Experimental Design and Statistical Analysis (EDSA)
The Graduate School of CAAS

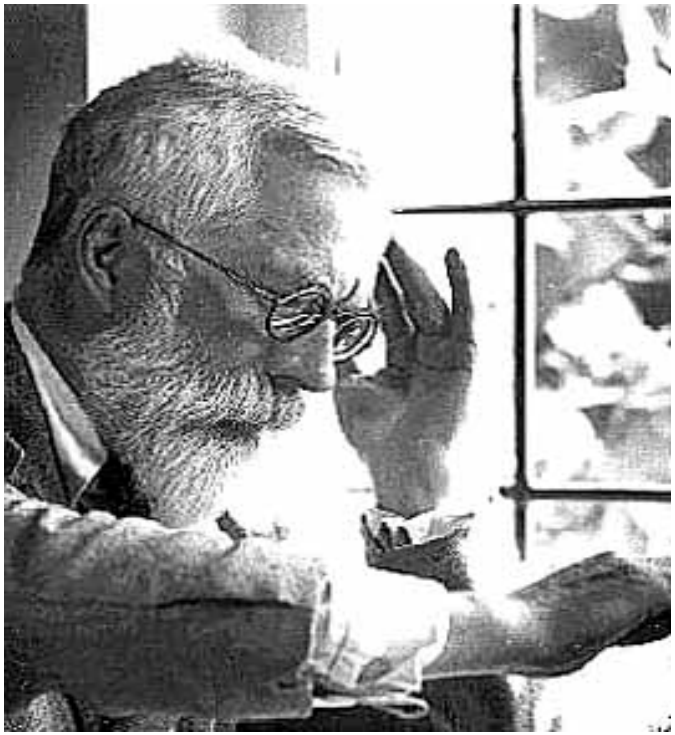
# Lecture 1

# Principles of design, probability and statistics

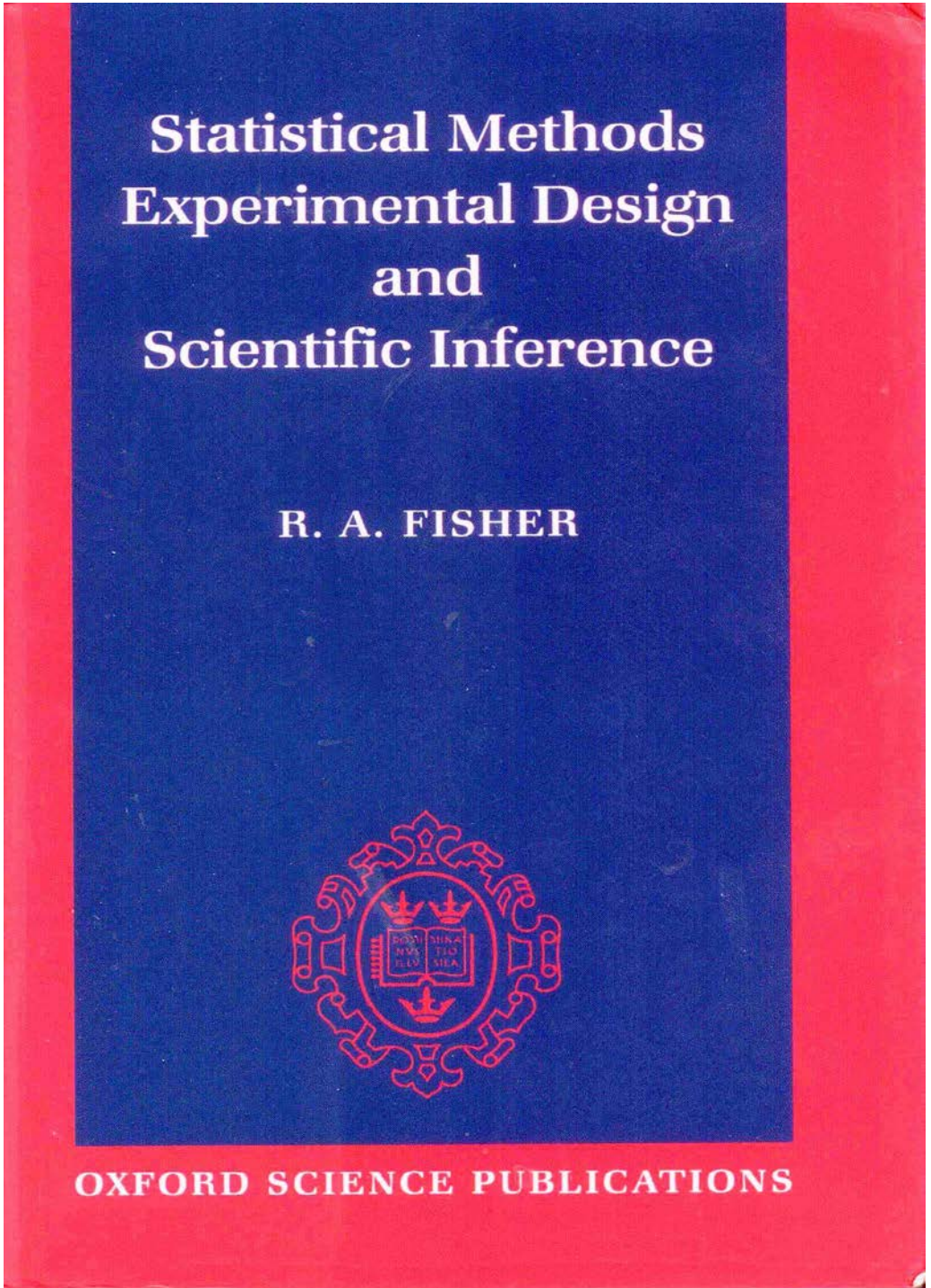
# Principles of experimental design

# The legacy of Sir Ronald A. Fisher

- He developed basic principles of design and analysis from 1919 to 1933 at Rothamsted Experimental Station, England
- In his paper “The Arrangement of Field Experiments (Fisher, 1926)”, he outlined and advanced three fundamental components for experiments in agricultural trials: **Local control, Replication and Randomization**
- Statistical Principles for Research Workers (Fisher, 1925)
- The Design of Experiments (Fisher, 1935)

**Sir Ronald Fisher**
**(1890-1962)**



# Fisher’s contributions to genetics and statistics

- His work on the theory of population genetics also made him one of the three great figures of that field, together with Sewall Wright and J. B. S. Haldane, and as such was one of the founders of the neo-Darwinian modern evolutionary synthesis. His 1918 paper *The Correlation Between Relatives on the Supposition of Mendelian Inheritance* was the start of the modern evolutionary synthesis — a synthesis which he would later contribute much to in his 1930 book *The Genetical Theory of Natural Selection*.
- Much of Fisher‘s contributions to statistics were based on biological data from Rothamsted. Fisher invented the techniques of maximum likelihood and analysis of variance (F-test), was a pioneer in the design of experiments, and originated the concepts of sufficiency, ancillarity, and Fisher information score, making him a major figure in 20th century statistics.

