

Data Collection and Statistical Analysis for Agricultural research



**Presented by:
Ejaz Ashraf, Ph.D.**



Quote of the day

- **The wealth of Nations lies not in precious metals, stones, and luxurious lifestyles, but in the capacities of its people to invent, to produce, and to organize.**

Adam Smith

Another quote of the day

Education and innovations are the currencies of the twenty first century

Barak Hussain Obama" addressing the Egyptian Parliament on June 4, 2009

One more quote of the day

- **Mind is beautiful and obedient servant**
- **However, body is dangerous master**
- **Take your mind in which direction you wish. That's up to you**

What is Statistics?

- **Originated from the word status (Latin) or statistia (Italian)**
- **It is a social science**
- **Art of dealing with facts and figures**

Purpose & goal of today's session

- **To understand tricky statistical techniques for applications in Agri. research**
- **Agriculture has wide applications of statistical techniques**
- **To accomplish ultimate goal of growing crops to get sustainable yield**

Topics

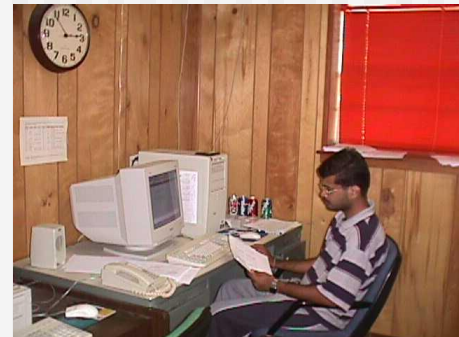
- **Data collection sources**
- **Determination of desired sample size and sampling techniques**
- **Testing of hypotheses**
- **Statistical significance**
- **Design of experiments**

Data Collection Sources



From Field

**From Office
record**



Two Sources of Data Collection

- **Primary Source**
- **Secondary Source**

Primary Source

- **Collected by the researcher or his representative from the field**
- **Methods for collecting data from primary source**
 - 1. Direct Personal Observation**
 - 2. Estimates Through Correspondents**
 - 3. Investigation Through Schedules/Questionnaires**

Direct Personal Observations



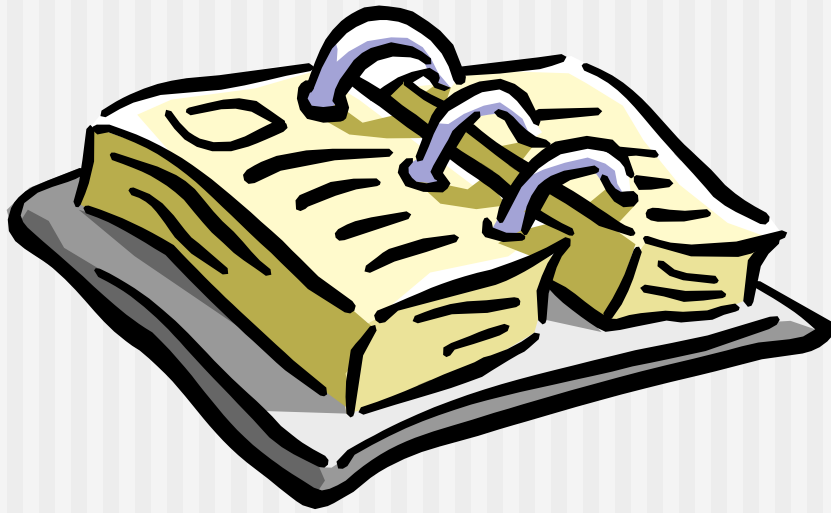
- **Researcher selects the area**
- **Collects data Personally or by his representative from the area**
- **This method is suitable for small area**

Estimates Through Correspondent



- **Agents are appointed in the area**
- **Agents collect information and transmit it to the researcher**

Investigation Through Schedules/Questionnaires



- **List of questions for specific subject**
- **Two methods**
 - a) Sent by post**
 - b) Sent through enumerators**

Sent by Post



- Questionnaires usually sent to respondent
- Respondent has to fill it out and send back to the researcher

