FIGURE 3. Gamma (r, λ) densities for $\lambda = 1$ and r a multiple of $1/4$, $0 < r \leq 1$.

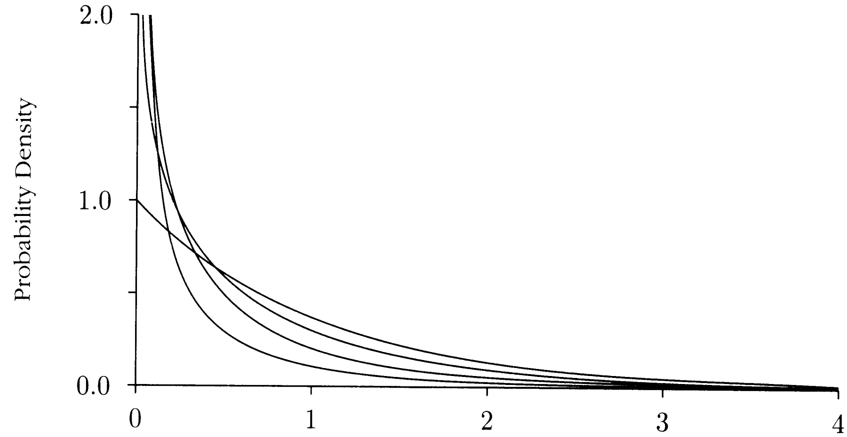


FIGURE 4. Gamma (r, λ) densities for $\lambda = 1$ and r a multiple of $1/4$, $1 \leq r \leq 2$.

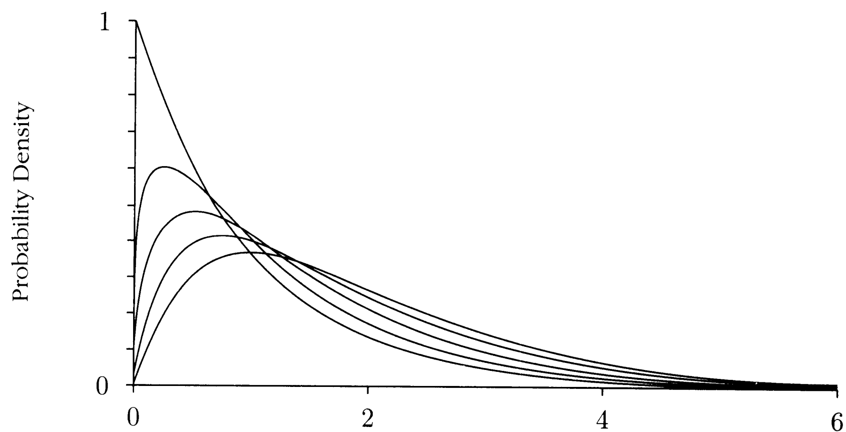
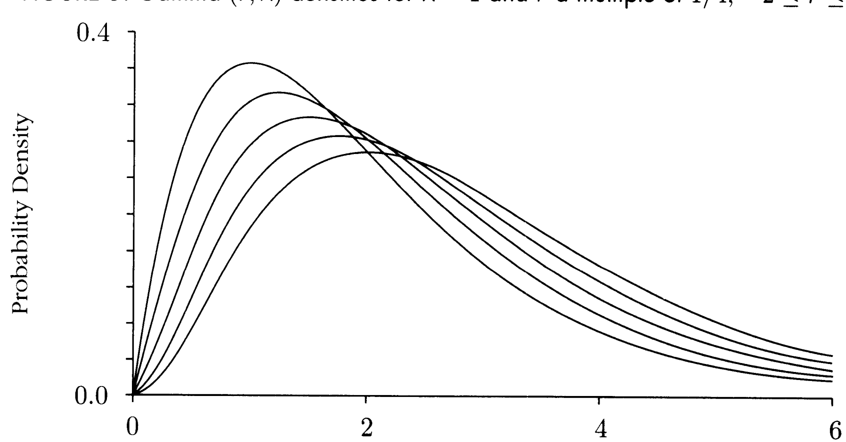


FIGURE 5. Gamma (r, λ) densities for $\lambda = 1$ and r a multiple of $1/4$, $2 \leq r \leq 3$.



Exercises 4.2

1. Suppose a particular kind of atom has a half-life of 1 year. Find:

a) the probability that an atom of this type survives at least 5 years;

b) the time at which the expected number of atoms is 10% of the original;

c) if there are 1024 atoms present initially, the time at which the expected number of atoms remaining is one;

d) the chance that in fact none of the 1024 original atoms remains after the time calculated in c).

2. A piece of rock contains 10^{20} atoms of a particular substance. Each atom has an exponentially distributed lifetime with a half-life of one century. How many centuries must pass before

a) it is most likely that about 100 atoms remain;

b) there is about a 50% chance that at least one atom remains. What assumptions are you making?

3. Suppose the time until the next earthquake in a particular place is exponentially distributed with rate 1 per year. Find the probability that the next earthquake happens within

a) one year; b) six months; c) two years; d) 10 years.

4. Suppose component lifetimes are exponentially distributed with mean 10 hours. Find:

a) the probability that a component survives 20 hours;

b) the median component lifetime;

c) the SD of component lifetime;

d) the probability that the average lifetime of 100 independent components exceeds 11 hours;

e) the probability that the average lifetime of 2 independent components exceeds 11 hours.

5. Suppose calls are arriving at a telephone exchange at an average rate of one per second, according to a Poisson arrival process. Find:

a) the probability that the fourth call after time $t = 0$ arrives within 2 seconds of the third call;

b) the probability that the fourth call arrives by time $t = 5$ seconds;

c) the expected time at which the fourth call arrives.

6. A Geiger counter is recording background radiation at an average rate of one hit per minute. Let T_3 be the time in minutes when the third hit occurs after the counter is switched on. Find $P(2 \le T_3 \le 4)$.

7. Let $0 < p < 1$. For the exponential distribution with rate λ, find a formula for the $100p$th percentile point t_p such that $P(T \le t_p) = 100p\%$.

8. Transistors produced by one machine have a lifetime which is exponentially distributed with mean 100 hours. Those produced by a second machine have an exponentially distributed lifetime with mean 200 hours. A package of 12 transistors contains 4 produced by the first machine and 8 produced by the second. Let X be the lifetime of a transistor picked at random from this package. Find:

a) $P(X \geq 200 \text{ hours})$; b) $E(X)$; c) $Var(X)$.

9. Gamma function and moments of the exponential distribution. Consider the gamma function $\Gamma(r) = \int_0^\infty x^{r-1}e^{-x}dx \quad (r > 0)$

a) Use integration by parts to show that $\Gamma(r+1) = r\Gamma(r) \quad (r > 0)$

b) Deduce from a) that $\Gamma(r) = (r-1)! \quad (r = 1, 2, \ldots)$

c) If T has exponential distribution with rate 1, then

$$E(T^n) = n! \qquad (n = 0, 1, 2, \ldots) \qquad \text{and} \quad SD(T) = 1$$

d) If T has exponential distribution with rate λ, then show λT has exponential distribution with rate 1, hence

$$E(T^n) = n!/\lambda^n \qquad (n = 0, 1, 2, \ldots) \qquad \text{and} \quad SD(T) = 1/\lambda$$

10. Geometric from exponential.

a) Show that if T has exponential distribution with rate λ, then int(T), the greatest integer less than or equal to T, has a geometric (p) distribution on $\{0, 1, 2, \ldots\}$, and find p in terms of λ.

b) Let $T_m = \text{int}(mT)/m$, the greatest multiple of $1/m$ less than or equal to T. Show that T has exponential distribution on $(0, \infty)$ for some λ, if and only if for every m there is some p_m such that mT_m has geometric (p_m) distribution on $\{0, 1, 2, \ldots\}$. Find p_m in terms of λ.

c) Use b) and $T_m \leq T \leq T_m + 1/m$ to calculate $E(T)$ and $SD(T)$, from the formulae for the mean and standard deviation of a geometric random variable.

11. Suppose the probability that a given kind of atom disintegrates in any particular microsecond, given that it was alive at the beginning of the microsecond, is $\lambda \times 10^{-6}$ where $\lambda > 0$ is a constant. Let T be the random lifetime of the atom in seconds.

a) Show that the distribution of T is approximately exponential with parameter λ. [*Hint*: Consider $P(T \geq t)$ for t a multiple of 10^{-6}.]

b) What is the chance that the atom has a lifetime of between 1 and 2 seconds?

12. Gamma distribution. Derive the following features of the gamma (r, λ) distribution for all positive r:

a) For $r \geq 1$ the mode (i.e., the value that maximizes the density) is $(r-1)/\lambda$. What if $0 < r < 1$?

b) For $k > 0$, the kth moment of T with gamma (r, λ) distribution is

$$E(T^k) = \frac{1}{\lambda^k}\frac{\Gamma(r+k)}{\Gamma(r)}$$

In particular $E(T) = r/\lambda$.

c) $SD(T) = \sqrt{r}/\lambda$ and Skewness$(T) = 2/\sqrt{r}$.

13. Suppose that under normal operating conditions the operating time until failure of a certain type of component has exponential (λ) distribution for some $\lambda > 0$. And suppose that the random variables representing lifetimes of different components of this type may be regarded as independent.

a) The average lifetime of 10,000 components is found to be 20 days. Estimate the value of λ based on this information.

b) Assuming the exponential lifetime model with $\lambda = 5\%$ per day, let N_d be the number of components among 10,000 components which survive more than d days. Find $E(N_d)$ and $SD(N_d)$ for $d = 10, 20, 30$.

14. Interpretation of the rate. In Exercise 13, the exponential model with $\lambda = 5\%$ per day implies the probability of a component failing in the first day of its use is:
a) exactly 5%; b) approximately 5%, but slightly less;
c) approximately 5%, but slightly more. Without doing any numerical calculations, pick out which of a), b), or c) is true, and explain your choice. Confirm your choice by numerical calculation of the exact probability.

15. Satellite problem. Suppose that a system using one of the components described in Exercise 13, with failure rate 5% per day, is sent up in a satellite together with three spare components of the same type. Assume that as soon as the original component fails, it is replaced by one of the spares, and when that component fails it is replaced by a second spare, and so on. The total operating time of the component plus three spares is then $T_{\text{total}} = T_1 + T_2 + T_3 + T_4$ where T_1 is the operating time of the first component, T_2 is the operating time of the first spare, and so on. Assuming that the satellite launch is successful, and normal operating conditions obtain once the satellite is in orbit, calculate:

a) $E(T_{\text{total}}])$; b) $SD(T_{\text{total}})$; c) $P(T_{\text{total}} \geq 60 \text{ days})$.

16. In the satellite problem of Exercise 15, how many spares would have to be provided to make $P(T_{\text{total}} \geq 60 \text{ days})$ at least 90%?

17. Another type of component has lifetime distribution which is approximately gamma $(2, \lambda)$ with $\lambda = 10\%$ per day.

a) Redo Exercise 15 for this type of component, making similar independence assumptions. After calculating the answers to a) and b), guess without calculation whether the answer to c) should be larger or smaller than under the original assumptions of the satellite problem. Confirm your guess by calculation.

b) Redo Exercise 16 for this type of component.

4.3 Hazard Rates (Optional)

Let T be a positive random variable with probability density $f(t)$, where t ranges over $(0, \infty)$. Think of T as the lifetime of some kind of component. The *hazard rate* $\lambda(t)$ is the probability per unit time that the component will fail just after time t, *given* that the component has survived up to time t. Thus

$$P(T \in dt \,|\, T > t) = \lambda(t)dt$$

where $(T \in dt)$ stands for the event $(t < T \leq t + dt)$ that the component fails in an infinitesimal time interval of length dt just after time t. As usual, this is shorthand for a limit statement:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P\,(T \in (t, t + \Delta t)\,|\,T > t)}{\Delta t}$$

Depending on what lifetime T represents in an application, the hazard rate $\lambda(t)$ may also be called a *death rate* or *failure rate*. For example, T might represent the lifetime of some kind of component. Then $\lambda(t)$ would represent the failure rate for components that have been in use for time t, estimated, for example, by the number of failures per hour among similar components in use for time t.

In practice, failure rates can be estimated empirically as suggested above. Often it is found that empirically estimated hazard rates based on large amounts of data tend to follow a smooth curve. It is then reasonable to fit an ideal model in which $\lambda(t)$ would usually be a continuous function of t. The exponential distribution of the previous section is the simplest possible model corresponding to *constant* failure rate $\lambda(t) = \lambda$ for some $\lambda > 0$. Other distributions with densities on $(0, \infty)$ correspond to time-varying failure rates. The following box summarizes the basic terminology and analytic relationships between the probability density, survival function, and hazard rate.

Formulae (1), (2), and (3) in the box are simply definitions, and (4) is the usual integral for the probability of an interval. Formulae (4) and (5) are equivalent by the fundamental theorem of calculus. Informally, (5) results from

$$\begin{aligned} f(t)dt &= P(T \in dt) \qquad \text{by (1)} \\ &= P(T > t) - P(T > t + dt) \qquad \text{by the difference rule} \\ &= G(t) - G(t + dt) \qquad \text{by (2)} \\ &= -dG(t) \end{aligned}$$

Random Lifetimes

Probability density:	$P(T \in dt) = f(t)dt$	(1)
Survival function:	$P(T > t) = G(t)$	(2)
Hazard rate:	$P(T \in dt \mid T > t) = \lambda(t)dt$	(3)
Survival from density:	$G(t) = \int_t^{\infty} f(u)du$	(4)
Density from survival:	$f(t) = -\dfrac{dG(t)}{dt}$	(5)
Hazard from density and survival:	$\lambda(t) = \dfrac{f(t)}{G(t)}$	(6)
Survival from hazard:	$G(t) = \exp\left(-\int_0^t \lambda(u)du\right)$	(7)

To obtain (6), use $P(A|B) = P(AB)/P(B)$, with $A = (T \in dt)$, $B = (T > t)$. Since $A \subset B$, $AB = A$,

$$\lambda(t)dt = P(T \in dt \mid T > t) = \frac{P(T \in dt)}{P(T > t)} = \frac{f(t)dt}{G(t)}$$

by (1) and (2).

The most interesting formula is (7). To illustrate, in case $\lambda(t) = \lambda$ is constant,

$$\int_0^t \lambda(u)du = \lambda t$$

so (7) becomes the familiar exponential survival probability

$$P(T > t) = e^{-\lambda t} \qquad \text{if } T \text{ has exponential } (\lambda) \text{ distribution.}$$

In general, the exponential of the integral in (7) represents a kind of continuous product obtained as a limit of discrete products of conditional probabilities. This is explained at the end of the section. Formula (7) follows also from (5) and (6) by calculus as you can check as an exercise.

Example 1. **Linear failure rate.**

Problem. Suppose that a component has linear increasing failure rate, such that after 10 hours the failure rate is 5% per hour, and after 20 hours 10% per hour.

(a) Find the probability that the component survives 20 hours.

(b) Calculate the density of the lifetime distribution.

(c) Find the mean lifetime.

Solution. By assumption,

$$\lambda(t) = (t/2)\% = t/200$$

(a) The required probability is by (7)

$$P(\text{survive 20 hours}) = G(20) = \exp\left(-\int_0^{20} \lambda(u)du\right)$$

The integral inside the exponent is

$$\int_0^{20} \frac{u\,du}{200} = \frac{1}{400}u^2\Big|_0^{20} = 1$$

Thus $P(\text{survive 20 hours}) = e^{-1} \approx 0.368$

(b) Put t instead of 20 above to get

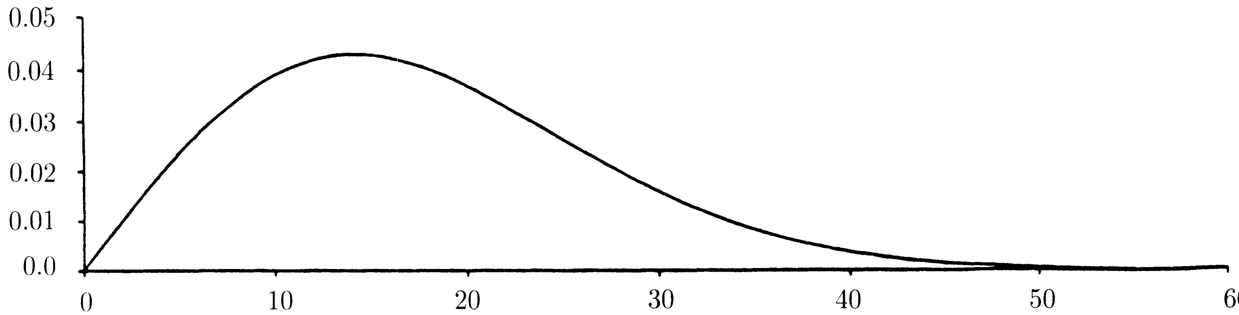
$$G(t) = \exp(-t^2/400)$$

Now by (5)

$$\begin{aligned} f(t) &= -\frac{d}{dt}G(t) \\ &= \frac{t}{200}\exp\left(-\frac{t^2}{400}\right) \end{aligned}$$

You can sketch the density by calculating a few points, as in the following table and graph:

t	0	5	10	15	20
$f(t)$	0	0.023	0.039	0.043	0.037



(c) The mean can be calculated from

$$E(T) = \int_0^\infty t f(t)\, dt$$

but there is a shortcut for examples like this where the survival function $G(t)$ is simpler than the density $f(t)$. This is to use the following formula:

Mean Lifetime from Survival Function

$$E(T) = \int_0^\infty G(t)\, dt \tag{8}$$

This follows by integration by parts from the previous formula for $E(T)$, using $\dfrac{dG(t)}{dt} = -f(t)$. It is the continuous analog of the formula

$$E(T) = \sum_{n=1}^{\infty} P(T \geq n)$$

valid for a random variable T with possible values $0, 1, 2, \ldots$. In the present example, (8) gives

$$E(T) = \int_0^\infty \exp\left(-t^2/400\right) dt \tag{9}$$

Now the problem is that you cannot integrate the function $\exp(-t^2/400)$ in closed form. But you should recognize this integral as similar to the standard Gaussian integral

$$\int_0^\infty e^{-z^2/2}\, dz = \frac{1}{2}\int_{-\infty}^\infty e^{-z^2/2}\, dz = \frac{1}{2}\sqrt{2\pi} = \sqrt{\frac{\pi}{2}}$$

Since $t^2/400 = \dfrac{1}{2}\left(\dfrac{t}{10\sqrt{2}}\right)^2$, make the change of variable $z = t/10\sqrt{2}$, $dz = dt/10\sqrt{2}$, $dt = 10\sqrt{2}dz$ in (9) to get

$$E(T) = 10\sqrt{2}\int_0^\infty e^{-z^2/2}\, dz = 10\sqrt{2}\sqrt{\frac{\pi}{2}} \approx 17.72$$

Derivation of the formula $G(t) = \exp\left(-\int_0^t \lambda(u)\,du\right)$. Recall that the exponential of a sum is the product of exponentials. An integral is a kind of continuous sum, so an exponential of an integral is a kind of continuous product. In this case, the continuous product is a limit of discrete products of conditional probabilities. To see how, divide the time interval $[0, t]$ into a very large number N of very small intervals of length say $\Delta = t/N$. Survival to time t means survival of each of the N successive intervals of length Δ between 0 and t

$$\begin{aligned}
G(t) &= P(T > t) = P(T > N\Delta) \\
&= P(T > \Delta, T > 2\Delta, \cdots, T > N\Delta) \\
&= P(T > \Delta)P(T > 2\Delta \,|\, T > \Delta) \cdots P(T > N\Delta \,|\, T > (N-1)\Delta) \\
&= [1 - P(T \le \Delta)]\,[1 - P(\Delta \le T \le 2\Delta \,|\, T > \Delta)] \cdots \\
&\approx [1 - \Delta\lambda(0)]\,[1 - \Delta\lambda(\Delta)]\,[1 - \Delta\lambda(2\Delta)] \cdots [1 - \Delta\lambda((N-1)\Delta)] \\
&\qquad \text{for small } \Delta \text{, by the definition of } \lambda(t) \\
&\approx e^{-\Delta\lambda(0)} e^{-\Delta\lambda(\Delta)} \cdots e^{-\Delta\lambda((N-1)\Delta)} \\
&\qquad \text{for small } \Delta \text{, by the approximation } 1 - x \approx e^{-x} \text{ for small } x \\
&= \exp\left[-\Delta \sum_{i=0}^{N-1} \lambda(i\Delta)\right] \\
&\approx \exp\left[-\int_0^t \lambda(u)\,du\right] \\
&\qquad \text{for small } \Delta \text{, by a Riemann sum approximation of the integral.}
\end{aligned}$$

As $\Delta \to 0$, the errors in each of the three approximations $\approx$ above tend to zero. So the approximate equality between the first and last expressions not involving Δ must in fact be an exact equality. This is (7).

Note how the exponential appears here, as always, as the limit of a product of more and more factors all approaching 1 in the limit.

Exercises 4.3

1. For T with survival function $G(t) = P(T > t)$, find:
a) $P(T \le b)$; b) $P(a \le T \le b)$.

2. Use the formulae of this section to show that the hazard rate $\lambda(t)$ is constant if and only if the distribution is exponential (λ) for some λ.

3. Business enterprises have the feature that the longer an enterprise has been in business, the less likely it is to fail in the next month. This indicates a decreasing failure rate. One that has been successfully fitted to empirical data of lifetimes of businesses is $\lambda(t) = a/(b + t)$, where a, b, and t are greater than 0. For this $\lambda(t)$:
a) find a formula for $G(t)$; b) find a formula for $f(t)$.

4. Weibull distribution. Show that the following are equivalent:

(i) $\lambda(t) = \lambda\alpha t^{\alpha-1}$ for constants $\lambda > 0$ and $\alpha > 0$

(ii) $G(t) = e^{-\lambda t^{\alpha}}$

(iii) $f(t) = \lambda\alpha t^{\alpha-1}e^{-\lambda t^{\alpha}}$

This is called the *Weibull* distribution with parameters λ and α. This family of distributions is widely used in engineering practice. It can be verified both theoretically and practically that the distribution of the lifetime of a component which consists of many parts, and fails when the first of these parts fails, can be well approximated by a Weibull distribution.

5. Moments of the Weibull distribution. Let T have the Weibull distribution described in Exercise 4. a) Show that $E(T^k) = \Gamma(1 + \frac{k}{\alpha})\,\lambda^{-\frac{k}{\alpha}}$ b) Find $E(T)$ and $Var(T)$.

6. Suppose that a component is subject to failure at constant rate 5% per hour for the first 10 hours in use. After 10 hours the component is subject to additional stress producing a failure rate of 10% per hour.

a) Find the probability that the component survives 15 hours.

b) Calculate and sketch the survival probability function.

c) Calculate and sketch the probability density function.

d) Find the mean lifetime.

7. Second moment from survival function.

a) Show that $E(T^2) = 2\int_0^\infty tG(t)\,dt$

b) Use this formula to calculate the SD of the component in Example 1.

c) If 100 components of this type operate independently, what approximately is the probability that the average lifetime of these components exceeds 20 hours?

8. Suppose the failure rate is $\lambda(t) = at + b$ for $t \geq 0$.

a) For what parameter values a and b does this make sense?

b) Find the formula for $G(t)$. c) Find the formula for $f(t)$.

d) Find the mean lifetime. e) Find the SD of the lifetime.

9. Calculus derivation of $G(t) = \exp\{-\int_0^t \lambda(u)du\}$ (Formula (7)).

a) Use (5) and (6) to show $\lambda(t) = -\frac{d}{dt}\log G(t)$.

b) Now derive (7) by integration from 0 to t.

10. Suppose a component has failure rate $\lambda(t)$ which is an increasing function of t.

a) For $s, t > 0$, is $P(T > s + t|T > s)$ larger or smaller than $P(T > t)$?

b) Prove your answer.

c) Repeat a) and b) for $\lambda(t)$ which is decreasing.

4.4 Change of Variable

Many problems require finding the distribution of some function of X, say $Y = g(X)$, from the distribution of X. Suppose X has density $f_X(x)$, where a subscript is now used to distinguish densities of different random variables. Then provided the function $y = g(x)$ has a derivative dy/dx which does not equal zero on any interval in the range of X, the random variable $Y = g(X)$ has a density $f_Y(y)$ which can be calculated in terms of $f_X(x)$ and the derivative dy/dx. How to do this calculation is the subject of this section.

Linear Functions

To see why the derivative comes in, look first at what happens if you make a *linear* change of variable. For a linear function $y = ax + b$, the derivative is the constant $dy/dx = a$. The function stretches or shrinks the length of every interval by the same factor of $|a|$.

Example 1. Uniform distributions.

Suppose X has the uniform $(0, 1)$ distribution, with density

$$f_X(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Then for $a > 0$, you can see that $Y = aX + b$ has the uniform $(b, b + a)$ density

$$f_Y(y) = \begin{cases} 1/a, & b < y < b + a \\ 0 & \text{otherwise} \end{cases}$$

Similarly, if $a < 0$, then $Y = aX + b$ has the uniform $(b + a, b)$ distribution

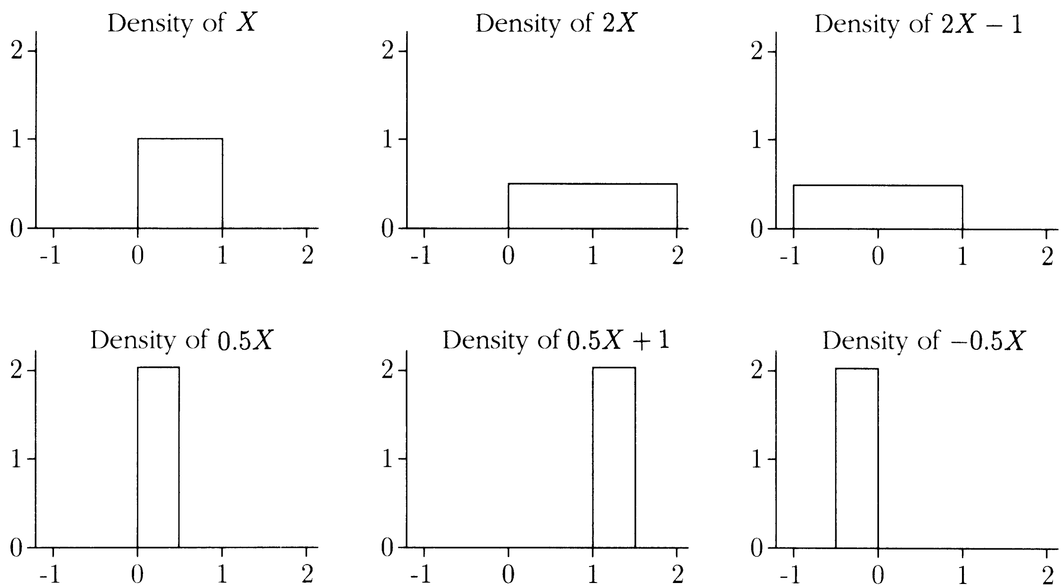
$$f_Y(y) = \begin{cases} 1/|a|, & b + a < y < b \\ 0 & \text{otherwise} \end{cases}$$

You might guess the density of $Y = aX + b$ at y was the density of X at the corresponding point $x = (y - b)/a$. But this must be divided by $|a|$, because the probability density gives probability per unit length, and the transformation from x to $ax + b$ multiplies lengths by a factor of $|a|$:

Linear Change of Variable for Densities

$$f_{aX+b}(y) = \frac{1}{|a|} f_X\left(\frac{y - b}{a}\right)$$

FIGURE 1. **Linear change of variable for uniform densities.** The graphs show the densities of $Y = aX + b$ for various a and b, where X has uniform $(0, 1)$ distribution. Notice how if $a > 1$ the range is spread out and the density decreased. And if $0 < a < 1$ the range is shrunk and the density increased. Adding $b > 0$ shifts to the right by b, and adding $b < 0$ shifts to the left by $-b$.



Example 2. **Normal distributions.**

Take X with standard normal density $\phi(x)$, $a = \sigma > 0$, and $b = \mu$. The linear change of variable formula then gives the density of the normal (μ, σ^2) distribution, displayed on page 267.

One-to-One Differentiable Functions

Let X be a random variable with density $f_X(x)$ on the range (a, b). Let $Y = g(X)$ where g is either strictly increasing or strictly decreasing on (a, b). For example, X might have an exponential distribution on $(0, \infty)$, and Y might be X^2, $\sqrt{X}$, or $1/X$. The range of Y is then an interval with endpoints $g(a)$ and $g(b)$.

