Ocean University of China

Fisheries College Of OUC

Bayesian Data Analysis

Author:
Han Fangcheng

Student ID:
18060013010

A note for Bayesian Data Analysis

Only through hard work can we become good steel

July 12, 2021

Contents

1 Probability and inference

1.1 The three steps of Bayesian data analysis

The process of Bayesian data analysis can be idealized by dividing it into the following three steps:

1. Setting up a *full probability model*-a joint probability distribution for all observable and unobservable quantities in a problem.

2. Conditioning on observed data: calculating and interpreting the appropriate *posterior distubrition*-the conditional probability distribution of the unobserved quantities of ultimate interest, given the observed data.

3. Evaluating the fit of the model and the implications of the resulting posterior distribution: how well does the model fit the data, are the substantive conclusions reasonable, and how sensitive are the results to the modeling assumptions in step 1?

We distinguish between two kinds of *estimands*-unobserved quantities for which statistical inferences are made-first, potentially observable quantities, such as future observations of a process, or the outcome under the treatment not received in the clinical trial examples; and second, quantities that are not directly observable, that is, parameters that govern the hypothetical process leading to the observed data(for example, regression coefficients).

1.2 General notation for statistical inference

Paramaters, data, and predictions

As general notation, we let θ denote unobservable vector quantities or population *parameters* of interest, y denote the observed data, and $\widetilde{y}$ denote unknown, but potentially observable, quantities. When using matrix notation, we consider vectors as column vector throughout; for example, if u is a vector with n components, then $u^T u$ is a scalar and uu^T an $n \times n$ matrix.

Exchangeability

The usual starting point of a statistical analysis is the assumption that the n values y_i may be regraded as *exchangeable*, meaning that we express uncertainty as a joint probability density $p(y_1, \cdots, y_n)$ that is invariant to permutations of the indexes. A nonexchangeable model would be appropriate if information relevant to the outcome were conveyed in the unit indexed rather than by explanatory variables.

The Exchangeability of data indicates that the sampling order has no influence on our model.

We commonly model data from an exchangeable distribution as independently and identically distributed(*iid*) given some unknown parameter vector θ with distribution $p(\theta)$.

Explanatory variables

We use x to denote *explanatory variables*. We use X to denote the entire set of explanatory variables for all n units; if there are k explanatory variables, then X is a matrix with n rows and k columns.

1.3 Bayesian inference

Bayesian statistical conclusions about a parameter $Theta$, or unobserved data $\widetilde{y}$, are made in terms of emphprobability statements. These probability statements are conditional on the observed value of y, and in our notation are written simply as $p(\theta|y)$ or $p(\widetilde{y}|y)$.

Probability notation

First, $p(\cdot|\cdot)$ denotes a conditional probability density with the arguments determined by the context, and similarly for $p(\cdot)$, which denotes a marginal dis-

tribution. To avoid confusion, we may use the notation $\Pr(\cdot)$ for the probability of an event; for example, $\Pr(\theta > 2) = \int_{\theta>2} p(\theta)d\theta$. When using a standard distribution, we use a notation based on the name of distribution; for example, if θ has a normal distribution with mean μ and variance σ^2, we write $\theta \sim N(\mu, \sigma^2)$ or $p(\theta) = N(\theta|\mu, \sigma^2)$ or, to be more explicit, $p(\theta|\mu, \sigma^2) = N(\theta|\mu, \sigma^2)$. Throughout, we use notation such as $N(\mu, \sigma^2)$ for random variables and $N(\theta|\mu, \sigma^2)$ for density functions.

We also occasionally ues the following expressions for random variables θ: the *coefficient of variation* is defined as $sd(\theta)/E(\theta)$, the *geometric mean* is $\exp(E[\log(\theta)])$, and the *geometric standard deviation* is $\exp(sd[\log(\theta)])$.

Bayes' rule

$$p(\theta, y) = p(\theta)p(y|\theta)$$

While we refer to $p(\theta)$ as *prior distribution* and $p(y|\theta)$ as *sampling distribution(or data distribution).*

Simply conditioning on the known value of the data y, using the basic property of conditional probability known as Bayes' rule, yields the *posterior* density:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)} \tag{1.1}$$

where $p(y) = \sum_\theta p(\theta)p(y|\theta)$, and in the case of continuous θ: $p(y) = \int p(\theta)p(y|\theta)d\theta$. If $p(y|\theta)$ is a function of θ, not of y, then $p(y)$ can be considered a constant, which does not depend on θ, yielding the *unnormalized posterior density*:

$$p(\theta|y) \propto p(\theta)p(y|\theta) \tag{1.2}$$

Prediction

To make inferences about an unknown observable, often called predictive inferences, we follow a similar logic. Before the data y are considered, the

distribution of the unknown but observable y is

$$p(y) = \int p(y, \theta) d\theta = \int p(\theta) p(y|\theta) d\theta \tag{1.3}$$

This has a informative name which is called *prior predictive distribution*: prior becaues it is not conditional on a previous observation of the process, and predictive because it is the distribution for a quantity that is observable.

