Research Design and Analysis

FALCON 2015 Conference
November 8, 2015

Darren James
USDA-ARS Jornada Experimental Range
New Mexico State University
Road Map

Part 1: The importance of having a plan

Part 2: Monitoring

Part 3: Experimentation

Part 4: Analysis
Part 1: The importance of having a plan

Part 2: Monitoring

Part 3: Experimentation

Part 4: Analysis
What is statistics?

• The science of learning from data

• The science of collecting, organizing, analyzing, interpreting, and presenting data.

What are data?

• Numerical facts

• Numbers with context
<table>
<thead>
<tr>
<th>76</th>
<th>72</th>
<th>75</th>
<th>73</th>
<th>90</th>
<th>81</th>
</tr>
</thead>
<tbody>
<tr>
<td>76</td>
<td>80</td>
<td>80</td>
<td>77</td>
<td>91</td>
<td>79</td>
</tr>
<tr>
<td>81</td>
<td>69</td>
<td>58</td>
<td>75</td>
<td>92</td>
<td>79</td>
</tr>
<tr>
<td>80</td>
<td>58</td>
<td>52</td>
<td>80</td>
<td>91</td>
<td>82</td>
</tr>
<tr>
<td>92</td>
<td>52</td>
<td>68</td>
<td>88</td>
<td>98</td>
<td>83</td>
</tr>
<tr>
<td>92</td>
<td>58</td>
<td>69</td>
<td>88</td>
<td>95</td>
<td>80</td>
</tr>
<tr>
<td>92</td>
<td>80</td>
<td>82</td>
<td>98</td>
<td>94</td>
<td>82</td>
</tr>
<tr>
<td>87</td>
<td>85</td>
<td>97</td>
<td>100</td>
<td>92</td>
<td>75</td>
</tr>
<tr>
<td>75</td>
<td>88</td>
<td>97</td>
<td>97</td>
<td>91</td>
<td>82</td>
</tr>
<tr>
<td>72</td>
<td>93</td>
<td>98</td>
<td>98</td>
<td>89</td>
<td>85</td>
</tr>
<tr>
<td>85</td>
<td>99</td>
<td>99</td>
<td>91</td>
<td>88</td>
<td>86</td>
</tr>
<tr>
<td>77</td>
<td>98</td>
<td>99</td>
<td>92</td>
<td>78</td>
<td>79</td>
</tr>
<tr>
<td>100</td>
<td>98</td>
<td>98</td>
<td>98</td>
<td>92</td>
<td>74</td>
</tr>
<tr>
<td>104</td>
<td>96</td>
<td>97</td>
<td>88</td>
<td>81</td>
<td>89</td>
</tr>
<tr>
<td>72</td>
<td>85</td>
<td>95</td>
<td>88</td>
<td>79</td>
<td>83</td>
</tr>
<tr>
<td>103</td>
<td>89</td>
<td>100</td>
<td>80</td>
<td>91</td>
<td>80</td>
</tr>
<tr>
<td>98</td>
<td>83</td>
<td>96</td>
<td>75</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>77</td>
<td>86</td>
<td>93</td>
<td>70</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>86</td>
<td>88</td>
<td>80</td>
<td>83</td>
<td></td>
</tr>
</tbody>
</table>

**Numbers without context**
Data: Numbers with context

Metadata: data about data

Data: Temperature

Metadata:

• Observed high temperature, in degrees Fahrenheit
• Temporal Location: August 24, 2014
• Spatial Location: 111 cities in lower 48 states
• Source: The Weather Channel
What data analysis looks like to most people
The larger context of research

Goal

Design → Protocol → Setup → Data Collection

Statistician

Raw Data → Clean Data → Data Analysis

~90% of “analysis” time

~10% of “analysis” time

Results
The larger context of research

Goal

**Magic**

Design → Protocol → Setup → Data Collection

Statistician

Raw Data → Clean Data → Data Analysis → Results
Two main points:

1. **The entire process is driven by the goal**
   - Without a goal, the process has no direction
Two main points:

1. **The entire process is driven by the goal**
   - Without a goal, the process has no direction
   - The goal provides the context

2. **The real “magic” happens with the design**
Bloom’s Taxonomy:
The different levels of thinking we use during the learning process

- **Remembering**: Retrieving, recognizing, and recalling relevant knowledge from long-term memory.
- **Understanding**: Constructing meaning from oral, written, and graphic messages through interpreting, exemplifying, classifying, summarizing, inferring, comparing, and explaining.
- **Applying**: Carrying out or using a procedure through executing, or implementing.
- **Analyzing**: Breaking material into constituent parts, determining how the parts relate to one another and to an overall structure.
- **Evaluating**: Making judgments based on criteria and standards through checking and critiquing.
- **Creating**: Putting elements together to form a coherent or functional whole; reorganizing elements into a new pattern or structure through generating, planning, or producing.
Bloom’s Taxonomy

• The pyramid depicts the different levels of thinking we use when learning.

