MOR2 DATA MANAGEMENT & ANALYSIS WORKSHOP

June 13-14 2013, Institute of Geo-Ecology, Ulaanbaatar, Mongolia

Agenda

- Welcome, Introductions, Pretest
- Part 1, June 13 AM: A Tour of the Data
- Part 2, June 13 PM: Introduction to Data Analysis: Hypothesis Testing with Simple Inferential Statistics
- Part 3, June 14 AM: Remote Sensing and GIS Analysis
- Part 4, June 14 AM: Analyzing Qualitative Data

Part 1: A Tour of the Data

Part 1: A Tour of the Data

1. Learning Objectives
2. MOR2 Data Sharing Policy (Maria)
3. Overview of MOR2 Data Types (Maria)
4. Data Entry and Checking (Chantsaa)
5. Introduction to Ecological Data (Khishig)
6. Introduction to Social Data (Maria)
7. Data Exploration, Descriptive Statistics, Graphic Views, Data Manipulation (Maria and Team)

1.1 Learning Objectives

Participants should be able to:

- Describe the different types of data in the MOR2 database and how they relate to the overall research questions and hypotheses in the project
- Understand the importance of accuracy in data collection, entry and management
- Explore simple data sets, calculate descriptive statistics, view data graphically, and manipulate data (e.g. create new calculated variables)

1.2 Data Sharing Guidelines

1. Data are collective property of the team, agreement of senior team members needed to use the data for individual projects.
2. To use the data, please make a proposal that includes:
    - Research question and objectives
    - Research hypotheses
    - Proposed data analysis methods
    - Timeframe for analysis and write-up
    - Expected outputs and products
    - Proposed authorship

1.2 Data Sharing Guidelines

3. The product should be reviewed by team members to ensure appropriate analysis and conclusions.
4. Publications and authorship should be discussed with the team and appropriate team members included as co-authors based on their involvement in the research design, data collection, analysis, interpretation and writing.
5. Ethical and professional standards should be followed in all analyses and writing, including appropriate acknowledgement of all literature and information sources used.

1.2 Data Sharing Guidelines

6. These guidelines apply to publications to peer reviewed and non-technical publications in Mongolian as well as English

1.3 MOR2 Data Types--Overview

	SPATIAL/SOCIAL SCALE OF DATA COLLECTION			
TYPE OF DATA	Individual/ Household/ Ecological Plot	CBRM Organization/ Neighborhood	Soum	Region
Physical and Spatial Data	GPS coordinates Remote sensing	GPS coordinates Remote sensing	Temp & Precip (some soums) Remote sensing	Temp & Precip Hydro data Remote sensing
Ecological Data	Species Cover Species Richness Biomass Soils Dung counts	1 uvuljaa & 3 ecol plots per org.	2-9 uvuljaa	
Social Data	Household Survey	Organizational Profile	Soum Focus Groups Soum Profile	
Livestock Data	Household herd size & composition		Soum herd size & composition	
Herder Observations	Observations of environmental change	Participatory mapping		

1.3 Types of Data

- **Continuous (“scale”) variables** have values that vary continuously, such as the percent cover of a plant species, or income in MNT.
- **Categorical (“nominal”) variables** are discrete categories, such as “Yes” and “No” responses on a survey, or categories of soil texture “Sandy Loam,” “Clay Loam,” etc.
- **Ordinal variables** are discontinuous variables that occur in a meaningful numerical order, for example, the responses to a survey question “1=strongly disagree, 2=somewhat disagree, 3=neutral, 4=somewhat agree, 5=strongly agree”

1.3 Types of Data

- The type of variable is important, because different kinds of analysis are required for different types of data.

1.4 Data Entry and Checking

CHANTSAA

1.5 Introduction to MOR2 Ecological Data

KHISHIG

1.6 Introduction to MOR2 Social Data

1.6.1 Household Survey Database

- 706 households surveyed using face to face, closed end questionnaire
- Data entered into MS Access
- Data exported to SPSS
- SPSS data can also be transferred to MS Excel, but labels associated with codes will be lost

Household Survey—Overview

I. Household demographics
II. Livestock production
III. Movements
IV. Pasture and livestock management practices
V. Land and water tenure
VI. Knowledge exchange and learning
VII. Social relationships
VIII. Household production, income and expenses
IX. Social capital—for formal CBRM members

1.6.2 Organizational Profile

- 140 CBRM and herder neighborhoods
- Data summarized and entered by social teams, based on qualitative data collected with each group:
  - Leader interviews
  - Group member focus groups
  - Non-member interviews

Organizational Profile--Overview

I. Group characteristics
II. Pasture tenure and resource use rules
III. Group activities and outcomes
IV. Conflicts and conflict management
V. Group leadership and social capital
VI. Knowledge exchange and learning
VII. Institutional linkages
VIII. Governance and organizational function

1.6.3 Soum-Level Data

- Existing soum statistics (lots of missing data!)
  - Livestock populations by species over time
  - Human population (people, households, herding households)
  - Employment, poverty
- Soum development questionnaire
  - Infrastructure, services, etc.
- Soum focus group questionnaire
  - Data on soum-level social capital, leadership effectiveness, perceived community assets and problems

1.7 Data Exploration

Data Files

Data Files

Data Checking

- Review Spreadsheets for potential errors
  - Look for unusual numbers (see LPICoverLitter for example). You may need to look in original access database to find source of error.
  - Look for missing values
    - If the value was zero, enter 0
    - If the data are missing, enter 999
  - Conduct random check, comparing original data from a subset of plots or surveys with the data in the spreadsheet

Re-coding

- In some cases, the data produced by the access database are words or letter abbreviations. These cannot be used in our statistical analyses, so they must be recoded using numbers.
- For example, the soil texture data was entered using letter codes, so we recoded them into numbers.

  Soil Texture 1=SL 2=LS 3=L 4=SiCL 5=SC 6=SCL 7=S 8=SIC

Import Data into SPSS

- SPSS is a statistical package we will use to explore and analyze our data
- It will import an Excel file
- Once data are in SPSS, we need to make sure that the file is properly set up.
  - Variable types
  - Variable names and labels
  - Values for categorical (nominal and ordinal) data
  - Missing values

Creating New Variables in SPSS

- Creating new variables through computation
  - EXAMPLE: create an index of winter preparedness in the household survey database by summing all the practices that that relate to winter preparations
  - EXAMPLE: create a new variable “Total Biomass” by summing grass, forb, shrub, sedge and Achnatherum biomass.

Descriptive Statistics

Descriptive Statistics, Exploratory Analysis & Data Transformations

Descriptive Statistics

- **Descriptive statistics** describe the characteristics of the sample units in terms of the central tendency (average) and variation (standard deviation).

