Fast Valid Statistical Inference when the Maximum Likelihood Estimate Does Not Exist in an Exponential Family Model and the "Usual Asymptotics" are Bogus

Charles J. Geyer

School of Statistics
University of Minnesota

Mini-Conference to Celebrate Elizabeth Thompson's Contributions to Statistics, Genetics and the University of Washington, June 19, 2018

Elevator Pitch

In exponential family models for discrete multivariate analysis (logistic regression, Poisson regression, categorical data analysis, Markov spatial lattice processes, Markov point processes, Markov random graphs, aster models), when the MLE does not exist in the conventional sense,

- currently available software does no valid inference,
- currently available software often fails to detect this situation,
- consequently, no one knows how much applied statistics is garbage because of this,
- but we know how to do valid point estimates, hypothesis tests, and confidence intervals in this situation,
- and we will release good software real soon now.

Time Line

Circa 1975 I independently invent Bradley-Terry models with ties and home field advantage (33 years after Bradley and Terry). I understand "solutions at infinity" in this context.

Circa 1978 my sister Ruth Shaw gives me Bishop, Fienberg, and Holland (1975) for a birthday present.

Circa 1980 my sister Ruth Shaw and I invent the first aster model. We don't publish, although in hindsight could have.

Early 1980's I somehow discover Barndorff-Nielsen (1978) and start reworking his theory of completion of exponential families.

Fall 1986. I start grad school at UW never having had a statistics course.

Time Line (cont.)

Ruth also knows Elizabeth, but I don't meet Elizabeth until the first day of class; she is the teacher for the 580's that year.

Spring 1987. Elizabeth posts ad for RA. I apply and am accepted. I start learning C so I can extend Alun Thomas's pedigree analysis software to do gene extinction.

Summer 1987. John Haslett's summer spatial statistics course introduces me to spatial lattice processes and MCMC.

Elizabeth becomes my thesis advisor even though I want to do a bunch of stuff outside her area.

Time Line (cont.)

Sometime in 1988 Ollie Ryder (research director, San Diego Zoo) brings Elizabeth the DNA fingerprinting problem. Elizabeth and I figure out how to do it with exponential families and MCMC. This eventually becomes RSS read paper (Geyer and Thompson, 1992). So MCMC gets into my thesis even though application of MCMC to probabilities on pedigrees is Nuala Sheehan's thesis topic. This paper also had "solutions at infinity".

PhD thesis Geyer (1990).

2005 I start R package `rcdd` (Geyer, Meeden, and Fukuda, 2017) which does computational geometry. Impetus is Glen Meeden wanting it for Bayesian finite population sampling with linear equality and inequality constraints on probabilities.

Time Line (cont.)

Geyer, Wagenius, and Shaw (2007). Aster models finally published. R package `aster` (Geyer, 2018) on CRAN since 2005. Package and first draft of paper written while on sabbatical at UW 2004–2005. Package detects "solutions at infinity".

Geyer (2009). My theory of completion of exponential families finally published. All the computations are in `Sweave` tech report and use R package `rcdd` extensively. New hypothesis tests and confidence intervals when MLE does not exist in conventional sense. Impetus is Steve Fienberg getting interested in subject. Hypothesis tests are due to Fienberg. He gave a talk on the subject at Minnesota. I asked about inference, and he sketched theory in his answer to my question. After the talk I told him my thesis had the complete solution to everything he talked about (except that answer to my question).

Time Line (cont.)

Summer 2012 Dan Eck writes R package `gdor` and puts on CRAN but doesn't keep it there (archived). It does some calculations from Geyer (2009).

I become Dan Eck's thesis advisor even though he wants to do a bunch of stuff outside my area. What goes around comes around. I particularly need Elizabeth's example of how to deal with this.

PhD thesis Eck (2017).

Eck and Geyer (submitted). Doesn't use R package `rcdd`. Calculations fast enough for users to tolerate. Also a bunch of new theory.

Binomial Example

First something everyone understands: the binomial distribution.

The binomial family of distributions is an exponential family of distributions with the usual data as canonical statistic but $\theta = \text{logit}(p)$ as canonical parameter.

