Introduction to Applied Statistics
STAT 5005
Introduction (Chapters 1 & 2)

Jingyu Sun

Fall 2021

Statistics and the Scientific Method (Chapter 1)

Using Surveys and Experimental Studies to Gather Data (Chapter 2)

Statistics and the Scientific Method (Chapter 1)

Statistics and the Scientific Method (Chapter 1)

In this section, we will discuss

- what statistics is
- why you need to study statistics

What is Statistics?

- Statistics is the science of **Learning from Data**, which consists of four steps:

1. Defining the problem
2. Collecting the data
3. Summarizing the data
4. Analyzing the data

- Example: The transportation department in a large city wants to assess the public's perception of the city's bus system in order to increase the use of buses within the city
- What would be the first three steps in the following problem?

Definitions

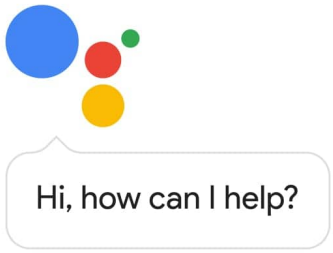
- For the results of a study to be applicable to a larger group than just the participants in the study, we must carefully define the **population** to which inferences are sought and design a study in which the **sample** has been appropriately selected from the designated population
- A **population** is the set of all measurements of interest to the sample collector
- A **sample** is any subset of measurements selected from the population

Why Study Statistics?

1. To become an informed and critical reader of data-based reports and articles
   - Statistics can be made to support almost anything
   - Misunderstandings of statistical results can lead to major errors by government policymakers, medical workers
2. Statistics plays an important role in almost all areas of science, business and industry
   - Your profession may require you to employ statistical methods and/or interpret their results
     Video: https://www.youtube.com/watch?v=wV0Ks7aS7YI

