Tutorial 7

Research Methods for Politcal Science A

Michele McArdle

24 & 25 October 2020

Outline

1. Central Limit Theorem
2. Standard Error
3. Confidence Intervalls

Central Limit Theoreme

The central limit theorem states that if **random samples** are taken from the population, then the sample means will be **normally distributed**.

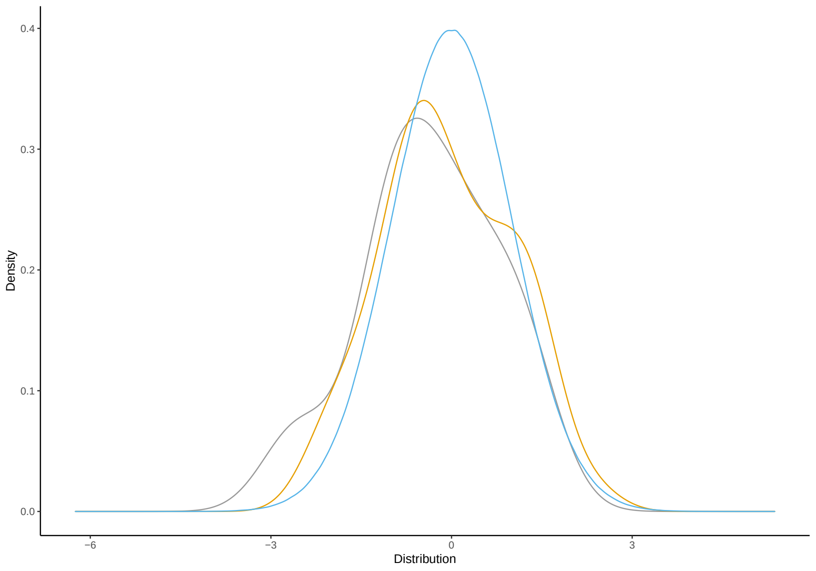
This will hold true regardless of whether the source population is normal or skewed, provided the **sample size is sufficiently large ($n > 30$)**.

Importance of the Sample Size

The larger the sample the better it is.

A larger sample means:
- that your sample means are more “normally” distributed.
- that your standard error will, be smaller.
- that your findings will be more accurate and more statistically significant.

Effect of Sample size on Sampeling distribution



Standard Normal Distributions

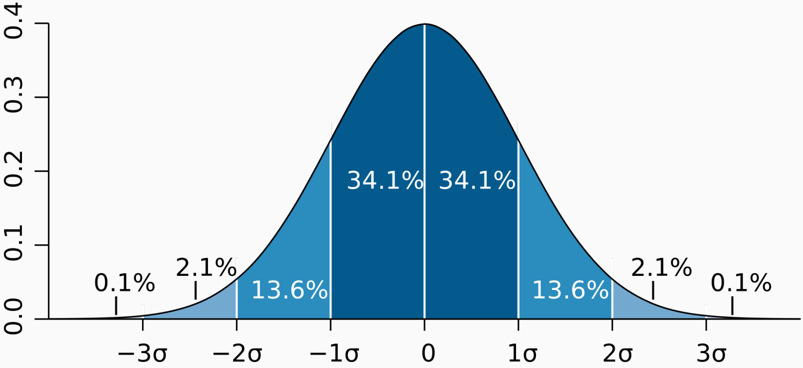


Figure 1: Std. Normal Distribution

Standard Error

The standard error is the standard deviation of the distribution of the sample means.

Standard Error

```r
# Generate Random IQ data for TCD undergraduates
# We treat this as our statistical population
population <- rnorm(n = 11718, mean = 100, sd = 15)
mean(population); sd(population)
```

```
## [1] 100.2399
```

```
## [1] 14.97848
```

Imagine this to be a dataset of the IQs of all TCD undergraduates. This is what in statistics we call the population.

This means that $\mu = 100.06$ and $\sigma = 14.96$. We would normally not know this as we can't observe the population in real life.

