Random Variables, Probability Distributions, & Maximum Likelihood

Topics

- Review Bernoulli and Binomial distribution
- Random Variables and their expected values
- Introduce 3 examples
  - Click-through rates in online advertizing
  - Simple genetics model for a population
  - Classification of spam, fraud, etc.
- Likelihood function

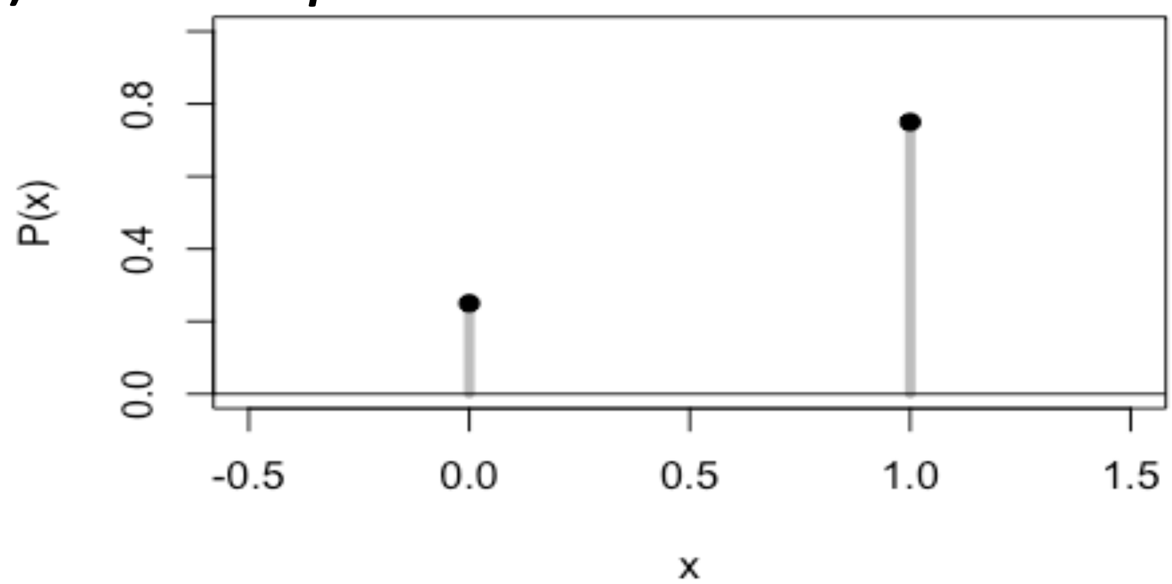
Probability Distributions

Bernoulli(p) Distribution

B = indicator for a success

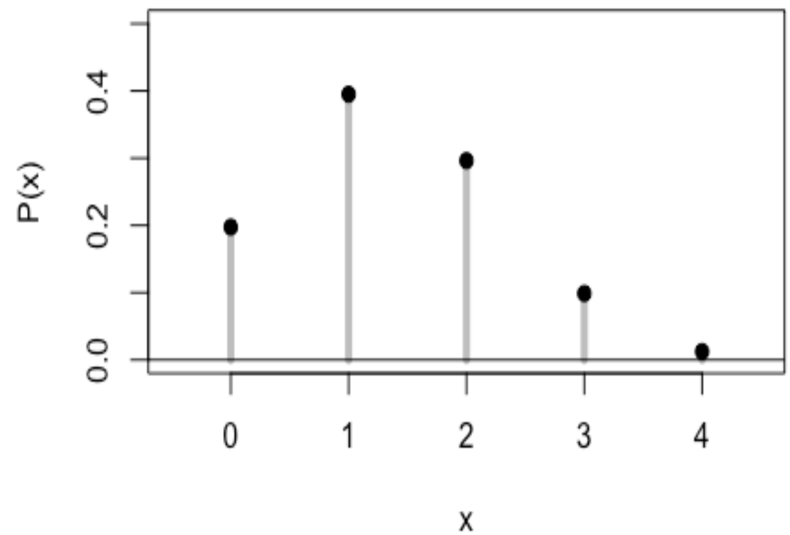
x	0	1
P(x)	1-p	p

Bernoulli(p) where *p* is the chance of success



Binomial(n,p) Distribution

- ***n*** trials
- *p* chance of success on a trial
- trials are independent
- Observe the number of successes



$$\mathbb{P}(k) \;=\; \binom{n}{k} p^k (1-p)^{n-k}$$

$\binom{n}{k}$ is $n!/k!(n-k)!$

Other Distributions

- Geometric(p)
  - Repeat independent trials with chance p of success until the first success
- Poisson(lambda)
  - Count for rare events
- Hypergeometric(n, N, M)
  - Draw n times without replacement from a population of N where M units in the population have a trait. Count those in sample with the trait

