Historical Highlights
in the Development of Categorical Data Analysis

Alan Agresti

Department of Statistics, University of Florida

UW Statistics 50th Anniversary

Karl Pearson (1857-1936)

Karl Pearson (1900) *Philos. Mag.*

Introduces chi-squared statistic

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$df = \text{no. categories} - 1$$

- testing values for multinomial probabilities

  (Monte Carlo roulette runs)

- testing fit of Pearson curves
- testing statistical independence in $r \times c$ contingency table ($df = rc - 1$)

Karl Pearson (1904)

Advocates measuring association in contingency tables by approximating the correlation for an assumed underlying continuous distribution

- tetrachoric correlation (2×2, assuming bivariate normality)
- contingency coefficient $\sqrt{\frac{X^2}{X^2+n}}$ based on X^2 for testing independence in $r \times c$ contingency table
- introduces term "contingency" as a "measure of the total deviation of the classification from independent probability."

George Udny Yule (1871-1951)

(1900) *Philos. Trans. Royal Soc. London*

(1912) *JRSS* (1903) *Biometrika* “Simpson’s paradox”

Advocates measuring association using odds ratio

n_{11}	n_{12}
n_{21}	n_{22}

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}} \qquad Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}} = (\hat{\theta} - 1)/(\hat{\theta} + 1)$$

“At best the normal coefficient can only be said to give us. . . a hypothetical correlation between supposititious variables. The introduction of needless and unverifiable hypotheses does not appear to me a desirable proceeding in scientific work.”

(1911) *An Introduction to the Theory of Statistics* (14 editions)

K. Pearson, with D. Heron (1913) *Biometrika*

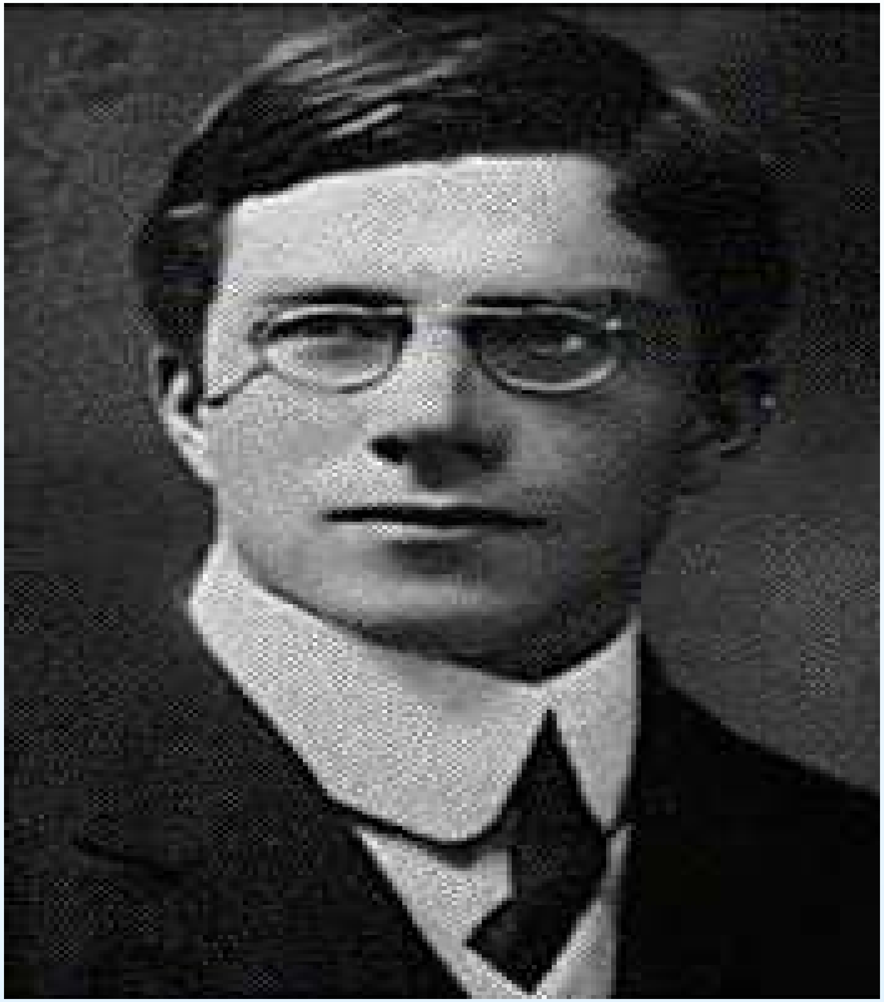
"Unthinking praise has been bestowed on a textbook which can only lead statistical students hopelessly astray."

