Chapter 4: Probability inequalities and theorems

Markov's inequality

Suppose that $g(Y)$ is a function of random variable Y such that $\mathbb{E}[|g(Y)|] < \infty$. For any $k > 0$, it can be shown that

$$P(|g(Y)| \geq k) \leq \frac{\mathbb{E}[|g(Y)|]}{k}.$$

or equivalently

$$P(|g(Y)| < k) \geq 1 - \frac{\mathbb{E}[|g(Y)|]}{k}.$$

This is *Markov's inequality*.

Chebychev's inequality

aka

- Chebyshev's inequality
- Bienaymé-Chebyshev inequality
- Tchebysheff's Theorem
- Chebysheff, Chebychov, Chebyshov, Tchebychev . . .

Chebychev's inequality (cont.)

Suppose that Y has finite mean μ and variance σ^2.

In Markov's inequality, consider the function

$$g(t) = (t - \mu)^2.$$

We have for any $k > 0$ that

$$P(|(Y - \mu)^2| < k) \geq 1 - \frac{\mathbb{E}[|(Y - \mu)^2|]}{k}.$$

which can be more simply written as

$$P((Y - \mu)^2 < k) \geq 1 - \frac{\mathbb{E}[(Y - \mu)^2]}{k}.$$

Chebychev's inequality (cont.)

But

$$\mathbb{E}[(Y-\mu)^2] = \sigma^2$$

and if we choose to write $k = c^2\sigma^2$, the inequality becomes

$$P((Y-\mu)^2 < c^2\sigma^2) \geq 1 - \frac{\mathbb{E}[(Y-\mu)^2]}{c^2\sigma^2} = 1 - \frac{1}{c^2}$$

Chebychev's inequality (cont.)

The left hand side of the inequality is equal to

$$P(\sqrt{(Y-\mu)^2} < \sqrt{c^2\sigma^2}) = P(|Y-\mu| < c\sigma)$$

so the inequality becomes

$$P(|Y-\mu| < c\sigma) \geq 1 - \frac{1}{c^2}.$$

or equivalently

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

This is *Chebychev's inequality*.

Chebychev's inequality (cont.)

This states that for *any* distribution with finite mean μ and variance σ^2, we have that

$$P(\mu - c\sigma < Y < \mu + c\sigma) \geq 1 - \frac{1}{c^2}$$

which bounds how much probability can be present in the tails of the distribution.

The sample mean

Suppose that $Y_1, \ldots, Y_n$ are independent that have the same distribution, with expectation μ and variance σ^2. If

$$U = \sum_{i=1}^{n} Y_i$$

then

$$\mathbb{E}[U] = n\mu \qquad \mathbb{V}[U] = n\sigma^2.$$

Let

$$\overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

be the *sample mean* random variable.

The sample mean (cont.)

Then

$$\mathbb{E}[\overline{Y}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} Y_i\right] = \frac{1}{n} n\mu = \mu$$

and it can be shown that

$$\mathbb{V}[\overline{Y}_n] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} Y_i\right] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

which follows by *independence*.

The sample mean (cont.)

Now recall Chebychev's inequality:

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

Applying this to $\overline{Y}_n$, we have that

$$P\left(|\overline{Y}_n - \mu| \geq \frac{c\sigma}{\sqrt{n}}\right) \leq \frac{1}{c^2}.$$

The sample mean (cont.)

We can rewrite this for any n by setting

$$\epsilon = \frac{c\sigma}{\sqrt{n}} \quad \text{so that} \quad c = \frac{\sqrt{n}\epsilon}{\sigma}.$$

Then the inequality becomes

$$P(|\overline{Y}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Convergence in probability

Consider a sequence of random variables $Y_1, Y_2, \ldots, Y_n, \ldots$, and real constant c. We say that the sequence *converges in probability* to c as $n \longrightarrow \infty$ if, for every $\epsilon > 0$,

$$\lim_{n \longrightarrow \infty} P(|Y_n - c| < \epsilon) = 1$$

that is, the probability

$$P(c - \epsilon < Y_n < c + \epsilon)$$

converges to 1 as $n \longrightarrow \infty$. We write

$$Y_n \xrightarrow{p} c$$

as $n \longrightarrow \infty$ to indicate convergence in probability.

Convergence in probability (cont.)

Note

- If, for a fixed $\epsilon > 0$,

$$p_n = P(|Y_n - c| < \epsilon)$$

the definition requires that the real sequence $\{p_n\}$ converges to 1 in the usual sense.

- There is no restriction on the joint distribution of the Y_is in the sequence (that is, they can be dependent).
- We can equivalently write the condition

$$\lim_{n \longrightarrow \infty} P(|Y_n - c| > \epsilon) = 0.$$

The weak law of large numbers

Suppose that $Y_1, Y_2, \ldots, Y_n,$ are independent with the same distribution, and expectation μ and variance σ^2.

Then

$$\overline{Y}_n \xrightarrow{p} \mu$$

as $n \longrightarrow \infty$.

The weak law of large numbers (cont.)

This follows from Chebychev's inequality: recall that for the sample mean

$$P(|\overline{Y}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

or equivalently

$$P(|\overline{Y}_n - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2}.$$

for any $\epsilon > 0$.

The weak law of large numbers (cont.)

But

$$1 - \frac{\sigma^2}{n\epsilon^2} \longrightarrow 1$$

as $n \longrightarrow \infty$, as σ and ϵ are fixed. Thus

$$\lim_{n \longrightarrow \infty} P(|\overline{Y}_n - \mu| < \epsilon) \geq 1$$

and hence $\overline{Y}_n \xrightarrow{p} \mu$.

The weak law of large numbers (cont.)

What this result means: under these conditions, the distribution of sample mean random variable becomes *degenerate* at expectation of the distribution of the Ys.

- For example, if you repeatedly and independently toss a fair coin, the proportion of heads will converge to 1/2 as the number of tosses increases.

Convergence in distribution

For cdfs, $F_1(y), F_2(y), \ldots, F_n(y), \ldots$, suppose that

$$\lim_{n \longrightarrow \infty} F_n(y) = F(y)$$

for some cdf $F(y)$ at all points where $F(y)$ is *continuous*.

Then $F(y)$ is the *limiting distribution* for the sequence.

Convergence in distribution (cont.)

If $Y_1, Y_2, \ldots, Y_n, \ldots$ are a sequence of random variables having cdfs $F_1(y), F_2(y), \ldots, F_n(y), \ldots$ respectively, and Y has cdf $F(y)$, then we say that the sequence $Y_1, Y_2, \ldots, Y_n, \ldots$ *converges in distribution* to Y,

$$Y_n \xrightarrow{d} Y$$

as $n \longrightarrow \infty$.

The central limit theorem

Suppose that $Y_1, Y_2, \ldots, Y_n$, are independent, have the same distribution, with finite expectation μ and variance σ^2.

For each n, let

$$\overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

be the sample mean random variable.

The central limit theorem (cont.)

Then the random variable

$$U_n = \frac{\sqrt{n}(\overline{Y}_n - \mu)}{\sigma}$$

converges in distribution to a *standard normal distribution*.

That is,

$$\lim_{n \longrightarrow \infty} F_{U_n}(u) = \lim_{n \longrightarrow \infty} P(U_n \leq u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$$

This result is the *Central Limit Theorem* (CLT).

The central limit theorem (cont.)

Note that we could write

$$U_n = \frac{(\overline{Y}_n - \mu)}{\sigma/\sqrt{n}}$$

or, multiplying numerator and denominator by n

$$U_n = \frac{\left(\sum_{i=1}^{n} Y_i - n\mu\right)}{\sqrt{n}\sigma}$$

The central limit theorem (cont.)

The result says that, provided the conditions on expectation and variance are met, when n is large, we can approximate the distribution of $\overline{Y}_n$, or of

$$S_n = \sum_{i=1}^{n} Y_i$$

using a Normal distribution, irrespective of the actual distribution of $Y_1, \ldots, Y_n$.

The central limit theorem (cont.)

That is,

- $\overline{Y}_n$ is approximately distributed as

$$Normal(\mu, \sigma^2/n)$$

- S_n is approximately distributed as

$$Normal(n\mu, n\sigma^2)$$

The central limit theorem (cont.)

Note

If $Y_1, \ldots, Y_n$ are independent and identically distributed Normal random variables with expectation μ and variance σ^2, then

- $\overline{Y}_n$ is exactly distributed as

$$Normal(\mu, \sigma^2/n)$$

- S_n is exactly distributed as

$$Normal(n\mu, n\sigma^2)$$

Applications of the central limit theorem

We can use the central limit theorem to approximate some standard distributions for which it is otherwise awkward to compute probabilities.

Suppose we wish to approximate the distribution of random variable Y. We must first attempt to decompose Y as

$$Y = \sum_{i=1}^{n} Y_i$$

where $Y_1, \ldots, Y_n$ are independent and identically distributed, and have expectation μ and variance σ^2 that can be easily computed.

Applications of the central limit theorem (cont.)

We then can say that Y is approximately distributed as

$$Normal(n\mu, n\sigma^2).$$

and compute, for example

$$F_Y(y) = \Pr(Y \leq y) = P\left(\frac{Y - n\mu}{\sqrt{n}\sigma} \leq \frac{y - n\mu}{\sqrt{n}\sigma}\right) \simeq \Phi\left(\frac{y - n\mu}{\sqrt{n}\sigma}\right)$$

where $\Phi(.)$ is the standard Normal cdf.

Applications of the central limit theorem (cont.)

Example (Binomial case)

If $Y \sim Binomial(n, p)$, the mgf of Y takes the form

$$m_Y(t) = (pe^t + q)^n$$

By the uniqueness of mgfs, we recognize that Y has the representation

$$Y = \sum_{i=1}^{n} Y_i$$

where $Y_i \sim Binomial(1, p) \equiv Bernoulli(p)$ are independent. We have

$$\mathbb{E}[Y_i] = p \qquad \mathbb{V}[Y_i] = pq.$$

Applications of the central limit theorem (cont.)

