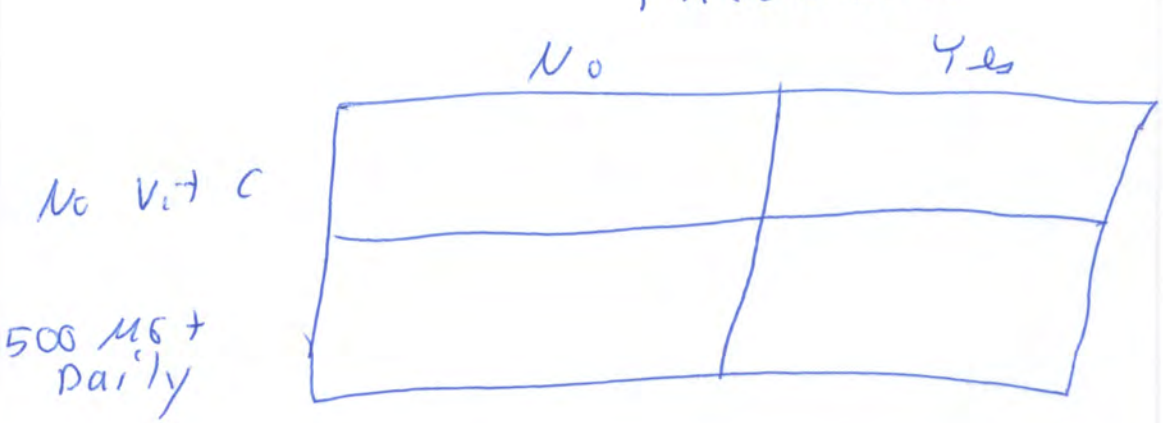
Contingency Tables: Part One

Maybe read 2.1 - 2.4 not 2.5

Contingency table is a <u>joint</u> frequency distribution. (2 or more variables)



X is rows I categories

Y is cols J categories

<u>Cell probabilities</u>

Course	Passed: No	Passed: Yes	Total	
Catch-up	π_{11}	π_{12}	π_{1+}	
Mainstream	π_{21}	π_{22}	π_{2+}	
Elite	π_{31}	π_{32}	π_{3+}	
Total	π_{+1}	π_{+2}	1	

Marginal Probabilities

$$P(X=i) = \sum_{j=1}^{J} \pi_{ij} = \pi_{i+}$$

$$P(Y=j) = \sum_{i=1}^{I} \pi_{ij} = \pi_{+j}$$

Conditional Probabilities

$$P(Y=j \mid X=i) = \frac{P(Y=j \cap X=i)}{P(X=i)} = \frac{\pi_{ij}}{\pi_{i+}}$$

Usually interest is in conditional distribution of response variable Y given explanatory variable X.

Cell frequencies

Course	Passed		Total
	No	Yes	
Catch-up	n_{11}	n_{12}	n_{1+}
Mainstream	n_{21}	n_{22}	n_{2+}
Elite	n_{31}	n_{32}	n_{3+}
Total	n_{+1}	n_{+2}	n

Course	Passed		Total
	No	Yes	
Catch-up	27	8	35
Mainstream	124	204	328
Elite	7	24	31
Total	158	236	394

Should we estimate π_{ij} with $p_{ij} = \frac{n_{ij}}{n}$? Sometimes.

Study Designs

Cross-sectional
Prospective
Retrospective (case-control)

Cross-sectional

Both vars are measured (assessed)
No assignment to experimental conditions
No selection based on variable values
total Sample size fixed by design
Multinomial, IJ
Estimate π_{ij} with p_{ij}
Estimating conditional probs is easy.

Prospective (looking forward from X to Y.)

Groups that define the explanatory variable are formed before response variable is observed.

Experimental studies.

Cohort studies

Stratified sampling

Marginal totals of explanatory variable are fixed by the design.

Assume random sampling within each category of X, and independence between categories.

