Chapter 4: Parametric Methods

Tyler J. Brough

Data Analytics & Information Systems

Introduction to Sources

This notebook is based on the following sources:

- *Chapter 4: Parametric Methods* of *Introduction to Machine Learning* by Alpaydin
- *Chapter 9: Point Estimation* of *Introduction to Probability and Mathematical Statistics*
- *Chapter 2: General Matters* of *Machine Learning: An Applied Mathematics Introduction*

See also:

- *Section 2.2.3 The Classification Setting* of *Chapter 2: Statistical Learning* of *An Introduction to Statistical Learning*
- *Section 4.4 Generative Models for Classification* of *Chapter 4: Classification* of *An Introduction to Statistical Learning*
- *Chapter 9: Statistical Pattern Recognition* of *Computational Statistics Handbook with MATLAB*

Defintion of an Estimator

Estimator $T = t(X_1, X_2, \ldots, X_n)$, that is used to estimate the value of $\tau(\theta)$ is called an **estimator** of $\tau(\theta)$, and an observed value of the statistic, $t = t(x_1, x_2, \ldots, x_n)$, is called an **estimate** of $\tau(\theta)$.

Likelihood Function

Likelihood Function The joint function of n random variables $X_1, \ldots, X_n$ evaluated at $x_1, \ldots, x_n$, say $f(x_1, \ldots, x_n)$ is referred to as a **likelihood function.** For fixed $x_1, \ldots, x_n$ the likelihood function is a function of θ and often is denoted by $L(\theta)$.

If $X_1, \ldots, X_n$ represents a random sample from $f(x; \theta)$, then

$$L(\theta) = f(x_1; \theta) \cdots f(x_n; \theta)$$

Maximum Likelihood Estimator (MLE)

Maximum Likelihood Estimator Let $L(\theta) = f(x_1, \ldots, x_n; \theta), \theta \in \Omega$, be the joint pdf of $X_1, \ldots, X_n$. For a given set of observations, $(x_1, \ldots, x_n)$, a value $\hat{\theta}$ in Ω at which $L(\theta)$ is a maximum is called a **maximum likelihood estimator** (MLE) of θ. That is, $\hat{\theta}$ is a value of θ that satisfies

$$f(x_1, \ldots, x_n; \hat{\theta}) = \max_{\theta \in \Omega} f(x_1, \ldots, x_n; \theta)$$

MLE Continued

- Notice that if each set of observations $(x_1, \ldots, x_n)$ corresponds to a unique value of $\hat{\theta}$, then this procedure defines a function, $\hat{\theta} = t(x_1, \ldots, x_n)$.
- This same function, when applied to random sample, $\hat{\theta} = t(X_1, \ldots, X_n)$, is called the **maximum likelihood estimator**, also denoted MLE.
- Usually, the same notation, $\hat{\theta}$, is used for both the ML estimate and the ML estimator.

MLE Continued

In most cases, $L(\theta)$ represents the joint pdf of a random sample, although the maximum likelihood principle also applies to other cases such as sets of order statistics.

If Ω is an open interval, and if $L(\theta)$ is differentiable and assumes a maximum on Ω, then the MLE will be a solution of the equation (maximum likelihood equation)

$$\frac{d}{d\theta} L(\theta) = 0$$

MLE Continued

- If one or more solutions to the above equation exist, then it should be verified which, if any, maximize $L(\theta)$
- Note that any value of θ that maximizes $L(\theta)$ also will maximize the log-likelihood, $\ln[L(\theta)]$
- So for computational convenience the alternate form of the maximum likelihood equation will often be used

$$\frac{d}{d\theta}\ln[L(\theta)] = 0$$

Example: Coin Tossing

Suppose you toss a coin n times and get h heads. What is the probability, p, of tossing a head next time?

The probability of getting h heads from n tosses is, assuming that the tosses are independent,