After the data y have been observed, we can predict an unknown observable, $\widetilde{y}$, from the same process.

$$\begin{aligned} p(\widetilde{y}|y) &= \int p(\widetilde{y}, \theta|y) d\theta \\ &= \int p(\widetilde{y}|\theta, y) p(\theta|y) d\theta \\ &= \int p(\widetilde{y}|\theta) p(\theta|y) d\theta \end{aligned} \tag{1.4}$$

The last step follows from the assumed conditional independence of y and $\widetilde{y}$ given θ.

Likelihood

Using Bayes' rule with a chosen probability model means that the data y affect the posterior inference(1.2) *only* through $p(y|\theta)$, which, when regarded as a function of θ, for fixed y, is called the *likelihood function*. In this way Bayesian inference is obeying what is sometimes called the *likelihood principle*.

Likelihood and odds ratios The ratio of the posterior density $p(\theta|y)$ evaluated at the points θ_1 and θ_2 under a given model is called the posterior *odds* for θ_1 compared to θ_2.

$$\frac{p(\theta_1|y)}{p(\theta_2|y)} = \frac{p(y|\theta_1)p(\theta_1)/p(y)}{p(y|\theta_2)p(\theta_2)/p(y)} = \frac{p(\theta_1)p(y|\theta_1)}{p(\theta_2)p(y|\theta_2)} \tag{1.5}$$

In words, the posterior odds are equal to the prior odds multiplied by the *likeihood ratio*, $p(y|\theta_1)/p(y|\theta_2)$.

1.4 Useful results from probability

We can *factor* a joint density as a product of marginal and conditional densities; for example, $p(u, v, w) = p(u|v, w)p(v|w)p(w)$.

In the interest of conciseness, however, our notation hides the conditioning on hypotheses that hold throughout-no probability judgement can be made in a vacuum-and to be more explicit one might use a notation such as the following:

$$p(\theta, y|H) = p(\theta|H)p(y|\theta, H)$$

where H refers to the set of hypotheses or assumptions used to define the model.

Means and variances of conditional distributions

It is often useful to express the mean and variance of a random variable u in terms of the conditional mean and variance given some related quantity v. The mean of u can be obtained by averaging the conditional mean over the marginal distribution of v,

$$E(u) = E(E(u|v)) \tag{1.6}$$

where the inner expectation averages over u, conditional on v, and the outer expectation averages over v. Identity(1.6) is easy to derive by writing the expectation in terms of the joint distribution of u and v and then factoring the joint distribution:

$$E(u) = \int\int up(u, v)dudv = \int\int up(u|v)dup(v)dv = \int E(u|v)p(v)dv$$

The corresponding result for the variance includes two terms, the mean of the conditional variance and the variance of the conditional mean:

$$var(u) = E(var(u|v)) + var(E(u|v)) \tag{1.7}$$

This result can be derived by expanding the terms on the right side of(1.7):

$$\begin{aligned} E(var(u|v)) + var(E(u|v)) &= E(E(u^2|v)) - (E(u|v))^2) + E((E(u|v))^2) - (E(E(u|v)))^2 \\ &= E(u^2) - E((E(u|v))^2) + E((E(u|v))^2) - (E(u))^2 \\ &= E(u^2) - (E(u))^2 \\ &= var(u) \end{aligned}$$

Identities(1.6) and (1.7) also hold if u is a vector, in which case $E(u)$ is a vector and $var(u)$ a matirx.

Transformation of variables

It is common to transform a probability distribution from one parameterization to another. Suppose $p_u(u)$ is the density of the vector u, and we transform to $v = f(u)$, where v has the same number of components as u.

If p_u is a discrete distribution, and f is a one-to-one function, then the density of v is given by

$$p_v(v) = p_u(f^{-1}(v))$$

If f is a many-to-one function, then a sum of terms appears on the right side of this expression for $p_v(v)$, with one term corresponding to each of the banches of the inverse function.

If p_u is a continuous distribution, and $v = f(u)$ is a one-to-one transformation, then the joint density of the transformed vector is

$$p_v(v) = |J| p_u(f^{-1}(v))$$

where $|J|$ is the absolute value of the determinant of the Jacobian of transformation $u = f^{-1}(v)$ as a function of v; the Jacobian J is the square matrix of partial derivatives, with the (i, j)th entry equal to $\partial u_i / \partial v_j$. Once again, if f is many-to-one, then $p_v(v)$ is a sum or integral of terms.

In one dimension, we commonly use the logarithm to transform the parameter space from $(0, \infty)$ to $(-\infty, \infty)$. When working with parameters defined on the open unit interval, $(0, 1)$, we often use the logistic transformation:

$$logit(u) = \log(\frac{u}{1-u}) \tag{1.8}$$

whose inverse transformation is

$$logit^{-1}(v) = \frac{e^v}{1+e^v}$$

Another common choice is the probit transformation, $\Phi^{-1}(u)$, where Φ is the standard normal cummulative distribution function, to transform from $(0,1)$ to $(-\infty, \infty)$.