When the observed data y is at either end of its range ($y = 0$ or $y = n$) the MLE for the canonical parameter does not exist. The MLE for the usual parameter does exist $\hat{p} = y/n$.

logit(0) and logit(1) are undefined or one can say $\text{logit}(0) = -\infty$ and $\text{logit}(1) = +\infty$, but these are not exponential family parameter values.

So the distributions corresponding to $p = 0$ or $p = 1$ are not in the binomial family considered as an exponential family. They are in the Barndorff-Nielsen completion of this exponential family.

Binomial Example (cont.)

The distributions in the completion but not in the original model ($p = 0$ and $p = 1$) are *degenerate*, concentrated at one point ($p = 0$ implies $Y = 0$ almost surely, and $p = 1$ implies $Y = n$ almost surely).

The usual asymptotics of maximum likelihood don't work on or near the boundary.

Intro stats books teach the rule of thumb that you need $np \geq 5$ and $n(1 - p) \geq 5$ for asymptotics to "work".

The Wald confidence intervals

$$\hat{p} \pm 1.96 \sqrt{\hat{p}(1 - \hat{p})/n}$$

are worthless, having width zero, when $\hat{p} = 0$ or $\hat{p} = 1$.

Binomial Example (cont.)

Textbooks recommend using Rao (a.k.a. score, a.k.a. Wilson) intervals done by

```
prop.test(x, n, correct = FALSE)
```

But Rao intervals are hard to do for multivariate problems.

Moreover, Rao intervals are justified by the same asymptotics as Wald intervals, so aren't valid near the boundary either. They just aren't as obviously completely worthless as Wald intervals.

