Topic 4: Sampling Distribution and Large-Sample Estimation

Optional Reading: Chapter 7 and 8

Xiner Zhou

Department of Statistics

University of California, Davis

- Central Limit Theorem for Sampling Distribution
- Large-Sample estimation (point estimation and confidence interval)
    - One population mean
    - Difference between two population means
    - One population proportion
    - Difference between two population proportions

Descriptive statistics

Probability

Random variable and its distribution

Where we have been

Inference problems:

- Decision making
    - The new vaccine is more effective than an old one?
    - A home-buyer wants to estiamte the market price for a house before putting out an offer
- Prediction
    - A financial analyst need to predict the behavior of stock market
    - Political scientists need to predict the outcome of an election

…….

It's the job of statistics to make objective suggestions for decision making and predictions, based on data, i.e. "let the data speak!"

First, we need to have a probability model for each problem, e.g. Binomial experiment

- The new vaccine is more effective than an old one?

p1: probability of getting sick who received new vaccine
p2: probability of getting sick who received the old vaccine

Sample from the population who received the new vaccine ~ Binomial(n1, p1)
Sample from the population who received the oldvaccine ~ Binomial(n1, p2)

Question: p1 <p2 ? Hypothesis testing problem

- A home-buyer wants to estiamte the market price for a house before putting out an offer

The houses belong to a population of houses with similar characteristics, with population has some mean μ
Question: $\mu = ?$

Estimation problem

Previous chapter: probability distribution

Binomial(n,p) ⟶ p=?

$Normal(\mu, \sigma^2)$ ⟶ $\mu = ?, \sigma^2 = ?$

The shape of distribution is determined by parameters

In practice when modeling specific problem,

we need to decide which type of probability distribution is appropriate as a model.

Political poll:
Do you support …. Yes/No

Binomial distribution might be appropriate

Student test scores in a class?

Normal distribution might be appropriate

Once the type of distribution is considered as appropriate,
Still we don't know the value of the parameters.

➡ We need to "Infer" what the parameters are, i.e. Inference about parameters

Estimation:
estimate the values of parameters (topic of this chapter)

Point estimation:
Using sample data, a single number based on certain formula that gives the best guess about the parameter of interest

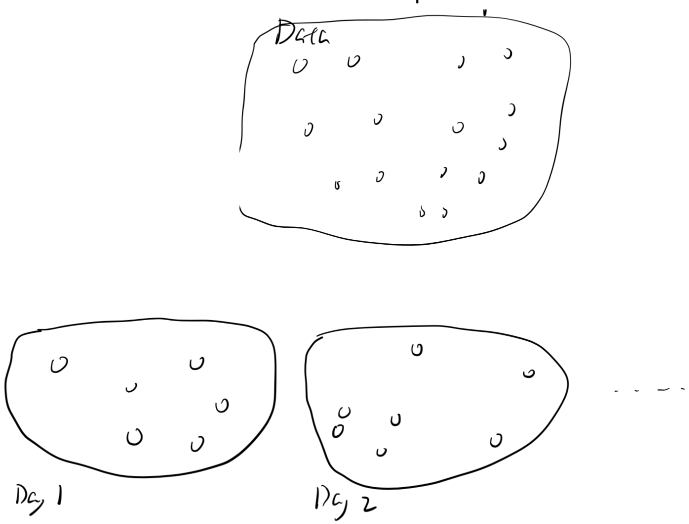
Confidence interval:
Using sample data, an interval based on certain formula that forms the range within which the parameters is expected to lie.

Hypothesis testing:
making decision about whether certain statement is true or false (topic of next chapter)

Sampling Distribution

Once the type of distribution is considered as appropriate,
Still we don't know the value of the parameters.

Use sampling information, i.e. data, to estimate parameters!



Since you select a random sample, the data or sample will be different each time you re-draw the sample or someone else conducting the same experiment.

➡ So, your estimates will be different, and the difference is the natural variation due to the random sampling process.

When a random sample is drawn from a population, any quantity calculated based on the sample are called a statistic. E.g. sample mean, sample variance, sample proportion…

These statistics change for each different random sample, so statistics themselves are random variables.

