Principles of Statistics

Xuming He
xmhe@umich.edu

Lecture 2

1 The World Cup example by using stratified sampling

Method: divide the population into subgroups and apply simple random sampling (SRS) to each subgroup. For the World Cup example, divide the students into two groups of boys and girls, respectively, denoted by B and G.

Additional information: the ratio of boys to girls in the population is 1 to 3.

Calculation of the overall proportion: assume that $p_1 = 0.5$ and $p_2 = 0.2$, where p_1 and p_2 are the proportions of people watching the World Cup in groups B and G, respectively.

Then the overall proportion is $p = 0.2 \times \frac{3}{4} + 0.5 \times \frac{1}{4} = \frac{11}{40}$.

However, in practice the true proportions are **unknown**!

If we estimate p by stratified sampling, we want the variance of the estimate to be as small as possible given the total sample size.

Analysis: With the total sample size $n = 100$, let n_1 and n_2 be the number of observations for groups B and G, and $\widehat{p}_1$ and $\widehat{p}_2$ be the estimators for p_1 and p_2, respectively.

A **point estimator** for p is $\widehat{p} = \frac{1}{4}\widehat{p}_1 + \frac{3}{4}\widehat{p}_2$.

The estimate $\widehat{p}$ is an **unbiased** estimator of p, since $E(\widehat{p}_1) = p_1$ and $E(\widehat{p}_2) = p_2$.

The variance of $\widehat{p}$ is

$$\begin{aligned} \mathrm{Var}(\widehat{p}) &= \frac{1}{16}\mathrm{Var}(\widehat{p}_1) + \frac{9}{16}\mathrm{Var}(\widehat{p}_2) \\ &= \frac{1}{16}\frac{\sigma_1^2}{n_1} + \frac{9}{16}\frac{\sigma_2^2}{n_2}, \end{aligned}$$

by assuming that we select boys and girls independently, where $\sigma_1^2 = p_1(1-p_1)$ and $\sigma_2^2 = p_2(1-p_2)$ which are unknown.

Consider the case of $\sigma_2^2 = k\sigma_1^2$. Then

$$\begin{aligned} \mathrm{Var}(\widehat{p}) &= \frac{1}{16}\frac{\sigma_1^2}{n_1} + \frac{9k}{16}\frac{\sigma_1^2}{n_2} \\ &= \frac{1}{16}\sigma_1^2\left[\frac{1}{n_1} + \frac{9k}{n_2}\right], \end{aligned}$$

in which $n_1 + n_2 = n$ (given).

We find n_1 and n_2 by minimizing $\mathrm{Var}(\widehat{p})$, which is equivalent to minimizing $\frac{1}{n_1} + \frac{9k}{n_2} = \frac{1}{n_1} + \frac{9k}{n-n_1}$.

Thus we have

$$n_2 = 3\sqrt{k}n_1.$$

If $k = \frac{1}{4}$, then $n_2 = \frac{3}{2}n_1$. Given $n_1 + n_2 = 100$, we have $n_1 = 40$ and $n_2 = 60$.

Homework 3. Compare the above stratified sampling plan with SRS to see which one has a smaller variance by using the sample size $n = 100$? How much can you actually gain by using the stratified sampling?

Proof of $E(\widehat{p}) = p$. Let $\widehat{p} = \frac{X}{n} = \frac{1}{n}\sum_{i=1}^n X_i$, where $X_i = \begin{cases} 0, & \text{not watching the World Cup} \\ 1, & \text{watching the World Cup} \end{cases}$.

Then

$$\begin{aligned} E(\widehat{p}) &= E(\frac{1}{n}\sum_{i=1}^n X_i) = \frac{1}{n}\sum_{i=1}^n E(X_i) = p. \\ \text{no bias} &\iff E(\widehat{p}) = p. \end{aligned}$$

Question: under what scenarios is there a bias?

Multiple sources of bias:

- non-response bias (some people do not respond);

- false-response (dishonest answers from the selected subjects);
- convenience sampling (e.g., walk into a dormitory, and take a sample of people who are watching TV in a common room – not random sampling).

2 More discussion about the cluster sampling

Principle: the statistical method needs to match the sampling plan.