⋮

"If Mr. Yule's views are accepted, irreparable damage will be done to the growth of modern statistical theory . . . Yule's Q has never been and never will be used in any works done under my supervision. . . . Yule must withdraw his ideas if he wishes to maintain any reputation as a statistician."

and so on, for 150 pages

Ronald A. Fisher (1890-1962)



R. A. Fisher (1922)

- Introduces concept of degrees of freedom with geometrical argument.
- Shows that when marginal proportions in $r \times c$ table are estimated, the additional $(r-1)+(c-1)$ constraints imply

$$df = (rc-1) - [(r-1)+(c-1)] = (r-1)(c-1)$$

K. Pearson (1922)

“Such a view is entirely erroneous. The writer has done no service to the science of statistics by giving it broad-cast circulation in the pages of JRSS. I trust my critic will pardon me for comparing him with Don Quixote tilting at the windmill; he must either destroy himself, or the whole theory of probable errors, for they are invariably based on using sample values for those of the sampled population unknown to us.”

Fisher uses data from Pearson's son Egon

E. S. Pearson (1925) *Biometrika* generated $>$ 12,000 "random" 2×2 tables, for paper about Bayes Theorem

df = 3 or df = 1?

Fisher (1926) *Eugenics Rev.*

$$\frac{\sum_{i=1}^{12,000} X_i^2}{12,000} = 1.00001$$

In a later volume of his collected works (1950), Fisher wrote of Pearson, "If peevish intolerance of free opinion in others is a sign of senility, it is one which he had developed at an early age."

Fisher's exact test

2nd ed. *Statistical Methods for Research Workers* (1934), *The Design of Experiments* (1935)

Tea-tasting lady: Dr. Muriel Bristol, Rothamsted

		GUESS		
		Milk	Tea	
ACTUAL	Milk	3	1	4
Poured first	Tea	1	3	4
		4	4	

$$P\text{-value} = \frac{\binom{4}{3}\binom{4}{1} + \binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = 0.243$$

Fisher (1936) *Annals of Science*

Analyzes data from Mendel (1865)
(experiments testing theories of natural inheritance)
For 84 separate 2×2 tables, $\sum X^2 = 42$ (df = 84)

$$P_{H_0}(\chi^2_{84} \leq 42) = 0.00004$$

"When data have been faked, . . . people underestimate the frequency of wide chance deviations; the tendency is always to make them agree too well with expectations. The data of most, if not all, of the experiments have been falsified so as to agree closely with Mendel's expectations."

Fisher also proposed *partitioning chi-squared* (SMRW), and used canonical correlation to assign scores to rows and columns of contingency table to maximize correlation (1940), which relates to later *correspondence analysis* methods.

Maurice Bartlett (1935 *JRSS*)

For probabilities in a $2 \times 2 \times 2$ cross-classification of (X, Y, Z), “no interaction” defined as identical XY odds ratio at each level of Z

Shows maximum likelihood (ML) approach to estimating cell probabilities satisfying this condition

(attributes idea to R.A. Fisher)

End of story? Lancaster (1951 *JRSS-B*) “Doubtless little use will ever be made of more than a three-dimensional classification.”

The *probit* model for binary data

Chester Bliss: (1934) *Science*
(1935) *Ann. Appl. Biol.*

Popularizes probit model for applications in toxicology

Binary y with $y = 1$ for death, x = dosage or log dosage.
Underlying latent variable model for tolerance implies

Model: $P(y = 1) = \Phi(\alpha + \beta x)$

for *cdf* Φ of $N(0, 1)$ r.v.