Sent Through Enumerators



- **Hired professional enumerators**
- **Train them on specific issue**

Secondary Sources

- **Two methods**

- a) Published**

- b) Unpublished**

Published Source

- **Research papers, journals, survey reports**
- **Monthly or annual reports of national or international agencies**

Unpublished Source

- **Records of government departments and other private organizations**

Sampling Techniques

- **Before Sample, make sure**
 - 1. Sampling method**
 - 2. Sample Size**
 - 3. Reliability of the estimates in terms of probability**

EXAMPLE: Sample Size Calculation

Yamane's formula:

$$*n = \frac{N}{1 + N(e)^2}$$

Where

n = Sample size

N = Population size

e = Level of precision or Sampling of Error

which is ±5%

*Reference: Yamane, Taro. 1967. Statistics, An Introductory Analysis, 2nd Ed.

New York: Harper and Row.

$$n = \frac{N}{1 + Ne^2}$$

Where n= desired sample size

N= total population (73,547)

E= acceptable error limit; by using the formula above

$$n = \frac{N}{1 + Ne^2}$$

N= 73,547

$$E= 0.05 \text{ or } \underline{0.0025}, \quad n = \frac{73547}{1+73547(0.05)^2}; \quad n = \frac{73547}{1+353.59} = \frac{73547}{354.59}$$

$$n = 207.41 \approx \underline{207}, \quad n = 207.$$

Selection of sample size for unknown population

- Kasley and Kumar 1989 used this formula for reaching the workable sample size for unknown population
- $n = Z^2 V^2 / d^2$

Selection of sample size for unknown population

■ Where,

Z=Reliability coefficient (Constant) follows normal distribution= 1.96

n= Sample size

V= assumed variation that could be 50% or may be more or less than 50% depends on the researcher's decision

d= assumed marginal error 5% or may be 1% or 10% again depends on the researcher's how large sample needs to be consider

$$n = \frac{(1.96)^2 (50)^2}{(5)^2} = 384$$

During Sampling

- **Following errors may occur**
 - 1. Sampling Error; one population unit has higher probability**
 - 2. Non-response Error; no response from subject or measurement recording**

Main Divisions of Sampling

- **Probability Sampling**
- **Non-Probability Sampling**

Probability Sampling

- 1. Random Sampling (with and without replacement)**
- 2. Stratified Random Sampling**
- 3. Cluster Sampling**
- 4. Systematic Sampling**

Non-Probability

- **Purposive Sampling is the main techniques in which the researcher's personal judgment plays an important role**
- **Convenient sampling, Snow-ball sampling, Quota sampling and etc.**

Example of Sample

- **Suppose we select a sample of 28 out of 560 total using systematic sampling**
- **We are using only 5% of the total in a sample such as n/N (percentage of a population representing in a sample) called sampling factor**
- **Sampling interval $K = N/n$ it means $560/28 = 20$ In sample, every 20th population unit will be included in the sample**

Simple Random Sampling

- **SRS with replacement**
 - **The easiest way is to make a draw to select a required sample (in our example 28 out of 560)**
 - **Elements of the population have equal probability**

Cont'd

- **If a sampling unit selected more than once, it means we are using SRSWR**
- **If a sampling unit selected only once, it means we are using SRSWOR**

Stratified Random Sampling

- Population is heterogeneous
- Total population N divided into k sub-population for homogeneous groups $N_1 + N_2 + \dots, N_k = N$
- For sample of size n , we have $n_1 + n_2 + \dots + n_k = n$

Cont'd

- **Stratified RS give us more precise estimates than SRS**
- **Give more precise info inside the sub-popn. about the variables under study**
- **However, sometimes it may be difficult to divide the popn. Into sub-popn.**

Cluster Sampling

- **In cluster Sampling;**
 - **Population is divided into units or groups**
 - **They should represent the quarters or groups of the popn.**
 - **Must be homogenous among each other**

Systematic Sampling

- Numbered the population from 1 to N
- Decide the interval by $K=N/n$
- Randomly start within this interval
e.g. if $K=5$ so we need to pick up starting point within 1-5 (usually lower than the interval) and add the same interval in the subsequent population units to get the required sample

Cont'd

- **Advantages**

- Extends the sample to all the population.
- Easy to apply without any extra effort

- **Disadvantages**

- Increase in variance

Testing of Hypotheses



- **Procedure to test an assumption**
- **Based on information obtained from sample**

Experimental Design & Technical Terms used (Agri)



- **Experiment:** Planned & systematic inquiry
- **Treatment:** Any procedure or thing whose effects on the experiment material is to be measured.

