4/9/2010 36-402/608 ADA-II H. Seltman

Handout #20: Categorical Outcomes, part 1

1. The basics of comparing proportions and using odds ratios

   (a) When outcomes are yes/no, success/fail, live/die DO NOT use methods for Normal outcomes.

   (b) For two groups we have the canonical contingency table:

	true success prob.	successes	failures	total
group 1	p_1	y_1	$n_1 - y_1$	n_1
group 2	p_2	y_2	$n_2 - y_2$	n_2

   (c) If Y Binomial(n, p), then $E(Y) = np$ and the variance of Y is $np(1-p)$ rather than having a separate σ^2 value.

   (d) If $Y \sim$ Binomial(n, p) then $E(Y/n) = p$ and $\text{Var}(Y/n) = p(1-p)/n$.

   (e) Estimate of p: $\hat{p} = y/n$. Estimate of $\text{Var}(\hat{p}) = \hat{p}(1-\hat{p})/n$.

   (f) If $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$ then $\hat{p} = Y/n \approx N(\hat{p}, \hat{p}(1-\hat{p})/n)$, which leads to a CI for p and a Z-score based test for $H_0 : p = p_0$.

   (g) Probability differences for two independent groups have sampling variance equal to the sum of the individual sampling variances. Use $\hat{p}_{\text{diff}} \pm 1.96\text{SE}_{\text{diff}}$ for a 95% CI, and $Z = \hat{p}_{\text{diff}}/\text{SE}_{\text{diff}}$ for a test of $H_0 : p_1 = p_2$.

   (h) Often a better scale is odds(success)=prob(success)/prob(failure). Interpret as number of successes for each failure.

   (i) A probability difference of zero corresponds to an odds ratio of 1.

   (j)
   $$\text{Odds ratio} = \frac{\frac{y_1}{n_1}}{\frac{n_1-y_1}{n_1}} \Big/ \frac{\frac{y_2}{n_2}}{\frac{n_2-y_2}{n_2}} = \frac{y_1(n_2-y_2)}{y_2(n_1-y_1)}$$

   (k) $E(\text{crossproduct})$ = true odds ratio.

   (l) E.g. success rates are 0.50 vs. 0.45: prob. diff=0.05, OR=1.22. E.g. success rates are 0.07 vs. 0.02: prob. dif.=0.05 and OR = 3.7.

   (m) SE(log odds ratio) = $\sqrt{1/y_1 + 1/(n_1 - y_1) + 1/y_2 + 1/(n_2 - y_2)}$, and the sampling distribution of the log odds ratio is Normal. Odds ratio 95% CI = exp(log odds ratio +/- 1.96SE).

   (n) The p-value for Z=LOR/SE(LOR) is used as a test of H_0 : LOR = 0.

2. Other tests for independence of two categorical random variables

   (a) Chi-square test: compare observed to expected under independence.

$$X^2 = \sum_{i=1}^{\text{cell count}} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(R-1)(C-1)} \text{ (asymptotically)}$$

   (b) Fisher Exact test: Assuming fixed row and column margin counts, use the hypergeometric distribution. Correct for small numbers of counts. Often only available for 2x2 tables without large counts.

   (c) Mantel-Haensel test: Test odds ratio = 1 in K 2x2 table with a common odds ratio and possibly very different odds values across tables. Can be used to "correct" for a covariate. (First need to test for a common odds ratio.)