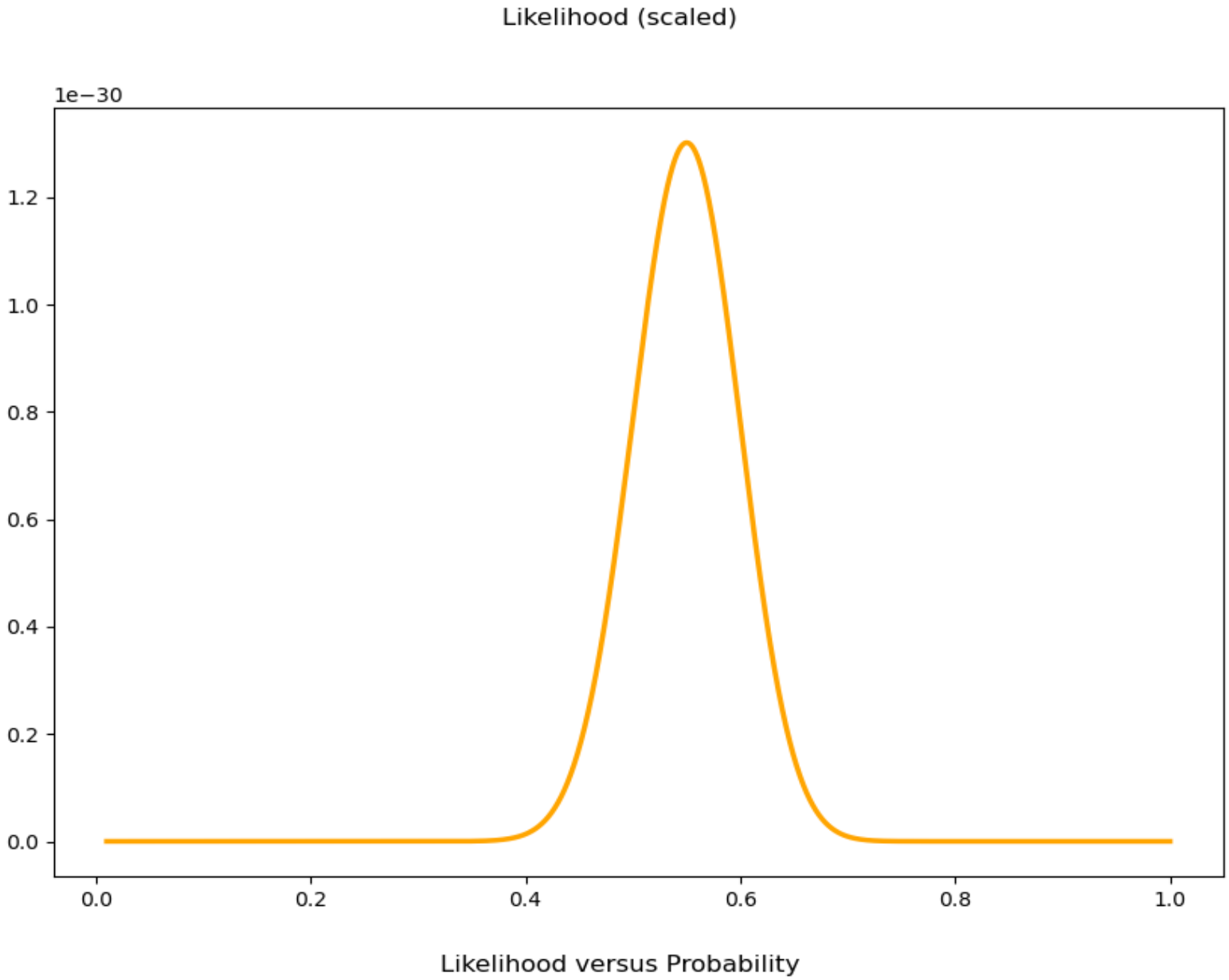
$$\frac{n!}{h!(n-h)!}p^h(1-p)^{n-h} = \binom{n}{h}p^h(1-p)^{n-h}$$

MLE Coin Tossing Example Continued

- Applying MLE is the same as maximizing this expression with respect to p.
- This likelihood function (without the coefficient in the front that is independent of p) is shown below for $n = 100$ and $h = 55$.
- There is a very obvious maximum:

```
## The MLE is:  0.55
```

