Assignment 4

Kyle Ligon

2018-10-31

Required Packages

```
library(tidyverse)
library(broom)
library(nortest)
```

Dataset #1: Using KS on a dataset to check Exponential Distribution with Rate == 1

```
air <- as.tibble(datasets::airquality)
air <- filter(air, !is.na(air$Ozone))

air
```

```
## # A tibble: 116 x 6
##    Ozone Solar.R  Wind  Temp Month   Day
##    <int>   <int> <dbl> <int> <int> <int>
##  1    41     190   7.4    67     5     1
##  2    36     118   8      72     5     2
##  3    12     149  12.6    74     5     3
##  4    18     313  11.5    62     5     4
##  5    28      NA  14.9    66     5     6
##  6    23     299   8.6    65     5     7
##  7    19      99  13.8    59     5     8
##  8     8      19  20.1    61     5     9
##  9     7      NA   6.9    74     5    11
## 10    16     256   9.7    69     5    12
## # ... with 106 more rows
```

Hypotheses

H_0: $F(x) = F^*(x)$
H_1: $F(x) \neq F^*(x)$

Test Statistic

```
ks_test <- ks.test(x = air$Ozone, 'pexp', rate = 1) %>%
  tidy()
```

We have 0.9803 as our test statistic.

P-value

We have 0 as our p-value.

Conclusion

With a p-value less than 0.05, we have enough evidence to reject the null hypothesis that our data fit the exponential distribution with a Rate == 1. There appears to be evidence that our data do not fit an exponential model with Rate == 1.

Dataset #2: Using the Chi-Square test to see if my data is Binomially Distributed

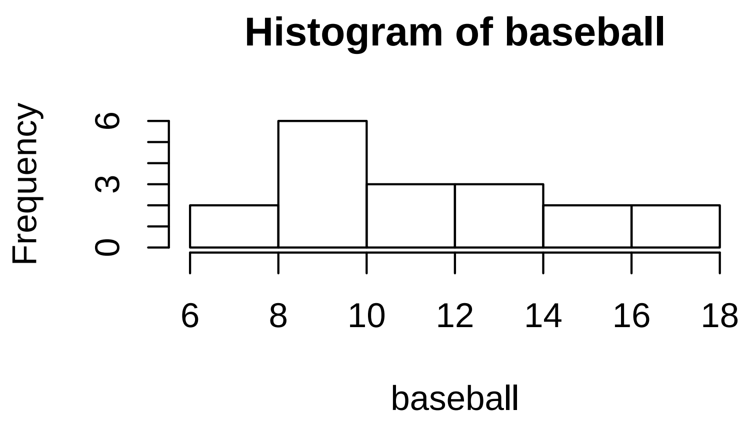
```
faithful <- as.tibble(datasets::faithful)

faithful

## # A tibble: 272 x 2
##    eruptions waiting
##  *     <dbl>   <dbl>
##  1      3.6       79
##  2      1.8       54
##  3      3.33      74
##  4      2.28      62
##  5      4.53      85
##  6      2.88      55
##  7      4.7       88
##  8      3.6       85
##  9      1.95      51
## 10      4.35      85
## # ... with 262 more rows
```

Hypotheses

$H_0 : P(X \text{ is in class } j) = p_j*$
$H_1 : P(X \text{ is in class } j) \neq$ for at least one class

Test Statistic

```
baseball <- c(18, 17, 16, 15, 14, 14, 13, 12, 11, 11, 10, 10, 10, 10, 10, 9, 8, 7)
hist <- hist(baseball)
```




```
breaks <- hist$breaks
counts <- hist$counts

binom_calculator <- function(left_val, right_val, n, p){
  if(right_val != 1){
    pbinom(right_val, p = p, size =  n) - pbinom(left_val, p = p, size = n)
  } else {
    1 - pbinom(left_val, p = p, size = n)
  }
}

p2 <- binom_calculator(6, 8.0, n = 45, p = 0.25)
p3 <- binom_calculator(8, 10, n = 45, p = 0.25)
p4 <- binom_calculator(10, 12, n = 45, p = 0.25)
p5 <- binom_calculator(12, 14, n = 45, p = 0.25)
p6 <- binom_calculator(14, 16, n = 45, p = 0.25)
p7 <- binom_calculator(16, 1, n = 45, p = 0.25)
```

```
prob_vec <- c(p2, p3, p4, p5, p6, p7)

baseball_test <- chisq.test(x = counts, prob_vec) %>%
  tidy()
```

Our test statistic is 12.

P-value

```
pvalue <- 1 - pchisq(baseball_test$statistic, df = length(prob_vec) - 3)
```

Degrees of Freedom Calculation = N - x - 1 = 6 - 2 - 1 = 3, where N is the number of classes and x is the number of parameters in the binomial distribution.

The p-value is $1.63176x10^{-5}$

Conclusion

With a p-value less than 0.05, we can reject the null hypothesis that this sample belongs to a binomially distributed Random Variable with $p = 0.25$ and $n = 45$. There does not seem to be evidence pointing to the fact that this is a binomially distributed sample.

Dataset #3: Running the Shapiro-Wilk test to see if the beaver1's `temp` variable is normally distributed.

```
beaver <- as.tibble(datasets::beaver1)

beaver
```

```
## # A tibble: 114 x 4
##      day  time  temp activ
##    <dbl> <dbl> <dbl> <dbl>
##  1   346   840  36.3     0
##  2   346   850  36.3     0
##  3   346   900  36.4     0
##  4   346   910  36.4     0
##  5   346   920  36.6     0
##  6   346   930  36.7     0
##  7   346   940  36.7     0
##  8   346   950  36.8     0
##  9   346  1000  36.8     0
## 10   346  1010  36.9     0
## # ... with 104 more rows
```

H_0: The random sample comes from a population with the normal distribution, with unknown mean and standard deviation.

H_1: The distribution function the X_i's is nonnormal.

Test Statistic

```
normal_test <- shapiro.test(x = beaver$temp) %>%
  tidy()
```

Our test statistic is 0.97.

P-Value

Our p-value is 0.012.

Conclusion

With a p-value less than 0.05, we have enough evidence to reject the null hypothesis that the random sample comes from a population with unknown mean and standard deviation. There seems to be evidence to support that the distribution function is nonnormal.