Chi Square Motivation, Polio

C. Durso

Test of Independence

Apply the χ^2 method to the polio data. One format in which chisq.test takes data is as a matrix. The default in this case is to do a test of independence of the rows and columns.

This is a χ^2 test of the null hypothesis that contracting paralytic polio is independent of vaccination status in the Randomized Control experiment. Note that since the "Population" column includes paralytic polio cases and non-cases, the number of paralytic polio cases must be subracted from the full population in each group.

```
data("PolioTrials")
dat<-PolioTrials
m<-as.matrix(dat[1:2,3:4],nrow=2)
m[,1]<-m[,1]-m[,2]
chisq.test(m)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  m
## X-squared = 44.153, df = 1, p-value = 3.038e-11
```

```
chisq.test(m, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  m
## X-squared = 45.252, df = 1, p-value = 1.733e-11
```

```
dimnames(m)[[1]]<-c("treatment", "control")
dimnames(m)[[2]]<-c("no polio","polio")
m
```

```
##           no polio polio
## treatment   200712    33
## control     201114   115
```

The very small p-values give us strong evidence against the null hypothesis.

The matrix of expected values under the null hypothesis is available from the fitted model.

```
model.chisq<-chisq.test(m)
model.chisq$expected
```

```
##           no polio   polio
## treatment 200671.1 73.9109
## control   201154.9 74.0891
```

The method for the χ^2 test of independence begins with the calculation of the values expected in each cell if the column proportions in each row are the same, and the row sums and column sums are preserved.

Compare this to the matrix of expected values from the fitted model.

```
p<-sum(m[,2])/sum(m)
(P<-matrix(c(1-p,p),ncol=2))
```

```
##           [,1]        [,2]
## [1,] 0.9996318 0.000368183
```

```
(T<-matrix(rowSums(m),nrow=2))
```

```
##        [,1]
## [1,] 200745
## [2,] 201229
```

```
(E<-T%*%P)
```

```
##          [,1]    [,2]
## [1,] 200671.1 73.9109
## [2,] 201154.9 74.0891
```

```
# check
rowSums(E)
```

```
## [1] 200745 201229
```

```
rowSums(m)
```

```
## treatment   control
##    200745    201229
```

```
colSums(E)
```

```
## [1] 401826    148
```

```
colSums(m)
```

```
## no polio    polio
##   401826      148
```

The test statistic is the sum of the values (Observed-Expected)^2/Expected in each cell. This may be adjusted for integer values by subtracting 1/2 from any cell with absolute value greater than 1/2 before squaring.

Note that these, respectively, equal the X-squared values from the uncorrected and the corrected tests above.

```
(chistat<-sum((m-E)^2/E))
```

```
## [1] 45.25191
```

```
(chistat_adj<-sum((abs(m-E)-.5)^2/E))
```

