- From Agresti (2013, 3rd ed.)

Table 2.11 Data for Exercise 2.15 on Graduate Admissions

Department	Male: Yes	Male: No	Female: Yes	Female: No
A	512	313	89	19
B	353	207	17	8
C	120	205	202	391
D	138	279	131	244
E	53	138	94	299
F	22	351	24	317
Total	1198	1493	557	1278

3-way contingency table

check table in LNp.2-4

825 = 512 + 313 | 108 = 89 + 19

373 = 22 + 351 | 341 = 24 + 317

fixed? random?

Y: resp. (Admitted); X: expl. (Gender), (Department)

1st Yes, 2nd Yes, ..., No — 2 levels; M, F, ..., M, F — 2 levels; A, A, ..., F, F — nominal

↓ sufficient statistics

N: count resp. (Admitted); X: expl. (Gender), (Department)

Yes, No — 2 levels; M, F, ..., M, F — 2 levels; A, A, ..., F, F — nominal

N_{11}, N_{12}; N_{21}, N_{22}; ...; N_{I1}, N_{I2}; N_{21}, N_{22}

binomial?

Poisson (LNp.2-7~8)

512 = N_{1A1} (M, A, Yes)
313 = N_{1A2} (M, A, No)
89 = N_{2A1} (F, A, Yes)
19 = N_{2A2} (F, A, No)

$N_{G,D,A} \overset{\text{indep}}{\sim} \text{Poisson}(\lambda_{G,D,A})$ — parameter

Q: how (G, D, A) influence $\lambda_{G,D,A}$'s?
or how (G, D) influence →

Note. $P_{G,D,1} = \frac{\lambda_{G,D,1}}{\lambda_{G,D,1} + \lambda_{G,D,2}}$

- From Agresti (2013, 3rd ed.)

- In data collection, collect information of X|Y

Table 2.5 Cross-Classification of Smoking by Lung Cancer ← Y(resp?expl.?) X(expl.? resp.?)

Smoker	Lung Cancer: Cases (Yes)	Lung Cancer: Controls (No)
Yes	688	650
No	21	59
Total	709	709

retro-spective study

cf. table3.1 in LNp.2-8

random → 1338 = 688 + 650; 80 = 21 + 59

fixed

resp. (smoker): 1st Yes, 2nd No, ..., Yes, No — random

expl. (cancer): Yes, Yes, ..., No, No — fixed

Suff. Stat. count: resp. X(Smoker), expl. Y(cancer)

	Yes	No
Yes (1)	N_{11}	N_{12}
No (2)	N_{21}	N_{22}

★ Statistical modeling:
N_{11} (Y=1) ~ binomial (709, P_1) ← P(X=Yes|Y=1)
N_{21} (Y=2) ~ binomial (709, P_2) ← P(X=Yes|Y=2)
⇒ can study problem related to X|Y

pro-: X is fixed in advance, like in DOE
retro-: y (categorical response) is fixed.

Under what circumstances, we can get useful information about Y|X from X|Y?

⊙ But, in data analysis, interested in problems related to Y|X, e.g., how smoking (X=expl) influence the probability of getting cancer (Y=resp.)

Table 3.9 Data for Fisher's Tea-Tasting Experiment

Poured First (PF)	Guess: Milk	Guess: Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	

fixed

resp. (Guess): 1st M, 2nd T, ..., 7th M, 8th T — random

expl. (PF): M, M, ..., T, T — fixed

always 4 M, 4 T

Sufficient statistics

count: resp. (Guess), expl. (PF)

	M	T
M (1)	N_{11}	N_{12}
T (2)	N_{21}	N_{22}

If N_{11} is known, the rest counts are known.

- statistical modeling under random guess: $N_{11} \sim$ hypergeometric (4, 4, 4)
  - 4: # of balls drawn (claim M)
  - 4: # of white balls (T)
  - 4: # of black balls (M)
  - # of black balls drawn

what if not?

