Statistical Inference - Assignemnt

Daniel Bader

Thu Jun 18 23:58:24 2015

In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

Exponential Distribution

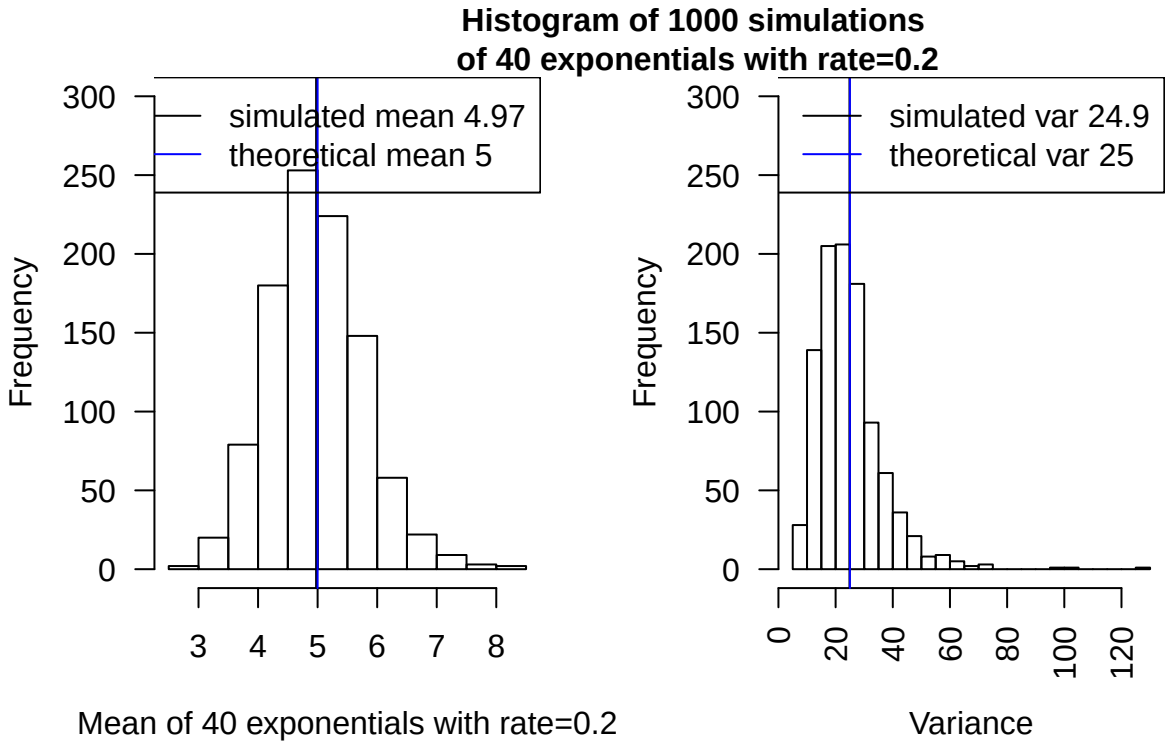
```
suppressPackageStartupMessages(library(geneplotter))
LAMBDA=0.2
nummeans=40
nsimul=1000
```

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials.

Mean and variance: Sample versus Theory

```
exp_means = exp_vars = NULL
# simul data
set.seed(42)
for (i in 1 : nsimul){
    exp_means = c(exp_means, mean(rexp(nummeans, rate=LAMBDA)))
    exp_vars = c(exp_vars, var(rexp(nummeans, rate=LAMBDA)))
}
limits=c(0,300)
par(mfrow=c(1,2))
# mean
hist(exp_means,
     xlab=paste0('Mean of ',nummeans,' exponentials with rate=',LAMBDA),
     main='', ylim=limits, las=1
     )
vvalues= c(mean(exp_means),1/LAMBDA)
vcols= c('black','blue')
abline(v= vvalues, col=vcols)
legend('topright', paste(c('simulated mean','theoretical mean'), signif(vvalues,3)),
       col=vcols, lwd=1)
# var
hist(exp_vars,
     xlab=paste0('Variance'),
     main='', breaks=30, ylim=limits, las=2
)
vvalues= c(mean(exp_vars), 1/LAMBDA^2)
```

```
abline(v= vvalues, col=vcols)
legend('topright', paste(c('simulated var','theoretical var'), signif(vvalues,3)),
       col=vcols, lwd=1
)
par(mfrow=c(1,1))
mtext(paste0('Histogram of ',nsimul,' simulations \nof ',
             nummeans,' exponentials with rate=',LAMBDA), cex=1, font=2)
```

