STAT 6502 HW #1

Daniel Park

January 6, 2016

Chapter 10: 5, 6, 10, 11, 15, 29

5.

Let $X_1, ..., X_n$ be a sample (i.i.d.) from a distribution function, F, and let F_n denote the ecdf. Show that

$$\text{Cov}[F_n(u), F_n(v)] = \frac{1}{n}[F(m) - F(u)F(v)]$$

where $m = \min(u, v)$. Conclude that $F_n(u)$ and $F_n(v)$ are positively correlated: If $F_n(u)$ overshoots $F(u)$, then $F_n(v)$ will tend to overshoot $F(v)$.

We begin with a property of the covariance:

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

and

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{(-\infty,x]}(X_i)$$

So

$$\begin{aligned}
\text{Cov}[F_n(u), F_n(v)] &= E[F_n(u)F_n(v)] - E[F_n(u)]E[F_n(v)] \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{I}_{(-\infty,u]}(U_i)\frac{1}{n}\sum_{j=1}^{n}\mathbf{I}_{(-\infty,v]}(V_j)\right] - E\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{I}_{(-\infty,u]}(U_i)\right]E\left[\frac{1}{n}\sum_{j=1}^{n}\mathbf{I}_{(-\infty,v]}(V_j)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[P(U_i \le u, V_j \le v)] - \frac{1}{n}nE[\mathbf{I}_{(-\infty,u]}(U_i)]\frac{1}{n}nE[\mathbf{I}_{(-\infty,v]}(V_j)] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[P(U_i \le u, V_j \le v)] - P(U_i \le u)P(V_j \le v) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[P(U_i \le u, V_j \le v)] - F(u)F(v) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[P(U_i \le u, V_j \le v)] - \sum_{i=1}^{n}\frac{1}{n}F(u)\sum_{j=1}^{n}\frac{1}{n}F(v) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[P(U_i \le u, V_j \le v)] - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}F(u)F(v) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[P(U_i \le u, V_j \le v) - F(u)F(v)] \\
&= \frac{1}{n^2}\sum_{i=j}[P(U_i \le u, V_j \le v) - F(u)F(v)] + \frac{1}{n^2}\sum_{i\ne j}[P(U_i \le u, V_j \le v) - F(u)F(v)]
\end{aligned}$$

Recall that if two variables are independent, the joint probability is equal to the product of their individual probabilities. In this case,

$$\begin{aligned}
P(U_i \le u, V_j \le v) &= P(U_i \le u)P(V_j \le v) \\
&= F(u)F(v)
\end{aligned}$$

Also, note that the joint cdf is equal to the cdf of the smaller parameter, u or v. If $u < v$, and $U_i < u$, then $U_i < v$.

So, returning back to the covariance,

$$
\begin{aligned}
\text{Cov}[F_n(u), F_n(v)] &= \frac{1}{n^2}\sum_{i=j}[P(U_i \leq u, V_j \leq v) - F(u)F(v)] + \frac{1}{n^2}\sum_{i\neq j}[P(U_i \leq u, V_j \leq v) - F(u)F(v)] \\
&= \frac{1}{n^2}\sum_{i=j}[P(U_i \leq u, V_j \leq v) - F(u)F(v)] + \frac{1}{n^2}\sum_{i\neq j}[F(u)F(v) - F(u)F(v)] \\
&= \frac{1}{n^2}\sum_{i=j}[P(U_i \leq u, V_j \leq v) - F(u)F(v)] + 0 \\
&= \frac{1}{n^2}n[F(m) - F(u)F(v)] \\
&= \frac{1}{n}[F(m) - F(u)F(v)]
\end{aligned}
$$

where $m = \min(u, v)$. Finally, we note that

$$\text{Cov}[F_n(u), F_n(v)] > 0$$

since $F(m), F(u)$, and $F(u)$ are less than or equal to 1, $F(m) = F(u)$ or $F(m) = F(v)$, which means that the product of two values less than 1 will be less than one of the values. In other words,

