lec14_1

Jiansong Xu

March 5, 2019

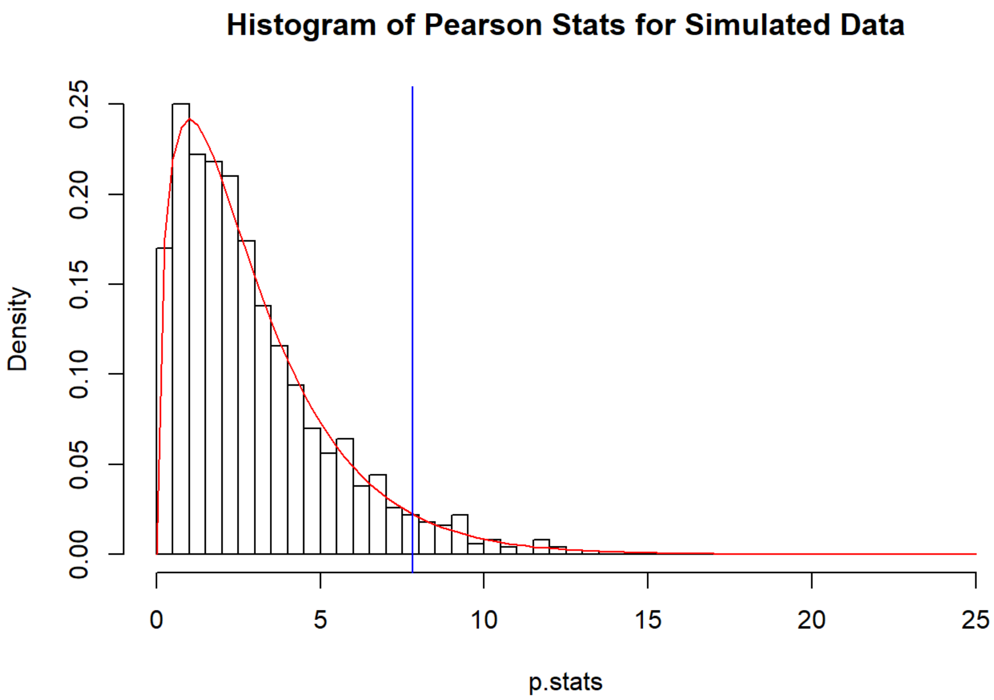
1

a.

```
pearson <- function(counts){chisq.test(x=counts, p=c(9,3,3,1)/16, correct=FALSE)$statistic}

samp.size <- 1611
set.seed(123)
rcount <- rmultinom(n=1000, size=samp.size, prob=c(9,3,3,1)/16)
p.stats <- apply(X=rcount, MARGIN=2, FUN=pearson)
summary(p.stats)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.04076  1.17739  2.31071  2.92777  4.00297 13.02262
```

```
hist(x=p.stats, breaks=c(0:50)/2, freq=FALSE, main="Histogram of Pearson Stats for Simulated Data")
curve(expr=dchisq(x=x, df=3), add=TRUE, col="red")

# Add line at the 0.05 critical value of chi-squared(3)
abline(v=qchisq(0.95, df=3), col="blue")
```



Chi-square seems like a good fit.

b.

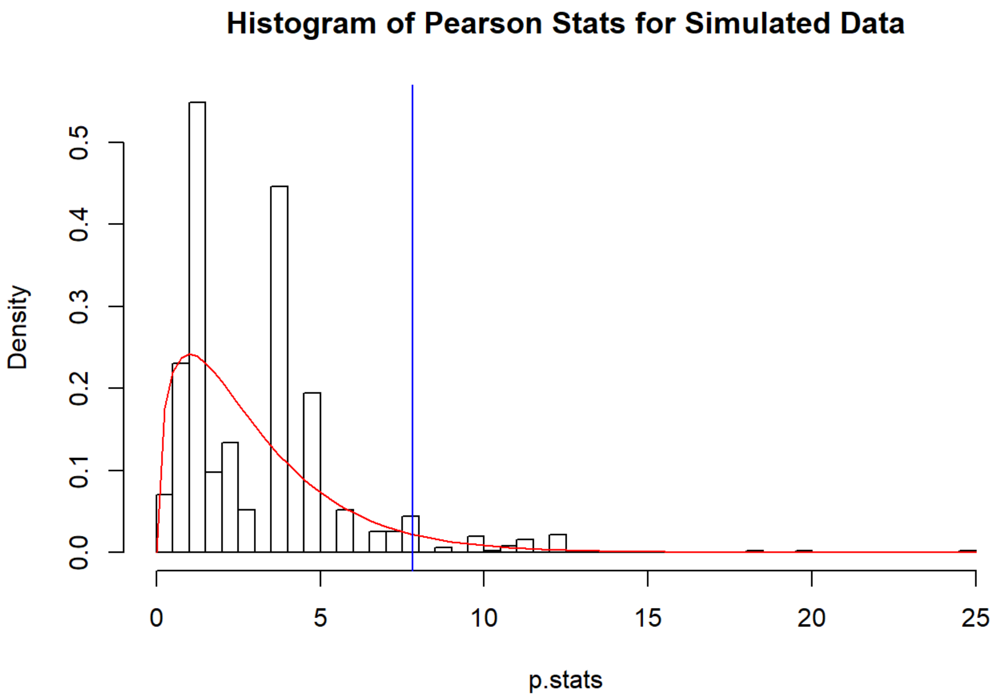
i

```
n <- 10
probs <- c(9,3,3,1)/16
n*probs -> counts10
list10 <- as.list(counts10)
names(list10) <- c("Tall cut-Leaf", "Tall potato-leaf", "Dwarf cut-leaf", "Dwarf potato-leaf")
data.frame(list10)
```

```
##   Tall.cut.Leaf Tall.potato.leaf Dwarf.cut.leaf Dwarf.potato.leaf
## 1         5.625            1.875          1.875             0.625
```

