Maximum Likelihood
& Bayes Rule

Topics

- Review Maximum Likelihood Concept
- Consider Prior Beliefs
- Continuous valued Random Variables
- Posterior probability and Maximum Likelihood

Click-Through Rates in Online Advertizing

Example from Xueri Wang et al

Results

Model:
X = Number of click-throughs in 200 views
X ~ Binomial(200, p)

In 200 views, 25 click-throughs occurred

Let's estimate p using the likelihood approach

Maximum Likelihood

Consider the chance of 25 successes if p=0.01

$P(X = 25|p=0.01) = C(200,25)\ 0.01^{25} 0.99^{200-25}$
$= 7.7e-20$

Let's consider other possible values for *p*,

Maximum Likelihood

If p=0.05, then P(X = 25|p=0.05) = 1.7e-05
If p=0.10, then P(X = 25|p=0.10) = 0.04
if p=0.15, then P(X = 25|p=0.10) = 0.05

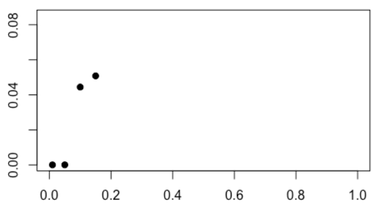
Let's place these values on a plot

Maximum Likelihood

If p=0.05, then P(X = 25|p=0.05) = 1.7e-05

If p=0.10, then P(X = 25|p=0.10) = 0.04

if p=0.15, then P(X = 25|p=0.10) = 0.05



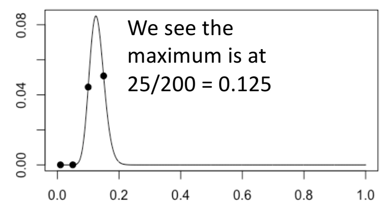
Let's Add the likelihood for all possible values of p

Maximum Likelihood

If p=0.05, then P(X = 25|p=0.05) = 1.7e-05

If p=0.10, then P(X = 25|p=0.10) = 0.04

if p=0.15, then P(X = 25|p=0.10) = 0.05



Likelihood

- These likelihoods can be viewed as a function of *p* given the data

$$L(p) = C(200,25)\, p^{25}(1-p)^{200-25}$$

Find the *p* that maximizes the likelihood for our data and use it to estimate *p*.

Likelihood

$L(p) = C(200,25)\, p^{25}(1-p)^{200-25}$

here C(200,25) is the factorial

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

log(L(p)) = C(200,25) + 25log(p) + (200-25)log(1-p)

We can differentiate the log-likelihood and set to 0 to solve for p to get 0.125 as our estimate

Practice Problem

Geometric(p)

- X = number of failures until first success
- Trials are independent with the same probability of success
- P(k) = P(k failures followed by a success)
  $= p(1-p)^k$ for k = 0, 1, 2, …

Suppose you observe the geometric n times, and record $k_1, \ldots, k_n$ Which value of p maximizes the likelihood of our data?

Find the Likelihood

$P(X_1=k_1, X_2=k_2, \ldots, X_n = k_n)$
$= P(X_1=k_1)P(X_2=k_2) \cdots P(X_n = k_n)$ independence
$= p(1-p)^{k1} \times p(1-p)^{k2} \times \ldots \times p(1-p)^{kn}$ geometric(p)

$= p^n(1-p)^{k1+k2+\ldots+kn}$ The likelihood function

Let's maximize the log likelihood

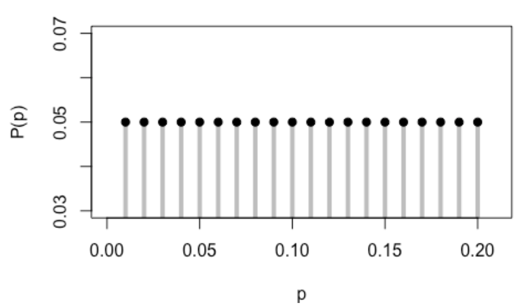
$\text{Log}(L(p)) = n\log(p) + (k_1+\ldots+k_n)\log(1-p)$