The aim now is to calculate the probability density function $f_Y(y)$ for y in the range of Y. For an infinitesimal interval dy near y, the event $(Y \in dy)$ is identical to the event $(X \in dx)$, where dx is an infinitesimal interval near the unique x such that $y = g(x)$. See Figure 2, where each of the two shaded areas represents the probability of the same event

$$P(Y \in dy) = P(X \in dx) \quad \text{where} \quad y = g(x)$$

This identity $P(Y \in dy) = P(X \in dx)$, where $y = g(x)$, makes

$$f_Y(y)dy = f_X(x)dx$$

and so
$$f_Y(y) = f_X(x)\frac{dx}{dy} = f_X(x)\Big/\frac{dy}{dx} \quad \text{where} \quad y = g(x)$$

The case of a decreasing function g is similar except that the calculus derivative dy/dx now has a negative sign. This sign must be ignored because it is only the *magnitude* of the ratio of lengths of small intervals which is relevant. To summarize:

One-to-One Change of Variable for Densities

Let X be a random variable with density $f_X(x)$ on the range (a, b).
Let $Y = g(X)$ where g is either strictly increasing or strictly decreasing on (a, b). The range of Y is then an interval with endpoints $g(a)$ and $g(b)$. And the density of Y on this interval is

$$f_Y(y) = f_X(x)\Big/\left|\frac{dy}{dx}\right| \qquad \text{where} \quad y = g(x)$$

The equation $y = g(x)$ must be solved for x in terms of y, and this value of x substituted into $f_X(x)$ and dy/dx. This will leave an expression for $f_Y(y)$ entirely in terms of y.

Example 3. **Square root of an exponential variable (illustrated by Figure 2)**

Problem. Let X have the exponential density, $f_X(x) = e^{-x} \qquad (x > 0)$
Find the density of $Y = \sqrt{X}$.

Solution. **Step 1.** Find the range of y: here $0 < x < \infty$, $y = \sqrt{x}$, so $0 < y < \infty$.

Step 2. Check the function is one-to-one by solving for x in terms of y: here $x = y^2$

Step 3. Calculate $\frac{dy}{dx}$: here $\frac{dy}{dx} = \frac{d}{dx}\sqrt{x} = \frac{1}{2\sqrt{x}}$

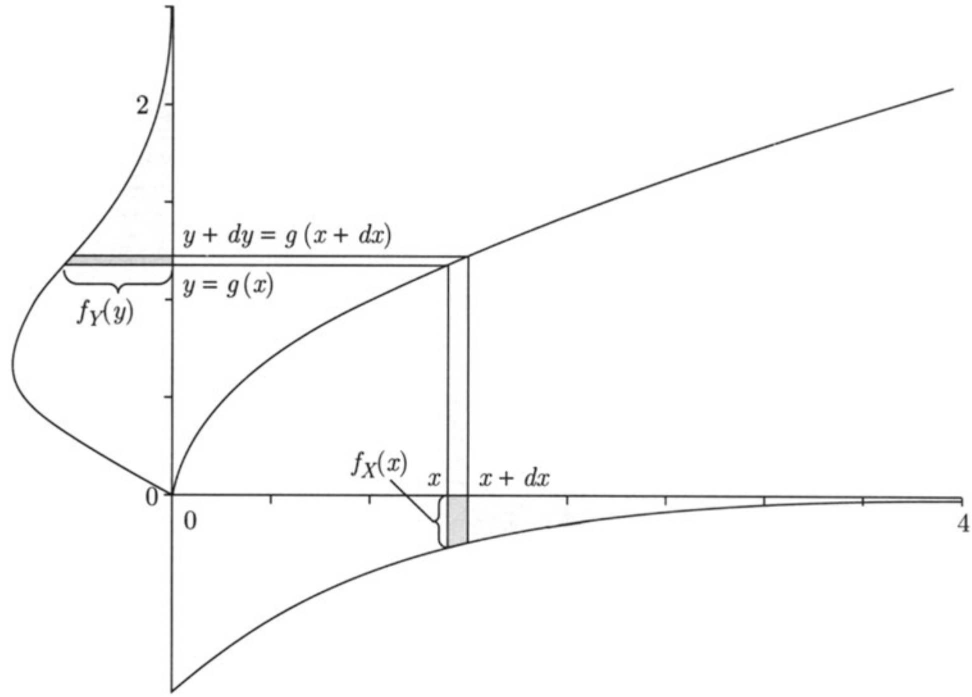
Step 4. Plug density of X and the result of Step 3 into $f_Y(y) = f_X(x)\Big/\left|\frac{dy}{dx}\right|$:

$$f_Y(y) = e^{-x}\Big/\frac{1}{2\sqrt{x}}$$

Step 5. Use result of Step 2 to eliminate x from the right side

$$f_Y(y) = e^{-y^2}\Big/\frac{1}{2\sqrt{y^2}} = 2ye^{-y^2} \qquad (y > 0)$$

FIGURE 2. Change of variable formula for densities. The diagram shows the graph of $y = g(x)$ for the increasing function $g(x) = \sqrt{x}$, $x > 0$. Density $f_X(x)$ is graphed upside down below the x-axis. Density $f_Y(y)$ is graphed on the side of the y-axis. The densities are as in Example 3.



Example 4. **Log of uniform.**

Let X have uniform $(0, 1)$ distribution.

Problem 1. Find the distribution of $Y = -\lambda^{-1} \log(X)$, where $\lambda > 0$.

Solution. This follows the steps of the previous example in a slightly different order. Here $y = -\lambda^{-1} \log x$ has

$$\frac{dy}{dx} = -\frac{1}{\lambda x} < 0 \quad \text{for} \quad 0 < x < 1$$

so y decreases from ∞ to 0 as x increases from 0 to 1. The density of Y is then

$$f_Y(y) = f_X(x) \Big/ \left| \frac{dy}{dx} \right| = 1 \Big/ \frac{1}{\lambda x} = \lambda x$$

where $\quad -\lambda^{-1} \log x = y, \quad$ or $\quad x = e^{-\lambda y}$, so

$$f_Y(y) = \lambda e^{-\lambda y} \qquad (y > 0)$$

Conclusion: Y is exponentially distributed with rate λ.

Discussion. This way of obtaining an exponential variable as a function of a uniform $(0,1)$ variable is a standard method of simulating exponential variables by computer. The next section shows how any distribution on the line can be obtained as the distribution of a function of a uniform variable.

Problem 2. Find the distribution of $-\lambda^{-1}\log(1-X)$, where $\lambda > 0$.

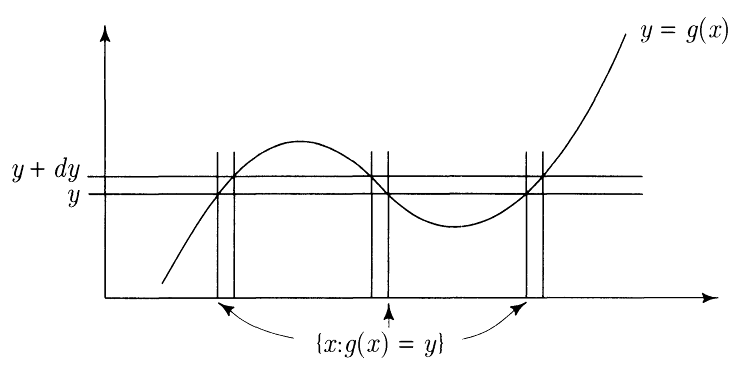
Solution. Clearly the technique used to solve Problem 1 could be repeated. But this is unnecessary. It is intuitively clear (and easy to check) that $X' = 1 - X$ is also a uniform $(0,1)$ random variable, so $-\lambda^{-1}\log(1-X) = -\lambda^{-1}\log(X')$ has the same distribution as $-\lambda^{-1}\log(X)$. Therefore, $-\lambda^{-1}\log(1-X)$ also has exponential (λ) distribution.

Discussion. The justification of the short argument in the last solution is the change of variable principle. This principle, stated for discrete random variables in Section 3.1, is worth restating here. The principle can often be used as in the last example to eliminate calculations by reducing a change of variables problem to one whose solution is already known:

Change of Variable Principle

If X has the same distribution as Y, then $g(X)$ has the same distribution as $g(Y)$, for any function g.

Many-to-one functions. Suppose the function $y = g(x)$ has a derivative that is zero at only a finite number of points. Now some values of y may come from more than one value of x. Consider $Y = g(X)$ for a random variable X. As shown in the diagram, Y will be in an infinitesimal interval dy near y when X is in one of possibly several infinitesimal intervals dx near points x such that $g(x) = y$.



Now

$$P(Y \in dy) = \sum_{\{x: g(x)=y\}} P(X \in dx)$$

This gives

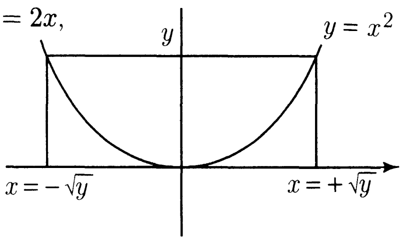
$$f_Y(y) = \sum_{\{x:\, g(x)=y\}} f_X(x) \bigg/ \left|\frac{dy}{dx}\right|$$

Example 5. **Density of the square of a random variable.**

Problem. Suppose X has density $f_X(x)$. Find a formula for the density of $Y = X^2$.

Solution. Here, for $y > 0$, there are two values x such that $x^2 = y$, namely, $x = \sqrt{y}$ and $x = -\sqrt{y}$. Since $dy/dx = 2x$,

$$f_Y(y) = \sum_{\{x=\pm\sqrt{y}\}} f_X(x)/|2x|$$
$$= [f_X(\sqrt{y}) + f_X(-\sqrt{y})]/2\sqrt{y}.$$



Expectation of a function of X. If you just want to calculate the expectation of $Y = g(X)$, it is not necessary to calculate the density of Y, and usually simpler not to. For instance, there is no need to use the linear change of variable formula for densities to calculate $E(Y)$ or $SD(Y)$ for $Y = aX + b$. Instead use the simple scaling rules

$$E(aX + b) = aE(X) + b \qquad \text{and} \qquad SD(aX + b) = |a|SD(X)$$

whenever $E(X)$ or $SD(X)$ are defined. More generally, if $Y = g(X)$, where both X and Y have densities, then

$$E(Y) = \int_{-\infty}^{\infty} y f_Y(y)\, dy = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$$

Often the second integral is easier to evaluate than the first. The equality of the two integrals is the density analog of the basic discrete formula for the expectation of a function of X that was derived in Section 4.1. The equality of integrals can also be checked by the calculus technique of substitution

$$y = g(x), \qquad dy = g'(x)dx.$$

Further Examples

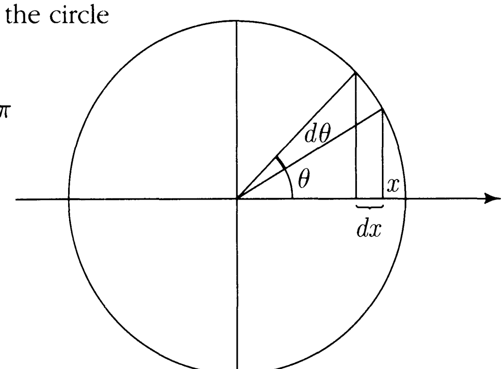
Here are some more geometric problems solved by the same basic technique of finding the probability in an infinitesimal interval by calculus.

Example 6. Projection of a uniform random variable on a circle.

A point is picked uniformly at random from the perimeter of a unit circle.

Problem 1. Find the probability density of X, the x-coordinate of the point.

Solution. From the diagram, since two places on the circle map to one x-value,



$$P(X \in dx) = 2|d\theta|/2\pi = |d\theta|/\pi$$

where $x = \cos\theta, \ 0 < \theta < \pi$. So

$$\frac{dx}{d\theta} = -\sin\theta = -\sqrt{1-x^2}$$

$$\frac{d\theta}{dx} = -\frac{1}{\sqrt{1-x^2}}$$

$$\frac{P(X \in dx)}{dx} = \frac{1}{\pi}\left|\frac{d\theta}{dx}\right| = \frac{1}{\pi\sqrt{1-x^2}} \qquad (-1 < x < 1)$$

Problem 2. Find $E(X)$.

Solution. Easily, $E(X) = 0$, since the density of X is symmetric about 0.

Problem 3. Find the probability density of $Y = |X|$, the absolute value of X.

Solution. Since two x values $+y$ and $-y$, with the same probability density, map to any given value of y with $0 < y < 1$, $P(Y \in dy) = 2 \times P(X \in dy)$, and so

$$f_Y(y) = \frac{2}{\pi\sqrt{1-y^2}} \qquad (0 < y < 1)$$

Problem 4. Find $E(Y)$.

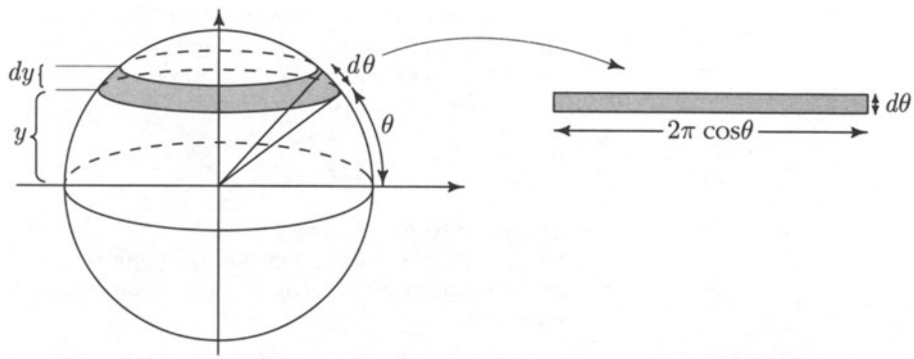
Solution. $E(Y) = \frac{2}{\pi}\int_0^1 \frac{y}{\sqrt{1-y^2}}dy = -\frac{2}{\pi}\sqrt{1-y^2}\Big|_0^1 = \frac{2}{\pi}$

Example 7. Projection of a uniform random variable on a sphere.

Let Θ be the latitude, between $-\pi/2$ and $\pi/2$, of a point chosen uniformly at random on the surface of a unit sphere.

Problem 1. Find the probability density of Θ.

Solution. From the diagram:



$$P(\Theta \in d\theta) = \frac{\text{Indicated Area}}{\text{Total Surface Area}} = \frac{2\pi \cos\theta d\theta}{4\pi}$$

$$f_\Theta(\theta) = \frac{\cos\theta}{2} \qquad (-\frac{\pi}{2} < \theta < \frac{\pi}{2})$$

Problem 2. Let Y be the vertical coordinate of the point on the sphere, between -1 and 1. Find the probability density of Y.

Solution. $P(Y \in dy) = P(\Theta \in d\theta)$ with $y = \sin\theta$, which implies that $dy = \cos\theta\, d\theta$ and

$$P(Y \in dy) = P(\Theta \in d\theta) = f_\Theta(\theta)d\theta = \frac{\cos\theta\, d\theta}{2} = \frac{dy}{2} \qquad (-1 < y < 1)$$

Conclusion: Y has uniform $(-1, 1)$ distribution.

Discussion. This calculation shows that the surface area of the sphere between two parallel planes cutting the sphere depends only on the distance between the planes, and not on exactly how they cut the sphere. This fact was discovered by Archimedes. The formula $4\pi r^2$ for the total surface area, used in Problem 1, is a consequence.

Exercises 4.4

1. Suppose X has an exponential (λ) distribution. What is the distribution of cX for a constant $c > 0$?

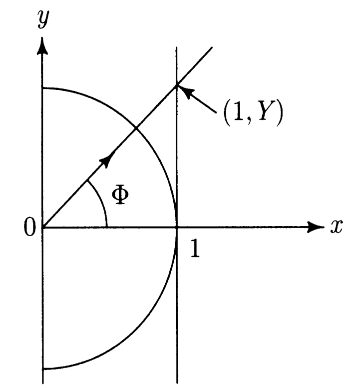
2. **Scaling of gamma distributions.** Show that a random variable T has gamma (r, λ) distribution, if and only if $T = T_1/\lambda$, where T_1 has gamma $(r, 1)$ distribution.

3. Suppose U has uniform $(0, 1)$ distribution. Find the density of U^2.

4. Suppose X has uniform distribution on $(-1, 1)$. Find the density of $Y = X^2$.

5. Suppose X has uniform $[-1, 2]$ distribution. Find the density of X^2.

6. **Cauchy distribution.** Suppose that a particle is fired from the origin in the (x, y)-plane in a straight line in a direction at random angle Φ to the x-axis, and let Y be the y-coordinate of the place where the particle hits the line $\{x = 1\}$. Show that if Φ has uniform $(-\pi/2, \pi/2)$ distribution, then



$$f_Y(y) = \frac{1}{\pi(1+y^2)}$$

This is called the *Cauchy distribution.* Show that the Cauchy distribution is symmetric about 0, but that the expectation of a Cauchy random variable is undefined.

7. Show that if U has uniform $(0, 1)$ distribution, then $\tan(\pi U - \frac{\pi}{2})$ has the Cauchy distribution, as in Exercise 6.

8. **Arcsine distribution.** Suppose that Y has the Cauchy distribution as in Exercise 6. Let $Z = 1/(1+Y^2)$.
   a) Show Z has density
   $$f_Z(z) = \frac{1}{\pi\sqrt{z(1-z)}} \qquad (0 < z < 1)$$
   b) Show $P(Z \le x) = (2/\pi)\arcsin(\sqrt{x}) \quad (0 < x < 1)$.
   c) Find $E(Z)$. d) Find $Var(Z)$.

   [This *arcsine distribution* of Z is the special case $r = s = 1/2$ of the beta(r, s) distribution. This distribution arises naturally in the context of random walks. If $S_n = X_1 + \cdots + X_n$ for X_i with values ± 1 determined by tosses of a fair coin, and L_n is the last time $k \le n$ such that $S_k = 0$, then the limit distribution of L_n/n as $n \to \infty$ is the arcsine distribution. See Feller, *An Introduction to Probability Theory and Its Applications*, Vol. I.]

9. **Weibull distribution.**
   a) Show that if T has the Weibull (λ, α) distribution, with density
   $$f(t) = \lambda\alpha t^{\alpha-1}e^{-\lambda t^\alpha} \qquad (t > 0)$$
   where $\lambda > 0$ and $\alpha > 0$, then T^α has an exponential (λ) distribution. (Note the special case when $\alpha = 1$.)
   b) Show that if U is a uniform $(0, 1)$ random variable, then $T = (-\lambda^{-1}\log U)^{\frac{1}{\alpha}}$ has a Weibull (λ, α) distribution.

10. Let Z be a standard normal random variable. Find formulae for the densities of each of the following random variables:
    a) $|Z|$; b) Z^2; c) $1/Z$; d) $1/Z^2$.

11. Explain how the calculations of Example 7 imply the formula $4\pi r^2$ for the surface area of a sphere of radius r.

4.5 Cumulative Distribution Functions

One way to specify a probability distribution on the line is to say how much probability is at or to the left of each point x. In terms of a random variable X with the given distribution, this probability is a function of x,

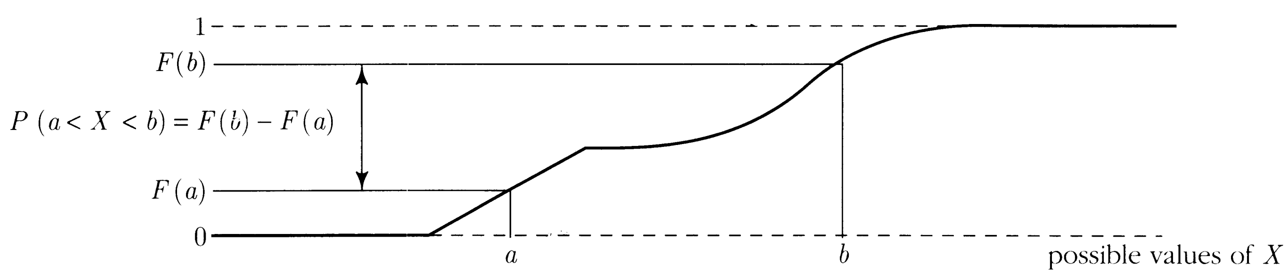
$$F(x) = P(X \leq x)$$

called the *cumulative distribution function (c.d.f.)* of X. For example, the standard normal c.d.f. is the function $F(x) = \Phi(x)$ used in calculations with the normal distribution. But the cumulative distribution function can be defined for any distribution of a random variable X over the line, whether continuous, discrete, or neither.

If you can define or calculate the c.d.f. of X then, by using the rules of probability, you can find the probability of any event determined by X, for example, the probability that X falls in an interval, or the probability that X is an even integer. To clarify terminology, the ***distribution*** of X refers broadly to the assignment of probabilities to all such events determined by X. Technically, this means probabilities defined for a collection of subsets of the line, satisfying the rules of probability, now including the infinite sum rule of Section 3.4. The c.d.f. just gives the probabilities of the intervals $(-\infty, x]$ as a function of the point x.

Interval probabilities. The formula $P(a < X \leq b) = F(b) - F(a)$, a consequence of the difference rule for probabilities, is familiar from the special case of the standard normal c.d.f. Because probabilities must be non-negative, this shows that a c.d.f. $F(x)$ must be a nondecreasing function of x

FIGURE 1. Graph of a continuous c.d.f



The distribution is called ***continuous*** if the c.d.f. is a continuous function. Then it can be shown that

$$P(X = x) = 0 \quad \text{for all } x$$

so it makes no difference in formulae involving the c.d.f. whether inequalities are strict or weak. For example, using the rule of complements,

$$P(X > x) = 1 - F(x) \qquad \text{whatever the distribution of } X$$

$$P(X \geq x) = 1 - F(x) \qquad \text{if the distribution of } X \text{ is continuous}$$

More generally, it can be shown that the c.d.f. determines the probability of every interval, and also the probability of more complicated sets by the addition rule. To summarize:

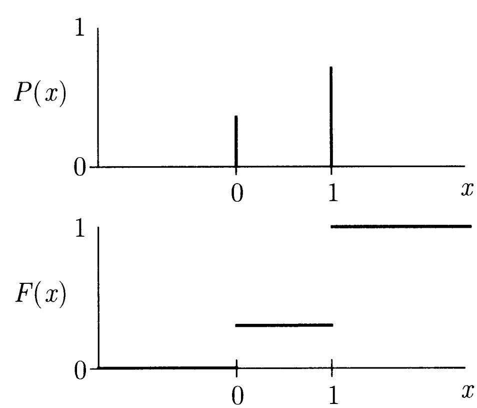
> A probability distribution over the line is completely determined by its c.d.f.

Most distributions of practical interest are either discrete or defined by densities. These two cases will now be discussed in more detail.

Discrete Case

Here is an illustration:

FIGURE 2. Individual probabilities and the c.d.f. for an indicator variable. Consider the c.d.f. of an indicator variable X which is 0 with probability 0.3 and 1 with probability 0.7. The value of $F(x)$ is 0 for $x < 0$ because there is no chance for $X \leq x$ for a negative x. The value of $F(x)$ is 0.3 for $0 \leq x < 1$, because for such an x the event $(X \leq x)$ is the same as the event $(X = 0)$, which has probability 0.3. And the value of $F(x)$ is 1 for $1 \leq x < \infty$, because for these x the event $(X \leq x)$ is certain. Thus $F(x)$ jumps by $0.3 = P(0)$ at $x = 0$ and by $1 - 0.3 = 0.7 = P(1)$ at $x = 1$.



In general, the c.d.f. of a discrete random variable X looks like a staircase with a rise of $P(x) = P(X = x)$ at each possible value x of X:

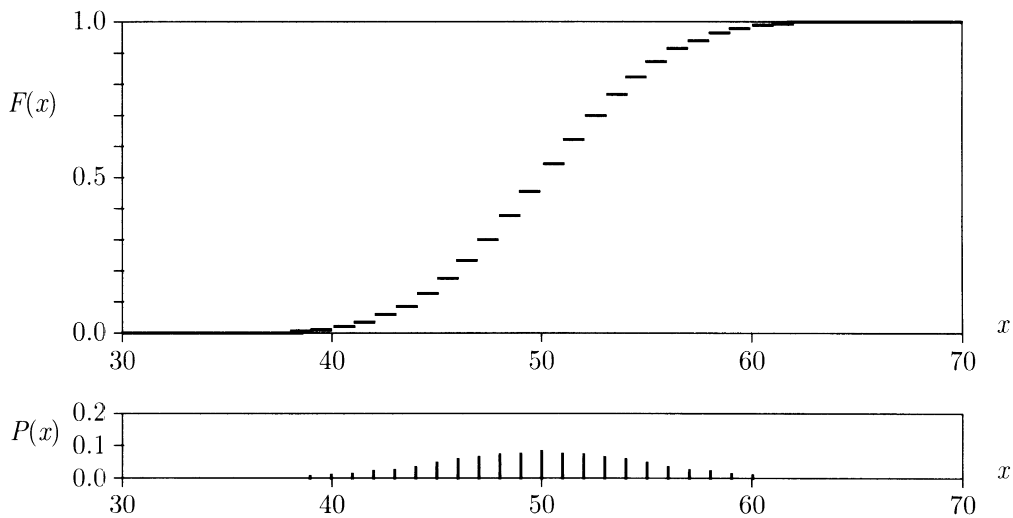
$$F(x) = \sum_{y \leq x} P(y)$$

and $P(x)$ is the jump of the c.d.f. at x:

$$P(x) = F(x) - F(x-)$$

where $F(x-) = P(X < x)$ is the limit of values of F approaching x from the left. Figure 3 gives a more interesting example.

FIGURE 3. **The c.d.f. and individual probabilities for the binomial $(100, 0.5)$ distribution.** Here $F(x)$ is the probability of getting x or less heads in 100 fair coin tosses, $P(x)$ is the probability of exactly x heads. The value of $F(x)$ is simply the sum of values $P(y)$ over all integers y less than or equal to x. Each integer x introduces a new term $P(x)$ into the sum. Thus the graph of F jumps by $P(x)$ at each integer x, and is flat between. Put another way, the probability $P(x)$ of an individual value x shows the *difference* between $F(x)$ and $F(x-)$, where $F(x-) = F(x-1)$ is the value of $F(y)$ for any y in the interval $[x-1, x)$.



Density Case

As usual in this case, sums become integrals. So if X has density $f(x)$, then $F(x)$ is the area under the density function to the left of x

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

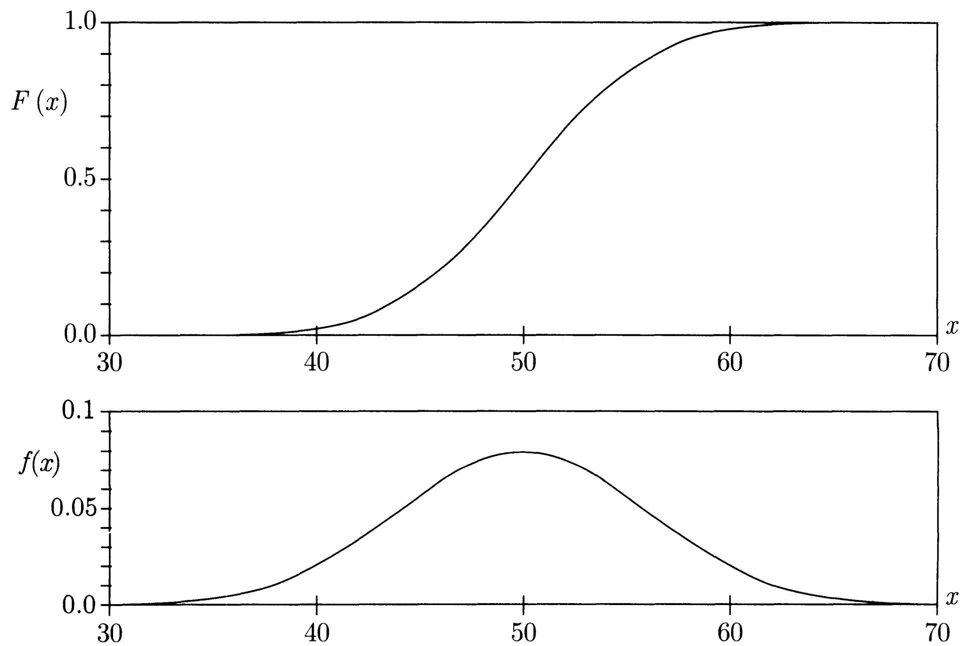
Similarly, discrete differences become derivatives,

$$dF(x) = F(x + dx) - F(x) = P(X \in dx) = f(x)dx$$

$$\text{so} \qquad f(x) = \frac{dF(x)}{dx} = F'(x)$$

That is to say, the density $f(x)$ is the slope at x of the c.d.f. This is an instance of the fundamental theorem of calculus. Conversely, it can be shown that if the c.d.f. is everywhere continuous, and differentiable at all except at perhaps a finite number of points, then the corresponding distribution has density $f(x) = F'(x)$. In this density case, $F(x)$ is a particular choice of an indefinite integral of $f(x)$, namely, the one which vanishes at $-\infty$.

