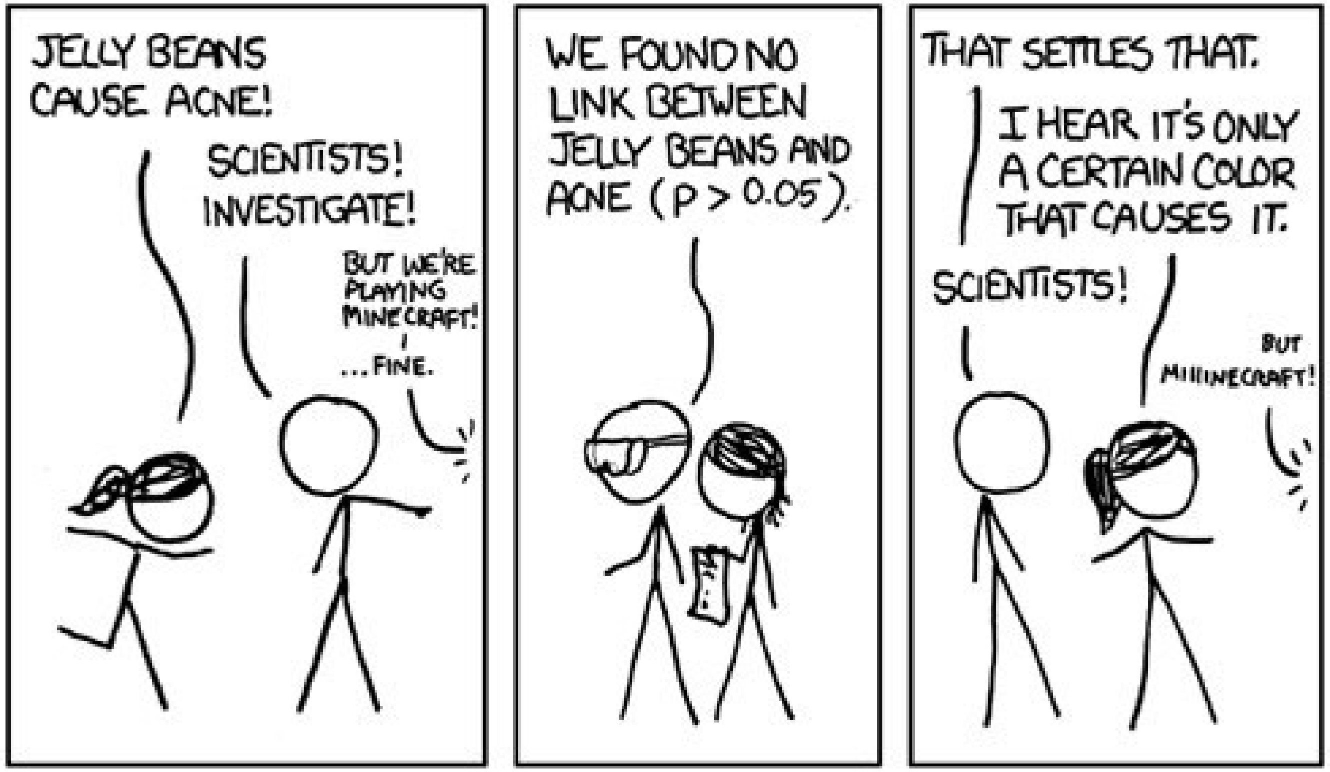
Design of Experiments: A gentle introduction

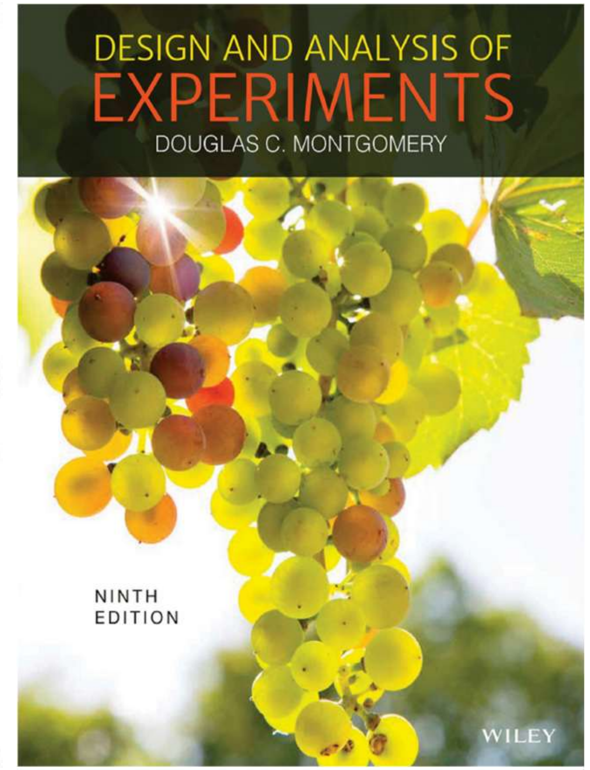
— INTRODUCTION —

Fabio Stella

University of Milano-Bicocca

fabio.stella@unimib.it





To consult the statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination.
He can perhaps say what the experiment died of.

Attributed to Ronald A. Fisher (1890-1962)

Why do we need Design of Experiments?

A well-planned research design helps ensure that your methods match your research aims, that you collect high-quality data, and that you use the right kind of analysis to answer your questions, utilizing credible sources.

This allows you to draw valid, trustworthy conclusions.

Fabio Antonio Stella

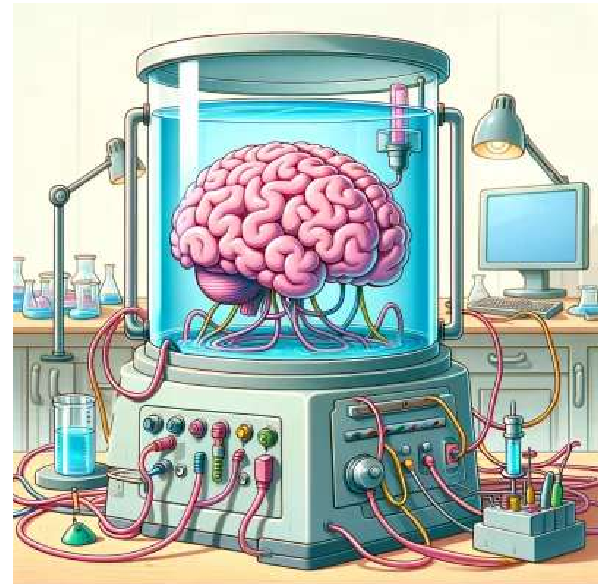
DISCo

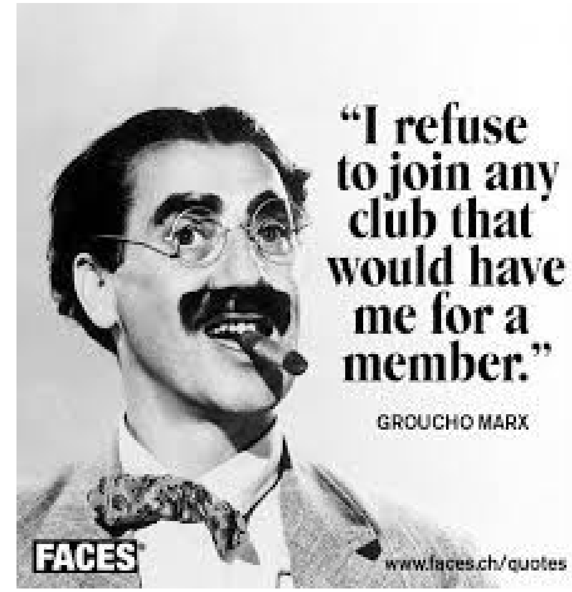
University of Milano-Bicocca

Building U14

Viale Sarca 336

20126 Milan





Research Interests

- Bayesian (causal) networks
- Causality
- Courterfactuals
- Medicine and Healthcare

FOUR MAIN LECTURES

1) INTRODUCTION
2) FACTORIAL DESIGN
3) 2^k FACTORIAL DESIGN
4) FITTING REGRESSION MODELS

INTRODUCTION

LECTURE LEARNING OBJECTIVES

1) Learn about the **objectives of experimental design** and the **role it plays in the knowledge discovery process**.

2) Learn about different **strategies of experimentation**.

