Basic Biostatistics

Statistics for Public Health Practice

B. Burt Gerstman

Chapter 18
Cross-Tabulated Counts

February 09 1

In Chapter 18:

- 18.1 Types of Samples
- 18.2 Naturalistic and Cohort Samples
- 18.3 Chi-Square Test of Association
- 18.4 Test for Trend
- 18.5 Case-Control Samples
- 18.6 Matched Pairs

February 09 2

§18.1 Types of Samples

This chapter continues our consideration of categorical data

We begin by considering how data may be collected

February 09 3

Types of Samples, cont.

I. Naturalistic Samples. SRS with data cross-classified according to the explanatory and response variables after selection.

II. Purposive Cohort Samples. Select a fixed number of individuals in groups defined by the explanatory factor

III. Case-Control Samples. Select a fixed number of individuals in groups according to the response variable

February 09 4

Naturalistic Sample

SRS ⇒ cross-classify explanatory and response

Illustrative Example: ***Naturalistic sample (Cytomegalovirus and coronary restenosis).*** Investigators prospectively studied 75 consecutive patients undergoing coronary atherectomy for symptomatic coronary artery disease. Before atherectomy was performed, blood levels of anti-cytomegalovirus (CMV) antibodies were measured. Forty-nine patients were seropositive for CMV, and 26 patients were seronegative. Subjects were then followed for 6 months after atherectomy to determine incidents of restenosis.[b] If we make the simplifying assumption that the sample represents a simple random sample of like patients, we can view this as a naturalistic sample of $N = 75$ in which $n_1 = 49$ are seropositive and $n_2 = 26$ are seronegatives for CMV. ■

Purposive Cohorts

Select by explanatory status ⇒ ascertain responses

Illustrative Example: ***Purposive cohort sample (Estrogen trial).*** Chapter 16 examined data from an experimental cohort that was part of the NIH-sponsored Women's Health Initiative study. This experimental cohort study involved 16,608 postmenopausal women. Roughly half these women were assigned conjugated estrogens ($n_1 = 8506$). The remaining subjects ($n_2 = 8102$) received a placebo. Coronary heart disease, invasive breast cancer, and other adverse outcomes were monitored in the groups on an ongoing basis.[c] ■

February 09 6

Case-Control

Select by response status ⇒ ascertain explanatory status

Illustrative Example: ***Case-control sample (Baldness and myocardial infarction).*** To examine the relationship between male-pattern baldness and myocardial infarction, investigators studied 655 men admitted to a hospital for a first nonfatal myocardial infarction (cases) and 772 admitted to the same hospitals with noncardiac diagnoses (controls). This is a case-control sample because the investigator selected fixed numbers of subjects based on the status of the response variable (myocardial infarction). The extent of baldness was assessed in subjects using several methods.[e] ■

February 09 7

§18.2 Naturalistic and Cohort

- **Example**: Smoking by education level
- *R* rows and *C* columns (*R*-by-*C* table)
- For uniformity, put explanatory var. in row and response var. in columns
- Totals in table **margins**

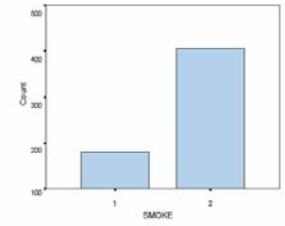
Degree	Smoke +	Smoke –	Tot
HighS	12	38	50
JC	18	67	85
Some	27	95	122
UG	32	239	271
Grad	5	52	57
Total	94	491	585

February 09 8

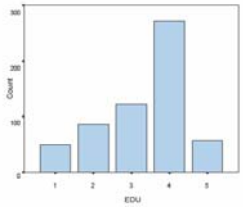
Marginal Distributions

Naturalistic samples only

Smoking Distribution (column variable)



Education Distribution (row variable)



9

Relationships

Conditional Percents

Relationship between row variable and column variable described by **row or column (conditional) percents**

Cohort and Naturalistic Samples

$$\text{row percent} = \frac{\text{cell count}}{\text{row total}} \times 100\%$$

Case - Control Sample

$$\text{column percent} = \frac{\text{cell count}}{\text{column total}} \times 100\%$$

