228	217	226	225
Flood severity index_norm	Pearson Correlation	.257**	.152*	.022	.713**	1	.282**	.122	.498**	.076	.130	.209**
	Sig. (2-tailed)	.003	.021	.745	.000		.002	.147	.000	.260	.051	.001
	N	129	230	226	257	281	114	143	232	222	225	229
Achang Preserve index_norm	Pearson Correlation	.400**	.294**	.216*	.426**	.282**	1	.757**	.412**	.432**	.234*	.452**
	Sig. (2-tailed)	.000	.003	.028	.000	.002		.000	.000	.000	.022	.000
	N	78	99	104	112	114	120	118	106	96	95	101
Management efficiency index_norm	Pearson Correlation	.254*	.399**	.391**	.204*	.122	.757**	1	.258**	.444**	.256**	.381**
	Sig. (2-tailed)	.014	.000	.000	.016	.147	.000		.004	.000	.005	.000
	N	93	124	128	138	143	118	153	125	118	120	122
Management Success index_norm	Pearson Correlation	.208*	.161*	.162*	.552**	.498**	.412**	.258**	1	-.059	.079	.166*
	Sig. (2-tailed)	.027	.019	.019	.000	.000	.000	.004		.405	.252	.018
	N	113	213	210	228	232	106	125	244	199	211	205
Volunteer index_norm	Pearson Correlation	.186	.134	.202**	.004	.076	.432**	.444**	-.059	1	.384**	.319**
	Sig. (2-tailed)	.055	.060	.005	.959	.260	.000	.000	.405		.000	.000
	N	107	199	195	217	222	96	118	199	238	197	192
EnvBehavior Frequency index_norm	Pearson Correlation	.277**	.272**	.368**	.139*	.130	.234*	.256**	.079	.384**	1	.371**
	Sig. (2-tailed)	.003	.000	.000	.036	.051	.022	.005	.252	.000		.000
	N	113	209	205	226	225	95	120	211	197	241	198
Management agree index_norm	Pearson Correlation	.191*	.382**	.311**	.257**	.209**	.452**	.381**	.166*	.319**	.371**	1
	Sig. (2-tailed)	.043	.000	.000	.000	.001	.000	.000	.018	.000	.000	
	N	113	204	200	225	229	101	122	205	192	198	246

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!

Correlations		flood_impact	Management agree index_norm
flood_impact	Pearson Correlation	1	.010
	Sig. (2-tailed)		.890
	N	243	192
Management agree index_norm	Pearson Correlation	.010	1
	Sig. (2-tailed)	.890	
	N	192	246

- 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!

Correlations

		Percent_Merizo	Condition index_norm
Percent_Merizo	Pearson Correlation	1	.307**
	Sig. (2-tailed)		.000
	N	283	230
Condition index_norm	Pearson Correlation	.307**	1
	Sig. (2-tailed)	.000	
	N	230	246

- 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 form 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!

Correlations

		threat_sedfire	Fire severity index_norm
threat_sedfire	Pearson Correlation	1	-.029
	Sig. (2-tailed)		.632
	N	305	280
Fire severity index_norm	Pearson Correlation	-.029	1
	Sig. (2-tailed)	.632	
	N	280	281

- 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*

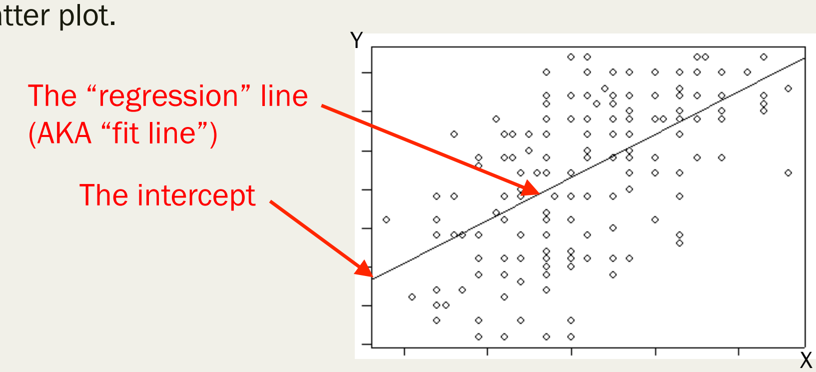
Save you output as "Manell_Geus_Output_correlations.spv"

Simple Linear Regression

Day 5: September 16, 2016

Recap – Simple Linear Regression

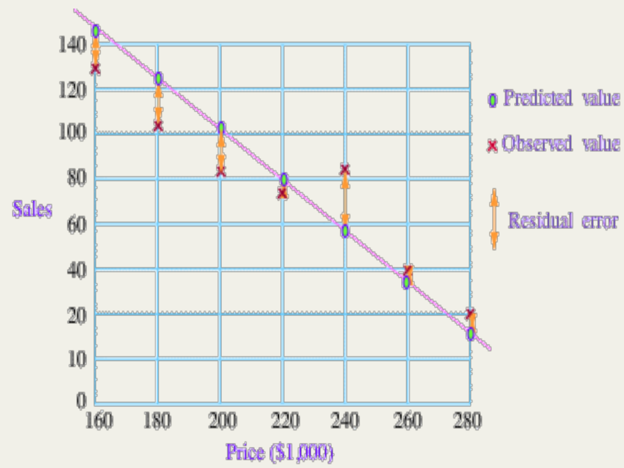
- One dependent variable (Y) and one independent variable (X)
  - *A regression model is a model that describes how a variable X influences the value of another variable Y*
- At the center of the regression analysis is the task of fitting a single line through a scatter plot.



The relationship between X and Y is positive

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

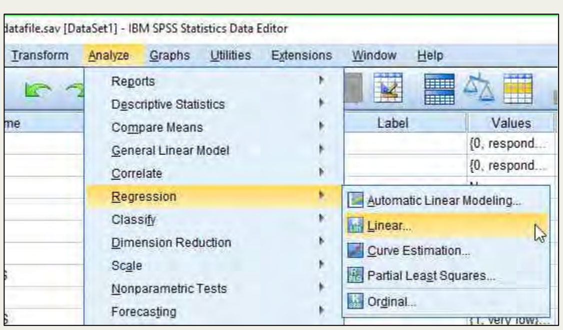
- A technique used to determine the linear relationship between two variables, and in turn, attempt to predict changes in Y due to changes in X
  - *Defined by the formula Y = c + b*X*
    where Y = estimated dependent variable
    c = constant (intercept)
    b = regression coefficient (slope)
    X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

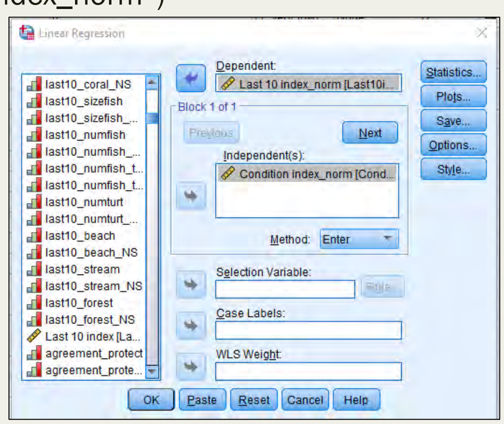
Assumptions of Linear Regression

- 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

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.828[a]	.686	.685	13.358

a. Predictors: (Constant), Condition index_norm

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	88885.279	1	88885.279	498.162	.000[b]
	Residual	40681.247	228	178.427		
	Total	129566.526	229			

a. Dependent Variable: Last 10 index_norm

b. Predictors: (Constant), Condition index_norm

Significance of X variable's predictive power of Y

Intercept

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	4.687	2.301		2.037	.043
	Condition index_norm	.881	.039	.828	22.320	.000

a. Dependent Variable: Last 10 index_norm

Slope

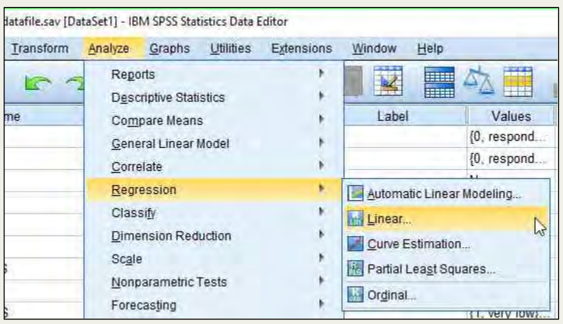
Linear Regression in SPSS

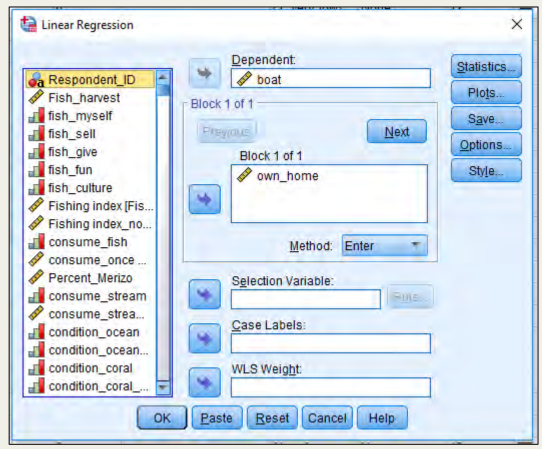
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	4.687	2.301		2.037	.043
	Condition index_norm	.881	.039	.828	22.320	.000

- 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")





Linear Regression in SPSS

- Conclusions?
  - Adjusted R square
    - 5.4% of the variation in boat ownership is explained by home ownership
  - Regression line
    - boat = 0.074 + (0.19)*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