Summarize a Distribution

- Recall, we summarize a data distribution with its average (center) and spread (SD)
- We can similarly summarize a probability distribution with its expected value and SD

$$E(X) = \sum_{i=1}^{m} x_i p_i$$

$$Var(X) = \sum_{i=1}^{m} (x_i - E(X))^2 p_i$$

$$SD(X) = \sqrt{Var(X)}$$

Bernoulli(p)

$$E(B) = 0(1 - p) + 1p = p$$

$$Var(B) = (0 - p)^2 (1 - p) + (1 - p)^2 p$$

$$Var(B) = p(1 - p)$$

Gambling Problem

- Recall the pot has \$64 in it
- Sam won 2 rounds and Andrew won 1
- W = Sam's winnings
- What is W's distribution?
- What is E(W)?

Sam's Winnings

W= winnings

w	0	64
P(w)	1/4	3/4

E(W) = 0*1/4 + 64*3/4 = 48

Does this distribution look familiar?

W = 64B

Properties of Expected Value

$$E(aX+b)=aE(X)+b$$

$$E(aX+b)=\sum_{i=1}^{m}(ax_i+b)p_i$$

$$E(X+Y)=E(X)+E(Y)$$

Properties of Variance

$$Var(aX+b)=a^2Var(X)$$

$$Var(aX+b)=\sum_{i=1}^{m}(ax_i+b-(aE(x)+b))^2\,p_i$$

$$Var(X+Y)=Var(X)+Var(Y),\ if\ independent$$

Click-Through Rates in Online Advertizing

Example from Xueri Wang et al

- An online experiment
- Visitors to a page are randomly selected to see a version of the page with a particular ad
- We are interested in the how successful the ad is in getting visitors to “click-through” to the advertisers page

Probability Model?

- What is a reasonable model for this process?
- What assumptions are you making?

Probability Model

- Visitors act independently
- Visitors have the same chance of clicking through to the site
- Any others?

Can you provide a Probability Distribution that captures this process?

Results

Model:

X = Number of click-throughs in 1000 views

X ~ Binomial(1000, p)

In 1000 views, 25 click-throughs occurred

What is your estimate for p?

Why?

X ~Binomial(n,p) distribution

- $X = B_1+B_2+\ldots+B_n$ where $B_i \sim$ Bernoulli(p)

- $E(B_i) = p$
- $E(X) = E(B_1+\ldots+B_n) = np$
- Observe X (sum of 1000 Bernoulli) to be 25
- Avg of 1000 Bernoulli should be close to E(X)
- p_hat = 25/1000
- This approach is called the Method of Moments (an average is a moment)

An Alternative Approach

- Consider the chance of 25 successes if p=0.01

$P(X = 25|p=0.01) = C\ 0.01^{25} 0.99^{1000-25}$

- Consider the chance of 25 successes if p=0.02

$P(X = 25|p=0.02) = C\ 0.02^{25} 0.98^{1000-25}$

- Consider the chance of 25 successes if p=0.025

$P(X = 25|p=0.025) = C\ 0.025^{25} 0.975^{1000-25}$

- Consider the chance of 25 successes if p=0.05

$P(X = 25|p=0.05) = C\ 0.05^{25} 0.95^{1000-25}$

Likelihood

- These quantities, e.g., C $0.05^{25}0.95^{1000-25}$, can be viewed as a function of *p* given the data

$L(p) = C\, p^{25}(1-p)^{1000-25}$

Find the *p* that maximizes this quantity and use it to estimate *p*. It has the highest likelihood for the data.

