Stat 471/571: Analysis of non-normal data

This course, and Stat 301/587 before it, have focused on the analysis of normally distributed data. That is because of prevalence of such data in many fields. But, some data are not normally distributed, and that data can be very common in some fields. We discuss the analysis of two common types of non-normal data: yes/no/proportion data and counts.

We will see that most principles carry over to non-normal data. The major changes are in the names of the methods and the details of the analysis.

Yes/No/Proportion data

Example study: roadway construction. Two different types of concrete. Study done in Iowa. Experimental unit is a 10m stretch of roadway. There are 20 eu's: 10 with type A and 10 with type B concrete. After 2 years of use, the concrete is assessed at 10 locations over the 10m stretch. Each location is scored acceptable or not. The ou is the location: the data are 1 = acceptable and 0 = not.

Dealing with the subsampling:

- locations are subsamples.
- before: average subsamples to get an eu average
- now: average is the proportion (or percent) acceptable

Comparing the two types:

- before: do a t-test, perhaps after transforming the data
  - Assumes independence, equal variance, normally distributed errors
  - proportions have unequal variance and may be far from normally distributed
- now: use methods designed for yes/no or proportion data

Properties of proportion data:

- # acceptable commonly assumed to have a Binomial distribution: $Y_i \sim Binomial(N_i, \pi)$
- for eu i, $p = \hat{\pi} =$ #acceptable/#examined $= Y_i/N_i$
- p is unbiased: theoretical average of $p = \pi$
- Variance of $p = \pi(1-\pi)/N$
- estimated by Var $p = p(1-p)/N$, so se of $p = \sqrt{p(1-p)/N}$

- non-constant!

A simple analysis:

- Ignore any differences between the 10 replicate eu's of the same type
- Have 100 locations for type A, 100 locations for type B
- Data is 68 acceptable for type A, 88 for type B
- Compute proportion for type A = 0.68, proportion for type B=0.88
- se's are: $\sqrt{0.68(1-0.68)/100} = 0.0466$ and $\sqrt{0.88(1-0.88)/100} = 0.0324$
- 95% confidence interval for each is $p \pm 1.96 \times se$: (0.59, 0.77) and (0.82, 0.94)
- What, no df? 1.96 is the 0.975 quantile of a normal distribution (∞ df)
- Do you have to estimate a sd (i.e. between observations)?
  - no - variance and se are determined by p
- The two samples are independent so variance of diff = Variance for A + Variance for B
  - $= 0.0466^2 + 0.0324^2 = 0.00323$, so se $= \sqrt{0.00323} = 0.0568$
- 95% ci for difference is $(0.88 - 0.68) \pm 1.96 \times 0.0568 = (0.088, 0.31)$
- Does not include zero, so $p < 0.05$

Test of equal proportions:

- Arrange the data as a contingency table: separately count acceptable and not

Type	Acceptable	Not	Total
B	88	12	100
A	68	32	100
total	156	44	200

- Usual test is Chi-square test:

$$C = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum \frac{(Y_{ij} - \hat{Y}_{ij})^2}{\hat{Y}_{ij}}$$

- Null hypothesis is that the two proportions are the same
- E_{ij} is the expected count when that null hypothesis is true
  - Compute the overall probability of acceptable $= (68 + 88)/(100 + 100) = 0.78$

  - Multiple that probability by each sample size $= 0.78 \times 100 = 78$ for both groups
  - Compute SS for each "cell" of the contingency table
  - Weight by expected count (more later)
  - Sum contributions from all four cells
- Connection to ideas in this course
  - F tests are a model comparison, based on SSE for two models
    - SSE quantifies how well a model fits the data
    - F test based on change SSE between two models
  - χ^2 tests are also a model comparison
    - Using a more appropriate (for proportion data) measures of how well a model fits the data
  - For χ^2 test: measure of fit is C, computed using expected counts given by the model
  - The two models are:
  - Null:
    - An additive effect (row effects + column effects) fits the data.
    - Gives the usual expected counts
    - fit $= C$ computed in the usual way
  - Alternative:
    - Need an interaction (row + column + interaction).
    - Expected counts are the observed values
    - fit $= 0$ (perfect fit)
  - The change in fit is $C - 0 = C$, i.e. the usual test statistic
- Just fine for simple studies, 2 groups or k groups with no structure
- Does not generalize easily to more complicated designs and not at all to regression

Extensions:

- Role of independence assumption
  - Key assumption in Chi-square test is independent observations
  - Here, assumes each of the 10 eu's for one type has same probability of acceptable
  - I.e., no variability between the 10 eu's
  - We'll see approaches to handle this soon
- Paired data:
  - Another violation of independence: pairs intended to be different from each other

- So we did a paired t-test for continuous responses
- Can't ignore pairing just because the data are proportions
- The simple test to account for pairing is McNemar's test

- First step towards more useful models
  - Difference between the two types is 0.20 = 20%.
  - Estimate for IA is that using type B concrete increases acceptability after 2 years by 20%.
  - Apply this to Arizona. Current use of type A concrete is 92% acceptable after 2 years.
  - So predict using type B will increase acceptability to 112% ???
  - Differences of proportions, percents, don't "transport" well to other baseline values
  - Because of 0%, 100% bounds

Odds and odds ratios:

- Goal: comparisons between proportions that apply more generally than differences
- Odds:

$$\begin{aligned} \text{population: } O &= \frac{\pi}{1-\pi} \\ \text{sample: } O &= \frac{p}{1-p} \\ \text{for sample percent: } O &= \frac{p}{100-p} \end{aligned}$$

  - Odds = ratio of P[event] / P[not event]
  - Odds of 2 means event twice as probable as "not event"
  - Converting Odds to probability

$$p = O/(1+O)$$

  - e.g., odds ratio of 2 $\Rightarrow p = 2/(1+2) = 2/3$
  - Betting, e.g. horse racing, X is the 3:1 favorite $\Rightarrow$ odds ratio = 1/3, P[win] = 0.25
    - Bet 1$, get 3$ if X wins: Expected return $= 0.25 \times 3 + 0.75 \times (-1) = 0$
  - Concrete data: Odds of acceptable for type B = 88/12 = 7.33, for type A = 2.12

- Odds ratio, 2nd way to compare two probabilities:
  - Odds in group 1 / odds in group 2
  - Value from 0 to ∞
  - Odds ratio = 1 $\Leftrightarrow$ equal probabilities in the two groups
  - Easy to compute from the contingency table counts

$$\hat{O} = \frac{n_{11}\, n_{22}}{n_{12}\, n_{21}}$$

  - Concrete data: estimated odds ratio = $\frac{88\,32}{68\,12} = 3.45$
  - Odds of acceptable type B is 3.45 times odds for type A
- Relative risk, 3rd way to compare two probabilities:
  - RR = P[event in group 1] / P[event in group 2]
  - Often easier to interpret
  - But doesn't work well for large probabilities (e.g., > ca 0.1 or 0.2)
    * e.g.: $p_1 = 0.95$, $p_2 = 0.9$, RR = 0.95/0.90 = 1.05.
    * But RR of the "non-event" = 0.05 / 0.10 = 0.5
  - Odds ratio is symmetric: $OR_e = 1/OR_{ne}$
  - When event rare (< 0.10 and especially < 0.01, RR $\approx$ OR
- Log odds ratio:
$$\log \frac{O_1}{O_2}$$
  - Value from $-\infty$ to ∞
  - Log OR = 0 $\Leftrightarrow$ equal probabilities in the two groups

Inference on odds ratios:

- In large samples, log OR normally distributed
- With se estimated by
$$\text{se } \log OR = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$
- Use for tests, usually $\log OR = 0$, and confidence intervals
- With quantiles of a normal distribution
- Concrete data:
  - Var $\log OR$ = 1/88 + 1/12 + 1/68 + 1/32 = 0.1407, se = $\sqrt{0.1407} = 0.375$
  - 95% ci for $\log OR = \log \hat{OR} \pm 1.96$ se $\log OR = \log 3.45 \pm 0.735 = (0.50,\ 1.97)$
  - Get ci for OR by exponentiating: = (1.65, 7.20)
  - Test of $\log OR = 0$: $(\log 3.45 - 0)/0.375 = 3.30$, p = 0.0010

Odds ratios are more easily transportable to different baseline values

- If OR = 3.45, what is the expected P[acceptable] for type B concrete in Arizona?
- Baseline P[acceptable], i.e. type A = 0.92, odds = 11.5, log odds = 2.44

- Predicted odds for type B $= 3.45 \times 11.5 = 39.68$, predicted log odds $= 2.44 + \log 3.45 = 3.68$
- Predicted probability $= 39.68/(1+39.68) = 0.975$
- Can calculate a confidence interval on this from the ci for $\log OR$ (and if baseline is estimated, also the se of the baseline log odds)