Differentiate wrt p
$n/p - (k_1+\ldots+k_n)/(1-p)$
Set to 0 and solve for p
$0 = n/p - (k_1+\ldots+k_n)/(1-p)$
$0 = (1-p)n - p(k_1+\ldots+k_n)$
p_hat = $n/(n+ k_1+\ldots+k_n) = 1/(1 + \text{avg})$
number of successes / number of trials

Return to our click-through example

- Suppose that you were certain that p was either 0.01, 0.02, …, 0.20
- You figure it might be any one of these values
- And you think of them as equally likely values for p
- This is similar to the vitamin problem where you reach into a barrel of coins that have different chances of success and pick one to flip

Here's a pmf for the possible values of p



Given the data: 25 click-throughs on 200 visits, how would you update your pmf for p?

For example, what is
$P(p = 0.01 \mid \text{data} = 25)$?

$P(p = 0.01 \mid data = 25)$

Let's apply the conditional probability rule

$= P(p = 0.01 \text{ and } data = 25)\ /\ P(data)$

Let's apply the rule again to the numerator

$= P(p = 0.01)\ P(data = 25 \mid p = 0.01)\ /\ P(data)$

$= 0.05\ C(200,25)\ 0.01^{25}\ 0.99^{200-25}\ /P(data)$

Bayes Rule

$P(A|B) = P(A \text{ and } B)/\ P(B)$

$P(B|A) = P(A \text{ and } B)/P(A)$

So $P(A \text{ and } B) = P(A)P(B|A)$

Plug this into the numerator above

Bayes RULE: $P(A|B) = P(A)P(B|A)/\ P(B)$

Bayes Rule

- We just used Bayes rule

  $P(A|B) = P(A)P(B|A)/\ P(B)$
- $A = \{p = 0.1\}$
- $B = \{data = 25\}$

$P(p = 0.01 \mid data = 25) =$

$P(p = 0.01)\ P(data = 25|p = 0.01)\ /\ P(data = 25)$

The denominator is the average over all possible values for p.

P(p| data = 25)

We can compute $P(p = x \mid data = 25)$ for all possible values of p.

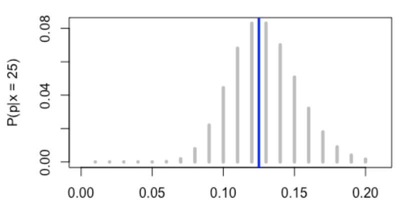
In general, for x = 0.01, 0.02, …, 0.20,

$P(p = x \mid data = 25) =$

$= P(p = x)\ P(data = 25 \mid p = x)\ /\ P(data)$

$= 0.05\ C(200,25)\ x^{25}\ (1-x)^{200-25}\ /P(data)$

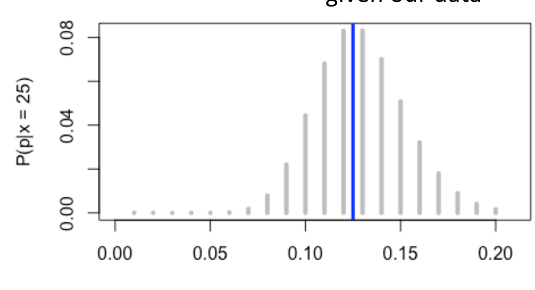
$P(p = x|\ data = 25) = C\ x^{25}\ (1-x)^{200-25}$



The sample avg of 0.125 is not one of the possible values.

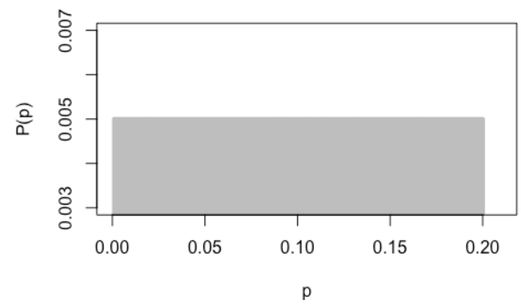
We have updated our probabilities for p given the data.

