Problem Set #3 - Florence

Azka, Jonathan, Jordan

Monday, February 15, 2016

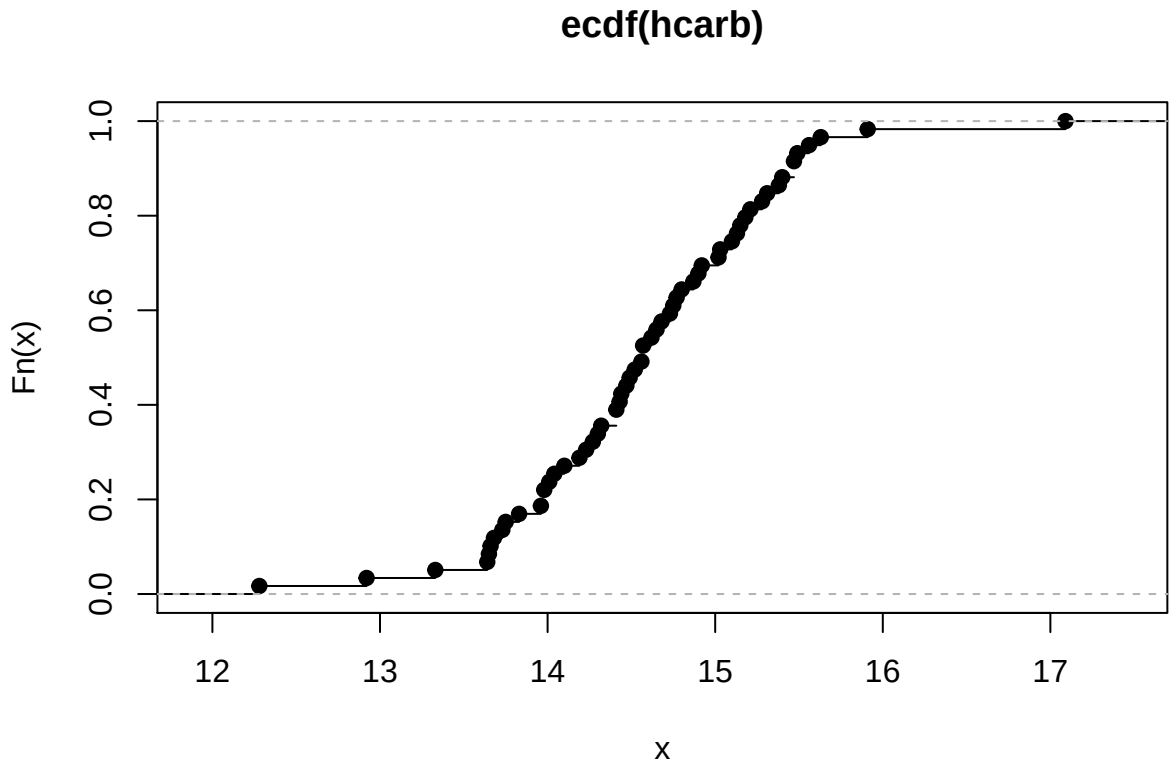
Problem 10.6

Various chemical tests were conducted on beeswax. In particular, the percentage of hydrocarbons in each sample of wax was determined.

