ECON10005 Quantitative Methods 1

Tutorial in Week 7

Zheng Fan

The University of Melbourne

Introduction

Zheng Fan

- Ph.D student in Economics at Unimelb

Don't be shy if you need help

- visit Ed Discussion Board
- Go for lecturer's consultation sessions: see Canvas
- In case of special considerations, consult Stop 1
- For admin issues contact Chin via qm1-economics@unimelb.edu.au
- Email me with subject code titled: fan.z@unimelb.edu.au

Announcement

Anzac Day (25th April):

- Next Tuesday is public holiday (no tutorials)
- Tutorial will be pre-recorded and sent to you
- No need to worry attendance, but you have to complete the quiz

MST-2 (10%):

- 4th May (Thursday in week 9)
- Register for a session by 28th April 5pm (Friday in week 8)

Group (or individual) assignment: Second draft (7%):

- Feedbacks for your first draft (3%) will be provided later
- Due in week 11 (details to be announced)

Statistical Inference

sample data ⟵ population data

$\bar{x}, s^2 \xrightarrow{\text{infer}} \mu, \sigma^2$

- Statistical inference: using a random sample to infer some aspect of a population (mean, variance, hypothesis test)
- Estimator: a rule (function) for calculating an estimate of a given quantity based on a sample $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ $\quad X_1 = X_1$
- (Desired) properties of estimators:
  - Unbiasedness: $E(\bar{X}) = \mu$ $\quad \bar{X}$ is an unbiased estimator of μ
  - Consistency: $\bar{X} \xrightarrow{p} \mu$ as $n \to \infty$ $\quad \Leftarrow \begin{cases} \text{① unbiased} \\ \text{② } var(\bar{x}) \to 0, \ n \to \infty \end{cases}$
  - Asymptotic normality: CLT $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ for large n

$\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$ $\quad \to z$

Part A Questions

for each $i = 1, 2, \cdots n$.

$var(aX+bY) = a^2 var(X) + b^2 var(Y) + 2ab\,cov(X,Y)$

$X_i \sim Ber(p)$.

X_i	0	1
Pr	$1-p$	p

A.1 Suppose we have a simple random sample of size n from 0/1 random variables that satisfy $P(X_i = 1) = p$, and $P(X_i = 0) = 1 - p$ where i can be 1, 2, 3, ... n.

Binomial $G = \sum_{i=1}^{n} X_i \sim Bi(n, p)$.

Show that $\bar{X}$ is an unbiased estimator of p.

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\left(\sum_{i=1}^{n} p\right) = \frac{1}{n}\cdot np = p.$$

extension:
show consistency:

$$var(\bar{X}) = var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} var(X_i) = \frac{1}{n^2}\cdot\sum_{i=1}^{n} p(1-p) = \frac{1}{n^2}\cdot np(1-p) = p(1-p)/n \xrightarrow{n\to\infty} 0$$

$E(X_i) = p$

$$var(X_i) = (1-p)\cdot(0-p)^2 + p(1-p)^2 = (1-p)p^2 + p(1-p)^2 = (1-p)p(p+1-p) = p(1-p)$$

$E(G) = np$.

Part A Questions

A.1 Suppose we have a simple random sample of size n from 0/1 random variables that satisfy $P(X_i = 1) = p$, and $P(X_i = 0) = 1 - p$ where i can be 1, 2, 3, ...

Show that $\bar{X}$ is an unbiased estimator of p.

$$\begin{aligned}
E(\bar{X}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) && \text{(addition rule for E)} \\
&= \frac{1}{n}\sum_{i=1}^{n} p \\
&= p
\end{aligned}$$

Part A Questions

A.2 Suppose we want to estimate $\mathrm{var}(\bar{X})$. Here are two estimators:

(a) s^2/n, where s^2 is the usual variance

(b) $\bar{X}(1-\bar{X})/n$

Discuss why these might be applicable. Are they equal to each other? How might this depend on the sample size n?

Part A Questions

A.2 Suppose we want to estimate $\text{var}(\bar{X})$. Here are two estimators:

(a) s^2/n, where s^2 is the usual variance

(b) $\bar{X}(1-\bar{X})/n$

Discuss why these might be applicable. Are they equal to each other? How might this depend on the sample size n?