MLE Coin Tossing Example Likelihood Function

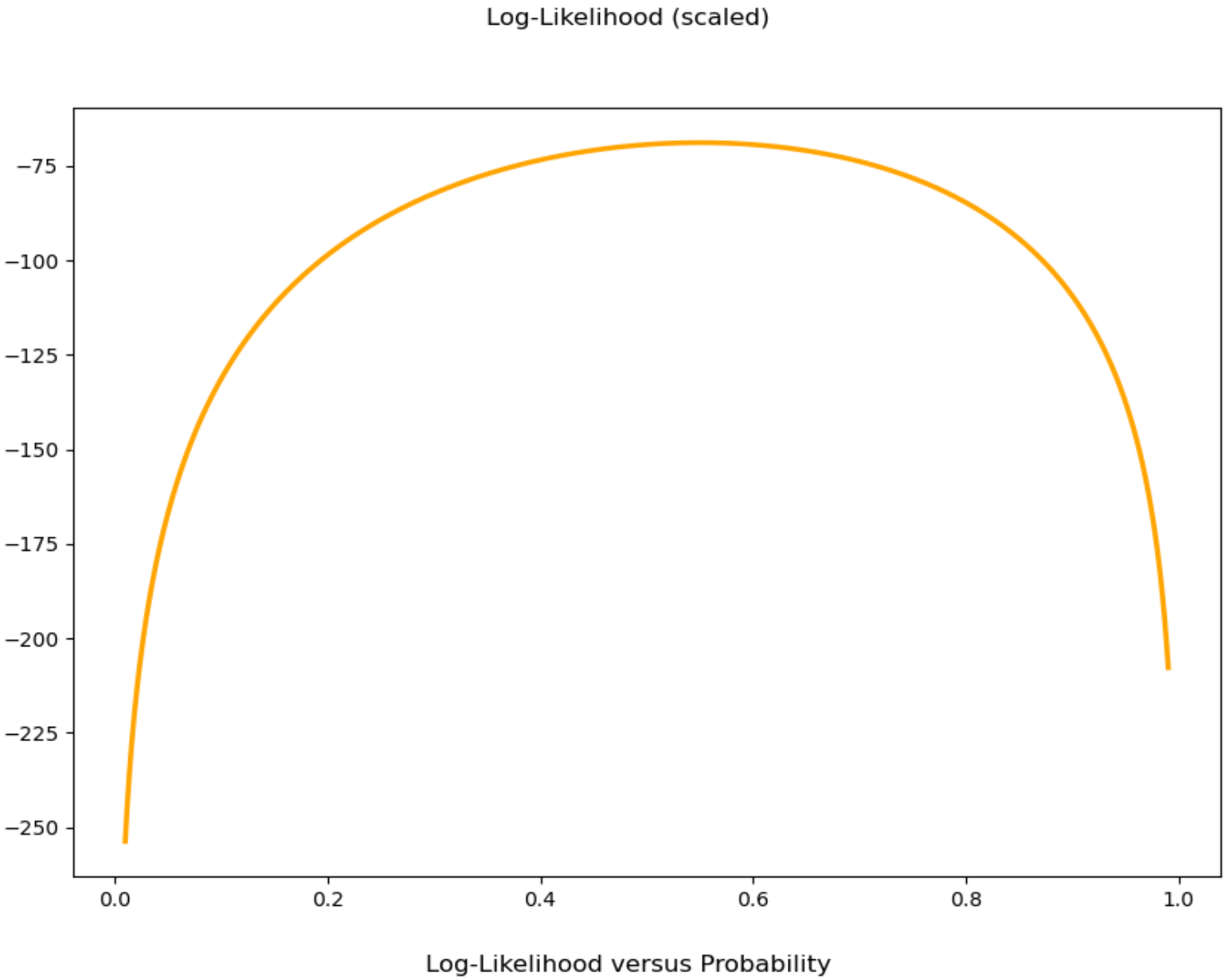


Log-Likelihood

- Often with MLE when multiplying probabilities, as here, you will take the logarithm of the likelihood and maximize that.
- This doesn't change the maximizing value but it does stop you from having to multiply many small numbers, which is going to be problematic with finite precision. (Look at the scale of the numbers on the vertical axis in the figure.)
- Since the first part of this expression is independent of p we maximize with respect to p

$$h \log p + (n - h) \log (1 - p)$$

Log-Likelihood Continued



The Solution

$$\frac{d}{d\theta} \ln [L(\theta)] = 0$$

$$= \frac{h}{p} - \frac{(n-h)}{(1-p)} = 0$$

The solution is

$$p = \frac{h}{n}$$

- We see that this is the sample mean $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ for the binary random variable x.
- For the case above ($n = 100$ and $h = 55$) we see that $\hat{p} = \frac{55}{100} = 0.55$.

A Poisson Random Variable Example

Consider a random sample from a Poisson distribution, $X \sim POI(\theta)$. The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \frac{\exp(-n\theta)\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

and the log-likelihood is

$$\ln\left[L(\theta)\right] = -n\theta + \sum_{i=1}^{n} x_i \ln(\theta) - \ln\left(\prod_{i=1}^{n} x_i!\right)$$

The Maximum Likelihood Equation

$$\frac{d}{d\theta} \ln [L(\theta)] = -n + \sum_{i=1}^{n} \frac{x_i}{\theta} = 0$$

which has the solution $\hat{\theta} = \sum_{i=1}^{n} \frac{x_i}{n} = \bar{x}$. It is possible to verify that this is a maximum by use of the second derivative,

$$\frac{d^2}{d\theta^2} \ln [L(\theta)] = -\sum_{i=1}^{n} \frac{x_i}{\theta^2}$$

which is negative when evaluated at $\bar{x}$, $-n/\bar{x} < 0$.

A Normal Random Variable Example

- Say we have draws from a normal distribution with unknown mean and standard deviation
- That's two parameters
- The probability of drawing a number x is

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- where μ is the mean and σ the standard deviation which are both to be estimated.

The Log-Likelihood Function for a Normal R.V.

The log-likelihood is then

$$\ln\left(p(x)\right) = -\frac{1}{2}\ln\left(2\pi\right) - \ln\left(\sigma\right) - \frac{1}{2\sigma^2}(x-\mu)^2$$

If the draws are independent then after N draws, x_n, the likelihood will be

$$p(x_1)p(x_2)\cdots p(x_N) = \prod_{n=1}^{N} p(x_n)$$

The Log-Likelihood Function for a Normal R.V. Continued

And so the log-likelihood function is

$$ln\left(\prod_{n=1}^{N} p(x_n)\right) = \sum_{n=1}^{N} \ln\left(p(x_n)\right)$$

This gives us a convenient form for the log-likelihood

$$-N \ln(\sigma) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2$$

Any terms that do not contain the parameter of interest can be dropped.

The MLE Solution for a Normal R.V.

- To find the MLE for μ you just differentiate with respect to μ and set to zero.

