Section 4.8 Beta distribution

Thursday, October 27, 2016 9:37 AM

Section 4.8
Beta distri...

4.8 Beta distribution

TERMINOLOGY: A random variable Y is said to have a **beta distribution** with parameters $\alpha > 0$ and $\beta > 0$ if its pdf is given by

$$f_Y(y) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

$\leftarrow$ pdf.

Since the support of Y is $R = \{y : 0 < y < 1\}$, the beta distribution is a popular probability model for proportions. Shorthand notation is $Y \sim \text{beta}(\alpha, \beta)$.

NOTE: Upon closer inspection, we see that the nonzero part of the beta(α, β) pdf

$$f_Y(y) = \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}}_{\text{Constant}} \underbrace{y^{\alpha-1}(1-y)^{\beta-1}}_{\text{Kernel}}$$

consists of two parts:

- the **kernel** of the pdf: $y^{\alpha-1}(1-y)^{\beta-1}$
- a **constant** out front: $\Gamma(\alpha+\beta)/\Gamma(\alpha)\Gamma(\beta)$.

Again, the kernel is the "guts" of the formula, while the constant out front is simply the "right quantity" that makes $f_Y(y)$ a valid pdf; i.e., the constant which makes $f_Y(y)$ integrate to 1. Note that because

$$\int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} dy = 1,$$ ✓

it follows immediately that

$$\int_0^1 y^{\alpha-1}(1-y)^{\beta-1} dy = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

BETA PDF SHAPES: The beta pdf is very flexible. That is, by changing the values of α and β, we can come up with many different pdf shapes. See Figure 4.13 for examples.

- When $\alpha = \beta$, the pdf is **symmetric** about the line $y = \frac{1}{2}$.
- When $\alpha < \beta$, the pdf is **skewed right** (i.e., smaller values of y are more likely).

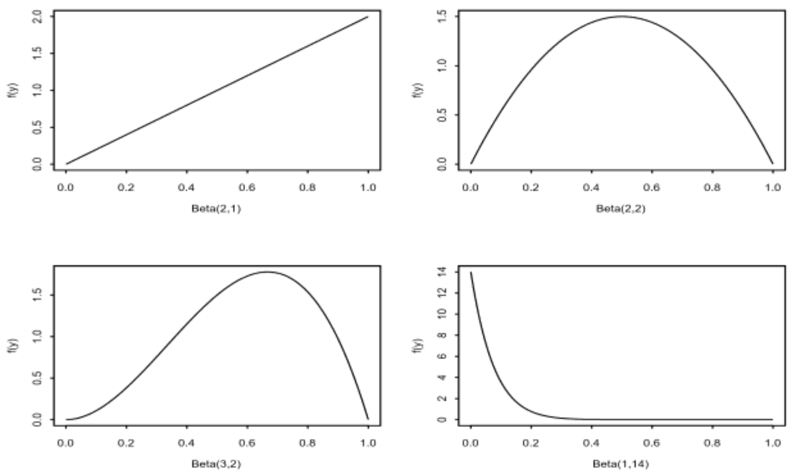

Figure 4.13: *Four beta pdfs. Upper left:* $\alpha = 2$, $\beta = 1$. *Upper right:* $\alpha = 2$, $\beta = 2$. *Lower left:* $\alpha = 3$, $\beta = 2$. *Lower right:* $\alpha = 1$, $\beta = 14$.

- When $\alpha > \beta$, the pdf is **skewed left** (i.e., larger values of y are more likely).
- When $\alpha = \beta = 1$, the beta pdf reduces to the $\mathcal{U}(0, 1)$ pdf!

$y^{\alpha-1}(1-y)^{\beta-1}$ $\alpha=\beta=1$

$y^0(1-y)^0 = 1$

BETA MGF: The beta(α, β) mgf exists, but not in closed form. Hence, we'll compute moments directly.

