Count data and nonparametric tests

Session 13

MATH 80667A: Experimental Design and Statistical Methods
HEC Montréal

Outline

Count data

Nonparametric tests

Count data

Tabular data

Aggregating binary responses gives counts.

Duke and Amir (2023) investigated the impact on sales of presenting customers with

- a sequential choice (first decide whether or not to buy, then pick quantity) versus
- an integrated decision (choose not to buy, or one of different quantities).

	quantity-integrated	**quantity-sequential**
did not purchase	100	133
purchased	66	26

Question: does the selling format increases sales?

Pearson chi-square test

Consider an $I \times J$ **contingency table**.

Denote the observed counts in the (i, j)th cell O_{ij}.

We compare these with expected counts under the null hypothesis, E_{ij}'s.

The test statistic is

$$P = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

Yate's correction for 2×2 tables involves subtract $1/2$ from the differences $O_{ij} - E_{ij}$.

Null distribution for Pearson chi-square test

In large samples (if $\min_{i,j} E_{ij} > 5$), the statistic behaves like a chi-square distribution with ν degrees of freedom, denoted $P \dot{\sim} \chi^2_\nu$.

The degrees of freedom are the difference between the number of cells $N = IJ$ and the number of parameters under $\mathscr{H}_0$.

```r
data(DA23_E2, package = "hecedsm")
tabs <- with(DA23_E2, table(purchased, format))
# Chi-square test for independence
chisq <- chisq.test(tabs)
```

The test statistic is 21.92, with 1 degree of freedom. The p-value is less than 10^{-4}, so strong evidence that there are differences between selling format.

Effect size

Effect sizes for contingency tables range from 0 (no association) to 1 (perfect association).

Measures include

- ϕ for 2×2 contingency tables, $\phi = \sqrt{P/n}$.
- Cramér's V, which is a renormalization, $V = \phi/\sqrt{\min(I-1, J-1)}$.

Small sample (bias) corrections are often employed.

We obtain $V = 0.2541$, a moderate effect size.

Poisson regression models

We cannot use ANOVA for counts, but analysis of deviance is similar.

Assume $Y_{ij} \sim \mathsf{Poisson}(\mu_{ij})$ where the mean is nonnegative.

For example, the main-effect model is of the form

$$\ln \mu_{ij} = \underbrace{\mu}_{\substack{\text{global}\\\text{mean}}} + \underbrace{\alpha_i}_{\substack{\text{row}\\\text{effect}}} + \underbrace{\beta_j}_{\substack{\text{column}\\\text{effect}}}, \qquad i = 1, \ldots, I; j = 1, \ldots, J$$

with sum-to-zero constraints for α_i, β_j.

Remarks

- Compared to linear regression and ANOVA, the variance of the cells is solely determined by the mean counts
- Each dimension of the contingency table (row, column, depth) is a factor
- Each cell is a response value. There are as many observations, $N = IJ$, as cells.

Tests for Poisson regression models

We can use a likelihood ratio test or score test (aka Pearson χ^2 statistic!)

We compare two nested models:

- typically, the alternative model is the **satured model**, which has as many averages as cells (model with an interaction) and for which the averages are given by observed counts, $\widehat{\mu}_{ij} = o_{ij}$.
- the null model, a simplification with k parameters
- large-sample distribution of tests is χ^2_ν, with $\nu = N - k$ degrees of freedom the difference in the number of parameters between alternative and null model.

Example 2 - frequency of elocution

We consider Elliot et al. (2021) multi-lab replication study on spontaneous verbalization of children when asked to identify pictures of objects.

```
data(MULTI21_D1, package = "hecedsm")
contingency <- xtabs( #pool data
  count ~ age + frequency,
  data = MULTI21_D1)
# No correction to get same result as Poisson regression model
(chisqtest <- chisq.test(contingency, correct = FALSE))
```

```
##
##      Pearson's Chi-squared test
##
## data:  contingency
## X-squared = 87.467, df = 6, p-value < 2.2e-16
```

Poisson regression analog

```
MULTI21_D1_long <- MULTI21_D1 |> # pool data by age freq
  dplyr::group_by(age, frequency) |>  # I=4 age group, J=3 freq
  dplyr::summarize(total = sum(count)) # aggregate counts
mod_main <- glm(total ~ age + frequency, # null model, no interaction
    family = poisson, data = MULTI21_D1_long)
mod_satur <- glm(total ~ age * frequency, # saturated model
    family = poisson, data = MULTI21_D1_long)
# Likelihood ratio test and Pearson chi-square
anova(mod_main, mod_satur, test = "LRT")  # deviance
anova(mod_main, mod_satur, test = "Rao")  # score test
```

The null model is the **main effect model** (no interaction, "independence between factors"). There are $(I-1) \times (J-1)$ interaction terms (6 degrees of freedom) for the tests.

Example 3 - racial discrimination

We consider a study from Bertrand and Mullainathan (2004), who study racial discrimination in hiring based on the consonance of applicants names.

The authors created curriculum vitae for four applicants and randomly allocated them a name, either one typical of a white person or a black person.

The response is a count indicating how many of the applicants were called back (out of 4 profiles: 2 black and 2 white), depending on their origin.

Testing symmetry

Under the null hypothesis of **symmetry**, the off-diagonal entries of the table have equal frequency.