We call $L(p)$ the likelihood

Likelihood

$L(p) = C\, p^{25}(1-p)^{1000-25}$

It is often easier to maximize the log of the likelihood function:

$\log(L(p)) = C + 25\log(p) + (1000-25)\log(1-p)$

We can differentiate the log-likelihood and set to 0 to solve for p

Maximize the Log-Likelihood

$\log(L(p)) = C + 25\log(p) + (1000-25)\log(1-p)$

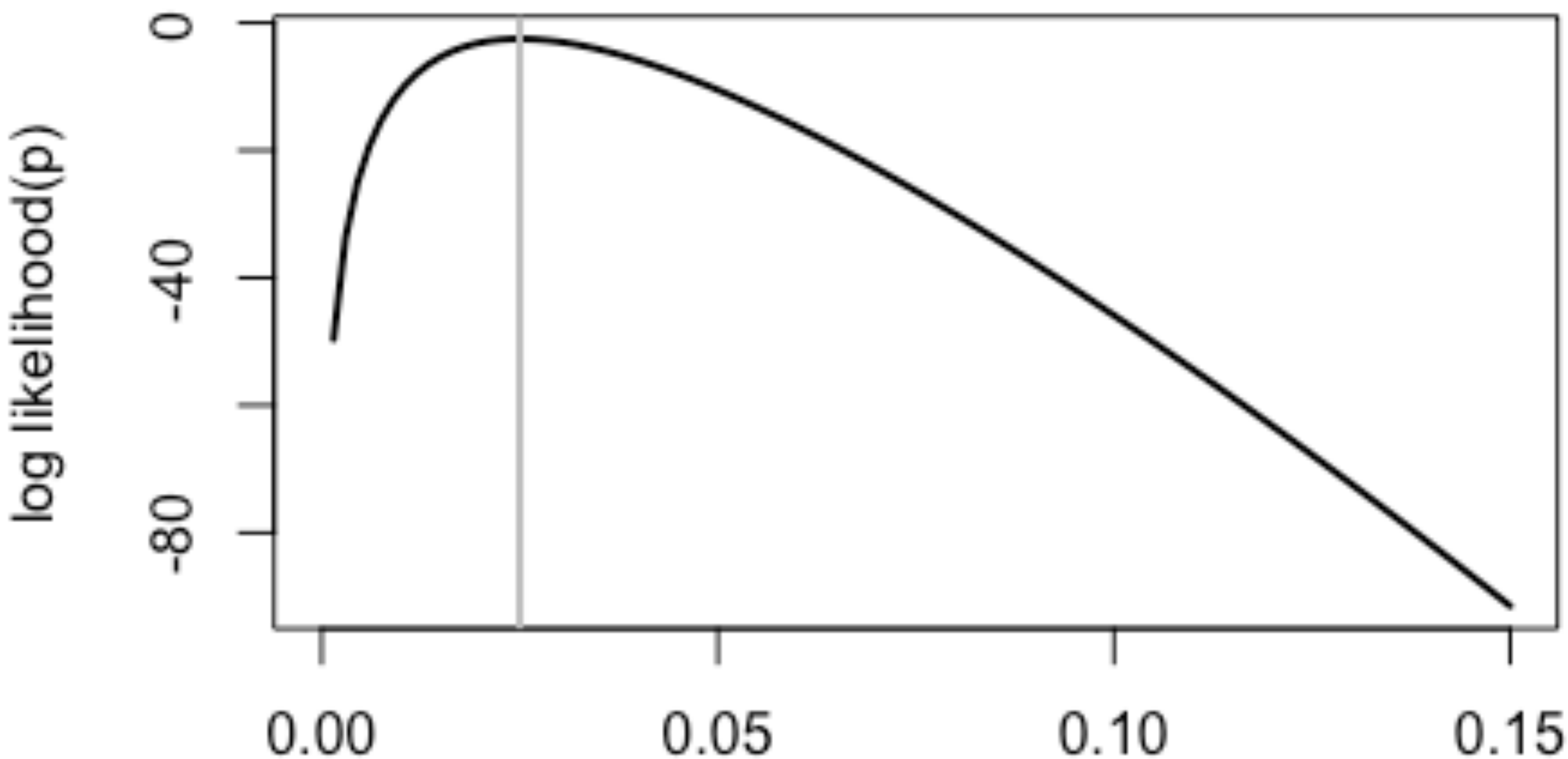
Differentiate wrt *p*:

$25/p - (1000-25)/(1-p)$

Set to 0 and solve for *p*: $0 = (1-p)25 - p(1000-25)$

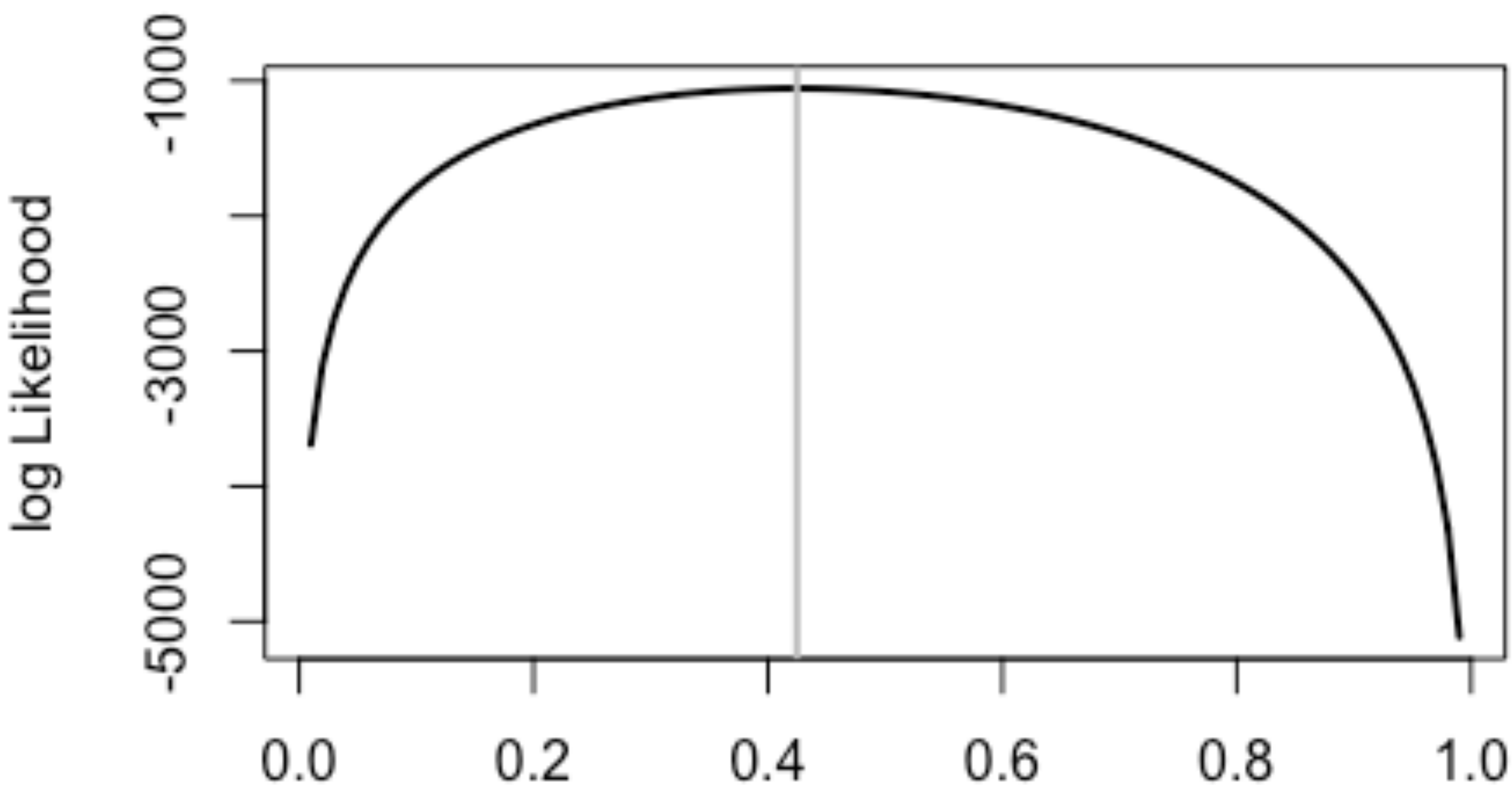
$\hat{p} = 25/1000$

Log – Likelihood function



Simple Genetics Example

On the Board


log Likelihood
-1000
-3000
-5000
0.0
0.2
0.4
0.6
0.8
1.0

Spam Detection

Spam

- Spam appears in our email, comments on blogs, reviews on Yelp, etc.
- We can develop detectors to help us programmatically identify spam
- In the case of email, Spam Assassin provided 9000 email messages that are hand-classified as spam or ham

