Statistical Inference Final Project

JJDV

January 25, 2018

Simulation and Central Limit Theorem Exercise

Overview

In this exercise/simulation we're gonna be taking a look at the exponential distribution, which can be simulated in R with the `rexp()` function. We're gonna be simulating distributions of 40 samples and getting their average. We will then analyze these averages and their properties.

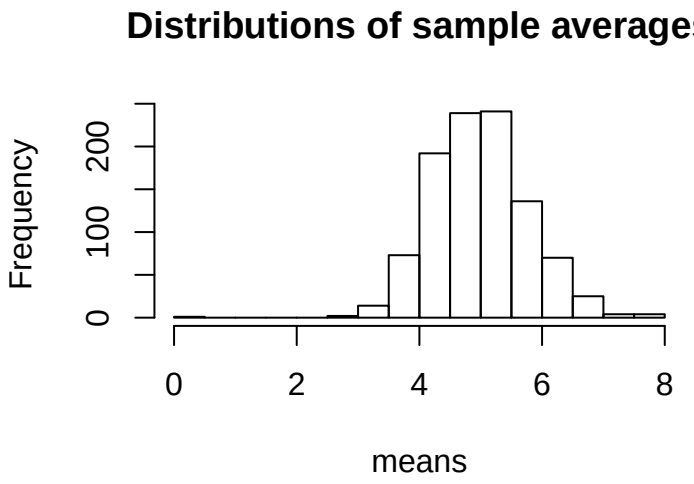
Simulations

We will be running 1000 simulations of exponential distributions, 40 samples each:

```
means <- 0
for (i in 1 : 1000) means <- c(means, mean(rexp(40, .2)))
```

After getting the averages for each of the distributions we graph a histogram of the means and we can then show the sample mean.

```
hist(means, main="Distributions of sample averages")
```

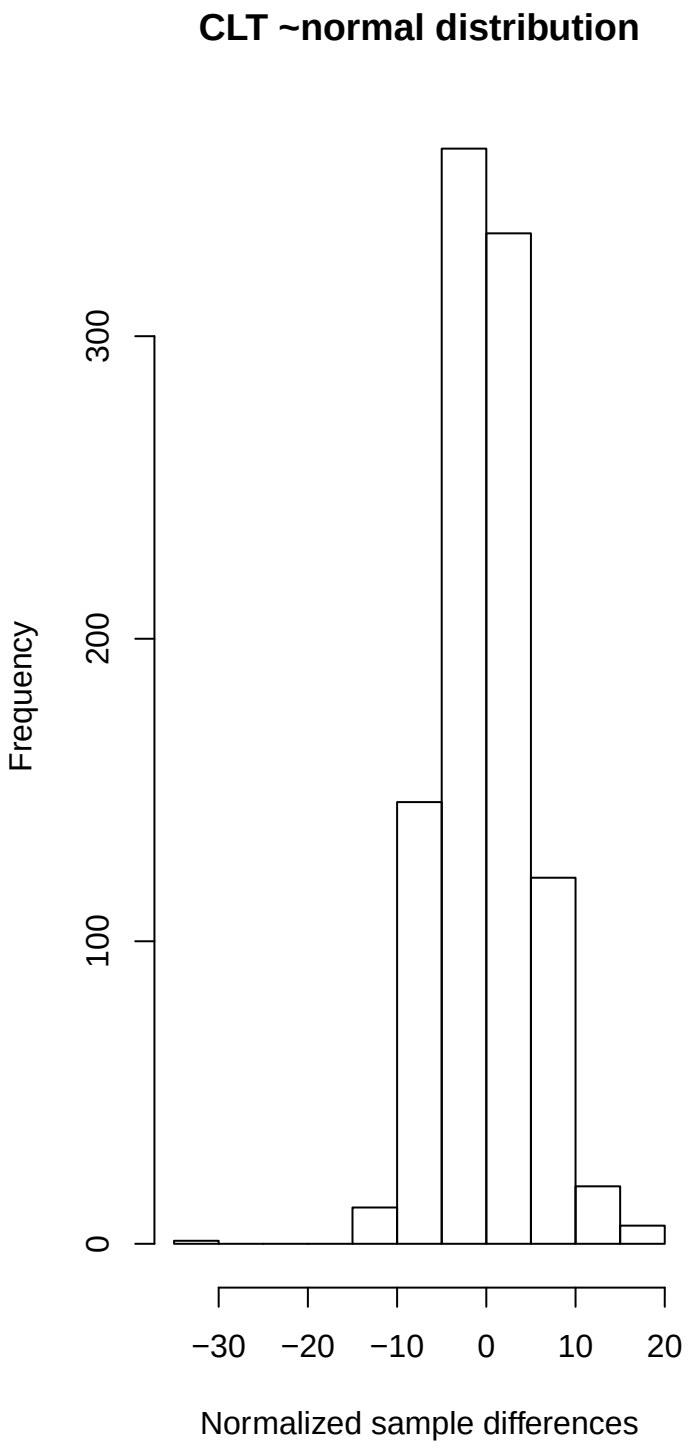


Sample vs Theory

The theoretical mean should be **5** as we used a `lambda` value of **0.2**. The sample mean we got is **4.9761363**. The variance of the distributions should have the same value as the mean, in our case 25, the sample variance in this example came out to be **0.6123093** and standard deviation of **0.7825019**.

