Modeling Queries: Different Assumptions

- Multi-Bernoulli: Modeling word presence/absence
  - $q = (x_1, \ldots, x_{|V|})$, $x_i = 1$ for presence of word w_i; $x_i = 0$ for absence

$$p(q=(x_1,\ldots,x_{|V|})\mid d)=\prod_{i=1}^{|V|} p(w_i=x_i\mid d)=\prod_{i=1,x_i=1}^{|V|} p(w_i=1\mid d)\prod_{i=1,x_i=0}^{|V|} p(w_i=0\mid d)$$

  - Parameters: $\{p(w_i=1|d), p(w_i=0|d)\}$ $p(w_i=1|d)+ p(w_i=0|d)=1$
- Multinomial (Unigram LM): Modeling word frequency
  - $q=q_1,\ldots q_m$, where q_j is a query word

$$p(q=q_1\ldots q_m\mid d)=\prod_{j=1}^{m} p(q_j\mid d)=\prod_{i=1}^{|V|} p(w_i\mid d)^{c(w_i,q)}$$

  - $c(w_i,q)$ is the count of word w_i in query q
  - Parameters: $\{p(w_i|d)\}$ $p(w_1|d)+\ldots\ p(w_{|v|}|d) = 1$

[Ponte & Croft 98] **uses Multi-Bernoulli; most other work uses multinomial**
Multinomial seems to work better [Song & Croft 99, McCallum & Nigam 98,Lavrenko 04]