431 Class 20

Thomas E. Love

2018-11-08

Today's Agenda

- Dealing with Larger two-way contingency tables (Notes Chapter 32)
    - Building a $J \times K$ Table
    - χ^2 Tests of Independence
- Dealing with an additional categorical variable (Notes Chapter 33)
    - The Cochran-Mantel-Haenszel Test
    - The Woolf test to check assumptions
    - Aggregation and Simpson's Paradox

and, if we get to it . . .

- Comparing 3 or more Population Means with ANOVA (Notes Chapter 28)

Today's R Setup

```r
source("Love-boost.R") # helps to load Hmisc explicitly

library(Hmisc); library(magrittr); library(vcd)
library(tidyverse) # always load tidyverse last

surd1 <- read.csv("data/surveyday1_2018.csv") %>% tbl_df
dm192 <- read.csv("data/dm192.csv") %>% tbl_df
```

Setting the Foundation: A 2x3 Contingency Table

Comparing a 3-Category Response to Either Active Treatment or Placebo

The table below, specifies the number of patients who show *complete*, *partial*, or *no response* after treatment with either **active** medication or a **placebo**.

Group	None	Partial	Complete
Active	16	26	29
Placebo	24	26	18

Is there a statistically significant association here? That is to say, is there a statistically significant difference between the treatment groups in the distribution of responses?

Getting the Table into R

To answer this, we'll put the table into a matrix in R. Here's one approach...

```r
T1 <- matrix(c(16,26,29,24,26,18),
             ncol=3, nrow=2, byrow=TRUE)
rownames(T1) <- c("Active", "Placebo")
colnames(T1) <- c("None", "Partial", "Complete")
T1
```

```
        None Partial Complete
Active    16      26       29
Placebo   24      26       18
```

Getting the Chi-Square Test Results

H_0: rows and columns are independent vs. H_A: rows and columns are associated

```r
chisq.test(T1)
```

```

    Pearson's Chi-squared test

data:  T1
X-squared = 4.1116, df = 2, p-value = 0.128
```

Chi-Square Assumptions

We assume that the expected frequency, under the null hypothesized model of independence, will be **at least 5** in each cell. If that is not the case, then the χ^2 test is likely to give unreliable results.

How do we calculate expected frequencies for a cell?

$$\text{Expected Frequency} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand Total}}$$

This assumes that the independence model holds - that the probability of being in a particular column is exactly the same regardless of what row we're looking at.

Expected Frequencies under Independence

```r
addmargins(T1)
```

```
         None Partial Complete Sum
Active     16      26       29  71
Placebo    24      26       18  68
Sum        40      52       47 139
```

- What is the expected frequency for the (Active, None) cell?
  - row total = 71, column total = 40, grand total = 139, and

$$\frac{71 \times 40}{139} = 20.43$$

- In fact, all of the expected frequencies turn out to be at least 5, so our χ^2 results should be reasonable.
- We could have run a **Fisher's exact test**, too...

Fisher's Exact Test Results

H_0: rows and columns are independent vs. H_A: rows and columns are associated

```r
fisher.test(T1)
```

```
    Fisher's Exact Test for Count Data

data:  T1
p-value = 0.1346
alternative hypothesis: two.sided
```

Large Two-Way Tables and the Day 1 Survey Data for 431

Working with Day 1 431 Survey Data

Suppose we want to study the association between two items. Each was measured on a (1 = Strongly Disagree, 5 = Strongly Agree) scale.

9. I prefer to learn from lectures than to learn from activities.
10. I prefer to work on projects alone than in a team.

Let's start by gathering the data we need.

```r
sur1 <- surd1 %>%
  filter(complete.cases(student, lecture, alone)) %>%
  select(student, lecture, alone) %>%
  mutate(lecture = factor(lecture),
         alone = factor(alone))
```

A 5x5 Contingency Table

```
sur1 %$% table(lecture, alone) %>% addmargins
```

```
       alone
lecture   1   2   3   4   5 Sum
    1     4   6   4   1   0  15
    2     7  18  27  14   2  68
    3     7  25  32  32   6 102
    4     1   8  19  20   6  54
    5     2   2   7   0   3  14
    Sum  21  59  89  67  17 253
```

The Chi-Square Test

For any contingency table, the Pearson chi-square (χ^2) test will assess:

- H_0: No association of rows (lecture rating) and columns (alone rating) vs.
- H_A: Rows (lecture) and columns (alone) are associated

Chi-Square testing

```r
sur1 %$% table(lecture, alone) %>% chisq.test
```

```
Warning in chisq.test(.): Chi-squared approximation may
be incorrect


    Pearson's Chi-squared test

data:  .
X-squared = 34.924, df = 16, p-value = 0.004071
```

Why the warning message?

