popkin_final

Ken Popkin

12/15/2020

Load the data

```
data = read.csv('C:\\Users\\Ken\\Documents\\00_Applications_DataScience\\CUNY\\DATA605\\Final\\house-prices-advanced-regression-techniques\\train.csv', header=TRUE)
```

Problem 2

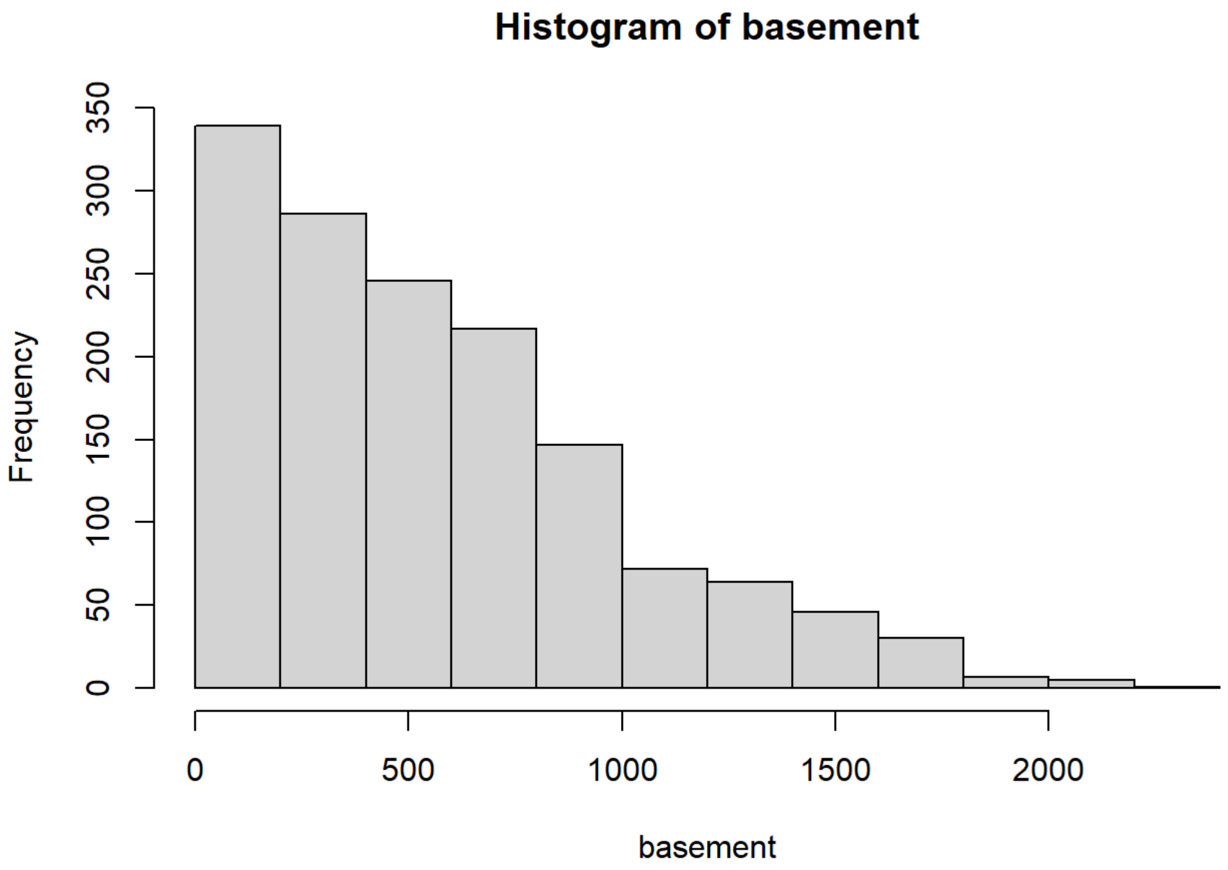
Calculus Based Probability and Statistics Many times, it makes sense to fit a closed form distribution to data.
1. Select a variable in the Kaggle.com training dataset that is skewed to the right.
2. Load the MASS package and run fitdistr to fit an exponential probability density function. (See https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/fitdistr.html (https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/fitdistr.html)).
3. Find the optimal value of lambda for this distribution 4. Take 1000 samples from this exponential distribution using this value (e.g., rexp(1000, lambda)).
5. Plot a histogram and compare it with a histogram of your original variable.
6. Using the exponential pdf, find the 5th and 95th percentiles using the cumulative distribution function (CDF).
7. Generate a 95% confidence interval from the empirical data, assuming normality.
8. Provide the empirical 5th percentile and 95th percentile of the data.
9. Discuss.

