Statistical Inference Project

Matthew Henderson

14 February 2021

Overview

This project aims to display concepts from statistical inference through data analsis of simulated and real-life data. The first section uses simulated data from an exponential distribution to show evidence of well-known results in statistical theory such as the central limit theorem. In the second section there is a data analysis of a ToothGrowth dataset which constructs a hypothesis test to distinguish between different groups in the data.

Simulations

The exponential distribution

The exponential distribution models the time between counts of the Poisson distribution. The exponential density function is:

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

where x is on the interval $[0, \infty)$. The expected value of this distribution is $E[X] = 1/\lambda$ and the variance is $Var[X] = 1/\lambda^2$.

Simulating data

In this project I will use simulated data from an exponential distribution with $\lambda = 0.2$. The PDF of this function is $0.2e^{0.2x}$ defined for values of x greater than 0. I will simulate 1000 samples of 40 random variables from this distribution.

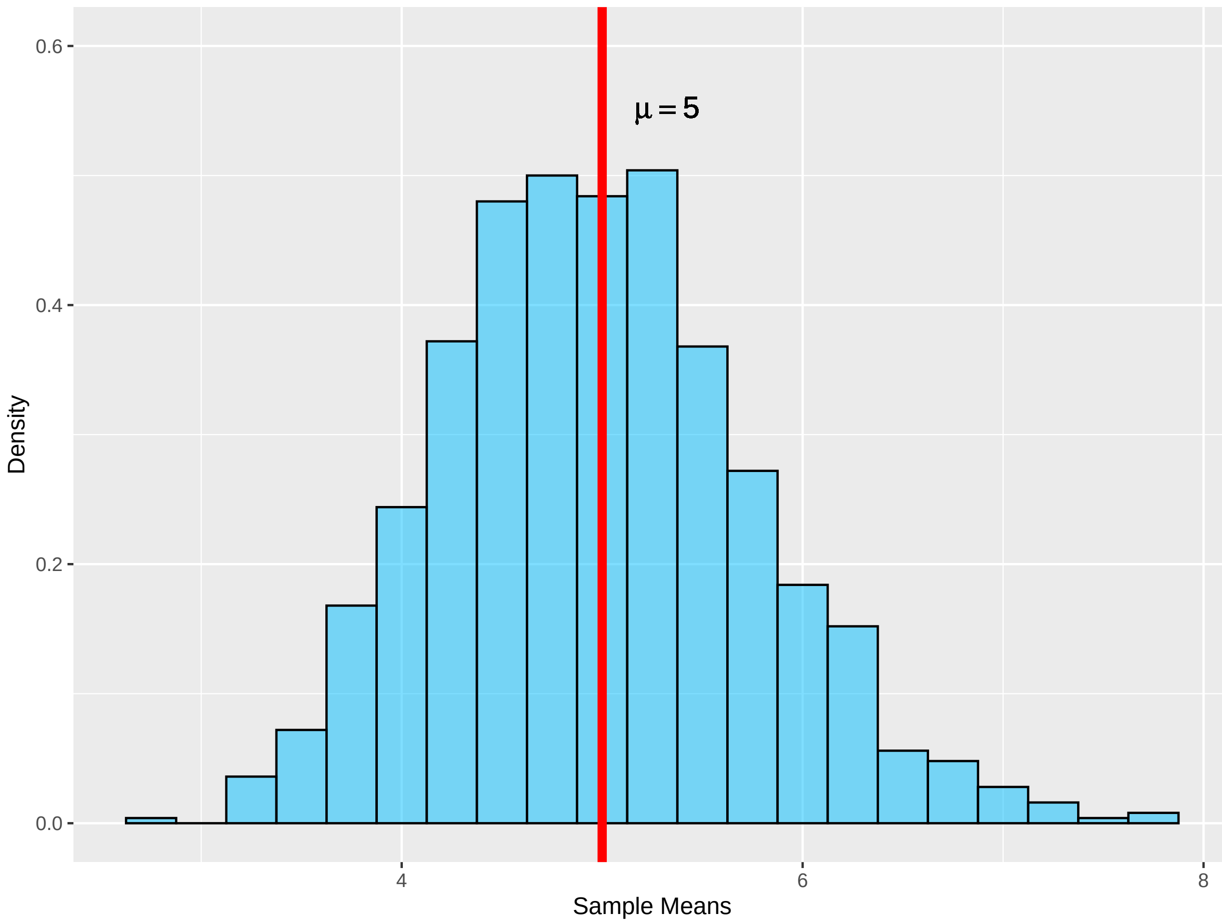
```
set.seed(1234)
lambda <- 0.2
n <- 40
nosim <- 1000
expMatrix <- matrix(rexp(n*nosim, rate = lambda), nrow = nosim)
mns <- apply(expMatrix, 1, mean)
sds <- apply(expMatrix, 1, sd)
```

