The steps of the process of learning and making inferences from data can be summarized as:

- Designing the data collection process
- Preparing the data for analysis
- Describe the information graphically and numerically
- Build statistical models for inference
- Test the validity of the models
- Report conclusions

In this class we’ll discuss experimental design issues and strategies.

Collecting data

In this class we will give definitions for randomized controlled experiments and observational studies. We will compare the two types of studies using some famous case studies. We will discuss how a causal effect can be established and how confounding factors can be the hidden cause of association between two variables. We will also consider the problem of reporting percentages for the different groups in a study.

Longitudinal versus cross-sectional studies

Suppose the university conducts a survey of all students, faculty and staff on campus and records the height, age and weight of each person.

Such a study is a cross section of the campus population. It gives a picture of the characteristics of the population at a given time.

Suppose you group your sample in three age groups: below 30, between 30 and 50 and above 50. And, you observe that

- The average weight in the first group is smaller than in the second.
- The average weight in the second group is smaller than the third group.

Can you conclude that age is responsible for an increase in weight?

No, you can not conclude that age is responsible for an increase in weight. This is because the effect may be confounded with the fact that eating habits may have changed during the last decades and this may have an effect in the average weight of the population.

Warning: You can not draw conclusions about the effects of age from such a study since you are comparing different people of (possibly) different ages.

To draw conclusions about the effect of age you need to conduct a longitudinal study. That is, you follow the evolution of a person’s weight in time, for each person in the sample.
Observational studies

In a controlled experiment the researcher decides who gets assigned to which group. But there are many situations in which the researcher can just watch what happens.

In an observational study the subjects assign themselves to the treatment and control groups

Studies related to accidents or smoking are examples of observational studies. People are not usually willing to be randomized to smoke or have an accident just to participate in a study.

Observational studies can find evidence of association between a treatment variable and an outcome (response) variable. For example between smoking and lung cancer.

But there may be hidden factors that make people smoke and also make them get sick.

Association is not causation

To reduce the effect of confounding factors in observational studies we have to make the control and the treatment groups as similar as possible, that is, we have to control for confounding factors.

In the case of smoking age and gender can be confounding factors. So the right thing to do is to compare subjects of the same age and gender who smoke and do not smoke.

Collecting data: experimental design

Suppose a new drug is introduced, how do we gather evidence that it is effective to treat a given disease?

The key idea is comparison.

A group of patients suffering from the disease is divided into two groups, a treatment group where patients get the drug, and a control group of patients that are not treated.

To eliminate bias, subjects are assigned to each group at random and the experiment is run double blind. That is, neither the patients nor the doctors know who is in the control and who is in the treatment.

This is called a Controlled Randomized Experiment and can establish a causal effect of the treatment on the response.

Case Study: The Salk vaccine

Polio was an epidemic in the US during forty years that started in 1916. In the ’50s several vaccines were developed and the most successful one was the Salk vaccine. When it was ready to be tried on humans a field trial was conducted.

Why not give the vaccine to every child? Because even after laboratory experimentation the effectiveness and the risks associated with the vaccine were unclear. Also, giving the vaccine to every child in a given year could confound the effect of the vaccine with the cycles of the epidemic.

Why not give the vaccine only to children of consenting parents? Because the incidence of polio was higher among higher-income families which were more likely to give their consent. So the effect of the vaccine would have been confounded with that of social class.

All children in the study had to receive an injection. Children
in the control group were given salty water to avoid confounding with the psychological effect of receiving a shot. This is called a **placebo**.

**Who decided which children were going to be under treatment?** This was decided at random, to avoid biases due to human judgment.

**Doctors in charge of diagnosing children were not told if they were vaccinated or not.** This, again, was done to avoid biases, since doctors may have had a preconception about the validity of the vaccine.

A different field study was conducted, giving the vaccine to all second graders whose parents consented and leaving all first and third graders unvaccinated in areas with high risk of polio.