# Why conduct experiments?

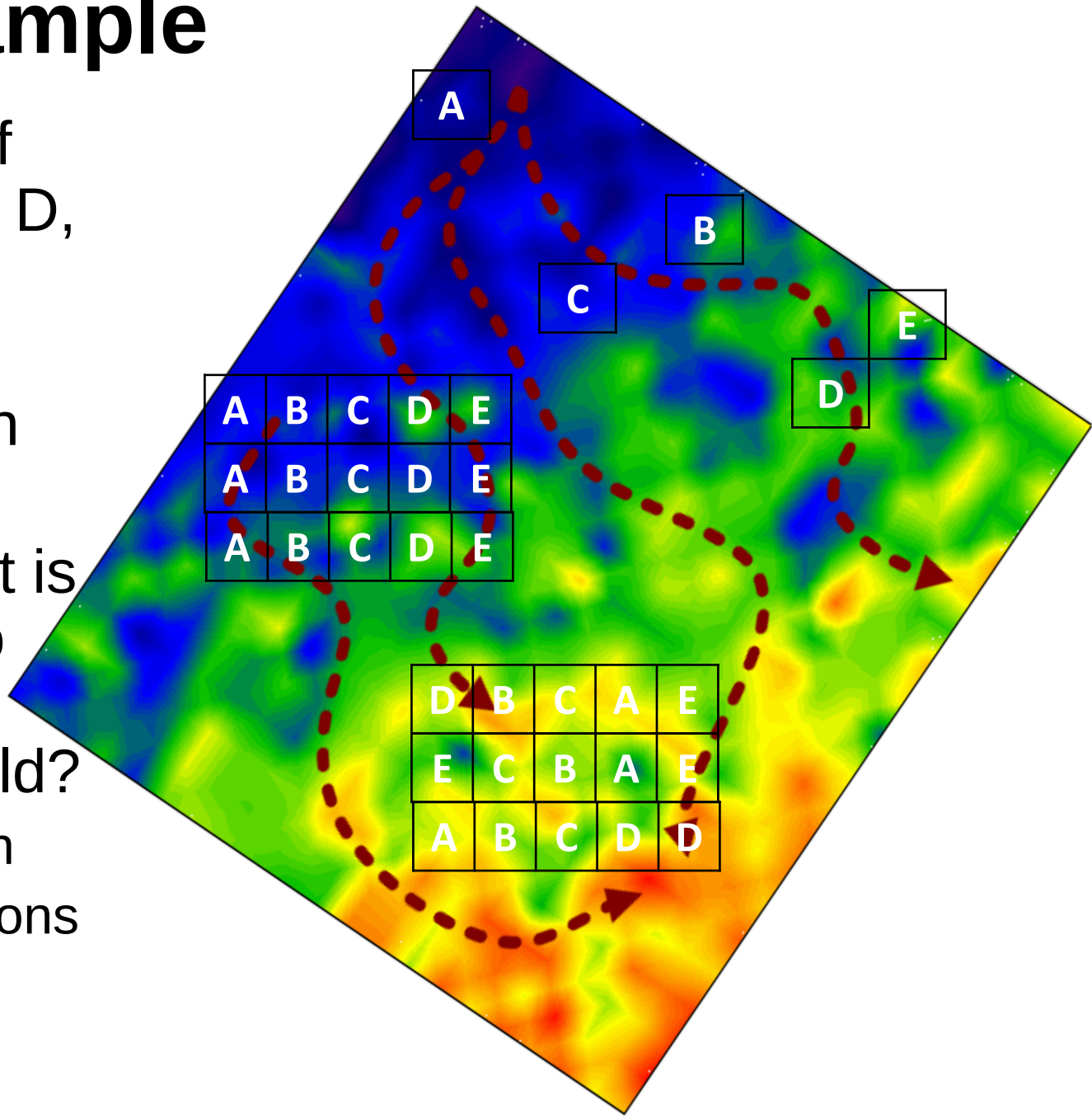
- To explore new technologies, new crops, and new areas of production
- To develop a basic understanding of the factors that control production
- To develop new technologies that are superior to existing technologies
- To study the effect of changes in the factors of production and to identify optimal levels
- To demonstrate new knowledge to growers and get feedback from end-users about the acceptability of new technologies

# Importance of the design

- An experimental design is a rule that determines the assignment of the experimental units to the treatments
- No method of analysis can save a badly designed experiment!

# An example

- Five mutants of wheat: A, B, C, D, and E
- Objective: To learn their grain yield in Beijing
- Question: What is the best way to identify their yields in the field?
  - One replication
  - Three replications



# Choice of the experimental site

- Site should be representative of the target population of environments (TPE) because of GE interaction
- Grower fields are better suited to applied research
- Suit the experiment to the characteristics of the site
  - make a sketch map of the site including differences in topography
  - minimize the effect of the site sources of variability
  - consider previous crop history
  - if the site will be used for several years and if resources are available, a uniformity test may be useful

# Source of errors in the field

- Plant variability
  - type of plant, larger variation among larger plants
  - competition, variation among closely spaced plants is smaller
  - plot to plot variation because of plot location (border effects)
- Seasonal variability
  - climatic differences from year to year
  - rodent, insect, and disease damage varies
  - conduct tests for several years before drawing firm conclusions
- Soil variability
  - differences in texture, depth, moisture-holding capacity, drainage, available nutrients
  - since these differences persist from year to year, the pattern of variability can be mapped with a uniformity trial

# Local control of experimental errors

- Selection of uniform experimental units
- Blocking to reduce experimental error variation
- Four major criteria for blocking
  - Proximity (neighboring field plots)
  - Physical characteristics (age and weight in animals)
  - Time
  - Management of tasks in experiment