Product multinomial: A product of I multinomial likelihoods

Each $\pi_{j|i} = P(Y=j \mid X=i)$

Product multinomial

- Take independent random samples sizes $n_{1+}, n_{2+} \cdots n_{I+}$ from I sub-populations
- In each, observe a multinomial with J categories. Compare.
- $\pi_{ij} = P(Y=j \mid X=i)$
- Likelihood

$$\prod_{i=1}^{I} \binom{n_{i+}}{n_{i1}, n_{i2}, \cdots, n_{iJ}} \pi_{i1}^{n_{i1}} \pi_{i2}^{n_{i2}} \cdots \pi_{i,J-1}^{n_{i,J-1}} \left(1 - \sum_{j=1}^{J-1} \pi_{ij}\right)^{n_{iJ}}$$

Retrospective design

Thurs. Sept. 29

Designs

Cross-sectional
Prospective
Retrospective

Incidence of common cold in a double blind study involving 279 French skiers (Pauling, 1971)

	Cold	No cold	
Placebo	31	109	140
Vit C	17	122	139
	48	231	279

Hypothetical Stratified Random Sample

	Related	Unrelated	Unemployed	
UTM	n_{11}	n_{12}	$n_{13} = 200 - n_1 - n_2$	$200 = n_{1+}$
SG	n_{21}	n_{22}	$n_{23} = 200 - n_{21} - n_{22}$	$200 = n_{2+}$
UTSC	n_{31}	n_{32}	$n_{33} = 200 - n_{31} - n_{32}$	$200 = n_{3+}$
	n_{+1}	n_{+2}	n_{+3}	n

MARGINAL FREQUENCIES for X variables are fixed by design.

Cell probabilities are conditional probabilities

	Related	Unrelated	Unemp	
VTM	π_{11}	π_{12}	$1-\pi_{11}-\pi_{12}$	1
SG	π_{21}	π_{22}	$1-\pi_{21}-\pi_{22}$	1
SC	π_{31}	π_{32}	$1-\pi_{31}-\pi_{32}$	1

Likelihood

$$\ell = \pi_{11}^{n_{11}} \pi_{12}^{n_{12}} (1-\pi_{11}-\pi_{12})^{n_{13}} \pi_{21}^{n_{21}} \pi_{22}^{n_{22}} (1-\pi_{21}-\pi_{22})^{n_{23}} \pi_{31}^{n_{31}} \pi_{32}^{n_{32}} (1-\pi_{31}-\pi_{32})^{n_{33}}$$

Independent - multiply to get likelihood

MLEs

$$\frac{\partial}{\partial \pi_{11}} \log \ell = \frac{\partial}{\partial \pi_{11}} \left(n_{11} \log \pi_{11} + n_{12} \log \pi_{12} + n_{13} \log (1-\pi_{11}-\pi_{12}) + \cdots + n_{33} \log (1-\pi_{31}-\pi_{32}) \right)$$

$$= \frac{n_{11}}{\pi_{11}} - \frac{n_{13}}{1-\pi_{11}-\pi_{12}} + 0$$

STOP!

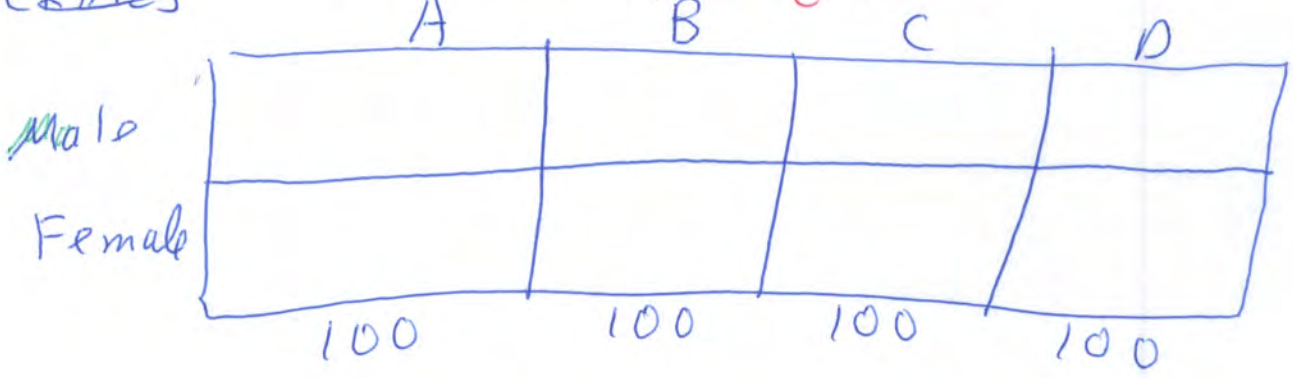
p_{ij} correspond to % of row total

Retrospective: Looking Backward

From response var Y to explanatory variable X

Marginal totals for Resp var Y are fixed by the design

~~CRIMES~~ type of crime



Retrospective design is natural for estimating P(X|Y). Often that's not what you want.