There are some very small cells in the table, with just 1 or even 0 subjects… Could we eliminate that problem?

Collapse Some Categories with `fct_recode`

9. I prefer lectures over activities. (1 = SD, 5 = SA)
10. I prefer to work alone instead of in a team. (1 = SD, 5 = SA)

```
sur2 <- sur1 %>% filter(complete.cases(lecture, alone)) %>%
  mutate(lec2 = fct_recode(lecture,
                           "Activities" = "1",
                           "Activities" = "2",
                           "Neutral" = "3",
                           "Lectures" = "4",
                           "Lectures" = "5"),
         alone2 = fct_recode(alone,
                              "Team" = "1",
                              "Team" = "2",
                              "Neutral" = "3",
                              "Alone" = "4",
                              "Alone" = "5"))
```

Result of Collapsing Categories

```
sur2 %$% table(lec2, alone2) %>% addmargins
```

```
            alone2
lec2          Team Neutral Alone Sum
  Activities    35      31    17  83
  Neutral       32      32    38 102
  Lectures      13      26    29  68
  Sum           80      89    84 253
```

Collapsed Contingency Table's Chi-Square test

Again, the Pearson chi-square (χ^2) test will assess:

- H_0: No association of rows (lecture rating) and columns (alone rating) vs.
- H_A: Rows (lecture) and columns (alone) are associated

What do we conclude now, from this collapsed cross-tabulation?

```
sur2 %$% table(lec2, alone2) %>% chisq.test()
```

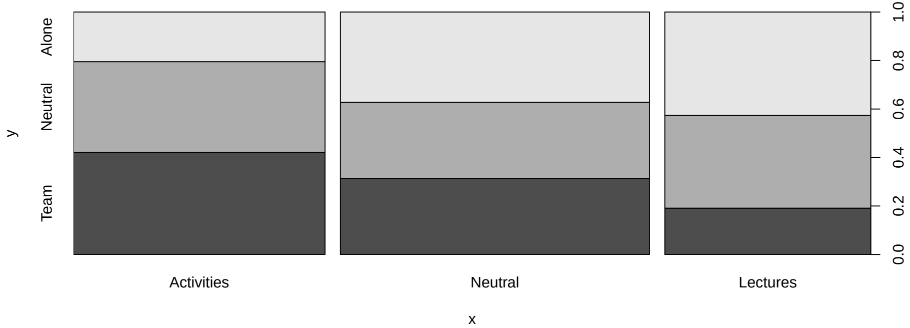
```

    Pearson's Chi-squared test

data:  .
X-squared = 13.373, df = 4, p-value = 0.009592
```

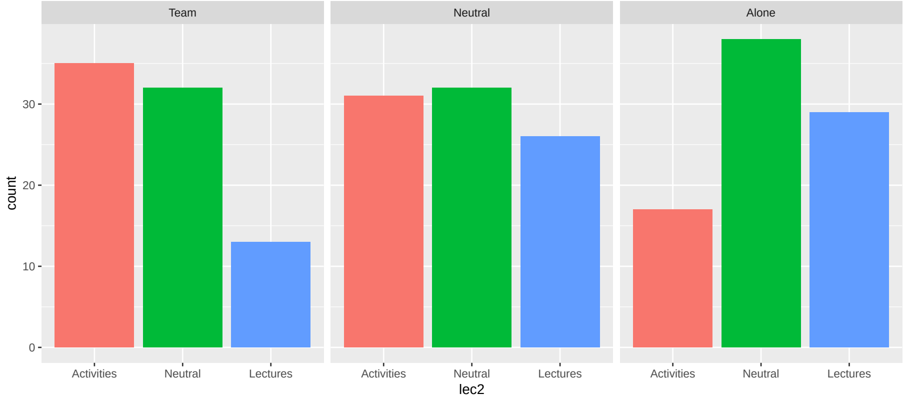
Default plot for Categorical Data (Mosaic plot)

```
plot(sur2$lec2, sur2$alone2)
```



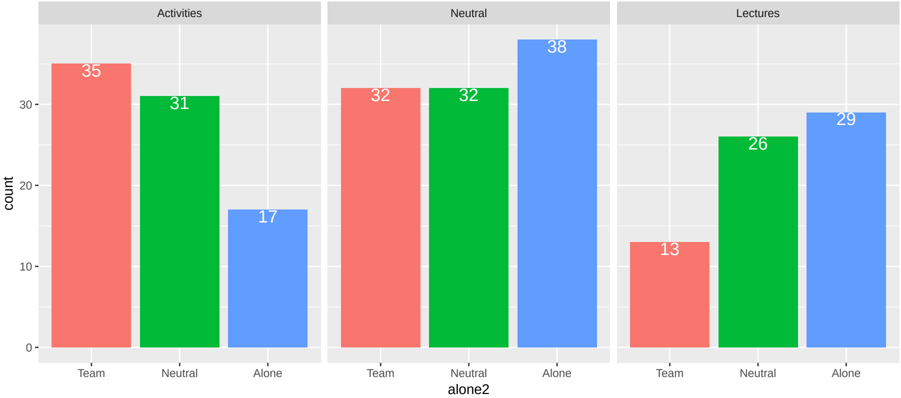
Plotting Categorical Data (Bars)

```
ggplot(sur2, aes(x = lec2, fill = lec2)) +
  geom_bar() + guides(fill = FALSE) +
  facet_wrap(~ alone2)
```