3) Understand the role that **statistical methods** play in **designing and analyzing experiments**.

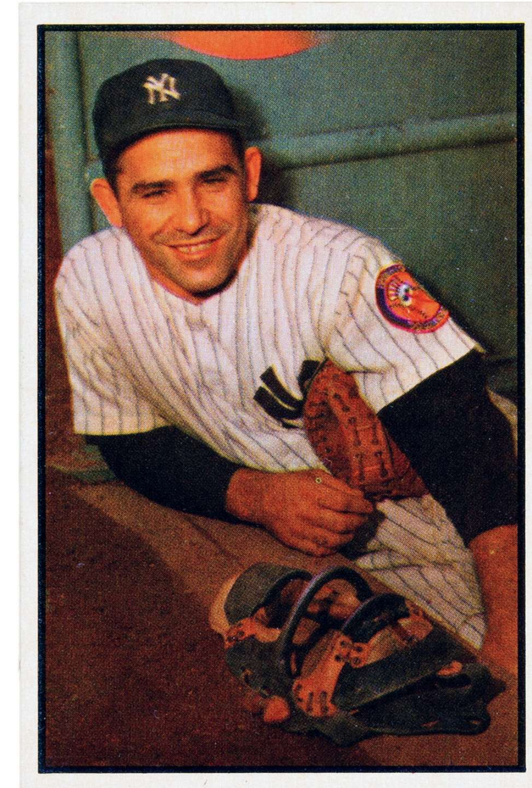
4) Understand the concepts of **main effects** of factors and **interaction** between factors.

5) Know about **factorial experiments**.

6) Know the practical **guidelines for designing and conducting experiments**.

Observing a system or process while it is in operation is an important part of the learning process and is an integral part of understanding and learning about how systems and processes work.

The great **New York Yankees** catcher **Yogi Berra** said that “*. . . you can observe a lot just by watching.*”



However, **to understand what happens to a process when you change certain input factors**, you have to do more than just watch—**you actually have to change the factors**.

This means that to really understand **cause-and-effect relationships** in a system you must deliberately change the input variables to the system and observe the changes in the system output that these changes to the inputs produce.

You need to conduct experiments on the system!!!

EXPERIMENT; a test or series of runs in which purposeful **changes** are made to the **input variables** of a process or system so that we may **observe and identify** the **reasons for changes** that may be observed **in the output response**.

We may want to

- Determine which input variables are responsible for the observed changes in the response,
- Develop a model relating the response to the important input variables,
- Use this model for process or system improvement or other decision-making.

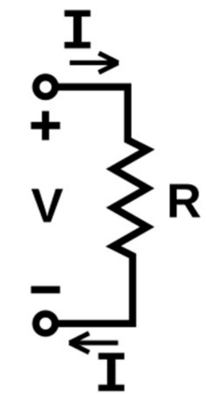
Experimentation plays an important role in **technology commercialization** and **product realization** activities, which consist of new product design and formulation, manufacturing process development, and process improvement.

The objective in many cases may be to **develop a robust process**, that is, a process **affected minimally by external sources of variability**.

Experimentation is a vital part of the scientific method.

Now there are certainly situations where the scientific phenomena are so well understood that useful results including mathematical models can be developed directly by applying these well-understood principles. The models of such phenomena that follow directly from the physical mechanism are usually called **mechanistic models**.

Ohm's Law



$$V = IR \quad \text{or} \quad I = \frac{V}{R} \quad \text{or} \quad R = \frac{V}{I}$$

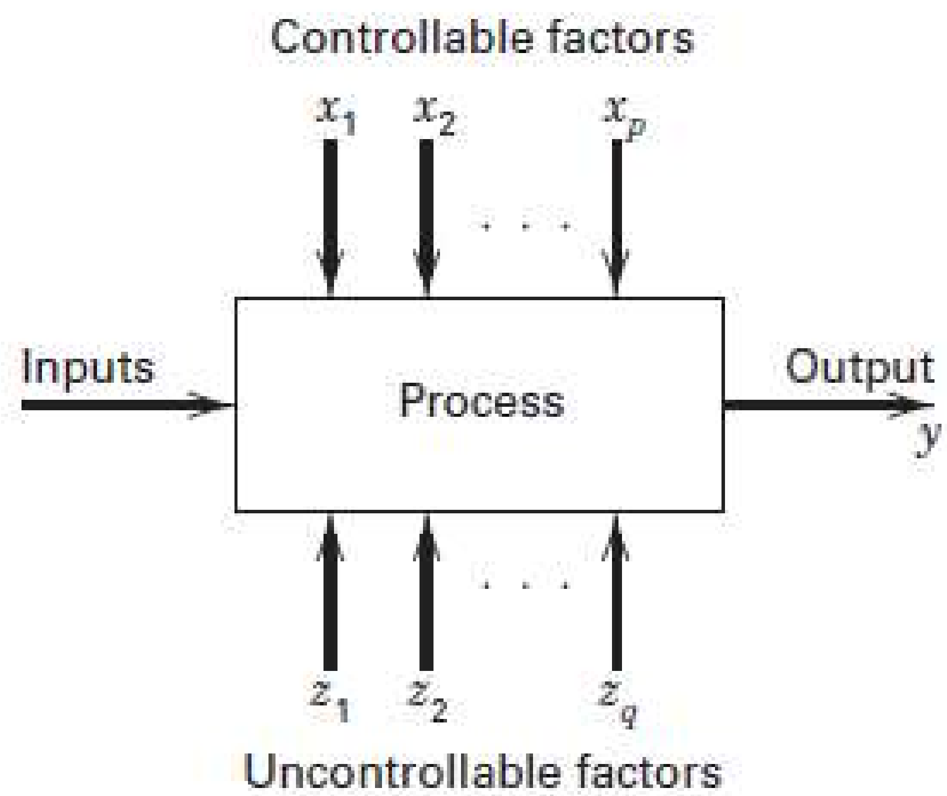
However, most problems in science and engineering require observation of the system at work and **experimentation to elucidate** information about **why and how it works**.

Well-designed experiments can often **lead to** a model of system performance; such experimentally determined models are called **empirical models**.

In general, **experiments are used to study the performance of processes and systems**.

The process or system can be represented by the model shown in figure to the right.

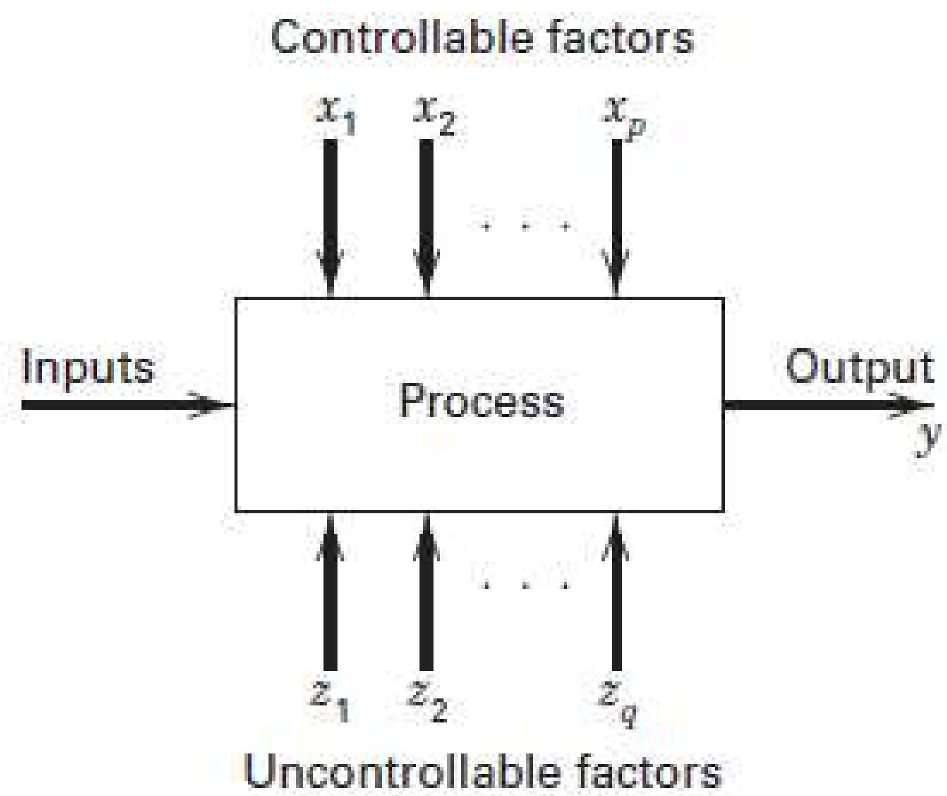
We can usually visualize the process as a combination of operations, machines, methods, people, and other resources that transforms some input (often a material) into an output that has one or more observable response variables.