Example (Binomial case)

Therefore for large n, Y is approximately distributed as

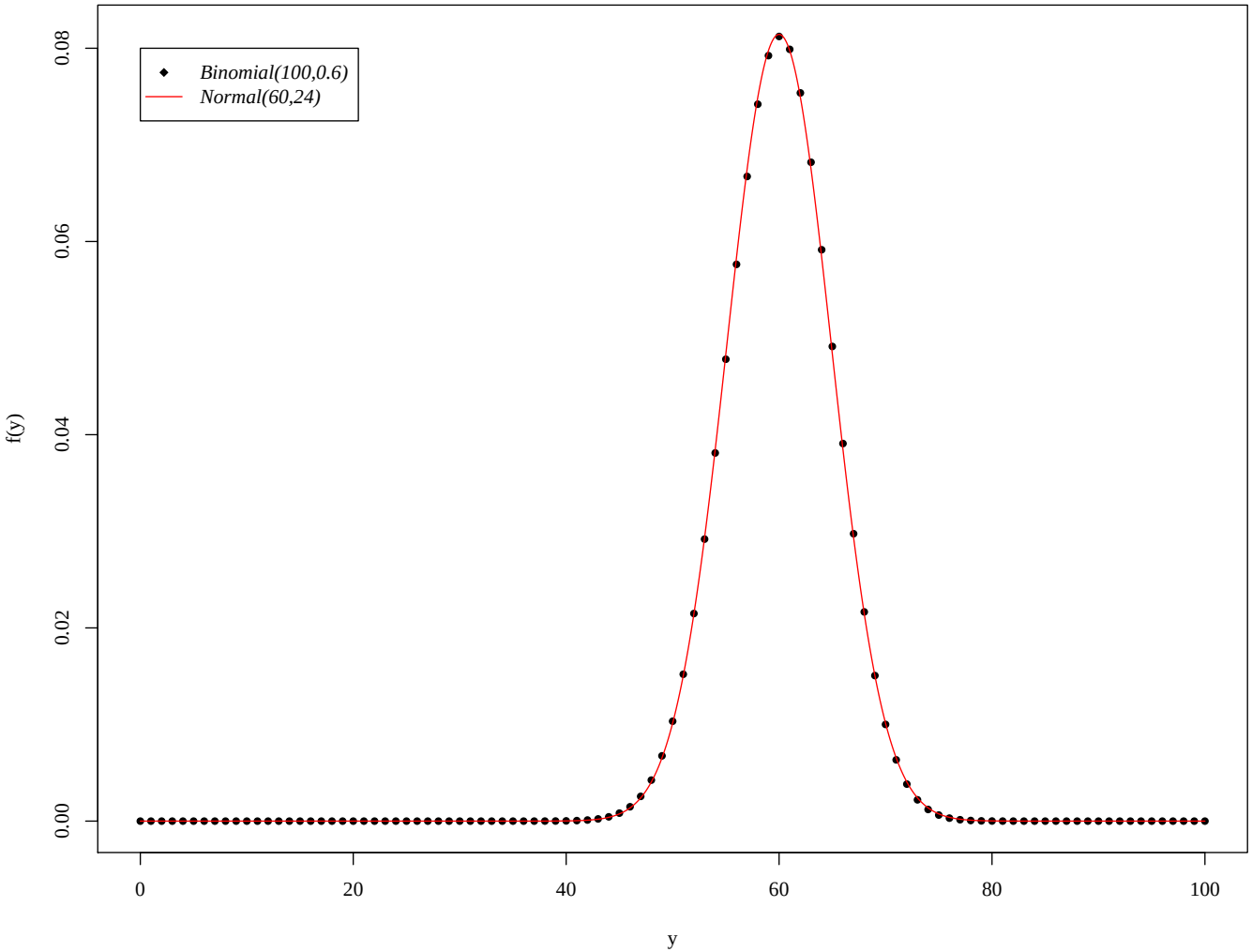
$$Normal(np, npq)$$

and

$$F_Y(y) = P(Y \leq y) \simeq \Phi\left(\frac{y - np}{\sqrt{npq}}\right)$$

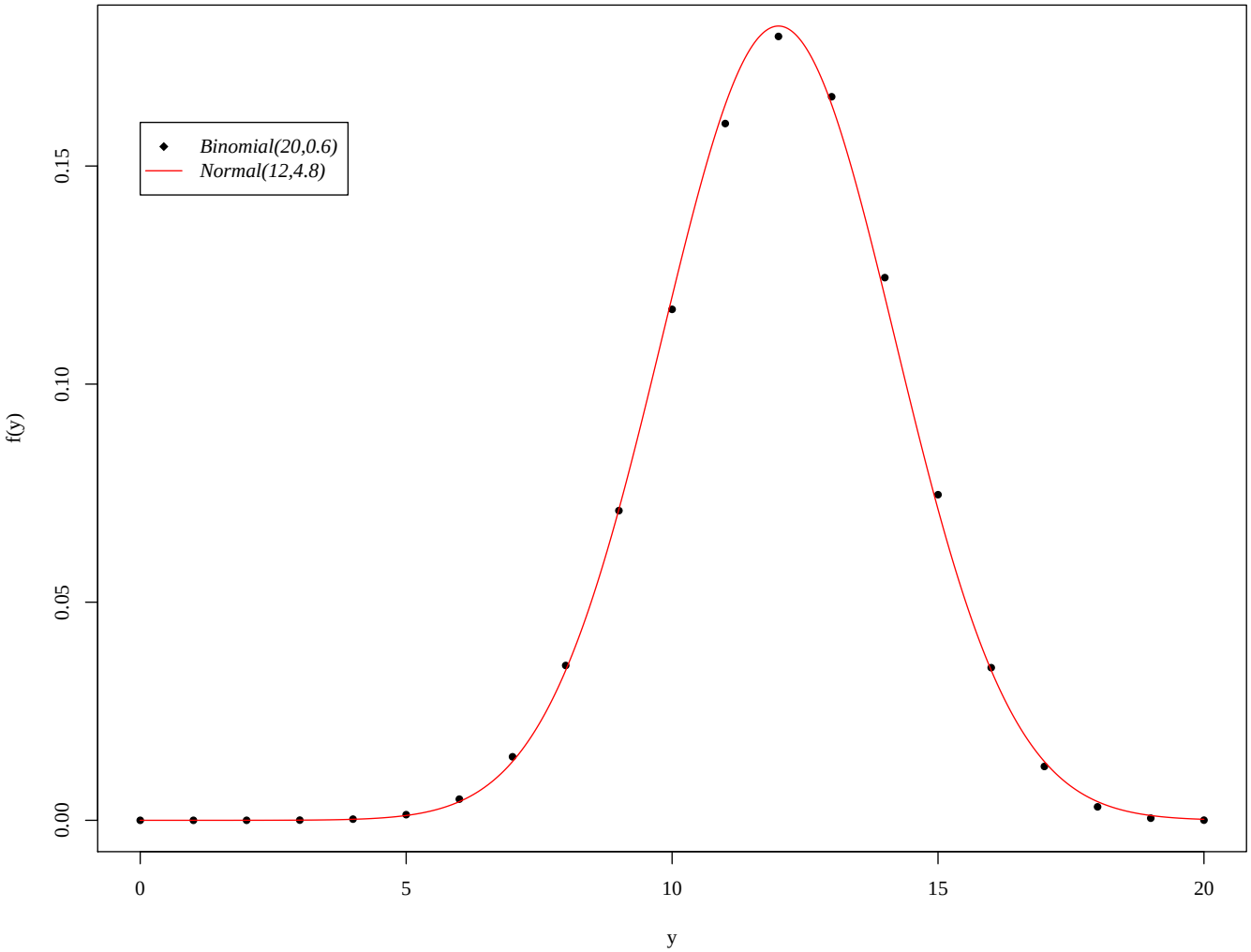
Applications of the central limit theorem (cont.)

Binomial: $n = 100$, with $p = 0.6$.



Applications of the central limit theorem (cont.)

Binomial: $n = 20$, with $p = 0.6$.



Applications of the central limit theorem (cont.)

Example (Poisson case)

If $Y \sim Poisson(\lambda)$, the mgf of Y takes the form

$$m_Y(t) = \exp\{\lambda(e^t - 1)\}.$$

By the uniqueness of mgfs, we recognize that Y has the representation

$$Y = \sum_{i=1}^{n} Y_i$$

where the $Y_i \sim Poisson(\lambda/n)$ are independent. We have

$$\mathbb{E}[Y_i] = \frac{\lambda}{n} \qquad \mathbb{V}[Y_i] = \frac{\lambda}{n}.$$

Applications of the central limit theorem (cont.)

Example (Poisson case)

Therefore Y is approximately distributed as

$$Normal(\lambda, \lambda)$$

and

$$F_Y(y) = P(Y \leq y) \simeq \Phi\left(\frac{y-\lambda}{\sqrt{\lambda}}\right)$$

Note that for this approximation to work, we need to have λ large.

Applications of the central limit theorem (cont.)

Continuity correction

When we use the central limit theorem to approximate a discrete distribution, a *continuity correction* is sometimes applied: that is, if Y is discrete we compute

$$F_Y(y) = P(Y \leq y) \simeq \Phi\left(\frac{y + \frac{1}{2} - n\mu}{\sqrt{n}\sigma}\right)$$

Applications of the central limit theorem (cont.)

Example (Gamma case)

If $Y \sim Gamma(\alpha, \beta)$, the mgf of Y takes the form

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

By the uniqueness of mgfs, we recognize that Y has the representation

$$Y = \sum_{i=1}^{n} Y_i$$

where the $Y_i \sim Gamma(\alpha/n, \beta)$ are independent. We have

$$\mathbb{E}[Y_i] = \frac{\alpha\beta}{n} \qquad \mathbb{V}[Y_i] = \frac{\alpha\beta^2}{n}.$$

Applications of the central limit theorem (cont.)

Example (Gamma case)

Therefore Y is approximately distributed as

$$Normal(\alpha\beta, \alpha\beta^2)$$

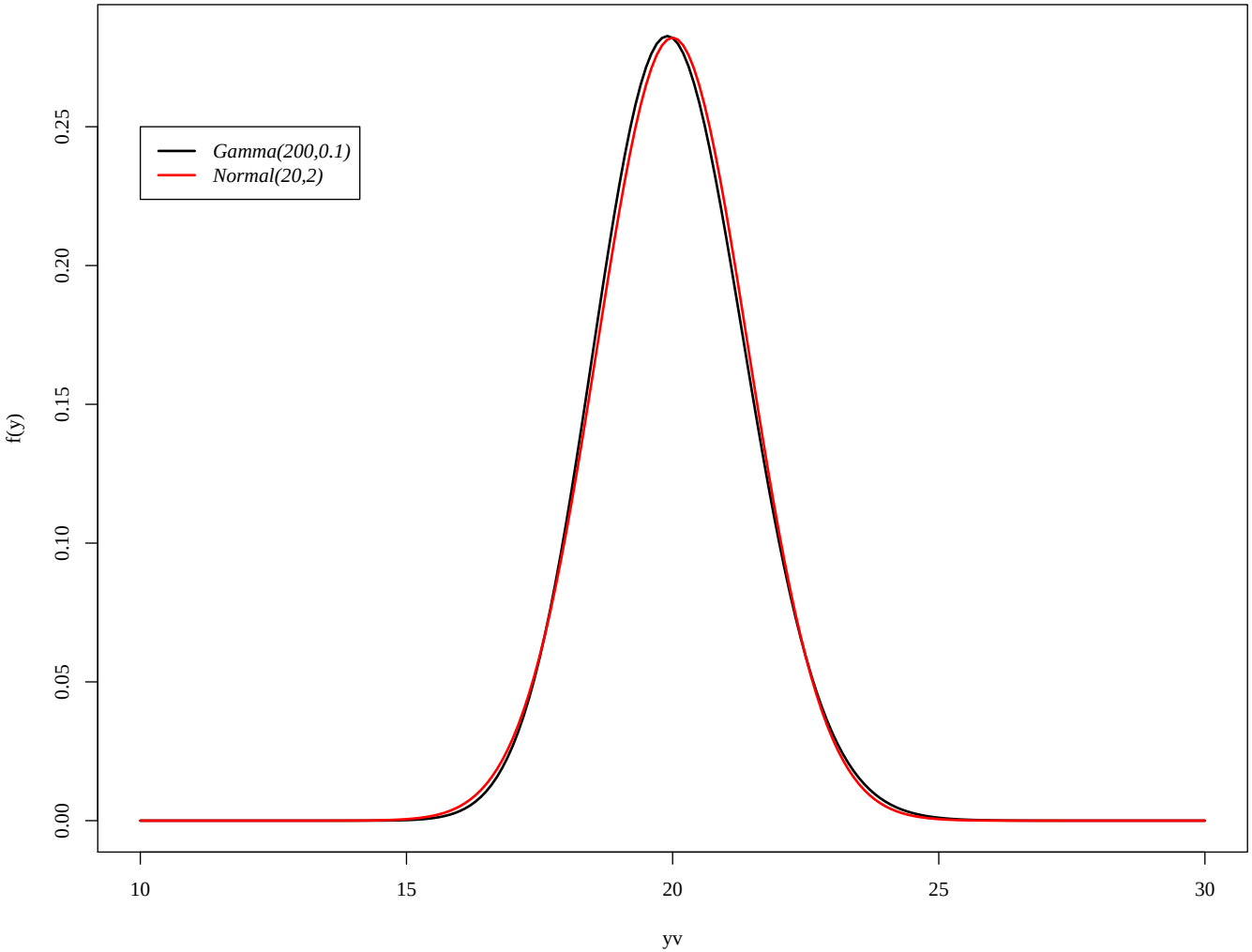
and

$$F_Y(y) = P(Y \leq y) \doteq \Phi\left(\frac{y - \alpha\beta}{\sqrt{\alpha\beta^2}}\right)$$

