Problem Set #3 - Florence

Azka, Jonathan, Jordan

Monday, February 15, 2016

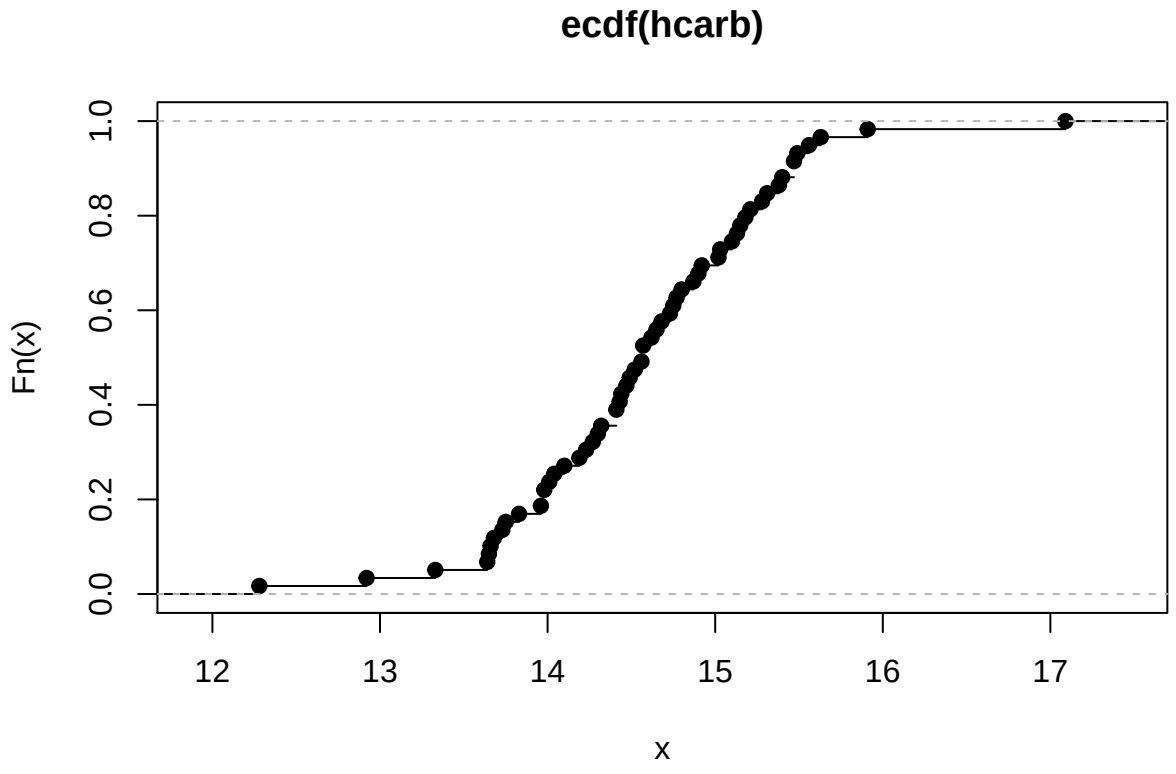
Problem 10.6

