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:

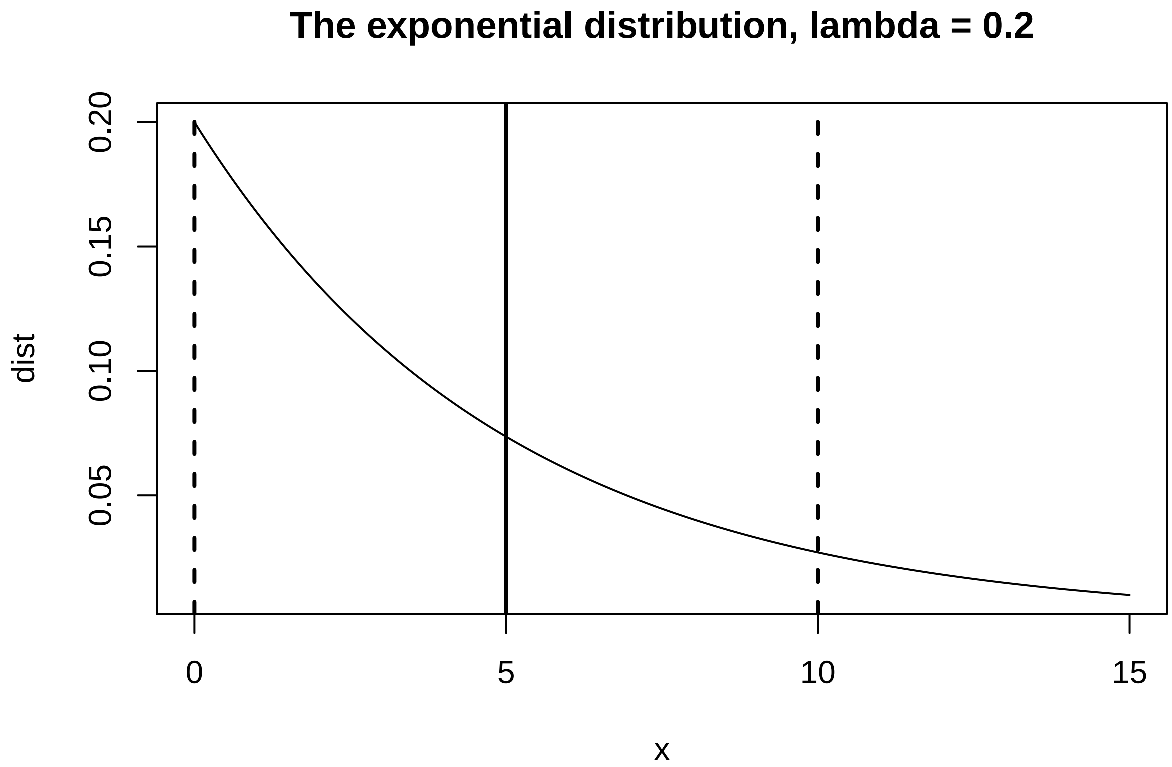
- the *sample mean* statistic can be a good estimator of the mean of the population (exponential) distribution,
- the distribution of the sample means approaches a `t` or `Z` distribution as the sample size increases as expected from the Central Limit Theorem,
- the `standard` *`error`* of this population 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.

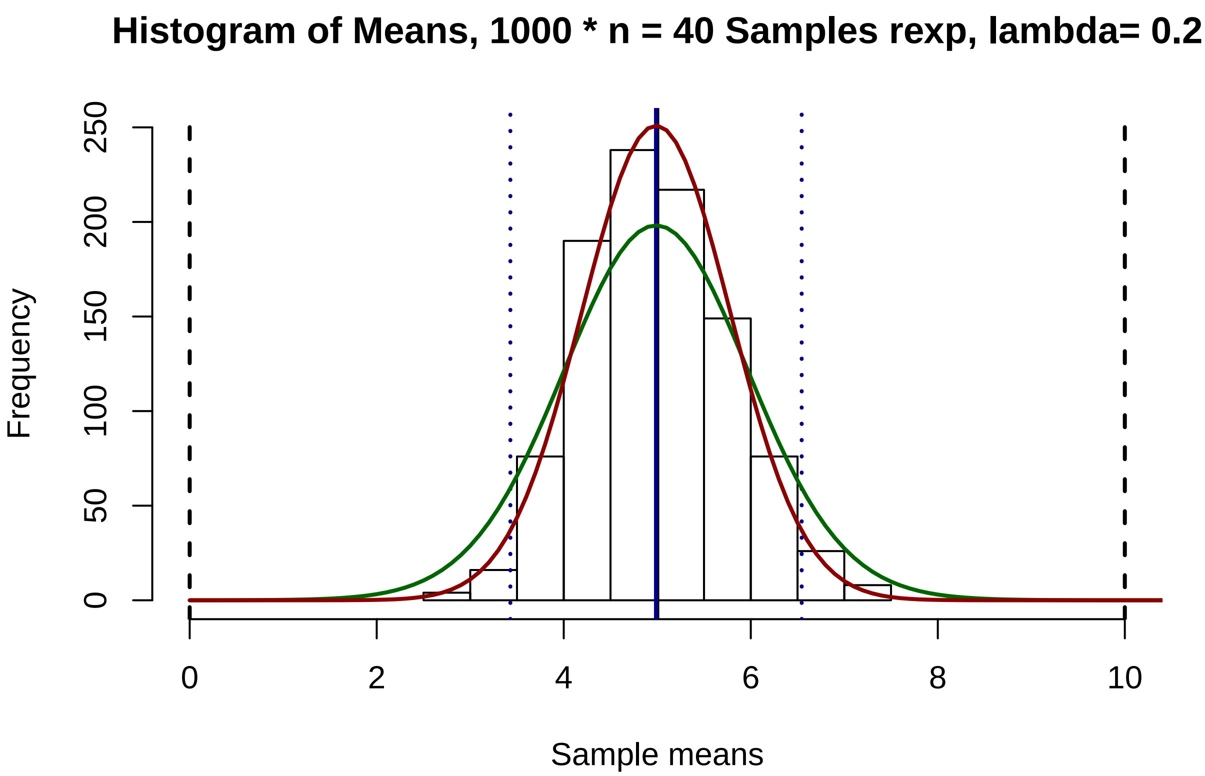
Simulation and the *sample mean* as an *estimator* of the *population mean*