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.240[a]	.058	.054	.379

a. Predictors: (Constant), own_home

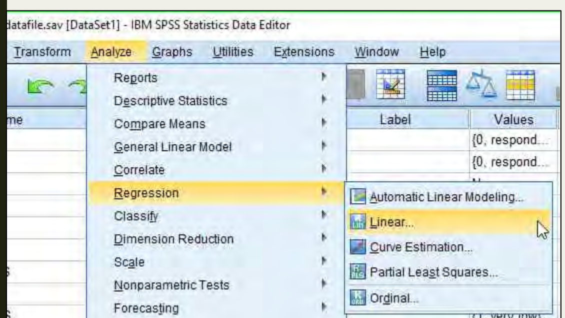
Coefficients[a]

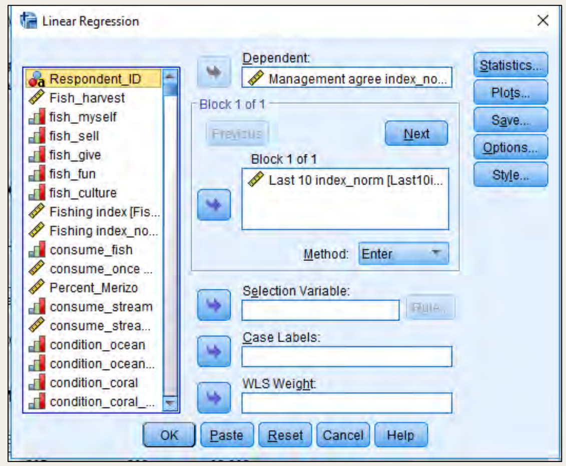
Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	.074	.034		2.157	.032
	own_home	.190	.045	.240	4.231	.000

a. Dependent Variable: boat

Linear Regression in SPSS

- 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")





Linear Regression in SPSS

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.311[a]	.097	.092	20.880

a. Predictors: (Constant), Last 10 index_norm

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	36.401	3.427		10.623	.000
	Last 10 index_norm	.279	.060	.311	4.612	.000

a. Dependent Variable: Management agree index_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!

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.128[a]	.016	.012	19.966

a. Predictors: (Constant), tenure

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	56.353	2.175		25.910	.000
	tenure	-.167	.084	-.128	-1.981	.049

a. Dependent Variable: EnvBehavior Frequency Index_norm

- 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!

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.043[a]	.002	-.002	19.222

a. Predictors: (Constant), resident_responsible

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	73.039	2.692		27.136	.000
	resident_responsible	-2.001	3.044	-.043	-.657	.512

a. Dependent Variable: Volunteer index_norm

- 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"*

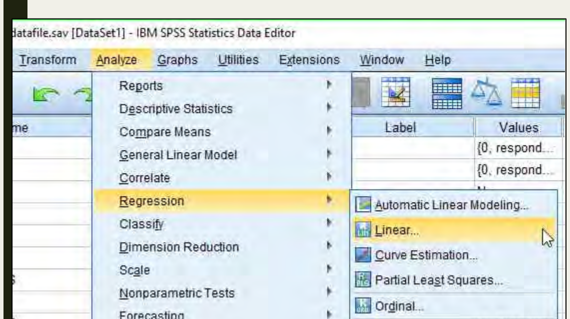
Recap

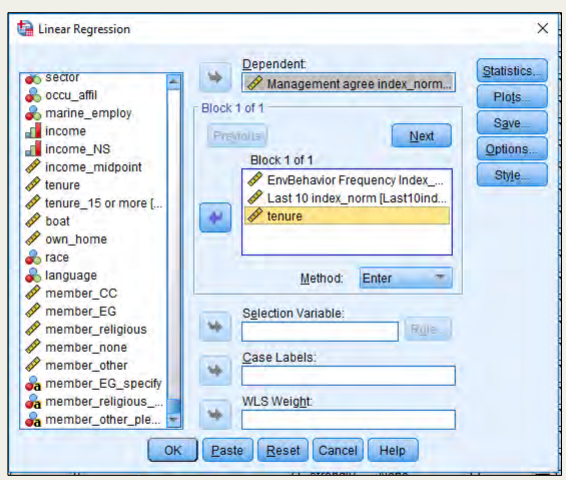
- Model given by:
  - *Y = c + + + + … +*
- Each coefficient (b) is tested for statistical significance using p-values
- Interpretation:
  - *Each predictor (X) variable is interpreted on its own (i.e. “all else held equal”)*
  - *Example: if is significant, that “all else held equal, significantly effects Y”*
  - *If changes by 1, we expect Y to change by*
  - *If changes by 1, we expect Y to change by*
  - *If = = = = 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 “Manell_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

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.476[a]	.227	.213	19.344

a. Predictors: (Constant), tenure, Last 10 index_norm, EnvBehavior Frequency Index_norm

Coefficients[a]

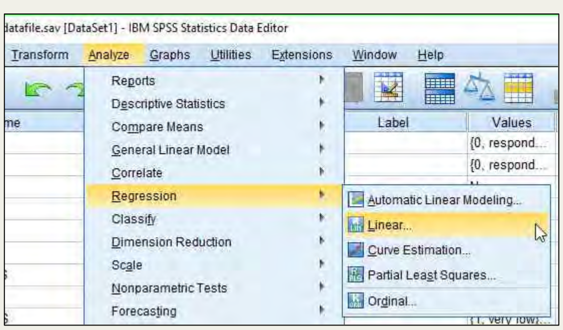
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	17.658	5.841		3.023	.003
	EnvBehavior Frequency Index_norm	.261	.088	.228	2.960	.004
	Last 10 index_norm	.317	.069	.342	4.570	.000
	tenure	.040	.106	.027	.379	.705

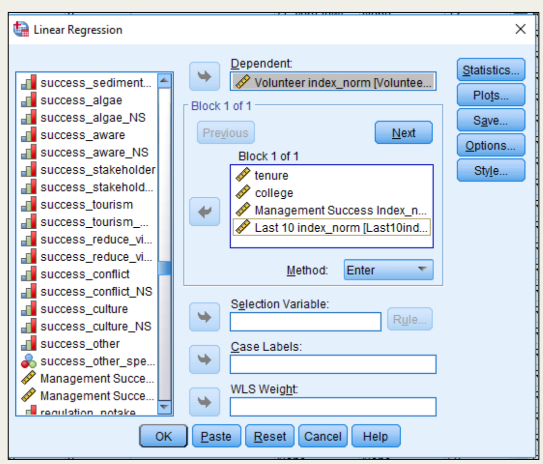
a. Dependent Variable: Management agree index_norm

- 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

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.233[a]	.054	.032	17.742

a. Predictors: (Constant), Last 10 index_norm, college, tenure, Management Success Index_norm

Coefficients[a]

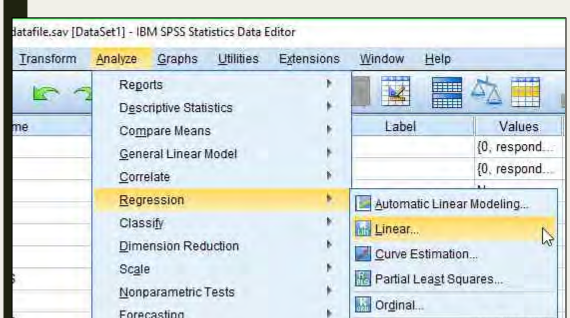
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	63.219	5.379		11.753	.000
	tenure	.139	.090	.119	1.552	.123
	college	2.646	3.297	.063	.803	.423
	Management Success Index_norm	-.044	.077	-.046	-.573	.567
	Last 10 index_norm	.153	.058	.207	2.657	.009

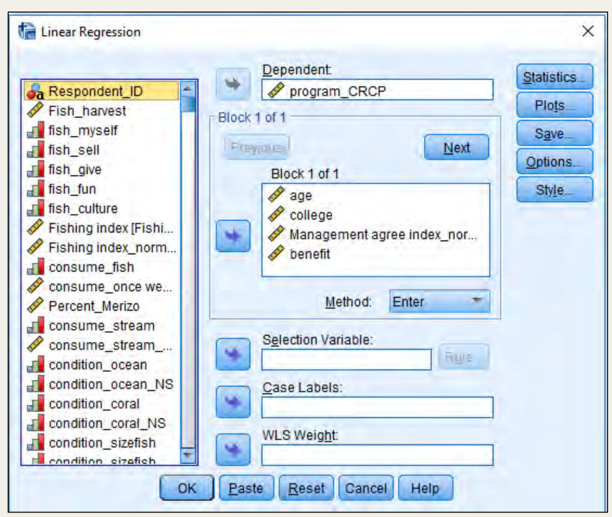
a. Dependent Variable: Volunteer index_norm

- 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

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.290[a]	.084	.055	.485

a. Predictors: (Constant), benefit, Management agree index_norm, college, age

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	.465	.152		3.062	.003
	age	-.002	.003	-.050	-.564	.574
	college	-.245	.111	-.195	-2.212	.029
	Management agree index_norm	.005	.002	.196	2.239	.027
	benefit	-.139	.107	-.114	-1.296	.197

a. Dependent Variable: program_CRCP

- 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 peoples 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*

