Statistical Modeling: Motivation and Foundations

Intermediate Social Statistics

Week 6 (23 February 2016)

Andy Eggers

So far

- Regression (OLS)
- RCTs
- Matching
- Instrumental variables
- RDD
- Diff-in-diff/panel

Also:

- Logistic regression

What else do we need?

Conventional approach: tour of Generalized Linear Models (GLMs)

If your dependent variable is …	…you need this model.	See this Stata command.
Continuous (and unbounded)	OLS	regress
Binary (e.g. join WTO or not)	Logit Probit	logit probit
A count (e.g. 0, 1, 10 wars)	Poisson Negative binomial	poisson nbreg
Ordered categories (e.g. "opposed", "neutral", in favor")	Ordinal logit Ordinal probit	ologit oprobit
Non-ordered categories (e.g. Tory, Labour, Lib Dem; Christian, Muslim, Jewish, atheist)	Multinomial logit, conditional logit	mlogit clogit
A measure of survival or duration (e.g. cabinet or war duration)	Survival or hazard model	stcox

See glm (generalized linear model) package for many of these.

Generalized linear models

Gailmard p. 146: "invertible function of the model parameter is expressed as a linear function of the covariate(s)"

Linear regression model:

$$\mathrm{E}(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Binary logistic models:

$$\log\left[\frac{P(Y=1)}{P(Y=0)}\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Multinomial logistic models:

$$\log\left[\frac{P(Y=j)}{P(Y=0)}\right] = \alpha_j + \beta_{j1} X_1 + \beta_{j2} X_2 + \cdots + \beta_{jk} X_k$$

Ordinal logistic models:

$$\log\left[\frac{P(Y \geq j)}{P(Y < j)}\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Count models:

$$\log\left[\mathrm{E}(Y)\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

What you need to know about GLMs

- Syntax (trivial)

  Stata: `[model name] [outcome] [covariates], [options]`
- Interpretation (not trivial)

Think about *what your model is supposed to help you understand* **(quantities of interest)**.

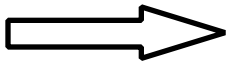
Especially with GLMs, this is usually not (quite) a regression coefficient.

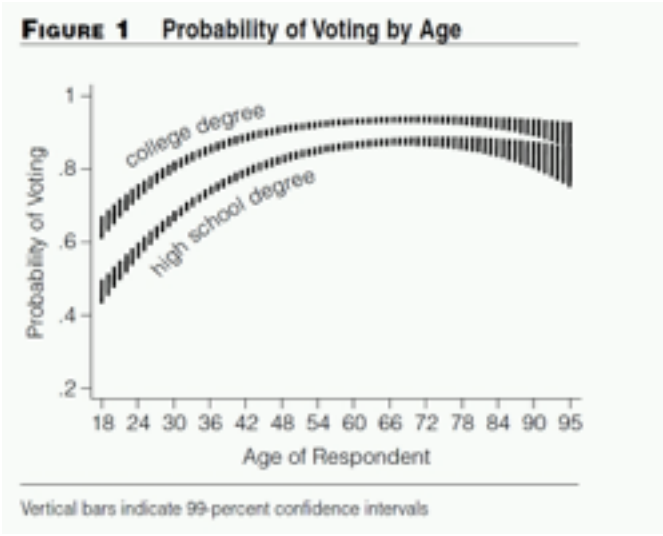
```
. oprobit gaymarriage female partyid highschool collegedegree

Iteration 0:   log likelihood = -2353.7283
Iteration 1:   log likelihood = -2315.4646
Iteration 2:   log likelihood = -2315.4564
Iteration 3:   log likelihood = -2315.4564

Ordered probit regression                       Number of obs   =       2176
                                                LR chi2(4)      =      76.54
                                                Prob > chi2     =     0.0000
Log likelihood = -2315.4564                     Pseudo R2       =     0.0163

------------------------------------------------------------------------------
 gaymarriage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |     .10073   .0498641     2.02   0.043     .0029981    .1984618
     partyid |  -.0810757   .0123165    -6.58   0.000    -.1052155   -.0569359
  highschool |  -.1840961   .0568693    -3.24   0.001    -.295558   -.0726342
collegedegree |   .1816416   .0686662     2.65   0.008     .0474503    .3158329
-------------+----------------------------------------------------------------
        /cut1 |  -.522102   .0636712                     -.6465033   -.3977007
        /cut2 |   .1277551   .0629013                     .0044709    .2510393
------------------------------------------------------------------------------
```





The next three weeks

- **Lab**: intuition and practice with GLMs in Stata
- **Lecture & reading**: broader perspective on the uses of statistical modeling; fundamentals of probability and estimation

The big picture

Learning to use/interpret other GLMs (probit, ordinal logit, etc) allows you to extend what you do with OLS

(prediction/description; causal inference under "selection on observables")

to other types of DVs. (What if we just use OLS?)

Understanding modeling principles behind GLMs, you get

- new measurement strategies (→ outcomes, covariates)
- structural models
- multilevel models and other dependence structures
- insights into inference (std errors, etc)

The big picture (2)

Building statistical models can (should) be fun!

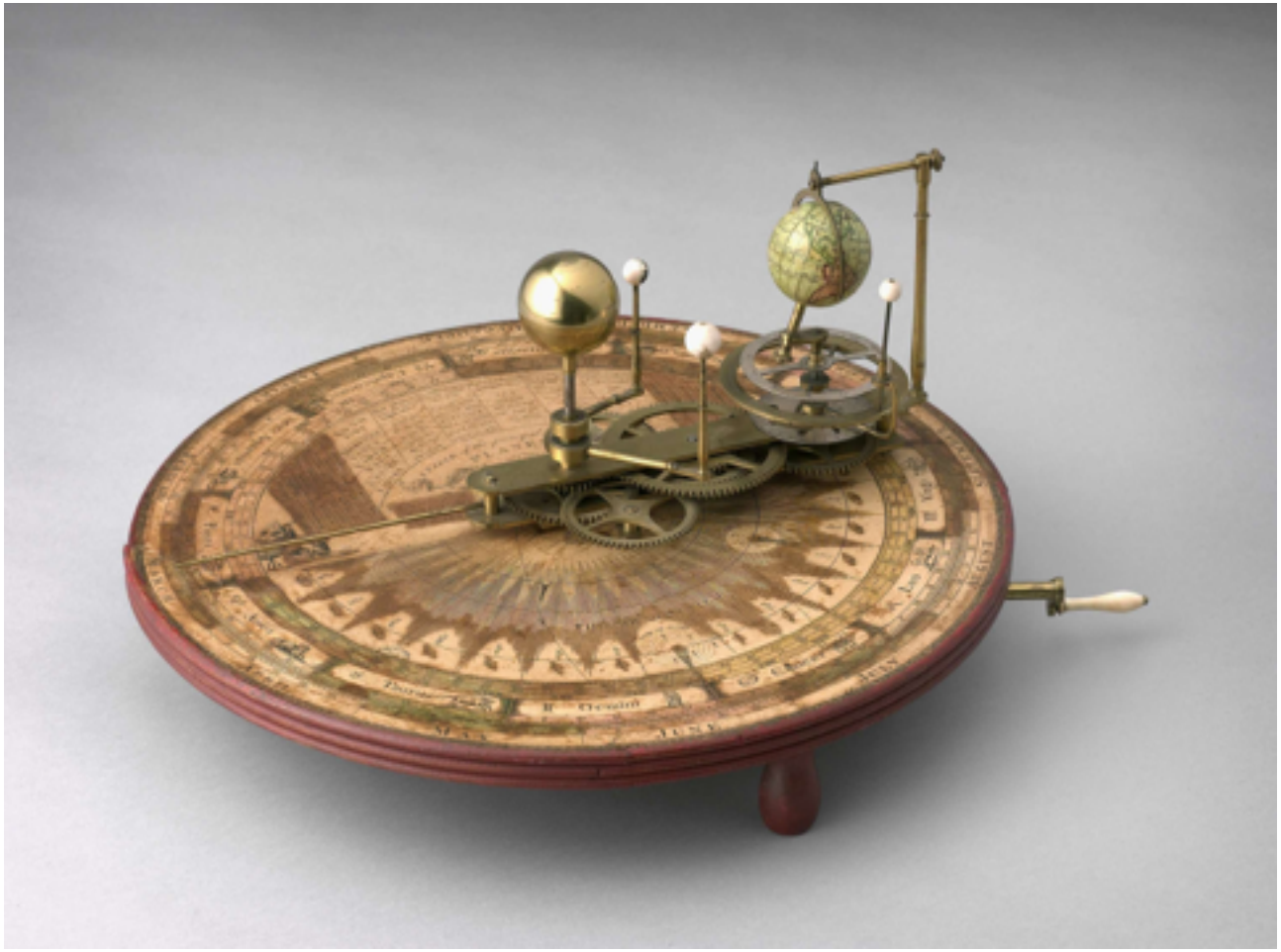
But don't lose sight of the goal:

"On one hand, we would like our models to be consistent with any functional form restrictions known to be true. On the other, attempts at this may complicate estimation and inference, for little payoff. Worst of all, the technical complexity of nonlinear models seems to cause authors to 'wax econometric,' an effort that may come at the expense of attention to substantive issues of importance."

Josh Angrist

from "Reply" to comments on "Estimation of Limited Dependent Variable Models With Dummy Endogenous Regressors: Simple Strategies for Empirical Practice" JBES 2001

What is a model?

