a1q3

```r
data<-read.csv("/Users/rudranibhadra/Downloads/class_data.csv")
#data
#View(data)
str(data)
```

```
## 'data.frame':    42 obs. of  6 variables:
##  $ random_digit : int  7 7 9 3 2 7 3 3 7 9 ...
##  $ student_digit: int  9 0 8 0 6 6 2 3 9 8 ...
##  $ green_card1  : Factor w/ 4 levels "a","b","c","d": 1 1 2 3 1 4 2 1 1 1 ...
##  $ green_card2  : Factor w/ 4 levels "a","b","c","d": 4 3 3 4 3 3 1 3 3 3 ...
##  $ red_card1    : Factor w/ 3 levels "a","b","c": 1 1 1 1 1 1 1 1 1 1 ...
##  $ red_card2    : Factor w/ 4 levels "a","b","c","d": 3 3 3 2 4 2 3 3 4 3 ...
```

```r
names(data)
```

```
## [1] "random_digit"  "student_digit" "green_card1"   "green_card2"
## [5] "red_card1"     "red_card2"
```

1

a. $E(D) = (1/10) * (0 + 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9) = 4.5$

b. $E(D^2) = \frac{1}{10} * (0^2 + 1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2 + 7^2 + 8^2 + 9^2) = 28.5$
   $Var(D) = E(D^2) - E(D)^2 = 28.5 - 20.25 = 8.25$
   $SD(D) = \sqrt{(}Var(D)) = 2.872$

c. median of D is 4.5

d.

e. Binomial distribution

ii. $Pr(X = x) = \binom{n}{x} \left(\frac{1}{10}\right)^x * \left(\frac{9}{10}\right)^{n-x}$

iii. $E(X) = n * \frac{1}{10}$

e.

f.

```r
mean(data$random_digit)
```

```
## [1] 5.5
```

```r
mean(data$student_digit)
```

```
## [1] 4.595238
```

ii.

```r
sd(data$random_digit)
```

```
## [1] 2.778401
```

```r
sd(data$student_digit)
```

```
## [1] 3.052864
```

iii.

```r
median(data$random_digit)
```

```
## [1] 7
```

```r
median(data$student_digit)
```

```
## [1] 5
```

iv. Random digit: The mean is greater than the corresponding theoritical value for D. The standard deviation is almost equal than the corresponding theoritical value for D. The median is greater than the corresponding theoritical value for D.

Student digit: The mean is smaller to the corresponding theoritical value for D. The standard deviation is greater to the corresponding theoritical value for D. The median is greater to the corresponding theoritical value for D.

v. The mean and standard deviation of random digit is close to the corresponding theoritical values for D. The median of student digit is close to the corresponding theoritical value for D.

w. For x = 0 $Pr(X = 0) = \binom{42}{0} * (1/10)^0 * (9/10)^{42} = 0.01197$
   For x = 5 $Pr(X = 5) = \binom{42}{5} * (1/10)^5 * (9/10)^{37} = 0.1724$
   For x = 10 $Pr(X = 10) = \binom{42}{10} * (1/10)^{10} * (9/10)^{32} = 0.0050$

2

a.

b.

```r
stem(data$student_digit)
```

```
##
##   The decimal point is at the |
##
##   0 | 000000
##   1 | 0000
##   2 | 000
##   3 | 0
```

```
##   4 | 000000
##   5 | 000
##   6 | 0000000
##   7 | 0
##   8 | 0000000
##   9 | 0000
```

ii.

```r
stem(data$random_digit)
```

```
##
##   The decimal point is at the |
##
##   0 | 000
##   1 | 0
##   2 | 000
##   3 | 000000
##   4 | 00
##   5 | 00
##   6 | 000
##   7 | 000000000000
##   8 | 0000
##   9 | 000000
```

iii.

```r
R<-c(0:9)
S<-sample(R,length(data$random_digit),replace = TRUE)
stem(S)
```

```
##
##   The decimal point is at the |
##
##   0 | 0000000
##   1 | 0000
##   2 | 000000000
##   3 | 0
##   4 | 000
##   5 | 000000000
##   6 | 0000
##   7 | 00
##   8 |
##   9 | 000
```

student digit looks like it might have come from a uniform on the digits. In random digits, 7 seems have to been chosen the maximum number of times.