$$F(m) - F(u)F(v)$$

We can then conclude that the correlation between u and v is positive since

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

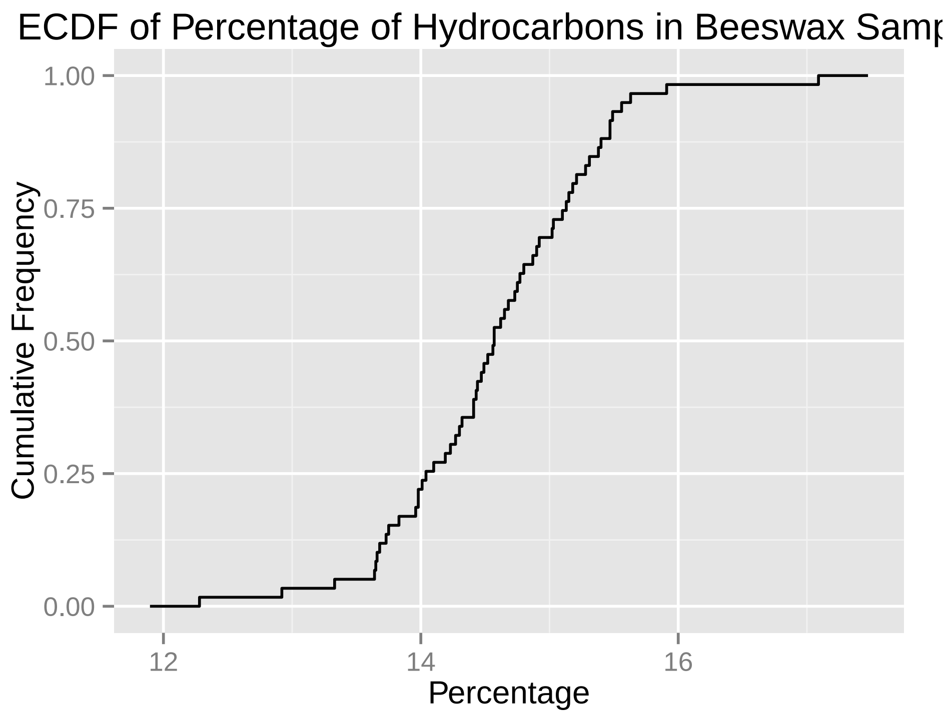
6.

Various chemical tests were conducted on beeswax by White, Riethof, and Kushnir (1960). In particular, the percentage of hydrocarbons in each sample of wax was determined.

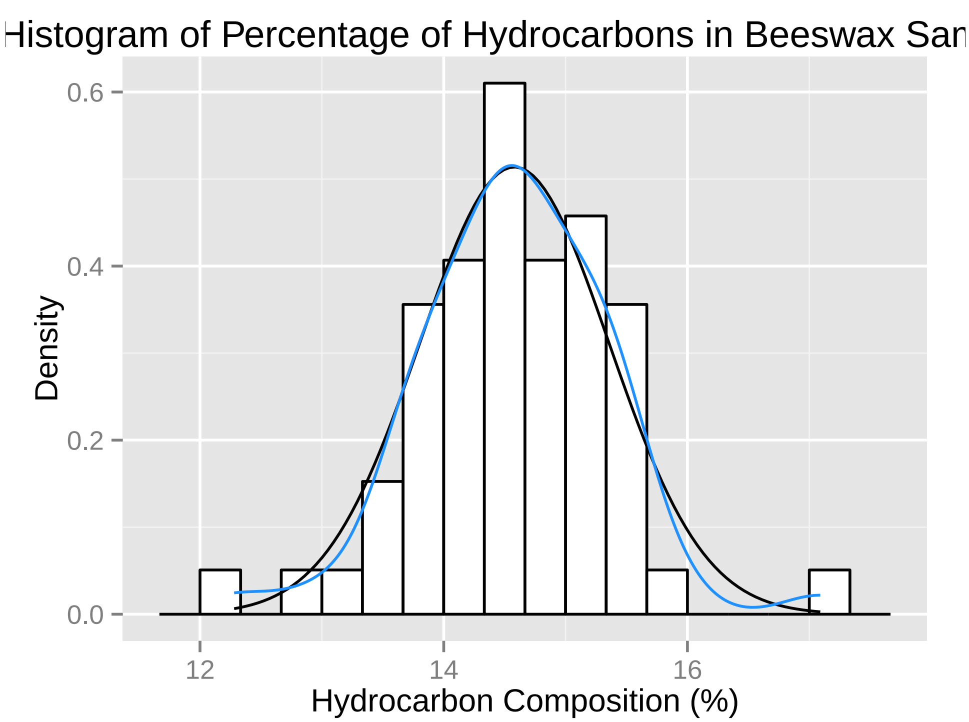
6a.

