

# Special Topics in Comp Bio

November 1, 2006

Stanley Sawyer, Department of Mathematics, WashU

The gamma distribution:

This is a distribution for  $x \geq 0$  with density

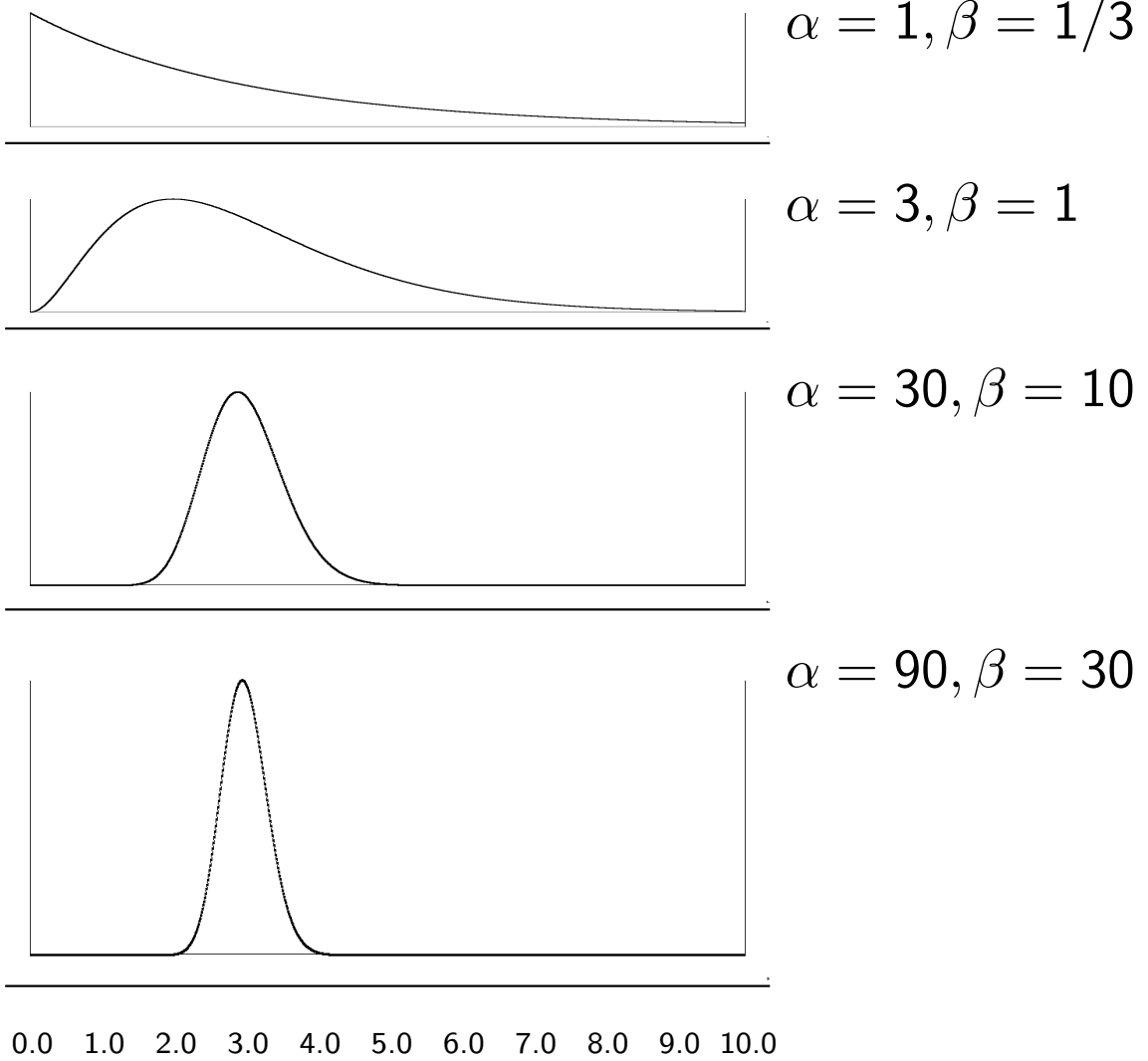
$$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \quad 0 \leq x < \infty$$

Here  $\alpha, \beta > 0$  and  $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} \exp(-\beta y) dy$ .

