# Statistical Inference Course Project

Mogens Yde-Andersen

## Overview

In this statistical inference course project, I use simulation to explore inference and do simple inferential data analysis. The project consists of an simulation exercise and a basic inferential data analysis and an appendix.

## 1. Simulation exercise

Investigate the exponential distribution and compare it with the Central Limit Theorem.

### 1.1. Simulations

3 simulation exercises are conducted. In each exercise each simulation is done 1000 times. In each simulation 40 exponentials are randomly generated with the R code "rexp(n, lambda)". The 1. simulation take the mean of the 40 random exponentials in each of the 1000 simulations and place it in the mean1000 variable. The 2.simulation take the variance of the 40 random exponentials in each of the 1000 simulations and place it in the var1000 variable. The 3.simulation run 1000 simulations generating 40 exponentials each. Place the 40000 random exponentials in the rexp40000 variable.

Assumptions in all 3 simulations exercises are that the random numbers are generated by "set.seed(1234)" to make the simulation exercises reproducible and that the rate parameter for the exponential function is $\lambda$ = 0.2.

### 1.2 Sample Mean versus Theoretical Mean

Show the sample mean and compare it to the theoretical mean of the distribution.

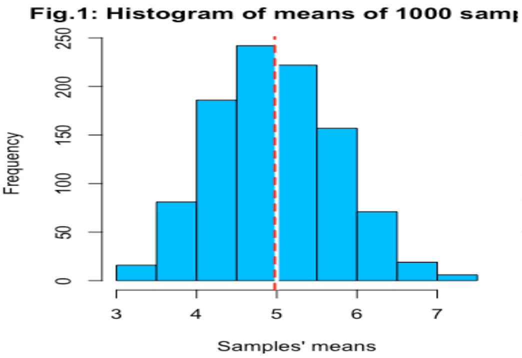


Fig.1: Histogram of means of 1000 sam

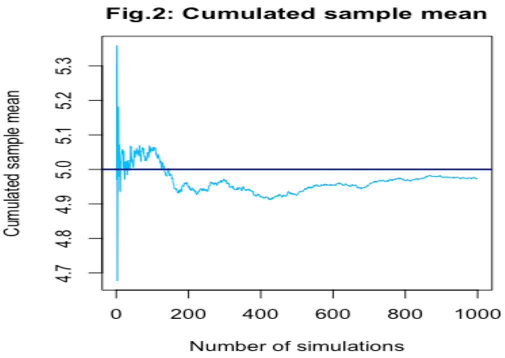


Fig.2: Cumulated sample mean

The theoretical mean of an exponential distribution = 1/lambda = 1/0.2 = 5.

The blue figure 1 shows the distribution of the outcome "mean" of each of a 1000 simulations, where each simulation has produced an outcome mean from 40 randomly generated data points from the function (rexp) in R. The white line shows the theoretical mean 5 of the distribution, where as the dotted red line shows the cumulated mean of the sample means of 4.974239. (See calc. below.) Even though the left "shoulder" is a little bit bigger, the sampled mean distribution centers elegantly around the theoretical mean.

Figure 2 shows the development of the cumulated mean, as the number of simulations goes toward 1000. The graph shows how the cumulated mean (of 4.97, see Appendix A1) approximates the theoretical mean of the underlying exponential distribution.

### 1.3 Sample Variance versus Theoretical Variance

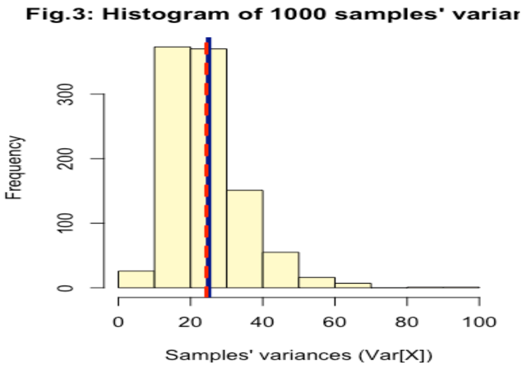
Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

