Georgian Implementation Science Fogarty Training (GIFT)

Ilia State University & Yale University

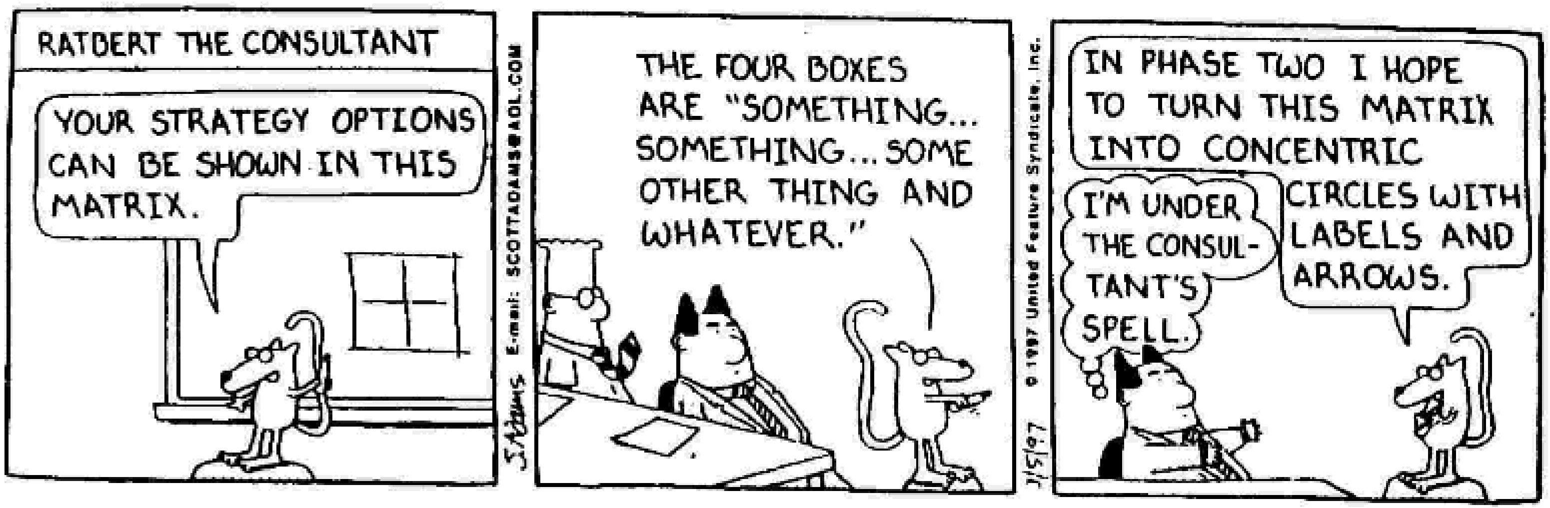
Summummer Bootcamp in Implementation Science & Biostatistics

July 21–24, 2026
Tbilisi, Georgia

Yale Center for Analytical Sciences

Categorical Data

DILBERT


RATBERT THE CONSULTANT
YOUR STRATEGY OPTIONS CAN BE SHOWN IN THIS MATRIX.
S. Adams E-mail: SCOTTADAMS@AOL.COM
THE FOUR BOXES ARE "SOMETHING... SOMETHING... SOME OTHER THING AND WHATEVER."
1/5/97 © 1997 United Feature Syndicate, Inc.
IN PHASE TWO I HOPE TO TURN THIS MATRIX INTO CONCENTRIC
I'M UNDER THE CONSUL-TANT'S SPELL.
CIRCLES WITH LABELS AND ARROWS.

