Biostatistics for Health Care Researchers: A Short Course

Analysis of Categorical Data

Presented by:

Siu L. Hui, Ph.D.

Department of Medicine, Division of Biostatistics

Indiana University School of Medicine

OBJECTIVES

- Know the basic principles, assumptions, and a few basic methods in analyzing categorical data
- Understand and interpret the results of categorical data analyses in the literature
- Know the assumptions needed for sample size estimation

OVERVIEW

Inference

	Estimation	(Hypothesis Testing) Comparison
Single Proportion	1	2
Two Proportions	3	2x2 tables

–Extensions

CATEGORICAL VARIABLES

- Each outcome can be classified at one of several levels, with no natural order to the levels,

  e.g.- blood type, race, treatment group

- Binary variables - Y/N - disease, exposure, response to treatment, M/F

A SINGLE SAMPLE

- In the POPULATION, the TRUE probability of an event is π.
- A single random sample proportion is p.
- Inference from $p \rightarrow \pi$
  - Estimating π
  - Hypothesis test -
    $H_0: \pi = \pi_0$ (pre-specified)

Examples of One-Sample Problems

- Unemployment rate in the US based on survey
  - *What is the true unemployment rate π ?*
  - *Is π greater than 10%?*
- National Immunization Survey by CDC (telephone)
  - *What percent of children are immunized in the US?*
  - *Is the immunization rate lower than reported by another country?*

ESTIMATION of π

- True probability π
- ***Point estimate*** of π: $p = x/n$,
  x = # positive outcomes
  n = total # in sample
- Need ***confidence interval*** (C.I.) for π
  [For large n, $p \sim N(\pi, \pi(1-\pi)/n)$]

CONFIDENCE INTERVAL for π

From Binomial,

$$\text{s.e. of } p = \sqrt{\frac{\pi(1-\pi)}{n}} \ ,$$

$$95\% \text{ C.I. for } \pi = p \pm 1.96 \text{ s.e.}$$

Example 1: Phase 2 Clinical Trial

A disease has no known treatment:
Spontaneous remission rate ≤ 0.4.
Experimental Rx given to 25 patients - 15 remissions

Estimated π (remission rate with Rx) :

$$p = \frac{15}{25} = 0.6$$

$$s.e. = \sqrt{\frac{0.6 \times 0.4}{25}} = 0.098$$

$$95\% \; C.I. = 0.6 \pm 1.96 \times 0.098 = (0.41, 0.79)$$

Example 1 (contd): Is the new treatment effective?

Hypothesis testing: Comparing to a given proportion

$$H_0 : \pi = 0.4 \qquad H_a : \pi > 0.4$$

$$Z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

$$= \frac{0.6 - 0.4}{\sqrt{\frac{0.4(1-0.4)}{25}}} = 2.04$$

$p<0.05$

One Sample Problem: Example 2

In whole population, pregnancy loss rate π=-.0095. A large consortium studied pregnancy loss in 3096 patients who underwent mid-trimester amniocentesis [Eddleman OB/GYN 2006].

Among 3096 with amniocentesis, 31 had spontaneous pregnancy loss < 24 weeks

$$\text{p}=\frac{31}{3096}=0.01$$

$$95\%\ \text{C.I.}=0.0100\pm1.96\sqrt{\frac{0.01(1-0.01)}{3096}}=(0.0065,0.0135)$$

Is it different from the population pregnancy loss rate?

$$H_0:\pi=\pi_0=0.0095 \qquad H_a:\pi>0.0095$$

$$z=\frac{0.0100-0.0095}{\sqrt{\frac{0.0095(1-0.0095)}{3096}}}=0.287 \qquad \text{p}>0.05$$

Conclusion: Midtrimester amniocentesis does not increase pregnancy loss.

OVERVIEW

Inference

	Estimation	(Hypothesis Testing) Comparison
Single Proportion	√	√
Two Proportions	3	2x2 tables

–Extensions

Examples of 2 Independent Proportions

- Prospective: Randomized controlled trial – subjects randomly assigned into 2 groups – compare their binary outcome (dead/alive)

- Retrospective: case-control study – compare the proportion of smokers among cancer cases and controls with no cancer

- Cross-sectional study: Comparing the prevalence of CHD between blacks and whites from a survey

DIFFERENCE of 2 INDEPENDENT PROPORTIONS

Estimation

$$\text{Estimate of } (\pi_2 - \pi_1) = p_2 - p_1$$

$$\text{s.e. } (p_2 - p_1) = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$$

$$95\% \text{ C.I.} = (p_2 - p_1) \pm 1.96 \text{ s.e.}$$

DIFFERENCE OF 2 INDEPENDENT PROPORTIONS

Example (Spiro JAMA 2006)