Some of the process variables and material properties $x_1, x_2, \dots, x_p$ are **controllable**, whereas other variables such as environmental factors or some material properties $z_1, z_2, \dots, z_q$ are **uncontrollable** (although they may be controllable for purposes of a test).

The **objectives of the experiment** may include the following:

- Determining which variables are most influential on the response y
- Determining where to set the influential $x's$ so that y is almost always near the desired nominal value
- Determining where to set the influential $x's$ so that variability in y is small
- Determining where to set the influential $x's$ so that the effects of the uncontrollable variables $z_1, z_2, \ldots, z_q$ are minimized.



As you can see from the foregoing discussion, **experiments** often **involve** several **factors**.

Usually, an objective of the experimenter is to **determine the influence that these factors have on the output response of the system**. The general approach to planning and conducting the experiment is called the **strategy of experimentation**. An experimenter can use several strategies.

I (Douglas) really like to play golf. Unfortunately, I do not enjoy practicing, so I am always looking for a simpler solution to lowering my score.

Some of the **factors** that I think may be important, or **that may influence** my **golf score**, are as follows:

1. The type of driver used (oversized or regular sized)
2. The type of ball used (balata or three piece)
3. Walking and carrying the golf clubs or riding in a golf cart
4. Drinking water or drinking “something else” while playing
5. Playing in the morning or playing in the afternoon
6. Playing when it is cool or playing when it is hot
7. The type of golf shoe spike worn (metal or soft)
8. Playing on a windy day or playing on a calm day.

I (Douglas) really like to play golf. Unfortunately, I do not enjoy practicing, so I am always looking for a simpler solution to lowering my score.

Some of the **factors** that I think may be important, or **that may influence** my **golf score**, are as follows:

1. The type of driver used (oversized or regular sized)
2. The type of ball used (balata or three piece)
3. Walking and carrying the golf clubs or riding in a golf cart
4. Drinking water or drinking "something else" while playing
5. ~~Playing in the morning or playing in the afternoon~~
6. ~~Playing when it is cool or playing when it is hot~~
7. ~~The type of golf shoe spike worn (metal or soft)~~
8. ~~Playing on a windy day or playing on a calm day.~~

their effects are so small that they have no practical value

I (Douglas) really like to play golf. Unfortunately, I do not enjoy practicing, so I am always looking for a simpler solution to lowering my score.

Some of the **factors** that I think may be important, or **that may influence** my **golf score**, are as follows:

1. The type of driver used (oversized or regular sized)
2. The type of ball used (balata or three piece)
3. Walking and carrying the golf clubs or riding in a golf cart
4. Drinking water or drinking “something else” while playing

Factors 1 through 4 could be experimentally tested to determine their effect on my golf score.

Suppose that a maximum of **eight rounds of golf can be played** over the course of the experiment.

One approach would be to select an arbitrary combination of these factors, test them, and see what happens.

1. The type of driver used (oversized or regular sized)
2. The type of ball used (balata or three piece)
3. Walking and carrying the golf clubs or riding in a golf cart
4. Drinking water or drinking “something else” while playing

For example, suppose the **oversized driver**, **balata ball**, **golf cart**, and **water** combination is selected, and the resulting **score is 87**.

During the round, however, I noticed several wayward shots with the big driver, and, as a result, I decide to play another round with the regular-sized driver, holding the other factors at the same levels used previously.

This approach could be continued almost indefinitely, **switching the levels of one or two (or perhaps several) factors for the next test, based on the outcome of the current test**.

This strategy of experimentation, which we call the **best-guess approach**, is frequently used in practice by engineers and scientists.

It often works reasonably well, too, because the experimenters often have a great deal of technical or theoretical knowledge of the system they are studying, as well as considerable practical experience.

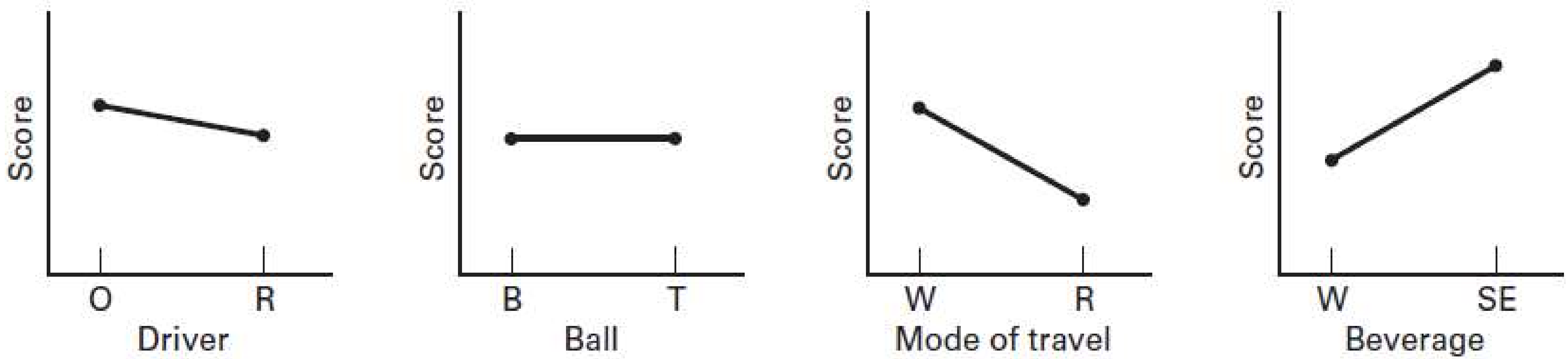
The **best-guess approach** has at least **two disadvantages**.

- Suppose the initial best-guess does not produce the desired results. Now the experimenter has to take another guess at the correct combination of factor levels. This could continue for a **long time, without any guarantee of success**.
- Suppose the initial best-guess produces an acceptable result. Now the experimenter is tempted to **stop testing**, although **there is no guarantee that the best solution has been found**.

Another strategy of experimentation that is used extensively in practice is the **one-factor-at-a-time** (OFAT) approach.

It consists of selecting a starting point, or **baseline set of levels**, for each factor, and then successively **varying each factor** over its range **with the other factors held constant at the baseline level**.

After all tests are performed, a series of graphs are usually constructed showing **how the response variable is affected by varying each factor with all other factors held constant**.

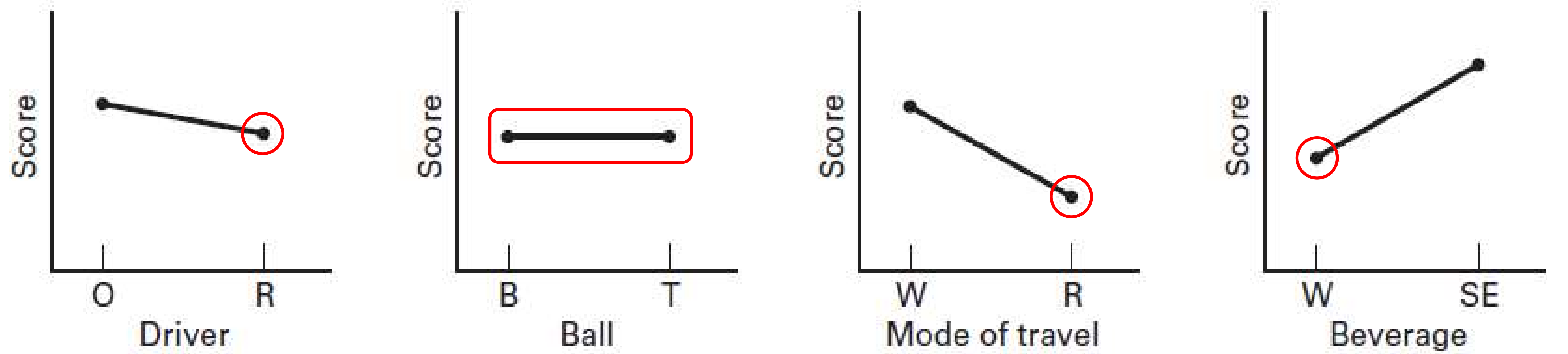


Another strategy of experimentation that is used extensively in practice is the **one-factor-at-a-time** (OFAT) approach.