• Notice how each level builds on the foundation that precedes it.

• We must learn the lower levels before we can use the higher levels effectively.
dollar
dice
tricycle
four-leaf clover
hand
six-pack
Seven-Up
octopus
cat lives
bowling pins
football team
egg
Friday
hour
piano
cleaner
equipment
aardvark
documentation
fund
malfunction
miracle
How many words or phrases do you remember?

Objective: in 10 seconds memorize as many of the following words or phrases as possible:
scientist
accountant
artist
carpenter
builder
taxi driver
police officer
referee
umpire
judge
juror
coach
mentor
facilitator
student
teammate
lawyer
teacher
mathematician
computer scientist
programmer
graphic designer
What were two major differences between the first attempt and the second attempt?

1. We knew what the task was  
2. We knew how the information was organized  

Key idea: when things get confusing, let the goal be your guide.
Ultimate Goals

• I want to design a process that works better.
• I want to make my community stronger.
• I want to improve people’s quality of life.
• I want to make the environment healthier.
• I want to live more sustainably.
• I want to reduce poverty in my community.
• I want to reduce crime in my community.
• I want to solve a particular problem.

Research Goals
These ultimate goals must be turned into research hypotheses.
A study only addresses a small component of the larger goals.
What is the goal of your study?

To monitor.
To measure change over time.
To test a hypothesis.
To demonstrate cause and effect.
To measure the size of an effect.
To show that a management plan has been put in place.
To show that a management plan is working.
Developing research goals

Important distinctions:

1. What goals are essential?
2. What goals would be nice to achieve but are not absolutely essential?

Then:
Are the essential goals actually achievable?
It could be that the monitoring project needed to reach your goal is so expensive that it can’t be carried out.
Case example: NEON

- NEON: National Ecological Observatories Network
- $433 million project
- Goal: collect ecological monitoring data on a massive continental scale
- Collect a continuous stream of data (soil, atmospheric, aquatic, etc.) from dozens of sites within 20 ecological regions.
- Construction began in 2011 and was expected to end in 2016

Many challenges; now **16 months behind schedule and $80 million over budget**

Government report: “It is important to remember that the ecological research community has no experience with a project of this scale.”
The larger context of research

Goal

Design → Protocol → Setup → Data Collection

Statistician

Raw Data → Clean Data → Data Analysis

Results
The importance of having a plan:
The goal will be your guide

Part 2: Monitoring

Part 3: Experimentation

Part 4: Analysis
Part 1: The importance of having a plan

Part 2: Monitoring

Part 3: Experimentation

Part 4: Analysis
Standard rain gauges: 34 locations across the Jornada Experimental Range

- Data: monthly precipitation, in inches
- Now replaced with automatic tipping bucket rain gauges
Examples of monitoring

Phenology: The timing of natural phenomena in relation to plant and animal life

Examples: Date of first green leaves in the spring
Dates of hibernation
Dates of ice in/ ice out
Examples of monitoring results

Month of first leaf in Honey mesquite (*Prosopis glandulosa*) and Black grama (*Bouteloua eriopoda*)
**Examples of monitoring results**

- 22 years of monthly observations
- 16 focal species
- 15 research sites
Key: Consistent goals
Consistent protocols

**PRGL FLeaf (exclude)**

**PRGL FLeaf (include)**
Vegetation quadrats at Jornada Experimental Range

Monitoring began in 1915 and still continues today.
Jornada Experimental Range
Permanent Quadrats

- Quadrat locations
- Quadrat A1
- Quadrat B1
Quadrat A1

- total area of perennial grasses
- total area of mesquite
Soil moisture
Monitoring

• Consistent data collection, analysis, and interpretation
• Identify indicator change over time
• Develop a baseline for making management decisions

How is the resource changing over time?
Assessment

- One point in time
- Compare an indicator or attribute to a threshold or reference condition

What is the condition of the resource compared to a threshold?
Inventory

How much of the resource do I have? Where is the resource located?