Acute Otitis Media (AOM): Randomized study to reduce antibiotics use

Standard Prescription vs. Wait-and-see prescription

Outcome: Proportion who did NOT fill the prescription

$$\text{Standard} \qquad p_1 = \frac{19}{145} = 0.13$$

$$\text{Wait-and-see} \qquad p_2 = \frac{87}{138} = 0.62$$

$$\text{Estimated difference} = p_2 - p_1 = 0.62 - 0.13 = 0.49$$

$$95\%\ \text{C.I.} = 0.49 \pm 1.96\sqrt{\frac{0.13(1\text{-}0.13)}{145} + \frac{0.62(1-0.62)}{138}}$$

$$= (0.39,\ 0.59)$$

AOM Example (contd): *Is the Intervention effective?*

$$H_0 : \pi_1 = \pi_2$$
$$H_a : \pi_1 \neq \pi_2$$

In AOM example,

Standard prescription: $p_1 = \dfrac{19}{145}$

Wait-and-see: $p_2 = \dfrac{87}{138}$

2 x 2 Tables

AOM Study

	Standard	Wait & See	Total
Filled Prescription	126	51	177
Did Not Fill Prescription	19	87	106
Total	145	138	283

$$H_0 : \pi_1 = \pi_2 \quad vs \quad H_0 : \pi_1 \neq \pi_2$$

Chi-Square Tests

In each cell, O ~ Observed #

E ~ Expected # under H_0

$$\text{Chi-square test statistic} = \sum \frac{(O - E)^2}{E}$$

Degree-of-freedom

df = (# rows – 1) x (# columns -1)

For 2x2 table (2-1) x (2-1) = 1df

COMPARISON BETWEEN 2 PROPORTIONS

	Standard	Wait & See	Total
Filled Prescription	126(90.7)	51(86.3)	177
Did Not Fill Prescription	19(54.3)	87(51.7)	106
Total	145	138	283

Chi-Square Tests – contd.

AOM example

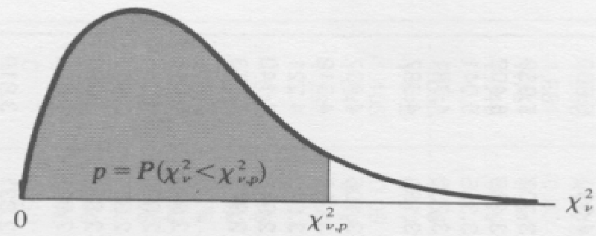
$$\text{Test Statistic} = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(126 - 90.7)^2}{90.7} + \frac{(51 - 86.3)^2}{86.3} + \frac{(19 - 54.3)^2}{54.3} + \frac{(87 - 51.7)^2}{51.7}$$