FIGURE 4. **The c.d.f. and density for the normal** $(50, 25)$ **distribution.** This distribution, with mean 50 and variance 25, is the usual normal approximation to the preceding binomial distribution. Its c.d.f. and density are just scale changes of the standard normal ones plotted in Section 2.2.



A distribution with a density can be specified by a formula for the density $f(x)$, or by a formula for the c.d.f. $F(x)$. Either of these functions can be obtained from the other by calculus.

You might think that every continuous distribution has a density, but this turns out not to be so. Still, you don't have to worry about continuous distributions without densities in this course. The famous mathematician Poincare thought such distributions "were invented by mathematicians to confound their ancestors". For a nice picture of one, see Mandelbrot's book, *The Fractal Geometry of Nature*.

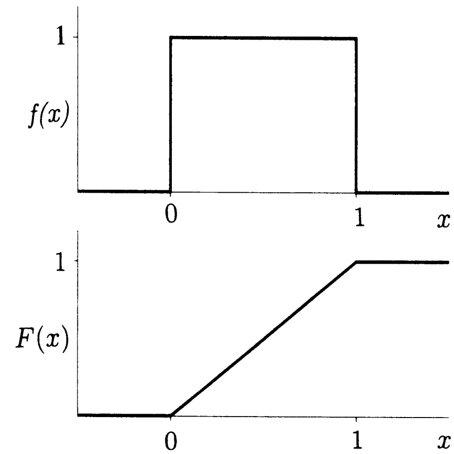
Example 1. **The uniform (0, 1) distribution.**

The density is

$$f(x) = \begin{cases} 1 & \text{for } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

and the c.d.f. is

$$F(x) = \begin{cases} x & \text{for } 0 \le x \le 1 \\ 0 & \text{for } x < 0 \\ 1 & \text{for } x > 1 \end{cases}$$



Here is an application: If U is uniform $(0,1)$, then so is $X = 2\left|U - \frac{1}{2}\right|$, because

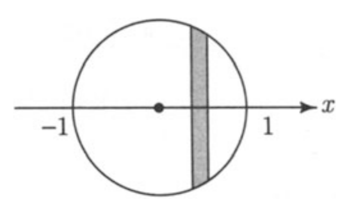
$$\begin{aligned} P(X \le x) &= P\left(2\left|U - \frac{1}{2}\right| \le x\right) \\ &= P\left(\frac{1}{2} - \frac{x}{2} \le U \le \frac{1}{2} + \frac{x}{2}\right) \\ &= F(x) \end{aligned}$$

as defined above. This technique is an alternative to the method of the previous section for calculating the distribution of a function of a random variable.

Example 2. **Uniform on a disc.**

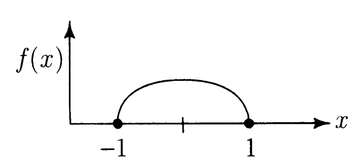
Let (X, Y) be a point chosen uniformly at random from the unit disc $\{(x, y) : x^2 + y^2 \le 1\}$. Calculate the c.d.f. and density function of X.

Solution. It is easiest to find the density function first. Suppose $|x| \le 1$. The event $(X \in dx)$ is shaded in the diagram. For small dx the event in question is approximately a rectangle with height $2\sqrt{1 - x^2}$ and width dx. Dividing by the total area π gives its probability, then dividing by dx gives the density



$$f(x) = \begin{cases} \frac{2}{\pi}\sqrt{1 - x^2} & |x| \le 1 \\ 0 & \text{otherwise} \end{cases}$$

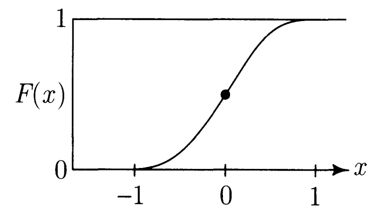
as graphed on the right. This is half an ellipse obtained by rescaling the upper semi-circle. The c.d.f. $F(x)$, which represents the relative area of the disc to the left of x, is now obtained by calculus



$$F(x) = \int_{-1}^{x} f(z)\,dz = \frac{1}{\pi}\int_{-1}^{x} 2\sqrt{1 - z^2}\,dz$$

This is not a very easy integral. Still, because $F(x)$ has derivative $f(x)$ which you know, and $F(x)$ is 0 for $x \le -1$ and 1 for $x \ge 1$, you should be able to sketch the graph of $F(x)$ and see it must have the shape shown below. Some more calculus (or consulting a table of integrals) gives

$$F(x) = \frac{1}{2} + \frac{1}{\pi}\left[x\sqrt{(1-x^2)} + \arcsin x\right] \qquad (|x| \le 1).$$



Maximum and Minimum of Independent Random Variables

Cumulative distribution functions make it easy to find the distribution of the maximum and minimum

$$X_{\text{max}} = \max(X_1, \ldots, X_n) \qquad \text{and} \qquad X_{\text{min}} = \min(X_1, \ldots, X_n)$$

of a collection of independent random variables $X_1, X_2, \ldots, X_n$. Let F_i denote the c.d.f. of X_i, $i = 1, \ldots, n$. The c.d.f. of either the maximum or the minimum of the X's can be written in terms of the individual distribution functions F_i, once you notice the following key facts:

> For any number x:
>
> (a) X_{max} is less than or equal to x if and only if all the X's are less than or equal to x;
>
> (b) X_{min} is greater than x if and only if all the X's are greater than x.

The c.d.f. of the maximum is then

$$\begin{aligned}
F_{\text{max}}(x) &= P(X_{\text{max}} \le x) \qquad (-\infty < x < \infty) \qquad \text{by definition}\\
&= P(X_1 \le x, X_2 \le x, \ldots, X_n \le x) \qquad \text{by (a)}\\
&= P(X_1 \le x)P(X_2 \le x)\cdots P(X_n \le x) \qquad \text{by independence}\\
&= F_1(x)F_2(x)\cdots F_n(x)
\end{aligned}$$

The c.d.f. of the minimum is

$$\begin{aligned}
F_{\min}(x) &= P(X_{\min} \le x) \quad (-\infty < x < \infty) \\
&= 1 - P(X_{\min} > x) \\
&= 1 - P(X_1 > x, X_2 > x, \ldots, X_n > x) \qquad \text{by (b)} \\
&= 1 - (1 - F_1(x))(1 - F_2(x)) \cdots (1 - F_n(x)).
\end{aligned}$$

It is best not to try and memorize these formulae. Just remember (a) and (b), and derive the formulae when you need them.

Example 3. Minimum of independent exponential variables is exponential.

Let $X_1, X_2, \ldots, X_n$ be independent random variables, and suppose X_i has exponential distribution with rate λ_i, $i = 1, \ldots, n$.

Problem. Find the distribution of $X_{\min}$ the minimum of $X_1, \ldots, X_n$.

Solution. For $i = 1, \ldots, n$, the c.d.f. of X_i is

$$F_i(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\lambda_i x} & \text{if } x \ge 0 \end{cases}$$

Since the X's are non-negative, so is their minimum. So $X_{\min}$ has c.d.f.

$$F_{\min}(x) = 0 \qquad (x < 0)$$

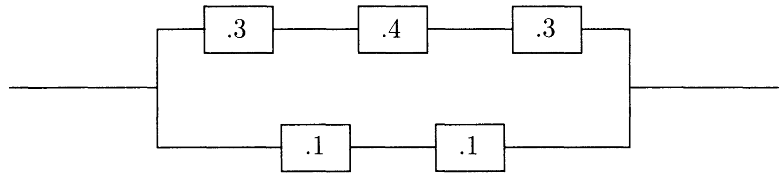
For $x \ge 0$,

$$\begin{aligned}
F_{\min}(x) &= 1 - e^{-\lambda_1 x} e^{-\lambda_2 x} \cdots e^{-\lambda_n x} \\
&= 1 - e^{-(\lambda_1 + \lambda_2 + \cdots + \lambda_n)x}
\end{aligned}$$

This is the c.d.f. of the exponential distribution with rate $\lambda_1 + \lambda_2 + \cdots + \lambda_n$. So the minimum of independent exponential variables with rates λ_i is simply a new exponential variable with rate the sum of the rates λ_i.

Example 4. Expected lifetime of a circuit.

An electrical circuit consists of five components, connected as in the following diagram. The lifetimes of the components, measured in days, have independent exponential distributions with rates indicated in the diagram.



Problem. What is the expected lifetime of the circuit?

Solution. We want $E(L)$, where L denotes the lifetime of the circuit. Let L_{top} and L_{bottom} denote the lifetimes of the top and bottom parts of the circuit. Then L_{top} and L_{bottom} are independent, and

$$L = \max(L_{\text{top}}, L_{\text{bottom}})$$

since the top and bottom parts are linked in parallel.

Now L_{top} is the minimum of three independent exponential lifetimes, since the top consists of three components linked in series. By Example 3, L_{top} has exponential distribution with rate $0.3 + 0.4 + 0.3 = 1$. So the top is expected to last about 1 day. By a similar argument, L_{bottom} has exponential (0.2) distribution, so the bottom is expected to last about $1/0.2 = 5$ days.

Since L is the maximum of L_{top} and L_{bottom}, its c.d.f. is

$$F_L(x) = \begin{cases} 0 & x < 0 \\ (1 - e^{-x})(1 - e^{-0.2x}) & x \geq 0 \end{cases}$$

Since L is a positive random variable

$$E(L) = \int_0^\infty (1 - F_L(x))dx$$

(See Exercise 9 .) For $x \geq 0$,

$$\begin{aligned} F_L(x) &= 1 - e^{-x} - e^{-0.2x} + e^{-1.2x} \\ 1 - F_L(x) &= e^{-x} + e^{-0.2x} - e^{-1.2x} \end{aligned}$$

so

$$E(L) = \int_0^\infty (e^{-x} + e^{-0.2x} - e^{-1.2x})dx = 1 + (1/0.2) - (1/1.2) = 5.17$$

So the circuit is expected to last about 5.17 days.

Note. Once you have the c.d.f. of L, you can, of course, compute its expectation by first differentiating to find the density, then using the density to find the expectation by integration. But that involves more work than the method used here.

Suppose now that in addition to being independent, the X's are continuous random variables with the same density. For example, the X's could be a sequence of random numbers produced by a uniform random number generator. Let f denote the common density function of the X's, and F the common c.d.f. The maximum X_{max} and minimum X_{min} are also continuous random variables, whose densities can be obtained by differentiating their c.d.f.'s

$$F_{\text{max}}(x) = (F(x))^n \qquad (-\infty < x < \infty)$$

$$f_{\text{max}}(x) = \frac{d}{dx}(F(x))^n = n\,(F(x))^{n-1} f(x) \qquad (-\infty < x < \infty)$$

by the chain rule of calculus. Similarly,

$$F_{\text{min}}(x) = 1 - (1 - F(x))^n \qquad (-\infty < x < \infty)$$

$$f_{\text{min}}(x) = n\,(1 - F(x))^{n-1} f(x) \qquad (-\infty < x < \infty)$$

These densities can also be found more directly by a differential calculation explained in the next section.

Percentiles and the Inverse Distribution Function

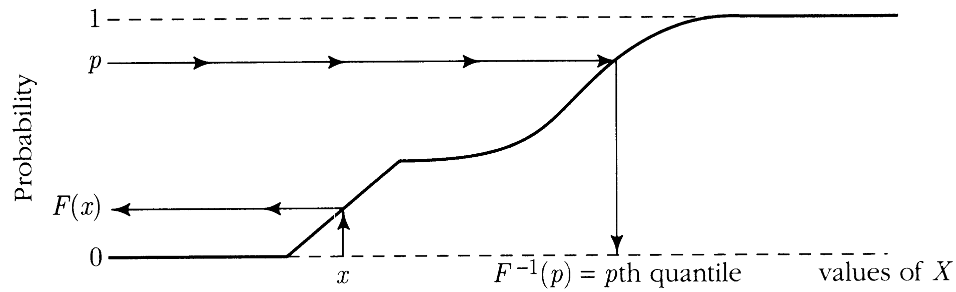
Given a distribution of X and a value x, the c.d.f. $F(x)$ gives the probability that X is less than or equal to x. Often the question gets turned around. For instance: For what value of x is there probability $1/2$ that X is less than or equal to x? Such an x is a *median* of the distribution. More generally, given a probability p, for what x is $P(X \leq x) = p$? By definition of the c.d.f. this x must solve the equation

$$F(x) = p$$

In the case of $F(x)$ given by a formula, the formula can usually be rearranged to express x in terms of p. In general, assuming this equation has a unique solution, as it does for most continuous distributions of interest and $0 < p < 1$, the solution of this equation defines the *inverse c.d.f.*

$$x = F^{-1}(p)$$

FIGURE 5. Relation between a c.d.f. and its inverse.



See Figure 5. This point x, such that $P(X \leq x) = p$, is called the pth *quantile* of the distribution of X. This term is a generalization of the more common *quartile, decile,* and *percentile* in case p is expressed as a multiple of $1/4$, $1/10$, or $1/100$.

Example 5. Finding percentiles.

Problem 1. For the exponential (λ) distribution, find a formula for the pth quantile, $0 < p < 1$.

Solution. Since the c.d.f. is $F(x) = 1 - e^{-\lambda x}$ for $x > 0$, the required point x is found from

$$1 - e^{-\lambda x} = p \qquad \text{so} \qquad x = -\frac{1}{\lambda}\log(1-p)$$

Problem 2. Find the 75th percentile point of the standard normal distribution.

Solution. This is $\Phi^{-1}(0.75)$ where Φ is the standard normal c.d.f. Just as there is no simple formula for Φ, there is none for Φ^{-1}. But numerical values of Φ^{-1} are easily found by backwards lookup in the table of values of Φ. Inspection of the table gives $\Phi(0.67) = 0.7486$ and $\Phi(0.68) = 0.7517$, so $\Phi^{-1}(0.75) \approx 0.675$.

Simulation via Inverse Distribution Function

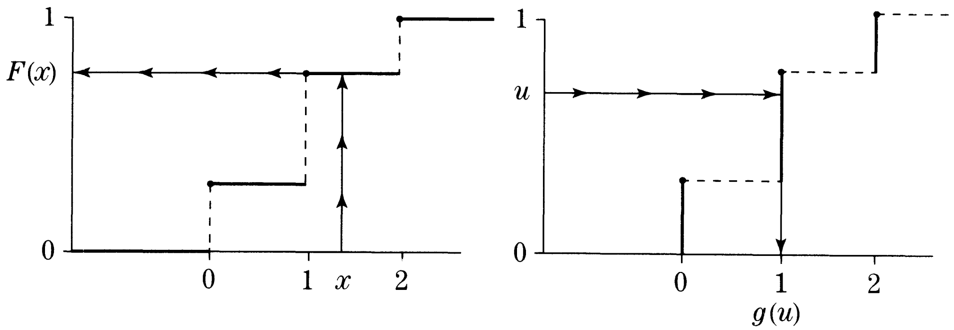
Given a distribution on the line, how can you create random variables with this distribution? This problem arises in computer simulation of random variables. The random number generator on a computer provides a sequence of numbers between 0 and 1, say $U_1, U_2, \ldots$, which behaves in most respects like a sequence of independent uniform $(0, 1)$ random variables. For example, the long-run proportion of values U_i in any subinterval of $[0, 1]$ will be very close to the length of the subinterval. How can these variables be transformed into a sequence simulating independent random variables with some other distribution? The problem is to find a function g such that if U has uniform $(0, 1)$ distribution, then $X = g(U)$ has a prescribed c.d.f., say $F(x)$:

$$P(g(U) \leq x) = F(x) \quad \text{for all } x$$

There are many ways to solve this problem by tricks depending on the desired distribution. Which method is best depends on considerations such as computational efficiency, not discussed here. One method will now be described which works no matter what the required distribution. Here is a simple example to illustrate the method.

Example 6. Simulating a binomial (2, 0.5) random variable.

The left graph shows the required c.d.f. The right graph shows a function g from $(0, 1)$ to $\{0, 1, 2\}$. This graph should be read on its side as a kind of inversion of the graph of the c.d.f. The staircase is the same in both graphs. Imagine U picked at random from the vertical unit interval. Then $g(U) \in \{0, 1, 2\}$ has the required distribution.



In detail, as it would be programmed on a computer, the rule for getting from the uniform $(0, 1)$ variable U to the binomial $(2, 0.5)$ variable $g(U)$ is

$$\begin{array}{lll} \text{if } 0 \le U \le 0.25 & \text{then} & g(U) = 0 \\ \text{if } 0.25 < U \le 0.75 & \text{then} & g(U) = 1 \\ \text{if } 0.75 < U \le 1.0 & \text{then} & g(U) = 2 \end{array}$$

This g does the job because by construction the intervals on which g takes the values 0, 1, and 2 have lengths 0.25, 0.5, and 0.25, respectively, as required by the binomial $(2, 0.5)$ distribution.

Simulation of a discrete distribution. The method of the previous example generalizes easily to any discrete distribution. For example, to get a random variable with discrete distribution on $1, 2, \ldots$ defined by probabilities $p_1, p_2, \ldots$ define

$$g(u) = k \quad \text{if} \quad p_1 + \cdots + p_{k-1} < u \le p_1 + \cdots + p_{k-1} + p_k$$

Then if U has uniform $(0, 1)$ distribution

$$P(g(U) = k) = P(p_1 + \cdots + p_{k-1} < U \le p_1 + \cdots + p_{k-1} + p_k) = p_k$$

since this is the length of the interval of U-values that make $g(U) = k$. This means $g(U)$ has the given discrete distribution.

The inverse distribution function. The function $g(u)$ defined in the discrete case above is always a kind of inverse of the c.d.f. $F(x)$, in the sense that

$$g(F(x)) = x \quad \text{for all possible values } x$$

Check this inverse relation in the example above for $x = 0, 1, 2$. Given any c.d.f. F, not necessarily continuous or strictly increasing, a function g satisfying the above inverse relation can be defined. Because of the inverse relation, $g(u)$ is usually denoted $F^{-1}(u)$, and called the *inverse c.d.f.* In general, the inverse c.d.f. $F^{-1}(u)$ can be defined as the least value x such that $F(x) \geq u$. This function has the following important property:

Inverse c.d.f. Applied to Standard Uniform

For any cumulative distribution function F, with inverse function F^{-1}, if U has uniform $(0, 1)$ distribution, then $F^{-1}(U)$ has c.d.f. F.

To restate this result more intuitively, if you pick a percentage uniformly at random on $(0, 100)$, then take that percentile point in a distribution, you get a random variable with that distribution.

Proof. The discrete case has already been treated. The continuous case is more interesting. Assume, for simplicity, that $F(x)$ is a continuous and strictly increasing function of x. Then $F^{-1}(u)$ is the usual inverse function of $F(x)$, as discussed earlier, and

$$w \leq x \quad \iff \quad F(w) \leq F(x)$$

The events $(F^{-1}(U) \leq x)$ and $(F(F^{-1}(U)) \leq F(x))$ are therefore identical. But since $F(F^{-1}(u)) = u$ for every u in $(0, 1)$, by definition of the inverse function, we can calculate

$$\begin{aligned} P(F^{-1}(U) \leq x) &= P(F(F^{-1}(U)) \leq F(x)) \\ &= P(U \leq F(x)) \\ &= F(x) \quad \text{from the c.d.f. of } U \end{aligned}$$

Thus the random variable $F^{-1}(U)$ has c.d.f. F. □

The method of generating random variables via F^{-1} is efficient computationally in simulations only if F^{-1} turns out to be a fairly simple function to compute, as it is for the uniform distribution on (c, d) for any $c < d$, or the exponential distribution. But F^{-1} is laborious to compute for the normal distribution. In this case it is quicker and nearly as accurate to approximate using the central limit theorem, using, for instance, a standardized sum of 12 independent uniform $(0, 1)$ variables. See also Exercise 5.3.13 for another method of generating normal variables from uniform ones.

Exercises 4.5

1. For the exponential (λ) distribution:

a) Show the c.d.f. is $F(x) = 1 - e^{-\lambda x}$ for $x \geq 0$. b) Sketch this c.d.f. for $\lambda = 1$.

2. Find and sketch the cumulative distribution functions of:

a) the binomial $(3, 1/2)$ distribution;

b) the geometric $(1/2)$ distribution on $\{1, 2, \ldots\}$.

3. Let (X, Y) be as in Example 2.

a) Find f_Y and F_Y. [*Hint*: No calculations required!]

b) Let $R = \sqrt{X^2 + Y^2}$. Sketch the event $\{R \leq r\}$ as a subset of the circle. Deduce a formula for the c.d.f. of R, and check by differentiating that you get the same density for R as in Example 4.1.2.

4. Let X be a random variable with c.d.f. $F(x)$. Find the c.d.f. of $aX + b$ first for $a > 0$, then for $a < 0$.

5. Find the c.d.f. of X with density function $f_X(x) = \frac{1}{2}e^{-|x|}$ $\quad (-\infty < x < \infty)$.

6. Let X be a random variable with c.d.f. $F(x) = x^3$ for $0 \leq x \leq 1$. Find:

a) $P(X \geq \frac{1}{2})$; b) the density function $f(x)$; c) $E(X)$.

d) Let Y_1, Y_2, Y_3 be three points chosen independently and uniformly on the unit interval, and let X be the rightmost point. Show that X has the distribution described above.

7. Let T have the exponential distribution with parameter λ, and let $Y = \sqrt{T}$.

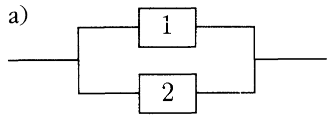
a) Find the density of Y.

b) Find the expectation of Y, correct to two decimal places, for $\lambda = 3$.

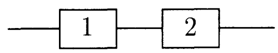
c) A random number generator produces uniform $[0, 1]$ random numbers. How could you use these to generate random numbers which have the distribution of Y?

8. Components in the following series-parallel systems have independent exponentially distributed lifetimes. Component i has mean lifetime μ_i. In each case, find a formula for the probability that the system operates for at least t units of time, and sketch the graph of this function of t in case $\mu_i = i$ for each i.

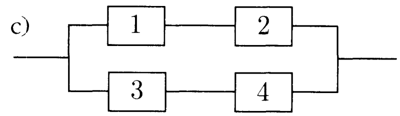
a)



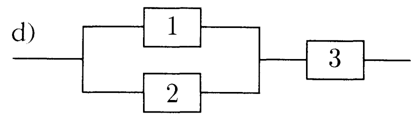
b)



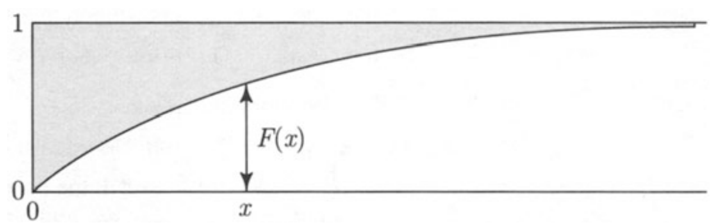
c)



d)



9. Expectation from c.d.f. Let X be a positive random variable, with c.d.f. F, as in the following diagram for example:



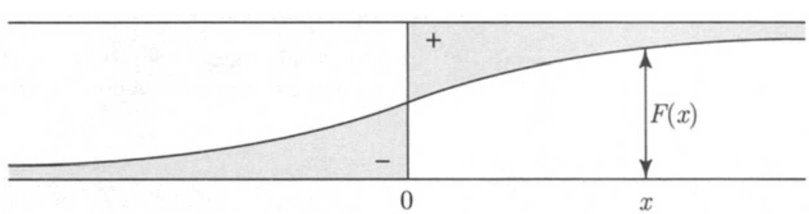
a) Show, using the representation $X = F^{-1}(U)$ for a uniform $[0, 1]$ random variable U, that $E(X)$ can be interpreted as the shaded area above the c.d.f. of X, both for X with a density, and for discrete X. Deduce that

$$E(X) = \int_0^\infty [1 - F(x)]\, dx = \int_0^\infty P(X > x)\, dx$$

b) Deduce that if X has possible values $0, 1, 2, \ldots$, then $E(X) = \sum_{n=1}^{\infty} P(X \geq n)$.

c) Use these formulae to rederive the means of the exponential and geometric distributions.

d) Show that for a random variable X with both positive and negative values (either discrete or with a density), $E(X) = E(X_+) - E(X_-)$ where $X_+ = XI(X > 0)$, and $X_- = (-X)I(X < 0)$, so $E(X)$ is area $(+)$ minus area $(-)$ defined in terms of the c.d.f. as indicated below:

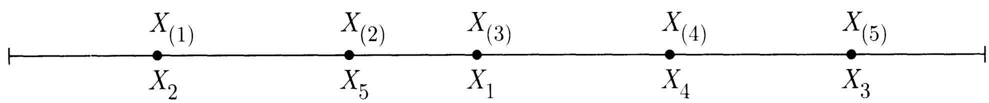


4.6 Order Statistics (Optional)

Let $X_1, X_2, \ldots, X_n$ be random variables. Let $X_{(1)}$ denote the smallest of the X's, $X_{(2)}$ the next smallest, and so on, so that

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

This relabeling of the X's corresponds to arranging them in increasing order, as shown below, for one particular ordering of five values $X_1, \ldots, X_5$.



Notice that

$$X_{(1)} = \min(X_1, \ldots, X_n)$$

$$X_{(n)} = \max(X_1, \ldots, X_n)$$

In general, $X_{(k)}$ is called the kth *order statistic* of $X_1, \ldots, X_n$.

This section deals with properties of order statistics of independent and identically distributed random variables. Beta distributions appear as the distributions of order statistics of independent uniform $(0, 1)$ random variables.

Let $X_1, X_2, \ldots, X_n$ be independent random variables, all with the same density function f and cumulative distribution function F. For example, the X's could be a sequence of random numbers produced by a uniform random number generator. The object is to find a formula for the density of the kth order statistic $X_{(k)}$. This has been done already in Section 4.5 in the case of the maximum $X_{(n)}$ and minimum $X_{(1)}$ by first finding the c.d.f., then differentiating. But here is another argument in these special cases which generalizes more easily. First of all, it can be shown that in a sequence $X_1, \ldots, X_n$ of independent continuous random variables, all n values are distinct with probability 1. Taking this for granted, here is a calculation of the density of the maximum $X_{(n)}$

$$\begin{aligned}
f_{(n)}(x)dx &= P(X_{(n)} \in dx) \\
&= P(\text{one of the } X\text{'s} \in dx, \text{ all others} < x) \\
&= P(X_1 \in dx, \text{ all others} < x) + P(X_2 \in dx, \text{ all others} < x) \\
&\qquad + \cdots + P(X_n \in dx, \text{ all others} < x) \\
&= nP(X_1 \in dx, \text{ all others} < x) \qquad \text{by symmetry} \\
&= nP(X_1 \in dx)P(\text{all others} < x) \qquad \text{by independence} \\
&= nf(x)dx\,(F(x))^{n-1}
\end{aligned}$$

in agreement with the previous calculation in Section 4.5. Similarly,

$$\begin{aligned} f_{(1)}(x)\,dx &= P(X_{(1)} \in dx) \\ &= P(\text{one of the } X\text{'s} \in dx, \text{ all others} > x) \\ &= nf(x)\,dx\,(1 - F(x))^{n-1} \end{aligned}$$

The same method can be used to derive a formula for the density of the kth order statistic of $X_1, \ldots, X_n$. Recall that $X_{(k)}$ is the kth smallest of $X_1, \ldots, X_n$. The density $f_{(k)}(x)$ of $X_{(k)}$ is found as follows. For $-\infty < x < \infty$

$$\begin{aligned} f_{(k)}(x)dx &= P(X_{(k)} \in dx) \\ &= P(\text{one of the } X\text{'s} \in dx, \text{ exactly } k-1 \text{ of the others} < x) \\ &= nP(X_1 \in dx, \text{ exactly } k-1 \text{ of the others} < x) \\ &= nP(X_1 \in dx)P(\text{exactly } k-1 \text{ of the others} < x) \\ &= nf(x)dx\binom{n-1}{k-1}(F(x))^{k-1}(1-F(x))^{n-k} \end{aligned}$$

using the binomial formula. To summarize:

Density of the kth Order Statistic

Let $X_{(k)}$ denote the kth order statistic of $X_1, X_2, \ldots, X_n$, where $X_1, \ldots, X_n$ are independent, identically distributed random variables with common density f and c.d.f. F. The density of $X_{(k)}$ is given by

$$f_{(k)}(x) = nf(x)\binom{n-1}{k-1}(F(x))^{k-1}(1-F(x))^{n-k} \qquad (-\infty < x < \infty)$$

It is best not to memorize the formula, but to remember how it is derived.