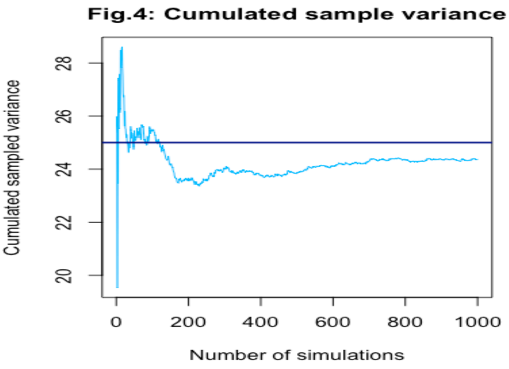
The yellow figure 3 shows the distribution of the outcome "variance" of each a 1000 simulations, where each simulation has produced an outcome variance from 40 randomly generated data points from the function (rexp) in R. The blue line shows the theoretical variance of the distribution. The dotted red line shows the cumulated mean of the sample variances of 24.37801. (See calc. in Appendix A2.)

Theoretical variance of an exponential distribution = 1/(lambda^2) = 1/(0.2^2) = 25.

The sampled variance distribution centers around the theoretical variance, but has a bigger left "shoulder." If the number of simulations were to be increased, the distribution of observed variances would center better around the theoretical variance mean, and the cumulated mean of the variances (of 24.38, se Appendix A2) will approximate the theoretical variance mean.

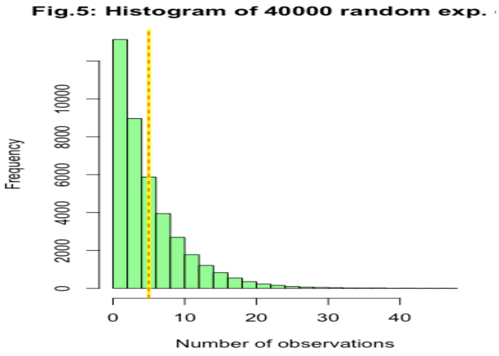
Figure 4 shows the development of the cumulated mean of variance, as the number of simulations goes toward 1000. The graph shows how the cumulated mean of variance approximates the theoretical variance of the underlying exponential distribution.

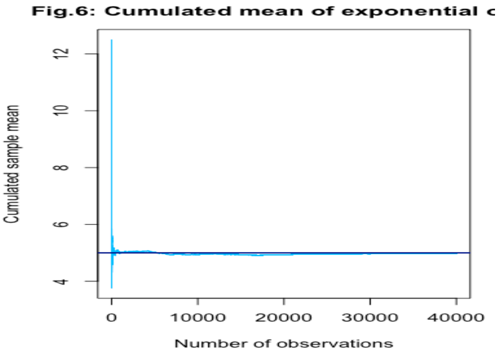




## 1.4 Distribution

Show that the distribution is approximately standard normal.





The Central Limit Theorem (CLT) states that the distribution of averages of iid variables becomes that of a standard normal distribution as the sample size increases. That is (Estimat of X(n) - mean of estimate)/Std. Error of Estimate has a distribution like that of a standard normal distribution for large numbers of n.

An approximation of the measured sample mean is N(mean,(sd^2)/n).

The green figure 5 shows the distribution of a 1000 simulations, where each simulation has randomly generated 40 datapoints from the function (rexp) in R. The 40.000 data points show the profile of a exponential distribution. The yellow line shows, where the theretical mean for exp(x) distribution is. The dotted red line shows the sample mean.

The observed distribution is heavily skewed to left due to the nature of the exponential function. It doesn't follow the bell-shaped curve of a normal distribution like figure 1 and 3. Figure 6 shows how quickly the cumulated sample mean centers around the theoretical mean 5.

