

Central Limit Theorem

Jos Elkind

October 14, 2008

1 Sampling

2 Central Limit Theorem

- Goal
- Assumptions
- Theorem
- Movie

Outline

1 Sampling

2 Central Limit Theorem

- Goal
- Assumptions
- Theorem
- Movie

Simple random sampling

Each subject from a population has the exact same chance of being selected in the sample.

The **sample probability** for each subject is the same.

Simple random sampling

Each subject from a population has the exact same chance of being selected in the sample.

The **sample probability** for each subject is the same.

Random sampling is necessary for valid statistical inferences.

Sampling error

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

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.

Non-random samples

For example:

- Internet poll

Non-random samples

For example:

- Internet poll
- Muslims in the Netherlands

Non-random samples

For example:

- Internet poll
- Muslims in the Netherlands
- Literary Digest

Non-random samples

Bias: when a sample statistic gives systematically the wrong estimate of the population parameter.

Non-random samples

Bias: when a sample statistic gives systematically the wrong estimate of the population parameter.

When the probability of inclusion in the sample varies across subjects, we have the risk of bias.

Non-random samples

Bias: when a sample statistic gives systematically the wrong estimate of the population parameter.

When the probability of inclusion in the sample varies across subjects, we have the risk of bias.

When the probability of inclusion correlates with the variable of interest, we have bias.

Other causes of bias

For example:

- Misreporting by respondents

Other causes of bias

For example:

- Misreporting by respondents
- Characteristics of interviewer

Other causes of bias

For example:

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

Other causes of bias

For example:

- Misreporting by respondents
- Characteristics of interviewer
- Question-ordering effects
- Other examples?

Other types of sampling

Other types of sampling procedures exist, such as stratified or clustering sampling

Other types of sampling

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.

Other types of sampling

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

1 Sampling

2 Central Limit Theorem

- Goal
- Assumptions
- Theorem
- Movie

Outline

- 1 Sampling
- 2 Central Limit Theorem
 - Goal
 - Assumptions
 - Theorem
 - Movie

From probability to statistics

Using **probability theory**, we can understand how samples behave once we know the parameters.

Usually, we do not know these variables.

From probability to statistics

Using **probability theory**, we can understand how samples behave once we know the parameters.

Usually, we do not know these variables.

The **Central Limit Theorem** helps us understand the sample data in terms of the random variables in probability theory.

Estimates and uncertainty

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

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.

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**.

Variables

In **descriptive statistics**, when we talk about a variable, we refer to a set of observations on one particular measure, for example “age” or “political interest”.

Variables

In **descriptive statistics**, when we talk about a variable, we refer to a set of observations on one particular measure, for example “age” or “political interest”.

In **inferential statistics**, we will have to make an important distinction:

Variables

In **descriptive statistics**, when we talk about a variable, we refer to a set of observations on one particular measure, for example “age” or “political interest”.

In **inferential statistics**, we will have to make an important distinction:

- **Random variable**: the “potential observations”, or the process from which the observations are drawn;

Variables

In **descriptive statistics**, when we talk about a variable, we refer to a set of observations on one particular measure, for example “age” or “political interest”.

In **inferential statistics**, we will have to make an important distinction:

- **Random variable**: the “potential observations”, or the process from which the observations are drawn;
- **Realised variable**: the observations generated by the random variable.

Variables

In **descriptive statistics**, when we talk about a variable, we refer to a set of observations on one particular measure, for example “age” or “political interest”.

In **inferential statistics**, we will have to make an important distinction:

- **Random variable**: the “potential observations”, or the process from which the observations are drawn;
- **Realised variable**: the observations generated by the random variable.

(Note that programming languages, including R, also have variables, which refers to simply any kind of storage of data in computer memory - anything on the left hand side of the assignment operator. This is unrelated to the above use of the word variable.)

Outline

- 1 Sampling
- 2 Central Limit Theorem
 - Goal
 - Assumptions
 - Theorem
 - Movie

Independent and identically distributed

We make two assumptions about our data to proceed:

- The observations are **independent**
- The observations are **identically distributed**

This is called an **i.i.d.** or **random** sample.

Independent observations

Knowing the value of one observation does not tell us anything about the value in another observation.

Independent observations

Knowing the value of one observation does not tell us anything about the value in another observation.

Examples of dependent observations:

Independent observations

Knowing the value of one observation does not tell us anything about the value in another observation.

Examples of dependent observations:

- Grades of students in three different classes;

Independent observations

Knowing the value of one observation does not tell us anything about the value in another observation.

Examples of dependent observations:

- Grades of students in three different classes;
- Time-series data;

Independent observations

Knowing the value of one observation does not tell us anything about the value in another observation.

