431 Class 20

thomaselove.github.io/431

2020-11-05

Today's Agenda

- Comparing Proportions using Independent Samples
    - Working with 2x2 Tables
    - Working with more general two-way tables

Today's Setup and Data

```
knitr::opts_chunk$set(comment = NA)
options(dplyr.summarise.inform = FALSE)

library(Epi) # new today
library(janitor)
library(knitr)
library(magrittr)
library(mosaic) # not usually something I load
library(broom)
library(tidyverse)

theme_set(theme_bw())

dm431 <- readRDS("data/dm431_2020.Rds")
source("data/Love-boost.R")
```

Comparing Two Proportions (A 2×2 table)

Using `twobytwo` from the `Love-boost.R` script

–	A1c $<$ 8	A1c $>=$ 8	*Total*
Never	22	12	*34*
Current	20	13	*33*
Total	*42*	*25*	*67*

Code we need is:

```
twobytwo(22, 12, 20, 13,  # note order of counts
      "Never", "Current", # names of the rows
      "A1c<8", "A1c>=8",  # names of the columns
      conf.level = 0.90)  # default is 95% confidence
```

Complete Output shown on the next slide

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : A1c<8
Comparing : Never vs. Current

        A1c<8 A1c>=8    P(A1c<8) 90% conf. interval
Never      22     12      0.6471    0.5040   0.7679
Current    20     13      0.6061    0.4613   0.7343

                                    90% conf. interval
             Relative Risk: 1.0676    0.7823   1.4571
         Sample Odds Ratio: 1.1917    0.5187   2.7377
Conditional MLE Odds Ratio: 1.1885    0.4625   3.0712
    Probability difference: 0.0410   -0.1486   0.2271

             Exact P-value: 0.8032
        Asymptotic P-value: 0.7288
------------------------------------------------------
```

Walking through the `twobytwo` Output

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : A1c<8
Comparing : Never vs. Current

        A1c<8 A1c>=8    P(A1c<8) 90% conf. interval
Never      22     12      0.6471    0.5040   0.7679
Current    20     13      0.6061    0.4613   0.7343
```

These are 90% confidence intervals for Pr(A1c<8) conditional on the exposure, and while we've seen five other methods for making this estimate, we use a sixth method here.

The computational details are shown on the next two slides.

90% CI for Pr(A1c $<$ 8 | Never) (`twobytwo`)

```
        A1c<8 A1c>=8    P(A1c<8) 90% conf. interval
Never      22     12      0.6471    0.5040   0.7679
```

This result is computed using the Normal approximation for log(odds), and I'll show the entire calculation on the next slide.

Key Facts:

- odds = prob / (1 - prob) lets you convert from probability to odds
- prob = odds / (1 + odds) lets you convert from odds to probability
- standard error of the log(odds) formula is

$$se_{log(odds)} = \sqrt{\frac{1}{np(1-p)}}$$

- 90% confidence interval (two-sided) requires $Z_{\alpha/2} = Z_{0.05} = 1.645$

Calculation for 90% CI for x = 22, n = 34

```
n <- 22 + 12; prob <- 22/(22+12)
odds <- prob / (1 - prob)
logodds <- log(odds)
se_logodds <- sqrt(1 / (n * prob * (1 - prob)))
ci_logodds <- c(logodds - 1.645*se_logodds,
                logodds + 1.645*se_logodds)
  # that is the 90% CI on the log(odds) scale
  # so we exponentiate to get CI on odds scale
ci_odds <- exp(ci_logodds) # ci on odds scale
  # then convert odds to probability scale
ci_prob <- ci_odds / (1 + ci_odds) # ci on prob scale
ci_prob
```

```
[1] 0.5039485 0.7678975
```

Returning to the `twobytwo` output

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : A1c<8
Comparing : Never vs. Current

                                90% conf. interval
           Relative Risk: 1.0676    0.7823   1.4571
       Sample Odds Ratio: 1.1917    0.5187   2.7377
Conditional MLE Odds Ratio: 1.1885    0.4625   3.0712
    Probability difference: 0.0410   -0.1486   0.2271
```

We get confidence intervals for four different measures comparing A1c<8 rates for Never to Current, but we'll only use three in 431.

- Relative Risk
- Odds Ratio (we'll use the sample version - the cross-product version)
- Probability Difference

Relative Risk

