Principles of Statistics

Xuming He
xmhe@umich.edu

1 What can we do with statistics?

We use **statistics** to

- summarize and describe data in the way that helps us make decisions
- use past data and other information to forecast
- draw inference – hypotheses testing;

  Examples: comparison of treatment with control.
- interpret results of statistical analysis;

Example: Correlation or causation?

Correlation describes the strength of statistical association between two random variables or measurements. Normally it refers to **Pearson's correlation coefficient,** one of the most common methods to measure linear association.

Types of correlation: negative, positive and no correlations.

Example of positive correlation: highest temperature of a given day and sales of ice cream on that day;

Example of negative correlation: HIV infection rate and condom use for a given population (e.g., sex workers in Thailand)

Correlation is not causation!

Example: there is a positive correlation between how high a bomber (airplane) is flying at the time of attack and the precision of the bombs. This seems counter-intuitive.

Reason: the pilots decide to fly low only when the visibility is low. Low visibility reduces accuracy of bombing. The visibility is a **confounding variable**. It relates to both the altitude of the plane and the precision of bombing.

Correlation of two variables may not tell the whole story, since there may exist confounding variables which relate to both variables. We must think hard about possible confounders.

Example: there is no correlation between the number of study hours per week of a high school student and the rank of the university he or she goes to after high school. Any confounding variable? What if we do the same study for students of similar IQ levels?

Homework Problem 1: Find an example of two random variables where one variable tends to rise with the other variable, but the correlation between the two variables is weak (close to 0).

Homework Problem 2: Use your personal experience to find one example to illustrate that the correlation is not causation.

Under what scenarios does correlation imply causation? To address this question, we first discuss two types of studies (data).

2 Nature of Studies

1. **Observational studies**: we observe data from existing record without any intervention.

2. **Controlled experiments**: we run experiments by controlling or varying certain conditions.

Example: selecting 1000 students to investigate the relationship between the number of study hours and the placement of the student in college.

Observational study: we simply observe the data from a sample of students.

Control experiment: we divide the students into several subgroups and put the students in each subgroup to study different number of hours per week. (Is this feasible?)

- If data are collected from a controlled experiment, correlation implies causation, but controlled experiments can sometimes be infeasible!

- Example: Tobacco companies knew (from observations) that the smokers had a higher rate of lung disease than nonsmokers, but they argued that this might be just correlation and not causation.

  Example 1: to see whether a new treatment is effective, a controlled experiment is more desirable – using placebo and treatment groups

  Example 2: to see which teaching method (the traditional method v.s. a new format of instruction) is better, how to design a controlled experiment?

  Procedure: first, we assign randomly the students into two classes with 50 students each. One class uses the traditional method and the other uses the new method. The **key point** here is that we have to use random assignment of the students to avoid bias in the comparisons!

  But with just observational studies, it is hard to make a comparison between the two methods. If the students sign up for a class out of their own preference, the assignment is not random.

3 Sampling

Sampling is often a necessary step for data collection. To perform statistical inference, we need to have a well-defined **population** and formulate a specific question about certain aspect of the population (parameter).

Example: population: students in this class; parameter: the proportion of students who do not major in statistics.

To find the proportion in the population may not be feasible, we need to do a **random sampling.** In a random sample of size $n = 5$, we find 3 students out of 5 who are not statistics majors. A **point estimate** of the population proportion p is $\widehat{p} = 3/5 = 0.6$.

To provide inference about the population proportion, we need an **interval estimate** – confidence interval (CI).

Constructing a 95% CI for p:

Assumption: $\frac{n}{N}$ is very small (say ¡ 0.05), where n is the sample size and N is the population size. If this assumption is violated, the statistical calculations may need to be adjusted.

$\text{Var}(\widehat{p}) = \frac{p(1-p)}{n}$

$\text{s.e.}(\widehat{p}) = \sqrt{\frac{p(1-p)}{n}} \leq \frac{1}{2\sqrt{n}}$ (s.e. denotes the standard error)

An approximate 95% CI for p is given as $\widehat{p} \pm 2\times\text{s.e.}(\widehat{p})$.

The **margin of error** : $2\times\text{s.e.}(\widehat{p}) \leq \frac{1}{\sqrt{n}}$.

Example: $n = 5$, the margin of error $= \frac{1}{\sqrt{5}} \approx 0.44$; a 95% CI for p is $(0.16, 1.00)$ when $\widehat{p} = 0.6$.

Question: how many people to survey in order to get the the margin of error of 1%?

We need $\frac{1}{\sqrt{n}} = 1\%$, and thus $n = 10^4$. Here the population size needs to be much larger than 10,000.

Common sampling plans

Example: what proportion of the students at SUFE watch World Cup? We wish to survey 40 students out of $15,000$.

- **Simple random sampling** (SRS): every observation has the equal probability of being selected

Requirement: we need to have a list of all the students and use random numbers to select 40.

- **Stratified sampling**: we may consider boys and girls separately. We take a random sample of boys and a random sample of girls.

  Assumption: within each subgroup (boys or girls), the observations have less variability (more homogeneous)

  Key idea: divide the data into subgroups with the idea that each group is more homogeneous than the whole population.

  Question for tomorrow: how many observations in each group should we take (if the total number of students is 40)?

- **Cluster sampling**: people may live in different places (e.g., dormitories).

  Strategy to use: we first randomly select two dormitories and then randomly select the students from each selected dorm. Here the dorms are the clusters. The larger clusters may have the higher chance to be selected.

  If we choose students from schools across the country, we may first select a few provinces randomly, and then select cities, and then schools, and then students. This is **multi-stage cluster sampling**. More to come tomorrow.

Just to go repeat some general principles:

1. Correlation is not causation. Think hard about confounding variables.
2. Understand the difference between observational studies and controlled experiments.
3. Understand the margin of error calculations and its use.

4. The population size might not be a factor in your sampling plan if the population size is sufficiently large.

5. A good sampling plan should be statistically valid and practically economical.