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

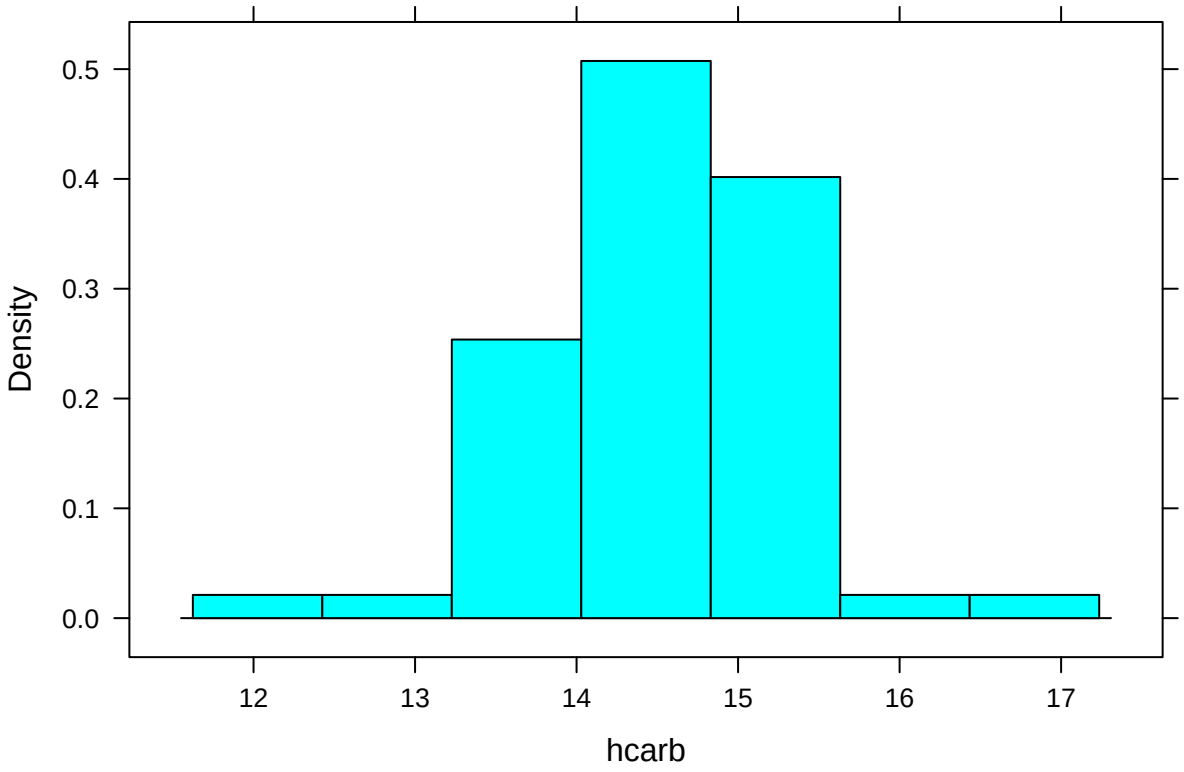
Part (a)

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?