If the distribution of the observations in variable mean1000 (Fig.1) and in variable var1000 (Fig.3) can be described by a normal distribution X ~ (mu, sigma^2), then a standarnd distribution Z vil follow the equation Z = (x-mu)/sigma ~ N(0,1). If Z is true, then X = mu + sigma*Z ~N(mu, sigma^¨2). This can be approximated with N(mean,(sd^2)/n).

Figure 1 and figure 2 show that the sample mean approximates 5, as the number of simulations over the underlying exponential distribution increases. Similarly does figure 3 and figure 4 show that the sample variance approximates 25, as the number of the same simulations over the same underlying distribution increases as well.

Therefore a distribution of averages can be described as X ~ N(5,25). According til CLT, if X ~ N(5,25) is true, then a distribution Z can be described as a standard normal distribution by the equation Z = (X-5)/25 ~ N(0,1). If Z is true, then X = 5 + 25*Z ~ N(5,25). Therefore the predicted mean of the distribution of averages shall lie in the following confidence interval of 97,5%

Since the sample mean of 4.974 (See Appendix A3) lies within the predicted 97,5% confidence interval [4.951,5.049] (See Appendix A3)), then the Central Limit Theorem holds, and the distribution of averages is approximately standard normal.

## 2. Basic inferential data analysis

### 2.1 Load the ToothGrowth data and perform some basic exploratory data analyses

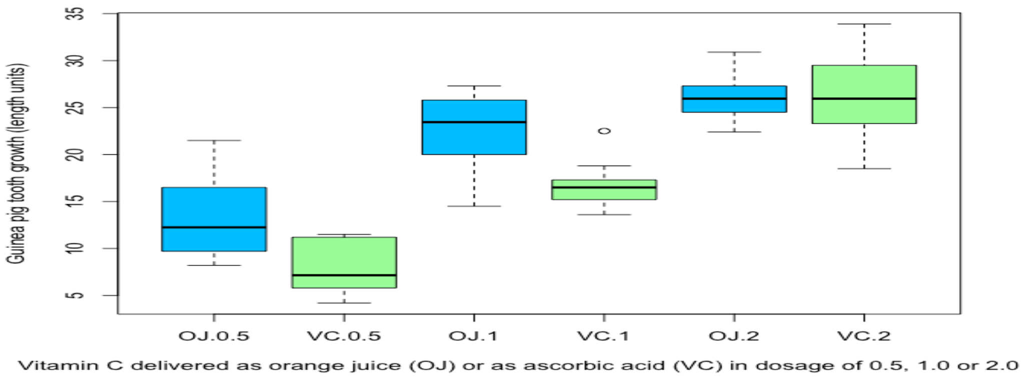


Fig.7: Results of guinea pig tooth growth experiment

As figure 7 shows, the dataset ToothGrowth contains the measured growth of the teeth of 60 guinea pigs grouped into 6 due to received supplement type and dosage.

### 2.2 Provide a basic summary of the data

The ToothGrowth dataset contains the response in 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 (OJ) or ascorbic acid (VC)).

The ToothGrowth dataset format is a data frame with 60 observations on 3 variables; 1) "len", a numeric variable for tooth growth length, 2) "supp", a factor variable for supplement type (OJ or VC) and 3) "dose", a numeric variable for dose in milligrams.

The observed means and variances of the guinea pig tooth growth length in each groups of 10 guinea pigs per supplement and per dosage are seen in the appendix A4.