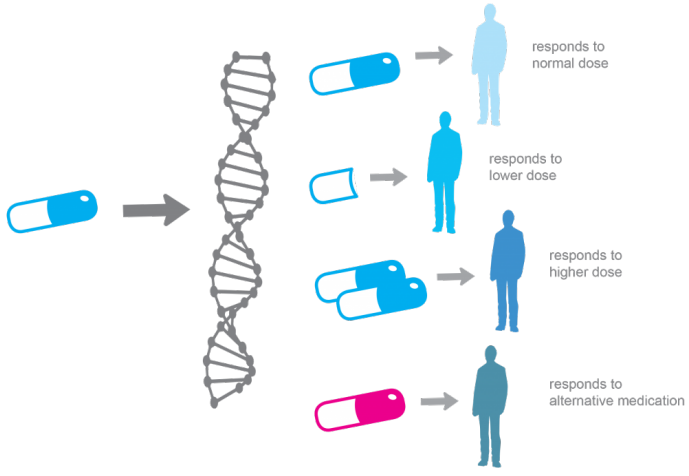
Some Current Applications of Statistics

- Speech recognition

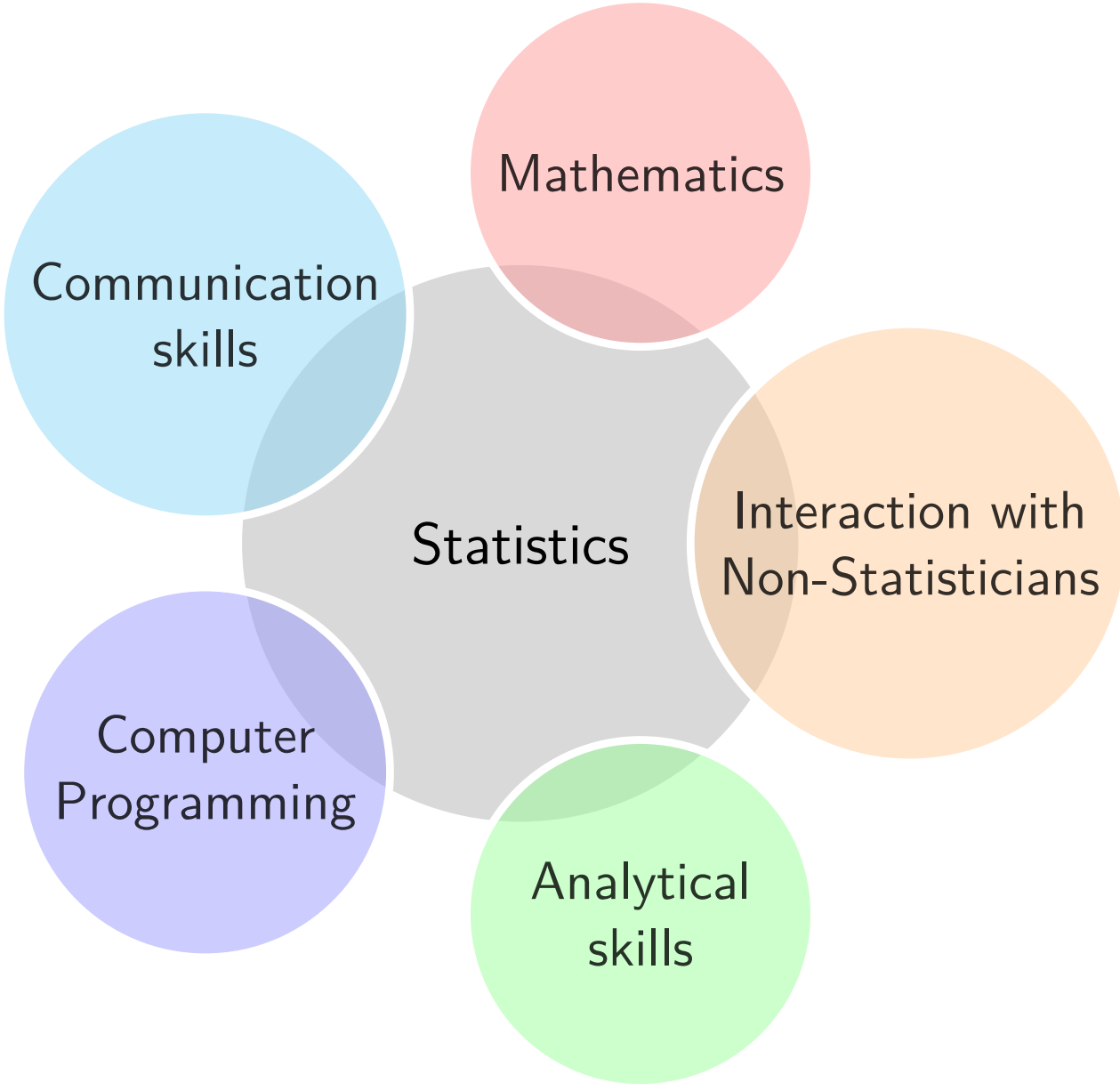


- Recommender system

- Personalized medicine



The Multiple Facets of Statistics



Using Surveys and Experimental Studies to Gather Data (Chapter 2)

Using Surveys and Experimental Studies to Gather Data (Chapter 2)

In this section, we will discuss

- intelligent data gathering
- various survey designs and experimental designs for scientific studies
- how to collect data on the variables of interest in order to address the stated objectives of the study
- the distinction between observational and experimental studies

The Data Collection Process

- Intelligent data gathering consists of the following steps:
  1. Specifying the objective of the study, survey, or experiment
  2. Identifying the variable(s) of interest
  3. Choosing an appropriate design for the survey or experimental design
  4. Collecting the data
- Example:

Source: https://bit.ly/2ZgQQXh

- Identify objective of the study and variables of interest

Surveys vs. Experiments

- The two major methods for collecting data are **surveys** and **experiments**
- In surveys or observational studies, data is gathered on existing conditions, attitudes, or behaviors
- Information on the subjects is recorded without any interference with the process that is generating the information