Notice that this study not only had the problem of consenting parents being confounded with social class it also had the problem that children were grouped in different grades and, since polio is a contagious disease, this can have an effect on the way it spreads.

| controlled randomized study on 1st, 2nd and 3rd graders |
| Sample Size | Polio? | Sample Size | Polio? |
| Treatment   | 200,000 | 56,000       | Grade 2 | 225,000 | 56,250 |
| Control     | 200,000 | **142,000**  | Grades 1 & 3 | 725,000 | **391,500** |
| No consent  | 350,000 | 161,000      | Grade 2 | 125,000 | 55,000 |

Notice that different groups have different sample sizes. It doesn’t make sense to compare the **numbers** of children who contracted polio.

What’s a better way to compare the numbers?

We can instead compare **rates** (number of polio cases per 100,000 in each group).

| controlled randomized experiment | study on 1st, 2nd and 3rd graders |
| Sample Size | Rate | Sample Size | Rate |
| Treatment   | 200,000 | 28 | Grade 2 | 225,000 | 25 |
| Control     | 200,000 | **71** | Grades 1 & 3 | 725,000 | **54** |
| No consent  | 350,000 | 46 | Grade 2 | 125,000 | 44 |

Notice that in the **randomized controlled double blind (RCDB)** experiment the rate drops from 71 to 28. This is a much higher drop than the 54 to 25 shown in the other case.

Since the other study had treatment and control groups which were **not comparable** it resulted in a bias against the vaccine.

The key of a good design is that both groups have to be as similar as possible.
Historical controls

Since randomized controlled trials are hard and expensive, sometimes other designs are used to assess the validity of a treatment. One possibility is to compare the treatment with historical data. That is, patients who were treated in an old way in the past compared with patients treated with a new drug or surgery.

The problem with this approach is that treatment and control may differ in important ways.

A controlled experiment begins with a well defined patient population. The first step is to decide which patients are eligible. Among these patients a control and a treatment group is chosen at random and contemporaneously.

If, for example all patients who are too sick to undergo surgery are treated as control, then the trial is biased towards the surgery.

The Coronary Drug Project was a randomized, controlled, double-blind experiment to compare five drugs for the prevention of heart attacks. Of the 8,341 middle age men with heart problems, 5,552 were assigned to one of the five treatments and 2,789 to the control group. The patients were followed for 5 years.

The group on Clofibrate, a drug that reduces the level of cholesterol in blood, did not do very well. The death rate was 20% compared to 21% in the control group. It was suggested that this was due to the fact that many subjects did not take their medicine. The following table reports the death rates of subjects who took and did not take the medicine.

<table>
<thead>
<tr>
<th></th>
<th>Clofibrate</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>taking number</td>
<td>708</td>
<td>1,813</td>
</tr>
<tr>
<td>deaths</td>
<td>15%</td>
<td>15%</td>
</tr>
<tr>
<td>not taking</td>
<td>357</td>
<td>882</td>
</tr>
<tr>
<td>deaths</td>
<td>25%</td>
<td>28%</td>
</tr>
<tr>
<td>total number</td>
<td>1,103</td>
<td>2,789</td>
</tr>
<tr>
<td>deaths</td>
<td>20%</td>
<td>21%</td>
</tr>
</tbody>
</table>

Comparing subjects that took the medicine with those that did not is an observational study.

This is because subjects assigned themselves to one of the ‘treatments’. Among the subjects under Clofibrate there is a drop from 25% to 15% when the subjects not taking the medicine are compared to those who really took it. This seems as strong evidence that Clofibrate works. Is this correct?

In fact, we see the same drop in deaths among subjects under placebo. Is there a confounding factor present? If so, what could it be?

The confounding factor in this observational study could be the level of health consciousness that made some subjects be more willing to follow the prescription than others.

The conclusions are:

1. Clofibrate does not have an effect.
2. The subjects that stick to the prescription are different from the ones that don’t.

Lifestyle or health consciousness may be the hidden factor that make people take their prescription regularly and also lowers their mortality rate.

If a variable is a confounding factor, it must affect BOTH the treatment and control variables.
Cervical cancer and circumcision

**Fact:** In the ’50s cervical cancer was found to be fairly rare among Jews in different countries. A similar pattern was observed among Muslim women. As a result of these observations, several researchers concluded that male circumcision was the protective factor.

