Parametric Models of Linguistic Count Data

Martin Jansche

ACL 2003

What this talk is about

- Accurately modeling discrete properties of text documents, such as length, word frequencies, etc.
- Focus on word frequency in documents.
- Claim 1: In addition to overdispersion, variation of word frequency across documents is partly due to zero-inflation.
- Claim 2: Modeling zero-inflation is sometimes preferable to modeling overdispersion.

What this talk is not about

- Estimating the probabilities of unseen words. Instead, focus on words that occur zero times in most documents (true of most words!), but do occur a few times in a small number of documents.
- State-of-the-art text classification. Text classification using an independent feature model is used merely for illustration, since it is simple and benefits from richer models for individual features.

Parametric models

- Encode all properties of a distribution in (typically) very few parameters.
- Easy to incorporate prior information about plausible values for parameters.
- Can work with very small amounts of data.
- Can work with sparse data.
- Often closed form expressions are available for moments, probabilities, percentiles, etc.

Linguistic count data

- Focus on modeling document length and word frequency in documents.
- Sample sizes are often small: most words are extremely rare and most documents are fairly short.
- Overdispersion: natural variation not well captured by simple models with very few parameters.
- Zero-inflation: most words occur zero times in a given document; not captured by standard models.

Claim 1

- Overdispersed models can capture increased variance of token frequency [Mosteller and Wallace 1964, 1984; Church and Gale 1995].
- Zero-inflation accounts for variation not captured by overdispersed models.
- Need to develop a zero-inflated extension of a robust, overdispersed model of token frequency.
- Zero-inflation can be observed in M&W's data.

Poisson family models

Start with the Poisson distribution with rate $\lambda > 0$:

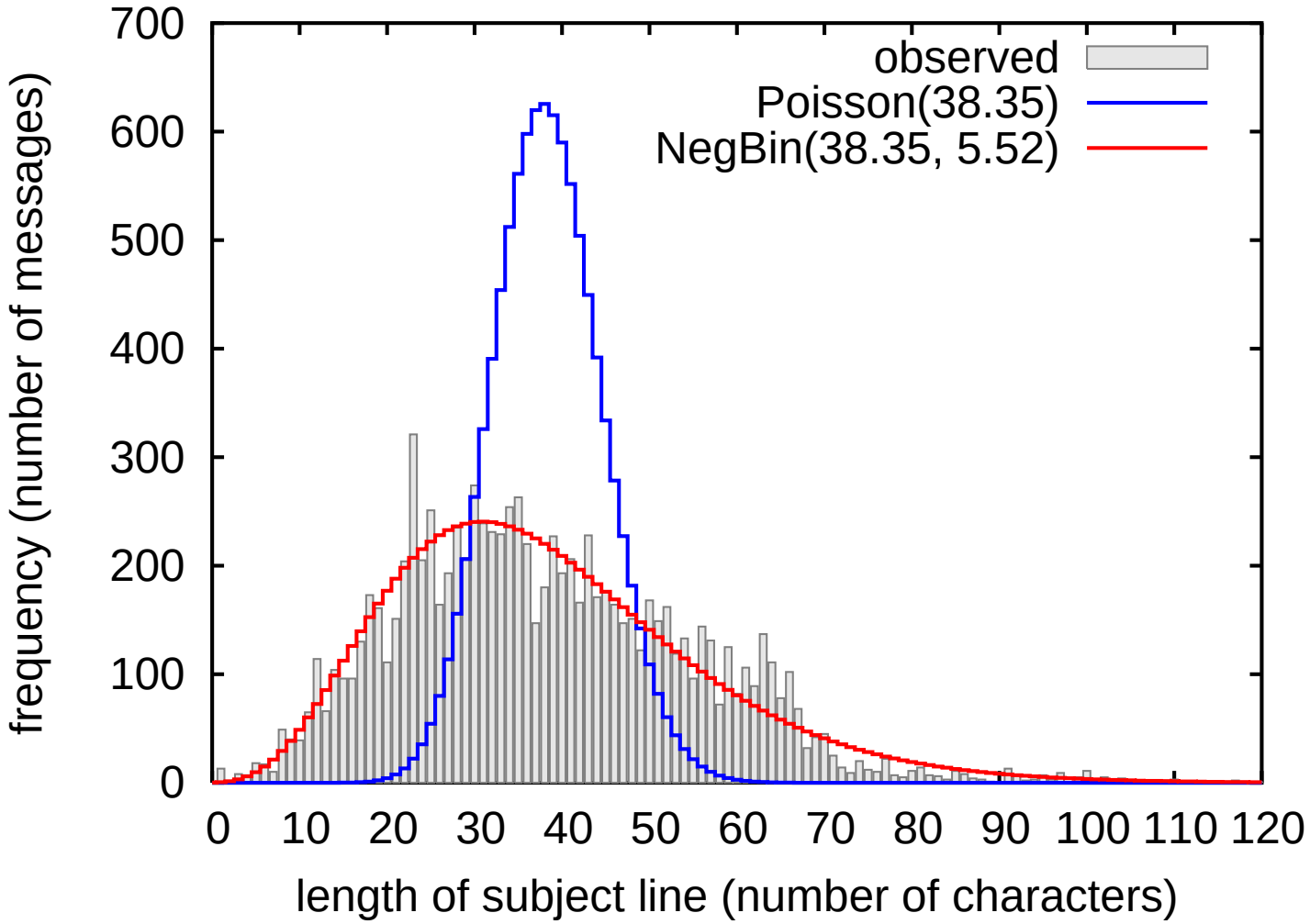
$$\text{Poisson}(\lambda)(x) = \frac{\lambda^x}{x!} \exp(-\lambda).$$