b.

c.

```r
my_digits<-c(0,1,3,4,7,1,4,9,7,4)
count_digits<-function(d){
  a<-c()
  n<-length(d)
  for(i in 1:10){
    co<-0
    for(j in 1:n){
      if(d[j]==i-1){
        co<-co+1}
    }
    a[i]<-co
  }
  a
}
count_digits(my_digits)
```

```
##  [1] 1 2 0 1 3 0 0 2 0 1
```

```r
#my_digits<-c(0,1,3,4,7,1,4,9,7,4)
#count_digits(my_digits)
```

ii.

```r
count_digits(data$student_digit)
```

```
##  [1] 6 4 3 1 6 3 7 1 7 4
```

```r
count_digits(data$random_digit)
```

```
##  [1]  3  1  3  6  2  2  3 12  4  6
```

iii.

```r
Pearson_chi_sq<-function(observed,expected){
 # observed<-count_digits(data$student_digit)
  expected<-sum(observed)/length(observed)
  e<-expected
  v<-c()
  s<-c()
  n<-length(observed)
  s<-0
  expected
  for(i in 1:n){
    v[i]<-e
  }
#  v
 # observed
  #v

  for(i in 1:n){
```

```
   # e<-
    s<-s+((observed[i]-e)^2)/e
  #print(s[i])
  }
  #sum(s)
 return(s)
 # val=((observed-v)^2)/v
#sum(val)
}

Pearson_chi_sq(count_digits(data$student_digit))
```

```
## [1] 10.85714
```

```
chisq.test(count_digits(data$student_digit))$statistic
```

```
## Warning in chisq.test(count_digits(data$student_digit)): Chi-squared
## approximation may be incorrect

## X-squared
##  10.85714
```

iv.

```
Pearson_chi_sq(count_digits(data$student_digit))
```

```
## [1] 10.85714
```

```
chisq.test(count_digits(data$student_digit))$statistic
```

```
## Warning in chisq.test(count_digits(data$student_digit)): Chi-squared
## approximation may be incorrect

## X-squared
##  10.85714
```

```
Pearson_chi_sq(count_digits(data$random_digit))
```

```
## [1] 21.80952
```

```
chisq.test(count_digits(data$random_digit))$statistic
```

```
## Warning in chisq.test(count_digits(data$random_digit)): Chi-squared
## approximation may be incorrect

## X-squared
##  21.80952
```

Both corresponding values of pchisq and chisq.test match.

v.

```r
pchisq(Pearson_chi_sq(count_digits(data$student_digit)), 9, lower=FALSE)
```

```
## [1] 0.2856284
```

```r
chisq.test(count_digits(data$student_digit))$p.value
```

```
## Warning in chisq.test(count_digits(data$student_digit)): Chi-squared
## approximation may be incorrect
```

```
## [1] 0.2856284
```

```r
pchisq(Pearson_chi_sq(count_digits(data$random_digit)), 9, lower=FALSE)
```

```
## [1] 0.009502653
```

```r
chisq.test(count_digits(data$random_digit))$p.value
```

```
## Warning in chisq.test(count_digits(data$random_digit)): Chi-squared
## approximation may be incorrect
```

```
## [1] 0.009502653
```