Complete Separation Example of Agresti

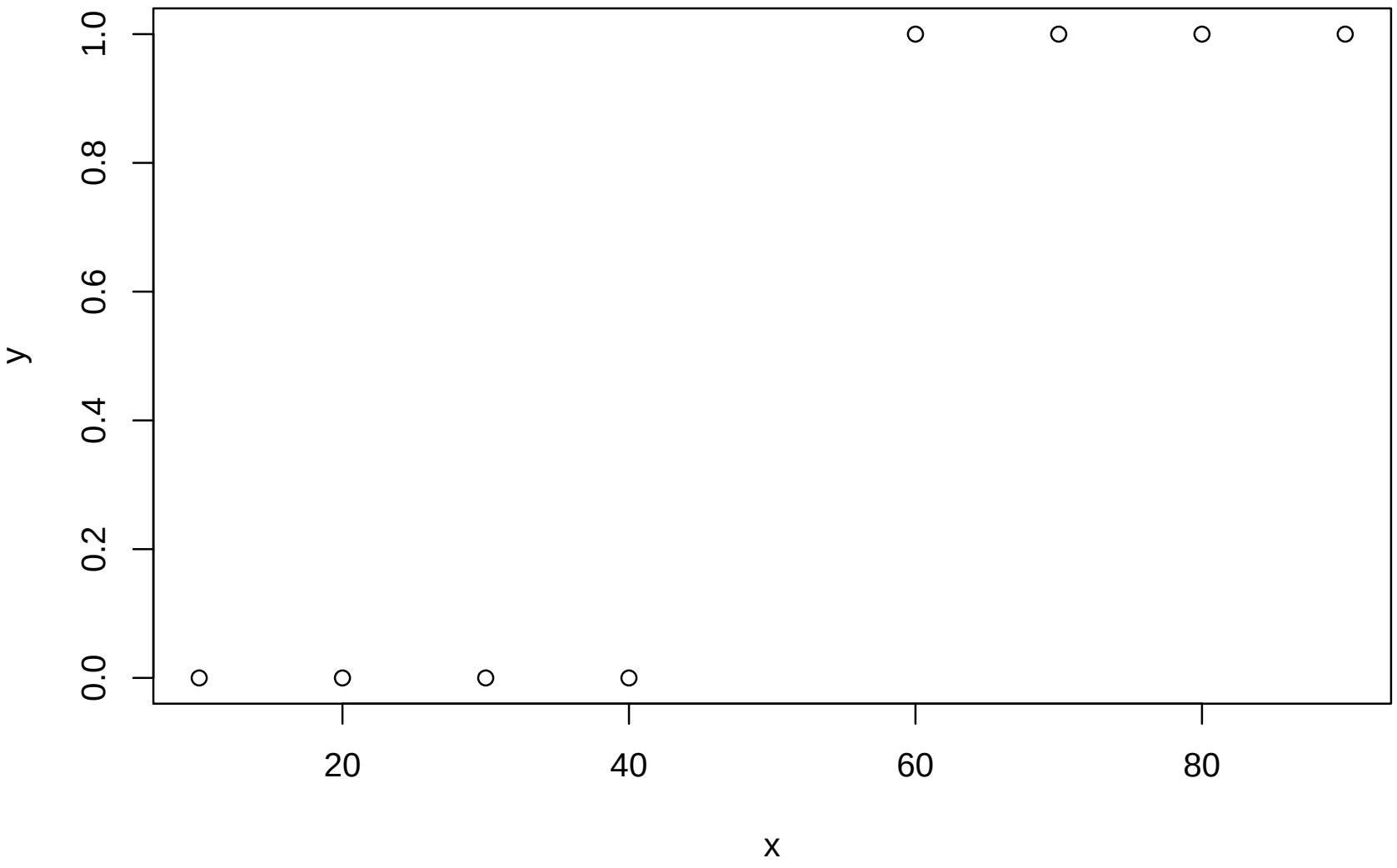

Figure: Scatterplot of Agresti complete separation toy data.

Complete Separation Example of Agresti (cont.)

```
gout <- glm(y ~ x, family = binomial, x = TRUE)


## Warning:  glm.fit:  fitted probabilities
numerically 0 or 1 occurred
```

R gives a warning when fitting, but

- The warning is based on inexact computer arithmetic. Both false positives and false negatives occur.
- R provides no methods for valid inference when the warning is correct.

Complete Separation Example of Agresti (cont.)

```
as.vector(predict(gout))
```

```
## [1] -94.52642 -70.89481 -47.26321 -23.63160
## [5]  23.63160  47.26321  70.89481  94.52642
```

```
as.vector(zapsmall(predict(gout, type = "response")))
```

```
## [1] 0 0 0 0 1 1 1 1
```

MLE for saturated model canonical parameter vector ("linear predictor" in GLM parlance) has all components nearly plus or minus infinity. MLE of mean value parameter has degenerate distribution for all components.

Complete Separation Example of Agresti (cont.)

Degeneracy is not a problem. The sample is not the population. Estimates are not the parameters they estimate.

There is a problem only if we are naive about statistical inference.

Geyer (2009) has a proposal for valid confidence intervals when the MLE does not exist in the conventional sense.

Complete Separation Example of Agresti (cont.)