We say  $X \approx \text{Gam}(\alpha, \beta)$  if a random variable  $X$  has this density. For  $\alpha = 1$ ,  $\text{Gam}(1, \beta)$  is the exponential distribution with rate  $\beta$ :

$$\beta \exp(-\beta x) \quad 0 \leq x < \infty$$

Some example densities:



If  $X \approx \text{Gam}(\alpha, \beta)$ ,

$$E(X) = \alpha/\beta, \quad \text{Var}(X) = \alpha/\beta^2$$

In general

$$\text{Gam}(\alpha, \beta) \approx (1/\beta) \text{Gam}(\alpha, 1)$$

(that is,  $\beta$  is a rate parameter).

Gamma distributions also can be parametrized by

$$X_v = \text{Gam}\left(\frac{1}{v}, \frac{1}{v}\right), \quad Y = \theta X_v$$

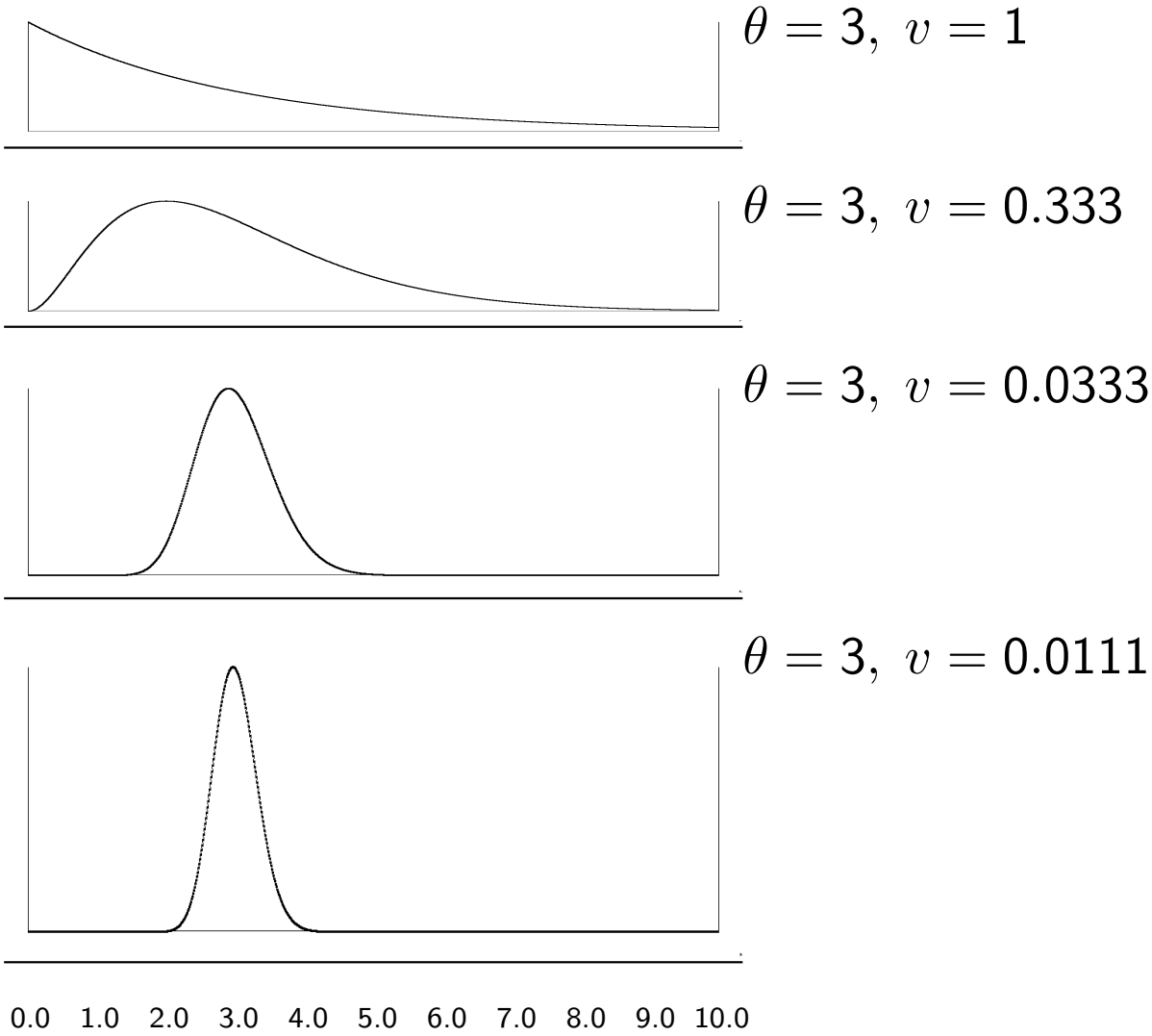
Then

$$E(X_v) = 1, \quad \text{Var}(X_v) = v,$$

$$E(Y) = \theta, \quad \text{Var}(Y) = \theta^2 v$$

This allows modeling arbitrary random  $X > 0$  in terms of  $E(X)$  and  $\text{Var}(X)$ :

The same densities in  $\theta$  and  $v$  coordinates:



Some other important properties of gamma variables:

(i) If  $X_1 \approx \text{Gam}(\alpha_1, \beta)$  and  $X_2 \approx \text{Gam}(\alpha_2, \beta)$  and  $X_1$  and  $X_2$  are independent, then

$$X_1 + X_2 \approx \text{Gam}(\alpha_1 + \alpha_2, \beta)$$

That means that we can view  $T_k = \text{Gam}(k, \beta)$  as the waiting time for  $k$  independent events, where the  $k$  events must occur in sequence and each has an exponential waiting time  $\text{Gam}(1, \beta)$ .