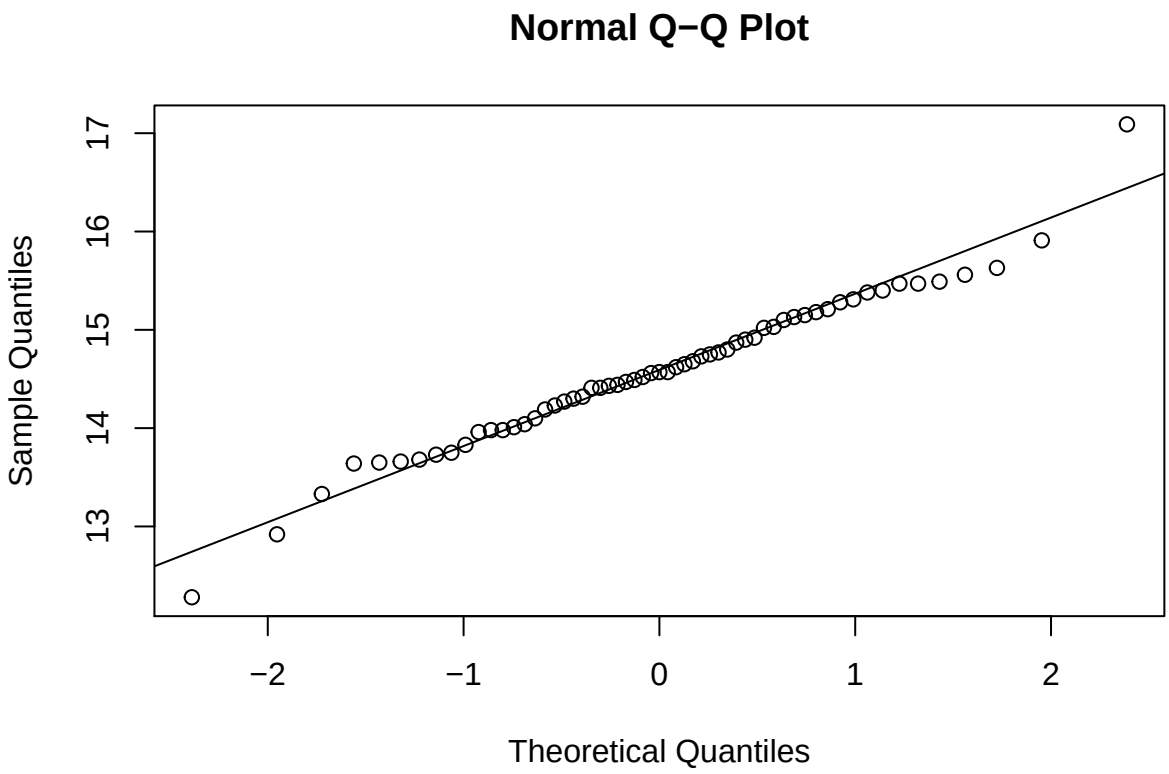
```
beeswax <- read.csv("http://www3.amherst.edu/~nhorton/rice/chapter10/beeswax.csv")
hcarb <- beeswax$Hydrocarbon

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.

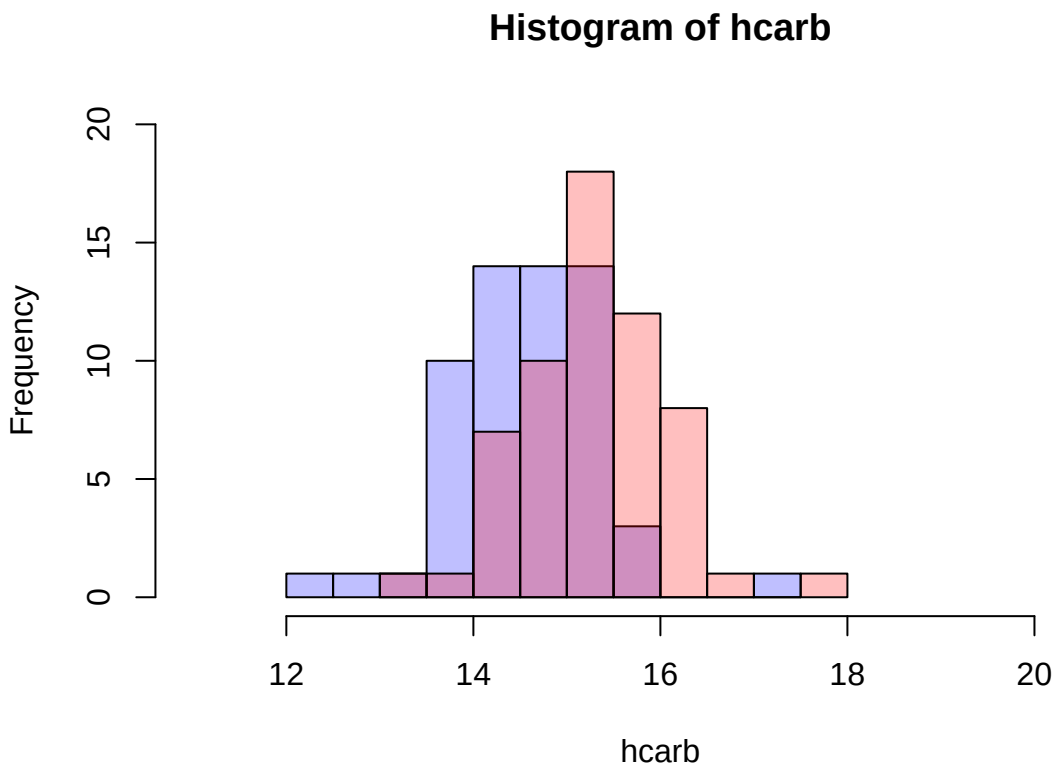
Part (b)

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?

To help illustrate these dilutions, we plot histograms of the original beeswax compared to histograms of the diluted beeswaxes.