Note that for this approximation to work, we need to have α large.

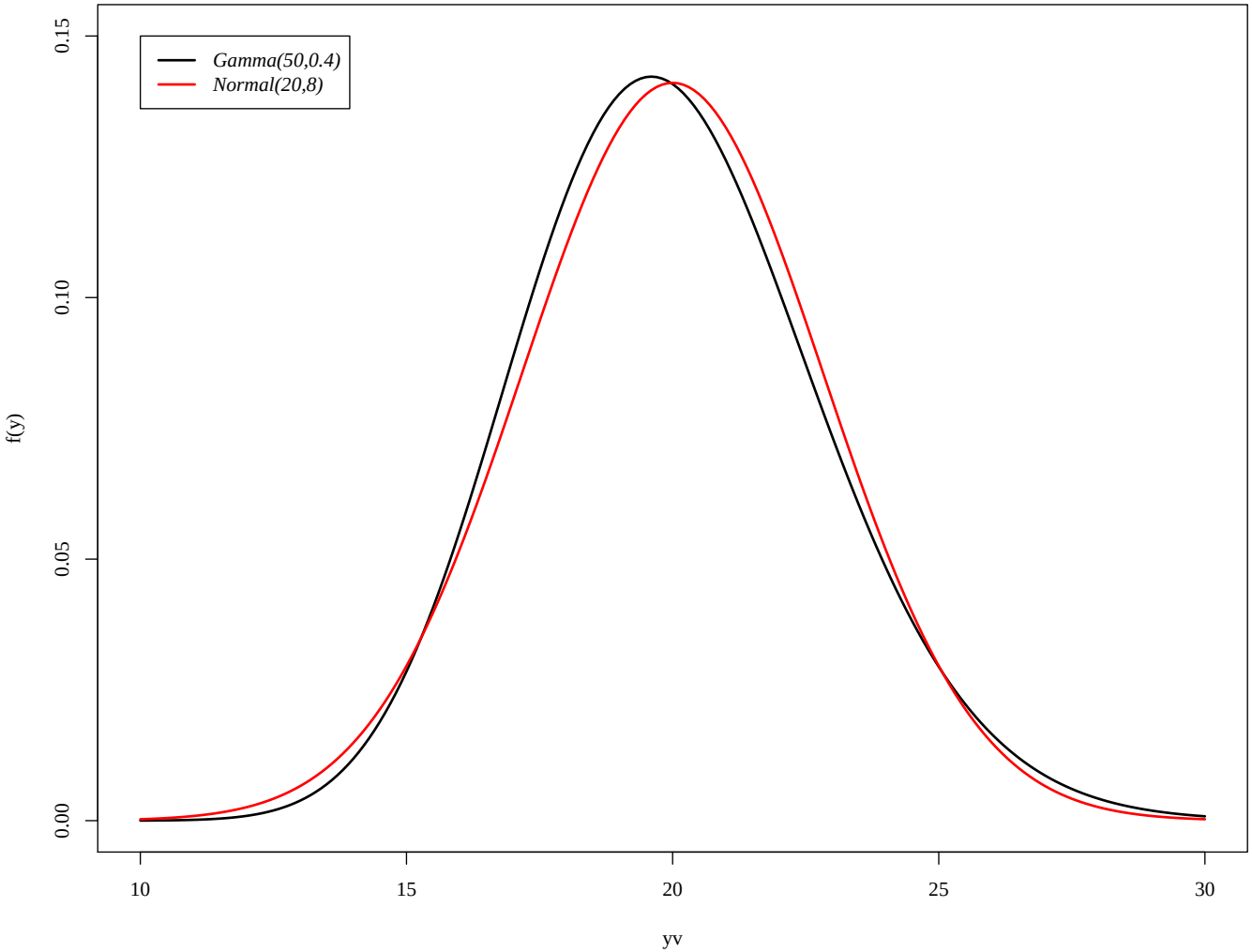
Applications of the central limit theorem (cont.)

Gamma: $\alpha = 200$, with $\beta = 0.1$.



Applications of the central limit theorem (cont.)

Gamma: $\alpha = 50$, with $\beta = 0.4$.



Chapter 5: Multivariate distributions

Multivariate probability distributions

We now consider probability specifications for a *vector* of random variables

$$(Y_1, Y_2, \ldots, Y_n).$$

We will mainly focus on the case where $n = 2$.

Note

The critical thing to remember is that when we distribute probability amongst the possible values of the random variables, *the total probability we assign must still equal one*.

Discrete bivariate distributions

A *joint pmf* for two variables (Y_1, Y_2) is the function

$$p(y_1, y_2) = P((Y_1 = y_1) \cap (Y_2 = y_2)) \equiv P(Y_1 = y_1, Y_2 = y_2)$$

for $(y_1, y_2) \in \mathbb{R}^2$. This function

- *"specifies probabilities"*

$$0 \leq p(y_1, y_2) \leq 1 \quad \text{for all } y_1, y_2$$

- *"sums to one"*

$$\sum_{y_1=-\infty}^{\infty} \sum_{y_2=-\infty}^{\infty} p(y_1, y_2) = 1.$$

although $p(y_1, y_2)$ may be zero for some arguments.

The marginal mass function

The *marginal pmf* for Y_1, $p_{Y_1}(.)$ is given by

$$p_{Y_1}(y_1) = P(Y_1 = y_1) = \sum_{y_2=-\infty}^{\infty} P(Y_1 = y_1, Y_2 = y_2)$$

that is

$$p_{Y_1}(y_1) = \sum_{y_2=-\infty}^{\infty} p_{Y_1,Y_2}(y_1, y_2).$$

This is a direct consequence of the partitioning result

$$(Y_1 = y_1) = \bigcup_{y_2=-\infty}^{\infty} (Y_1 = y_1) \cap (Y_2 = y_2).$$

The marginal mass function (cont.)

If $p_{Y_1,Y_2}(y_1, y_2)$ specifies the values in a probability table, we compute the marginal pmf

- for Y_1 by summing across the rows of the table;
- for Y_2 by summing down the columns of the table.

		Y_2			
		1	2	3	$p_{Y_1}(.)$
Y_1	1	$p_{Y_1,Y_2}(1,1)$	$p_{Y_1,Y_2}(1,2)$	$p_{Y_1,Y_2}(1,3)$	$p_{Y_1}(1)$
	2	$p_{Y_1,Y_2}(2,1)$	$p_{Y_1,Y_2}(2,2)$	$p_{Y_1,Y_2}(2,3)$	$p_{Y_1}(2)$
	3	$p_{Y_1,Y_2}(3,1)$	$p_{Y_1,Y_2}(3,2)$	$p_{Y_1,Y_2}(3,3)$	$p_{Y_1}(3)$
	4	$p_{Y_1,Y_2}(4,1)$	$p_{Y_1,Y_2}(4,2)$	$p_{Y_1,Y_2}(4,3)$	$p_{Y_1}(4)$
	5	$p_{Y_1,Y_2}(5,1)$	$p_{Y_1,Y_2}(5,2)$	$p_{Y_1,Y_2}(5,3)$	$p_{Y_1}(5)$
	$p_{Y_2}(.)$	$p_{Y_2}(1)$	$p_{Y_2}(2)$	$p_{Y_2}(3)$	1

The marginal mass function (cont.)

Note

The marginal pmfs $p_{Y_1}(y_1)$ and $p_{Y_2}(y_2)$ have all the properties of single variable pmfs; specifically

$$\sum_{y_1=-\infty}^{\infty} p_{Y_1}(y_1) = \sum_{y_1=-\infty}^{\infty} \sum_{y_2=-\infty}^{\infty} p_{Y_1,Y_2}(y_1, y_2) = 1.$$

The conditional mass function

Once we have the joint pmf

$$p_{Y_1,Y_2}(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2)$$

and the marginal pmf

$$p_{Y_1}(y_1) = P(Y_1 = y_1)$$

we can consider *conditional* pmfs.

The conditional mass function (cont.)

We have that if $P(Y_1 = y_1) > 0$, then the conditional probability that $Y_2 = y_2$, given that $Y_1 = y_1$, is

$$P(Y_2 = y_2 | Y_1 = y_1) = \frac{P(Y_1 = y_1, Y_2 = y_2)}{P(Y_1 = y_1)}$$

For a *fixed* value of y_1, we can consider how this conditional probability varies as argument y_2 varies.

The conditional mass function (cont.)

The *conditional pmf* for Y_2, given that $Y_1 = y_1$, is denoted

$$p_{Y_2|Y_1}(y_2|y_1)$$

and defined by

$$p_{Y_2|Y_1}(y_2|y_1) = P(Y_2 = y_2|Y_1 = y_1)$$

whenever $P(Y_1 = y_1) > 0$.

The conditional mass function (cont.)

The conditional pmfs are obtained by taking 'slices' through the joint pmf, and then standardizing the slice so that the probabilities sum to one.

Recall that

$$p_{Y_2|Y_1}(y_2|y_1) = \frac{P(Y_1 = y_1, Y_2 = y_2)}{P(Y_1 = y_1)} = \frac{p_{Y_1,Y_2}(y_1, y_2)}{p_{Y_1}(y_1)}$$

$$= \frac{p_{Y_1,Y_2}(y_1, y_2)}{\sum_{t_2=-\infty}^{\infty} p_{Y_1,Y_2}(y_1, t_2)}.$$

The conditional mass function (cont.)

Given $Y_1 = 2$, we define the conditional pmf $p_{Y_2|Y_1}(y_2|2)$ by examining the second row of the table.