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

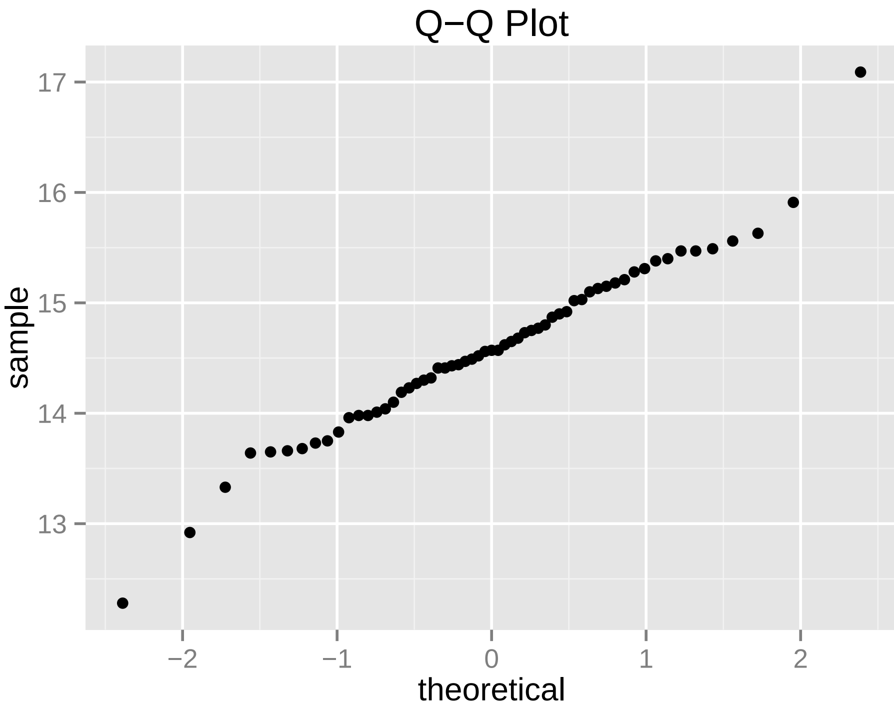
```r
hydrobject <- ggplot(wax, aes(Hydrocarbon))
hydrobject +
  stat_ecdf(geom = "step") +
  labs(title="ECDF of Percentage of Hydrocarbons in Beeswax Samples",
       x='Percentage',y='Cumulative Frequency')
```




```r
hydrobject +
  geom_histogram(mapping=aes(y=..density..),color="black", fill = "white", binwidth=1/3) +
  stat_function(fun=dnorm, args=list(mean=mean(wax$Hydrocarbon), sd=sd(wax$Hydrocarbon))) +
  geom_line(mapping=aes(color="Hydro Density"), stat="density", color='dodgerblue') +
  labs(title="Histogram of Percentage of Hydrocarbons in Beeswax Samples",
    x='Hydrocarbon Composition (%)',y='Density')
```



```
ggplot(wax, aes(sample=Hydrocarbon)) +
  stat_qq() +  labs(title="Q-Q Plot")
```




```
quantile(wax$Hydrocarbon, c(.1, .25, .5, .75, .9))
```

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

The graphs make the data appear to be normally distributed.

6b.

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? (Such questions were one of the main concerns of the beeswax study.)

Following the Example A from the book, we can examine the effect of dilution in relation to the range of the observations.

Diluting the beeswax with 1% microcrystalline wax would race the hydrocarbon percentage by $85 \cdot 0.01 = 0.85$.

```
range(wax$Hydrocarbon)
```

```
## [1] 12.28 17.09
```

With a range of $17.09 - 12.28 = 4.81$, it would be difficult to detect the dilution if it were done to beeswax that had a low hydrocarbon composition.

