



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Latent Variables and Your Future

IMS Junior Researchers Workshop
Baltimore, MD

Karen Bandeen-Roche, Ph.D.
Hurley-Dorrier Professor and Chair of Biostatistics
Johns Hopkins Bloomberg School of Public Health

July 29, 2009

Latent [Variables and Your Future]?

Latent Variables AND Your Future?

[Latent Variables] and [Your Future]

- Future of our field
- Latent variable modeling
 - Opportunity to learn something
 - Context of our field's future
- Your future

Most important career advice

*Back up your
computer!!*

Part I

Field's future

Field's Future: Some references

- Bartholomew DJ. What is statistics? *JRSSA* 1995; 158:1-20.
- Cox DR. Applied statistics: A review. *Ann Appl Stat* 2007; 1:1-16.
- Demets DL, ... Louis TA, et al. Training the next generation of biostatisticians: A call to action in the U.S. *Stat Med* 2006; 25:3415-29.
- Efron B. The future of statistics. White paper, 2007.
- Lindsay BG, Kettenring J, Siegmund DO. A report on the future of statistics (with discussion). *Statist Sci* 2004; 3-387-413.
- Zelen M. Biostatisticians, biostatistical science, and the future. *Stat Med* 2006; 25:3409-14.

“Statistics”—two connotations

- Discipline – Two aspects (NSF)
 - “Core”—subset of statistical activity focused inward
 - The rest

“Statistics”—two connotations

- Discipline – Two aspects (NSF)
 - “Core”—subset of statistical activity focused inward
 - The rest
- Profession
 - *“How individuals work most fruitfully in what” (DRC)*
 - Education
 - *“Statisticians have... something to contribute at the highest level[s]... which is often not recognized either among ourselves or others” (DJB)*

On Statistics

Some definitions

- *(Pearson, per DJB; Kendall, 1950): “The study of collective characters in populations.”*

On Statistics

Some definitions

- *(Pearson, per DJB; Kendall, 1950): “The study of collective characters in populations.”*
- *(Grenander & Miller, 1994; NSF): The science of learning from data.*

On Statistics

Some definitions

- *(Pearson, per DJB; Kendall, 1950): “The study of collective characters in populations.”*
- *(Grenander & Miller, 1994; NSF): The science of learning from data.*
- *(Stevens, 1968): ... “a straightforward discipline designed to amplify the power of common sense in the discernment of order amid uncertainty.”*

On Statistics

Some definitions

- (Pearson, per DJB; Kendall, 1950): *“The study of collective characters in populations.”*
- (Grenander & Miller, 1994; NSF): *The science of learning from data.*
- (Stevens, 1968): ... *“a straightforward discipline designed to amplify the power of common sense in the discernment of order amid uncertainty.”*
- (DRC/NSF/DJB): *“The discipline concerned with the study of variability, with the study of uncertainty, and with the study of decision making in the face of uncertainty.”*

On Statistics

- “... *Statistical Science [is] the particular aspect of human progress which gives the 20th century its special character.... It is to the statistician that the present age turns for what is most essential in all its more important activities.*”

Sir R. A. Fisher, 1952 address to IBS

A few views

- “... *why do so many people, who suppose that they do the very things that we do, prefer **not** to call themselves **statisticians** but operational researchers, software engineers, QA experts, forecasters or astrologers?*”

RSS News, Agent Provocateur, 1993

A Changing Ecology?

- Information age
- “Big science”
- Society
- Generational shift?
 - “Old way outmoded”
 - “Nothing new under the sun”

Future of statistics: Research

- Highlighted disciplinary futures:
 - Data: technology revolution; massive size (BE, NSF)
 - “Compromise” methods (BE / NSF)
 - “Errors of the third kind”: model selection (BE)
 - Modeling complex systems (NSF)
 - Dealing with uncertainty (NSF)
 - Middle ground between proof and computation (NSF)
 - Analysis techniques arising outside statistics (NSF)
 - Trends: more general vs. more specific models (DRC)
 - Design (DRC)
 - Causality (DRC)

Future of Statistics: Education

- Some issues
 - Decline of young people entering the field (NSF)
 - Offering sufficient depth over wide range of tools (NSF)
 - Postdoctoral training (NSF; Demets)
 - Non-traditional areas: consulting, ethics, communication, leadership, management (Demets)
 - Balance between core & science training (Demets)

Future of Statistics: Society

- How can / should statisticians impact on science and society over the coming decades?
 - Increasing demand for statistical collaboration in science (NSF)
 - What organizational structures and rules of engagement will allow our field to flourish in what many see as a changed and changing organizational environment? (TAL)
 - What needs to be done to increase the degree to which statistics contributes at the highest levels (DJB)?