Plotting Categorical Data (Bars with Counts)

```
ggplot(sur2, aes(x = alone2, fill = alone2)) +
  geom_bar() + guides(fill = FALSE) +
  geom_text(stat = 'count', aes(label = ..count..),
            col = "white", vjust = 1, size = 5) +
  facet_wrap(~ lec2)
```



Example: Treatment of Kidney Stones

Kidney Stone Treatment Example

Suppose we compare the success rates of two treatments for kidney stones.

- Treatment A (all open surgical procedures): 273/350 patients (78%) had a successful result.
- Treatment B (percutaneous nephrolithotomy - less invasive): 289/350 were successful (83%).

Kidney Stones	Successful Outcome	Bad Outcome
A (open surgery)	273 (78%)	77 (22%)
B (less invasive)	289 (83%)	61 (17%)

Which approach would you choose?

- Sources: https://en.wikipedia.org/wiki/Simpson%27s_paradox and Charig CR et al. (1986) PMID 3083922.

Kidney Stones, twobytwo results

```r
twobytwo(273, 77, 289, 61, "A", "B", "Success", "Bad")
```

```
2 by 2 table analysis:
--------------------------------------------------
Outcome   : Success        Comparing : A vs. B

  Success Bad     P(Success) 95% conf. interval
A     273  77         0.7800    0.7336   0.8203
B     289  61         0.8257    0.7823   0.8620

                                95% conf. interval
         Relative Risk:  0.9446    0.8776   1.0168
     Sample Odds Ratio:  0.7483    0.5146   1.0883
Probability difference: -0.0457   -0.1045   0.0133

Exact P-value: 0.154    Asymptotic P-value: 0.1292
--------------------------------------------------
```

Kidney Stones: A Third Variable

But this comparison may be misleading.

Some kidney stones are small, and some are large.

- Treatment A was used in 87 small stones, and 263 large ones.
- Treatment B was used in 270 small stones, and 80 large ones.

Could that bias our results? Should we account for this difference in “size mix”?

Kidney Stone results stratified by stone size

- For small stones, the odds ratio for a successful outcome comparing A to B is 2.08 (95% CI 0.84, 5.11)

Small Stones	Successful Outcome	Bad Outcome
A (open surgery)	81 (93%)	6 (7%)
B (less invasive)	234 (87%)	36 (13%)

- For large stones, that odds ratio is 1.23 (95% CI 0.71, 2.12)

Large Stones	Successful Outcome	Bad Outcome
A (open surgery)	192 (73%)	71 (27%)
B (less invasive)	55 (69%)	25 (31%)

Aggregated Data: % with Successful Outcome

- 78% of Treatment A subjects, 83% of Treatment B

What We Have Here is a Three-Way Table

- rows: which treatment was received (A or B)
- columns: was the outcome Successful or Bad?
- *strata* or *layers*: was the stone Small or Large?

```
Size  Treatment  Good  Bad  Total  % Good
----- ---------  ----  ---  -----  ------
Small     A        81    6     87     93
Small     B       234   36    270     87
Large     A       192   71    263     73
Large     B        55   25     80     69
```

We'll talk about three-way and larger contingency tables more in 432, but in 431, we focus on the situation where a 2x2 table is repeated over multiple strata (categories in a third variable.)

A Meta-Analysis of Niacin vs. Placebo in Coronary Artery Disease

The Niacin and Heart Disease Meta-Analysis

Duggal et al (2010) did a meta-analysis[1] of 5 placebo-controlled studies (AFREGS, ARBITER2, CLAS1, FATS and HATS) of niacin and heart disease, where the primary outcome was the need to do a coronary artery revascularization procedure.

For example, the FATS study had these results:

FATS	Revascularization	No Revasc.
Niacin	2	46
Placebo	11	41

FATS is just one of the five studies, and this table exists in each!