1.5 Computation and software

Sampling using the inverse cumulative distribution function

Theorem 1.1 *If X is a continuous random variable, and it's cdf is F_X, then $U = F_X^{-1}(X) \sim U(0,1)$.*

Definition 1.1

$$F_X^{-1}(u) = \inf\{x : F_X(x) = u\}, 0 < u < 1$$

If random varibale $U \sim U(0,1)$, then for all $x \in R$:

Proof 1.1

$$\begin{aligned} P(F_X^{-1}(U) \leq x) &= P(\inf\{t : F_X(t) = U\} \leq x) \\ &= P(U \leq F_X(x)) \\ &= F_U(F_X(x)) \\ &= F_X(x) \end{aligned}$$

2 Single-parameter models

In this chapter, we consider four fundamental and widely used one-dimensional models-the binomial, normal, Poisson, and exponential.

2.1 Estimating a probability from binomial data

The binomial sampling model is:

$$p(y \mid \theta) = \text{Bin}(y \mid n, \theta) = \begin{pmatrix} n \\ y \end{pmatrix} \theta^y (1-\theta)^{n-y} \tag{2.1}$$

Example 2.1 ***Estimating the probability of a female birth***

Let y be the number of girls in n recorded births. For this example we define the parameter θ to be the proportion of female births. To perform Bayesian inference in the binomial model, we must specify a prior distribution for θ, we assume that the prior distribution for θ is uniform on the interval $[0, 1]$.

Elementary application of Bayes' rule as displayed in (1.2), applied to (2.1), then gives the posterior density for θ as

$$p(\theta|y) \propto \theta^y(1-\theta)^{n-y} \tag{2.2}$$

In the present case, we can recognize (2.2) as the unnormalized form of the *beta* distribution [1]

$$\theta|y = \text{Beta}(y+1, n-y+1) \tag{2.3}$$

A Gamma, Beta distribution

A.1 Magical gamma function

The Gamma is

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt, \quad \alpha > 0$$

Through the Integration by parts, the following properties can be obtained:

$$\begin{aligned}
\Gamma(\alpha+1) &= \int_0^\infty t^\alpha(e^{-t})dt \\
&= -\int_0^\infty t^\alpha d(e^{-t}) \\
&= -[t^\alpha e^{-t}|_0^\infty - \alpha\int_0^\infty e^{-t}t^{\alpha-1}dt] \\
&= \alpha\Gamma(\alpha)
\end{aligned}$$

[1] The Probability density function of Beta distribution is

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} = \frac{1}{\text{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

It is easy to prove that

$$\Gamma(n+1) = n!, \Gamma(1) = 1, \Gamma(\frac{1}{2} = \sqrt{\pi})$$

Visualization of gamma function

Listing 1: **Gamma.py**

```
1  import numpy as np
2  from scipy.special import gamma
3  import matplotlib.pyplot as plt
4
5  fig = plt.figure(figsize=(12,8))
6
7  x = np.linspace(-5, 5, 1000)
8  plt.plot(x, gamma(x), ls='-', c='k', label='$\Gamma(x)$')
9
10 x2 = np.linspace(1,6,6)
11 y = np.array([1, 1, 2, 6, 24, 120])
12 plt.plot(x2, y, marker='*', markersize=12, markeredgecolor='r',
       markerfacecolor='r',ls='',c='r',label='$(x-1)!$')
13 plt.title('Gamma Function')
14 plt.ylim(-50, 50)
15 plt.xlim(-5, 5)
16 plt.xlabel('$x$')
17 plt.legend()
18 plt.show()
```

From A.1 we can see that gamma function is a Convex function, and at the same time $\log \Gamma(x)$ (A.2) is also a convex function.

Listing 2: **log gamma.py**

```
1  import numpy as np
2  from scipy.special import gamma
3  import matplotlib.pyplot as plt
4
5  fig = plt.figure(figsize=(12,8))
6  # The Gamma function
7  x = np.linspace(0, 15, 1000)
8  plt.plot(x, np.log(gamma(x)), ls='-', c='k', label='$\log\Gamma(x)$')
9
10 plt.title('$Log\Gamma(x)$ Fuction')
11 plt.ylim(-1, 50)
12 plt.xlim(-1, 15)
13 plt.xlabel('$x$')
14 plt.legend()
15 plt.show()
```