- From Agresti (2002, 2rd ed.)

Total (K) fixed ② ? or random ③ ?

row sum ($K_{\underset{\sim}{X}}$) fixed ① ? or random ② ?

Information: ① < ② < ③

cf. table3.1 (LNp.2-8) table2.11 (LNp.2-10)

TABLE 7.1 Primary Food Choice of Alligators

Lake	Gender	Size (m)	Fish	Invertebrate	Reptile	Bird	Other	
Hancock	Male	≤ 2.3 1	7	1	0	0	5	= 13
		> 2.3 2	4	0	0	1	2	
	Female	≤ 2.3 3	16	3	2	2	3	+
		> 2.3 4	3	0	1	2	3	
Oklawaha	Male	≤ 2.3 5	2	2	0	0	1	
		> 2.3 6	13	7	6	0	0	
	Female	≤ 2.3 7	3	9	1	0	2	+
		> 2.3 8	0	1	0	1	0	
Trafford	Male	≤ 2.3 9	3	7	1	0	1	
		> 2.3 10	8	6	6	3	5	
	Female	≤ 2.3 11	2	4	1	1	4	+
		> 2.3 12	0	1	0	0	0	
George	Male	≤ 2.3 13	13	10	0	2	2	
		> 2.3 14	9	0	0	1	2	
	Female	≤ 2.3 15	3	9	1	0	1	+
		> 2.3 16	8	1	0	0	1	= 10

= 219

count resp.	expl. (X)	(Food)
$N_{X_1,E}$	X_1	E
$N_{X_1,O}$	X_1	O

③ $N_{\underset{\sim}{X},Food} \overset{indep}{\sim} Poisson(\lambda_{\underset{\sim}{X},Food})$

Q: how ($\underset{\sim}{X}$, Food) influence $\lambda_{\underset{\sim}{X},Food}$'s ?

② $N_{\underset{\sim}{X},Food} \sim$ multinomial (80×1 vector) $(K, P_{\underset{\sim}{X},Food})$, $\sum_{\underset{\sim}{X}}\sum_{Food} P_{\underset{\sim}{X},Food} = 1$

- data collection: if data from survey sampling, $\underset{\sim}{X}$ & Y are random
- data analysis: in regression-type analysis, Y: random; $\underset{\sim}{X}$: fixed (conditioned)

Suff. Stat. → $N_{\underset{\sim}{X}}$ → count resp. (Food)

nominal

expl. $\underset{\sim}{X}$ (Lake)(Gender)(Size)

E I R B O

$N_{X_1} = (N_{1E}, N_{1I}, N_{1R}, N_{1B}, N_{1O})$

H M ≤ X_1

Y resp. (Food) | expl. $\underset{\sim}{X}$ (Lake) (Gender) (Size)

	Food	Lake	Gender	Size
1st	F	O	M	≤
2nd	R	G	F	>
	nominal	nominal	2 levels	2 levels

① $N_{\underset{\sim}{X}} \sim$ multinomial $(k_{\underset{\sim}{X}}, P_{\underset{\sim}{X}} \equiv (P_{\underset{\sim}{X}E}, P_{\underset{\sim}{X}I}, P_{\underset{\sim}{X}R}, P_{\underset{\sim}{X}B}, P_{\underset{\sim}{X}O}))$

5×1 vector $P_{\underset{\sim}{X}E} + P_{\underset{\sim}{X}I} + P_{\underset{\sim}{X}R} + P_{\underset{\sim}{X}B} + P_{\underset{\sim}{X}O} = 1$ ($\underset{\sim}{X}$ fixed)

Q: how $\underset{\sim}{X}$, Food influence $P_{\underset{\sim}{X},Food}$?

- From Agresti (2002, 2rd ed.)