# Uniformity trials —

## Identify the heterogeneous patterns in the field

- The area is planted uniformly to a single crop
- The trial is partitioned into small units and harvested individually
- Determine suitability of the site for the experiment
- Group plots into blocks to reduce error variance within blocks; blocks do not have to be rectangular
- Determine size, shape and orientation of the plots

| | | | | | |
|---|---|---|---|---|---|
| 49 | 49 | 48 | 48 | 44 | 37 |
| 49 | 46 | 44 | 44 | 44 | 37 |
| 46 | 44 | 40 | 44 | 42 | 38 |
| 44 | 40 | 40 | 42 | 40 | 38 |
| 35 | 39 | 39 | 39 | 39 | 38 |
| 35 | 39 | 39 | 39 | 39 | 40 |
| 42 | 39 | 39 | 39 | 39 | 40 |
| 43 | 41 | 38 | 38 | 38 | 40 |
| 45 | 45 | 38 | 38 | 38 | 40 |
| 45 | 45 | 43 | 44 | 44 | 44 |
| 42 | 44 | 43 | 44 | 44 | 45 |
| 42 | 42 | 40 | 40 | 44 | 44 |
| 45 | 42 | 39 | 39 | 43 | 44 |
| 45 | 42 | 39 | 40 | 43 | 44 |
| 41 | 39 | 39 | 41 | 43 | 44 |
| 39 | 39 | 39 | 41 | 41 | 37 |
| 32 | 33 | 39 | 41 | 41 | 37 |
| 32 | 33 | 37 | 43 | 43 | 38 |

# Steps to conduct an experiment

| | |
|---|---|
| Formulation of an hypothesis | ✓Definition of the problem<br>✓Statement of objectives |
| Planning an experiment to objectively test the hypothesis | ✓Selection of treatments<br>✓Selection of experimental material<br>✓Selection of experimental design<br>✓Selection of the unit for observation and the number of replications<br>✓Control of the effects of the adjacent units on each other<br>✓Consideration of data to be collected<br>✓Outlining statistical analysis and summarization of results |
| Careful observation and collection of data from the experiment | ✓Conducting the experiment |
| Interpretation of the experimental results | ✓Analyzing data and interpreting results<br>✓Preparation of a complete, readable, and correct report |

# A well-planned experiment should…

- Be of simplicity
  - don’t attempt to do too much
  - write out the objectives, listed in order of priority
- Have degree of precision
  - appropriate design
  - sufficient replication
- Be of absence of systematic error
- Have range of validity of conclusions
  - well-defined reference population
  - repeat the experiment in time and space
  - a factorial set of treatments also increases the range
- Be able to calculate the degree of uncertainty

# Terminology in experimental design

- **Experiment**: investigations that establish a particular set of circumstances under a specified protocol to observe and evaluate implications of the resulting observations, e.g. Yield test of five wheat mutants
- **Treatment or factor**: the set of circumstances created for the experiment in response to research hypotheses, e.g. Wheat mutant (or genotype, or line)
- **Level:** states of a treatment or a factor, e.g. five
- **Experimental unit**: the physical entity or subject exposed to the treatment independent of other units, i.e. plot

# Terminology in experimental design

- **Block:** group of homogeneous experimental units, e.g. 3 replications
- **Sampling unit:** part of experimental unit that is measured, e.g. a mutant at a plot
- **Variable**: measurable characteristic of a plot
- **Replications:** experimental units that receive the same treatment, e.g. yield of a mutant at a plot
- **Experimental error**: the variation among identically and independently treated experimental units, e.g. yield of a mutant at the $1^{st}$ replication is not the same as those at the $2^{nd}$ and $3^{rd}$ replication

# Three principles in design!

- **I: Local control**
  - Implemented by blocking
  - Reduce or control experimental error
  - Increase accuracy of observations
  - Establish the inference base of a study

- **Accuracy vs precision:** true value 10 for example
  - High accuracy: 9.8, 9.9, 10.3; close to true values
  - Low accuracy: 9.2, 9.7, 10.9;
  - High precision: 9.5, 9.6, 9.7; small standard errors
  - Low precision : 9.0, 9.6, 10.4;

# Three principles in design!

- **II: Replication**
  - Demonstrate the results to be reproducible
  - Provide a degree of insurance against aberrant results in the experiment due to unforeseen accidents
  - Provide the means to estimate experimental error variance
  - Provide the capacity to increase the precision for estimates of treat means

# Three principles in design!

- **III: Randomization**
  - Randomization is the random assignment of treatments to experimental units
  - Randomization provides a valid estimate of error variance for justifiable statistical inference methods of estimation and tests of hypothesis

# Completely randomized design (CRD)

For studying the effects of one primary factor without the need to take other nuisance variables into account

Statistical Model:

Response = Constant + Treatment Effect + Error Effect

# Assembling the research design

- All completely randomized designs with one primary factor are defined by 3 numbers:
- k = number of factors
- L = number of levels
- n = number of replications
- and the total sample size (number of runs, or experimental units) is $N = k \times L \times n$.
- Each factor $\times$ level combination is also called a treatment.

# An example

- Consider only one factor with 4 levels, and 3 replications per level
- Then N=12
- Note that in this example there are 12!/(3!*3!*3!*3!) = 369,600 ways to run the experiment, all equally likely to be picked by a randomization procedure.
- For example, the randomized sequence of trials might look like: 3, 1, 4, 2, 2, 1, 3, 4, 1, 2, 4, 3

# Randomly complete block design (RBD)

Blocking to increase precision by grouping the experimental units into homogeneous blocks to compare treatments within a more uniform environment

Statistical Model:

Response = Constant + Block Effect + Treatment Effect + Error Effect

# Observations of the $n = 5$ mutants in $r = 3$ replications

| Mutant | Real yield | Observations | | | Mean across replications |
|---|---|---|---|---|---|
| | | Rep I | Rep II | Rep III | |
| A | $\mu_1$ | $y_{11}$ | $y_{12}$ | $y_{13}$ | $\bar{y}_{1\cdot}$ |
| B | $\mu_2$ | $y_{21}$ | $y_{22}$ | $y_{23}$ | $\bar{y}_{2\cdot}$ |
| C | $\mu_3$ | $y_{31}$ | $y_{32}$ | $y_{33}$ | $\bar{y}_{3\cdot}$ |
| D | $\mu_4$ | $y_{41}$ | $y_{42}$ | $y_{43}$ | $\bar{y}_{4\cdot}$ |
| E | $\mu_5$ | $y_{51}$ | $y_{52}$ | $y_{53}$ | $\bar{y}_{5\cdot}$ |

