Midterm Take-Home

Donald Lee Beadle

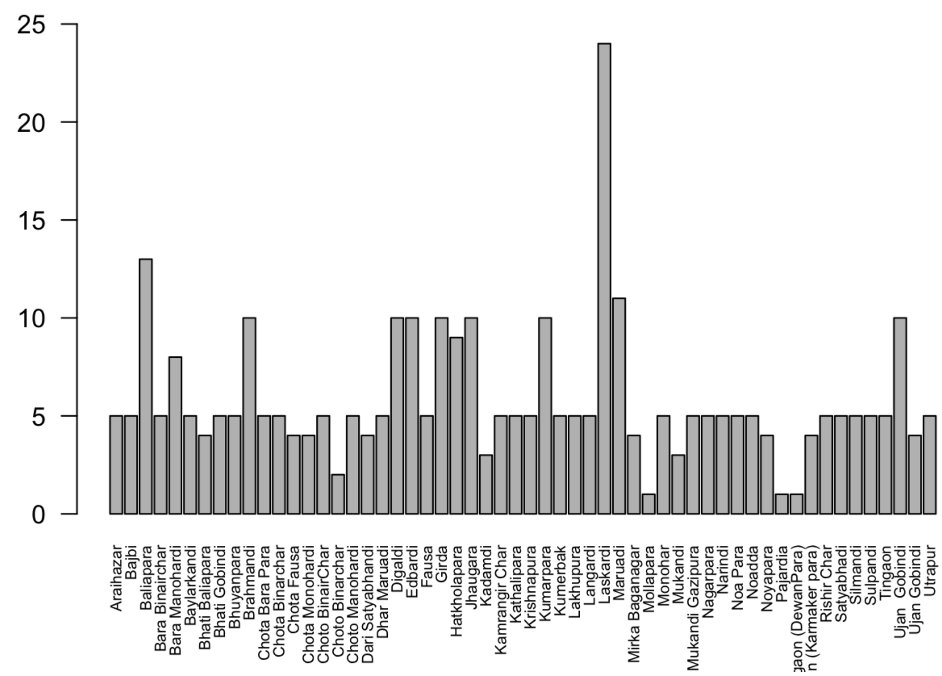
11/03/17

Question 1

Using the **ASWELLS** data set, which deals with arsenic in groundwater wells in Bangladesh (*Environmental Science & Technology*),

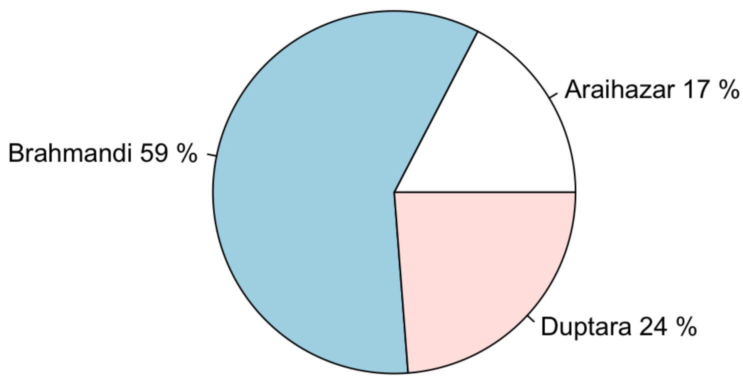
1. Construct a bar graph (with labels) of the *Village* data.
    - Note to have all the labels showing, add the command *las=2* in your bar graph command.

```
c<-count(aswells.village)
data<-c[,2]
names(data)<-c[,1]
barplot(data, cex.names=.6, ylim = c(0,25), las=2)
```