Posterior probability of p given our data

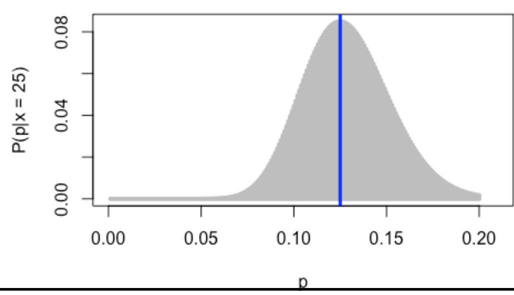


What happens if we change our initial (prior) probabilities on p to be discrete uniform on 0.001, 0.002, …, 0.200?

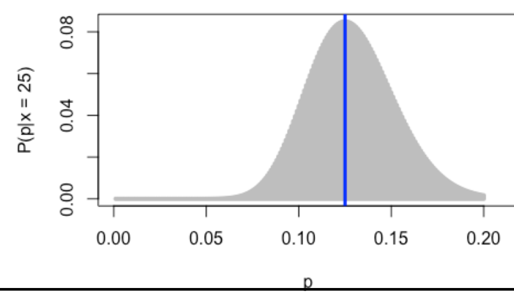
What probability mass function is so dense that we can't distinguish the individual values for p.



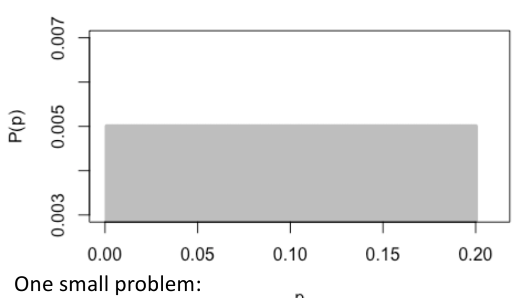
This posterior distribution for p places the highest probability on 0.125, but the small values between 0 and 0.05 still have a chance (a small chance) of being the true p.



We can use this posterior distribution to estimate p (choose the one with the highest chance).
We can provide credible regions for p, e.g., interval with a 95% posterior probability.



This pmf is so dense that we wonder, can we just say that p can be equally likely to be any value in (0, 1)?



One small problem:
There are uncountably many possible values in (0,1)

Continuous Random Variables

Continuous Uniform Distribution

- Uniform(0, 1) distribution
- We say the random variable X has a Uniform(0, 1) distribution, if

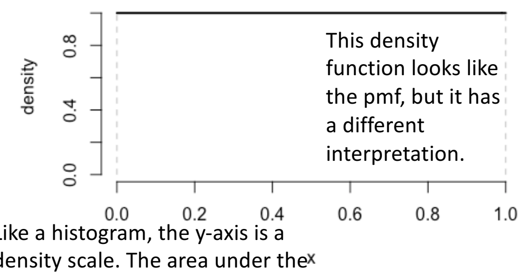
$P(a < X < b) = (b-a)$ for any a & b in (0,1)

For example,

$P(1/4 < X < 1/2) = P(1/8 < X < 3/8)$
$= P(3/4 < X < 1)$
$= 1/4$

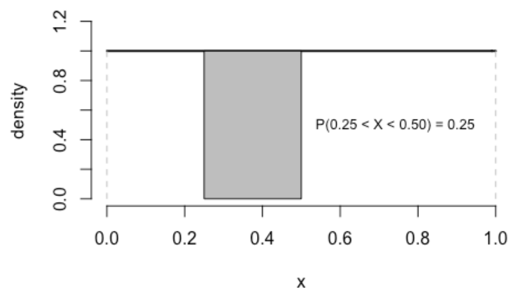
We compute probabilities on intervals, rather than exact values

Probability Density Function



This density function looks like the pmf, but it has a different interpretation.

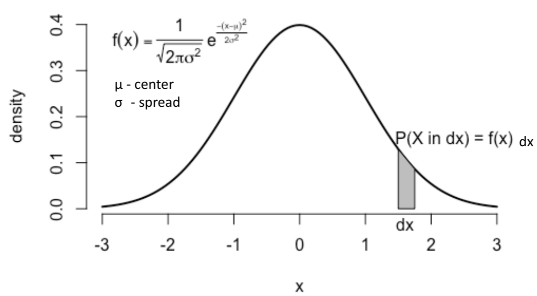
Like a histogram, the y-axis is a density scale. The area under the curve corresponds to chance.