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.451[a]	.204	.168	20.439

a. Predictors: (Constant), Condition index_norm, college, tenure, program_MC

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	38.132	7.664		4.975	.000
	college	-12.646	5.559	-.222	-2.275	.025
	tenure	.186	.136	.143	1.367	.175
	program_MC	11.548	4.780	.255	2.416	.018
	Condition index_norm	.249	.100	.251	2.481	.015

a. Dependent Variable: Achang Preserve Index_norm

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)*

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.451[a]	.204	.168	20.439

a. Predictors: (Constant), Condition index_norm, college, tenure, program_MC

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	38.132	7.664		4.975	.000
	college	-12.646	5.559	-.222	-2.275	.025
	tenure	.186	.136	.143	1.367	.175
	program_MC	11.548	4.780	.255	2.416	.018
	Condition index_norm	.249	.100	.251	2.481	.015

a. Dependent Variable: Achang Preserve Index_norm

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*

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.470[a]	.221	.210	.441

a. Predictors: (Constant), Percent_Merizo, tenure, Management agree index_norm

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	.443	.091		4.869	.000
	Management agree index_norm	-.003	.001	-.139	-2.287	.023
	tenure	.003	.002	.078	1.305	.193
	Percent_Merizo	.007	.001	.464	7.646	.000

a. Dependent Variable: familiar_Achang

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%*

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.470[a]	.221	.210	.441

a. Predictors: (Constant), Percent_Merizo, tenure, Management agree index_norm

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	.443	.091		4.869	.000
	Management agree index_norm	-.003	.001	-.139	-2.287	.023
	tenure	.003	.002	.078	1.305	.193
	Percent_Merizo	.007	.001	.464	7.646	.000

a. Dependent Variable: familiar_Achang

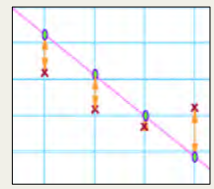
Save your output as "Manell_Geus_Output_multiple regression.spv"

(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)*



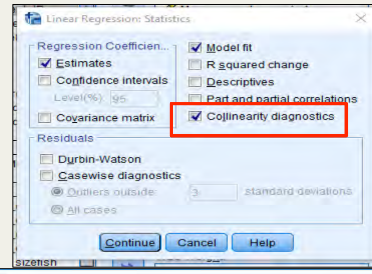
- Autocorrelation
  - *A phenomenon in which a variable's value is a function of its past values*
  - *Usually only a problem with time series data*
- If your model exhibits ANY of the above characteristics, it is INVALID

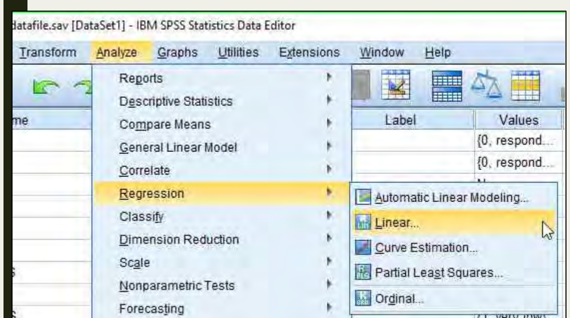
Tests to Determine Validity

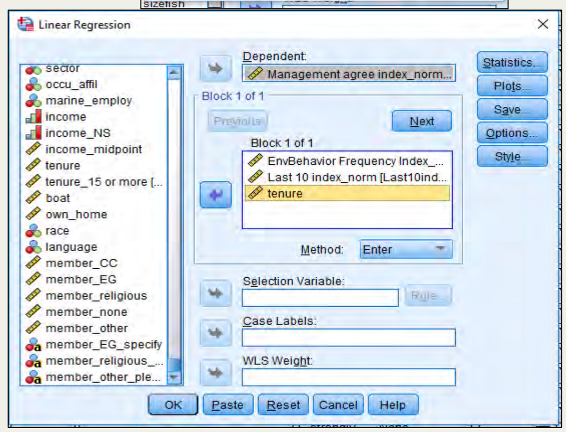
- Multicollinearity
  - *Collinearity diagnostics (under "Statistics" in our regression box)*
  - *Use the VIF (variance inflation factor)*
- Heteroscedasticity
  - *Breusch-Pagan test*
    - Not as easy to do in SPSS, but there is a way
- Autocorrelation
  - *Durbin-Watson test*
  - *However, since we are not dealing with time series data (we are dealing with survey data),* ***we will only focus on diagnosing multicollinearity and heteroscedasicity***

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"*







Testing for Multicollinearity in SPSS

Coefficients[a]

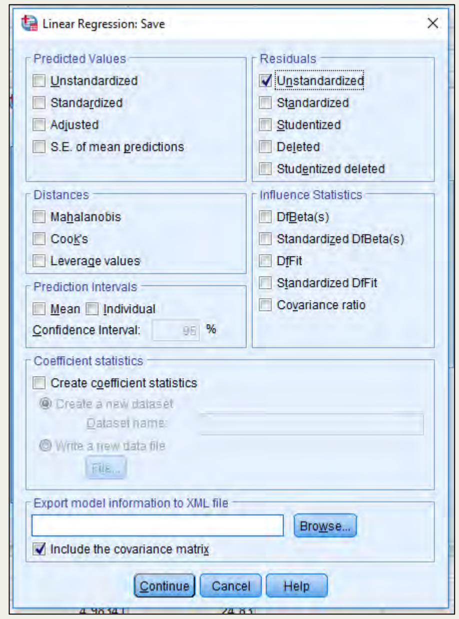
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	17.658	5.841		3.023	.003		
	EnvBehavior Frequency index_norm	.261	.088	.228	2.960	.004	.793	1.261
	Last 10 index_norm	.317	.069	.342	4.570	.000	.839	1.192
	tenure	.040	.106	.027	.379	.705	.938	1.066

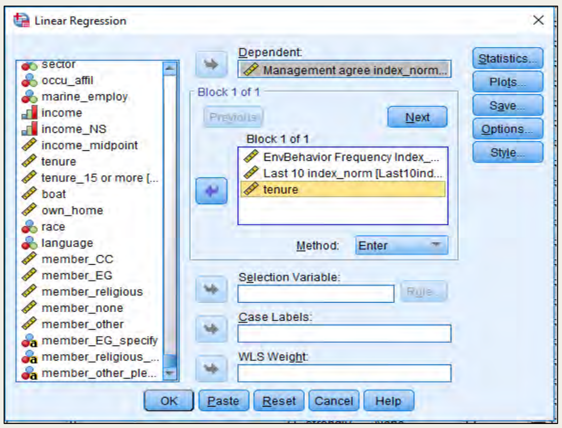
a. Dependent Variable: Management agree index_norm

- 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

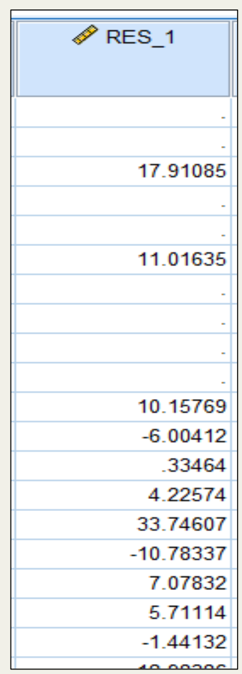
- To perform the Breusch-Pagan test, we need to "save" a new "residuals variable"
  - *In the "save" box, check the box for "unstandardized residuals"*

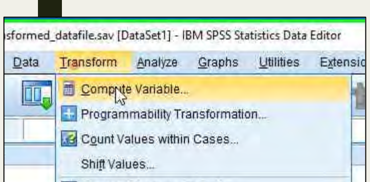


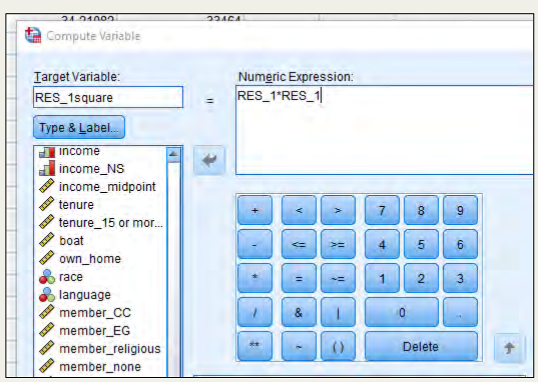


Testing for Heteroscedasticity in SPSS

- Now, a new variable was created in our data set
  - *RES_1 = unstandardized residuals*
- From here, we must compute the squared residuals

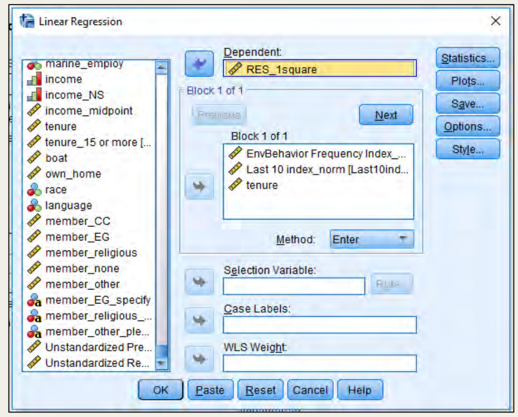


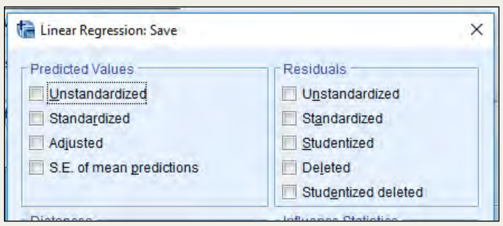




Testing for Heteroscedasticity in SPSS

- 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"





Testing for Heteroscedasticity in SPSS

- 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**

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.169[a]	.029	.011	443.18870

a. Predictors: (Constant), tenure, Last 10 index_norm, EnvBehavior Frequency Index_norm

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	958547.565	3	319515.855	1.627	.185[b]
	Residual	32408677.62	165	196416.228		
	Total	33367225.18	168			

a. Dependent Variable: RES_1square

b. Predictors: (Constant), tenure, Last 10 index_norm, EnvBehavior Frequency Index_norm

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	305.030	133.819		2.279	.024
	EnvBehavior Frequency Index_norm	-1.774	2.016	-.076	-.880	.380
	Last 10 index_norm	1.365	1.591	.072	.858	.392
	tenure	4.195	2.428	.137	1.728	.086

a. Dependent Variable: RES_1square

Testing for multicollinearity and heteroscedasticity in SPSS: Practice!