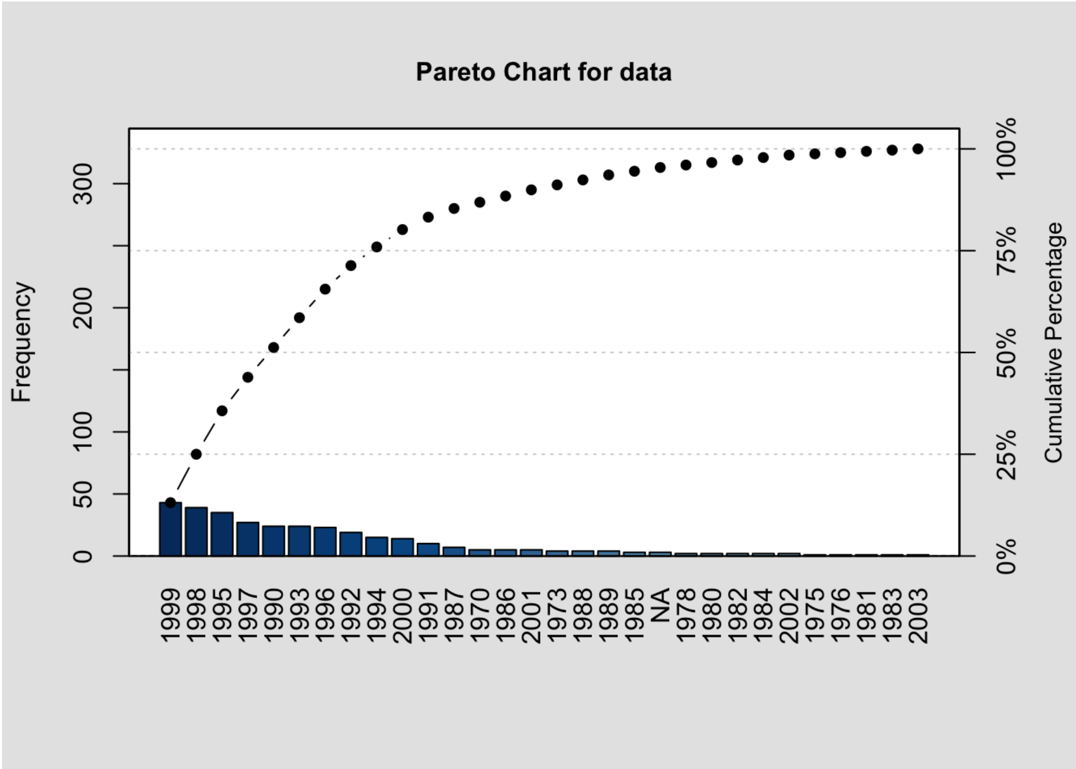
2. Construct a pie chart (with labels, percentages, and % symbol) of the *Union* data.

```
c<-count(aswells.union)
data<-c[,2]
names(data)<-c[,1]
cumm.freq<-data/sum(data)
pie(data,labels = paste(paste(c[,1],round(cumm.freq *100)),"%"))
```



3. Construct a Pareto diagram (with labels) of the *Year* data.
   - Note that after your name your individual data after counting the occurrences, the last entry will be ⟨NA⟩. Change this by forcing it's name to be "NA" by using *names(your_data_name)[30]<-"NA"*.

```
c<-count(aswells.year)
data<-c[,2]
names(data)<-c[,1]
names(data)[30]<-"NA"
pareto.chart(data)
```


Pareto Chart for data
Frequency
0
50
100
200
300
Cumulative Percentage
0%
25%
50%
75%
100%
1999
1998
1995
1997
1990
1993
1996
1992
1994
2000
1991
1987
1970
1986
2001
1973
1988
1989
1985
NA
1978
1980
1982
1984
2002
1975
1976
1981
1983
2003

```
##
## Pareto chart analysis for data
##         Frequency   Cum.Freq.  Percentage Cum.Percent.
##   1999  43.0000000  43.0000000  13.1097561   13.1097561
##   1998  39.0000000  82.0000000  11.8902439   25.0000000
##   1995  35.0000000 117.0000000  10.6707317   35.6707317
##   1997  27.0000000 144.0000000   8.2317073   43.9024390
##   1990  24.0000000 168.0000000   7.3170732   51.2195122
##   1993  24.0000000 192.0000000   7.3170732   58.5365854
##   1996  23.0000000 215.0000000   7.0121951   65.5487805
##   1992  19.0000000 234.0000000   5.7926829   71.3414634
##   1994  15.0000000 249.0000000   4.5731707   75.9146341
##   2000  14.0000000 263.0000000   4.2682927   80.1829268
##   1991  10.0000000 273.0000000   3.0487805   83.2317073
##   1987   7.0000000 280.0000000   2.1341463   85.3658537
##   1970   5.0000000 285.0000000   1.5243902   86.8902439
##   1986   5.0000000 290.0000000   1.5243902   88.4146341
##   2001   5.0000000 295.0000000   1.5243902   89.9390244
##   1973   4.0000000 299.0000000   1.2195122   91.1585366
##   1988   4.0000000 303.0000000   1.2195122   92.3780488
##   1989   4.0000000 307.0000000   1.2195122   93.5975610
##   1985   3.0000000 310.0000000   0.9146341   94.5121951
##   NA     3.0000000 313.0000000   0.9146341   95.4268293
##   1978   2.0000000 315.0000000   0.6097561   96.0365854
##   1980   2.0000000 317.0000000   0.6097561   96.6463415
##   1982   2.0000000 319.0000000   0.6097561   97.2560976
##   1984   2.0000000 321.0000000   0.6097561   97.8658537
##   2002   2.0000000 323.0000000   0.6097561   98.4756098
##   1975   1.0000000 324.0000000   0.3048780   98.7804878
##   1976   1.0000000 325.0000000   0.3048780   99.0853659
##   1981   1.0000000 326.0000000   0.3048780   99.3902439
##   1983   1.0000000 327.0000000   0.3048780   99.6951220
##   2003   1.0000000 328.0000000   0.3048780  100.0000000
```