There is one large outlier, at 17.09. The next largest value is 15.91. If we were to subtract 0.85 from 15.91,

```r
sum(wax$Hydrocarbon < (sort(wax$Hydrocarbon)[58]-0.85))
```

```
## [1] 43
```

we could dilute 43 of the samples, and their hydrocarbon percentage would still be less than the second largest value

Diluting the beeswax with 3% microcrystalline wax would race the hydrocarbon percentage by $85 \cdot 0.03 = 2.55$. Using the same approach,

```r
sum(wax$Hydrocarbon < (sort(wax$Hydrocarbon)[58]-2.55))
```

```
## [1] 3
```

we find that are only 3 samples that could be diluted that would still have a lower hydrocarbon percentage than the second largest value. The likelihood of detection would be relatively high.

Diluting the beeswax with 5% microcrystalline wax would race the hydrocarbon percentage by $85 \cdot 0.05 = 4.25$. This dilution would be easy to detect.

10.

Let $X_1, ..., X_n$ be a sample from cdf F and denote the order statistics by $X_{(1)}, X_{(2)}, ..., X_{(n)}$. We will assume that F is continuous, with density function f. From Theorem A in Section 3.7, the density function of $X_{(k)}$ is

$$f_k(x) = n\binom{n-1}{k-1}[F(x)]^{k-1}[1-F(x)]^{n-k}f(x)$$

10a.

Find the mean and variance of $X_{(k)}$ from a uniform distribution on $[0,1]$. You will need to use the fact that the density of X_k integrates to 1. Show that

$$\begin{aligned}
\text{Mean} &= \frac{k}{n+1} \\
\text{Variance} &= \frac{1}{n+2}\left(\frac{k}{n+1}\right)\left(1-\frac{k}{n+1}\right)
\end{aligned}$$

For $U[0,1]$, $f(x) = 1$. So

$$\begin{aligned}
f_k(x) &= n\binom{n-1}{k-1}[F(x)]^{k-1}[1-F(x)]^{n-k} \cdot 1 \\
&= n\frac{(n-1)!}{(k-1)!(n-k)!}[F(x)]^{k-1}[1-F(x)]^{n-k} \\
&= n\frac{\Gamma(n)}{\Gamma(k)\Gamma(n-k+1)}[F(x)]^{k-1}[1-F(x)]^{n-k} \\
&= n\frac{\Gamma(n+1)}{n\Gamma(k)\Gamma(n-k+1)}[F(x)]^{k-1}[1-F(x)]^{(n-k+1)-1} \\
&= \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)}[F(x)]^{k-1}[1-F(x)]^{(n-k+1)-1}
\end{aligned}$$

This resembles the beta distribution, $B(\alpha, \beta)$, where $\mu = \frac{\alpha}{\alpha+\beta}$ and $\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

So

$$\mu = \frac{k}{n+1}$$

and

$$\begin{aligned}
\sigma^2 &= \frac{k(n-k+1)}{(n+1)^2(n+2)} \\
&= \frac{1}{n+2}\frac{k}{n+1}\frac{n+1-k}{n+1} \\
&= \frac{1}{n+2}\frac{k}{n+1}\left(\frac{n+1}{n+1} - \frac{k}{n+1}\right) \\
&= \frac{1}{n+2}\left(\frac{k}{n+1}\right)\left(1 - \frac{k}{n+1}\right)
\end{aligned}$$

10b.

Find the approximate mean and variance of $Y_{(k)}$, the kth-order statistic of a sample of size n from F. To do this, let

$$X_i = F(Y_i)$$

or

$$Y_i = F^{-1}(X_i)$$

The X_i are a sample from a $U[0,1]$ distribution (why?). Use the propagation of error formula,

$$\begin{aligned}
Y_{(k)} &= F^{-1}(X_{(k)}) \\
&\approx F^{-1}\left(\frac{k}{n+1}\right) + \left(X_{(k)} - \frac{k}{n+1}\right)\frac{d}{dx}F^{-1}(x)\bigg|_{k/(n+1)}
\end{aligned}$$

and argue that

$$\begin{aligned}
E(Y_{(k)}) &\approx F^{-1}\left(\frac{k}{n+1}\right) \\
\operatorname{Var}(Y_{(k)}) &\approx \frac{k}{n+1}\left(1 - \frac{k}{n+1}\right)\frac{1}{(f\{F^{-1}[k/(n+1)]\})^2}\left(\frac{1}{n+2}\right)
\end{aligned}$$