- Let's examine another of our previously ran regression models
  - *AchangPreserveIndex_norm = 38.132 – (12.646)*college + (0.186)*tenure + (11.548)*Program_MC + (0.249)*ConditionIndex_norm*

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	.443	.091		4.869	.000
	Management agree index_norm	-.003	.001	-.139	-2.287	.023
	tenure	.003	.002	.078	1.305	.193
	Percent_Merizo	.007	.001	.464	7.646	.000

a. Dependent Variable: familiar_Achang

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.470[a]	.221	.210	.441

a. Predictors: (Constant), Percent_Merizo, tenure, Management agree index_norm

Testing for multicollinearity in SPSS: Practice!

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	30.256	6.072		4.983	.000		
	Management agree index_norm	.445	.094	.433	4.714	.000	.957	1.044
	tenure	-.043	.129	-.030	-.332	.740	.995	1.005
	Percent_Merizo	.122	.061	.183	1.997	.049	.961	1.041

a. Dependent Variable: Achang Preserve Index_norm

- All VIFs<10, our model does not exhibit multicollinearity

Testing for heteroscedasticity in SPSS: Practice!

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	781343.943	3	260447.981	.955	.418[b]
	Residual	25089468.30	92	272711.612		
	Total	25870812.25	95			

a. Dependent Variable: RES_2square

b. Predictors: (Constant), Percent_Merizo, tenure, Management agree index_norm

- F test p-value = 0.418 > 0.05
- Fail 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*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*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*

threat_bleach * male gender Crosstabulation

			male gender		
			female	male	Total
threat_bleach	respondent did not chose as top 3	Count	159	130	289
		% within male gender	95.8%	95.6%	95.7%
	respondent chose as top 3	Count	7	6	13
		% within male gender	4.2%	4.4%	4.3%
Total		Count	166	136	302
		% within male gender	100.0%	100.0%	100.0%

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*

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.007[a]	1	.934		
Continuity Correction[b]	.000	1	1.000		
Likelihood Ratio	.007	1	.934		
Fisher's Exact Test				1.000	.576
Linear-by-Linear Association	.007	1	.934		
N of Valid Cases	302				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 5.85.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Kendall's tau-b	.005	.058	.083	.934
N of Valid Cases		302			

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

	Female	Male	Total
Does not believe that coral bleaching is a top threat to coral reefs	95.8%	95.6%	289
Believes that coral bleaching is a top threat to coral reefs	4.2%	4.4%	13
Total	166	136	302
Chi-Square statistic			0.007
Chi-square p-value			0.934
Tau-b association measure			0.005

- ❖ Does Gender have an effect on the belief that coral bleaching is a top threat to coral reefs?
  - ❖ No, it does not
  - ❖ 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

activity_hookline * boat Crosstabulation

			boat		
			no	yes	Total
activity_hookline	no	Count	45	7	52
		% within boat	22.4%	15.2%	21.1%
	yes, in Cocos Lagoon	Count	78	24	102
		% within boat	38.8%	52.2%	41.3%
	yes, in Achang Preserve	Count	31	1	32
		% within boat	15.4%	2.2%	13.0%
	Yes, in both places	Count	47	14	61
		% within boat	23.4%	30.4%	24.7%
Total		Count	201	46	247
		% within boat	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	8.360[a]	3	.039
Likelihood Ratio	10.467	3	.015
Linear-by-Linear Association	.207	1	.649
N of Valid Cases	247		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 5.96.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.184	.039
	Cramer's V	.184	.039
N of Valid Cases		247	

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

Does owning a boat affect where someone uses a hookline?

	Does not own a boat	Owns a boat	Total
Does not use a hookline	22.4%	15.2%	52
Uses a hookline in Cocos Lagoon	38.8%	52.2%	102
Uses a hookline in Achang Preserve	15.4%	2.2%	32
Uses a hookline in Cocos Lagoon and Achang Preserve	23.4%	30.4%	61
Total	201	46	247
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

condition_numfish_NS * Fish_harvest Crosstabulation

			Fish_harvest no	Fish_harvest yes	Total
condition_numfish_NS	very bad	Count	13	8	21
		% within Fish_harvest	9.6%	5.4%	7.4%
	bad	Count	31	13	44
		% within Fish_harvest	23.0%	8.8%	15.5%
	neither good nor bad	Count	25	42	67
		% within Fish_harvest	18.5%	28.4%	23.7%
	good	Count	51	61	112
		% within Fish_harvest	37.8%	41.2%	39.6%
	very good	Count	15	24	39
		% within Fish_harvest	11.1%	16.2%	13.8%
Total		Count	135	148	283
		% within Fish_harvest	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	15.272[a]	4	.004
Likelihood Ratio	15.540	4	.004
Linear-by-Linear Association	7.323	1	.007
N of Valid Cases	283		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 10.02.

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Kendall's tau-c	.163	.065	2.503	.012
N of Valid Cases		283			

Making “Report Ready” Contingency Tables: Practice

	Does not fish or harvest for marine resources	Does fish or harvest for marine resources	Total
Believes that the condition of the number of fish is very bad	9.6%	5.4%	21
Believes that the condition of the number of fish is bad	23.0%	8.8%	44
Believes that the condition of the number of fish is neither good nor bad	18.5%	28.4%	67
Believes that the condition of the number of fish is good	37.8%	41.2%	112
Believes that the condition of the number of fish is very good	11.1%	16.2%	39
Total	135	148	283
Chi-Square statistic			15.272
Chi-square p-value			0.004
Tau-b association measure			0.163

- 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)

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

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Weight Before	200.56	50	58.208	8.232
	Weight After	187.40	50	52.594	7.438

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Weight Before & Weight After	50	.956	.000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Weight Before - Weight After	13.160	17.286	2.445	8.248	18.072	5.383	49	.000

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

Sample Size	Average weight before weight loss treatment	Average weight after weight loss treatment	Mean Difference	T statistic	P value
50	200.56	187.4	-13.16	5.383	0.000

- ❖ 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

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Cholesterol Before	183.44	50	44.613	6.309
	Cholesterol After	170.64	50	38.681	5.470

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Cholesterol Before & Cholesterol After	50	.926	.000

Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	Cholesterol Before - Cholesterol After	12.800	17.087	2.416	7.944	17.656	5.297	49	.000

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

Sample Size	Average cholesterol before weight loss treatment	Average cholesterol after weight loss treatment	Mean Difference	T statistic	P value
50	183.44	170.64	-12.8	5.297	0.000

- ❖ 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

Group Statistics

	college	N	Mean	Std. Deviation	Std. Error Mean
condition_beach_NS	did not complete college	227	2.89	1.073	.071
	completed college	64	3.00	1.234	.154

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means					95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
condition_beach_NS	Equal variances assumed	1.560	.213	-.701	289	.484	-.110	.157	-.419	.199
	Equal variances not assumed			-.648	91.547	.519	-.110	.170	-.448	.227

- 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?

	Did not complete college		Completed college		Mean difference	T statistic	P value
	Sample size	Mean	Sample size	Mean			
Condition of beach/shoreline	227	2.89	64	3.00	-0.11	-0.701	0.484

- ❖ 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

Visualizing Independent Samples T-test Results: Practice

Group Statistics

	fish_sell	N	Mean	Std. Deviation	Std. Error Mean
mng_tradfish_NS	>= 2	106	3.57	1.235	.120
	< 2	28	3.32	1.249	.236

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality o			
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference
mng_tradfish_NS	Equal variances assumed	.515	.474	.930	132	.354	.245
	Equal variances not assumed			.924	42.020	.361	.245

- Create a report ready table based off of this output

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

	Does not fish to sell		Does fish to sell		Mean difference	T statistic	P value
	Sample size	Mean	Sample size	Mean			
Agreement with creating areas for only traditional fishing	106	3.57	28	3.32	0.245	0.93	0.354

- ❖ 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

Last 10 index_norm

	N	Mean
no	50	56.44
yes, in Cocos Lagoon	88	46.09
yes, in Achang Preserve	16	56.51
Yes, in both places	55	57.65
Total	209	52.41

Multiple Comparisons

Dependent Variable: Last 10 index_norm

Tukey HSD

(I) activity_beach	(J) activity_beach	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
no	yes, in Cocos Lagoon	10.359	4.157	.064	-.41	21.13
	yes, in Achang Preserve	-.066	6.742	1.000	-17.53	17.40
	Yes, in both places	-1.207	4.586	.994	-13.09	10.67
yes, in Cocos Lagoon	no	-10.359	4.157	.064	-21.13	.41
	yes, in Achang Preserve	-10.425	6.379	.362	-26.95	6.10
	Yes, in both places	-11.566*	4.034	.024	-22.02	-1.12
yes, in Achang Preserve	no	.066	6.742	1.000	-17.40	17.53
	yes, in Cocos Lagoon	10.425	6.379	.362	-6.10	26.95
	Yes, in both places	-1.141	6.667	.998	-18.41	16.13
Yes, in both places	no	1.207	4.586	.994	-10.67	13.09
	yes, in Cocos Lagoon	11.566*	4.034	.024	1.12	22.02
	yes, in Achang Preserve	1.141	6.667	.998	-16.13	18.41

*. The mean difference is significant at the 0.05 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?

	(1) Does not participate in beach recreation		(2) Participates in beach recreation in Cocos Lagoon		(3) Participates in beach recreation in Achang Preserve		(4) Participates in beach recreation in both places		Significance Between Groups	p-value
	Sample Size	Mean	Sample Size	Mean	Sample Size	Mean	Sample Size	Mean		
Last 10 Years Index	50	56.44	88	46.09	16	56.51	55	57.65	4>2**	0.024

** = 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 amore 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

Condition index_norm

	N	Mean
would not do	4	35.42
would consider	82	47.59
would do	152	58.11
Total	238	54.11

Multiple Comparisons

Dependent Variable: Condition index_norm

Tukey HSD

(I) volunteer_protect_NS	(J) volunteer_protect_NS	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
would not do	would consider	-12.178	11.063	.515	-38.27	13.92
	would do	-22.697	10.944	.097	-48.51	3.12
would consider	would not do	12.178	11.063	.515	-13.92	38.27
	would do	-10.519*	2.960	.001	-17.50	-3.54
would do	would not do	22.697	10.944	.097	-3.12	48.51
	would consider	10.519*	2.960	.001	3.54	17.50

*. 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?