It consists of selecting a starting point, or **baseline set of levels**, for each factor, and then successively **varying each factor** over its range **with the other factors held constant at the baseline level**.

After all tests are performed, a series of graphs are usually constructed showing **how the response variable is affected by varying each factor with all other factors held constant**.

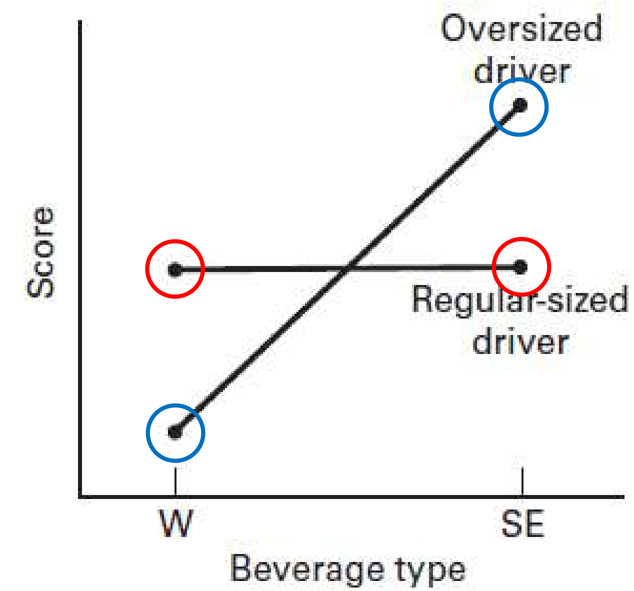
Using these **one-factor-at-a-time graphs**, we would select the **optimal combination** to be the **regular-sized driver (R)**, **riding (R)**, and **drinking water (W)**. The **type of golf ball** seems unimportant.



Disadvantage of the OFAT is that it **fails to consider** any possible **interaction between the factors**.

An **interaction** is the **failure of one factor to produce the same effect on the response at different levels of another factor**.

An interaction between the type of driver and the beverage factors for the golf experiment.



Notice that if I use the **regular-sized driver**, the **type of beverage** consumed **has virtually no effect on the score**, but if I use the **oversized driver**, much **better results** are obtained by **drinking water (W) instead of "something else (SE)."**

Interactions between factors are very common, and if they occur, the **one-factor-at-a-time** strategy will usually **produce poor results**. Many people do not recognize this, and, consequently, OFAT experiments are run frequently in practice. (*Some individuals actually think that this strategy is related to the scientific method*).

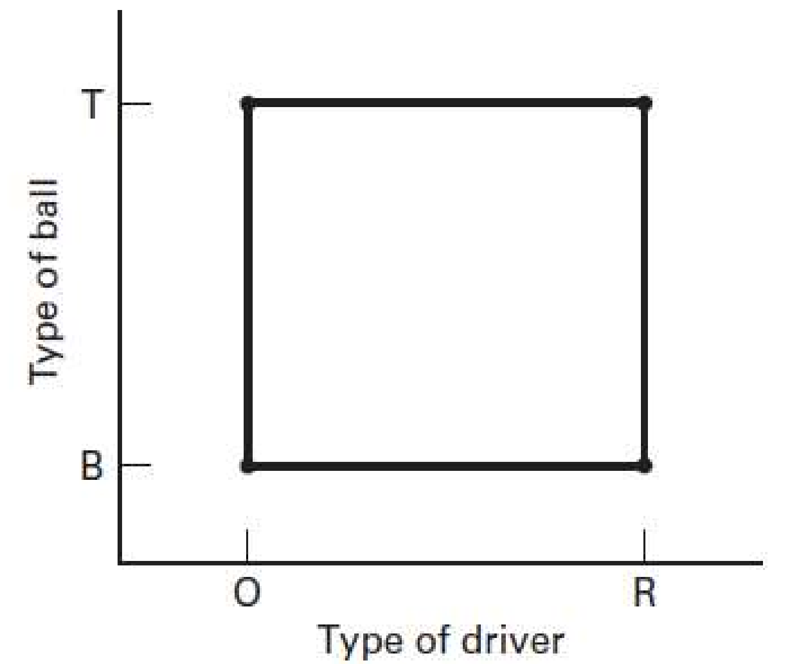
One-factor-at-a-time experiments are always less efficient than other methods based on a statistical approach to design.

The **correct approach to dealing with several factors** is to conduct a **factorial experiment**.

FACTORIAL EXPERIMENT; factors are varied together, instead of one at a time.

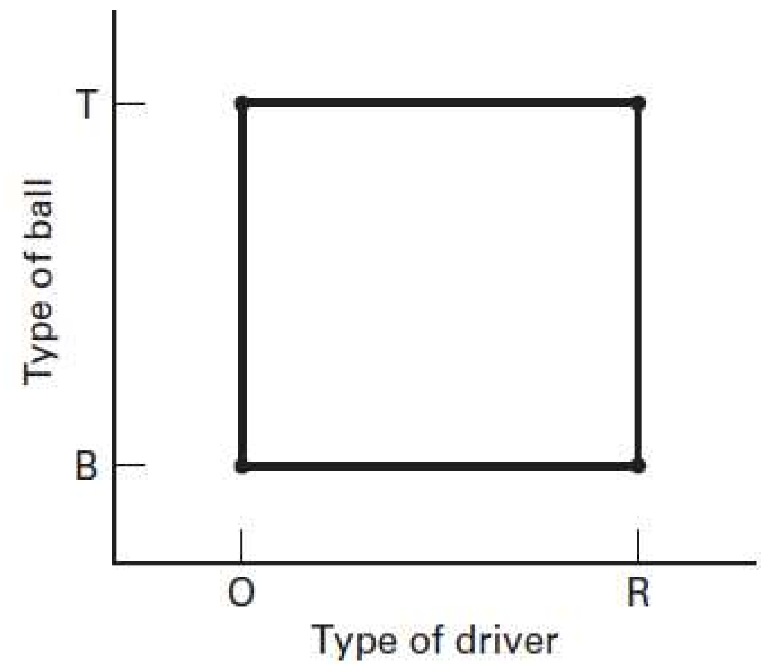
The **factorial experimental design** concept is extremely important.

To illustrate how **a factorial experiment** is conducted, consider the golf experiment and suppose that only **two factors**, **type of driver** and **type of ball**, are of interest. The figure to the right shows a **two-factor factorial experiment** for studying the joint effects of these two factors on my golf score.



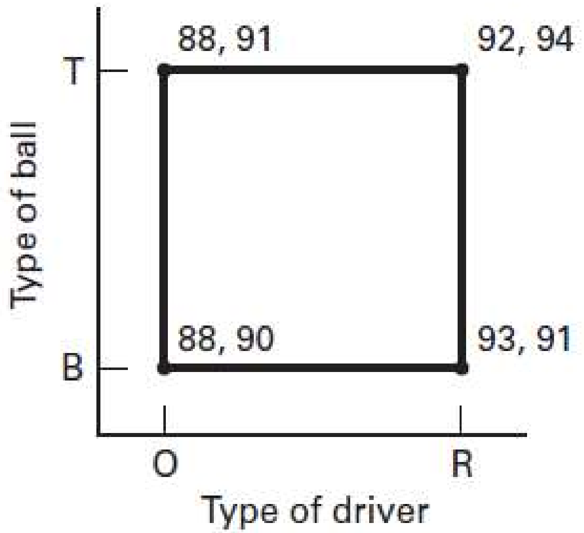
Notice that this factorial experiment has both factors at two levels and that all possible combinations of the two factors across their levels are used in the design.

Geometrically, the four runs form the corners of a **square**. This particular type of factorial experiment is called **a 2^2 factorial design** (two factors, each at two levels).