```
Outcome   : A1c<8
Comparing : Never vs. Current

        A1c<8 A1c>=8    P(A1c<8) 90% conf. interval
Never      22     12      0.6471    0.5040   0.7679
Current    20     13      0.6061    0.4613   0.7343

                                    90% conf. interval
             Relative Risk: 1.0676    0.7823   1.4571
```

$$RR = \frac{0.6471}{0.6061} = 1.0676$$

- What does RR = 1 imply about the probabilities we are comparing?

Relative Risk

```
Outcome   : A1c<8
Comparing : Never vs. Current

        A1c<8 A1c>=8    P(A1c<8) 90% conf. interval
Never      22     12      0.6471    0.5040   0.7679
Current    20     13      0.6061    0.4613   0.7343

                                   90% conf. interval
             Relative Risk: 1.0676    0.7823   1.4571
```

$$RR = \frac{0.6471}{0.6061} = 1.0676$$

- What does RR = 1 imply about the probabilities we are comparing?
- How about RR > 1?

Relative Risk

```
Outcome   : A1c<8
Comparing : Never vs. Current

        A1c<8 A1c>=8    P(A1c<8) 90% conf. interval
Never      22     12      0.6471    0.5040   0.7679
Current    20     13      0.6061    0.4613   0.7343

                                    90% conf. interval
             Relative Risk: 1.0676    0.7823   1.4571
```

$$RR = \frac{0.6471}{0.6061} = 1.0676$$

- What does RR = 1 imply about the probabilities we are comparing?
- How about RR > 1?
- What about RR < 1?

Odds Ratio (Sample Odds Ratio)

```
Outcome   : A1c<8
Comparing : Never vs. Current

        A1c<8 A1c>=8    P(A1c<8) 90% conf. interval
Never      22     12      0.6471    0.5040   0.7679
Current    20     13      0.6061    0.4613   0.7343

                                   90% conf. interval
         Sample Odds Ratio: 1.1917    0.5187   2.7377
```

$$OR = \frac{22 \times 13}{12 \times 20} = 1.1917$$

- What does OR = 1 imply about the probabilities being compared?

Odds Ratio (Sample Odds Ratio)

```
Outcome   : A1c<8
Comparing : Never vs. Current

        A1c<8 A1c>=8    P(A1c<8) 90% conf. interval
Never      22     12      0.6471    0.5040   0.7679
Current    20     13      0.6061    0.4613   0.7343

                                    90% conf. interval
         Sample Odds Ratio: 1.1917    0.5187   2.7377
```

$$OR = \frac{22 \times 13}{12 \times 20} = 1.1917$$

- What does OR $=$ 1 imply about the probabilities being compared?
- How about OR $>$ 1?

Odds Ratio (Sample Odds Ratio)

```
Outcome    : A1c<8
Comparing : Never vs. Current

        A1c<8 A1c>=8    P(A1c<8) 90% conf. interval
Never      22     12      0.6471    0.5040   0.7679
Current    20     13      0.6061    0.4613   0.7343

                                     90% conf. interval
         Sample Odds Ratio: 1.1917    0.5187   2.7377
```

$$OR = \frac{22 \times 13}{12 \times 20} = 1.1917$$

- What does OR = 1 imply about the probabilities being compared?
- How about OR > 1?
- What about OR < 1?

Probability Difference (also called Risk Difference)

```
Outcome   : A1c<8
Comparing : Never vs. Current

        A1c<8 A1c>=8    P(A1c<8) 90% conf. interval
Never      22     12      0.6471    0.5040   0.7679
Current    20     13      0.6061    0.4613   0.7343

                                    90% conf. interval
    Probability difference: 0.0410   -0.1486   0.2271
```

$$\Delta = 0.6471 - 0.6061 = 0.0410$$

- What will the probability difference be if the probabilities are the same?

Probability Difference (also called Risk Difference)

```
Outcome   : A1c<8
Comparing : Never vs. Current

        A1c<8 A1c>=8    P(A1c<8) 90% conf. interval
Never      22     12      0.6471    0.5040   0.7679
Current    20     13      0.6061    0.4613   0.7343