Examples of dependent observations:

- Grades of students in three different classes;
- Time-series data;
- GDP in European countries;
- etc.

Identically distributed

All the observations are draws from the same **random variable** with the same **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 behavior of our sample data in the case of a large, random sample.

Central Limit Theorem: assumptions

- Sample is i.i.d.
- Population mean and variance exist (are finite)

Law of large numbers

If these two assumptions are satisfied, the sample mean will approach the population mean with probability one if the sample is infinitely large.

Outline

- 1 Sampling
- 2 Central Limit Theorem
 - Goal
 - Assumptions
 - Theorem
 - Movie

Central Limit Theorem

If these two assumptions are satisfied,

Central Limit Theorem

If these two assumptions are satisfied,

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

Central Limit Theorem

If these two 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 (Law of Large Numbers).

Central Limit Theorem

If these two 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 (Law of Large Numbers).
- The standard deviation (**standard error**) of the sample mean is $\frac{\sigma}{\sqrt{n}}$.

Central Limit Theorem: use

The CTL is the foundation of many statistical (**frequentist**) procedures, because even when the distribution of the variable under study is not normal, the **average** will be. Our knowledge of the normal distribution will now allow us to draw **inferences** from our sample data about the population parameters.

Central Limit Theorem: use

The CTL is the foundation of many statistical (**frequentist**) procedures, because even when the distribution of the variable under study is not normal, the **average** will be. Our knowledge of the normal distribution will now allow us to draw **inferences** from our sample data about the population parameters.

Because the sampling distribution of the mean will be normally distributed, we can sensibly talk of the sample standard deviation of the mean, which is called the **standard error**.

Central Limit Theorem: unknown σ

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

Central Limit Theorem: unknown σ

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

Standard error is $\frac{s}{\sqrt{n}}$.

Example

```
nes <- read.table("nes.Rdata")
age <- 2002 - nes$birth.year

xbar <- mean(age, na.rm=TRUE)
s <- sd(age, na.rm=TRUE)
n <- length(age)
std.err <- s / sqrt(n)
```

Example

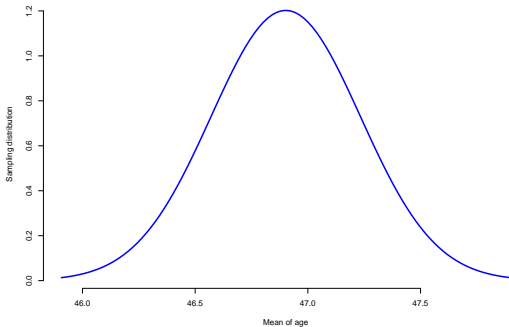


Figure: Sampling distribution of mean(age)

Outline

- 1 Sampling
- 2 Central Limit Theorem
 - Goal
 - Assumptions
 - Theorem
 - **Movie**

Central Limit Theorem: an R movie

```
sample.mean <- NULL
for (i in 1:200) {
  age.sample <- sample(age, 100)
  sample.mean[i] <- mean(age.sample)
  hist(sample.mean)
}
```

Central Limit Theorem: an R movie

```
> sd(sample.mean)
[1] 1.685797
```

Central Limit Theorem: an R movie

```
> sd(sample.mean)
[1] 1.685797
```

```
> sd(age)
[1] 17.12367
```

Central Limit Theorem: an R movie

```
> sd(sample.mean)
[1] 1.685797
```

```
> sd(age)
[1] 17.12367
```

```
> sd(age)/sqrt(100)
[1] 1.712367
```

Simulation: CLT example

```
plot(0,0,xlim=c(40,54),ylim=c(0,1),yaxt="n",bty="n",  
     xlab="Density",ylab="Sample mean of age",type="n")
```

Simulation: CLT example

```
plot(0,0,xlim=c(40,54),ylim=c(0,1),yaxt="n",bty="n",
     xlab="Density",ylab="Sample mean of age",type="n")

for (sample.size in c(10,20,100,1000)) {
  mean.sample <- NULL
  for (i in 1:100) {
    age.sample <- sample(age, sample.size)
    mean.sample[i] <- mean(age.sample)
  }
  lines(density(mean.sample), lwd=2)
}
```

Simulation: CLT example

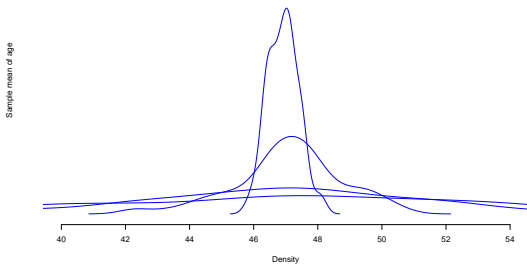


Figure: Demonstration of effect of sample size on CLT