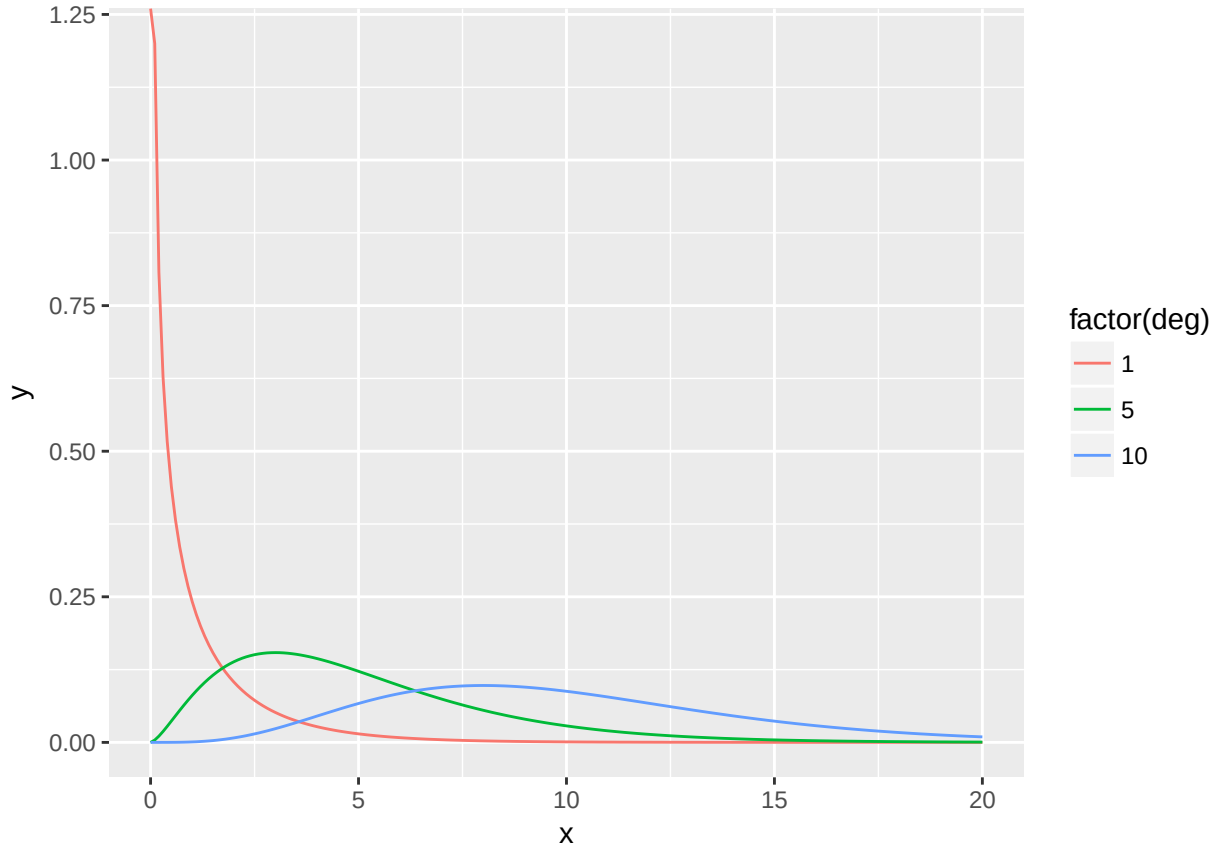
```
## [1] 44.15256
```

The degrees of freedom for the χ^2 is the number of cells that can be assigned freely while preserving the row and column sums. Here, df=1.

*Visualize the χ^2 curves for 1,5, and 10 degrees of freedom.$

```
x=seq(0,20,by=.1)
dat<-bind_rows(data.frame(x=x,y=dchisq(x,1),deg=1),
               data.frame(x=x,y=dchisq(x,5),deg=5),
               data.frame(x=x,y=dchisq(x,10),deg=10)
              )

ggplot(group_by(dat,deg),aes(x=x,y=y,color=factor(deg)))+geom_line()
```



A one-tailed test is typical: we're usually not concerned with an improbably good fit of observed to expected.

```
pchisq(chistat,df=1,lower.tail=FALSE)
```

```
## [1] 1.732517e-11
```

```
chisq.test(m, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  m
## X-squared = 45.252, df = 1, p-value = 1.733e-11
```

```
pchisq(chistat_adj,df=1,lower.tail=FALSE)
```

```
## [1] 3.037545e-11
```

```
chisq.test(m)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  m
## X-squared = 44.153, df = 1, p-value = 3.038e-11
```

Practice

In the work below, you will conduct a χ^2 test of the null hypothesis that contracting paralytic polio is independent of membership in the "Grade2NotInoculated" or "Controls" in the Observed Control experiment. First, please conduct the χ^2 test using the "chisq.test" function on an appropriately constructed matrix. Please use the corrected test. Hint: the number of paralytic polio cases must be subracted from the full population.

```
dat<-PolioTrials
m<-as.matrix(dat[6:7,3:4],nrow=2)
m[,1]<-m[,1]-m[,2]
dimnames(m)[[1]]<-c("control", "not.innoculated")
dimnames(m)[[2]]<-c("no polio","polio")
(model.chisq<-chisq.test(m))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  m
## X-squared = 2.5232, df = 1, p-value = 0.1122
```

Now please calculate the expected values under the null hypothesis in each portion of the contingency table using matrix algebra. Compare your results to the expected values from the fitted model.

```
p<-sum(m[,2])/sum(m)
P<-matrix(c(1-p,p),ncol=2)
T<-matrix(rowSums(m),nrow=2)
(E<-T%*%P)
```

```
##          [,1]      [,2]
## [1,] 724854.3 318.68113
## [2,] 123550.7  54.31887
```

```
model.chisq$expected
```

```
##                 no polio     polio
## control         724854.3 318.68113
## not.innoculated 123550.7  54.31887
```