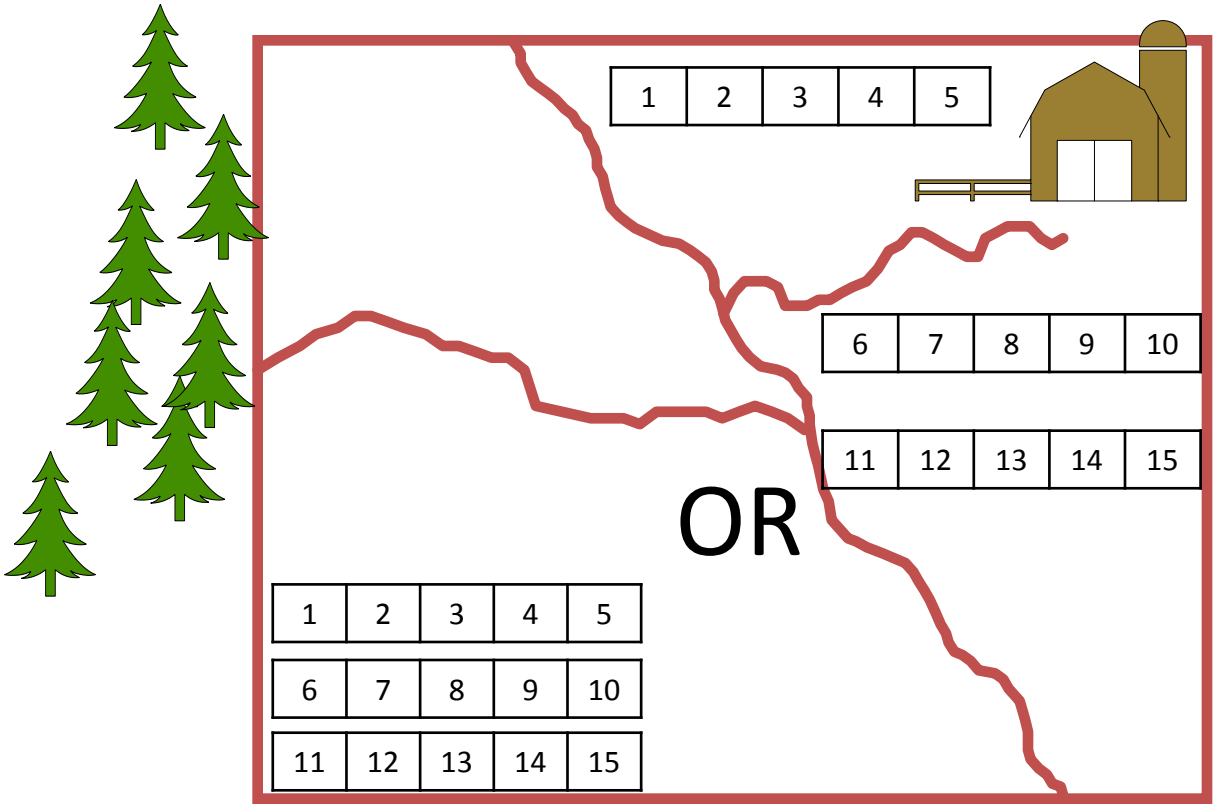
# Local control of experimental errors by blocking

- Blocking stratifies experiment units into homogeneous groups or units or plots
- Four major criteria for blocking
  - Proximity (neighboring field plots in agriculture)
  - Physical characteristics (age and weight in animals)
  - Time (a single batch of manufactured material in engineering)
  - Management of tasks (in laboratory experiment)

# Assembling the research design

- The research hypothesis
  - $H_0$: no yield difference in the 5 wheat mutants
  - $H_a$: yields are different in the 5 wheat mutants
- Treatment design
  - Number of mutants: $n = 5$
  - Number of replications (or blocks): $r = 3$
  - A total of $n \times r = 15$ plots
- Experiment design
  - Randomly complete block design (RBD)

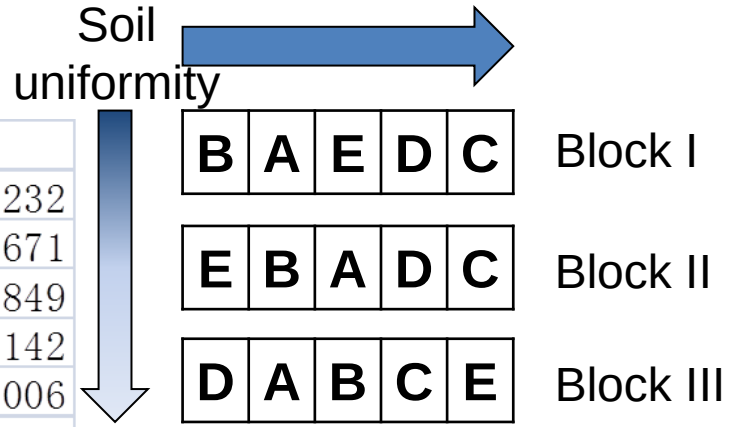
# Experimental site for the 15 plots



# How to randomize?

- Each mutant used for the experiment was a relatively homogeneous experimental unit, e.g. homogeneous in genotypic constitution
- Every plot is equally likely to be assigned to any treatment

| Mutant | Rep 1 | Mutant | Rep 2 | Mutant | Rep 3 |
|---|---|---|---|---|---|
| B | 0.033996 | E | 0.250736 | D | 0.112232 |
| A | 0.146851 | B | 0.696279 | A | 0.213671 |
| E | 0.356414 | A | 0.727923 | B | 0.234849 |
| D | 0.724533 | D | 0.736666 | C | 0.834142 |
| C | 0.86423 | C | 0.79693 | E | 0.905006 |
| | =RAND() | | =RAND() | | =RAND() |



# Other ways to minimize variation within blocks in RBD

- Field operations should be completed in one block before moving to another
- If plot management or data collection is handled by more than one person, assign each to a different block

# Advantages of the RBD

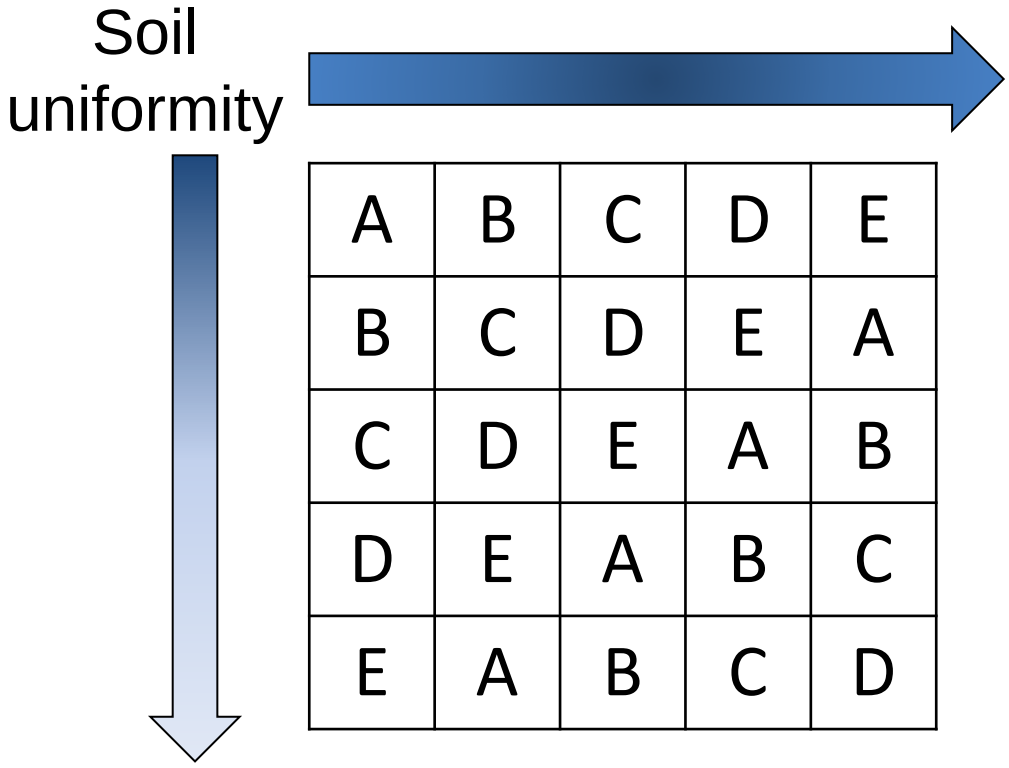
- Can remove site variation from experimental error and thus increase precision
- By placing blocks under different conditions, it can broaden the scope of the trial
- Can accommodate any number of treatments and any number of blocks, but each treatment must be replicated the same number of times in each block
- Statistical analysis is fairly simple