February 09 10

Incidence and Prevalence

(Cohort & Naturalistic Only)

R-by-2 Table

	+	–	Total
Grp 1	a_1	b_1	n_1
Grp 2	a_2	b_2	n_2
↓	↓	↓	↓
Grp R	a_R	b_R	n_R
Total	m_1	m_2	N

Incidence or prevalence

Group i

$$\hat{p}_i = \frac{a_i}{n_i}$$

February 09 11

Example

Prevalence of smoking by education:

Education group	Smoker 1 (Yes)	Smoker 2 (No)	Total	Prevalence ($\hat{p}_i$)
1 (High school graduate)	12	38	50	0.2400[a]
2 (Associate degree)	18	67	85	0.2118[b]
3 (Some college)	27	95	122	0.2213
4 (Undergraduate degree)	32	239	271	0.1181
5 (Graduate degree)	5	52	57	0.0877

Example, prevalence group 1:

$$\hat{p}_1 = \frac{a_1}{n_1} = \frac{12}{50} = 0.24$$

February 09 12

Relative Risk
R-by-2 Tables

Let group 1 represent least exposed group
Relative risks, group *i*:

$$\hat{RR}_i = \frac{\hat{p}_i}{\hat{p}_1}$$

February 09 13

Illustration: *RRs*

Education group	Smoker 1 (Yes)	Smoker 2 (No)	Total	Prevalence ($\hat{p}_i$)	"Relative risk" ($\hat{RR}_i$)
1 (High school graduate)	12	38	50	0.2400[a]	1.00
2 (Associate degree)	18	67	85	0.2118[b]	0.88[c]
3 (Some college)	27	95	122	0.2213	0.92
4 (Undergraduate degree)	32	239	271	0.1181	0.49
5 (Graduate degree)	5	52	57	0.0877	0.37

Example, RR group 2 :

$$\hat{RR}_2 = \frac{\hat{p}_2}{\hat{p}_1} = \frac{0.2118}{0.2400} = 0.88$$

Downward dose-response in RRs

February 09

More than Two Levels of Response

Efficacy of Echinacea. A randomized controlled clinical trial pitted echinacea vs. placebo in the treatment of upper respiratory symptoms in children. The response variable was severity of illness classified as: mild, moderate or severe.

	Mild	Moderate	Severe	Total
Echinacea	153	128	48	329
Placebo	170	157	40	367
Total	323	285	88	696

Source: JAMA 2003. *290*(21), 2824-30

February 09 17

Echinacea Example

(B) Conditional distributions (row percents)

	Parental assessment: Mild	Moderate	Severe	Total
Echinacea	46.5	38.9	14.6	100.0
Placebo	46.3	42.8	10.9	100.0
Total	46.4	40.9	12.6	100.0

- Row percents determine incidence of each outcome
- Example (data prior slide) % severe w/ echinacea = 48 / 329 = .1459 = 14.6%
- Example of calculation, % severe w/placebo = 40 / 367 = .1090 = 10.9%
- Treatment group fared slightly worse than control group

February 09 18

§18.3 Chi-Square Test of Association

A. Hypotheses.
H_0: no association in population
H_a: association in population

B. Test statistic – by hand or computer

$$X^2_{stat} = \sum_{\text{all cells}} \frac{(O_i - E_i)^2}{E_i} \quad \text{where } O_i \equiv \text{observed count, cell } i$$

$$\text{and } E_i \equiv \text{expected count in cell } i \text{ calculated } E_i = \frac{\text{row total} \times \text{column total}}{\text{table total}}$$

$$df = (R-1)(C-1)$$

C. *P*-value. Use Table E or software

February 09 19

Chi-Square Example

H_0: no association in the population
H_a: association in the population

Data

Degree	Smoke +	Smoke –	Tot
HighS	12	38	50
JC	18	67	85
Some	27	95	122
UG	32	239	271
Grad	5	52	57
Total	94	491	585

February 09 20

Expected Frequencies

(under H_0)

$$\text{Expected frequencies } E_i = \frac{\text{row total} \times \text{column total}}{\text{table total}}$$