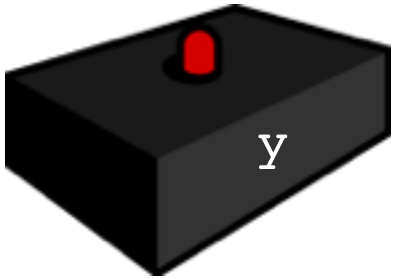


Jones's "New Portable Orrery" (1794)

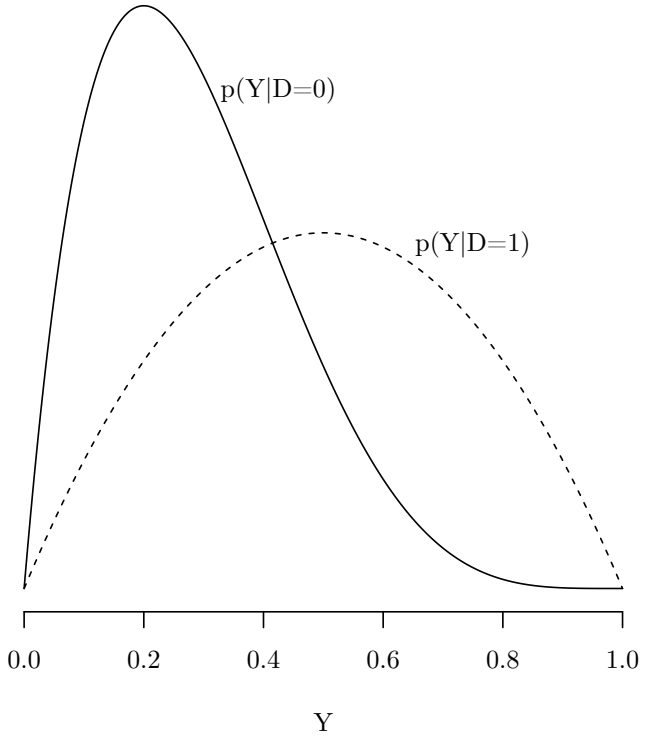
"All models are wrong, but some are useful." George Box

What is a statistical model?

A statistical model describes how a dependent variable (Y) is thought to have been generated.



More formally, a statistical model describes a **set of probability distributions** for a random variable (Y).



In any interesting statistical model, **different units have different distributions**, depending on the features of the unit (e.g. exposure to treatment vs. control, values of covariates).

Random variables and probability distributions

Gailmard 4.3

A **random variable** Y takes one of multiple possible (numerical) values depending on the outcome of an "experiment".

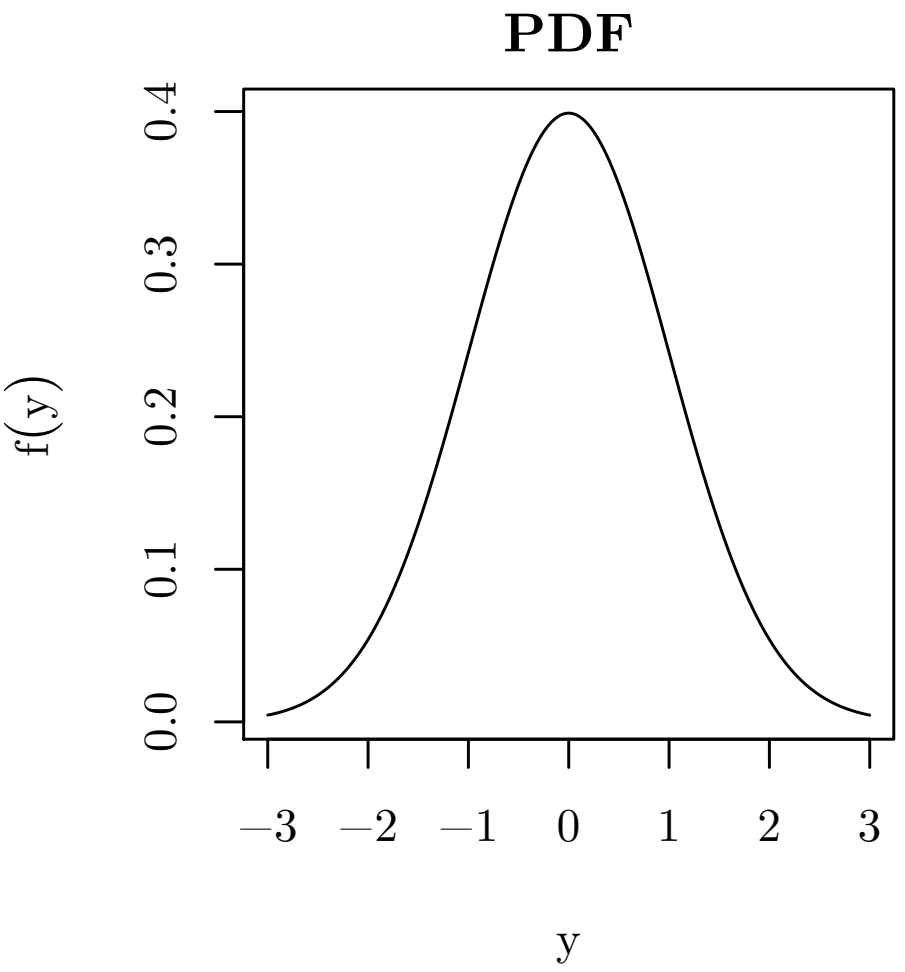
Conventional notation: Y is the RV; y is a particular value.

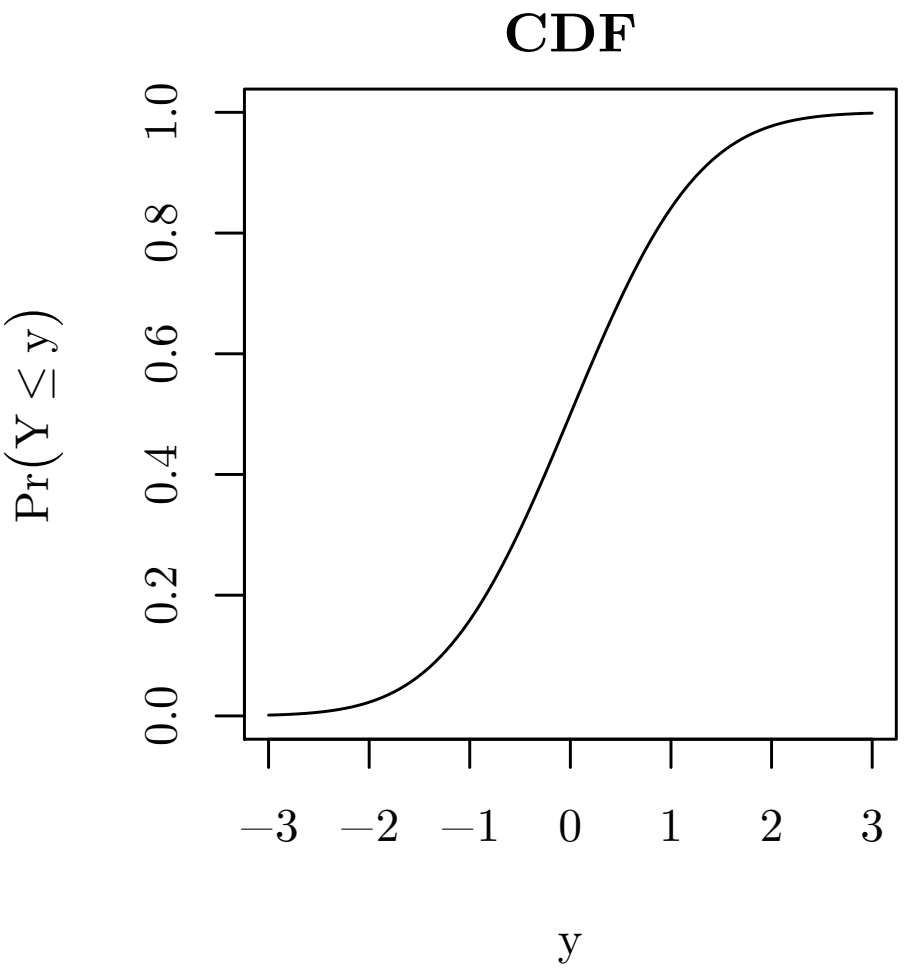
Gailmard 4.4

The probability distribution of a random variable Y can be summarized by

- a **cumulative distribution function** (CDF) gives $\Pr(Y \leq y)$
- (if discrete) a **probability mass function** (PMF) gives $\Pr(Y=y)$
- (if continuous) a **probability density function** (PDF) gives the derivative of the CDF at y

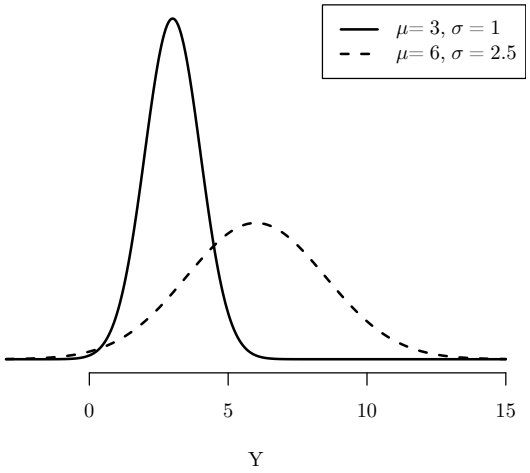
Normal PDF and CDF



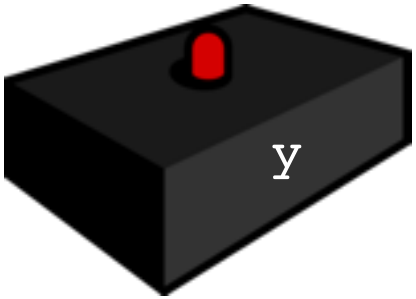


How probability works

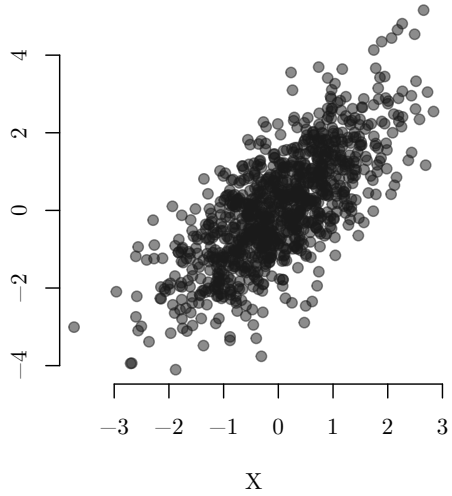
Given a set of probability distributions…



…that characterize a **data generating process** (DGP)…

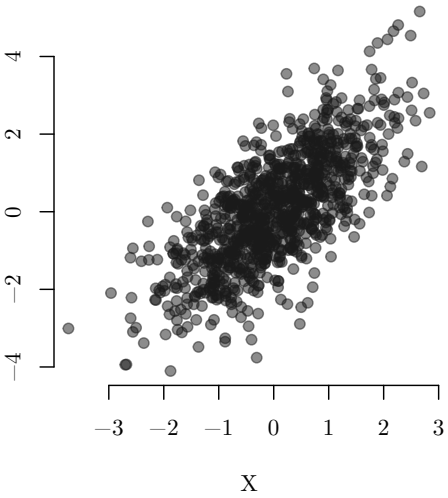


…what data should we expect to observe?