So now I have a 40×1000 matrix of simulated data from the $exp(\lambda = 0.2)$ distribution. A 1000-dim vector of the mean of each row and a 1000-dim vector of the standard deviation of each row.

Sample Mean vs Theoretical Mean

The theoretical mean of the exponential distribution is $E[X] = 1/\lambda$ so for this distribution the expected values is $E[X] = 1/0.2 = 5$. We can compare this theoretical value with the simulated data by creating a histogram of the sample means.

```
library(ggplot2)
## Create a density histogram for the mean data
g <- ggplot(data.frame(mns = mns), aes(x = mns)) + geom_histogram(aes(y = ..density..), colour = "black
## add a vertical line at the theoretical mean = 5
g <- g + geom_vline(xintercept = 5, colour = "red", size = 2) +
    geom_text(aes(x = 5, y = 0.55), label = expression(mu==5), hjust = -0.5, size = 5)
g
```

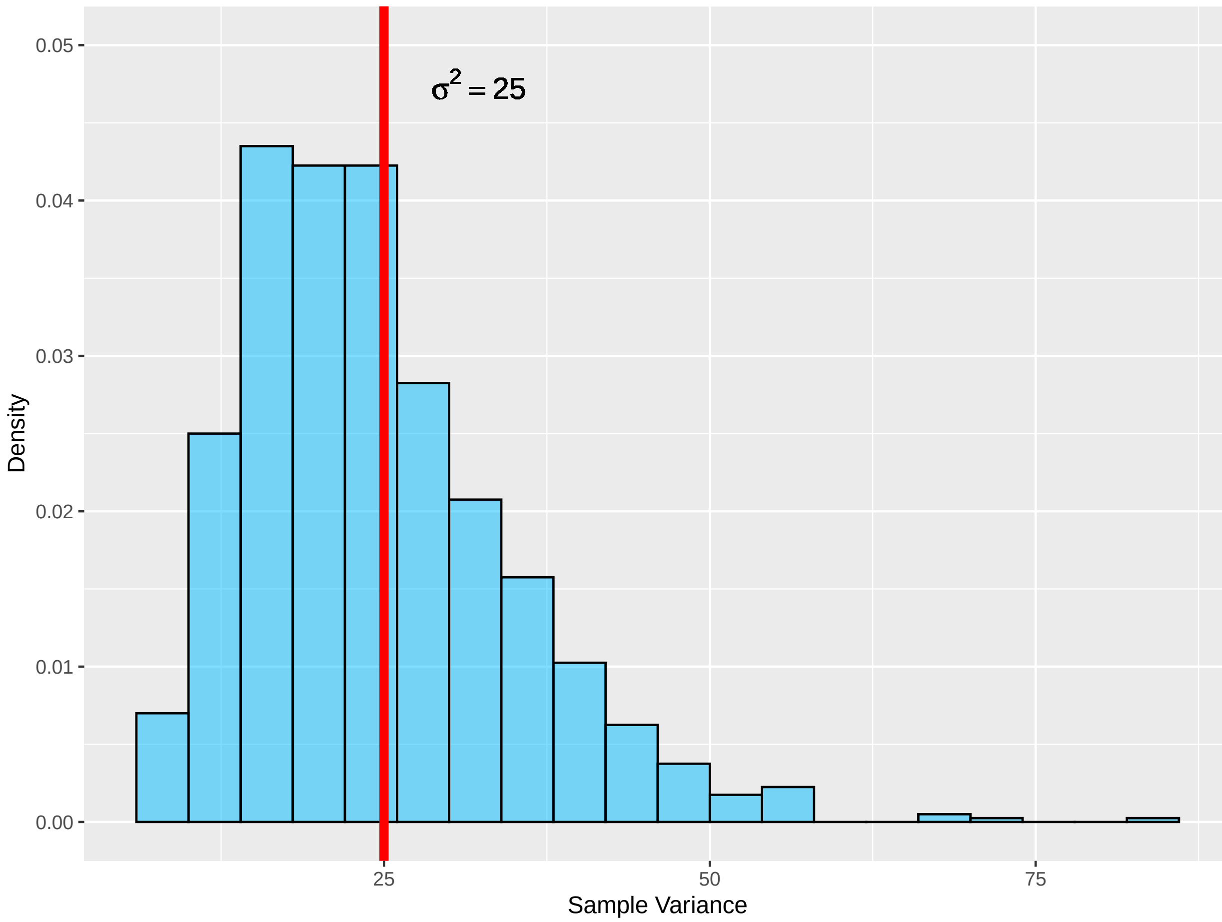


The sample means create a bell-shaped curve around the true population mean and look to be a very good estimator of the true mean.

Sample Variance vs Theoretical Variance

The theoretical variance of the exponential distribution is $Var[X] = 1/\lambda^2$ hence, for this distribution $Var[X] = 1/0.2^2 = 25$. We can compare the simulated data to the theoretical variance.

```
## create a density histogram for the variance data
g2 <- ggplot(data.frame(var = sds ^ 2), aes(x = var)) + geom_histogram(aes(y = ..density..), colour = "
## add a vertical line at the theoretical variance = 25
g2 <- g2 + geom_vline(xintercept = 25, colour = "red", size = 2) +
    geom_text(aes(x = 25, y = 0.0475), label = expression(sigma^{2}==25), hjust = -0.5, size = 5)
g2
```



The sample variance is consentrated round the true population variance however, it is noticable that the sample variances have a more skewed looking distribution than the sample means with a higher chance of more extreme values to the right side of the distribution with some more than three standard deviations away from the mean. For this distribution we probably want a higher sample size n for the sample variance to be a good approximator of the population variance.