### 2.3 Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.

The hypothesis is that orange juice is better than ascorbin acid at all measured dosages for guinea pig tooth growth length (tgl).
$H_{01}$: E[X($tgl.oj05$)] ≥ E[X($tgl.vc05$)] => H01: E[X($tgl.oj05$)] - E[X(tgl.vc05)] ≥ 0
$H_a1$: E[X($tgl.oj05$)] < E[X($tgl.vc05$)] => H01: E[X($tgl.oj05$)] - E[X($tgl.vc05$)] < 0
$H_{02}$: E[X($tgl.oj10$)] ≥ E[X($tgl.vc10$)] => H01: E[X($tgl.oj10$)] - E[X($tgl - vc10$)] ≥ 0
$H_a2$: E[X($tgl.oj10$)] < E[X($tgl.vc10$)] => H01: E[X($tgl.oj10$)] - E[X($tgl - vc10$)] < 0
$H_{03}$: E[X($tgl.oj20$)] ≥ E[X($tgl.vc20$)] => H01: E[X($tgl.oj20$)] - E[X($tgl - vc20$)] ≥ 0
$H_a3$: E[X($tgl.oj20$)] < E[X($tgl.vc20$)] => H01: E[X($tgl.oj20$)] - E[X($tgl - vc20$)] < 0
The t-test will test the 3 null hypotheses against their alternatives. See calculations in the appendix A5.

### 2.4 State your conclusions and the assumptions needed for your conclusions

Assumptions: The tooth growth length of each guinea pig is the measured length difference at the end and at the start of the experiment. That is repeat measurement, and thus the data shall be treated in at an individual level, and the t-test er calculated with paired=TRUE.

Conclusions:

$H_{01}$ versus $H_a1$: Since the t-statistic = 2.98 in the t-test lies within 95% confidence interval of [1.26,9.24], and since the confidence interval is positive, i fail to reject the $H_{01}$ hypothesis. This hypothesis is broken into 3 sub-hypotheses.

$H_{02}$ versus $H_a2$: Since the t-statistic = 3.37 in the t-test lies within 95% confidence interval of [1.95,9.91], and since the confidence interval is positive, i fail to reject the $H_{02}$ hypothesis.

$H_{03}$ versus $H_a3$: The t-statistic = -0.0426 in the t-test lies within 95% confidence interval of [-4.33,4.17]. But since the confidence interval contains 0, i con not rule out that $H_a3$ is true, but i fail to reject the $H_{03}$ hypothesis.

Orange juice as supplement in dosage 0.5 mg and 1.0 mg has a better impact on guinea pig tooth growth than ascorbin acid. In dosage 2.0 mg it is unproven that either orange juice or ascorbin acid is better at growing guinea pig teeth. Orange juice is not proven to bebetter than ascorbin acid at all dosages.

## Appendix

A1: Calculation of the cumulated mean of the 1.000 samples' means

```
## [1] 4.974239
```

A2: Calculation of the cumulated mean of variance of the 1.000 samples' variances

```
## [1] 24.37801
```

A3: Calculation of the sample mean and of the 97,5% confidence interval

```
## [1] 4.951001 5.048999

## [1] 4.974239
```

A4: Extracting the observed means of the guinea pig tooth growth length in each groups of 10 guinea pigs per supplement and per dosage are

```
##   Supplement Dosage Mean length Standard deviation length
## 1         OJ    0.5       13.23                  4.459709
## 2         VC    0.5        7.98                  2.746634
## 3         OJ    1.0       22.70                  3.910953
## 4         VC    1.0       16.77                  2.515309
## 5         OJ    2.0       26.06                  2.655058
## 6         VC    2.0       26.14                  4.797731
```

A5: Creating the 6 relevant subgroups and calculating observed means and variances of the guinea pig tooth growth length in each groups of 10 guinea pigs per supplement and per dosage

```
## 
##  Paired t-test
## 
## data:  oj05$len and vc05$len
## t = 2.9791, df = 9, p-value = 0.01547
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.263458 9.236542
## sample estimates:
## mean of the differences 
##                    5.25

## 
##  Paired t-test
## 
## data:  oj10$len and vc10$len
## t = 3.3721, df = 9, p-value = 0.008229
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.951911 9.908089
## sample estimates:
## mean of the differences 
##                    5.93

## 
##  Paired t-test
## 
## data:  oj20$len and vc20$len
## t = -0.0426, df = 9, p-value = 0.967
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.328976  4.168976
## sample estimates:
```

```
## mean of the differences
##                    -0.08
```