A natural generalization of the Poisson is the Negative Binomial distribution, with an additional parameter $\kappa > 0$ that controls non-Poissonness:

$$\text{NegBin}(\lambda, \kappa)(x) = \frac{\lambda^x}{x!} \frac{\Gamma(\kappa + x)}{\Gamma(\kappa)(\lambda + \kappa)^x} \frac{\kappa^\kappa}{(\lambda + \kappa)^\kappa}$$

	Poisson(λ)		NegBin(λ, κ)
Mean μ	λ	$=$	λ
Variance σ^2	λ	$\leq$	$\lambda\,(1 + \lambda/\kappa)$
Skewness γ_1	$\dfrac{1}{\sqrt{\lambda}}$	$\leq$	$\dfrac{1}{\sqrt{\lambda}}\,\dfrac{\kappa + 2\lambda}{\sqrt{\kappa\,(\lambda + \kappa)}}$
Kurtosis γ_2	$\dfrac{1}{\lambda}$	$\leq$	$\dfrac{1}{\lambda}\,\dfrac{\kappa}{\lambda + \kappa} + \dfrac{6}{\kappa}$
Mode	$\lfloor \lambda \rfloor$	$\geq$	$\lfloor \lambda\,(1 - 1/\kappa) \rfloor$ if $\kappa \geq 1$ 0 if $0 < \kappa < 1$