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- And differentiating with respect to σ gives

$$\hat{\sigma} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2$$

- These results make eminent sense

Bias and Mean Squared Error

If T is an estimator of $\tau(\theta)$, the the **bias** is given by

$$b(T) = E(T) - \tau(\theta)$$

and the **mean squared error** (MSE) of T is given by

$$MSE(T) = E[T - \tau(\theta)]^2$$

MSE

If T is an estimator of $\tau(\theta)$, then

$$MSE(T) = Var(T) + [b(T)]^2$$

Proof

$$\begin{aligned} MSE(T) &= E[T - \tau(\theta)]^2 \\ &= E[T - E(T) + E(T) - \tau(\theta)]^2 \\ &= E[T - E(T)]^2 + 2[E(T) - \tau(\theta)] \\ &\times [E(T) - E(T)] + [E(T) - \tau(\theta)]^2 \\ &= Var(T) + [b(T)]^2 \end{aligned}$$

Bayes' Estimator

- Treat θ as a random variable with prior $p(\theta)$
- Bayes' rule: $p(\theta|\mathcal{X}) = p(\mathcal{X}|\theta)p(\theta)/p(\mathcal{X})$
- Full: $p(x|\mathcal{X}) = \int p(x|\theta)p(\theta|\mathcal{X})d\theta$
- Maximum a Posteriori (MAP):

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|\mathcal{X})$$

- Maximum likelihood (MLE):

$$\hat{\theta}_{MLE} = \arg\max_{\theta} p(\mathcal{X}|\theta)$$

- Bayes':

$$\hat{\theta}_{Bayes'} = E[\theta|\mathcal{X}] = \int \theta p(\theta|\mathcal{X})d\theta$$

MAP vs MLE Example

- Let's compare the MAP vs the MLE for a specific example
- Let's look at artificial data generated from a binomial likelihood function

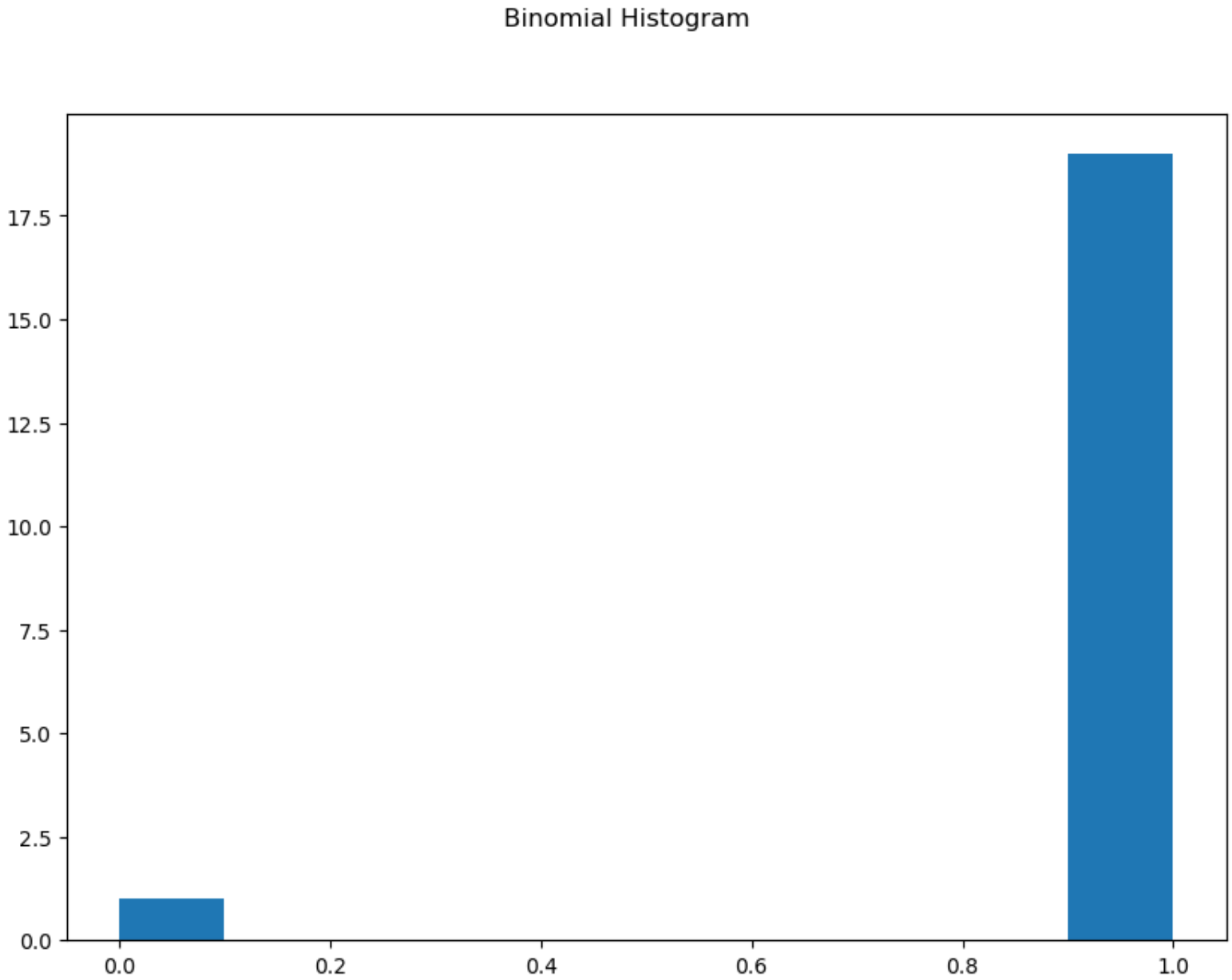
```
np.random.seed(123456789)
theta = 0.85
m = 20
D = np.random.binomial(1, theta, m)
```