• Systematic
• Can include monitoring and assessment data
• Describe the amount, location, and condition of a resource
Attributes

Component that provides information about the functional status of the ecological processes
Attribute: Soil/Site Stability

The capacity of an area to limit redistribution and loss of soil resources (including nutrients and organic matter) by wind and water. (NRC 1994 Rangeland Health Report)

Desert grassland-
good stability

Desert grassland-
loss of stability
Attribute: Hydrologic Function

The capacity of an area to capture, store, and safely release water from rainfall, run-on, and snowmelt (where relevant), to resist a reduction in this capacity and to recover this capacity when a reduction does occur. (NRC 1994 Rangeland Health Report)

Sagebrush “captures” snow

Grasses have reduced ability (structure) to “capture” snow
Monitoring is often used to address a problem.

Case study: eutrophication

Strategy: Determine baseline levels of nutrients.

Compare new samples to the baseline to anticipate state changes and to see if management practices are working.
Monitoring manuals for rangelands:

http://jornada.nmsu.edu/monit-assess/manuals/monitoring

Advance copies of the newest edition:
http://www.landscapetoolbox.org/manuals/monitoring-manual/
Part 1: The importance of having a plan

Part 2: Monitoring

Part 3: Experimentation

Part 4: Analysis
Part 1: The importance of having a plan

Part 2: Monitoring

Part 3: Experimentation

Part 4: Analysis
Experimentation always requires monitoring.

Monitoring does not require experimentation.
In general, there are 2 types of scientific studies:

1. Observational Studies
2. Experimentation

Observational studies are the same as monitoring—they follow a sampling design but do not force the experimental units to follow a certain treatment.

Experimentation involves having a treatment that is forced upon the experimental units.

This is the mechanism by which cause and effect is demonstrated.

In general, observation studies do not prove cause and effect. Experimental studies can prove cause and effect if implemented correctly.
Sampling methods

Convenience sampling: Just ask whoever is around.

- Example: “Person on the street” survey (cheap, convenient, often quite opinionated, or emotional => now very popular with TV “journalism”)

- Which person, and on which street?

- Ask about gun control or legalizing marijuana “on the street” in Berkeley or in some small town in Idaho and you would probably get totally different answers.

- Even within an area, answers would probably differ if you did the survey outside a high school or a country western bar.

- Bias: Opinions limited to individuals present.
**Voluntary Response Sampling:**

- Individuals choose to be involved. These samples are very susceptible to being biased because different people are motivated to respond or not. Often called “public opinion polls,” these are not considered valid or scientific.

- **Bias:** Sample design systematically favors a particular outcome.

---

Magazine article summarizing responses of readers:

70% of (10,000) parents wrote in to say that having kids was not worth it—if they had to do it over again, they wouldn’t.

**Bias:** Most letters to magazines are written by disgruntled people. A random sample showed that 91% of parents WOULD have kids again.
In contrast:

**Probability or random sampling:**

- Individuals are randomly selected. No one group should be over-represented.

Sampling randomly gets rid of bias.

Random samples rely on the absolute objectivity of random numbers. There are tables and books of random digits available for random sampling.

Statistical software can generate random digits (Excel `=random()`).
Simple random samples

A Simple Random Sample (SRS) is made of randomly selected individuals. Each individual in the population has the same probability of being in the sample. All possible samples of size $n$ have the same chance of being drawn.

Note: the pool from which the SRS is drawn is called the **sampling frame**.

The sampling frame is critical to the inference of the study.
Population versus sample

- **Population**: The entire group of individuals in which we are interested but can’t usually assess directly.
  
  Example: All humans, all working-age people in California, all people in your community.

- **Sample**: The part of the population we actually examine and for which we do have data.
  
  How well the sample represents the population depends on the sample design.

- A **parameter** is a number describing a characteristic of the population.

- A **statistic** is a number describing a characteristic of a sample.
POPULATION
Your whole customer base

SAMPLING FRAME
The group of customers you have the ability to contact with your survey

SAMPLE
The group of customers you actually contact with your survey and who actually fill it out
To assess the opinion of college students about campus safety, a reporter interviews 15 students he meets walking on a college campus late at night who are willing to give their opinion.

