

## Index of 9-L

Page	Title
1	Practical information
2	Mid-term exam practical info
3	Data example: Avadex for mice
4	Data example: Health habits of students
5	Data example: Music and wine purchase
6	Multinomial distribution
7	2-way tables: Notation
8	2-way tables: Models and estimation
9	2-way tables: Hypotheses
10	2-way tables: Test
11	Chi-square distribution
12	Music and wine purchase — Analysis
13	Health habits — Analysis
14	$2 \times 2$ tables
15	Fisher's exact test
16	Avadex data: Fisher's exact test
17	Simpson's paradox
18	Summary notes

## PRACTICAL INFORMATION

### Lecture contents:

- \* summary worksheet (Chapter 7) on sample size calc.,
- \* brief follow-up on sample size calc. (8L–17),
- \* 2-sample proportions revisited (7L–15/16), incl.  $z$ -tests,
- 2-way (contingency) tables and chi-square tests<sup>1</sup>  
(also Simpson's paradox and 2-way table calculations),<sup>2</sup>
- include in course: the distinction between the 2 models,
- two additional topics:
  - \* multinomial distribution (*in* course curriculum),
  - \* Fisher's exact test (*not in* course curriculum).

### Schedule news:

- lab review 1-2pm today (incl. midterm 2012),
- midterm exam next Wednesday: March 7th, 9:00-10:00am, AVC 278N,
- home assignment II returned on Monday, with solution,
- home assignment III to be posted on March 12th.

---

<sup>1</sup> PSLS 3e: Chapter 23; S: Section 11.2; IPS 7e: Sections 9.1-2.

<sup>2</sup> PSLS 3e: Chapter 5; IPS 7e: Section 2.5; not covered in S (note that S discusses Simpson's paradox for quantitative data only).

## MID-TERM EXAM PRACTICAL INFO

- mark: *optional* for 15% of course mark; that is, you decide *after you have received your mark* if you want to use it,
- all aids (books and notes) are allowed — except computer or computer-like device (tablet or smartphone),
- duration: 1 hour sharp,
- note: you must bring a calculator and statistical tables.

1 question/assignment with possible types of (sub)questions:

- choice of statistical model and analysis —
  - \* carry out analysis when calculations manageable (see below),
  - \* or base analysis on Minitab print + extra calculations,
  - \* or outline analysis if calculations not manageable,
- probability calculations manageable (see below),
- multiple choice (one or several correct answers),
- important note: (sub)questions can usually be answered independently of each other.

Calculations manageable by hand (calculator):

- simple probabilities (e.g.,  $1-p$ ,  $(1-p)^n$ , simple binomial),
- probabilities in normal distribution (incl. standardization),
- $z/t$ -tests and CIs (given estimates or calculated statistics),
- one-sample proportion and two-sample proportions (only CI for two-sample),
- no data entry into calculator or large summations.

DATA EXAMPLE: AVADEx FOR MICE
-------------------------------

A clinical trial was conducted to assess a possible carcinogenic effect of Avadex (a fungicide).

- Data: control and treatment (Avadex in feed) groups of mice; number of mice with lung tumors recorded:

Outcome	Avadex group	control group	Total
tumors	4	5	9
no tumors	12	74	86
Total	16	79	95

- Model:  $X \sim B(16, p_1)$  and  $Y \sim B(79, p_2)$ , where  $X, Y$  are no. of mice with tumors in Avadex/control groups,
- Estimation:

$$\hat{p}_1 = 4/16 = 0.250, \quad SE_{\hat{p}_1} = 0.108,$$

$$\hat{p}_2 = 5/79 = 0.063, \quad SE_{\hat{p}_2} = 0.027,$$

$$\hat{p}_1 - \hat{p}_2 = 0.187, \quad SE_{\hat{p}_1 - \hat{p}_2} = 0.112,$$

- Confidence intervals: (95%, plus four method)

$$p_1 : 0.30 \pm 0.20, \quad p_2 : 0.084 \pm 0.060, \quad p_1 - p_2 : 0.204 \pm 0.215.$$

