# Inference Project

*Bernhard Suhm*

*Tuesday, October 28, 2014*

Task: explore the exponential distribution f(x) = l {e}^{-l*x} by simulating the average of 40 executions 1000 times.

Here's some R code to return the average of n outcomes each, n_sim times simulated:

```
# simulate exponential distribution exp(lambda) n_sim times
# computing averages over n simulations each
phats_exp <- function(n_sim,lambda,n) {

  # as prep for using sapply, build up vector of number of executions
  ns<-c(n)
  for (i in 2:n_sim) ns<-c(n,ns)

  # perform n_sim applications of rexp(n,lambda) and gather outcome of each
  outcomes <- sapply(ns, function(x) rexp(x, lambda))

  # computer average of each set of n
  avg <- numeric(n_sim)
  for (i in 1:n_sim) {
    #print(outcomes[,i])
    avg[i] <- mean(outcomes[,i])
  }
  avg
}
```

## Question 1:

Show where the distribution is center at and compare it to the theoretical center of the distribution.

To this end, let's use this code to simulate averages of 40 outcomes 1000 times, and calculate its mean:

```
z <- phats_exp(1000,0.2,40)
m_exp <- mean(z)
m_exp
```

```
## [1] 4.978053
```

That should be pretty close to the expected mean, which is 5.

## Question 2:

Show how variable the distribution is and compare it to its theoretical variance.

```
## [1] 0.5927814
```

That should be pretty close to its expected variance, which is 1/{lambda^2*n} = 0.625

## Question 3:

We can use the Shapiro-Wilk test to check for normality. The p value is very small, indicating the distribution is essentially normal. No wonder since after 1000 simulations the empirical distribution should be approximately normal per CLT.

```
phats<-phats_exp(1000,0.2,40)
shapiro.test(phats)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  phats
## W = 0.9861, p-value = 3.934e-08
```

Not quite after just 100 simulations

```
phats<-phats_exp(100,0.2,40)
shapiro.test(phats)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  phats
## W = 0.9897, p-value = 0.6403
```

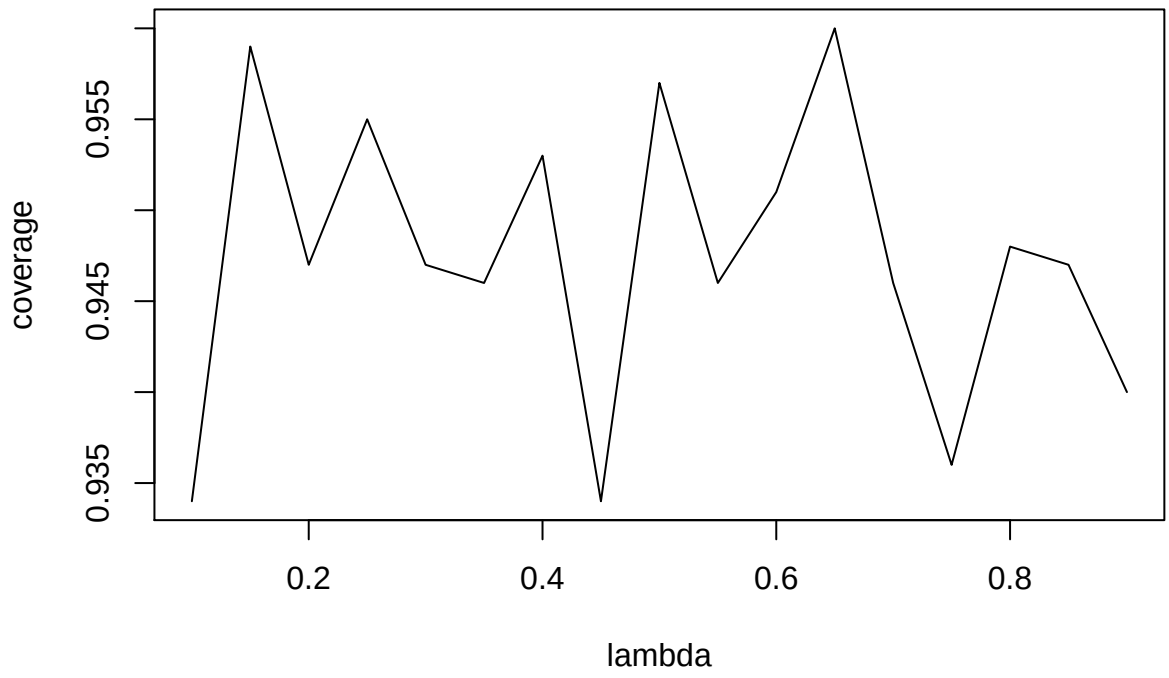
## Question 4:

Evaluate the coverage of the standard 95% confidence interval for lambda.