The resulting distribution

$$\text{Gam}(k, \beta) = \frac{\beta^k}{(k-1)!} x^{k-1} \exp(-\beta x)$$

is called the *Erlang* distribution in queueing theory.

(ii) If  $z \approx N(0, 1)$ , then  $z^2$  is  $\text{Gam}(1/2, 1/2)$ . Thus

$$\chi_n^2 \approx z_1^2 + z_2^2 + \dots + z_n^2 \approx \text{Gam}(n/2, 1/2)$$

This means that chi-square distributions in statistics are special cases of gamma distributions.

(iii) An interesting use of gamma distributions is Fisher's method of combining the results of different experiments. (Nowadays this would be called "meta-analysis".)

Suppose that you conducted four different experiments and concluded that none were significant, with P-values

$$P_1 = 0.07, \quad P_2 = 0.18, \quad P_3 = 0.09, \quad P_4 = 0.14$$

Taken together, are these enough to conclude significance, assuming that you are not able to combine all the data and analyze them together? What is the combined P-value?

Fishers idea is as follows. The first step is to combine the four P-values into a single score, for which one can assign a single P-value. A natural choice is

$$T = P_1 P_2 P_3 P_4$$

Then  $T_{\text{obs}} = (0.07)(0.18)(0.09)(0.14) = 0.0001430$

Is this significantly small, given that it is the product of P-values for 4 experiments?