- Hypotheses:  $H_0: p_1 = p_2 (= p)$  vs.  $H_a: p_1 \neq p_2$ ;  
estimate common  $p$  under  $H_0$ :  $\hat{p} = 9/95 = 0.095$ , and  
 $SE_{D_p} = \sqrt{\hat{p}(1-\hat{p})((1/16) + (1/79))} = 0.080$ ,

- Test: (approximate, based on normal distribution)  
 $z = (\hat{p}_1 - \hat{p}_2)/SE_{D_p}$ ,  $z_{\text{obs}} = 0.187/0.080 = 2.326$ ,  
which gives  $P\text{-value} = 2 \cdot P(Z > 2.326) = 0.020$ .

## DATA EXAMPLE: HEALTH HABITS OF STUDENTS

A survey<sup>3</sup> obtained information on the levels of physical activity and consumptions of fruit — is there a link between these or they are independent?

Data: responses obtained for 1184 college students:

Fruit consumption	Physical activity			Total
	low	moderate	vigorous	
low	69	206	294	569
medium	25	126	170	321
high	14	111	169	294
Total	108	443	633	1184

2 response variables, because none of the variables are fixed (known) in advance.

Descriptive statistics: (*for response variables*)

- Marginal distributions — looking at each variable separately: fruit consumption:  $569/1184=48\%$  low,  $321/1184=27\%$  medium,  $294/1184=25\%$  high; physical activity:  $108/1184=9\%$  low,  $443/1184=37\%$  moderate,  $633/1184=53\%$  vigorous,
- Conditional distributions — looking at one variable when the other is fixed: e.g. fruit consumption in low physical activity group:  $69/108=64\%$  low,  $25/108=23\%$  medium,  $14/108=13\%$  high.

---

<sup>3</sup> Data from Seo D-C et al. (2007), *Amer. J. College Health* **56**, 187-197; also Example 9.8 of IPS 7e.

## DATA EXAMPLE: MUSIC AND WINE PURCHASE

Experimental study on music's impact on wine purchase (no. of bottles sold of categories French, Italian and Other) in a supermarket<sup>4</sup> under different music conditions (None, French accordion music, Italian string music).

Data: 243 bottles sold categorized by wine and music types:

# bottles	Music			
Wine	None	French	Italian	Total
French	30	39	30	99
Italian	11	1	19	31
Other	43	35	35	113
Total	84	75	84	243

1 response variable — the type of wine purchased,  
1 explanatory variable — the type of music played (controlled by the store)  $\sim$  3 separate time periods and therefore independent samples.

Descriptive statistics:

- Conditional distributions — proportions of wine sold for the three samples: e.g., Italian wine  $\sim 19/84 = 23\%$  for Italian music, but only  $1/75 = 1\%$  for French music,
- Marginal wine type distribution — pooled across music type: e.g., Italian wine  $\sim 31/243 = 13\%$  of bottles sold.

---

<sup>4</sup> Study carried out in Northern Ireland in 1990s; Ryan et al. (1998), Proc. Nutrition Soc. **57**, 169A; also Example 9.8 in IPS 6e.

MULTINOMIAL DISTRIBUTION
--------------------------

Example: wine purchase when no music is played:

type of wine	French	Italian	Other	Total
count	30	11	43	84
symbol $N_i$	$N_1$	$N_2$	$N_3$	$n$
rel. frequency	0.357	0.131	0.519	1
symbol $p_i$	$p_1$	$p_2$	$p_3$	1

Multinomial Distribution:

$$(N_1, N_2, \dots, N_q) \sim \text{multinomial}(n; p_1, p_2, \dots, p_q)$$

where  $n$  is the total number of observations,  $q$  is the number of classes, and  $p_i$  is the (population) probability of each class (so that  $p_1 + p_2 + \dots + p_q = 1$ ), if

- mathematical definition:

$$P(N_1 = n_1, \dots, N_q = n_q) = \binom{n}{n_1 \dots n_q} p_1^{n_1} \dots p_q^{n_q},$$

where *multinomial coefficient*  $\binom{n}{n_1 \dots n_q} = \frac{n!}{n_1! \dots n_q!}$ ,

- conceptual definition (“multinomial setting”):
  - \*  $n$  trials; in each, one of  $q$  categories is observed,
  - \* *independent* trials, with *same probabilities* of the categories across all trials.

