Statistical Inference Course Project - Part I

Steven Michiels

4/16/2020

Overview

This document reports on the simulation project of the Johns Hopkins Statistical Inference Course.

The project essentially is a direct application of the important **central limit theorem (CLT)**. The CLT states that when drawing a simple random sample of sufficiently large size n from any population with mean μ and standard deviation σ, the sampling distribution of the sample mean is approximately normally distributed N(μ, σˆ2/n) (Moore et al., Introduction to the Practics of Statistics).

In this project, a sufficiently large sample (n=40) was drawn many times (1000 samples) from an exponential distribution. The resulting sampling distribution of the mean indeed is approximately normally distributed with the mean and variance very close to the expected values. In practice this theorem allows us to estimate the population mean and standard deviation from any pouplation distribution based on a sufficiently large sample.

Simulations

We define the **rate** λ of the exponential distribution, as well as the **sample size n**. The **population mean and standard deviation** are **known** from the rate λ.

```
set.seed(3422)
n=40
lambda=.2
population_mean=1/lambda
population_variance=(1/lambda)^2
```

We then **compare** the **mean of a drawn sample** with the **known population mean**.

```
sample1=setNames(data.frame(rexp(n,lambda)), "value")
sample_mean1=mean(sample1$value)
```

We **plot** the **drawn sample**, together with the **actual distribution**.

```
g1=ggplot(sample1,aes(x=value))
g1=g1+geom_histogram(aes(y=..density..),color="orange", fill="white",binwidth=2)
g1=g1+xlab("value")+ggtitle("Histogram for a single sample of size 40")
g1=g1+geom_vline(aes(xintercept=sample_mean1),
           color="orange", linetype="dashed", size=2)
g1=g1+geom_vline(aes(xintercept=population_mean),
           color="blue", linetype="dashed", size=2)
g1=g1+stat_function(fun = dexp, args = list(rate=lambda),color="blue", size=1)
g1=g1+annotate("text", x = 10.5, y = .25, label = "Sampled exponential (n=40)", color="orange", size=5)
g1=g1+annotate("text", x = 10.5, y = .2, label = "Theoretical exponential (rate = 0.2)", color="blue", 
plot(g1)
```

The previously drawn **large sample** provides us with an **estimate for the population mean**. We now **show** that the **sample mean is approximately normally distributed N(μ, σˆ2/n)**, by drawing a **large number (1000 samples) of samples** of large size (n=40).

```
simul=1000
sample2=setNames(data.frame(rep(0,simul)), "value")
for (i in 1 : simul) sample2[i,] = mean(rexp(n,lambda))
```

We **compare** the **mean of the sampling distribution** with the **population mean**. We also calculate the variance of the sample mean and **derive** from this an **estimate for the variance of the population**.

```
sample_mean2=(mean(sample2$value))
sample_variance=var(sample2$value)
population_variance_estimate=sample_variance*n
```

We calculate the **95% confidence intervals** for the **mean of the sampling distribution** and for a **perfect Gaussian N(μ, σ^2/n)**.

```
sample_lower_95percent_bound=sample_mean2 - qnorm(.975)*sqrt(sample_variance)
sample_upper_95percent_bound=sample_mean2 + qnorm(.975)*sqrt(sample_variance)
gaussian_upper_95percent_bound=population_mean + qnorm(.975)*sqrt(population_variance/n)
gaussian_lower_95percent_bound=population_mean - qnorm(.975)*sqrt(population_variance/n)
```

And we create a **density plot** for the sampling distribution and for a perfect Gaussian N(μ, σ^2/n).