- The expected counts E are the average of two cells $E_{ij} = (O_{ij} + O_{ji})/2$ for $i \neq j$.

black	white	sym	O	E
0	0	0:0	1103	1103.0
1	0	0:1	33	53.5
2	0	0:2	6	12.5
0	1	0:1	74	53.5
1	1	1:1	46	46.0
2	1	1:2	7	12.5
0	2	0:2	19	12.5
1	2	1:2	18	12.5
2	2	2:2	17	17.0

Fitting Poisson models

- Null model: Poisson model with `sym` as factor
- Alternative model: saturated model (observed counts)

```
data(BM04_T2, package = "hecedsm")
# Symmetric model with 6 parameters (3 diag + 3 upper triangular)
mod_null <- glm(count ~ gnm::Symm(black, white),
                data = BM04_T2,
                family = poisson)
# Compare the two nested models using a likelihood ratio test
pchisq(deviance(mod_null), lower.tail = FALSE,
       df = mod_null$df.residual) # 9 cells - 6 parameters = 3
PearsonX2 <- sum(residuals(mod_null, type = "pearson")^2)
pchisq(PearsonX2, df = mod_null$df.residual, lower.tail = FALSE)
```

Nonparametric tests

Why nonparametric tests?

Nonparametric tests refer to procedures which make no assumption about the nature of the data (e.g., normality)

Rather than considering numeric response $Y_{(1)} \leq \cdots \leq Y_{(n)}$, we substitute them with ranks $1, \ldots, n$ (assuming no ties), where

$$R_i = \text{rank}(Y_i) = \#\{j : Y_i \geq Y_j, j = 1, \ldots, n\}$$

- e.g., numbers $(8, 2, 1, 2)$ have (average) ranks $(4, 2.5, 1, 2.5)$

Understanding rank-based procedures

Many tests could be interpreted (roughly) as linear/regression or ANOVA

- but with the values of the rank R_i rather than that of the response Y_i

Ranks are not affected by outliers (more robust)

- this is useful for continuous data, less for Likert scales (lots of ties, bounded scales)

Wilcoxon's signed rank test

For paired data with differences $D_i = Y_{i2} - Y_{i1}$, we wish to know if the average rank is zero.

- remove zero differences
- rank absolute values $R_i = \mathrm{rank}(|D_i|)$ of the remaining observations
- compute the test statistic $T = \sum_{i=1}^n \mathrm{sign}(D_i) R_i$
- compare with reference under hypothesis of symmetry of the distribution.

The latter is analogous to a one-sample t-test for $\mu_D = 0$.

Kruskal–Wallis test

Roughly speaking

- rank observations of the pooled sample (abstracting from K group labels)
- compare average ranks in each group.
- compare with reference

For $K = 2$, the test is called Mann–Whitney–Wilcoxon or Mann–Whitney U or Wilcoxon rank sum test.

Analogous to running two-sample t-test or one-way ANOVA with ranks.

Null distributions and benchmarks

Since ranks are discrete (assuming no ties), we can derive explicit expression for values that the statistic can take in small samples.

- Zero differences and ties mess up things.
- With more than 15 observations by group, large-sample approximations (normal, Student-t or F distribution) from linear regression/ANOVA are valid.

Example 1 - Virtual communications

Brucks and Levav (2022) measure the attention of participants during exchanges using an eyetracker in

- face-to-face meetings
- videoconference meetings

Data suggests that videoconferencing translates into longer time spent gazing at the partner than in-person meetings.

Code for Wilcoxon rank-sum test

The `coin` package function reports Hodges–Lehmann estimate of location.

Intervals and estimates of difference in mean are in seconds (-37 seconds).

```
data(BL22_E, package = "hecedsm")
(mww <- coin::wilcox_test( # rank-based test
  partner_time ~ cond,
  data = BL22_E,
  conf.int = TRUE)) # values and intervals are times in seconds
welch <- t.test(partner_time ~ cond,
  data = BL22_E, # compare results with two sample t-test
  conf.int = TRUE)
```

Example 2 - Smartwatches distractions

We consider a within-subject design from Tech3Lab (Brodeur et al., 2021).

Each of the 31 participants was assigned to four distractions while using a driving simulator

- phone
- using a speaker
- texting while driving
- smartwatch

Task order was randomized and data are balanced

The response is the number of road safety violations conducted on the segment.

Friedman and Quade tests

We use Quade's test, which ranks responses of each participants $1, 2, 3, 4$ separately.

```
data(BRLS21_T3, package = "hecedsm")
coin::friedman_test(nviolation ~ task | id,
                    data = BRLS21_T3)
```

```
##
##      Asymptotic Friedman Test
##
## data:  nviolation by
##       task (phone, watch, speaker, texting)
##       stratified by id
## chi-squared = 18.97, df = 3, p-value = 0.0002774
```

Pairwise differences

Since there are overall differences, we can follow-up by looking at all pairwise differences using Wilcoxon rank-sum test

```
# Transform to wide format - one line per id
smartwatch <- tidyr::pivot_wider(
  data = BRLS21_T3,
  names_from = task,
  values_from = nviolation)
# Wilcoxon signed-rank test
coin::wilcoxsign_test(phone ~ watch,
                      data = smartwatch)
```

There are $\binom{4}{2} = 6$ pairwise comparisons, so we should adjust p-values for multiple testing using, e.g., Holm–Bonferroni.