		Y_2			
		1	2	3	$p_{Y_1}(.)$
	1	$p_{Y_1,Y_2}(1,1)$	$p_{Y_1,Y_2}(1,2)$	$p_{Y_1,Y_2}(1,3)$	$p_{Y_1}(1)$
	2	$p_{Y_1,Y_2}(2,1)$	$p_{Y_1,Y_2}(2,2)$	$p_{Y_1,Y_2}(2,3)$	$p_{Y_1}(2)$
Y_1	3	$p_{Y_1,Y_2}(3,1)$	$p_{Y_1,Y_2}(3,2)$	$p_{Y_1,Y_2}(3,3)$	$p_{Y_1}(3)$
	4	$p_{Y_1,Y_2}(4,1)$	$p_{Y_1,Y_2}(4,2)$	$p_{Y_1,Y_2}(4,3)$	$p_{Y_1}(4)$
	5	$p_{Y_1,Y_2}(5,1)$	$p_{Y_1,Y_2}(5,2)$	$p_{Y_1,Y_2}(5,3)$	$p_{Y_1}(5)$
	$p_{Y_2}(.)$	$p_{Y_2}(1)$	$p_{Y_2}(2)$	$p_{Y_2}(3)$	1

The conditional mass function (cont.)

Given $Y_2 = 3$, we define the conditional pmf $p_{Y_1|Y_2}(y_1|3)$ by examining the third column of the table.

		Y_2			
		1	2	3	$p_{Y_1}(.)$
Y_1	1	$p_{Y_1,Y_2}(1,1)$	$p_{Y_1,Y_2}(1,2)$	$p_{Y_1,Y_2}(1,3)$	$p_{Y_1}(1)$
	2	$p_{Y_1,Y_2}(2,1)$	$p_{Y_1,Y_2}(2,2)$	$p_{Y_1,Y_2}(2,3)$	$p_{Y_1}(2)$
	3	$p_{Y_1,Y_2}(3,1)$	$p_{Y_1,Y_2}(3,2)$	$p_{Y_1,Y_2}(3,3)$	$p_{Y_1}(3)$
	4	$p_{Y_1,Y_2}(4,1)$	$p_{Y_1,Y_2}(4,2)$	$p_{Y_1,Y_2}(4,3)$	$p_{Y_1}(4)$
	5	$p_{Y_1,Y_2}(5,1)$	$p_{Y_1,Y_2}(5,2)$	$p_{Y_1,Y_2}(5,3)$	$p_{Y_1}(5)$
	$p_{Y_2}(.)$	$p_{Y_2}(1)$	$p_{Y_2}(2)$	$p_{Y_2}(3)$	1

The conditional mass function (cont.)

Note

We have the fundamental relationship

$$p_{Y_1,Y_2}(y_1, y_2) = p_{Y_1}(y_1) p_{Y_2|Y_1}(y_2|y_1)$$

whenever $p_{Y_1}(y_1) > 0$.

Independence

Random variables Y_1 and Y_2 are *independent* if

$$p_{Y_1,Y_2}(y_1, y_2) = p_{Y_1}(y_1)p_{Y_2}(y_2)$$

for *all* values y_1, y_2. Equivalently, they are independent if

$$p_{Y_1|Y_2}(y_1|y_2) = p_{Y_1}(y_1) \qquad p_{Y_2|Y_1}(y_2|y_1) = p_{Y_2}(y_2)$$

for *all* values y_1, y_2 where the conditional pmfs are defined.

If Y_1 and Y_2 are not independent, they are *dependent*.

Independence (cont.)

Generalization: $Y_1, \ldots, Y_n$ are independent if

$$p_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = \prod_{i=1}^{n} p_{Y_i}(y_i)$$

for *all* vectors $(y_1, \ldots, y_n)$.

In terms of the joint cdf, this is equivalent to

$$F_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = P\left(\bigcap_{i=1}^{n}(Y_i \leq y_i)\right) = \prod_{i=1}^{n} F_{Y_i}(y_i)$$

Continuous case

If Y_1 and Y_2 are two *continuous* random variables, then we can still consider statements of the form

$$P((Y_1 \leq y_1) \cap (Y_2 \leq y_2))$$

and hence define the *joint cdf*

$$F_{Y_1,Y_2}(y_1, y_2) = P((Y_1 \leq y_1) \cap (Y_2 \leq y_2))$$

for any pair of real numbers (y_1, y_2).

Continuous case (cont.)

The joint cdf has the following properties:

- *"starts at zero"*

$$\lim_{y_1 \longrightarrow -\infty} F_{Y_1,Y_2}(y_1, y_2) = \lim_{y_2 \longrightarrow -\infty} F_{Y_1,Y_2}(y_1, y_2) = 0$$

- *"ends at one"*

$$\lim_{y_1 \longrightarrow \infty} \lim_{y_2 \longrightarrow \infty} F_{Y_1,Y_2}(y_1, y_2) = 1$$

- *"non-decreasing in both y_1 and y_2"*

$$F_{Y_1,Y_2}(y_1, y_2) \leq F_{Y_1,Y_2}(y_1 + c, y_2)$$

$$F_{Y_1,Y_2}(y_1, y_2) \leq F_{Y_1,Y_2}(y_1, y_2 + c)$$

for all y_1, y_2, and any $c > 0$.

Continuous case (cont.)

Furthermore, we have that

$$\lim_{y_1 \longrightarrow \infty} F_{Y_1,Y_2}(y_1, y_2) = P(Y_1 < \infty, Y_2 \leq y_2) \equiv P(Y_2 \leq y_2)$$

so that

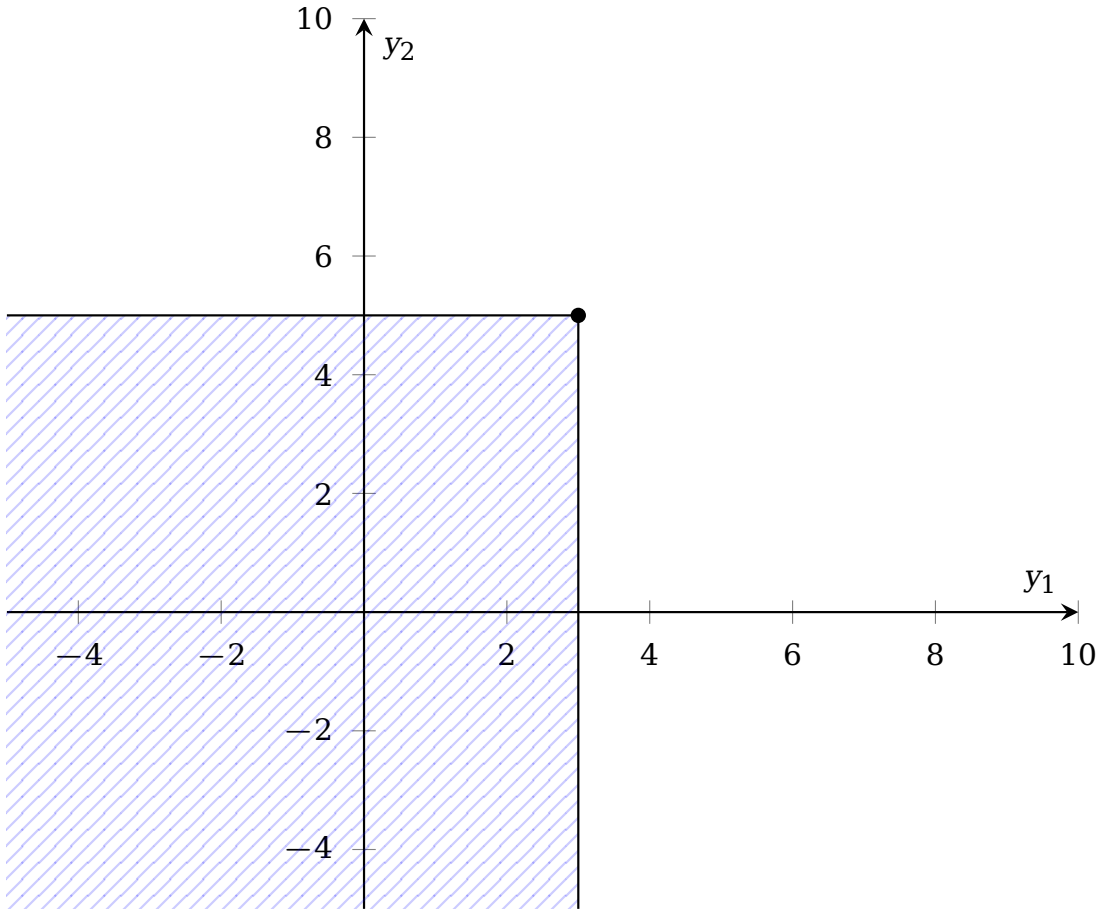
$$\lim_{y_1 \longrightarrow \infty} F_{Y_1,Y_2}(y_1, y_2) = F_{Y_2}(y_2)$$

and similarly

$$\lim_{y_2 \longrightarrow \infty} F_{Y_1,Y_2}(y_1, y_2) = F_{Y_1}(y_1).$$

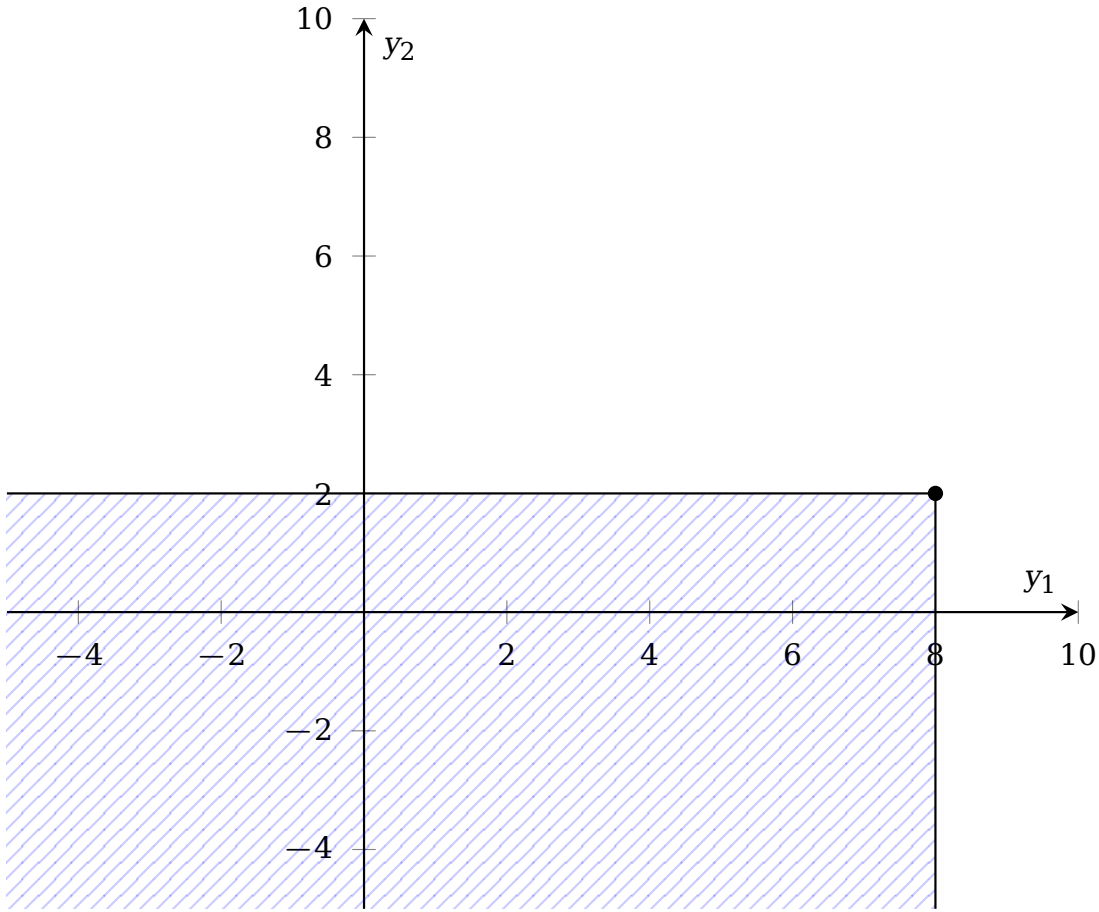
Continuous case (cont.)