Note: As  $\text{multinomial}(n; p_1, p_2) = B(n, p_1)$ , the multinomial distribution extends the binomial to  $> 2$  categories.

2-WAY TABLES: NOTATION
------------------------

$I \times J$  table:

$n$  observations grouped (cross-classified) according to two criteria A and B with  $I$  and  $J$  levels, respectively:

Counts	$j \sim$ criterion B							sum
	1	2	...	$j$	...	$J-1$	$J$	
1	$N_{11}$	$N_{12}$	...	$N_{1j}$	...	$N_{1,J-1}$	$N_{1J}$	$N_{1.}$
2	$N_{21}$	$N_{22}$	...	$N_{2j}$	...	$N_{2,J-1}$	$N_{2J}$	$N_{2.}$
$i \sim$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
crit. $i$	$N_{i1}$	$N_{i2}$	...	$N_{ij}$	...	$N_{i,J-1}$	$N_{iJ}$	$N_{i.}$
A	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
$I-1$	$N_{I-1,1}$	$N_{I-1,2}$	...	$N_{I-1,j}$	...	$N_{I-1,J-1}$	$N_{I-1,J}$	$N_{I-1.}$
$I$	$N_{I1}$	$N_{I2}$	...	$N_{Ij}$	...	$N_{I,J-1}$	$N_{IJ}$	$N_{I.}$
sum	$N_{.1}$	$N_{.2}$	...	$N_{.j}$	...	$N_{.,J-1}$	$N_{.J}$	$n$

Notes:

- always:  $i \sim$  rows, and  $I =$  no. of rows,  
 $j \sim$  columns, and  $J =$  no. of columns,
- each  $(i, j)$  combination corresponds to a table cell,
- the  $N_{i.}$ 's are row totals, and the  $N_{.j}$ 's are column totals,
- textbook chapters are “notation-less”, but double subscript notation will be used later on (for ANOVA).

## 2-WAY TABLES: MODELS AND ESTIMATION

Model I: Independent multinomials<sup>5</sup> over columns (or rows):

$$\begin{aligned} (N_{11}, \dots, N_{I1}) &\sim \text{multinomial}(N_{\cdot 1}; p_{11}, \dots, p_{I1}), \\ &\vdots \\ (N_{1J}, \dots, N_{IJ}) &\sim \text{multinomial}(N_{\cdot J}; p_{1J}, \dots, p_{IJ}), \end{aligned}$$

where  $p_{ij}$  = probability of group  $i$  in  $j$ th column, and all probability column sums equal 1 (e.g.,  $p_{11} + \dots + p_{I1} = 1$ ).

- Examples: Wine purchase, Avadex in mice,
- Assumptions: multinomial setting in each column, and independence between columns,
- Estimation:  $\hat{p}_{ij} = N_{ij}/N_{\cdot j}$  — sample proportions within each column,
- Interpretation: one response variable (rows), one explanatory variable (columns).

Model II: a Single multinomial<sup>6</sup> on  $IJ$  classes:

$$(N_{11}, \dots, N_{ij}, \dots, N_{IJ}) \sim \text{multinomial}(n; p_{11}, \dots, p_{ij}, \dots, p_{IJ}),$$

where  $p_{ij}$  = probability of group (cell)  $(i, j)$ , and all probabilities sum to 1 (i.e.,  $p_{11} + p_{12} + \dots + p_{1J} + p_{21} + \dots + p_{2J} + \dots + p_{IJ} = 1$ ),

- Example: Health habits,
- Assumptions: multinomial setting for table ( $IJ$  cells),
- Estimation:  $\hat{p}_{ij} = N_{ij}/n$  — table sample proportions,
- Interpretation: 2 response variables (rows and columns).

