Design of Experiments and Data Handling

(WPS 415 lecture notes)

Fikret Isik

fisik@ncsu.edu

Department of Forestry and Environmental Resources, NCSU

1019 Biltmore Hall
Lecture Outline

1. Statistical Components of a Research
2. Some Definitions in the Experiment and Survey Data
3. Design of Experiments (design tools, sample size)
4. Data Management Using a Software
5. Data cleaning (outliers, typo errors, odd values)
6. Descriptive Statistics
7. Data Summary (plots, charts)
8. Goodness of fit?
1. Statistical components of a research

* A plan to *conduct* the experiment/s to collect the right type of data, and enough of it. Talk to a statistician before designing an experiment

* *Prepare specific questions* and construct hypothesis. Planning an experiment begins with carefully considering what the goals are.

* Prioritize objectives. That helps you decide which direction to go with regard to the selection of the factors, responses and the particular design.

* Have a timetable and follow it, write activities in detail (research diary).
Example 1. (From a WPS student project)

The paper industry is constantly looking for ways to operate more efficiently. One way to increase the efficiency of papermaking is using additives. The effects of starch, PAE, and latex on paper strength and elasticity were tested in an experiment. The possible benefits could be greater paper strength values, reduced additives and fiber usage.

Possible questions:
1. Which additive is the most efficient to increase the paper strength and elasticity?
2. If we add different levels of an additive, how does the fiber usage change? Does it decrease?
3. At which stage of the paper making is the additive more efficient?
Consider the following points:

1. What are the factors? How many levels are needed?
2. Is the measuring process simple?
3. Does the study produce reliable data?
4. Is the cost reasonable?
5. Can the study be completed in timely manner?
6. Are there extraneous variable effecting measurements?
7. How to summarize data? Appropriate statistical methods to answer questions?
8. What kind of inferences can be made from the results? We use results from statistical analysis based on the sample to make inference about the population.
2. Some Definitions in the Experiment and Survey Data

* **Factor:** A factor of an experiment is a controlled independent variable whose levels are set by the experimenter. The effect of starch on paper strength experiment-Starch is a factor.

* **Treatment:** A treatment is a level (amount) of factor applied to the experimental units. For example, an experiment is divided into four units; each part is 'treated' with 10mg, 20mg, 30mg and 40mg of the same starch N. Each amount of N is a treatment.

* **Variable:** All the factors, their levels and all the measured traits (responses) are variables

* **Independent or class variables:** Factors applied to the experimental units are independent variables

* **Response or dependent variables:** Measured or observed traits in the experiment. The strength of the paper measured after ‘treated’ with starch is a response (or dependent) variable.
Main effects: This is the simple effect of a factor on a dependent variable. It is the effect of the factor alone averaged across the levels of other factors. Example: In the starch experiment, it was found that on average, the starch increased the strength (main effect) compared to the control (no starch).

Interaction effect: An interaction is the variation among the differences between means for different levels of one factor over different levels of another factor. Example: Let's say starch and latex both significantly increased the paper strength (main effects). However, when starch and latex applied to the same experimental unit, it was found that the increase in paper strength was even higher. The paper got the benefits of both factors, plus a bonus.

Experimental (or Sampling) Unit: A unit is a person, animal, plant or thing, which is studied by a researcher; the basic objects upon which the study or experiment is carried out. For example, a person; a sample of soil; a pot of seedlings are units.
* **Population:** A population is any entire collection of people, animals, plants from which we may collect data. It is the entire group we are interested in, which we wish to describe or draw conclusions about. *Example,* we are interested in average GPA values of all CNR students. All the students registered in CNR is a population.

* **Sample:** A sample is a group of units selected from a larger group (the population). By studying the sample it is hoped to draw valid conclusions about the larger group. A sample is generally selected for study because the population is too large to study in its entirety.

* **Randomization:** The sample should be representative of the general population. This is often best achieved by random sampling. For each population there are many possible samples.

* **Parameter:** A parameter is a value, usually unknown (and which therefore has to be estimated). It is used to represent a certain population characteristic. *For example,* the population mean is a parameter that is often used to indicate the average value of a quantity. Parameters are often assigned Greek letters (e.g. sigma).
* **Statistic:** A statistic is a quantity that is calculated from a sample of data. *For example,* the mean, a variance and a standard deviation calculated from a sample are statistics. Statistics are assigned Roman letters (e.g. \( s \) for standard deviation). The value of a statistic will vary from sample to sample.

* **Sampling Distribution:** The sampling distribution is the probability distribution or probability density function of the statistic. Derivation of the sampling distribution is the first step in calculating a confidence interval or carrying out a hypothesis test for a parameter. Example, suppose that \( x_1, x_2, \ldots, x_n \) are a simple random sample from a normally distributed population with expected value \( \mu \) and known variance \( \sigma^2 \). Then the sample mean \( \bar{x} \) is normally distributed with expected value \( \mu \) and variance \( \sigma^2/n \)

* **Estimate:** An estimate is an indication of the value of an unknown quantity based on observed data. Example, suppose we want to know the average height of CNR students (population). We can use an estimate of this population mean by calculating the mean of a sample
of students. Estimators of population parameters are sometimes
distinguished from the true value by using the symbol 'hat'. For
example, $\sigma = \text{true population standard deviation}$. $\overline{\sigma} = \text{estimated (from a sample) population standard deviation}$.