$$F_{Y_1,Y_2}(3,5) = P(Y_1 \leq 3, Y_2 \leq 5)$$



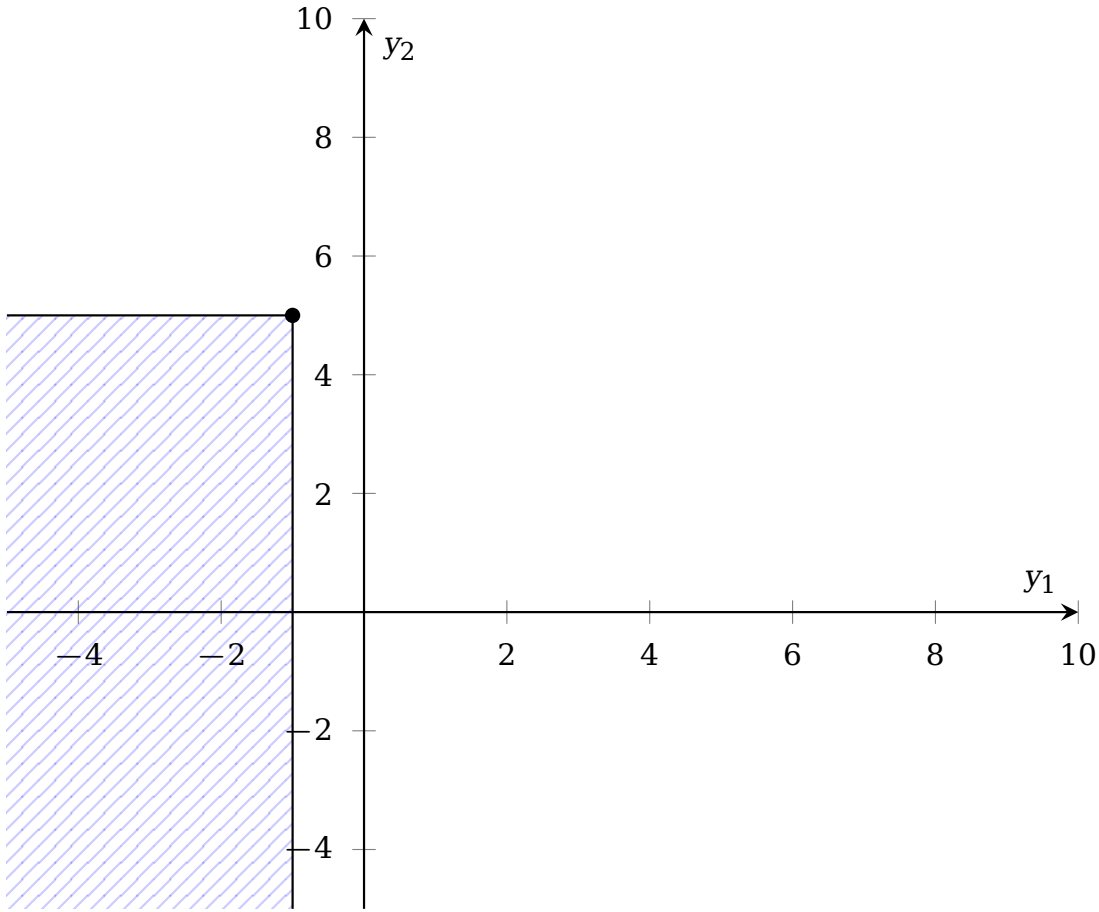
Continuous case (cont.)

$$F_{Y_1,Y_2}(8,2) = P(Y_1 \leq 8, Y_2 \leq 2)$$



Continuous case (cont.)

$$F_{Y_1,Y_2}(-1,5) = P(Y_1 \leq -1, Y_2 \leq 5)$$



Continuous case (cont.)

We accumulate probability over the shaded rectangle

$$(-\infty, y_1] \times (-\infty, y_2],$$

to compute $F_{Y_1,Y_2}(y_1, y_2)$.

However we must still have

$$P(Y_1 = y_1, Y_2 = y_2) = 0$$

for all y_1 and y_2.

Continuous case (cont.)

As in the single variable case, we introduce the *joint pdf*

$$f_{Y_1,Y_2}(y_1, y_2)$$

to describe how probability is spread around, where

$$F_{Y_1,Y_2}(y_1, y_2) = \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} f_{Y_1,Y_2}(t_1, t_2)\, dt_2 dt_1$$

that is, to compute $F_{Y_1,Y_2}(y_1, y_2)$ we integrate $f_{Y_1,Y_2}(y_1, y_2)$ over the rectangle

$$(-\infty, y_1] \times (-\infty, y_2].$$

Continuous case (cont.)

We can compute the *double* integral as follows: writing

$$\int_{-\infty}^{y_1} \left\{ \int_{-\infty}^{y_2} f_{Y_1,Y_2}(t_1, t_2)\, dt_2 \right\} dt_1$$

- fix t_1, and perform the first (*inner*) integration

$$\int_{-\infty}^{y_2} f_{Y_1,Y_2}(t_1, t_2)\, dt_2$$

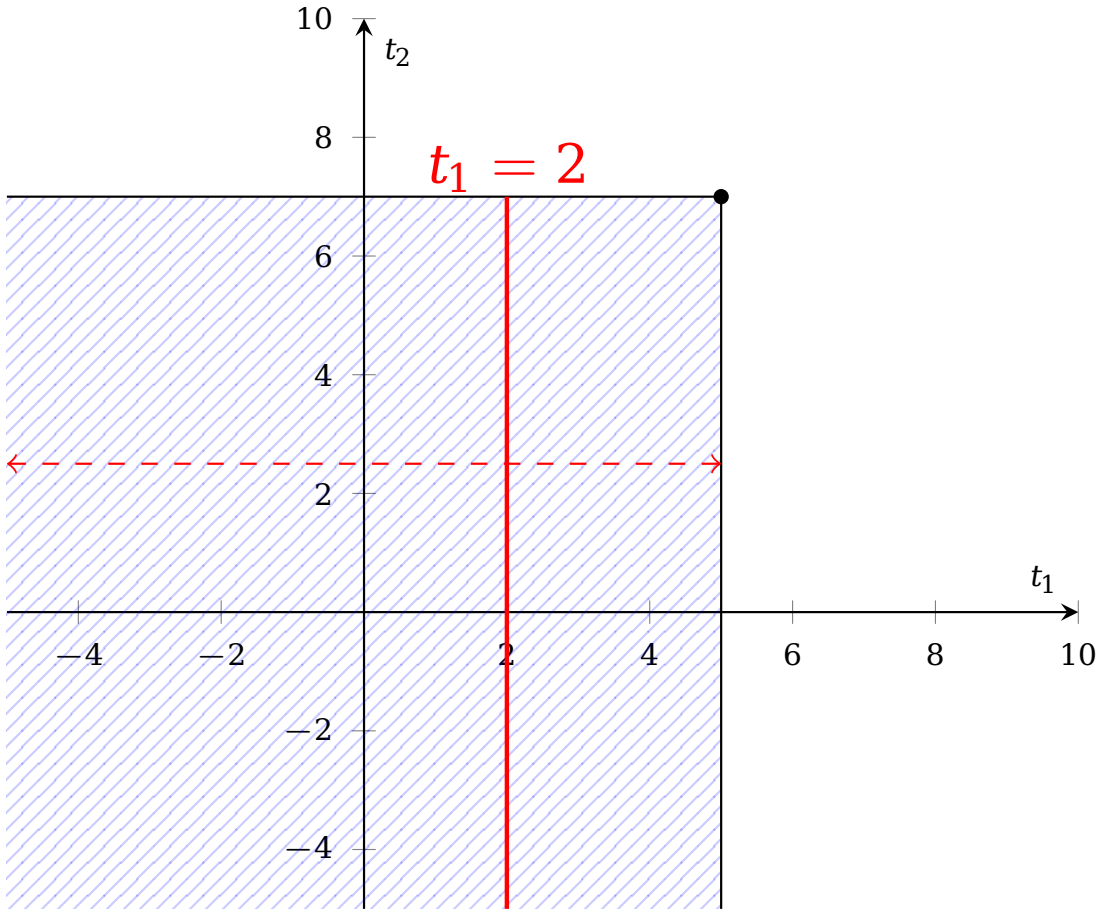
  in the 'strip' at t_1 to obtain a function $g(t_1, y_2)$, say;
- perform the second (*outer*) integration

$$\int_{-\infty}^{y_1} g(t_1, y_2)\, dt_1.$$

  to obtain the joint cdf.

Continuous case (cont.)

$$F_{Y_1,Y_2}(5,7) = P(Y_1 \leq 5, Y_2 \leq 7)$$



Continuous case (cont.)

The joint pdf describes how the probability is spread 'point-by-point' across $\mathbb{R}^2$. We must have that

- the joint pdf is non-negative

$$f_{Y_1,Y_2}(y_1, y_2) \geq 0 \qquad (y_1, y_2 \in \mathbb{R}^2$$

(as the joint cdf is non-decreasing in both y_1 and y_2);

- the joint pdf integrates to 1

$$\int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2)\, dy_2 \right\} dy_1 = 1$$

(as the probability must accumulate to 1 over $\mathbb{R}^2$).

Continuous case (cont.)

We compute $f_{Y_1,Y_2}(y_1, y_2)$ from $F_{Y_1,Y_2}(y_1, y_2)$ using partial differentiation:

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{\partial^2}{\partial y_1 \partial y_2} \{F_{Y_1,Y_2}(y_1, y_2)\}$$

Step 1: differentiate $F_{Y_1,Y_2}(y_1, y_2)$ with respect to y_2 *holding y_1 constant*;

Step 2: differentiate the *result of **Step 1*** with respect to y_1.

Continuous case (cont.)