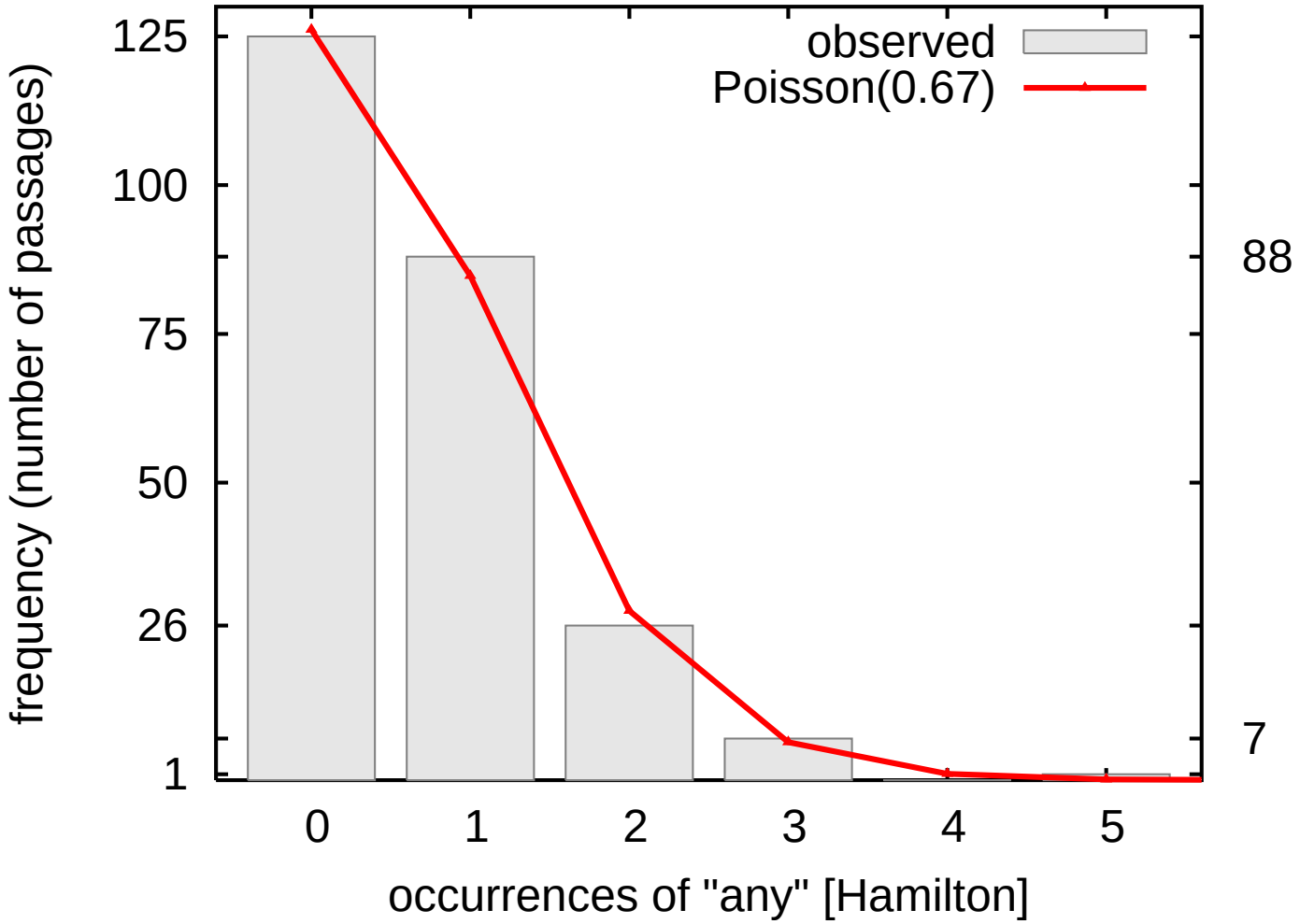
Example: Subject lines of spam email



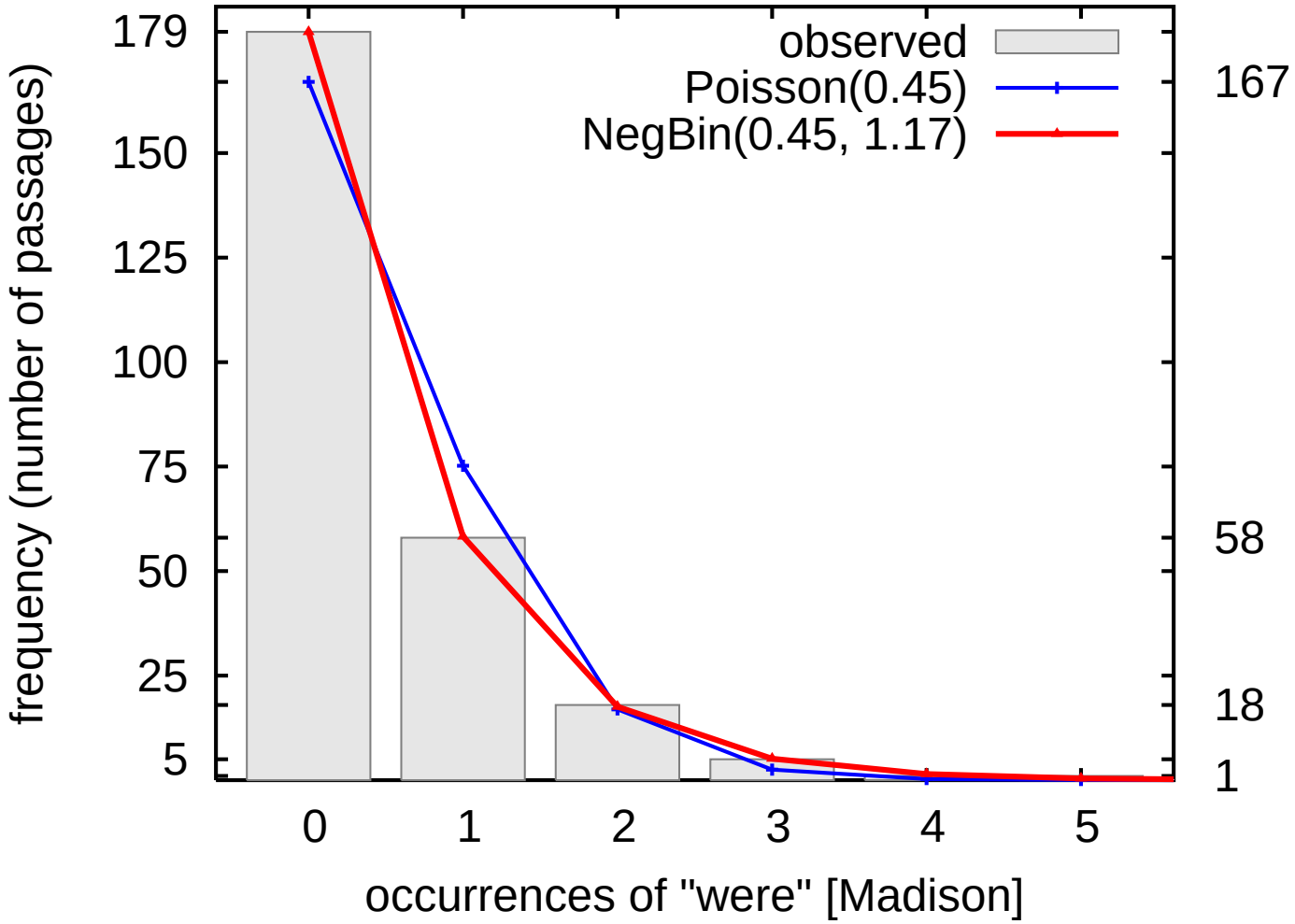
Detailed examples

- Mosteller and Wallace's [1964, 1984] data, taken from *The Federalist* papers.
- Essays by Alexander Hamilton and James Madison (and John Jay) on the shape of the proposed US constitution.
- M&W sampled approx. 250 contiguous passages of equal length for each of the two main authors.

Some words follow the Poisson



Some words follow the Neg. Binomial



And some words are special

For example 'his' (Hamilton and Madison pooled) in Mosteller and Wallace's data.

The method of maximum likelihood leads to NegBin$(0.54, 0.15)$. Here's what that model has to say:

	0	1	2	3	4	5	6	7	8	14
obsrvd	405	39	26	18	5	4	5	3	3	1
expctd	404	48	22	12	7	5	3	2	2	0

Alternatively, we could have estimated the parameters based on: (a) the number of documents with zero occurrences of 'his'; and (b) the number of documents with one occurrence of 'his'. Not surprisingly, the resulting model, NegBin$(0.76, 0.11)$, is worse:

	0	1	2	3	4	5	6	7	8	14
obsrvd	405	39	26	18	5	4	5	3	3	1
expctd	405	39	19	12	8	6	4	3	2	1

Adaptation, burstiness and all that

Church [2000]: "The first mention of a word obviously depends on frequency, but surprisingly, the second does not. Adaptation [the degree to which the probability of a word encountered in recent context is increased] depends more on lexical content than frequency[.]"