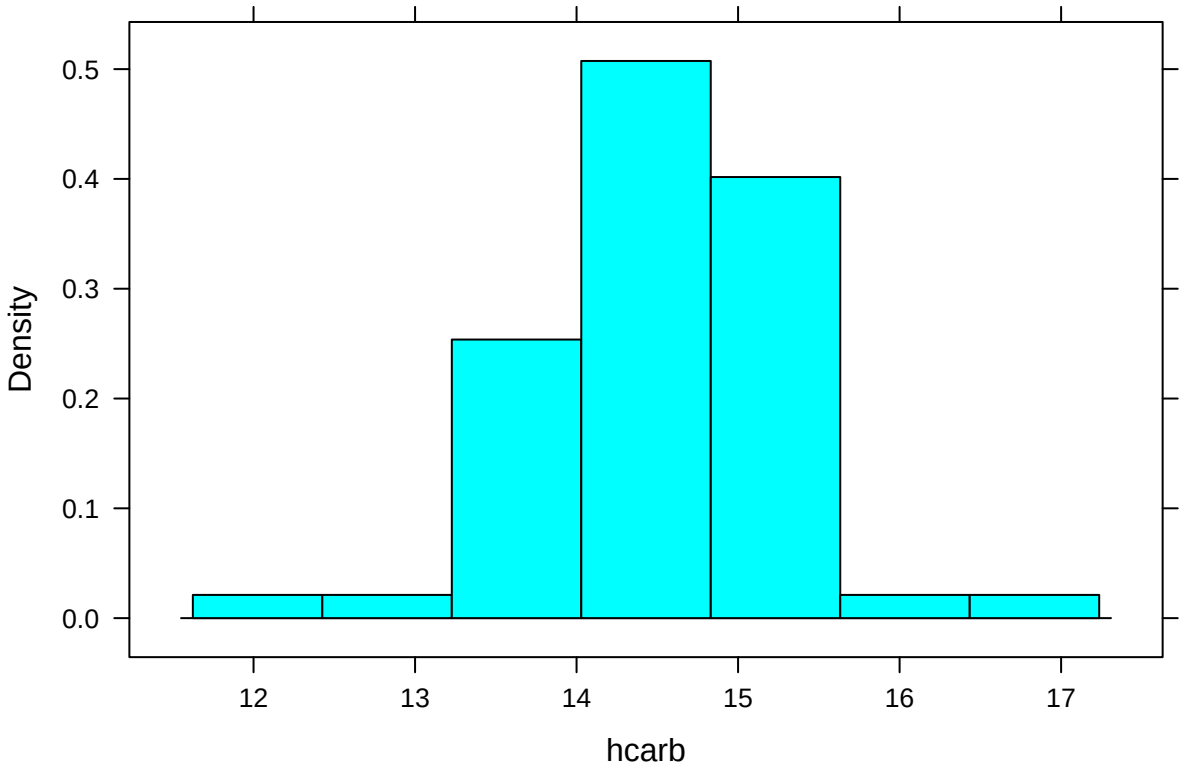
Part (a)

Plot the ecdf, the histogram, and a normal probability plot of the percentages of hydrocarbons. Find the .90, .75, .50, .25, and .10 quantiles. Does the distribution appear Gaussian?