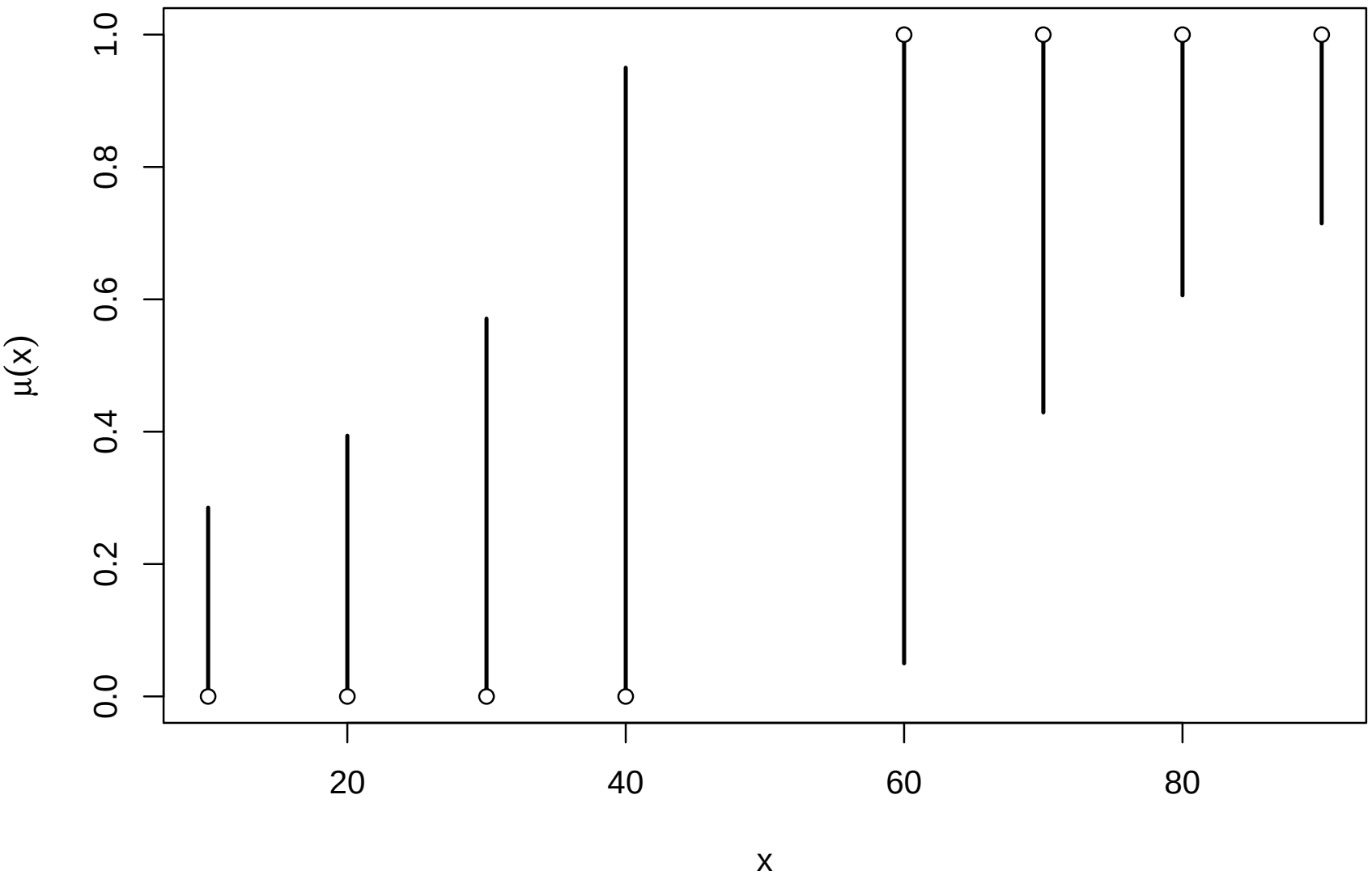

Figure: 95% confidence intervals for Agresti complete separation toy data.

Complete Separation Example of Agresti (cont.)

```
options(show.signif.stars = FALSE)
drop1(gout, ~ x, test = "LRT")
```

```
## Single term deletions
##
## Model:
## y ~ x
##        Df Deviance    AIC    LRT  Pr(>Chi)
## <none>       0.00   4.00
## x       1    11.09 13.09 11.09 0.0008678
```

Complete Separation Example of Agresti (cont.)

Even though the MLE does not exist for one of the models being compared, the LRT is valid because the MLE does exist (and is not close to the boundary) for the null hypothesis.

Conclusion: the only model that fits the data has "solution at infinity" and we need this theory to analyze it.

Because the MLE in the Barndorff-Nielsen completion is completely degenerate, it fits the data perfectly and no larger model can fit better.

General Theory

Saturated model, with canonical parameter vector θ (a.k.a. linear predictor) and canonical statistic vector y (a.k.a. response vector) is an exponential family.

Submodel given by $\theta = M\beta$, where M is model matrix, is also exponential family. Because of

$$\langle y, \theta \rangle = \langle y, M\beta \rangle = \langle M^T y, \beta \rangle$$

- $M^T y$ is the submodel canonical statistic vector, and
- β is the submodel canonical parameter vector.

General Theory (cont.)