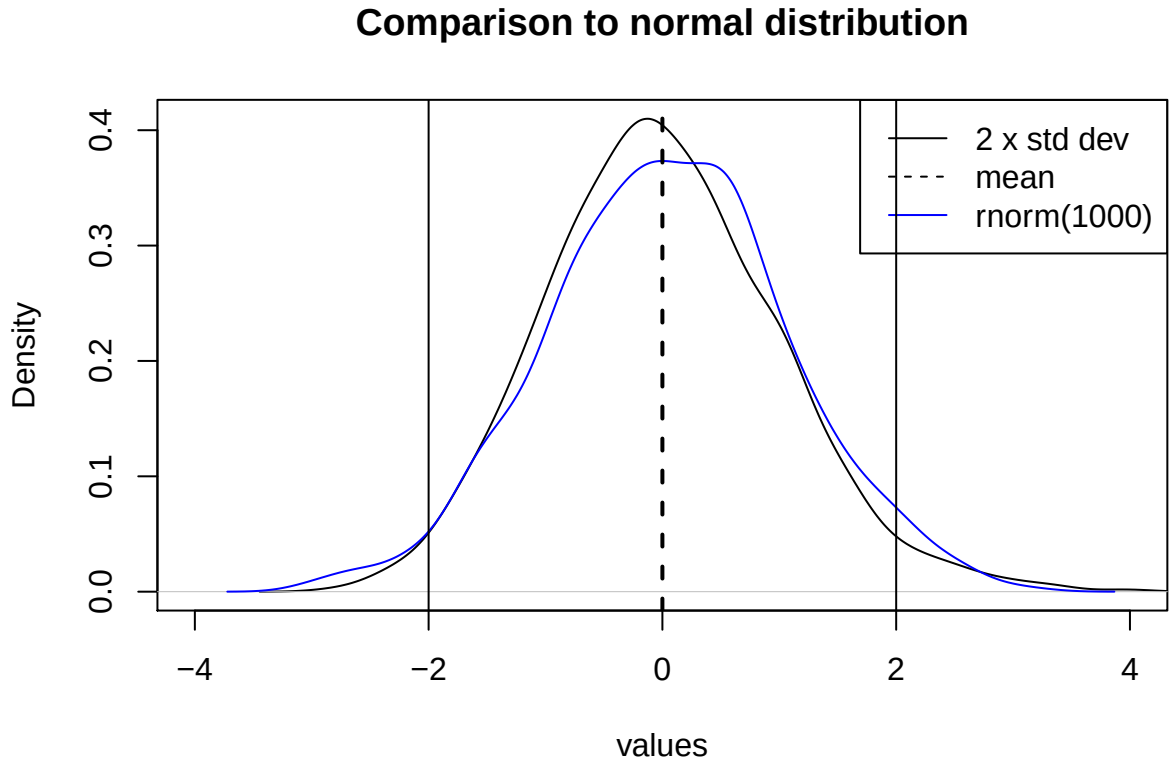


The above code is computing 1000 times the mean and variance of 40 random draws from an exponential distribution with rate=0.2. The simulated mean is very close to the theoretical mean=1/rate. The simulated mean of variances is very close to the theoretical var=1/rate^2. This leads to the conclusion that 1000 simulations are enough to see the Gaussian distribution predicted by the central limit theorem.

Approximately Normal Distribution

```
quasi_normal= (exp_means - mean(exp_means))/sd(exp_means)
multidensity( list(quasi_normal, rnorm(nsimul)),
     main='Comparison to normal distribution',
     xlab='values',
     xlim=c(-4,4),
     col=c('black', 'blue'),
     legend= list(x='topright',
                  legend=c('2 x std dev', 'mean', 'rnorm(1000)'),
                  lty=c(1,2,1),
```

```
                  col=c('black', 'black', 'blue') )
    )
tmplines= c(1,1,2)
abline(v=c(c(-2,2)*sd(quasi_normal), mean(quasi_normal)), lwd=tmplines, lty=tmplines)
```



The above code transforms the distribution of means (X) for 40 exponentials into a standard normal by the following equation:

$$\frac{X - \mu}{SE}$$

where μ and **SE** are the mean and standard deviation of the simulated distribution, respectively.

The Effect of Vitamin C on Tooth Growth in Guinea Pigs

```
library(datasets)
library(beeswarm)
data(ToothGrowth)
tg= ToothGrowth
```

Basic data summary and exploratory data analyses

The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid).

I would not assume that the order of the pigs is the same when applying another dose or vitamin source, so I will use unpaired tests.