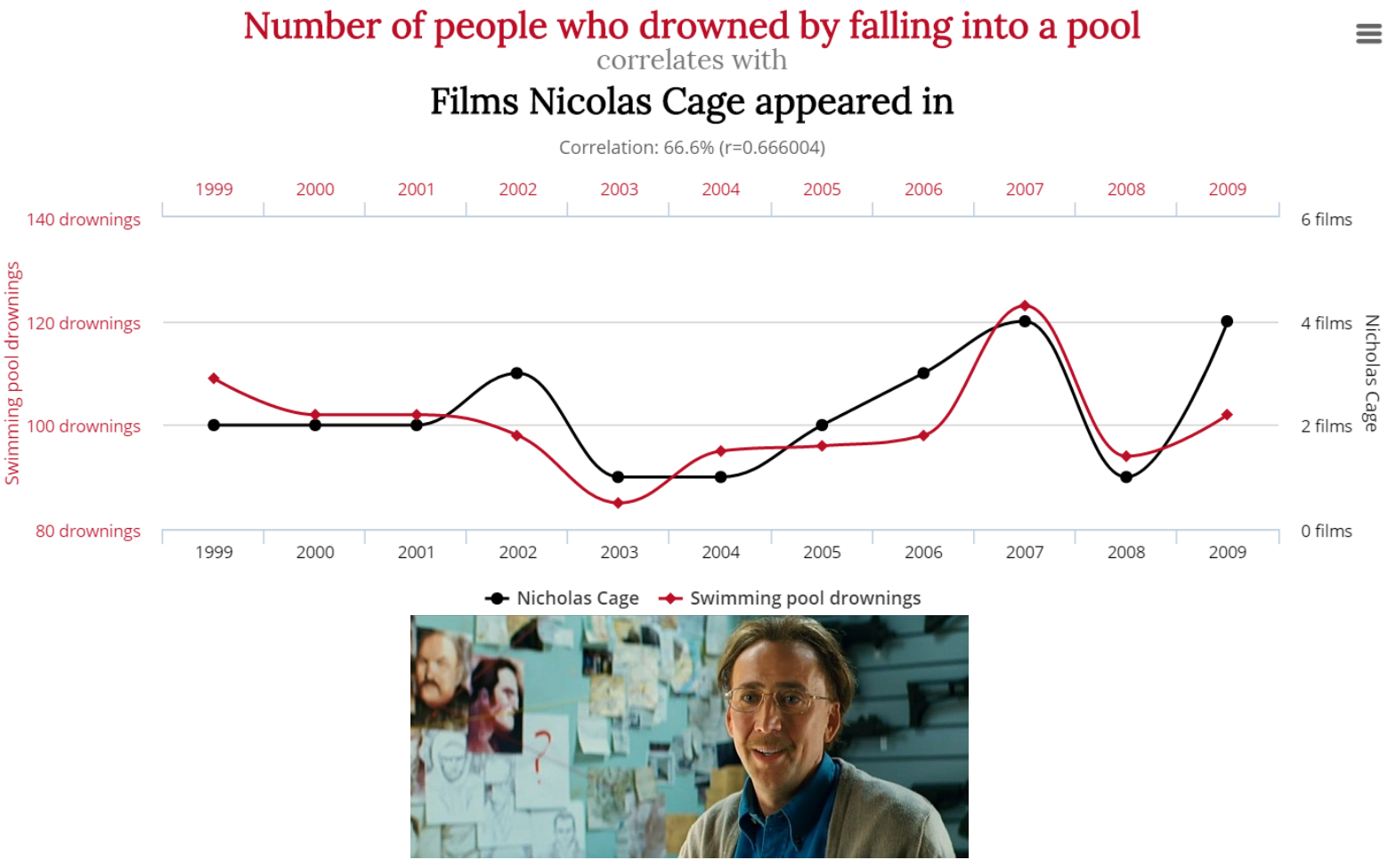
- Two types of observational studies: **comparative study** and **descriptive study**
- Descriptive study: purpose is to characterize a population based on certain attributes
  - Example: Who are the hikers in the US?
- Comparative study: two or more methods of achieving a result are compared for effectiveness
  - Example: How well are states educating pre-college students?

- In experimental studies, we vary the experimental conditions to study the effect of the conditions on the outcome of the experiment
- The researcher actively manipulates or *controls* certain variables associated with the study (the **explanatory variables** or **factors**), and then records their effects on the **response variables** associated with the experimental subjects
- Example: Effect of animal protein based diet on health
- In this chapter, we will consider sampling designs for surveys and some designs for experimental studies

Limitation of observational studies

- The recorded values of the response variables may be affected by variables other than explanatory variables
- These variables that are not under the control of the researcher are called **confounding variables**
- Observational studies are subject to many biases and sampling problems, leading to assigning **cause-and-effect relationships** to spurious associations between factors
- Observational studies can only establish **associations** between factors, **NOT** causal relationships
- Observed differences in responses between treatment groups could very well be due to these other hidden mechanisms, rather than the treatments themselves

Correlation is not causation !



Several concepts are needed to support a causal relationship:

- **Consistency:** all other things being equal, the relationship between two variables is consistent across populations in direction and maybe in amount
- **Gradient:** A relationship is more likely to be causal if a greater exposure to the suspected cause leads to a greater effect
- **Plausibility:** there is a step-by-step mechanism leading from cause to effect

Sampling Designs for Surveys

- Example: The marketing department of a kitchen appliance company develops a new website
- At a meeting, the managers are asked to assess whether the new website is an improvement over the current website
- How is the sample selected from the population in this survey? What biases may affect the study's results?