	(1) Would not volunteer for an activity that will help to protect the reefs at least once a year		(2) Would consider volunteering for an activity that will help to protect the reefs at least once a year		(3) Would volunteer for an activity that will help to protect the reefs at least once a year		Significance Between Groups	p-value
	Sample Size	Mean	Sample Size	Mean	Sample Size	Mean		
Condition Index	4	35.42	82	47.59	125	58.11	3>2***	0.001

*** = significant at the 1% 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)

Correlations

		Fish_harvest	age	tenure	high school	college	male gender	family	own_home	income_midpoint
Fish_harvest	Pearson Correlation	1	-.003	-.019	.062	-.060	.096	.060	-.056	-.128
	Sig. (2-tailed)		.961	.741	.290	.304	.096	.301	.333	.273
	N	303	301	295	296	296	300	300	297	75
age	Pearson Correlation	-.003	1	.532**	-.040	.154**	.036	-.151**	.324**	.337**
	Sig. (2-tailed)	.961		.000	.493	.008	.537	.008	.000	.003
	N	301	304	297	297	297	302	303	299	74
tenure	Pearson Correlation	-.019	.532**	1	.029	-.038	.072	-.015	.397**	-.122
	Sig. (2-tailed)	.741	.000		.620	.523	.217	.801	.000	.299
	N	295	297	298	291	291	296	296	295	74
high school	Pearson Correlation	.062	-.040	.029	1	.135*	.065	-.043	.090	.200
	Sig. (2-tailed)	.290	.493	.620		.020	.265	.458	.124	.088
	N	296	297	291	299	299	297	296	293	74
college	Pearson Correlation	-.060	.154**	-.038	.135*	1	.056	-.030	.064	.698**
	Sig. (2-tailed)	.304	.008	.523	.020		.335	.609	.273	.000
	N	296	297	291	299	299	297	296	293	74
male gender	Pearson Correlation	.096	.036	.072	.065	.056	1	-.059	.027	-.027
	Sig. (2-tailed)	.096	.537	.217	.265	.335		.308	.643	.819
	N	300	302	296	297	297	303	301	298	74
family	Pearson Correlation	.060	-.151**	-.015	-.043	-.030	-.059	1	-.117*	-.316**
	Sig. (2-tailed)	.301	.008	.801	.458	.609	.308		.044	.006
	N	300	303	296	296	296	301	303	298	74
own_home	Pearson Correlation	-.056	.324**	.397**	.090	.064	.027	-.117*	1	.060
	Sig. (2-tailed)	.333	.000	.000	.124	.273	.643	.044		.613
	N	297	299	295	293	293	298	298	300	74
income_midpoint	Pearson Correlation	-.128	.337**	-.122	.200	.698**	-.027	-.316**	.060	1
	Sig. (2-tailed)	.273	.003	.299	.088	.000	.819	.006	.613	
	N	75	74	74	74	74	74	74	74	75

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

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

Correlation Matrix – Visualizing Correlations

- Copy and paste SPSS output into excel and work from there

	Fish or Harvests for marine resources	Age	Number of years lived in Merizo	Completed high school	Completed college	Male gender	Number of family members	Owns a home	Annual household income
Fish or Harvests for marine resources	1.000								
Age	-0.003	1.000							
Number of years lived in Merizo	-0.019	0.532**	1.000						
Completed high school	0.062	-0.040	0.029	1.000					
Completed college	-0.060	0.154**	-0.038	0.135*	1.000				
Male gender	0.096	0.036	0.072	0.065	0.056	1.000			
Number of family members	0.060	-0.151**	-0.015	-0.043	-0.030	-0.059	1.000		
Owns a home	-0.056	0.324**	0.397**	0.090	0.064	0.027	-0.117*	1.000	
Annual household income	-0.128	0.337**	-0.122	0.200	0.698**	-0.027	-0.316**	0.060	1.000

**. Correlation is significant at the 0.01 level

*. Correlation is significant at the 0.05 level

Correlation Matrix – Visualizing Correlations: Practice

Correlations

		Fishing index_norm	Condition index_norm	Last 10 index_norm	Flood severity index_norm	Flood severity index_norm	Achang Preserve index_norm	Management efficiency index_norm	Management Success index_norm	Volunteer index_norm	EnvBehavior Frequency index_norm	Management agree index_norm
Fishing index_norm	Pearson Correlation	1	.078	.026	.401**	.257**	.400**	.254*	.208*	.186	.277**	.191*
	Sig. (2-tailed)		.400	.775	.000	.003	.000	.014	.027	.055	.003	.043
	N	133	120	121	128	129	78	93	113	107	113	113
Condition index_norm	Pearson Correlation	.078	1	.828**	.180**	.152*	.294**	.399**	.161*	.134	.272**	.382**
	Sig. (2-tailed)	.400		.000	.007	.021	.003	.000	.019	.060	.000	.000
	N	120	246	230	226	230	99	124	213	199	209	204
Last 10 index_norm	Pearson Correlation	.026	.828**	1	.087	.022	.216*	.391**	.162*	.202**	.368**	.311**
	Sig. (2-tailed)	.775	.000		.192	.745	.028	.000	.019	.005	.000	.000
	N	121	230	241	224	226	104	128	210	195	205	200
Flood severity index_norm	Pearson Correlation	.401**	.180**	.087	1	.713**	.426**	.204*	.552**	.004	.139*	.257**
	Sig. (2-tailed)	.000	.007	.192		.000	.000	.016	.000	.959	.036	.000
	N	128	226	224	272	257	112	138	228	217	226	225
Flood severity index_norm	Pearson Correlation	.257**	.152*	.022	.713**	1	.282**	.122	.498**	.076	.130	.209**
	Sig. (2-tailed)	.003	.021	.745	.000		.002	.147	.000	.260	.051	.001
	N	129	230	226	257	281	114	143	232	222	225	229
Achang Preserve index_norm	Pearson Correlation	.400**	.294**	.216*	.426**	.282**	1	.757**	.412**	.432**	.234*	.452**
	Sig. (2-tailed)	.000	.003	.028	.000	.002		.000	.000	.000	.022	.000
	N	78	99	104	112	114	120	118	106	96	95	101
Management efficiency index_norm	Pearson Correlation	.254*	.399**	.391**	.204*	.122	.757**	1	.258**	.444**	.256**	.381**
	Sig. (2-tailed)	.014	.000	.000	.016	.147	.000		.004	.000	.005	.000
	N	93	124	128	138	143	118	153	125	118	120	122
Management Success index_norm	Pearson Correlation	.208*	.161*	.162*	.552**	.498**	.412**	.258**	1	-.059	.079	.166*
	Sig. (2-tailed)	.027	.019	.019	.000	.000	.000	.004		.405	.252	.018
	N	113	213	210	228	232	106	125	244	199	211	205
Volunteer index_norm	Pearson Correlation	.186	.134	.202**	.004	.076	.432**	.444**	-.059	1	.384**	.319**
	Sig. (2-tailed)	.055	.060	.005	.959	.260	.000	.000	.405		.000	.000
	N	107	199	195	217	222	96	118	199	238	197	192
EnvBehavior Frequency index_norm	Pearson Correlation	.277**	.272**	.368**	.139*	.130	.234*	.256**	.079	.384**	1	.371**
	Sig. (2-tailed)	.003	.000	.000	.036	.051	.022	.005	.252	.000		.000
	N	113	209	205	226	225	95	120	211	197	241	198
Management agree index_norm	Pearson Correlation	.191*	.382**	.311**	.257**	.209**	.452**	.381**	.166*	.319**	.371**	1
	Sig. (2-tailed)	.043	.000	.000	.000	.001	.000	.000	.018	.000	.000	
	N	113	204	200	225	229	101	122	205	192	198	246

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

Correlation Matrix – Visualizing Correlations: Practice

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

**. 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

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.476[a]	.227	.213	19.344

a. Predictors: (Constant), tenure, Last 10 index_norm, EnvBehavior Frequency Index_norm

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	17.658	5.841		3.023	.003
	EnvBehavior Frequency Index_norm	.261	.088	.228	2.960	.004
	Last 10 index_norm	.317	.069	.342	4.570	.000
	tenure	.040	.106	.027	.379	.705

a. Dependent Variable: Management agree index_norm

Visualizing Regression Results

Dependent Variable: Management Agreement Index	Rsqaure adj = 0.213		
Independent Variable	Coefficient	T Statistic	P value
Constant	17.658		
Environmental Behavior Frequency Index	0.261***	2.960	0.004
Last 10 Index	0.317***	4.570	0.000
Amount of years lived in Merizo	0.040	0.379	0.705

*** = 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

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.290[a]	.084	.055	.485

a. Predictors: (Constant), benefit, Management agree index_norm, college, age

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	.465	.152		3.062	.003
	age	-.002	.003	-.050	-.564	.574
	college	-.245	.111	-.195	-2.212	.029
	Management agree index_norm	.005	.002	.196	2.239	.027
	benefit	-.139	.107	-.114	-1.296	.197

a. Dependent Variable: program_CRCP

Visualizing Regression Results: Practice

Dependent Variable: Familiarity with CRCP — Rsqaure adj = 0.055

Independent Variable	Coefficient	T Statistic	P value
Constant	0.465		
Age	-0.002	-0.564	0.574
Completed College	-0.245**	-2.212	0.029
Management Agreement Index	0.005**	2.239	0.027
Receives benefit form Achang Preserve	-0.139	-1.296	0.197

** = 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

Quiz #9

Day 5: September 16, 2016

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?