Descriptive Statistics

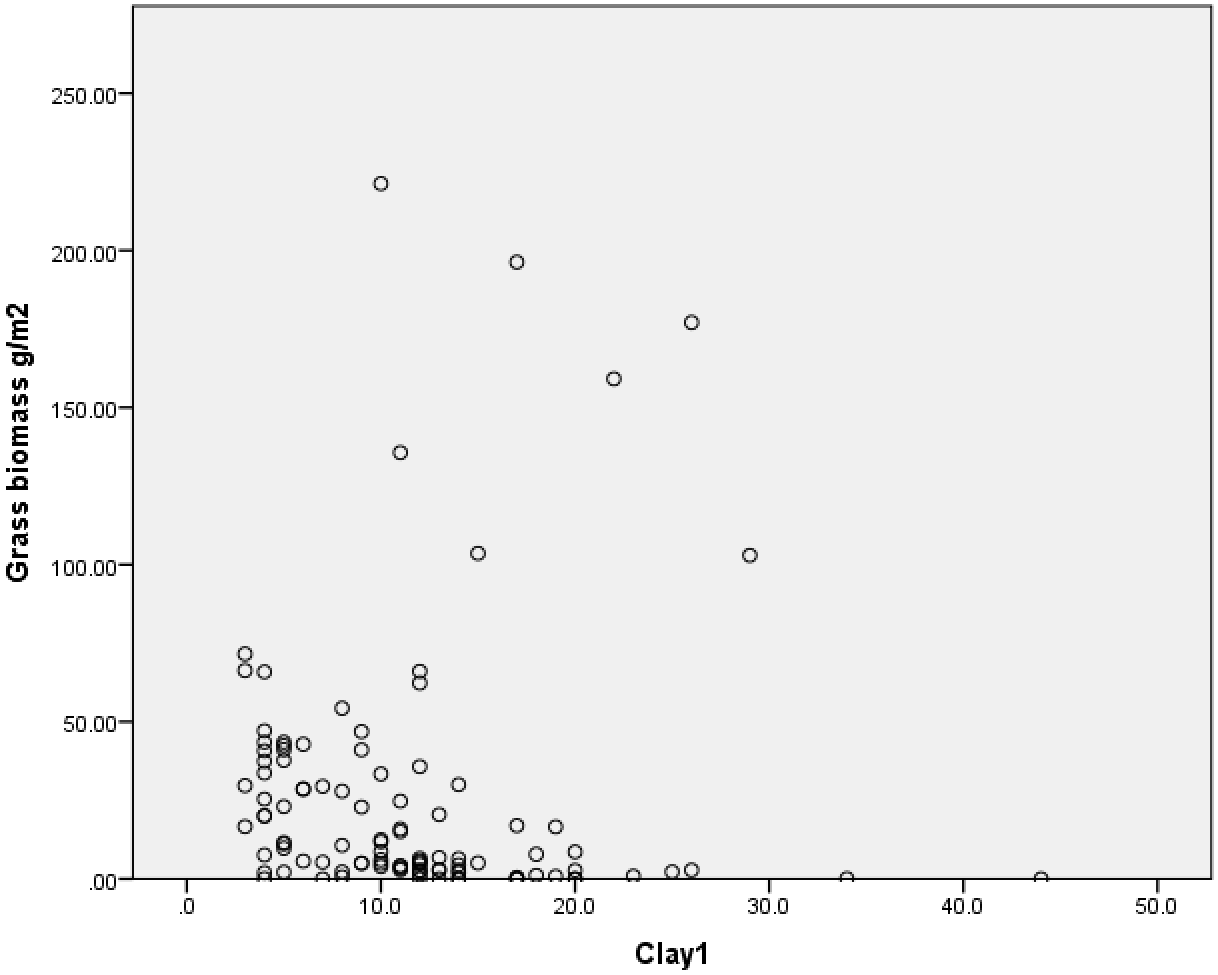
- Mean (average) $\bar{X}$= Sum of observations/number of observations
- Median = the middle observation
- Variance (s^2)= $\sum(X_i-\bar{X})^2/n-1$
- Standard Deviation (s)= $\sqrt{s^2}$
- Standard Error (SE) = $2/\sqrt{n}$
- Coefficient of Variation (CV) = $(s/\bar{X})*100$

Descriptive Statistics in SPSS

- Analyze
  - Descriptives
  - Explore
    - Identifies outliers
    - Tests for normality