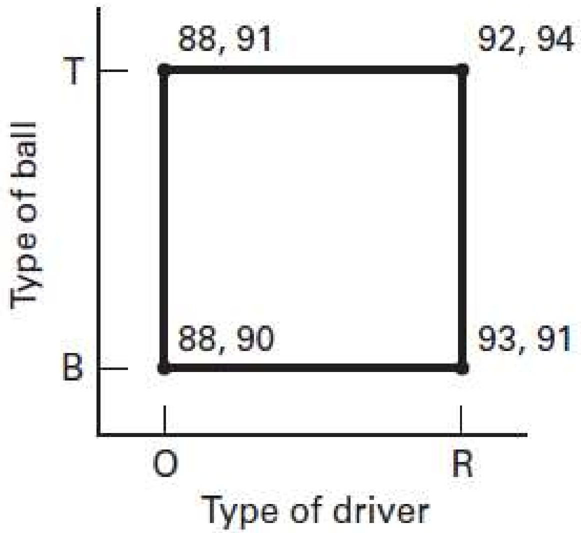


Because I can reasonably expect to **play eight rounds of golf** to investigate these factors, a reasonable plan would be to **play two rounds of golf at each combination of factor levels** shown in the figure.

An experimental designer would say that **we have replicated the design twice**. This experimental design would enable the experimenter to **investigate the individual effects of each factor** (or the main effects) **and to determine whether the factors interact**.

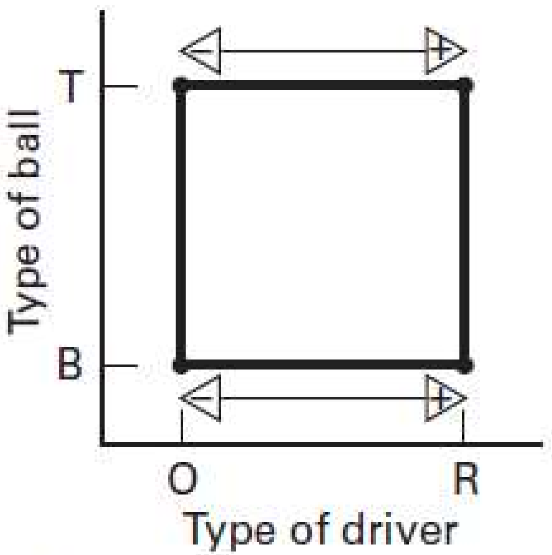


Notice that there are four rounds of golf that provide information about using the regular-sized driver (R) and four rounds that provide information about using the oversized driver (O).

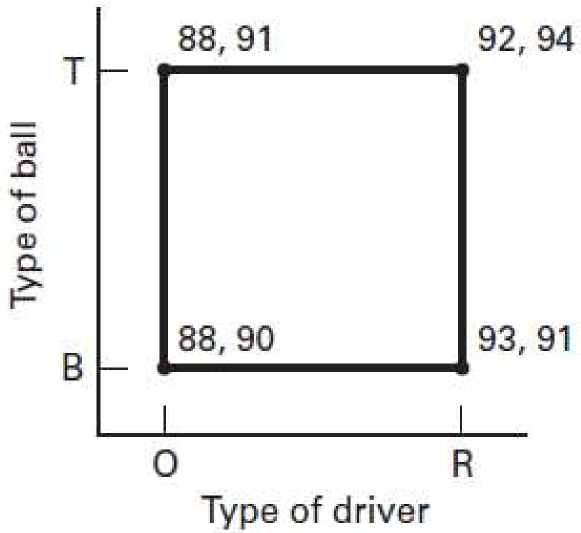


By finding the average difference in the scores on the right- and left-hand sides of the square we have a measure of the **effect of switching from the oversized driver to the regular-sized driver**

$$\text{Driver effect} = \frac{92 + 94 + 93 + 91}{4} - \frac{88 + 91 + 88 + 90}{4} = 3.25$$



(*b*) Comparison of scores leading to the driver effect

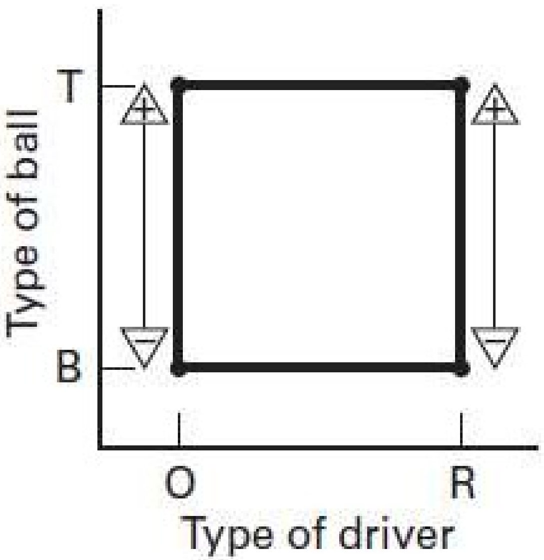


By finding the average difference in the scores on the right- and left-hand sides of the square we have a measure of the **effect of switching from the oversized driver to the regular-sized driver**

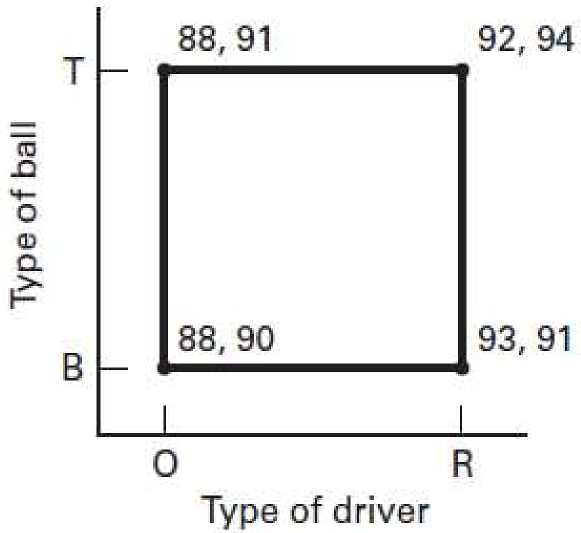
$$\text{Driver effect} = \frac{92+94+93+91}{4} - \frac{88+91+88+90}{4} = 3.25$$

By finding the average difference in the four scores at the top of the square and the four scores at the bottom measures the **effect of the type of ball** used

$$\text{Ball effect} = \frac{88+91+92+94}{4} - \frac{88+90+93+91}{4} = 0.75$$



(c) Comparison of scores leading to the ball effect

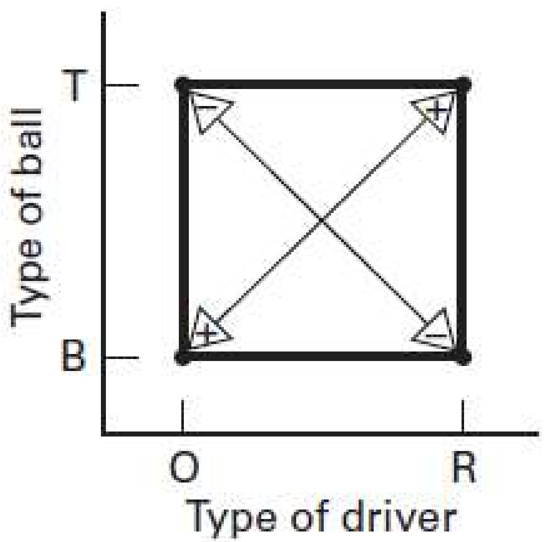


By finding the average difference in the scores on the right- and left-hand sides of the square we have a measure of the **effect of switching from the oversized driver to the regular-sized driver**

$$\text{Driver effect} = \frac{92+94+93+91}{4} - \frac{88+91+88+90}{4} = 3.25$$

By finding the average difference in the four scores at the top of the square and the four scores at the bottom measures the **effect of the type of ball** used

$$\text{Ball effect} = \frac{88+91+92+94}{4} - \frac{88+90+93+91}{4} = 0.75$$



(*d*) Comparison of scores leading to the ball–driver interaction effect

A measure of the **interaction effect between the type of ball and the type of driver** can be obtained by subtracting the average scores on the left-to-right diagonal in the square from the average scores on the right-to-left diagonal

$$\text{Ball–driver interaction effect} = \frac{92+94+88+90}{4} - \frac{88+91+93+91}{4} = 0.25$$

The results of this factorial experiment indicate that **driver effect is larger than either the ball effect or the interaction**.