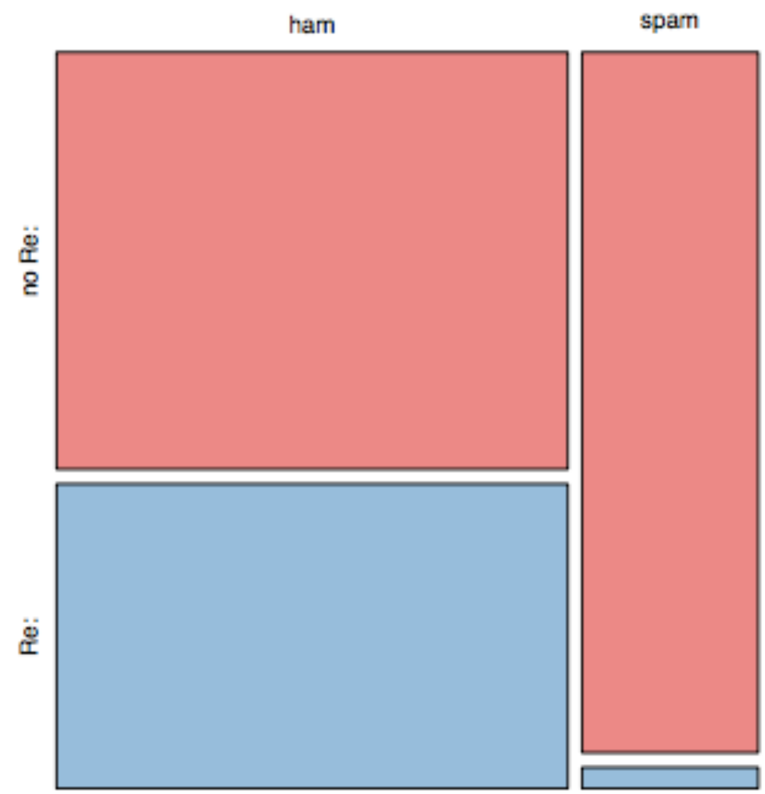
Email Corpus

- Later in the semester we will discuss how to build classifiers
- We look at a simple example today
- From the 9000 email messages we determined
  - whether or not the subject line starts Re:
  - the percentage of capital letters in the email

Re: in the subject line of the email

	ham	spam	
Re:	2400	300	2700
No Re:	3600	2700	6300
	6000	3000	9000

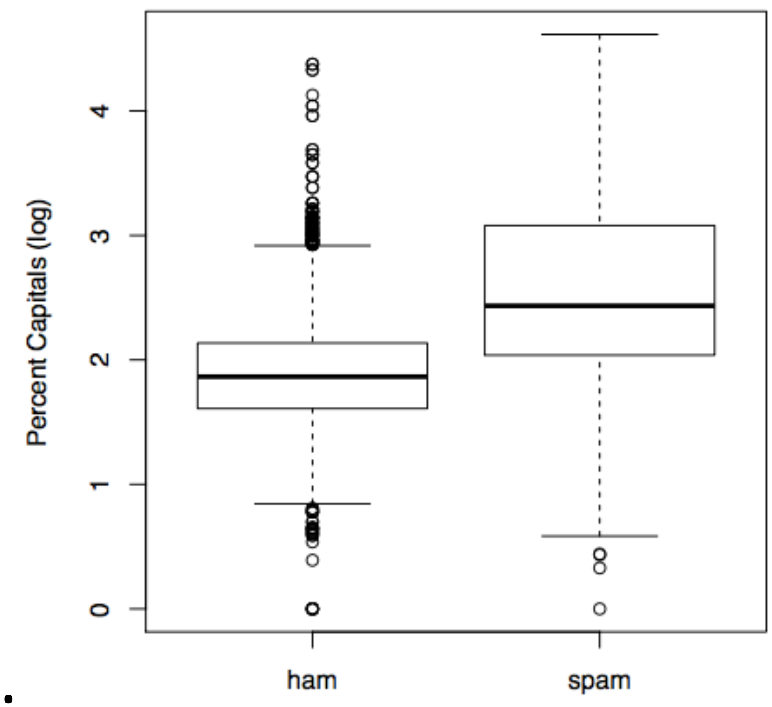
A new email arrives, is has an Re: in its subject line.
What is the chance it is spam?



