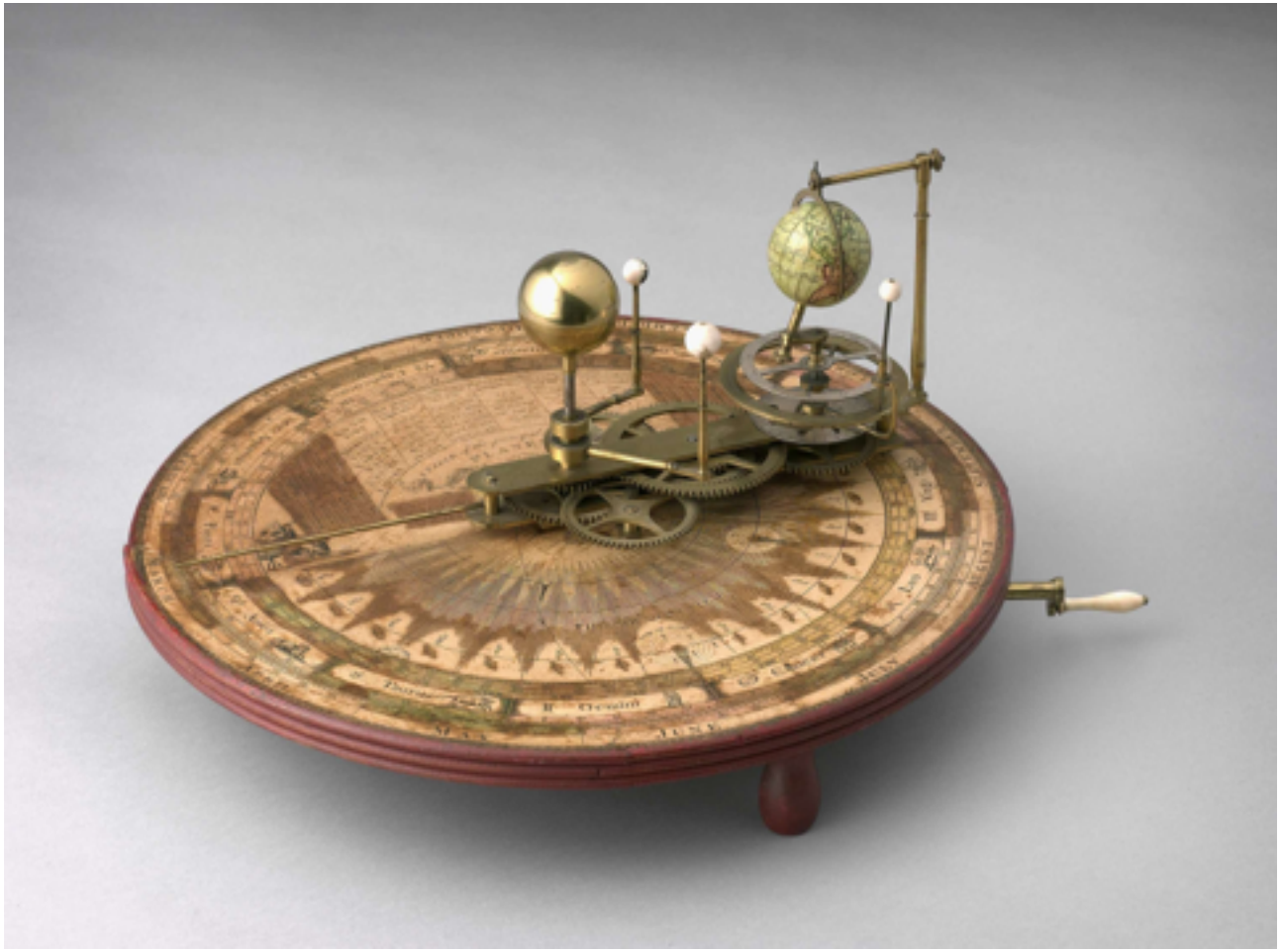
DPIR SPRING SCHOOL

Foundations of statistical modeling for text analysis

13 April 2016

Andy Eggers

What is a model?

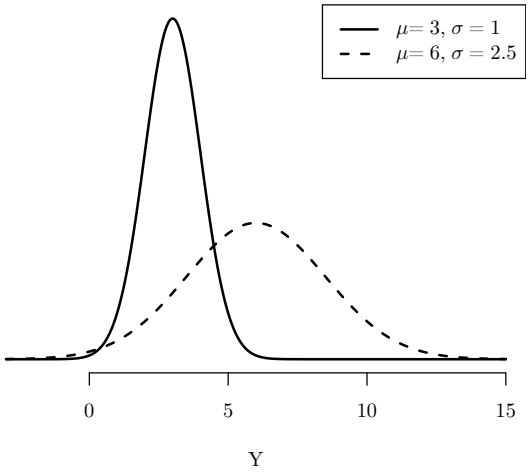


Jones's "New Portable Orrery" (1794)

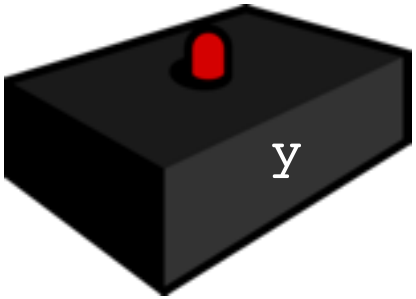
"All models are wrong, but some are useful." George Box