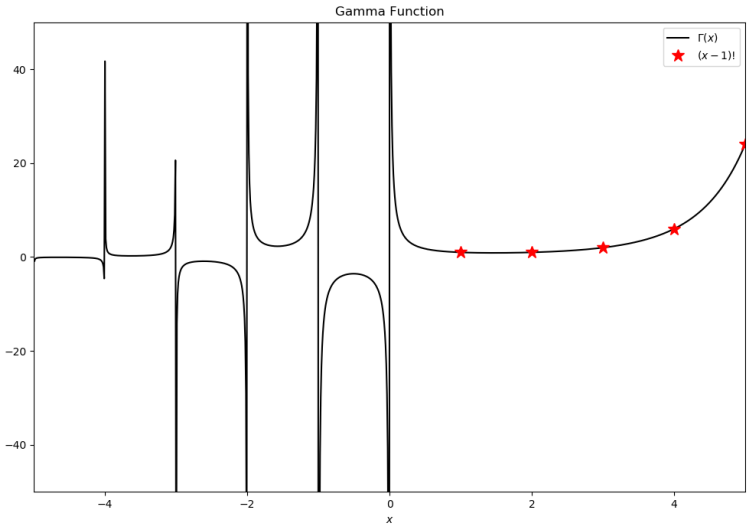

Figure A.1: Gamma function plot

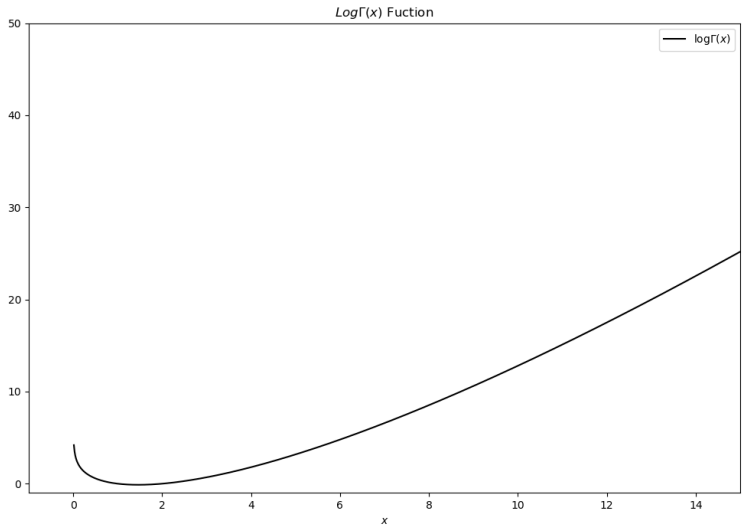

Figure A.2: log gamma function plot

The belowing function is called *Digamma* function:

$$\Psi = \frac{d\log(\Gamma(x))}{dx}$$

The *Digamma* function has the belowing property:

$$\Psi(x+1) = \frac{d\log(\Gamma(x+1))}{dx} = \frac{d\log(x\Gamma(x))}{dx} = \frac{d\log(\Gamma(x))}{dx} + \frac{d\log(x)}{dx} = \Psi(x) + \frac{1}{x}$$

From binomial distribution to gamma function

By deforming the gamma function, you can get the following formula:

$$\int_0^\infty \frac{t^{\alpha-1}e^{-t}dt}{\Gamma(\alpha)} = 1$$

Taking the function in the integral as the probability density, a simple density function of Gamma distribution[2] is obtained:

$$\text{Gamma}(\text{t}|\alpha) = \frac{t^{\alpha-1}e^{-t}}{\Gamma(\alpha)}$$

[2]I think this is just a definition, and it has nothing with gamma function itself

If you make a transformation $t = \beta x$, you will get a more general form of Gamma distribution:

$$\text{Gamma}(x|\alpha,\beta) = \frac{\beta^{\alpha} x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$$

, α is called shape parameter, it mainly determines the shape of the distribution curve. And β is called rate parameter or inverse scale parameter($\frac{1}{\beta}$ scale parameter), it mainly determines how steep the curve is.

Listing 3: **gamma distribution.py**

```
import numpy as np
from scipy.stats import gamma
from matplotlib import pyplot as plt

alpha_values = [1, 2, 3, 3, 3]
beta_values = [0.5, 0.5, 0.5, 1, 2]
color = ['b', 'r', 'g', 'y', 'm']
x = np.linspace(1E-6, 10, 1000)

fig, ax = plt.subplots(figsize=(12, 8))

for k, t, c in  zip(alpha_values, beta_values, color):
        dist = gamma(k, 0, t)
        plt.plot(x, dist.pdf(x), c=c, label=r'$\alpha=%.1f,\theta=%.1f$' %(k
              ,t))

plt.xlim(0, 10)
plt.ylim(0, 2)

plt.xlabel('$x$')
plt.ylabel(r'$p(x|\alpha,\beta)$')
plt.title('Gamma Distribution')

plt.legend(loc=0)
plt.show()
```