```
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)
```

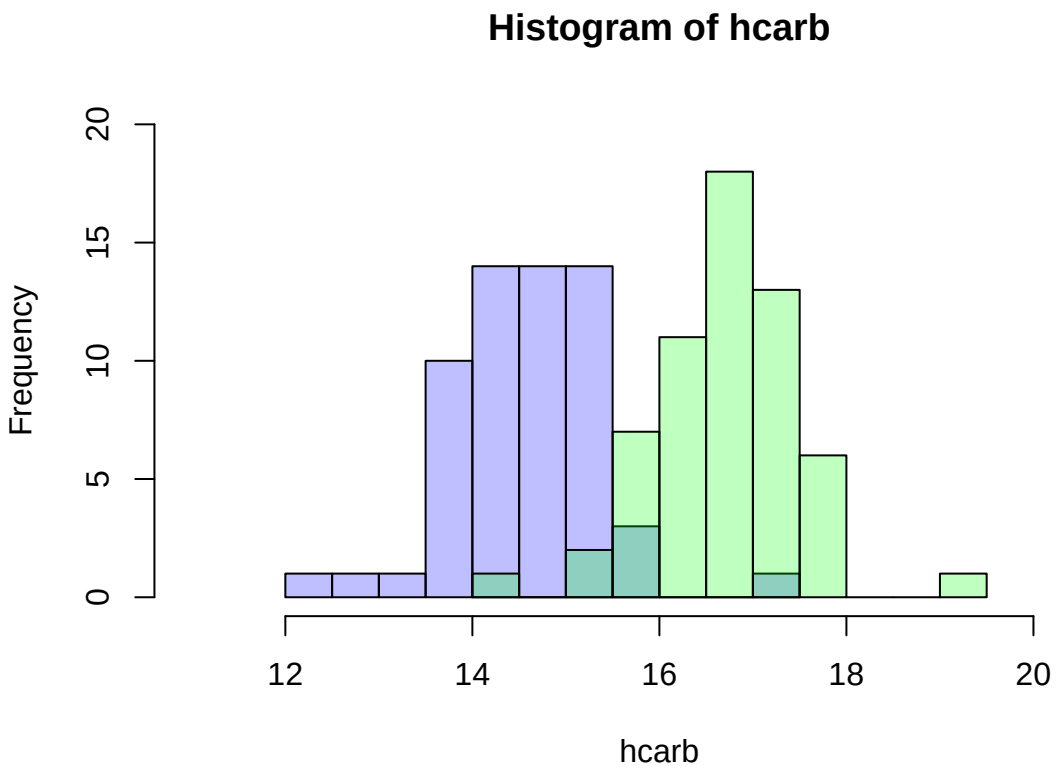



```
(mean(hcarb*.99 + 85*.01) - mean(hcarb))
```

```
## [1] 0.7042
```

If we examine the mean of the 1% dilutions, we notice that it is only .70 standard deviations away from the mean of the originals, i.e. within 1 standard deviation of the original mean. As the histograms above show, it would be difficult to detect a 1% dilution.

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




```
(mean(hcarb*.97 + 85*.03) - mean(hcarb))
```

```
## [1] 2.1126
```

If we examine the mean of the 3% dilutions, we notice that it is 2.11 standard deviations away from the mean of the originals, which is not much more than 2 standard deviations. As the histograms above show, it would be somewhat difficult to detect a 3% dilution.

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


```
(mean(hcarb*.95 + 85*.05) - mean(hcarb))
```

```
## [1] 3.521
```

If we examine the mean of the 5% dilutions, we notice that it is 3.52 standard deviations away from the mean of the originals, which is more than 3 standard deviations away from the original mean. As the histograms above show, it would be relatively easy to detect a 5% dilution.

Problem 10.10

Let's take a sample of size n from a Normal distribution with mean μ and standard deviation σ. We want to determine the variance of the sample median and compare it to the variance of the sample mean.

To make life easy, let's use $n = 99$, $\mu = 0$, and $\sigma = 1$. We will simulate draws of size 99 from a Normal(0,1) and examine the sample medians and means (This is part (d)).