Sampling Designs for Surveys

- Example: The marketing department of a kitchen appliance company develops a new website
- At a meeting, the managers are asked to assess whether the new website is an improvement over the current website
- How is the sample selected from the population in this survey? What biases may affect the study's results?
- A good sample must accurately reflect the population as a whole to ensure the credibility and applicability of the study's results

Sampling Designs for Surveys: Definitions

- **Target population:** The complete collection of objects whose description is the major goal of the study
- **Sample:** A subset of the target population
- **Sampled population:** The population from which the sample is *actually* selected
  - Example: A telephone survey of people on the property tax list
- **Observation unit:** The object about which data are collected
- **Sampling unit:** The object that is actually sampled

Sampling Designs for Surveys: Designs

- **Simple random sampling:** select a group of n units in such a way that each sample of size n has the same chance of being selected
- **Stratified random sampling:** divide the population into two or more groups (*strata*) according to some variable and select a simple random sample from each group
- **Systematic sampling:** units are selected according to a random starting point and a fixed, periodic interval
- Apply these three sampling designs to assess voters' views on social programs in the United States

Problems Associated with Surveys

- **Survey nonresponse:** A portion of the individuals sampled cannot or will not participate in the survey. This may result in a biased survey
  - Remedies:
    1. Offering an inducement for participating in the survey
    2. Sending reminders
    3. Using statistical techniques to adjust the survey findings

- **Measurement problems:** The respondents' answers do not provide the type of data that the survey was designed to obtain
  - Reasons:
    1. Inability to recall answers to questions. Example: 'How many books did you read during the past year?'
    2. Leading questions: The fashion in which an opinion question is posed may be biased in the direction in which the question is slanted. Compare "Do you support the city's bus fleet renewal program, which may increase efficiency and passenger comfort and reduce transport related emissions?" vs "Do you support the city's bus fleet renewal program, which may increase taxes by 3%?"
    3. Unclear wording of questions: Different definitions of important words or phrases in survey questions may greatly reduce the accuracy of results. Example: "How many vehicles do you have?"

Terminology for Experiments

- A **designed experiment** is an investigation in which a specified framework is provided in order to observe, measure, and evaluate groups with respect to a designated response. The researcher controls the elements of the framework during the experiment in order to obtain data from which statistical inferences can provide valid comparisons of the groups of interest
- Controlled variables called **factors** are selected by the researchers for comparison
- Response variables are **measurements** or **observations** that are recorded but not controlled by the researcher

- The **treatments** in an experimental study are the conditions constructed from the factors
- In some experiments, there may only be a single factor, and hence the treatments and levels of the factor would be the same
- In most cases, we will have several factors and the treatments are formed by combining levels of the factors. This type of treatment design is called a **factorial treatment design**

Example

- A researcher is studying the conditions under which commercially raised shrimp reach maximum weight gain.
- Three water temperatures (25°, 30°, 35°) and four water salinity levels (10%, 20%, 30%, 40%) were selected for study
- Shrimp were raised in containers with specified water temperatures and salinity levels. The weight gain of the shrimp in each container was recorded after a 6-week study period.
- The experiment was conducted as follows: 24 containers were available for the study. A specific variety and size of shrimp was selected for study. The density of shrimp in the container was fixed at a given amount. One of the three water temperatures and one of the four salinity levels were randomly assigned to each of the 24 containers
- Identify the response variable, factors, and treatments in this example

Terminology for Experiments (ctd')

- In some experiments, there may be a large number of factors and hence the number of treatments may be so large that only a subset of all possible treatments would be examined
- Example: baking chocolate cakes
- An experiment where only a fraction of the possible treatments are actually used is called a **fractional factorial treatment structure**
- Among the treatments, the **control treatment** is the benchmark to which the effectiveness of the remaining treatments are compared
- 3 situations in which a control treatment is particularly necessary:
  - no treatment
  - standard method
  - placebo

- The **experimental unit** is the physical entity to which the treatment is randomly assigned or the subject that is randomly selected from one of the treatment populations
- In general, we will randomly assign several experimental units to each treatment.We will thus obtain several independent observations on any particular treatment and hence will have several **replications** of the treatments
- The **measurement unit** is the physical entity upon which a measurement is taken
- Do not confound experimental unit and measurement unit!