What is the sample here? What is the population?

- All those students walking on campus late at night
- All students at universities with safety issues
- The 15 students interviewed
- All students approached by the reporter
Toward statistical inference

The techniques of inferential statistics allow us to draw inferences or conclusions about a population in a sample.

- Your estimate of the population is only as good as your sampling design. → Work hard to eliminate biases.
- Your sample is only an estimate—and if you randomly sampled again you would probably get a somewhat different result.
- The bigger the sample the better.
Undercoverage:

Occurs when parts of the population are left out in the process of choosing the sample.

Because the U.S. Census goes “house to house,” homeless people are not represented. Illegal immigrants also avoid being counted. Geographical districts with a lack of coverage tend to be poor. Representatives from wealthy areas typically oppose statistical adjustment of the census.

Historically, clinical trials have avoided including women in their studies because of their periods and the chance of pregnancy. This means that medical treatments were not appropriately tested for women. This problem is slowly being recognized and addressed.
Need for statistically-valid designs

Selecting a *statistically representative* sample from a *population* allows us to *estimate* attributes of the population in an *unbiased* manner with known *level of confidence*.
Sampling variability

Each time we take a random sample from a population, we are likely to get a different set of individuals and a calculate a different statistic. This is called sampling variability.

The good news is that, if we take lots of random samples of the same size from a given population, the variation from sample to sample—the sampling distribution—will follow a predictable pattern. All of statistical inference is based on this knowledge.
Stratified samples
There is a slightly more complex form of random sampling:

A **stratified random sample** is essentially a series of SRSs performed on subgroups of a given population. The subgroups are chosen to contain all the individuals with a certain characteristic. For example:

- Divide the population of college students into males and females.
- Divide the population of California by major ethnic group.
- Divide the counties in America as either urban or rural based on criteria of population density.

The SRS taken within each group in a stratified random sample need not be of the same size. For example:

- A stratified random sample of 100 male and 150 female college students
- A stratified random sample of a total of 100 Californians, representing proportionately the major ethnic groups
**Goal:** Produce area-based estimates of indicators (i.e., proportion of bare soil)

- Sampling should be as **efficient** as possible
  - Maximum amount of information from smallest number of points

**Problem:** The study area is not uniform with respect to bare soil

- Some areas have relatively homogeneous amounts of bare soil
  - A low density of sampling points can capture this variability
- Some areas have very heterogeneous amounts of bare soil
  - A high density of sampling points is needed to capture this variability
Solution: Divide up the study area into similar units (strata)
General approach: Sample the yellow area more intensively than the other areas in order to capture the variability and produce more accurate estimates.
Stratification

Generate Candidate Points (spatially balanced within each stratum)
Sample Design

- Points are uniquely identified with a number
- Numbers are in random order within a stratum

<table>
<thead>
<tr>
<th>Point</th>
<th>Stratum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>21</td>
<td>3</td>
</tr>
<tr>
<td>22</td>
<td>3</td>
</tr>
<tr>
<td>23</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Area</th>
<th>Points</th>
<th>Prop. Area</th>
<th>Prop. Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>6</td>
<td>0.300</td>
<td>0.261</td>
</tr>
<tr>
<td>2</td>
<td>35</td>
<td>8</td>
<td>0.350</td>
<td>0.348</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>6</td>
<td>0.300</td>
<td>0.261</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>3</td>
<td>0.050</td>
<td>0.130</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>12</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Density of points = \[rac{\text{points}}{\text{area}}\]

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Area</th>
<th>Points</th>
<th>Density of points</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>6</td>
<td>0.200</td>
</tr>
<tr>
<td>2</td>
<td>35</td>
<td>8</td>
<td>0.229</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>6</td>
<td>0.200</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>3</td>
<td>0.600</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

- Stratum 4 will be sampled about three times as intensively as the other strata
**Multistage samples** use multiple stages of stratification. They are often used by the government to obtain information about the U.S. population.

Example: Sampling both urban and rural areas, people in different ethnic and income groups within the urban and rural areas, and then within those strata individuals of different ethnicities.

Data are obtained by taking an SRS for each substrata.

Statistical analysis for multistage samples is more complex than for an SRS.
Caution about sampling surveys

• **Nonresponse**: People who feel they have something to hide or who don’t like their privacy being invaded probably won’t answer. Yet they are part of the population.