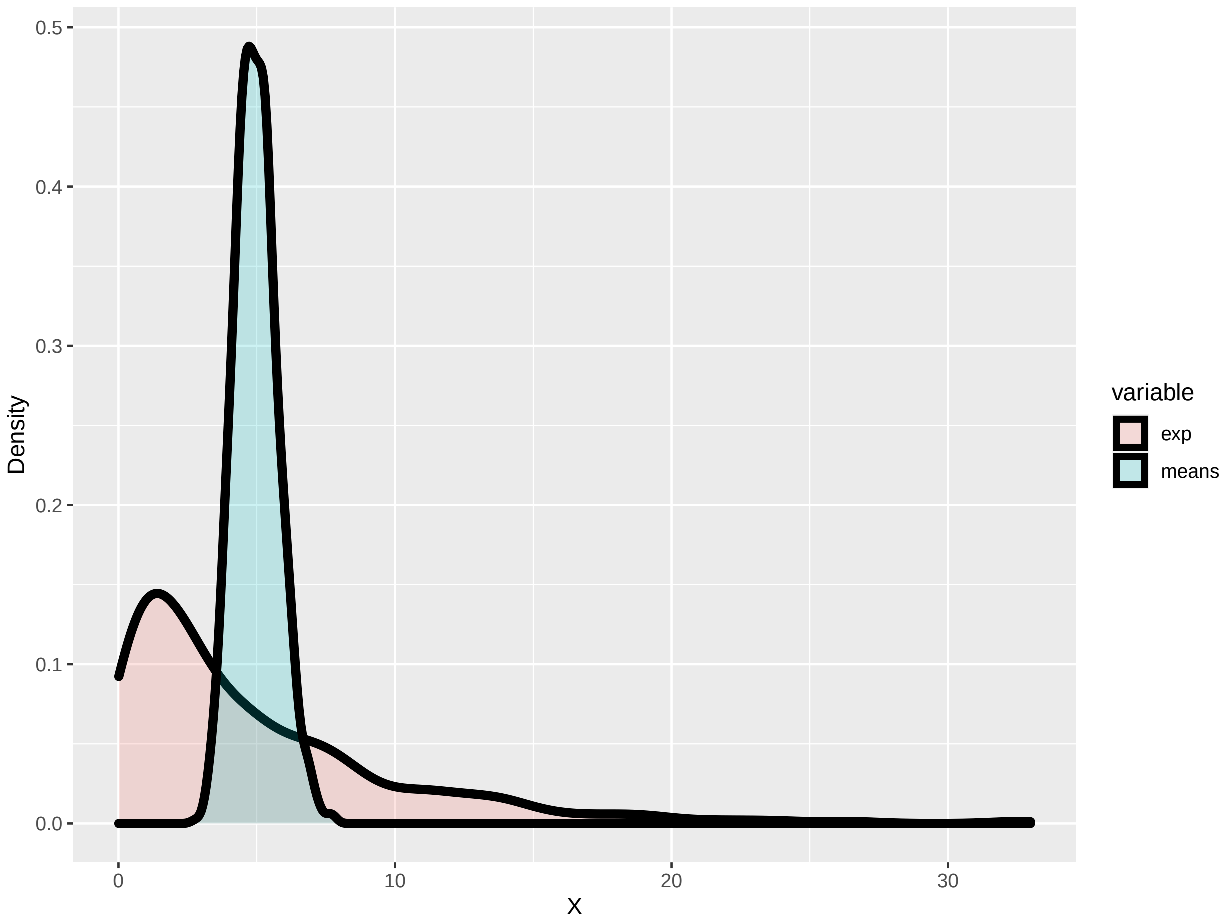
Sample Means are approximetely Normal

Plotting a density curve of a large number of random variables from an exponential distribution should be an exponential looking curve with the center of mass around its true mean. If we take the average of 40 exponential random variables a large number of times we see that the distribution of averages has a a gaussian looking density which is far more concentrated around the population mean.

```r
library(reshape2)
df <- data.frame(exp = rexp(nosim, lambda), means = mns)
df <- melt(df)
```

```
## No id variables; using all as measure variables
```

```r
g3 <- ggplot(df, aes(x = value, fill = variable)) + geom_density(alpha = 0.2, size = 2) + labs(x = "X",
g3
```



This results can be explained by the Central Limit Theorem which tells us that the distribution of averages of any independent and identically distributed (iid) random variable becomes increasingly more normal as the sample size increases.

In our case with have taken the average of 40 exponentials ($\lambda = 0.2$) 1000 times, hence each average is independant of the last with the exact same probability distribution. The central limit theorem states:

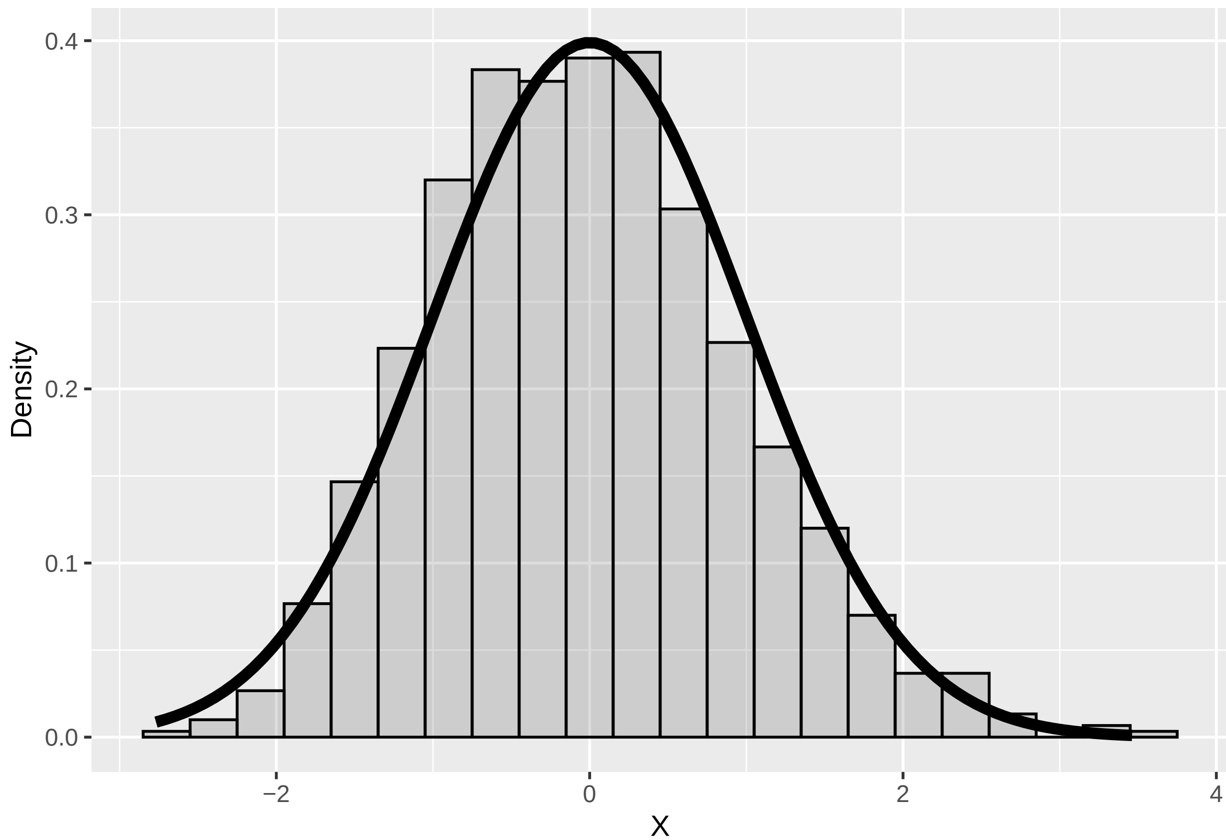
$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{sigma} = \frac{\text{Estimate} - \text{Mean of Estimate}}{\text{Std. Err. of estimate}}$$

follows a distribution like that of a standard normal. We can apply the CLT to our simulated data by applying the transformation:

$$\frac{\sqrt{4}0(\bar{X}_n - 5)}{5}$$

since $\mu = 5$, $\sigma = 5$ and the sample size $n = 40$.

```
# normalise the data using variables from the population
cltFunc = function(x, n) sqrt(n) * (mean(x) - 5) / 5
df2 <- data.frame(x = apply(expMatrix, 1, cltFunc, 40))
g4 <- ggplot(df2, aes(x = x)) + geom_histogram(aes(y = ..density..), alpha = 0.2, binwidth = 0.3, colou
g4 <- g4 + stat_function(fun = dnorm, size = 2)
g4
```



The approximation of the standard normal distribution from the distribution of averages of 40 random variables from the $exp(\lambda = 0.2)$ is a very good fit.