- The MLE

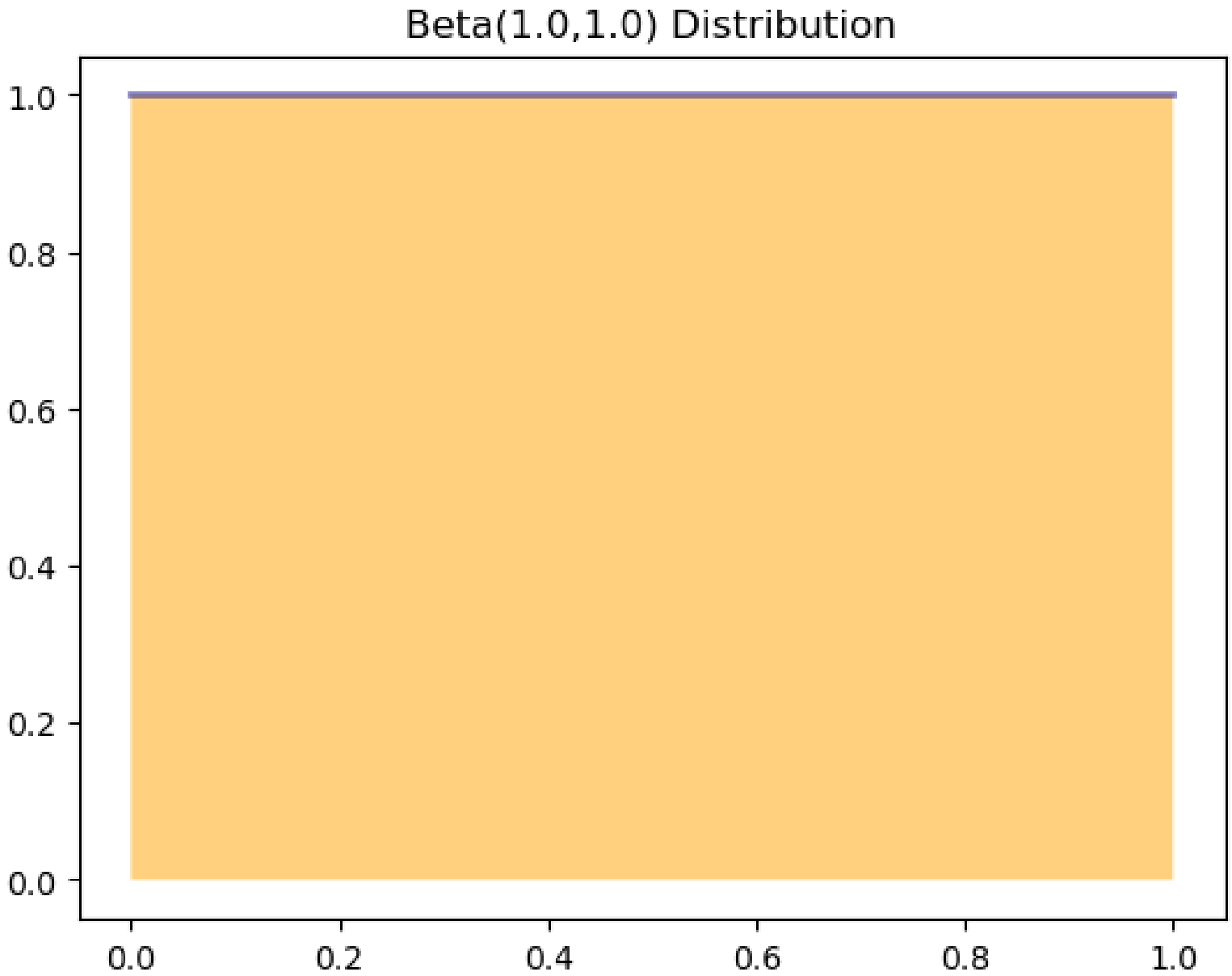
```
mle = np.mean(D)
print(f"The MLE is: {mle: 0.4f}")
```

```
## The MLE is:  0.9500
```