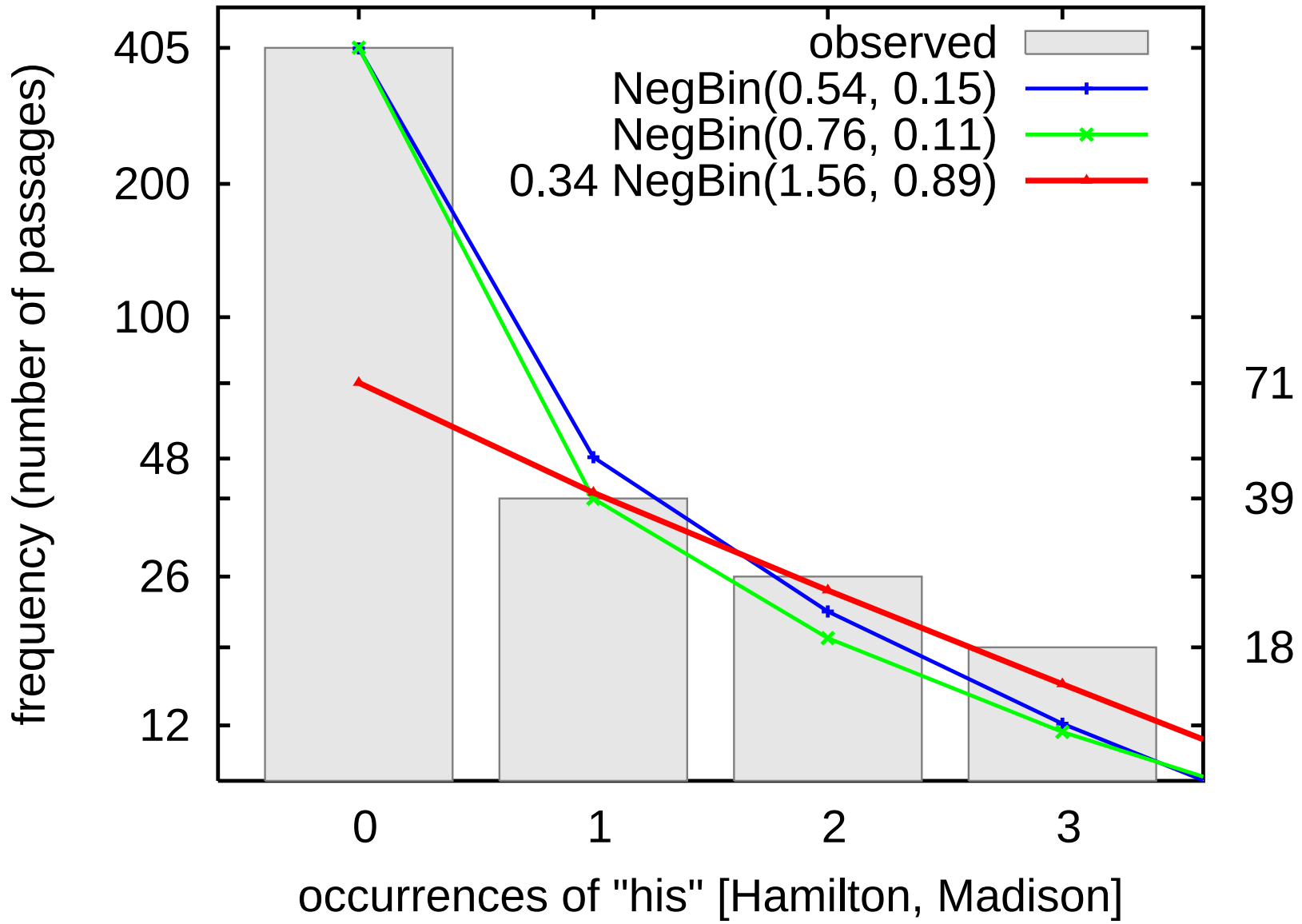
Church, concerned mostly with empirical exploration, used nonparametric methods. How can his findings be incorporated into a parametric setting?

A modest proposal

Whether a given word appears at all in a document is one thing. How often it appears, if it does, is another thing.

Not all words are appropriate in a given context (taboo words, technical jargon, proper names). A writer's/speaker's active vocabulary is limited and idiosyncratic ('(tom/pot)atos'/'(tom/pot)atoes').

We insist on capturing non-zero occurrences with parametric models, but treat zeroes specially.


observed
NegBin(0.54, 0.15)
NegBin(0.76, 0.11)
0.34 NegBin(1.56, 0.89)
405
200
100
48
26
12
71
39
18
0
1
2
3
frequency (number of passages)
occurrences of "his" [Hamilton, Madison]

A concrete modest proposal

Two-component mixture: first component is a degenerate distribution at zero (or possibly a geometric distribution starting at zero); second component a standard distribution, e. g. from the Poisson or Binomial family.

$$\mathsf{ZI}\mathcal{F}(z,\theta)(x) = z(x \equiv 0) + (1-z)\mathcal{F}(\theta)(x)$$

where $0 \leq z \leq 1$ ($z < 0$ may be allowable).

Properties of ZI$\mathcal{F}$

If $\mathcal{F}(\theta)$ has mean μ and variance σ^2, then ZI$\mathcal{F}(z, \theta)$ has mean

$$(1 - z)\, \mu$$

and variance

$$(1 - z)\, (\sigma^2 + z\, \mu^2).$$

Furthermore, ZI$\mathcal{F}(z, \theta)$ has the same modes as $\mathcal{F}(\theta)$ plus potentially an additional mode at zero.

Zero-inflated distributions

Straightforward interpretation of generative process: flip a z-biased coin; on heads, generate 0; on tails, generate according to $\mathcal{F}$.