$$= 75.2$$

$$\sim \chi^2 \text{ with 1 df}$$

TABLE A-3 *Percentiles of the chi-square distribution*



(b) χ^2 *distribution*

df \ %	0.5	1	2.5	5	10	20	30	40	50	60	70	80	90	95	97.5	99	99.5	99.95
1	0.0001	0.0002	0.001	0.004	0.016	0.064	0.148	0.275	0.455	0.708	1.074	1.642	2.706	3.841	5.024	6.635	7.879	12.116
2	0.010	0.020	0.051	0.103	0.211	0.446	0.713	1.022	1.386	1.833	2.408	3.219	4.605	5.991	7.378	9.210	10.597	15.202
3	0.072	0.115	0.216	0.352	0.584	1.005	1.424	1.869	2.366	2.946	3.665	4.642	6.251	7.815	9.348	11.345	12.838	17.730
4	0.207	0.297	0.484	0.711	1.064	1.649	2.195	2.753	3.357	4.045	4.878	5.989	7.779	9.488	11.143	13.277	14.860	19.997
5	0.412	0.554	0.831	1.145	1.610	2.343	3.000	3.655	4.351	5.132	6.064	7.289	9.236	11.070	12.833	15.086	16.750	22.105
6	0.676	0.872	1.237	1.635	2.204	3.070	3.828	4.570	5.348	6.211	7.231	8.558	10.645	12.592	14.449	16.812	18.548	24.103
7	0.989	1.239	1.690	2.167	2.833	3.822	4.671	5.493	6.346	7.283	8.383	9.803	12.017	14.067	16.013	18.475	20.278	26.018
8	1.344	1.646	2.180	2.733	3.490	4.594	5.527	6.423	7.344	8.351	9.524	11.030	13.362	15.507	17.535	20.090	21.955	27.868
9	1.735	2.088	2.700	3.325	4.168	5.380	6.393	7.357	8.343	9.414	10.656	12.242	14.684	16.919	19.023	21.666	23.589	29.666
10	2.156	2.558	3.247	3.940	4.865	6.179	7.267	8.295	9.342	10.473	11.781	13.442	15.987	18.307	20.483	23.209	25.188	31.420
11	2.603	3.053	3.816	4.575	5.578	6.989	8.148	9.237	10.341	11.530	12.899	14.631	17.275	19.675	21.920	24.725	26.757	33.137
12	3.074	3.571	4.404	5.226	6.304	7.807	9.034	10.182	11.340	12.584	14.011	15.812	18.549	21.026	23.337	26.217	28.300	34.821
13	3.565	4.107	5.009	5.892	7.042	8.634	9.926	11.129	12.340	13.636	15.119	16.985	19.812	22.362	24.736	27.688	29.819	36.478
14	4.075	4.660	5.629	6.571	7.790	9.467	10.821	12.078	13.339	14.685	16.222	18.151	21.064	23.685	26.119	29.141	31.319	38.109
15	4.601	5.229	6.262	7.261	8.547	10.307	11.721	13.030	14.339	15.733	17.322	19.311	22.307	24.996	27.488	30.578	32.801	39.719
16	5.142	5.812	6.908	7.962	9.312	11.152	12.624	13.983	15.338	16.780	18.418	20.465	23.542	26.296	28.845	32.000	34.267	41.308
17	5.697	6.408	7.564	8.672	10.085	12.002	13.531	14.937	16.338	17.824	19.511	21.615	24.769	27.587	30.191	33.409	35.718	42.879
18	6.265	7.015	8.231	9.390	10.865	12.857	14.440	15.893	17.338	18.868	20.601	22.760	25.989	28.869	31.526	34.805	37.156	44.434
19	6.844	7.633	8.907	10.117	11.651	13.716	15.352	16.850	18.338	19.910	21.689	23.900	27.204	30.144	32.852	36.191	38.582	45.973
20	7.434	8.260	9.591	10.851	12.443	14.578	16.266	17.809	19.337	20.951	22.775	25.038	28.412	31.410	34.170	37.566	39.997	47.498
21	8.034	8.897	10.283	11.591	13.240	15.445	17.182	18.768	20.337	21.991	23.858	26.171	29.615	32.671	35.479	38.932	41.401	49.011
22	8.643	9.542	10.982	12.338	14.041	16.314	18.101	19.729	21.337	23.031	24.939	27.301	30.813	33.924	36.781	40.289	42.796	50.511
23	9.260	10.196	11.689	13.091	14.848	17.187	19.021	20.690	22.337	24.069	26.018	28.429	32.007	35.172	38.076	41.638	44.181	52.000
24	9.886	10.856	12.401	13.848	15.659	18.062	19.943	21.752	23.337	25.106	27.096	29.553	33.196	36.415	39.364	42.980	45.559	53.479
25	10.520	11.524	13.120	14.611	16.473	18.940	20.867	22.616	24.337	26.143	28.172	30.675	34.382	37.652	40.646	44.314	46.928	54.947
26	11.160	12.198	13.844	15.379	17.292	19.820	21.792	23.579	25.336	27.179	29.246	31.795	35.563	38.885	41.923	45.642	48.290	56.407
27	11.808	12.879	14.573	16.151	18.114	20.703	22.719	24.544	26.336	28.214	30.319	32.912	36.741	40.113	43.195	46.963	49.645	57.858
28	12.461	13.565	15.308	16.928	18.939	21.588	23.647	25.509	27.336	29.249	31.391	34.027	37.916	41.337	44.461	48.278	50.993	59.300
29	13.121	14.256	16.047	17.708	19.768	22.475	24.577	26.475	28.336	30.283	32.461	35.139	39.087	42.557	45.722	49.588	52.336	60.735
30	13.787	14.953	16.791	18.493	20.599	23.364	25.508	27.442	29.336	31.316	33.530	36.250	40.256	43.773	46.979	50.892	53.672	62.162
35	17.192	18.509	20.569	22.465	24.797	27.836	30.178	32.282	34.336	36.475	38.859	41.778	46.059	49.802	53.203	57.342	60.275	69.199
40	20.707	22.164	24.433	26.509	29.051	32.345	34.872	37.134	39.335	41.622	44.165	47.269	51.805	55.758	59.342	63.691	66.766	76.095
45	24.311	25.901	28.366	30.612	33.350	36.884	39.585	41.995	44.335	46.761	49.452	52.729	57.505	61.656	65.410	69.957	73.166	82.876
50	27.991	29.707	32.357	34.764	37.689	41.449	44.313	46.864	49.335	51.892	54.723	58.164	63.167	67.505	71.420	76.154	79.490	89.561
60	35.534	37.485	40.482	43.188	46.459	50.641	53.809	56.620	59.335	62.135	65.227	68.972	74.397	79.082	83.298	88.379	91.952	102.695
70	43.275	45.442	48.758	51.739	55.329	59.898	63.346	66.396	69.334	72.358	75.689	79.715	85.527	90.531	95.023	100.425	104.215	115.578
80	51.172	53.540	57.153	60.391	64.278	69.207	72.915	76.188	79.334	82.566	86.120	90.405	96.578	101.879	106.629	112.329	116.321	128.261
90	59.196	61.754	65.647	69.126	73.291	78.558	82.511	85.993	89.334	92.761	96.524	101.054	107.565	113.145	118.136	124.116	128.299	140.782
100	67.328	70.065	74.222	77.929	82.358	87.945	92.129	95.808	99.334	102.946	106.906	111.667	118.498	124.342	129.561	135.807	140.169	153.167
120	83.852	86.923	91.573	95.705	100.624	106.806	111.419	115.465	119.334	123.289	127.616	132.806	140.233	146.567	152.211	158.950	163.648	177.603
140	100.655	104.034	109.137	113.659	119.029	125.758	130.766	135.149	139.334	143.604	148.269	153.854	161.827	168.613	174.648	181.840	186.847	201.683
160	117.679	121.346	126.870	131.756	137.546	144.783	150.158	154.856	159.334	163.898	168.876	174.828	183.311	190.516	196.915	204.530	209.824	225.481
180	134.884	138.820	144.741	149.969	156.153	163.868	169.588	174.580	179.334	184.173	189.446	195.743	204.704	212.304	219.044	227.056	232.620	249.048
200	152.241	156.432	162.728	168.279	174.835	183.003	189.049	194.319	199.334	204.434	209.985	216.609	226.021	233.994	241.058	249.445	255.264	272.423