Statistical testing could be used to determine whether any of these effects differ from zero.

$$\text{Driver effect} = 3.25$$
$$\text{Ball effect} = 0.75$$
$$\text{Ball–driver interaction effect} = 0.25$$

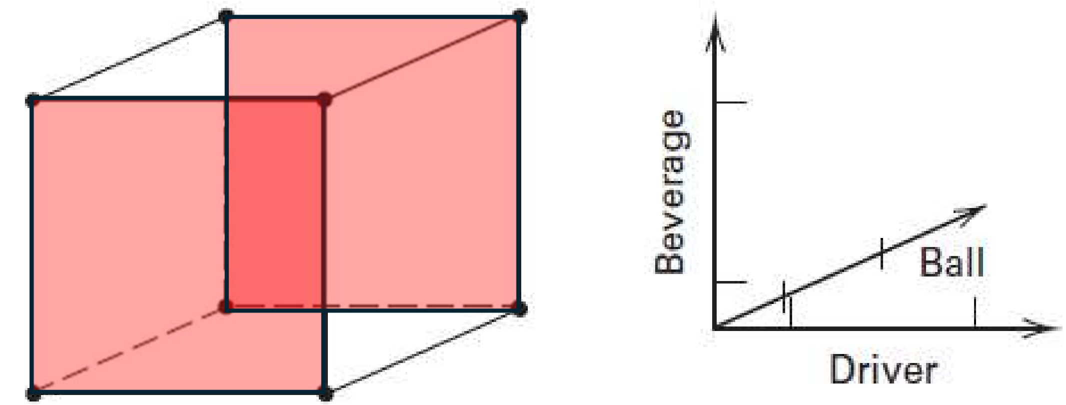
In fact, it turns out that there is reasonably strong statistical evidence that the driver effect differs from zero and the other two effects do not. Therefore, **this experiment indicates that I should always play with the oversized driver**.

One very important feature of the factorial experiment is evident from this simple example; namely, **factorials make the most efficient use of the experimental data**. Notice that this experiment included eight observations, and **all eight observations are used to calculate the driver, ball, and interaction effects**. No other strategy of experimentation makes such an efficient use of the data. This is an **important and useful feature of factorials**.

We can extend the factorial experiment concept to **three factors**.

Suppose that I wish to study the effects of **type of driver**, **type of ball**, and the **type of beverage** consumed on my golf score.

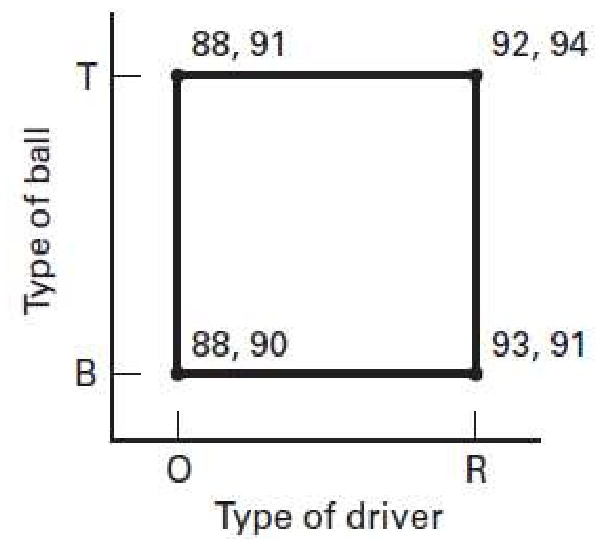
Assuming that **all three factors have two levels**, a **factorial design** can be set up as shown in figure.



This is an example of a 2^3 **factorial design**. Because I only want to play eight rounds of golf, this experiment would require that one round be played at each combination of factors represented by the eight corners of the cube.

The 2^3 **factorial design** provides the same information about the factor effects.

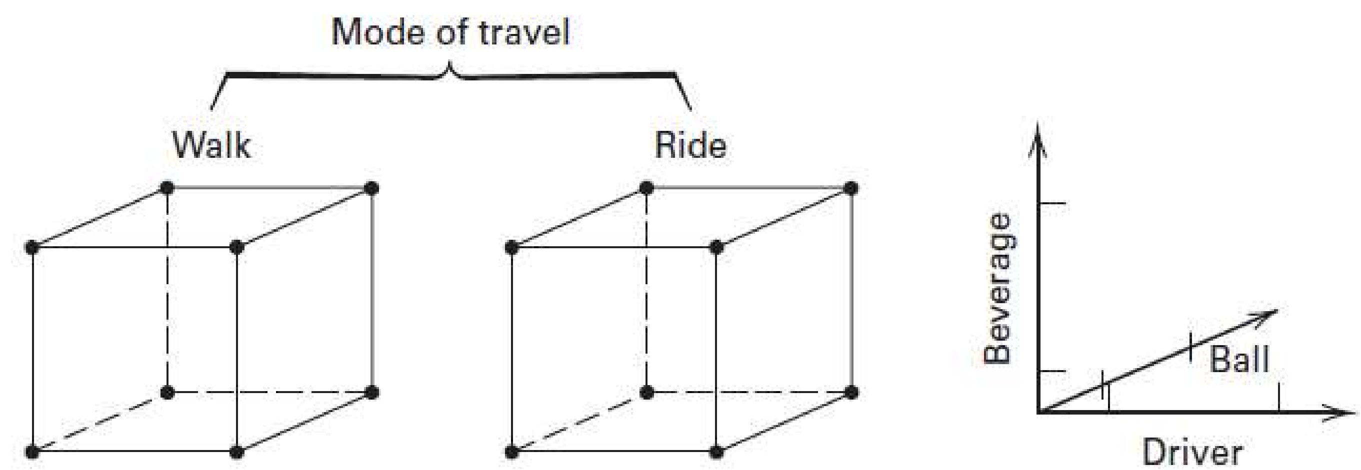
For example, there are four tests in both designs that provide information about **type of ball (B)** and four tests that provide information about **type of ball (T)**, assuming that each run in the 2^2 **factorial design** is replicated twice.



The figure at the right illustrates how all four factors—driver, ball, beverage, and mode of travel (walking or riding)—could be investigated in a 2^3 **factorial design**.

As **in any factorial design, all possible combinations of the levels of the factors are used**.

Because all four factors are at two levels, this experimental design can still be represented geometrically as a cube (actually a hypercube).



Generally, if there are k factors, each at two levels, the factorial design would require 2^k runs.

For example, the experiment in the figure above requires 16 runs.

Clearly, as the number of factors of interest increases, the number of runs required increases rapidly; for instance, a **10-factor experiment with all factors at two levels would require 1024 runs**.

This quickly becomes infeasible from a time and resource viewpoint.

In the golf experiment, I can only play eight rounds of golf, so even the experiment in the figure above is too large.

Fortunately, if there are four to five or more factors, it is usually unnecessary to run all possible combinations of factor levels.

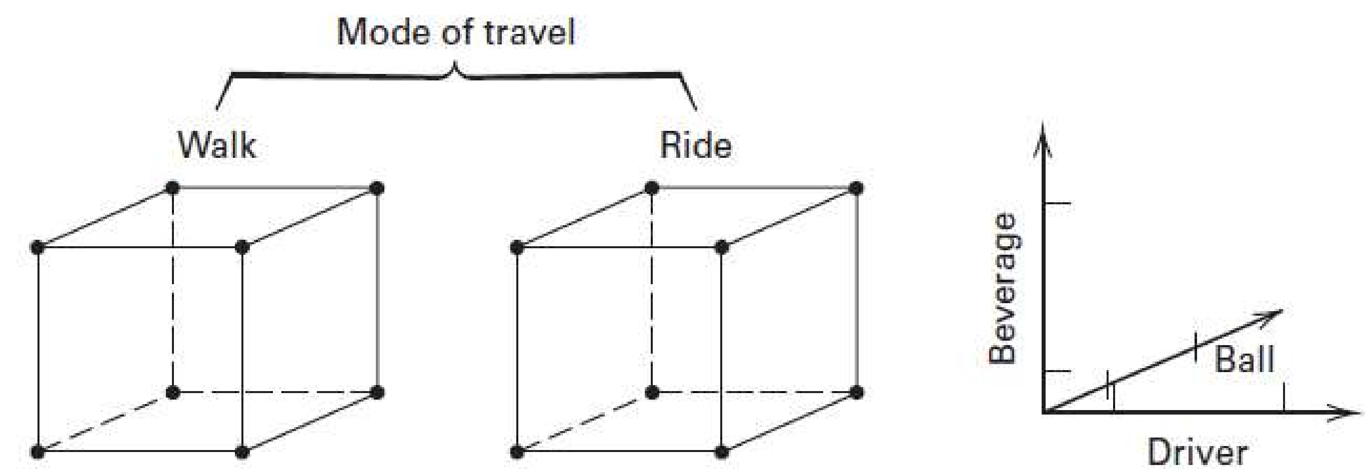
A **fractional factorial experiment** is a variation of the basic factorial design in which **only a subset of the runs is used**.