If parameter vector θ of $\mathcal{F}$ can be estimated straightforwardly, use EM to estimate z and θ. Otherwise use multidimenisional maximization algorithms.

ZINB model for 'his'

Recall that a NegBin model can already account for the fact that most of the probability mass is concentrated at zero. Can a zero-inflated NegBin (ZINB) model do better?

Note that the maximum likelihood models for the distribution of 'his' in M&W's data say very different things, even though the net effects may be superficially similar.

The NegBin model claims that 'his' occurs much less than once on average (0.54 expected occurrences) and that it has large variance.

The ZINB model claims that 'his' occurs in only a third of all passages, but within those its expected number of occurrences is 1.56 and its variance is less than that predicted by the NegBin model.

		NegBin	ZINB
	obsrvd	expctd	expctd
0	405	403.853	405.000
1	39	48.333	40.207
2	26	21.686	24.206
3	18	12.108	14.868
4	5	7.424	9.223
5–6	9	8.001	9.361
7–14	7	6.996	5.977
χ^2 q-value		0.832	0.601
$-\log L(\hat{\theta})$		441.585	439.596

Comparison of Poisson models

$$x \sim \text{Poisson}(\lambda)$$
$$\mu = \lambda$$
$$\sigma^2 = \lambda = \mu$$

$$x \sim \text{NegBin}(\lambda, \kappa)$$
$$\mu = \lambda,$$
$$\sigma^2 = \lambda\,(1 + \tfrac{\lambda}{\kappa})$$

$$x \sim \text{ZIPoisson}(z, \lambda)$$
$$\mu = (1 - z)\,\lambda,$$
$$\sigma^2 = \mu\,(1 + z\,\lambda)$$

$$x \sim \text{ZINegBin}(z, \lambda, \kappa)$$

Comparison of Binomial models

$$x \mid n \sim \text{Binom}(p)$$
$$\mu = n\,p$$
$$\sigma^2 = n\,p\,(1-p) = \mu\,q$$

$$x \mid n \sim \text{BetaBin}(p, \gamma)$$
$$\mu = n\,p$$
$$\sigma^2 = \mu\,q\,(1 + (n-1)\gamma)$$

$$x \mid n \sim \text{ZIBinom}(z, p)$$
$$\mu = (1-z)\,n\,p$$
$$\sigma^2 = \mu\,(q + z\,n\,p)$$