                                    90% conf. interval
    Probability difference: 0.0410   -0.1486   0.2271
```

$$\Delta = 0.6471 - 0.6061 = 0.0410$$

- What will the probability difference be if the probabilities are the same?
- What does a positive risk difference imply?

Probability Difference (also called Risk Difference)

```
Outcome   : A1c<8
Comparing : Never vs. Current

        A1c<8 A1c>=8    P(A1c<8) 90% conf. interval
Never      22     12      0.6471    0.5040   0.7679
Current    20     13      0.6061    0.4613   0.7343

                                    90% conf. interval
    Probability difference: 0.0410   -0.1486   0.2271
```

$$\Delta = 0.6471 - 0.6061 = 0.0410$$

- What will the probability difference be if the probabilities are the same?
- What does a positive risk difference imply?
- How about a negative risk difference?

Hypothesis Testing?

At the bottom of the `twobytwo` output, we have two p values...

```
Exact P-value: 0.8032        Asymptotic P-value: 0.7288
```

The `Exact P-value` comes from Fisher's exact test, and is technically exact only if we treat the row and column totals as being fixed. The `Asymptotic P-value` comes from a Pearson χ^2 test. These test:

- H_0: Pr(A1c $<$ 8 $|$ Never) $=$ Pr(A1c $<$ 8 $|$ Current) vs.
- H_A: Pr(A1c $<$ 8 $|$ Never) $\neq$ Pr(A1c $<$ 8 $|$ Current).

We usually state this as:

- H_0: rows and columns of the table are *independent* (Pr(A1c$<$8) is the same regardless of which row you're in) vs.
- H_A: the rows and columns of the table are *associated*.

Bayesian Augmentation in a 2x2 Table?

Original command:

```
twobytwo(22, 12, 20, 13,
      "Never", "Current",
      "A1c<8", "A1c>=8", conf.level = 0.90)
```

Bayesian augmentation approach: Add two successes and add two failures in each row...

```
twobytwo(22+2, 12+2, 20+2, 13+2,
      "Never", "Current",
      "A1c<8", "A1c>=8", conf.level = 0.90)
```

Output shown on next slide.

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : A1c<8
Comparing : Never vs. Current

        A1c<8 A1c>=8    P(A1c<8) 90% conf. interval
Never      24     14      0.6316    0.4965   0.7488
Current    22     15      0.5946    0.4582   0.7178

                                   90% conf. interval
             Relative Risk: 1.0622    0.7851   1.4371
         Sample Odds Ratio: 1.1688    0.5355   2.5513
Conditional MLE Odds Ratio: 1.1664    0.4837   2.8243
    Probability difference: 0.0370   -0.1437   0.2148

             Exact P-value: 0.8147
        Asymptotic P-value: 0.7424
------------------------------------------------------
```

Example B: Statin use in Medicaid vs. Uninsured

In the dm431 data, suppose we want to know whether statin prescriptions are more common among Medicaid patients than Uninsured subjects. So, we want a two-way table with "Medicaid", "Statin" in the top left.

```
dm431 %>%
  filter(insurance %in% c("Medicaid", "Uninsured")) %>%
  tabyl(insurance, statin)
```

```
  insurance  0  1
 Commercial  0  0
   Medicaid 17 83
   Medicare  0  0
  Uninsured 15 29
```

But we want the tabyl just to show the levels of insurance we're studying...

Obtaining a 2x2 Table from a tibble

We want to know whether statin prescriptions are more common among Medicaid patients than Uninsured subjects.. So, we want a two-way table with "Medicaid", "Uninsured" in the top left.

```
dm431 %>%
  filter(insurance %in% c("Medicaid", "Uninsured")) %>%
  droplevels() %>%
  tabyl(insurance, statin)
```

```
 insurance  0  1
  Medicaid 17 83
 Uninsured 15 29
```

But we want Medicaid in the top row (ok) and "statin = yes" in the left column (must fix)...