Part II

Latent variables

Latent Variables

Why statisticians should care

- Highlighted disciplinary futures:
 - Data: technology revolution; massive size (BE, NSF)
 - “Compromise” methods (BE / NSF)
 - “Errors of the third kind”: model selection (BE)
 - **Modeling complex systems (NSF)**
 - **Dealing with uncertainty (NSF)**
 - Middle ground between proof and computation (NSF)
 - Analysis techniques arising outside statistics (NSF)
 - Trends: more general vs. more specific models (DRC)
 - Design (DRC)
 - Causality (DRC)

Latent Variables

Why statisticians should care

- They may indeed yield novel insights
 - to operationalize / test **theory**
 - to learn about **measurement problems**
 - they **summarize** multiple measures **parsimoniously**
 - to describe population **heterogeneity**
- They are ripe for misuse
 - their **modeling assumptions** may determine scientific conclusions
 - their **interpretation** may be ambiguous
 - Nature of latent variables?
 - What if very different models fit comparably?
 - Seeing is believing
- They are widely used

Latent Variables: What?

- *Underlying*: not directly measured. Existing in hidden form but capable of being measured indirectly by observables
 - Ex/ Pollution source contributions to an airshed
 - Ex/ Syndromal type
 - Ex/ Integrity of physiological regulation of systemic inflammation
- Some favorite books: Bartholomew (1988), Bollen (1989), McCutcheon (1987), Skrondal & Rabe-Hesketh (2004)
- Model: A framework linking latent variables to observables

Latent Variables: What?

Integrands in a hierarchical model

- Observed variables ($i=1,\dots,n$): Y_i =M-variate; x_i =P-variate
- Focus: response (Y) distribution = $G_{Y|x}(y|x)$; x-dependence
- Model:

— Y_i generated from latent (underlying) U_i :

$$F_{Y|U,x}(y|U=u,x;\pi) \quad (\textit{Measurement})$$

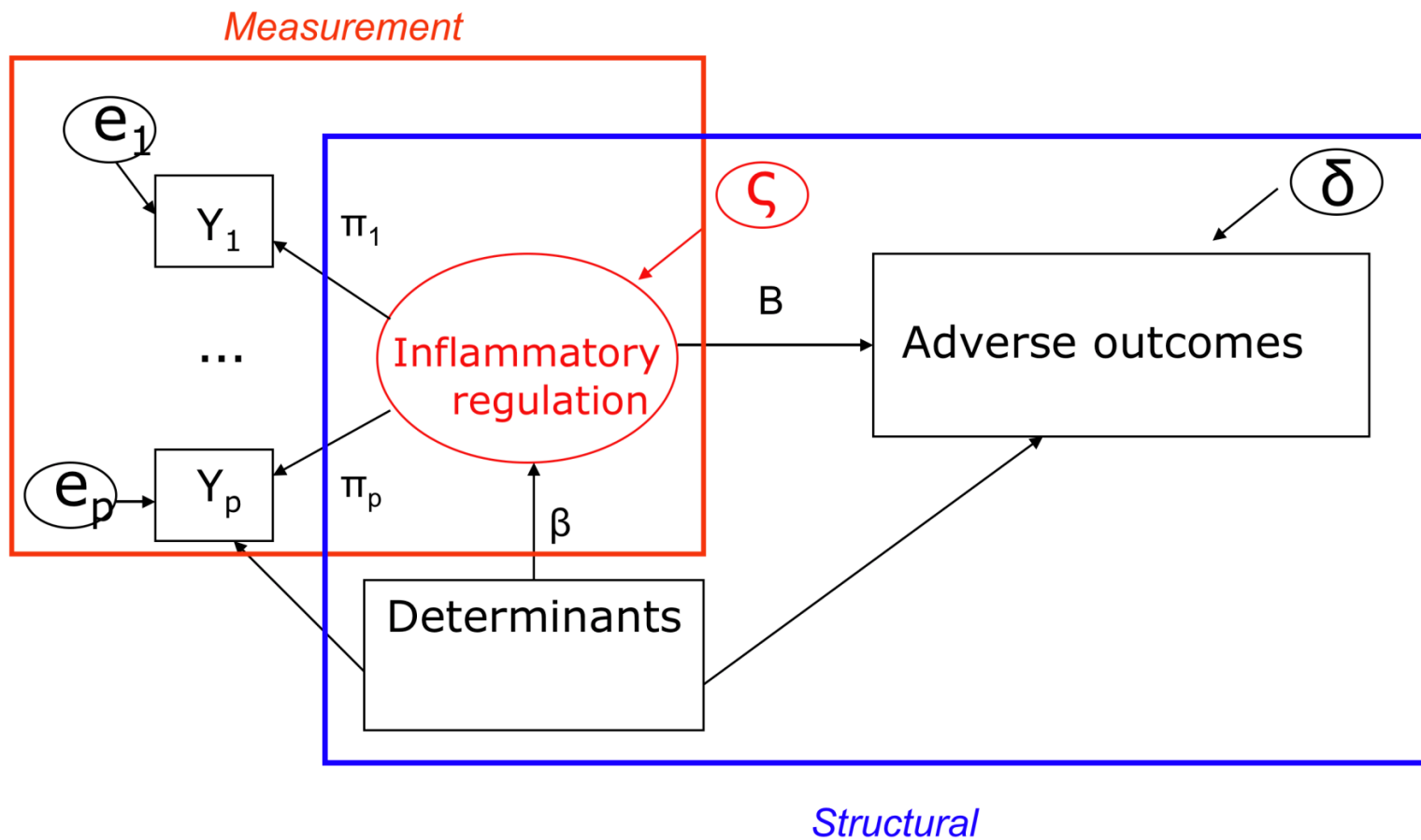
— Focus on distribution, regression re U_i :