In finding the mean of the approximate value of $Y_{(k)}$, we note that many of the terms are constants:

$$\begin{aligned}
E\big(Y_{(k)}\big) &\approx E\left[F^{-1}\left(\frac{k}{n+1}\right) + \left(X_{(k)} - \frac{k}{n+1}\right)\frac{d}{dx}F^{-1}(x)\bigg|_{k/(n+1)}\right] \\
&= E\left[F^{-1}\left(\frac{k}{n+1}\right)\right] + E\left[\left(X_{(k)} - \frac{k}{n+1}\right)\frac{d}{dx}F^{-1}(x)\bigg|_{k/(n+1)}\right] \\
&= F^{-1}\left(\frac{k}{n+1}\right) + \frac{d}{dx}F^{-1}(x)\bigg|_{k/(n+1)} \cdot E\left[X_{(k)} - \frac{k}{n+1}\right] \\
&= F^{-1}\left(\frac{k}{n+1}\right) + \frac{d}{dx}F^{-1}(x)\bigg|_{k/(n+1)} \cdot 0 \\
&= F^{-1}\left(\frac{k}{n+1}\right)
\end{aligned}$$

Note that $E\big[X_{(k)} - k/(n+1)\big] = 0$ since it is essentially expressing the property $E(X - \mu_X) = 0$.

To solve for $\mathrm{Var}(Y_{(k)})$, we require taking the derivative of an inverse function:

$$\frac{d}{dx}f^{-1}(x) = \frac{1}{\frac{d}{dx}f(f^{-1}(x))}$$

Also, we will use the property

$$\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$$

So

$$\begin{aligned}
\mathrm{Var}(Y_{(k)}) &= \mathrm{Var}\left[F^{-1}\left(\frac{k}{n+1}\right) + \left(X_{(k)} - \frac{k}{n+1}\right)\frac{d}{dx}F^{-1}(x)\bigg|_{k/(n+1)}\right] \\
&= \mathrm{Var}\left[\left(X_{(k)} - \frac{k}{n+1}\right)\frac{d}{dx}F^{-1}(x)\bigg|_{k/(n+1)}\right] \\
&= \left(\frac{d}{dx}F^{-1}(x)\bigg|_{k/(n+1)}\right)^2 \mathrm{Var}\left(X_{(k)} - \frac{k}{n+1}\right) \\
&= \left(\frac{d}{dx}F^{-1}(x)\bigg|_{k/(n+1)}\right)^2 \mathrm{Var}(X_{(k)}) \\
&= \frac{1}{\left(\frac{d}{dx}F(F^{-1}(x))\Big|_{k/(n+1)}\right)^2}\mathrm{Var}(X_{(k)}) \\
&= \frac{1}{\left(f\{F^{-1}(x)\}\Big|_{k/(n+1)}\right)^2}\mathrm{Var}(X_{(k)}) \\
&= \frac{1}{(f\{F^{-1}(k/(n+1)\})^2}\mathrm{Var}(X_{(k)}) \\
&= \frac{1}{(f\{F^{-1}(k/(n+1)\})^2}\frac{1}{n+2}\left(\frac{k}{n+1}\right)\left(1 - \frac{k}{n+1}\right)
\end{aligned}$$

10c.

Use the results of parts (a) and (b) to show that the variance of the pth sample quantile is approximately

$$\frac{1}{nf^2(x_p)}p(1-p)$$

where x_p is the pth quantile.

I'd rather not do this problem.

10d.

Use the result of part (c) to find the approximate variance of the median of a sample of size n from a $N(\mu, \sigma^2)$ distribution. Compare to the variance of the sample mean.

The pdf of the normal distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

In a normal distribution, the median is equal to the mean, and $p = 0.5$.