The Histogram of the Generated Data



MAP with a Flat Prior

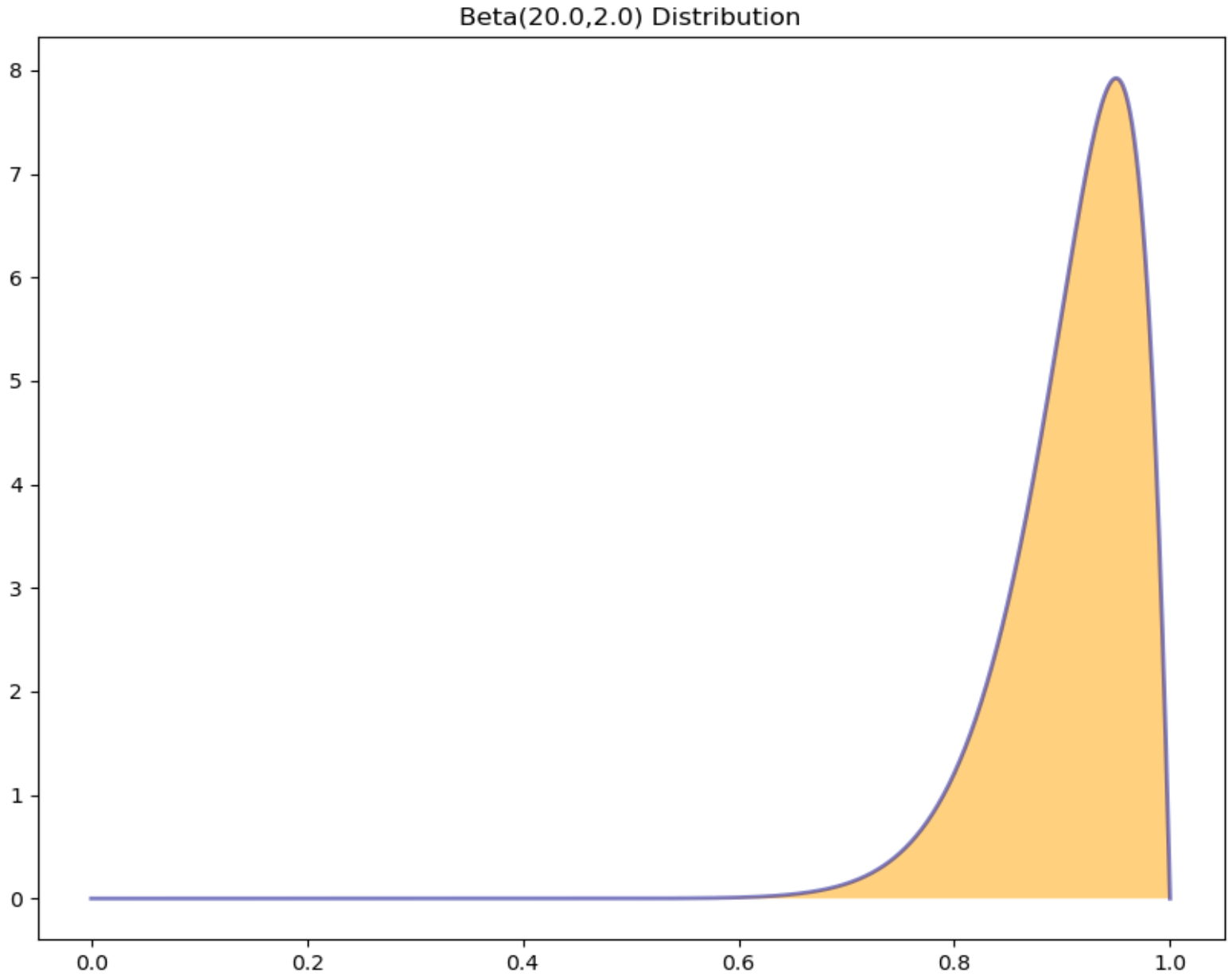


The MAP Continued

```
## Conjugate model will be Beta(a*, b*) via Bayes' Rule
N1 = np.sum(D)
N0 = m - N1
a_post = a_prior + N1
b_post = b_prior + N0
a_post, b_post
```

```
## (20.0, 2.0)
```