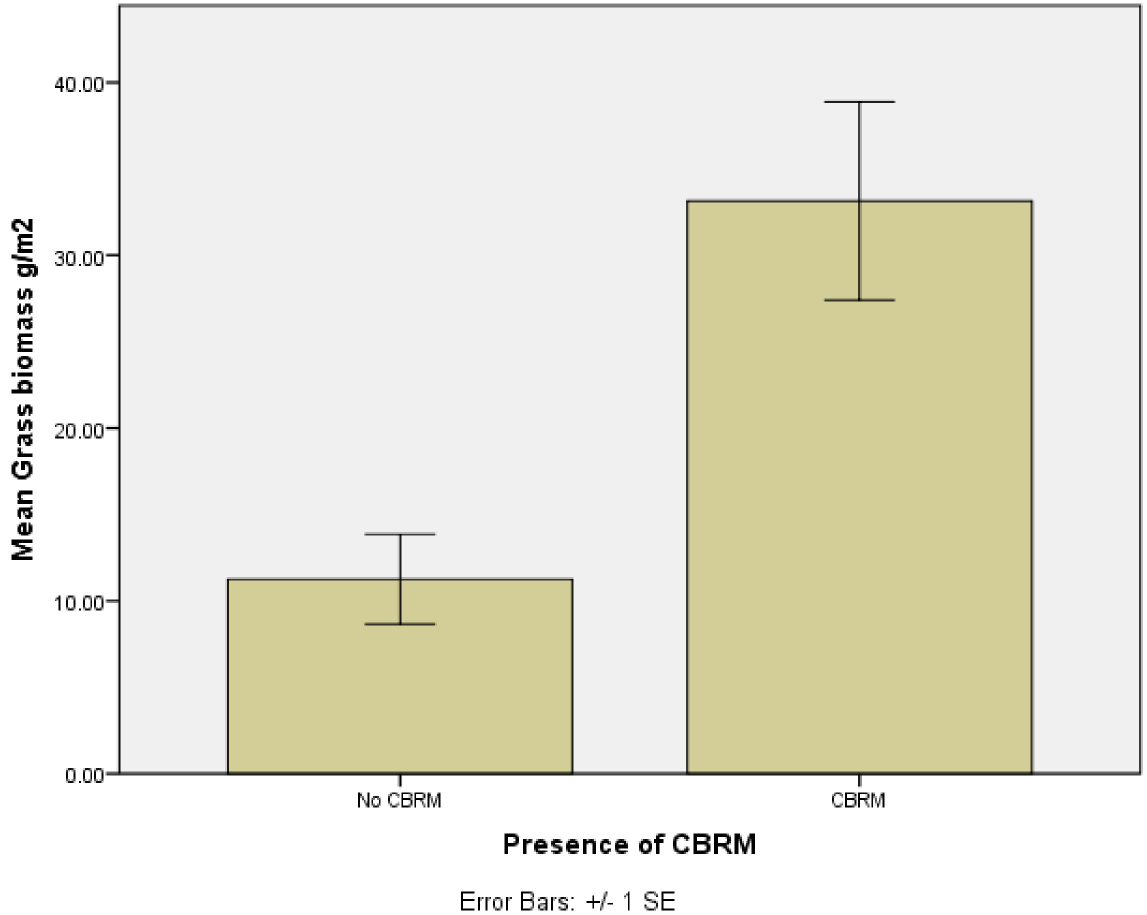
Data Exploration in SPSS

Scatter Plots

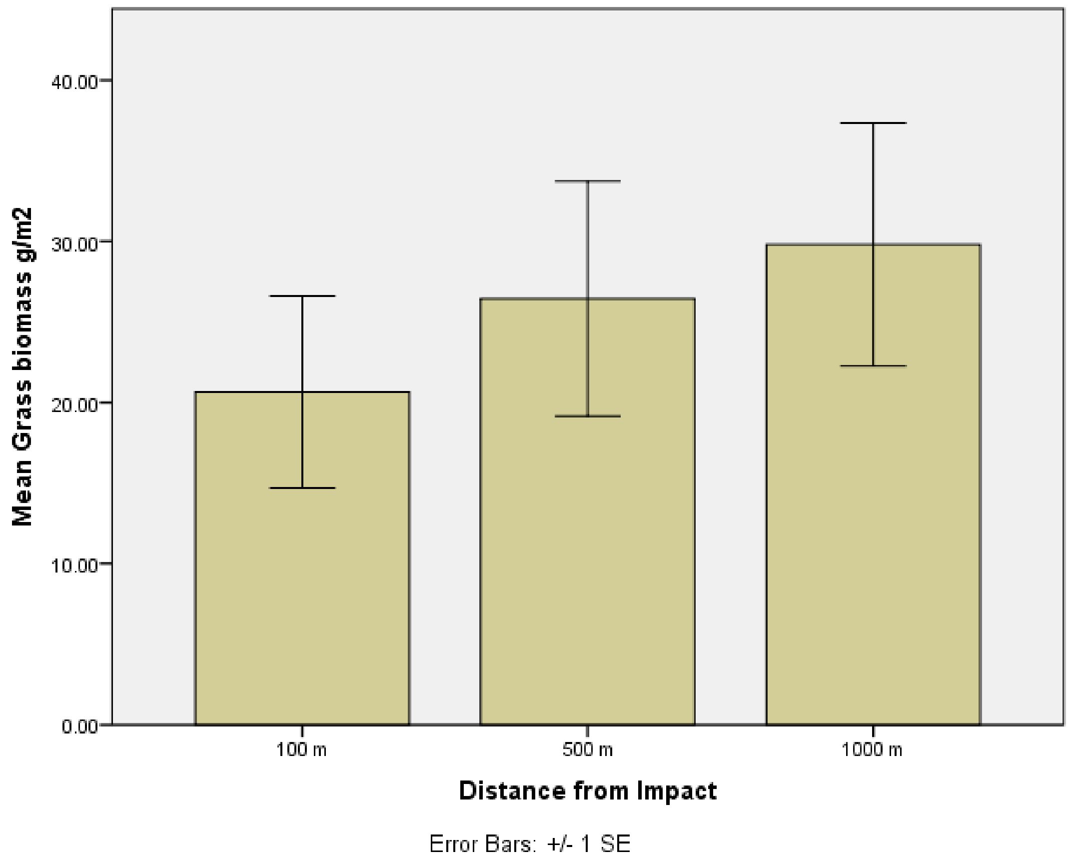


Data Exploration in SPSS

Bar Graphs

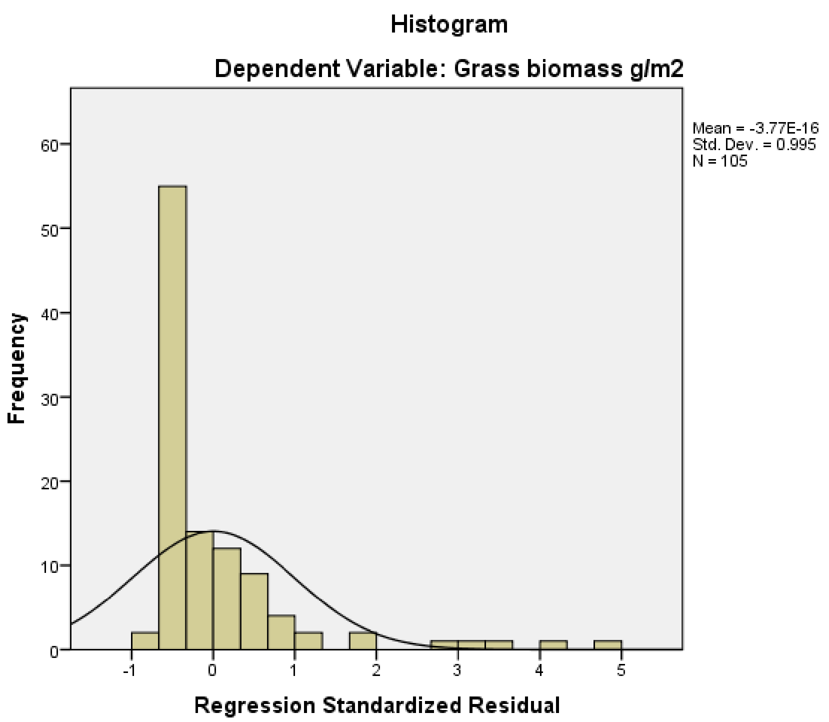


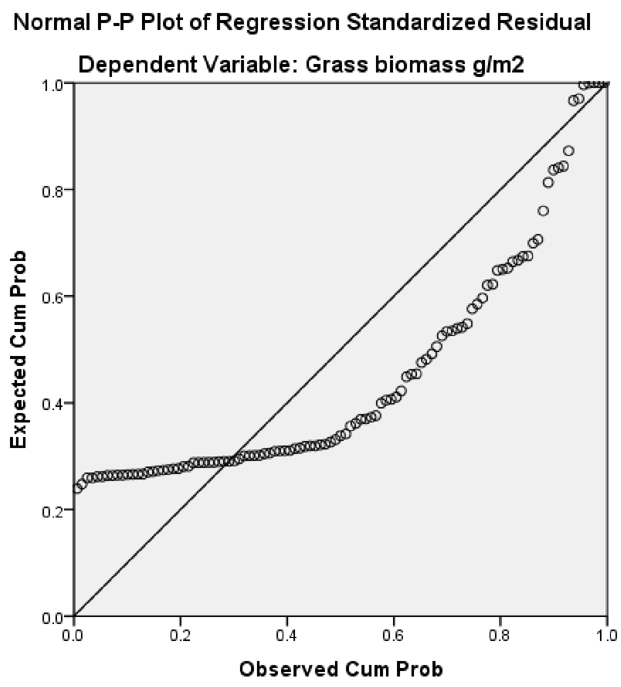
Data Exploration in SPSS



Testing Statistical Assumptions

- Data come from a normal distribution
  - Plot a histogram
  - Normal probability plot