[1] Duggal JK et al. 2010. Effect of niacin therapy on cardiovascular outcomes in patients with coronary artery disease. J Cardiovasc Pharmacology & Therapeutics 15: 158-166. My Source: http://www.biostathandbook.com/cmh.html

Exploring the FATS study

FATS	Revascularization	No Revasc.
Niacin	2	46
Placebo	11	41

- Pr(revascularization | Niacin) $= \frac{2}{2+46} = 0.042$
- Odds(revascularization | Niacin) $= \frac{2}{46} = 0.043$
- Pr(revascularization | Placebo) $= \frac{11}{11+41} = 0.212$
- Odds(revascularization | Placebo) $= \frac{11}{41} = 0.268$

and so the Odds Ratio $= \frac{2*41}{11*46} = 0.16$.

But that is just the result for the FATS study.

Building the Meta-Analysis Table

```r
study <- c(rep("FATS", 4), rep("AFREGS", 4),
           rep("ARBITER2", 4), rep("HATS", 4),
           rep("CLAS1", 4))
treat <- c(rep(c("Niacin", "Niacin",
                 "Placebo", "Placebo"),5))
outcome <- c(rep(c("Revasc.", "No Rev."), 10))
counts <- c(2, 46, 11, 41, 4, 67, 12, 60, 1, 86,
            4, 76, 1, 37, 6, 32, 2, 92, 1, 93)
meta <- data.frame(study, treat, outcome, counts) %>% tbl_df
meta$treat <- fct_relevel(meta$treat, "Niacin")
meta$outcome <- fct_relevel(meta$outcome, "Revasc.")
meta.tab <- xtabs(counts ~ treat + outcome + study,
                  data = meta)
```

Five Studies in the Meta-Analysis

```
ftable(meta.tab)
```

```
                 study AFREGS ARBITER2 CLAS1 FATS HATS
treat   outcome
Niacin  Revasc.             4        1     2    2    1
        No Rev.            67       86    92   46   37
Placebo Revasc.            12        4     1   11    6
        No Rev.            60       76    93   41   32
```

The three variables we are studying are:

- `treat` (2 levels: Niacin/Placebo),
- `outcome` (2 levels: Revascularization or No Revascularization) across
- `study` (5 levels: AFREGS, ARBITER2, CLAS1, FATS, HATS)

Cochran-Mantel-Haenszel Test

The Cochran-Mantel-Haenszel test is designed to test whether the rate of revascularization is the same across the two levels of the treatment (i.e. Niacin or Placebo).

- We *could* do this by simply adding up the results across the five studies, but that wouldn't be wise, because the studies used different populations and looked for revascularization after different lengths of time.
- But we can account for the differences between studies to some extent by adjusting for study as a stratifying variable in a CMH test.
- The big assumption we'll have to make, though, is that the odds ratio for revascularization given Niacin instead of Placebo does not change across the studies. Is this reasonable in our case?

Looking at the Study-Specific Odds Ratios

We'll calculate the odds ratios, comparing revascularization odds with niacin vs. placebo, within each separate study.

Study	Rev N	Rev P	NoRev N	NoRev P	Odds Ratio
AFREGS	4	67	12	60	$\frac{4*60}{67*12} = 0.3$
ARBITER2	1	86	4	76	0.22
CLAS1	2	92	1	93	2.02
FATS	2	46	11	41	0.16
HATS	1	37	6	32	0.14

The table shows patient counts for the categories in each of the respective two-by-two tables (Rev N = Revascularization and Niacin, NoRev P = No Revascularization and Placebo, etc.)

Can we assume a Common Odds Ratio?

The Woolf test checks a key assumption for the Cochran-Mantel-Haenszel test. The Woolf test assesses the null hypothesis of a common odds ratio across the five studies.

```
woolf_test(meta.tab)
```

```

    Woolf-test on Homogeneity of Odds Ratios (no
    3-Way assoc.)

data:  meta.tab
X-squared = 3.4512, df = 4, p-value = 0.4853
```

Our conclusion from the Woolf test is that we are able to retain the null hypothesis of homogeneous odds ratios. So it's not crazy to fit a test that requires that all of the odds ratios be the same in the population.

Running the Cochran-Mantel-Haenszel test

So, we can use the Cochran-Mantel-Haenszel test to make inferences about the population odds ratio (for revascularization given niacin rather than placebo) accounting for the five studies. We'll use a 90% confidence interval, and the results appear on the next slide.

```
mantelhaen.test(meta.tab, conf.level = .90)
```

Complete CMH output

```r
mantelhaen.test(meta.tab, conf.level = .90)
```

```
Mantel-Haenszel chi-squared test with continuity correction

data:  meta.tab
Mantel-Haenszel
X-squared = 12.746, df = 1, p-value = 0.0003568

alt. hypothesis: true common odds ratio is not equal to 1

90 percent confidence interval: 0.1468942 0.4968686
sample estimates: common odds ratio 0.2701612
```

What can we conclude in this case?