ii

```
samp.size <- 10
set.seed(123)
rcount <- rmultinom(n=1000, size=samp.size, prob=c(9,3,3,1)/16)
p.stats <- apply(X=rcount, MARGIN=2, FUN=pearson)
hist(x=p.stats, breaks=c(0:50)/2, freq=FALSE, main="Histogram of Pearson Stats for Simulated Data")
invisible(curve(expr=dchisq(x=x, df=3), add=TRUE, col="red"))
# Add line at the 0.05 critical value of chi-squared(3)
abline(v=qchisq(0.95, df=3), col="blue")
```



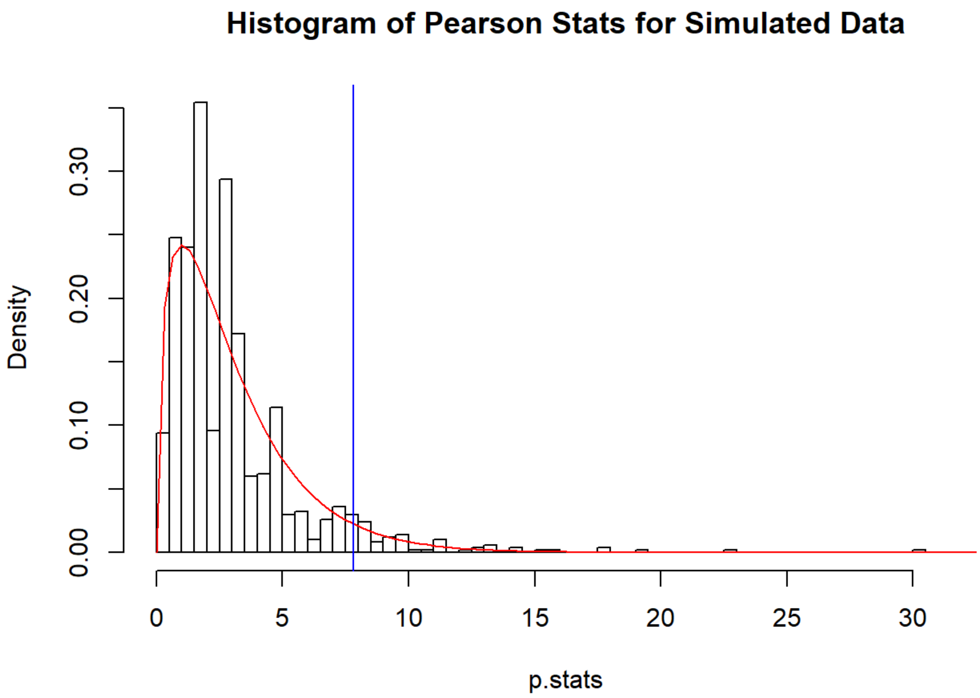
The curve dose not fit histogram very well, although there is a hardly observable Chi-square shape. The right-tail of histogram has been stretched very long, as we can see even at the right-end of Chisq curve there are still statistics observed, because a larger portion of statistics are far away beyond the critical value comparing to the large sample size we had in the last part.

c.

```
##Make a function which calculates expected counts and draws graphs for convenience##
probs <- c(9,3,3,1)/16

draw <- function(ss) {

    n <- ss
    n*probs -> counts
    list <- as.list(counts)
    names(list) <- c("Tall cut-Leaf", "Tall potato-leaf", "Dwarf cut-leaf", "Dwarf potato-leaf")
    list
    rcount <- rmultinom(n=1000, size=ss, prob=probs)
    p.stats <- apply(X=rcount, MARGIN=2, FUN=pearson)
    hist(x=p.stats, breaks=c(0:65)/2, freq=FALSE, main=paste0("Histogram of Pearson Stats for Simulated Data
"))
    curve(expr=dchisq(x=x, df=3), add=TRUE, col="red")
    abline(v=qchisq(0.95, df=3), col="blue")
    return(data.frame(list))

  }
```