How probability works

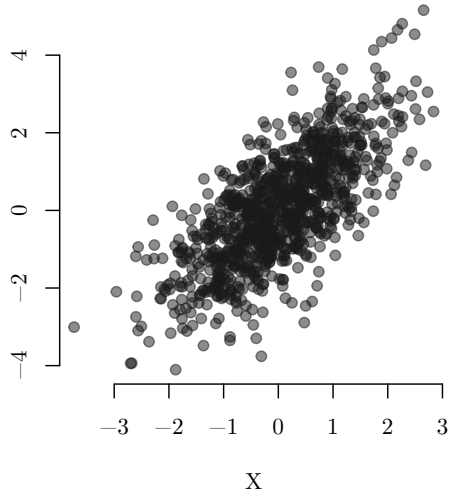
Given a set of probability distributions…



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

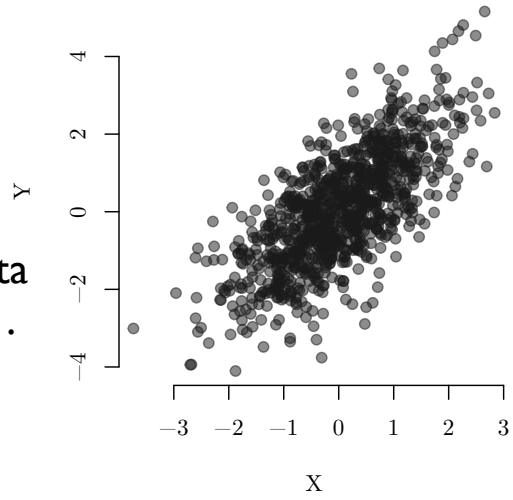


…what data should we expect to observe?

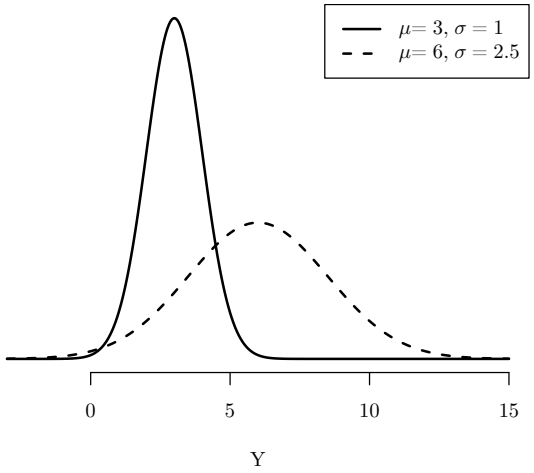


How statistics works

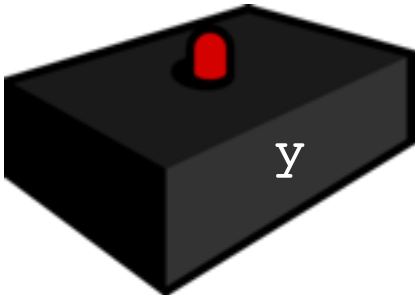
Given the data we observe…



…what set of probability distributions…



characterize the DGP?



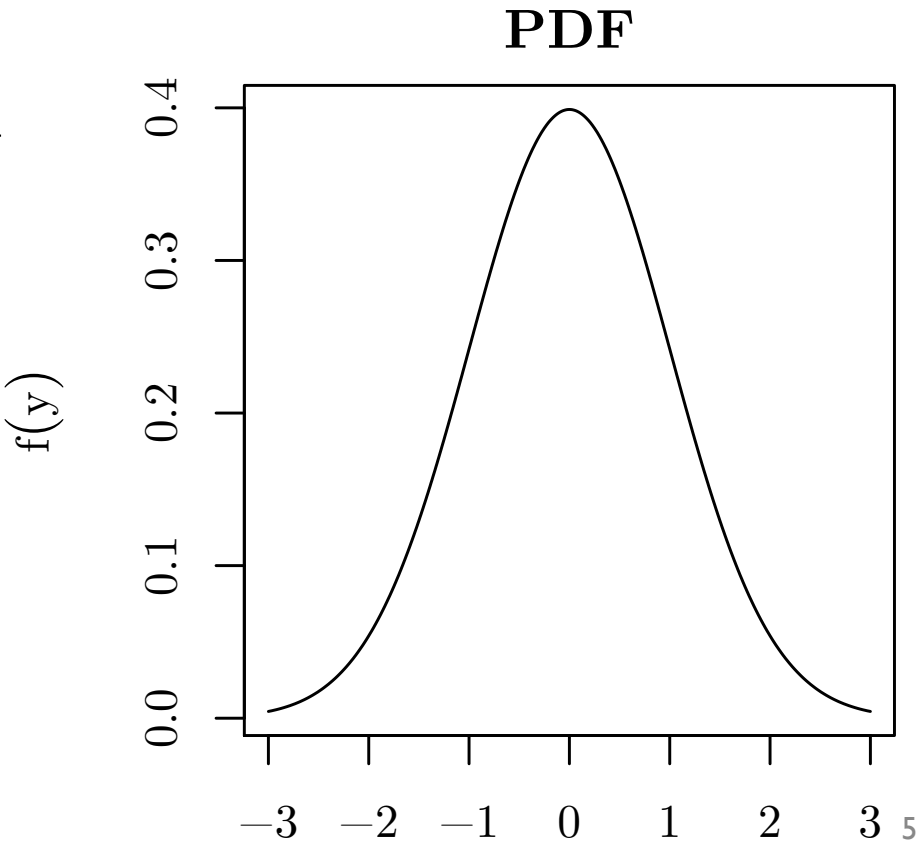
Probability distributions (PDF/PMF)

Continuous random variables are described by probability density function (PDF). (Discrete RVs by probability mass function, PMF.)

$$f(y|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Two parameters in normal PDF: mean (μ) and variance (σ^2).