---

<sup>5</sup> IPS: model for comparing several populations or independent SRSs.

<sup>6</sup> IPS: model for examining independence or for a single SRS.

## 2-WAY TABLES: HYPOTHESES

Model I: Independent multinomials over columns:

- Hypothesis  $H_0$ : homogeneity among columns (same distribution in all columns):

$$H_0 : p_{ij} = p_{i.} \text{ for all } j, \quad \text{and } H_a : \text{not } H_0,$$

and  $H_0$  corresponds to using the marginal distribution across columns (row totals),

- Estimation under  $H_0$ :  $\hat{p}_{i.} = N_{i.}/n$ ,
- Expected value of cell  $(i, j)$  under  $H_0$ :  
 $e_{ij} = \text{row total} \times \text{column total} / \text{overall total}.$

Model II: a Single multinomial on  $IJ$  classes:

- Hypothesis  $H_0$ : independence between row and column classification:

$$H_0 : p_{ij} = p_{i.} p_{.j} \text{ for all } i \text{ and } j, \quad \text{and } H_a : \text{not } H_0,$$

and  $H_0$  corresponds to using the marginal distribution across both rows and columns,

- Interpretation:

$$\begin{aligned} p_{ij} &= P(\text{row}=i \text{ and column}=j) \\ &= (\textit{independence}) P(\text{row}=i) P(\text{column}=j) = p_{i.} p_{.j}, \end{aligned}$$

- Estimation under  $H_0$ :  $\hat{p}_{i.} = N_{i.}/n$ , and  $\hat{p}_{.j} = N_{.j}/n$ ,
- Expected value of cell  $(i, j)$  under  $H_0$ :  
 $e_{ij} = \text{row total} \times \text{column total} / \text{overall total}.$

## 2-WAY TABLES: TEST

Result: In *both* of the models I and II, we test  $H_0$  (homogeneity or independence) by the (Pearson chi-square) statistic,

$$X^2 = \sum_{i,j} \frac{(N_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i,j} \frac{(\text{observed}_{ij} - \text{expected}_{ij})^2}{\text{expected}_{ij}}$$

$\sim \chi^2$  distribution(df) under  $H_0$ ,

where  $\text{df} = (I - 1)(J - 1)$ .<sup>7</sup>

Notes:

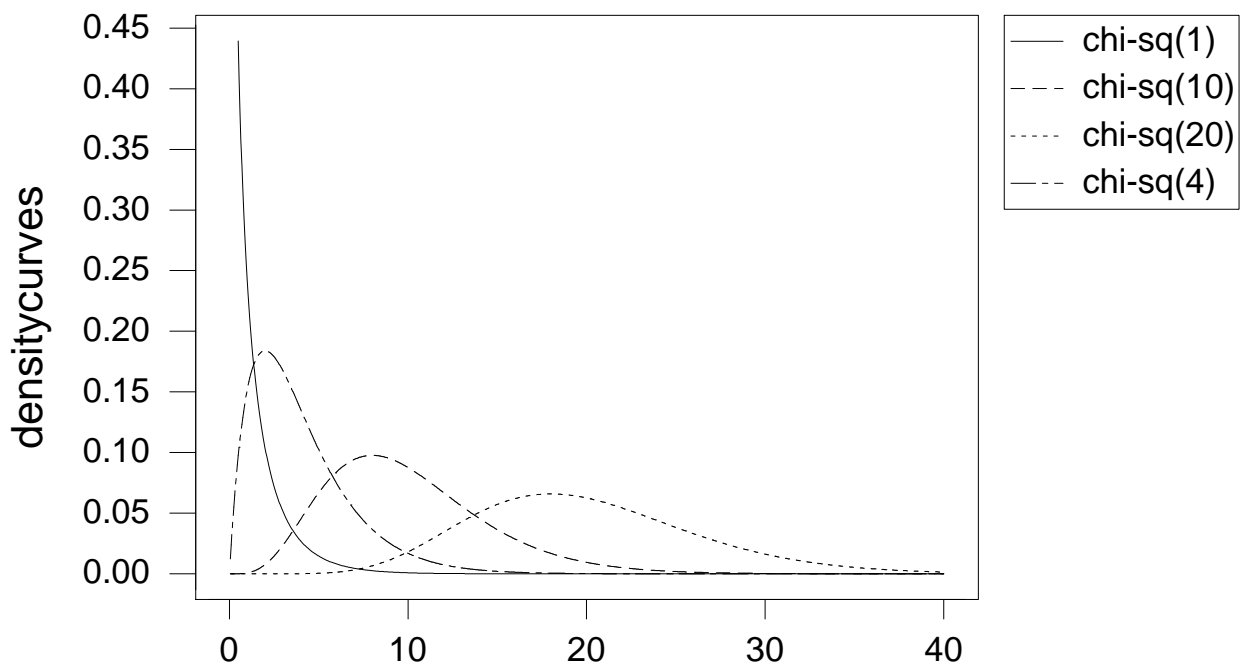
- *same* statistic, but *different* models and hypotheses  $\Rightarrow$  different interpretations!
- one-sided test: large values are critical, and  $P = P(\chi^2(\text{df}) > X_{\text{obs}}^2)$ ,
- distribution of  $X^2$  under  $H_0$  is approximate and best for large  $n$ ; guidelines for use:
  - \*  $e_{ij} > 1$  in all cells  $(i, j)$ , and
  - \*  $e_{ij} > 5$  in at least 80% of cells  $(i, j)$ ,
- other test for  $H_0$  exist and may be better...

---

<sup>7</sup> Technical note: the degrees of freedom can in both models be calculated as number of free parameters in model minus number of free parameters under  $H_0$ .

## $\chi^2$ DISTRIBUTIONS

- “chi-square” distribution,
- a new distribution — to be used for test statistics in tables of categorical data (*not for modelling*):  
— reference distribution for  $X^2$  (or  $\chi^2$ ) statistics,
- a single parameter df (df = 1, 2, 3, ...), the degrees of freedom (determined by the statistical design),
- denoted  $\chi^2(\text{df})$  to indicate degrees of freedom,
- distribution on  $(0, \infty)$  — only positive values,
- mean = df, standard deviation =  $\sqrt{2\text{df}}$ ,
- tail probabilities: software, or tables: PSLS Table D; S Table 5; IPS Table F.



MUSIC AND WINE PURCHASE — ANALYSIS
------------------------------------

- Model: three independent multinomial distributions on 3 classes,
- Hypothesis  $H_0$ : same distribution of wine purchases for all types of music,
- Test: see table below for computation of  $X^2$ :
  - \*  $df = (3 - 1) \cdot (3 - 1) = 4$ ,
  - \*  $X_{\text{obs}}^2 = 18.28$ ,  $P = P(\chi^2(4) > 18.28) = 0.0011$ ,
- Conclusion: very clear evidence against  $H_0 \Rightarrow$  wine purchase depends on music played,
- Presentation and estimation: observed proportions separately for the three samples.

Table of computations for  $X^2$ :

count ( $e_{ij}$ )	Music			
Wine	None	French	Italian	Total
French	30 (34.2)	39 (30.6)	30 (34.2)	99
Italian	11 (10.7)	<b>1 (9.6)</b>	<b>19 (10.7)</b>	31
Other	43 (39.1)	35 (34.9)	35 (39.1)	113
Total	84	75	84	243

$$X^2 = \frac{(30-34.2)^2}{34.2} + \frac{(39-30.6)^2}{30.6} + \dots + \frac{(35-39.1)^2}{39.1} = 18.28.$$

