SimulationReport

bbneo

03/18/2015

Simulation Exercise:

The distribution of a sample mean statistic with reference to the population mean and standard deviation and its relation to the `t` and `Z` distributions.

Overview

In this report, we will illustrate the nature of the distribution of the `n =` 40 *sample mean* statistic from an exponential population distribution. What we will see is that:

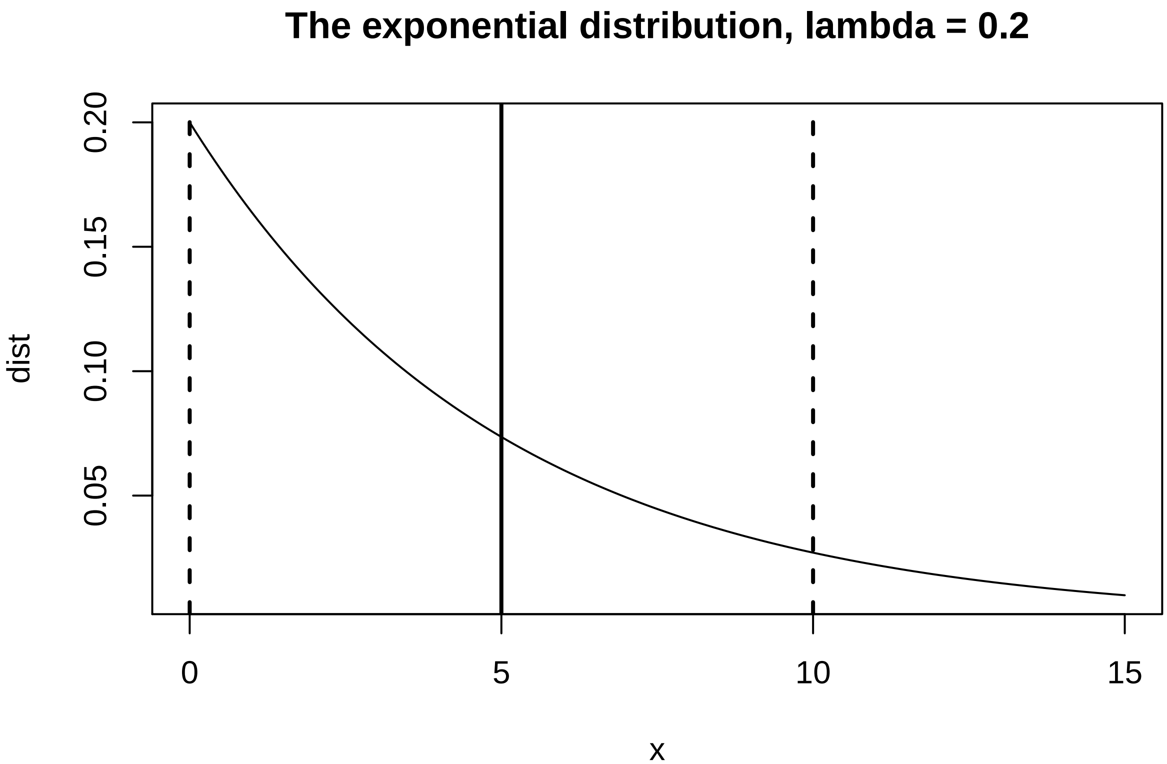
- the *sample mean* statistic can be a good estimator of the mean of the population (exponential) distribution,
- the distribution of the sample means approaches a `t` or `Z` distribution as the sample size increases as expected from the Central Limit Theorem,
- the `standard` *`error`* of this population mean estimator (`SE`*`mean`*) depends on the `standard deviation` of the sample means and the sample size, and
- the `standard error` of this *mean estimator* has a much narrower confidence interval for the **`population mean`** estimate than either
  - the *standard deviation* of the distribution of the sample means, or
  - the *standard deviation* of the population distribution itself

The `exponential` distribution

The population distribution we will use for this illustration is the *exponential* distribution. The `exponential` distribution is implemented in R as `dexp/rexp`.

The distribution has one parameter, the `rate` or `lambda`, and both the theoretical `mean` **`and`** `standard deviation` (of random *single* samples) of the (*population*) distribution are `1 / lambda`.

For `lambda =` 0.2, `rexp` looks like this:



The population `mean` is represented as solid vertical line at 5, and the dashed lines represent `+/- 1 standard deviation` of the distribution at 0, 10. Notice that this distribution is skewed and does not have a tail approaching zero on the left.