Is the presence of Re: a useful indicator of ham?

Capitalization in the email

	ham	spam	
<20	4000	750	4750
20-30	1500	1500	3000
>30	500	750	1250
	6000	3000	9000



A new email arrives, is has more than 30% capital letters. What is the chance it is spam?

Is the percent capitals a useful indicator of ham?

Re: in the subject line of the email

What are we assuming to answer this question?

	ham	spam	
Re:	2400	300	2700
No Re:	3600	2700	6300
	6000	3000	9000

New email has a similar distribution of Re: within spam and ham as the corpus

We have enough data to accurately estimate this probability

Prop(spam|Re:) =
Prop(spam and Re:)/prop(Re:)
= 750/1250 = 0.6

- A new email arrives.
- It has an Re: in the subject line and fewer than 20% of the letters are capitalized
- What is the chance it is spam?
- Can we answer this question?

P(spam|Re: and <20% caps)

		ham	spam	
Re:	<20			
	20-30			
	>30			
No Re:	<20			
	20-30			
	>30			
		6000	3000	9000

$P(spam|Re; \text{ and } <20\%)=$
$P(spam \,\&\, Re: \,\&\, <20\%)/$
$P(Re: \,\&\, <20\%)$
$= 50/2050$

P(spam|Re: and <20% caps)

		ham	spam	
Re:	<20	2000	50	2050
	20-30	300	100	400
	>30	100	150	250
No Re:	<20	2000	700	2700
	20-30	1200	1400	2600
	>30	400	600	1000
		6000	3000	9000

In practice

- We have many features $X_1, X_2, \ldots, X_m$
- We observe $x_1, x_2, \ldots, x_m$
- We want $P(\text{spam} | x_1, x_2, \ldots, x_m)$
- Building a probability model is quite complex
- We don't have enough data to estimate the joint distribution of m random variables

$P(\text{spam} | x_1, x_2, \ldots, x_m)$

$= P(\text{spam and } x_1, x_2, \ldots, x_m) / P(x_1, x_2, \ldots, x_m)$

Why? Definition of conditional probability

$= P(\text{spam})P(x_1, x_2, \ldots, x_m | \text{spam}) / P(x_1, x_2, \ldots, x_m)$

Why?

Bayes Rule $P(A|B) = P(A)P(B|A)/ P(B)$

Naively assume independence

$P(spam|x_1, x_2, \ldots, x_m)$

$= P(spam)P(x_1, x_2, \ldots, x_m|spam) / P(x_1, x_2, \ldots, x_m)$

$= P(spam)P(x_1|spam) * \ldots * P(x_m|spam) / P(x_1, x_2, \ldots, x_m)$

Naïve Bayes Estimation of $P(spam|x_1, x_2, \ldots, x_m)$

Computational Considerations

Take log to turn product of small probabilities into sums

$\text{Log}(P(\text{spam})) = \log(P(\text{spam})) + \Sigma\log(P(x_i|\text{spam}))$
$- \log(P(x_1, x_2, \ldots, x_m))$

Examine the likelihood ratio,

$\text{Log}(P(\text{spam})/P(\text{ham}))$

We don't need to compute $P(x_1, x_2, \ldots, x_m)$

Take Aways

- Named probability distributions are defined in terms of parameters
- Given the data, we maximize the likelihood of the data over the possible parameter values
- In practice,
  - We might not be able to analytically solve for the parameters
  - We might not have the complete data
  - Computational considerations can be important for accuracy and efficiency