	Smoke +	Smoke −	Total
HighS	(50 × 94) ÷ 585 = 8.034	41.966	50
JC	13.658	71.342	85
Some	19.603	102.397	122
UG	43.545	227.455	271
Grad	9.159	47.841	57
Total	94	491	585

Chi-Square Hand Calc.

$$X^2_{stat} = \sum\left[\frac{(O_i - E_i)^2}{E_i}\right] = \frac{(12 - 8.034)^2}{8.034} + \frac{(38 - 41.966)^2}{41.966} +$$
$$\frac{(18 - 13.658)^2}{13.658} + \frac{(67 - 71.342)^2}{71.342} + \frac{(27 - 19.603)^2}{19.603} +$$
$$\frac{(95 - 102.397)^2}{102.397} + \frac{(32 - 43.545)^2}{43.545} + \frac{(239 - 227.455)^2}{227.455} +$$
$$\frac{(5 - 9.159)^2}{9.159} + \frac{(52 - 47.841)^2}{47.841} = 1.958 + 0.375 + 1.380 +$$
$$0.264 + 2.791 + 0.534 + 3.061 + 0.586 + 1.889 +$$
$$0.362 = 13.20$$

Degrees of freedom = $(R - 1) \times (C - 1) = (5 - 1) \times (2 - 1) = 4 \times 1 = 4$

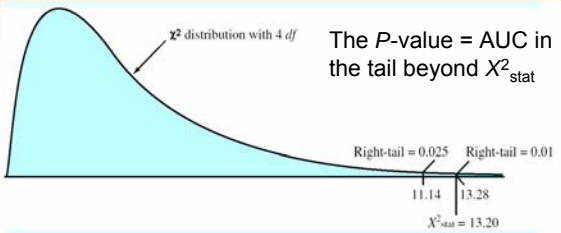
Chi-Square ⇒ *P*-value

- X^2_{stat}= 13.20 with 4 *df*
- Table E, find the row for 4 *df*, then find chi-square critical values that bracket test statistic
- Example: bracketing values are 11.14 (*P* = .025) and 13.28 (*P* = .01) ⇒ thus .025 < *P* < .01

	Probability in right tail								
df	0.98	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.01
4	0.48	5.39	5.99	6.74	7.78	9.49	11.14	13.28	14.86

February 09 23

Illustration: X^2_{stat}= 13.20 with 4 df

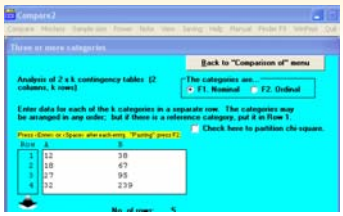


The *P*-value = AUC in the tail beyond X^2_{stat}

FIGURE 18.5 Sampling distribution for X^2_{stat} from illustrative example. The *P*-value is about equal to 0.01.

February 09 24

WinPEPI > Compare2 > F. Categorical Data (2 x k)



Input screen row 5 not visible

Output

```
Chi-square tests (DF = 4):

 Pearson chi-sq.          = 13.199  P = 0.010
```

February 09 25

Continuity Corrected Chi-Square

- Two different chi-square statistics
- Both used in practice
- Pearson's ("uncorrected") chi-square

$$X^2_{stat} = \sum_{\text{all cells}} \frac{(O_i - E_i)^2}{E_i}$$

- Yates' continuity-corrected chi-square:

$$X^2_{stat,c} = \sum_{\text{all cells}} \frac{\left(|O_i - E_i| - \frac{1}{2}\right)^2}{E_i}$$

February 09 26

Chi-Square, cont.

1. **How the chi-square works.** When observed values = expected values, the chi-square statistic is 0. When the observed minus expected values gets large $\Rightarrow$ evidence against H_0 mounts
2. **Avoid chi-square tests in small samples.** Do not use a chi-square test when more than 20% of the cells have expected values that are less than 5.

Chi-Square, cont.

3. **Supplement chi-squares with measures of association.** Chi-square statistics do *not* quantify effects (need RR, RD, or OR)
4. Chi-square and *z* tests (Ch 17) produce identical *P*-values. The relationship between the statistics is:

$$\sqrt{X^2_{\text{stat with 1 df}}} = z_{\text{stat}}$$