We first need to know what we are trying to estimate.

$$\begin{aligned}\text{var}(\bar{X}) &= \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(X_i) \\ &= \frac{1}{n^2} n\,\text{var}(X_i) = \frac{\text{var}(X_i)}{n} \\ &= \frac{p(1-p)}{n}\end{aligned}$$

Part A Questions

$E(X_i) = E(\bar{X}) = p.$ $E(\bar{X}^2) \neq (E(\bar{X}))^2.$

$E\left(\frac{\bar{X}}{n} - \frac{\bar{X}^2}{n}\right) = \frac{E(\bar{X})}{n} - \frac{E(\bar{X}^2)}{n}.$

$\hookrightarrow E\left(\frac{\bar{X}(1-\bar{X})}{n}\right) \overset{?}{=} \frac{p(1-p)}{n}.$

A.2 Suppose we want to estimate $\boxed{\text{var}(\bar{X}) = \frac{\text{var}(X_i)}{n} = \frac{p(1-p)}{n}}$. Here are two estimators:

- (a) s^2/n, where s^2 is the usual variance
- (b) $\bar{X}(1-\bar{X})/n$

$\frac{s^2}{n}$ is an unbiased estimator of $\text{var}(\bar{X})$.

Discuss why these might be applicable. Are they equal to each other? How might this depend on the sample size n?

(a) is pretty standard, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ ← sample variance

so s^2 is good to estimate $\text{var}(X_i)$, in other words $\boxed{E(s^2) = \text{var}(X_i)}$ ⇐ proof.

Hence, $E(s^2/n) = \text{var}(X_i)/n = \text{var}(\bar{X})$

Part A Questions

A.2 Suppose we want to estimate $\text{var}(\bar{X}) = \frac{\text{var}(X_i)}{n} = \frac{p(1-p)}{n}$. Here are two estimators:

(a) s^2/n, where s^2 is the usual variance

(b) $\bar{X}(1-\bar{X})/n$

Discuss why these might be applicable. Are they equal to each other? How might this depend on the sample size n?

(b) recall that p can be estimated by $\bar{X}$, namely $E(\bar{X}) = p$

so $\bar{X}(1-\bar{X})/n$ seems to be a good match with $p(1-p)/n$

Part A Questions

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$\Rightarrow n\bar{X} = \sum_{i=1}^{n} X_i$

(a) (b)

Let's compare s^2/n and $\bar{X}(1-\bar{X})/n$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2)$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + \sum_{i=1}^{n}\bar{X}^2\right)$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - 2\bar{X}n\bar{X} + n\bar{X}^2\right)$$

$-2n\bar{X}^2 + n\bar{X}^2$

$X_i = 0 \Rightarrow X_i^2 = 0$

$X_i = 1 \Rightarrow X_i^2 = 1$,

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i - n\bar{X}^2\right)$$

$$= \frac{1}{n-1}(n\bar{X} - n\bar{X}^2) = \frac{n}{n-1}\bar{X}(1-\bar{X})$$

so

$$(a)\ \frac{s^2}{n} = \frac{n}{n-1}\frac{\bar{X}(1-\bar{X})}{n}\ (b).$$

$n \to \infty$, $\frac{n}{n-1} \to 1$

Part B Questions

B.1 Calculate the two variance estimators in Part A question 2.

B.2 Compare their values.

The data is from pre-quiz 6 part B Ghosts.xlsx.

- sample variance $s^2 = \text{VAR.S(B12:B145)} \approx 0.2483$
- sample size $n = 134$
- $\bar{X} = \text{AVERAGE(B12:B145)} = 0.4403$

(c) $\boxed{\dfrac{\bar{X}(1-\bar{X})}{n-1}}$

$E((c)) =$

The estimates are:

$$(a)\ \frac{s^2}{n} = \frac{0.2483}{134} = 0.00185$$

$$(b)\ \frac{\bar{X}(1-\bar{X})}{n} = \frac{0.4403 \times (1 - 0.4403)}{134} = 0.00184 \times \frac{134}{133} = 0.00185$$

Part C Questions

CLT: $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = Z \overset{a}{\sim} N(0,1)$

$n \to$ large enough

$\bar{X}(\hat{p}) \xrightarrow{\text{infer}} \mu(p)$

$\mu = 42\%$

1. Suppose it is true that 42% of the population believes in ghosts. Use the Central Limit Theorem to work out (approximately) the probability that an estimate of this proportion based on a sample of size 134 will differ from the population proportion by less than 2%.
2. How does your answer change if the sample sizes were doubled to 268?