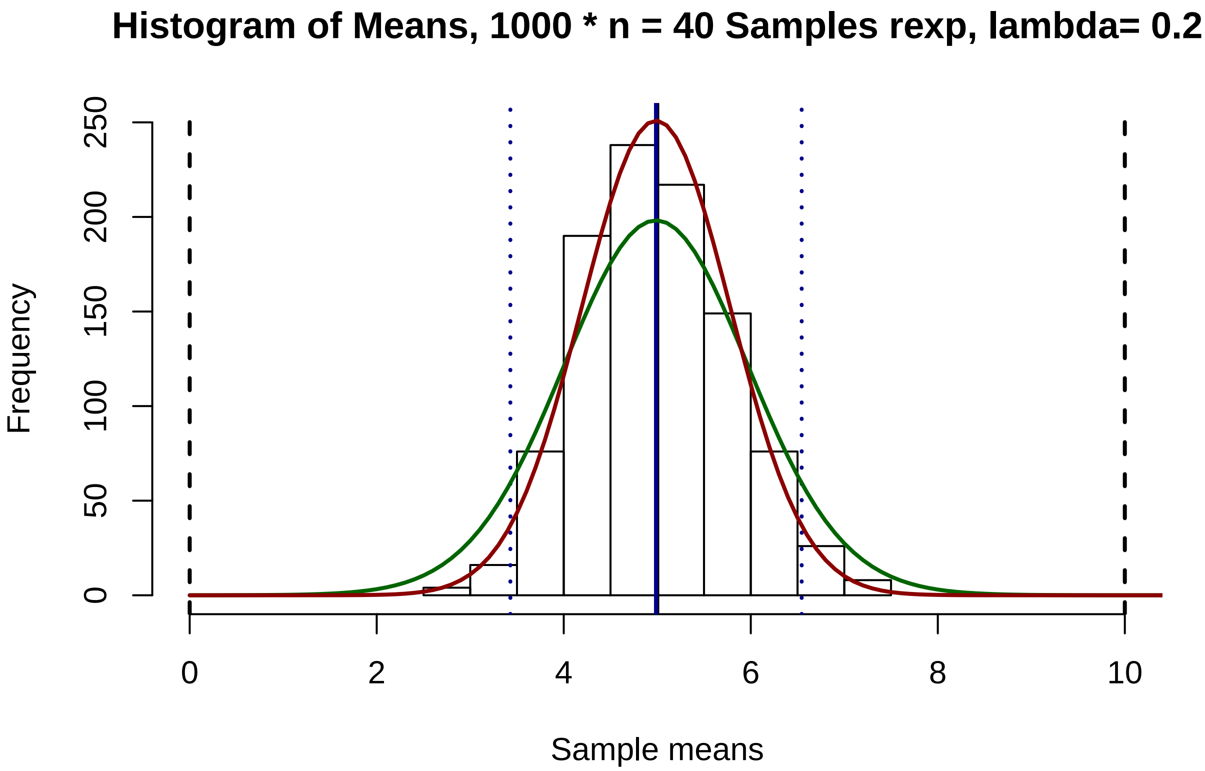
Simulation and the *sample mean* as an *estimator* of the *population mean*

We will take `1000 *` (`n =` 40) samples from the popultation distribution `rexp`, and use these samples to estimate the *population mean* and the *standard error* of that *mean* estimate (`SE`*`mean`*). We create these samples in R by successively adding `n =` 40 *sample means* to a vector `mns`.

```
mns <- NULL
for (i in 1:1000) mns <- c(mns, mean(rexp(n,lambda)) )
```

The distribution of the *sample means*

The distribution of these `sample means` approaches a `t` or `Z` distribution as is shown in the figure below:



The `mean` of the *sample means* is 4.987 (represented as a solid vertical dark blue line) with a 95% confidence interval for the distribution of the *sample means* represented by the dotted dark blue vertical lines on the figure at:

```
sampleMean + c(-1,1)*qnorm(0.975)*sampleSd
```

```
## [1] 3.43 6.54
```

The *population mean* and standard deviation are represented by a vertical line at `x = `5, and dashed lines at `x = `0, 10. The population distribution for **single** random iid samples from an exponential distribution would give the very wide `95%` *standard* quartile -1.96, 1.96 `* theoreticalSd` interval of **-4.8, 14.8 !!**