Method: randomly select 50 dormitory rooms and then randomly choose 2 students in each room.

A point estimator of p by the **cluster sampling** is

$$\widetilde{p} = \frac{1}{100}\sum_{i=1}^{50} X_i,$$

where X_i is the number of students who watch the World Cup in the i^{th} room.

A point estimator of p by the **SRS** is

$$\widehat{p} = \frac{X}{100},$$

where X is the number of students who watch the World Cup in the whole sample.

Question: which estimator is better? $\widetilde{p}$ or $\widehat{p}$? – need to calculate the variance! Both of them are unbiased estimators.

Let $X_i = \begin{cases} 0, & \text{not watching the World Cup} \\ 1, & \text{watching the World Cup} \end{cases}$, $i = 1, \ldots, n$. Then

$$\widetilde{p} = \frac{1}{100}\sum_{i=1}^{100} X_i = \left[(X_1 + X_2) + (X_3 + X_4) + \cdots + (X_{99} + X_{100})\right].$$

Assume that there are positive correlations within pairs (X_{2i-1}, X_{2i}), $i = 1, \ldots, 50$, such that $\mathrm{Cov}(X_{2i-1}, X_{2i}) > 0$. Then

$$\begin{aligned} \mathrm{Var}\left(X_{2i-1} + X_{2i}\right) &= \mathrm{Var}\left(X_{2i-1}\right) + \mathrm{Var}\left(X_{2i}\right) + 2\mathrm{Cov}(X_{2i-1}, X_{2i}) \\ &> \mathrm{Var}\left(X_{2i-1}\right) + \mathrm{Var}\left(X_{2i}\right). \end{aligned}$$

Therefore, $\text{Var}(\widehat{p}) < \text{Var}(\widetilde{p})$.

Conclusion: cluster sampling is more convenient, but it may lead to a larger variance unless the sample size is increased.

Bottom line: to compare different sampling methods, we need to consider both **statistical efficiency** and **cost of implementation**.

3 Probability

If the probability (chance) of rain tomorrow is 60%, we then believe that on average 6 out of 10 such days will have rain. If we can calculate probabilities, we can often use the calculations to demonstrate how unlikely something will happen.

A Broker Example: the government agency tries to convict a high level broker of cheating to make money for himself. The broker has an account for himself, which is set up to cover accidental errors he might make when taking orders from his customers. The broker sells any investments shortly after he makes a purchase. The agency suspects that the broker is making a large number of profitable trades which is unlikely to happen without cheating.

Data: in the first month, the broker makes 7 winning trades out of 8 trades.

If the trades are made for accidental purposes, the chance of a winning trade is 0.5. Having 7 out of 8 winning trades is too good to be true? Here is how we do the calculations:

$$P(\text{7 wins in 8 trades}) = 8 \times 0.5^8 = 3.125\%.$$

To show that the broker's record in unlikely, we also need to consider more extreme cases than 7

wins in 8 trades, which is 8 wins in 8 trades in this example. Together, we have

$$P(7 \text{ or } 8 \text{ wins in } 8 \text{ trades}) = 8 \times 0.5^8 + 0.5^8 = 3.515\%.$$

The reason that we should not focus on one specific record of 7 wins out of 8 trades is that any single record may have a small probability even if no cheating takes place. Think about having 50 wins in 100 trades! Is this probability small?

The above probability may not be small enough to show that the broker's record is unlikely. Let's use data from the second month, which happens to be 7 wins in 8 trades again. Then we calculate

$$\begin{aligned} & P(7+ \text{ wins in 8 trades in two consecutive months}) \\ = \; & P(7+ \text{ wins in 8 trades}) \times P(7+ \text{ wins in 8 trades}) \\ = \; & (3.515\%)^2 = 0.0012. \end{aligned}$$

If similar records are there for 12 consecutive months, the probability of having such a profitable record will then be very very small. For example,

$$\begin{aligned} & P(7+ \text{ wins in 8 trades in twelve consecutive months}) \\ = \; & P(7+ \text{ wins in 8 trades})^{12} \\ = \; & (3.515\%)^{12} = 3.6 \times 10^{-18}. \end{aligned}$$

Question: if you want to defend this broker, how to challenge the above calculation?