Another Example: Admissions to Departments at the University of California at Berkeley

The UC Berkeley Student Admissions Example

The UCBAdmissions data set contains aggregate data on applicants to graduate school at Berkeley for the six largest departments in 1973, classified by whether the applicant was admitted, and their sex.

```r
ftable(UCBAdmissions)
```

```
                Dept   A   B   C   D   E   F
Admit    Gender
Admitted Male        512 353 120 138  53  22
         Female       89  17 202 131  94  24
Rejected Male        313 207 205 279 138 351
         Female       19   8 391 244 299 317
```

Do the data show evidence of sex bias in admission practices?

Summarizing Department D

In Department D, we have

Department D	Males	Females
Admitted	138	131
Not Admitted	279	244
Applicants	417	375

$$\text{Pr(Admitted if Male)} = \frac{138}{138+279} = 0.331$$

$$\text{Odds(Admitted if Male)} = \frac{138}{279} = 0.49$$

$$\text{Pr(Admitted if Female)} = \frac{131}{131+244} = 0.349$$

$$\text{Odds(Admitted if Female)} = \frac{131}{244} = 0.54$$

$$\text{Odds Ratio (Admit for Male vs Female)} = \frac{138*244}{131*279} = 0.92$$

Can we use the Cochran-Mantel-Haenszel test?

Are the odds ratios similar across departments?

Department	A	B	C	D	E	F
Admitted Males	512	353	120	138	53	22
Male Applicants	825	560	325	417	191	373
Admitted Females	89	17	202	131	94	24
Female Applicants	108	25	593	375	393	341
Pr(Admit if Male)	0.62	0.63	0.37	0.33	0.28	0.06
Pr(Admit if Female)	0.82	0.68	0.34	0.35	0.24	0.07
Odds(Admit if Male)	1.64	1.71	0.59	0.49	0.38	0.06
Odds(Admit if Female)	4.68	2.12	0.52	0.54	0.31	0.08
Odds Ratio	0.35	0.8	1.13	0.92	1.22	0.83

Can we use a Cochran-Mantel-Haenszel test?

A Cochran-Mantel-Haenszel test describes a single combined odds ratio accounting for department. This assumes that the population odds ratio for admission by sex is identical for each of the six strata (departments).

- Does that seem reasonable?
- Or is there a three-way interaction here, where the odds ratios for admission by sex differ significantly across departments?

Department	A	B	C	D	E	F
Odds Ratio	0.35	0.8	1.13	0.92	1.22	0.83

How can we test this?

Woolf Test for Interaction in UCB Admissions

- H_0: There is no three-way interaction.
  - Odds ratios are homogenous, and we may proceed with the CMH test.)
- H_A: There is a meaningful three-way interaction.
  - CMH test is inappropriate because there are significantly different odds ratios across the departments.

```
woolf_test(UCBAdmissions)
```

```

    Woolf-test on Homogeneity of Odds Ratios (no
    3-Way assoc.)

data:  UCBAdmissions
X-squared = 17.902, df = 5, p-value = 0.003072
```

What's Going On Here?

Department	A	B	C	D	E	F
Pr(Admit if Male)	0.62	0.63	0.37	0.33	0.28	0.06
Pr(Admit if Female)	0.82	0.68	0.34	0.35	0.24	0.07
Pr (Admitted, regardless of sex)	0.64	0.63	0.35	0.34	0.25	0.06
% of Applicants who are Female	11.6	4.3	64.6	47.3	67.3	47.8

- Females used to apply more to departments with lower admission rates.
- This is a famous example related to what is called Simpson's Paradox.

A NEW TOPIC: Comparing 3 or more Population Means, with the Analysis of Variance (ANOVA)

Analysis of Variance to Compare More Than Two Population Means using Independent Samples

Suppose we want to compare more than two population means, and we have collected three or more independent samples.

This is analysis of a continuous outcome variable on the basis of a single categorical factor. In fact, it's often called **one-factor** ANOVA or **one-way** ANOVA to indicate that the outcome is being split up into the groups defined by a single factor.

- H_0: population means in each group are the same
- H_A: H_0 isn't true; at least one μ differs from the others

When there are just two groups, then this boils down to an F test that is equivalent to the Pooled t test.

One-Way ANOVA

If we have a grouping factor with k levels, then we are testing:

- H_0: $\mu_1 = \mu_2 = ... = \mu_k$ vs.
- H_A: At least one of the population means $\mu_1, \mu_2, ..., \mu_k$ is different from the others.

Returning to the `dm192` Example

- Our outcome is the `a1c` value (measured as a percentage),
- Factor is the insurance group (we'll compare 3 categories).