• **Response bias**: Fancy term for lying when you think you should not tell the truth, or forgetting. This is particularly important when the questions are very personal (e.g., “How much do you drink?”) or related to the past.

• **Wording effects**: Questions worded like “Do you agree that it is awful that...” are prompting you to give a particular response.
The Scientific Method

1. Define the problem
2. Formulate a hypothesis
3. Design the study
4. Make observations / collect data
5. Interpret the data
6. Draw conclusions

Arrive at a new problem
How does statistics fit into the scientific method?

Research issue--not statistical issue

For the statistical analysis, hypotheses are formulated as a mathematical statement

The study design determines the appropriate data analysis

Data collection must follow the design.

Interpreting the data always involves fitting a model of some kind

1. Define the problem
2. Formulate a hypothesis
3. Design the study
4. Make observations / collect data
5. Interpret the data
6. Draw conclusions
However, the practice if science is not a linear process.
The hierarchy of evidence in science

Meta Analysis

Multi-site Replication

Randomized experiments

Quasi-experiments

Single-case experiments

Descriptive Studies (e.g. correlations, pre test/post test without control group)

Other Data (e.g., Anecdotal case reports and observations)
Some common mistakes in scientific experimentation

• Having a preference for one outcome over others

• Mistake the hypothesis for an explanation of a phenomenon without performing experimental tests

• Common sense and logic may tempt us into thinking no experiment is needed

• Ignore or rule out data which do not support the hypothesis.
  You must:
  Consider all the plausible hypotheses.
  Consider all the plausible scenarios.
  Consider all the plausible outcomes.
Indicator ‘status’ versus ‘condition’

*Condition*: how a measured indicator value (or range of values) compares to a benchmark

Condition thresholds can change with physiographic conditions.
Science is a process that works by trial and error.

Scientific progress cannot be made without **falsification**.

General strategy: set up investigation/study as best as possible with respect to goals.

Learn from failures and mistakes.

There are no easy answers in scientific research.

Scientific studies serve two purposes:
1. To test hypotheses
2. To demonstrate results

You research study is both a test and a demonstration.
It is possible to conduct a study the “right” way and still end up with the wrong answer.

*What causes this?*

- The multiple areal unit problem: scaling
- Changed objectives
- Some data handled differently
- Failure to estimate quantitatively systematic errors (and all errors)
- Inaccurate *key terms*
Is your design a plausible reflection of the real world?
In designing a scientific study:

• Details matter
• Intention matters
• Integrity matters
• Clarity matters
• Planning (forethought) matters
• Assumptions matter
• Continuity and Consistency matter
• Keep good records
• Get feedback

*The worst scientific failures happen when:*

Knowledge is lost
Knowledge is locked up
Researchers do not learn from their mistakes

Perfection is not required
Deception is not allowed
Steps to success in scientific research

1. Be clear
2. Trust the process
3. Do your homework
4. Communicate and participate
5. Pay attention
6. Be honest
7. Be open to the possibility that you are mistaken
8. Handle all data the same way
9. Actively experiment
10. Have open communication
Part 1: The importance of having a plan

Part 2: Monitoring

Part 3: Experimentation

Part 4: Analysis

Key idea: The design dictates the analysis
Data structure: how the data is organized in a computer file

- tabular: summary table
- vertical / horizontal
- univariate / multivariate
- separators
- what does a line represent?
- what does a column represent?
- file type

Best practices:
- Use the simplest possible data structure.
- For data entry, organize the data by the observational unit. This maximizes the information contained in the data.
Metadata

Metadata is data about data. Minimally, metadata should include:

- **Who** collected the data? Who owns the data?
- **What** does the data file consist of (variable names, units, etc.)
- **When** were the data collected?
- **How** were the data collected (methods)?
- **Why** were the data collected (research questions)?
Metadata

Metadata is crucial to the data analysis process

- Are the data appropriate for the research question?
  - What is the inference space?
- Can the data be combined with others?
  - If so, how?
- Are some observations more reliable than others?
Data entry and error checking

- Data types can be mixed (numeric and character)
- Variables can have the same name
- Missing data can mean different things
- Links can break
- Spaces in names and data can create errors
- Many other possible problems
Standard rain gauges: 34 locations across the Jornada Experimental Range
- Data: monthly precipitation, in inches
- Now replaced with automatic tipping bucket rain gauges
Combining data sets by Stacking

Notice that we have to add a Year variable.