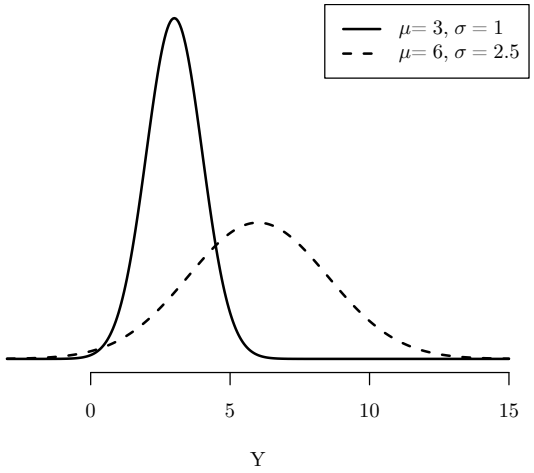


How (classical) statistics works

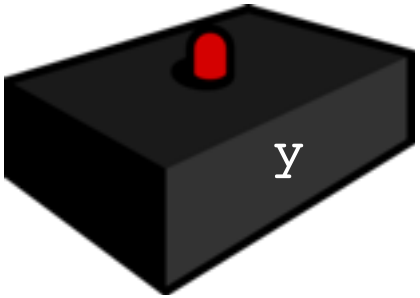
Given the data we observe…



…what set of probability distributions…



characterize the DGP?



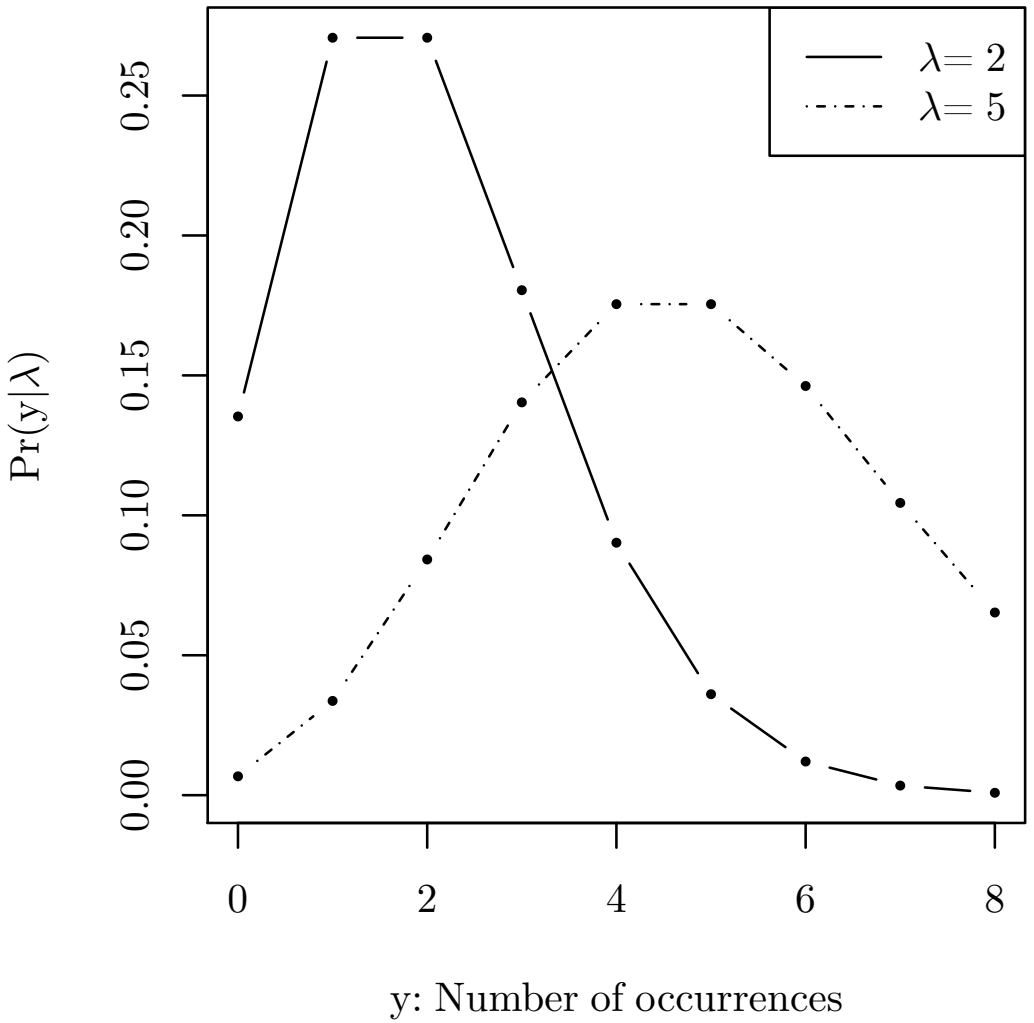
Poisson PMF

Gailmard 6.3

$$\Pr(Y = y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

Characterizes count of events (e.g. false convictions, horse kicks) observed in a fixed interval when

- events are independent
- rate of occurrence (probability per unit time) is constant

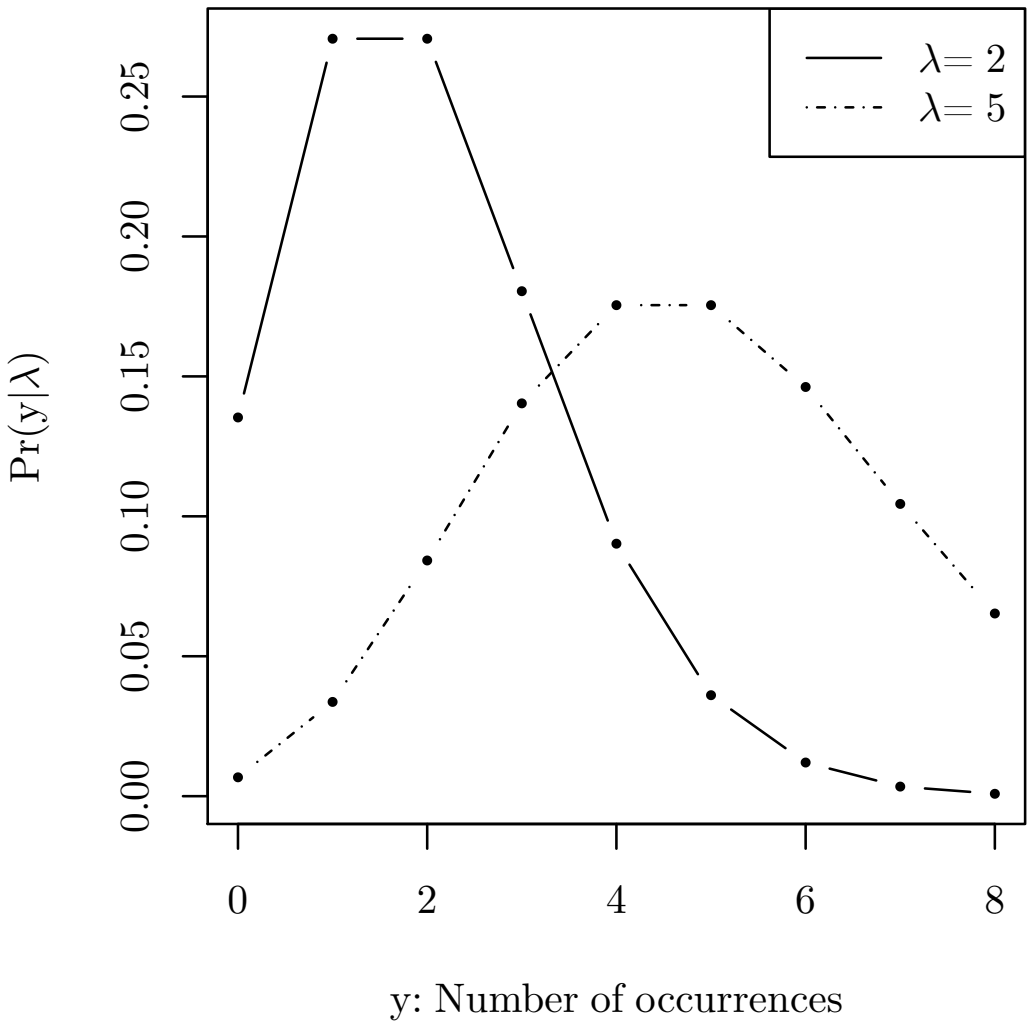


Single count

ACTIVITY!!!

Suppose we view the number of students sitting in row 3 as a Poisson random variable.
(Reasonable?)

1) If λ = 2, how likely is the observed outcome?
2) If λ = 5, how likely is the observed outcome?