The key idea is that, if a null hypothesis is true, then the P-value itself is uniformly distributed in  $(0, 1)$ . Then

$$P(-\log(P_i) \geq t) = P(P_i \leq e^{-t}) = e^{-t}$$

This means that each  $-\log(P_i) \approx \text{Gam}(1, 1)$  given  $H_0$ . Then given  $H_0$

$$-\log(T) = -\sum_{i=1}^4 \log(P_i) \approx \text{Gam}(4, 1)$$

In Fisher's day, there were tables of  $\chi^2$  P-values but no computers or statistical calculators. However

$$\text{Gam}(4, 1) \approx (1/2) \text{Gam}(8/2, 1/2) \approx (1/2)\chi_8^2$$

Hence the overall P-value is

$$P = \Pr(\chi_8^2 \geq -2 \log(T)) = \Pr(\chi_8^2 \geq 17.71) = 0.024$$

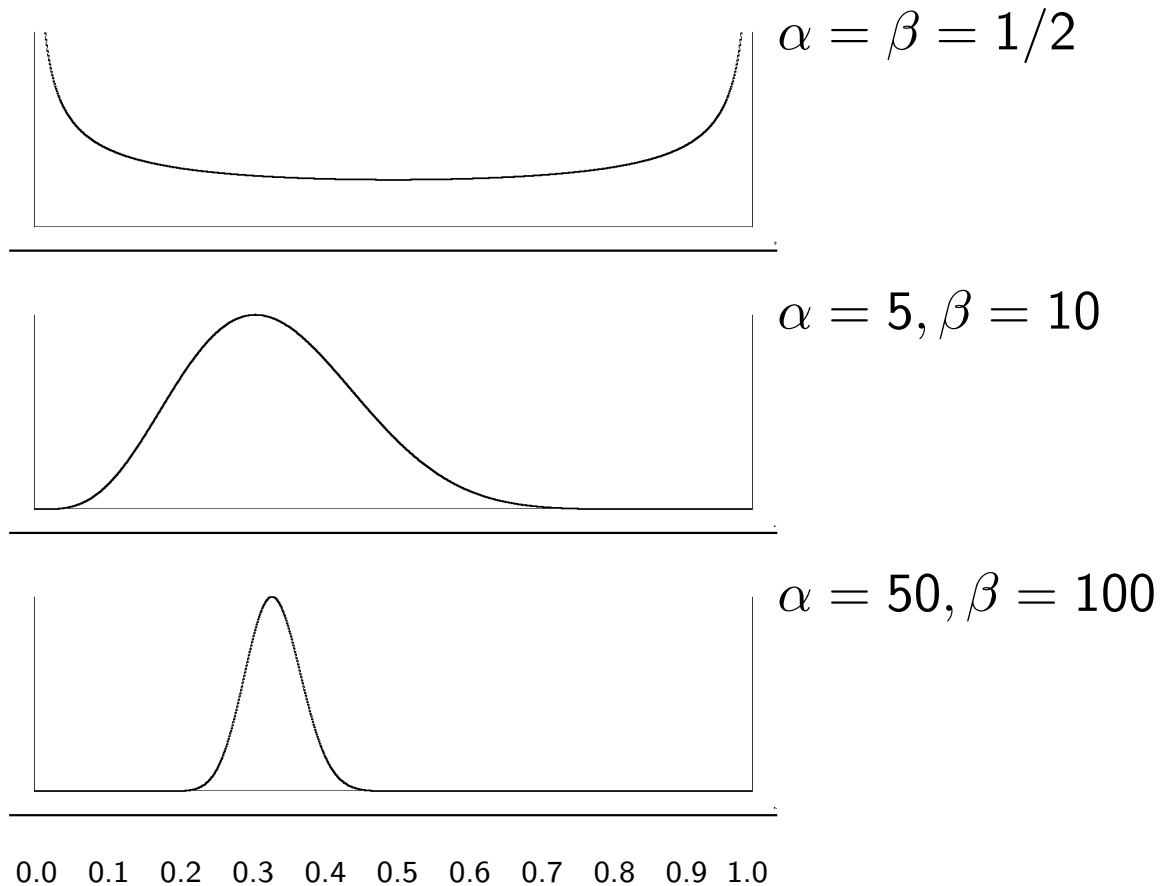
Thus the combined effect of the four experiments is significant.



