

# Central Limit Theorem

Jos Elkind

October 10, 2007

1 Central Limit Theorem

2 Simulation and bootstrapping

# 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**.

# 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.

Typical example of non-independent observations: **time-series** data.

## Identically distributed

All the observations are draws from a **random variable** with the **same distribution and parameters**.

## 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.

# 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: 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**.

Most statistics in political science are frequentist in nature, but other methods of inference exist, such as **Bayesian** inference.

## Central Limit Theorem: unknown $\sigma$

When the population variance,  $\sigma$ , is unknown, we can use the sample estimate:

## Central Limit Theorem: unknown $\sigma$

When the population variance,  $\sigma$ , is unknown, we can use the sample estimate:

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

## Central Limit Theorem: an R movie

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

## Central Limit Theorem: an R movie

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

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 and bootstrapping

Used for:

- Gaining **intuition** about distributions and sampling

# Simulation and bootstrapping

Used for:

- Gaining **intuition** about distributions and sampling
- Providing **distribution** information not directly available

# Simulation and bootstrapping

Used for:

- Gaining **intuition** about distributions and sampling
- Providing **distribution** information not directly available
- Acquiring **uncertainly** estimates

# Simulation and bootstrapping

Used for:

- Gaining **intuition** about distributions and sampling
- Providing **distribution** information not directly available
- Acquiring **uncertainly** estimates

Both simulation and bootstrapping are **numerical approximations** of the quantities we are interested in. Run the same code twice, and you get different answers!

## 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

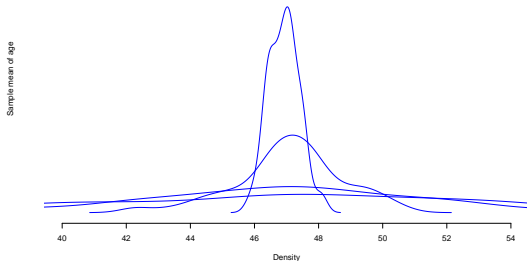


Figure: Demonstration of effect of sample size on CLT

# Bootstrapping

In the above we used simulations to **sample** from a population and get an idea of how the Central Limit Theorem operates.

# Bootstrapping

In the above we used simulations to **sample** from a population and get an idea of how the Central Limit Theorem operates.

This is only one way of using simulations; other examples will follow later in the course.

# Bootstrapping

In the above we used simulations to **sample** from a population and get an idea of how the Central Limit Theorem operates.

This is only one way of using simulations; other examples will follow later in the course.

The **bootstrap** is similar to this type of simulations, whereby the sampling takes place **with replacements**.

# Bootstrapping

The purpose of bootstrapping is to get an estimate of the variation, i.e. the **standard error**, of an estimate.

# Bootstrapping

The purpose of bootstrapping is to get an estimate of the variation, i.e. the **standard error**, of an estimate.

For example, the variance around the estimate of a mean.

# Bootstrapping

The purpose of bootstrapping is to get an estimate of the variation, i.e. the **standard error**, of an estimate.

For example, the variance around the estimate of a mean.

(The bootstrap is a **nonparametric** approach to estimating the standard error on your estimate, since no assumptions are made about the distribution of the underlying data.)

## Bootstrapping: example

```
mean.age <- NULL
for (i in 1:1000) {
  age.bootstrap <- sample(age, 100, replace=TRUE)
  mean.age[i] <- mean(age.bootstrap)
}
summary(mean.age)
```

## Bootstrapping: example

```
mean.age <- NULL
for (i in 1:1000) {
  age.bootstrap <- sample(age, 100, replace=TRUE)
  mean.age[i] <- mean(age.bootstrap)
}
summary(mean.age)

quantile(mean.age, c(.05, .95))
```

## Bootstrapping: exercise

- Using the example code, calculate the mean and standard error of this mean for the thermometer scores of Bertie Ahern.
- Calculate the correlation between age and the thermometer score of Fianna Fail and the standard error on this correlation.

Basic estimators:

```
mean(2002 - nes$year.birth, na.rm=TRUE)  
cor(nes$year.birth, nes$thermo.ff, use="complete.obs")
```