R. A. Fisher (appendix to Bliss 1935) provides “Fisher scoring” algorithm for ML fitting of probit model. Later book by Finney.

The *logit*

- Maurice Bartlett (1937 *JRSS*) uses $\log[y/(1-y)]$ to transform continuous proportions for use in regression and ANOVA

- R. A. Fisher and Frank Yates (1938) *Statistical Tables* suggest transformation $\log\left[\frac{P(y=1)}{P(y=0)}\right]$ of binomial parameter

- Joseph Berkson (1944 *JASA*) of Mayo Clinic introduces term *logit* for $\log\left[\frac{P(y=1)}{P(y=0)}\right]$, shows similarity in shape of probit model and logistic regression model $P(y=1) = F(\alpha + \beta x)$ for logistic *cdf* $F(z) = \frac{e^z}{1+e^z}$

- D. R. Cox - Influential paper (1958 *JRSS-B*) and book (*Analysis of Binary Data*, 1970) on logistic regression

Later advances using logistic regression

- Case-control (Cornfield 1951, Mantel 1973, Prentice 1976)
- *Ordinal* data: McKelvey and Zavoina (1975) probit for cumulative probabilities, P. McCullagh (1980) arbitrary link

  Nominal data: Substantial econometric literature on baseline-category logit models and related discrete choice models (Theil 1970, McFadden 1974 – Nobel prize in 2000)
- Conditional logistic regression to eliminate nuisance parameters (Breslow, Prentice and others in late 1970s)
- Cyrus Mehta and Nitin Patel (1983) Develop network algorithm for exact conditional logistic regression
- Marginal models for clustered data (GEE approach: Kung-Yee Liang and Scott Zeger 1986)
- Random effects models: D. Pierce and B. Sands (1975)
- Item response models (Rasch 1961)

Jerzy Neyman (1949) *Berkeley symposium*

Cell probabilities $\{p_i\}$

Sample proportions $\{\hat{p}_i\}$

Model: $p_i = p_i(\theta)$

Develops BAN theory for estimators such as

- minimum chi-squared

$$\tilde{\theta} \text{ that minimizes } \sum_i \frac{(\hat{p}_i - p_i(\theta))^2}{p_i(\theta)}$$

- minimum modified chi-squared

$$\tilde{\theta} \text{ that minimizes } \sum_i \frac{(\hat{p}_i - p_i(\theta))^2}{\hat{p}_i}$$

Only mention of Fisher is disparaging comment that Fisher had claimed (not very clearly) that only ML estimators could be asymptotically efficient.

William Cochran (1909-1980)

(1940) *Annals*: ANOVA for Poisson and binomial responses

(1943) *JASA*: Dealing with overdispersion

(1950) *Biometrika*: Cochran's Q for comparing proportions in several matched samples, generalizes McNemar (1947) test

(1954) *Biometrics*: Methods for strengthening χ^2

- Guidelines on using X^2 for small n
  ("all expected frequencies $\geq$ 5" too strict)
- Partitioning X^2, such as a $df = 1$ test for a linear trend in proportions in a $r \times 2$ table with ordered rows
  (Cochran - Armitage test)
- Test of XY conditional independence in $2 \times 2 \times K$ tables
  Compare $\sum_k n_{11k}$ to $E_{H_0}(\sum_k n_{11k})$, df = 1
  (similar to Mantel - Haenszel (1959) test)

Goodman and Kruskal measures of association

Leo Goodman and William Kruskal
(1954, 1959, 1963, 1972 *JASA*)

> Introduce measures of association for contingency tables, emphasize interpretability of proportional reduction in error (PRE) measures

e.g. for ordinal classifications, discrete version of Kendall's tau for concordant (C) and discordant pairs (D)

$$\hat{\gamma} = \frac{C - D}{C + D}$$

Later extensions by social scientists to ordinal models predicting sign($y_i - y_j$) using sign($x_i - x_j$) for pairs (i, j) of observations

Loglinear models

Leo Goodman and others make explicit loglinear model formulation of multiplicative categorical data relationships