$$x \mid n \sim \text{ZIBetaBin}(z, p, \gamma)$$

Document classification

20 Newsgroups data set, stratified so that all classes are equally likely a priori, therefore 5% baseline accuracy.

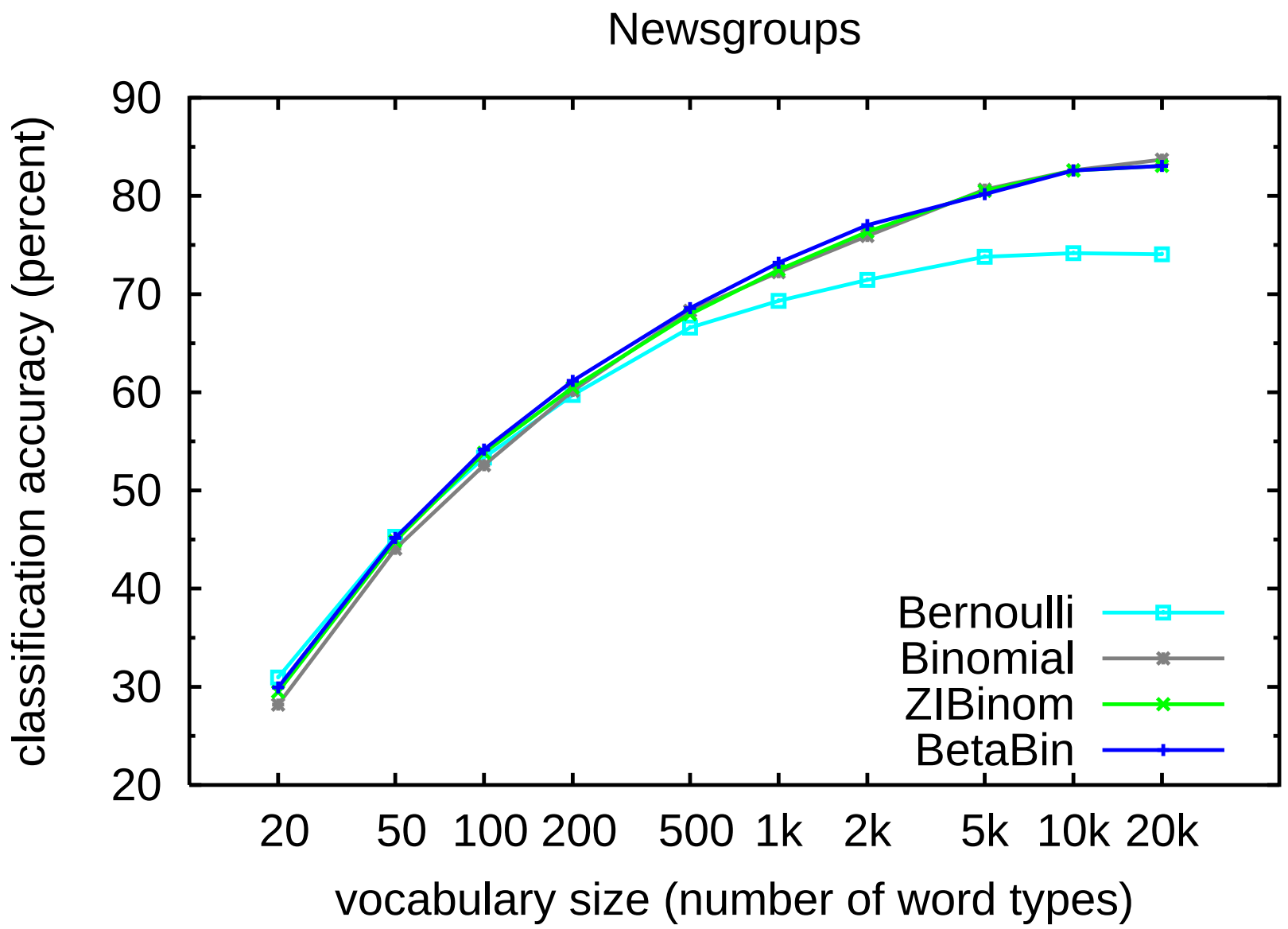
McCallum and Nigam [1998] compared multivariate Bernoulli and multinomial models. We compare (joint independent) Bernoulli, binomial, beta-binomial, and zero-inflated binomial models.

Bernoulli model can be interpreted as binning

(nonparametric historgram method) into two dominant classes: zero and nonzero. Zero-inflated binomial should be able to combine advantages of Bernoulli and detail of binomial model.

"naive"		*"standard"*	
Poisson	1	Negative Binomial	2
Binomial	1	Beta-Binomial	2
Multinomial	k	Dirichlet-Multinomial	k+1

McCallum and Nigam recommended Bernoulli for small vocabulary sizes; we recommend ZIBinomial.


Newsgroups
classification accuracy (percent)
90
80
70
60
50
40
30
20
20
50
100
200
500
1k
2k
5k
10k
20k
vocabulary size (number of word types)
Bernoulli
Binomial
ZIBinom
BetaBin

	Binom	ZIB	McNemar
10	21.61	23.00	**7.99**
20	28.19	29.93	**9.57**
50	44.04	45.15	**6.51**
100	52.57	54.16	**13.12**
200	60.15	61.16	**4.69**
500	68.30	68.58	0.36
1000	72.24	73.20	**5.00**
2000	75.92	77.03	**6.38**
5000	80.64	80.19	1.07
10000	82.61	82.58	0.00
20000	83.70	83.06	2.68

Longer documents

'Tom' in Project Gutenberg books (15k–25k words).

No surprises initially:

0	1	2	3	4	5	6	7	8
313	54	19	11	8	7	5	3	1

But the tail is very long:

71	74	78	102	620
1	1	1	1	1

Document lengths

Document length in newsgroup data is non-negative, heavily skewed to the right, and seems to be unimodal (unlike newswire). Approximated well by log-logistic density:

$$\text{LogLogistic}(\mu, \sigma, \delta)(x) = \frac{\delta \left(\frac{x-\mu}{\sigma}\right)^{\delta-1}}{\sigma \left[1 + \left(\frac{x-\mu}{\sigma}\right)^{\delta}\right]^2}$$