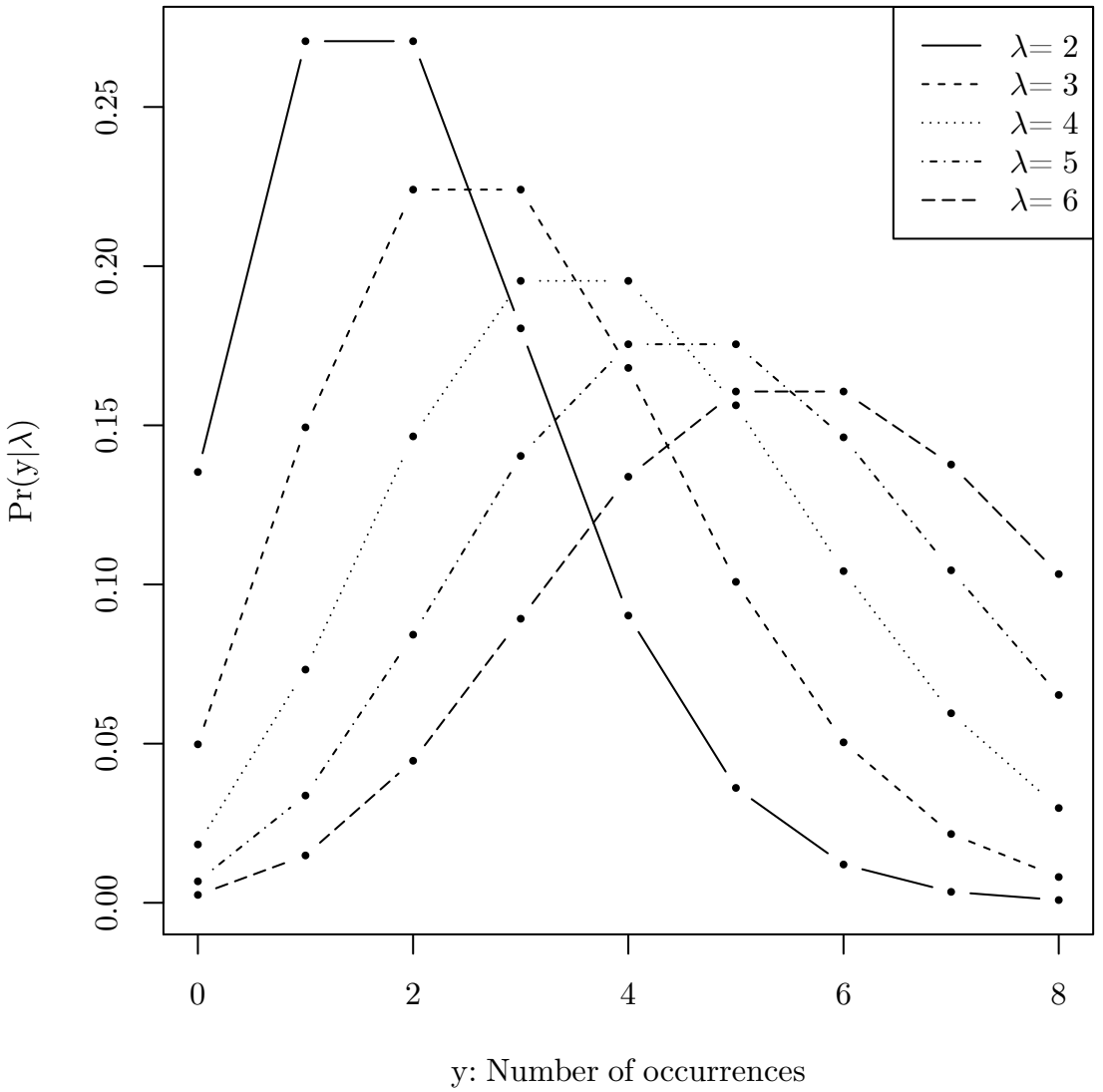


Single count (2)

ACTIVITY!!!

Suppose we view the number of students sitting in row 3 as a Poisson random variable.

For what value of λ is the observed outcome most likely?



Joint & conditional probability and independence

For two events E and F, the probability of both events happening is written

$$\mathrm{P}(E, F) \qquad \text{or} \qquad P(E \cap F)$$

joint probability

The probability of E happening given F is written

$$\mathrm{P}(E|F)$$

conditional probability

If E and F are independent,

$$\mathrm{P}(E|F) = \mathrm{P}(E)$$

and:

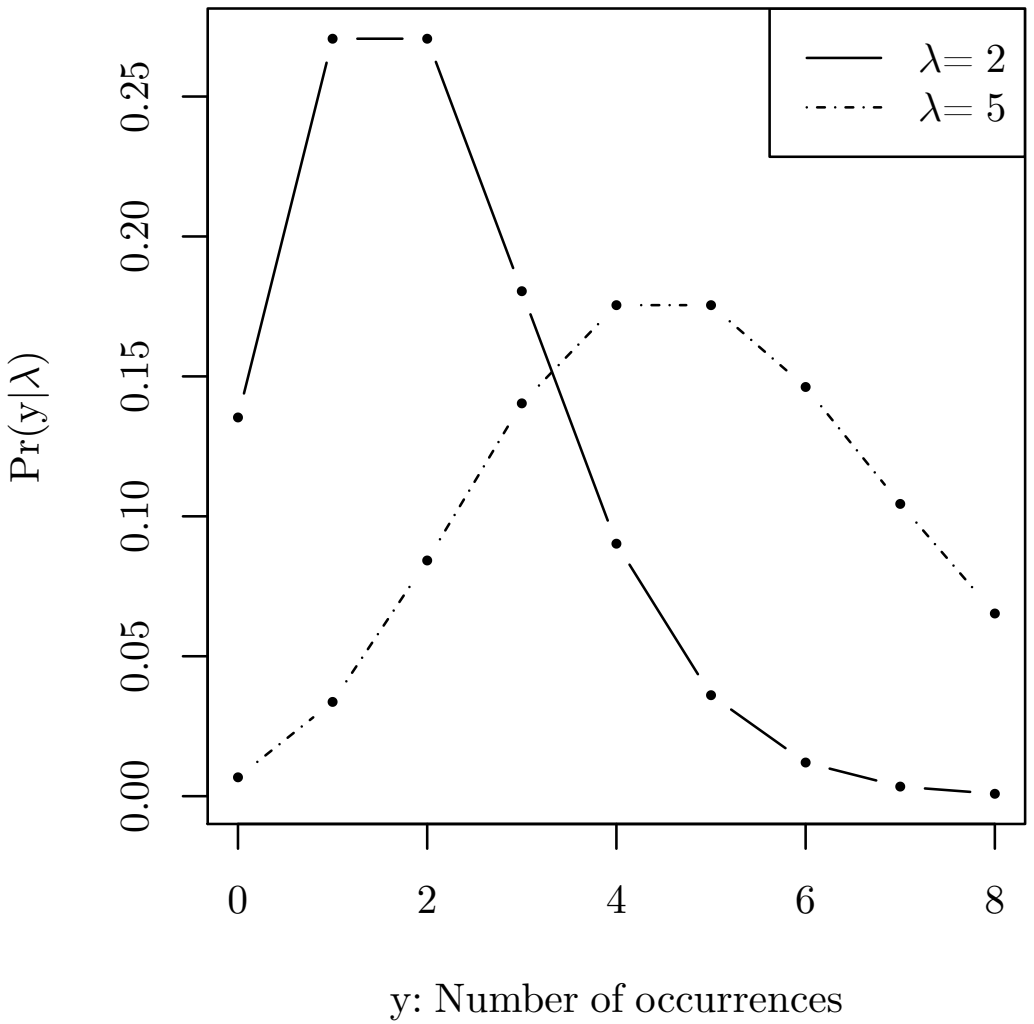
$$P(E, F) = P(E) \times P(F)$$

Vector of counts

ACTIVITY!!!

Suppose we view the number of students sitting **in each row** as an independent Poisson random variable. (Reasonable?)

1) If λ = 2, how likely is the observed outcome for rows 3-6?
2) If λ = 5, how likely is the observed outcome for rows 3-6?

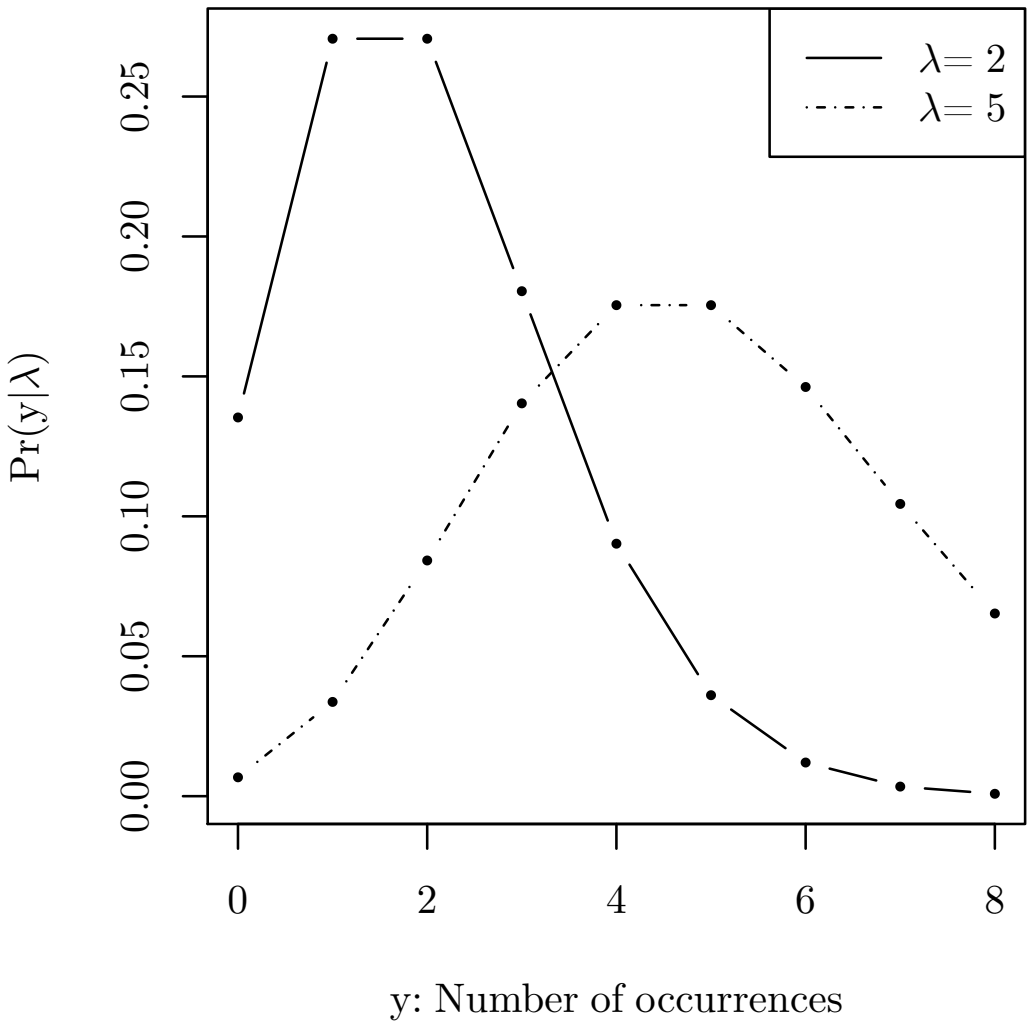


Vector of counts

SOLUTIONS

Suppose we have 5, 2, 7, 4 students in these rows.

