Comparing Counts (Chapter 22)

Patrick Frenett, Vickie Ip, Sarah McDonald, and Nicholas Horton (nhorton@amherst.edu)

June 29, 2018

Introduction and background

This document is intended to help describe how to undertake analyses introduced as examples in the Fourth Edition of *Intro Stats* (2013) by De Veaux, Velleman, and Bock. More information about the book can be found at http://wps.aw.com/aw_deveaux_stats_series. This file as well as the associated R Markdown reproducible analysis source file used to create it can be found at https://nhorton.people.amherst.edu/is4.

This work leverages initiatives undertaken by Project MOSAIC (http://www.mosaic-web.org), an NSF-funded effort to improve the teaching of statistics, calculus, science and computing in the undergraduate curriculum. In particular, we utilize the `mosaic` package, which was written to simplify the use of R for introductory statistics courses. A short summary of the R needed to teach introductory statistics can be found in the mosaic package vignettes (http://cran.r-project.org/web/packages/mosaic). A paper describing the mosaic approach was published in the *R Journal*: https://journal.r-project.org/archive/2017/RJ-2017-024.

Note that some of the figures in this document may differ slightly from those in the IS4 book due to small differences in datasets. However in all cases the analysis and techniques in R are accurate.

Chapter 22: Comparing Counts

Section 22.1: Goodness-of-fit tests

Here we verify the calculations of expected counts for ballplayers by month (page 611).

```
ballplayer <- c(137, 121, 116, 121, 126, 114,
                102, 165, 134, 115, 105, 122)
national <- c(0.08, 0.07, 0.08, 0.08, 0.08, 0.08,
              0.09, 0.09, 0.09, 0.09, 0.08, 0.09)
n <- sum(~ ballplayer)
n
```

```
## [1] 1478
```

```
sum(~ national)
```

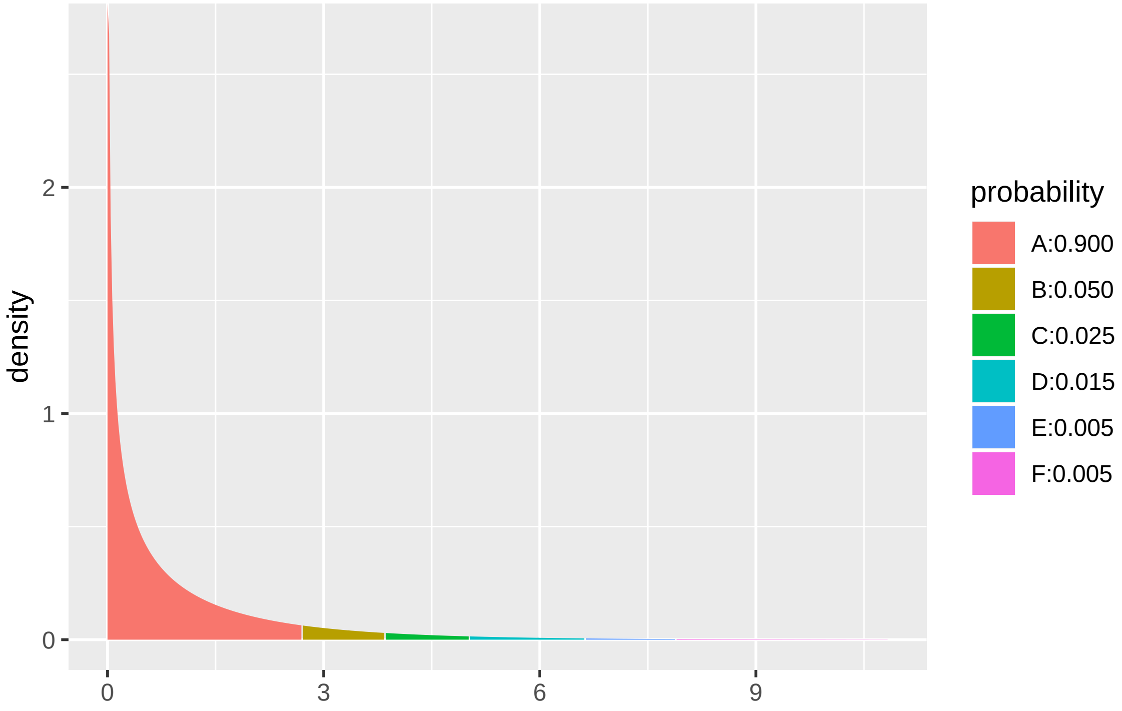
```
## [1] 1
```

```
expect <- n * national
cbind(ballplayer, expect)
```

```
##       ballplayer expect
##  [1,]        137 118.24
##  [2,]        121 103.46
##  [3,]        116 118.24
##  [4,]        121 118.24
##  [5,]        126 118.24
##  [6,]        114 118.24
##  [7,]        102 133.02
##  [8,]        165 133.02
##  [9,]        134 133.02
## [10,]        115 133.02
## [11,]        105 118.24
## [12,]        122 133.02
```