Example

- Four types of protective coatings for frying pans are to be evaluated
- Five frying pans are randomly assigned to each of the four coatings
- A measure of the abrasion resistance of the coating is measured at three locations on each of the 20 pans
- Identify the following items for this study: treatments, replications, experimental unit, measurement unit, and total number of measurements

Terminology for Experiments (ctd')

- The term **experimental error** is used to describe the variation in the responses among experimental units that are assigned the same treatment and are observed under the same experimental conditions
- Some reasons why the experimental error is not zero:
  - (a) the natural differences in the experimental units prior to their receiving the treatment,
  - (b) the variation in the devices that record the measurements,
  - (c) the variation in setting the treatment conditions,
  - (d) the effect on the response variable of all extraneous factors other than the treatment factors
- Example: 10 rats are assigned a single dose of an experimental drug

Randomization in Experimental Studies

- In experimental studies, the researcher controls the crucial factors by one of two methods:
  - **Method 1:** The subjects in the experiment are randomly assigned to the treatments. The researcher randomly selects experimental units from a homogeneous population of experimental units and then has complete control over the assignment of the units to the various treatments

  Example: Testing new ice cream flavors

  - **Method 2:** Subjects are randomly selected from different populations of interest. The researcher has control over the random sampling from the treatment populations but not over the assignment of the experimental units to the treatments

  Example: Flu vaccine effectiveness

- In experimental studies, it is crucial that the scientist follows a systematic plan established **prior to running the experiment**
- There may be extraneous factors present that may affect the experimental units
- Randomization, i.e. either the assignment of experimental units to treatments or the selection of units from the treatment populations, ensures that, on the average, any large differences observed in the responses of the experimental units in different treatment groups can be attributed to the differences in the groups and not to factors that were not controlled during the experiment

Conducting an Experimental Study

Aspects to consider when conducting the experiment

1. The research objectives of the experiment
2. The selection of the factors that will be varied (the treatments)
3. The identification of extraneous factors that may be present in the experimental units or in the environment of the experimental setting (the blocking factors)
4. The characteristics to be measured on the experimental units (response variable)

5. The method of randomization, either randomly selecting from treatment populations or the random assignment of experimental units to treatments
6. The procedures to be used in recording the responses from the experimental units
7. The selection of the number of experimental units assigned to each treatment may require designating the level of significance and power of tests or the precision and reliability of confidence intervals
8. A complete listing of available resources and materials

Limitation of Experiments

- Does soap really kill 99.9% of germs?

Limitation of Experiments

- Does soap really kill 99.9% of germs?

- Yes, under laboratory conditions. In a study that looked at the hand washing practices of eighth graders in Hamilton, Ontario, researchers found a typical hand washing removed between 46-60% of germs. Source: https://www.wsj.com/articles/SB126092257189692937
- The greater the control in artificial settings, the less likely the experiment is portraying the true state of nature

Designs for Experimental Studies

- In the remainder of this chapter, we will have an overview of different experimental designs

Designs for Experimental Studies: Example

- A consumer testing agency decides to evaluate the wear characteristics of four major brands of tires
- The agency selects four cars of a standard car model and four tires of each brand
- The tires will be placed on the cars and then driven 30,000 miles on a 2-mile racetrack
- The decrease in tread thickness over the 30,000 miles is the variable of interest in this study
- Four different drivers will drive the cars, but the drivers are professional drivers with comparable training and experience
- The weather conditions, smoothness of track, and the maintenance of the four cars will be essentially the same for all four brands over the study period. All extraneous factors that may affect the tires are nearly the same for all four brands

- There should be a recorded tread wear for each of the sixteen tires, four tires for each brand
- The methods presented in subsequent chapters could be used to summarize and analyze the sample tread wear data in order to make comparisons (inferences) among the four tire brands
- One possible inference of interest could be the selection of the brand having minimum tread wear
- Can the best-performing tire brand in the sample data be expected to provide the best tread wear if the same study is repeated?
- Are the results of the study applicable to the driving habits of the typical motorist?

- How to assign the tires to the cars?
- We could randomly assign a single brand to each car. What is the drawback?
- A better design is to randomly assign the sixteen tires to the four cars

Car 1	Car 2	Car 3	Car 4
Brand B	Brand A	Brand A	Brand D
Brand B	Brand A	Brand B	Brand D
Brand B	Brand C	Brand C	Brand D
Brand C	Brand C	Brand A	Brand D

- The assignment shown in the above table is an instance of a **completely randomized design**

- In a **completely randomized design** (CRD), we are interested in comparing t treatments. For each of the treatments, we obtain a sample of observations which are assumed to be the result of a simple random sample of observations from the hypothetical population of possible values that could have resulted from that treatment
- Now, assume that the wear on tires imposed by Car 4 was less severe than that of the other three cars, would our design take this effect into account?