# Disadvantages of the RBD

- Missing data can cause some difficulty in the analysis
- Assignment of treatments by mistake to the wrong block can lead to problems in the analysis
- If there is more than one source of unwanted variation, the design is less efficient
- If the plots are uniform, then RBD is less efficient than CRD (completely randomized design)
- As treatment or entry numbers increase, more heterogeneous area is introduced and effective blocking becomes more difficult. Split plot or lattice designs may be better suited

# Other experimental designs

# 5×5 Latin square design use two blocking criteria




| | | | | |
|---|---|---|---|---|
| A | B | C | D | E |
| B | C | D | E | A |
| C | D | E | A | B |
| D | E | A | B | C |
| E | A | B | C | D |

# Incomplete block design -- Balanced

5 mutants, 3 replications, 5 blocks

| | | | |
|---|---|---|---|
| Block 1 | **A** | **B** | **C** |
| Block 2 | **B** | **C** | **D** |
| Block 3 | **C** | **D** | **E** |
| Block 4 | **D** | **E** | **A** |
| Block 5 | **E** | **A** | **B** |

# Factorial treatment design

- Factor A: Planting date
  - A1 Early, A2 Medium, A3 Late (i=1, 2, 3), m=3
- Factor B: Mutant
  - B1, B2, B3, B4, B5 (j=1, 2, 3, 4, 5), n=5
- Replication: r=3

# Split-plot design

- Factor A: Planting date
  - A1 Early, A2 medium, A3 late (i=1, 2, 3) , m=3
- Factor B: Mutant
  - B1, B2, B3, B4, B5 (j=1, 2, 3, 4, 5), n=5
- Replication: r=3

| A1 | A2 | A3 |
|---|---|---|

| B3 | B5 | B4 |
|---|---|---|
| B2 | B2 | B1 |
| B5 | B1 | B2 |
| B4 | B4 | B3 |
| B1 | B3 | B5 |

# Alpha (α) lattice design

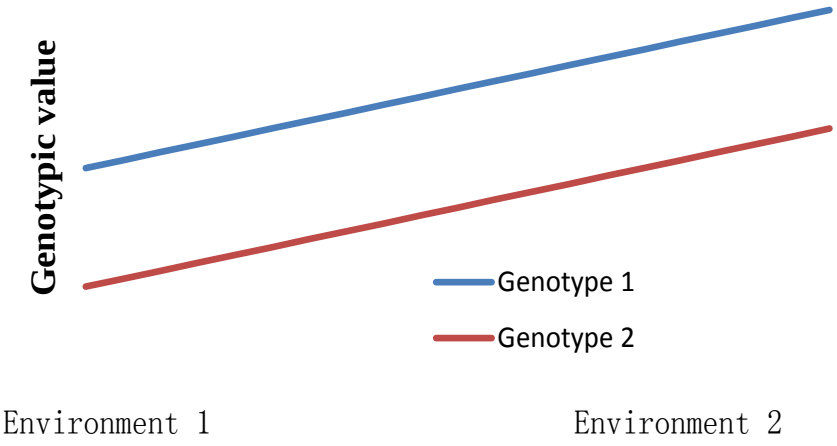
12 treatments in 3 replications

| Replicate | I | | | II | | | III | | |
|---|---|---|---|---|---|---|---|---|---|
| Block | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| | 4 | 5 | 6 | 4 | 5 | 6 | 4 | 5 | 6 |
| | 7 | 8 | 9 | 7 | 8 | 9 | 7 | 8 | 9 |
| | 10 | 11 | 12 | 10 | 11 | 12 | 10 | 11 | 12 |

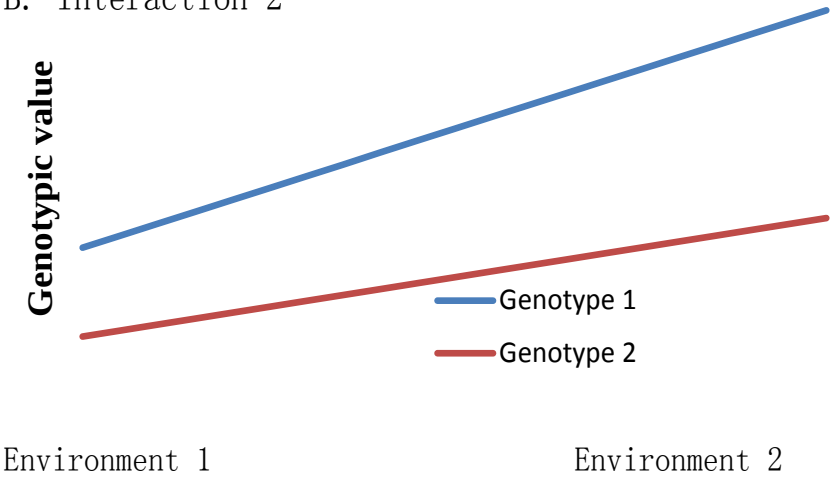
# Multi-environmental trials (MET)

