Sampling Statistics

Diego-MX

March 22, 2015

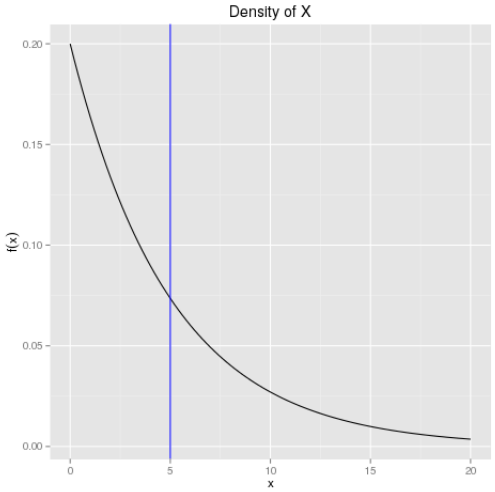
In this project we will illustrate via sampling the differences between a random variable X and the corresponding variable $\bar{X}_n$, that is the mean of an n-sample of X. The main property about $\bar{X}_n$ is that it approximates a normal distribution as $n \to \infty$. We will see that we don't need to go too far -in fact we stick with $n = 40$- in order for the resemblance to be evident.

0.1 Exponential Variable X

Let $X \sim \text{Exp}(\lambda)$, its density of X has the formula

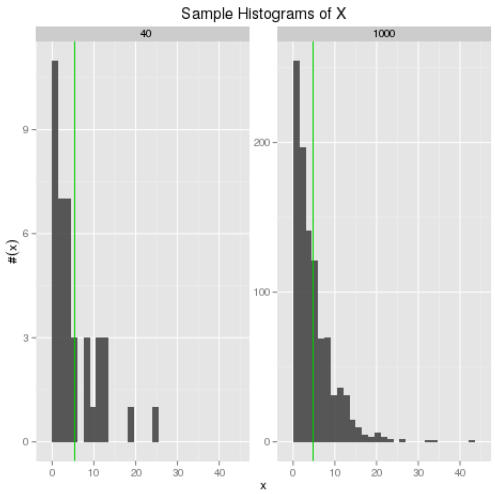
$$f(x) = \lambda e^{-\lambda x}.$$

By fixing $\lambda = 0.2$ the density looks like this.



where we have indicated the mean $\mu = \frac{1}{\lambda} = 5$ with a blue vertical line. Being an exponential this looks nothing normal... obviously.

As we take samples of X, the shape of the histogram resembles that density.

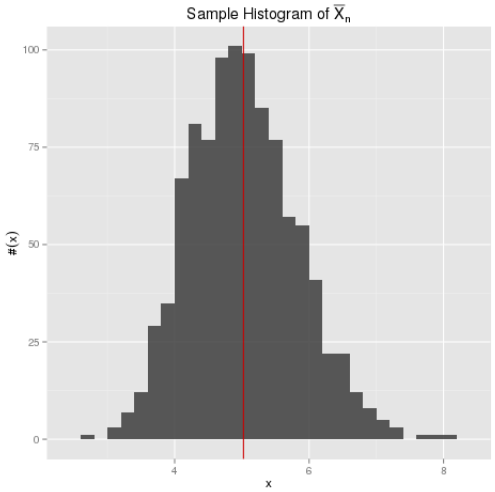


The number in the top indicates the sample size, and we can see that their shape is similar and resembles that of the density $f(x)$. We point out that their means -marked in green- are also similar, and close to the original mean $\mu = 5$.

0.2 Exponential Variable Sample Mean $\bar{X}_n$

Having shown what we did, we are ready for the real deal here. We will consider the random variable of the sample mean $\bar{X}_n$. And even more, to give a simplistic approach of how this "becomes" normal as $n \to \infty$. What we mentioned earlier is that $n = 40$ will be convincing enough.

Each realization of $\bar{X}_n$ is the average of $n = 40$ realizations of X; in our previous figure this corresponds to the green line of the left plot. We do this repeatedly 1000 times to get the following histogram; and where we indicate the mean in red.



