Statistics for Data Science

Lecture 9

Gamma and Beta Distributions

Ken Cameron

Class Test

- 1.15pm Thursday 2nd Nov
- 1W 2.104
- 60 minutes
- Multiple choice (10 question)

Class Test

- Test will start promptly at 1.15pm.

- Desks will not be numbered.
  - You may sit where you wish.

- Collect question paper and script on arrival.
  - Don’t open question paper until instructed to do so.

- Calculators will be provided.
  - Collect on arrival.

Class Test

- You should hand in your answer script and calculator before leaving.
    - Don’t leave in the first 20 minutes.
    - Don’t leave in the last 10 minutes.
- No admittance after the first 20 minutes.
- You will need pencil + eraser.

Content

- Gamma Distribution
- Chi-squared Distribution
- Beta Distribution
- CLT

Gamma Distribution

- A family of distributions in $[0, \infty)$.

- If α is a positive constant, the following integral is finite:

$$\int_0^\infty t^{\alpha-1} e^{-t} dt$$

  - If α is a positive constant, the integral can be expressed in closed form.

Gamma Function

- This defines the gamma function (Γ)

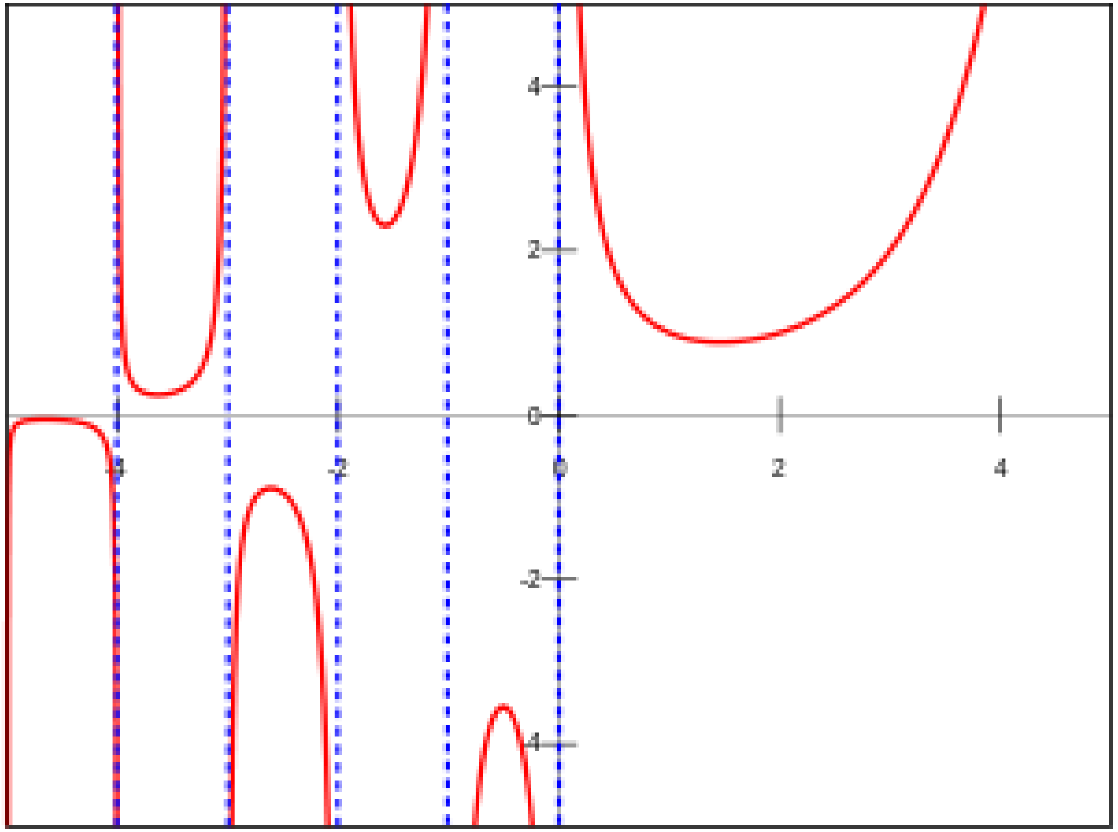
$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt \qquad [1]$$

- It's an extension of the factorial function to real (and complex) numbers.