- We can challenge the assumption that the probability of winning in one trade is 50%. If the broker is knowledgable and experienced, the probability might be larger than 50%. To be more convincing, we may do the calculations by assuming that the broker has 0.7 probability to have a winning trade.

- We often calculate the probability based on the **least favorable criterion** to make the argument more convincing.

Probability calculations can be tricky. Often, intuitions might not lead to correct answers.

Example: we have two outcomes T (tail) and H (head) when flipping a coin. We flip a coin repeatedly until we get one of the two sequences below.

$$\text{sequence 1} \quad : \text{HHT},$$

$$\text{sequence 2} \quad : \text{THH}.$$

Which sequence is more likely to be reached in such experiments? Your intuition might say that there is equal probability for each sequence to be reached first, but this is far from being true.

We did real time experiments in class and found that empirically sequence 2 is twice as likely to be reached than sequence 1. We may find it convenient to use computer simulations to calculate such probabilities.

4 Hypothesis testing

We have a **population** and some quantity about the population that we want to know (**parameter**). We use θ as a generic notation for such a parameter.

For statistical inference, we set up the hypotheses

$$\text{null hypothesis} \quad : \quad \theta = 0,$$

$$\text{alternative hypothesis} \quad : \quad \theta = 0 \text{ or } \theta > 0.$$

Question: how to choose the null and alternative hypotheses?

Example 1: we want to know whether the population in Shanghai is changing (or increasing or decreasing) in in the past few years. Since we do not have census data every year, we have to rely on estimated population numbers each year.

Let θ be the slope of a regression line by regressing the estimated population numbers on time (year). Or we can use θ as the correlation between the two.

Principles : (1) never use the sample to formulate your hypothesis! You may formulate the hypotheses before the data are collected. (2) The null and alternative hypotheses should cover all the possible situations under consideration.

If you think the population in Shanghai is either constant or increasing, then you can use $\theta = 0$ as the null hypothesis and $\theta > 0$ as the alternative hypothesis.

Example 2: Let θ be the defective rate of a large shipment of products from a manufacturer to a store (reseller). The manufacturer claims that $\theta \leq 0.01$. The receiving side (the store) uses a random sample to test whether the defective rate is indeed so small. In this case, we use the null hypothesis $\theta = 0.01$ and the alternative hypothesis $\theta > 0.01$.

Example 3: If there is a dispute between the store and the manufacturer about the defective rate of the shipment, and the manufacturer sues the store for returning the shipment. Which hypotheses should be tested by the court?

Principle: the null hypothesis should be the one that the system (court) wants to protect.

Here, the court wants to protect the innocent. It may use

$$\begin{aligned} \text{null hypothesis} &: \quad \theta > 0.01, \\ \text{alternative hypothesis} &: \quad \theta \leq 0.01. \end{aligned} \tag{1}$$

When conducting the tests, we want the probability of Type I error to be controlled, that is,

$$P(\text{Type I error}) = P(\text{Reject the null hypothesis when it is true})$$

should be small. That is, if the store is correct in claiming that $\theta > 0.01$, we do not want to reject the null hypothesis easily. That is why $\theta > 0.01$ is placed as the null hypothesis.

The hypothesis (1) is not commonly used; for convenience, we consider

$$\begin{aligned} \text{null hypothesis} &: \quad \theta = 0.01, \\ \text{alternative hypothesis} &: \quad \theta < 0.01. \end{aligned} \tag{2}$$

The hypotheses (1) and (2) lead to the same conclusions, but the latter is easier to handle.

Question: What if the store is suing the manufacturer? What hypotheses would the court use? (figure it out after class)

Principle: in scientific studies, we put what you want to discover under the alternative hypothesis. For the Shanghai population size change example, if we want to show that the population size is increasing, then the alternative hypothesis should be $\theta > 0$.

Let's repeat what we have learned.

1. In stratified sampling how many to sample from each stratum (given the total sample size) depends on two factors: the size of the stratum and the variance of the stratum relative to the other strata.

2. Probability calculations are very useful in showing how unlikely something is under certain assumptions.

3. Intuitions may not work well in probability calculations. Computer simulations can be useful.

4. In hypothesis testing, there are some simple principles to follow in setting up the null and the alternative hypotheses.