For this to work correctly, the variable names in the two data sets must match.
Combining data sets by Merging

Rain_gauge Season Precip

Rain_gauge Elevation

merge BY Rain_gauge

Precip elevation

The Rain_gauge names must match in both files.

- Using consistent names is essential for maintaining long term data
What about missing observations in long term data records?
What about gaps in data files?

• Some statistical procedures do not work if any observations are missing

• Gap-filling
  • Temporal imputation
  • Spatial imputation
  • Time + Space imputation
Managing Long-term data

- Metadata is critical
- Organization and consistency are critical
- Monitor the data as they are coming in
- Keep data in the finest resolution possible
- Document changes
- Be flexible and realistic
### Simple Linear Regression

The graph shows the demand for widgets with the relationship between quantity and price. The equation of the line of best fit is:

\[ y = -0.1664x + 13.675 \]

The coefficient of determination, \( R^2 \), is 0.6309, indicating that 63.09% of the variability in the data is explained by the linear relationship between quantity and price.
<table>
<thead>
<tr>
<th>Quantity</th>
<th>Category</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>Low</td>
<td>$12.66</td>
</tr>
<tr>
<td>24</td>
<td>Low</td>
<td>$7.61</td>
</tr>
<tr>
<td>27</td>
<td>Low</td>
<td>$11.08</td>
</tr>
<tr>
<td>30</td>
<td>Low</td>
<td>$7.84</td>
</tr>
<tr>
<td>24</td>
<td>Low</td>
<td>$8.65</td>
</tr>
<tr>
<td>34</td>
<td>Low</td>
<td>$6.32</td>
</tr>
<tr>
<td>39</td>
<td>Low</td>
<td>$7.04</td>
</tr>
<tr>
<td>41</td>
<td>Low</td>
<td>$8.45</td>
</tr>
<tr>
<td>20</td>
<td>Low</td>
<td>$9.89</td>
</tr>
<tr>
<td>39</td>
<td>Low</td>
<td>$4.47</td>
</tr>
<tr>
<td>45</td>
<td>High</td>
<td>$9.38</td>
</tr>
<tr>
<td>45</td>
<td>High</td>
<td>$4.91</td>
</tr>
<tr>
<td>46</td>
<td>High</td>
<td>$6.02</td>
</tr>
<tr>
<td>49</td>
<td>High</td>
<td>$7.12</td>
</tr>
<tr>
<td>49</td>
<td>High</td>
<td>$2.77</td>
</tr>
<tr>
<td>52</td>
<td>High</td>
<td>$5.18</td>
</tr>
<tr>
<td>35</td>
<td>High</td>
<td>$10.25</td>
</tr>
<tr>
<td>57</td>
<td>High</td>
<td>$5.68</td>
</tr>
<tr>
<td>58</td>
<td>High</td>
<td>$3.59</td>
</tr>
<tr>
<td>61</td>
<td>High</td>
<td>$1.69</td>
</tr>
<tr>
<td>66</td>
<td>High</td>
<td>$3.45</td>
</tr>
</tbody>
</table>

*Mean price by quantity category (with standard error bars)*

**P = 0.015**

*T-Test (ANOVA)*
The 3 assumptions of Analysis of Variance (ANOVA) are:

1. Independent samples
2. Normal residuals
3. Equal variance between groups

The most important assumption is independent samples.

This can be tricky when the units of observation are different from the experimental unit.
When implementing Long term research, maximize the experimental units

There are 2 replications of treatment at each site

But there is only 1 replication of treatment in the study
If I want to test for differences between sites then I must have more than 1 replication of each level of site.

There are 2 replications of treatment at each site.

There are 2 replications of site in the study.
Any sampling within an experimental unit does not gain more information about treatment.
Any sampling within an experimental unit does not gain more information about treatment.
How does the modeling process work?

Example: Vegetation Cover Data

- 3 groups
- Each observation is a proportion between 0 and 1
- Find out if the groups are different
Density curves come in any imaginable shape.

Some are well known mathematically and others aren’t.
Median and mean of a density curve

The **median** of a density curve is the equal-areas point: the point that divides the area under the curve in half.