$$\Gamma(n) = (n-1)!, \ for\ n \in \{1, 2, 3, \ldots\}$$

Plot

Gamma function



[Wikipedia]

Gamma Function

- Some properties of the gamma function

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha),\ \alpha > 0$$

$$\Gamma(n) = (n - 1)!, for\ n \in \{1, 2, 3, \ldots\}$$

$$\Gamma(1) = 1$$

$$\Gamma(1/2) = \sqrt{\pi}$$

Recursion

- The recursive nature of those relationships:
  - Allows us to calculate the value of the gamma function from knowing only
  - The values of $\Gamma(c)$, $0 < c <= 1$.

pdf

- Since the integrand of [1] is positive, it follows that

$$f(t) = \frac{t^{\alpha-1}e^{-t}}{\Gamma(\alpha)}, 0 < t < \infty$$

is a pdf.

The full Gamma

- The full gamma family has two parameters (α, β)

- Define pdf of random variable X = βT, where β is a positive constant.

$$f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}, 0 < x < \infty, \alpha > 0, \beta > 0$$

- $(f_X(x) = \frac{1}{c} f_T(\frac{x}{c}))$

The full Gamma

$$f(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}, 0 < x < \infty, \alpha > 0, \beta > 0 \qquad [2]$$

- The parameter α is known as the shape parameter.
  - Mostly influences the peaked-ness of the distribution.

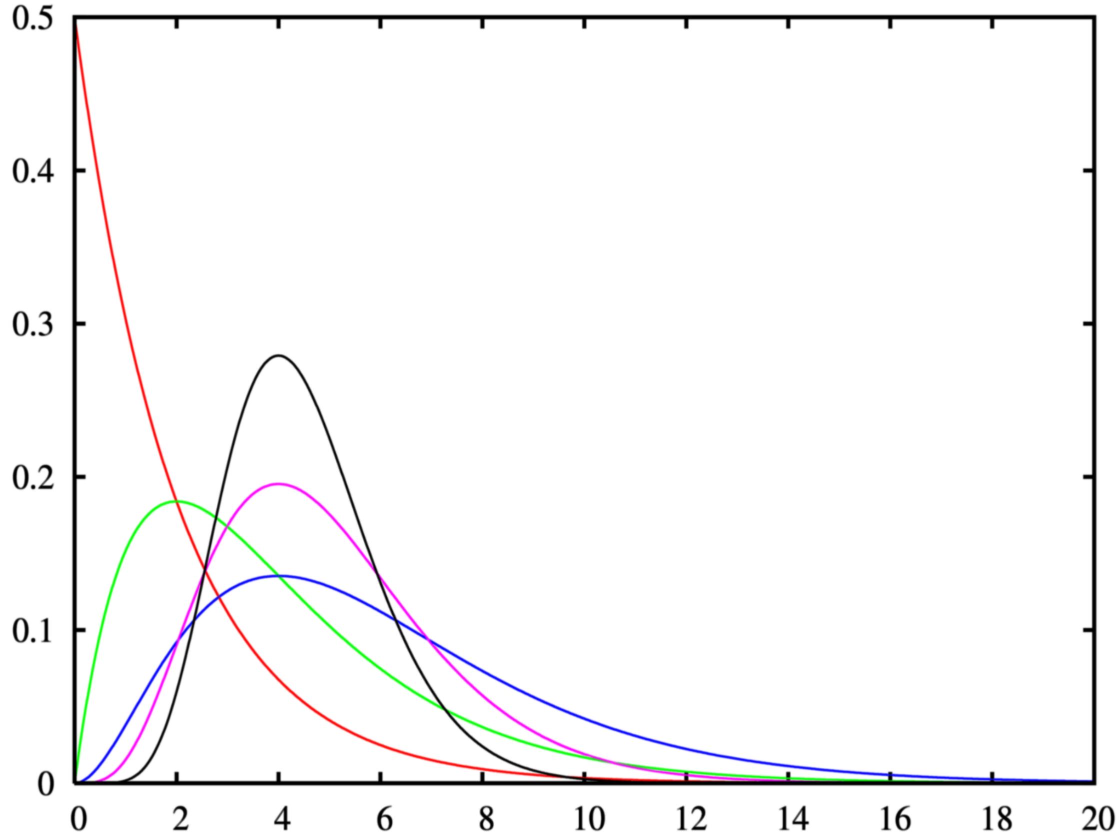
- The parameter β is known as the scale parameter.
  - Mostly influences the spread of the distribution.