We can regard the calculation as

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{\partial}{\partial y_1}\left\{\frac{\partial F_{Y_1,Y_2}(y_1,y_2)}{\partial y_2}\right\}$$

Note

We could choose to differentiate with respect to y_1 in Step 1, and then differentiate the result with respect to y_2: we will get the same answer

$$\frac{\partial^2}{\partial y_1 \partial y_2}\left\{F_{Y_1,Y_2}(y_1,y_2)\right\} = \frac{\partial^2}{\partial y_2 \partial y_1}\left\{F_{Y_1,Y_2}(y_1,y_2)\right\}.$$

This is a general property of partial derivatives that holds provided the result of each side is continuous.

Constant pdfs

If the joint pdf is constant over a finite region $\mathcal{Y} \in \mathbb{R}^2$

$$f_{Y_1,Y_2}(y_1, y_2) = c \qquad (y_1, y_2) \in \mathcal{Y}$$

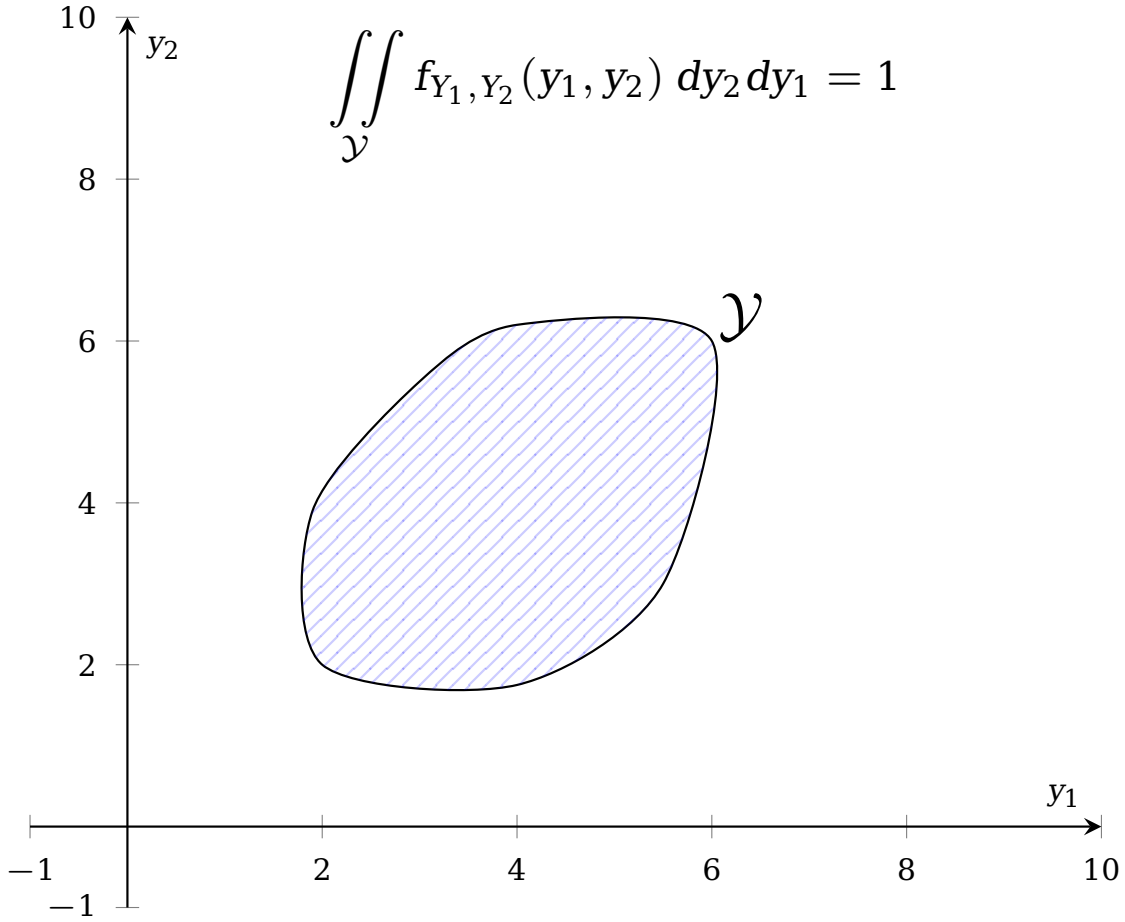
with $f_{Y_1,Y_2}(y_1, y_2)$ zero otherwise, then to compute probabilities associated with this pdf, for example

$$P((Y_1, Y_2) \in \mathcal{A}) = \iint_{\mathcal{A}} f_{Y_1,Y_2}(y_1, y_2)\, dy_2\, dy_1$$

we must compute the area of $\mathcal{A}$ and divide it by the area of $\mathcal{Y}$.

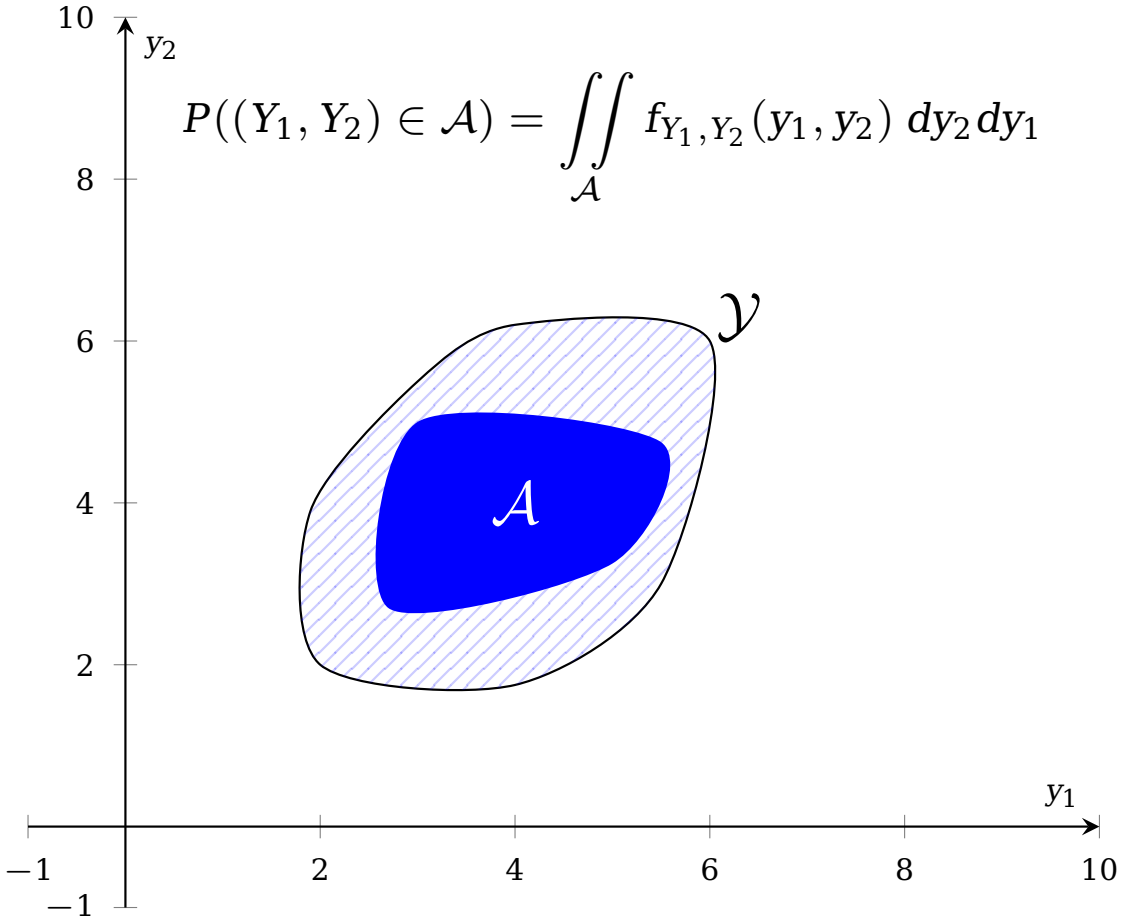
Constant pdfs (cont.)

$$\iint_{\mathcal{Y}} f_{Y_1,Y_2}(y_1, y_2)\, dy_2\, dy_1 = 1$$

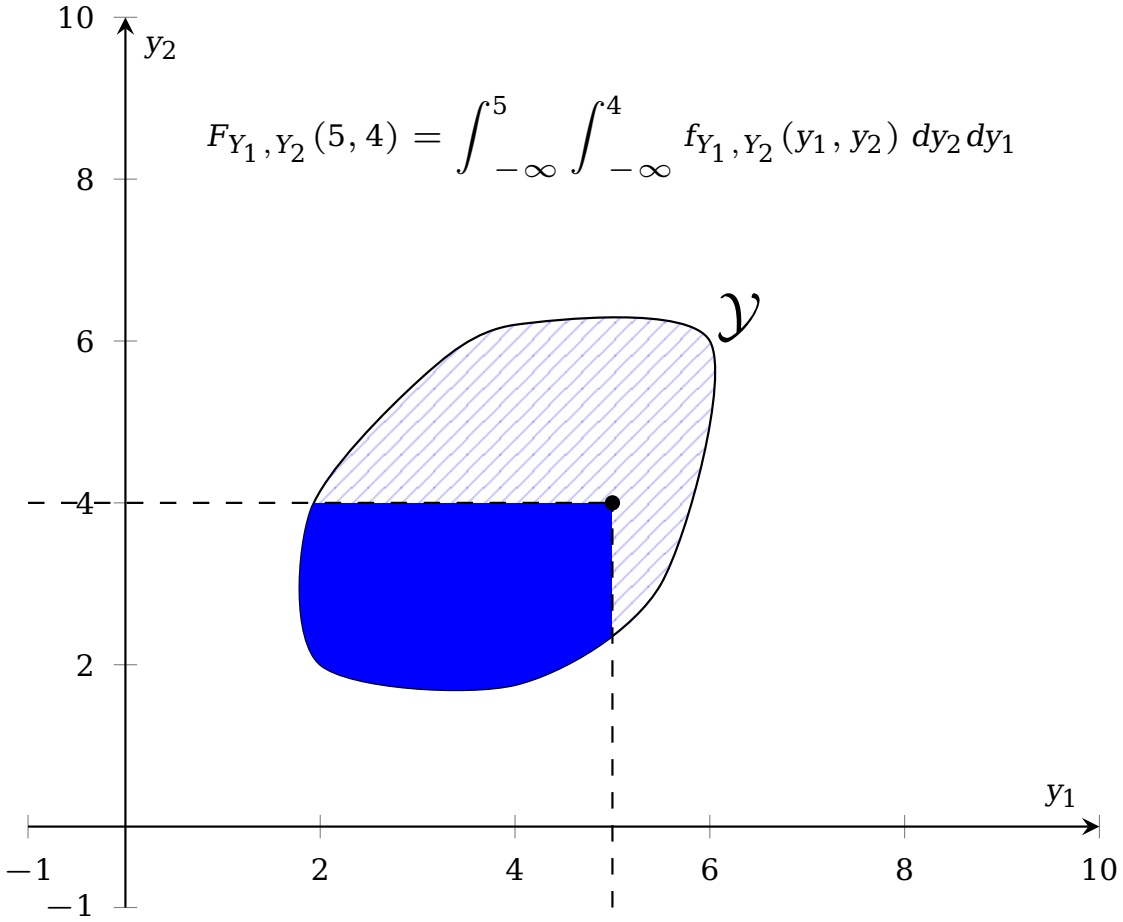


Constant pdfs (cont.)