Percentiles of the chi-square distribution

% / df	………..	95	97.5	99	99.5	99.95
1	………..	3.841	5.024	6.635	7.879	12.116
2	………..	5.991	7.378	9.210	10.597	15.202
3	………..	7.815	9.348	11.345	12.838	17.730
4	………..	9.488	11.143	13.277	14.860	19.997
5	………..	11.070	12.833	15.066	16.750	22.105

Recall chi-square statistic=75.2 in AOM study
P<0.0005 for df=1
Conclusion: Proportion of filled prescription is different between groups.

RXC TABLES

Example: *Fusarium keritities* & contact lens solution (Chang JAMA 2006)

	Cases	Controls	Totals
ReNu/MoistureLock	20(9)	7(18)	27
MultiPlus (all brands)	9(11)	24(22)	33
Others	0(9)	27(18)	27
Totals	29	58	87

$$\text{Test Statistic} = \frac{(20-9)^2}{9} + \ldots = 34.2 \quad df = (3-1)(2-1) = 2$$

$$\sim \chi_{(2)}^{2} \quad > 15.202, \quad p < 0.0005$$

Conclusion: *Fusarium keritities* is significantly associated with the type of contact lens solution used.

Percentiles of the chi-square distribution

df \ %	………..	95	97.5	99	99.5	99.95
1	………..	3.841	5.024	6.635	7.879	12.116
2	………..	5.991	7.378	9.210	10.597	15.202
3	………..	7.815	9.348	11.345	12.838	17.730
4	………..	9.488	11.143	13.277	14.860	19.997
5	…………	11.070	12.833	15.066	16.750	22.105

ASSUMPTIONS for Chi-square Tests

1. Independent units
2. $E \geq 5$ (Rule-of-thumb)

What if assumptions are not met?

1. Fisher's Exact Test for small numbers
2. McNemar's test for matched pairs
3. Go see a statistician

SAMPLE SIZE for Detecting Difference between Two Proportions

Need: α, β, π_1, π_2

where $(\pi 1 - \pi 2)$ is the minimum clinically significant effect you want to detect

Put into formula or software to calculate require n per group.

SAMPLE SIZE - Example

Design a randomized trial: Treatment vs placebo.

Placebo response = 0.2

Hypothesized treatment response = 0.4

(based on clinically important difference)

$\alpha = 0.05$ $\quad Z_{\alpha/2}(\text{two tailed}) = 1.96$

$\beta = 0.1$ $\quad \text{Power} = .9 \quad Z_{\beta}(\text{one-tailed}) = -1.28$

$\pi_1 = 0.2, \quad \pi_2 = 0.4$

Substitute in sample size formula:

For each group n=92

SUMMARY

1. Estimation and C.I. for:
    a. A single proportion
    b. Two proportions

2. Chi-square tests
    a. Large sample (RxC)
    b. Fisher’s exact test
    c. McNemar’s test