Car 1	Car 2	Car 3	Car 4
Brand B	Brand A	Brand A	Brand D
Brand B	Brand A	Brand B	Brand D
Brand B	Brand C	Brand C	Brand D
Brand C	Brand C	Brand A	Brand D

- In some situations, the objects being observed have **existing differences prior to their assignment to the treatments**
- We thus resort to a **randomized block design** (RBD) to "block" out any differences in the units to obtain a precise comparison of the treatments
- In an RBD, **each treatment appears in every block**
- In the tire wear example, we would use the four cars as the blocks and randomly assign one tire of each brand to each of the four cars, as shown in the table below

Car 1	Car 2	Car 3	Car 4
Brand A	Brand A	Brand A	Brand A
Brand B	Brand B	Brand B	Brand B
Brand C	Brand C	Brand C	Brand C
Brand D	Brand D	Brand D	Brand D

- If there are any differences in the cars that may affect tire wear, that effect will be equally applied to all four brands

- What happens if the position of the tires on the car affects the wear on the tire?
- The positions on the car are right front (RF), left front (LF), right rear (RR), and left rear (LR)
- In this type of situation, the effect of brand and the effect of position on the car would be confounded
- Thus, we now need two blocking variables: the 'car' and the 'position' on the car. A design having two blocking variables is called a **Latin square design** (LSD), see example below

Position	Car 1	Car 2	Car 3	Car 4
RF	Brand A	Brand B	Brand C	Brand D
RR	Brand B	Brand C	Brand D	Brand A
LF	Brand C	Brand D	Brand A	Brand B
LR	Brand D	Brand A	Brand B	Brand C

- Note that the RBD and LSD are both extensions of a CRD in which the objective is to compare t treatments
- The analysis of data for a CRD and for block designs and the inferences made from such analyses are discussed in later chapters

Factorial Treatment Structure in a Completely Randomized Design

- Consider an experiment where treatments are constructed with several factors rather than just being t levels of a single factor
- Here, we examine the effect of two or more independent variables on a response variable
- Example: Scientists want to study how animal protein (chicken, beef, pork) and plant protein (soy and beans) affect health. Identify factors and treatments

Factorial Treatment Structure in a Completely Randomized Design

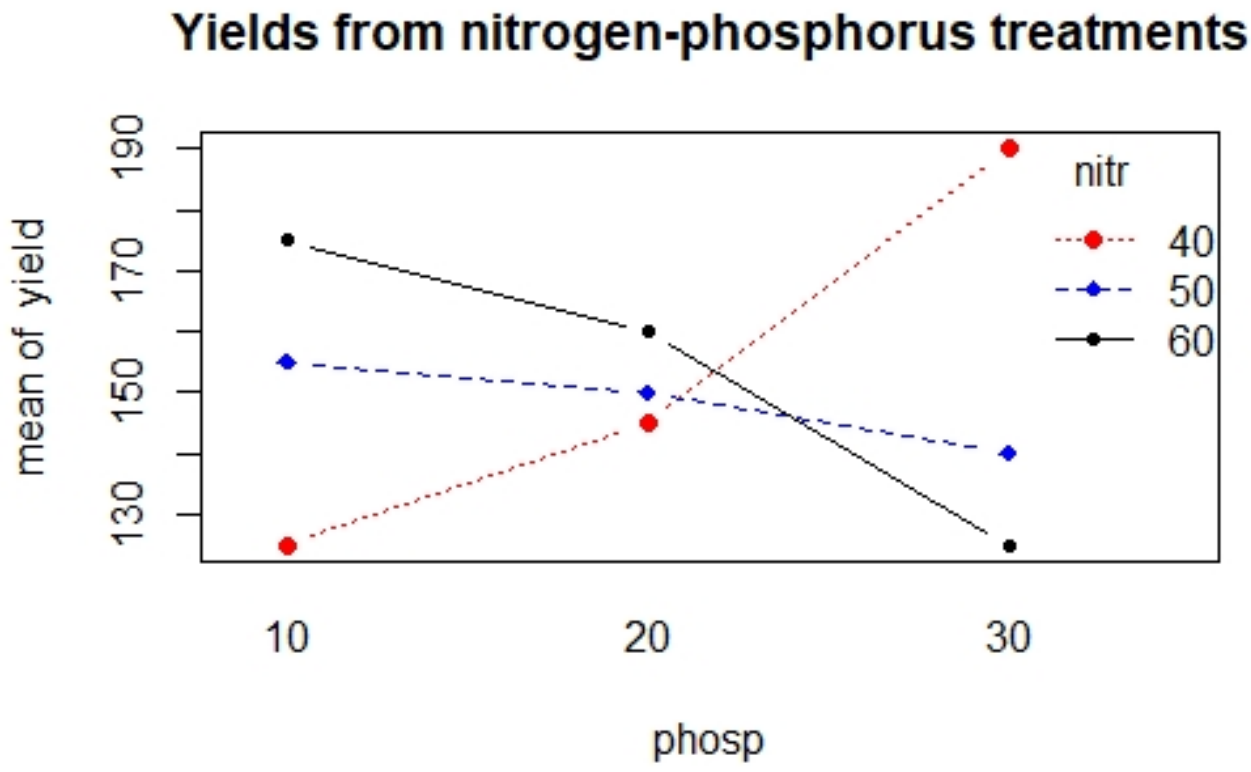
- Consider an experiment where treatments are constructed with several factors rather than just being t levels of a single factor
- Here, we examine the effect of two or more independent variables on a response variable
- Example: Scientists want to study how animal protein (chicken, beef, pork) and plant protein (soy and beans) affect health. Identify factors and treatments
- When the treatments are formed by combining levels of the factors, this type of treatment design is called a **factorial treatment design**
- Construction of treatments depends on budget, time to complete the study, and, most important, the experimenter's knowledge of the physical situation under study