The Posterior



The MAP Point Estimate

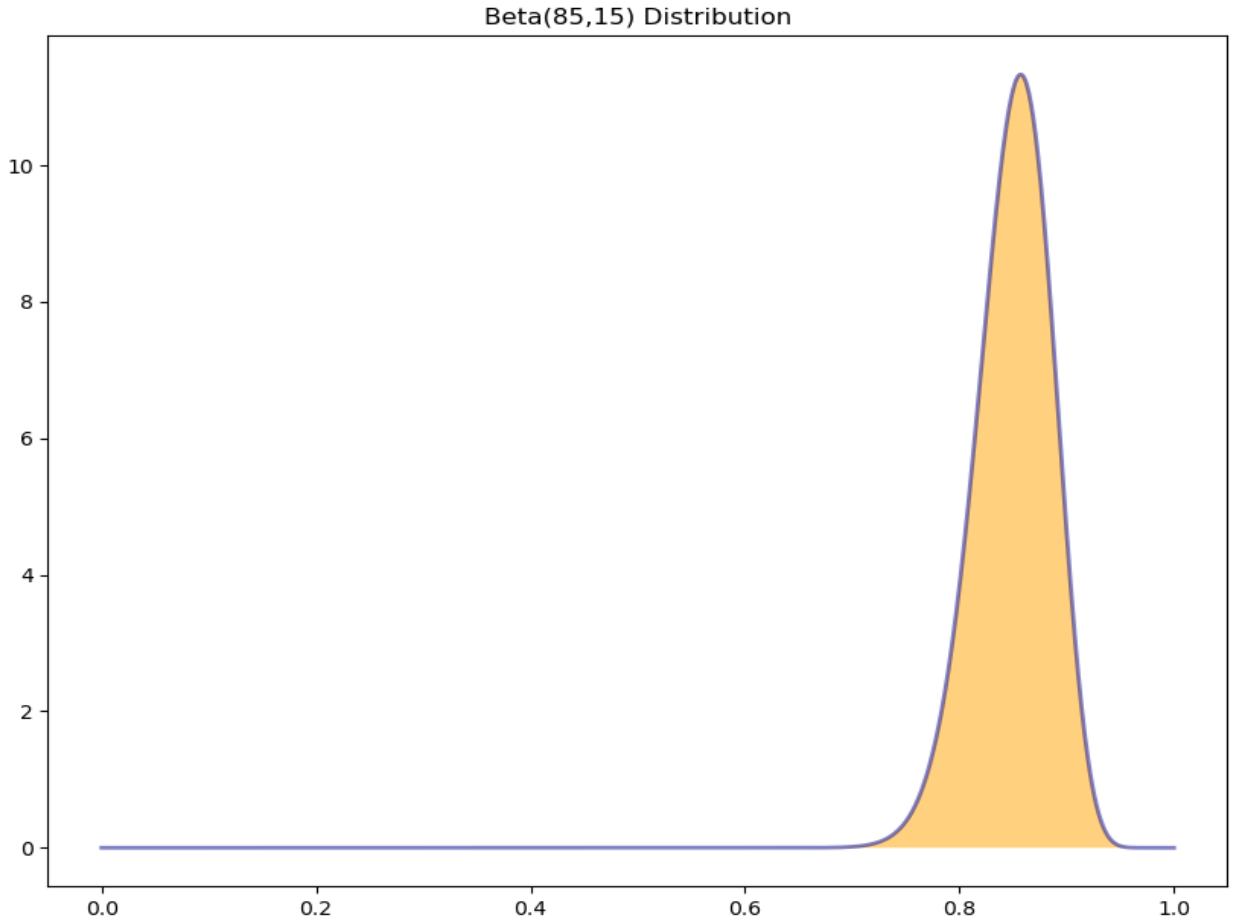
- We see that, at least for this prior, the MAP is identical to the MLE!

```
map = (a_post - 1.0) / (a_post + b_post - 2.0)
print(f"The MAP is: {map : 0.4f}")
```

```
## The MAP is:  0.9500
```