The **mean** of a density curve is the balance point, at which the curve would balance if it were made of solid material.

The median and mean are the same for a symmetric density curve. The mean of a skewed curve is pulled in the direction of the long tail.
Simulated Vegetation Cover data as proportions on [0,1]
3 Hypothetical “treatments”

Overall density

Density by group

Is there 1 large group? Or multiple (up to 3) distinct groups?
Use variance to decide
The **median** of a density curve is the equal-areas point: the point that divides the area under the curve in half.

The **mean** of a density curve is the balance point, at which the curve would balance if it were made of solid material.

The median and mean are the same for a symmetric density curve. The mean of a skewed curve is pulled in the direction of the long tail.
Comparing the mean and the median

The mean and the median are the same only if the distribution is symmetrical. The median is a measure of center that is resistant to skew and outliers. The mean is not.
The mean, on the other hand, is only slightly pulled to the right by the outliers (from 3.4 to 3.6).

The mean is pulled to the right a lot by the outliers (from 3.4 to 4.2).

The median, on the other hand, is only slightly pulled to the right by the outliers (from 3.4 to 3.6).
Measure of spread: the **quartiles**

The **first quartile**, $Q_1$, is the value in the sample that has 25% of the data at or below it (\(\Leftrightarrow\) it is the median of the lower half of the sorted data, excluding $M$).

The **third quartile**, $Q_3$, is the value in the sample that has 75% of the data at or below it (\(\Leftrightarrow\) it is the median of the upper half of the sorted data, excluding $M$).

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>1.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>2.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>2.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>2.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>2.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>2.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>2.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>3.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>3.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>3.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>3.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>3.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>4</td>
<td>3.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>5</td>
<td>4.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>6</td>
<td>4.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>7</td>
<td>4.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>4.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>4.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>3</td>
<td>5.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>4</td>
<td>5.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>5</td>
<td>6.1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$M = \text{median} = 3.4$

$Q_1 = \text{first quartile} = 2.2$

$Q_3 = \text{third quartile} = 4.35$
### Five-number summary and boxplot

<table>
<thead>
<tr>
<th>Years until death</th>
<th>Disease X</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td>2</td>
<td>1.2</td>
</tr>
<tr>
<td>3</td>
<td>1.6</td>
</tr>
<tr>
<td>4</td>
<td>1.9</td>
</tr>
<tr>
<td>5</td>
<td>1.5</td>
</tr>
<tr>
<td>6</td>
<td>2.1</td>
</tr>
<tr>
<td>7</td>
<td>2.3</td>
</tr>
<tr>
<td>8</td>
<td>2.3</td>
</tr>
<tr>
<td>9</td>
<td>2.5</td>
</tr>
<tr>
<td>10</td>
<td>2.8</td>
</tr>
<tr>
<td>11</td>
<td>2.9</td>
</tr>
<tr>
<td>12</td>
<td>3.3</td>
</tr>
<tr>
<td>13</td>
<td>3.4</td>
</tr>
<tr>
<td>14</td>
<td>3.6</td>
</tr>
<tr>
<td>15</td>
<td>3.7</td>
</tr>
<tr>
<td>16</td>
<td>3.8</td>
</tr>
<tr>
<td>17</td>
<td>3.9</td>
</tr>
<tr>
<td>18</td>
<td>4.1</td>
</tr>
<tr>
<td>19</td>
<td>4.2</td>
</tr>
<tr>
<td>20</td>
<td>4.5</td>
</tr>
<tr>
<td>21</td>
<td>4.7</td>
</tr>
<tr>
<td>22</td>
<td>5.3</td>
</tr>
<tr>
<td>23</td>
<td>5.6</td>
</tr>
<tr>
<td>24</td>
<td>5.6</td>
</tr>
<tr>
<td>25</td>
<td>6.1</td>
</tr>
</tbody>
</table>

Largest = max = 6.1

Smallest = min = 0.6

\[ M = \text{median} = 3.4 \]

\[ Q_1 = \text{first quartile} = 2.2 \]

\[ Q_3 = \text{third quartile} = 4.35 \]
\[ Q_3 = 4.35 \]
\[ Q_1 = 2.2 \]

Individual #25 has a value of 7.9 years, which is 3.55 years above the third quartile. This is more than 3.225 years, \(1.5 \times \text{IQR}\). Thus, individual #25 is a suspected outlier.