- **One-at-a-time approach**: To examine the effect of a single variable, an experimenter varies the levels of this variable while holding the levels of the other independent variables fixed. This process is continued until the effect of each variable on the response has been examined

- Example: determine the combination of nitrogen and phosphorus that produces the maximum amount of corn per plot
- The values in the table below would be unknown to the experimenter

	Phosphorus		
Nitrogen	10	20	30
40	125	145	190
50	155	150	140
60	175	160	125

- This type of experimentation may produce incorrect results whenever the effect of one factor on the response does not remain the same at all levels of the second factor. In this situation, the factors are said to **interact**


Yields from nitrogen-phosphorus treatments
mean of yield
130
150
170
190
10
20
30
phosp
nitr
40
50
60

- Factorial treatment structures are useful for examining the effects of two or more factors on a response, whether or not interaction exists

Controlling Experimental Error

- There are many potential sources of experimental error in an experiment
- When the variance of experimental errors is large, the precision of our inferences will be greatly compromised
- Thus the need to potential sources of experimental errors. These sources include:
  - (1) the procedures under which the experiment is conducted,
  - (2) the choice of experimental units and measurement units,
  - (3) the procedure by which measurements are taken and recorded,
  - (4) the blocking of the experimental units,
  - (5) the type of experimental design,
  - (6) the use of ancillary variables (called covariates)

Experimental Procedures

- Differences in the responses may be due to changes in the experimental conditions and not due to treatment differences (personnel training, quality of the equipment, ...)

Selecting Experimental and Measurement Units

- Ideally, the experimental units are randomly selected from a population of interest and then randomly assigned to the treatments
- The researcher then determines whether there is a difference in the mean responses of experimental units receiving different treatments

Selecting Experimental and Measurement Units

- Ideally, the experimental units are randomly selected from a population of interest and then randomly assigned to the treatments
- The researcher then determines whether there is a difference in the mean responses of experimental units receiving different treatments
- In practice, the researcher is somewhat limited in the selection of experimental units by cost, availability, and ethical considerations

Selecting Experimental and Measurement Units: Example

- A sales campaign to market children's products uses TV commercials
- A marketing firm wants to determine whether the attention span of children is different depending on the type of product being advertised among four types of products: sporting equipment, healthy snacks, shoes, and video games
- The firm selected 100 fourth-grade students from a New York City public school. Twenty-five students were randomly assigned to view a commercial for each of the four types of products
- The attention spans of the 100 children were then recorded. The firm thought that by selecting participants of the same grade level and from the same school system it would achieve a homogeneous group of subjects
- What problems exist with this selection procedure?