Building and Releveling Factors in the tibble

```
exampleB <- dm431 %>%
  filter(insurance %in% c("Medicaid", "Uninsured")) %>%
  droplevels() %>%
  mutate(insur_f = fct_relevel(insurance, "Medicaid"),
         statin_f = fct_recode(factor(statin),
                               statin = "1", no_statin = "0"),
         statin_f = fct_relevel(statin_f, "statin"))

exampleB %>% tabyl(insur_f, statin_f)
```

```
   insur_f statin no_statin
  Medicaid     83        17
 Uninsured     29        15
```

Since Medicaid was already on top, we didn't *have to* set `insur_f`.

Adding percentages

```
exampleB %>% tabyl(insur_f, statin_f) %>%
  adorn_totals(where = c("row", "col")) %>%
  adorn_percentages(denom = "row") %>%
  adorn_pct_formatting(digits = 1) %>%
  adorn_ns(position = "front") %>%
  adorn_title(row = "Insurance", col = "Statin Status")
```

```
           Statin Status
 Insurance        statin  no_statin         Total
  Medicaid   83 (83.0%) 17 (17.0%) 100 (100.0%)
 Uninsured   29 (65.9%) 15 (34.1%)  44 (100.0%)
     Total  112 (77.8%) 32 (22.2%) 144 (100.0%)
```

Running twoby2 against a data table

The twoby2 function from the Epi package can operate with tables (but not, alas, tabyls) generated from data.

Original Data

```
twoby2(exampleB %$% table(insur_f, statin_f))
```

(output on next slide)

With Bayesian Augmentation

```
twoby2(exampleB %$% table(insur_f, statin_f) + 2)
```

(output on the slide after that)

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : statin
Comparing : Medicaid vs. Uninsured

          statin no_statin    P(statin) 95% conf. interval
Medicaid      83        17       0.8300    0.7434   0.8916
Uninsured     29        15       0.6591    0.5090   0.7829

                                    95% conf. interval
             Relative Risk: 1.2593    1.0003   1.5854
         Sample Odds Ratio: 2.5254    1.1202   5.6933
Conditional MLE Odds Ratio: 2.5074    1.0252   6.1298
    Probability difference: 0.1709    0.0218   0.3307

             Exact P-value: 0.0299
        Asymptotic P-value: 0.0255
------------------------------------------------------
```

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : statin
Comparing : Medicaid vs. Uninsured

          statin no_statin    P(statin) 95% conf. interval
Medicaid      85        19       0.8173    0.7312   0.8803
Uninsured     31        17       0.6458    0.5023   0.7671

                                    95% conf. interval
             Relative Risk: 1.2655    1.0071   1.5901
         Sample Odds Ratio: 2.4533    1.1327   5.3136
Conditional MLE Odds Ratio: 2.4375    1.0464   5.6838
    Probability difference: 0.1715    0.0245   0.3261

             Exact P-value: 0.0251
        Asymptotic P-value: 0.0228
------------------------------------------------------
```

Measuring Association using Categorical Variables with Chi-Squared Tests: Working with Two-Way Tables

A Two-Way Table

```
dm431 %>% tabyl(insurance, tobacco)
```

```
  insurance Current Former Never
 Commercial      35     60    69
   Medicaid      33     33    34
   Medicare      17     58    48
  Uninsured      11     13    20
```

Is tobacco use status associated with insurance type?

- Two factors here (insurance, tobacco)
- This is a 4×3 table, with 4 rows and 3 columns.

Is tobacco use associated with insurance type?

H_0: Tobacco status is independent of insurance type

```
dm431 %>% tabyl(insurance, tobacco) %>%
  adorn_totals(where = "row") %>%
  adorn_percentages() %>% adorn_pct_formatting() %>%
  adorn_ns(position = "front") %>% adorn_title()
```

```
             tobacco
  insurance  Current        Former         Never
 Commercial 35 (21.3%)  60 (36.6%)  69 (42.1%)
   Medicaid 33 (33.0%)  33 (33.0%)  34 (34.0%)
   Medicare 17 (13.8%)  58 (47.2%)  48 (39.0%)
  Uninsured 11 (25.0%)  13 (29.5%)  20 (45.5%)
      Total 96 (22.3%) 164 (38.1%) 171 (39.7%)
```