Then using the formula in 10c,

$$\begin{aligned}
\frac{1}{nf^2(x_p)}p(1-p) &= \frac{1}{n\frac{1}{(\sigma\sqrt{2\pi})^2}}0.5(1-0.5) \\
&= \frac{2\sigma^2\pi}{4n} \\
&= \frac{\sigma^2\pi}{2n}
\end{aligned}$$

A simulation

```r
set.seed(123)
trials <- 10000
n <- 100
x<-numeric(trials)
for (i in 1:trials){
  x[i] <- median(rnorm(n)) # mu=0, sigma=1
}
var(x) # simulated variance
```

```
## [1] 0.01555409
```

```r
1*pi/(2*n) # estimated variance
```

```
## [1] 0.01570796
```

In comparison to the sample mean, $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$, we see that the variance of the sample median is slightly larger.

11.

Calculate the hazard function for

$$F(t) = 1 - e^{-\alpha t^\beta}, \quad t \geq 0$$

The hazard function is defined as

$$h(t) = \frac{f(t)}{1-F(t)}$$

The pdf of our provided cdf is

$$\begin{aligned}
f(t) &= F'(t) \\
&= 0 - e^{-\alpha t^\beta}(-\alpha\beta)t^{\beta-1} \\
&= \alpha\beta e^{-\alpha t^\beta}t^{\beta-1}
\end{aligned}$$

So the hazard function is

$$\begin{aligned}
h(t) &= \frac{\alpha\beta\exp\big(-\alpha t^\beta\big)t^{\beta-1}}{1-\big[1-\exp\big(-\alpha t^\beta\big)\big]} \\
&= \frac{\alpha\beta\exp\big(-\alpha t^\beta\big)t^{\beta-1}}{\exp\big(-\alpha t^\beta\big)} \\
&= \alpha\beta t^{\beta-1}
\end{aligned}$$

15.

A prisoner is told that he will be released at a time chosen uniformly at random within the next 24 hours. Let T denote the time that he is released. What is the hazard function for T? For what values of t is it smallest and largest? If he has been waiting for 5 hours, is it more likely that he will be released in the next few minutes than if he has been waiting for 1 hour?

pdf: $f(t) = \frac{1}{24}$
cdf: $F(t) = \frac{t}{24}$

Hazard function:

$$\begin{aligned} h(t) &= \frac{\frac{1}{24}}{1 - \frac{t}{24}} \\ &= \frac{1}{24 - t} \end{aligned}$$

29.

Of the 26 measurements of the heat of sublimation of platinum, 5 are outliers (see Figure 10.10). Let N denote the number of these outliers that occur in a bootstrap sample (sample with replacement) of the 26 measurements.

29a.

Explain why the distribution of N is binomial.

If N denotes the number of outliers, each measurement is considered an outlier or not an outlier. Therefore, each observation is a Bernoulli random variable, and N is binomially distributed.

29b.

Find $P(N \geq 10)$.

Based on our data, $p = 5/26$.

The probability that $N = x$ is given by the binomial pdf:

$$P(N = x) = \binom{26}{n} \left(\frac{5}{26}\right)^x \left(1 - \frac{5}{26}\right)^{26-x}$$

We can use the `pbinom()` function in `R` to determine $P(N \geq 10)$:

```
pbinom(q=9, size=26, prob=5/26, lower.tail=FALSE)
```

```
## [1] 0.01787622
```

29c.

In 1000 bootstrap samples, how many would you expect to contain 10 or more of these outliers? (Also, run a bootstrap to compare.)

```r
# Smallest outlier in sample:
threshold <- sort(platinum, decreasing=TRUE)[5]
set.seed(123)
mean(replicate(n=100,
               expr=sum(replicate(n=1000,
                                  expr=sum(sample(x=platinum,
                                                  size=length(platinum),
                                                  replace=TRUE)>= threshold) >= 10))))
```

```
## [1] 18.55
```

29d.

What is the probability that a bootstrap sample is composed entirely of these outliers?

With a 5/26th chance that any given observation is an outlier, the projected probability that an entire sample is entirely composed of outliers is

$$\left(\frac{5}{26}\right)^{26} = 2.420544 \times 10^{-19}$$