Order Statistics of Uniform Random Variables

Let $X_1, \ldots, X_n$ be independent random variables each with uniform distribution on $(0, 1)$. The common density of the X's is

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Their common c.d.f. is

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \le x \le 1 \\ 1 & x > 1 \end{cases}$$

By the boxed formula above, the density of the kth order statistic of the n uniform random variables is

$$f_{(k)}(x) = \begin{cases} n\binom{n-1}{k-1}x^{k-1}(1-x)^{n-k} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Some of these densities are graphed in Figure 1 on the next page.

Notice how as n increases, the density for the minimum gets more concentrated near 0, the density for the maximum gets more concentrated near 1, and the density for the middle value of the X's gets more concentrated near $1/2$. This is what you would expect intuitively.

Notice also the functional form of the density: a constant, times x raised to a power, times $1 - x$ raised to a power. This simple form for a density on $(0, 1)$ appears in many settings. Here is a general definition:

Beta (*r*,*s*) Distribution

For $r, s > 0$, the *beta* (r, s) distribution on $(0, 1)$ is defined by the density

$$\frac{1}{B(r,s)}x^{r-1}(1-x)^{s-1} \qquad (0 < x < 1)$$

where

$$B(r,s) = \int_0^1 x^{r-1}(1-x)^{s-1}dx$$

is the normalizing constant which makes the density integrate to 1.

Viewed as a function of r and s, $B(r, s)$ is called the *beta function*.

A comparison of the last two boxes shows the following:

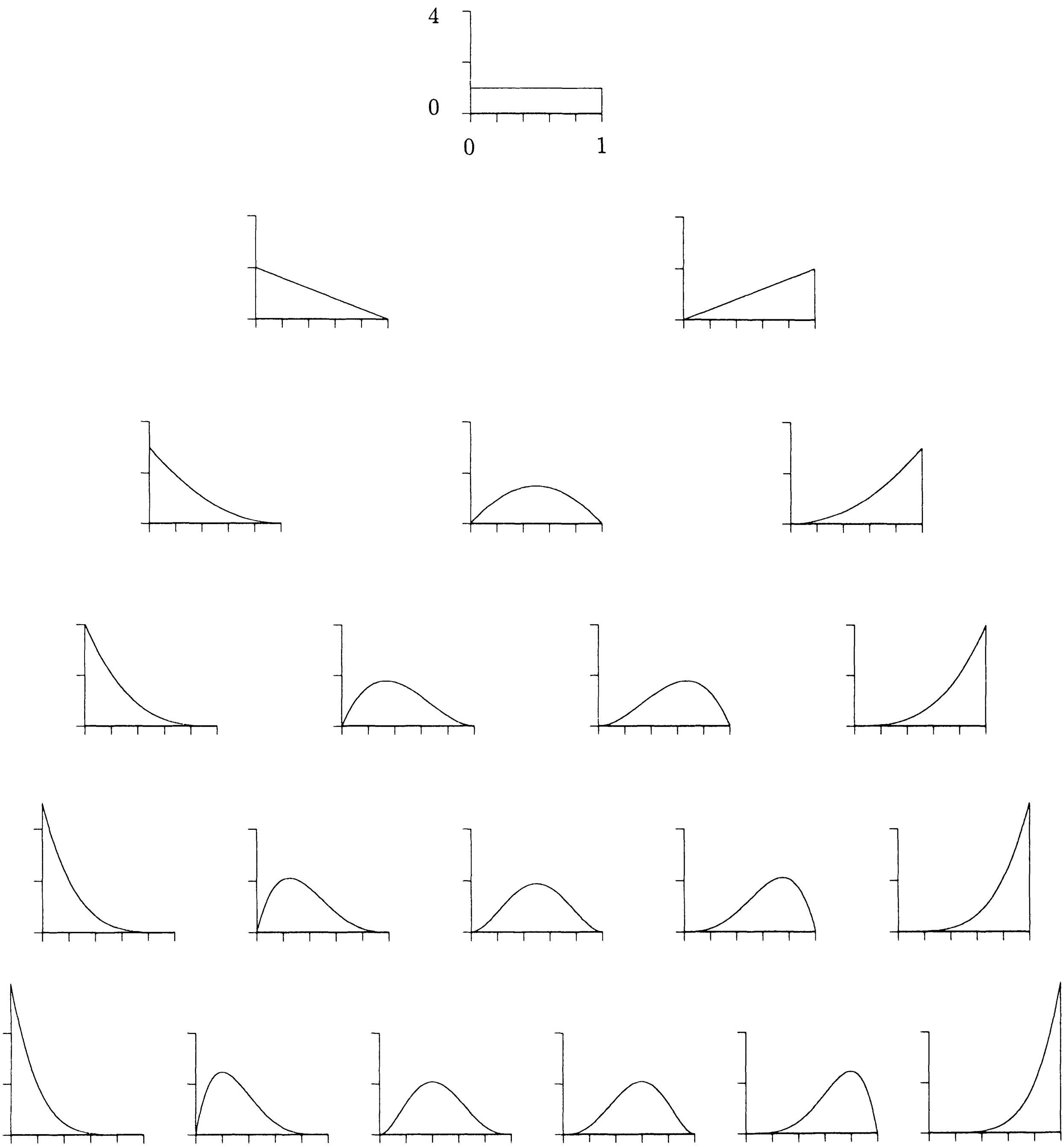
Beta Distribution of Uniform Order Statistics

The kth order statistic of n independent uniform $(0, 1)$ random variables has beta $(k, n - k + 1)$ distribution.

A nice corollary of the formula for the density of $X_{(k)}$ derived above is that for integers r and s, the beta function $B(r, s)$ is evaluated. Since $f_{(k)}$ is a density it must integrate to 1 over $[0, 1]$. So

$$\int_0^1 x^{k-1}(1-x)^{n-k}dx = \frac{1}{n\binom{n-1}{k-1}} = \frac{(k-1)!(n-k)!}{n!}$$

FIGURE 1. Densities of order statistics of independent uniform variables. For $n = 1, 2, \ldots, 6$ and $k = 1, 2, \ldots, n$, the density of the kth order statistic of n independent uniform $(0, 1)$ random variables, which is the beta density with parameters k and $n - k + 1$, is plotted as the kth graph in the nth row of the diagram.



Substitute $r = k$ and $s = n-k+1$ and recall that $\Gamma(r) = (r-1)!$ for positive integers r to get the following result for integers r and s:

Evaluation of the Beta Integral

For positive r and s

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

The beta (r, s) distribution is defined, and the above evaluation of the beta integral is valid, for all positive r and s, not necessarily integers. See Section 5.4, especially Exercise 5.4.19 for a proof of this and explanation of the connection between the beta and gamma distributions.

Moments of the beta distribution. The expectation and variance of a beta random variable with integer parameters are now easy to calculate. If X has beta distribution with positive integer parameters r and s,

$$\begin{aligned}
E(X) &= \int_0^1 x \cdot \frac{1}{B(r,s)} x^{r-1}(1-x)^{s-1}dx \\
&= \frac{1}{B(r,s)} \int_0^1 x^{(r+1)-1}(1-x)^{s-1}dx \\
&= \frac{B(r+1,s)}{B(r,s)} \\
&= \frac{r!(s-1)!}{(r+s)!} \cdot \frac{(r+s-1)!}{(r-1)!(s-1)!} \\
&= \frac{r}{r+s}
\end{aligned}$$

$E(X^2)$ can be calculated in the same way, and used to find a formula for the variance of X. This is left as an exercise.

The kth order statistic of n independent uniform $(0,1)$ random variables has beta distribution with parameters k and $n-k+1$, so

$$E(X_{(k)}) = \frac{k}{n+1}$$

Thus the smallest of four uniform random numbers is expected to be around $1/5$, the next smallest around $2/5$, the third smallest around $3/5$, and the largest around

4/5. In other words, if you think of picking four points at random from $[0,1]$ as cutting the interval into five pieces

all the pieces are expected to have the same length. In fact, more is true: It can be shown that when an interval is split at random like this by any number of independent uniform random points, the length of each piece has the same beta distribution as the length of the first piece. See Chapter 6 Review Exercise 32.

Exercises 4.6

1. Four people agree to meet at a cafe at noon. Suppose each person arrives at a time normally distributed with mean 12 noon and SD 5 minutes, independently of all the others.

a) What is the chance that the first person to arrive at the cafe gets there before 11:50?

b) What is the chance that some of the four have still not arrived at 12:15?

c) Approximately what is the chance that the second person to arrive gets there within ten seconds of noon?

2. Let X have beta (r, s) distribution.

a) Find $E(X^2)$, and use the formula for $E(X)$ given in this section to find $Var(X)$.

b) Find a formula for $E(X^k)$, for integers $k \geq 1$.

3. Let $U_{(1)}, \ldots, U_{(n)}$ be the values of n independent uniform $(0,1)$ variables arranged in increasing order. Let $0 \leq x < y \leq 1$. Find simple formulae for:

a) $P(U_{(1)} > x \text{ and } U_{(n)} < y)$; b) $P(U_{(1)} > x \text{ and } U_{(n)} > y)$;

c) $P(U_{(1)} < x \text{ and } U_{(n)} < y)$; d) $P(U_{(1)} < x \text{ and } U_{(n)} > y)$;

e) $P(U_{(k)} < x \text{ and } U_{(k+1)} > y)$ for $1 \leq k \leq n-1$;

f) $P(U_{(k)} < x \text{ and } U_{(k+2)} > y)$ for $1 \leq k \leq n-2$.

4. Let $X = \min(S, T)$ and $Y = \max(S, T)$ for independent random variables S and T with a common density f. Let Z denote the indicator of the event $S < T$.

a) What is the distribution of Z?

b) Are X and Z independent? Are Y and Z independent? Are (X, Y) and Z independent?

c) How can these conclusions be extended to the order statistics of three or more independent random variables with the same distribution?

5. C.d.f. of the beta distribution for integer parameters.

a) Let $X_1, X_2, \ldots, X_n$ be independent uniform $(0,1)$ random variables, and let $X_{(k)}$ be the kth order statistic of the X's. Find the c.d.f. of $X_{(k)}$ by expressing the event $X_{(k)} \leq x$ in terms of the number of X_i that are $\leq x$.

b) Use a) to show that for positive integers r and s, the c.d.f. of the beta (r, s) distribution is given by

$$\sum_{i=r}^{r+s-1} \binom{r+s-1}{i} x^i (1-x)^{r+s-i-1} \qquad (0 \leq x \leq 1)$$

c) Expand the power of $(1-x)$ in the beta density using the binomial theorem, and then integrate, to obtain the following alternative formula for the c.d.f. of the beta (r, s) distribution:

$$\frac{x^r}{B(r,s)} \sum_{i=0}^{s-1} \binom{s-1}{i} (-1)^i x^i / (r+i) \qquad (0 \leq x \leq 1)$$

[Equating the results of these two calculations yields an algebraic identity that is not easy to prove directly.]

Continuous Distributions: Summary

For a random variable X with probability density $f(x)$:

Differential formula: $P(X \in dx) = f(x)dx$.

Integral formula: $P(a \leq X \leq b) = \int_a^b f(x)dx$.

Interpretation: $f(x)$ is the chance per unit length for values of X near x.

Properties of $f(x)$: Non-negative, total integral 1.

Expectation of a function g of X

$$E\left((g(X)\right) = \int_{-\infty}^{\infty} g(x)f(x)dx \quad \text{provided} \quad \int_{-\infty}^{\infty} |g(x)|f(x)dx < \infty$$

Uniform, exponential, normal distributions: See Distribution Summaries.

Hazard rates

Let T be a positive random variable with probability density f. Think of T as the lifetime of a component. The *hazard rate* (or *failure rate*, or *death rate*) function $\lambda(t)$ is the probability per unit time that the component will fail just after time t, given that it has survived up to time t

$$P(T \in dt \,|\, T > t) = \lambda(t)\, dt$$

For relations between λ and the density, survival function, etc., of T, see the table "Random Lifetimes" on page 297.

Expectation from the survival function: For a non-negative random variable T,

$$E(T) = \int_0^{\infty} G(t)dt$$

where $G(t) = P(T > t)$ is the survival function of T.

One-to-one change of variable for densities

Let X be a random variable with density $f_X(x)$ in the range (a, b).

Let $Y = g(X)$ where g is either strictly increasing or strictly decreasing on (a, b). The range of Y is then an interval with endpoints $g(a)$ and $g(b)$. And the density of Y on this interval is

$$f_Y(y) = f_X(x) \Big/ \left|\frac{dy}{dx}\right| \qquad \text{at} \quad x = g^{-1}(y)$$

where dy/dx is the derivative of $y = g(x)$, and g^{-1} is the inverse function of g.

Linear change of variable for densities:

$$f_{aX+b}(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

Change of variable principle: If X has the same distribution as Y, then $g(X)$ has the same distribution as $g(Y)$, for any function g.

Cumulative distribution function of X: $F(x) = P(X \leq x)$

If the distribution has a *density* $f(x)$, then

$$F(x) = \int_{-\infty}^{x} f(y)\, dy$$

and the density function at x is the derivative of the c.d.f. at x

$$f(x) = \frac{dF(x)}{dx} = F'(x)$$

provided $F'(x)$ is continuous at x.

Percentiles

The kth percentile point of a distribution is the value x such that $F(x) = k/100$, written $x = F^{-1}(k/100)$, where F^{-1} is the *inverse c.d.f.*

Transformation by the inverse c.d.f.

If U has uniform $(0,1)$ distribution, then $F^{-1}(U)$ has c.d.f. F.

Order statistics

If $X_1, \ldots, X_n$ are independent with common density f and c.d.f. F, then the kth *order statistic* $X_{(k)}$, that is, the kth smallest value among the $X_1, \ldots, X_n$, has density

$$f_{X_{(k)}}(x) = nf(x)\binom{n-1}{k-1}(F(x))^{k-1}(1-F(x))^{n-k}$$

If the X_i have uniform $(0,1)$ distribution, then $X_{(k)}$ has beta $(k, n-k+1)$ distribution.

Review Exercises

1. Suppose atoms of a given kind have an exponentially distributed lifetime with rate λ. Let X_t be the number of atoms still present at time $t \geq 0$, starting from $X_0 = n$. Find formulae in terms of n, t, and λ for a) $E(X_t)$; b) $Var(X_t)$.

2. Find the constant c which makes the function $f(x) = c(x+x^2)$ for $0 < x < 1$ the density of a probability distribution on $(0, 1)$. Find the corresponding c.d.f. $F(x)$. Sketch the graphs of $f(x)$ and $F(x)$. Find the expectation μ and standard deviation σ of a random variable X with this distribution. Mark the points μ, $\mu + \sigma$ on your graphs.

3. Let Y_1, Y_2, and Y_3 be three points chosen independently and uniformly from $(0, 1)$, and let X be the rightmost (largest) point. Find the c.d.f., density function, and expectation of X.

4. Let X be a random variable with density $f(x) = 0.5e^{-|x|}$ $(-\infty < x < \infty)$. Find:

a) $P(X < 1)$; b) $E(X)$ and $SD(X)$; c) the c.d.f. of X^2.

5. An ambulance station, 30 miles from one end of a 100-mile road, services accidents along the whole road. Suppose accidents occur with uniform distribution along the road, and the ambulance can travel at 60 miles an hour. Let T minutes be the response time (between when accident occurs and when ambulance arrives).

a) Find $P(T > 30)$.

b) Find $P(T > t)$ as a function of t. Sketch its graph.

c) Calculate the density function of T.

d) Calculate the mean and standard deviation of T.

e) What would be a better place for the station? Explain.

6. Electrical components of a particular type have exponentially distributed lifetimes with mean 48 hours. In one application the component is replaced by a new one if it fails before 48 hours, and in case it survives 48 hours it is replaced by a new one anyway. Let T represent the potential lifetime of a component in continuous use, and U the time of such a component in use with the above replacement policy. Sketch the graphs of:

a) the c.d.f. of T; b) the c.d.f. of U. Is U discrete, continuous, or neither?

c) Find $E(U)$. [*Hint*: Express U as a function of T.]

d) Does the replacement policy serve any good purpose? Explain.

7. **Two-sided exponential distribution.** Suppose X with range $(-\infty, \infty)$ has density $f(x) = \alpha e^{-\beta|x|}$ where α and β are positive constants.

a) Express α in terms of β. b) Find $E(X)$ and $Var(X)$ in terms of β.

c) Find $P(|X| > y)$ in terms of y and β. d) Find $P(X \leq x)$ in terms of x and β.

8. **The principle of ignoring constants.** In calculating the density of a random variable X, a quick method is to ignore constant factors as you go along, to end up with an answer of the form $P(X \in dx)/dx = f(x)$ with $f(x) = c\,h(x)$ for a known function $h(x)$ and mystery constant c. The point is that provided your calculation has been consistent with the basic rules of probability, the density of X must integrate to 1, so

$$\int c\,h(x)\,dx = \int f(x)\,dx = 1$$

a) Use this identity to evaluate c in terms of $\int h(x)\,dx$.

b) You can often recognize at the end of a calculation that $h(x) = c_1 f_1(x)$ for some named density $f_1(x)$ (e.g. one of the densities displayed in the table on page 477 and some constant c_1. Deduce that then $c = 1/c_1$ and $f(x) = f_1(x)$.

Use this method to evaluate the constant factor c that makes $c\,h(x)$ a probability density for each of the following functions $h(x)$, assumed to be zero except for the indicated range of x, and find $E(X)$ and $Var(X)$ in each case from the table on page 477.

c) $e^{-\frac{1}{2}x^2}$ $(-\infty < x < \infty)$ d) x $(0 < x < 1)$

e) 1 $(0 < x < 10)$ f) e^{-5x} $(x > 0)$

9. Use the method of Exercise 8 to evaluate the constant factor c that makes $f(x) = c\,h(x)$ a probability density for each of the following functions $h(x)$, assumed to be zero except for the indicated range of x, where a and b are positive parameters. Also find $E(X)$ and $Var(X)$ in each case:

a) $e^{-(x-a)^2}$ $(-\infty < x < \infty)$; b) $e^{-(x-a)^2/b^2}$ $(-\infty < x < \infty)$;

c) $e^{-ax}x^5 (x > 0)$; d) $e^{-a|x|}$ $(-\infty < x < \infty)$;

e) $x^7(1-x)^9$ $(0 < x < 1)$; f) $x^7(b-x)^9$ $(0 < x < b)$.

10. Evaluate the following integrals:

a) $\displaystyle\int_0^\infty e^{-x^2}dx$; b) $\displaystyle\int_0^1 e^{-x^2}dx$; c) $\displaystyle\int_0^\infty x\,e^{-x^2}dx$; d) $\displaystyle\int_0^\infty x^2\,e^{-x^2}dx$.

11. Evaluate the following integrals:

a) $\displaystyle\int_0^\infty z^3e^{-z^2}dz$; b) $\displaystyle\int_0^\infty x^7e^{-2x}dx$; c) $\displaystyle\int_0^{100} x^2(100-x)^2dx$.

12. A Geiger counter is recording background radiation at an average rate of 2 hits per minute; the hits may be modeled as a Poisson process. Let T be the time (in minutes) of the third hit after the machine is switched on. Find $P(1 < T < 3)$.

13. Local calls are coming into a telephone exchange according to a Poisson process with rate λ_{loc} calls per minute. Independently of this, long-distance calls are coming in at a rate of λ_{dis} calls per minute. Write down expressions for probabilities of the following events:

a) exactly 5 local calls and 3 long-distance calls come in a given minute;

b) exactly 50 calls (counting both local and long distance) come in a given three-minute period;

c) starting from a fixed time, the first ten calls to arrive are local.

14. Particles arrive at a Geiger counter according to a Poisson process with rate 3 per minute.

a) Find the chance that less than 4 particles arrive in the time interval 0 to 2 minutes.

b) Let T_n minutes denote the arrival time of the nth particle. Find

$$P(T_1 < 1,\ T_2 - T_1 < 1,\ T_3 - T_2 < 1)$$

c) Find the conditional distribution of the number of arrivals in 0 to 2 minutes, given that there were 10 arrivals in 0 to 4 minutes. Recognize this as a named distribution, and state the parameters.

15. Two Geiger counters record arrivals of radioactive particles. Particles arrive at Counter I according to a Poisson process, at an average rate of 3 per minute. Independently, particles arrive at Counter II at an average rate of 4 per minute, also according to a Poisson process. In a particular one-minute period, the counters recorded at total of 8 arrivals. Given this, what is the chance that each counter recorded four arrivals?

16. Cars arrive at a toll booth according to a Poisson process at a rate of 3 arrivals per minute.

a) What is the probability that the third car arrives within three minutes of the first car?

b) Of the cars arriving at the booth, it is known that over the long run 60% are Japanese imports. What is the probability that in a given ten-minute interval, 15 cars arrive at the booth, and 10 of these are Japanese imports? State your assumptions clearly.

17. Show that T has exponential distribution with rate λ if and only if

$$P(T \leq t) = 1 - e^{-\lambda t} \quad \text{for all} \quad 0 \leq t < \infty$$

18. Bus lines A, B, and C service a particular stop. Suppose the lines come as independent Poisson processes with rates λ_A, λ_B, and λ_C buses per hour respectively. Find expressions for the following probabilities:

a) exactly one A bus, two B buses, and one C bus come to the stop in a given hour;

b) a total of 7 buses come to the stop in a given two hour time period;

c) starting from a fixed time, the first A bus arrives after t hours.

19. A piece of rock contains 10^{20} atoms of a particular substance, each with a half-life of one century. How many centuries must pass before:

a) most likely about 100 atoms remain;

b) there is about a 50% chance that at least one atom remains.

20. **Hazard rates (refers to Section 4.3)**. Suppose a component with constant failure rate λ is backed up by a second similar component. When the first component burns out the second is installed, and is thereafter subject to failure at the same rate λ, independently of when it was installed and how long it has been in use. Let T be the total time to failure of both components. Find for T:

a) the density function; b) the survival function; c) the hazard rate function.

d) Suppose $\lambda = 1$ per hour. Given $T \geq 2$ hours, what is the approximate probability of failure in the next minute?

21. Suppose R_1 and R_2 are two independent random variables with the same density function $f(x) = x \exp\left(-\frac{1}{2}x^2\right)$ for $x \geq 0$. Find

(a) the density of $Y = \min\{R_1, R_2\}$; b) the density of Y^2; c) $E(Y^2)$.

22. Let X be a random variable that has a uniform distribution on the interval $(0, a)$.

a) Find the c.d.f. of $Y = \min(X, a/2)$.

b) Is the distribution of Y continuous? Explain. c) Find $E(Y)$.

23. An earthquake of magnitude M releases energy X such that $M = \log X$. For earthquakes of magnitude greater that 3, suppose that $M - 3$ has an exponential distribution with mean 2.

a) Find $E(M)$ and $Var(M)$ for an earthquake of magnitude greater than 3.

b) For an earthquake as in part a), find the density of X.

c) Consider two earthquakes, both of magnitude greater than 3. What is the probability that the magnitude of the smaller earthquake is greater than 4? Assume that the magnitudes of the two earthquakes are independent of each other.

24. Suppose stop lights at an intersection alternately show green for one minute, red for one minute (ignore amber). Suppose a car arrives at the lights at a time distributed uniformly at random relative to this cycle. Let X be the delay of the car at the lights, neglecting any delay due to traffic congestion.

a) Find a formula for the c.d.f. of X, and sketch its graph.

b) Is X discrete, continuous, or neither? c) Find $E(X)$ and $Var(X)$.

d) Suppose that the car encounters a succession of ten such stop lights. Make an independence assumption and use the normal approximation to estimate the probability that the car will be delayed more than four minutes by the lights.

25. Suppose the random variable U is distributed uniformly on the interval $(0, 1)$. Find:

a) the density of the random variable $Y = \min\{U, 1 - U\}$ (indicate where the density is positive);

b) the density of $2Y$; c) $E(Y)$ and $Var(Y)$.

26. Suppose that the weight W_t of a tumor after time t is modeled by the formula $W_t = Xe^{tY}$ where X and Y are independent random variables, X distributed according to a gamma distribution with mean 2 and variance 1, and Y distributed uniformly on 1 to 1.5. Find formulae for: a) $E(W_t)$; b) $SD(W_t)$.

27. Suppose $U_1, U_2, \ldots$ are independent uniform $(0, 1)$ variables, and let N be the first $n \geq 2$ such that $U_n > U_{n-1}$. Show that for $0 \leq u \leq 1$:

a) $P(U_1 \leq u \text{ and } N = n) = \dfrac{u^{n-1}}{(n-1)!} - \dfrac{u^n}{n!} \quad n \geq 2;$

b) $P(U_1 \leq u \text{ and } N \text{ is even}) = 1 - e^{-u}$.

c) $E(N) = e$.

28. A point is chosen uniformly at random from the circumference of a circle of diameter 1. Let X be the length of the chord joining the random point to an arbitrary fixed point on the circumference. Find: a) the c.d.f. of X; b) $E(X)$; c) $Var(X)$.

29. A gambling game works as follows. A random variable X is produced; you win \$1 if $X > 0$ and you lose \$1 if $X < 0$. Suppose first that X has a normal $(0, 1)$ distribution. Then the game is clearly "fair". Now suppose the casino gives you the following option. You can make X have a normal $(b, 1)$ distribution, but to do so you have to pay $\$cb$ which is not returned to you even if you win. Here $c > 0$ is set by the casino, but you can choose any $b > 0$.

a) For what values of c is it advantageous for you to use this option?

b) For these values of c, what value of b should you choose?

30. A manufacturing process produces ball bearings with diameters which are independent and normally distributed with mean 0.250 inches and SD 0.001 inches. In a high-precision application, 16 bearings are arranged in a ring. The specifications are that:

(i) each bearing must be between 0.249 and 0.251 inches in diameter;

(ii) the sum of the diameters of the 16 bearings must be between 3.995 and 4.005 inches.

a) What is the expected number of bearings which must be produced by the process to obtain 16 satisfying specification (i)?

b) Given 16 bearings obtained like this, what is the chance that they meet specification (ii)?

[*Hint* for b): Write $x^2\phi(x) = x[x\phi(x)]$ and use integration by parts to show that

$$\int_{-z}^{z} x^2\phi(x)dx = 2\Phi(z) - 1 - 2z\phi(z).]$$

31. The skew-normal pseudo-density. Referring to the end of Section 3.3, let

$$\Phi_\theta(z) = \Phi(z) - \frac{\theta}{6}(z^2 - 1)\phi(z)$$

This is the substitute for the normal c.d.f. $\Phi(z)$ which for $\frac{\theta}{6} \neq 0$ typically gives a better approximation than $\Phi(z)$ to the c.d.f. of a random variable with mean zero, variance 1 and third moment θ.

a) Let $\phi_\theta(z) = \frac{d}{dz}\Phi_\theta(z)$. Show $\phi_\theta(z) = [1 - \frac{\theta}{6}(3z - z^3)]\phi(z)$.

b) Show that for every θ

$$\int_{-\infty}^{\infty} \phi_\theta(z)dz = 1; \int_{-\infty}^{\infty} z\phi_\theta(z)dz = 0; \int_{-\infty}^{\infty} z^2\phi_\theta(z)dz = 1; \int_{-\infty}^{\infty} z^3\phi_\theta(z)dz = \theta$$

[So $\phi_\theta(z)$ is very like the probability density of a distribution with mean zero, variance 1 and third moment θ. This explains the choice $\theta = \text{Skewness}(X) = E(X_*^3)$ in the skew-normal approximation to the distribution of a standardized variable $X_* = (X - \mu)/\sigma$.]

c) Show that ϕ_θ is negative for large negative z if $\theta > 0$, and negative for large positive z if $\theta < 0$. So for $\theta \neq 0$, $\phi_\theta(z)$ is in fact *not* a probability density. It may be called instead a *pseudo-density*.

d) Find a probability in the Poisson(9) distribution whose normal approximation with continuity and skewness corrections is a negative number.

e) Explain carefully why, despite c) and d) the functions $\phi_{1/3}(z)$ and $\Phi_{1/3}(z)$ provide practically useful approximations to the Poisson(9) and other distributions which are roughly normal in shape but slightly skewed.