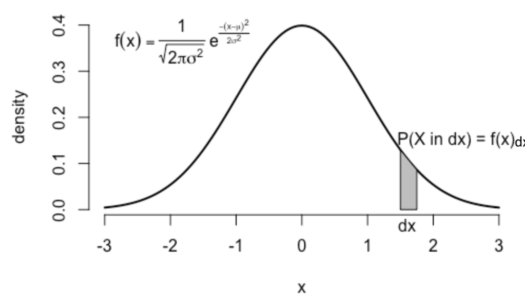
Another Continuous Distribution

The Normal(μ, σ^2)

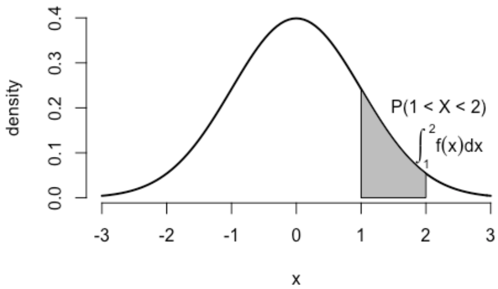
Normal Distribution



For a small region of width dx, P(X in dx) = f(x)dx



We find probabilities by integration

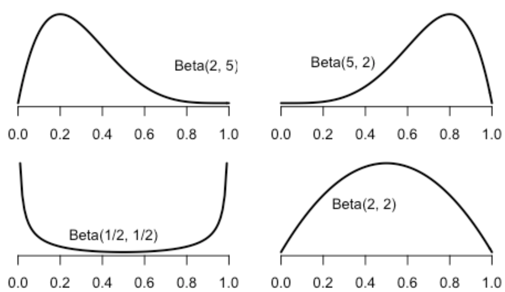


This integral doesn't have a closed form solution so we need the computer to approximate the area.

Another Continuous Distribution

The Beta(α,β)

The Family of Beta Distributions



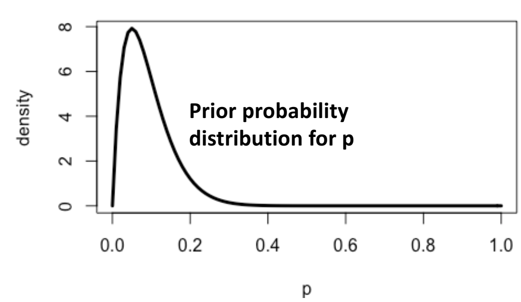
Beta(α,β)

- The Beta distribution is for random variables in (0, 1)
- The Beta(1,1) is the Uniform(0,1) distribution
- The Beta includes symmetric, skewed, U-shaped distributions
- The probability density function is
  $f(p) = B(\alpha,\beta)\, p^{\alpha-1}(1-p)^{\beta-1}$ for p in (0, 1)

Beta(α,β)

- Let's return to our click-through example
- Let's consider another prior distribution for p, such as the Beta(2, 10)
- The Beta(2, 10) probability density function is
  $f(p) = B(\alpha,\beta)\, p^{2-1}(1-p)^{10-1}$ for p in (0, 1)

Beta(2, 10)



Posterior for p given the data

Recall that we used Bayes rule

to compute the

P(p = 0.01 | data = 25) =

P(p = 0.01) P(data = 25|p = 0.01) / P(data = 25)

We use it again with the Beta density giving us the probability P(p in dp) = f(p) dp

Posterior for p given the data

If we have a continuous distribution for the possible values of p, then

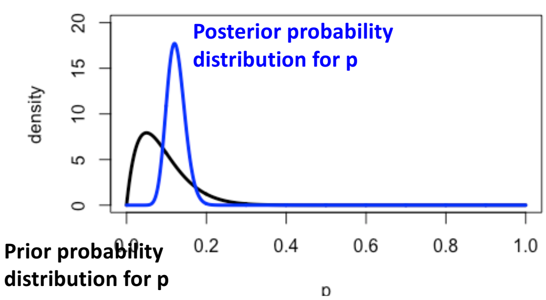
P(p in dp| data = 25)