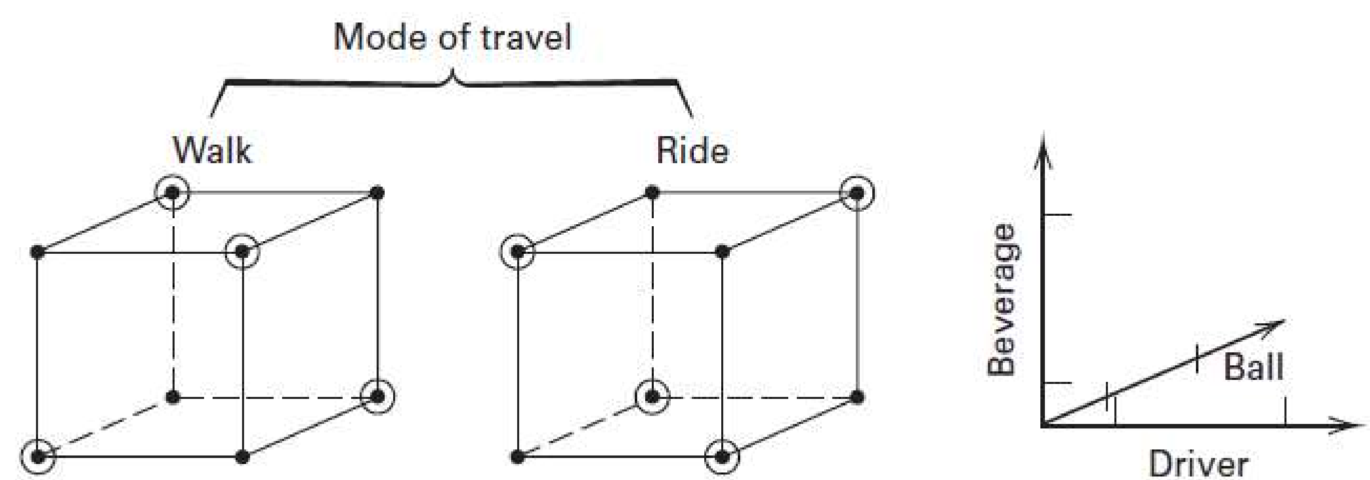
The figure to the low right shows a **fractional factorial design for the four-factor** version of the golf experiment.

This design requires **only 8 runs instead of the original 16** and would be called a **one-half fraction**.

If I can play only eight rounds of golf, this is an excellent design in which to study all four factors.

It will provide good information about the main effects of the four factors as well as some information about how these factors interact.





DESIGNING A WEB PAGE

A lot of business today is conducted via the World Wide Web.

Consequently, the design of a business' web page has potentially important economic impact.

Suppose that the website has the following components:

(1) a photoflash image,

(2) a main headline,

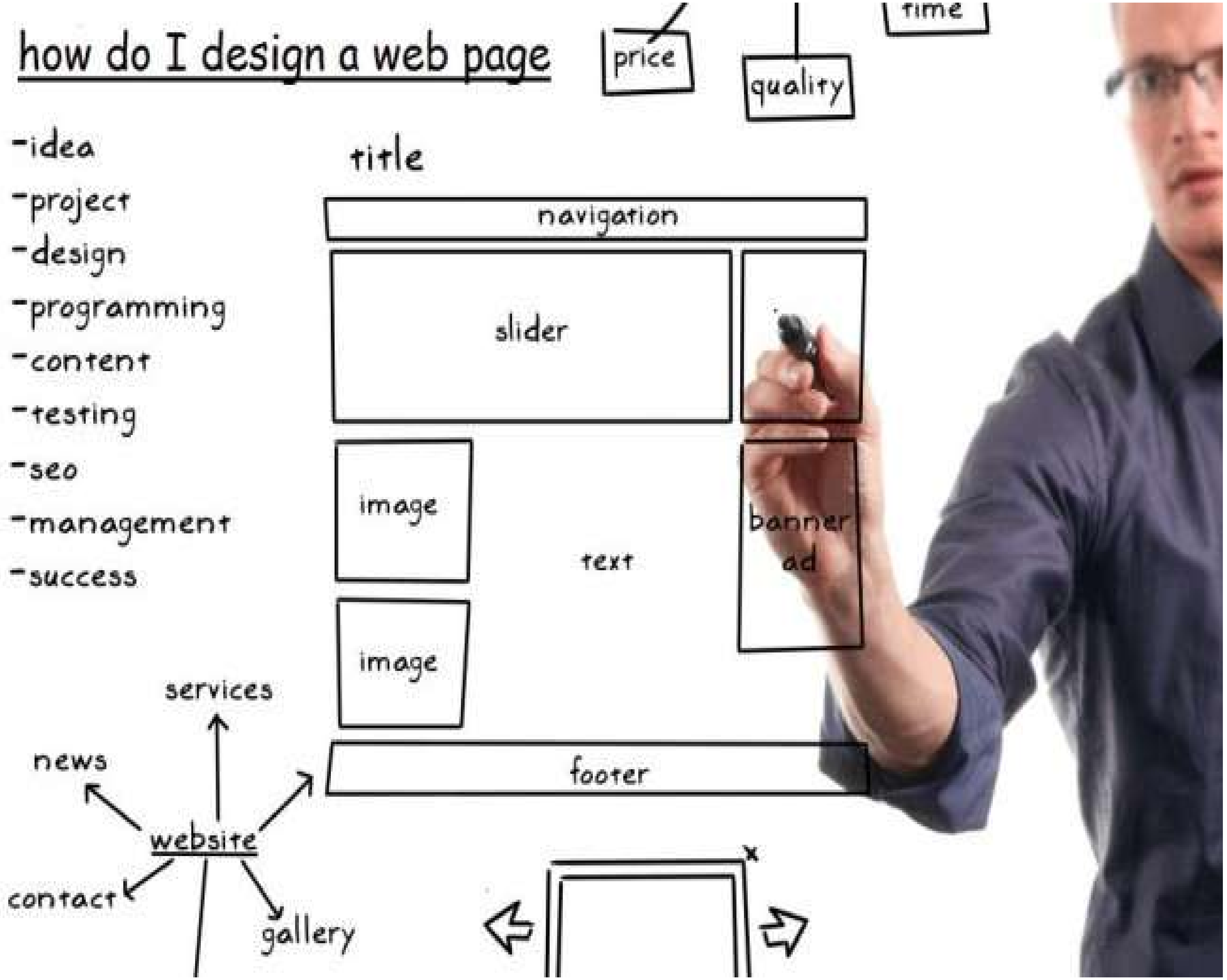
(3) a subheadline,

(4) a main text copy,

(5) a main image on the right side,

(6) a background design,

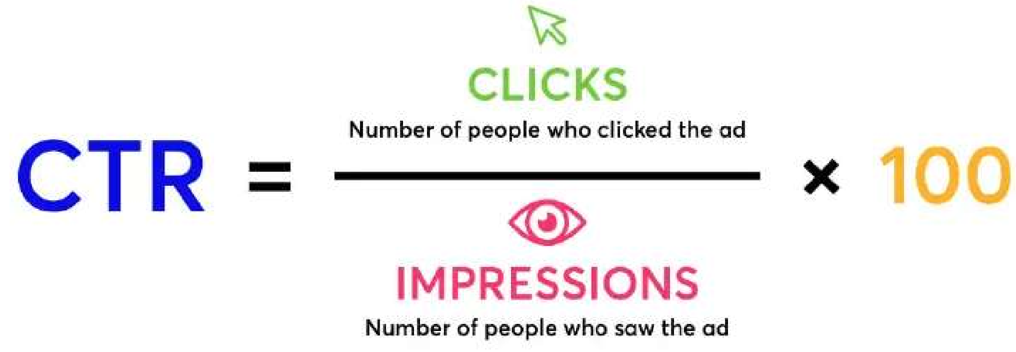
(7) a footer.



DESIGNING A WEB PAGE

We are interested in **finding the factors that influence** the **click-through rate**; that is, the *number of visitors who click through into the site divided by the total number of visitors to the site*.