$$F_{U|x}(u|x;\beta) \quad (\textit{Structural})$$

> Overall, **hierarchical model**:

$$F_{Y|x}(y|x) = \int F_{Y|U,x}(y|U=u,x) dF_{U|x}(u|x)$$

Latent Variable Model Schematic

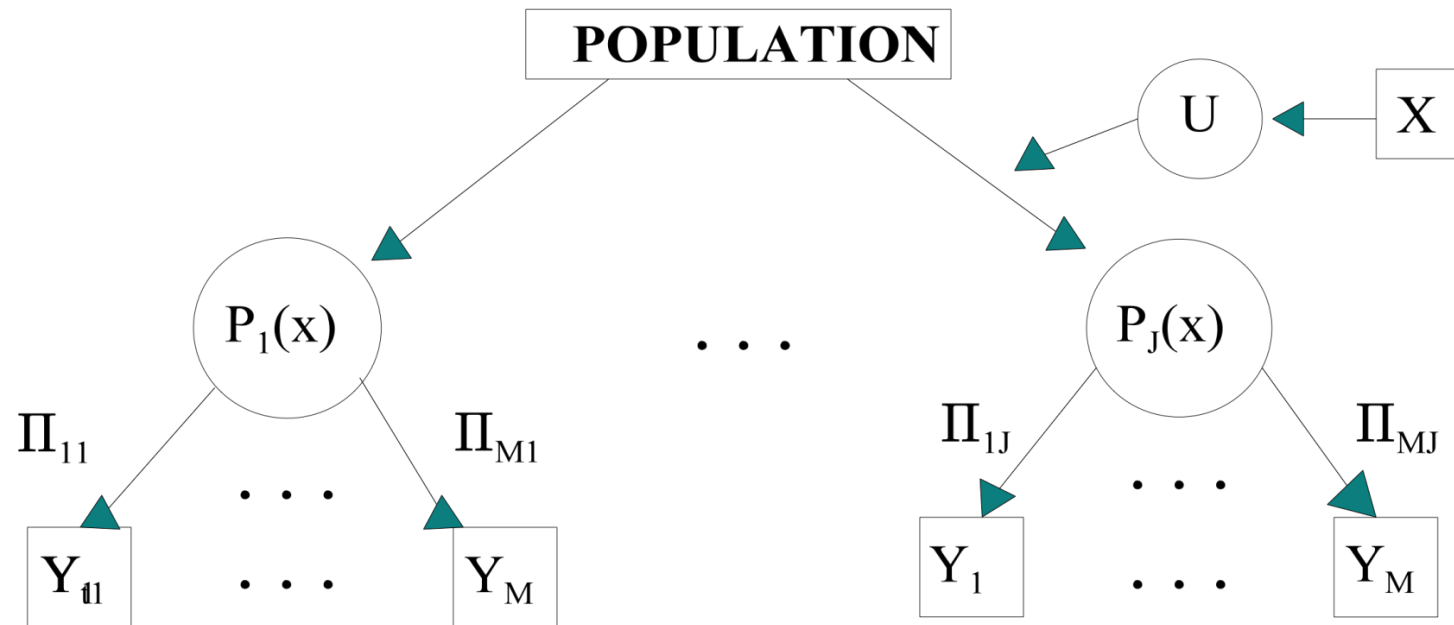


Application: Post-traumatic Stress Disorder Ascertainment

- PTSD
 - Follows a qualifying traumatic event
 - **This study: personal assault, other personal injury/trauma, trauma to loved one, sudden death of loved one = “x”, along with sex**
 - Criterion endorsement of symptoms related to event => diagnosis
 - Binary report on 17 symptoms = “Y”
- Study (Chilcoat & Breslau, *Arch Gen Psych*, 1998)
 - Telephone interview in metropolitan Detroit
 - N=1927 with a qualifying event
 - Analytic issues
 - Nosology
 - Does diagnosis differ by trauma type or gender?
 - *Are female assault victims particularly at risk?*

Model 1

Latent Class Regression



- > $P_j(x) = \Pr\{U = j|x\}$
- > $\pi_{mj} = \Pr\{Y_m=1|U = j\}$

References: Dayton & Macready 1988, van der Heidjen et al., 1996; Bandeen-Roche et al., 1997

Latent Variable Models: What / How

Latent Class Regression (LCR) Model

- **Model:**

$$f_{Y|x}(y|x) = \sum_{j=1}^J P_j(x, \beta) \prod_{m=1}^M \pi_{mj}^{y_m} (1 - \pi_{mj})^{1-y_m}$$

- **Structural model:** $[U_i|x_i] = Pr\{U_i=j|x_i\} = P_j(x_i, \beta)$

— $RPR_j = Pr\{U_i = j|x_i\} / Pr\{U_i = J|x_i\}; j=1, \dots, J$

- **Measurement assumptions** : $[Y_i|U_i]$

— conditional independence

— nondifferential measurement