Drawing a sample

```r
sample.df <- sample(population, size = 500)
mean(sample.df); sd(sample.df)
```

```
## [1] 100.6606
```

```
## [1] 14.78655
```

Imagine this is a sample we took testing 500 students. This is the data set we will be working with. How can we test whether the mean of our sample is equal to the mean of the population.

Calculating the Standard Error from one sample

The formula to calculate the standard error from a single sample is:

$$se = \frac{s}{\sqrt{n}} = \frac{std.\ dev.\ of\ the\ sample}{\sqrt{number\ of\ observations}}$$

Calculate the Standard Error in R

```r
se <- sd(sample.df)/sqrt(500)

se
```

```
## [1] 0.6612745
```

Confidence Intervalls

We can now use the SE to calculate the confidence interval for our population mean. The CI is the range of possible values the population mean can take. We usually use the 95%-CI.

Confidence Intervalls

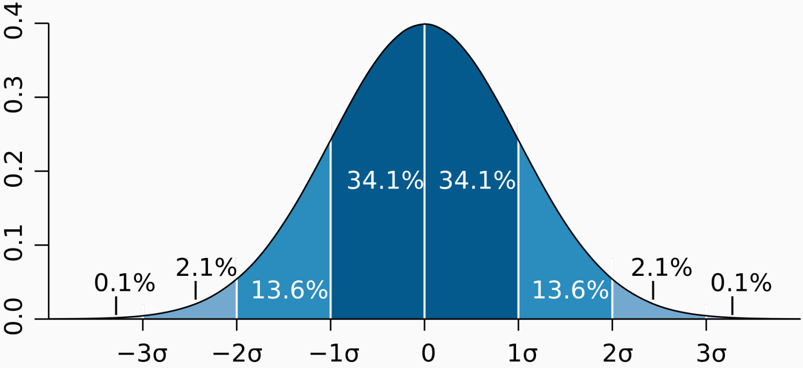


Figure 2: Std. Normal Distribution

Confidence Intervalls

$$CI_{68\%} = \bar{x} \ \pm se$$

$$CI_{95\%} = \bar{x} \ \pm 2se$$

$$CI_{99\%} = \bar{x} \ \pm 3se$$

Confidence Intervalls in R

Is the mean of the population 105?

```
sample1 <- sample(population, size = 50)
sample2 <- sample(population, size = 500)

avg1 <- mean(sample1)
se1  <- sd(sample1)/sqrt(50)

CI68_low  <- avg1 - se1
CI68_high <- avg1 + se1

CI68_low; CI68_high
```

```
## [1] 95.67509
```

```
## [1] 99.42594
```

Confidence Intervalls in R

```r
avg2 <- mean(sample2)
se2  <- sd(sample2)/sqrt(500)

CI68_low  <- avg2 - se2
CI68_high <- avg2 + se2

CI68_low; CI68_high
```

```
## [1] 99.8131
```

```
## [1] 101.1282
```

Exercise

Calculate the CI-95% and CI-99%.

Exercise

```r
CI95_low  <- avg1 - 2*se1
CI95_high <- avg1 + 2*se1

CI95_low; CI95_high
```

```
## [1] 93.79967
```

```
## [1] 101.3014
```

```r
CI99_low  <- avg1 - 3*se1
CI99_high <- avg1 + 3*se1

CI99_low; CI99_high
```

```
## [1] 91.92425
```

```
## [1] 103.1768
```

Exercise

```r
CI95_low  <- avg2 - 2*se2
CI95_high <- avg2 + 2*se2

CI95_low; CI95_high
```

```
## [1] 99.15556
```

```
## [1] 101.7857
```

```r
CI99_low  <- avg2 - 3*se2
CI99_high <- avg2 + 3*se2

CI99_low; CI99_high
```

```
## [1] 98.49802
```

```
## [1] 102.4433
```