4. Using the *Arsenic* data, find
    - the mean,
    - the median,
    - the variance,
    - the standard deviation,
    - the $12^{th}, 25^{th}, 65^{th}, 94^{th}$ percentiles.
    - Construct a histogram.
    - Is the data normally distributed?
    - Construct a horizontal box and whiskers plot. Label Q_1, Q_3, the median, and the fences with their respective value.
    - Are there any outliers?

```
mean(aswells.arsenic)
```

```
## [1] 95.35976
```

```
median(aswells.arsenic)
```

```
## [1] 54.5
```

```
var(aswells.arsenic)
```

```
## [1] 12538.81
```

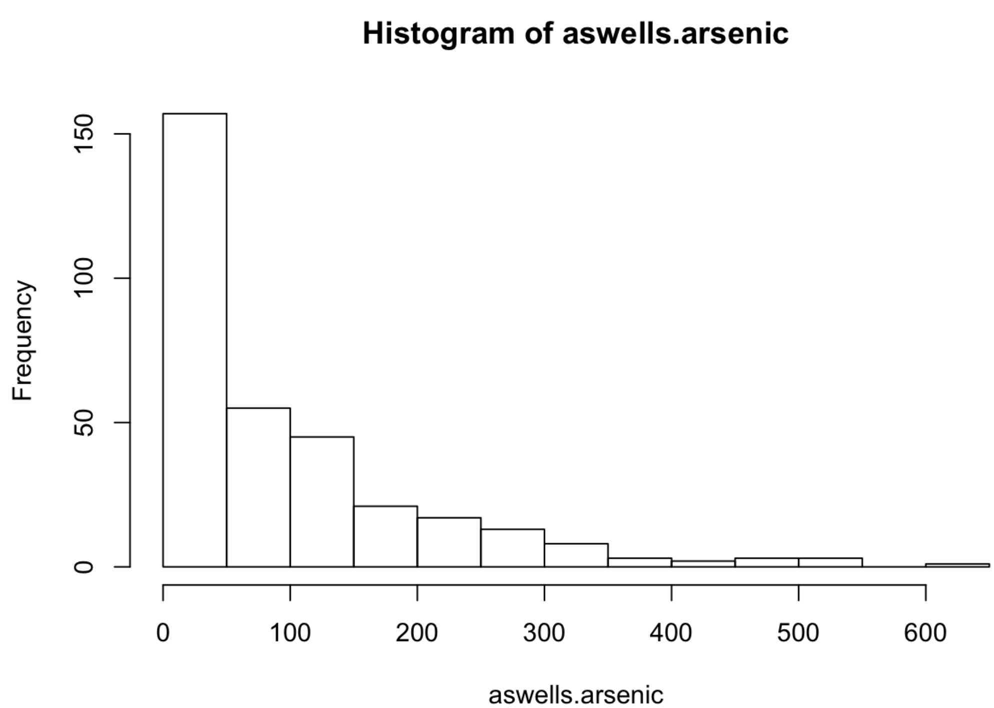
```
sd(aswells.arsenic)
```

```
## [1] 111.9768
```