```
numsim <- 10000

n <- 99 # Number of elements in each sample
mean <- 0
sd <- 1

medVals <- do(numsim)*median(rnorm(n, mean = mean, sd = sd))
var(medVals[,1])
```

```
## [1] 0.01567084
```

So, the variance of the sample median is 0.0156708.

```
meanVals <- do(numsim)*mean(rnorm(n, mean = mean, sd = sd))
var(meanVals[,1])
```

```
## [1] 0.01023234
```

So, the variance of the sample means is 0.0102323.

If we look at the ratio between the variances, we get 1.5315021, which looks suspiciously like $\frac{\pi}{2} \approx 1.571$. In fact, it is $\frac{\pi}{2}$, as we will show with the following (heavily simplified) analytical excursion.

To find the analytical ratio between the variance of the sample median and the sample mean, we first need to find the actual variances of the sample median and the sample mean. Rice gives us a formula for the approximate variance of the kth order statistic (see part (b)). If we plug in k = 50 for the median of a sample of n = 99, we get:

$$Var(Y_{50}) \approx \frac{1}{2}(1-\frac{1}{2})(\frac{1}{f(F^{-1}(\frac{1}{2}))^2})(\frac{1}{101})$$

Because we are using a Normal(0,1), we know that $F^{-1}(1/2)$ is just the mean, which is just the median (See part (a) for more generality). So, we can just substitute in $\mu = 0$.

$$Var(Y_{50}) \approx \frac{1}{2}(1-\frac{1}{2})(\frac{1}{f(0)^2})(\frac{1}{101})$$

Now, we evaluate the pdf of the Normal(0,1) at 0.

So, we have found the variance of the median to be $\frac{\pi}{2n}$.

From Rice (page 207), we know that the variance of the sample mean is $\frac{\sigma^2}{n}$, which for us is just $\frac{1}{n}$. Thus, the ratio of the variance of the sample median to the variance of the sample mean is just $\frac{\pi}{2}$.

Problem 10.28

For a sample of size n = 3, median = η, we know that:

$$P(x_1 < \eta < x_2) = 1 - P(\eta < x_1) - P(\eta > x_2)$$

$$= 1 - P(0 \textit{ observations less than } \eta) - P(0 \textit{ observations greater than } \eta) - P(1 \textit{ observation greater than } \eta)$$

Using binomial distributions with p = .5 (since the probability that any random observation is greater than or less than the median is .5), we get:

$$P(x_1 < \eta < x_2) = 1 - \frac{1}{8} - \frac{1}{8} - \frac{3}{8} = \frac{3}{8}$$

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

```
isBetween12 <- function() {
  sample <- rnorm(3, 0, 1)
  order1 <- min(sample)
  order2 <- median(sample)
  return(order1<0 & order2>0)
}
x <- do(10000) * isBetween12()
tally(~x, format='prop')
```

```
##
##   TRUE  FALSE
## 0.3698 0.6302
```

We get an empirical solution of about $\frac{3}{8}$.

For the second part:

$$P(x_1 < \eta < x_3) = 1 - P(\eta < x_1) - P(\eta > x_3)$$
$$= 1 - P(0 \textit{ observations less than } \eta) - P(0 \textit{ observations greater than } \eta)$$

Again using binomial distributions with p = .5, we get:

$$P(x_1 < \eta < x_2) = 1 - \frac{1}{8} - \frac{1}{8} = \frac{6}{8} = \frac{3}{4}$$

Empirically, we use a standard Normal again (median = 0).

```
isBetween13 <- function() {
  sample <- rnorm(3, 0, 1)
  order1 <- min(sample)
  order3 <- max(sample)
  return(order1<0 & order3>0)
}
x <- do(10000) * isBetween13()
tally(~x, format='prop')
```

```
##
##   TRUE  FALSE
## 0.7421 0.2579
```

We get an empirical solution of about $\frac{3}{4}$.

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.