Describes sum/average of many random variables.



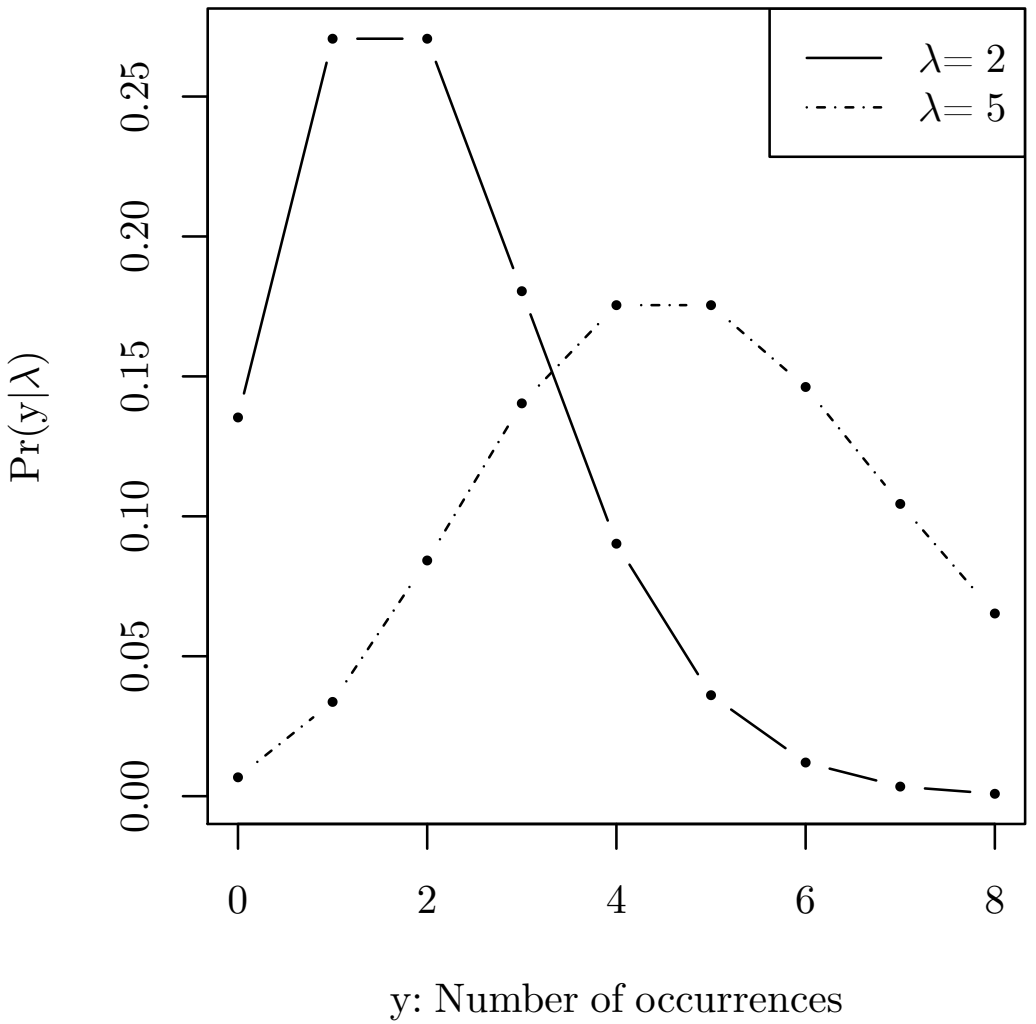
Poisson PMF

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

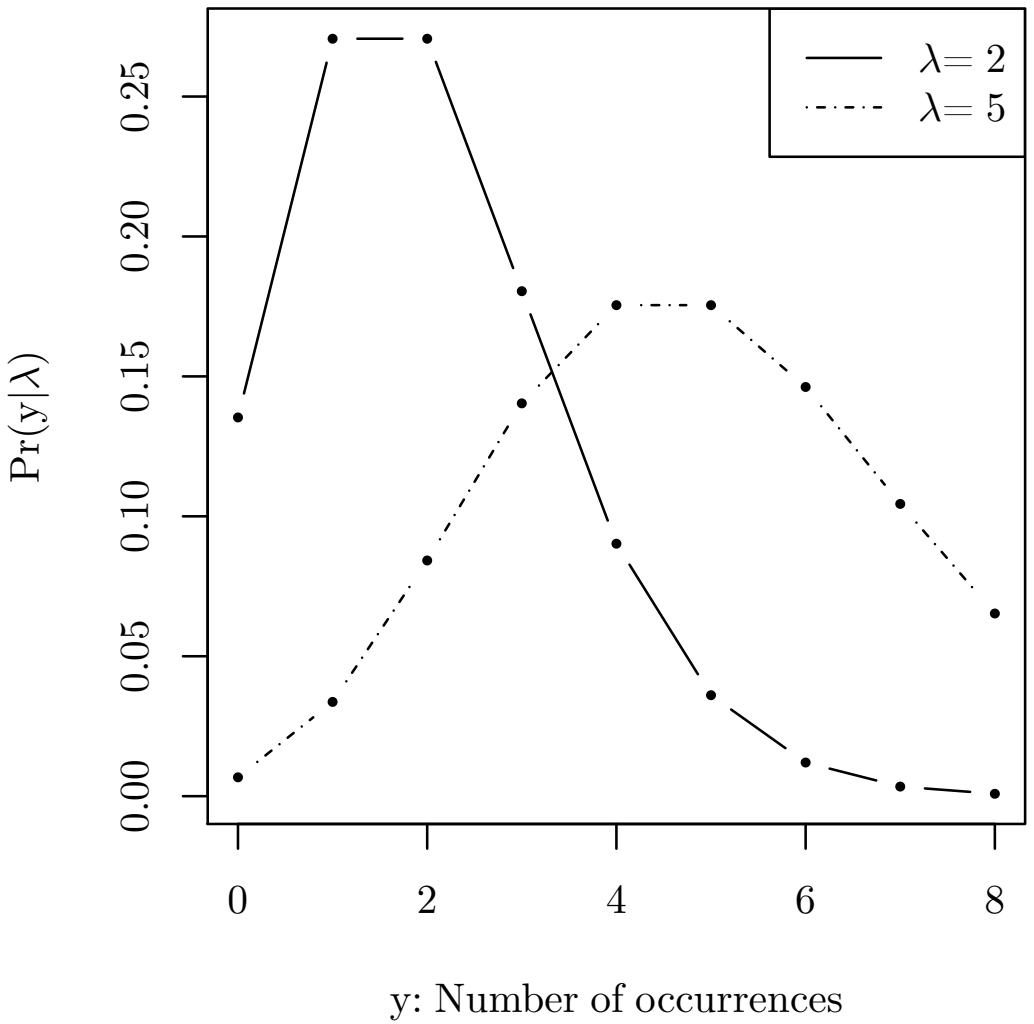
- events are independent