Independence model: tobacco rates are 22.3%, 38.1%, 39.7% in each row

Independence Model for Insurance and Tobacco

Table shows observed counts + (expected under independence model)

- expected count = (row total) $\times$ (col total) / (grand total)
- for example, Medicaid/Current expected count is 100*96/431 = 22.3

	Current	Former	Never	Total
Commercial	35 (36.5)	60 (62.4)	69 (65.1)	164
Medicaid	33 (22.3)	33 (38.1)	34 (39.7)	100
Medicare	17 (27.4)	58 (46.8)	48 (48.8)	123
Uninsured	11 (9.8)	13 (16.7)	20 (17.5)	44
Total	96	164	171	431

Since all of these expected counts exceed 10, the Pearson χ^2 test should provide a reasonably accurate approximate p value for H_0: rows and columns are independent.

dm431 association of tobacco with insurance

```r
tab43 <- dm431 %$% table(insurance, tobacco)
tab43
```

```
            tobacco
insurance     Current Former Never
  Commercial       35     60    69
  Medicaid         33     33    34
  Medicare         17     58    48
  Uninsured        11     13    20
```

```r
chisq.test(tab43)
```

```

	Pearson's Chi-squared test

data:  tab43
X-squared = 15.033, df = 6, p-value = 0.02
```

Can obtain observed + expected counts

```
chisq.test(tab43)$expected
```

```
             tobacco
insurance       Current   Former    Never
  Commercial 36.529002 62.40371 65.06729
  Medicaid   22.273782 38.05104 39.67517
  Medicare   27.396752 46.80278 48.80046
  Uninsured   9.800464 16.74246 17.45708
```

```
chisq.test(tab43)$observed
```

```
             tobacco
insurance     Current Former Never
  Commercial       35     60    69
  Medicaid         33     33    34
  Medicare         17     58    48
  Uninsured        11     13    20
```

Can obtain residuals (observed - expected)

```
chisq.test(tab43)$residuals
```

```
            tobacco
insurance       Current     Former      Never
  Commercial -0.2529818 -0.3042827  0.4875409
  Medicaid    2.2727394 -0.8188378 -0.9009896
  Medicare   -1.9863151  1.6367193 -0.1145855
  Uninsured   0.3831686 -0.9146343  0.6086218
```

```
chisq.test(tab43)$stdres
```

```
            tobacco
insurance       Current     Former      Never
  Commercial -0.3645763 -0.4911829  0.7975282
  Medicaid    2.9416467 -1.1871498 -1.3237207
  Medicare   -2.6651867  2.4599172 -0.1745200
  Uninsured   0.4586584 -1.2263477  0.8269565
```

Augmented test using `mosaic::xchisq.test()`

(see full output on next slide)

```
xchisq.test(tobacco ~ insurance, data = dm431)
```

Note that flipping the rows and columns here changes the table, but not the conclusions of the χ^2 test.

```
	Pearson's Chi-squared test
X-squared = 15.033, df = 6, p-value = 0.02

   35        33        17        11
(36.53)   (22.27)   (27.40)   ( 9.80)
[0.064]   [5.165]   [3.945]   [0.147]
<-0.25>   < 2.27>   <-1.99>   < 0.38>
                                          key:
   60        33        58        13        observed
(62.40)   (38.05)   (46.80)   (16.74)     (expected)
[0.093]   [0.670]   [2.679]   [0.837]     [X-square contribution]
<-0.30>   <-0.82>   < 1.64>   <-0.91>     <Pearson residual>

   69        34        48        20
(65.07)   (39.68)   (48.80)   (17.46)
[0.238]   [0.812]   [0.013]   [0.370]
< 0.49>   <-0.90>   <-0.11>   < 0.61>
```