## HEALTH HABITS — ANALYSIS

- Model: a single multinomial distribution on 9 classes,
- Hypothesis  $H_0$ : independence between levels of fruit consumption and exercise,
- Test: see table below for computation of  $X^2$ :
  - \*  $df = (3-1) \cdot (3-1) = 4$ ,
  - \*  $X_{\text{obs}}^2 = 14.15$ ,  $P = P(\chi^2(4) > 14.15) = 0.007$ ,
- Conclusion: very clear evidence against  $H_0 \Rightarrow$  dependence between fruit consumption and exercise level,
- Presentation and estimation: conditional distributions for fruit consumption given physical activity, or conversely; the main discrepancies are in cells (low,low) and (high,low).

Table of computations for  $X^2$ :

count ( $e_{ij}$ )	Physical activity			
Fruit	low	moderate	vigorous	Total
low	<b>69 (51.9)</b>	206 (212.9)	294 (304.2)	569
medium	25 (29.3)	126 (120.1)	170 (171.6)	321
high	<b>14 (26.8)</b>	111 (110.0)	169 (157.2)	294
Total	108	443	633	1184

$$X^2 = \frac{(69-51.9)^2}{51.9} + \frac{(206-212.9)^2}{212.9} + \dots + \frac{(169-157.2)^2}{157.2} = 14.15.$$

## 2×2 TABLES

2×2 tables = special case of 2-way tables:

- simplest case, and very common in practice,
- special relations for chi-square statistic:
  - \*  $X^2 = z^2$ , where  $z$  is statistic for comparing two binomial distributions,  
⇒ methods equivalent (lead to same  $P$ -value),  
and guideline for use of  $X^2$  applies to  $z$  as well!
  - \* easier computational formula for  $X^2$ :

$$X^2 = \frac{(N_{11}N_{22} - N_{12}N_{21})^2}{N_{1.}N_{2.}N_{.1}N_{.2}} \times n$$

- an almost endless selection of methods / procedures:
  - \* alternative tests for same hypothesis:
    - Fisher's "exact" test (next slide),
    - continuity-correction for  $X^2$  (better approximated by  $\chi^2$ -distribution); use Fisher's test instead,
  - \* other measures for comparison of probabilities than differences: relative risk and odds-ratio (epi course),
  - \* tests for other hypotheses (e.g. McNemar's test) . . . ,
  - \* other statistics (e.g. kappa values) . . .
- simple advice: use your common sense and use only procedures that you understand (the rationale and assumptions behind).

## FISHER'S EXACT TEST

Fisher's exact test for 2 independent binomial distributions:

---

- test of null hypothesis  $H_0: p_1 = p_2$  against one- or two-sided alternative  $H_a$ ,
- test "statistic" = observed table (one cell of table),
- idea: compare observed table with other tables that have the *same margins* (row and column sums),
- technical: under  $H_0$ , tables with same margins  $\sim$  hypergeometric distribution  $\Rightarrow$  table probabilities computable,
- P-value for one-sided  $H_a$  = sum of table probabilities for tables more indicative for  $H_a$  than for  $H_0$ ,
- P-value for two-sided  $H_a$  = twice the smallest one-sided  $P$ , or the sum of table probabilities less than for observed table (most commonly used: Minitab/Stata/R).

Why / When use Fisher's test?

- + works also if  $\chi^2$ -approximation not good  
 $\Rightarrow$  recommended test if  $\chi^2$ -guidelines violated,
- + allows one-sided alternative hypothesis (as  $z$ -tests),
- requires software, and for large samples computing time may be very long,
- \* also for multinomial model and test of independence,
- \* version of test also for larger tables than  $2 \times 2$ .