$Z_{\underset{\sim}{X}}$: continuous lifetime distribution

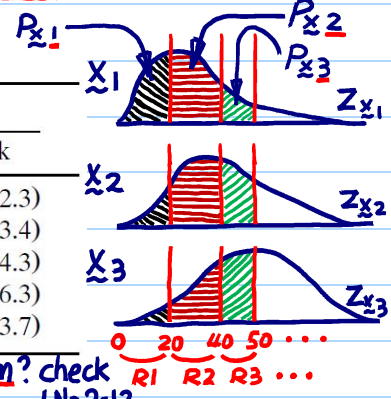


TABLE 7.7 Life-Length Distribution of U.S. Residents (Percent),[a] 1981

	Life Length	Males — White (race)	Males — Black (race)	Females — White (race)	Females — Black (race)
R1	0–20	2.4 (2.4)	3.6 (4.4)	1.6 (1.2)	2.7 (2.3)
R2	20–40	3.4 (3.5)	7.5 (6.4)	1.4 (1.9)	2.9 (3.4)
R3	40–50	3.8 (4.4)	8.3 (7.7)	2.2 (2.4)	4.4 (4.3)
R4	50–60	17.5 (16.7)	25.0 (26.1)	9.9 (9.6)	16.3 (16.3)
R5	Over 65	72.9 (73.0)	55.6 (55.4)	84.9 (84.9)	73.7 (73.7)

100% (+)

assume column sums are fixed. Q: What if they are random? check LNp2-12

Y resp. (Life) | expl. $\underset{\sim}{X}$ (Gender) (Race)

	Life	Gender	Race
1st	R3	M	W
2nd	R5	F	W
	...	...	...
	R4	F	B
	R1	M	B
	disc. inte.	2 levels	2 levels

Suff. Stat. → $N_{\underset{\sim}{X}}$ → count resp. (Life)

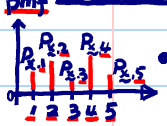
disc. inte.

expl. $\underset{\sim}{X}$ (Gender)(Race)

$R1, R2, R3, R4, R5$

$N_{\underset{\sim}{X}1} = (N_{11}, N_{12}, N_{13}, N_{14}, N_{15})$

M	W ← $\underset{\sim}{X}_1$
F	W ← $\underset{\sim}{X}_2$
M	B ← $\underset{\sim}{X}_3$
F	B ← $\underset{\sim}{X}_4$



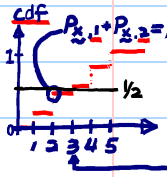
- statistical modeling:

$N_{\underset{\sim}{X}} \sim$ multinomial $(K_{\underset{\sim}{X}}, P_{\underset{\sim}{X}} \equiv (P_{\underset{\sim}{X}1}, P_{\underset{\sim}{X}2}, P_{\underset{\sim}{X}3}, P_{\underset{\sim}{X}4}, P_{\underset{\sim}{X}5}))$

has some structure ∵ discrete interval categories

$\underset{\sim}{X}$ relationship $Z_{\underset{\sim}{X}}$

$\underset{\sim}{X}$ relationship $P_{\underset{\sim}{X}}$



Q: how $\underset{\sim}{X}$ influence $P_{\underset{\sim}{X}}$ (pmf) or cdf of $P_{\underset{\sim}{X}}$? (cf.) continuous, ordinal, nominal

e.g. median is meaningful

- From Agresti (2013, 3rd ed.)

Table 8.7 Outcomes for Pregnant Mice in Developmental Toxicity Study

Concentration (mg/kg per day)	Response (abnormality) Nonlive		Malformation		Normal		
0 (controls)	15	+	1	+	281	=	297
62.5	17	•	0	•	225	•	•
125	22	•	7	•	283	•	•
250	38	•	59	•	202	•	•
500	144	+	132	+	9	=	285

assume fixed.