Any random variable has probability distribution to describe:

- What values can occur
- How often each value occur

We call the probability distribution for a statistic: the sampling distribution of a statistic

Our goal: sampling distributions for

Sample mean $\longrightarrow$ μ

Sample proportion $\longrightarrow$ p

There are 3 ways to find the sampling distribution of a statistic:

1. Derive the distribution mathematically

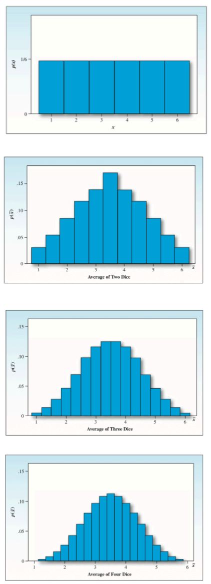
2. Use a simulation to approximate the distribution.

   Draw a large number of samples of size n, calculating the value of the statistic for each sample, tabulate the results in a histogram. When the number of repeated sampling is large, the histogram will be very close to the true sampling distribution.

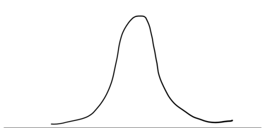
3. Use central limit theorem to derive approximate sampling distribution

Central Limit Theorem (CLT) for Sampling Distribution

What's the distribution of average number when you toss a fair die many times?



- Symmetric, bell-shaped curve
- Spread of the distribution slowly decreases when we increase n
    - i.e. the distribution becomes thinner, more concentrated around the center



Normal

This is true in general:

Average of random samples of measurements drawn from a population tend to have an approximately Normal distribution

Central Limit Theorem

If random sample of n observations are drawn from a population (any distributions, not necessarily Normal), with mean μ and standard deviation σ, then, when n is large (rule of thumb $n \geq 30$):

The sampling distribution of sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is approximately normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$

$$\bar{X} \sim N(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2)$$

- When the population is actually Normal, then the sampling distribution of sample mean is exactly Normal regardless of how many sample we took
- CLT will be used extensively in the following “inference” problems
- Application of CLT to sampling distribution of the sample mean

The Sampling Distribution of the Sample Mean

If random sample of n observations are drawn from a population (any distributions, not necessarily Normal), with mean μ and standard deviation σ, then, when n is large (rule of thumb $n \geq 30$):

Directly due to the central limit theorem,

The sampling distribution of sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is approximately normally distributed

with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$:

$$\bar{X} \sim N(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2)$$

$$\bar{X} \sim N(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2)$$

The standard deviation of a statistic used as an estimator of a population parameter is called the standard error of the estimator, abbreviated as SE

e.g. $SE(\bar{X}) = \frac{\sigma}{n}$

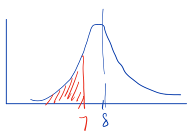
- SE measures: how precise we can estimate the parameter using this statistic, i.e. the precision of the estimator

The duration of Alzheimer's disease from the onset of symptoms until death ranges from 3 to 20 years; the average is 8 years with a standard deviation of 4 years. The administrator of a large medical center randomly selects the medical records of 30 deceased Alzheimer's patients from the medical center's database, and records the average duration. Find the approximate probabilities for these events:

1. The average duration is less than 7 years.
2. The average duration exceeds 7 years.
3. The average duration lies within 1 year of the population mean $\mu = 8$.

Solution: $X_1 \cdots X_{30} \overset{iid}{\sim}$ some unknown dist with $\mu = 8$, $\sigma = 4$

CLT $\Rightarrow \bar{X} \sim N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$

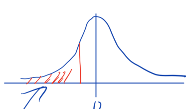


$\mu = 8$

$\frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{30}}$

We like to standardize to standard normal dist:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$



$P(\underbrace{\bar{X} < 7}_{})$

$\Leftrightarrow Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{7-8}{\frac{4}{\sqrt{30}}} = -1.37$