AVADEx EXAMPLE: FISHER'S EXACT TEST
-------------------------------------

Data table (with expected values) and general notation:

Outcome	Avadex	control	Total		Avadex	control	Total
tumors	4 (1.5*)	5 (7.5)	9		$a$	$b$	$K$
no tumors	12 (14.5)	74 (74.5)	86		$c$	$d$	$N - K$
Total	16	79	95		$n$	$N - n$	$N$

\* expected value  $< 5 \Rightarrow$  conditions for  $X^2$ -test violated

Idea: under  $H_0$ , we consider all table margins as fixed and focus only on the distribution of the  $K = 9$  tumor cases among the two samples,

- similar to sampling of  $n$  elements from a finite population of size  $N$  of which  $K$  are “cases”  $\rightarrow$  a hypergeometric distribution  $(N, K, n, a)$  for the number of cases  $a$  in the sample.

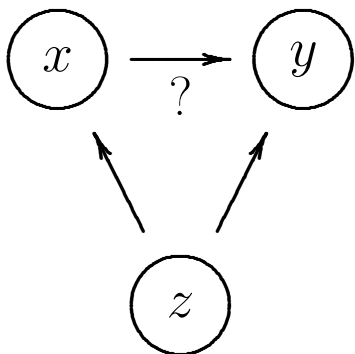
Scenario	Cell counts				Prob. under $H_0$	One/two-sided tail prob.		
	$a$	$b$	$c$	$d$		one- ( $>$ )	one- ( $<$ )	two- ( $\neq$ )
1	0	9	16	70	.175		.175	
2	1	8	15	71	.335		.335	
3	2	7	14	72	.296		.296	
4	3	6	13	73	.132		.132	
<b>5</b>	<b>4</b>	<b>5</b>	<b>12</b>	<b>74</b>	<b>.035</b>	.035	.035	.035
6	5	4	11	75	.006	.006		.006
7	6	3	10	76	.001	.001		.006
8	7	2	9	77	.000	.000		.000
9	8	1	8	78	.000	.000		.000
10	9	0	7	79	.000	.000		.000
$P$ -value = sum of tail prob.						.041	.994	.041

Note: two-sided  $P$ -value computed by adding up prob. for tables with prob.  $\leq$  observed table; here, it equals one of the one-sided  $P$ -values.

## SIMPSON'S PARADOX

- = an extreme affect of ignoring a lurking variable,
- at closer look, no paradox at all.

Example: Airline punctuality (IPS7e Suppl. Exercise 9.19):



$x$  = airline (Alaska Airlines/America West)  
 $y$  = on time (yes/no)  
 $z$  = airport (Los Angeles/Phoenix)

Summary 2-way tables:

	All airports		Los Angeles		Phoenix	
On time/airline	AA	AW	AA	AW	AA	AW
on time	718	5534	497	694	221	4840
delayed	74	532	62	117	12	415
sum	792	6066	559	811	233	5255
prop. delayed	0.093	0.088	0.111	0.144	0.052	0.079

Comments and conclusions:

- Simpson's paradox:
  - \* overall, airline AW better (less delays) than AA,
  - \* in both airports, airline AA better than AW,
- explanation: airport Los Angeles has more flights delayed, and AA has more flights at this airport,
- conclusion: "paradox" may happen whenever both effects  $z \rightarrow x$  and  $z \rightarrow y$  are strong.

## SUMMARY NOTES

### Key words and concepts:

- two-way table (of counts),  $2 \times 2$ -table (and larger),
- marginal and conditional distributions in two-way tables, Simpsons paradox,
- multinomial distribution for counts in  $> 2$  categories,
- 2 models/hypotheses for two-way tables of counts:
  - \* independent multinomial distributions (for comparing several populations/samples), with hypothesis of *homogeneity*,
  - \* single multinomial distribution for entire table (for single population/sample), with hypothesis of *independence*,
- $X^2$ -test statistic:
  - \* computation from observed and expected counts,
  - \* reference  $\chi^2$ -distribution, and its degrees of freedom,
  - \* guideline for use of  $X^2$ -test (and its  $\chi^2$  reference distribution),
  - \* relationship with 2-sample  $z$ -test (for  $2 \times 2$ -table),
- Fisher's exact test for sparse tables.