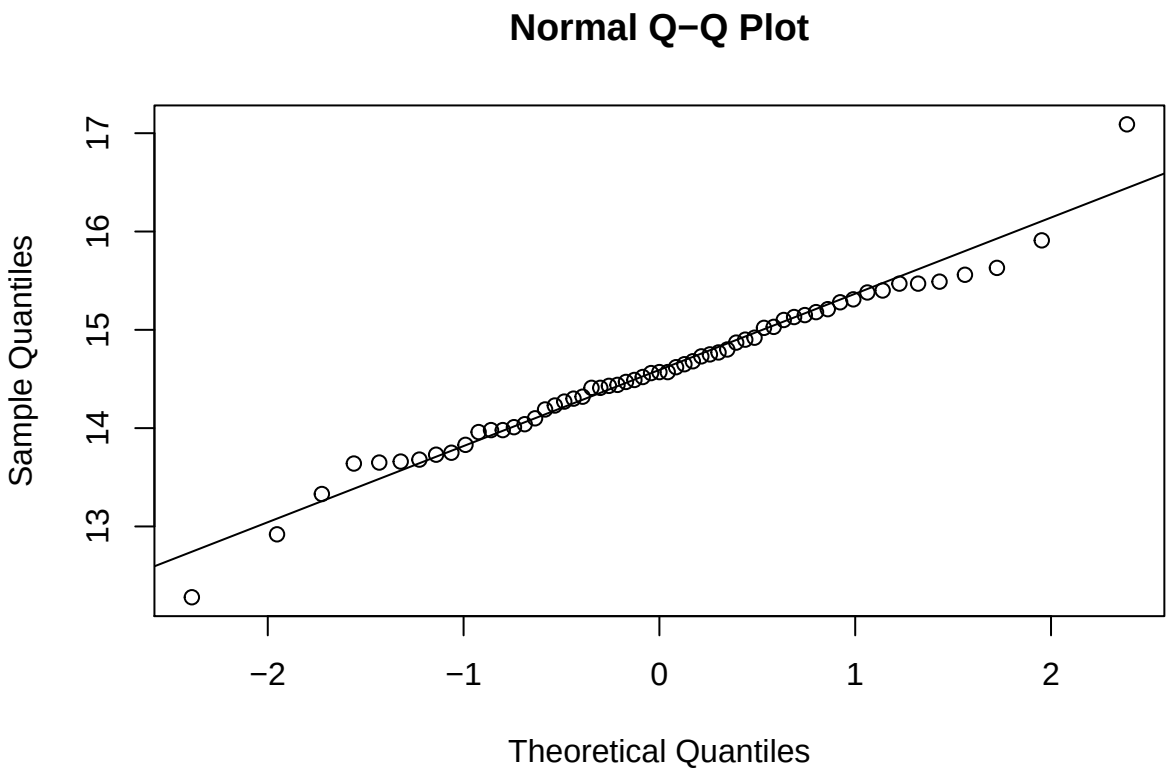
```
beeswax <- read.csv("http://www3.amherst.edu/~nhorton/rice/chapter10/beeswax.csv")
hcarb <- beeswax$Hydrocarbon

hcarb.ecdf <- ecdf(hcarb)
plot(hcarb.ecdf)
```



```
histogram(hcarb)
```




```
qqnorm(hcarb)
qqline(hcarb)
```




```
quantile(hcarb, c(.90, .75, .50, .25, .10))
```

```
##    90%    75%    50%    25%    10%
## 15.470 15.115 14.570 14.070 13.676
```

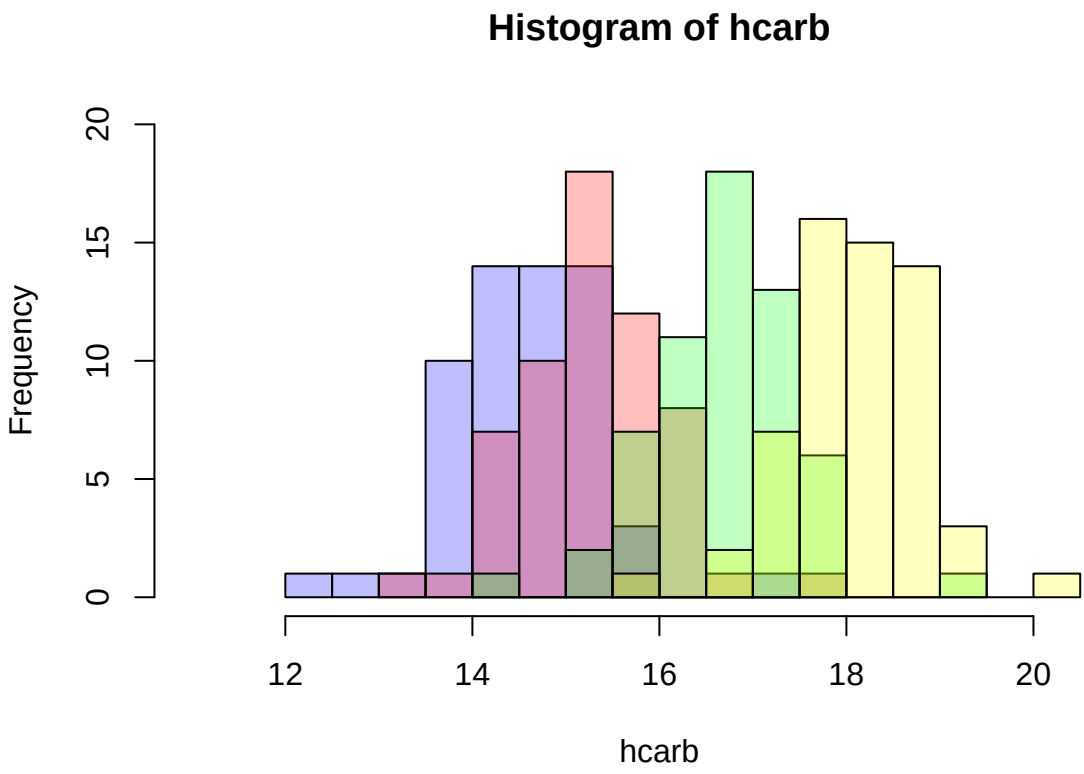
The distribution does appear Gaussian.