For sample size 20

```
set.seed(123)
draw(20)
```

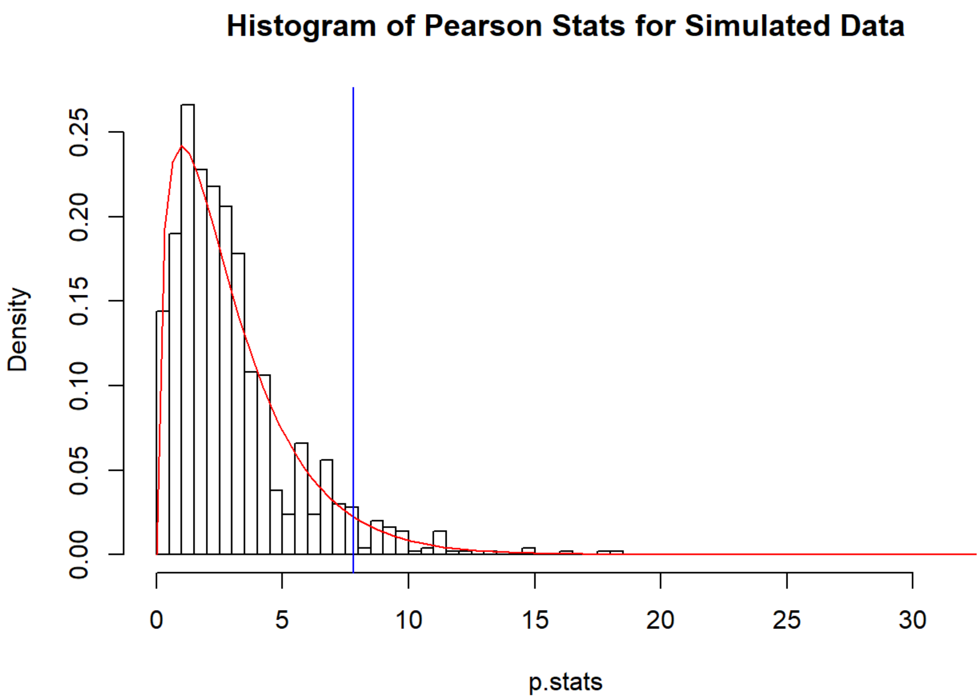



```
##   Tall.cut.Leaf Tall.potato.leaf Dwarf.cut.leaf Dwarf.potato.leaf
## 1         11.25              3.75           3.75              1.25
```

Chi-square curve fits the histogram not so well, but better than sample size 10. There is still a considerable amount of statistics that fall out of critical value.

For sample size 40

```
set.seed(123)
draw(40)
```

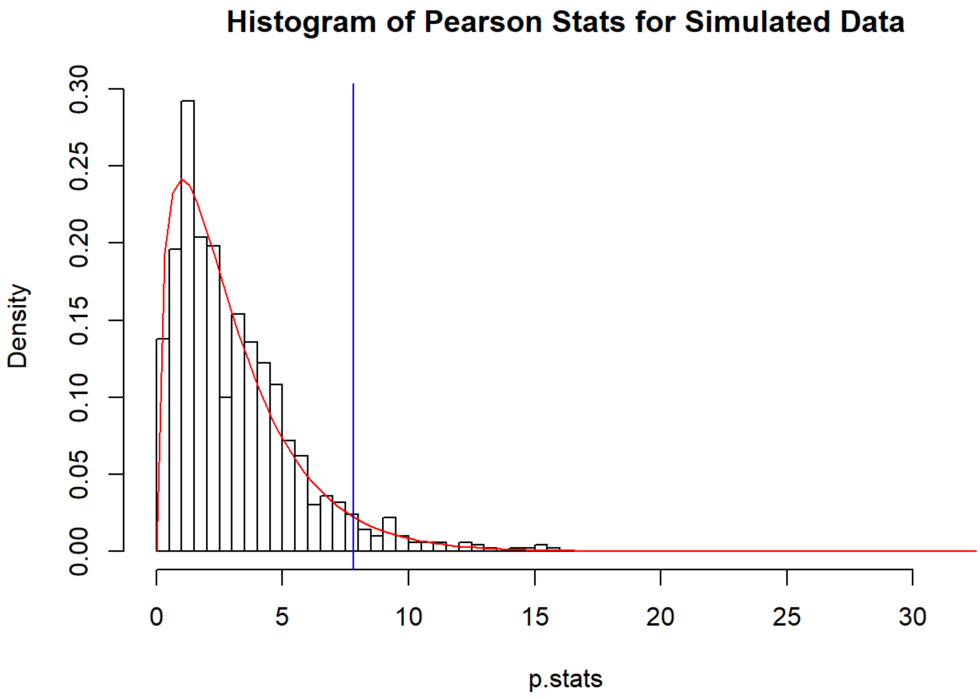



```
##   Tall.cut.Leaf Tall.potato.leaf Dwarf.cut.leaf Dwarf.potato.leaf
## 1          22.5               7.5            7.5               2.5
```

The curve fits the histogram obviously better than sample size 20. We can see less statistics fall beyond the critical value.

For sample size 80

```r
set.seed(123)
draw(80)
```




```
##   Tall.cut.Leaf Tall.potato.leaf Dwarf.cut.leaf Dwarf.potato.leaf
## 1            45               15             15                 5
```

The curve fits the histogram mostly well. It is hard to see but fewer statistics are greater than critical value comparing to the last one.