**But:** There are differences between Jews or Muslims and members of other communities besides circumcision, so there are many confounding factors to consider.

It turns out that cervical cancer has a causal agent in the human papilloma virus, which is a sexually transmitted disease. So women who are more sexually active are more exposed to the virus and thus possibly more likely to develop cervical cancer.

So the confounding factor was probably sexual behavior in the ’30s and ’40s.

---

**Simpson’s Paradox**

The following table shows the admission rates of the six largest majors in UC Berkeley.

<table>
<thead>
<tr>
<th>Major</th>
<th>Men applicants</th>
<th>Men percent</th>
<th>Women applicants</th>
<th>Women percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>825</td>
<td>62</td>
<td>108</td>
<td>82</td>
</tr>
<tr>
<td>B</td>
<td>560</td>
<td>63</td>
<td>25</td>
<td>68</td>
</tr>
<tr>
<td>C</td>
<td>325</td>
<td>37</td>
<td>593</td>
<td>34</td>
</tr>
<tr>
<td>D</td>
<td>417</td>
<td>33</td>
<td>375</td>
<td>35</td>
</tr>
<tr>
<td>E</td>
<td>191</td>
<td>28</td>
<td>393</td>
<td>24</td>
</tr>
<tr>
<td>F</td>
<td>373</td>
<td>6</td>
<td>341</td>
<td>7</td>
</tr>
<tr>
<td>totals</td>
<td>2,691</td>
<td><strong>44</strong></td>
<td>1,835</td>
<td><strong>30</strong></td>
</tr>
</tbody>
</table>

The totals suggest that there is strong sex bias in the admission system with 30% of women against 44% of men.

When we look at the percents major by major we observe that they are pretty comparable, actually in some cases, men have a substantially lower percent, like in major A.

**What is going on?**

Let’s have a look at the table showing the percent of women and men applying for each major.
Notice that women apply to the majors that have lower acceptance rates, whilst men apply to the ‘easy’ ones. This effect is confounded with gender.

**Simpson’s Paradox** can be stated as:

> Relationships between percentages in subgroups can be reversed when the subgroups are combined

Tables with many entries are hard to read. Can we produce a number that represents the information contained in the acceptance ratios table?

If instead of calculating the total percent of male and female applicants we consider an average of the ratios per major we could get such a number. The trouble is that we have to give a different weight to each major, corresponding to the number of applicants to that major.

<table>
<thead>
<tr>
<th>major</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>applicants</td>
<td>933</td>
<td>585</td>
<td>918</td>
<td>792</td>
<td>584</td>
<td>714</td>
</tr>
</tbody>
</table>

The weighted average admission rates are

**Men:**

\[
\frac{.62 \times 933 + .63 \times 585 + .37 \times 918 + .33 \times 792 + .28 \times 584 + .06 \times 714}{4,562}
\]

**Women:**

\[
\frac{.82 \times 933 + .68 \times 585 + .34 \times 918 + .35 \times 792 + .24 \times 584 + .07 \times 714}{4,562}
\]

The results are 39% for the men and 43% for the women.

Suggesting a completely different conclusion than the one based on the total admission rate.

**University Problem**

A university has two departments, A and B. There are 2,000 in-state applicants, of whom half apply to each department. There are 1,100 out-of-state applicants: 100 apply to department A and 1,000 to department B. Department A admits 60% of the in-state and 60% of the out-of-state who apply. Department B admits 30% of the in-state and 30% of the out-of-state who apply. “Since for each department the % of in-state admitted is equal to the % of out-of-state, the same must be true for the two departments together”. Answer yes or no, and explain briefly.
Oral Contraceptives Problem

According to a study, users of oral contraceptives have higher risks of cervical cancer than non-users, even after adjusting for age, education and marital status. Investigators conclude that the pill causes cervical cancer

- Is this a controlled experiment or an observational study?
- Why did the investigators adjust for age, education and marital status?
- Women on the pill were likely to differ from non-users on another factor which affects cervical cancer. What factor is that?
- Were the conclusions justified by the data? Answer yes or no, and explain briefly.