Part (b)

The average percentage of hydrocarbons in microcrystalline wax (a synthetic commercial wax) is 85%. Suppose that beeswax was diluted with 1% microcrystalline wax. Could this be detected? What about a 3% or a 5% dilution?

```
undiluted <- hist(hcarb)
synth.01 <- hist(hcarb*.99 + 85*.01)
synth.03 <- hist(hcarb*.97 + 85*.03)
synth.05 <- hist(hcarb*.95 + 85*.05)
```

```
plot(undiluted, col=rgb(0,0,1,1/4), xlim=c(11, 21), ylim = c(0, 20))
plot(synth.01, col=rgb(1,0,0,1/4), xlim=c(11, 21), add = TRUE)
plot(synth.03, col=rgb(0,1,0,1/4), xlim=c(11, 21), add = TRUE)
plot(synth.05, col=rgb(1,1,0,1/4), xlim=c(11, 21), add = TRUE)
```



As the plot shows, a 5% dilution should be quite detectable. A 1% or 3% dilution, though, would be more difficult to detect.

Problem 10.10

Part (a)

Find the mean and variance of $X_(k)$ from a uniform distribution on [0,1].

```
numsim <- 10000

n <- 99 # Number of elements in each sample
k <- 50 # Arbitrary order statistic to examine; here we use median

getOrderStat <- function(){
  x <- runif(n, 0, 1)
  return(sort(x)[k])
}

vals <- do(numsim)*getOrderStat()
mean(vals[,1]); k/(n+1)
```

```
## [1] 0.5006884
```

```
## [1] 0.5
```

```
var(vals[,1]); (1/(n+2))*(k/(n+1))*(1-k/(n+1))
```

```
## [1] 0.002446582
```

```
## [1] 0.002475248
```

Part (b)

Find the approximate mean and variance of $Y_(k)$ from an arbitrary distribution with cdf F.

We use an exponential distribution with rate parameter 1:

```
n <- 99  # Number of elements in each sample
k <- 50  # Arbitrary order statistic to examine; here we use median
lambda <- 1

getOrderStat <- function() {
    x <- rexp(n, rate = lambda)
    return(sort(x)[k])
}

vals <- do(numsim) * getOrderStat()
mean(vals[, 1])
```

```
## [1] 0.697843
```

```
(-1/lambda) * log(1 - k/(n + 1))
```

```
## [1] 0.6931472
```

```
var(vals[, 1])
```

```
## [1] 0.01022147
```

```
(k/(n + 1)) * (1 - k/(n + 1)) * (1/(n + 2)) * (1/(lambda * exp(-lambda * ((-1/lambda) *
    log(1 - k/(n + 1)))))^2)
```

```
## [1] 0.00990099
```

Part (c)

Select some arbitrary pth quantile. For the numsim samples in part (b), show that the variance of the pth quantile is the given formula.

Part (d)

Sample n from rnorm with some arbitrary mean and standard deviation. Do this numsim times. Find the variance of the median, and match it to the derived formula. Find the variance of the mean, and match it to the formula for the variance of the sample mean.