5 Continuous Joint Distributions

The *joint distribution* of a pair of random variables X and Y is the probability distribution over the plane defined by

$$P(B) = P((X,Y) \in B)$$

for subsets B of the plane. So $P(B)$ is the probability that the random pair (X,Y) falls in the set B. Joint distributions for discrete random variables were considered in Section 3.1. This chapter shows how these ideas for discrete random variables are extended to two or more continuously distributed random variables with sums replaced by integrals.

Section 5.1 concerns the simplest kind of continuous joint distribution, a *uniform* distribution defined by relative areas. Section 5.2 introduces the concept of a *joint density function*. Joint probabilities are then defined by volumes under a density surface. The important special case of independent normal variables is studied in Section 5.3. Then Section 5.4 deals with a general technique for finding the distribution of a function of two variables.

5.1 Uniform Distributions

The uniform distribution on an interval was discussed in Section 4.1. The idea extends to higher dimensions with relative lengths ıeplaced by relative areas or relative volumes. For example, a random point (X, Y) in the plane has uniform distribution on D, where D is a region of the plane with finite area, if:

(i) (X, Y) is certain to lie in D;

(ii) the chance that (X, Y) falls in a subregion C of D is proportional to the area of C

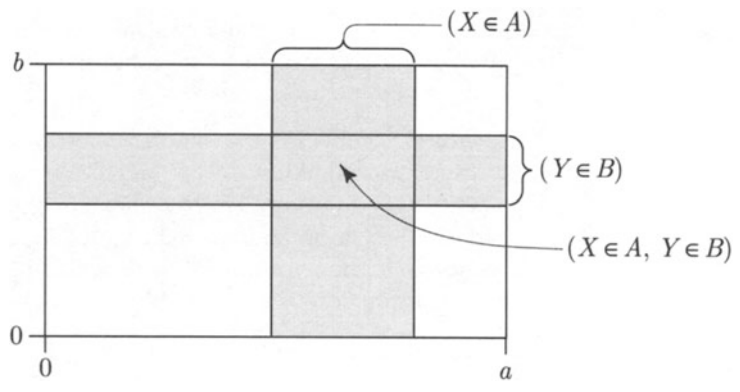
$$P((X, Y) \in C) = \frac{\text{area } (C)}{\text{area } (D)} \qquad \text{for} \quad C \subset D$$

Here is an important observation:

Independent Uniform Variables

If X and Y are independent random variables, each uniformly distributed on an interval, then (X, Y) is uniformly distributed on a rectangle.

To see why, suppose X and Y are independent and uniformly distributed on, say, $(0, a)$ and $(0, b)$, respectively. For intervals A and B the event $(X \in A, Y \in B)$ is the event that (X, Y) falls in the rectangle $A \times B$, as shown in the following Venn diagram:



So for any rectangle $A \times B$

$$
\begin{aligned}
P((X,Y) \in A \times B) &= P(X \in A, Y \in B) \\
&= P(X \in A)P(Y \in B) \quad \text{by independence of } X \text{ and } Y \\
&= \frac{\text{length}(A)}{a} \cdot \frac{\text{length}(B)}{b} \quad \text{by assumed uniform distributions of } X \text{ and } Y \\
&= \frac{\text{area}(A \times B)}{ab}
\end{aligned}
$$

Thus the probability that $(X,Y) \in C$ is the relative area of C in $(0,a) \times (0,b)$ for every rectangle C. The same must then be true for finite unions of rectangles, by the addition rule of probability and for area, hence also for any set C whose area can be defined by approximating with unions of rectangles. *Conclusion*: (X,Y) has uniform distribution on the rectangle $(0,a) \times (0,b)$.

The above observation allows probabilities involving two independent uniform variables X and Y to be found geometrically in terms of areas. The key step is correct identification of areas in the plane corresponding to events in question. Skill at doing this is essential for all further work in this chapter.

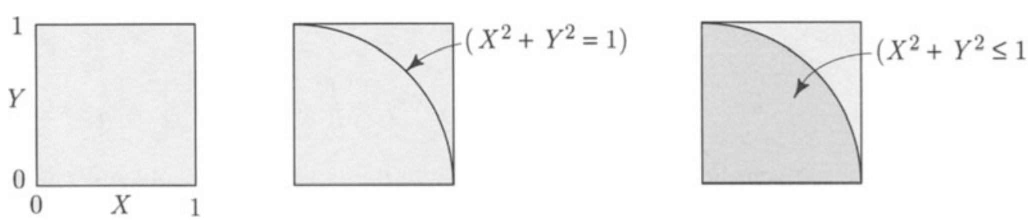
Example 1. Probabilities for two independent uniform random variables.

Suppose X and Y are independent uniform $(0,1)$ random variables.

Problem 1. Find $P(X^2 + Y^2 \le 1)$.

Solution. Proceed by 3 steps as in the diagram below:

- Draw a unit square with coordinates X, Y.
- Notice that $X^2 + Y^2 = 1$ gives the equation of a circle of radius 1.
- Recognize $(X^2 + Y^2 \le 1)$ as the region inside both the square and circle.
- Use the formula for the area of a circle to get $P(X^2 + Y^2 \le 1) = \frac{\pi}{4}$.

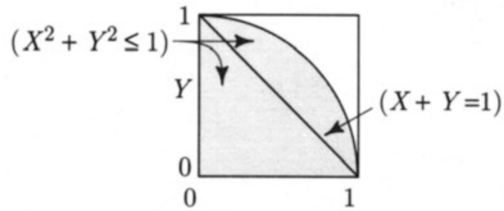


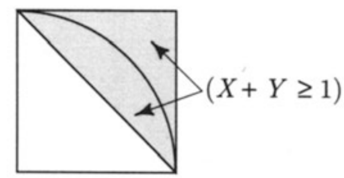
Problem 2. Find the conditional probability $P(X^2 + Y^2 \le 1 | X + Y \ge 1)$.

Solution. After first identifying $X^2 + Y^2 \le 1$ as above, next:

- Recognize $(X + Y = 1)$ as the line through the points $(0,1)$ and $(1,0)$.

– Deduce that $(X + Y \geq 1)$ is the shaded region above this line.





– Now compute the required relative area:

$$P(X^2 + Y^2 \leq 1 | X + Y \geq 1) = \frac{P(X^2 + Y^2 \leq 1, X + Y \geq 1)}{P(X + Y \geq 1)}$$
$$= \frac{\pi/4 - 1/2}{1/2} = \frac{\pi}{2} - 1$$

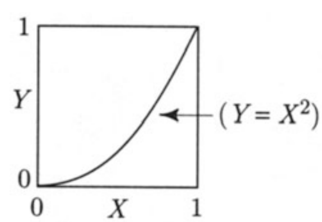
Problem 3. Find $P(Y \leq X^2)$.

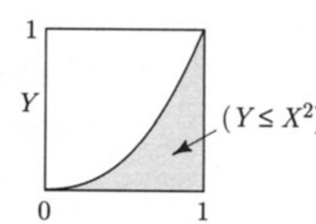
Solution.

– Graph $Y = X^2$.

– Recognize $(Y \leq X^2)$ as the region under this graph.

– Compute the area of this region by calculus.





$$P(Y \leq X^2) = \int_0^1 x^2 dx = \frac{1}{3} x^3 \Big|_0^1 = \frac{1}{3}$$

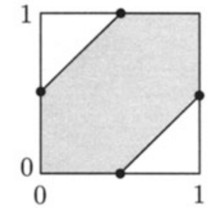
Discussion. Note well how only in the last of these problems was it necessary to resort to calculus to find the area. *Always* sketch the relevant regions first, then look out for familiar shapes, rectangles, triangles, and circles. If all else fails, use calculus.

Example 2. **More probabilities for two independent uniform variables.**

Let X and Y be independent random variables, each uniformly distributed on $(0, 1)$. Calculate the following probabilities:

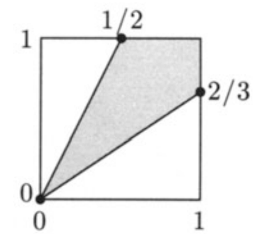
a)

$$P(|X - Y| \le 0.5) = \text{indicated area}$$
$$= 1 - \frac{1}{4} = 0.75$$



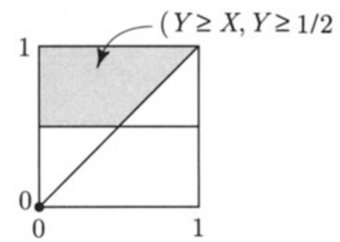
b)

$$P\left(\left|\frac{X}{Y} - 1\right| \le 0.5\right) = P\left(\frac{2}{3}X \le Y \le 2X\right)$$
$$= \text{indicated area}$$
$$= 1 - \frac{1}{2}\left(\frac{1}{2} + \frac{2}{3}\right) = \frac{5}{12}$$



c)

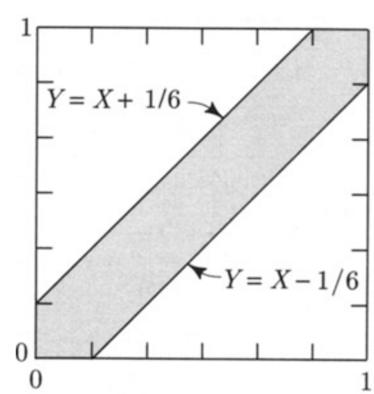
$$P\left(Y \ge X | Y \ge \frac{1}{2}\right) = \text{indicated area}/\frac{1}{2}$$
$$= \left(\frac{1}{2} - \frac{1}{8}\right)/\frac{1}{2} = \frac{3}{4}$$



Example 3. Probability of meeting.

Problem. Two people try to meet at a certain place between 5:00 P.M. and 5:30 P.M. Suppose that each person arrives at a time distributed uniformly at random in this time interval, independent of the other, and waits for the other at most 5 minutes. What is the probability that they meet?

Solution. Let X and Y be the arrival times measured as fractions of the 30 minute interval, starting from 5:00 P.M. Then X and Y are independent uniform $(0, 1)$ random variables. The people meet if and only if $|X - Y| \le 1/6$.



Desired probability = indicated area = $1 - (\frac{5}{6})^2 = \frac{11}{36}$

Uniform Distribution over a Volume

This is the extension of the idea of relative lengths in one dimension and relative areas in two dimensions to relative volumes in three and higher dimensions. If $U_1, \ldots, U_n$ are n independent random variables, with U_i uniformly distributed on an interval (a_i, b_i), then the same argument given earlier for the case $n = 2$ shows that the joint distribution of $(U_1, \ldots, U_n)$ is the uniform distribution defined by relative volumes within the n-dimesional box

$$(a_1, b_1) \times (a_2, b_2) \times \cdots \times (a_n, b_n)$$

whose n-dimensional volume is the product $(b_1 - a_1)(b_2 - a_2) \cdots (b_n - a_n)$ of the lengths of its sides.

To illustrate, a random point in the *unit cube* $(0,1) \times (0,1) \times (0,1)$, with approximately independent coordinates, is obtained by three successive calls of a pseudo-random number generator, say $(\mathrm{RND}_1, \mathrm{RND}_2, \mathrm{RND}_3)$. For any subvolume B of the unit cube bounded by a reasonably smooth surface (e.g., the portion of a box, pyramid, or sphere that lies inside the unit cube) the long-run frequency of times that $(\mathrm{RND}_1, \mathrm{RND}_2, \mathrm{RND}_3)$ is in B will be approximately the volume of B, that is $P(B)$ for the uniform distribution on the unit cube. For example, the long-run frequency of triples $(\mathrm{RND}_1, \mathrm{RND}_2, \mathrm{RND}_3)$ with

$$(\mathrm{RND}_1 - \tfrac{1}{2})^2 + (\mathrm{RND}_2 - \tfrac{1}{2})^2 + (\mathrm{RND}_3 - \tfrac{1}{2})^2 < 1/4$$

is approximately the volume of the subset of the unit cube

$$\{(x, y, z) : 0<x<1,\ 0<y<1,\ 0<z<1,\ (x - \tfrac{1}{2})^2 + (y - \tfrac{1}{2})^2 + (z - \tfrac{1}{2})^2 < 1/4\}$$

This is the volume of a sphere of radius $\frac{1}{2}$ centered at $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$, which is $\frac{4}{3}\pi \left(\frac{1}{2}\right)^3 = \frac{\pi}{6}$.

Exercises 5.1

1. Let (X, Y) have uniform distribution on the set

$$\{(x, y) : 0 < x < 2 \text{ and } 0 < y < 4 \text{ and } x < y\}.$$

Find: a) $P(X < 1)$; b) $P(Y < X^2)$.

2. A metal rod is l inches long. Measurements on the length of this rod are equal to l plus random error. Assume that the errors are uniformly distributed over the range -0.1 inch to $+0.1$ inch, and are independent of each other.

a) Find the chance that a measurement is less than $1/100$ of an inch away from l.

b) Find the chance that two measurements are less than $1/100$ of an inch away from each other.

3. Suppose X and Y are independent and uniformly distributed on the unit interval $(0, 1)$. Find:

$$P(Y \geq \frac{1}{2} \,|\, Y \geq 1 - 2X).$$

4. Let X and Y be independent random variables each uniformly distributed on $(0, 1)$. Find:

a) $P(|X - Y| \leq 0.25)$; b) $P(|X/Y - 1| \leq 0.25)$; c) $P(Y \geq X \,|\, Y \geq 0.25)$.

5. A very large group of students takes a test. Each of them is told his or her percentile rank among all students taking the test.

a) If a student is picked at random from all students taking the test, what is the probability that the student's percentile rank is over 90%?

b) If two students are picked independently at random, what is the probability that their percentile ranks differ by more than 10%?

6. A group of 10 people agree to meet for lunch at a cafe between 12 noon and 12:15 P.M. Assume that each person arrives at the cafe at a time uniformly distributed between noon and 12:15 P.M., and that the arrival times are independent of each other.

a) Jack and Jill are two members of the group. Find the probability that Jack arrives at least two minutes before Jill.

b) Find the probability of the event that the first of the 10 persons to arrive does so by 12:05 P.M., and the last person arrives after 12:10 P.M.

7. Let X and Y be two independent uniform $(0, 1)$ random variables. Let M be the smaller of X and Y. Let $0 < x < 1$.

a) Represent the event $(M \geq x)$ as the region in the plane, and find $P(M \geq x)$ as the area of this region.

b) Use your result in a) to find the c.d.f. and density of M. Sketch the graph of these functions.

8. Let $U_{(1)}, \ldots, U_{(n)}$ be the values of n independent uniform $(0, 1)$ random variables arranged in increasing order. Let $0 \leq x < y \leq 1$.

a) Find and justify a simple formula for $P(U_{(1)} > x \text{ and } U_{(n)} < y)$.

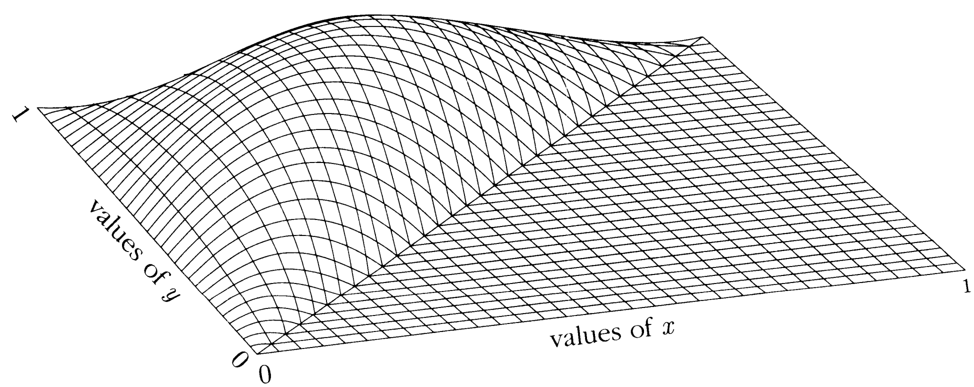
b) Find a formula for $P(U_{(1)} \leq x \text{ and } U_{(n)} < y)$.

9. A triangle problem. Suppose a straight stick is broken in three at two points chosen independently at random along its length. What is the chance that the three sticks so formed can be made into the sides of a triangle?

5.2 Densities

The concept of a *joint probability density function* $f(x, y)$ for a pair of random variables X and Y is a natural extension of the idea of a one-dimensional probability density function studied in Chapter 4. The function $f(x, y)$ gives the density of probability per unit area for values of (X, Y) near the point (x, y).

FIGURE 1. A joint density surface. Here a particular joint density function given by the formula $f(x, y) = 5!\,x(y - x)(1 - y) \qquad (0 < x < y < 1)$, is viewed as the height of a surface over the unit square $0 \le x \le 1, 0 \le y \le 1$. As explained later in Example 3, two random variables X and Y with this joint density are the second and fourth smallest of five independent uniform $(0, 1)$ variables. But for now the source and special form of this density are not important. Just view it as a typical joint density surface.

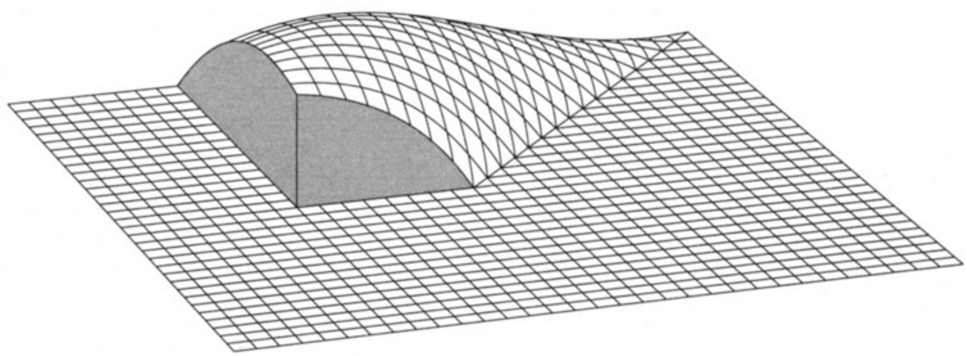


Examples in the previous section show how any event determined by two random variables X and Y, like the event $(X > 0.25 \text{ and } Y > 0.5)$, corresponds to a region of the plane. Now instead of a uniform distribution defined by relative areas, the probability of region B is defined by the volume under the density surface over B. This volume is an integral

$$P((X, Y) \in B) = \iint_B f(x, y)\,dx\,dy$$

This is the analog of the familiar area under the curve interpretation for probabilities obtained from densities on a line. Examples to follow show how such integrals can be computed by repeated integration, change of variables, or symmetry arguments. Uniform distribution over a region is now just the special case when $f(x, y)$ is constant over the region and zero elsewhere. As a general rule, formulae involving joint densities are analogous to corresponding formulae for discrete joint distributions described in Section 3.1. See pages 348 and 349 for a summary.

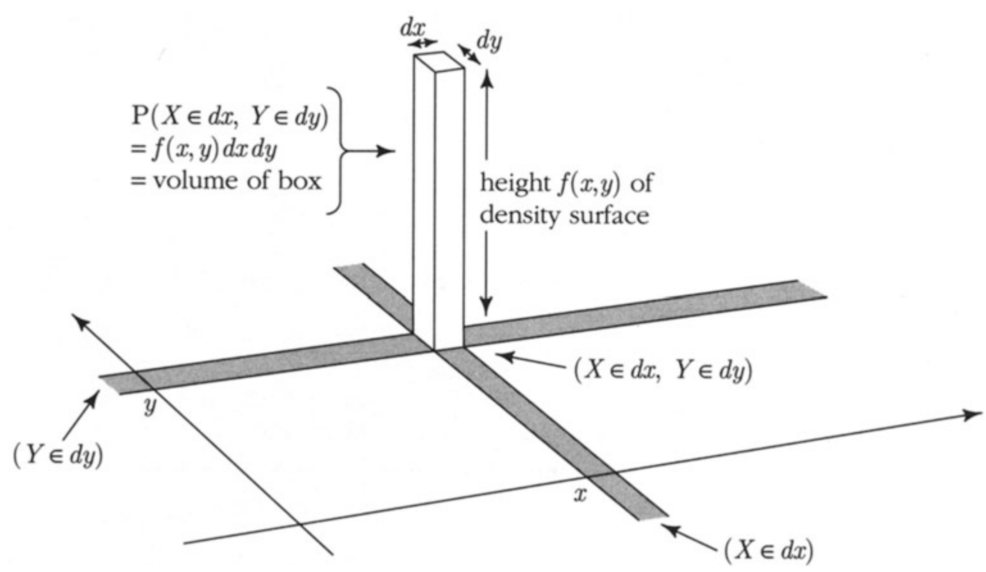
FIGURE 2. **Volume representing a probability.** The probability $P(X > 0.25$ and $Y > 0.5)$, for random variables X and Y with the joint density of Figure 1. The set B in this case is $\{(x, y) : x > 0.25$ and $y > 0.5\}$. You can see the volume is about half the total volume under the surface. The exact value, found later in Example 3, is $27/64$.



Informally, if (X, Y) has joint density $f(x, y)$, then there is the ***infinitesimal probability formula***

$$P(X \in dx, Y \in dy) = f(x, y)dx\,dy$$

This means that the probability that the pair (X, Y) falls in an infinitesimal rectangle of width dx and height dy near the point (x, y) is the probability density at (x, y) multiplied by the area $dx\,dy$ of the rectangle.



Discrete Joint Distribution

Probability of a point:

$$P(X = x, Y = y) = P(x, y)$$

The joint probability $P(x, y)$ is the probability of the single point (x, y).

Probability of a set B: The sum of probabilities of points in B

$$P((X, Y) \in B) = \sum_{(x,y)\in B} P(x, y)$$

Constraints: Non-negative with total sum 1

$$P(x, y) \geq 0 \qquad \text{and} \qquad \sum_{\text{all } x}\sum_{\text{all } y} P(x, y) = 1$$

Marginals:

$$P(X = x) = \sum_{\text{all } y} P(x, y)$$

$$P(Y = y) = \sum_{\text{all } x} P(x, y)$$

Independence: $\quad P(x, y) = P(X = x)P(Y = y) \quad$ (for all x and y)

Expectation of a function g of (X,Y), e.g., XY,

$$E(g(X, Y)) = \sum_{\text{all } x}\sum_{\text{all } y} g(x, y)P(x, y)$$

provided the sum converges absolutely.

Joint Distribution Defined by a Density

Infinitesimal probability:

$$P(X \in dx, Y \in dy) = f(x,y)\,dx\,dy$$

The joint density $f(x,y)$ is the probability per unit area for values near (x,y).

Probability of a set B: The volume under the density surface over B

$$P((X,Y) \in B) = \iint_B f(x,y)\,dx\,dy$$

Constraints: Non-negative with total integral 1

$$f(x,y) \geq 0 \qquad \text{and} \qquad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$$

Marginals:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

Independence: $f(x,y) = f_X(x)f_Y(y)$ (for all x and y)

Expectation of a function g of (X,Y), e.g., XY

$$E(g(X,Y)) = \iint g(x,y)f(x,y)dx\,dy$$

provided the integral converges absolutely.

The infinitesimal probability formula

$$P(X \in dx, Y \in dy) = f(x,y)dx\,dy$$

is really shorthand for a limiting statement about the ratio of probability per unit area for small areas, which, strictly speaking, holds only at points (x,y) such that the joint density is continuous at (x,y). But the infinitesimal formula conveys the right intuitive idea, and can be manipulated to obtain useful formulae which turn out to be valid even without assuming that the joint density is continuous.

Marginal densities. If (X,Y) has a joint density $f(x,y)$ in the plane, then each of the random variables X and Y has a density on the line. These are called the *marginal densities*. As shown in the preceding display, the marginal densities can be calculated from the joint density by integral analogs of the discrete formulae for marginal probabilities as row and column sums in a joint distribution table. Probabilities of discrete points are replaced by densities, and sums by integrals.

Independence. In general, random variables X and Y are called independent if

(1) $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ for all choices of sets A and B.

Joint Density for Independent Variables

Random variables X and Y with joint density $f(x,y)$ are independent if and only if the joint density is the product of the two marginal densities:

(2) $$f(x,y) = f_X(x)f_Y(y) \qquad \text{(for all } x \text{ and } y\text{)}$$

Intuitively (2) follows from (1) by taking A to be a small interval $(x, x+dx)$ near x, B a small interval $(y, y+dy)$ near y, to obtain

(3) $$P(X \in dx, Y \in dy) = P(X \in dx)P(Y \in dy)$$

$$\text{so} \quad f(x,y)\,dx\,dy = f_X(x)\,dx\,f_Y(y)\,dy$$

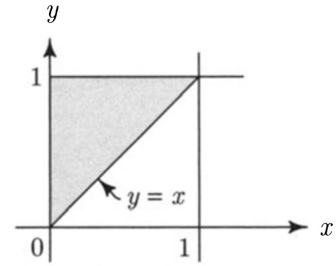
Cancelling the differentials dx and dy leaves the product formula for densities. Conversely, (1) is obtained from (2) by integration.

Example 1. Uniform on a triangle.

Suppose (X, Y) is uniformly distributed over the region $\{(x, y) : 0 < x < y < 1\}$.

Problem 1. Find the joint density of (X, Y).

Solution. By the assumption, $f(x, y) = c$ for $0 < x < y < 1$ and 0 elsewhere. Because the triangle has area $\frac{1}{2}$, $c = 2$.



Problem 2. Find the marginal densities $f_X(x)$ and $f_Y(y)$.

Solution.

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)\,dy \\ &= \int_{y=x}^{y=1} 2dy \qquad \text{since } f(x,y) = 2 \text{ for } 0 < x < y < 1, \quad 0 \text{ elsewhere} \\ &= 2(1-x) \qquad \text{for } 0 < x < 1 \quad \text{and 0 elsewhere.} \end{aligned}$$

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)\,dx \\ &= \int_{x=0}^{x=y} 2dx \qquad \text{since } f(x,y) = 2 \text{ for } 0 < x < y < 1 \quad 0 \text{ elsewhere} \\ &= 2y \qquad \text{for } 0 < y < 1 \quad \text{and 0 elsewhere.} \end{aligned}$$

Problem 3. Are X and Y independent?

Solution. No, since $f(x, y) \neq f_X(x) f_Y(y)$.

Problem 4. Find $E(X)$ and $E(Y)$.

Solution.

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_0^1 2x(1-x)\,dx = \frac{1}{3}$$

$$E(Y) = \int_{-\infty}^{\infty} y f_Y(y)\,dy = \int_0^1 2y^2\,dy = \frac{2}{3}$$

Problem 5. Find $E(XY)$.

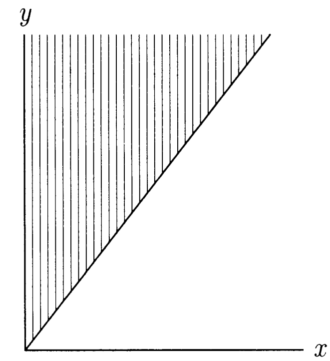
Solution.

$$E(XY) = \iint_{R^2} xy f(x,y)\,dx\,dy = 2\int_{y=0}^{1} dy \int_{x=0}^{y} xy\,dx = 2\int_{y=0}^{1} \frac{y^3}{2}\,dy = \frac{1}{4}$$

Remark. You can show that the joint distribution of X and Y considered here is that of $X = \min(U, V)$, $Y = \max(U, V)$, where U and V are independent uniform $(0, 1)$ variables. Example 3 gives a more difficult derivation of this kind.

Example 2. Independent exponential variables.

Problem. Let X and Y be independent and exponentially distributed random variables with parameters λ and μ, respectively. Calculate $P(X < Y)$.

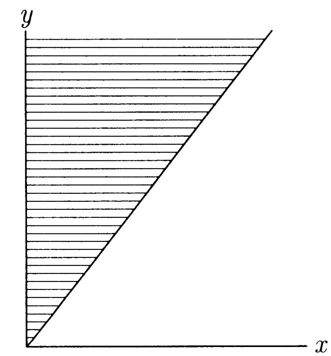


Solution. The joint density is

$$f(x, y) = (\lambda e^{-\lambda x})(\mu e^{-\mu y}) = \lambda\mu e^{-\lambda x - \mu y}$$

by independence. And $P(X < Y)$ is found by integration of this joint density over the set $\{(x, y) : x < y\}$:

$$\begin{aligned} P(X < Y) &= \iint_{x<y} \lambda\mu\, e^{-\lambda x - \mu y} dx\, dy \\ &= \int_{x=0}^{\infty} dx \int_{y=x}^{\infty} \lambda\mu\, e^{-\lambda x - \mu y} dy \\ &= \int_{x=0}^{\infty} \lambda e^{-\lambda x - \mu x} dx \\ &= \frac{\lambda}{\lambda + \mu} \end{aligned}$$



Remark. Done in the other order, the integral is

$$\int_{y=0}^{\infty} dy \int_{x=0}^{y} \lambda\mu\, e^{-\lambda x - \mu y} dx$$

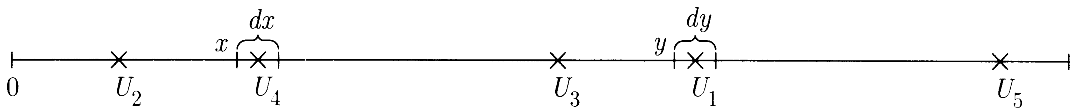
which simplifies to the same answer. As a general rule, provided the integrand is positive, as always when finding probabilities, double integrals done in either order produce the same result.