Bayes' Theorem

	Test Neg	Pos No Hosp	Pos Hospitalized	Pos Died
No vaccine	π_{11}	π_{12}	π_{13}	π_{14}
Vaccine	π_{21}	π_{22}	π_{23}	π_{24}
Boosted	$\pi_{31} = 1-\pi_{11}-\pi_{21}$	$\pi_{32} = 1-\pi_{12}-\pi_{22}$	$\pi_{33} = 1-\pi_{13}-\pi_{23}$	$\pi_{34} = 1-\pi_{14}-\pi_{24}$

$a = P(Y=1)$ $\quad b = P(Y=2)$ $\quad c = P(Y=3)$ $\quad d = P(Y=4)$

Probably known.

$$P(\text{Died} \mid \text{No Vaccine}) = P(Y=4 \mid X=1)$$

$$= \frac{P(Y=4 \cap X=1)}{P(X=1)} = \frac{P(X=1 \mid Y=4)\,P(Y=4)}{\sum_{j=1}^{4} P(X=1 \mid Y=j)\,P(Y=j)}$$

Bayes' Thm

$$= \frac{\pi_{@14} \cdot P(Y=4)}{\pi_{11}\,P(Y=1) + \pi_{12}P(Y=2) + \pi_{13}P(Y=3) + \pi_{14}P(Y=4)}$$

3 meanings of "unrelated"

- Conditional distribution of Y given $X = x$ does not depend on x. Prospective
- Conditional distribution of X given $Y = y$ does not depend on y. Retrospective
- X & Y are independent (Both random). Cross-sectional

Tuesday Oct 4

Prospective Design Linus Pauling's Vit C

	Cold	No cold	
Placebo	31 π_{11}	109 $1-\pi_{11}$	140
Vit C	17 π_{21}	122 $1-\pi_{21}$	139

Estimates

	Cold	No Cold	
Plac	$p_{11}=\frac{n_{11}}{n_{11}+n_{12}}$	$p_{12}=\frac{n_{12}}{n_{11}+n_{12}}$	$n_{11}+n_{12}$
Vit C	$p_{21}=\frac{n_{21}}{n_{21}+n_{22}}$	$p_{22}=\frac{n_{22}}{n_{21}+n_{22}}$	$n_{21}+n_{22}$

Under H_0: $P(\text{Cold} \mid \text{Placebo}) = P(\text{Cold} \mid \text{Vit C})$

$H_0: \pi_{11} = \pi_{21}$

π_{11}	$(1-\pi_{11})$
π_{21}	$1-\pi_{21}$

$$H_0 : \pi_{11} = \pi_2$$

$$\ell = \pi_{11}^{n_{11}} (1-\pi_{11})^{n_{12}} * \pi_{21}^{n_{21}} (1-\pi_{21})^{n_{22}}$$

$$\ell_0 = \pi^{n_{11}} (1-\pi)^{n_{12}} \pi^{n_{21}} (1-\pi)^{n_{22}}$$

$$= \pi^{n_{11}+n_{21}} (1-\pi)^{n_{12}+n_{22}}$$

$$\frac{\partial}{\partial \pi} \log \ell_0 = \frac{\partial}{\partial \pi} \left((n_{11}+n_{21}) \log \pi + (n_{12}+n_{22}) \log (1-\pi) \right)$$

$$= \frac{n_{11}+n_{21}}{\pi} + \frac{n_{12}+n_{22}}{1-\pi} (-1) \stackrel{!}{=} 0$$

$$\frac{n_{11}+n_{21}}{\pi} = \frac{n_{12}+n_{22}}{1-\pi}$$

$$\Leftrightarrow n_{11}+n_{21} - \pi(n_{11}+n_{21}) = \pi(n_{12}+n_{22})$$

$$\Leftrightarrow n_{11}+n_{21} = \pi(n_{11}+n_{21}+n_{12}+n_{22}) = \pi n$$

$$\Rightarrow \pi = \frac{n_{11}+n_{21}}{n}$$

	Cold	No cold
Plac	n_{11}	n_{12}
Vit C	n_{21}	n_{22}

$$\hat{\pi} = \frac{n_{11} + n_{21}}{n} = \text{Proportion who got cold}$$

The restricted MLE pools data from the two rows.