= P(p in dp) P(data = 25|p in dp) / P(data = 25)

= $B(\alpha,\beta)\ p(1-p)^{9}\ C(200,25)\ p^{25}(1-p)^{175}$

= $B(\alpha,\beta)\ C(200,25)\ p^{25+1}(1-p)^{175+9}$

This is the Beta(26, 184) distribution for p



Using the prior to provide a single point estimate for p

- Our posterior distribution for p is B(26,184)
- If we want to provide an estimate for p, then we could provide the p with the highest posterior probability, i.e., we would maximize

  $B(\alpha,\beta)\ C(200,25)\ p^{25+1}(1-p)^{175+9}$
- We know this is p_hat = (25+1)/(200+10)

Laplace Smoothing

- p_hat = (25+1)/(200+10)
- Notice that we have taken the MLE of 25/200 and added a small bit to the numerator and a small bit to the denominator
- This is especially useful for rare events, i.e., when p is small
- We can re-express p_hat as a weighted average of the MLE and the Ev of the prior:

$$\frac{200}{210}\frac{25}{200}+\frac{10}{210}\frac{1}{10}$$

Step Back

- Real world phenomena can often be modeled with a probability distribution
- Many probability distributions can be expressed as probability mass (or density) functions that depend on parameter(s)
- We can use Maximum Likelihood to estimate these parameters, i.e., to find the parameters that are most likely to have produced our data

- We may have prior information about the parameters
- This prior information can be expressed as a probability distribution on the parameter values
- We can use Bayes rule to find the posterior distribution of our parameters given our data
- Sometimes the prior fits nicely with the data distribution, e.g., the Beta and Binomial
- Other times we use computational methods to compute the posterior

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

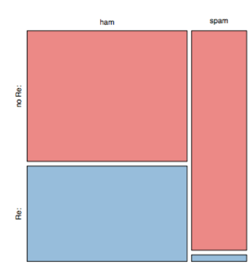
Email Corpus

- 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, it has an Re: in its subject line.
What is the chance it is spam?



Is the presence of Re: 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

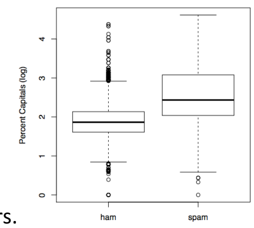
New email has a similar distribution of Re: within spam and ham as the corpus

We have enough data to accurately estimate this probability

Prop(spam|Re:) =
Prop(spam and Re:)/prop(Re:)
= 300/2700 = 0.11

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

Is the percent capitals a useful indicator of ham?

> 30% capital letters in the email

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

Prop(spam| > 30%) =
Prop(spam and >30%)/prop(>30%)
= 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; 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 $P(spam|x_1, x_2, \ldots, x_m)$
- Building a probability model is quite complex
- We don't have enough data to estimate the joint distribution of m random variables

$P(spam|x_1, x_2, \ldots, x_m)$
$= P(spam \text{ and } x_1, x_2, \ldots, x_m) / P(x_1, x_2, \ldots, x_m)$
Why? Definition of conditional probability

$= P(spam)P(x_1, x_2, \ldots, x_m|spam) / P(x_1, x_2, \ldots, x_m)$
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
$\text{Log}(P(spam)) = \log(P(spam)) + \Sigma\log(P(x_i|spam)) - \log(P(x_1, x_2, \ldots, x_m))$
approx $\log(3/9) + \Sigma\log(\#x_i \text{ in spam})/\#spam) – C$
approx $\log(3/9) + \Sigma\log(\#x_i \text{ in spam } + 1)/(\#spam + 1)) – C$

Examine the likelihood ratio,
$\text{Log}(P(spam)/P(ham))$
We don't need to compute $P(x_1, x_2, \ldots, x_m)$
Values above 0 indicate $P(spam) > P(ham)$

Take Aways

- In practice,
  - We might not have a named probability distribution so we resort to estimating probabilities with proportions
  - We might not have enough data so we smooth our proportions and make naïve assumptions
  - Computational considerations can be important for accuracy and efficiency