1 Select a variable in the Kaggle.com training dataset that is skewed to the right.

```
basement = data$BsmtUnfSF

basement = as.vector(basement)
basement = gsub(",", "", basement)   # remove comma
basement = as.numeric(basement)      # turn into numbers

hist(basement)
```



2 Run fitdistr to fit an exponential probability density function.

```
fitdistr(basement,densfun = 'exponential')
```

```
##        rate
##   1.762921e-03
##  (4.613775e-05)
```

3 Find the optimal value of lambda for this distribution

```
#lambda = 1/mean
mean = mean(basement)
lambda = 1/mean
lambda
```

```
## [1] 0.001762921
```

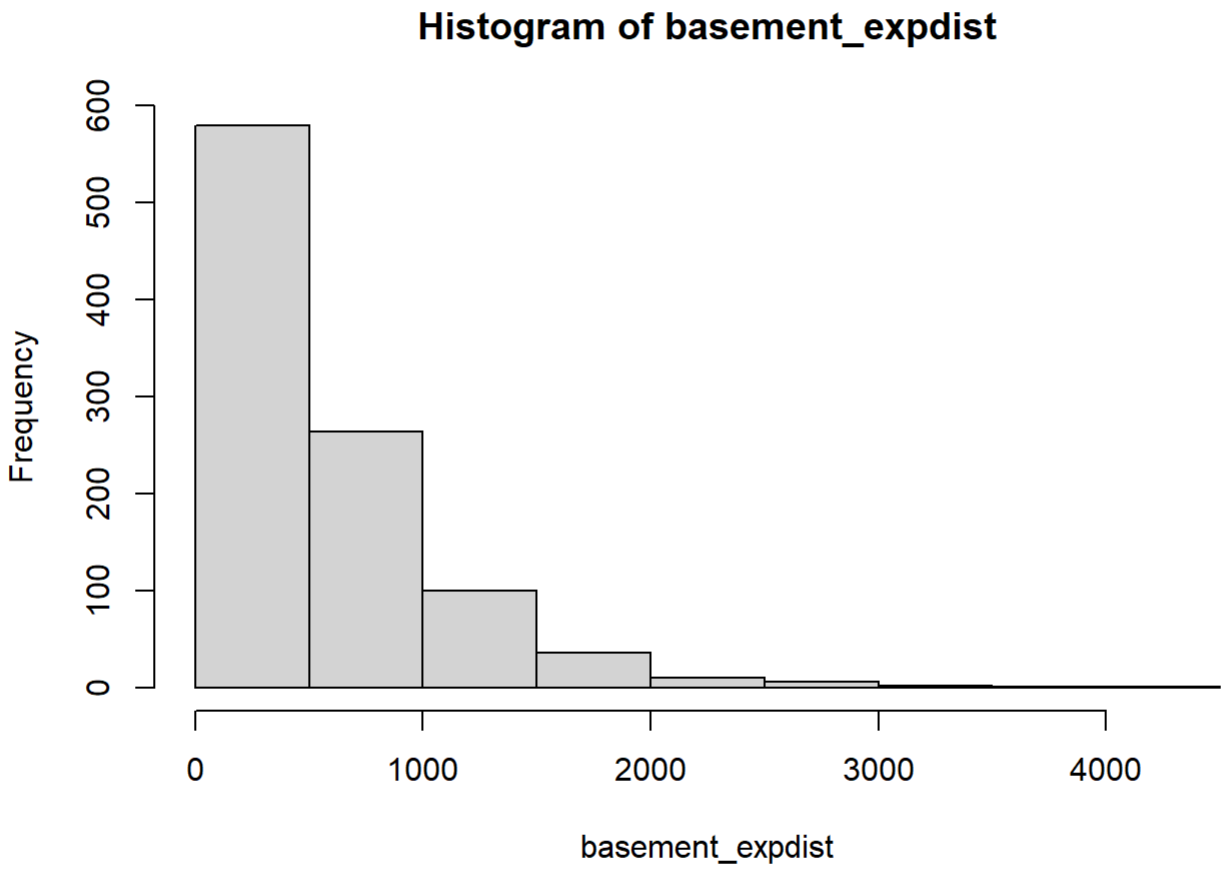
4 Take 1000 samples from this exponential distribution using this value (e.g., rexp(1000, lambda)).

```
basement_expdist = rexp(1000,lambda)
```

5 Plot a histogram and compare it with a histogram of your original variable.

Comparing the exponential distribution histogram below with the original histogram in step 1, shows similar charactercs in that both are right skewed. The exponential distribuition appears more right skewed, with more bins and a lower count per bin across the histogram.

```
hist(basement_expdist)
```