- rate of occurrence (probability per unit time) is constant



Single count

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

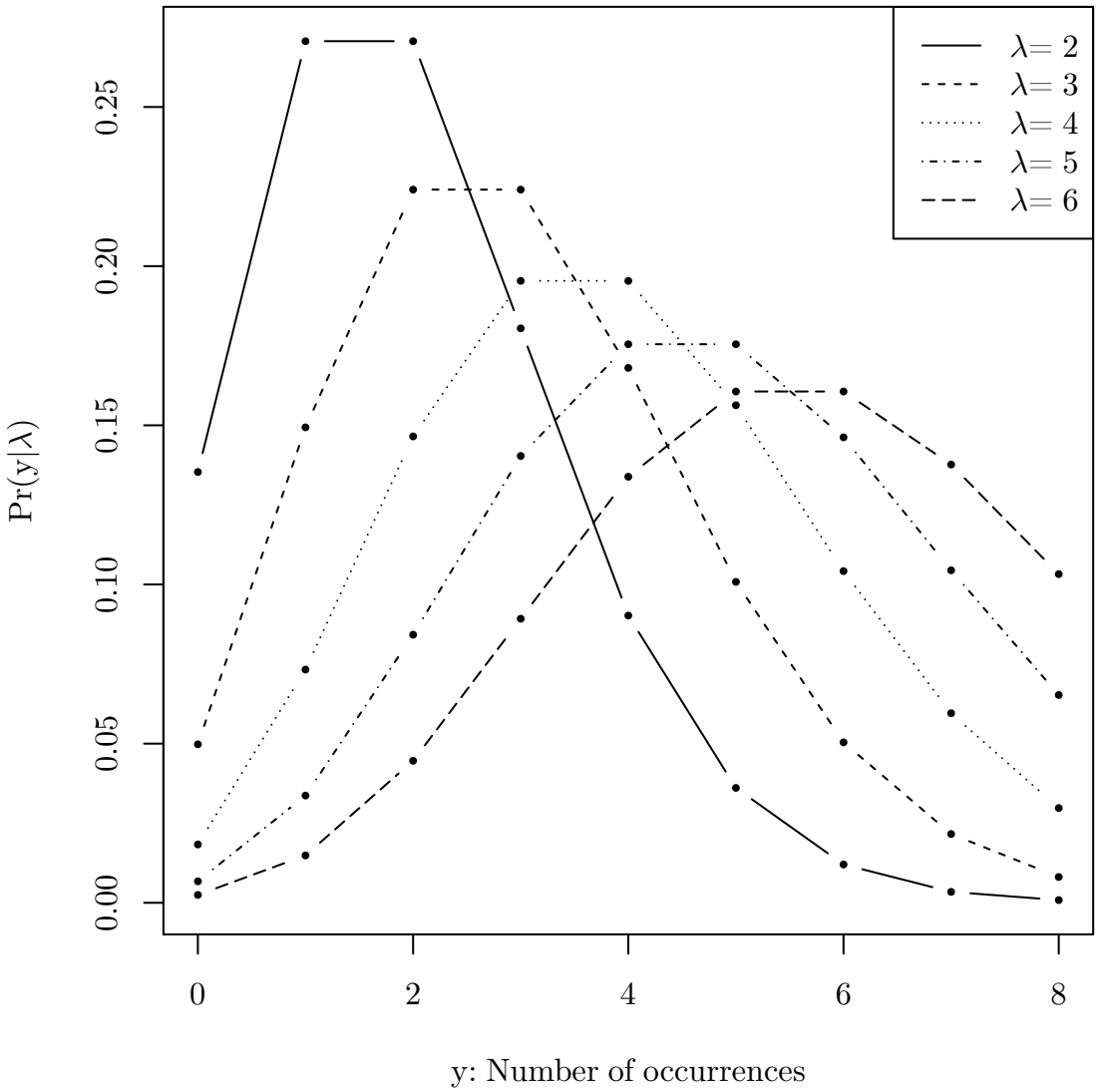
1) If λ = 2, how likely is the observed outcome?
2) If λ = 5, how likely is the observed outcome?