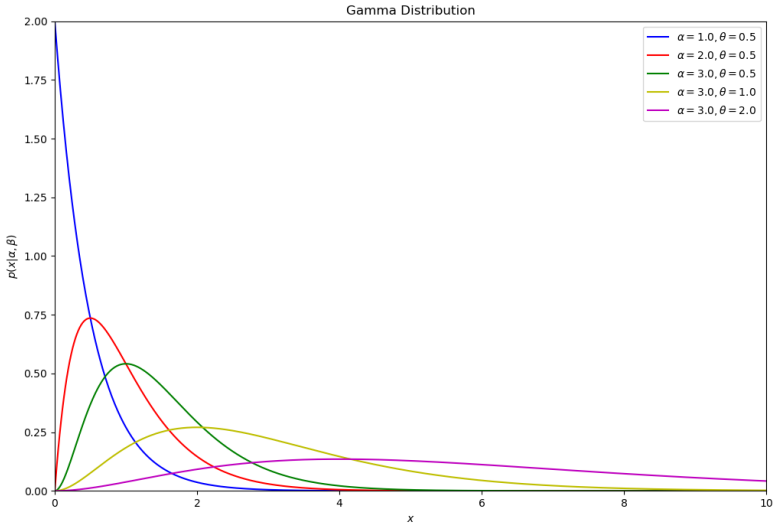

Figure A.3: gamma distribution

We can find that the probability density of Gamma distribution (A.3) is highly consistent with the mathematical form of Poisson distribution. The Poisson distribution of the parameter λ, the probability is

$$\text{Poisson}(X = k|\lambda) = e^{-\lambda}\frac{\lambda^k}{k!}$$

If we take $\alpha = k + 1, \beta = 1$ from the density function of Gamma distribution, we can get

$$\text{Gamma}(x|\alpha = k + 1, \beta = 1) = \frac{x^k e^{-x}}{\Gamma(k + 1)} = \frac{x^k e^{-x}}{k!}$$

We can see that the two distributions are consistent in mathematical form, but Poisson distributed discrete and Gamma distributed continuous. It can be intuitively considered that Gamma distribution is a continuous version of Poisson distribution on a set of positive real numbers.

We have learned in the courses of probability theory and mathematical statistics that Poisson(λ) distribution can be regarded as the limit distribution of binomial

distribution $\mathrm{B}(n,p)$ under the condition of $np=\lambda, n\to\infty$:

$$\mathrm{B}(k;n,p)=C_n^k p^k(1-p)^{n-k}\xrightarrow{np=\lambda,n\to\infty}\mathrm{Poisson}(X=k|np=\lambda)=\frac{\lambda^k e^{-\lambda}}{k!}$$

The binomial distribution also satisfies the following wonderful equation:

$$\mathrm{P}(x\le k)=\frac{n!}{k!(n-k-1)!}\int_p^1 t^k(1-t)^{n-k-1}dt$$

This equation reflects the relationship between Binomial distribution and Beta distribution, we will prove it later.

We make a change $t=\frac{x}{n}$ in the right equation

$$\begin{aligned}
P(x\le K)&=\frac{n!}{k!(n-k-1)!}\int_p^1 t^k(1-t)^{n-k-1}dt\\
&=\frac{n!}{k!(n-k-1)!}\int_{np}^n(\frac{x}{n})^k(1-\frac{x}{n})^{n-k-1}d\frac{x}{n}\\
&=\frac{(n-1)!}{k!(n-k-1)!}\int_{np}^n(\frac{x}{n})^k(1-\frac{x}{n})^{n-k-1}dx\\
&=\int_{np}^n\begin{pmatrix}n-1\\k\end{pmatrix}\left(\frac{x}{n}\right)^k\left(1-\frac{x}{n}\right)^{n-k-1}dx\\
&=\int_{np}^n\mathrm{Binomial}(Y=k|n-1,\frac{x}{n})dx
\end{aligned}$$

The left side of the above formula is binomial distribution $\mathrm{B}(n,p)$, while the right side is the integral summation of infinite binomial distribution $\mathrm{B}(n-1,\frac{x}{n})$, so it can be written as

$$\mathrm{Binomial}(X\le k|n,p)=\int_{np}^n\mathrm{Binomial}(Y=k|n-1,\frac{x}{n})dx$$

If you take the limit on both sides under the condition $np=\lambda, n\to\infty$, then there is a $\mathrm{B}(n,p)\to\mathrm{Poisson}(\lambda)$ on the left and $\mathrm{B}(n-1,\frac{x}{n})\to\mathrm{Poisson}(x)$ on the right, so you get:

$$\mathrm{Poisson}(X\le k|\lambda)=\int_\lambda^\infty\mathrm{Poisson}(Y=k|x)dx$$

, expand the Poisson distribution and get

$$\mathrm{Poisson}(X\le k|\lambda)=\int_\lambda^\infty\frac{x^k e^{-x}}{k!}dx$$

, this is *Possion-Gamma duality.*

We take the limit $\lambda \to 0$ on both sides of the above formula, and on the left is the probability of Poisson happening at most k. When $\lambda \to 0$, there can be no more events, so $\mathrm{P}(X \leq k) = 1$, so:

$$1 = \lim_{\lambda \to 0} \int_{\lambda}^{\infty} \frac{x^k e^{-x}}{k!} dx = \int_{0}^{\infty} \frac{x^k e^{-x}}{k!} dx$$

The integral formula shows that $\frac{x^k e^{-x}}{k!}$ is a probability distribution function on the set of real numbers, and this function happens to be the Gamma distribution.