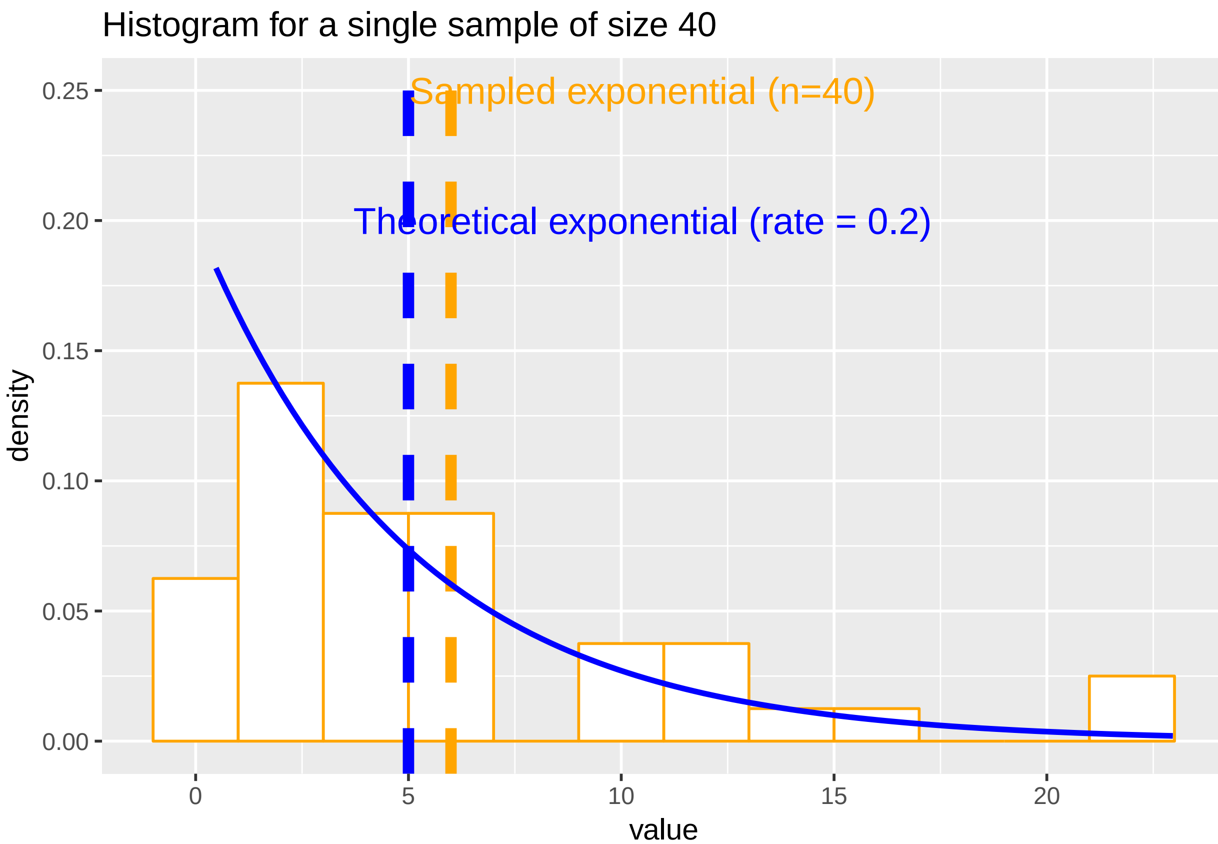
```
g2=ggplot(sample2,aes(x=value))
g2=g2+geom_histogram(aes(y=..density..),color="orange", fill="white",binwidth=.3)
g2=g2+geom_density(colour="orange", size=1)
g2=g2+xlab("value")+ggtitle("Histogram of the mean for a 1000 samples of size 40")
g2=g2+geom_vline(aes(xintercept=sample_mean2),
            color="orange", linetype="dashed", size=1)
g2=g2+geom_vline(aes(xintercept=population_mean),
            color="blue", linetype="dashed", size=1)
g2=g2+annotate("text", x = 7, y = .55, label = "Distribution for sample (n=40)", color="orange", size=5)
g2=g2+annotate("text", x = 7, y = .5, label = "N((1/lambda),(1/lambda)^2/n)", color="blue", size=5)
g2=g2+stat_function(fun = dnorm, args = list(sd=sqrt(population_variance/n),mean=population_mean),color=
plot(g2)
```

Sample Mean versus Theoretical Mean

We compare the mean of a drawn sample with the known population mean. Later on we will determine the interval that will contain the mean of such a random sample with 95% confidence.

```
##      sample_mean1 population_mean
## [1,]     5.999797               5
```

We plot the drawn sample, together with the actual distribution.



Sample Variance versus Theoretical Variance

Performing a **large number of samples** with a large size allows us to **determine the variance of the sampling distribution of the mean**. The **mean of the sampling distribution of the mean** is **very close to the population mean**. Moreover, the **estimated population variance**, estimated from the variance of the sampling distribution of the mean, is **very close to the actual population variance**.