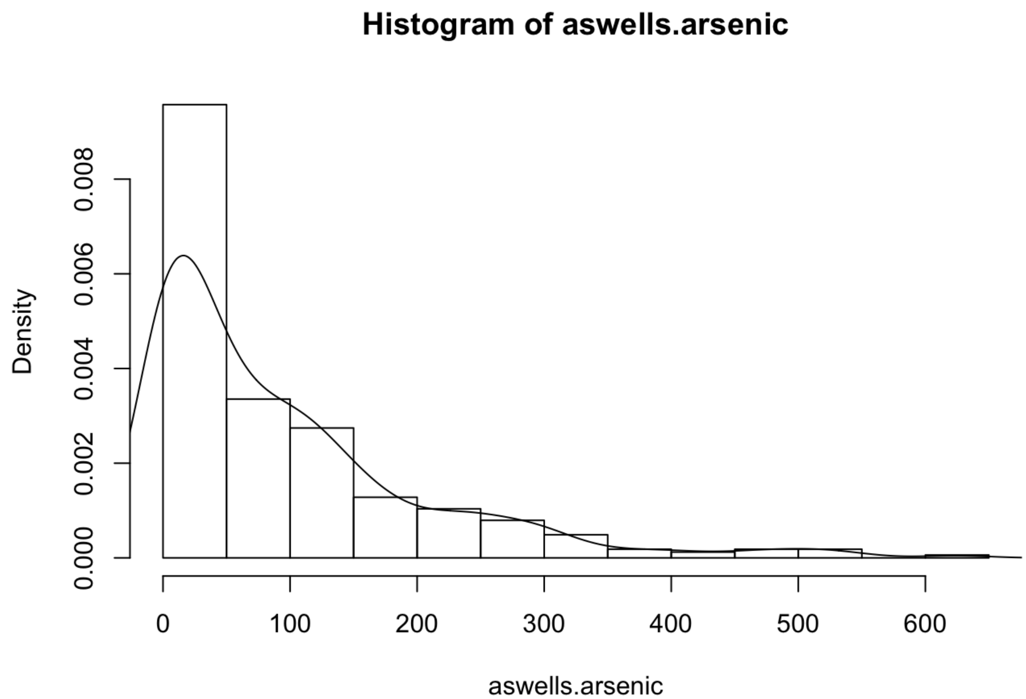
```
quantile(aswells.arsenic,c(.12,.25,.65,.94))
```

```
##    12%    25%    65%    94%
##   1.00   9.00 105.00 297.28
```

```
hist(aswells.arsenic)
```

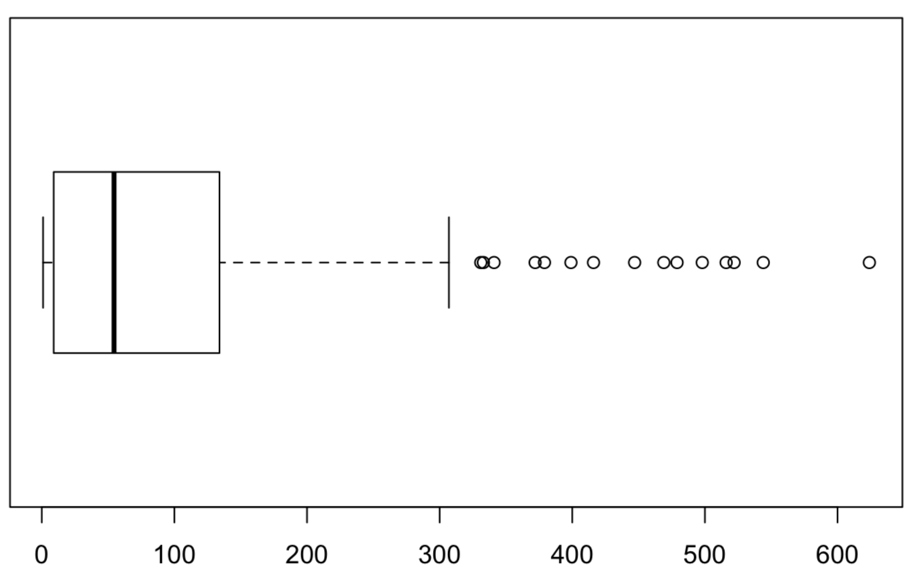



```
hist(aswells.arsenic,breaks=14,prob=T)
lines(density(aswells.arsenic))
```



The data is right-skewed.