Technical Terms

- **Experimental Unit:** Piece of experimental material
- **Observation:** a measurement that is made for some output(s) of interest (yield, plant stand, nutrient status, disease, insect infestation, etc.).
- **Block:** Homogeneous experimental unit
- **Yield:** Measurable quantity

Technical Terms

- **Replication: Repeated treatment in a study**
- **Experimental Design: A set of rules for assigning treatments**

Statistical significance

- Often mentioned but seldom explained
- When experiments are conducted
- Experimental errors are computed
- These errors are used to assess whether or not treatments are different significantly from one another

Statistical significance

- Statistics are based on probability theory
- Researchers have to decide what level of probability constitutes significance for them
- In scientific studies, probability level often referred to as “p” value
- Selected solely at the discretion of the researcher

Statistical significance

- Scientific community in general prefers a probability level of 90% or 95%
- Means that the difference between the two treatments did not occur by chance
- If the 95% probability level in terms of p-value met then treatments are significantly different

Statistical significance

- Here some gray area enters into research
- What is appropriate probability level?
- Each researcher has his or her own set of criteria

Statistical significance

- In short:
- Next time when you attend any Extension workshop, and the speaker is discussing some research data
- You need to think about what level of probability is being used to evaluate treatment differences

Statistical significance

- Statistics allows researchers to the error associated with the experiment
- To separate real treatment differences from differences caused by uncontrollable environmental factors

Statistical significance

- It must be used properly and effectively
- Like to separate grain from the chaff
- That is why the next time you need to learn the significance of randomization and replication in experimental research

Example

- **At harvest the yield levels of the three treated plots are 54, 53, and 52. The three untreated plots yielded 50, 52, and 48. The average yield levels for the treated and untreated plots are 53 and 50 bushels per acre, respectively.**
- **Statistical analysis reveals that the probability of the fungicide treatment resulting in a 3 bushel per acre yield increase by sheer chance is 8% ($p = 0.08$). Thus, we are comfortable stating that the treated plots yielded significantly higher than the untreated plots.**

Why do we use Experimental Design in Agriculture?

- **To develop new crop varieties**
- **To study effects of changes in the factor of production**
- **To explore new technologies, new crops, and new areas**

Why do we use Experimental Design in Agriculture?

- **To demonstrate new knowledge**
- **To develop understanding of production control factors**

Principles of Experimental Design

- **Randomization:** A random process for assigning treatment
- **Replication:** Repetition of the basic experiment
- **Local Control:** Control of extraneous sources of variation

Basic Experimental Designs

- **The completely Randomized Design (CRD)**
- **The randomized Complete Block Design (RCBD)**

The Completely Randomized Design (CRD)

- **Simplest type of the basic designs**
- **Least restrictive experimental design**

- **Advantages of CRD: Flexible, simple in statistical analysis, and not affected by missing plots**

Completely Randomized Design Field Plan

- **Three treatments A,B,C**
- **Replicated four times**
- **Total plots are $3 \times 4 = 12$**

- **Allocation of Treatments: By some random process e.g. could be lottery method or random number table**