# Four modes of GE interaction

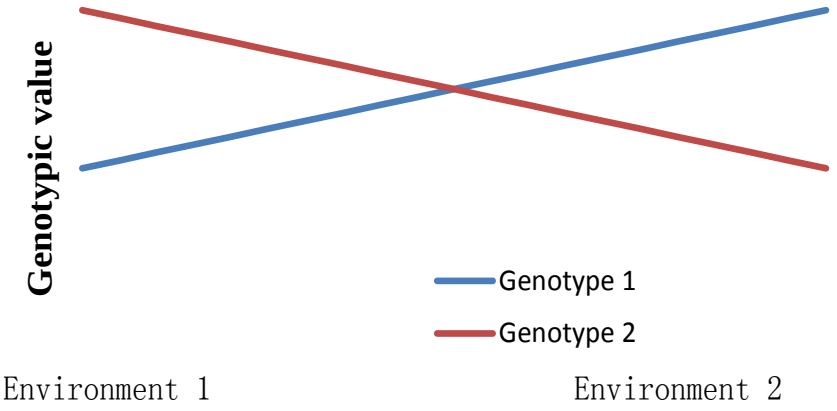
A. Interaction 1



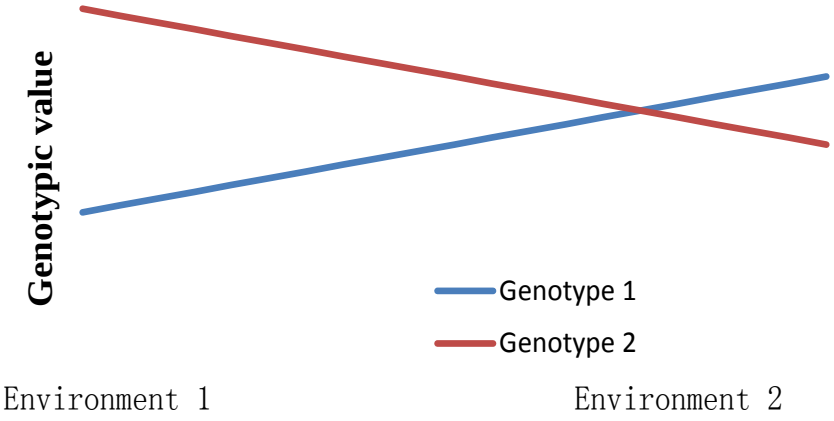

B. Interaction 2




C. Interaction 3




D. Interaction 4



# Purpose of MET

- How do varieties change in response to differences in soil and weather throughout a region?
- Detect and quantify interactions of varieties and locations and interactions of varieties and seasons in a target population of environments (TPE)
- Combined estimates are valid only if locations are randomly chosen within target area
  - Experiments often carried out on experiment stations
  - Generally use sites that are most accessible or convenient
  - Can still analyze the data, but consider possible bias due to restricted site selection when making interpretations

# An example

5 mutants tested in 2 environment, each of 3 replications

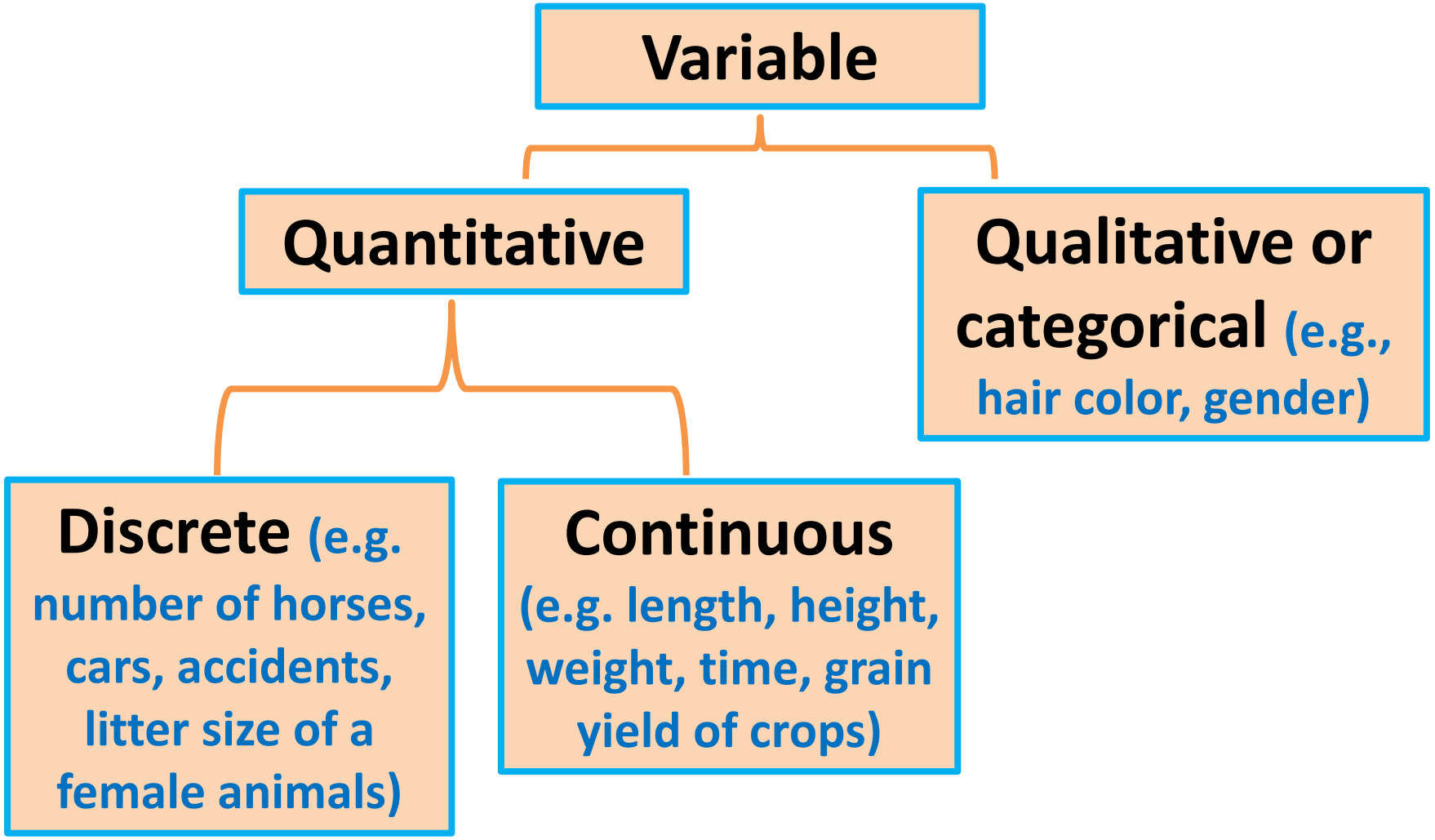
| Environment | E1 | | | E2 | | |
|---|---|---|---|---|---|---|
| Replication | R1 | R2 | R3 | R1 | R2 | R3 |
| Mutant A | 4.9 | 5.8 | 6.3 | 5.9 | 6.3 | 6.9 |
| Mutant B | 5.3 | 6.2 | 6.5 | 5.7 | 6.5 | 6.8 |
| Mutant C | 4.3 | 5.9 | 4.1 | 5.5 | 6.2 | 5.7 |
| Mutant D | 5.1 | 4.8 | 5.6 | 3.8 | 4.9 | 5.2 |
| Mutant E | 3.5 | 4.9 | 3.9 | 4.6 | 5.2 | 4.7 |

# Variables

# Variables and constants

- A **variable** is a characteristic under study that assumes different values for different elements.
- In contrast to a variable, the value of a constant is fixed, such as planting date
- The value of a variable for an element is called an observation or measurement.
- A data set is a collection of observations on one or more variables.