Example 3. Joint distribution of order statistics.

Suppose $U_{(1)} < U_{(2)} < \cdots < U_{(5)}$ are the order statistics of 5 independent uniform $(0, 1)$ variables $U_1, \ldots, U_5$, so $U_{(i)}$ is the ith smallest of $U_1, \ldots, U_5$, as, for example, in the following diagram:

Problem 1. Find the joint density of $U_{(2)}$ and $U_{(4)}$.

Solution. This is very like the calculation of the density of $U_{(i)}$ done in Section 4.6. The following diagram shows one way of getting $U_{(2)}$ in dx and $U_{(4)}$ in dy for $0 < x < y < 1$:



$$
\begin{aligned}
&P(U_{(2)} \in dx, U_{(4)} \in dy) \\
&= P(\text{ one } U_i \text{ in } (0,x), \text{ one in } dx, \text{ one in } (x,y), \text{ one in } dy, \text{ one in } (y,1)\ ) \\
&= 5!\, P(U_2 \in (0,x), U_4 \in dx, U_3 \in (x,y), U_1 \in dy, U_5 \in (y,1)) \\
&= 5!\, x\, dx(y-x)\, dy(1-y)
\end{aligned}
$$

Here the 5! is the number of different ways of deciding which variables fall in which intervals. The conclusion is that the joint density of $U_{(2)}$ and $U_{(4)}$ is

$$
P(U_{(2)} \in dx, U_{(4)} \in dy)/dx\, dy = \begin{cases} 5!\, x(y-x)(1-y) & \text{for } 0 < x < y < 1 \\ 0 & \text{elsewhere} \end{cases}
$$

This is the density surface shown in Figure 1 on page 346.

Problem 2. Find $P(U_{(2)} > 1/4 \text{ and } U_{(4)} > 1/2)$.

Solution. The volume representing this probability is shown in Figure 2 on page 347. This is the volume under the density surface over the area shaded in the diagram at right. This area is the intersection of:

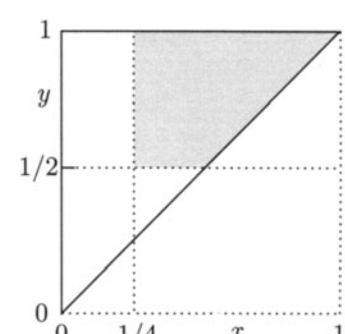
(i) the region representing the event; and

(ii) the region where the density is strictly positive.

This determines the ranges of integration. The required probability is thus

$$
\begin{aligned}
&5! \int_{y=1/2}^{1} \int_{x=1/4}^{y} x(y-x)(1-y)\, dx\, dy \\
&= 5! \int_{y=1/2}^{1} (1-y)\, dy \left[\frac{1}{2}x^2 y - \frac{1}{3}x^3\right]\Bigg|_{1/4}^{y} \\
&= 5! \int_{y=1/2}^{1} (1-y)\, dy \left[\frac{y^3}{6} - \frac{y}{2^5} - \frac{1}{3 \times 2^6}\right] = \frac{27}{64}
\end{aligned}
$$

by straightforward integration of the polynomial.



Exercises 5.2

1. Suppose that (X, Y) is uniformly distributed over the region $\{(x, y) : 0 < |y| < x < 1\}$. Find:

a) the joint density of (X, Y); b) the marginal densities $f_X(x)$ and $f_Y(y)$.

c) Are X and Y independent? d) Find $E(X)$ and $E(Y)$.

2. Repeat Exercise 1 for (X, Y) with uniform distribution over $\{(x, y) : 0 < |x| + |y| < 1\}$.

3. A random point (X, Y) in the unit square has joint density $f(x, y) = c(x^2 + 4xy)$ for $0 < x < 1$ and $0 < y < 1$, for some constant c.

a) Evaluate c. b) Find $P(X \le a)$, $0 < a < 1$. c) Find $P(Y \le b)$, $0 < b < 1$.

4. For random variables X and Y with joint density function

$$f(x, y) = 6e^{-2x-3y} \qquad (x, y > 0)$$

and $f(x, y) = 0$ otherwise, find:

a) $P(X \le x, Y \le y)$; b) $f_X(x)$; c) $f_Y(y)$.

d) Are X and Y independent? Give a reason for your answer.

5. Let X be exponentially distributed with rate λ, independent of Y, which is exponentially distributed with rate μ. Find $P(X \ge 3Y)$.

6. Let X and Y have joint density

$$f(x, y) = \begin{cases} 90(y - x)^8 & 0 < x < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

a) Find $P(Y > 2X)$. b) Find the marginal density of X.

c) Fill in the blanks (explain briefly):
The joint density f above is the joint density of the ________ and ________ of ten independent uniform (0, 1) random variables.

7. Two points are picked independently and uniformly at random from the region inside a circle. Let R_1 and R_2 be the distances of these points from the center of the circle. Find $P(R_2 \le R_1/2)$.

8. Random variables X and Y have joint density

$$f_{X,Y}(x, y) = \begin{cases} c(y^2 - x^2)e^{-y} & -y \le x \le y, \quad y > 0 \\ 0 & \text{otherwise} \end{cases}$$

Here c is a constant.

a) Show that Y has a gamma density, and hence deduce that $c = 1/8$.

b) Find the density of $4Y^3$.

c) Explain why $E(|X|)$ is at most 4.

9. **Minimum and maximum of two independent exponentials.** Let $X = \min(S, T)$ and $Y = \max(S, T)$ for independent exponential(λ) variables S and T. Let $Z = Y - X$.

a) Find the joint density of X and Y. Are X and Y independent?

b) Find the joint density of X and Z. Are X and Z independent?

c) Identify the marginal distributions of X and Z.

10. Minimum and maximum of n independent exponentials. Let $X_1, X_2, \ldots, X_n$ be independent, each with exponential (λ) distribution. Let $V = \min(X_1, X_2, \ldots, X_n)$ and $W = \max(X_1, X_2, \ldots, X_n)$. Find the joint density of V and W.

11. Suppose X and Y are independent random variables such that X has uniform $(0, 1)$ distribution, Y has exponential distribution with mean 1. Calculate:

a) $E(X+Y)$; b) $E(XY)$; c) $E[(X-Y)^2]$; d) $E(X^2e^{2Y})$.

12. Let T_1 and T_5 be the times of the first and fifth arrivals in a Poisson process with rate λ, as in Section 4.2. Find the joint density of T_1 and T_5.

13. Uniform spacings. Let $X = \min(U, V)$ and $Y = \max(U, V)$ for independent uniform$(0, 1)$ variables U and V. Find the distributions of

a) X; b) $1-Y$; c) $Y-X$.

14. Let U_1, U_2, U_3, U_4, U_5 be independent, each with uniform distribution on $(0, 1)$. Let R be the distance between the minimum and the maximum of the U_i's. Find

a) $E(R)$;

b) the joint density of the minimum and maximum of the U_i's;

c) $P(R > 0.5)$

15. C.d.f.'s in two dimensions. The *cumulative joint distribution function* of random variables X and Y is the function of x and y defined by $F(x, y) = P(X \le x, Y \le y)$.

a) Find a formula in terms of $F(x, y)$ for $P(a < X \le b,\ c < Y \le d)$.

b) For X and Y with joint density $f(x, y)$, express $F(x, y)$ in terms of f.

c) For X and Y with joint density $f(x, y)$, express $f(x, y)$ in terms of F.

These are analogs of formulae of Section 4.5 for cumulative distribution functions in one dimension. They are not used much, as there are few joint distributions for which there is an explicit formula for $F(x, y)$. But here are two examples.

d) Find $F(x, y)$ in terms of the marginal c.d.f's for independent X and Y.

e) Find $F(x, y)$ for X the minimum and Y the maximum of n independent uniform $(0, 1)$ variables, and $0 < x < y < 1$. Deduce the joint density of X and Y.

16. Suppose X_1, X_2, X_3 are independent exponential random variables with parameters $\lambda_1, \lambda_2, \lambda_3$ respectively. Evaluate $P(X_1 < X_2 < X_3)$.

17. Let (X, Y) be picked uniformly from the unit disc $R^2 \le 1$, where $R^2 = X^2 + Y^2$. Find:

a) the joint density of R and X;

b) repeat a) for a point (X, Y, Z) picked at random from inside the unit sphere $R^2 \le 1$, where now $R^2 = X^2 + Y^2 + Z^2$.

18. Suppose X_1, X_2 are independent random variables with the same density function.

a) Evaluate $P(X_1 < X_2)$.

b) Continuing, suppose X_1, X_2, X_3 are independent random variables with the same density function. Evaluate $P(X_{i_1} < X_{i_2} < X_{i_3})$ where (i_1, i_2, i_3) is a given permutation of $(1, 2, 3)$.

19. Let Lat be the latitude, Lon the longitude of the point of impact of the next meteorite that strikes the Earth's surface. Measure Lat in degrees from $-90°$ (South Pole) to $+90°$ (North Pole), and measure Lon similarly from $-180°$ to $+180°$. Assuming the point of impact is uniformly distributed over the Earth's surface, find

a) the density of Lon; b) the density of Lat;

c) the joint density of Lat and Lon. d) Are Lat and Lon independent?

20. Let X and Y be independent and uniform $(0, 1)$ and let $R = \sqrt{X^2 + Y^2}$. Show that:

a) $$f_R(r) = \begin{cases} \dfrac{\pi}{2} r & 0 \le r \le 1 \\ 2r\left[\dfrac{\pi}{4} - \arccos(1/r)\right] & 1 \le r \le \sqrt{2} \end{cases}$$

b) $$F_R(r) = \begin{cases} \dfrac{1}{4}\pi r^2 & 0 \le r \le 1 \\ \sqrt{r^2 - 1} + \left[\dfrac{\pi}{4} - \arccos(1/r)\right] r^2 & 1 \le r \le \sqrt{2} \end{cases}$$

c) Show without explicitly calculating $E(R)$ that

$$\sqrt{\frac{1}{2}} < E(R) < \sqrt{\frac{2}{3}}$$

d) (*Hard.*) Show that $E(R) \approx 0.765$.

21. Suppose two points are picked at random from the unit square. Let D be the distance between them. The main point of this problem is to find $E(D)$. This is hard to do exactly by calculus. But some information about $E(D)$ can be obtained as follows.

a) It is intuitively clear that $E(D)$ must be greater than $E(D_{\text{center}})$, where D_{center} is the distance from one point picked at random to the center of the square, and less than $E(D_{\text{corner}})$, the expected distance of one point from a particular corner of the square. Assuming this to be the case, find the values of these bounds on $E(D)$ using the results of Exercise 20.

b) Compute $E(D^2)$ exactly.

c) Deduce from b) a better upper bound for $E(D)$.

d) Computer simulation of 10,000 pairs of points gave mean distance 0.5197, and mean square distance 0.3310. Use these results to find an approximate 95% confidence interval for the unknown value of $E(D)$.

5.3 Independent Normal Variables

The most important properties of the normal distribution involve two or more independent normal variables. Suppose first that X and Y are independent, each with standard normal density function

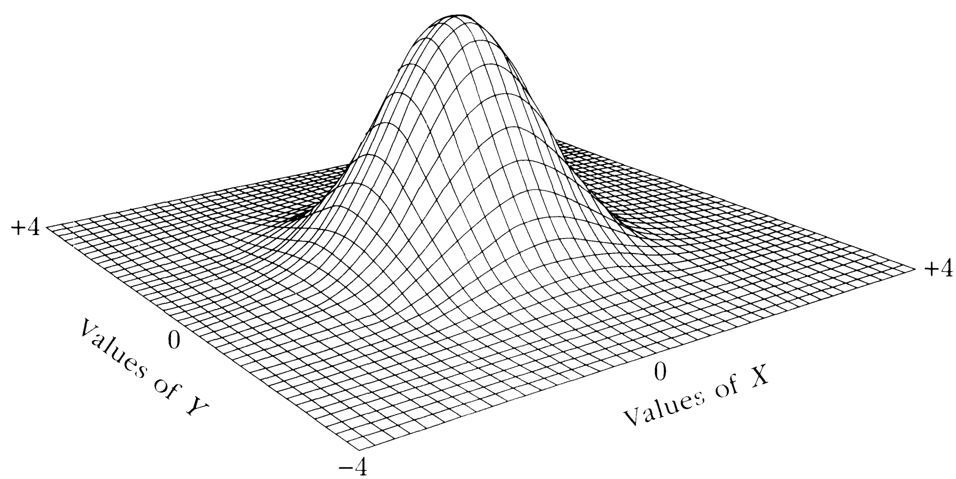
(a) $$\phi(z) = ce^{-\frac{1}{2}z^2} \qquad \text{where the formula} \qquad c = \frac{1}{\sqrt{2\pi}}$$

taken for granted up to now, will be verified in this section. The joint density of X and Y is given by

(b) $$f(x,y) = \phi(x)\phi(y) = c^2 e^{-\frac{1}{2}(x^2+y^2)}$$

The key property of this joint density is that it is a function of $r^2 = x^2 + y^2$, where r is the radial distance from the origin of the point (x, y). This makes the graph of this joint density a round bell-shaped surface over the (x, y) plane, with cross sections proportional to the standard normal curve.

FIGURE 1. Perspective plot of the joint density of X and Y.



The rotational symmetry of this bivariate distribution obtained from two independent normal variables is a very special property. It can be shown that this property distinguishes the normal distribution from all other probability distributions on the line. And this rotational symmetry is the key to understanding several important properties of the normal distribution, now considered in turn.

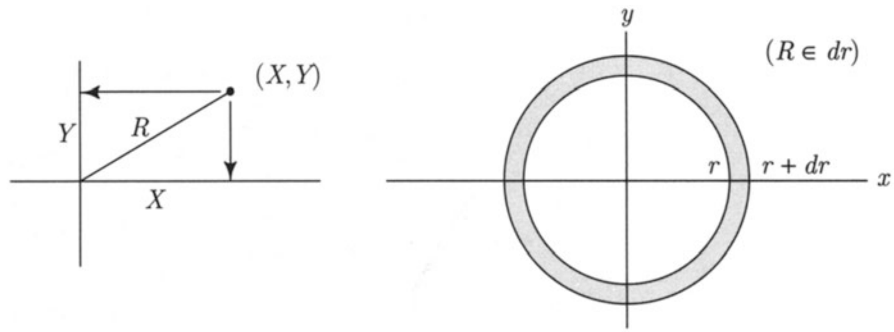
Evaluation of the Constant of Integration

The value of the constant c in the normal density (a) is found as a byproduct of calculating the distribution of the random variable

$$R = \sqrt{X^2 + Y^2}$$

which is the distance from the origin of a random point (X, Y) with joint density $\phi(x)\phi(y)$.

FIGURE 2. Geometry of X, Y, and R.



The event $(R \in dr)$ corresponds to (X, Y) falling in an annulus of infinitesimal width dr, radius r, circumference $2\pi r$, and area $2\pi r\, dr$, as in Figure 2. And $P(R \in dr)$ is the volume over this infinitesimal annulus beneath the joint density. But on the annulus the joint density has nearly constant value

$$\phi(x)\phi(y) = c^2 e^{-\frac{1}{2}(x^2+y^2)} = c^2 e^{-\frac{1}{2}r^2}$$

so the volume in question is just this nearly constant value times the area of the annulus. Thus

$$P(R \in dr) = 2\pi\, r\, dr\, c^2 e^{-\frac{1}{2}r^2} \qquad (r > 0)$$

This shows that R has probability density function

$$f_R(r) = 2\pi r c^2 e^{-\frac{1}{2}r^2}$$

The integral of this density from 0 to ∞ must be 1:

$$1 = \int_0^\infty 2\pi r c^2 e^{-\frac{1}{2}r^2}\, dr = -2\pi c^2 e^{-\frac{1}{2}r^2}\Big|_0^\infty = 2\pi c^2$$

This makes

$$2\pi c^2 = 1 \qquad \text{and} \qquad c = 1/\sqrt{2\pi}$$

So the constant of integration in the normal density involves π, due to the fact that the joint density of two independent standard normal variables is constant on circles centered at the origin.

The distribution of R appearing here, with density function

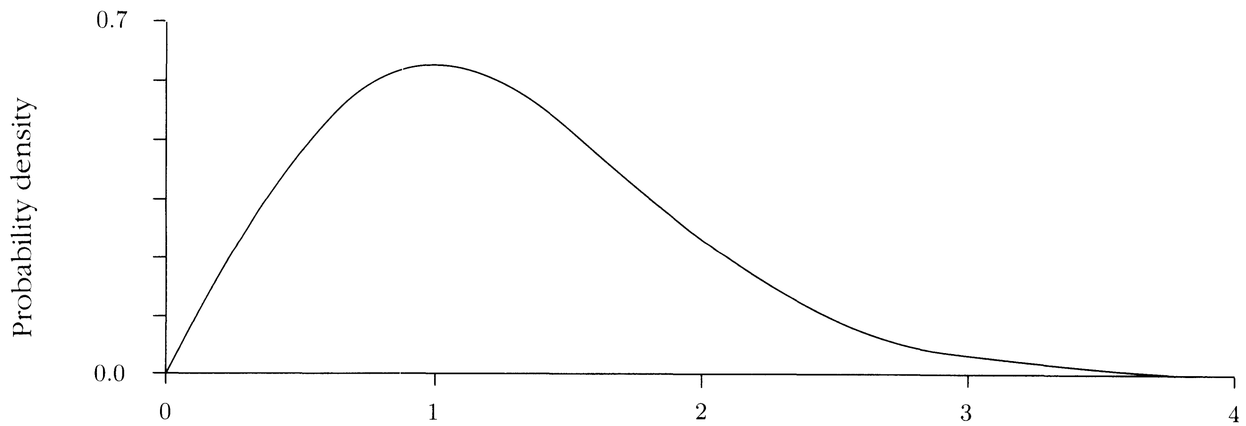
(c1) $$f_R(r) = r\,e^{-\frac{1}{2}r^2} \qquad (r > 0)$$

and c.d.f.

(c2) $$F_R(r) = \int_0^r s e^{-\frac{1}{2}s^2} ds = 1 - e^{-\frac{1}{2}r^2} \qquad (r > 0)$$

is called the *Rayleigh distribution.*

FIGURE 3. Density of the Rayleigh distribution of R.



Calculating the Variance of the Standard Normal Distribution

Since $E(X) = 0$ by symmetry, the variance of a standard normal random variable X is

$$\sigma^2 = E(X^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\frac{1}{2}x^2} dx$$

This integral can be reduced by an integration by parts to the integral of the standard normal density (exercise). But two independent standard normal variables X and Y

can also be used to show that $\sigma^2 = 1$. This, too, involves the radial random variable R. Because $R^2 = X^2 + Y^2$,

$$E(R^2) = E(X^2) + E(Y^2) = 2E(X^2)$$

using the fact that X and Y have the same distribution. So

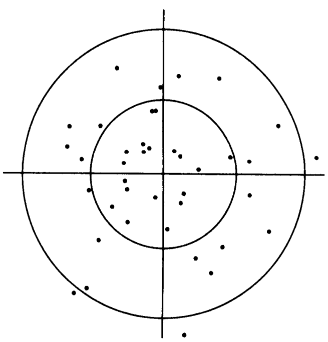
$$\sigma^2 = E(X^2) = \frac{1}{2}E(R^2)$$

But $S = R^2$ has density given by the change of variable formula

$$\begin{aligned} f_S(s) &= f_R(r) \Big/ \frac{ds}{dr} && (s = r^2 > 0) \\ &= re^{-\frac{1}{2}r^2}/2r && (s = r^2 > 0) \quad \text{by (c1)} \\ &= \frac{1}{2}e^{-\frac{1}{2}s} && (s > 0) \end{aligned}$$

Since this is the exponential density with parameter $\lambda = 1/2$,

$$E(R^2) = E(S) = 1/\lambda = 2 \qquad \text{so} \quad \sigma^2 = 1$$



Example 1. **Shots at a target.**

An expert marksman firing at a target produces a random scatter of shots which is roughly symmetrically distributed about the center of the bull's eye, with approximately 50% of the shots in the bull's eye, as in the diagram.

Problem 1. What is the approximate fraction of shots inside a circle with the same center as the bull's eye, but twice the radius?

Solution. Suppose that the marksman's shots are distributed approximately like (X, Y), where X and Y are independent normal random variables with mean 0 and variance σ^2. This would give such a symmetric distribution. By measuring distances in standard units, that is, relative to σ, we may as well assume $\sigma = 1$. Then the formulae obtained above for the distribution of $R = \sqrt{X^2 + Y^2}$ apply directly. Let r denote the radius of the bull's eye, measured in standard units. Using the normal approximation, the probability of each shot hitting the bull's eye would be

$$F_R(r) = 1 - e^{-\frac{1}{2}r^2}$$

from formula (c2) on page 359. Estimating this probability as 50% from the empirical data gives

$$e^{-\frac{1}{2}r^2} = 1/2 \qquad \text{so} \qquad r = \sqrt{2\log(2)} = 1.177\ldots \text{standard units}$$

Similarly, the fraction of shots inside a circle of twice the radius of the bull's eye should be approximately

$$F_R(2r) = 1 - e^{-\frac{1}{2}(2r)^2} = 1 - (e^{-\frac{1}{2}r^2})^4 = 1 - (1/2)^4 = \frac{15}{16} = 0.9375$$

Problem 2. What is the approximate average distance of the marksman's shots from the center of the bull's eye?

Solution. Using the law of large numbers, this average should be approximately

$$\begin{aligned}
E(R) &= \int_0^\infty r f_R(r)dr = \int_0^\infty r^2 e^{-\frac{1}{2}r^2} dr \qquad \text{by (c1) on page 359} \\
&= \frac{1}{2}\int_{-\infty}^{\infty} x^2 e^{-\frac{1}{2}x^2} dx \qquad \text{by symmetry} \\
&= \frac{\sqrt{2\pi}}{2}\int_{-\infty}^{\infty} x^2 \phi(x) dx \qquad \text{by definition of } \phi(x) \\
&= \sqrt{\frac{\pi}{2}} \qquad \text{because standard normal variance is 1} \\
&\approx 1.253 \text{ standard units} \\
&\approx 1.253/1.177 = 1.065 \text{ times the bull's eye radius } r
\end{aligned}$$

Linear Combinations and Rotations

Linear combinations of independent normal variables are always normally distributed. This important fact is another consequence of the rotational symmetry of the joint distribution of independent standard normal random variables X and Y. To see why, let X_θ be the first coordinate of (X, Y) relative to new coordinate axes set up at angle θ relative to the original X and Y axes, as in Figure 4.

As the diagram shows,

$$X_\theta = X\cos\theta + Y\sin\theta$$

But due to the rotational symmetry of the joint distribution, it is clear without calculation that the probability distribution of X_θ must be the same as that of X, namely, standard normal, no matter what the angle θ of rotation. For example, the event $x \le X_\theta \le x + \Delta x$ corresponding to (X, Y), falling in the area shaded in the left diagram of Figure 5, must have the same probability as the event $x \le X \le x + \Delta x$ corresponding to (X, Y), falling in the area shaded in the right diagram, because the shape of the bivariate normal density is the same over the two shaded regions. So,

$$P(x \le X_\theta \le x + \Delta x) = P(x \le X \le x + \Delta x)$$

FIGURE 4. Projection X_θ onto axis at angle θ to X-axis: $X_\theta = X\ \cos\theta + Y\ \sin\theta$.

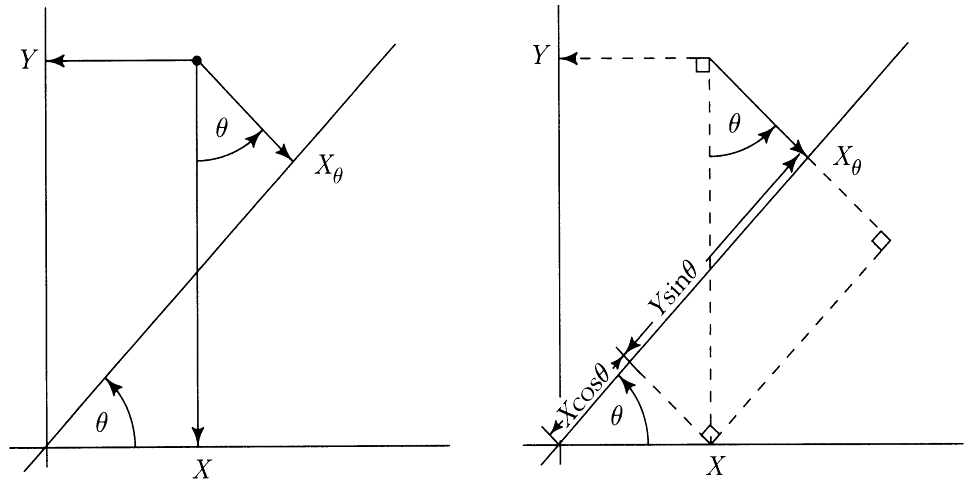
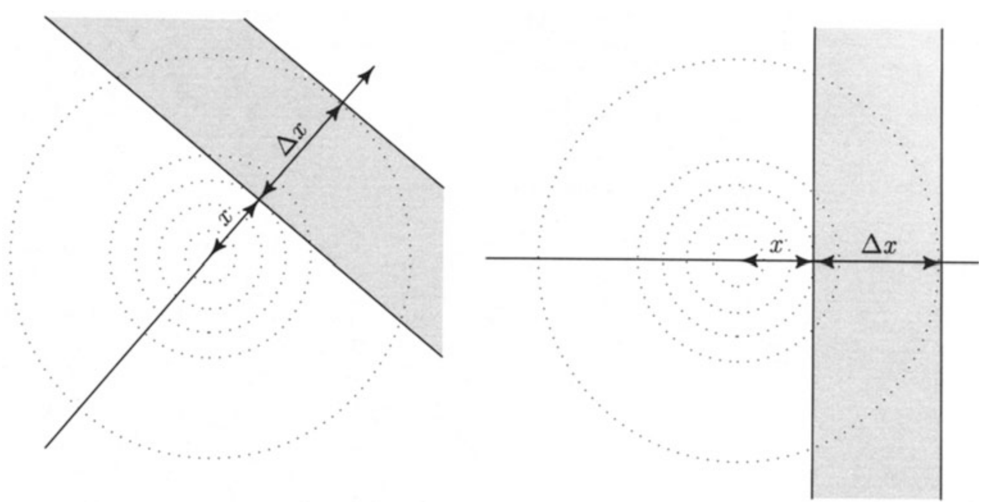


FIGURE 5. Events $(x \le X_\theta \le x + \Delta x)$ and $(x \le X \le x + \Delta x)$. Rotational symmetry of the joint density implies these two events have the same probability.



for every x and Δx. This shows that X_θ has normal $(0,1)$ distribution, for every θ. Since $\cos\theta$ and $\sin\theta$ may be arbitrary numbers α and β, subject only to the constraint that $\alpha^2+\beta^2=1$, the rotational symmetry of the joint distribution of two independent normal variables X and Y implies that:

(d) *If X and Y are two independent normal $(0,1)$ random variables, then $\alpha X+\beta Y$ has normal $(0,1)$ distribution for all α and β with $\alpha^2+\beta^2=1$.*

In particular, taking $\alpha=\beta=1/\sqrt{2}$, corresponding to rotation by 45°:

(e) *If X and Y are independent normal $(0,1)$ random variables, then $(X+Y)/\sqrt{2}$ has normal $(0,1)$ distribution.*

If Z has normal $(0,1)$ distribution, then σZ has normal $(0,\sigma^2)$ distribution. Taking $\sigma=\sqrt{2}$, (e) implies:

(f) *If X and Y are independent normal $(0,1)$ random variables, then $X+Y$ has normal $(0,2)$ distribution.*

This argument extends to give the following general conclusion, which includes (d), (e), and (f), as special cases.