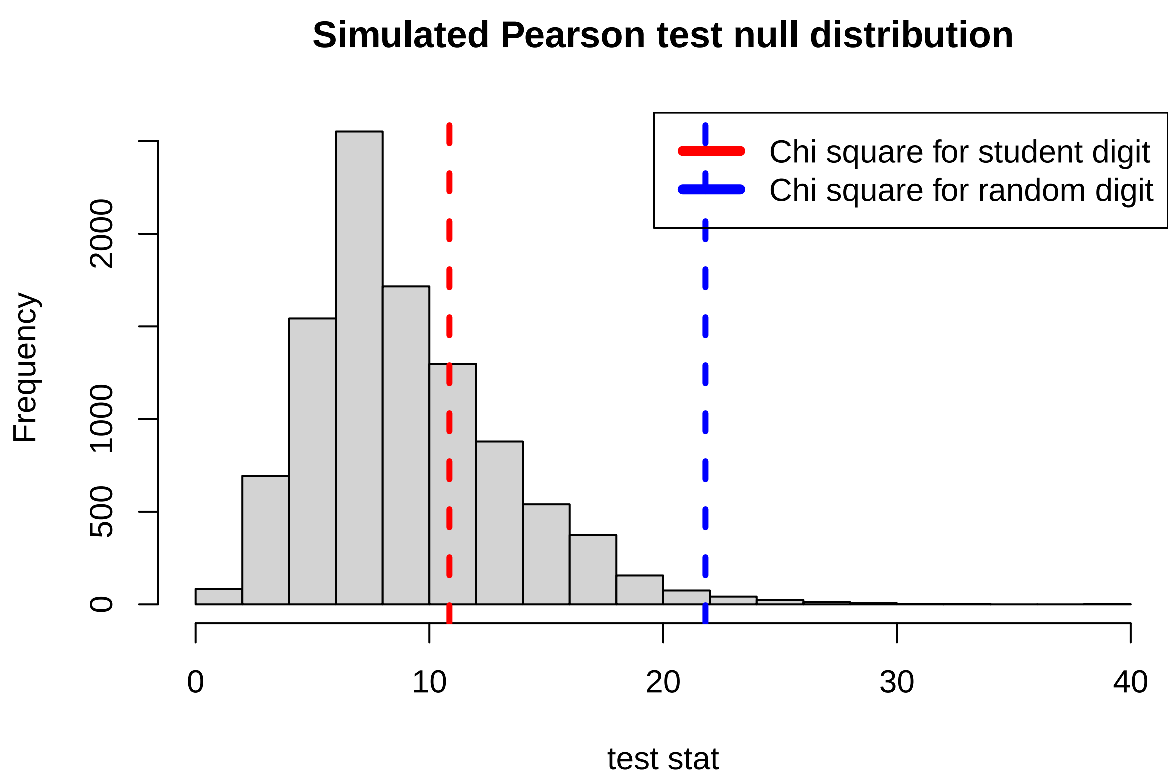
Both corresponding values of pchisq and chisq.test match.

vi.

```r
get_chisqs<-function(n,B=10){

  #p<-Pearson_chi_sq(count_digits(data$student_digit))
  r<-c(0:9)
  #print(r)
  #print(sample(r,10))
  v<-c()
  co<-c()
  #print(r)
  for(i in 1:B){

   s<- sample(r,n,replace=TRUE)
   #print(s)
   #c<-count_digits(s)
   b<-count_digits(s)
   #print(b)
   co[[i]]<-b
  #x<- Pearson_chi_sq(b)

  }
  #print(a) #sapply(s,Pearson_chi_sq(count_digits))
  a<- sapply(co,Pearson_chi_sq)
   #print(a)
}
n<-nrow(data)
results<-get_chisqs(n=n,B=100)
```

vii.

```r
n<-nrow(data)
B<-10000
set.seed(314159)
chisq_stats<-get_chisqs(n=n,B=B)
hist(chisq_stats,col="lightgrey",main="Simulated Pearson test null distribution", xlab="test stat")
abline(v=Pearson_chi_sq(count_digits(data$student_digit)),lwd=3,lty=2,col="red")
abline(v=Pearson_chi_sq(count_digits(data$random_digit)),lwd=3,lty=2,col="blue")
legend("topright", c("Chi square for student digit", "Chi square for random digit"),
col=c("red", "blue"), lwd=5)
```



Random digit seems less likely to have been generated as a random sample.

Reason: Its chi square value is at the edge of the histogram

viii.

```r
mean(chisq_stats>=Pearson_chi_sq(count_digits(data$student_digit)))
```

```
## [1] 0.2811
```

```r
mean(chisq_stats>=Pearson_chi_sq(count_digits(data$random_digit)))
```

```
## [1] 0.0095
```

ix. The hypothesis is true for student digit as its p value is large (>0.05). The hypothesis is false for random digit as its p value is very small (<0.05).