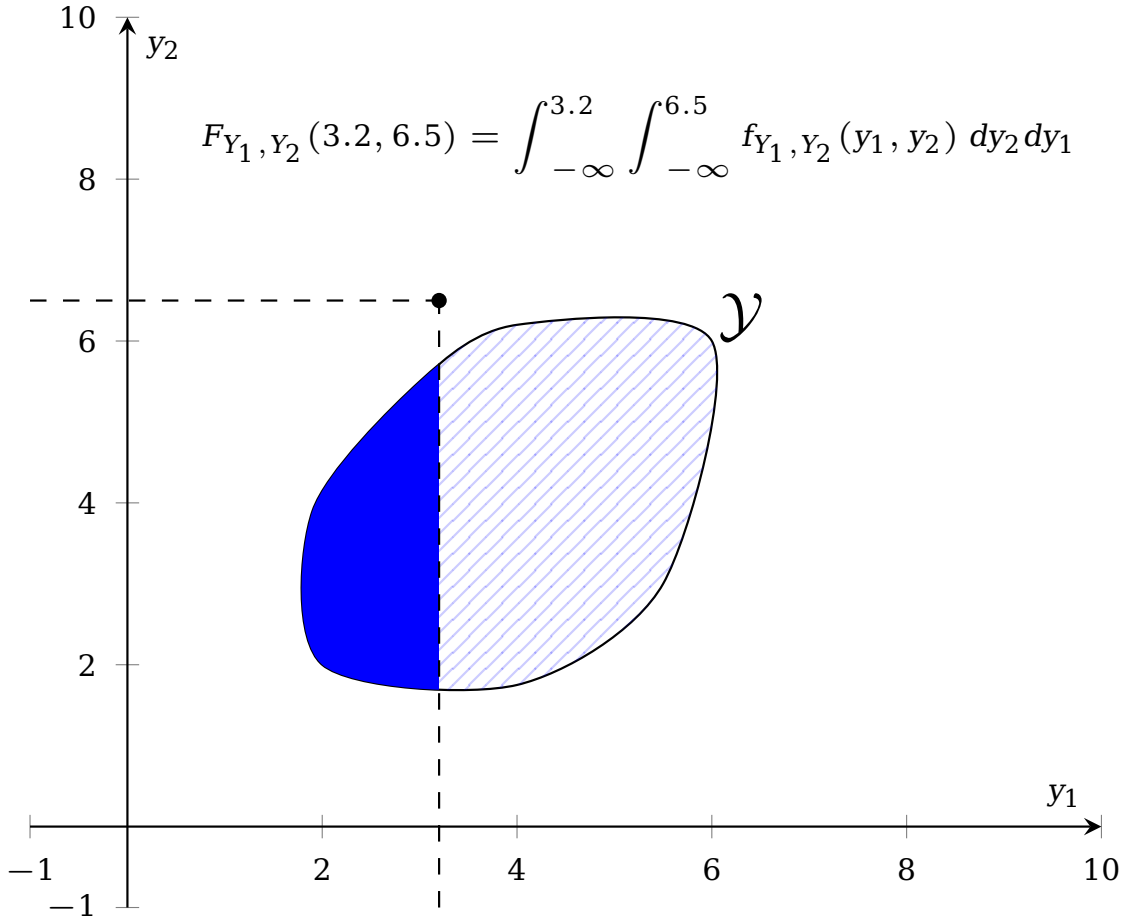
$$P((Y_1, Y_2) \in \mathcal{A}) = \iint_{\mathcal{A}} f_{Y_1,Y_2}(y_1, y_2)\, dy_2\, dy_1$$



Constant pdfs (cont.)



Constant pdfs (cont.)



The marginal density function

For Y_1: the *marginal pdf* $f_{Y_1}(y_1)$ for each fixed y_1 by

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2)\, dy_2$$

For Y_2: the *marginal pdf* $f_{Y_2}(y_2)$ for each fixed y_2 by

$$f_{Y_2}(y_2) = \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2)\, dy_1$$

The conditional density function

Once we have the joint pdf

$$f_{Y_1,Y_2}(y_1, y_2)$$

and the marginal pdf

$$f_{Y_1}(y_1)$$

we can consider *conditional* pdfs.

The conditional density function (cont.)

The *conditional pdf* for Y_2, given that $Y_1 = y_1$, is denoted

$$f_{Y_2|Y_1}(y_2|y_1)$$

and defined by

$$f_{Y_2|Y_1}(y_2|y_1) = \frac{f_{Y_1,Y_2}(y_1, y_2)}{f_{Y_1}(y_1)}$$

whenever $f_{Y_1}(y_1) > 0$.

The conditional density function (cont.)

The function $f_{Y_2|Y_1}(y_2|y_1)$ is a pdf in y_2 for every fixed y_1. That is, *for every fixed y_1 where $f_{Y_1}(y_1) > 0$*,

- the conditional pdf is non-negative

$$f_{Y_2|Y_1}(y_2|y_1) \geq 0 \qquad -\infty < y_2 < \infty;$$

- the conditional pdf is integrates to 1 over y_2

$$\int_{-\infty}^{\infty} f_{Y_2|Y_1}(y_2|y_1)\, dy_2 = 1.$$

We can define $f_{Y_1|Y_2}(y_1|y_2)$ in an identical fashion.

Conditional pdfs are obtained by taking a 'slice' through the joint pdf at $Y_1 = y_1$, and then standardizing the slice so that the density integrates to one.

The conditional density function (cont.)

As in the discrete case, we have the *chain rule factorization*

$$f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1}(y_1) f_{Y_2|Y_1}(y_2|y_1)$$

whenever $f_{Y_1}(y_1) > 0$.

As ever, we can exchange the roles of Y_1 and Y_2 to obtain that

$$f_{Y_1,Y_2}(y_1, y_2) = f_{Y_2}(y_2) f_{Y_1|Y_2}(y_1|y_2).$$

whenever $f_{Y_1}(y_1) > 0$.

Multivariate distributions

All of the above ideas extend naturally to more than two variables. If $(Y_1, \ldots, Y_n)$ form an n-dimensional random vector, we can consider the *joint pdf*

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n)$$

and *joint cdf*

$$F_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n).$$

Multivariate distributions (cont.)

We can consider *marginalization* by integrating out $n - d < n$ of the dimensions to leave a d-dimensional probability distribution: for example if $n = 4$ and $d = 2$, we have

$$f_{Y_1,Y_2}(y_1, y_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{Y_1,Y_2,Y_3,Y_4}(y_1, y_2, y_3, y_4)\, dy_3 dy_4$$

or

$$f_{Y_2,Y_4}(y_2, y_4) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{Y_1,Y_2,Y_3,Y_4}(y_1, y_2, y_3, y_4)\, dy_1 dy_3$$

and so on. This construction works with any n and any d.

Independence

Continuous random variables $(Y_1, \ldots, Y_n)$ are *independent* if

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = \prod_{i=1}^{n} f_{Y_i}(y_i)$$

for *all* vectors $(y_1, \ldots, y_n) \in \mathbb{R}^n$. Equivalently they are independent if

$$F_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = \prod_{i=1}^{n} F_{Y_i}(y_i)$$

for *all* vectors $(y_1, \ldots, y_n) \in \mathbb{R}^n$.

Independence (cont.)

In the bivariate case, Y_1 and Y_2 are independent if

$$f_{Y_2|Y_1}(y_2|y_1) = f_{Y_2}(y_2)$$

for all (y_1, y_2) where the conditional pdf is defined, or equivalently if

$$f_{Y_1|Y_2}(y_1|y_2) = f_{Y_1}(y_1)$$

for all (y_1, y_2).

Multivariate expectations

For two variables, we can define the *joint expectation*

$$\mathbb{E}[g(Y_1, Y_2)] = \sum_{y_1=-\infty}^{\infty} \sum_{y_2=-\infty}^{\infty} g(y_1, y_2) p_{Y_1,Y_2}(y_1, y_2).$$

or

$$\mathbb{E}[g(Y_1, Y_2)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(y_1, y_2) f_{Y_1,Y_2}(y_1, y_2)\, dy_1\, dy_2.$$

Multivariate expectations (cont.)

The properties of expectations for more than one variable match those in the single variable case.

In particular, if a_1, a_2 are constants, and $g_1(.,.)$ and $g_2(.,.)$ are two functions, then

$$\mathbb{E}[a_1 g_1(Y_1, Y_2) + a_2 g_2(Y_1, Y_2)]$$

$$= a_1 \mathbb{E}[g_1(Y_1, Y_2)] + a_2 \mathbb{E}[g_2(Y_1, Y_2)]$$

provided all the expectations are finite.

Covariance and correlation

For two random variables Y_1 and Y_2, with

$$\mathbb{E}[Y_1] = \mu_1 \qquad \mathbb{E}[Y_2] = \mu_2$$

- the *covariance* between Y_1 and Y_2 is

$$\text{Cov}[Y_1, Y_2] = \mathbb{E}[(Y_1 - \mu_1)(Y_2 - \mu_2)]$$

- the *correlation* between Y_1 and Y_2 is

$$\text{Corr}[Y_1, Y_2] = \frac{\text{Cov}[Y_1, Y_2]}{\sqrt{\mathbb{V}[Y_1]\mathbb{V}[Y_2]}}$$

Covariance and correlation (cont.)

Note that

$$\begin{aligned}
\mathrm{Cov}[Y_1, Y_2] &= \mathbb{E}[(Y_1 - \mu_1)(Y_2 - \mu_2)] \\
&= \mathbb{E}[Y_1 Y_2 - \mu_1 Y_2 - Y_1 \mu_2 + \mu_1 \mu_2] \\
&= \mathbb{E}[Y_1 Y_2] - \mu_1 \mathbb{E}[Y_2] - \mu_2 \mathbb{E}[Y_1] + \mu_1 \mu_2 \\
&= \mathbb{E}[Y_1 Y_2] - \mu_1 \mu_2 - \mu_2 \mu_1 + \mu_1 \mu_2 \\
&= \mathbb{E}[Y_1 Y_2] - \mu_1 \mu_2 .
\end{aligned}$$

Covariance and correlation (cont.)

Both covariance and correlation are *measures of dependence*.
We have that for any pair of random variables,

$$-1 \leq \text{Corr}[Y_1, Y_2] \leq 1$$

If

$$\text{Cov}[Y_1, Y_2] = \text{Corr}[Y_1, Y_2] = 0$$

we say that Y_1 and Y_2 are *uncorrelated*.

Independent random variables

If Y_1 and Y_2 are independent then

$$\mathbb{E}[Y_1 Y_2] = \mathbb{E}[Y_1]\mathbb{E}[Y_2]$$

More generally, if $g_1(Y_1)$ and $g_2(Y_2)$ are two functions of the independent random variables, then

$$\mathbb{E}[g_1(Y_1)g_2(Y_2)] = \mathbb{E}[g_1(Y_1)]\mathbb{E}[g_2(Y_2)]$$

Independent random variables (cont.)