$$\Pr\left(|\bar{X} - \mu| < 2\% \right) = ?$$

Part C Questions

$\rightarrow \text{var}(X_i) = p(1-p)$

$\rightarrow \sigma = \sqrt{p(1-p)}$

$\mu = p = 0.42$

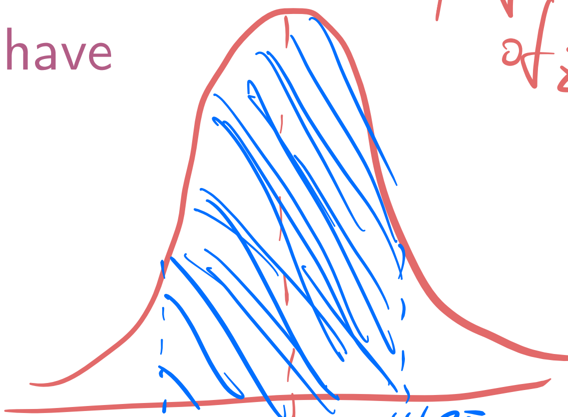
CLT states that

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \overset{a}{\sim} N(0, 1)$$

Using $\mu = 0.42$, $\sigma = \sqrt{p(1-p)} = 0.4936$, n = 134, we have

$$\frac{\bar{X} - 0.42}{0.4936/\sqrt{134}} = \boxed{\frac{\bar{X} - 0.42}{0.0426}} \overset{a}{\sim} N(0, 1) \rightarrow Z$$

pdf of Z



-0.4695 0 0.4695

sd = 1

We want to know the probability

$$P(|\bar{X} - \mu| < 0.02) = P(\underset{0.0426}{-0.02} < \underset{0.0426}{\bar{X} - \mu} < \underset{0.0426}{0.02})$$

$$= P\left(-0.4695 < \frac{\bar{X} - 0.42}{0.0426} < 0.4695\right)$$

$$= P(-0.4695 < Z < 0.4695)$$

$$= P(Z < 0.469) - P(Z < -0.469)$$

In Excel, use "=NORM.DIST(0.469,0,1,TRUE) - NORM.DIST(-0.469,0,1,TRUE)"

Part D Questions

Discuss how each of the following sampling situations may not produce a simple random sample. Also discuss any issues with the representativeness of the sample.

1. The aim is to estimate the voting intentions of the Australian population in the next Federal election.

   (a). All customers at this car dealership are surveyed for their voting intentions.

   (b). All customers at this cafe are surveyed for their voting intentions.

Part D Questions

Discuss how each of the following sampling situations may not produce a simple random sample. Also discuss any issues with the representativeness of the sample.

1. The aim is to estimate the voting intentions of the Australian population in the next Federal election.

   (a). All customers at this car dealership are surveyed for their voting intentions.

   (b). All customers at this cafe are surveyed for their voting intentions.

The aim is to estimate the voting intentions of the entire electorate of Australia, both these will give non-representative samples and hence biased results.

Part D Questions

Discuss how each of the following sampling situations may not produce a simple random sample. Also discuss any issues with the representativeness of the sample.

2. The aim is to investigate the academic performance of school students in Melbourne. Data is gathered for 1,000 students by randomly sampling 100 students from each of 10 schools in Melbourne.

Part D Questions

Discuss how each of the following sampling situations may not produce a simple random sample. Also discuss any issues with the representativeness of the sample.

2. The aim is to investigate the academic performance of school students in Melbourne. Data is gathered for 1,000 students by randomly sampling 100 students from each of 10 schools in Melbourne.

(1). It is not made clear how the 10 schools were selected.

(2). The 100 students within each school may have been selected independently of each other, but they may tend to exhibit school-specific attributes

A two step sampling scheme like this (i.e. first choose schools, then choose students) is quite possibly statistically sound, but is generally more complicated to analyse than a simple random sample.

Part D Questions

Discuss how each of the following sampling situations may not produce a simple random sample. Also discuss any issues with the representativeness of the sample.

3. Data is gathered to form a time series of stock market returns for Facebook shares over the last 12 months (as in Tutorial 2).

The end

Thanks for your attention! 😄

Feel free to leave and see you next week (on video) due to public holiday!