Transforming Data in SPSS

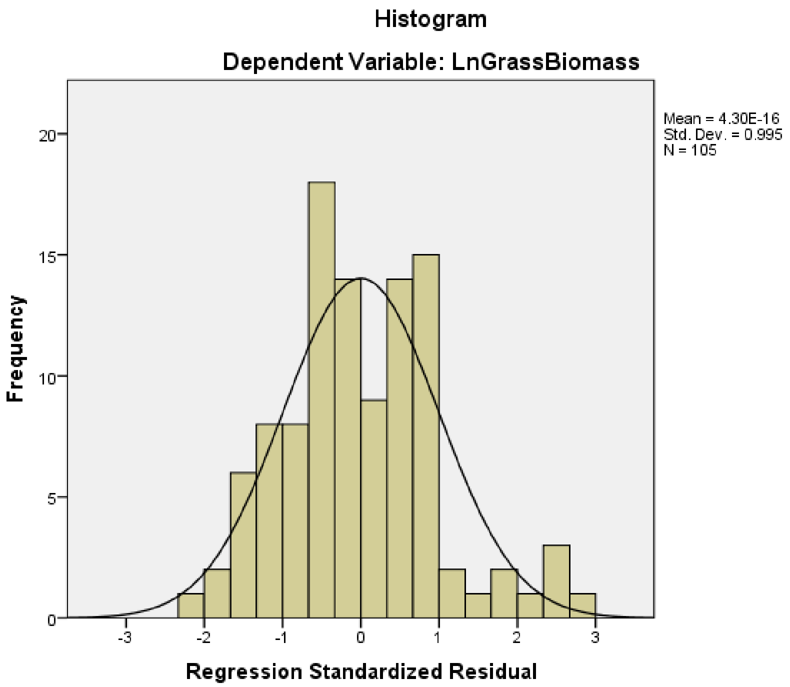
- Transform
  - Compute Variable
    - Target Variable Name [Type LNGrassBiomass]
    - Numeric Expression [Type: LN (Grass_gm2 + 1)
    - Click OK
- A new column of data will appear at the end of your dataset, labelled "LNGrassBiomass"

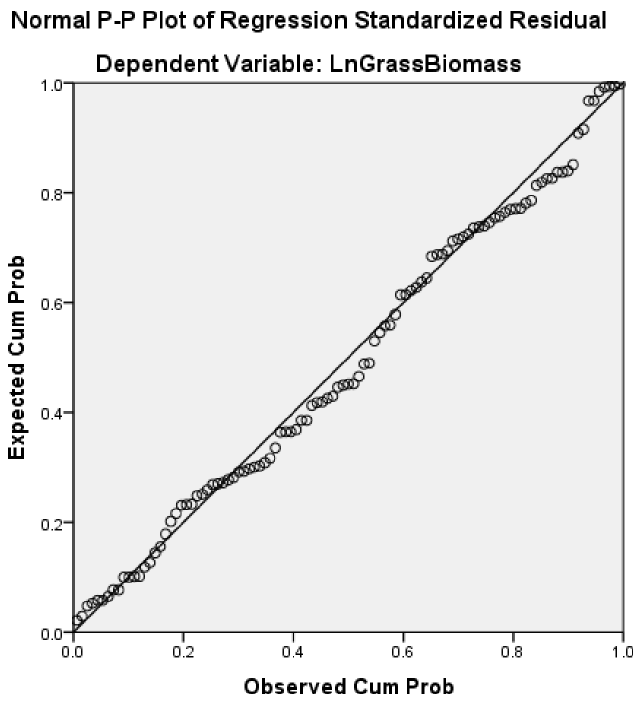
Transforming Data in SPSS

- Inferential statistics assume that
  - the data come from a normal distribution, and
  - the variances from the groups being compared are roughly equal.
- If the data do not meet these assumptions, they need to be transformed.
- Also, if there are a large number of low or zero values, it is helpful to transform the data.
- Useful transformation for cover and biomass data is:
  - Ln (Biomass +1)

Transforming Data in SPSS

- Transformed grass biomass data:





Log transformed grass biomass data are normally distributed.

Testing Statistical Assumptions

- Variances are homogeneous
  - Look at a scatter plot
  - In SPSS: Analyze > General Linear Model > Univariate > Options > Homogeneity tests

Levene's Test of Equality of Error Variances[a]

Dependent Variable: Grass biomass g/m2

F	df1	df2	Sig.
.401	2	102	.671

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + Dist_m

No difference in variances at different distances from Impact point.

You do it!

- Identify some continuous (scale) variables that you might want to analyze
- Calculate descriptive statistics for them
- Explore them graphically
- Determine if they meet assumptions of normality
- Transform them if they do not meet the assumption (create a new variable)