**The beta distribution:** This is a distribution with density

$$C x^{\alpha-1} (1-x)^{\beta-1}, \quad 0 \leq x \leq 1$$

where  $C = \Gamma(\alpha + \beta) / (\Gamma(\alpha)\Gamma(\beta))$ . Some examples are:



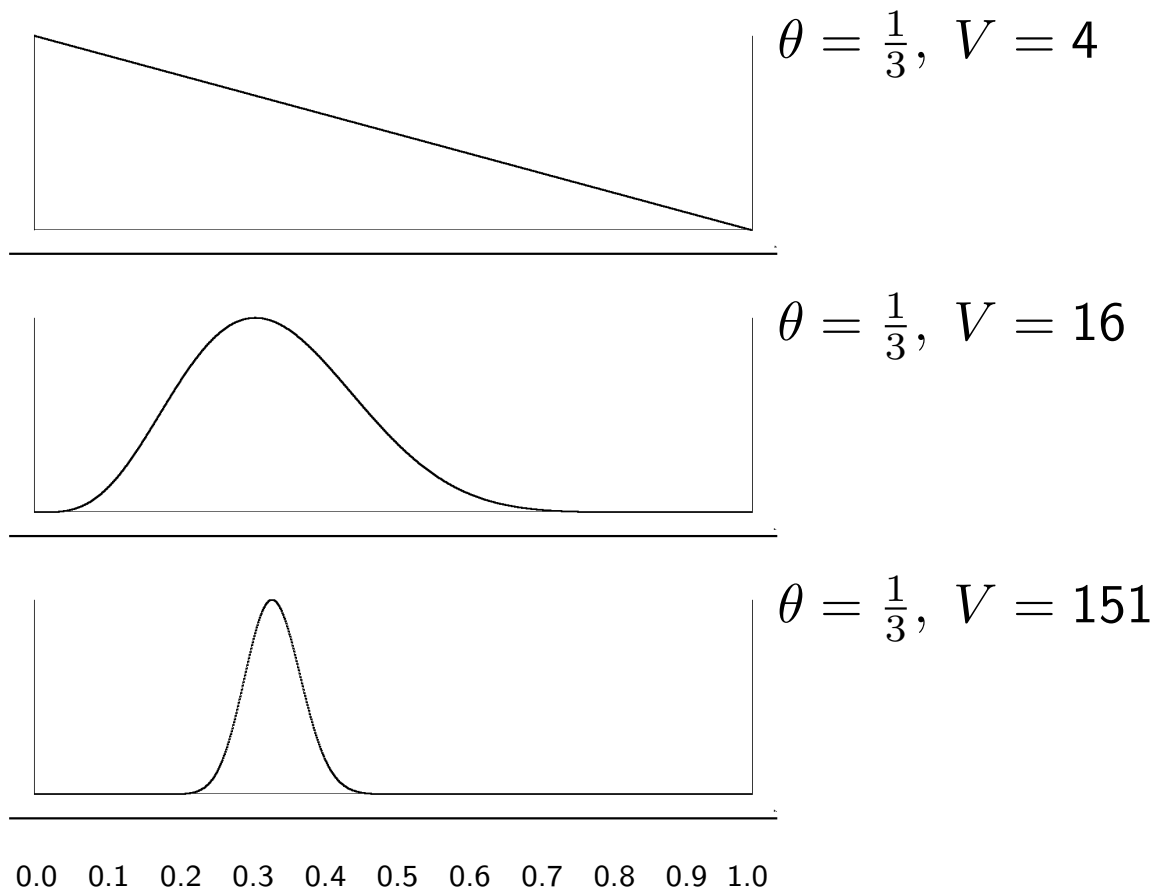
We say  $X \approx \text{Beta}(\alpha, \beta)$  if  $X$  has this density. Then  $\text{Beta}(1, 1)$  is uniform and