The dm192 data: Comparing Insurance Groups on Hemoglobin A1c

```r
dm_ins <- select(dm192, pt.id, insurance, a1c)
summary(dm_ins)
```

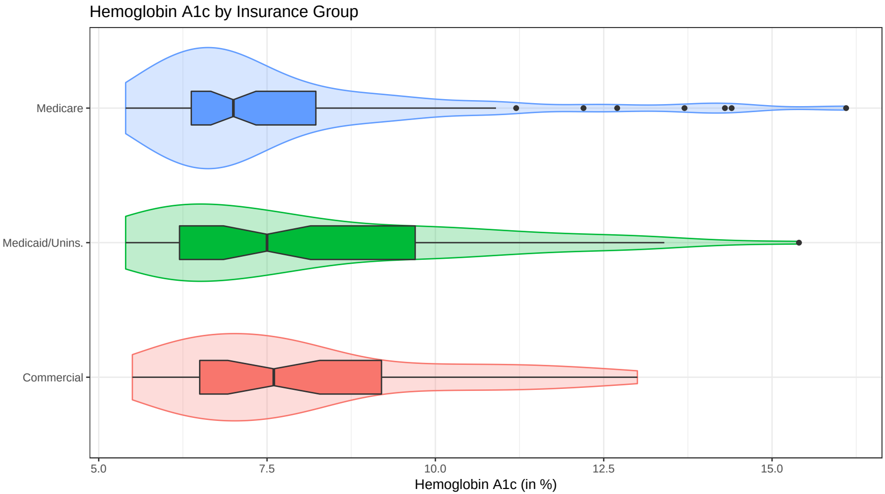
```
     pt.id                 insurance        a1c        
 Min.   :  1.00   commercial:39   Min.   : 5.400  
 1st Qu.: 48.75   medicaid  :67   1st Qu.: 6.300  
 Median : 96.50   medicare  :80   Median : 7.300  
 Mean   : 96.50   uninsured : 6   Mean   : 7.973  
 3rd Qu.:144.25                   3rd Qu.: 9.000  
 Max.   :192.00                   Max.   :16.100  
                                  NA's   :4       
```

- For now, we'll collapse the 6 uninsured in with the Medicaid patients, and we'll drop the four cases without an A1c value.

Collapse Medicaid and Uninsured, drop missing a1c

```
dm_ins <- dm_ins %>%
    mutate(insur_3 = fct_recode(insurance,
            "Commercial" = "commercial",
            "Medicare" = "medicare",
            "Medicaid/Unins." = "medicaid",
            "Medicaid/Unins." = "uninsured")) %>%
  filter(complete.cases(insur_3, a1c))
```

Hemoglobin A1c by Insurance Group



Hemoglobin A1c by Insurance Group

```
mosaic::favstats(a1c ~ insur_3, data = dm_ins)
```

```
          insur_3 min    Q1 median    Q3  max     mean
1      Commercial 5.5 6.500    7.6 9.200 13.0 8.100000
2 Medicaid/Unins. 5.4 6.200    7.5 9.700 15.4 8.121918
3        Medicare 5.4 6.375    7.0 8.225 16.1 7.764474
        sd  n missing
1 2.033276 39       0
2 2.350369 73       0
3 2.264962 76       0
```

One-Way ANOVA for the dm_ins Data

If we have a grouping factor (insurance) with *3* levels, then we are testing:

- H_0: $\mu_{Comm.} = \mu_{Medicare} = \mu_{Medicaid/Unins.}$ vs.
- H_A: At least one of the population means is different from the others.

```
anova(lm(a1c ~ insur_3, data = dm_ins))
```

```
Analysis of Variance Table

Response: a1c
           Df Sum Sq Mean Sq F value Pr(>F)
insur_3     2   5.55  2.7763  0.5466 0.5798
Residuals 185 939.60  5.0789
```

Elements of the ANOVA Table

The ANOVA table breaks down the variation in the outcome explained by the k levels of the factor of interest, and the variation in the outcome which remains (the Residual, or Error).

```
Analysis of Variance Table

Response: a1c
           Df Sum Sq Mean Sq F value Pr(>F)
insur_3     2   5.55  2.7763  0.5466 0.5798
Residuals 185 939.60  5.0789
```

- Df = degrees of freedom, Sum Sq = Sum of Squares,
- Mean Sq = Mean Square (Sum of Squares / df)
- F value = F test statistic, Pr(>F) = *p* value

The Degrees of Freedom

```
           Df
insur_3     2
Residuals 185
```

- The **degrees of freedom** attributable to the factor of interest (here, `insur_3`) is the number of levels of the factor minus 1.
    - Here, we have three insurance category levels, so df(`insur_3`) = 2.
- The total degrees of freedom are the number of observations (across all levels of the factor) minus 1.
    - We have 188 patients left in our `dm_ins` study after removing the four with missing A1c, so df(Total) = 187, although the Total row isn't shown here.
- Residual df = Total df - Factor df = 187 - 2 = 185.