Sums of Independent Normal Variables

If X and Y are independent with normal (λ,σ^2) and normal (μ,τ^2) distributions, then $X+Y$ has normal $(\lambda+\mu,\sigma^2+\tau^2)$ distribution.

Proof. Recall that X has normal (λ,σ^2) distribution if and only if $(X-\lambda)/\sigma$ has normal $(0,1)$ distribution. Transform all the variables to standard units by letting

$$U=(X-\lambda)/\sigma \quad \text{and} \quad V=(Y-\mu)/\tau \quad \text{and} \quad W=\frac{X+Y-(\lambda+\mu)}{\sqrt{\sigma^2+\tau^2}}$$

Then U and V are independent normal $(0,1)$ random variables. By algebra,

$$W=\alpha U+\beta V \quad \text{where} \quad \alpha^2=\frac{\sigma^2}{\sigma^2+\tau^2} \quad \text{and} \quad \beta^2=\frac{\tau^2}{\sigma^2+\tau^2} \quad \text{so} \quad \alpha^2+\beta^2=1$$

Apply (d) above with (U,V) instead of (X,Y) to deduce that W has normal $(0,1)$ distribution. So $X+Y=(\lambda+\mu)+\sqrt{\sigma^2+\tau^2}W$ has normal $(\lambda+\mu,\sigma^2+\tau^2)$ distribution.
□

Several Independent Normal Variables

The result that the sum of two independent normal variables is normal extends to sums and linear combinations of several independent normal random variables, by repeated applications of the result for two variables. For example, if $X_1, \ldots, X_n$ are independent and normal $(0, 1)$, then $X_1 + \cdots + X_n$ has normal $(0, n)$ distribution, with standard deviation $\sqrt{n}$.

Example 2. Linear combinations of normals.

For $\sigma = 1, 2, 3$ suppose X_σ has normal $(0, \sigma^2)$ distribution, and these three random variables are independent.

Problem 1. Find $P(X_1 + X_2 + X_3 < 4)$.

Solution. Let $S = X_1 + X_2 + X_3$. Then S has normal $(0, 1^2 + 2^2 + 3^2)$ distribution, and if $Z = S/\sqrt{14}$ is S standardized, the problem is just to find

$$P(S < 4) = P(Z < 4/\sqrt{14}) = \Phi(4/\sqrt{14}) \approx 0.857$$

Problem 2. Find $P(4X_1 - 10 < X_2 + X_3)$.

Solution. Rearranging the statement of the inequality shows this is the same as

$$P(4X_1 - X_2 - X_3 < 10) = P(L < 10) \qquad \text{where } L = 4X_1 - X_2 - X_3$$

Since the linear combination L has normal distribution with mean 0 and variance $4^2 \times 1^2 + (-1)^2 \times 2^2 + (-1)^2 \times 3^2 = 29$, the probability is

$$P(L < 10) = \Phi(10/\sqrt{29}) \approx 0.968$$

The Chi-Square Distribution

By the same calculation as in two dimensions, the joint density of n independent normal variables at every point on the sphere of radius r in n-dimensional space is $(1/\sqrt{2\pi})^n \exp\left(-\frac{1}{2}r^2\right)$. This joint density is symmetric with respect to arbitrary rotations of the coordinates in n-dimensional space, or *spherically symmetric*. So a cloud of points (or a galaxy of stars), in ordinary 3-dimensional space, with approximately independent normally distributed coordinates with common variance, appears spherical when viewed at a distance, from any perspective. For independent standard normal Z_i let

$$R_n = \sqrt{Z_1^2 + \cdots + Z_n^2}$$

denote the distance of $(Z_1, \ldots, Z_n)$ from the origin in n-dimensional space. The n-dimensional volume of a thin spherical shell of thickness dr at radius r is $c_n r^{n-1}\, dr$

where c_n is the $(n-1)$-dimensional volume of the "surface" of a sphere of radius 1 in n dimensions. (For $n = 3$, $c_3 = 4\pi$, by the formula $4\pi r^2$ for the surface area of a sphere of radius r in 3 dimensions.) The same argument used in two dimensions shows that

$$P(R_n \in dr) = c_n\, r^{n-1} \left(1/\sqrt{2\pi}\right)^n e^{-\frac{1}{2}r^2} dr \qquad (r > 0) \tag{1}$$

A change of variable allows the constant c_n to be evaluated by recognizing that the density of $R_n^2 = Z_1^2 + \cdots + Z_n^2$ is the gamma $(n/2, 1/2)$ density introduced in Section 4.2:

$$f_{R_n^2}(t) = (2^{n/2}\Gamma(n/2))^{-1} t^{(n/2)-1} e^{-t/2} \qquad (t > 0) \tag{2}$$

Exercise 15 and Chapter 5 Review Exercise 26 give formulae for c_n and $\Gamma(n/2)$.

Statisticians call this gamma $(n/2, 1/2)$ distribution of R_n^2 the *chi-square* distribution with n *degrees of freedom*. The chi-square distribution provides a useful test of *goodness of fit*, that is, how well data from an empirical distribution of n observations conform to the model of random sampling from a particular theoretical distribution. If there are only two categories, say success and failure, the model of independent trials with probability p of success is tested using the normal approximation to the binomial distribution. But for data in several categories the problem is how to combine the tests for different categories in a reasonable way. This problem was solved as follows by the statistician Karl Pearson (1857–1936). For a finite number of categories m, let N_i denote the number of results in category i. Under the hypothesis that the N_i are counting results of independent trials with probability p_i for category i on each trial, it turns out that no matter what the probabilities p_i, for large enough n the so-called *chi-square statistic*

$$\sum_{i=1}^{m} \frac{(N_i - n p_i)^2}{n p_i}$$

that is the sum over categories of (observed − expected)^2/expected, has distribution that is approximately chi-square with $m-1$ degrees of freedom. In statistical jargon, a value of the statistic higher than the 95th percentile point on the chi-square distribution with $m-1$ degrees of freedom would "reject the hypothesis at the 5% level". Unusually small values of the chi-square statistic are sometimes taken as evidence to suggest that an observer fudged the data to suit the hypothesis. The exact joint distribution of the N_i under the hypothesis of randomness is multinomial with parameters n and $p_1, \ldots, p_m$. The above result can be derived from a multivariate form of the normal approximation to the binomial. The joint distribution of $N_1, \ldots, N_m$ is essentially $m-1$ dimensional due to the constraint $N_1 + \cdots + N_m = n$. This is why the relevant chi-square distribution has $m-1$ degrees of freedom.

For tables of the chi-square distribution, and similar chi-square tests of other hypotheses such as independence, consult a statistics book. The mean, standard deviation and skewness of the chi-square distribution of R_n^2 with n degrees of freedom are easily calculated (Exercise 15):

$$E(R_n^2) = n, \qquad SD(R_n^2) = \sqrt{2n} \qquad \text{and} \qquad \text{Skewness}(R_n^2) = 4/\sqrt{2n}$$

For large n the chi-square distribution is approximately normal, by the central limit theorem. Because the skewness is quite large even for moderate values of n, the normal approximation with skewness correction gives the better approximation

$$P(R_n^2 \le x) \approx \Phi(z) - \frac{\sqrt{2}}{3\sqrt{n}}(z^2 - 1)\phi(z) \quad \text{where} \quad z = (x - n)/\sqrt{2n} \text{ and } x > 0$$

TABLE 1. **Distribution of radial distance in three dimensions.** The probability that a point with independent standard normal coordinates in three dimensions lies inside a sphere of radius r, that is, $P(R_3 \le r) = P(R_3^3 \le r^2)$, was obtained by numerical integration of the density. These probabilities are shown along with their approximations obtained using the skew–normal approximation to the chi-square (3) distribution of R_3^2. The approximations are surprisingly good considering the small value of n.

radius r	1	2	3	4
probability $P(R_3 \le r)$	0.199	0.739	0.971	0.999
skew-normal approximation	0.233	0.741	0.966	1.000

Exercises 5.3

1. Continuing Example 1, calculate the following, where all distances are measured in standard units:

a) the probability of a shot falling inside a circle of radius 1/2;

b) the probability of a shot falling in the region of the positive quadrant between radii 1 and 2;

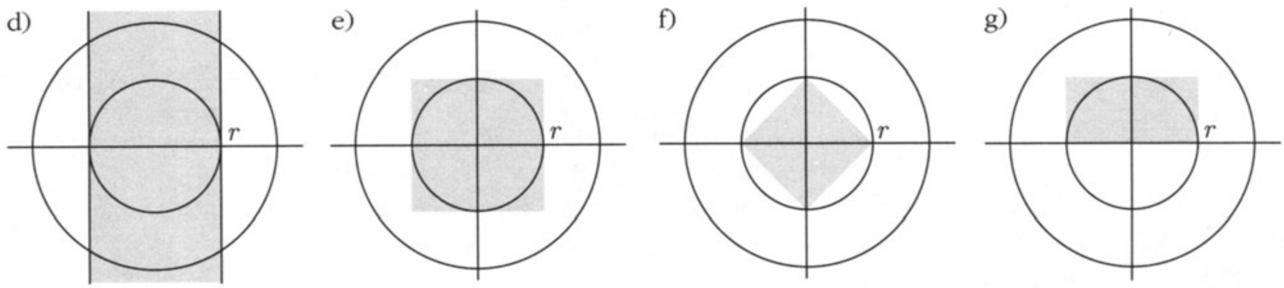
c) the approximate average absolute distance of the shots from the horizontal line through the center of the bull's eye;

d) the probability that a shot hit within distance r of the vertical axis through the center (r = radius of bull's eye in standard units);

e) the probability of hitting a square touching the outside of the bull's eye;

f) the probability of hitting a square touching the inside of the bull's eye;

g) the probability of hitting a rectangle of sides r and $2r$ positioned as shown relative to the bull's eye.



2. Let X and Y be independent random variables, with $E(X) = 1$, $E(Y) = 2$, $Var(X) = 3$, and $Var(Y) = 4$.

a) Find $E(10X^2 + 8Y^2 - XY + 8X + 5Y - 1)$.

b) Assuming all variables are normally distributed, find $P(2X > 3Y - 5)$.

3. W, X, Y and Z are independent standard normal random variables. Find (no integrations are necessary!)

a) $P(W + X > Y + Z + 1)$; b) $P(4X + 3Y < Z + W)$;

c) $E(4X + 3Y - 2Z^2 - W^2 + 8)$; d) $SD(3Z - 2X + Y + 15)$.

4. Suppose the true weight of a standard weight is 10 grams. It is weighed twice independently. Suppose that the first measurement is a normal random variable X with $E(X) = 10$ g and $SD(X) = 0.2$ g, and that the second measurement is a normal random variable Y with $E(Y) = 10$ g and $SD(Y) = 0.2$ g.

a) Compute the probability that the second measurement is closer to 10 g than the first measurement.

b) Compute the probability that the second measurement is smaller than the first, but not by more than 0.2 g.

5. Let X and Y be independent and normally distributed, X with mean 0 and variance 1, Y with mean 1. Suppose $P(X > Y) = 1/3$. Find the standard deviation of Y.

6. Let X and Y be independent standard normal variables. Find:

a) $P(3X + 2Y > 5)$; b) $P(\min(X, Y) < 1)$;

c) $P(|\min(X, Y)| < 1)$; d) $P(\min(X, Y) > \max(X, Y) - 1)$.

7. Suppose the AC Transit bus is scheduled to arrive at my corner at 8:10 A.M., but its actual arrival time is a normal random variable with mean 8:10 A.M., and standard deviation 40 seconds. Suppose I try to arrive at the corner at 8:09, but my arrival time is actually normally distributed with mean 8:09 A.M., and standard deviation 30 seconds.

a) What percentage of the time do I arrive at the corner before the bus is scheduled to arrive?

b) What percentage of the time do I arrive at the corner before the bus does?

c) If I arrive at the stop at 8:09 A.M. and the bus still hasn't come by 8:12 A.M., what is the probability that I have already missed it?

(State your assumptions carefully.)

8. Peter and Paul agree to meet at a restaurant at noon. Peter arrives at a time normally distributed with mean 12:00 noon, and standard deviation 5 minutes. Paul arrives at a time normally distributed with mean 12:02 P.M., and standard deviation 3 minutes. Assuming the two arrival times are independent, find the chance that

a) Peter arrives before Paul; b) both men arrive within 3 minutes of noon;

c) the two men arrive within 3 minutes of each other.

9. Suppose heights in a large population are approximately normally distributed with a mean of 5 feet 10 inches and an SD of 2 inches. Suppose a group of 100 people is picked at random from this population.

a) What is the probability that the tallest person in this group is over 6 feet 4 inches tall?

b) What is the probability that the average height of people in the group is over 5 feet 10.5 inches?

c) Suppose instead that the distribution of heights in the population was not normal, but some other distribution with the given mean and SD. To which of the problems a) and b) would the answer still be approximately the same? Explain carefully.

10. In a large corporation, people over age thirty have an annual income whose distribution can be approximated by a normal distribution with mean $60,000 and standard deviation $10,000. The incomes of those under age thirty are also approximately normal, but with mean $40,000 and standard deviation $10,000.

a) Two people are selected at random from those over age thirty. What is the chance that the average of their two incomes is over $65,000?

b) One person is selected at random from those over thirty, and independently, one person is selected at random from those under thirty. What is the chance that the younger's income exceeds the older's?

c) What is the chance that the smaller of the two incomes in b) exceeds $50,000?

11. **Einstein's model for Brownian motion.** Suppose that the X coordinate of a particle performing Brownian motion has normal distribution with mean 0 and variance σ^2 at time 1. Let X_t be the X displacement after time t. Assume the displacement over any time interval has a normal distribution with parameters depending only on the length of the interval, and that displacements over disjoint time intervals are independent.

a) Find the distribution of X_t.

b) Let (X_t, Y_t) represent the position at time t of a particle moving in two dimensions. Assume that X_t and Y_t are independent Brownian motions starting at 0 at time $t = 0$. Find the distribution of $R_t = \sqrt{X_t^2 + Y_t^2}$, and give the mean and standard deviation in terms of σ and t.

c) Suppose a particle performing Brownian motion (X_t, Y_t) as in b) has an X coordinate after one second which has mean 0 and standard deviation one millimeter (mm). Calculate the probability that the particle is more than 2 mm from the point $(0,0)$ after one second.

12. Suppose two shots are fired at a target. Assume each shot hits with independent normally distributed coordinates, with the same means and equal unit variances.

a) Find the mean of the distance between the points where the two shots strike.

b) Find the variance of the same random variable.

13. Independence of radial and angular parts. Let X and Y be independent normal $(0, \sigma^2)$ random variables. Let (R, Θ) be (X, Y) in polar coordinates, so $X = R\cos\Theta$, $Y = R\sin\Theta$.

a) Show that R and Θ are independent, and that Θ has uniform $(0, 2\pi)$ distribution.

b) Let R and Θ now be arbitrary random variables such that R/σ has the Rayleigh distribution (c1), Θ has uniform $(0, 2\pi)$ distribution, and R and Θ are independent. Explain why the random variables $X = R\cos\Theta$ and $Y = R\sin\Theta$ must be independent normal $(0, \sigma^2)$.

c) Find functions h and k such that if U and V are independent uniform $(0,1)$ random variables, then $X = \sigma h(U)\cos[k(V)]$ and $Y = \sigma h(U)\sin[k(V)]$ are independent normal $(0, \sigma^2)$. [This gives a means of simulating normal random variables using a computer random number generator. Try generating a random scatter of independent bivariate normally distributed pairs if you have random numbers available. It should look like the scatter in Example 1.]

14. Let X and Y be independent standard normal variables. Suppose they are transformed into polar coordinates, $X = R\cos\Theta$ and $Y = R\sin\Theta$ with $0 < \Theta < 2\pi$ and $0 < R < \infty$, as in Exercise 13.

a) Derive the distribution of $2\Theta \ mod \ 2\pi$. [The quantity $x \ mod \ a$ denotes the remainder when x is divided by a.]

b) Derive the joint distribution of $R\cos 2\Theta$ and $R\sin 2\Theta$.

c) Show that both

$$\frac{2XY}{\sqrt{X^2+Y^2}} \quad \text{and} \quad \frac{X^2-Y^2}{\sqrt{X^2+Y^2}}$$

have the standard normal distribution. Are they independent?

15. Chi-square distributions. These are the special case of half-integer gamma distributions which come from sums of squares of independent standard normal variables. Show:

a) If Z has standard normal distribution, then Z^2 has gamma $(1/2, 1/2)$ distribution, and $\Gamma(1/2) = \sqrt{\pi}$.

b) If n is an odd integer, then $\quad \Gamma(n/2) = \dfrac{\sqrt{\pi}(n-1)!}{2^{n-1}(\frac{n-1}{2})!}$

c) If X has normal $(0, \sigma^2)$ distribution, then X^2 has gamma $(1/2, 1/2\sigma^2)$ distribution.

d) If $Z_1, \ldots, Z_n$ are independent standard normal random variables, then ${Z_1}^2 + \cdots + {Z_n}^2$ has gamma $(n/2, 1/2)$ distribution, also known as the *chi-square distribution with n degrees of freedom*, or chi-square (n) distribution.

e) If $Y_1, \ldots, Y_n$ are independent chi-square random variables with $k_1, \ldots, k_n$ degrees of freedom, respectively, then $Y_1 + \cdots + Y_n$ has chi-square $(k_1 + \cdots + k_n)$ distribution.

f) The mean, variance and skewness of the chi-square (n) distribution are as stated on page 366.

16. Poisson formula for the chi-square $(2m)$ c.d.f. For $m = 1, 2, \ldots$ let R_{2m}^2 have chi-square $(2m)$ distribution. Use the connection between the gamma distribution and the Poisson process to find formulae in terms of appropriate Poisson probabilities for:

a) the c.d.f. of R_{2m}^2; b) the c.d.f. of R_{2m}.

c) Check that your formulae agree with the formulae in the text for $m = 1$. Now make a table of $P(R_4 \leq r)$ for $r = 1, \ldots 5$.

17. Skew-normal approximation to the chi-square distribution. Let R_n^2 have chi-square (n) distribution.

a) Find the approximation to $P(R_4 \leq r)$ for $r = 1, \ldots 5$ obtained from the skew-normal approximation to the distribution of R_4^2. Compare to the exact results found in Exercise 16.

b) Find both the plain normal approximation and the skew-normal approximation to $P(R_{10}^2 \leq 9.34) = 0.500$. Which approximation is better?

18. Suppose a large number n identical molecules are distributed independently at random in a box with sides of 1 centimeter. Let X, Y, Z be the coordinates in centimeters of the center of mass of the n molecules at a particular instant, relative to the center of the box. Thus,

$$X = (X_1 + \cdots + X_n)/n$$

and so on, where (X_i, Y_i, Z_i) are the coordinates of the ith molecule in centimeters. Let $R = \sqrt{X^2 + Y^2 + Z^2}$ be the distance of the center of mass of the n molecules from the center of the box. Given that for the chi-square distribution with 3 degrees of freedom the 95th percentile is at 7.82, find approximately the value of r such that R is 95% sure to be smaller than r.

5.4 Operations (Optional)

Many applications require calculation of the distribution of some random variable Z which is a function of X and Y, where X and Y are random variables with some joint density $f(x,y)$. Here the function of X and Y might be, for example, $X+Y$, XY, X/Y, $\max(X,Y)$, $\min(X,Y)$, or $\sqrt{X^2+Y^2}$. This kind of calculation has been done in special cases in previous sections. For example, maxima and minima in Section 4.5, sums and $\sqrt{X^2+Y^2}$ for normal variables in Section 5.3. This section gives a general technique for computing such distributions by integration.

Calculating the whole distribution of a function of X and Y can sometimes be tedious. So keep in mind that for some purposes it may be enough to calculate an expectation. The expectation of a function of X and Y can always be expressed as an integral with respect to the density of (X,Y). For example, for the product XY,

$$\begin{aligned} E(XY) &= \iint xyf(x,y)dx\,dy \\ &= E(X)E(Y) \qquad \text{if } X \text{ and } Y \text{ are independent} \end{aligned}$$

despite the fact that there are very few examples where the whole distribution of a product of independent random variables can be found explicitly.

One method of finding the distribution of $Z=g(X,Y)$ is to find the c.d.f. $P(Z \le z)$ by integration of $f(x,y)$ over the region in the (x,y) plane where $g(x,y) \le z$. Provided this integral can be evaluated fairly explicitly, the density of Z can then be found by differentiation of the c.d.f. Usually a quicker method of finding the distribution of Z is to anticipate that Z will have a density function f_Z, and to find this density $f_Z(z) = P(Z \in dz)/dz$ by integrating the joint density of X and Y over the subset $(Z \in dz)$ in the (X,Y) plane. This technique gives integral formulae for the density for the sum $X+Y$, for other linear combinations like $X-Y$, and for the product XY, and ratio X/Y. The formulae for sums and ratios will now be worked out in detail. Results for other operations are similar and left as exercises.

Distribution of Sums

A good deal has already been said on this topic. Recall the addition rule for expectation

$$E(X+Y) = E(X)+E(Y) \qquad \text{whatever the joint distribution of } X \text{ and } Y$$

the addition rule for variances in the case of independence, and the central limit theorem governing the asymptotic distribution for the sum of a large number of independent and identically distributed terms. Also, the exact distribution of sums has been computed in special cases by a variety of methods. The following table reviews some important examples:

Distribution of terms	Distribution of sum	See Section
n independent Bernoulli (p)	binomial (n, p)	2.1
independent Poisson (μ_i)	Poisson $(\Sigma\mu_i)$	3.5
independent normal (μ_i, σ_i^2)	normal $(\Sigma\mu_i, \Sigma\sigma_i^2)$	5.3
r independent geometric (p)	negative binomial (r, p)	3.4
r independent exponential (λ)	gamma (r, λ)	4.2

In the discrete case the distribution of the sum of random variables is determined by the formula

$$P(X+Y=z) = \sum_{\text{all } x} P(X=x, Y=z-x)$$

found in Section 4.1. The following display gives the corresponding formula for densities:

Density of $X + Y$

If (X, Y) has density $f(x, y)$ in the plane, then $X + Y$ has density on the line

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f(x, z-x)dx$$

Density Convolution Formula

If X and Y are independent, then

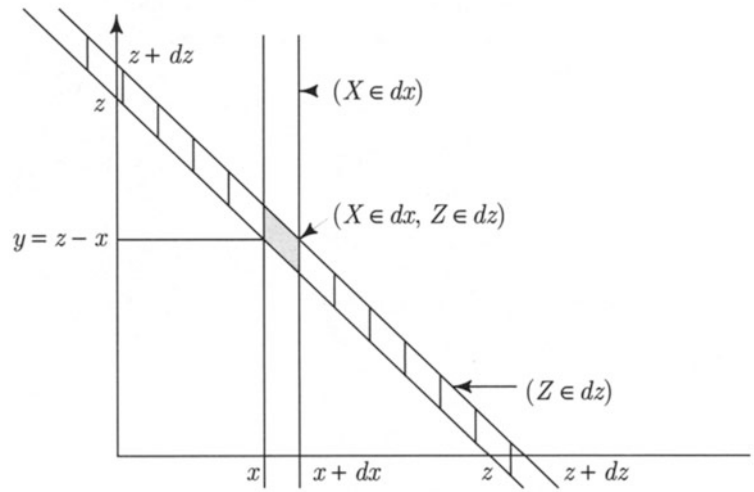
$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)dx$$

Note: If the random variables X and Y are non-negative, then the lower limit of integration in the convolution formula can be changed from $-\infty$ to 0, since $f_X(x) = 0$ for all $x < 0$, and the upper limit can be changed from ∞ to z, since $f_Y(z-x) = 0$ for $x > z$.

The convolution formula is the special case of the formula for the density of $X + Y$ when $f(x, y) = f_X(x) f_Y(y)$ by independence. This operation on probability density functions f_X and f_Y is called *convolution*. It leads to a new density, the density of the sum of random variables X and Y, assumed independent.

To avoid confusion about limits of integration in particular examples, sketch the subset of the plane where the joint density is strictly positive, and the line of integration corresponding to $X + Y = z$, as in examples below.

Derivation of the density of $X + Y$. Let $Z = X + Y$. The event $(Z \in dz)$ is shaded in the following diagram:



The event $(Z \in dz)$ can be broken up into vertical slices according to the values of X, as suggested by the vertical shading in the diagram. The heavily shaded parallelogram contained in the event $(Z \in dz)$ near the point $(x, z - x)$, represents the intersection of the events $(X \in dx)$ and $(Z \in dz)$, and has area $dx\,dz$. The probability density near this little parallelogram is $f(x, z - x)$, so

(a1) $$P(X \in dx, Z \in dz) = f(x, z - x)dx\,dz$$

This formula gives the joint density of X and Z. The marginal density of $Z = X + Y$ is therefore obtained by integrating out the x-variable

(a2) $$P(Z \in dz) = \left[\int_{-\infty}^{\infty} f(x, z - x)dx\right] dz$$

This gives the boxed formula for the density of $Z = X + Y$. Intuitively, you can think of (a2) as obtained by summing over infinitesimal parallelograms as in (a1).□

Example 1. Sums of independent exponential variables.

In Section 4.2 a Poisson process argument was used to show that the distribution of the sum of r independent exponential (λ) random variables is gamma (r, λ): If $f_{r,\lambda}(t)$ denotes the density of such a sum, then

$$f_{r,\lambda}(t) = \frac{1}{(r-1)!}\lambda^r t^{r-1} e^{-\lambda t} \qquad (t \geq 0)$$

This fact can also be derived using the convolution formula. Here is the calculation for $r = 2$.

Suppose T and U are independent, each exponentially distributed with rate λ. By independence, the joint density of T and U at (t, u) is

$$f(t,u) = f_T(t)f_U(u) = \lambda e^{-\lambda t}\lambda e^{-\lambda u} = \lambda^2 e^{-\lambda(t+u)} \qquad (t, u \geq 0)$$

Note how this joint density is a function of $t + u$. You can see the effect of this in Figure 1.

The density of $S = T + U$ at s is given by the convolution formula

$$\begin{aligned}
f_S(s) &= \int_{-\infty}^{\infty} f_T(t)f_U(s-t)dt \\
&= \int_0^s f_T(t)f_U(s-t)dt \qquad \text{since } f_T(t) = 0 \text{ if } t < 0 \\
&\qquad\qquad\qquad\qquad\qquad \text{and} \quad f_U(s-t) = 0 \text{ if } t > s \\
&= \int_0^s \lambda e^{-\lambda t}\lambda e^{-\lambda(s-t)}dt \\
&= \int_0^s \lambda^2 e^{-\lambda s}dt \\
&= \lambda^2 s e^{-\lambda s} \qquad (s \geq 0)
\end{aligned}$$

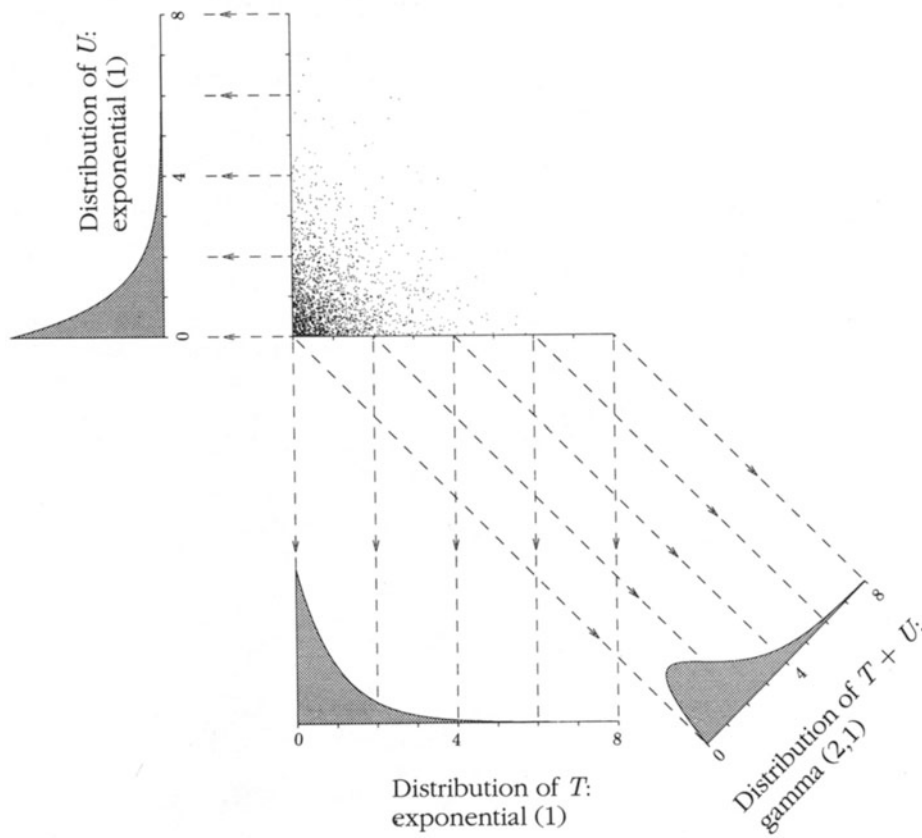
See Figure 1. For small s the factor of s makes the density grow linearly near zero. For large s the exponential factor $e^{-\lambda s}$ brings the density down to zero very rapidly.