> *reporting heterogeneity unrelated to measured, unmeasured characteristics*

- **Fitting:** ML w EM (Goodman, 1974) or Bayesian

- *Posterior* latent outcome information: $Pr\{U_i=j|Y_i, x_i; \theta=(\pi, \beta)\}$

Latent Variables Modeling

Do data bear out theoretic predictions?

- **Commonly used methods for adjudicating fit**
 - Global fit statistics (*many references*)
 - **Thresholds sensitive to study design; black box**
 - Relative fit statistics (*Akaike, 1974; Schwarz, 1978; Lo et al., 2001*)
 - **they're relative**
 - Comparisons of observed and predicted frequencies, associations
 - Cross-validation (*Cudeck & Browne, 1983; Collins & Wugalter, 1992*)
 - Pearson / correlation residuals (*Hagenaars, 1988; Bollen, 1989*)
 - Posterior predictive distributions (*Gelman et al., 1996*)
 - Bayesian graphical displays (*Garrett & Zeger, 2000*)
 - **whether fit fails, now how fit fails**
- Common wisdom: LV model assumptions are hard to check

Do data bear out theoretic predictions?

Part 1: Checking empirical reasonableness of the theory

- Rationale
 - If model correct and latent status known, measurement model “easy” to “explicate”
 - If persons can be partitioned into groups such that measurement model holds, model must correctly describe data distribution
- Research question: Suppose we estimate latent status.
 - Might the same idea work?
 - Seems circular?
 - Scientific intuition: Best shot = to randomize

Do data bear out theoretic predictions?