Wait-and-See Prescription for the Treatment of Acute Otitis Media

A Randomized Controlled Trial

Table 2. Clinical Outcomes According to Group Designation*

Outcome	WASP Group (n = 132)	SP Group (n = 133)	Unadjusted Difference (95% CI)	Adjusted Difference (95% CI)†	*P* Value‡
	4- to 6-Day Follow-up				
Parent did not fill the antibiotic prescription, No. (%)	82 (62)	17 (13)	4.86 (3.06 to 7.73)‡	4.80 (3.57 to 5.85)‡	<.001
Days postenrollment prescription was filled, mean (SD)	2.0 (0.8)	1.2 (0.7)	0.77 (0.53 to 1.06)	0.75 (0.53 to 1.03)	<.001
Otalgia, No. (%)	85 (64)	89 (67)	0.96 (0.81 to 1.15)‡	1.01 (0.83 to 1.17)‡	.96
Total days of otalgia, mean (SD)	2.4 (1.2)	2.0 (1.2)	0.42 (0.07 to 0.78)	0.43 (0.07 to 0.80)	.02
Use of otic analgesia, No. (%)	123 (93)	120 (90)	1.03 (0.96 to 1.11)‡	1.04 (0.94 to 1.08)‡	.34
Total days of otic analgesia use, mean (SD)	2.9 (1.3)	2.8 (1.5)	0.11 (−0.24 to 0.46)	0.11 (−0.25 to 0.46)	.56
Fever, No. (%)	42 (32)	46 (35)	0.92 (0.65 to 1.30)‡	1.04 (0.70 to 1.44)‡	.85
Total days of fever, mean (SD)	2.0 (1.1)	1.7 (1.0)	0.24 (−0.22 to 0.63)	0.33 (−0.13 to 0.73)	.20
Use of ibuprofen or acetaminophen, No. (%)	118 (89)	110 (83)	1.08 (0.98 to 1.19)‡	1.09 (0.98 to 1.14)‡	.11
Total days of ibuprofen or acetaminophen, mean (SD)	2.6 (1.3)	2.4 (1.3)	0.18 (−0.16 to 0.52)	0.22 (−0.13 to 0.58)	.22
Diarrhea, No. (%)	10 (8)	31 (23)	0.33 (0.17 to 0.64)‡	0.30 (0.14 to 0.64)‡	<.001
Total days of diarrhea, mean (SD)	2.3 (1.4)	2.0 (1.3)	0.33 (−0.60 to 1.19)	0.20 (−0.80 to 1.15)	.76
Vomiting, No. (%)	15 (11)	15 (11)	1.01 (0.51 to 1.98)‡	1.24 (0.59 to 2.41)‡	.56
Total days of vomiting, mean (SD)	1.5 (0.9)	1.2 (0.6)	0.33 (−0.20 to 0.80)	0.60 (0.06 to 1.15)	.02
Unscheduled visit(s) to a clinician, No. (%)	13 (10)	11 (8)	1.19 (0.55 to 2.56)‡	1.17 (0.51 to 2.51)‡	.70

Measures of Association

- Indicate the strength of the relation between 2 variables (example. recurrence and treatment)

In a prospective study to assess the relationship between smoking and the subsequent risk of lung cancer, a cohort of 2000 people were followed.

	Lung Cancer	**No Lung Cancer**	
Smoking	450 (A)	150 (B)	600 (A+B)
No Smoking	150 (C)	1250 (D)	1400 (C+D)
	600 (A+C)	1400 (B+D)	2000 (n)

Binary Outcome

		Recurrence	No Recurrence	
Trt	**A**	a=31	b=148	r_1=179
	B	c=283	d=330	r_2=613
		c_1=314	c_2=478	N=792

Measures of Association
Prospective Studies

- Risk Ratio

$$A = \frac{a}{a+b} = \frac{31}{179} = 0.17 \qquad B = \frac{c}{c+d} = \frac{283}{613} = 0.46$$

$$RR = \frac{\frac{\mathrm{a}}{\mathrm{a+b}}}{\frac{\mathrm{c}}{\mathrm{c+d}}} = \frac{0.17}{0.46} = 0.37$$

$$\frac{1}{0.37} = 2.7$$

Measures of Association
Prospective Studies

- Risk Difference

$$\frac{a}{a+b}-\frac{c}{c+d}=0.17-0.46=-0.29$$

Absolute Risk Reduction (ARR) of 0.29

- Number Needed to Treat (NNT)

$$\mathrm{NNT}=\frac{1}{\mathrm{ARR}}=3.45$$

  - How many subjects do we have to treat to prevent 1 recurrence

Measures of Association
Prospective Studies

- If risk difference is reversed

$$\frac{c}{c+d} - \frac{a}{a+b} = 0.46 - 0.17 = 0.29$$

  – Absolute Risk Increase (ARI)