The chi-square quantile values in the table on the bottom of page 658 can be verified using the `xqt()` function.

```
xqchisq(c(.90, .95, .975, .99, .995), df = 1)
```



```
## [1] 2.705543 3.841459 5.023886 6.634897 7.879439
```

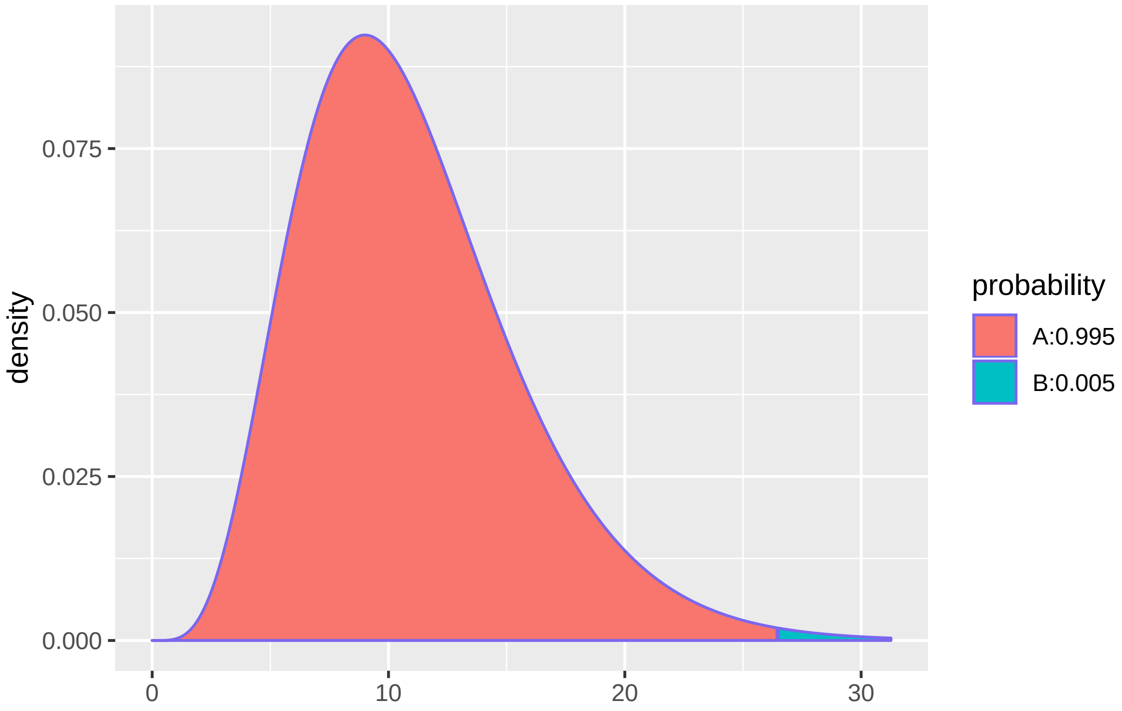
These results match the first row: other values can be calculated by changing the `df` argument.

The goodness of fit test on page 614 can be verified by calculating the chi-square statistic.

```
chisq <- sum((ballplayer - expect)^2/expect)
chisq
```

```
## [1] 26.48442
```

```
1 - xpchisq(chisq, df = 11, col = "slateblue2")
```



```
## [1] 0.005494028
```