Part 1: Checking empirical reasonableness of the theory

1. FIT MODEL
2. ESTIMATE posterior probabilities Θ_i of membership from fit (“hats”)
3. **RANDOMLY** ALLOCATE INDIVIDUALS INTO “PREDICTED,” I.E. “*PSEUDO-*” CLASSES C_i ACCORDING TO $\Theta_{i1}, \Theta_{i2}, \dots, \Theta_{iJ}$
4. ASSESS ASSUMPTIONS WITHIN PREDICTED CLASSES
 - > Y_{i1}, \dots, Y_{im} not highly associated
 - > Y_i, x_i not highly associated

Bandeen-Roche, Miglioretti, Zeger & Rathouz, 1997;
Huang & Bandeen-Roche, 2004; Wang, Brown & Bandeen-Roche, 2005

Checking empirical reasonableness of the theory

- Does the scheme work?
 - Hardest part: **how to formulate what it means** for scheme to work
- Notation
 - R_j : “Reasonable” class of LCR models; $\{\pi, \beta\} = \phi \in \Phi$
- Formal statement of diagnostic premise: define

$$Z_{i\phi} = \prod_{m=1}^M \pi_{mj}^{y_m} (1 - \pi_{mj})^{1-y_m} \text{ with prob. } P(x, \beta), j=1, \dots, J$$

— Then (Theorem)

$$\boxed{\text{Pr}\{Y_i=y | C_i, x_i\}} \stackrel{D}{\rightarrow} Z_{i\phi} \text{ for some } \phi$$

if and only if $f_{Y_i}(y) = f_Y(y) \in R_j$ for each i

PTSD Study: Descriptive Statistics

Gender	Trauma Type: percentage distribution				n
	<i>Personal Assault</i>	<i>Other Injury</i>	<i>Trauma to loved one</i>	<i>Sudden death</i>	
<i>Male</i>	14.2	37.7	26.9	21.3	964
<i>Female</i>	14.3	26.3	32.2	27.2	863
Total	14.2	32.3	29.4	24.1	1827

- PTSD symptom criteria met: 11.8% (n=215)
 - By gender: 8.3% of men, 15.6% of women
 - By trauma: *assault (26.9%), sudden death (14.8%), other injury (8.1%), trauma to loved one (6.0%)*
 - Interactions: female x assault (↑), female x other (↓)
 - Criterion issue? 60% reported symptoms short of diagnosis

Latent Class Model for PTSD: 9 items

SYMPTOM CLASS	SYMPTOM (prevalence)	SYMPTOM PROBABILITY (π)		
		Class 1 - NO PTSD	Class 2 - SOME SYMPTOMS	Class 3 - PTSD
RE-EXPERIENCE	Recurrent thoughts (.49)	.20	.74	.96
	Distress to event cues (.42)	.12	.68	.88
	Reactivity to cues (.31)	.05	.51	.77
AVOIDANCE/NUMBING	Avoid related thoughts (.28)	.08	.37	.75
	Avoid activities (.24)	.05	.34	.66
	Detachment (.15)	.01	.14	.64
INCREASED AROUSAL	Difficulty sleeping (.19)	.02	.18	.78
	Irritability (.21)	.02	.22	.83
	Difficulty concentrating (.25)	.03	.30	.89
MEAN PREVALENCE-BASELINE		.52	.33	.14

[Omitted: nightmares, flashback; **amnesia**, ↓ **interest**, ↓ **affect**, **short future**; hypervigilance, startle]

PTSD: DIAGNOSIS, LCR MEASUREMENT MODEL

- Method: Regress item responses on covariates “controlling” for class
 — For simplicity: non-assaultive traumas merged into “other trauma”

Variable	Odds Ratio or Interaction Ratio (CI)	By-item Odds Ratio MODEL 2
Female	1.07 (0.93,1.22)	1.07 (0.93,1.22)
Trauma =other than assault (recur.)	3.19 (1.89,5.40)	3.19 (1.89,5.40)
Cue distress x other trauma	0.18 (0.09,0.38)	0.58 (0.36,0.92)
Cue reactivity x other trauma	0.14 (0.07,0.28)	0.44 (0.27,0.72)
Avoid thoughts x other trauma	0.21 (0.11,0.41)	0.68 (0.44,1.05)
Avoid activities x other trauma	0.11 (0.05,0.22)	0.35 (0.21,0.58)
Detachment x other trauma	0.27 (0.13,0.58)	0.88 (0.51,1.49)
Difficulty sleep x other trauma	0.43 (0.21,0.90)	1.37 (0.78,2.42)
Irritability x other trauma	0.28 (0.13,0.61)	0.91 (0.52,1.59)
Concentration x other trauma	0.73 (0.36,1.47)	2.33 (1.35,4.03)