Reducing Experimental Error through Blocking

- Blocking may prove to be highly effective in reducing the experimental error variance
- The experimental units are placed into groups based on their similarity with respect to characteristics that may affect the response variable
- This results in sets or blocks of experimental units that are homogeneous within the block
- The treatments are randomly assigned separately within each block
- The comparison of the treatments is within the groups of homogeneous units
- The blocking design will enable us to separate the variability associated with the characteristics used to block the units from the experimental error

- There are many criteria used to group experimental units into blocks; they include the following:
  1. Physical characteristics such as age, weight, sex, health, and education of the subjects
  2. Units that are related such as twins or animals from the same litter
  3. Spatial location of experimental units such as neighboring plots of land or position of plants on a laboratory table
  4. Time at which experiment is conducted such as the day of the week, because the environmental conditions may change from day to day
  5. Person conducting the experiment, because if several operators or technicians are involved in the experiment they may have some differences in how they make measurements or manipulate the experimental units

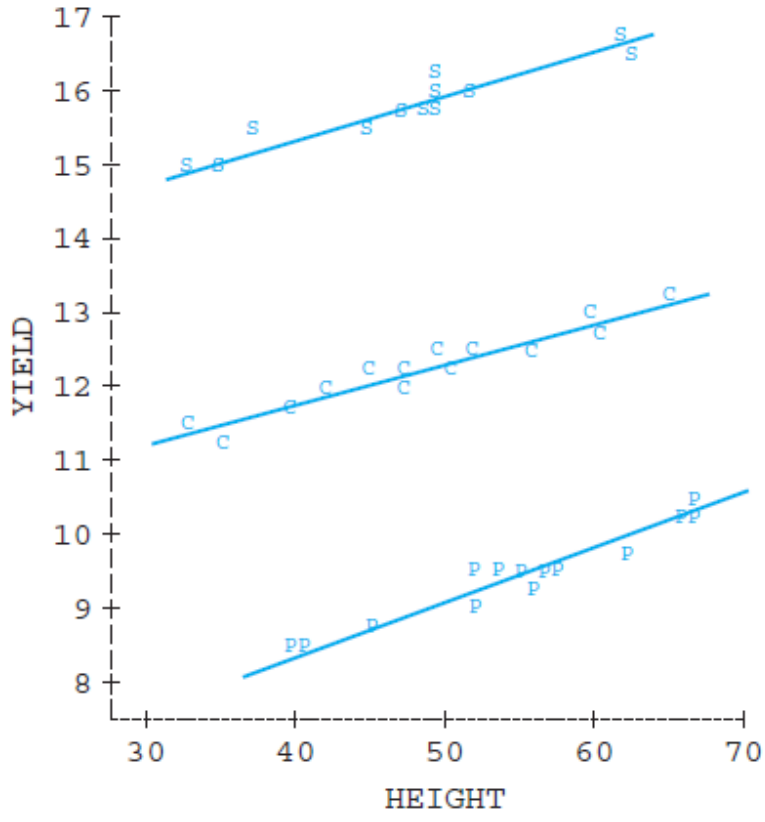
Using Covariates to Reduce Variability

- A **covariate** is a variable that is related to the response variable
- Example: compare the effectiveness of a new diet to a control diet in reducing the weight of dogs, dogs available for the study varied in age from 1 year to 12 years
- Propose a blocking strategy
- A more exacting methodology records the age of the dog and then incorporates the age directly into the model when attempting to assess the effectiveness of the diet. The response variable would be adjusted for the age of the dog prior to comparing the new diet to the control diet
- The covariate needs to have a relationship to the response variable, it must be measurable, and it cannot be affected by the treatment

- If no relationship exists between the response variable and the covariate, then the covariate need not be used in the analysis
- If the two variables are related, then we must use the techniques of **analysis of covariance** (ANCOVA) to properly adjust the response variable prior to comparing the treatment means

- Example: we study the effects of two treatments, supplemental lighting (SL) and partial shading (PS), on the yield of soybean plants were compared with normal lighting (NL)
- Each type of lighting was randomly assigned to 15 soybean plants and the plants were grown in a greenhouse study
- The experimenter knows that the plants were of differing size and maturity
- The height of the plant at the beginning of the study will serve as a covariate
- To determine whether the covariate has an effect on the response variable, we plot the two variables to assess any possible relationship

- Plot of plant height versus yield: S=Supplemental Lighting, C=Normal Lighting, P=Partial Shading



- In there a relationship between the covariate and the response variable?