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

If  $\theta = \alpha/(\alpha + \beta)$  and  $V = \alpha + \beta + 1$ , then

$$E(X) = \theta, \quad \text{Var}(X) = \frac{\theta(1 - \theta)}{V}$$

The first distribution below has  $\alpha = 1$  and  $\beta = 2$ , so that  $f(x) = Cx^{\alpha-1}(1-x)^{\beta-1} = C(1-x)$ . In general, some beta densities in  $\theta$  and  $V$  coordinates are



One can show that if  $Z \approx \text{Beta}(\alpha, \beta)$ , then

$$Z \approx \frac{X_1}{X_1 + X_2}$$

where  $X_1 \approx \text{Gam}(\alpha, r)$ ,  $X_2 \approx \text{Gam}(\beta, r)$ , and  $X_1$  and  $X_2$  are independent. This implies

$$\frac{Z}{1-Z} \approx \frac{X_1}{X_2} \approx \frac{\text{Gam}(\alpha, r)}{\text{Gam}(\beta, r)} \approx \frac{\chi^2(2\alpha)}{\chi^2(2\beta)}$$

Thus if  $Z \approx \text{Beta}(\alpha, \beta)$

$$\frac{Z}{1-Z} \approx \frac{\alpha}{\beta} \frac{\chi^2(2\alpha)/2\alpha}{\chi^2(2\beta)/2\beta} \approx \frac{\alpha}{\beta} F(2\alpha, 2\beta)$$

so that  $Z \approx \text{Beta}(\alpha, \beta)$  can be written in terms of an  $F$ -distribution and vice versa. This is in fact how  $F$ -distribution  $P$ -values are calculated in many statistical packages, since the  $F$ -distribution density itself has polynomial decay at infinity.

The beta density can be thought of as

$$f(x) = C x_1^{\alpha-1} x_2^{\beta-1}$$

where  $(x_1, x_2)$  are on the line  $x_1 + x_2 = 1$  for  $x_1 \geq 0, x_2 \geq 0$ . This is an equivalent way of looking at a beta density, as long as you are careful about how you are doing the integration.

**The Dirichlet distribution:** Once one gets used to this, one can generalize the beta density to more than two variables: For example, with a three-dimensional density

$$f(x) = C x_1^{\alpha-1} x_2^{\beta-1} x_3^{\gamma-1} x_4^{\delta-1}$$

on the simplex  $x_1 + x_2 + x_3 + x_4 = 1, x_i \geq 0$ , for parameters  $\alpha, \beta, \gamma, \delta > 0$ . Here

$$C = \frac{\Gamma(\alpha + \beta + \gamma + \delta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)\Gamma(\delta)}$$

This is called a *Dirichlet density* and has very similar properties to a beta density. For example if  $X_1, X_2, X_3, X_4$  have the Dirichlet density

$$f(x) = C x_1^{\alpha_1-1} x_2^{\alpha_2-1} x_3^{\alpha_3-1} x_4^{\alpha_4-1}$$

where  $X_1 + X_2 + X_3 + X_4 = 1$ , then each  $X_i$  is  $\text{Beta}(\alpha_i, \alpha - \alpha_i)$  for  $\alpha = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4$ , and

$$E(X_i) = \frac{\alpha_i}{\alpha} \quad \text{Var}(X_i) = \frac{\alpha_i(\alpha - \alpha_i)}{\alpha^2(\alpha + 1)}$$

$$\text{Cov}(X_i, X_j) = -\frac{\alpha_i \alpha_j}{\alpha^2(\alpha + 1)} \quad \text{if } i \neq j$$

If random variables  $X_1, X_2, X_3, X_4$  have the above Dirichlet distribution, then the  $X_i$  can be represented

$$X_i \approx \frac{Y_i}{Y_1 + Y_2 + Y_3 + Y_4} \quad (1 \leq i \leq 4)$$

where the  $Y_i \approx \text{Gam}(\alpha_i, r)$  are independent gamma-distributed random variables where.