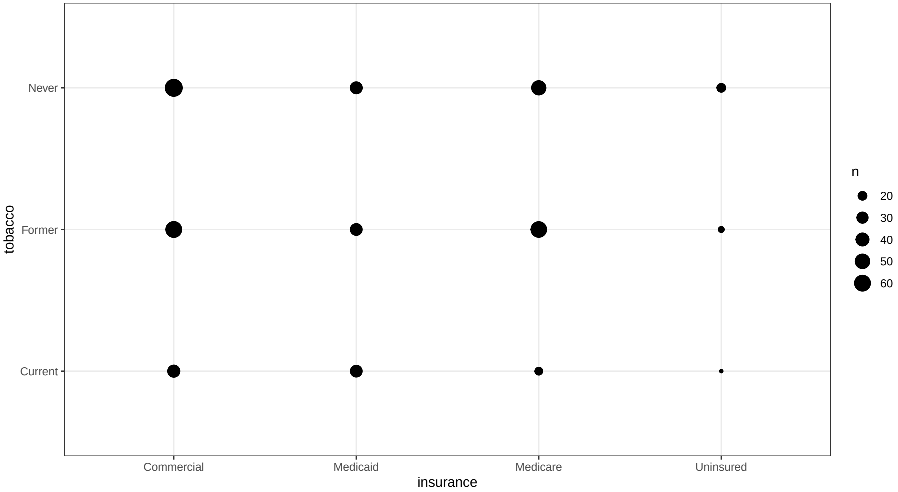
Visualizing the Association

We have cell counts for a 4×3 table here. How to visualize?

1. Use a tabyl (or table).
2. Consider the use of `geom_count`?
3. Consider a `mosaicplot`? The `mosaicplot` is a feature of the `graphics` package in base R, and has nothing to do with the `mosaic` package.
4. Consider an `assocplot`?

Using geom_count

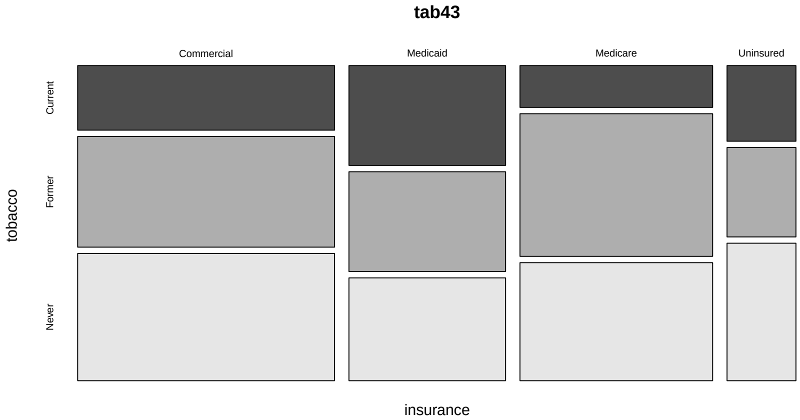
```
ggplot(dm431, aes(x = insurance, y = tobacco)) +
  geom_count()
```



Using `mosaicplot` to show what's in the table

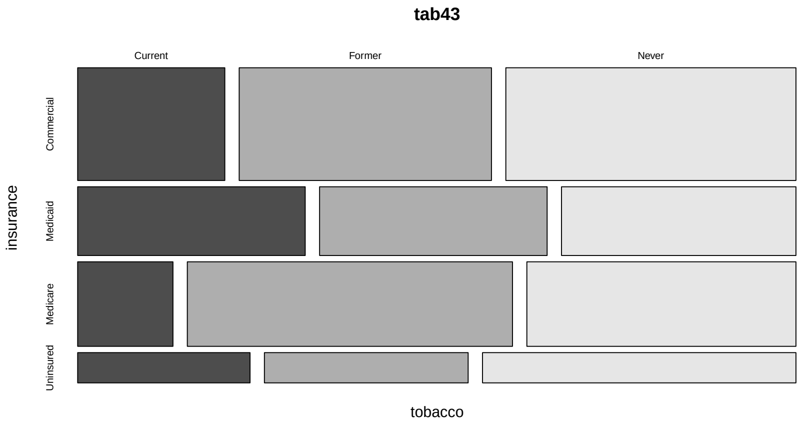
- Area of each box corresponds to its observed count

```
mosaicplot(tab43, color = TRUE)
```