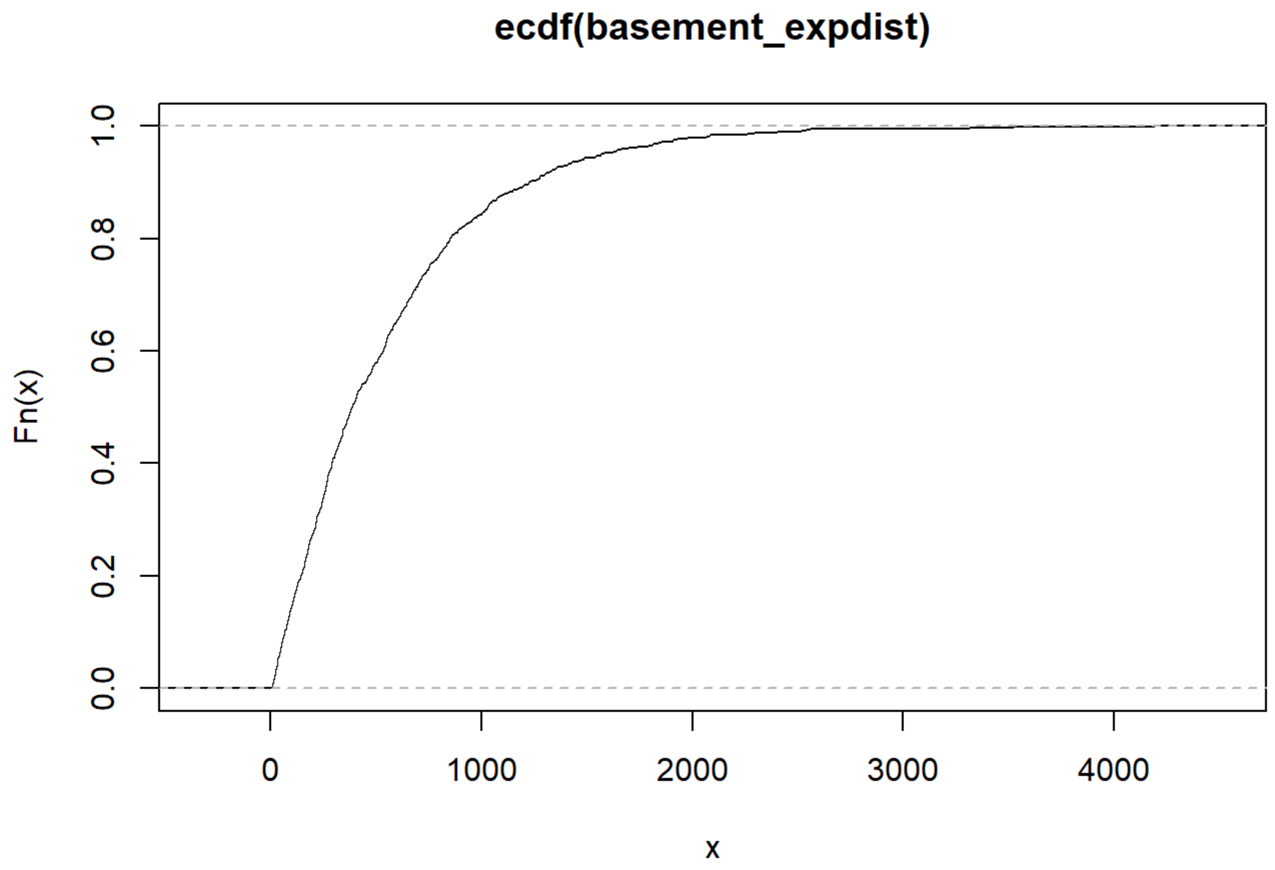


6 Using the exponential pdf, find the 5th and 95th percentiles using the cumulative distribution function (CDF).

```
my_cdf = ecdf(basement_expdist)

#From the chart below, the 5th percentile is 0 and the 95th percentile is 1.0

plot(my_cdf)
```



7 Generate a 95% confidence interval from the empirical data, assuming normality.

```
confint = 1.96

sem = sd(basement)/sqrt(length(basement))
mean = mean(basement)

lowercf = mean - (confint * sem)
uppercf = mean + (confint * sem)

cat('confidence level', lowercf, ' < ', mean, ' < ', uppercf)
```

```
## confidence level 544.5746  <  567.2404  <  589.9062
```

8 Provide the empirical 5th percentile and 95th percentile of the data.

```
fifth = quantile(basement,0.05)
ninety_fifth = quantile(basement, 0.95)

cat('5th percentile is', fifth, '\n')
```

```
## 5th percentile is 0
```

```
cat('95th percentile is', ninety_fifth)
```

```
## 95th percentile is 1468
```

9 Discuss

Applying a closed form distribution to the data did not reveal new information to me. Given a choice, I would not venture into applying closed form distribution to explore and derive information from the basement dataset. I much prefer staying with the original data to derive the various confidence intervals and percentiles that this problem asks for.