	C	N cold
$\tilde{\pi}$ = P	$\frac{n_{11} + n_{21}}{n}$	$1 - \frac{n_{11} + n_{21}}{n}$
VC	$\frac{n_{11} + n_{21}}{n}$	$1 - \frac{n_{11} + n_{21}}{n}$

$\hat{\mu}_0 = n\hat{\pi}$?? NO multiply by Row totals

	Cold	No cold	
Plac	$\frac{(n_{11}+n_{21})(n_{11}+n_{12})}{n}$	$\frac{(n_{12}+n_{22})(n_{11}+n_{12})}{n}$	$n_{11} + n_{12}$
Vit C	$\frac{(n_{11}+n_{21})(n_{21}+n_{22})}{n}$	$\frac{(n_{12}+n_{22})(n_{21}+n_{22})}{n}$	$n_{21} + n_{22}$

$$\hat{\mu}_{ij} = \frac{(\text{Row total})(\text{Col total})}{\text{total} \quad \text{total}} = \frac{n_{i+}\, n_{+j}}{n}$$

	Cold	No cold	
Placebo	31 (24.086)	109 (115.914)	140
Vit C	17 (23.914)	122 (115.086)	139
	48	231	279

Likelihood ratio test

$$G^2 = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log \frac{n_{ij}}{\hat{\mu}_{ij}}$$

$$= 2(31 \log \frac{31}{24.086} + \cdots + 122 \log \frac{122}{115.086})$$

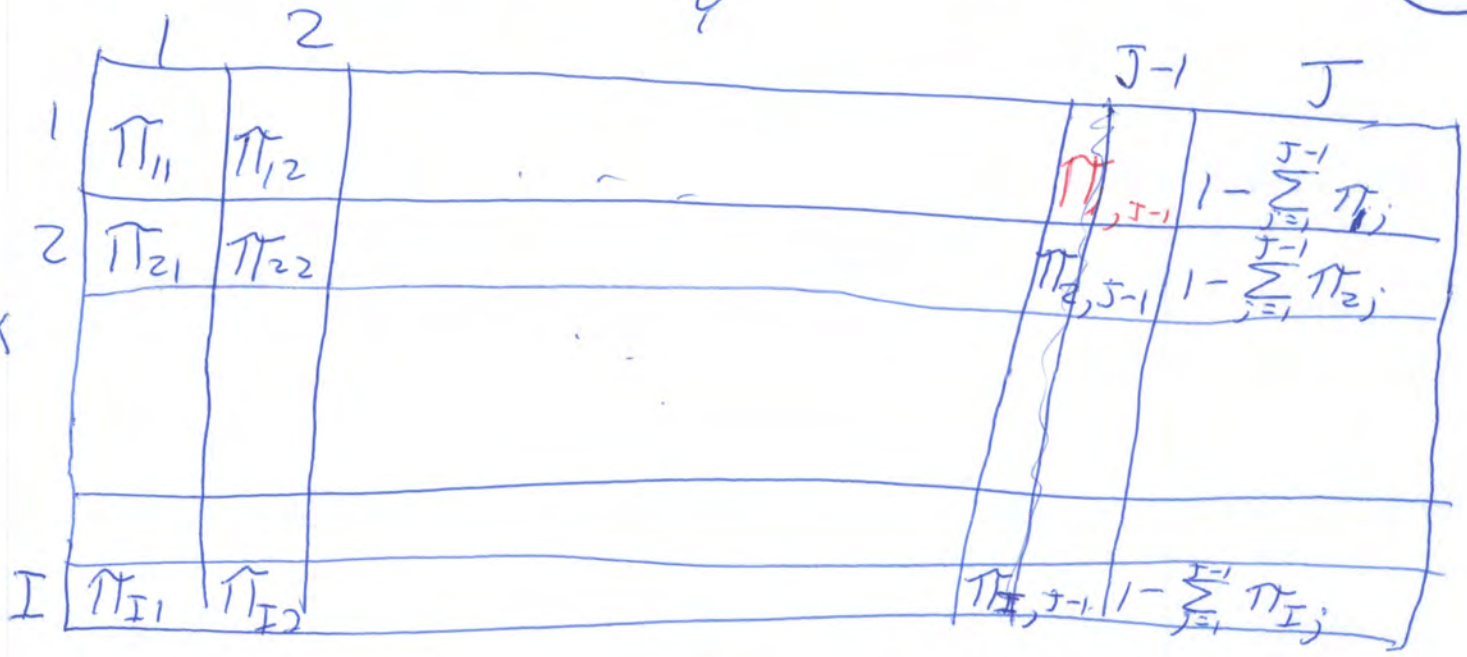
$$= 4.87 > \text{critical value of } 3.84 \text{ with 1 df}$$