Q: What if they're random? Check LN p.2-12

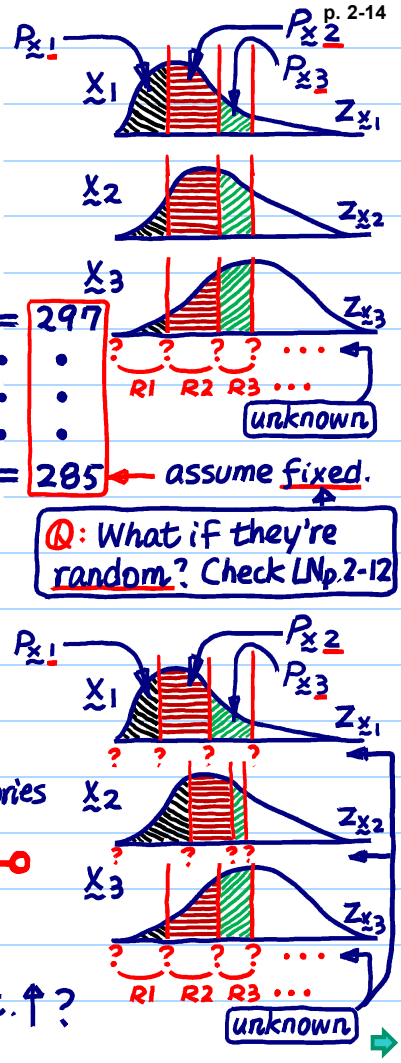
Y → resp. (Abnormality): 1st, 2nd, ..., Mal, Non — ordinal

X → expl. (Concentration): 0, 0, ..., 500, 500 — continuous

Suff. Stat. $N_{\sim}$ → count resp. (Abnormality) — ordinal

$N_{x_1} = (N_{11}, N_{12}, N_{13})$ for Non, Mal, Nor

expl. X (Conc.): x_1 = 0, 62.5, ..., 500



- statistical modeling:

$N_x \sim \text{multinomial}(K_x, P_x \equiv (P_{x1}, P_{x2}, P_{x3}))$

has some structure ∵ ordinal categories

cf. disc. inte. / nominal

Q: how X influence P_x, or cdf of P_x?

└ say, whether abnormality ↑ when conc. ↑ ?

- From Agresti (2002, 2nd ed.)

TABLE 2.8 Cross-Classification of Job Satisfaction by Income

	Income (dollars)	Job Satisfaction: Very Dissatisfied		Little Dissatisfied		Moderately Satisfied		Very Satisfied		
R1	< 15,000	1	+	3	+	10	+	6	=	20
R2	15,000–25,000	2	•	3	•	10	•	7	•	22
R3	25,000–40,000	1	•	6	•	14	•	12	•	33
R4	> 40,000	0	+	1	+	9	+	11	=	21

assume fixed.

Q: What if they're random? Check LN p.2-12

Y → resp. (Satisfaction): 1st, 2nd, ..., MS, VS — ordinal

X → expl. (Income): R1, R3, ..., R2, R4 — disc. inte.

Suff. Stat. $N_{\sim}$ → count resp. (Satisfaction) — ordinal

VD, LD, MS, VS

$N_{R1} = (N_{11}, N_{12}, N_{13}, N_{14})$

R1, R2, R3, R4

expl. X (Income) ← disc. inte.

Q: use what base functions?

- statistical modeling:

$N_x \sim \text{multinomial}(K_x, P_x \equiv (P_{x1}, P_{x2}, P_{x3}, P_{x4}))$

Q: how X influence P_x (pmf) or cdf of P_x?

└ say, whether Satisfaction ↑ when Income ↑ ?

disc. inte. X, e.g.

- continuous base function on 7500, 20000, ... — cont. X
- Helmert codings on R1, ..., R4 — ordi. X
- Treatment or Sum codings on R1, ..., R4 — nomi. X

❖ **Reading**: Agresti (2013), 1.1; Faraway (2006, 1st ed.), Preface (page v)