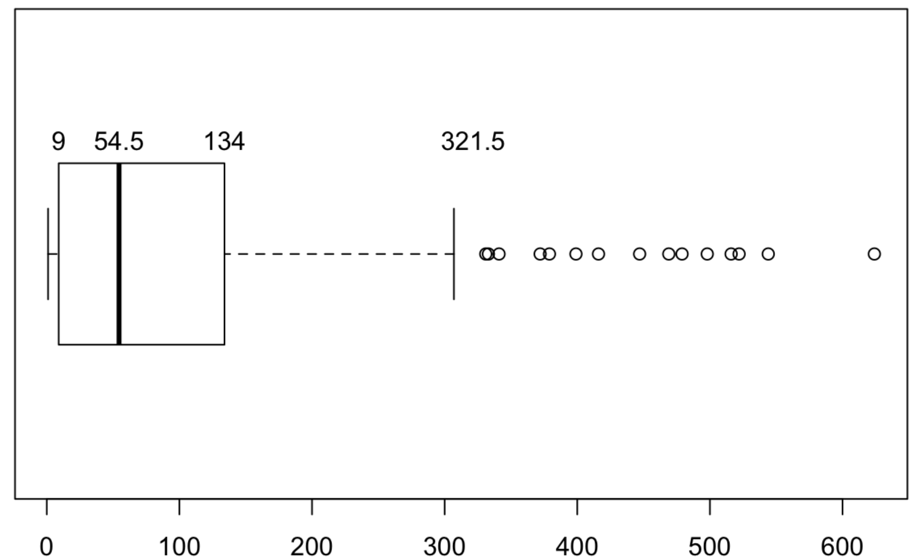
```
boxplot(aswells.arsenic,horizontal = T)
```




```
summary(fivenum(aswells.arsenic))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0     9.0    54.5   164.5   134.0   624.0
```

```
IQR<-quantile(aswells.arsenic)[4]-quantile(aswells.arsenic)[2]
lbs<-c(quantile(aswells.arsenic)[2]-1.5*IQR,quantile(aswells.arsenic)
[2],median(aswells.arsenic),quantile(aswells.arsenic)[4],quantile(aswell
s.arsenic)[4]+1.5*IQR)
boxplot(aswells.arsenic,horizontal = T)
text(lbs,labels=lbs,y = 1.25)
```



Yes, the outlier at 321.5.

Question 2

A random variable x has the following discrete probability distribution

x	-2	-1	0	1	2
$p(x)$	.10	.15	.40	.30	.05

1. Find μ.
2. Find σ.
3. Find $P(x \leq 0)$.
4. Find $P(-1 < x < 1)$.

```
x<--2:2
p.x<-c(.10,.15,.40,.30,.05)
rbind(x,p.x)
```

```
##     [,1]  [,2] [,3] [,4] [,5]
## x   -2.0 -1.00  0.0  1.0 2.00
## p.x  0.1  0.15  0.4  0.3 0.05
```

```
weighted.mean(x, p.x)
```

```
## [1] 0.05
```

```
weighted.sd(x, p.x)
```

```
## [1] 1.023474
```

```
pbinom(0, 5, .4)
```

```
## [1] 0.07776
```

```
pbinom(1,5,.30) - pbinom(-1,5,.15)
```

```
## [1] 0.52822
```

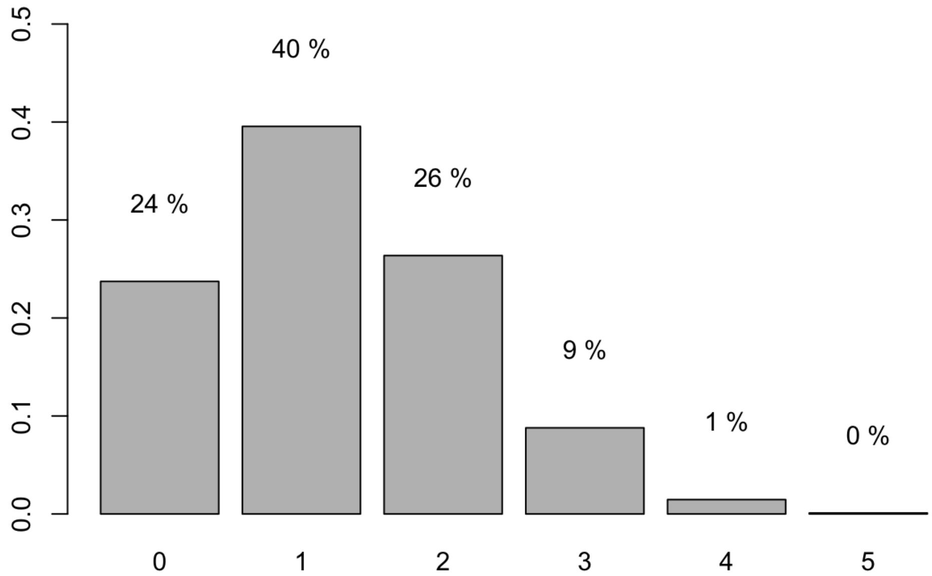
Question 3

A study of various brands of bottled water found that 25% of bottled water is just tap water packaged in a bottle. Consider a sample of five bottled-water brands and let x equal the number of these brands that use tap water.

