Sampling distributions and the Central Limit Theorem

Johan A. Elkink

University College Dublin

7 March 2013

Outline

Sampling

Statistical inference (or inductive statistics) concerns drawing conclusions regarding a population of cases on the basis of a sample, a subset.

Sampling refers to the selection of an appropriate subset of the population.

Sampling frame

The sampling frame refers to the identifiable list of members of the population, from which the sample can be selected.

Simple random sampling

Each subject from a population has the exact same chance of being selected in the sample, i.e. the sampling probability for each subject is the same.

Sampling bias

When the sampling probability correlates with a variable of interest, we are likely to get biased results.

Sampling bias

When the sampling probability correlates with a variable of interest, we are likely to get biased results.

Other causes of bias:

- Misreporting by respondents
- Characteristics of interviewer
- Question-ordering effects

Exercise

What is wrong with the following scenarios?

- Students in a class are asked to raise their hands if they have cheated on an exam one or more times within the past year.

Exercise

What is wrong with the following scenarios?

- Students in a class are asked to raise their hands if they have cheated on an exam one or more times within the past year.
- To get information on opinions among students, 100 students are surveyed at the start of a 9 am class.

Exercise

What is wrong with the following scenarios?

- Students in a class are asked to raise their hands if they have cheated on an exam one or more times within the past year.
- To get information on opinions among students, 100 students are surveyed at the start of a 9 am class.
- To get information on public opinion, you stand at the entrance of the Apple Store in a shopping street and interview passers-by randomly.

Weighting

Other types of sampling procedures exist, such as stratified or clustering sampling, whereby subsequent weighting of the data can recover the necessary unbiasedness for statistical inference.

Generally, the weight would be the inverse of the probability of inclusion in the sample.

Outline

Parameters

A parameter is number that describes a feature of the population. A parameter is generally fixed and not observable.

Parameters

A parameter is number that describes a feature of the population. A parameter is generally fixed and not observable.

A statistic is a number that describes a feature of a sample and is fixed for a given sample, but varies across samples.

Parameters

A parameter is number that describes a feature of the population. A parameter is generally fixed and not observable.

A statistic is a number that describes a feature of a sample and is fixed for a given sample, but varies across samples.

We can use statistics to estimate parameters.

(Moore, McCabe & Craig 2012: 198)

From probability to statistics

Using probability theory, we can understand how samples behave on average, given some assumptions.

From probability to statistics

Using probability theory, we can understand how samples behave on average, given some assumptions.

By comparing the sample at hand to samples on average, we can draw probabilistic conclusions about the population parameters.

Sampling distribution

"The sampling distribution of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population."

(Moore, McCabe & Craig 2012: 201)

Example

- Take 10 samples of size $n = 4$ from the class.
- Calculate average length.
- Draw histogram.

Sampling error

The amount of error when a population parameter is estimated or predicted by a sample estimate.

The bigger the sample, the lower the sampling error.

Estimates and uncertainty

When we estimate a parameter, we are uncertain what the true value is.

Besides an estimate of the parameter, we also need an estimate of how certain we are of this estimate.

The typical indicator of this is the standard error.

Outline

i.i.d.

We make three assumptions about our data to proceed:

- The observations are independent
- The observations are identically distributed
- The population has a finite mean and a finite variance

A variable for which the first two assumptions hold is called *iid*.

Independent observations

Intuitively: the value for one case does not affect the value for another case on the same variable.

More formally: $P(x_1 \cap x_2) = P(x_1)P(x_2)$.

Independent observations

Intuitively: the value for one case does not affect the value for another case on the same variable.
More formally: $P(x_1 \cap x_2) = P(x_1)P(x_2)$.

Examples of dependent observations:

- grades of students in different classes;
- stock values over time;
- economic growth in neighbouring countries.

Identically distributed

All the observations are drawn from the same random variable with the same probability distribution.

Identically distributed

All the observations are drawn from the same random variable with the same probability distribution.

An example where this is not the case would generally be panel data. E.g. larger firms will have larger variations in profits, thus their variance differs, thus these are not observations from the same probability distribution.

Random sample

A proper random sample is i.i.d.

The law of large numbers and the Central Limit Theorem help us to predict the behaviour of our sample data.

Law of large numbers

The law of large numbers (LLN) states that, if these three assumptions are satisfied, the sample mean will approach the population mean with probability one if the sample is infinitely large.

Central Limit Theorem

If these three assumptions are satisfied,

Central Limit Theorem

If these three assumptions are satisfied,

- The sample mean is normally distributed, *regardless of the distribution of the original variable.*

Central Limit Theorem

If these three assumptions are satisfied,

- The sample mean is normally distributed, *regardless of the distribution of the original variable.*
- The sample mean has the same expected value as the population mean (LLN).

Central Limit Theorem

If these three assumptions are satisfied,

- The sample mean is normally distributed, *regardless of the distribution of the original variable.*
- The sample mean has the same expected value as the population mean (LLN).
- The standard deviation (standard error) of the sample mean is: $S.E.(\bar{x}) = \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$.

Sample and population size

Note that the standard error depends only on the sample size, *not on the population size.*

Central Limit Theorem: unknown σ

When the population variance, σ, is unknown, we can use the sample estimate:

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}_x}{\sqrt{n}}$$

$$\hat{\sigma}_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Aside: variance of proportion

Note that the variance of x that of which a proportion of p cases are 1 and all others 0 can be calculated as:

$$\sigma_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = p(1-p)$$

Central Limit Theorem: example

Suppose we have a random sample of 100 individuals and ask each what their first preference vote would be if there were elections today. If 30 of them say they would vote Fianna Fail, what is the standard error of the estimate that the proportion is $p = .3$?

Central Limit Theorem: example

Suppose we have a random sample of 100 individuals and ask each what their first preference vote would be if there were elections today. If 30 of them say they would vote Fianna Fail, what is the standard error of the estimate that the proportion is $p = .3$?

$$\sigma_{\hat{p}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \frac{\sqrt{0.21}}{\sqrt{100}} = 0.0458$$

Exercises

Calculate the standard errors:

- A sample of 20 students has an average grade of 60.
- Out of a sample of 100 road accidents, 10 were fatal.
- Of the 1300 respondents in a survey, 48% voted "Yes" on the Lisbon Treaty referendum.
- The average score on a 5-point political knowledge scale in the same survey is 2.34.

Regression

Open `demdev.dta` and look at the standard errors for:

- The mean of `laggdppc` and `polity2`.
- The correlation between `laggdppc` and `polity2`.
- The regression coefficients for regressing `polity2` on `laggdppc`.
- The regression coefficients for regressing `polity2` on `log(laggdppc)`.