Single count (2)

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,

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

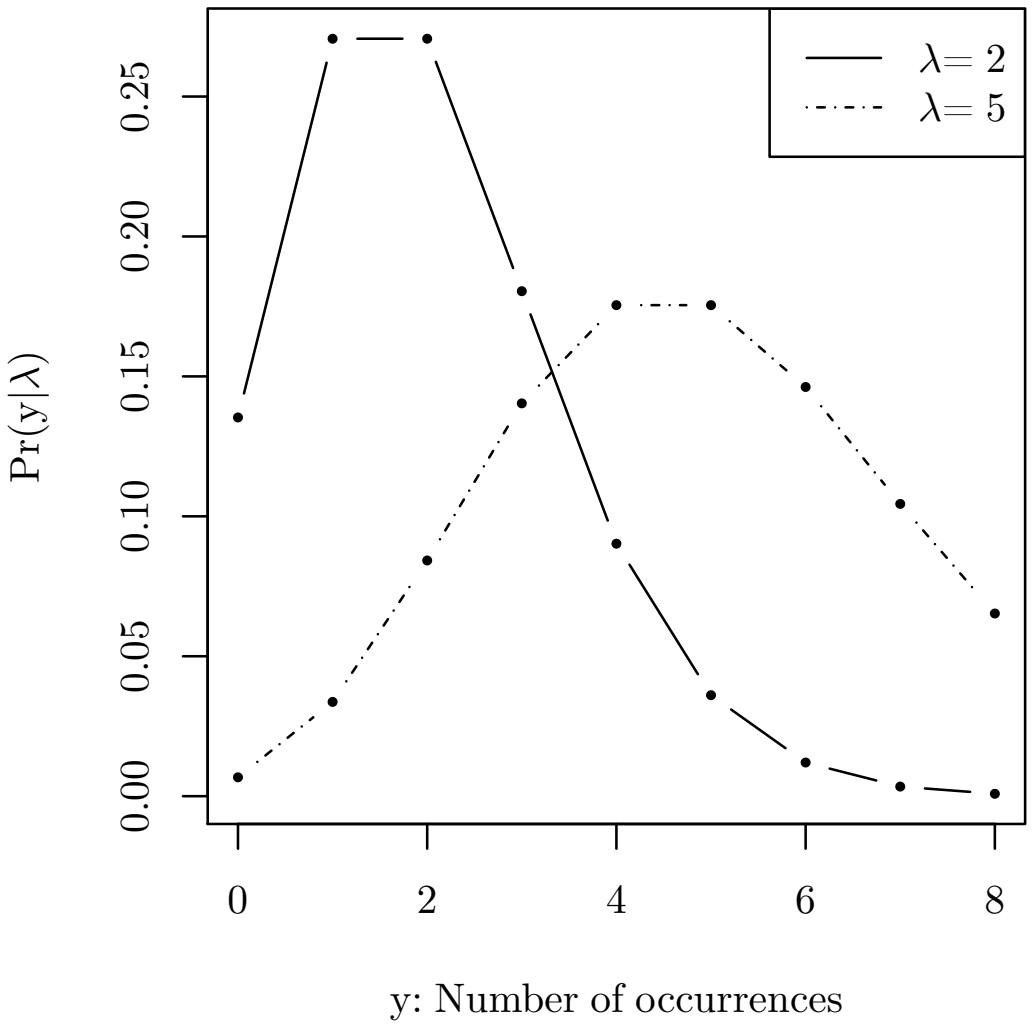
and:

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

Vector of counts

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

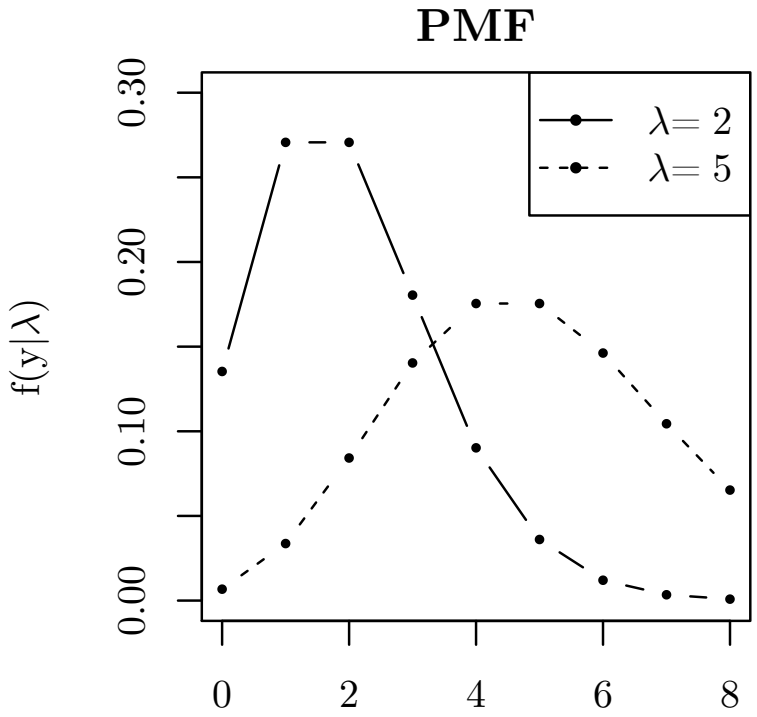


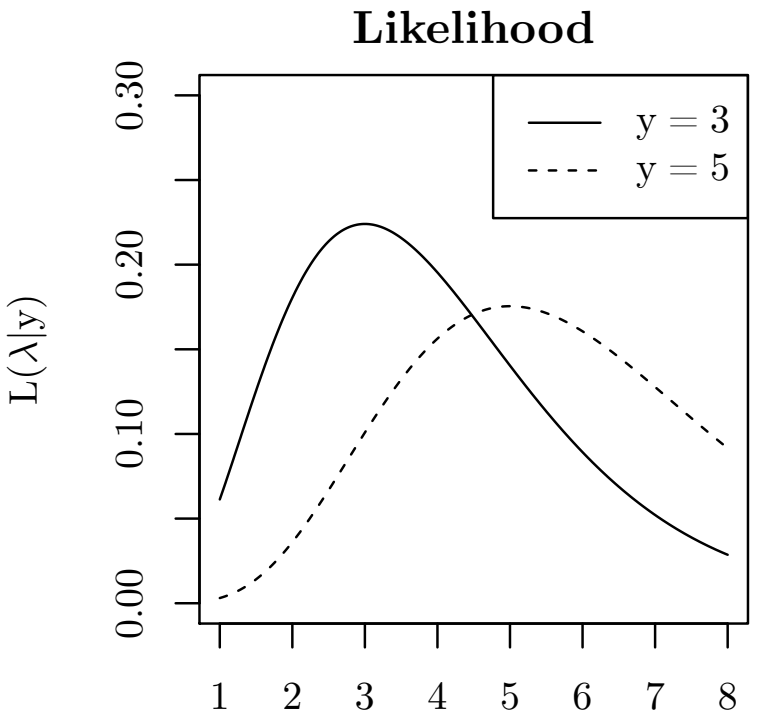
Likelihood vs probability

We postulate a model for how this data was generated (Poisson trials).

For one value of model parameter λ, we can consider different possible outcomes and calculate the probability of them occurring. (That is the PMF/PDF!)

Given one set of observed data, we can consider different model parameters λ and calculate the probability of observing the data for each one. (This is the likelihood!)