Proper selection of the important factors can lead to an optimal web page design.



Number of levels for each factor:	#levels
(1) a photoflash image,	4
(2) a main headline,	8
(3) a subheadline,	6
(4) a main text copy,	5
(5) a main image on the right side,	4
(6) a background design,	3
(7) a footer.	7

If we use a **factorial design**, web pages for *all possible combinations* of these factor levels must be constructed and tested.
Total of 4 × 8 × 6 × 5 × 4 × 3 × 7 = **80,640 web pages**.

Obviously, it is not feasible to design and test this many combinations of web pages, so a **complete factorial experiment cannot be considered**.

However, a **fractional factorial experiment** that uses a small number of the possible web page designs would likely be successful. (fractional factorial where the factors have different numbers of levels.)

Statistical design of experiments; process of planning the experiment so that appropriate data will be collected and analyzed by statistical methods, resulting in valid and objective conclusions.

Two aspects to any experimental problem:

- the **design of the experiment**, and
- the **statistical analysis of the data**.

The three basic principles of experimental design are

- **Randomization;** allocation of the experimental material and the order in which the individual runs of the experiment are to be performed are randomly determined. By properly *randomizing the experiment*, we also assist in "*averaging out*" the *effects of extraneous factors* that may be present.
- **Replication;** independent repeat run of each factor combination. Two important properties. *i)* it allows the experimenter to obtain an *estimate of the experimental error*. This estimate of error becomes a basic unit of measurement for determining whether observed differences in the data are really statistically different. *ii)* if the sample mean ($\bar{y}$) is used to estimate the true mean response for one of the factor levels in the experiment, replication permits the experimenter to obtain a more precise estimate of this parameter
- **Blocking;** design technique used to improve the precision with which comparisons among the factors of interest are made. Often used to *reduce or eliminate the variability transmitted from nuisance factors*—that is, factors that may influence the experimental response but in which we are not directly interested.

Guidelines for Designing an Experiment

1. Recognition of and statement of the problem 2. Selection of the response variable[a]	Pre-experimental Planning
3. Choice of factors, levels, and ranges[a]	
4. Choice of experimental design	
5. Performing the experiment	
6. Statistical analysis of the data	
7. Conclusions and recommendations	

[a] In practice, steps 2 and 3 are often done simultaneously or in reverse order.

Guidelines for Designing an Experiment

1. Recognition of and statement of the problem	Pre-experimental
2. Selection of the response variable[a]	Planning
3. Choice of factors, levels, and ranges[a]	
4. Choice of experimental design	
5. Performing the experiment	
6. Statistical analysis of the data	
7. Conclusions and recommendations	

[a] In practice, steps 2 and 3 are often done simultaneously or in reverse order.

- Prepare a **list of specific problems or questions that are to be addressed** by the experiment.
- Keep the overall objectives of the experiment in mind.
- Each type of experiment will generate its own list of **specific questions** that need to be addressed.

  Factor screening or characterization. *Optimization*, *Confirmation*, *Discovery*, *Robustness*, …

Guidelines for Designing an Experiment

1. Recognition of and statement of the problem	Pre-experimental
2. Selection of the response variable[a]	Planning
3. Choice of factors, levels, and ranges[a]	
4. Choice of experimental design	
5. Performing the experiment	
6. Statistical analysis of the data	
7. Conclusions and recommendations	

[a] In practice, steps 2 and 3 are often done simultaneously or in reverse order.

- Be certain that the **response variable** really provides useful information about the process under study. Most often, the **average or standard deviation** (or both) of the measured characteristic **will be the response variable**.
- Multiple responses are not unusual.
- Decide **how each response will be measured**, and address issues such as **how will any measurement system be calibrated** and **how this calibration will be maintained during the experiment**.

Guidelines for Designing an Experiment

1. Recognition of and statement of the problem	Pre-experimental
2. Selection of the response variable[a]	Planning
3. Choice of factors, levels, and ranges[a]	
4. Choice of experimental design	
5. Performing the experiment	
6. Statistical analysis of the data	
7. Conclusions and recommendations	

[a] In practice, steps 2 and 3 are often done simultaneously or in reverse order.

- When considering the **factors that may influence the performance** of a process or system, the experimenter usually discovers that these factors can be classified **as either potential design factors or nuisance factors**.
  - **Potential design factors;** they are classified as *design factors*, *held-constant factors*, and *allowed-to-vary*
  - **Nuisance factors;** they are classified as *controllable, uncontrollable, or noise factors.*
- **Range selection;** choose the ranges over which design factors will be varied and the specific levels at which runs will be made.

Guidelines for Designing an Experiment

1. Recognition of and statement of the problem	Pre-experimental
2. Selection of the response variable[a]	Planning
3. Choice of factors, levels, and ranges[a]	
4. Choice of experimental design	
5. Performing the experiment	
6. Statistical analysis of the data	
7. Conclusions and recommendations	

[a] In practice, steps 2 and 3 are often done simultaneously or in reverse order.

- Consideration of **sample size** (number of replicates), selection of a suitable **run order for the experimental trials**, and determination of whether or not **blocking** or other **randomization** restrictions are involved.
- Thinking about and **selecting a tentative empirical model** to describe the results

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \varepsilon$$

Guidelines for Designing an Experiment

1. Recognition of and statement of the problem	Pre-experimental
2. Selection of the response variable[a]	Planning
3. Choice of factors, levels, and ranges[a]	
4. Choice of experimental design	
5. Performing the experiment	
6. Statistical analysis of the data	
7. Conclusions and recommendations	

[a]In practice, steps 2 and 3 are often done simultaneously or in reverse order.

- Monitor the process carefully to ensure that everything is being done according to plan.
- The people conducting the experiment failed to set the variables to the proper levels on some runs.
- Someone should be assigned to check factor settings before each run.
- It is easy to underestimate the logistical and planning aspects of running a designed experiment in a complex manufacturing or research and development environment.

Guidelines for Designing an Experiment

1. Recognition of and statement of the problem	Pre-experimental
2. Selection of the response variable[a]	Planning
3. Choice of factors, levels, and ranges[a]	
4. Choice of experimental design	
5. Performing the experiment	
6. Statistical analysis of the data	
7. Conclusions and recommendations	

[a] In practice, steps 2 and 3 are often done simultaneously or in reverse order.

- **Statistical methods** should be used to analyze the data so that results and conclusions are objective rather than judgmental in nature.
- Because many of the questions that the experimenter wants to answer can be cast into an hypothesis-testing framework, **hypothesis testing and confidence interval estimation** procedures are very useful in analyzing data from a designed experiment.
- **Residual analysis and model adequacy checking** are also important analysis techniques.

Guidelines for Designing an Experiment

1. Recognition of and statement of the problem	Pre-experimental
2. Selection of the response variable[a]	Planning
3. Choice of factors, levels, and ranges[a]	
4. Choice of experimental design	
5. Performing the experiment	
6. Statistical analysis of the data	
7. Conclusions and recommendations	

[a] In practice, steps 2 and 3 are often done simultaneously or in reverse order.

- **Follow-up runs and confirmation testing** should also be performed to validate the conclusions from the experiment.
- **Experimentation is iterative**. It is usually a major mistake to design a single, large, comprehensive experiment at the start of a study.
- **We usually experiment sequentially**, and as a general rule, no more than about 25 percent of the available resources should be invested in the first experiment.

- **Use your nonstatistical knowledge of the problem**. Experimenters are usually highly knowledgeable in their fields. In some fields, there is a large body of physical theory on which to draw in explaining relationships between factors and responses. *Using a designed experiment is no substitute for thinking about the problem*.
- **Keep the design and analysis as simple as possible**. Don't be overzealous in the use of complex, sophisticated statistical techniques. Relatively simple design and analysis methods are almost always best.
- **Recognize the difference between practical and statistical significance**. Just because two experimental conditions produce mean responses that are statistically different, there is no assurance that this difference is large enough to have any practical value.
- **Experiments are usually iterative**. Remember that in most situations it is unwise to design too comprehensive an experiment at the start of a study. Successful design requires the knowledge of important factors, the ranges over which these factors are varied, the appropriate number of levels for each factor, and the proper methods and units of measurement for each factor and response. Generally, we are not well equipped to answer these questions at the beginning of the experiment, but we learn the answers as we go along.