Theory (Barndorff-Nielsen, 1978; Geyer, 1990, 2009; Eck and Geyer, submitted) says the MLE exists in the conventional sense if and only if the observed value of the canonical statistic is not on the boundary of the convex hull of its support.

For a GLM, this means the *submodel* canonical statistic vector $M^T y$.

For a two-parameter model with two-dimensional canonical statistic vector, we can visualize this support.

Otherwise, we cannot and need computational geometry R package `rcdd` or the new methods in Eck and Geyer (submitted).

General Theory (cont.)

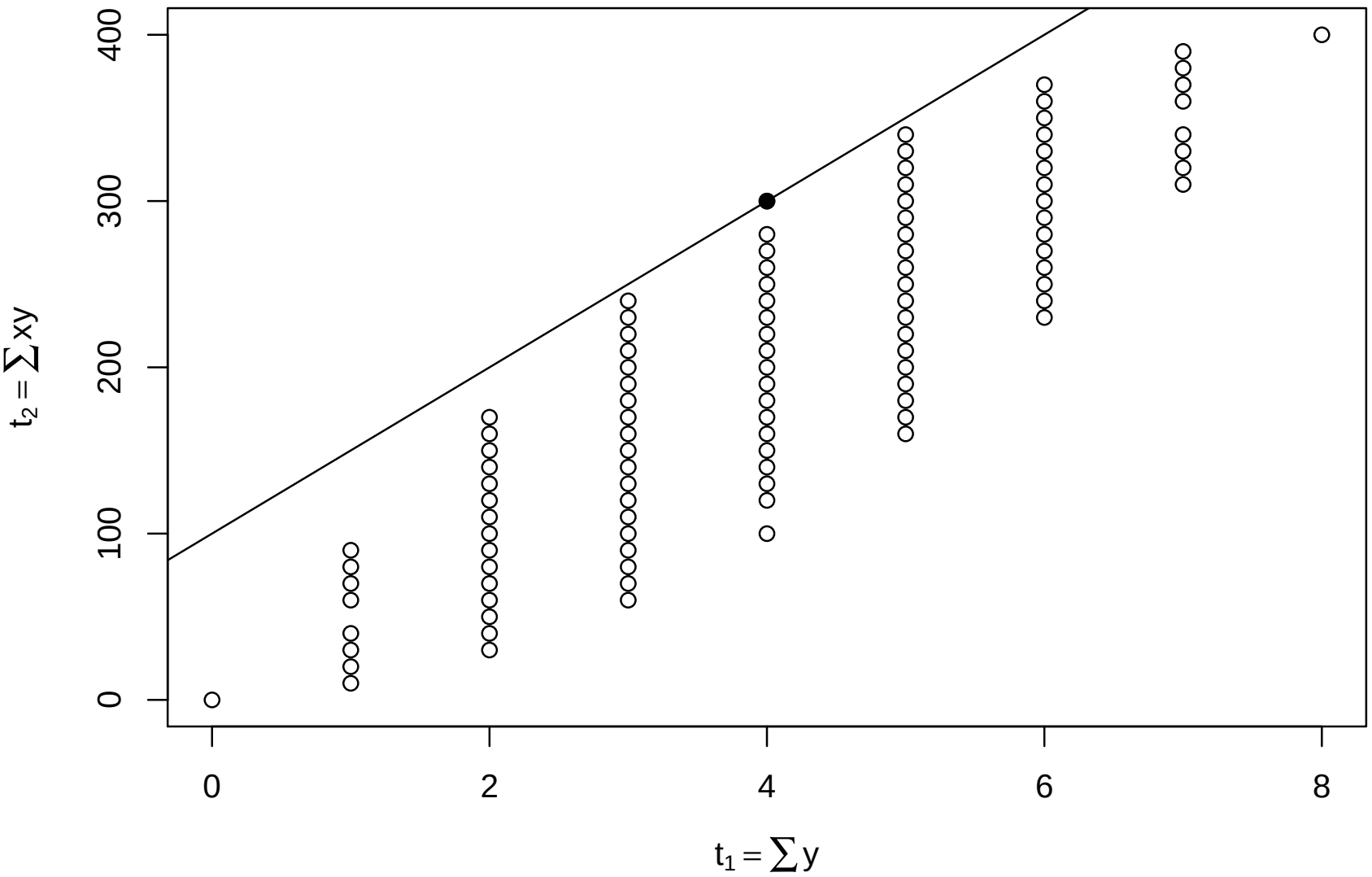

Figure: Support of $t = M^T y$ for Agresti complete separation toy data. Dots are possible values. Black dot is observed value.