The Sums of Squares

```
          Df Sum Sq
insur_3    2   5.55
Residuals 185 939.60
```

- The **sum of squares** (SS) represents variation explained.
- SS(Factor) is the sum across all levels of the factor of the sample size for the level multiplied by the squared difference between the level mean and the overall mean across all levels. SS(insur_3) = 5.55
- SS(Total) = sum across all observations of the square of the difference between the individual values and the overall mean.
  - Here SS(Total) = 5.55 + 939.60 = 945.15
- Residual SS = Total SS / Factor SS.

η^2, the Proportion of Variation Explained by ANOVA

```
          Df Sum Sq
insur_3     2   5.55
Residuals 185 939.60
```

- η^2 ("eta-squared") is equivalent to R^2 in a linear model.
    - η^2 = SS(Factor) / SS(Total) = the proportion of variation in our outcome (here, hemoglobin A1c) explained by the variation between levels of our factor (here, our three insurance groups)
    - In our case, $\eta^2 = 5.55 / (5.55 + 939.60) = 5.55 / 945.15 = 0.0059$
- So, insurance group accounts for about 0.59% of the variation in hemoglobin A1c observed in these data.

The Mean Square

```
           Df Sum Sq Mean Sq
insur_3     2   5.55  2.7763
Residuals 185 939.60  5.0789
```

- The Mean Square is the Sum of Squares divided by the degrees of freedom, so MS(Factor) = SS(Factor)/df(Factor).
- MS(insur_3) = SS(insur_3)/df(insur_3) = 5.55 / 2 = 2.78.
- MS(Residuals) = SS(Residuals) / df(Residuals) = 939.60 / 185 = 5.08.
  - MS(Residuals) estimates the residual variance, corresponds to σ^2 in the underlying linear model
  - MS(Residuals) = 5.0789, so Residual standard error = $\sqrt{5.0789}$ = 2.25 percentage points.

The F Test Statistic and p Value

```
Analysis of Variance Table

Response: a1c
           Df Sum Sq Mean Sq F value Pr(>F)
insur_3     2   5.55  2.7763  0.5466 0.5798
Residuals 185 939.60  5.0789
```

- F value = MS(insur_3) / MS(Residuals) = 2.78 / 5.08 = 0.55
- For an F distribution with 2 and 185 degrees of freedom, this F value yields $p = 0.58$

What is our conclusion regarding our test of our ANOVA hypotheses?

- H_0: $\mu_{Commercial} = \mu_{MedicaidorUninsured} = \mu_{Medicare}$ vs.
- H_A: H_0 is not true

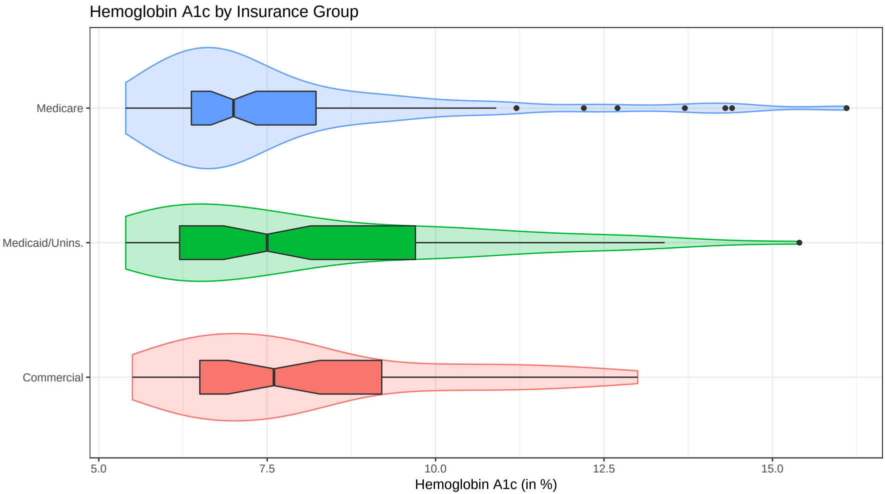
ANOVA Assumptions

The assumptions behind analysis of variance are the same as those behind a linear model. Of specific interest are:

- The samples obtained from each group are independent.
- Ideally, the samples from each group are a random sample from the population described by that group.
- In the population, the variance of the outcome in each group is equal. (This is less of an issue if our study involves a balanced design.)
- In the population, we have Normal distributions of the outcome in each group.

Happily, the F test is fairly robust to violations of the Normality assumption.

Can we assume population A1c levels are Normal?