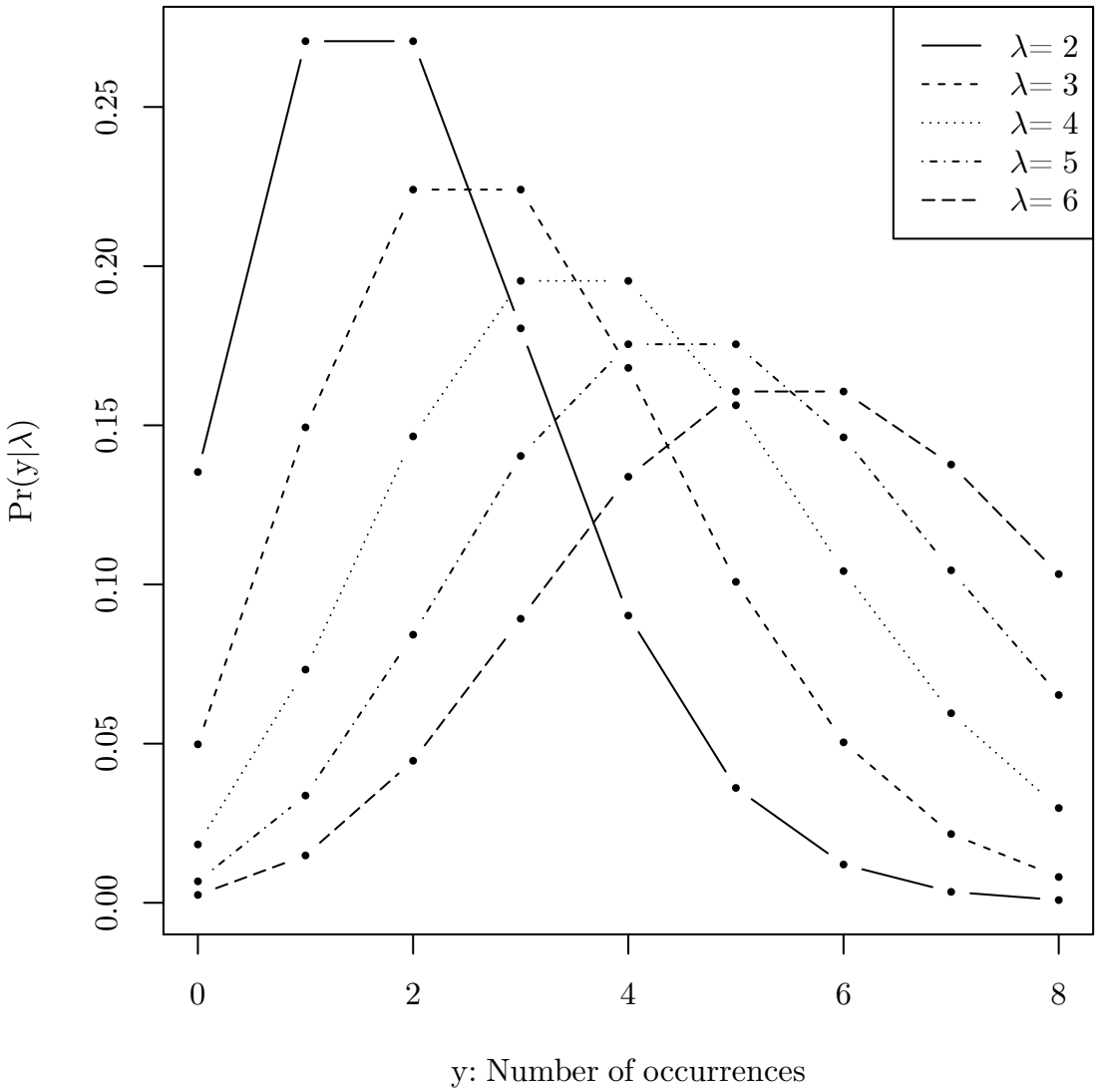


Vector of counts with a covariate

Suppose we view the number of students sitting **in each row** as an independent Poisson random variable, and we assume

$$\lambda_i = \beta \text{row}_i$$

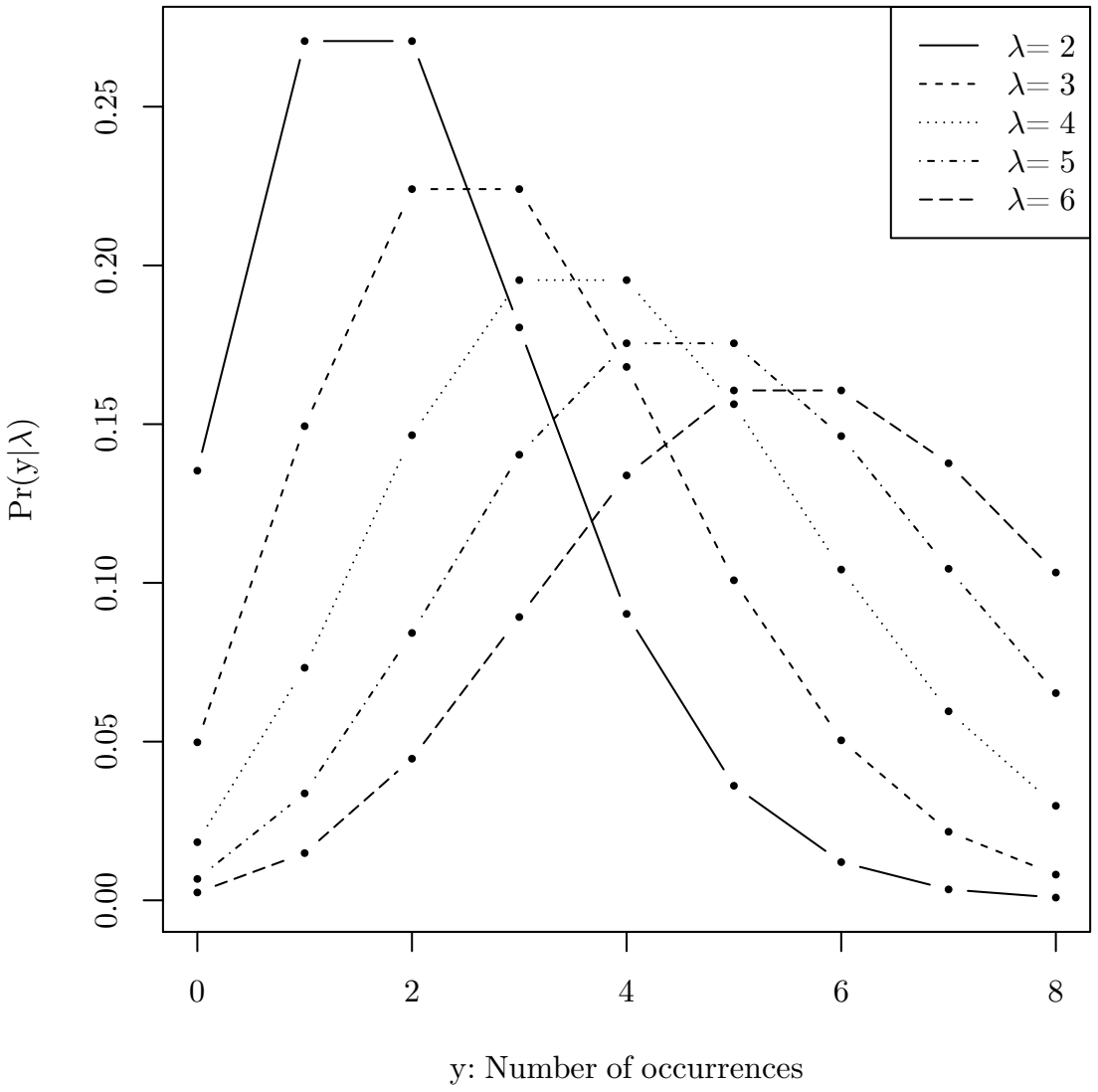
How likely is the observed outcome for rows 3-5 if β = 1?



Vector of counts with a covariate

Suppose we observe 5, 2, and 4 students.

row	# students	$\lambda = row_i$
3	5	0.1
4	2	0.15
6	4	0.13
		0.00195



Vector of counts with a covariate (2)

We assumed

$$\lambda_i = \beta \mathrm{row}_i$$

so now we want to choose the β that yields the vector of λ_i that maximizes the probability of seeing the data we observed.

The same as searching for the λ that makes the data most likely; just one intervening step.

Maximum likelihood

Using θ to refer to parameters (e.g. λ, β), 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: a function of the outcome, conditional on parameters); if we think of this as a function of the parameters, conditional on the outcome, we have L(θ|y) (the likelihood)
- observe data (y: actual outcomes)
- find parameters that maximize L(θ|y): the MLE!