- Number needed to harm (NNH)

$$\mathrm{NNH} = \frac{1}{\mathrm{ARI}}$$

  – How many patients do we need to treat to observe 1 recurrence

Measures of Association
Prospective Studies

Odds $\rightarrow$ Probability of recurrence/Probability of no recurrence

$$\frac{\mathrm{P(R \mid Trt\ A)}}{1 - \mathrm{P(R \mid Trt\ A)}} \longleftarrow \mathrm{P(\overline{R} \mid Trt\ A)}$$

- Odds Ratio – the odds in one group compared to the odds in another group
  - Example

$$OR = \frac{\mathrm{P(R \mid Trt\ A)} / \mathrm{P(\overline{R} \mid Trt\ A)}}{\mathrm{P(R \mid Trt\ B)} / \mathrm{P(\overline{R} \mid Trt\ B)}} \rightarrow \frac{ad}{bc} = \frac{31 \times 330}{283 \times 148} = 0.24$$

Measures of Association
Case Control Studies

- Since we select subjects based on "disease status", $^{a}/_{a+b}$ and $^{c}/_{c+d}$ are meaningless

	Recurrence	No Recurrence	
Trt **A**	a	b	r_1
B	c	d	r_2
	c_1	c_2	N

Measures of Association
Case Control Studies

- Estimate an exposure OR

$$OR = \frac{P(\text{Trt A} \mid R) / P(\text{Trt B} \mid R)}{P(\text{Trt A} \mid \bar{R}) / P(\text{Trt B} \mid \bar{R})} \quad \rightarrow \frac{ad}{bc}$$

- The flexibility of the OR across study designs as well as its benefits in multivariable analyses makes it an attractive and much utilized measure

OR Approximates RR When Rare Disease

- If rare disease,
  - a is small relative to b

$$\frac{a}{a+b} \approx \frac{a}{b}$$

  - c is small relative to d

$$\frac{c}{c+d} \approx \frac{c}{d}$$

Confidence Interval for RR

- When expected cell frequencies are >5 then we can use large sample approximation

$$100(1-\alpha)\%\ \mathrm{CI}$$

$$\ln \mathrm{RR} \ \pm z_{\alpha/2} \mathrm{SE}(\ln \mathrm{RR})$$

Where

$$\mathrm{SE}(\ln \mathrm{RR}) \ = \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}$$

Confidence Interval for RR

- 95% CI for Example

$$\ln 0.37 \pm 1.96\sqrt{\frac{1}{31}-\frac{1}{179}+\frac{1}{283}-\frac{1}{613}}$$

$$(-1.32, -0.66)$$

$$\downarrow$$

$$\left(e^{-1.32}, \mathrm{e}^{-0.66}\right) = (0.27, 0.52)$$

- 95% sure this interval covers the true population RR

Confidence Interval for OR

$$100(1-\alpha)\%\ \text{CI}$$

$$\ln\text{OR} \pm z_{\alpha/2}\text{SE}(\ln\text{OR})$$

Where

$$\text{SE}(\ln\text{OR}) = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

- 95% CI for Example

$$\ln 0.24 \pm 1.96\sqrt{\frac{1}{31}+\frac{1}{148}+\frac{1}{283}+\frac{1}{330}}$$

$$(-1.84, \text{-}1.02) \xrightarrow{\exp} (0.16, 0.36)$$

An OR > 1 means:

A. Controls had the same exposure as cases

B. Controls were not selected correctly

C. Cases were more likely to have had the exposure of interest

D. Exposed individuals developed disease at a greater frequency

Hypothesis Testing for Contingency Tables

- In the 2X2 table, the probability of recurrence:

		Recurrence	No Recurrence	
Trt	**A**	a=31	b=148	r_1=179
	B	c=283	d=330	r_2=613
		c_1=314	c_2=478	N=792

$$P(R) = 314/792 = 0.40$$

Hypothesis Testing for Contingency Tables

- If there is no relation between treatment and recurrence – expect 40% recurrence for treatment A and 40% recurrence for treatment B

$$P(R \mid \text{Trt } A) = P(R \mid \text{Trt } B) = P(R)$$

Chi-square Test of Independence

- Under the assumption of independence

$$P(\text{R and Trt A}) = P(\text{R}) \times P(\text{Trt A})$$

  - For each cell, we calculate the expected number of counts under the assumption of independence

$$E(a) = N[P(R) \times P(\text{Trt A})] = N\left[\frac{c_1}{N} \times \frac{r_1}{N}\right] = \frac{r_1 c_1}{N}$$

$$= \frac{314 \times 179}{792} = 70.97$$

Chi-square Test of Independence

$$E(b) = \frac{r_1 c_2}{N} = \frac{179 \times 478}{792} = 108.03$$

$$E(c) = \frac{r_2 c_1}{N} = \frac{613 \times 314}{792} = 243.03$$

$$E(d) = \frac{r_2 c_2}{N} = \frac{613 \times 478}{792} = 369.97$$

- These expected counts are what would have happened if independence was true (i.e. if trt not related to recurrence)

Chi-square Test of Independence

- Compare observed to expected

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

$$= \frac{(31 - 70.79)^2}{70.97} + \frac{(148 - 108.03)^2}{108.03} + \frac{(283 - 243.03)^2}{243.03} + \frac{(330 - 369.97)^2}{369.97}$$

$$= 48.19$$

- Compare chi-square statistic to critical chi-square:

$$\chi^2_\alpha \text{ with } (r-1)(c-1) \text{ d.f.}$$

Chi-square Test of Independence

- From table of critical chi-square values (p 356)

$$\chi^2_{0.05,1} = 3.84$$

  - Reject Null

Assumptions for Chi-square Test of Independence

- Random and independent sampling of subjects
- All cells should have expected counts > 5
  - If don’t satisfy – use Fisher’s exact test

Evaluation and Adjustment for Confounding

- Example – investigation of maternal tranquilizer use and the risk of malformation
  - Smoking is associated with increased risk for malformation
  - Tranquilizer users are more likely to smoke

Evaluation and Adjustment for Confounding

- Stratum-specific OR

			DISEASE		Total	
SMOKE			No	Yes		
Non	TRANQ	No	1916	848	2764	→ 1.44
		Yes	47	30	77	
	Total		1963	878	2841	
1-20	TRANQ	No	826	361	1187	→ 3.70
		Yes	13	21	34	
	Total		839	382	1221	
21+	TRANQ	No	122	75	197	→ 3.66
		Yes	8	18	26	
	Total		130	93	223	

Mantel-Haenszel Odds Ratio

- Summary OR that adjusts for a confounder

$$\text{OR}_{\text{MH}} = \frac{\sum_k a_k d_k / n_k}{\sum_k b_k c_k / n_k}$$

Estimate			2.154
ln(Estimate)			.767
Std. Error of ln(Estimate)			.176
Asymp. Sig. (2-sided)			.000
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	1.525
		Upper Bound	3.041
	ln(Common Odds Ratio)	Lower Bound	.422
		Upper Bound	1.112

Mantel-Haenszel Odds Ratio

- OR_{MH} (common odds ratio) is valid under the assumption that the stratum-specific ORs are similar
  - Can test this using the Breslow-Day Test for Homogeneity of Odds Ratios

Statistics		Chi-Squared	df	Asymp. Sig. (2-sided)
Conditional Independence	Cochran's	19.619	1	.000
	Mantel-Haenszel	18.776	1	.000
Homogeneity	Breslow-Day	6.626	2	.036
	Tarone's	6.625	2	.036

MAYBE AN R EXAMPLE HERE Chi-square and MH

Multiple Logistic Regression

- Used to evaluate the relation between a binary outcome and multiple independent variables of all types
- Example – an ED wants to identify using a number of different variables, which patients were most likely to have a BAC>50 mg/dl

