The Central Limit Theorem, a Swiss Army Knife of Statistics

Linda Angulo Lopez

29/12/2020

Introduction

The project consists of two parts (i) a simulation exercise and (ii) a basic inferential data analysis, the former is presented here.

Asymptotics form the basis for frequency interpretation of probabilities, where the behavior of statistics depends on the sample size or some other relevant quantity of limits to infinity or to zero. These limits are the the swiss army knives of statistics, Brian Caffo.

Simulations were made to investigate the asymptotic distributions of exponential distributions, a discreet case, and compared to test statistics which are expected to be Gaussian, a strong form of the Central Limit Theorem, in R4.0. Results show that like with greater $n$in the CLT, with an increase the unit of time,λ, the coverage improves and adheres to the CLT.

```
library(knitr) # creating the document
library(ggplot2) # making plots
```

`key words:`Asymptotics, The Central Limit Theorem, Sample Mean versus Theoretical Mean, Sample Variance, versus Theoretical Variance, Distributions.

Asymptotics

The Law of Large Numbers, is intuitive it says that if we collect an infinite amount of data we get close to the right answer. That is our sample means, the sample variance and the sample standard deviation of the iid random variables are consistent.

The Central Limit Theorem, CTL

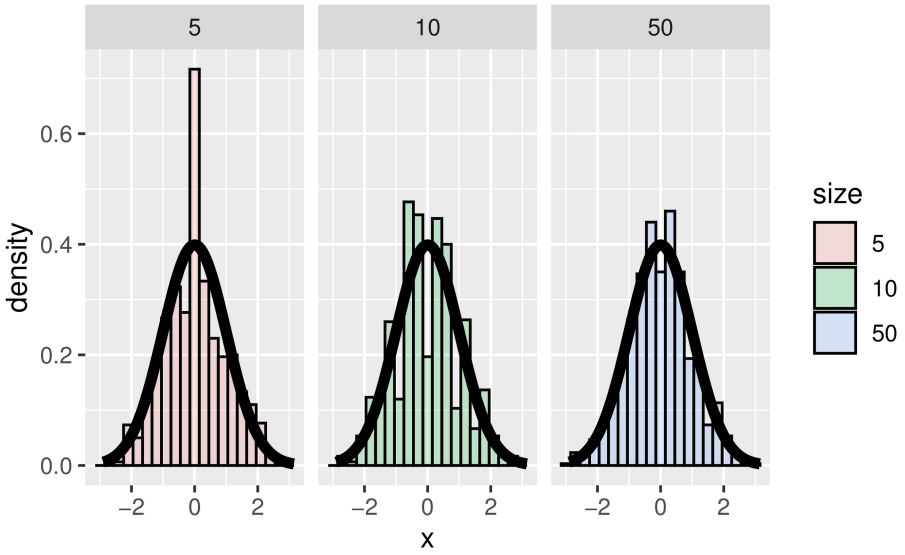
The CLT establishes that in a frequency plot, the properly normalized sum tends toward a normal distribution, a bell curve, even if the original variables themselves are not normally distributed, which implies that `probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions`.

The CLT says, for a large n, this normalized variable, $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ is almost normally distributed with a `mean of 0 and variance of 1`:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\text{Estimate} - \text{Mean of estimate}}{\text{Std. Err. of estimate}}.$$

- Here is the case of an unbiased coin flip experiment, p =.5 is presented.

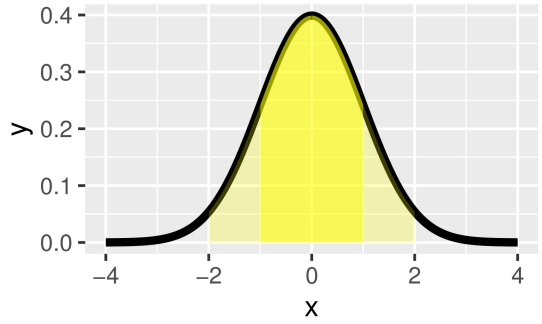
Simulations of Coverage in the CLT



The output depends on R's random number generator, but as the sample size increases from n= (5,10,50) the density curve for the blue plots n=50 adheres to the CLT, it is bell shaped, and so it is consistent as it converges to what it's trying to estimate, this is `a strong case for the CLT`.

Confidence intervals:

From the CLT we also know that 95% of the area under a normal curve is within two standard deviations of the mean.



The entire shaded portion of the plot depicts the area within 2 standard deviations of the mean and the darker yellow portion shows the 68% of the area within 1 standard deviation, we use the R function qnorm to find the 95th quantile for a standard normal distribution.

Simulations of Coverage for an exponential distribution

When dealing with distributions which apply to counts or rates, λ, the rate per unit of time or space at which the event occurs, is the limiting factor. Taking the course example, suppose a nuclear pump failed 5 times out of 94.32 days and we want a 95% confidence interval for the failure rate per day, we can estimate the failure rate = 5/94.32:

$$\frac{\lambda}{t}$$

where λ is our estimated mean and t is time, and we can verify this in R.

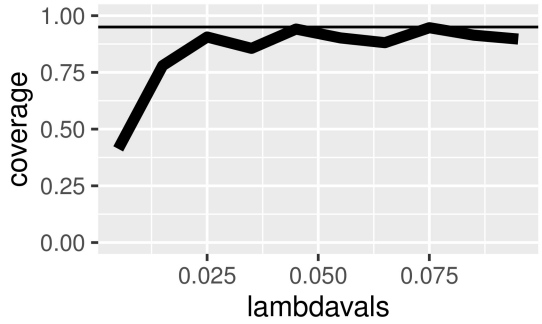
```
poisson.test(5,94.32)$conf
```

```
## [1] 0.01721254 0.12371005
## attr(,"conf.level")
## [1] 0.95
```

```
#The formula we've used to calculate a 95% confidence interval is
#est mean + c(-1,1)*qnorm(.975)*sqrt(est var).
```

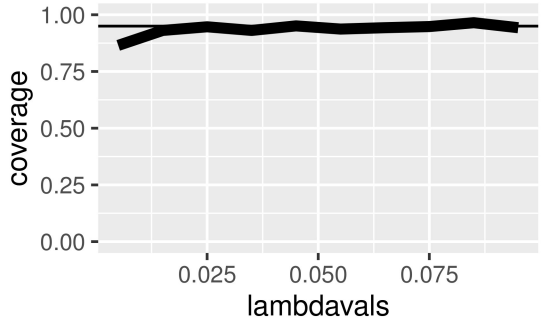
To check the coverage of our estimate we can run a simulation experiment and vary λ from .005 to .1 with steps of .01 (so we have 10 of them), and for each one we'll generate 1000 Poisson samples with mean λ times t. Calculate sample means and use them to compute 95% confidence intervals. We'll then count how often out of the 1000 simulations the true mean (our lambda) was contained in the computed interval.

- t =100



We see that like with greater $n$in the CLT, with the coverage improves when we increase the unit of time, λ. In the above plot we used t=100 (rounding the 94.32 up). Below is a plot of the same experiment setting t=1000. We see that the coverage is much better for almost all the values of lambda, except for the smallest ones.

- t = 1000



Afternotes:

- This is an R Markdown document,created as a submission for the Statistical Inference Course by Johns Hopkins University on Coursera.
- Note that the `echo = FALSE` parameter was added to several code chunks to prevent printing of the R code that generated the plots, you can find similar code in the swirl repo on Github.