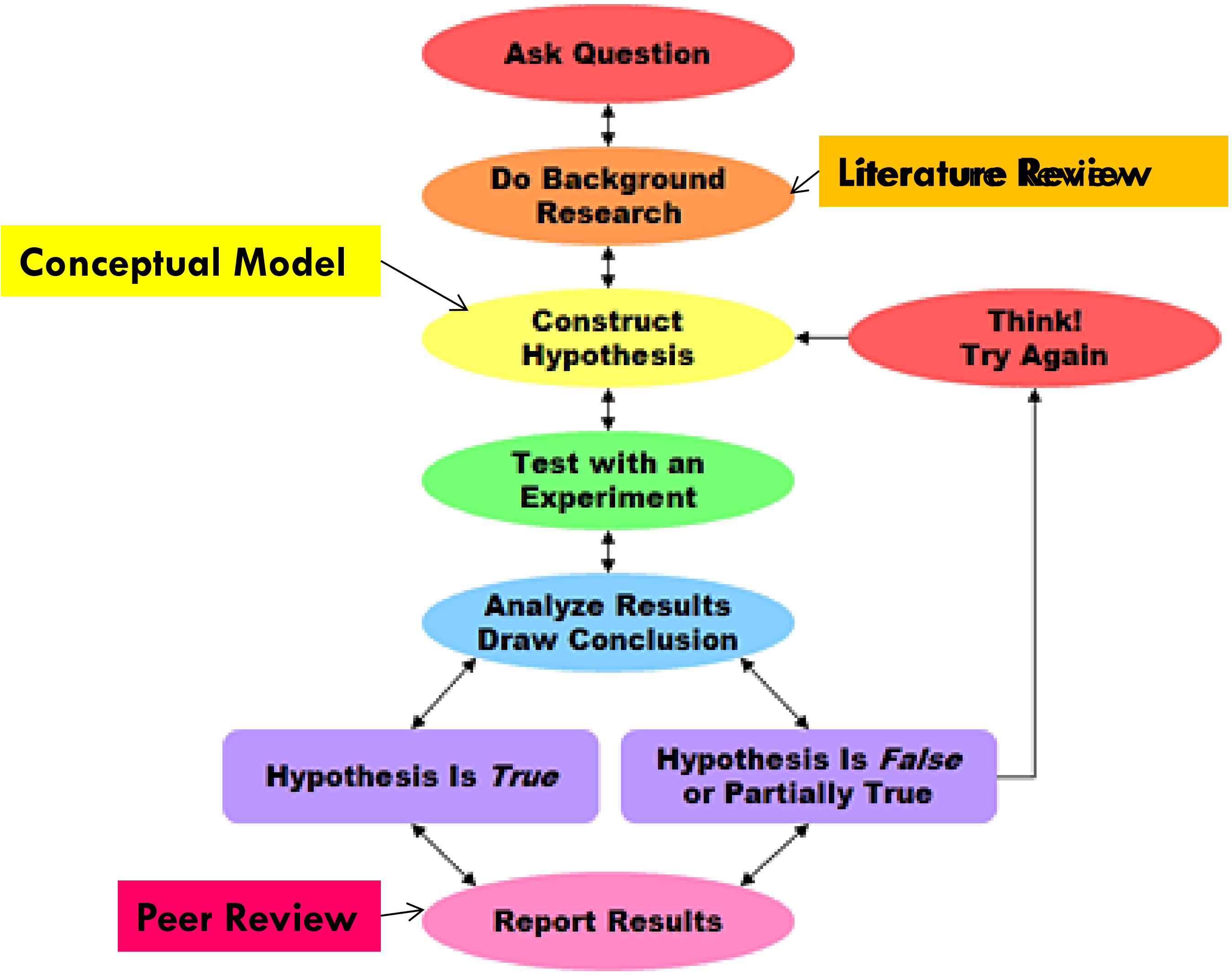
Part 2: Introduction to Data Analysis and Hypothesis Testing

Objectives

Participants should be able to:

1. Describe the process of the scientific method
2. Write a testable hypothesis
3. Identify different data types: nominal, ordinal & scale variables
4. Distinguish inferential from descriptive statistics
5. Learn to test hypotheses using SPSS
   - Regression (continuous x continuous)
   - ANOVA (categorical x continuous)
   - Cross-tabulations (categorical x categorical)

Scientific method



Key terms

- Hypothesis:
  - An educated guess about how things work. A statement about cause and effect.
  - Example: If we put fertilizer on a field, the yield will increase.
- Theory:
  - A confirmed hypothesis.

A good hypothesis

- Is testable and falsifiable
- Sets the research direction
- Identifies the important variables
  - Independent (explanatory)
  - Dependent (outcome)
- Is grounded in theory
- Is short and clear

Hypothesis

- Goal is to **test** a hypothesis in a way that it can be shown to be false.
- The goal IS NOT to "prove" the hypothesis.

Independent & Dependent Variables

- The factors you will measure to test your hypothesis
- A **dependent variable** is the outcome or response variable that is affected in the experiment or study.
- An **independent variable** is the explanatory variable that causes the effect. In an experiment, it is the variable that you change.

Example

- Hypothesis: Total plant cover increases with increasing distance from a winter camp.
- Distance from camp is the independent variable
- Total plant cover is the dependent variable

Which is independent variable? Which is the dependent variable?

- Hypothesis 1: Shrubs prefer sandy soils to soils with high clay content.
- Hypothesis 2: The cover of annual forbs increases closer to winter camps.
- Hypothesis 3: Households that are CBRM members are more likely to reserve winter pastures than non-CBRM households.

Testable and Falsifiable

- You can **disprove** or reject the hypothesis by performing an experiment, or collecting data in an observational study.
- Example:
  - Hypothesis: Adding fertilizer will increase hay yield.
  - We have a uniform hay field divided into treatment plots, and add fertilizer to 10 randomly selected plots, leaving 10 plots untreated.
  - If yield does not increase in the treated plots, our hypothesis is rejected.

Some Broad Hypotheses from our Project

- Winter pastures grazed by CBRM members are more productive, diverse, and healthy than those grazed by non-members.
- Households that belong to CBRM groups are more innovative and have higher social capital.

You do it! Create a hypothesis

- Using your knowledge of the available data sets:
  - Write a testable hypothesis
  - Identify the dependent variable(s)
  - Identify the independent variable(s)
  - What existing theory or knowledge is this hypothesis based on? (Explain WHY you think this hypothesis is true)
- Discuss your draft hypothesis with another person
- Give the other person feedback on their hypothesis
- Report back to the group

Hypothesis Testing with Inferential Statistics

Inferential Statistics

- **Inferential statistics** use data from the sample to draw conclusions about the population from which the sample was drawn. **To determine if two samples come from different populations, we need to use inferential statistics.** **<u>To test our hypotheses, we must use inferential statistics.</u>**

What are inferential statistics for?

- Are the populations different?
- How confident are we that they differ?

Hypothesis Testing

- To determine whether data from two samples are different, we must use inferential statistics.
- Example: To compare the difference between the distance moved by poor herders and wealthy herders.
- We CANNOT draw conclusions about differences or test our hypotheses with descriptive statistics.

% Native Grass Cover

	A. Heavy Grazing	B. Moderate Grazing	C. No Grazing
Plot 1	15	50	10
Plot 2	21	37	19
Plot 3	5	63	7
Plot 4	13	41	38
Plot 5	8	47	51
Mean	12.4	47.6	25
Variance	38.8	99.8	357.5

Null Hypothesis

- Mean cover for A = B = C

How do inferential statistics work?

- It's all about the <u>variance</u>
- Ratio of the variation between groups vs. within groups
- As this ratio becomes larger, the likelihood of a real difference increases

How do inferential statistics work? It's all about the variance

t = (difference between sample means)/

[2 x (average variance of the samples)/(sample size)]

F = between-groups variance/

within-groups variance

How to interpret P-values

When the test statistic (t, F) is large, then the probability of obtaining this result by chance is small.