We continue to move the $k!$ in the right side of the above formula to the left, so we get:

$$k! = \int_{0}^{\infty} x^k e^{-x} dx$$

So we get a way to express $k!$ as an integral.

Let's change the formula $\mathrm{Poisson}(X \leq k|\lambda) = \int_{\lambda}^{\infty} \frac{x^k e^{-x}}{k!} dx$:

$$\mathrm{Poisson}(X \leq k|\lambda) + \int_{0}^{\lambda} \frac{x^k e^{-x}}{k!} dx = 1$$

We can see that there is a complementary relationship between the probability density accumulation function of Poisson distribution and the probability density accumulation function of Gamma distribution.

Let's make a summary: starting from the equation of binomial distribution and using that the limit of binomial distribution is Poisson distribution, we derive Gamma distribution and express $k!$ in integral form at the same time.

A.2 Beta distribution

We will draw several distributions from a few questions:

Question 1:

1. $X_1, X_2, \cdots, X_n \sim iid\, \mathrm{Uniform}(0, 1)$
2. Sort these n random variables and get the order statistics $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$
3. What is $X_{(k)}$'s distribution

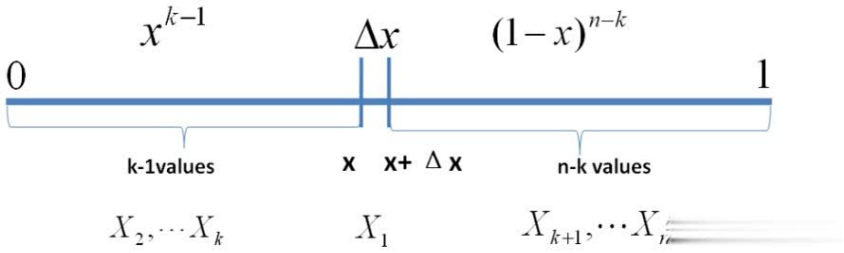

Figure A.4: situation one

First we try to calculate the probability that $X_{(k)}$ falls in the interval $[x, x+\Delta x]$, that is

$$P(x \leq X_{(k)} \leq x + \Delta x) = ?$$

We can divide $[0, 1]$ into three parts: $[0, x), [x, x + \Delta x], (x + \Delta x, 1]$
We talk about the first situation(A.4): Assuming that only one of the n number falls in the interval $[x, x + \Delta x]$, then the number $X_{(k)}$ in this interval is the kth largest, so there must be $k-1$ numbers in $[0, x)$ and $n-k$ numbers in $(x+\delta x, 1]$, we discribe this situation as event E:

$$E = \{X_1 \in [x, x+\Delta x], X_i \in [0, x)(i = 2, \cdots, k), X_j \in (x+\Delta x, 1](j = k+1, \cdots, n)\}$$

then:

$$\begin{aligned} P(E) &= \prod_{i=1}^{n} P(X_i) \\ &= x^{k-1}(1 - x - \Delta x)^{n-k}\Delta x \\ &= x^{k-1}(1 - x)^{n-k} + o(\Delta x) \end{aligned}$$

$o(\Delta x)$ is the higher-order infinitesimal of Δx. It is obvious that there are n ways for 1 number to fall in the interval of $[x, x + \Delta x]$, there are $\begin{pmatrix} n-1 \\ k-1 \end{pmatrix}$ kinds of combination for $k - 1$ numbers falling in the interval $[0, x)$ of the remaining $n - 1$ numbers, so there are a total of $n \begin{pmatrix} n-1 \\ k-1 \end{pmatrix}$ events which is equivalent to event E.

Now we talk about the second situation(A.5): Assuming that two numbers falls in interval $[x, x + \Delta x]$:

$$E' = \{X_1, X_2 \in [x, x+\Delta x], X_i \in [0, x)(i = 3, \cdots, k), X_j \in (x+\Delta x, 1](j = k+1, \cdots, n)\}$$

then:

$$\begin{aligned} P(E) &= \prod_{i=1}^{n} P(X_i) \\ &= x^{k-2}(1 - x - \Delta x)^{n-k}(\Delta x)^2 \\ &= o(\Delta x) \end{aligned}$$

From the above analysis, it can be concluded that as long as more than one

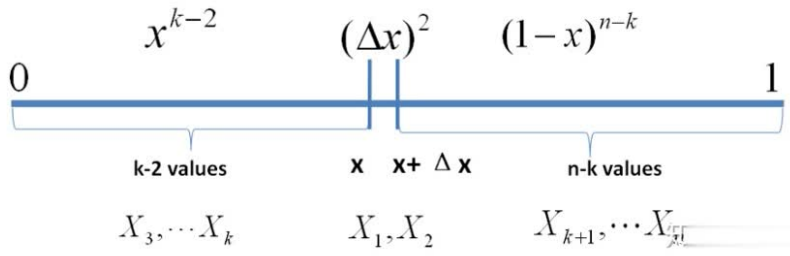

Figure A.5: Situation two

number falls within the $[x, x + \Delta x]$, the probability of the corresponding event is $o(\Delta x)$. So:

$$\begin{aligned} &P(x \le X_{(k)} \le x + \Delta x) \\ &\qquad = n \begin{pmatrix} n-1 \\ k-1 \end{pmatrix} P(E) + o(\Delta x) \\ &\qquad = n \begin{pmatrix} n-1 \\ k-1 \end{pmatrix} x^{k-1}(1-x)^{n-k}\Delta x + o(\Delta x) \end{aligned}$$