MEAN AND VARIANCE: If $Y \sim \text{beta}(\alpha, \beta)$, then

$$E(Y) = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad V(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Proof. We will derive $E(Y)$ only. From the definition of expected value, we have

$$\begin{aligned} E(Y) = \int_0^1 y f_Y(y)dy &= \int_0^1 y \left[\boxed{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}} y^{\alpha-1}(1-y)^{\beta-1}\right] dy \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \underbrace{y^{(\alpha+1)-1}(1-y)^{\beta-1}}_{\text{beta}(\alpha+1,\beta) \text{ kernel}} dy. \end{aligned}$$

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

Note that the last integrand is a beta kernel with parameters $\alpha + 1$ and β. Because integration is over $R = \{y : 0 < y < 1\}$, we have

$$\int_0^1 y^{(\alpha+1)-1}(1-y)^{\beta-1} = \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)}$$

and thus

$$\begin{aligned} E(Y) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cancel{\Gamma(\beta)}}\frac{\Gamma(\alpha+1)\cancel{\Gamma(\beta)}}{\Gamma(\alpha+1+\beta)} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)}\frac{\Gamma(\alpha+1)}{\Gamma(\alpha+1+\beta)} \\ &= \frac{\cancel{\Gamma(\alpha+\beta)}}{\cancel{\Gamma(\alpha)}}\frac{\alpha\cancel{\Gamma(\alpha)}}{(\alpha+\beta)\cancel{\Gamma(\alpha+\beta)}} = \frac{\alpha}{\alpha+\beta}. \end{aligned}$$

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

$\Gamma(\alpha+1+\beta) = (\alpha+\beta)\Gamma(\alpha+\beta)$

To derive $V(Y)$, first find $E(Y^2)$ using similar calculations. Use the variance computing formula $V(Y) = E(Y^2) - [E(Y)]^2$ and simplify. $\square$

Example 4.17. At a health clinic, suppose that Y, the proportion of individuals infected with a new flu virus (e.g., H1N1, etc.), varies daily according to a beta distribution with pdf

$$f_Y(y) = \begin{cases} 20(1-y)^{19}, & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

$$\left.\begin{array}{l} y^0(1-y)^{19} \\ y^{\alpha-1}(1-y)^{\beta-1} \end{array}\right\} \text{match}$$

$\alpha - 1 = 0$

$\beta - 1 = 19$

$\Rightarrow \alpha = 1$

$\beta = 20$

$Y \sim \text{Beta}(\alpha=1, \beta=20)$

This distribution is displayed in Figure 4.14.

QUESTIONS.

(a) What are the parameters in this distribution; i.e., what are α and β?

(b) What is the mean proportion of individuals infected?

(c) Find $\phi_{0.95}$, the 95th percentile of this distribution.

(d) Treating daily infection counts as independent (from day to day), what is the probability that during any given 5-day span, there is are at least 2 days where the infection proportion is above 10 percent?

_ _ _ _ _

# of days out 5

SOLUTIONS.

(a) $\alpha = 1$ and $\beta = 20$.

(b) $E(Y) = 1/(1+20) \approx 0.048$.

(c) The 95th percentile $\phi_{0.95}$ solves

$$P(Y \leq \phi_{0.95}) = \int_0^{\phi_{0.95}} 20(1-y)^{19}dy = 0.95.$$

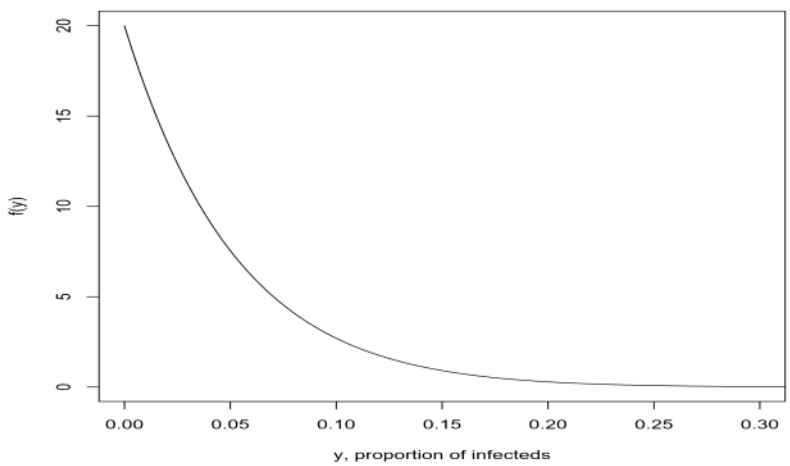

Figure 4.14: *The probability density function, $f_Y(y)$, in Example 4.17. A model for the proportion of infected individuals.*