With the above defined function *phats_exp* that simulates n_sim times computing the average of n values of the exponential distribution we can apply the same code as used during the lecture on a vector of different lambdas, and use 1/lambda as the standard deviation of the exponential distribution (we could also compute the empirical sample standard error, but I have already demonstrated in question 2 above that's about the same):

```
lambdas <- seq(0.1,0.9,by=0.05)
n <- 40
nosim <- 1000
coverage <- sapply(lambdas, function(p) {
    phats <- phats_exp(nosim,p,n)
    ll <- phats - qnorm(0.975)/(p*sqrt(n))
    ul <- phats + qnorm(0.975)/(p*sqrt(n))
    mean(ll<1/p & ul>1/p)
})

plot(lambdas,coverage, type="l", xlab="lambda")
```



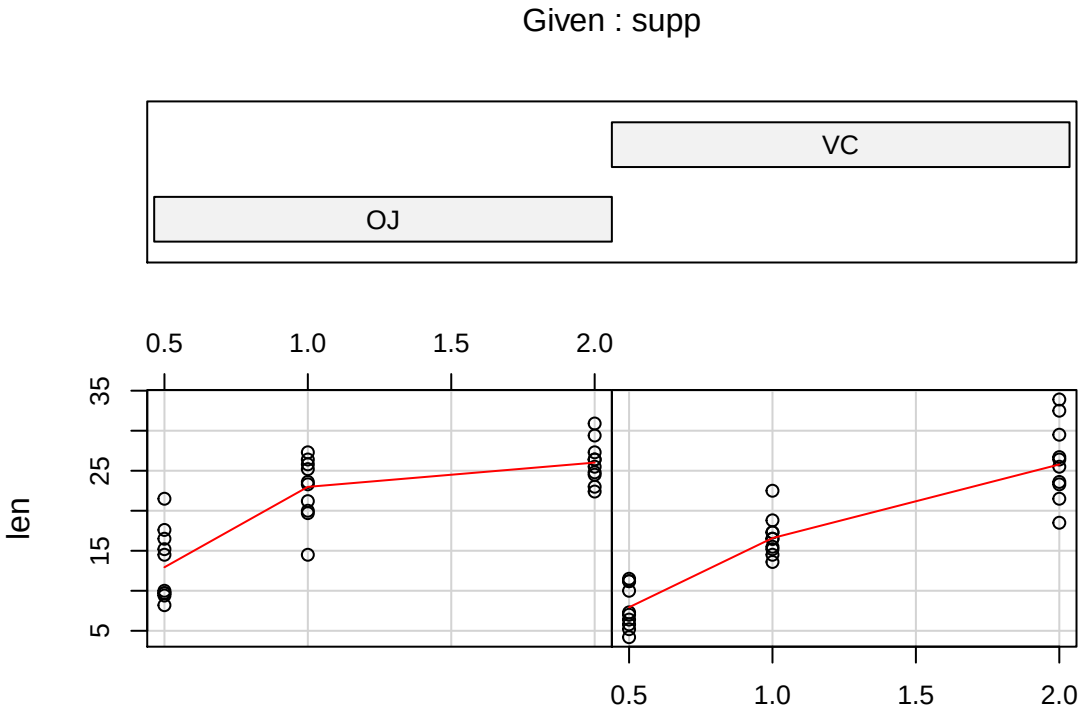

### Exploring the Toothgrowth data

In this data set, the average tooth length is 18.8. The plot suggests that the dose has an impact on tooth growth, whereas visually it's not clear whether the delivery method has an impact. Calculating the mean per factor dose vs. supplement suggest both may have an impact.

```
attach(ToothGrowth)
summary(ToothGrowth)
```

```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

```
coplot(len ~ dose | supp, data = ToothGrowth, panel = panel.smooth,
       xlab = "ToothGrowth data: length vs dose, given type of supplement")
```




ToothGrowth data: length vs dose, given type of supplement

```
tapply(len,dose,mean)
```

```
##    0.5      1      2
## 10.605 19.735 26.100
```

```
tapply(len,supp,mean)
```

```
##       OJ       VC
## 20.66333 16.96333
```

So let's do some hypothesis testing. First we test the null that the mean length for both delivery methods are the same against the hypothesis that it's higher for OJ (one-sided test).

```
x<-len[31:60]
y<-len[1:30]
t.test(x,y,alternative="greater",conf.level=0.95,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  x and y
## t = 1.9153, df = 58, p-value = 0.0302
## alternative hypothesis: true difference in means is greater than 0
```

```
## 95 percent confidence interval:
##  0.4708204       Inf
## sample estimates:
## mean of x mean of y
##  20.66333  16.96333
```

We can reject the null in favor of believing that receiving vitam C in form of orange juice helps tooth growth more than as supplement (for guinnea pigs, anyway). - I applied a t-test because the number of observations is rather small. We assume the administration of vitam C wasn't biased, or at least, drawing the observations wasn't. We get the same result whether or not we assume the variance of the two subsets to be equal.

To investigate the effect of dose on tooth growth I'll apply a regression model. Since we already know (believe) supplement has an impact, I include it in the model as a factor.

```
fit<-lm(len~dose+factor(supp))
summary(fit)
```

```
##
## Call:
## lm(formula = len ~ dose + factor(supp))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.600 -3.700  0.373  2.116  8.800
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.2725     1.2824   7.231 1.31e-09 ***
## dose             9.7636     0.8768  11.135 6.31e-16 ***
## factor(supp)VC  -3.7000     1.0936  -3.383   0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.236 on 57 degrees of freedom
## Multiple R-squared:  0.7038, Adjusted R-squared:  0.6934
## F-statistic: 67.72 on 2 and 57 DF,  p-value: 8.716e-16
```

Not surprisingly, the regression confirms that - the factor supplement has an impact - But also the variable dose is highly significant.