1. What type of random variable is x?
2. Construct a histogram for the distribution of x.
3. Find $P(x = 2)$.
4. Find $P(x \leq 3)$.

X is a binomial random variable.

```
water<-c(0:5)
p.water<-dbinom(water, 5, .25)
text(barplot(p.water,names=water,ylim=c(0,.50)),labels = paste(round(p.wa
ter*100),"%"),y=p.water+.08)
```




```
dbinom(2, 5, .25)
```

```
## [1] 0.2636719
```

```
pbinom(3, 5, .25)
```

```
## [1] 0.984375
```

Question 4

A report on a spare line replacement unit (LRU) states that the number of LRU's that fail in any 10,000-hour period is assumed to follow a discrete distribution with a mean and standard deviation equal to 1.2.

1. What type of random variable is x?
2. Find the probability that there is at most two LRU failures during the next 10,000 hours of operation.

x is a Poisson random variable.

```
ppois(2, 1.2)
```

```
## [1] 0.8794871
```

Question 5

Using the data set **CRASH**, consider the *drivhead* data, which concerns driver-side head injury ratings for each car, as a continuous random variable.

1. What type of random variable is x?
2. Find the probability that the a randomly selected car will have a rating between 500 and 700 points.
3. What rating will only 10% of the crash-tested cars exceed?

x is a continuous random variable with normal distribution.

```
drivehead.mean<-mean(crash.drivehead)
drivehead.sd<-sd(crash.drivehead)
pnorm(700, drivehead.mean, drivehead.sd) - pnorm(500, drivehead.mean, dri
vehead.sd)
```

```
## [1] 0.4103739
```

```
qnorm(.10, drivehead.mean, drivehead.sd)
```

```
## [1] 366.1953
```

Question 6

A random sample of $n = 100$ observations is selected from a population with $\mu = 30$ and $\sigma = 16.$

1. Find $\mu_{\bar{x}}$
2. Find $\sigma_{\bar{x}}$
3. Find $P(\bar{x} \geq 28)$
4. Find $P(22.1 \leq \bar{x} \leq 26.8)$

$\mu_{\bar{x}} = 30$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{16}{\sqrt{100}} = 1.6$$

```
pnorm(28, 30, 16/sqrt(100), lower.tail = F)
```

```
## [1] 0.8943502
```

```
pnorm(26.8, 30, 16/sqrt(100)) - pnorm(22.1, 30, 16/sqrt(100))
```

```
## [1] 0.02274974
```

Question 7

Consider the data set **PHISHING** as a continous random variable. The data concerns how long employees notified management if they suspected an email phishing attack (to see if there was phishing occuring from an “inside source”.)

1. What type of random variable is x?
2. What is the probability of observing an interarrival time of at least 2 minutes?

x is a continuous random variable with exponential distribution.

```
phishing.mean<-mean(phishing.intime)
phishing.sd<-sd(phishing.intime)
avg<-(phishing.mean+phishing.sd)/2
rate<-1/avg
pexp(120, rate, lower.tail = F)
```

```
## [1] 0.2772069
```

Question 8

In a sample of 158 cartridges, 36 were found to be contaminated and 122 where “clear.” If you randomly select 5 cartridges (without replacement), what is the probability that all 5 will be “clean?”

```
dhyper(5, 122, 36, 5)
```

```
## [1] 0.2692909
```

Question 9

Researchers estimate that the trace amount of uranium x in reservoirs follows a uniform distribution ranging between 1 and 3 parts per million. Find the probability that a randomly selected reservoir will have an amout of uranium between 2 and 2.5 parts per million.

```
c<-1
d<-3
uranium.pd<-1/(d-c)
uranium.mean<-(c+d)/2
uranium.sd<-(d-c)/sqrt(12)
uranium.solution<-(2.5-2)/(d-c)
uranium.solution
```

```
## [1] 0.25
```