Let $u = 1 - y$ so that $du = -dy$. The limits on the integral must change:

$$\int_0^{\phi_{.95}} 20(1-y)^{19} dy$$

$$\Downarrow$$

$$y: \quad 0 \longrightarrow \phi_{0.95}$$

$$u: \quad 1 \longrightarrow 1 - \phi_{0.95}$$

$$-\int_1^{1-\phi_{.95}} 20 \times u^{19} du$$

$$=$$

Thus, we are left to solve

$$0.95 = -\int_1^{1-\phi_{0.95}} 20u^{19} du = u^{20}\Big|_{1-\phi_{0.95}}^{1} = 1 - (1 - \phi_{0.95})^{20}$$

$$\int_{1-\phi_{.95}}^{1} 20u^{19} du$$

for $\phi_{0.95}$. We get

$$\phi_{0.95} = 1 - (0.05)^{1/20} \approx 0.139.$$

(d) First, we compute

$u > 1 - y$

$$P(Y > 0.1) = \int_{0.1}^{1} 20(1-y)^{19} dy = \int_0^{0.9} 20u^{19} du = u^{20}\Big|_0^{0.9} = (0.9)^{20} \approx 0.122.$$

This is the probability that the infection proportion exceeds 0.10 on any given day. Now, we treat each day as a "trial," and let X denote the number of days where "the

$n=5,\ p=.122$

infection proportion is above 10 percent" (i.e., a "success"). Because days are assumed independent, $X \sim b(5, 0.122)$ and

$1-P(X<2) = 1-P(X\leq 1) = 1-\text{binomcdf}(\ \ )$

$$\begin{aligned} P(X \geq 2) &= 1 - P(X=0) - P(X=1) \\ &= 1 - \binom{5}{0}(0.122)^0(1-0.122)^5 - \binom{5}{1}(0.122)^1(1-0.122)^4 \approx 0.116. \ \square \end{aligned}$$

4.9 Chebyshev's Inequality

MARKOV'S INEQUALITY: Suppose that X is a nonnegative random variable with pdf (pmf) $f_X(x)$ and let c be a positive constant. Markov's Inequality puts a bound on the upper tail probability $P(X > c)$; that is,

$$P(X > c) \leq \frac{E(X)}{c}.$$

Proof. First, define the event $B = \{x : x > c\}$. We know that

$$\begin{aligned} E(X) = \int_0^\infty x f_X(x)dx &= \int_B x f_X(x)dx + \int_{\overline{B}} x f_X(x)dx \\ &\geq \int_B x f_X(x)dx \\ &\geq \int_B c f_X(x)dx = cP(X > c). \ \square \end{aligned}$$

CHEBYSHEV'S INEQUALITY: Let Y be any random variable, discrete or continuous, with mean μ and variance $\sigma^2 < \infty$. For $k > 0$,

$$P(|Y - \mu| > k\sigma) \leq \frac{1}{k^2}.$$

Proof. Applying Markov's Inequality with $X = (Y-\mu)^2$ and $c = k^2\sigma^2$, we have

$$P(|Y-\mu| > k\sigma) = P[(Y-\mu)^2 > k^2\sigma^2] \leq \frac{E[(Y-\mu)^2]}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}. \ \square$$

REMARK: The beauty of Chebyshev's result is that it applies to any random variable Y. In words, $P(|Y-\mu| > k\sigma)$ is the probability that the random variable Y will differ from the mean μ by more than k standard deviations. If we do not know how Y is distributed,