Multivariable Model, Derivation Sample

Table 2. Multivariable Model, Derivation Sample

Patient Characteristic	% With HN	Adjusted Odds Ratio (95% CI, Adjusted): Model 1, Primary	P Value	Model 2	P Value
Race					
Nonblack	53.7	2.1 (1.0-4.4)	.06	2.2 (1.0-4.6)	.046
Black	39.2	1 [Reference]		1 [Reference]	
History of recurrent urinary tract infections					
Yes	76.0	2.7 (0.8-8.5)	.10	2.3 (0.7-7.1)	.16
No	46.3	1 [Reference]		1 [Reference]	
Diagnosis consistent with possible obstruction[a]					
Yes	67.4	2.4 (1.2-4.6)	.01	2.4 (1.2-4.7)	.009
No	36.0	1 [Reference]		1 [Reference]	
History of HN[b]					
Yes	90.3	11.1 (3.0-41.3)	<.001	11.7 (3.0-45.2)	<.001
No	42.6	1 [Reference]		1 [Reference]	
History of CHF					
No	52.7	2.1 (0.8-5.2)	.12	2.0 (0.8-5.0)	.14
Yes	37.1	1 [Reference]		1 [Reference]	
History of prerenal AKI, use of pressors or history of sepsis					
No	53.0	2.3 (0.9-6.2)	.10	NA	NA
Yes	35.3	1 [Reference]		NA	NA
History of prerenal AKI, use of pressors, history of sepsis, or hypotension					
No	60.2	NA	NA	2.1 (0.9-3.6)	.04
Yes	40.2	NA	NA	1 [Reference]	
Exposure to nephrotoxic medications prior to AKI[c]					
No	62.2	2.1 (1.0-3.85)	.053	1.8 (0.9-3.6)	.09
Yes	38.2	1 [Reference]		1 [Reference]	
Model characteristic					
AIC, score		237		235	
Accuracy, %		74		73	
C statistic		0.79		0.80	

Abbreviations: AIC, Akaike information criterion; AKI, acute kidney injury; CHF, congestive heart failure; CI, confidence interval; HN, hydronephrosis; NA, not applicable.

[a] Diagnosis consistent with possible obstruction: benign prostatic hyperplasia, abdominal or pelvic cancer, neurogenic bladder, single functional kidney, or previous pelvic surgery.

[b] History of HN: documented history of HN in the medical record or any imagining history of HN in the 2 years prior to the current RUS.

[c] Nephrotoxic medications: aspirin (>81 mg/d), diuretic, angiotensin-converting enzyme inhibitor, or intravenous vancomycin.

Licurse, A. et al. Arch Intern Med 2010;170:1900-1907.

ARCHIVES OF INTERNAL MEDICINE

Multiple Logistic Regression

- The probability of observing a particular outcome is defined as:

$$p = \frac{1}{1 + e^{\left[-(\beta_0 + \beta_1 x_1 + \dots \beta_k x_k)\right]}}$$

- Using the logistic regression results in straightforward model interpretation

$$\ln odds = \beta_0 + \beta_1 x_1 + \dots \beta_k x_k$$

Multiple Logistic Regression

- Interpretation
  - Example 1– let x_1 be a 0/1 indicator of gender
    - 0=M; 1=F
    - $\ln(\text{odds}_M)=\beta_0+\beta_1(0)=\beta_0$
    - $\ln(\text{odds}_F)=\beta_0+\beta_1(1)=\beta_0+\beta_1$

$$\text{OR} = \frac{\text{odds}_F}{\text{odds}_M} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Multiple Logistic Regression

- Interpretation
  - Example 2– let x_1 be a continuous variable such as age
    - β_{age}=0.083
    - OR for a 40y compared to a 25y old

$$e^{\beta_{age}(x_i - x_j)} = e^{0.083(40-25)}$$

$$OR = 3.47$$

Multiple Logistic Regression

- Parameters are estimated using maximum likelihood (as opposed to least squares)
- CI for OR

$$e^{\beta_1 \pm z_{\alpha/2} se(\beta_1)}$$

- Hypothesis Test – Wald statistic
  - H_0: $\beta_1=0$

$$\left(\frac{\hat{\beta}_k}{se_{\beta_k}} \right)^2 \sim \chi^2_{1d.f}$$

MAYBE AN R EXAMPLE HERE LOGISTIC REGRESSION