By convention, when $P < .05$ we reject the null hypothesis that the means are the same.

We conclude that the means come from populations that are significantly different.

t-Test: Two-Sample Assuming Equal Variances

Site A and Site B

	Site A	*Site B*
Mean	12.4	47.6
Variance	38.8	99.8
Observations	5	5
Pooled Variance	69.3	
Hypothesized Mean Difference	0	
df	8	
t Stat	6.685687152	
P(T<=t) one-tail	7.7489E-05	
t Critical one-tail	1.85954832	
P(T<=t) two-tail	0.000154978	
t Critical two-tail	2.306005626	

t-Test: Two-Sample Assuming Equal Variances

Site A and Site B

	Site A	*Site B*
Mean	12.4	47.6
Variance	38.8	99.8
Observations	5	5
Pooled Variance	69.3	
Hypothesized Mean Difference	0	
df	8	
t Stat	6.685687152	
P(T<=t) one-tail	7.7489E-05	
t Critical one-tail	1.85954832	
P(T<=t) two-tail	0.000154978	
t Critical two-tail	2.306005626	

Calculated t-value

P-value

Critical value for t

So, we can …

- We reject the null hypothesis that A = B
- Conclude that cover of native grasses on site B (moderately grazed) is significantly greater than on site A (heavily grazed).

t-Test: Two-Sample Assuming Unequal Variances

Site B and Site C

	Site B	*Site C*
Mean	47.6	25
Variance	99.8	357.5
Observations	5	5
Hypothesized Mean Difference	0	
df	6	
t Stat	2.363158442	
P(T<=t) one-tail	0.028020894	
t Critical one-tail	1.943180905	
P(T<=t) two-tail	0.056041789	
t Critical two-tail	2.446913641	

t-Test: Two-Sample Assuming Unequal Variances

Site A and Site C

	Site A	*Site C*
Mean	12.4	25
Variance	38.8	357.5
Observations	5	5
Hypothesized Mean Difference	0	
df	5	
t Stat	-1.41528372	
P(T<=t) one-tail	0.108070664	
t Critical one-tail	2.015049176	
P(T<=t) two-tail	0.216141329	
t Critical two-tail	2.570577635	

Overall Conclusions

- Moderately grazed (B) > Heavily grazed (A)
- Moderately grazed (B) = No grazing (C)*
- Heavily grazed (A) = No grazing (C)

*If you use Type I error rate of 5%

Types of Error

	NO DIFFERENCE	**REAL DIFFERENCE**
DIFFERENCE DETECTED	False-change error (Type I) α	No error (Power) $1-\beta$
NO DIFFERENCE DETECTED	No Error	Missed-change error (Type II) β

P-value

- The **P-value** or **significance value** of a given statistical test is the probability of making a false-change (or Type I) error. It is the probability that the test determined there was a significant difference between means, but the means really do not differ. We generally want to minimize this chance of a Type I error, and by convention, most scientists use a cut-off of 5% ($p=<0.05$) as the acceptable false-change error rate.

Types of Data

- **Continuous (“scale”) variables** have values that vary continuously, such as the percent cover of a plant species, or income in Tg.
- **Categorical (“nominal”) variables** are discrete categories, such as “Yes” and “No” responses on a survey, or categories of soil texture “Sandy Loam,” “Clay Loam,” etc.
- **Ordinal variables** are discontinuous variables that occur in a meaningful numerical order, for example, the responses to a survey question “1=strongly disagree, 2=somewhat disagree, 3=neutral, 4=somewhat agree, 5=strongly agree”

Types of Analysis

Type of Data	Type of Analysis	Example
2 continuous (scale) variables	Linear regression	Grass biomass x % Clay in soil
1 nominal x 1 continuous	T-test (if 2 groups) ANOVA (more than 2 groups)	CBREM membership x % foliar cover Distance from impact x species richness
2 nominal (or ordinal)	Chi-square Cross-tabulation	CBRM member x Fall Otor

Example: Linear Regression

- Ecological Hypothesis: Grass Biomass increases with increased soil clay content because clay soils hold water close to the soil surface
- Dependent variable = LN Grass Biomass
- Independent variable = % Clay in Horizon 1
- Analyze
  - Regression
    - Linear
      - Statistics > Model fit, R square change, Estimates
- Statistical null hypothesis: No relationship, line flat

Example: Linear Regression

LN Grass Biomass x %Clay in Horizon 1

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.323[a]	.104	.095	1.38253	.104	11.965	1	103	.001

a. Predictors: (Constant), Clay1

b. Dependent Variable: LnGrassBiomass

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	22.869	1	22.869	11.965	.001[b]
	Residual	196.872	103	1.911		
	Total	219.741	104			

a. Dependent Variable: LnGrassBiomass

b. Predictors: (Constant), Clay1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.112	.261		11.917	.000
	Clay1	-.067	.019	-.323	-3.459	.001

a. Dependent Variable: LnGrassBiomass

Can we reject the null hypothesis?

If so, is the relationship the one that we hypothesized?

Example: Linear Regression

LN Grass Biomass x %Clay in Horizon 1

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.323[a]	.104	.095	1.38253	.104	11.965	1	103	.001

a. Predictors: (Constant), Clay1

b. Dependent Variable: LnGrassBiomass

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	22.869	1	22.869	11.965	.001[b]
	Residual	196.872	103	1.911		
	Total	219.741	104			

a. Dependent Variable: LnGrassBiomass

b. Predictors: (Constant), Clay1

Can we reject the null hypothesis? Yes!