# Types of variables



# Qualitative traits in Mendel's hybridization experiments (1866)

- Seed shape: 5474 round vs 1850 wrinkled
- Cotyledon color: 6022 yellow vs 2001 green
- Seed coat color: 705 grey-brown vs 224 white
- Pod shape: 882 inflated vs 299 constricted
- Unripe pod color: 428 green vs 152 yellow
- Flower position: 651 axial vs 207 terminal
- Stem length: 787 long (185-230cm) vs 277 short (20-50cm)

Each trait on the right is controlled by one pair of genes.

The seven loci are independent in genetics, i.e. no linkage between them.

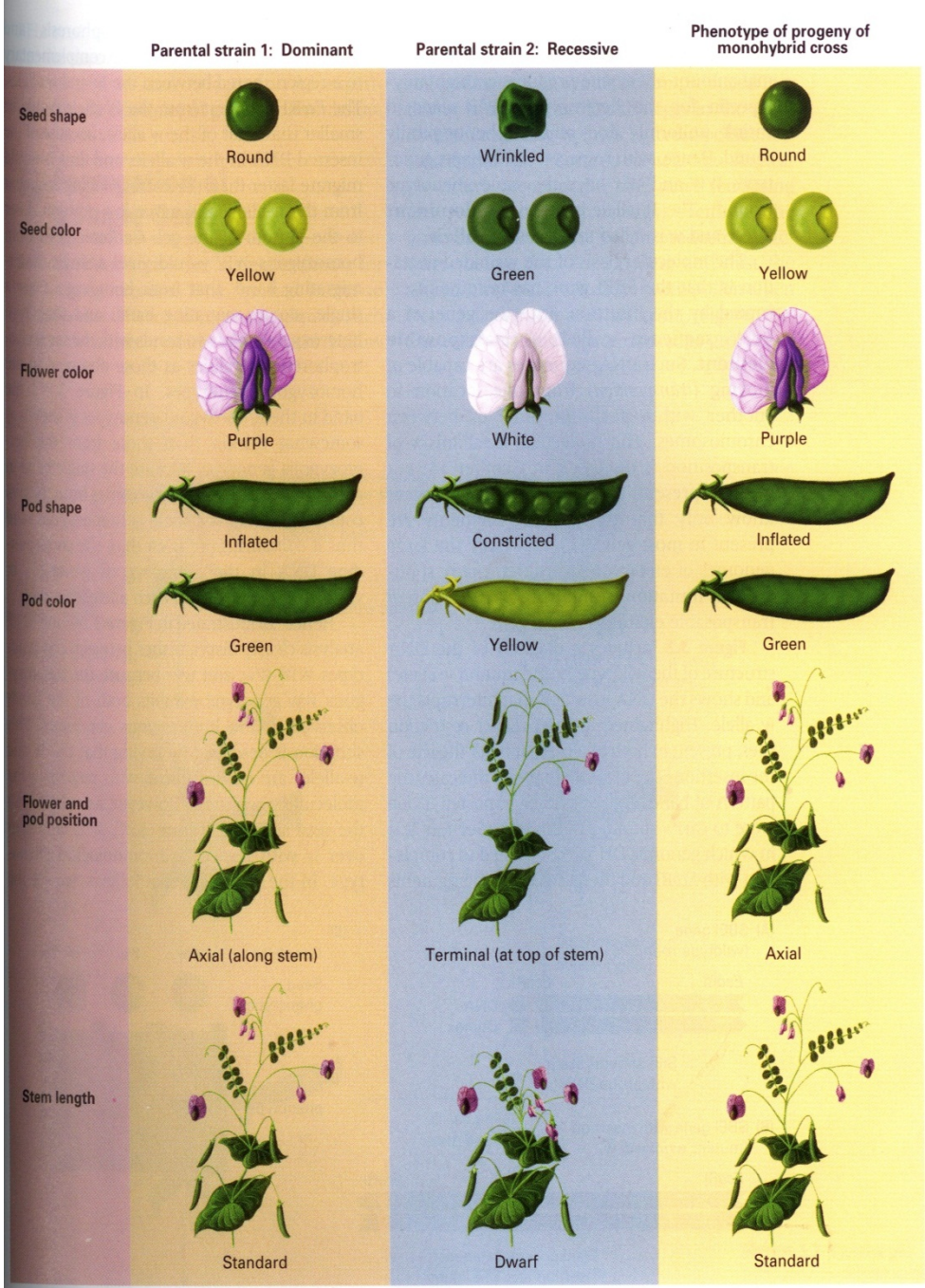
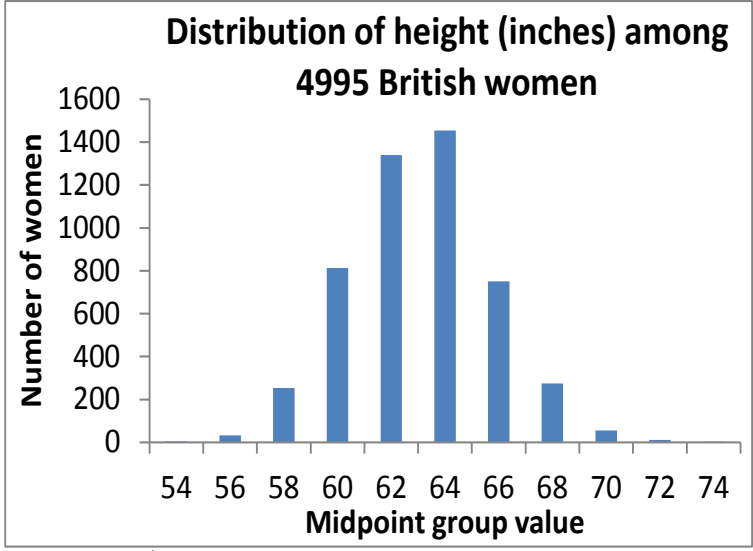
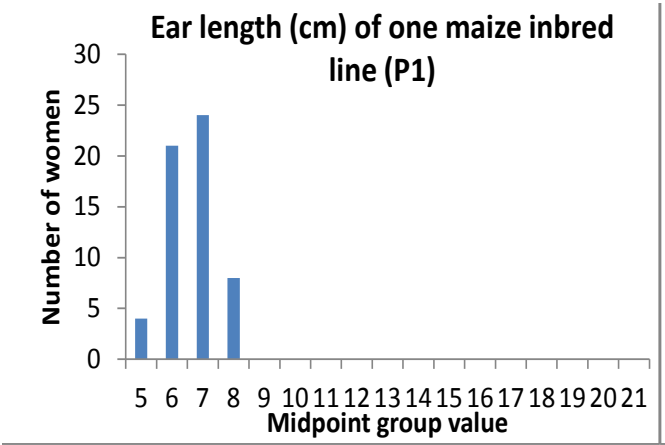
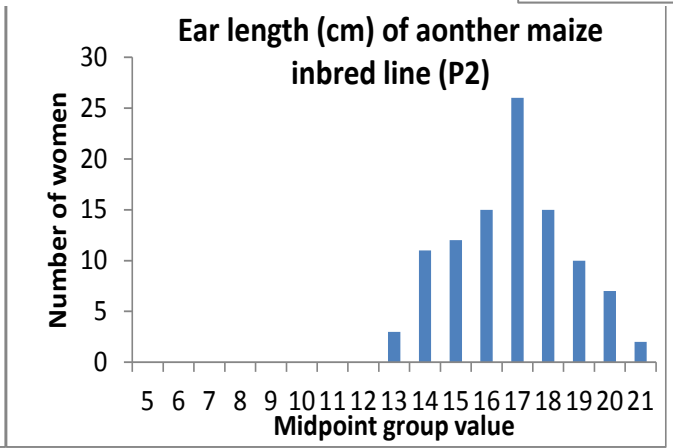