Non-Parametric Alternative: Kruskal-Wallis Test

```r
kruskal.test(a1c ~ insur_3, data = dm_ins)
```

```
	Kruskal-Wallis rank sum test

data:  a1c by insur_3
Kruskal-Wallis chi-squared = 1.7809, df = 2,
p-value = 0.4105
```

Rank Sum test for

- H_0: Center of Commercial distribution = Center of Medicaid or Uninsured distribution = Center of Medicare distribution vs.
- H_A: H_0 not true.

Another Way to get our ANOVA Results

H_0: H_0: $\mu_{Commercial} = \mu_{MedicaidorUninsured} = \mu_{Medicare}$ vs. H_A: H_0 not true.

```r
summary(aov(a1c ~ insur_3, data = dm_ins))
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
insur_3       2    5.6   2.776   0.547   0.58
Residuals   185  939.6   5.079
```

Indicator Variable Regression

Regression on Indicator Variables = Analysis of Variance

Yet another way to obtain an even more complete analog to the pooled t test is to run a linear regression model to predict the outcome (here, a1c) on the basis of the categorical factor, insurance group. We run the following …

```
summary(lm(a1c ~ insur_3, data = dm_ins))
```

Linear Model Summary Output

```
Call: lm(formula = a1c ~ insur_3, data = dm_ins)

Residuals:     Min      1Q  Median      3Q     Max
           -2.7219 -1.6000 -0.6432  1.0855  8.3355

Coefficients:          Estimate Std. Err. t value Pr(>|t|)
(Intercept)             8.10000   0.3609  22.446  <2e-16 ***
insur_3Medicaid/Unins.  0.02192   0.4470   0.049   0.961
insur_3Medicare        -0.33553   0.4439  -0.756   0.451
---
Sig. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.254 on 185 degrees of freedom
Multiple R-squared: 0.0059, Adjusted R-squared:  -0.0049
F-statistic: 0.5466 on 2 and 185 DF,  p-value: 0.5798
```

Indicator Variable Regression

The linear model uses two **indicator variables**, sometimes called **dummy variables**.

- Each takes on the value 1 when its condition is met, and 0 otherwise.
- With three insurance categories, we need two indicator variables (we always need one fewer indicator than we have levels of the factor).
- Here, we have a baseline category (which is taken to be `Commercial` in this case) and then indicators for `Medicaid or Uninsured` and for `Medicare`.

K-1 indicators specify K categories

These two indicator variables completely specify the insurance category for any subject, as follows:

Insurance Category	var1	var2
Commercial	0	0
Medicaid/Unins.	1	0
Medicare	0	1

- `var1` is `insur_3Medicaid/Unins.`
- `var2` is `insur_3Medicare`

The Regression Equation

What is the regression equation here?

```
Call: lm(formula = a1c ~ insur_3, data = dm_ins)

Coefficients:            Estimate Std. Err. t value Pr(>|t|)
(Intercept)               8.10000   0.3609  22.446  <2e-16 ***
insur_3Medicaid/Unins.    0.02192   0.4470   0.049   0.961
insur_3Medicare          -0.33553   0.4439  -0.756   0.451
```

Equation specifies the three sample means

- A1c = 8.1 + 0.02 [Medicaid or Uninsured] - 0.34 [Medicare]
- [group] is 1 if the patient is in that group, and 0 otherwise

The Model predictions are Sample Means

```
Coefficients:             Estimate Std. Err. t value Pr(>|t|)
(Intercept)                8.10000   0.3609  22.446  <2e-16 ***
insur_3Medicaid/Unins.     0.02192   0.4470   0.049   0.961
insur_3Medicare           -0.33553   0.4439  -0.756   0.451
```

Model Predictions:

- A1c = 8.1 if in the Commercial group
- A1c = 8.1 + 0.02192 = 8.12 if in the Medicaid or Uninsured group
- A1c = 8.1 - 0.33553 = 7.76 if in the Medicare group

K-Sample Study Design, Comparing Means

1. What is the outcome under study?
2. What are the (in this case, K $>$ 2) treatment/exposure groups?
3. Were the data in fact collected using independent samples?
4. Are the data random samples from the population(s) of interest? Or is there at least a reasonable argument for generalizing from the samples to the population(s)?
5. What is the significance level (or, the confidence level) we require here?
6. Are we doing one-sided or two-sided testing?
7. What does the distribution of each individual sample tell us about which inferential procedure to use?
8. Are there statistically meaningful differences between population means?
9. If an overall test is significant, can we identify pairwise comparisons of means that show significant differences using an appropriate procedure that protects against Type I error expansion due to multiple comparisons?

Coming up . . .

What happens if the ANOVA gives a statistically significant result? How do we then decide where the action is?

- ANOVA and the problem of multiple comparisons
- Building Models with Regression