Completely Randomized Design Field Plan

SITE MAP WITH 12 NUMBERED PLOTS

1	2	3	4
5	6	7	8
9	10	11	12

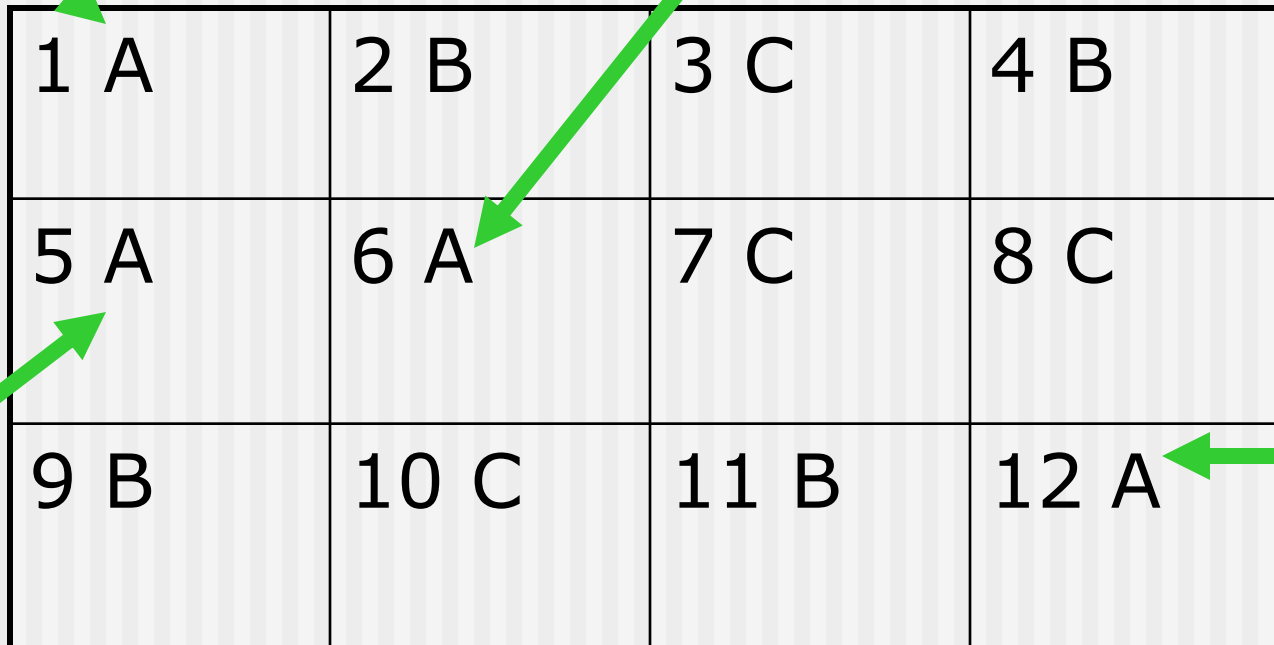
Completely Randomized Design Field Plan

SITE MAP WITH 12 NUMBERED PLOTS

1 A	2 B	3 C	4 B
5 A	6 A	7 C	8 C
9 B	10 C	11 B	12 A

Completely Randomized Design Field Plan

SITE MAP WITH 12 NUMBERED PLOTS



1 A	2 B	3 C	4 B
5 A	6 A	7 C	8 C
9 B	10 C	11 B	12 A

The table shows a 3x4 grid of 12 numbered plots. Each plot is labeled with a number and a treatment letter (A, B, or C). Green arrows point to plots 1, 6, 5, and 12.

The Randomized Complete Block Design (RCBD)

- **Advantages:**

- Mostly used in agriculture research
- Increases precision over the CRD
- Flexible and simple in analysis

Randomized Complete Block Design Field Plan

- **In this Plan:** Six treatments A,B,C,D,E, and F
- Can be tested in 3 blocks.
- Total number of plots required for this plan are $3 \times 6 = 18$
- Treatments are allocated at random within each block