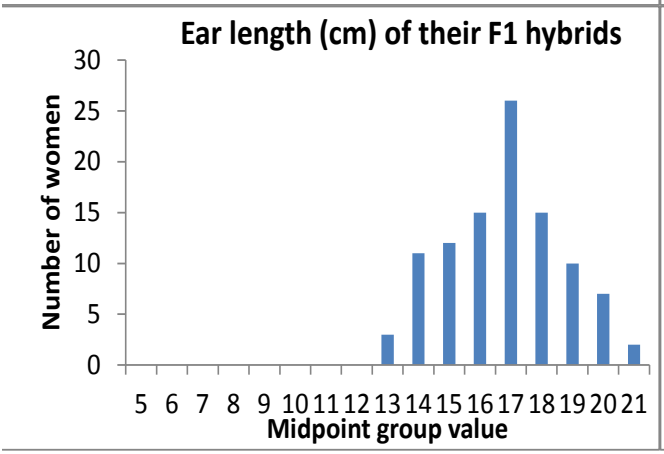

**Figure 3.1** The seven different traits in peas studied by Mendel. The phenotype shown at the far right is the dominant trait, which appears in the hybrid produced by crossing.

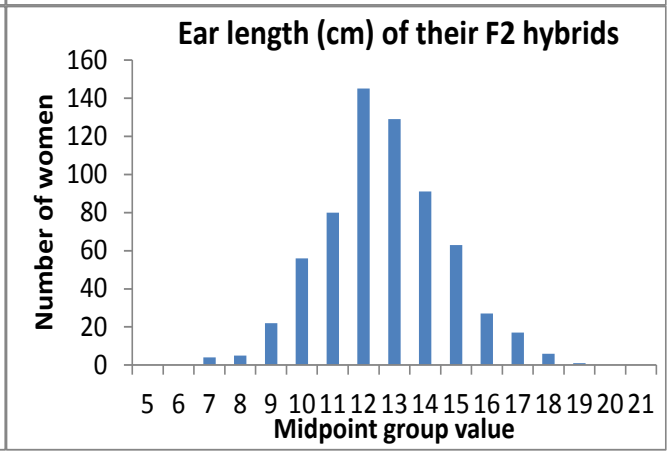
# Quantitative traits



# Quantitative trait does not have to be "continuous"

- Categorical traits: traits in which the phenotype corresponds to any one of a number of discrete categories
  - Number of skin ridges forming the fingerprints
  - Number of kernels on an ear of corn
  - Number of puppies in a litter
- Threshold traits: traits that have only two, or a few, phenotypic classes, but their inheritance is determined by the effects of multiple genes acting together with the environment
  - Liability to express the trait, which is not directly observable.
  - When liability is high enough (above a "threshold"), the trait will be expressed; Otherwise, the trait is not expressed.

# Variables

- The **independent variable** is the variable that is purposely changed. It is the manipulated variable. E.g. amount of fertilizer, planting date and density, genotype etc.
- The **dependent variable** changes in response to the independent variable. It is the responding variable. E.g. grain yield and quality.

# Single Variable vs. Multiple Variables

- Single Variable:
  - Only one independent variable
  - Cannot look at interactions
- Multiple Variables:
  - Two or more independent variables
  - If use factorial design, can look at interactions
  - Can require a lot of participants (between) or time (within)

# Independence of data

- Samples tell us about Populations
- This is only true if the data in a sample are drawn randomly from the population
- The true difficulty of non-independent data is that we do not know how it influences the sample (could be positively or negatively correlated)

# Interactions

- Two independent variables **interact** when the effect of one depends on the level of the other

# General objectives of the course

- Introduction (Lecture 1)
- Probability and Statistics (Lectures 2~4)
- Experimental designs and analysis (Lectures 5~9)
- Test of fitness, and Regression (Lectures 10~11)
- Genetic data analysis (Lectures 12~15)

- Data analysis in Excel
- Genetic analysis in QTL IciMapping

# Class exercises

- Calculate sample mean and sample variance in Excel
- Frequency distribution in Excel

| Population | Sample size | Plant height (cm) |
|---|---|---|
| Inbred A | 10 | 155, 161, 150, 164, 165, 161, 160, 158, 166, 164 |
| Inbred B | 10 | 97, 109, 92, 103, 109, 104, 98, 106, 102, 110 |
| $F_1$ | 10 | 156, 148, 140, 150, 148, 147, 146, 155, 148, 150 |
| $F_2$ | 30 | 89, 157, 149, 169, 123, 158, 151, 83, 167, 154, 152, 167, 116, 146, 97, 147, 162, 159, 111, 143, 144, 124, 137, 156, 80, 169, 157, 152, 157, 116 |

# Class exercises

- Draw the normal distributions of N(160, 27), N(149, 27), and N(103, 27)
- Draw the mixture distribution of N(160, 27), N(149, 27), and N(103, 27), with the proportions 0.25, 0.5 and 0.25