MAP with an Informative Prior: The Prior

- Let $\alpha = 85$ and $\beta = 15$

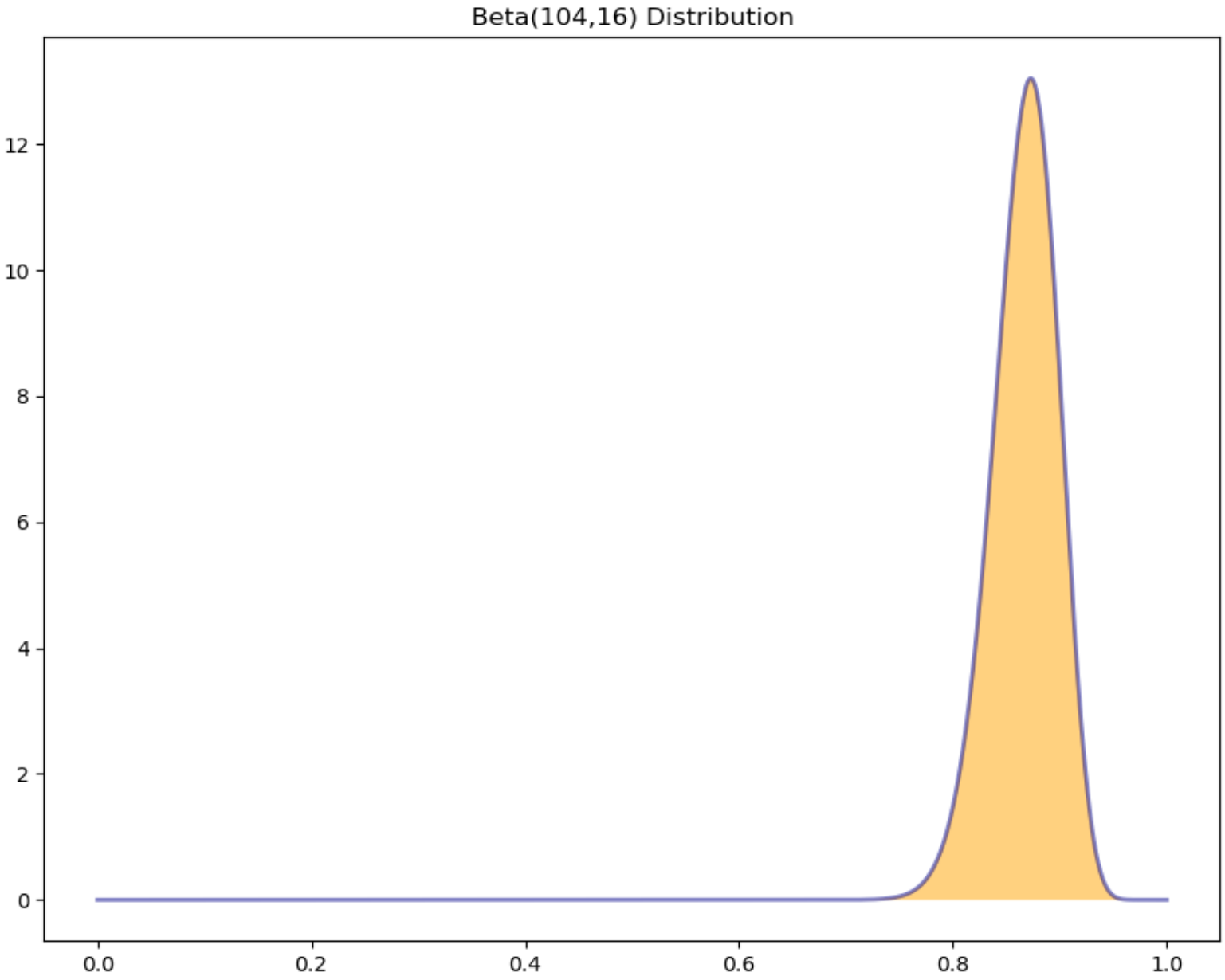


MAP with an Informative Prior: The Point Estimate

```
alpha_post = alpha + N1
beta_post = beta + N0
map_inf = (alpha_post - 1.0) / (alpha_post + beta_post - 2.0)
print(f"The Informative MAP: {map_inf : 0.4f}")
```

```
## The Informative MAP:  0.8729
```

MAP with an Informative Prior: The Posterior



Parametric Classification

- Using Bayes' Rule we can write the probability of class C_i as

$$p(C_i|x) = \frac{p(x|C_i)p(C_i)}{p(x)} = \frac{p(x|C_i)p(C_i)}{\sum_k p(x|C_k)p(C_k)}$$

- We can use this as a discriminant function (where we just care about the numerator):

$$g_i(x) = p(x|C_i)p(C_i)$$

OR

$$g_i(x) = \log P(x|C_i) + \log p(C_i)$$

Parametric Classifier Continued

- If we can assume that $p(x|C_i)$ are Normal then we will have the following likelihood:

$$p(x|C_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right]$$

- The discriminant function then simplifies to the following:

$$-\frac{1}{2}\log 2\pi - \log \sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log p(C_i)$$

An Example

- Imagine we are a car company selling K different types of cars
- Let's assume that the only factor that affects a customer's choice of car is her annual income, denoted x
- $p(C_i)$ is the proportion of customers who buy car type i
- If the annual income of car buyers can be approximated with a Normal distribution, then $p(x|C_i)$ is Normal
- The probability that a customer will purchase car type i who has income x will be $\mathcal{N}(\mu_i, \sigma_i^2)$
  - Where μ_i is the mean income of such customers
  - And σ_i is their income variance
- Using this model, we could then make predictions about which type of car a given customer might purchase given their income

- We typically will not know $p(C_i)$ nor $p(x|C_i)$, so we will have to estimate them from sample data

- We can then plug our estimates into the discriminant function to make predictions

- We are given the following sample

$$\mathcal{X} = \{x_j, y_j\}_{j=1}^{N}$$

- Where $x \in \mathcal{R}$ is one-dimensional and $\mathbf{y} \in \{0, 1\}^K$, such that

- For each class, the estimates for the means and variances are

$$m_i = \frac{\sum_j x_j y_{i,j}}{\sum_j y_{i,j}}$$

$$s_i^2 = \frac{\sum_j (x_j - m_i)^2 y_{i,j}}{\sum_j y_{i,j}}$$

- And the estimates for the priors are

$$\hat{p}(C_i) = \frac{1}{N} \sum_j y_{i,j}$$

- Plugging these estimates into the discriminant function above gives us

$$-\frac{1}{2}\log 2\pi - \log s_i - \frac{(x-m_i)^2}{2s_i^2} + \log \hat{p}(C_i)$$

- The first term can be dropped because it is common to all in $g_i(x)$

- If the priors are equal, then they too can be dropped

- If, for some reason, variances are equal then $g_i(x) = -(x-m_i)^2$, and

$$\text{Choose } C_i \text{ if } |x-m_i| = \min_k |x-m_k|$$