Therefore if two random variables Y_1 and Y_2 are independent,

$$\begin{aligned}\text{Cov}[Y_1, Y_2] &= \mathbb{E}[Y_1 Y_2] - \mu_1 \mu_2 \\ &= \mathbb{E}[Y_1]\mathbb{E}[Y_2] - \mu_1 \mu_2 \\ &= \mu_1 \mu_2 - \mu_1 \mu_2 = 0\end{aligned}$$

and Y_1 and Y_2 are uncorrelated.

Linear combinations of random variables

Suppose that $Y_1, \ldots, Y_n$ and $X_1, \ldots, X_m$ be random variables with

$$\mathbb{E}[Y_i] = \mu_i \qquad \mathbb{E}[X_j] = \xi_j.$$

Define

$$U_1 = \sum_{i=1}^{n} a_i Y_i \qquad U_2 = \sum_{j=1}^{m} b_j X_j$$

for real constants $a_1, \ldots, a_n$ and $b_1, \ldots, b_m$.

Linear combinations of random variables (cont.)

(a) Expectations:

$$\mathbb{E}[U_1] = \sum_{i=1}^{n} a_i \mathbb{E}[Y_i] = \sum_{i=1}^{n} a_i \mu_i$$

(b) Variances:

$$\mathbb{V}[U_1] = \sum_{i=1}^{n} a_i^2 \mathbb{V}[Y_i] + 2 \sum_{j=2}^{n} \sum_{i=1}^{j-1} a_i a_j \mathrm{Cov}[Y_i, Y_j]$$

(c) Covariances:

$$\mathrm{Cov}[U_1, U_2] = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \mathrm{Cov}[Y_i, X_j]$$

Sums of random variables (revisited)

For random variables Y_1 and Y_2 consider the distribution of

$$Y = Y_1 + Y_2$$

which we studied previously in the context using the convolution calculation in Chapter 3.

Sums of random variables (revisited) (cont.)

We now see that the convolution formula can be viewed in the context of a joint distribution calculation:

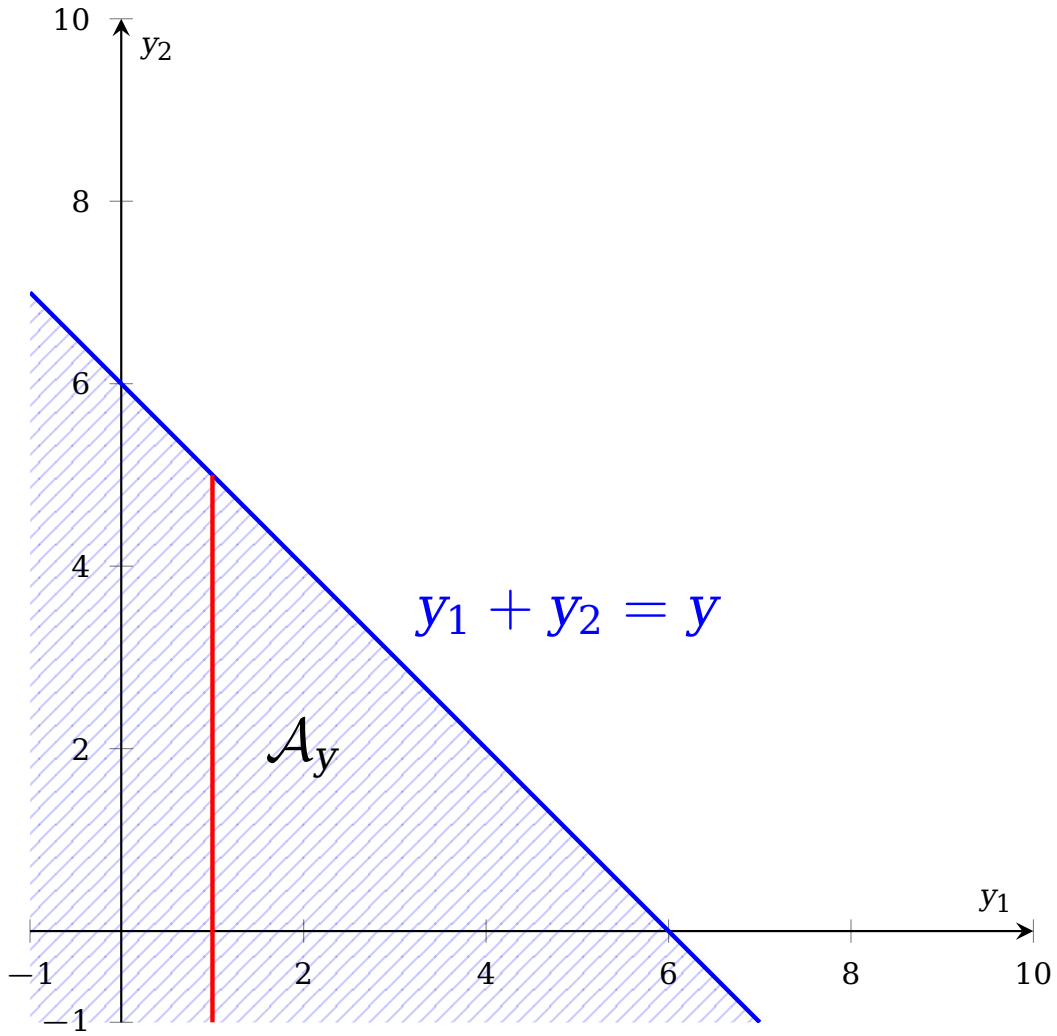
$$F_Y(y) = P(Y \leq y) = P(Y_1 + Y_2 \leq y) = \iint_{\mathcal{A}_y} f_{Y_1,Y_2}(y_1, y_2)\, dy_2\, dy_1$$

where

$$\mathcal{A}_y = \left\{(y_1, y_2) \in \mathbb{R}^2 : y_1 + y_2 \leq y\right\}.$$

Sums of random variables (revisited) (cont.)

Need to integrate over the region below the line $y_1 + y_2 = y$.



Sums of random variables (revisited) (cont.)

Therefore we can write

$$F_Y(y) = \int_{-\infty}^{\infty} \int_{-\infty}^{y-y_1} f_{Y_1,Y_2}(y_1, y_2)\, dy_2\, dy_1.$$

By differentiating with respect to y under the integral sign, using the Fundamental Theorem of Calculus, we see that

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y - y_1)\, dy_1.$$

Sums of random variables (revisited) (cont.)

We now consider an alternate method of calculation: recall the definition of the moment generating function:

$$m_Y(t) = \mathbb{E}\left[e^{tY}\right]$$

for $-b \leq t \leq b$, some $b > 0$.

We can define a similar quantity for the random variable U which is a transformation of Y, $U = g(Y)$ say, as

$$m_U(t) = \mathbb{E}\left[e^{tU}\right] \equiv \mathbb{E}\left[e^{tg(Y)}\right].$$

Sums of random variables (revisited) (cont.)

For two random variables Y_1 and Y_2, with

$$Y = a_1 g_1(Y_1) + a_2 g_2(Y_2),$$

we have

$$m_Y(t) = \mathbb{E}\left[e^{tY}\right] = \mathbb{E}\left[e^{t\{a_1 g_2(Y_1) + a_2 g_2(Y_2)\}}\right]$$

This is the mgf for Y. We can write

$$m_Y(t) = \mathbb{E}\left[e^{tY}\right] = \mathbb{E}\left[e^{ta_1 g_1(Y_1)} e^{ta_2 g_2(Y_2)}\right].$$

Sums of random variables (revisited) (cont.)

If Y_1 and Y_2 are independent we can further write

$$m_Y(t) = \mathbb{E}\left[e^{ta_1 g_1(Y_1)}\right] \mathbb{E}\left[e^{ta_2 g_2(Y_2)}\right].$$

– this follows by the earlier result on expectations.

Sums of random variables (revisited) (cont.)

Now consider the special case $Y = Y_1 + Y_2$:

$$m_Y(t) = \mathbb{E}\left[e^{t\{Y_1+Y_2\}}\right].$$

It follows immediately that if Y_1 and Y_2 are independent,

$$m_Y(t) = \mathbb{E}\left[e^{tY_1}\right]\mathbb{E}\left[e^{tY_2}\right] = m_{Y_1}(t)m_{Y_2}(t).$$

This gives us a way to compute the distribution of sums of independent random variables.

Sums of random variables (revisited) (cont.)

For example, if Y_1 and Y_2 and independent, and $Y = Y_1 + Y_2$ we have

Y_1	Y_2	$Y_1 + Y_2$
$Binomial(n_1, p)$	$Binomial(n_2, p)$	$Binomial(n_1 + n_2, p)$
$Poisson(\lambda_1)$	$Poisson(\lambda_2)$	$Poisson(\lambda_1 + \lambda_2)$
$Normal(\mu_1, \sigma_1^2)$	$Normal(\mu_2, \sigma_2^2)$	$Normal(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

Multivariate transformations

It is possible to consider general multivariate transformations of variables. For example if Y_1 and Y_2 are two variables with joint pdf $f_{Y_1,Y_2}(y_1, y_2)$, we can consider

$$U_1 = h_1(Y_1, Y_2) \qquad U_2 = h_2(Y_1, Y_2)$$

and attempt to compute

$$f_{U_1,U_2}(u_1, u_2).$$

Multivariate transformations (cont.)

For example, we might have

$$U_1 = Y_1 + Y_2 \qquad U_2 = Y_1 - Y_2$$

or

$$U_1 = Y_1 Y_2 \qquad U_2 = Y_1 / Y_2$$

etc.

Multivariate transformations (cont.)

We can approach this from a 'change of variable' perspective. If the transformation

$$(Y_1, Y_2) \longrightarrow (U_1, U_2)$$

is a *1-1 transformation*, then the inverse transformation

$$(U_1, U_2) \longrightarrow (Y_1, Y_2)$$

is well-defined. Suppose we write

$$Y_1 = h_1^{-1}(U_1, U_2) \qquad Y_2 = h_2^{-1}(U_1, U_2).$$

Multivariate transformations (cont.)

Then we may derive that for $(u_1, u_2) \in \mathbb{R}^2$,

$$f_{U_1,U_2}(u_1, u_2) = f_{Y_1,Y_2}(h_1^{-1}(u_1, u_2), h_2^{-1}(u_1, u_2))|J(u_1, u_2)|$$

where $J(u_1, u_2)$ is the *Jacobian* of the transformation, that is

$$J(u_1, u_2) = \det \begin{bmatrix} \dfrac{\partial h_1^{-1}(u_1, u_2)}{\partial u_1} & \dfrac{\partial h_1^{-1}(u_1, u_2)}{\partial u_2} \\ \dfrac{\partial h_2^{-1}(u_1, u_2)}{\partial u_1} & \dfrac{\partial h_2^{-1}(u_1, u_2)}{\partial u_2} \end{bmatrix}$$

where $\det[.]$ denotes determinant, and $|.|$ denotes absolute value.

Multivariate transformations (cont.)

This result holds provided that all the partial derivatives exist and are continuous, and that

$$|J(u_1, u_2)| > 0.$$

The result extends beyond two dimensions.