Maximum likelihood (common notation)

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

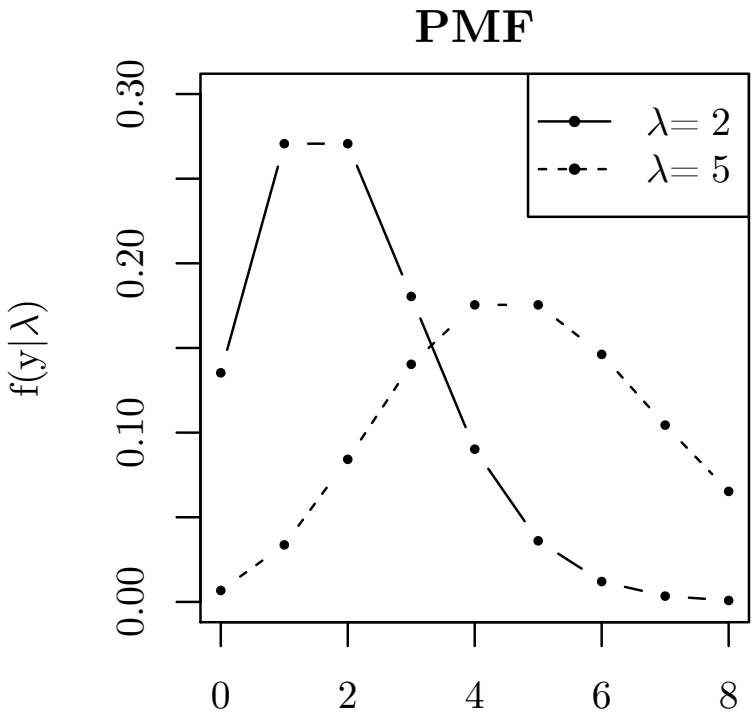
iid assumption

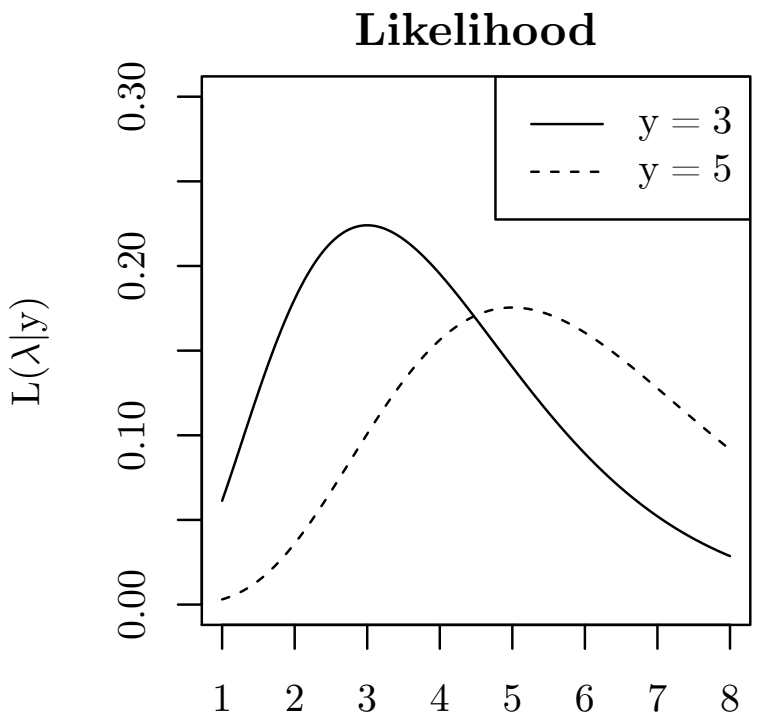
Likelihood is not probability

Finding parameters that maximize the likelihood is a good idea.

But likelihood is **not** probability:

- given λ=2, what is the probability of observing y=4
- given y=4, what is the probability that λ=2?





Maximum likelihood and Bayesian analysis

$$L(\theta|y) = \Pr(y|\theta)$$

$$\Pr(\theta|y) = \frac{\Pr(y|\theta)\Pr(\theta)}{\Pr(y)}$$

Likelihood | Prior

Bayes Law

$$\Pr(\theta|y) \propto \Pr(y|\theta)\Pr(\theta)$$

Proportional to

So: Bayesian analysis requires prior distributions on parameters, and allows us to make probability statements about the parameters (conditional on the model).

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

A topic model: what is each speech in Congress about?

First they describe the generative model, i.e. the likelihood or "sampling density":

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

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

Then they state the prior distributions for the parameters:

To complete a Bayesian specification of this model we need to determine prior distributions for θ and π. We assume a semiconjugate Dirichlet prior for θ. More specifically, we assume

$$\boldsymbol{\theta}_k \sim \mathcal{D}irichlet(\boldsymbol{\lambda}_k) \quad k = 1, \ldots, K.$$

For the data analysis below we assume that $\lambda_{kw} = 1.01$ for all k and w. This corresponds to a nearly flat prior over $\boldsymbol{\theta}_k$. This prior was chosen before looking at the data.

The prior for π is more complicated. Let $\boldsymbol{\pi}_t \in \Delta^{K-1}$ denote the vector of topic probabilities at time t. The model assumes that a priori

$$\mathbf{z}_d \sim \mathcal{M}ultinomial(1, \boldsymbol{\pi}_{s(d)}).$$