	λ = 2	λ = 5
5	0.04	0.18
2	0.27	0.08
7	0.003	0.10
4	0.10	0.18
Product X IM	3.03	270.84

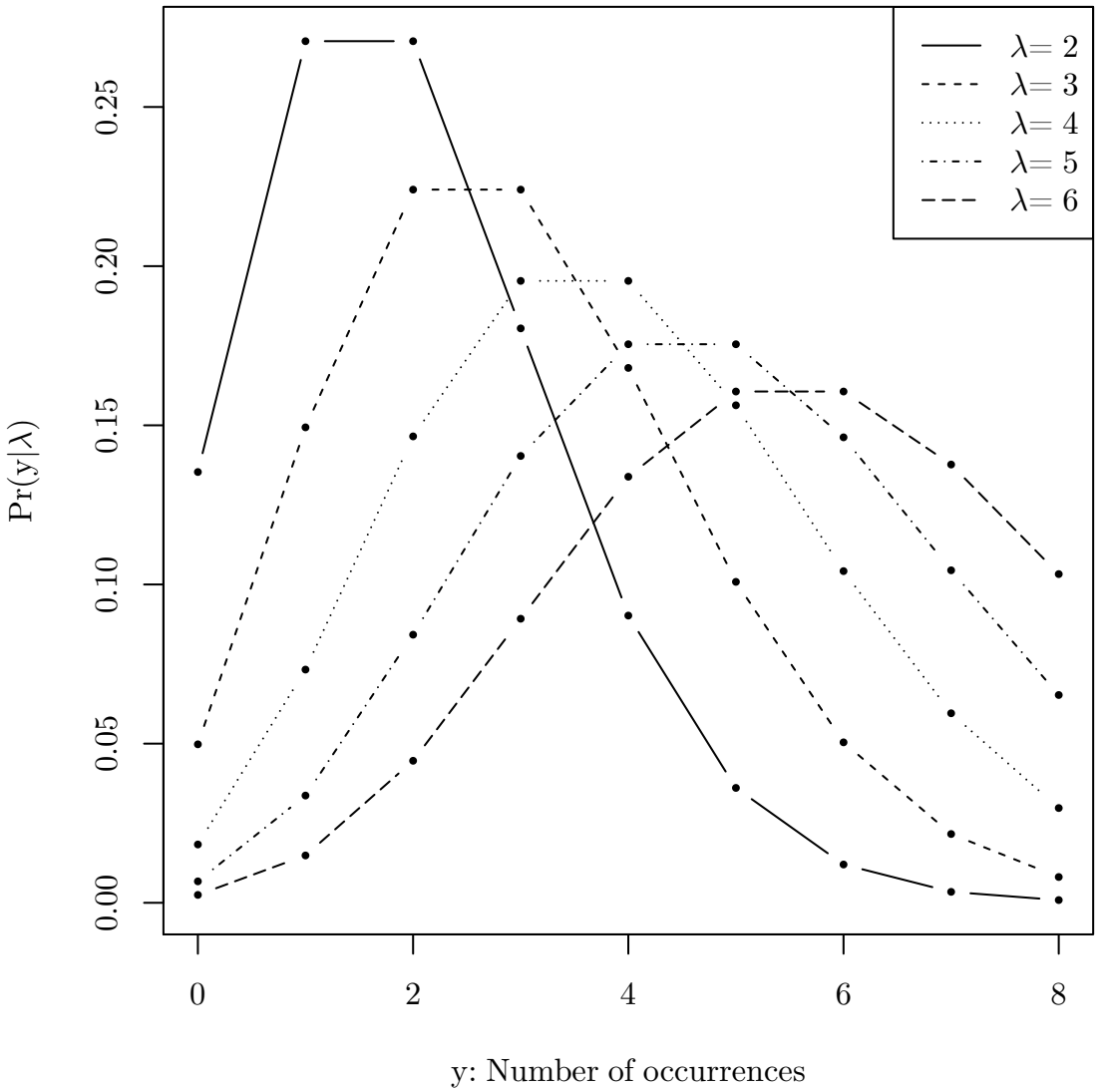


Vector of counts (2)

ACTIVITY!!!

Suppose we view the number of students sitting **in each row** as an independent **(iid)** Poisson random variable.

For what value of λ is the observed outcome for rows 3-6 most likely?

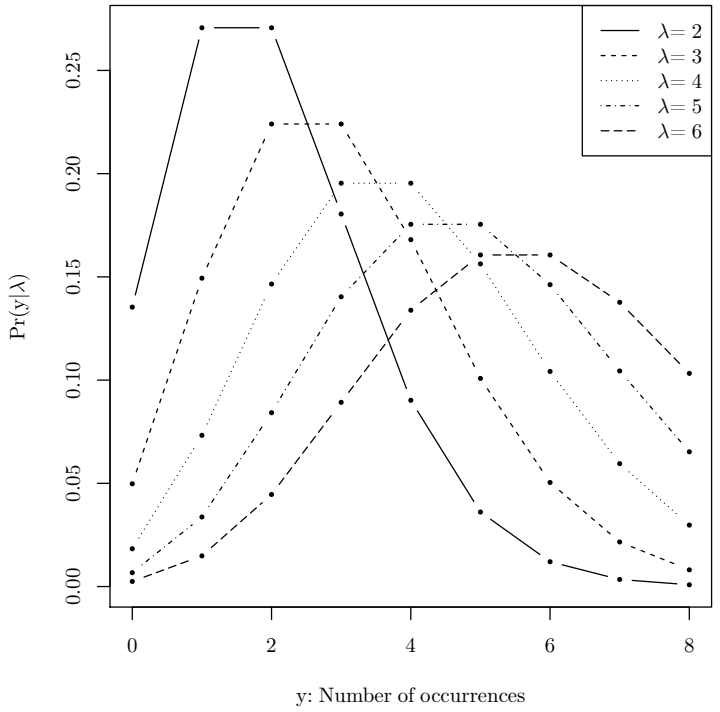


Vector of counts

SOLUTIONS

Suppose we have 5, 2, 7, 4 students in these rows. Then approximately:

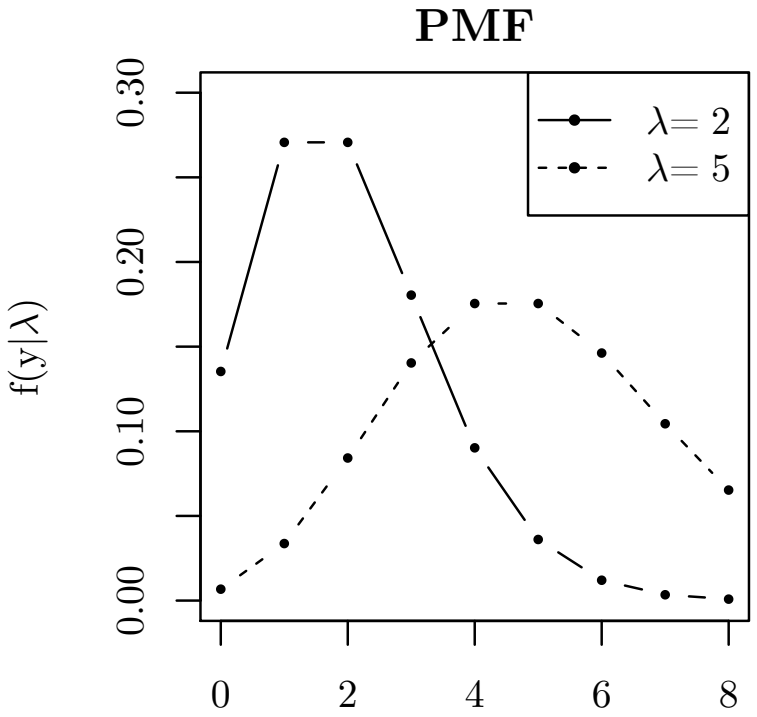
	λ = 2	λ = 3	λ = 4	λ = 5	λ = 6
5	0.04	0.10	0.16	0.18	0.16
2	0.27	0.22	0.15	0.08	0.04
7	0.003	0.02	0.06	0.10	0.14
4	0.10	0.17	0.20	0.18	0.13
Product X IM	3.03	82.00	266.39	270.84	132.07

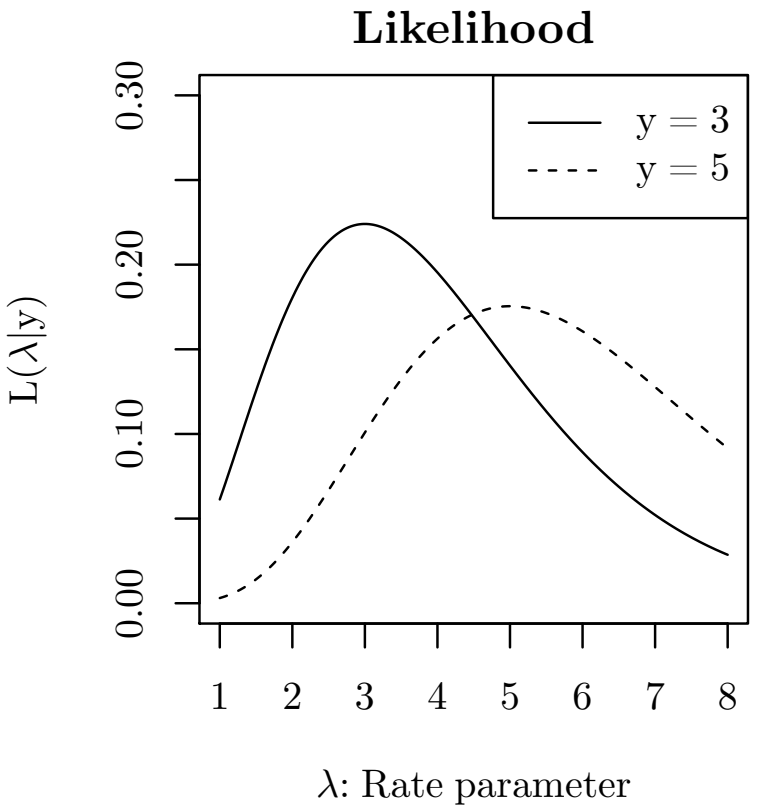


PMF vs Likelihood

The pmf can be written f(y|λ): a function of y (the observed data) whose shape depends on λ (the parameter).

But from that pmf we can derive L(λ|y): a function of λ (the parameter) whose shape depends on y (the observed data).





Maximum likelihood

Using θ to refer to parameters, consider:

$$\hat{\theta}(\mathbf{y}) = \text{argmax}_\theta L(\theta|\mathbf{y})$$

The **maximum likelihood estimate** (MLE) is the θ that makes the observed data (y) most likely.

A general approach to statistical modeling:

- write down f(y|θ) (pdf/pmf: probability of outcomes conditional on parameters), which is also L(θ|y)
- observe data (y: actual outcomes)
- find parameters that maximize L(θ|y): the MLE!