Dimensions of the data set:

```
dim(tg)
```

```
## [1] 60  3
```

Tooth length distribution:

```
summary(tg$len)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20   13.08   19.25   18.81   25.28   33.90
```

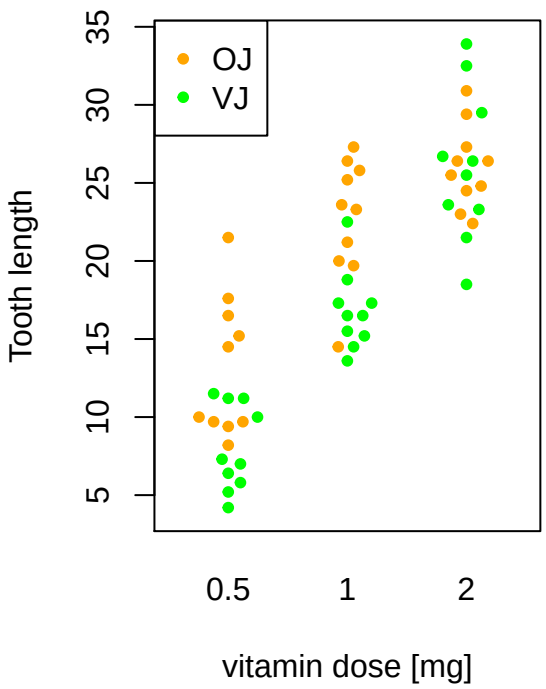
Contingency table for vitamin `dose` and vitamin `supp`lementary methods (OJ=orange juice, VC=ascorbic acid):

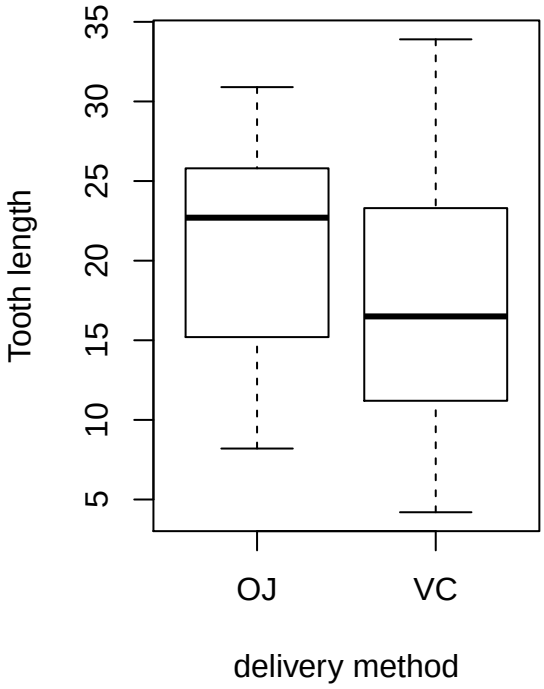
```
with(tg, table(dose, supp))
```

```
##      supp
## dose  OJ VC
##   0.5 10 10
##   1   10 10
##   2   10 10
```

Length comparison by dose and delivery methods:

```
par(mfrow=c(1,2))
beeswarm(len~ dose, data=tg,
        ylab='Tooth length',
        xlab='vitamin dose [mg]',
        pwcol=ifelse(tg$supp=='OJ', 'orange', 'green'),
        pch=20
        )
legend('topleft', c('OJ','VJ'), col=c('orange', 'green'), pch=20)
boxplot(len~ supp, data=tg,
        ylab='Tooth length',
        xlab='delivery method'
)
par(mfrow=c(1,1))
```





Compare tooth growth by supp and dose

I will perform a T-test to decide whether tooth length is different between the two delivery methods. Based on the analysis above I choose unequal variance for `OJ` and `VC`.

```
res_supp= t.test(len~ supp, data=tg, var.equal=FALSE)
res_supp
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

I do not reject the Null-hypothesis that the mean length for pigs fed on acid is the same than for those fed on juice with a Pvalue=0.061> 5% and a confidence interval [-0.1710156, 7.5710156] including the zero.

Additionally, I perform 3 T-tests with unequal variance between doses of 0.5, 1, 2.

```
res_dose= lapply(unique(tg$dose),
     function(d)
         t.test(len~ dose, data=tg[tg$dose!=d, ], var.equal=F)
)
res_dose
```

```
## [[1]]
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996481 -3.733519
## sample estimates:
## mean in group 1 mean in group 2
##          19.735          26.100
##
##
## [[2]]
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean in group 0.5   mean in group 2
##            10.605            26.100
##
##
## [[3]]
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983781  -6.276219
## sample estimates:
## mean in group 0.5   mean in group 1
##            10.605            19.735
```

All 3 tests reject the Null-hypothesis at type 1 error of alpha<5%.

Conclusions

Tooth growth in Guinea Pigs does correlate with dose of vitamin C, but not with delivery method of the vitamin. However, the latter effect is at the borderline of signifcance and a larger cohort should be used to

distinguish the effect of the delivery method.