	Fish or Harvests for marine resources	Age	Number of years lived in Merizo	Completed high school	Completed college	Male gender	Number of family members	Owns a home	Annual household income
Fish or Harvests for marine resources	1	-0.003	-0.019	0.062	-0.060	0.096	0.060	-0.056	-0.128
Age	-0.003	1	.532**	-0.040	.154**	0.036	-.151**	.324**	.337**
Number of years lived in Merizo	-0.019	.532**	1	0.029	-0.038	0.072	-0.015	.397**	-0.122
Completed high school	0.062	-0.040	0.029	1	.135*	0.065	-0.043	0.090	0.200
Completed college	-0.060	.154**	-0.038	.135*	1	0.056	-0.030	0.064	.698**
Male gender	0.096	0.036	0.072	0.065	0.056	1	-0.059	0.027	-0.027
Number of family members	0.060	-.151**	-0.015	-0.043	-0.030	-0.059	1	-.117*	-.316**
Owns a home	-0.056	.324**	.397**	0.090	0.064	0.027	-.117*	1	0.060
Annual household income	-0.128	.337**	-0.122	0.200	.698**	-0.027	-.316**	0.060	1

**. 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"?

Dependent Variable: Familiarity with CRCP	Rsqaure adj = 0.055	
Independent Variable	Coefficient	T Statistic
Constant	0.465	
Age	-0.002	-0.564
Completed College	-0.245**	-2.212
Management Agreement Index	0.005**	2.239
Receives benefit form Achang Preserve	-0.139	-1.296

** = 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

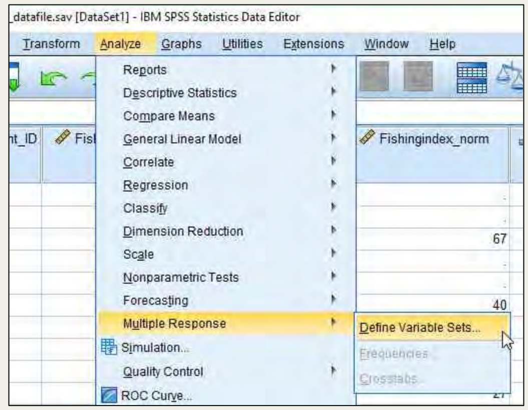
Multiple response analysis

- 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*

threat_sewage	threat_typhoon	threat_runoff	threat_shoreerosion	threat_algal
1	0	0	0	0
0	0	1	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	1	0
0	1	0	0	0
0	1	0	0	0
0	1	0	0	0
0	0	0	1	0
0	0	1	0	1
0	0	0	0	0
0	1	0	0	1
0	0	0	0	0
0	0	1	0	0

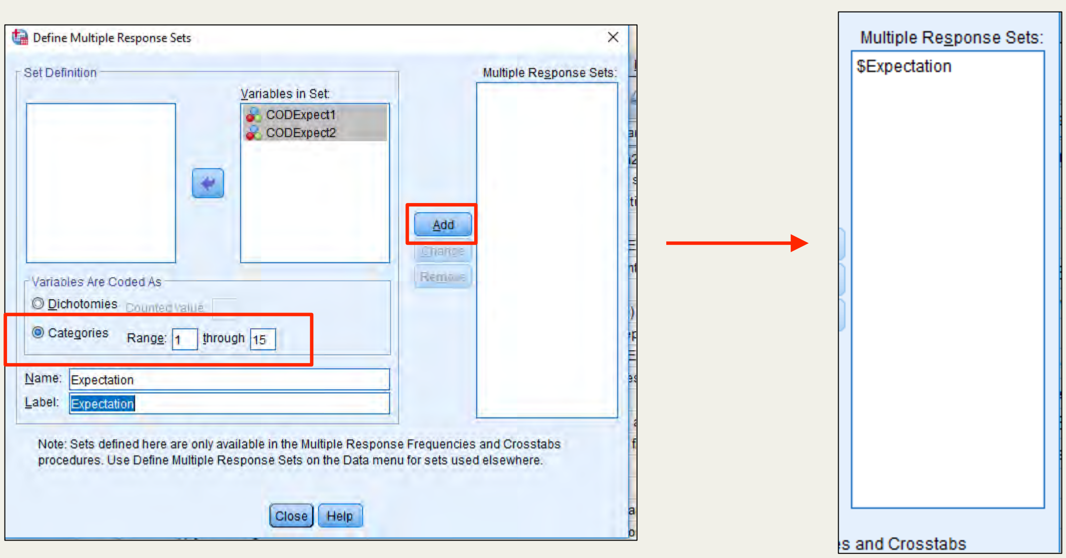
What if respondents are asked to rank?

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

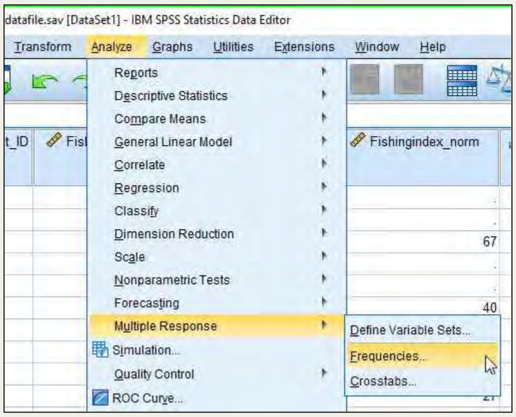


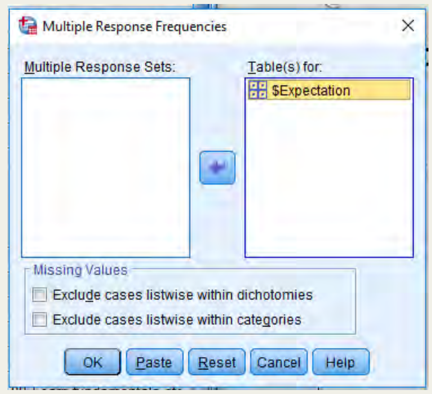
4. From a job perspective, what are your top 2 main expectations from the training course?
1. ______________________________
2. ______________________________

Defining Variable Sets



Frequencies of Multiple Responses





Output

Case Summary

	Cases: Valid N	Valid Percent	Missing N	Missing Percent	Total N	Total Percent
$Expectation[a]	69	55.6%	55	44.4%	124	100.0%

$Expectation Frequencies

		Responses N	Responses Percent	Percent of Cases
Expectation[a]	Knowledge fisheries management, effective fisheries management, governance	8	6.2%	11.6%
	Knowledge EAFM, EAFM process, EAFM principles	34	26.4%	49.3%
	Understanding of issues, threats, expectations, fishers	5	3.9%	7.2%
	Maximize stakeholder engagement, participation, co-management, collaboratiion with others	7	5.4%	10.1%
	Start up/prepare	1	0.8%	1.4%
	(Develop) Fisheries policies, management plans and actions	10	7.8%	14.5%
	Fisheries rules and regulations, enforcement	6	4.7%	8.7%
	EAFM application to local settings or in my responsibility/area, implementation, hands-on experience, coordination	29	22.5%	42.0%
	EBM, integrated approach to resource management	4	3.1%	5.8%
	Improve aquaculture	2	1.6%	2.9%
	Application to climate change, environmental risk analysis	5	3.9%	7.2%
	Facilitation skills/learning/training skills/conflict management/communications	7	5.4%	10.1%
	Evaluation of fisheries resource, environmental services, MCS	2	1.6%	2.9%
	Sustainable financing	1	0.8%	1.4%
	Others	8	6.2%	11.6%
Total		129	100.0%	187.0%

- 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%)*

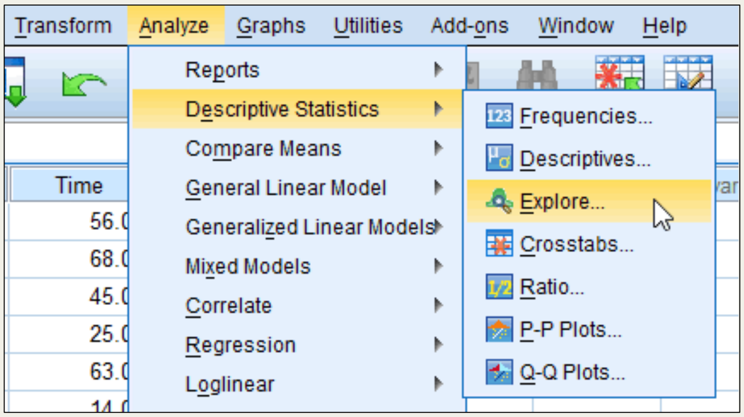
Normality and Non Parametric Tests

Day 6: September 17, 2016

The Central Limit Theorem

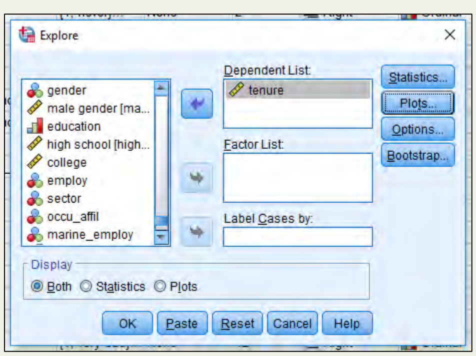
- Given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined (finite) expected value and finite variance, will be approximately normally distributed, regardless of the underlying distribution.
- In English:
  - *If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