So the probability density function of $X_{(k)}$ is:

$$\begin{aligned} f(x) &= \lim_{\Delta x \to 0} \frac{P(x \leq X_{(k)}) \leq x + \Delta x)}{\Delta x} \\ &= n \begin{pmatrix} n-1 \\ k-1 \end{pmatrix} x^{k-1}(1-x)^{n-k} \\ &= \frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k}, x \in [0,1] \end{aligned}$$

We know that many mathematical concepts can be extended from the set of integers to the set of real numbers by using Gamma function. So we let $\alpha = k, \beta = n - k + 1$ in the above equation:

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

, This is the *Beta* distribution.

Visualization of beta function

Listing 4: **beta distribution.py**

```
import numpy as np
from scipy.stats import beta
from matplotlib import pyplot as plt

alpha_values = [1/3, 2/3, 1, 1, 2, 2, 4, 10, 20]
beta_values = [1, 2/3, 3, 1, 1, 6, 4, 30, 20]
colors =  ['tab:blue', 'tab:orange', 'tab:green', 'tab:red', 'tab:purple',
           'tab:brown', 'tab:pink', 'tab:gray', 'tab:olive']

x = np.linspace(0, 1, 1002)[1:-1]
fig, ax = plt.subplots(figsize=(14,9))

for a, b, c in  zip(alpha_values, beta_values, colors):
        dist = beta(a, b)
        plt.plot(x, dist.pdf(x), c=c, label=r'$\alpha=%.1f, \beta=%.1f$' %(a
             , b))
```

```
16
17 plt.xlim(0, 1)
18 plt.ylim(0, 6)
19
20 plt.xlabel('$x$')
21 plt.ylabel(r'$p(x|\alpha,\beta)$')
22 plt.title('Beta distribution')
23
24 ax.annotate('Beta(1/3,1)', xy=(0.014, 5), xytext=(0.04, 5.2),
25             arrowprops= dict(facecolor='black', arrowstyle='-'))
26 ax.annotate('Beta(10,30)', xy=(0.276, 5), xytext=(0.3, 5.4),
27             arrowprops= dict(facecolor='black', arrowstyle='-'))
28 ax.annotate('Beta(20,20)', xy=(0.5, 5), xytext=(0.52, 5.4),
29             arrowprops= dict(facecolor='black', arrowstyle='-'))
30 ax.annotate('Beta(1,3)', xy=(0.06, 2.6), xytext=(0.07, 3.1),
31             arrowprops= dict(facecolor='black', arrowstyle='-'))
32 ax.annotate('Beta(2,6)', xy=(0.256, 2.41), xytext=(0.2, 3.1),
33             arrowprops= dict(facecolor='black', arrowstyle='-'))
34 ax.annotate('Beta(4,4)', xy=(0.53, 2.15), xytext=(0.45, 2.6),
35             arrowprops= dict(facecolor='black', arrowstyle='-'))
36 ax.annotate('Beta(1,1)', xy=(0.8, 1), xytext=(0.7, 2),
37             arrowprops= dict(facecolor='black', arrowstyle='-'))
38 ax.annotate('Beta(2,1)', xy=(0.9, 1.8), xytext=(0.75, 2.6),
39             arrowprops= dict(facecolor='black', arrowstyle='-'))
40 ax.annotate('Beta(2/3,2/3)', xy=(0.99, 2.4), xytext=(0.86, 2.8),
41             arrowprops= dict(facecolor='black', arrowstyle='-'))
42
43 plt.show()
```

As can be seen from the figure A.6, the Beta distribution can be concave, convex, monotonously rising, monotonously decreasing; it can be a curve or a straight line, and the uniform distribution is also a special Beta distribution. You can try to change the parameters to see the various forms of Beta distribution.