Maximum likelihood (common notation)

$$
\begin{aligned}
\mathcal{L}(\theta|\mathbf{Y}) &= f(y_1, y_2, \ldots, y_n|\theta) \\
&= f(y_1|\theta) f(y_2|\theta) \ldots f(y_n|\theta) \\
&= \prod_{i=1}^{n} f(y_i|\theta)
\end{aligned}
$$

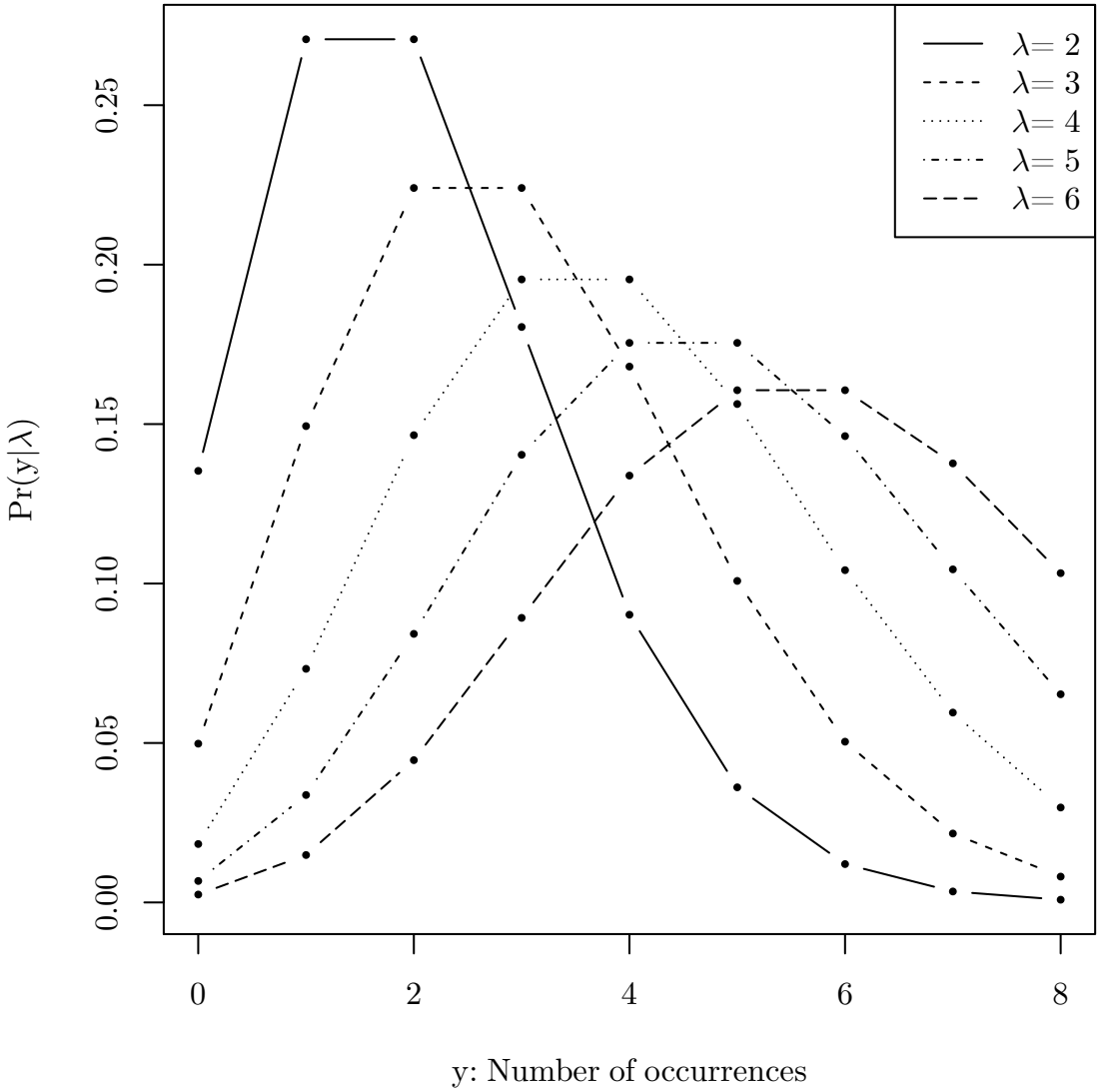
iid assumption

Vector of counts with a covariate

ACTIVITY!!!

Suppose we view the number of students sitting **in each row** as an independent Poisson random variable, with $\lambda = x_i$, where x_i is the number of the row.

How likely is the observed outcome for rows 3-6?

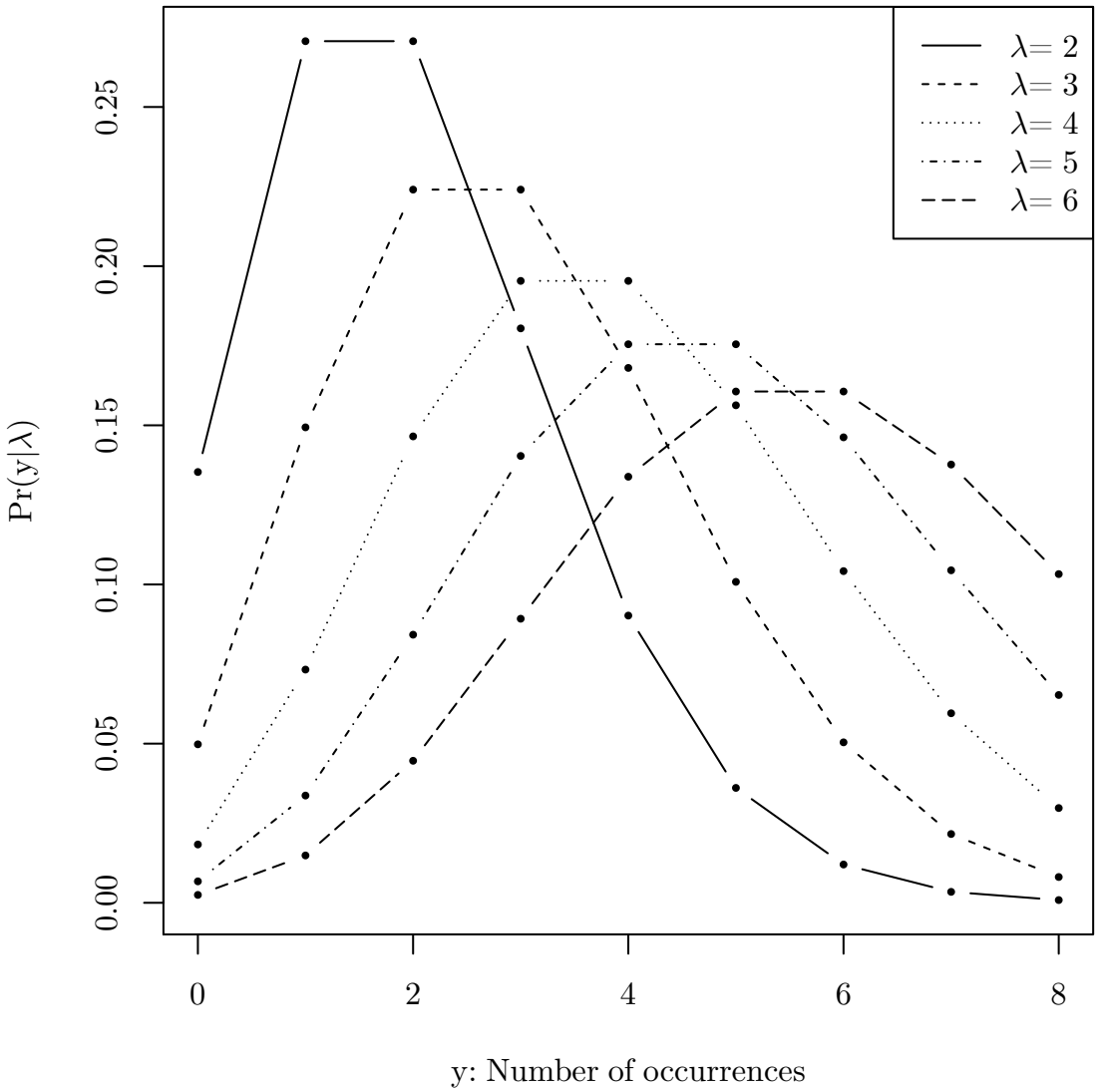


Vector of counts with a covariate

SOLUTIONS

Suppose we observe 5, 2, 7, and 4 students.