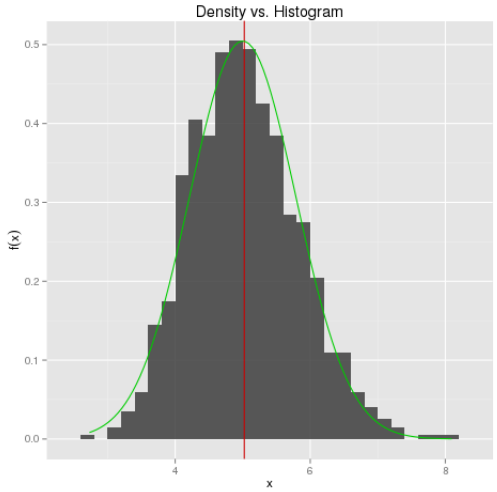
Finally, we perform the computations for the dessignated normal distribution and compare to a scaled histogram.

For the exponential variable $X \sim \text{Exp}(\lambda)$, we can compute (or look in course notes, Wolfram Alpha or Wikipedia) its mean to be $\text{E}(X) = \lambda^{-1}$ and variance $\text{Var}(X) = \lambda^{-2}$, therefore we have $\text{std}(X) = \lambda^{-1}$.

For the sample mean $\bar{X}_n = \frac{1}{n}\sum_i X_i$ we similarly have $\text{E}(\bar{X}_n) = \lambda^{-1}$ and $\text{Var}(\bar{X}_n) = \frac{1}{n}\lambda^{-2}$. In the following table we compare these to the ones obtained from the sample.

	Theoretical	Empirical
Mean	5	5.023
Variance	$\frac{25}{40} \approx 0.625$	0.637

Moreover, according to the Central Limit Theorem $\bar{X}_n$ "approaches" a normal distribution with said parameters; in or our case with $\lambda = 0.2$, $n = 40$. The following plot shows that.



Wow! This looks very neat! It is approximately normal.
Regards from Mexico.

0.3 Appendix

This is the code for the graphs above.

```
setwd(paste0("~/Sync/Dropbox/R/Coursera/",
"6 Statistical Inference/Project/src")) library(ggplot2) library(dplyr)
lambda <- 0.2
size   <- 40
reps   <- 1000
;
# Density of X
png("../output/Exp Density.png")
ggplot(data.frame(x=c(0,20)), aes(x)) +
    stat_function(fun=function(x)dexp(x,lambda), geom="line") +
    labs(title=expression("Density of"~X), y=expression(f(x))) +
    geom_vline(xintercept=1/lambda, color="blue")
nil <- dev.off()
;
# Two samples of X and their mean
xsample <- rexp(reps+size, lambda)
xsmplDF <- data.frame(x=xsample,
    n=c(rep(size,size), rep(reps,reps)))
by_n <- xsmplDF %>%
    group_by(n) %>%
    summarize(Mean=mean(x))
;
png("../output/Exp Histograms.png")
```

```
ggplot(xsmplDF, aes(x)) +
    geom_histogram(binwidth=1.5, alpha=0.8) +
    facet_wrap(~ n, scales='free_y') +
    labs(title=expression("Sample Histograms of"~X),
        y=expression("#"(x))) +
    geom_vline(data=by_n, mapping=aes(xintercept=Mean),
        color="green3")
nil <- dev.off()
;
# 1000 samples of X_bar.
expSim <- rexp(size*reps, lambda)
expMat <- matrix(expSim, nrow=size, ncol=reps)
meansDF <- data.frame(x_bar=apply(expMat, 2, mean))
;
png("../output/Exp_Bar Histogram.png")
ggplot(meansDF, aes(x_bar)) +
    geom_histogram(binwidth=0.2, alpha=0.8) +
    labs(title=expression("Sample Histogram of"~bar(X)[n]),
        y=expression("#"(x)), x="x") +
    geom_vline(data=meansDF, mapping=aes(xintercept=mean(x_bar)),
        color="red3")
nil <- dev.off()
;
png("../output/HistogramDensity.png")
ggplot(meansDF, aes(x=x_bar)) +
    geom_histogram(aes(y = ..density..), binwidth=0.2,
        alpha=0.8) +
    geom_vline(aes(xintercept=mean(x_bar)), color="red3") +
    stat_function(color="green3", fun=function(x)
        dnorm(x,1/lambda,1/lambda/sqrt(size))) +
    labs(title="Density vs. Histogram",
        y=expression(f(x)), x="x")
nil <- dev.off()
```