Flipping the coordinates of the mosaicplot

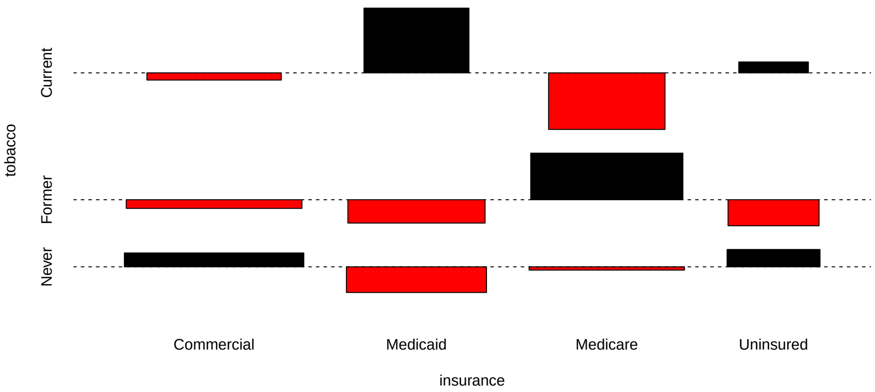
```
mosaicplot(tab43, color = TRUE, dir = c("h", "v"))
```



assocplot to show deviations from independence

- Area of each box is proportional to (observed - expected count)

```r
assocplot(tab43)
```



A Second Example (Statin Use by Practice)

```r
dm431 %>% tabyl(practice, statin)
```

```
  practice  0   1
 Arlington 20 101
   Bristol 26 107
   Chester 16  34
     Dover 14  49
  Franklin 10  54
```

Is there an association between rates of statin usage and practice?

Include the Percentages

```
dm431 %>% tabyl(practice, statin) %>%
  adorn_totals(where = "row") %>%
  adorn_percentages() %>% adorn_pct_formatting() %>%
  adorn_ns(position = "front") %>% adorn_title()
```

```
              statin
  practice         0            1
 Arlington 20 (16.5%) 101 (83.5%)
   Bristol 26 (19.5%) 107 (80.5%)
   Chester 16 (32.0%)  34 (68.0%)
     Dover 14 (22.2%)  49 (77.8%)
  Franklin 10 (15.6%)  54 (84.4%)
     Total 86 (20.0%) 345 (80.0%)
```

Does H_0 hold up well?

Chi-Square Test

```
dm431 %>% tabyl(practice, statin) %>% chisq.test()
```

```
    Pearson's Chi-squared test

data:  .
X-squared = 6.3987, df = 4, p-value = 0.1713
```

The association we see isn't strong enough to be detectable by this method. Is this because of the sample sizes?

Expected Counts for Practice - Statin Association

This is a 5×2 two-way table.

```
tab52 <- dm431 %>% tabyl(practice, statin) %>% chisq.test()
tab52$expected
```

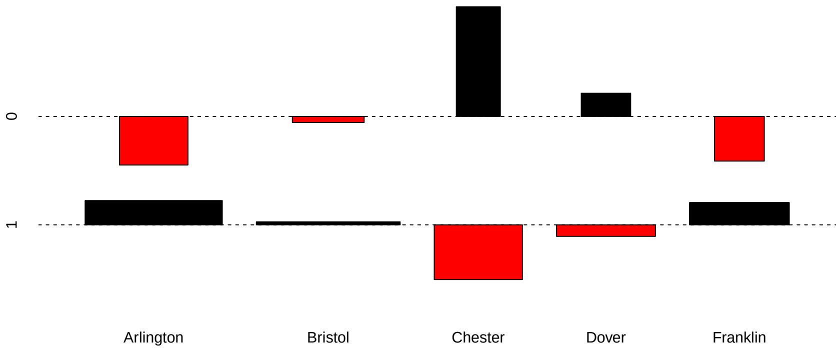
```
  practice         0         1
 Arlington 24.143852  96.85615
   Bristol 26.538283 106.46172
   Chester  9.976798  40.02320
     Dover 12.570766  50.42923
  Franklin 12.770302  51.22970
```

The expected frequencies look like they are 10 or more (with one slight exception.)

`assocplot` for practice × statin

Which practice might we look at most closely?

```
assocplot(table(dm431$practice, dm431$statin))
```



What's Next?

- Mantel-Haenszel test for comparing a series of 2x2 tables.
- I'll skip paired comparisons of proportions in 431 this year.

Then we'll move on to power and sample size considerations.