Exercise: What are the factors, treatments and independent
variables in the additive experiment?

The Source of the above definitions is (modified)
http://www.stats.gla.ac.uk/steps/glossary/anova.html#maineff
3. Design of Experiments

Tools for developing experimental designs

Randomization

Treatments should be allocated to the experimental units randomly. In the additive experiment, we would like to treat 10 samples per additive. The additives should be assigned to 10 paper samples randomly. Why is it important?

1. First, randomization makes sure that the sample is a representative of the population (whole paper produced in the mill).

2. Secondly, randomization eliminates the effect of systematic biases. Lets say the humidity in the mill increases during the day. If we do not randomly assign additives to the sample papers during the day, the humidity may have a confounding effect on the paper strength.
Replication
This is the number of experimental units (paper samples) measured for each treatment. Increasing the number of replications means collecting more information about the treatments. In the additives example, we tested 10 paper samples for each additive. Thus the number of replications per treatment is 10.

Replication provides the variability in the response variable that is not associated with the treatment effects. Increasing the number of replications increases the reliability of the outcome.

Let $\bar{x}$ is the mean, $S$ is the variance and $n$ is the number of observations of the response variable. The standard error of the strength mean ($std$) is

$$Std \left( \bar{x} \right) = \frac{S}{\sqrt{n}}$$

The standard error of the mean decreases as $n$ (number of observations) increases.
**Blocking**

Blocking is a technique used to eliminate the effect of a confounding factor. Blocks are group of experimental units sharing a common level of a confounding variable.

For example, in the additives experiment, we assume that humidity is an external factor and affecting the paper strength. If we treat all 10 samples with starch in the morning, and treat another 10 random samples with latex in the afternoon, we cannot eliminate the affect of humidity. The effect of treatments will be confounded.

There are different blocking techniques. The most commonly blocking is the Randomized Complete Blocks. It is random, because the experimental units are randomly assigned to each block. It is complete, because each treatment is included in every block. Assignment of experimental units to the blocks and to the treatments must be random.
Sample size

The decision between the sample size and the cost is a compromise. More measurements increase the reliability of the results but also increase the cost. In general, studying whole population is prohibitively expensive. We take a sample from the population to make inference.

How many observations do I need (sample size)?
The answer varies depending on the accuracy sought. Let us say the true population standard deviation of paper strength is $\sigma = 0.12$. If we are satisfied with a $d=0.05$ margin of error, then for a 95% confidence level we need to have at least

$$n = \left( \frac{t_{\alpha/2, n-1} \cdot \sigma}{d} \right)^2$$

where $\alpha/2$ is the error rate (typically .025)

$$= (1.96 \cdot 0.12/0.05)^2 = 22$$

as sample size, where 1.96 is the Z value for 95% level. For the same margin of error, what is the number of samples required for a population standard deviation 0.17? If $d=0.10$ is an acceptable margin of error, what is the sample size needed for $\sigma = 0.17$?
**Variance:** Distribution of individual observations around the means

\[ s^2 = \sum \frac{(x_i - \bar{x})^2}{n-1} \]

**Standard deviation:** Square root of variance. Distribution of individual observations around the mean

\[ s = \sqrt{\sum \frac{(x_i - \bar{x})^2}{n-1}} \]

**Standard error of the mean:** Distribution of sample means around the population mean

\[ se = \frac{s}{\sqrt{n}} \]

**Confidence interval:** How confident are we about the sample mean \( \bar{x} \) representing the population mean \( \mu \)? For 95% CI,

\( (\bar{x} - 1.96*se \leq \mu \leq \bar{x} + 1.96*se) \)
4. Data Management Using a Software

Statistical analysis software packages are essential tools to analyze data. With some practice, it will be simple to use a package. There are many packages in the market, such as S-Plus, Minitab, SPSS, and SAS, JMP etc. SAS is one of the most powerful and commonly used packages for data summary and analysis. The software is freely available to students in all unity labs.

A brief intro to the SAS for research data summary: SAS windows, data entry, import wizards, file management.
5. Data cleaning (outliers, typo errors, odd values)

Example 30.1: (Source: http://v8doc.sas.com/sashtml/)

The following example investigates how snapdragons grow in various soil types. To eliminate the effect of local fertility variations, the experiment is run in blocks, with each soil type sampled in each block.