$$\chi^2 = \sum\sum \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} = \frac{(31-24.086)^2}{24.086} + \cdots + \frac{(122-115.086)^2}{115.086}$$

$$= 4.81 > 3.84$$

Conclusion: Vit C reduced the chances of getting a cold.

For a general I by J table



Null hypothesis is that all the conditional distributions of Y are same for each $X = x$

$H_0: \pi_{11} = \pi_{21} = \cdots \pi_{I1} = \pi_1$ col 1

and $\pi_{12} = \pi_{22} = \cdots \pi_{I2} = \pi_2$ col 2

$\pi_{1,J-1} = \pi_{2,J-1} = \cdots = \pi_{I,J-1} = \pi_{J-1}$ col I

Likelihood $\pi_1^{n_{+1}} \pi_2^{n_{+2}} \cdots \pi_{J-1}^{n_{+J-1}} \left(1 - \sum_{j=1}^{J-1} \pi_j\right)^{n_{+J}}$

Multinomial

Likelihood is multinomial. MLEs

$$p_j = \frac{n_{+j}}{n} = \hat{\pi}_{i|j}, \quad i = 1, \ldots, I$$

To get estimated expected frequencies, multiply by row total n_{i+}, so

$$\hat{\mu}_{ij} = \hat{\pi}_{i|j} n_{i+} = \frac{n_{+j}}{n} n_{i+} = \frac{n_{i+} n_{+j}}{n}$$

$$\frac{(\text{Row total})(\text{Col total})}{\text{Total total}}$$

For a RETROSPECTIVE design in which column probabilities add to one, just exchange rows & columns and get same result

Same expected frequencies for Prospective & Retrospective design.

Same tests.

df: $I-1$ equals signs in each of $J-1$ columns, so

$$df = (I-1)(J-1)$$

Testing Independence of X & Y for Cross-sectional data

Course	Passed: No	Passed: Yes	Total
Catch-up	π_{11}	π_{12}	π_{1+}
Mainstream	π_{21}	π_{22}	π_{2+}
Elite	π_{31}	π_{32}	$1-\pi_{1+}-\pi_{21}$
Total	π_{+1}	$1-\pi_{+1}$	1

$H_0: \pi_{ij} = \pi_{i+}\,\pi_{+j}$

There are $(I-1)+(J-1)$ free parameters under H_0

MLEs of Marginal probabilities are

$$\hat{\pi}_{i+} = p_{i+} = \frac{n_{i+}}{n} \text{ and } \hat{\pi}_{+j} = p_{+j} = \frac{n_{+j}}{n}$$

Restricted MLEs of π_{ij} are

$$\hat{\pi}_{ij} = p_{i+}\,p_{+j}, \quad \hat{\mu}_{ij} = p_{i+}\,p_{+j}\cdot n$$

$$\hat{\mu}_{ij} = n\,\frac{n_{i+}}{n}\,\frac{n_{+j}}{n} = \frac{n_{i+}\,n_{+j}}{n}$$

$= (\text{Row total})(\text{col total})/\text{total}$ SAME

df # of parameters under unrestricted model is $IJ-1$

Under H_0, there are $I-1+J-1$ parameters

df is the difference in # of parameters

$$df = IJ-1-(I-1+J-1)$$

$$= (I-1)(J-1) \quad \text{same df}$$

So test statistics, df, p-values are SAME for

Prospective

Retrospective

Cross-sectional

designs