Problem 10.28

We use a standard Normal as our continuous probability distribution. So, the median = 0.

```
# Insert Jordan's simulation
```

Problem 10.50

Part (a)

For each station, plot flow and occupancy versus time. Explain the patterns you see. Can you deduce from the plots what the days of the week were?

```
p1 = xyplot(Lane.1.Flow ~ Timestamp, data=flow)
p2 = xyplot(Lane.1.Occ ~ Timestamp, data=flow)
print(p1, position = c(0, 0.5, 1, 1), more = TRUE)
print(p2, position = c(0, 0, 1, 0.5))
```

```
p3 = xyplot(Lane.2.Flow ~ Timestamp, data=flow)
p4 = xyplot(Lane.2.Occ ~ Timestamp, data=flow)
print(p3, position = c(0, 0.5, 1, 1), more = TRUE)
print(p4, position = c(0, 0, 1, 0.5))
```

```
p5 = xyplot(Lane.3.Flow ~ Timestamp, data=flow)
p6 = xyplot(Lane.3.Occ ~ Timestamp, data=flow)
print(p5, position = c(0, 0.5, 1, 1), more = TRUE)
print(p6, position = c(0, 0, 1, 0.5))
```

There is a sin-curve-like pattern to the plots. The second and third "bumps" look to be slightly less variable than the other bumps, so the days of the week could be Friday, Saturday, Sunday, Monday, Tuesday, Wednesday.

Part (b)

Compare the flows in the three lanes by making parallel boxplots. Which lane typically serves the most traffic?

```
boxplot(flow$Lane.1.Flow, flow$Lane.2.Flow, flow$Lane.3.Flow, names = c("Lane 1", "Lane 2", "Lane 3"))
```

Lane 2 typically serves the most traffic.

Part (c)

```
xyplot(Lane.1.Flow + Lane.2.Flow + Lane.3.Flow ~ Timestamp, data=flow, auto.key = TRUE)
```

The flow in lane 2 looks to typically be greater than the flow in lane 1, which looks to typically by greater than the flow in lane 3.

Part (d)

```
mean(flow$Lane.1.Occ); median(flow$Lane.1.Occ)
```

```
## [1] 0.06120776
```

```
## [1] 0.0478
```

```
mean(flow$Lane.2.Occ); median(flow$Lane.2.Occ)
```

```
## [1] 0.06118247
```

```
## [1] 0.05505
```

```
mean(flow$Lane.3.Occ); median(flow$Lane.3.Occ)
```

```
## [1] 0.05123862
```

```
## [1] 0.0413
```

In all three lanes, the medians are less than the means. This suggests that the data are skewed right, with some high-valued outliers that increase the mean.

Part (e)

```
histogram(flow$Lane.1.Occ, breaks = 10)
histogram(flow$Lane.1.Occ, breaks = 20)

histogram(flow$Lane.2.Occ, breaks = 10)
histogram(flow$Lane.2.Occ, breaks = 20)

histogram(flow$Lane.3.Occ, breaks = 10)
histogram(flow$Lane.3.Occ, breaks = 20)
```

20 bins looks to give a good representation of the shape of the distribution (skewed right, slightly bimodal). The bimodal nature can be explained by rush hours. Most of the time, when traffic is light, cars do not occupy the same area for very long. Sometimes, though, when traffic is heavy, cars tend to occupy the same area for longer times.

Part (f)

```
xyplot(Lane.1.Occ + Lane.2.Occ + Lane.3.Occ ~ Timestamp, data=flow, auto.key = TRUE)
```

From this plot, we can tell that all 3 lanes tend to have high occupancies at around the same times.

Part (g)

```
xyplot(Lane.1.Flow ~ Lane.1.Occ, data = flow)
xyplot(Lane.2.Flow ~ Lane.2.Occ, data = flow)
xyplot(Lane.3.Flow ~ Lane.3.Occ, data = flow)
```

The first part of the conjecture seems plausible. When occupancy is low, it tends to increase with flow. Beyond a certain point (around Occ=0.2), however, flow begins to decrease with greater occupancy.

Part (h)

Part (i)

Part (j)

Part (k)

Section i

Section ii

Section iii

Part (l)