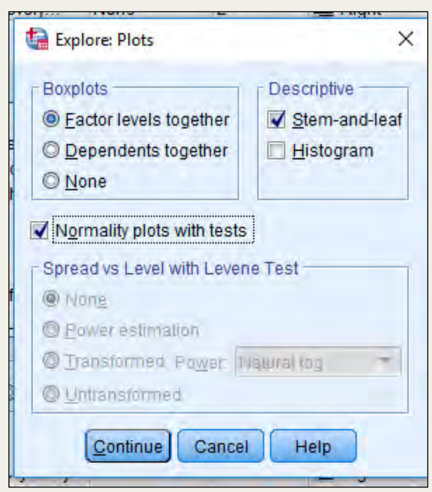
Assessing Normality in SPSS



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

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
tenure	.173	298	.000	.914	298	.000

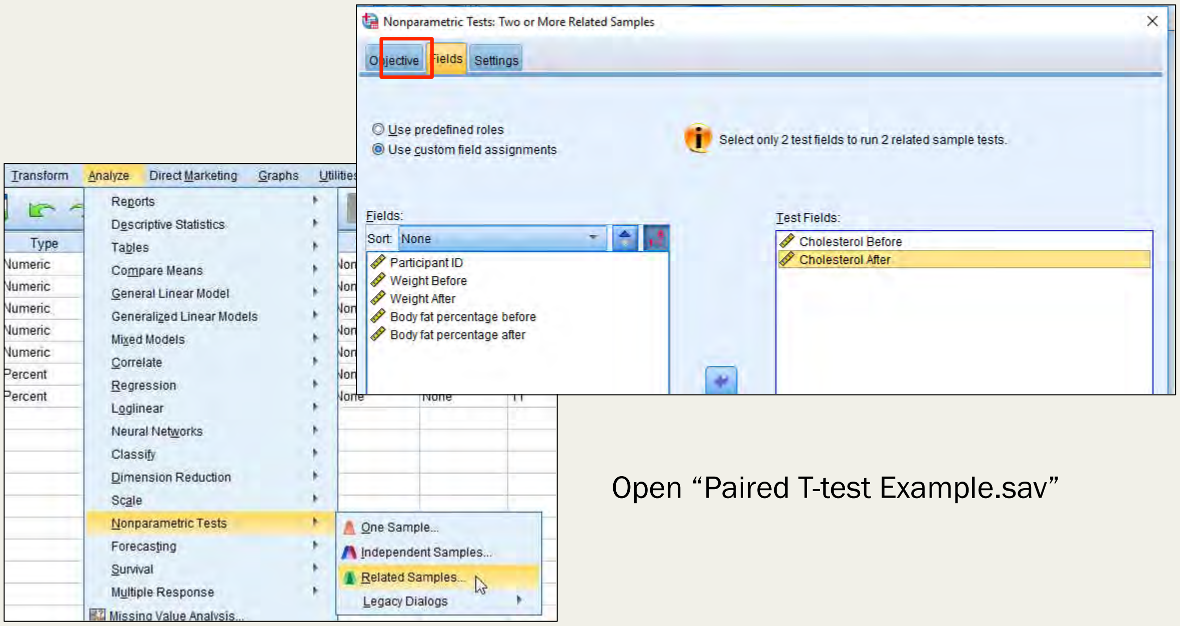
a. Lilliefors Significance Correction

- **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

What to do if data is not normal?

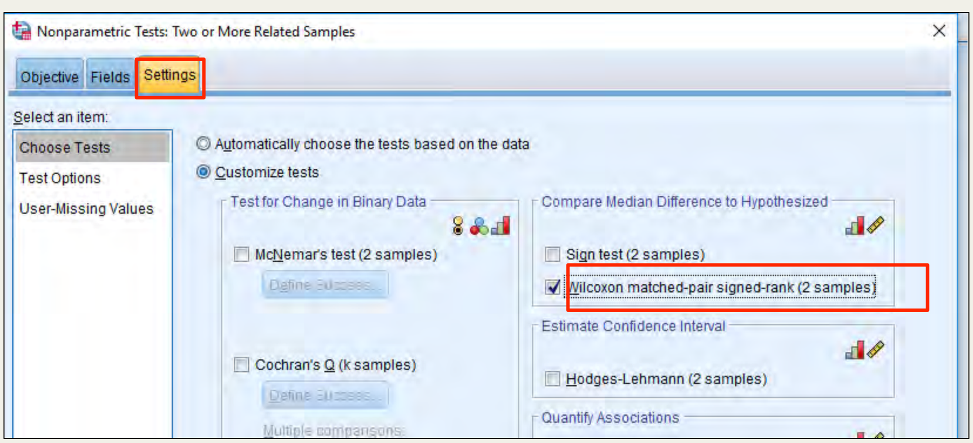
- Use non-parametric tests
  - *Non parametric tests do not assume normality*
- Types of non parametric tests:
  - *Wilcoxon signed-rank test (non parametric equivalent of paired samples t test)*
  - *Mann-Whitney test (non parametric equivalent of independent samples t test)*
  - *Kruskall Wallis test (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



Wilcoxon signed-rank test in SPSS

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between Cholesterol Before and Cholesterol After equals 0.	Related-Samples Wilcoxon Signed Rank Test	.000	Reject the null hypothesis.

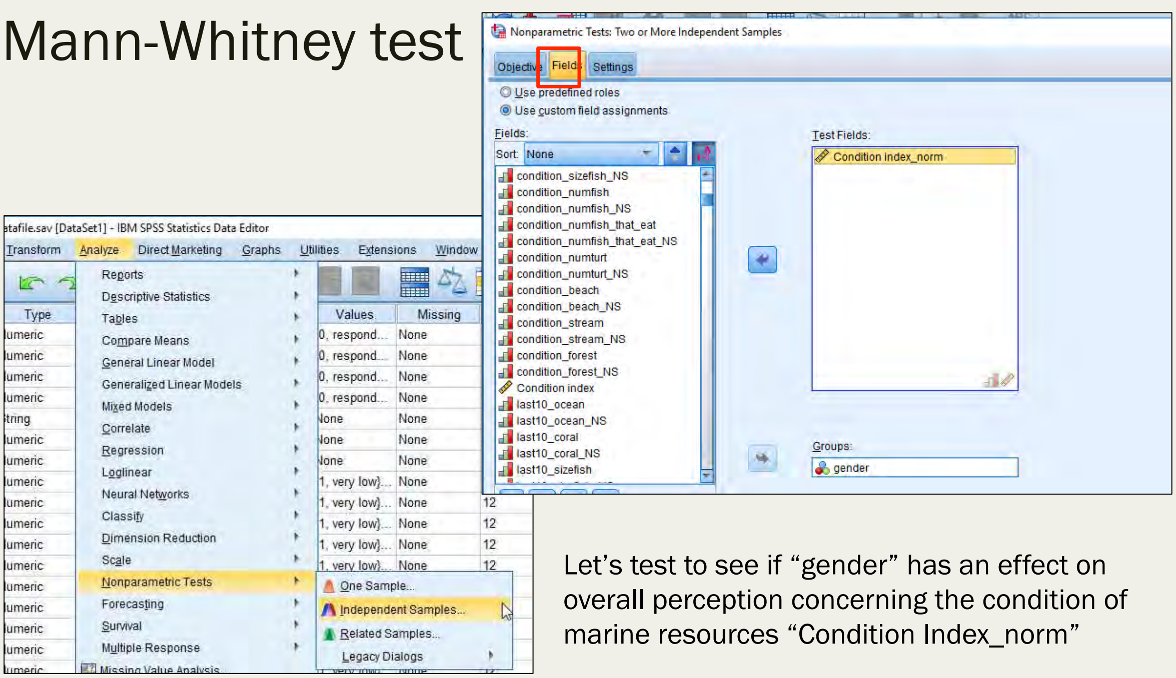
Asymptotic significances are displayed. The significance level is .05.

Statistics

		Cholesterol Before	Cholesterol After
N	Valid	50	50
	Missing	0	0
Median		182.00	168.50

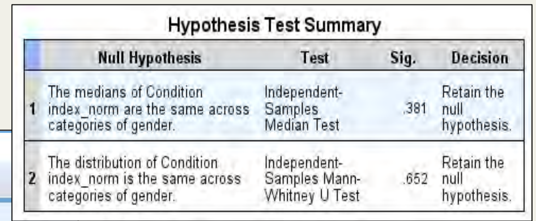
- P-value = 0.000 < 0.05
- The median "Cholesterol After" is significantly less than the median "Cholesterol before"
- There has been a statistically significant reduction in cholesterol since the weight loss treatment

Mann-Whitney test

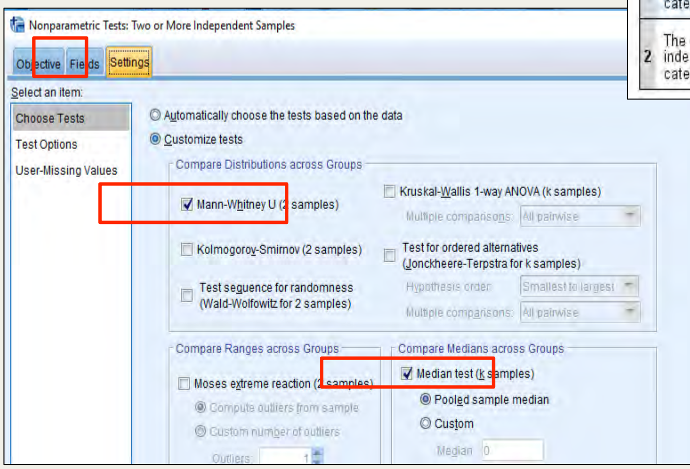


Let's test to see if "gender" has an effect on overall perception concerning the condition of marine resources "Condition Index_norm"