<table>
<thead>
<tr>
<th>Type</th>
<th>Block</th>
<th>StemLength</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinton</td>
<td>1</td>
<td>32.1</td>
</tr>
<tr>
<td>Clinton</td>
<td>2</td>
<td>29.7</td>
</tr>
<tr>
<td>Clinton</td>
<td>3</td>
<td>29.1</td>
</tr>
<tr>
<td>Knox</td>
<td>1</td>
<td>35.7</td>
</tr>
<tr>
<td>Knox</td>
<td>2</td>
<td>35.9</td>
</tr>
<tr>
<td>Knox</td>
<td>3</td>
<td>33.1</td>
</tr>
<tr>
<td>Compost</td>
<td>1</td>
<td>31.8</td>
</tr>
<tr>
<td>Compost</td>
<td>2</td>
<td>28.0</td>
</tr>
<tr>
<td>Compost</td>
<td>3</td>
<td>29.2</td>
</tr>
<tr>
<td>Webster</td>
<td>1</td>
<td>32.5</td>
</tr>
<tr>
<td>Webster</td>
<td>2</td>
<td>31.1</td>
</tr>
<tr>
<td>Webster</td>
<td>3</td>
<td>29.7</td>
</tr>
</tbody>
</table>
/* Typing data in the SAS Editor Window */

data plants;
    input Type $ Block StemLength ;
datalines;
Clinton 1 32.1
Knox 1 37.7
Compost 1 33.8
Webster 1 32.5
Clinton 2 29.7
Knox 2 35.9
Compost 2 32.0
Webster 2 31.1
Clinton 3 29.1
Knox 3 36.1
Compost 3 29.2
Webster 3 29.7
;
proc print ;
run;
/* Checking errors in the data */

proc univariate data=plants plot;
   var StemLength;
run;

* By treatments;
proc univariate data=plants plot;
   by block ;
   var stemLength;
run;
6. Descriptive Statistics

Mean, Variance, Standard deviation, Standard error, Coefficient of variation. Explanatory data analysis will give you an overall picture. You will get to know data. When analyzing data, do not skip this step.

/* Descriptive statistics */

proc means mean std stderr ;
class type ;
   var StemLength;
run;
7. Data Summary (plots, charts)

/* Using SAS-GRAPH to Summarize Data */

title 'VBAR charts with Means and 95% CL';
proc gchart;
   vbar type /
      TYPE=mean
      SUMVAR=StemLength
      coutline=black width=8
      errorbar=bars;
run;

8. ANOVA and goodness of fit example

ANOVA

/* ANOVA and comparisons of Means */

               proc glm order=data;
               class Block Type;
               model StemLength = Block Type ;
               means Type / waller regwq;
               run;
Goodness of fit

Are hair color and eye color have a significant relationship?

<table>
<thead>
<tr>
<th>Eyes (Eye Color)</th>
<th>Hair (Hair Color)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>black</td>
<td>dark</td>
</tr>
<tr>
<td>blue</td>
<td>6</td>
<td>51</td>
</tr>
<tr>
<td>brown</td>
<td>16</td>
<td>94</td>
</tr>
<tr>
<td>green</td>
<td>0</td>
<td>37</td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>182</td>
</tr>
</tbody>
</table>
/* Goodness of fit */

data color;
    input Region Eyes $ Hair $ Count @@;
    label eyes='Eye Color'
        hair='Hair Color'
            region='Geographic Region';
datalines;
1 blue  fair    23 1 blue  red      7 1 blue  medium 24
1 blue  dark    11 1 green fair    19 1 green red  7
1 green medium 18 1 green dark    14 1 brown fair 34
1 brown red    5 1 brown medium 41 1 brown dark 40
1 brown black  3 2 blue  fair   46 2 blue  red 21
2 blue medium 44 2 blue  dark   40 2 blue  black 6
2 green fair 50 2 green red    31 2 green medium 37
2 green dark 23 2 brown fair 56 2 brown red 42
2 brown medium 53 2 brown dark 54 2 brown black 13
;
proc freq data=color;
    weight count;
    tables eyes*hair/ chisq nocol norow ;
run;
Homework

A treatment is applied to two groups in the following data. The objectives are 1) to determine the effect of treatment on the response, 2) to investigate differences between two groups.

1. What are the independent, and response variables in the data?
2. Write specific questions according to the objectives of the study.
3. Copy the following lines and paste into a SAS Editor.

```
data fulldata;
   input group $ treat response @@;
   cards;
A 1 7.6  A 1 8.3  A 1 7.6  
A 2 8.5  A 2 8.7  A 2 7.7  A 2 8.3  A 2 8.7  
A 3 6.8  A 3 6.7  A 3 6.6  A 3 6.4  
A 4 7.4  A 4 6.5  A 4 6.8  
B 1 15.5  B 1 13.8  B 1 14.2  B 1 17.3  
B 3 20.5  B 3 17.7  B 3 191  B 3 21.1  B 3 16.9  B 3 18.7  
B 4 16.4  B 4 13.8  B 4 17.4  B 4 18.8  B 4 19.1  
B 5 16.1  B 5 14.4  B 5 13.0  
;```

a. Using the UNIVARIATE procedure, check for errors in the data.
b. Produce bar charts for Groups and for treatments separately (Use Proc GCHART).
c. Calculate descriptive statistics using the MEANS procedure
d. Are there differences between groups for response variable? (Use t-test in SAS)
e. Using ANOV, test the null hypothesis of H0=Group(A)=Group(B)

Put the outputs in a Word Document and send the homework to Dr. Lucia.