$= P(Z < -1.37)$

find the cumulative area under the normal curve

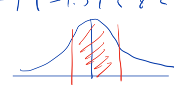
$= 0.0853$

$P(\bar{X} > 7) = 1 - P(\bar{X} < 7) = 0.9147$

$P(\underbrace{7 < \bar{X}}_{} < 9)$

$\Leftrightarrow -1.37 < \frac{7-8}{\frac{4}{\sqrt{30}}} < Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{9-8}{\frac{4}{\sqrt{30}}} = 1.37$

$= P(-1.37 < Z < 1.37)$



$= P(Z < 1.37) - P(Z < -1.37)$

$= 0.9147 - 0.0853 = 0.8294$

To avoid difficulties with the Federal Trade Commission or state and local consumer protection agencies, a beverage bottler must make reasonably certain that 12-ounce bottles actually contain 12 ounces of beverage. To determine whether a bottling machine is working satisfactorily, one bottler randomly samples 10 bottles per hour and measures the amount of beverage in each bottle. The mean $\bar{x}$ of the 10 fill measurements is used to decide whether to readjust the amount of beverage delivered per bottle by the filling machine.

If records show that the amount of fill per bottle is normally distributed, with a standard deviation of .2 ounce, and if the bottling machine is set to produce a mean fill per bottle of 12.1 ounces, what is the approximate probability that the sample mean $\bar{x}$ of the 10 test bottles is less than 12 ounces?

$$X_1 \dots X_{10} \overset{\text{sample from}}{\sim} N(\mu = 12.1,\ \sigma^2 = 0.2^2)$$

$\Rightarrow$ sampling distribution of sample mean is

$$\bar{X} = \frac{X_1 + \dots + X_{10}}{10} \sim N\Big(\mu = 12.1,\ \underbrace{\left(\frac{\sigma}{n}\right)^2 = \left(\frac{0.2}{10}\right)^2}_{SE^2}\Big)$$

$$\Rightarrow P(\bar{X} < 12) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{12 - \mu}{\sigma/\sqrt{n}}\right)$$

$$= P\Big(Z < \underbrace{\frac{12 - 12.1}{0.2/\sqrt{10}}}_{-1.59}\Big)$$

$$= 0.0559$$

The Sampling Distribution of the Sample Proportion

There are many practical situations, such as opinion polls, where we randomly sample n people to estimate the proportion p of people in the population who have a specific characteristic.

X= total number of sampled individual who have this characteristic

Sample proportion $\hat{p} = \frac{X}{n}$

Then directly due to the central limit theorem, then the sampling distribution of sample proportion can be approximated by a normal distribution:

$$\hat{p} = \frac{X}{n} \sim N(p, (\sqrt{\frac{p(1-p)}{n}})^2)$$

In a survey, 500 mothers and fathers were asked about the importance of sports for boys and girls. Of the parents interviewed, 60% agreed that the genders are equal and should have equal opportunities to participate in sports. Describe the sampling distribution of the sample proportion $\hat{p}$ of parents who agree that the genders are equal and should have equal opportunities.

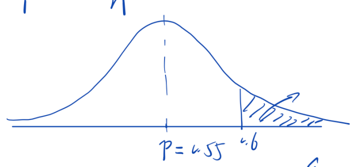
If the actual p=0.55. What is the probability of observing a sample proportion as large as or even larger than the sample proportion 0.6?

500 parents are a random sample of all parents in U.S.

We're interested in the true proportion of parents who think that both genders have equal opportunities to participate in sports, p

So the sample proportion $\hat{p}$ of parents who agree on this point has sampling distribution

$$\hat{p} = \frac{X}{n} \overset{\text{approximately}}{\sim} N\left(p, \underbrace{\sqrt{\frac{p(1-p)}{n}}}_{SE}{}^{2}\right)$$



$$P(\hat{p} \geq .6) = P\left(\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \geq \frac{.6 - p}{\sqrt{\frac{p(1-p)}{n}}}\right)$$

$$= P\left(Z \geq \frac{.6 - .55}{\sqrt{\frac{.55(1-.55)}{500}}}\right)$$

$$= P(Z \geq 2.25)$$

$$= 1 - .9875$$

$$= .0122$$