Section 22.2: Chi-square test of homogeneity

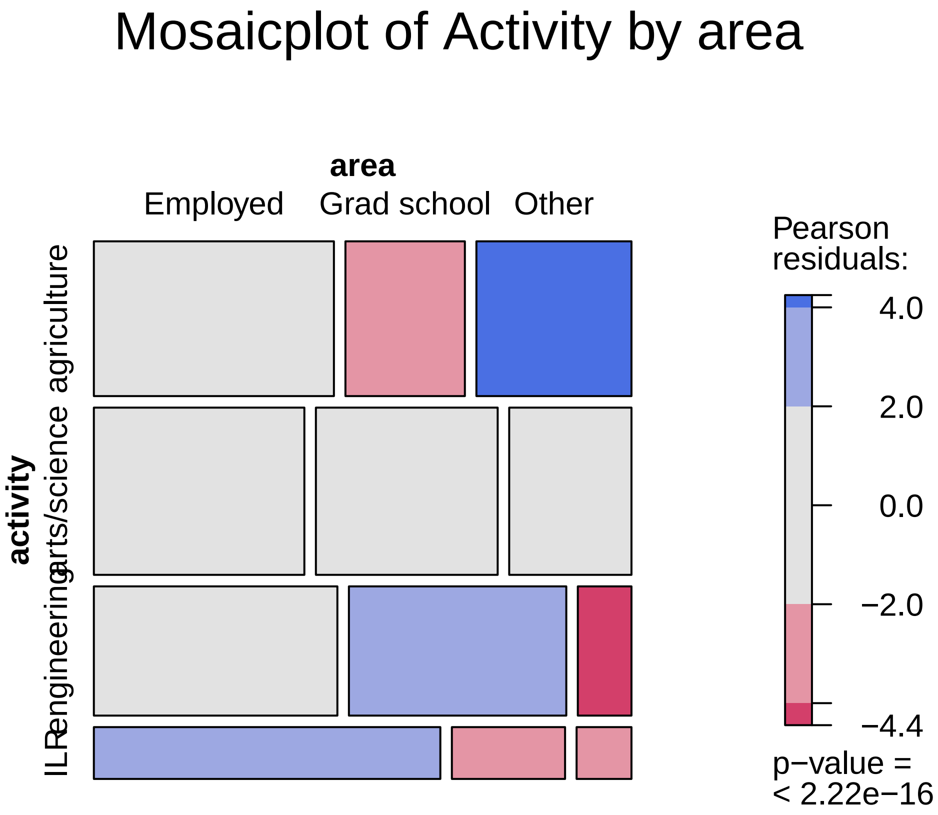
Data from one university regarding the association between postgraduation activity and area of study is displayed in Table 22.1 (page 618).

```
schooldata <- rbind(
  do(209) * data.frame(activity = "agriculture",  area = "Employed"),
  do(198) * data.frame(activity = "arts/science", area = "Employed"),
  do(177) * data.frame(activity = "engineering",  area = "Employed"),
  do(101) * data.frame(activity = "ILR",          area = "Employed"),
  do(104) * data.frame(activity = "agriculture",  area = "Grad school"),
  do(171) * data.frame(activity = "arts/science", area = "Grad school"),
  do(158) * data.frame(activity = "engineering",  area = "Grad school"),
  do(33)  * data.frame(activity = "ILR",          area = "Grad school"),
  do(135) * data.frame(activity = "agriculture",  area = "Other"),
  do(115) * data.frame(activity = "arts/science", area = "Other"),
  do(39)  * data.frame(activity = "engineering",  area = "Other"),
  do(16)  * data.frame(activity = "ILR",          area = "Other")
)

tally(~ activity + area, margins = TRUE, data = schooldata)
```

```
##                 area
## activity         Employed Grad school Other Total
##   agriculture         209         104   135   448
##   arts/science        198         171   115   484
##   engineering         177         158    39   374
##   ILR                  101           33     16    150
##   Total                685          466    305   1456
```

```
vcd::mosaic(tally(~ activity + area, data = schooldata), main = "Mosaicplot of Activity by area",
  shade = TRUE, ylab = "Area", xlab = "Activity")
```



```
xchisq.test(tally(~ activity + area, data = schooldata))
```

```
##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 93.657, df = 6, p-value < 2.2e-16
##
##   209       104       135
## (210.77) (143.38) ( 93.85)
## [ 0.0149] [10.8181] [18.0470]
## <-0.122> <-3.289> < 4.248>
##
##   198       171       115
## (227.71) (154.91) (101.39)
## [ 3.8754] [ 1.6720] [ 1.8277]
## <-1.969> < 1.293> < 1.352>
##
##   177       158        39
## (175.95) (119.70) ( 78.34)
## [ 0.0062] [12.2543] [19.7590]
## < 0.079> < 3.501> <-4.445>
##
##   101       33       16
## ( 70.57) ( 48.01) ( 31.42)
## [13.1215] [ 4.6918] [ 7.5689]
## < 3.622> <-2.166> <-2.751>
##
## key:
##  observed
##  (expected)
##  [contribution to X-squared]
##  <Pearson residual>
```

Section 22.3: Examining the residuals

Note that the `xchisq.test()` function displays the standardized residuals as the last item in each cell of the table (and these match the results in Table 22.4 (page 623).