row	# students	λ = row
3	5	0.1
4	2	0.15
5	7	0.10
6	4	0.13
	Product X IM	206.52

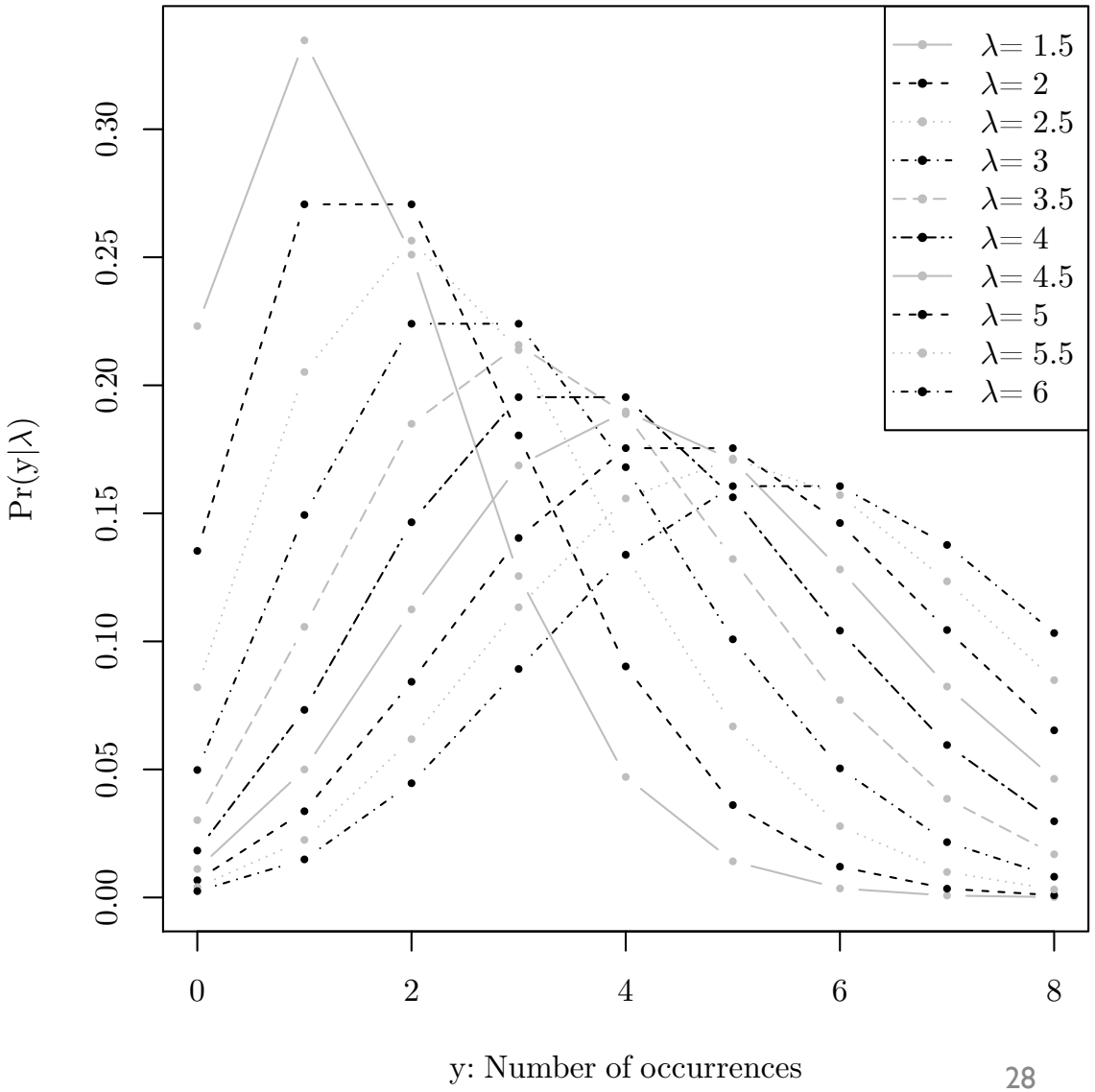


Vector of counts with a covariate (2)

ACTIVITY!!!

Suppose we view the number of students sitting **in each row** as an independent Poisson random variable, with $\lambda = \beta x_i$, where x_i is the number of the row.

1) If $\beta = 1$, how likely is the observed outcome for rows 3-6?
2) If $\beta = .5$, how likely is the observed outcome for rows 3-6?

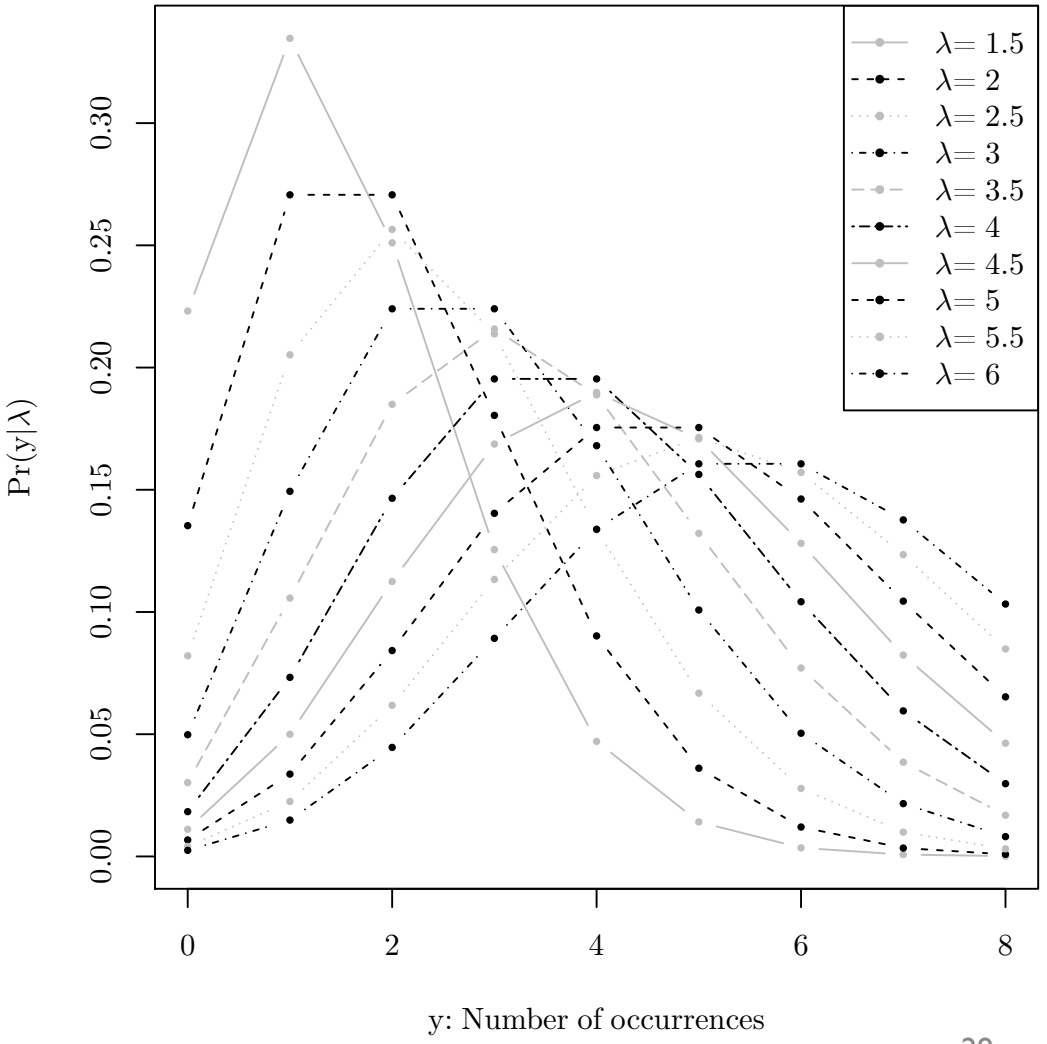


Vector of counts with a covariate (2)

SOLUTIONS

Suppose we observe
5, 2, 7, and 4 students.

row	# students	β = 1 (λ = row)	β = 1/2 (λ = row/2)
3	5	0.1	0.01
4	2	0.15	0.27
5	7	0.10	0.01
6	4	0.13	0.17
	Product X IM	206.52	6.38



Maximum likelihood (common notation)

$$\begin{aligned}\mathcal{L}(\theta|\mathbf{Y}) &= f(y_1, y_2, \ldots, y_n|\theta) \\ &= f(y_1|\theta) f(y_2|\theta) \ldots f(y_n|\theta) \\ &= \prod_{i=1}^{n} f(y_i|\theta)\end{aligned}$$

iid assumption

The likelihood function for the last MLE problem you just solved was:

$$\begin{aligned}\mathcal{L}(\theta|\mathbf{y}) &= f(y_3, y_4, y_5, y_6|\theta) \\ &= f(y_3|\theta) f(y_4|\theta) f(y_5|\theta) f(y_6|\theta) \\ &= \prod_{i=3}^{6} f(y_i|\theta) \\ &= \prod_{i=3}^{6} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} = \prod_{i=3}^{6} \frac{(x_i\beta)^{y_i} e^{-x_i\beta}}{y_i!}\end{aligned}$$

How statistical models look in research papers

A Statistical Method for Empirical Testing of Competing Theories

Kosuke Imai Princeton University
Dustin Tingley Harvard University

the model specified in equation (1) yields the following observed-data likelihood function where the latent variable Z_i has been integrated out,

$$L_{obs}(\Theta, \Pi | \{X_i, Y_i\}_{i=1}^N) = \prod_{i=1}^{N} \left\{ \sum_{m=1}^{M} \pi_m f_m(Y_i | X_i, \theta_m) \right\}. \quad (2)$$

Comparing Interest Group Scores across Time and Chambers: Adjusted ADA Scores for the U.S. Congress

TIM GROSECLOSE *Stanford University*
STEVEN D. LEVITT *University of Chicago*
and JAMES M. SNYDER, JR. *Massachusetts*

Given this representation, we can estimate a_t^c's, b_t^c's, and x_i's by maximizing the following likelihood function:

$$L(\bar{a}, \bar{b}, \bar{x}, \sigma; \bar{y}) = \prod_{t \in T} \prod_{c \in \{H,S\}} \prod_{i \in I_t^c} \phi\left(\frac{y_{it} - a_t^c - b_t^c x_i}{\sigma}\right) \frac{1}{\sigma},$$

How statistical models look in research papers

Ideology and Interests in the Political Marketplace

Adam Bonica Stanford University

Assuming independence across candidates and contributors, the log-likelihood to be maximized is,

$$LL(Y|\lambda,\sigma) = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{t=1}^{T}\sum_{g=0}^{1}(1-d_{ijt_g})\,ln\,(NB \times (y_{ijt_g}|\lambda_{ijt_g},\sigma_{it_g})) + (d_{ijt_g}) \qquad (3.3)$$
$$ln\left(1-\sum_{k=0}^{9}NB(k|\lambda_{ijt_g},\sigma_{it_g})\right)$$

where Y is an $n \times m$ matrix of observed contribution counts with y_{ijt_g} being the contribution amount of PAC i to candidate j in period t_g.

How statistical models look in research papers

How to Analyze Political Attention with Minimal Assumptions and Costs

Kevin M. Quinn University of California, Berkeley
Burt L. Monroe The Pennsylvania State University
Michael Colaresi Michigan State University
Michael H. Crespin University of Georgia
Dragomir R. Radev University of Michigan

As will become apparent later, it will be useful to write this sampling density in terms of latent data $\mathbf{z}_1, \ldots, \mathbf{z}_D$. Here $\mathbf{z}_d$ is a K-vector with element z_{dk} equal to 1 if document d was generated from topic k and 0 otherwise. If we could observe $\mathbf{z}_1, \ldots, \mathbf{z}_D$ we could write the sampling density above as

$$p(\mathbf{Y}, \mathbf{Z} \mid \boldsymbol{\pi}, \boldsymbol{\theta}) \propto \prod_{d=1}^{D} \prod_{k=1}^{K} \left(\pi_{s(d)k} \prod_{w=1}^{W} \theta_{kw}^{y_{dw}} \right)^{z_{dk}}.$$

Surveying a suite of algorithms that offer a solution to managing large document archives.