The distribution that results from substracting the theoretical mean from the sample means and then dividing by the standard error of the estimate can be represented by the following figure:

```r
cltdist <- (1000^(1/2)*(means - 5))/5
hist(cltdist, main="CLT ~normal distribution", xlab = "Normalized sample differences")
```



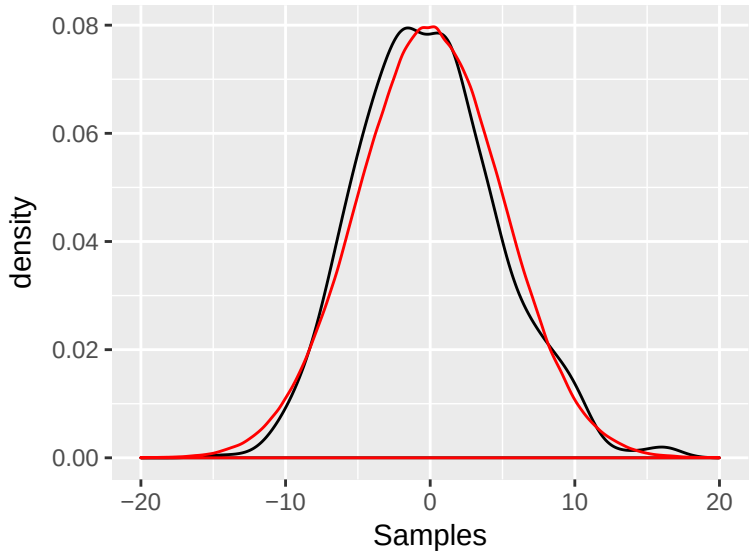
This new distribution has mean -0.1509272 and variance 24.4923702. We could also take a look at the main quantiles for this new distribution and get a hint as to whether it truly is approximately normal.

```r
quantile(cltdist)
```

```
##          0%         25%         50%         75%        100%
## -31.6227766  -3.5649999  -0.3728374   2.8766578  16.7402597
```

Finally with a confidence value of 5% we get an interval and check if our distribution respects the interval.

```r
# 5 + c(-1,1) * qnorm(.975) * 5 / sqrt(1000)
# mean(means) + c(-1,1) * qnorm(.975) * sd(means) / sqrt(length(means))
ggplot() + geom_density(aes(x = cltdist)) +
    geom_density(aes(x = rnorm(800000, 0, 5)), colour = "#FF0000") +
    xlim(c(-20,20)) + labs(xlab("Samples"), ylab("Density"))
```



Basic Inferential Data Analysis Instructions

In this section we are gonna work with the `ToothGrowth` dataset. We are gonna be looking at the growth and comparing the `supp` and `dose` variables. The dataset has a total of 60 data points and 3 variables.

Table 1: Mean and standard deviation of growth by supplement and dosage

supp	dose	length	sd
OJ	0.5	13.23	4.459708
OJ	1.0	22.70	3.910953
OJ	2.0	26.06	2.655058
VC	0.5	7.98	2.746634
VC	1.0	16.77	2.515309
VC	2.0	26.14	4.797731

We will therefore hypothesize that OJ supplementation is just as effective as ascorbic acid (`VC` in our dataset) in affecting tooth growth. We will test if this hypothesis holds true by doing an interval test with a confidence level of 95%.

We decide to do separate T tests for these samples. We do three T tests one for each comparison between

dosages and supplement, e.g. a dose of 0.5 for both OJ and VC compared. These T tests will be non-paired and with differing variances. Results are as (see the appendix page for details):

Table 2: P values for the different hypothesis and dosages

Test	P.value
0.5 Dosage	0.006359
1 Dosage	0.001038
2 Dosage	0.963900

Conclusion

We can therefore conclude that the hypothesis can be rejected for the does `0.5`, and `1`. That is, the `OJ` isn't as effective as the `VC` supplement. The hypothesis however cannot be rejected for the `2` dosage.

Appendix

Table 2 full data

```
# 0.5 Dosage
t.test(len ~ supp, paired = F, var.equal = FALSE,
       data=ToothGrowth[ToothGrowth$dose == 0.5,],
       alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean in group OJ mean in group VC
##            13.23             7.98
```

```
# 1 Dosage
t.test(len ~ supp, paired = F, var.equal = FALSE,
       data=ToothGrowth[ToothGrowth$dose == 1,],
       alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean in group OJ mean in group VC
##            22.70            16.77
```

```
# 2 Dosage
t.test(len ~ supp, paired = F, var.equal = FALSE,
       data=ToothGrowth[ToothGrowth$dose == 2,],
       alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```