General Theory (cont.)

Geyer (2009) shows that when the MLE does not exist in the conventional sense the MLE in the Barndorff-Nielsen completion is the MLE for the GLM that conditions on $M^T y$ lying in the hyperplane H that separates its observed value from other possible values (the line in the figure).

That conditional model is called the *limiting conditional model* (LCM).

General Theory (cont.)

Fisher information matrix

$$I(\hat{\beta}) = \text{var}_{\hat{\beta}}(M^T Y)$$

is close to correct (theorem in Dan's thesis).

If V is the null space of the Fisher information matrix, then the hyperplane that supports the LCM is

$$H = M^T y + V$$

Whether $I(\hat{\beta})$ has null eigenvectors is a much better test than what `glm` uses. Finding V and H allows all of the rest of our theory to go though: maximum likelihood, hypothesis tests, and confidence intervals.

This is also the test R package `aster` has used since 2005, but with no theory to back it up until now.

Fienberg's Theory of Hypothesis Tests

When the MLE does not exist in the conventional sense for the null hypothesis, the LRT statistic is still approximately chi-squared but not with the degrees of freedom conventional theory says.

One has to understand the degeneracy of the MLE model in the completion to calculate degrees of freedom correctly. Condition both null and alternative on $M^T Y \in H$, where H is the hyperplane for the null hypothesis. Then calculate LRT as usual with these conditional models.

My Theory of Confidence Intervals

A simple argument in Geyer (2009) says that the region of the parameter space which puts probability at least α on the support of the LCM is a $100(1 - \alpha)\%$ confidence region for the parameter.

Binomial(n, p), observe $x = 0$, confidence interval is $[0, 1 - \alpha^{1/n}]$

Binomial(n, p), observe $x = n$, confidence interval is $[\alpha^{1/n}, 1]$

Poisson(μ), observe $x = 0$, confidence interval is $[0, -\log(\alpha)]$

For $\alpha = 0.05$ Poisson bound is 2.9957323.

Same recipe works in general and produces complicated intervals like shown in the figure.

Slides for this Talk

http://users.stat.umn.edu/~geyer/ElizabethFest/

References I

Barndorff-Nielsen, O. (1978). *Information and Exponential Families*. Wiley, Chichester, U. K.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis*. MIT Press, Cambridge, Massachusetts. First textbook on categorical data analysis.

Eck, D. J. (2017). Statistical Inference in Multivariate Settings. PhD thesis, University of Minnesota. http://hdl.handle.net/11299/190441

Eck, D. J., and Geyer, C. J. Computationally efficient likelihood inference in exponential families when the maximum likelihood estimator does not exist. Submitted to *Annals of Statistics*. https://arxiv.org/abs/1803.11240

Geyer, C. J. (1990). Likelihood and Exponential Families. PhD thesis, University of Washington. http://hdl.handle.net/11299/56330

References II

Geyer, C. J. (2009). Likelihood inference in exponential families and directions of recession. *Electronic Journal of Statistics*, 3, 259–289.

Geyer, C. J. (2018). R package `aster` (Aster Models), version 0.9.1.1. http://www.stat.umn.edu/geyer/aster/ and http://cran.r-project.org/package=aster

Geyer, C. J., Meeden, G. D., and Fukuda, K. (2017). R package `rcdd` (C Double Description for R), version 1.2. http://cran.r-project.org/package=rcdd

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data, (with discussion). *Journal of the Royal Statistical Society, Series B*, 54, 657–699.

Geyer, C. J., Wagenius, S., and Shaw, R. G. (2007). Aster models for life history analysis. *Biometrika*, 94, 415–426.