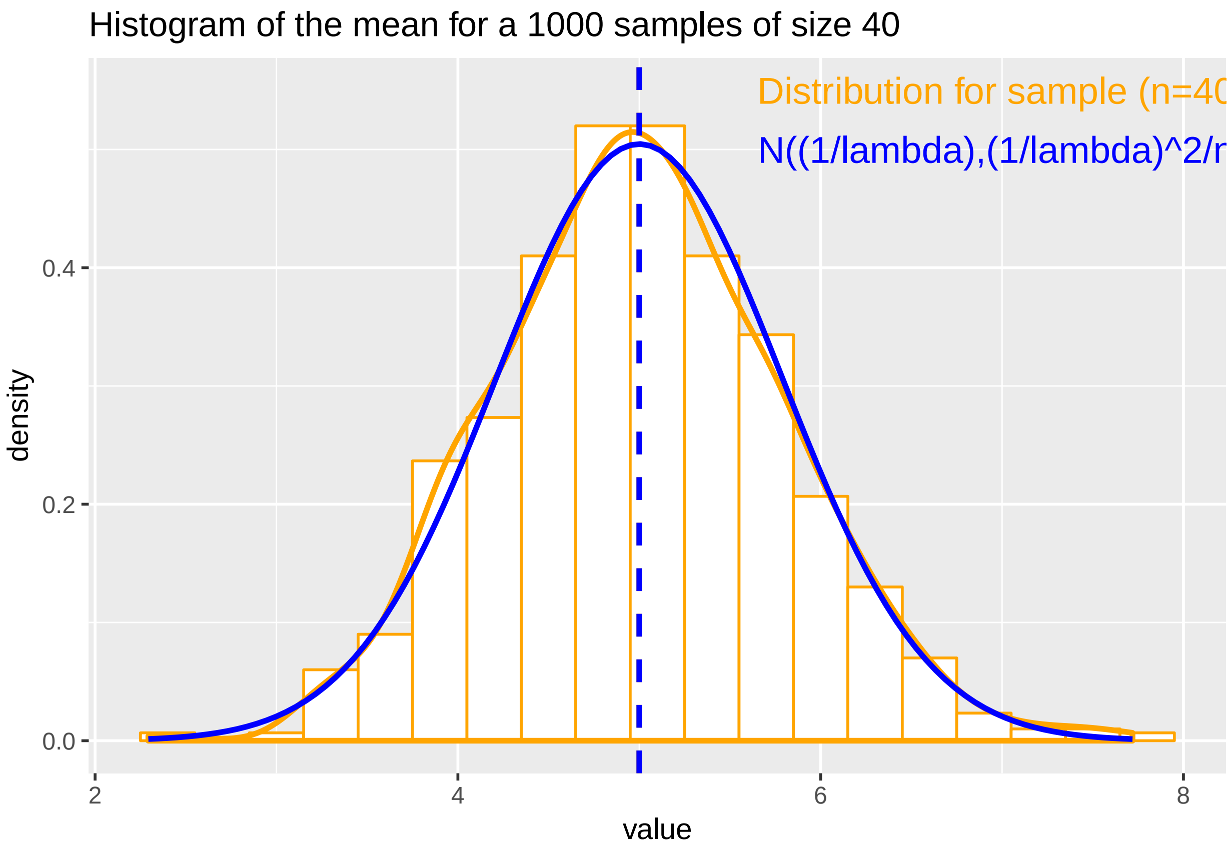
```
##      sample_mean2 population_mean
## [1,]     5.000119               5
```

```
##      sample_variance
## [1,]       0.6265141
```

```
##      population_variance_estimate population_variance
## [1,]                     25.06056                  25
```

Distribution

The **obtained distribution of the sample mean** looks **very similar to a perfect Gaussian N(μ, σ^2/n)**.



In addition, the **95% confidence interval for the obtained distribution of the sample mean** is **very close to the 95% confidence interval for a perfect Gaussian N(μ, σ^2/n)**.

```
##      population_variance population_variance_estimate
## [1,]                  25                     25.06056
```

```
##      sample_lower_95percent_bound gaussian_lower_95percent_bound
## [1,]                     3.448755                       3.450512
```

```
##      sample_upper_95percent_bound gaussian_upper_95percent_bound
## [1,]                     6.551482                       6.549488
```

We therefore **confirmed the central limit theorem with an example**: when drawing a simple random sample of sufficiently large size n from any population with mean μ and standard deviation σ, the sampling distribution of the sample mean indeed is approximally normally distributed N(μ, σ^2/n).