Mann-Whitney test

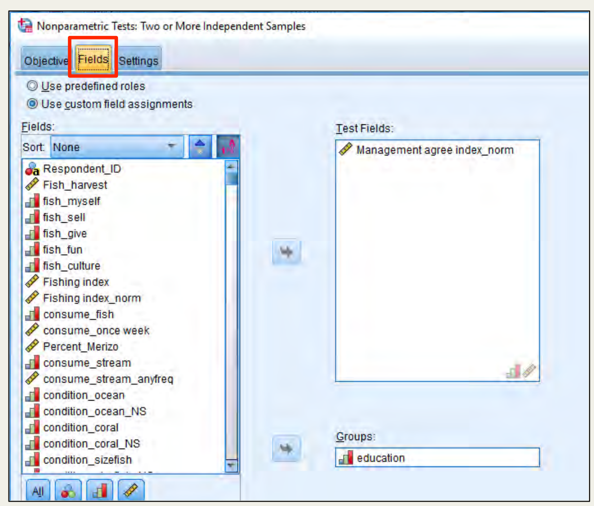

Hypothesis Test Summary

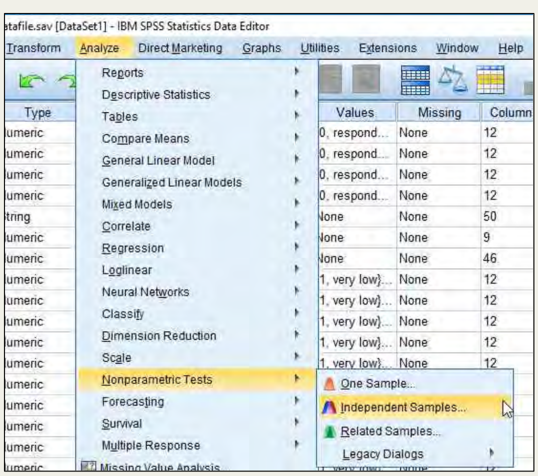
	Null Hypothesis	Test	Sig.	Decision
1	The medians of Condition index_norm are the same across categories of gender.	Independent-Samples Median Test	.381	Retain the null hypothesis.
2	The distribution of Condition index_norm is the same across categories of gender.	Independent-Samples Mann-Whitney U Test	.652	Retain the null hypothesis.



- 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

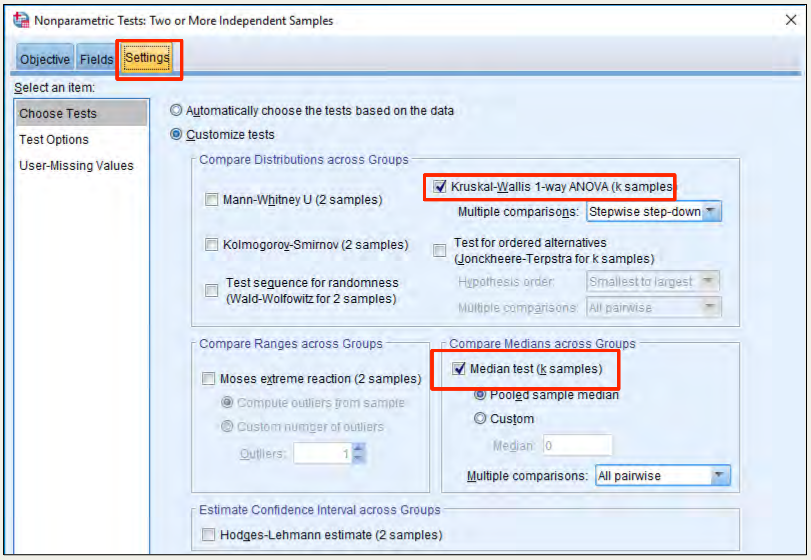
Kruskall Wallis Test





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

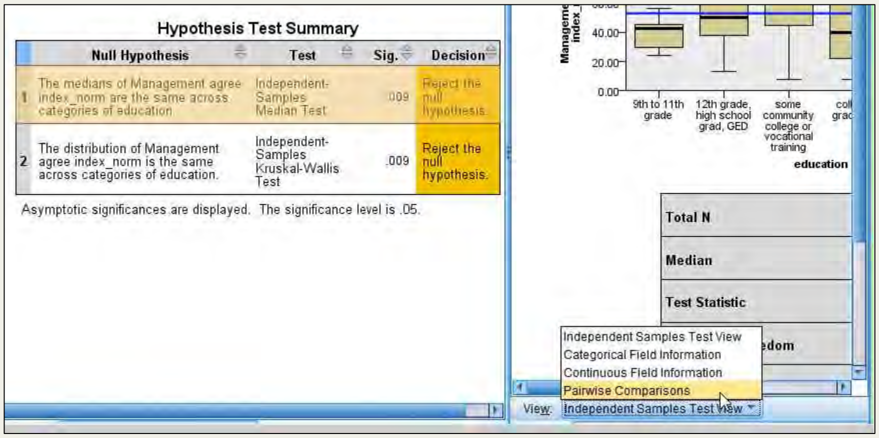
Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The medians of Management agree index_norm are the same across categories of education.	Independent-Samples Median Test	.009	Reject the null hypothesis.
2	The distribution of Management agree index_norm is the same across categories of education.	Independent-Samples Kruskal-Wallis Test	.009	Reject the null hypothesis.

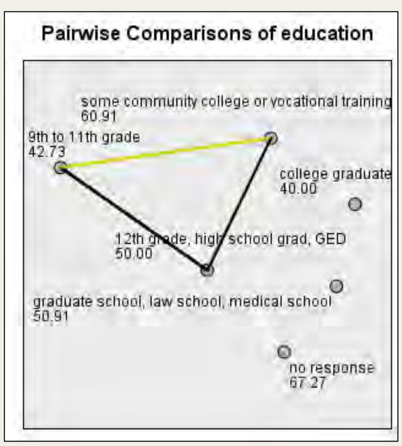
Asymptotic significances are displayed. The significance level is .05.

- 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



Kruskall Wallis 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 use 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

National Institute of Health (NIH), Office of Human Subject Research. 1979. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research.*

http://videocast.nih.gov/pdf/ohrp_belmont_report.pdf

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?

Quality & Quantity **36:** 43–53, 2002.

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?

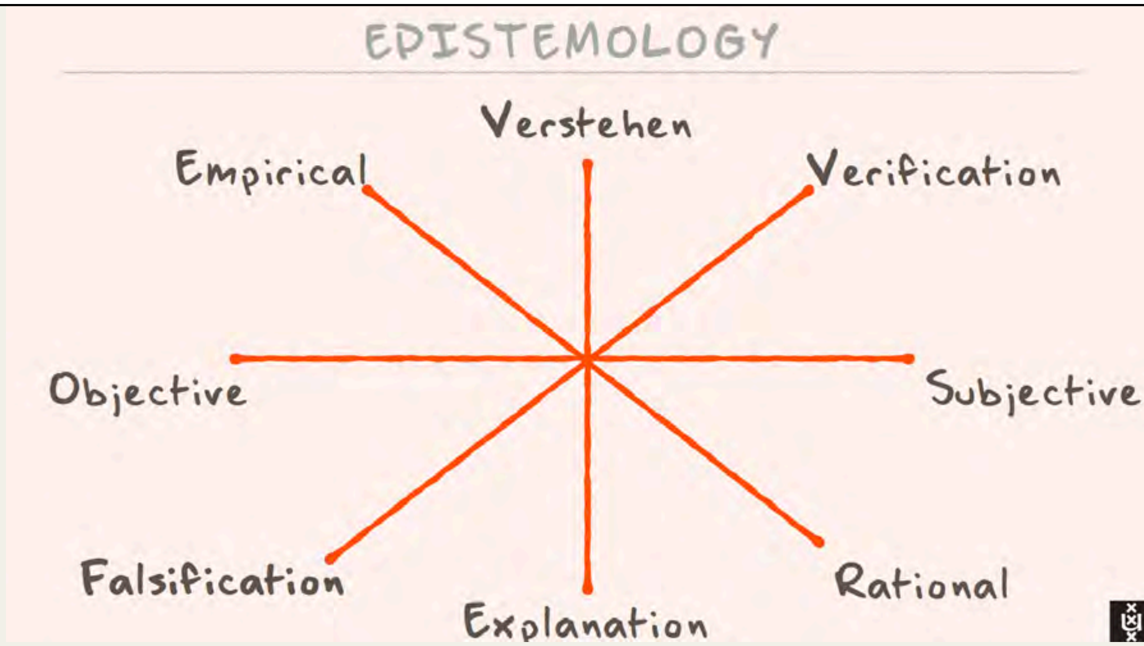
- The quantitative paradigm is based on positivism
  - *All phenomena can be reduced to empirical indicators which represent the truth*
  - *The position of the quantitative paradigm is that there is only one truth, an objective reality that exists independent of human perception*
  - *NUMBERS*
- The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



Source: Coursera, University of Amsterdam

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

(Rocheleau 1995)

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

Presenter(s), Department(s):
John Creswell
Professor
Department of Educational Psychology
University of Nebraska-Lincoln

Title:
Steps in Conducting a Scholarly Mixed Methods Study

Abstract:
Mixed methods research is a rapidly expanding methodology in the social and human sciences in the US and around the world. In this presentation I will first define mixed methods research (combining both quantitative and qualitative methods of research) and discuss what it is and what it is not. Then I will review a brief history of its development, and why it is important today. I will discuss several of the scientific developments in mixed methods that have occurred over the last ten years, such as the specification of types of designs, the formation of mixed methods questions, and the use of innovative approaches to jointly display quantitative and qualitative results. Finally, I will talk about the future of this methodology - where it is headed and some important worldwide developments that have encouraged mixed methods research.

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*

Number of Dissertations and Theses with "Mixed Methods" in the Title

Year Range	Number
2005-2009	2524
2000-2004	532
1995-1999	100
1990-1994	26
1985-1989	17
1980-1984	3

(Haines 2011)

(Creswell 2013)

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

DAY3

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