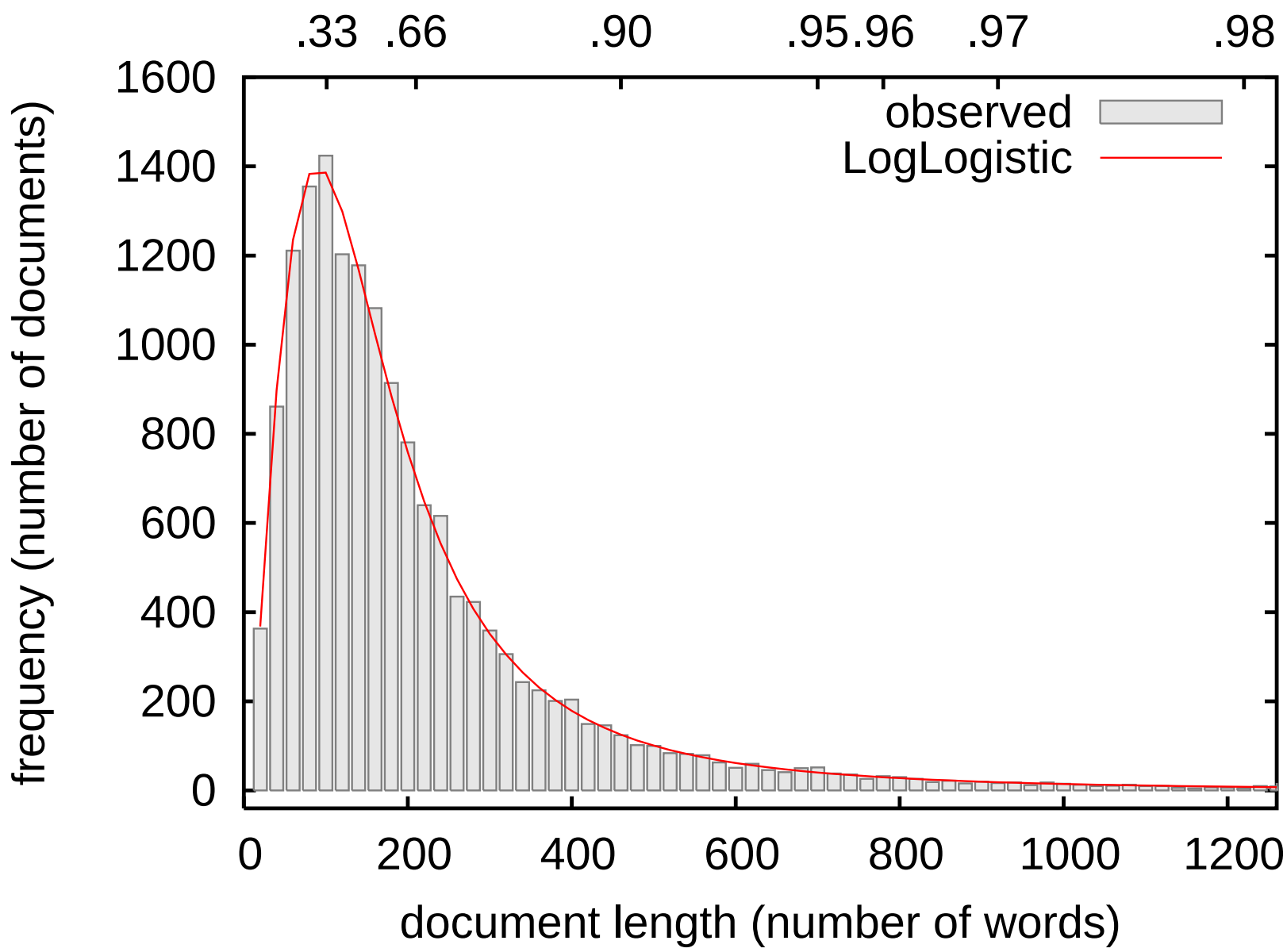
CDF easy to invert (unlike log-normal), pth

percentile point is:

$$\mu + \sigma \left(\frac{p}{1-p} \right)^{1/\delta}$$

Leave μ fixed, estimate remaining two parameters from tertile points:

$$\begin{aligned} \hat{\sigma} &= \sqrt{t_1 - \mu}\,\sqrt{t_2 - \mu} \\ \hat{\delta} &= \frac{2\log 2}{\log(t_2-\mu) - \log(t_1-\mu)} \end{aligned}$$


.33
.66
.90
.95
.96
.97
.98
observed
LogLogistic
frequency (number of documents)
1600
1400
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200
document length (number of words)