Interquartile range: \[ Q_3 - Q_1 = 4.35 - 2.2 = 2.15 \]

Distance to \(Q_3\): \[ 7.9 - 4.35 = 3.55 \]
Example study: You randomly assign subjects from your target population into two groups: control and experiment. Then you apply the treatment and measure the response. The results are below.

Have you proved that the treatment caused a change?

No—you did not measure the baseline (response variable before the treatment was applied)
- Winter grazing resulted in biggest decline
- Nongrazed plots showed dramatic increases, producing bimodality in 2002
- Shrubs exascerbated pulse, but did not affect response
- Varying levels of saturation and recovery rates lead to loss of bimodality by 2009
Data analysis

<table>
<thead>
<tr>
<th>Plot</th>
<th>Block</th>
<th>Shrub treatment</th>
<th>Grazing treatment</th>
<th>Percent composition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1996</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>intact</td>
<td>none</td>
<td>10.03</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>intact</td>
<td>summer</td>
<td>10.73</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>intact</td>
<td>winter</td>
<td>6.08</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>removed</td>
<td>none</td>
<td>11.33</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>removed</td>
<td>summer</td>
<td>12.43</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>removed</td>
<td>winter</td>
<td>12.38</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>intact</td>
<td>none</td>
<td>12.74</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>intact</td>
<td>summer</td>
<td>10.77</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>intact</td>
<td>winter</td>
<td>11.21</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>removed</td>
<td>none</td>
<td>14.18</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>removed</td>
<td>summer</td>
<td>16.77</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>removed</td>
<td>winter</td>
<td>10.44</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>intact</td>
<td>none</td>
<td>10.63</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>intact</td>
<td>summer</td>
<td>12.88</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td>intact</td>
<td>winter</td>
<td>7.46</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>removed</td>
<td>none</td>
<td>13.91</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>removed</td>
<td>summer</td>
<td>12.47</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>removed</td>
<td>winter</td>
<td>11.75</td>
</tr>
</tbody>
</table>
One strategy for repeated measures data: Gain scores

### Percent composition of Bouteloua eriopoda

<table>
<thead>
<tr>
<th>Plot</th>
<th>Block</th>
<th>Shrub treatment</th>
<th>Grazing treatment</th>
<th>Percent composition</th>
<th>Gain Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>intact</td>
<td>none</td>
<td>10.03</td>
<td>25.7</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>intact</td>
<td>summer</td>
<td>10.73</td>
<td>6.97</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>intact</td>
<td>winter</td>
<td>6.08</td>
<td>1.55</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>removed</td>
<td>none</td>
<td>11.33</td>
<td>27.32</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>removed</td>
<td>summer</td>
<td>12.43</td>
<td>7.75</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>removed</td>
<td>winter</td>
<td>12.38</td>
<td>4.38</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>intact</td>
<td>none</td>
<td>12.74</td>
<td>23.07</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>intact</td>
<td>summer</td>
<td>10.77</td>
<td>7.28</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>intact</td>
<td>winter</td>
<td>11.21</td>
<td>3.06</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>removed</td>
<td>none</td>
<td>14.18</td>
<td>29.11</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>removed</td>
<td>summer</td>
<td>16.77</td>
<td>6.91</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>removed</td>
<td>winter</td>
<td>10.44</td>
<td>2.71</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>intact</td>
<td>none</td>
<td>10.63</td>
<td>24.06</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>intact</td>
<td>summer</td>
<td>12.88</td>
<td>5.29</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td>intact</td>
<td>winter</td>
<td>7.46</td>
<td>4.57</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>removed</td>
<td>none</td>
<td>13.91</td>
<td>34.57</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>removed</td>
<td>summer</td>
<td>12.47</td>
<td>10.87</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>removed</td>
<td>winter</td>
<td>11.75</td>
<td>8.58</td>
</tr>
</tbody>
</table>

The sign of the gain score (+ or -) reflects the direction of the change.

Calculate the gain score at the level of the experimental unit.

To detect change, test whether the gain scores are different from 0.
Rainfall events can result in large increases in cover

- Select a covariance structure that can model these phenomena

2006 = < 0.5% grass cover

2009 = 5.8 - 33.8% grass cover
Road Map

Part 1: The importance of having a plan

Part 2: Monitoring

Part 3: Experimentation

Part 4: Analysis