BY DAVID M. BLEI

Probabilistic Topic Models

With this notation, the generative process for LDA corresponds to the following joint distribution of the hidden and observed variables,

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right). \quad (1)$$

Some questions about statistical modeling

We said: a general approach to statistical modeling:

- write down $f(y|\theta)$ (pdf/pmf: probability of outcomes conditional on parameters), which is also $L(\theta|y)$
- observe data (y: actual outcomes)
- find parameters that maximize $L(\theta|y)$: the MLE!

Some questions:

- What interesting likelihoods could you write down?
- Why is the data generating process $f(y|\theta)$ random?
- What does this have to do with OLS?
- How do we know/choose $f(y|\theta)$ — and thus $L(\theta|y)$?

What does the randomness in statistical models mean?

Gailmard, Ch. 3; p. 178-179

The data we observe is stochastic because of:

1. Sampling uncertainty (units' outcomes are deterministic, but which units we see isn't)
2. Theoretical uncertainty/incompleteness (the stuff we don't know about or ignore)
3. Fundamental uncertainty (things would happen differently even in a "do-over"*)

Sean Gailmard

*Quantum uncertainty? People playing mixed strategies? see p. 182

How does OLS fit in?

OLS has attractive predictive/descriptive features *independent of a statistical model*. Most importantly, the solution to

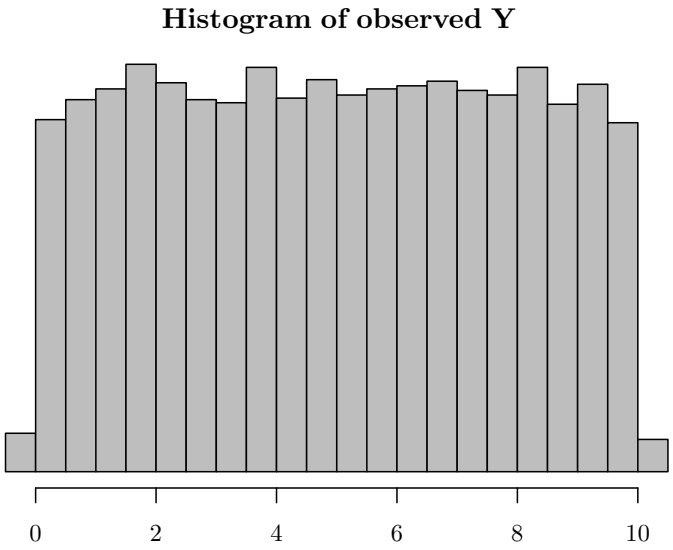
$$\text{argmin}_{\alpha,\beta} \sum_{i=1}^{n} \left(y_i - \alpha - \beta x_i \right)^2$$

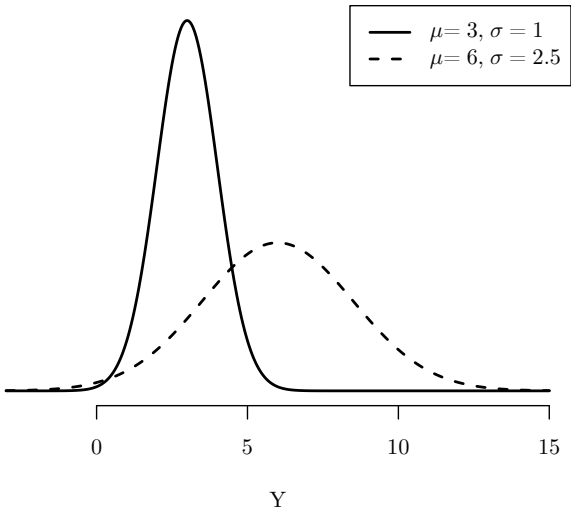
will give the best (minimum mean squared error) linear approximation of Y|X and E[Y|X] (the CEF) **regardless of linearity of CEF, distribution of errors**.* (Inference also works asymptotically.)

But the OLS coefficients are also the MLE in a statistical model where $Y \sim N(\alpha + \beta X, \sigma^2)$.

*This is main message of MHE 3.1 and Gailmard 132-135; see also Gailmard 314 ff.

How do we know/choose the likelihood?





Q: What type of probability distribution produced the outcome at left?

A: Normal distribution, with μ uniformly distributed from 0 to 10 and small σ^2

How do we know/choose the likelihood (2)?

Gailmard, 6.1.2

Researchers choose parametric family (Normal, Poisson, etc) of the **stochastic part**

- for good theoretical reasons (e.g. Y is a count in a period, with constant rate; Y is the sum of random variables)
- to avoid predicting impossible values of Y (range matching)
- for flexibility
- convenience/interpretability (e.g. logit)

The **systematic part** is more important: how X relates to parameters of the distribution.

In standard GLMs, this is just like OLS regression.

More generally, statistical models can reflect any assumption about what factors matter and how.

How do we know/choose the likelihood? (3)

'Statistics cannot tell a substantive researcher whether a particular model is "right" or "wrong" for a particular application. Statistics works well as a complement to researchers who figure that out for themselves based on their knowledge of what they are studying; statistics then offers a way to learn as much as possible about the DGP and to estimate the uncertainty about what was actually learned.'

Gailmard, p. 316

Sean Gailmard

Statistical models and research questions

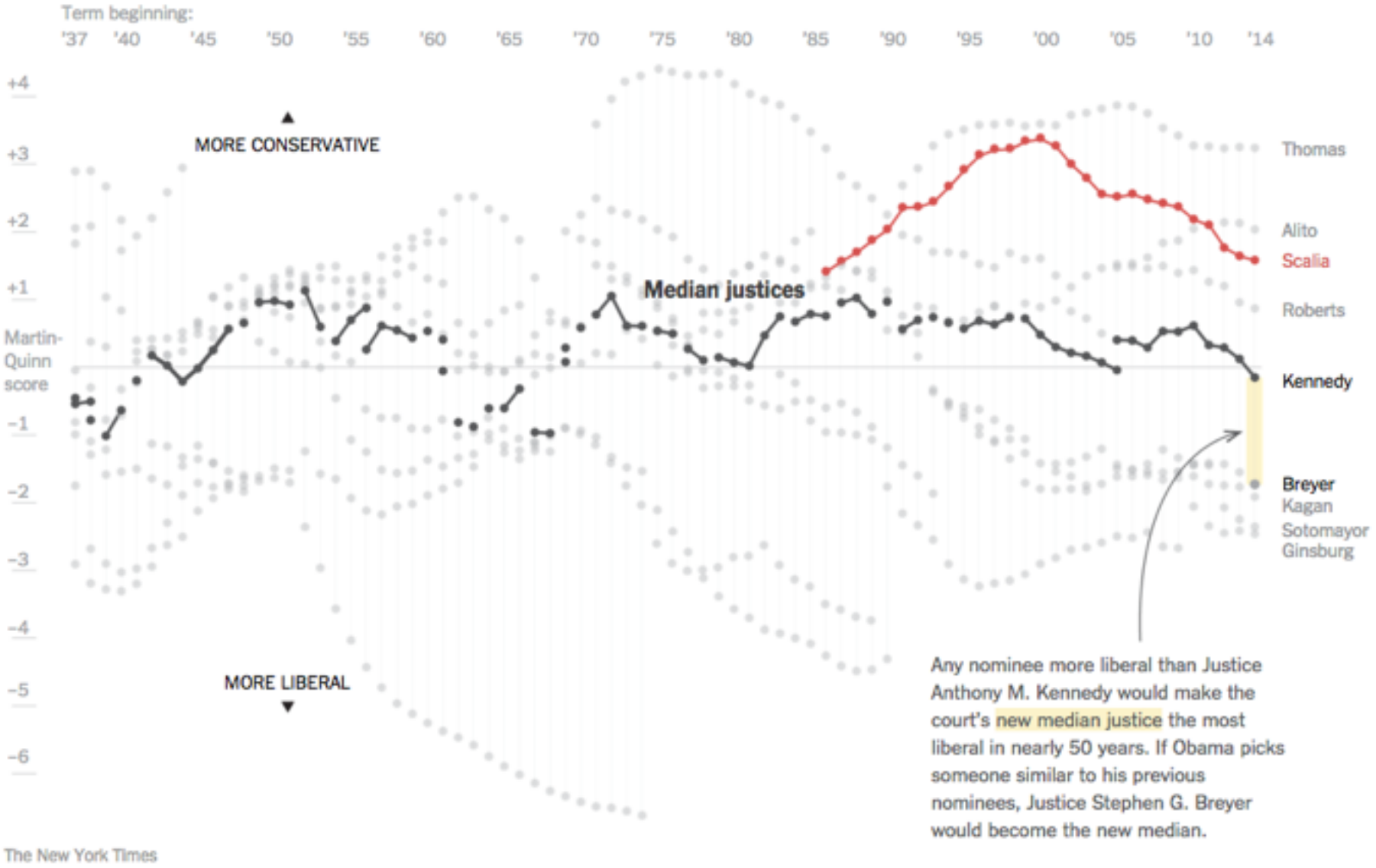
Research questions in social science are never directly about statistical models.

And if questions are about **observed data**, no model is necessary. For example:

- *On this survey question, is the proportion of German respondents answering 'Yes' higher than the proportion of French respondents answering 'Yes'?*

But they might be about elements estimated in statistical models:

- Is the German public ideologically more conservative than the French public?
- Is the congruence between legislators' policy preferences and public opinion closer in German states than in French regions?



Analysis on nytimes.com using Martin-Quinn scores (http://mqscores.berkeley.edu/) See "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953-1999", Andrew D. Martin and Kevin M. Quinn 2002.

Statistical models and causal inference

We ignored statistical models in weeks 1-5.

But OLS regression connects to a model:

- MHE view: OLS provides linear approximation to the DGP, where standard errors approach true sampling distribution as n increases
- classical view: under assumptions including normal errors, OLS provides unbiased estimate of linear DGP parameters, good standard errors even in small samples

Alternative approaches:

- **Randomization inference** in Fisher's exact test: inference about the "sharp null" with no assumptions about statistical model; the "sample" is the population of interest
- "**Finite Population Causal Standard Errors**" (Abadie, Susan Athey, Imbens, Wooldridge working paper): causal inference requires a different interpretation of standard errors because it is inference about missing potential outcomes, not DGP parameters

Coming up

Lab: practice with GLMs

Next two lectures:

- Ordinal probit model (in application)
- Can I just use OLS?
- Measurement/scaling models