Summary

PTSD Analysis

- Symptoms appeared differentially sensitive to different traumas

Within classes: those who had a non-assaultive trauma were

- **less prone** to report distress to cues, reactivity to cues, avoiding thoughts and avoiding activities
 - **more prone** to report recurrent thoughts and difficulty concentrating
- Concern: Criteria may better detect psychiatric sequelae to assault than to traumas other than assault

Latent Variable Modeling

Areas of Needed Discovery

- Scientifically relevant models
 - Area is increasingly active
 - Syndromes in older adults
 - Genetics, genomics
 - Toxicology
 - Virology, bacteriology
 - Economics
 - Criminology
 - Causal inference
 - Ecology
 - Application-driven

Areas needing discovery

- Falsifiability of competing hypotheses (*identification / estimability*)
 - Subject-area science: What are the relevant hypotheses (models)?
 - Mathematical statistics: Model characterization
 - Study design
 - Sensitivity analysis / “intervals” of plausible models
 - Computing
 - Visualization
 - Cross-validation

Latent Variable Modeling

Areas of Needed Advancement

- Computation
 - Bayesian / EM – intensive
 - Bayesian / Penalized LH: Prior / penalty choice
 - Estimating equations
 - Hybrid approaches

Part III

Your Future

Closing Thoughts

Future of statistics: In your hands

- NSF comment: “... *the most important advances will be unpredictable. For this reason we need to maintain an underlying philosophy that is flexible enough to adapt to change. At the same time is important that this future research not degenerate into a disparate collection of techniques.*”

Future of statistics: In your hands

- NSF comment: “... *the most important advances will be unpredictable. For this reason we need to maintain an underlying philosophy that is flexible enough to adapt to change. At the same time is important that this future research not degenerate into a disparate collection of techniques.*”
- BE comment: “*Statistics is in a period of rapid expansion and ... change. During such times it pays to concentrate on basics and not tie oneself too closely to any one technology or analysis fad.*”

Future of statistics: In your hands

- NSF comment: “... *the most important advances will be unpredictable. For this reason we need to maintain an underlying philosophy that is flexible enough to adapt to change. At the same time is important that this future research not degenerate into a disparate collection of techniques.*”
- BE comment: “*Statistics is in a period of rapid expansion and ... change. During such times it pays to concentrate on basics and not tie oneself too closely to any one technology or analysis fad.*”
- DRC comment: “*[though] certainly open to dispute, ... the broad approaches desirable to analysis and interpretation have not been radically changed by the capacity to handle very large amounts of data, however much approaches to implementation have been and are being revolutionized.*”

Future of statistics: In your hands

- NSF comment: “... *the most important advances will be unpredictable. For this reason we need to maintain an underlying philosophy that is flexible enough to adapt to change. At the same time is important that this future research not degenerate into a disparate collection of techniques.*”
- BE comment: “*Statistics is in a period of rapid expansion and ... change. During such times it pays to concentrate on basics and not tie oneself too closely to any one technology or analysis fad.*”
- DRC comment: “*[though] certainly open to dispute, ... the broad approaches desirable to analysis and interpretation have not been radically changed by the capacity to handle very large amounts of data, however much approaches to implementation have been and are being revolutionized.*”
- DJB comment: *Our view of statistics is too small. Let us seek to broaden it.*

Closing thoughts

- *How individuals work most fruitfully in what*
- *Future of statistics*

Closing thoughts

- *How individuals work most fruitfully in what*
- *Future of statistics*
 - “How should I know?”

Closing thoughts

- *How individuals work most fruitfully in what*
- *Future of statistics*
 - “How should I know?”
 - Napoleon Dynamite version:

Closing thoughts

- *How individuals work most fruitfully in what*
- *Future of statistics*
 - “How should I know?”
 - Napoleon Dynamite version:
“Like anyone can know that...”

Closing thoughts

- *How individuals work most fruitfully in what*
- *Future of statistics*
 - “How should I know?”
 - Napoleon Dynamite version:
“Like anyone can know that...”
- We have a great field
- Have a wonderful career

Future of statistics: In your hands

- *“This is the information age, statistics is the prime information science, and there is every reason to believe in a greatly increased statistical presence in the academy of the future. Or maybe not. Ideas are the coin of the realm in the intellectual world. Our continued growth and influence depends on the same thing that powered the last century, the continued production of useful new ideas and techniques.”*

Professor Bradley Efron, 2007