Randomized Complete Block Design Field Plan

1
2
3
4
5
6

Block 1

7
8
9
10
11
12

Block 2

13
14
15
16
17
18

Block 3

Randomized Complete Block Design Field Plan

1	D
2	B
3	A
4	C
5	F
6	E

Block 1

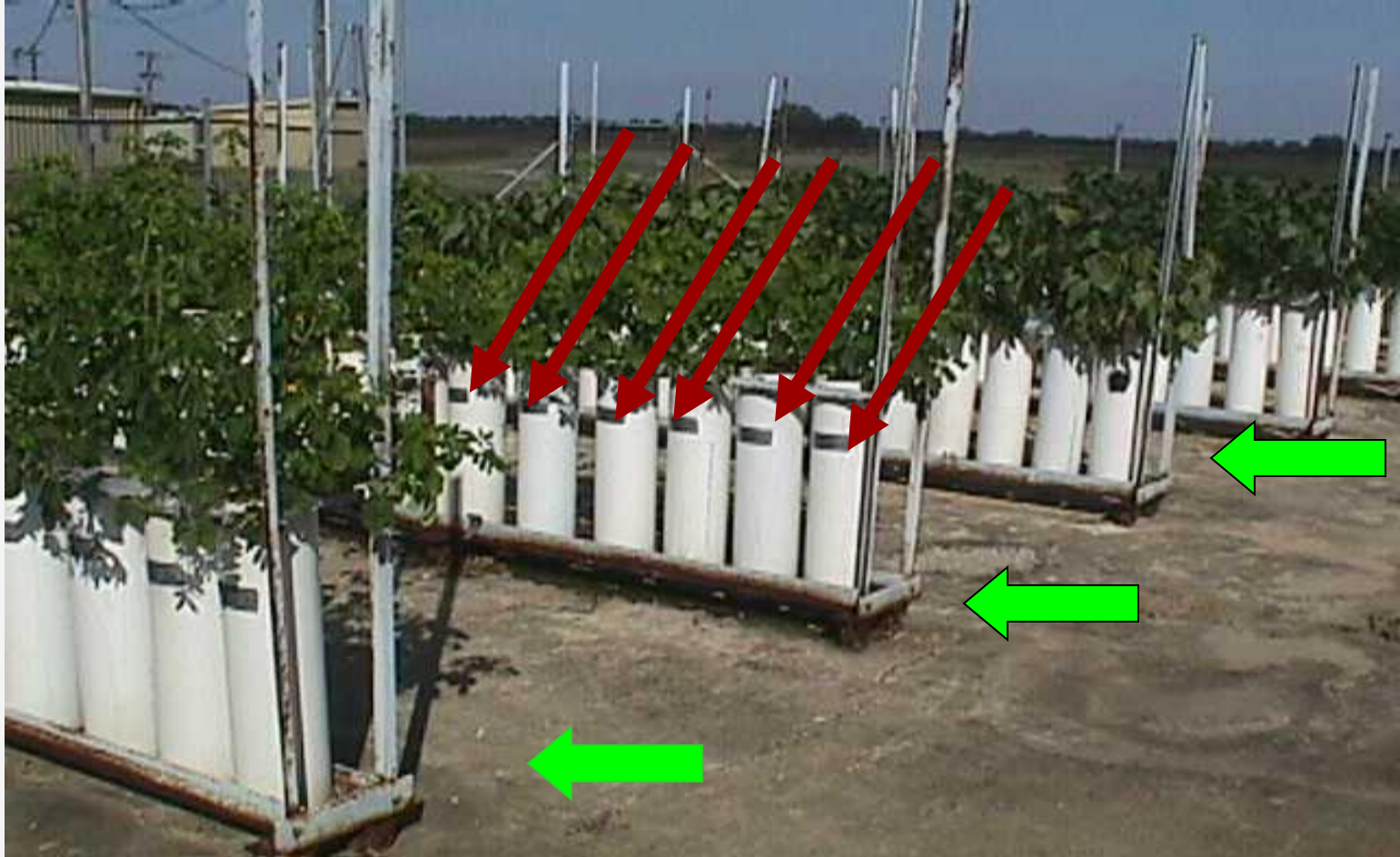
7	C
8	B
9	E
10	F
11	D
12	A

Block 2

13	C
14	F
15	B
16	D
17	A
18	E

Block 3

Randomize Complete Block Design in The Field



Important Tools in Agriculture Research

- Data Collection
- Testing of Hypotheses
- Design of Experiments

**Presented & Produced
by**

Ejaz Ashraf, Ph.D.