Week Fourteen

Last Week

- Class Recap
- Introduction to Correlated Data

This Week: More Correlated Data

- Tuesday: Activity
- Thursday: Lab

Next Week: Fall Break

Next-Next Week: Classification Algorithms + Final Exam Review

- Tuesday:
  - Video Lecture
  - In Class Lab
- Thursday:
  - Final Exam Review
  - Take Home Final Assigned

McNemar's Test

Recall McNemar's Test where we have repeated measurements on a sampling unit.

We previously framed this in the context of asking a respondent two related policy questions to evaluate whether one would be more amenable.

Another scenario would be where multiple treatments are given to a single patient. For example, we might compare a medical treatment with a placebo.

```
outcomes <- matrix(c(40, 6, 8, 46), nrow = 2,
                dimnames = list("Treatment" = c("Better", "Not Better"),
                                "Placebo" = c("Better", "Not Better")))
outcomes
```

```
            Placebo
Treatment    Better Not Better
  Better         40          8
  Not Better      6         46
```

```
mcnemar.test(outcomes)
```

```
    McNemar's Chi-squared test with continuity correction

data:  outcomes
McNemar's chi-squared = 0.071429, df = 1, p-value = 0.7893
```

Q: What are the null hypothesis and alternative hypothesis in this situation? How do and outcome of the test in this situation?

Analyzing Rater Agreement

In a similar setting, we can assess similarities of raters (or tests).

- Cohen's Kappa Measure of Agreement

$$\kappa \frac{\sum \pi_{ii} - \sum \pi_{i+}\pi_{+i}}{1 - \sum \pi_{i+}\pi_{+i}}$$

κ is 0 expected under independence, κ is 1 perfect agreement.

```
outcomes <- matrix(c(40, 6, 8, 46), nrow = 2,
               dimnames = list("Test 1" = c("Sick", "Not Sick"),
                               "Test 2" = c("Sick", "Not Sick")))
outcomes
```

```
          Test 2
Test 1     Sick Not Sick
  Sick       40        8
  Not Sick    6       46
```

```
CohenKappa(outcomes)
```

```
[1] 0.7191011
```

```
statement <- data.frame(
  A=c(2,3,1,3,1,2,1,2,3,3,3,3,3,2,1,3,3,2,2,1,
      2,1,3,3,2,2,1,2,1,1,2,3,3,3,3,3,1,2,1,1),
  B=c(2,2,2,1,1,2,1,2,3,3,2,3,1,3,1,1,3,2,1,2,
      2,1,3,2,2,2,3,2,1,1,2,2,3,3,3,3,2,2,2,3),
  C=c(2,2,2,1,1,2,1,2,3,3,2,3,3,3,3,2,2,2,2,3,
      2,2,3,3,2,2,3,2,2,2,2,3,3,3,3,3,3,2,2,2),
  D=c(2,2,2,1,1,2,1,2,3,3,2,3,3,3,3,3,2,2,2,2,
      3,1,3,2,2,2,1,2,2,1,2,3,3,3,3,3,3,2,2,1)
)

KappaM(statement)
```

```
[1] 0.5036937
```

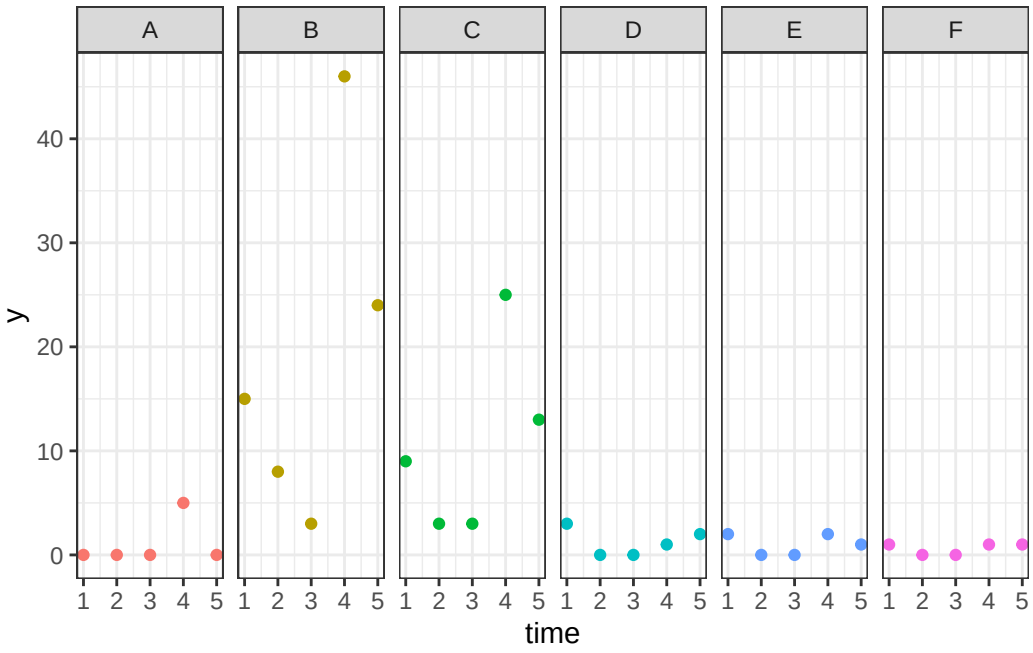
Q: Which two raters are most in agreement?

Modeling Correlated Data: GLMMs

Modeling Correlated Data:

- longitudinal studies or matched sets (case - control or family clusters)

Recall the bike rental data, now let's assume we have counts at different bike rental stations.



How might we model this data?

$y_{it} \sim Poisson(\mu_{it})$

$log(\mu_{it}) = \alpha_i + \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 + \gamma_5$

fixed effects

$y_{it} \sim Poisson(\mu_{it})$

$log(\mu_{it}) = \alpha_i + \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 + \gamma_5$

Maybe if there are a few stations in the dataset, we use a categorical variable in the modeling framework. In general there may be specific interest in those stations.

random effects

An alternative approach is to use random effects.

- If there are a large number of stations in the dataset and we are less interested in the stations themselves, we can use a random effect model.
- Instead of directly estimating each we use a categorical variable in the modeling framework. In general there may be specific interest in those stations.

$y_{it} \sim Poisson(\mu_{it})$

$log(\mu_{it}) = \alpha_i + \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 + \gamma_5$

$\alpha_i \sim N(\alpha, \sigma_\alpha^2)$ and

$\alpha_i = \alpha + u_i$, $u_i \sim N(0, \sigma_\alpha^2)$

- We can use `glmer` in `lme4` or `stan_glmer` in `rstanarm`. We will use the notation `(1|variable)` to denote a random effect on the variable.

GLMMs for Categorical Data

1. Logistic Regression

Write out this model and simulate data from this setting