(multiplicative relationships, so log transform yields linearity, ANOVA-like models)

ex. Statistical independence of X and Y

$$\begin{aligned} P(X=i,\ Y=j) &= P(X=i)P(Y=j) \\ \log P(X=i,\ Y=j) &= \log P(X=i) + \log P(Y=j) \\ &= \alpha_i + \beta_j \end{aligned}$$

ex. Conditional independence models in multiway tables

Darroch, Lauritzen, and Speed (1980 *Annals Statist.*) later showed graphical modeling connections

Loglinear models (more)

Rapid advances in loglinear methodology during late 60's and early 70's at

Chicago	**Harvard**	**N. Carolina**
Leo Goodman	students of	Gary Koch
Shelby Haberman	F. Mosteller and	and colleagues
	W. Cochran	

Chicago

Haberman Ph.D. thesis (1970) - outstanding theoretical development of loglinear models

Chicago: Leo Goodman

Tremendous contributions to loglinear methodology (and related logit models for contingency tables) starting in 1964

(1968, 1970) *JASA*:	Good surveys
	Fisher memorial lecture
	“quasi independence”
(1971) *Technometrics*:	Model-building, stepwise procedures
(1974) *Biometrika*:	Latent class model
	EM fitting, extends Lazarsfeld
(1979) *JASA*:	Association models for ordinal variables
(1986) *Int. Statist. Rev.*:	Inference for correspondence analysis

Simultaneously, applications articles in social science journals

Harvard: Fred Mosteller

ASA presidential address (1968) *JASA*

"I fear that the first act of most social scientists upon seeing a contingency table is to compute chi-square for it."

Paper describes influential work at this time by students of Mosteller (e.g., Bishop, Fienberg) and Cochran at Harvard

National Halothane Study

(Is halothane more likely than other anesthetics to cause death due to liver damage?)

impetus for Bishop, Fienberg, and Holland (1975)
Discrete Multivariate Analysis

Several articles on loglinear models by these authors in early 1970s

U. North Carolina: The "GSK method"

Grizzle, Starmer and Koch (1969) *Biometrics*

Apply weighted least squares methods to logit, loglinear, and other regression-type models for categorical data

Later papers by Gary Koch and students applying WLS to variety of problems, such as models for repeated categorical measurement (*Biometrics* 1977)

Vasant Bhapkar (1966) *JASA*

When model can be specified by constraint equations that are linear in $\{p_i\}$, WLS estimator = Neyman's minimum modified chi-squared estimator

Generalized linear models

John Nelder and Robert Wedderburn (1972) *JRSS-A*

Logistic, loglinear, probit models are special cases of generalized linear models for exponential family response distributions using various “link functions.”

- Unites categorical methods with ANOVA and regression methods for normally distributed responses
- ML fitting of all models achieved by same Fisher scoring algorithm, which is iterative WLS (GLIM)
- Wedderburn (1974) generalized to *quasi likelihood*

Bayesian approaches for categorical data analysis

- Using beta distribution (especially, uniform) as prior for binomial goes back to Bayes (1763), Laplace (1774)
- I. J. Good (1956, 1965, et al.) smooths proportions in sparse contingency tables using Dirichlet prior (outgrowth of intelligence work during WWII at Bletchley Park with Turing), also uses empirical Bayes and hierarchical approaches
- Pat Altham (1969) considers Bayesian analyses for 2×2 tables and shows connections with frequentist results
- Steve Fienberg and Paul Holland (1970, 1973) use empirical Bayes to smooth tables
- Tom Leonard (1970s) et al. generalize Dennis Lindley's work using normal prior dist's for logit, loglinear parameters
- Arnold Zellner and Peter Rossi (1984) and later papers use simulation methods to fit binary regression models

Final comments

I did not learn about categorical data methods at UW, but the program gave me the foundation to (like many in Statistics) work in other areas.

I'm proud to be a UW grad, and very honored to be here today.

Grateful thanks to Box, Draper, Gurland, Johnson, Wahba, Roussas, Tiao, Klotz, Van Ryzin, Watts, Harris, Hunter, Ney/Athreya/Kurtz, and especially Steve Stigler!