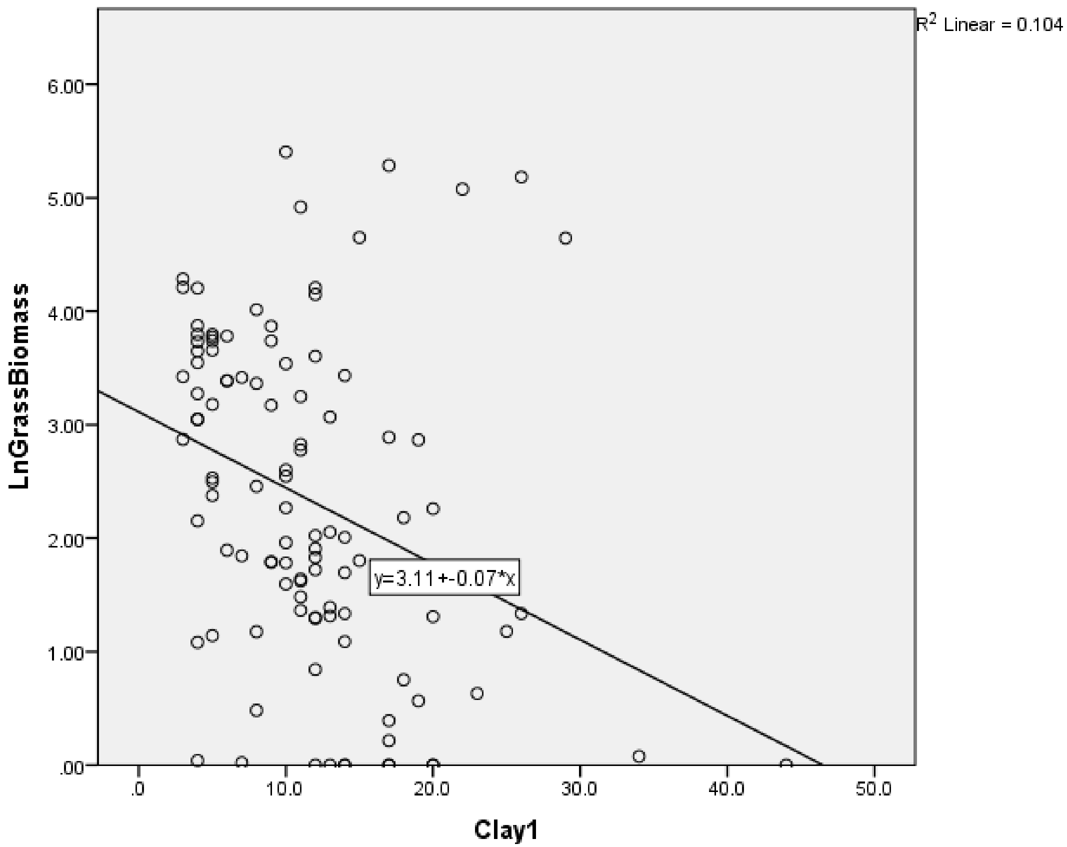
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.112	.261		11.917	.000
	Clay1	-.067	.019	-.323	-3.459	.001

a. Dependent Variable: LnGrassBiomass

Example: Linear Regression

If so, is the relationship the one that we hypothesized?



Example: ANOVA

- Ecological Hypothesis: Grass Biomass increases with increased increasing distance from uvuljaa
- Dependent variable = LN Grass Biomass
- Independent variable = Distance from Impact
- Analyze
  - General Linear Model
    - Univariate
      - Dependent variable = LNGrassBiomass
      - Random factor = Distance from Impact
- Statistical null hypothesis: mean 100-m = mean 500-m = mean 1000-m

Example: ANOVA

Tests of Between-Subjects Effects

Dependent Variable: LnGrassBiomass

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[c]
Intercept	Hypothesis	574.378	1	574.378	312.994	.003	.994	312.994	1.000
	Error	3.670	2	1.835[a]					
Dist_m	Hypothesis	3.670	2	1.835	.866	.424	.017	1.733	.195
	Error	216.071	102	2.118[b]					

a. MS(Dist_m)

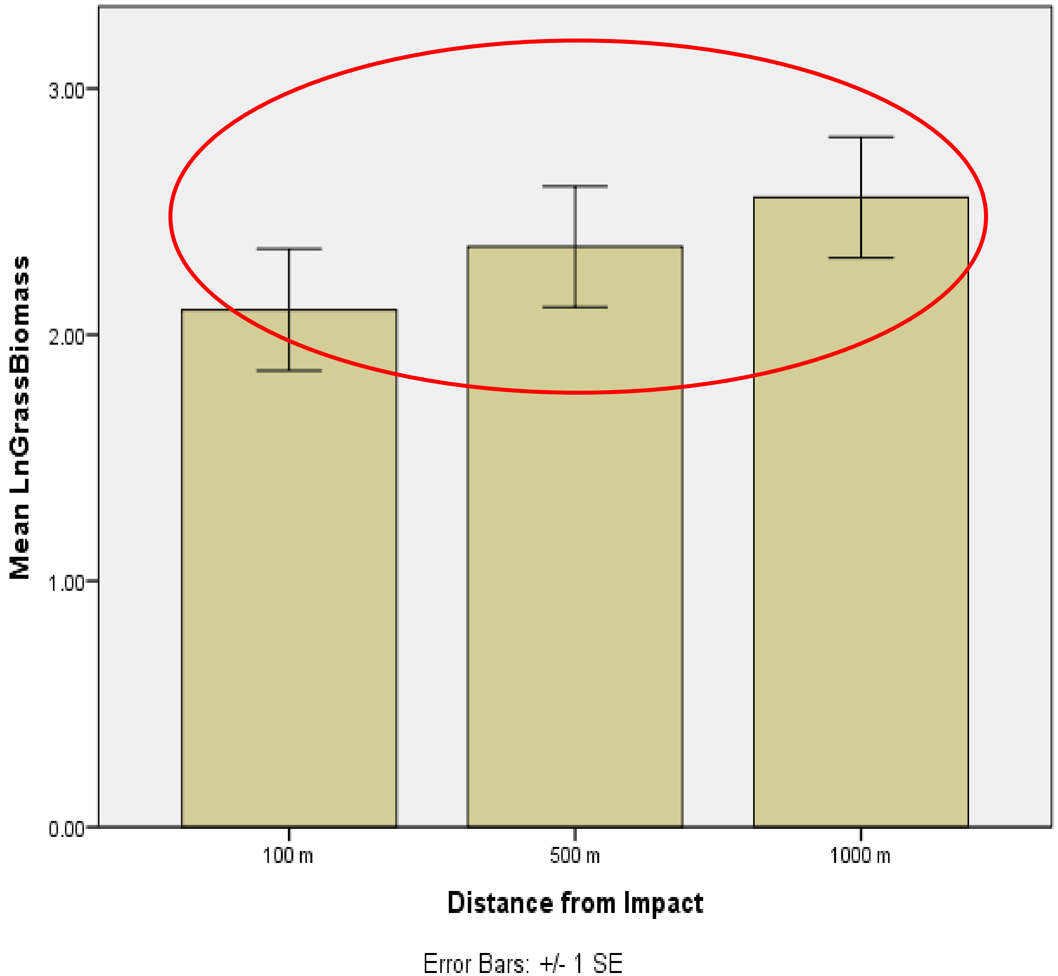
b. MS(Error)

c. Computed using alpha = .05

Do we accept or reject our null hypothesis that the means are the same?

Example: ANOVA

If the SE bars overlap, there is no statistical difference



There is an apparent trend in the data, but no statistically significant difference among the means.

We accept the statistical null hypothesis that the means are the same.

We reject our ecological hypothesis that grass biomass increases with distance from impact.