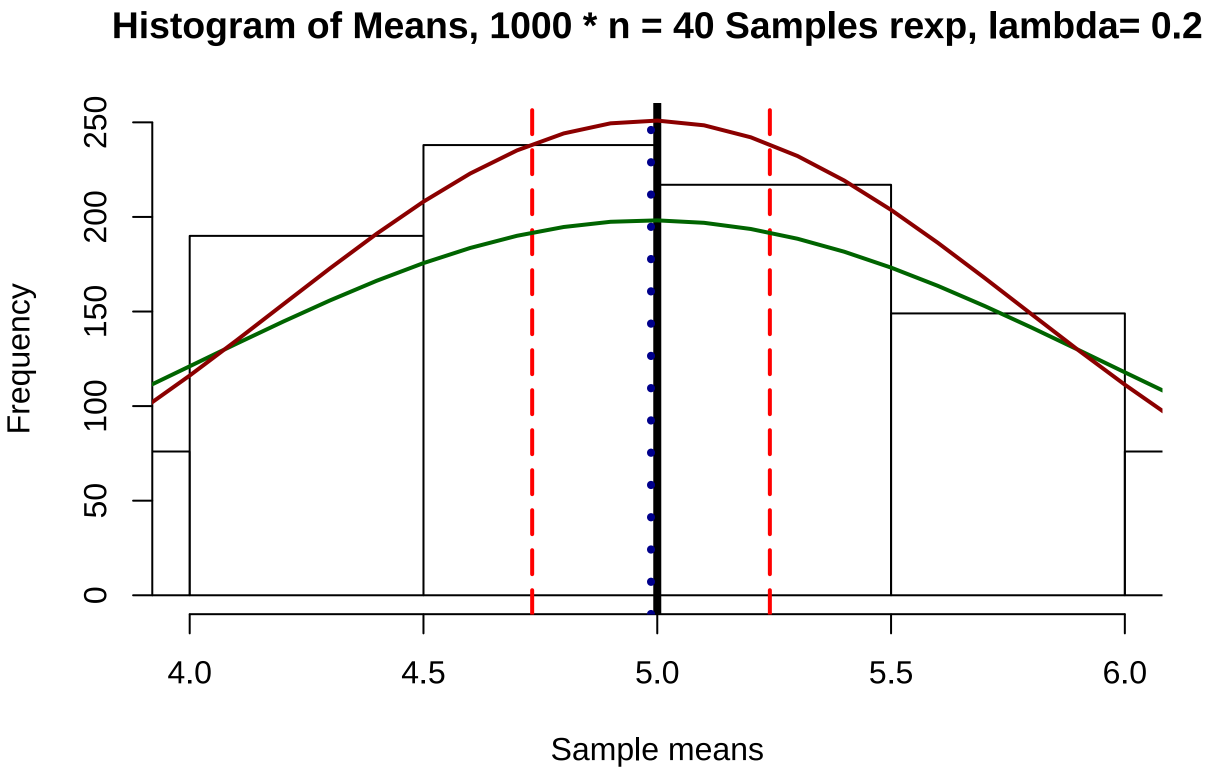
But the skewness of the `exponential` distribution makes the normal distribution a poor tool for describing its variation.

A `t density distribution` with `df=` 39 is overlaid in dark green.

A `normal` density distribution with mean and sd corresponding to the sample distribution is overlaid in dark red.

A closer look at the confidence interval for the population mean estimate

We will zoom in on the 95% confidence interval of the *sample mean* and its `SE`***`mean`*** as an estimate of the *population mean* of `rexp`.



The 95% confidence interval for the estimation of the `population (theoretical) mean` from this set of samples using a `t` statistic is:

```
SEmean <- sampleSd/sqrt(n)
SEmean
## [1] 0.126

qt(0.975,n-1)
## [1] 2.02
sampleMean + c(-1,1)*qt(0.975,n-1)*SEmean
## [1] 4.73 5.24
```

which is represented in the figure by vertical red dashed lines.

The 95% CI for the *sample mean* as an estimator of the *population mean* is (using a Z quartile):

```
qnorm(0.975)
## [1] 1.96
sampleMean + c(-1,1)*qnorm(0.975)*SEmean
## [1] 4.74 5.23
```

Which is very close to the 95% CI obtained using a `t` statistic, just slightly more narrow.

In Summary

For this report, for `n = 40 > 20-30`, the distribution of sample means approaches a normal `Z` distribution:

statistic	mean	sd	95% CI
population theoretical	5	5	–
dist. of sample mean	4.987	0.795	3.429, 6.544
pop. mean *estim*	4.987		4.74, 5.233

- the *sample mean* statistic can be a good estimator of the mean of the population (exponential) distribution,
- the distribution of the sample means approaches a `t` or `Z` distribution as the sample size increases as expected from the Central Limit Theorem,
- the `standard` *`error`* of this population mean estimator (`SE`*`mean`*) depends on the `standard deviation` of the sample means and the sample size, and
- the `standard error` of this *mean estimator* has a much narrower confidence interval for the **`population mean`** estimate than either:
    - the *standard deviation* of the distribution of the *sample* means, or
    - the *standard deviation* of the *population distribution itself* (the distribution of a large collection of exponentially distributed random numbers)