An alternative parameterisation

$$f(x|k,\theta) = \frac{\theta^{\alpha}}{\Gamma(\alpha)} x^{k-1} e^{-x\theta}, 0 < x < \infty, k > 0, \theta > 0$$

- The parameter k is known as the shape parameter.
  - Same as α
- The parameter θ is known as the rate parameter.
  - Same as 1/β
- You may see the choice of symbols reversed in some places!

Plot



[Wikipedia]

Mean

- The mean of the gamma(α, β) is

$$EX = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \int_0^{\infty} x x^{\alpha-1} e^{-x/\beta} dx$$

  - To evaluate, notice that the integrand is the kernel of a gamma(α+1, β) pdf.

  - From [2] we know that for any α, β > 0.

$$\int_0^{\infty} x^{\alpha-1} e^{-x/\beta} dx = \Gamma(\alpha)\beta^{\alpha}$$

Mean

$$EX = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \int_0^{\infty} x x^{\alpha-1} e^{-x/\beta} dx$$

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

$$= \frac{\alpha\Gamma(\alpha)\beta}{\Gamma(\alpha)}$$

EX = αβ

Variance, mgf

- Variance

$$\text{Var } X = \alpha\beta^2$$

- Moment Generating Function

$$M_x(t) = \left(\frac{1}{1-\beta t}\right)^{\alpha}, t < \frac{1}{\beta}$$

CDF

- When X follows the standard gamma distribution the its cdf is:

$$F(x|\alpha) = \int_0^x \frac{y^{\alpha-1}e^{-y}}{\Gamma(\alpha)} dy, x > 0$$

- If X ~ Γ(α, β) then:

$$F(x|\ \alpha, \beta) = P(X <= x) = F(x/\beta|\ \alpha)$$

Chi-squared distribution

- A special case of gamma distribution where α = p/2 and β = 2.

$$f(x|p) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{\left(\frac{p}{2}\right)-1} e^{-x/2}, 0 < x < \infty$$

- The mean, variance and mgf can all be calculated using the gamma formula.

- The Chi-squared distribution plays an important role when sampling from a normal distribution.

Beta Distribution

- This family of distributions is continuous on (0,1) index by the parameters α, β.