Example: Chi-square

- Use Household Survey Data Set
- Research Hypothesis: CBRM member households will do more to prepare for winter than non-CBRM households
- Statistical Hypothesis: frequencies of winter preparation actions do not differ between CBRM and non-CBRM households

Example: Chi-square

- Dependent variable(s): Reserve Winter Pasture, Cut Hay, Deworm Livestock
- Independent variable: CBRM organization member
- Analyze
  - Cross-tabs
    - Row = CBRM Org
    - Column = Reserve Winter Pasture
    - Statistics: Chi-square
    - Cells: row, column and total percent

Example: Chi-square

Crosstab

			4.a. Reserve winter pasture		Total
			no	yes	
CBRM org status	CBRM	Count	43	52	95
		% within CBRM org status	45.3%	54.7%	100.0%
		% within 4.a. Reserve winter pasture	53.8%	74.3%	63.3%
		% of Total	28.7%	34.7%	63.3%
	non-CBRM	Count	37	18	55
		% within CBRM org status	67.3%	32.7%	100.0%
		% within 4.a. Reserve winter pasture	46.3%	25.7%	36.7%
		% of Total	24.7%	12.0%	36.7%
Total		Count	80	70	150
		% within CBRM org status	53.3%	46.7%	100.0%
		% within 4.a. Reserve winter pasture	100.0%	100.0%	100.0%
		% of Total	53.3%	46.7%	100.0%

Accept or reject our statistical hypothesis?

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	6.780[a]	1	.009	.011	.007	
Continuity Correction[b]	5.924	1	.015			
Likelihood Ratio	6.888	1	.009	.011	.007	
Fisher's Exact Test				.011	.007	
Linear-by-Linear Association	6.735[c]	1	.009	.011	.007	.005
N of Valid Cases	150					

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 25.67.

b. Computed only for a 2x2 table

c. The standardized statistic is -2.595.

Questions?

You do it! Select Your Test

- Review and revise your scientific hypothesis based on the feedback you received
- Write a statistical "null hypothesis" to test your scientific hypothesis
- Identify whether the variables are continuous (scale), nominal, or ordinal
- Determine which analysis to use to test your hypothesis
- Discuss your choice with another person and review their choice
- Report back to the group

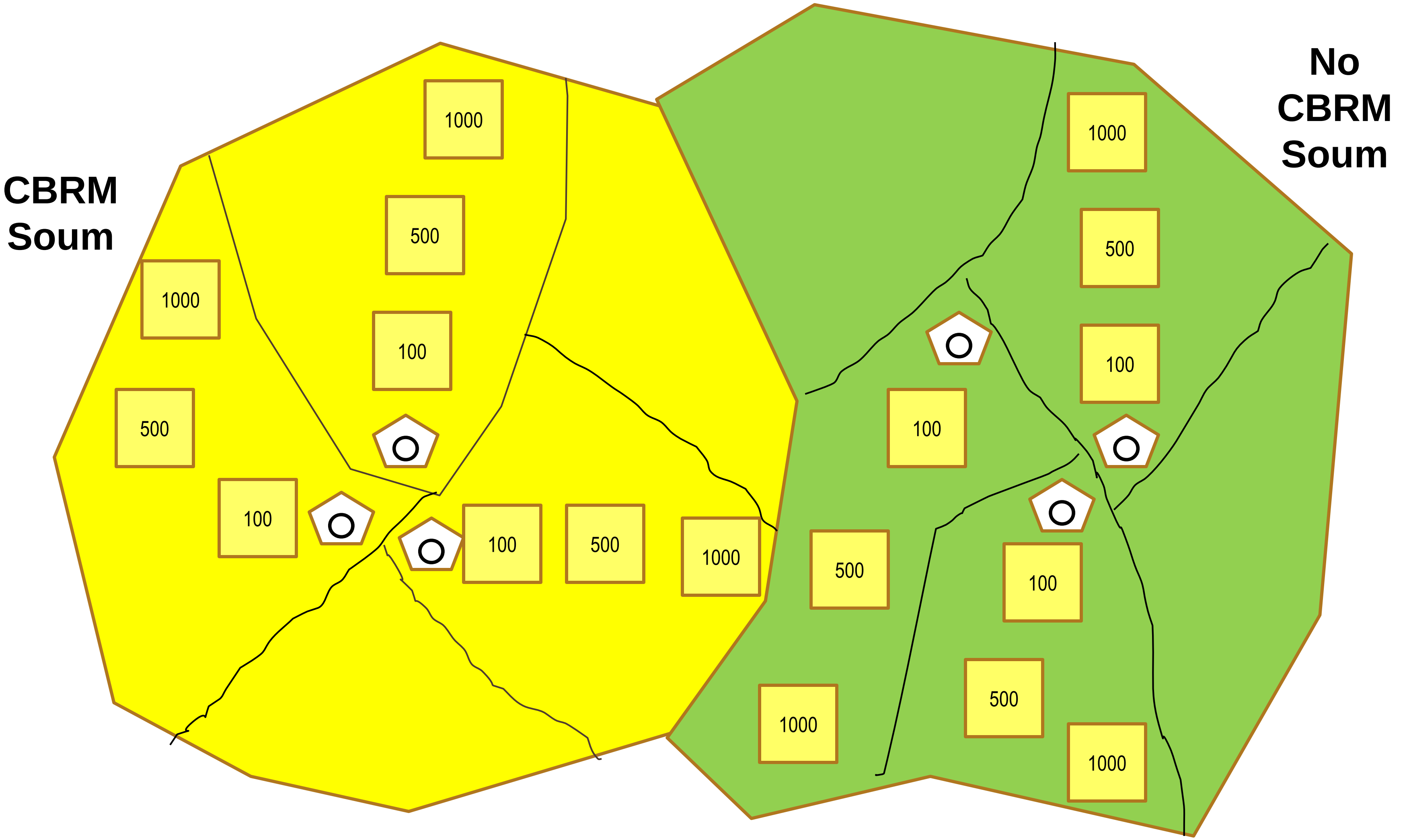
You do it! Test and Interpret

- Use SPSS to test your statistical hypothesis
- Based on the results, decide whether to accept or reject the statistical null hypothesis
- Do the results support or reject your scientific hypothesis?
- How do you explain your results? (What do you think you results mean?)
- Discuss with another person in the group.
- Report back

Thank you!

- Please fill out your evaluation for this workshop!

Our study design



Our Study Design

	With CBRM	Without CBRM	Total
Mountain Steppe 100 m			
Mountain Steppe 500 m			
Mountain Steppe 1000 m			
Steppe 100 m			
Steppe 500 m			
Steppe 1000 m			
Desert Steppe 100 m			
Desert Steppe 500 m			
Desert Steppe 1000 m			
Total			