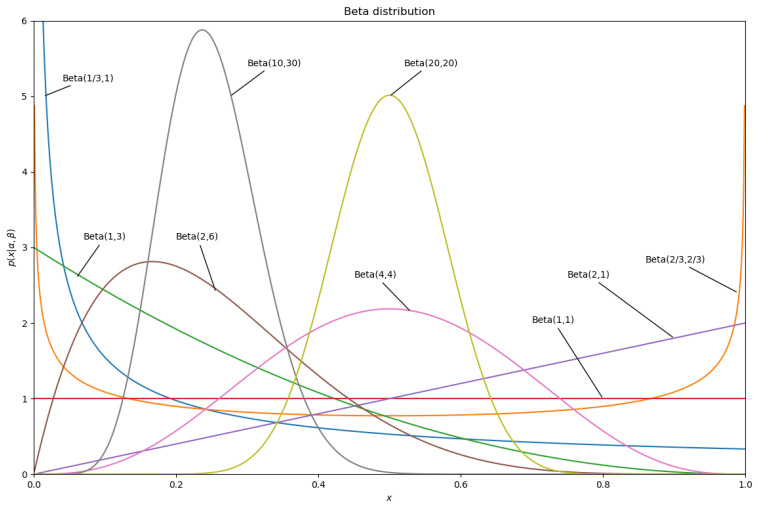

Figure A.6: Beta distribution

Beta-binomial conjugation

Question 2:

1. $X_1, X_2, \cdots, X_n \sim iid\,\text{Uniform}(0,1)$, the corresponding order statistics after sorting:$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$, we assumpt $p = X_{(k)}$

2. $Y_1, Y_2, \cdots, Y_m \sim iid\,\text{Uniform}(0,1)$, m_1 of Y_i are smaller than p, m_2 of Y_i are larger than p.

3. What is the distribution of $P(p|Y_1, Y_2, \cdots, Y_m)$?

Because $p = X_{(k)}$ is the kth larger among $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$, we can easily prove that $p = X_{(k)}$ is the $k + m_1$th larger among $m + n$ iid variables $X_{(1)}, X_{(2)}, \cdots, X_{(n)}, Y_1, Y_2, \cdots, Y_m \sim idd\,\text{Uniform}(0,1)$. Depending on the conclusion we get in last section, we can get that the probability density of $p = X_{(k)}$ is $\text{Beta}(p|k + m_1, n - k + 1 + m_2)$.

According to the logic of Bayesian reasoning:

1. $p = X_{(k)}$ is the parameter that we will estimate. We have proved that the probability density of p is $f(p) = \text{Beta}(p|k, k, n-k+1)$, which is called the prior distribution of p.

2. In the data Y_i, there are m_1 numbers that are smaller than p, and m_2 numbers are larger than p. Y_i is equivalent to doing m times Bernoulli experiment , so m_1 obeys binomial distribution $\text{B}(n, p)$

3. Given the (m_1, m_2) knowledge provided from the data, the posterior distribution is $f(p|m_1, m_2) = \text{Beta}(p|k + m_1, n - k + 1 + m_2)$.

The basic process of Bayesian parameter estimation is: **Prior distribution + data knowledge = posterior distribution**
So we can get:

$$\text{Beta}(p|k, n-k+1) + \text{BinomCount}(m_1, m_2) = \text{Beta}(p|k + m_1, n - k + 1 + m_2)$$

More generally, for non-negative real numbers α, β, we have the following relationship:

$$\text{Beta}(p|\alpha, \beta) + \text{BinomCount}(m_1, m_2) = \text{Beta}(p|\alpha + m_1, \beta + m_2)$$

In the previous derivation of the Gamma distribution from the binomial distribution, the following equation was used:

$$P(x \leq K) = \frac{n!}{k!(n-k-1)!} \int_p^1 t^k (1-t)^{n-k-1} dt$$

On the left is the probability accumulation of binomial distribution, and on the right is the probability accumulation of $\text{Beta}(t|k+1, n-k)$ distribution. Now let's prove this equation.
We construct the following binomial distribution(A.7) and take random variables $X_1, X_2, \cdots, X_n \sim iid\, \text{Uniform}(0, 1)$. A successful Bernoulli experiment is X_i We can get:

$$P(C \leq k) = P(X_{(k+1)} > p)$$

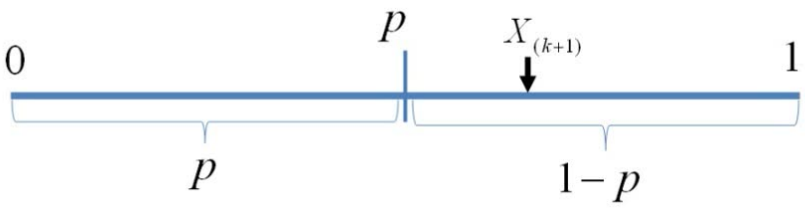

Figure A.7: Situation three

Here $P(X_{(k+1)})$ is the order statistic, which is the $k+1$th largest number. The above equation means that the number of success at most k times equal to $k+1$th larger number is bound to fail (that is, failure at least $n-k$ times). Because of $X_{(t+1)} \sim \text{Beta}(t|k+1, n-k)$, so

$$\begin{aligned} P(C \leq k) &= P(X_{k+1} > p) \\ &= \int_p^1 \text{Beta}(t|k+1, n-k)dt \\ &= \frac{n!}{k!(n-k-1)!} \int_p^1 t^k (1-t)^{n-k-1} dt \end{aligned}$$