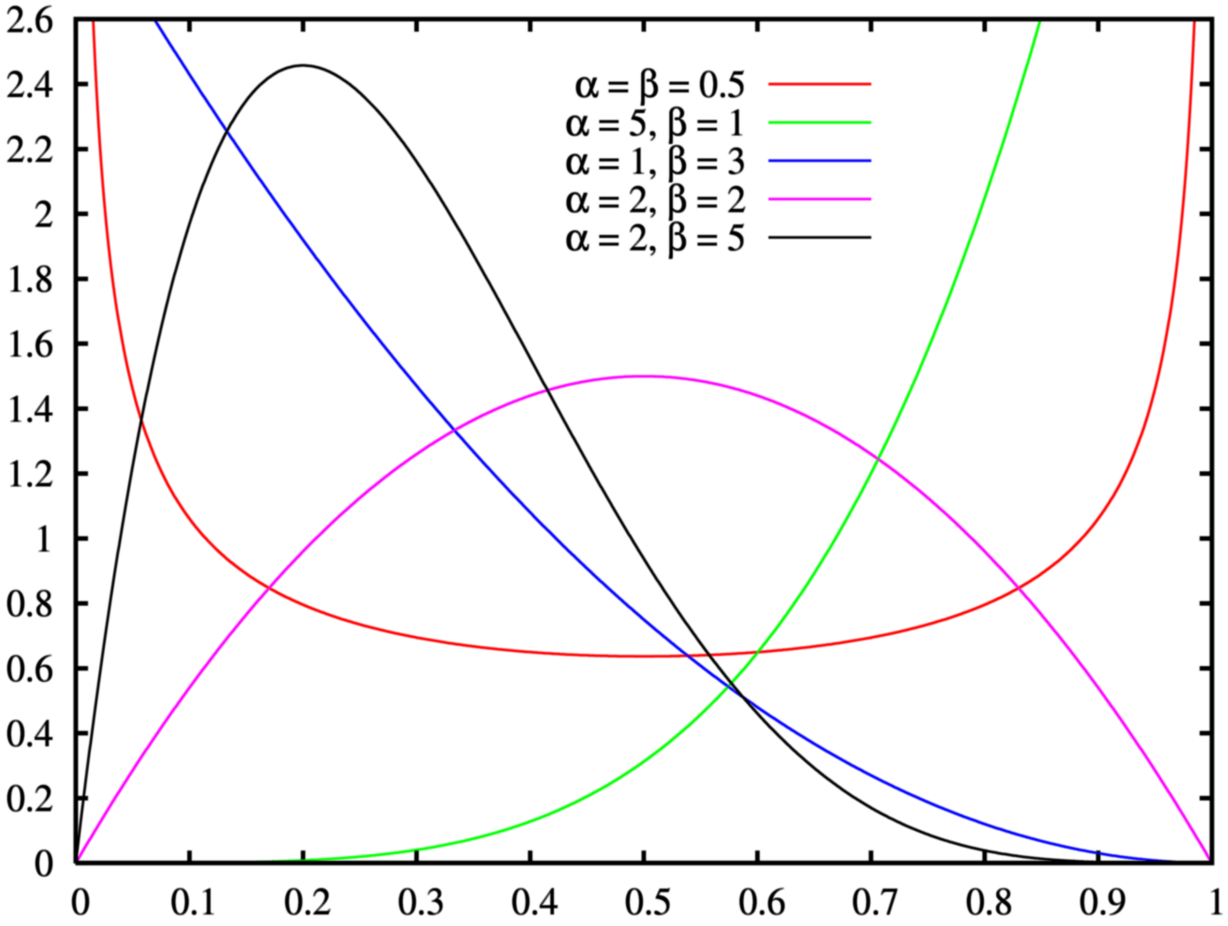
$$f(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0 < x < 1, \alpha > 0, \beta > 0$$

- Where,

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

- We can avoid using the beta function and use the gamma instead.

Plot



[Wikipedia]

Useful feature

- One of the few ‘named’ distributions that:
  - Give probability 1 to a finite interval, here taken to be (0,1).
- It is often used to model proportions
  - These naturally lie between 0 and 1.

Moments

- The form of the pdf makes the calculations of its moments straightforward.

$$EX^n = \frac{1}{B(\alpha,\beta)} \int_0^1 x^n x^{\alpha-1}(1-x)^{\beta-1}dx$$

$$= \frac{1}{B(\alpha,\beta)} \int_0^1 x^{\alpha+n-1}(1-x)^{\beta-1}dx$$

$$= \frac{B(\alpha+n,\beta)}{B(\alpha,\beta)}$$

Mean and Variance

$$EX^n = \frac{B(\alpha+n,\beta)}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+n)\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+n)\Gamma(\alpha)}$$

- With n = 1 and n=2:

$$EX = \frac{\alpha}{\alpha+\beta}$$

$$Var\ X = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Sum of random variables

- Let X be a discrete random variable with pmf $p_X(x)$ and support R_x
- Let Y be a discrete random variable with pmf $p_Y(y)$ support R_y
- Let Z = X + Y
- The following holds:

$$p_Z(z) = \sum_{y \in R_Y} p_X(z-y)p_Y(y)$$

$$p_Z(z) = \sum_{x \in R_X} p_Y(z-x)p_X(x)$$

These two summations called convolutions of two probability mass functions.

Sum of n independent random variables

- The distribution of the sum of a pair of continuous random variables can be dealt with in a similar way.

- We can extend this to n random variables by recursion.

$$Z = X_1 + \ldots + X_n.$$

$$Y_2 = X_1 + X_2$$

$$Y_3 = Y_2 + X_3$$

$$Z = Y_n = Y_{n-1} + X_n$$

Central Limit Theorem

- The CLT established that (in most situations) when independent random variables are added:
  - Their properly normalised sum tends towards a normal distribution,
  - Even if the original variables themselves are not normally distributed.
- This is useful because:
  - It means methods that work for normal distributions can be applied to many problems involving other types of distributions.