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 4.9866197 (represented as a solid vertical dark blue line) with a 95% confidence interval for the *sample means* represented by the dotted dark blue vertical lines on the figure at:

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

```
## [1] 3.428875 6.544364
```

The *population mean* and standard deviation are represented by a vertical line at `x = 5`, and dashed lines at `x = 0, 10`.

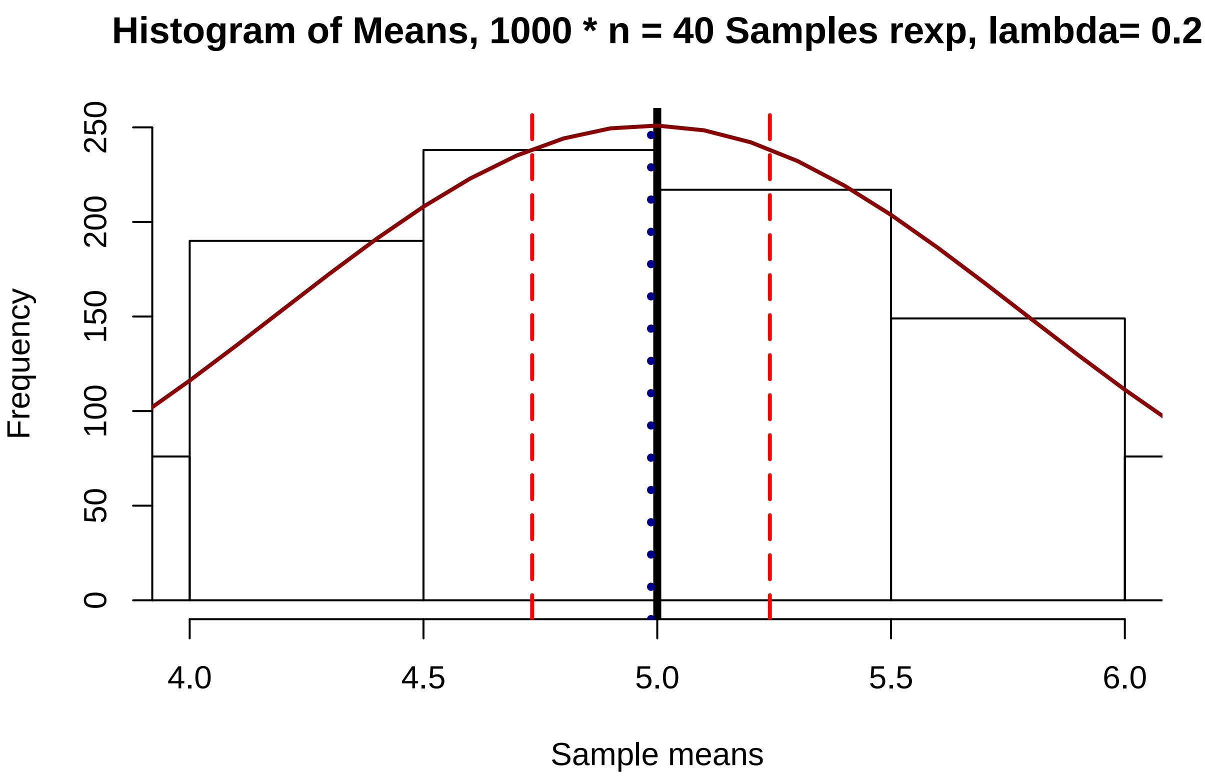
Notice that these population variation lines are just **one** population `sd` about the mean… as opposed to a normal distribution `95%` quartile -1.959964, 1.959964 `* sd` interval.

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:

```
sampleMean + c(-1,1)*qt(0.975,n-1)*sampleSd/sqrt(n)
```

```
## [1] 4.732436 5.240803
```

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

In Summary

For this report, `n` = 40:

statistic	mean	sd	95% CI
population theoretical	5	5	–
sample mean	4.9866197	0.7947823	3.428875, 6.5443643
pop. mean *estim*	4.9866197		4.7403186, 5.2329207

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

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

```
## [1] 4.740319 5.232921
```

`qnorm(0.975)` = 1.959964, and the `SE`*`mean`* is:

```
sampleSd/sqrt(n)
```

```
## [1] 0.1256661
```

- 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