Another way to derive this density is to argue infinitesimally: Let $s \geq 0$. The probability of $(S \in ds)$ is the integral of the joint density over the infinite strip $((t, u) : s \leq t + u \leq s + ds)$. We need only integrate over the (approximately) rectangular segment $((t, u) : s \leq t + u \leq s + ds,\ t \geq 0, u \geq 0)$, where the joint density is nonzero. This segment has length $\sqrt{2}s$ and width $ds/\sqrt{2}$, and the joint density has nearly constant value $\lambda^2 e^{-\lambda(t+u)} = \lambda^2 e^{-\lambda s}$ for points (t, u) in this segment; so the desired probability is

$$P(S \in ds) = \sqrt{2}s \cdot ds/\sqrt{2} \cdot \lambda^2 e^{-\lambda s} = \lambda^2 s e^{-\lambda s} ds \qquad (s \geq 0)$$

The fact that the sum of r independent exponential (λ) variables has gamma (r, λ) distribution can be derived from the convolution formula by mathematical induction on r.

FIGURE 1. **Distribution of the sum of two independent exponential variables.** Here is a random scatter of points suggesting the joint density of independent exponential variables T and U, along with graphs of the densities of T, U, and $S = T + U$.



Example 2. **Sums of independent gamma variables.**

Recall from Section 4.2 that the gamma (r, λ) distribution is defined for every real $r > 0$ by the density

$$f_{r,\lambda}(t) = \begin{cases} [\Gamma(r)]^{-1}\lambda^r t^{r-1} e^{-\lambda t} & t > 0 \\ 0 & t \leq 0 \end{cases} \qquad \text{where} \quad \Gamma(r) = \int_0^\infty t^{r-1} e^{-t} dt$$

If T_r and T_s are independent random variables with gamma (r, λ) and gamma (s, λ) distributions, respectively, then $T_r + T_s$ has gamma $(r + s, \lambda)$ distribution.

Proof for positive integers r and s. This case follows from the representation of a gamma variable as the sum of independent exponential variables. To see how, note first that the density of an independent sum $T_r + T_s$ is determined by the densities of T_r and T_s, by the convolution formula. So it is enough to derive the result for any convenient pair of independent random variables with gamma (r, λ) and gamma (s, λ) distributions. But the conclusion is obvious if we consider

$$T_r = W_1 + \cdots + W_r \quad \text{and} \quad T_s' = W_1' + \cdots + W_s'$$

defined by $r + s$ independent exponentials $W_1 \ldots, W_r, W_1', \ldots, W_s'$. Because then

$$T_r + T_s' = W_1 + \cdots + W_r + W_1' + \cdots + W_s'$$

is the sum of $r + s$ independent exponentials, with gamma $(r + s, \lambda)$ distribution. □

Proof for positive half-integers r and s. The case $r = n/2$ and $s = m/2$ for positive integers n and m can be derived almost the same way, using the result found in Section 5.3 that the gamma $(n/2, 1/2)$ distribution is the chi-square distribution of the sum of squares of n independent standard normal variables. Adding the sum of squares of n variables to the sum of squares of m variables gives the sum of squares of $n+m$ variables. Changing the rate parameter $1/2$ to a general λ is just a matter of multiplying of the chi-square variables by $1/(2\lambda)$. (See Exercises 5.3.15 and 4.4.2). □

Proof for general positive r and s. For $r > 0$, $s > 0$, let T_r and T_s be independent, with gamma (r, λ) and gamma (s, λ) distributions, and let $Z = T_r + T_s$. Then by the convolution formula

$$\begin{aligned}
f_Z(z) &= \int_0^z f_{T_r}(x) f_{T_s}(z - x)dx \\
&= \int_0^z \frac{1}{\Gamma(r)} \cdot \lambda^r x^{r-1} e^{-\lambda x} \cdot \frac{1}{\Gamma(s)} \lambda^s (z - x)^{s-1} e^{-\lambda(z-x)} dx \\
&= \int_0^z \frac{1}{\Gamma(r)\Gamma(s)} \lambda^{r+s} x^{r-1} (z - x)^{s-1} e^{-\lambda z} dx \\
&= \int_0^1 \frac{1}{\Gamma(r)\Gamma(s)} \lambda^{r+s} (zu)^{r-1} (z - zu)^{s-1} e^{-\lambda z} z\, du \qquad (x = zu, dx = z\, du) \\
&= \frac{1}{\Gamma(r + s)} \lambda^{r+s} z^{r+s-1} e^{-\lambda z} \int_0^1 \frac{\Gamma(r + s)}{\Gamma(r)\Gamma(s)} u^{r-1} (1 - u)^{s-1} du \\
&= f_{r+s,\lambda}(z) \int_0^1 \frac{\Gamma(r + s)}{\Gamma(r)\Gamma(s)} u^{r-1} (1 - u)^{s-1} du
\end{aligned}$$

where $f_{r+s,\lambda}(z)$ is the gamma $(r+s, \lambda)$ density. The integral on the right is a constant which does not depend on z. Since both $f_Z(z)$ and $f_{r+s,\lambda}(z)$ are probability densities on $(0, \infty)$, integrating both sides with respect to z from 0 to ∞ gives

$$1 = 1 \times \int_0^1 \frac{\Gamma(r + s)}{\Gamma(r)\Gamma(s)} u^{r-1} (1 - u)^{s-1} du$$

So the integral must equal 1. Therefore Z has the gamma $(r+s, \lambda)$ density. □

The last line of the previous argument evaluates an important integral:

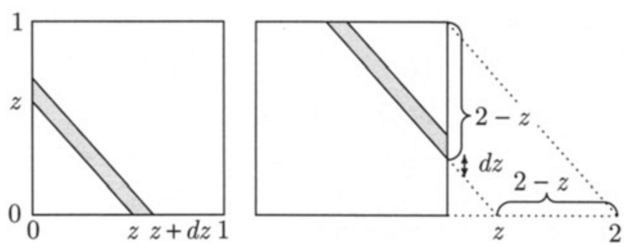
The Beta Integral

$$B(r,s) = \int_0^1 u^{r-1}(1-u)^{s-1}du = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)} \qquad (r>0, s>0)$$

This evaluation of $B(r,s)$ in terms of the gamma function agrees with the evaluation in Section 4.6 for integer r and s because $\Gamma(r) = (r-1)!$ for positive integers r.

Example 3. **Sums of independent uniform variables.**

Two terms. Suppose X and Y are independent, each with uniform $(0,1)$ distribution. To find the density of $X+Y$ it is simpler to work directly with a diagram than to use the convolution formula. Here (X,Y) has uniform distribution on the unit square. See Figure 2 on page 380.



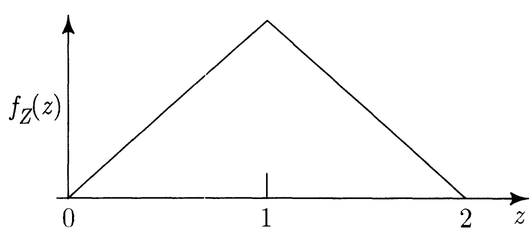
For $0 < z < 1$, the event $(X+Y \in dz)$ is represented as in the diagram by a shape of area $z\,dz + \frac{1}{2}(dz)^2$, by splitting the area into a parallelogram with altitude z perpendicular to sides of length dz, plus half a square of side dz. Ignoring the $(dz)^2$ as negligible in comparison to dz, gives simply

$$P(Z \in dz) = z\,dz$$

since the total area is 1. Similarly, for $1 \le z < 2$,

$$P(Z \in dz) = (2-z)dz$$

Thus $Z = X+Y$ has a tent-shaped density,



$$f_Z(z) = \begin{cases} z & 0 < z < 1 \\ 2-z & 1 \le z < 2 \\ 0 & \text{otherwise} \end{cases}$$

Three terms. Consider now $T = X+Y+W$ where X, Y, and W are independent uniform $(0,1)$. The joint distribution of (X,Y,W) is now uniform on a unit cube, and the density of T is proportional to the areas of slices through the cube perpendicular to an axis passing through the long diagonal. As you can convince yourself by handling a real cube, there are now several cases depending on which faces of the cube cut the slicing plane. This 3-dimensional geometry is tricky, but it reduces to two simpler two-dimensional problems.

To compute the density of $T = X + Y + W$ where X, Y, and W are independent uniform $(0, 1)$, write $T = X+Y+W = Z+W$, say, where the density of $Z = X+Y$ was found before. The convolution formula gives the density of $T = Z + W$ as an integral

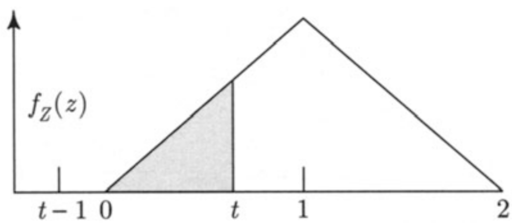
$$\begin{aligned} f_T(t) &= \int_{-\infty}^{\infty} f_Z(z) f_W(t-z)dz \\ &= \int_{t-1}^{t} f_Z(z)dz \qquad \text{since} \quad f_W(t-z) = 1 \text{ for } t-1 < z < t, \quad 0 \text{ else} \\ &= P(t-1 < Z < t) \qquad \text{by definition of } f_Z \end{aligned}$$

So the probability density of T at t turns out in this case to be a probability defined in terms of Z and t. This probability is represented by the shaded areas under the density $f_Z(z)$ in the diagrams that follow. There are 3 cases to consider.

Case 1. $0 < t < 1$. Then $t - 1 < 0$, so

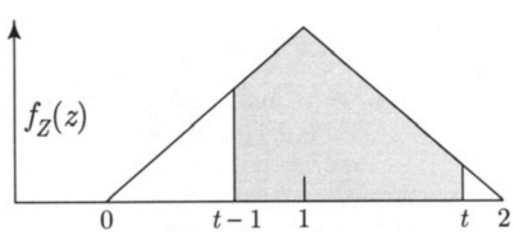
$$f_T(t) = P(t-1 < Z < t) = \frac{1}{2}t^2$$

by the formula for area of a triangle.



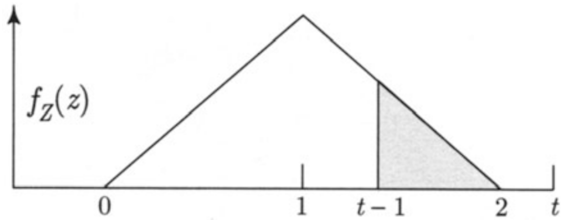
Case 2. $1 < t < 2$. Then $0 < t - 1 < 1$. The relevant area is a unit square less two triangles, hence

$$\begin{aligned} f_T(t) &= P(t-1 < Z < t) \\ &= 1 - \frac{1}{2}(2-t)^2 - \frac{1}{2}(t-1)^2 \\ &= -t^2 + 3t - \frac{3}{2} \end{aligned}$$



Case 3. $2 < t < 3$. Then $1 < t - 1 < 2$. The relevant area is now another triangle

$$f_T(t) = P(t - 1 < Z < t) = \frac{1}{2}(3 - t)^2$$



To summarize, the density of the sum $T = X + Y + W$ of three independent uniform $(0, 1)$ variables is $f_T(t)$, as defined above by quadratic functions of t, on each of the intervals $(0, 1)$, $(1, 2)$, and $(2, 3)$, and zero elsewhere. See Figures 2 and 3. Note the symmetric bell shape of the density of T.

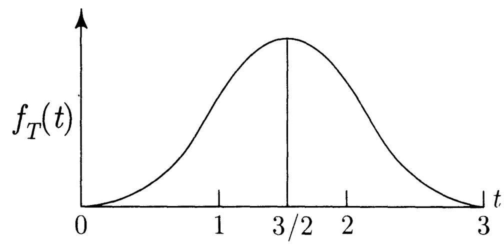


Illustration of Example 3 by numerical calculations. Let $T = X + Y + W$ where X, Y, and W are independent with uniform $(0, 1)$ distribution. Let us find:

a) $P(T < 3/2) = 1/2$ by symmetry of the density of T about $3/2$,

b) $P(1/2 < T < 3/2) = P(T < 3/2) - P(T \le 1/2)$
$= 1/2 - \int_0^{1/2} \frac{t^2}{2} dt = \frac{23}{48}$ by a) and Case 1 on page 378

c) $P(T > 5/2) = P(T \le 1/2) = 1/48$ by integral evaluated in b);

d) $E(T) = 3/2$ by symmetry;

e) $SD(T) = \sqrt{3} SD(W)$ where W is uniform $(0, 1)$, by the square root law
$= \sqrt{3} \cdot 1/\sqrt{12}$ by calculation done in Section 4.1
$= 1/2$

FIGURE 2. Distribution of the sum of two independent uniform $(0, 1)$ variables X and Y. The joint density of (X, Y) is suggested by a scatter, along with graphs of the densities of X, Y, and $X + Y$.

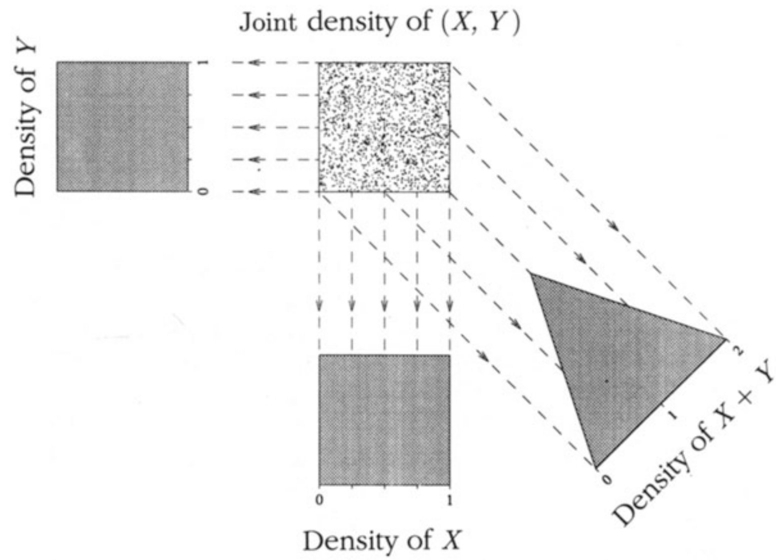


FIGURE 3. Distribution of the sum of three independent uniform $(0, 1)$ variables X, Y, and W. The joint density of $(X + Y, W)$ is suggested by a scatter, along with graphs of the densities of $X + Y$, W, and $X + Y + W$.

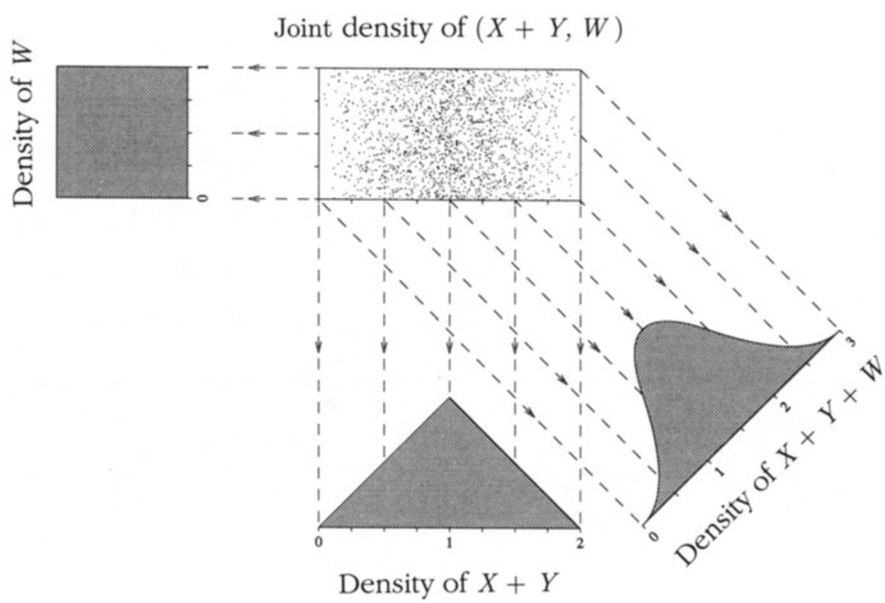
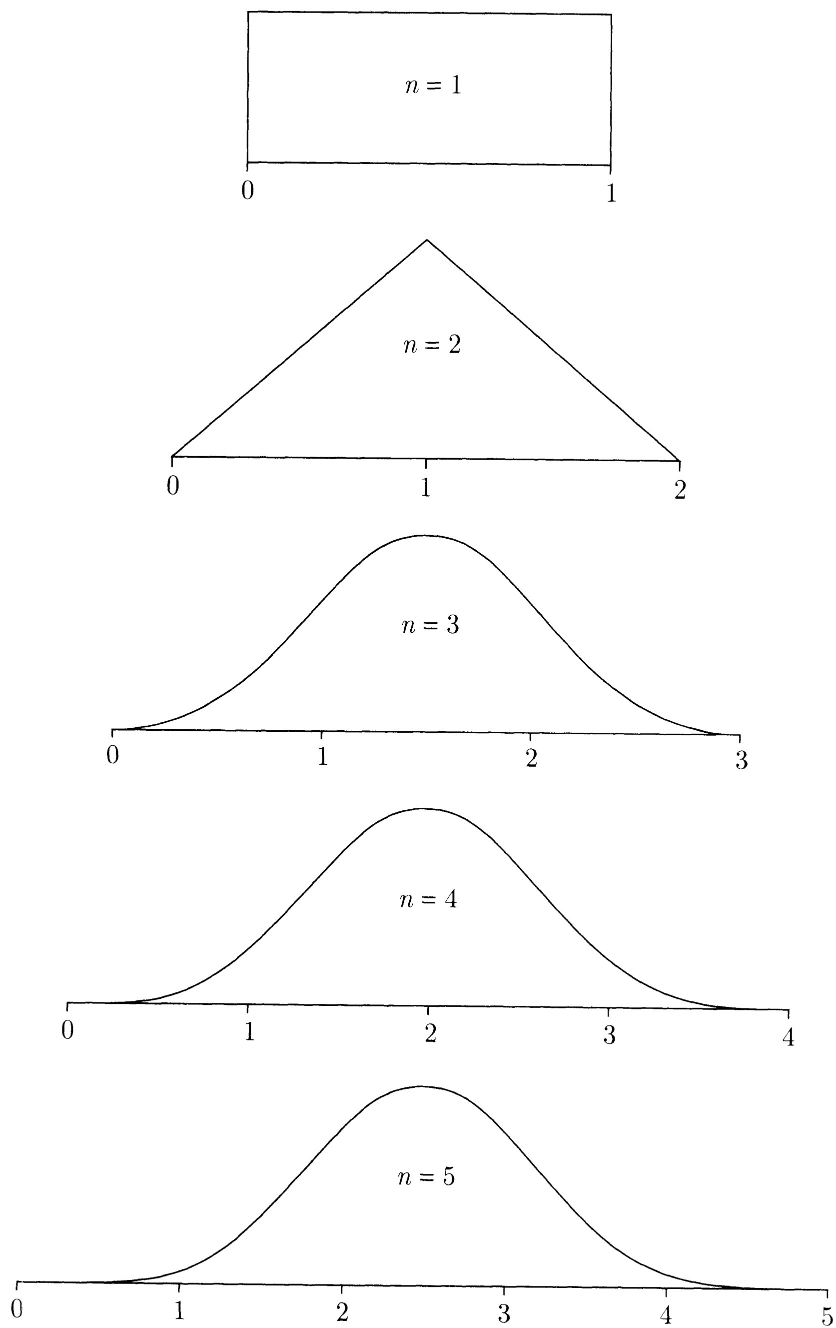


FIGURE 4. Density of the sum of n independent uniform $(0, 1)$ variables. The graphs are all centered at the mean with a constant horizontal distance on the page representing one standard unit in each graph. This shows how rapidly the shape of the distribution becomes normal as n increases.



Example 4. Roundoff errors.

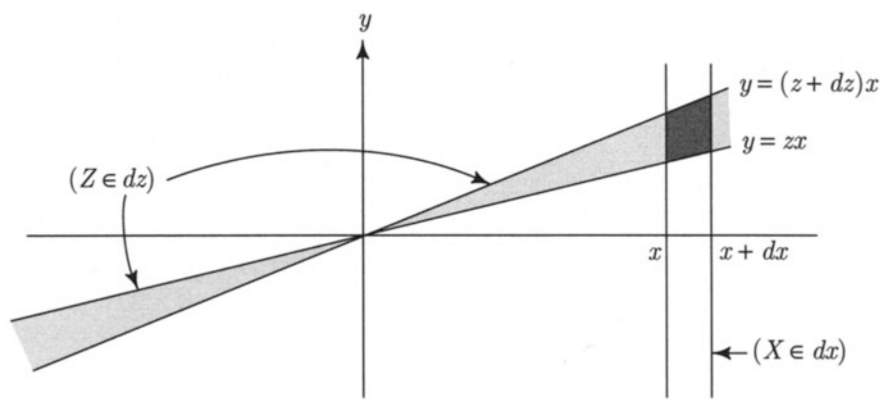
Problem. Suppose three numbers are computed, each with a roundoff error known to be smaller than 10^{-6} in absolute value. If the roundoff errors are assumed independent and uniformly distributed, what is the probability that the sum of the rounded numbers differs from the true sum of the numbers by more than 2×10^{-6}?

Solution. Let X_i be the error in the ith number in multiples of 10^{-6}, so the X_i are independent uniform $(-1, 1)$. To reduce to previous calculations, let $U_i = (X_i + 1)/2$, so the U_i are independent uniform $(0, 1)$. The problem is to find

$$\begin{aligned} P(|X_1 + X_2 + X_3| > 2) &= 2P(X_1 + X_2 + X_3 > 2) \quad \text{by symmetry} \\ &= 2P(2U_1 - 1 + 2U_2 - 1 + 2U_3 - 1 > 2) \\ &= 2P(U_1 + U_2 + U_3 > 5/2) \\ &= 2/48 \quad \text{by numerical calculation c) of Example 3 above.} \end{aligned}$$

Distribution of Ratios

Let $Z = Y/X$. The event $(Z \in dz)$ is shaded in the following diagram, for $z > 0$.



The event $(Z \in dz)$ is broken up into vertical slices according to values of X. The heavily shaded region, near (x, xz), represents the event $(X \in dx, Z \in dz)$. For small dx and dz this region is approximately a parallelogram. The left side has length $|x|dz$, and there is distance dx between the two vertical sides, so the area of the parallelogram is approximately $|x|dz\,dx = |x|dx\,dz$. The probability density over the small parallelogram can be taken to be $f(x, xz)$, so as dx and dz tend to zero we obtain the formula

$$P(X \in dx, Y/X \in dz) = f(x, xz)|x|dx\,dz$$

This works just as well for $z < 0$, though the picture looks a little different. Integrating out x yields

(f) $$f_{Y/X}(z) = \int_{-\infty}^{\infty} |x| f(x, xz)\,dx$$

As a special case, if X and Y are independent positive random variables, (f) reduces to $f_{Y/X}(z) = 0$ for $z \le 0$, and

(g) $$f_{Y/X}(z) = \int_{0}^{\infty} x f_X(x) f_Y(xz)\,dx \qquad (0 < z < \infty)$$

Example 5. **Ratio of independent normal variables.**

Suppose that X and Y are independent and normally distributed with mean 0 and variance σ^2.

Problem. Find the distribution of X/Y.

Solution. We may assume $\sigma = 1$, since

$$\frac{X}{Y} = \frac{X/\sigma}{Y/\sigma} \quad \text{and both} \quad X/\sigma \quad \text{and} \quad Y/\sigma \qquad \text{are standard normal.}$$

By symmetry between X and Y and (f) above

$$\begin{aligned} f_{X/Y}(z) = f_{Y/X}(z) &= \int_{-\infty}^{\infty} |x| f_{X,Y}(x, xz)\,dx \\ &= \int_{-\infty}^{\infty} |x| \cdot \frac{1}{2\pi} e^{-\frac{x^2 + x^2 z^2}{2}}\,dx \\ &= \int_{0}^{\infty} \frac{1}{\pi} x e^{-\frac{x^2(z^2+1)}{2}}\,dx \\ &= \frac{1}{\pi} \cdot \frac{(-1)}{z^2+1} \cdot e^{-\frac{x^2(z^2+1)}{2}} \Big|_0^{\infty} \\ &= \frac{1}{\pi(z^2+1)} \end{aligned}$$

That is, X/Y has Cauchy distribution (see Exercise 4.4.6).

Remark. This calculation illustates the general method, but is a bit heavy handed. In fact the distribution of X/Y is Cauchy whenever the joint distribution of X and Y is symmetric under rotations. See Exercise 14 below.

Exercises 5.4

1. Let X_1 be uniform $(0,1)$ independent of X_2, that is, uniform $(0,2)$. Find:

a) $P(X_1 + X_2 \le 2)$; b) the density of $X_1 + X_2$; c) the c.d.f. of $X_1 + X_2$.

2. Let S_n be the sum of n independent uniform $(0,1)$ random variables. Find

a) $P(S_2 \le 1.5)$; b) $P(S_3 \le 1.5)$; c) $P(S_3 \le 1.1)$;

d) $P(1.0 \le S_3 \le 1.001)$ approximately.

3. A computer job must pass through two queues before it is processed. Suppose the waiting time in the first queue is exponential with rate α, and the waiting time in the second queue is exponential with rate β, independent of the first.

a) Find the density of the total time the job spends waiting in the two queues. Sketch the density in case $\alpha = 1$ and $\beta = 2$.

b) Find the expected total waiting time in terms of α and β.

c) Find the SD of the total waiting time in terms of α and β.

4. A system consists of two components. Suppose each component is subject to failure at constant rate λ, independently of the other, up to when the first component fails. After that moment the remaining component is subject to additional load and to failure at constant rate 2λ.

a) Find the distribution of the time until both components have failed.

b) What are the mean and variance of this distribution?

c) Find the 90th percentile of this distribution.

5. Let X be the number on a die roll, between 1 and 6. Let Y be a random number which is uniformly distributed on $[0,1]$, independent of X. Let $Z = 10X + 10Y$.

a) What is the distribution of Z? Explain.

b) Find $P(29 \le Z \le 58)$.

6. Suppose $X_1, X_2, \ldots, X_n$ are independent and X_i has gamma (r_i, λ) distribution. What is the distribution of $X_1 + X_2 + \cdots + X_n$? Explain.

7. Let X and Y have joint density $f(x,y)$. Find formulae for the densities of each of the following random variables: a) XY; b) $X - Y$; c) $X + 2Y$.

8. Let X and Y be independent exponential variables with rates α and β. Find the c.d.f. of X/Y.

9. Find the density of $X = UV$ for independent uniform $(0,1)$ variables U and V.

10. Find the density of $Y = U/V$ for independent uniform $(0,1)$ variables U and V.

11. Find the distribution of $\min(U,V)/\max(U,V)$ for independent uniform$(0,1)$ variables U and V.

12. Let U, V be independent random variables, each uniform on $(0,1)$.

(a) Find the density of $X = -\log\{U(1-V)\}$. b) Compute $E(X)$ and $Var(X)$.

13. Find the density of $Z = X - Y$ for independent exponential (λ) variables X and Y.

14. Let X and Y have a joint distribution which is symmetric under rotations (e.g., uniform on a circle around 0, or uniform on a disc centered at 0). By changing to polar coordinates, show that

a) the distribution of X/Y is Cauchy (see Exercise 4.4.6);

b) the distribution of $X^2/(X^2 + Y^2)$ is arcsine (see Exercise 4.4.8).

15. Let $Z = \min(X, Y)/\max(X, Y)$ for independent exponential (λ) variables X and Y.

a) Explain with little calculation why the distribution of Z does not depend on λ

b) Let $0 < z < 1$. Identify the set $(Z \leq z)$ as a subset of the (x, y) plane, and calculate $P(Z \leq z)$ by integration of the joint density over this subset.

c) Find the density of Z at z for $0 < z < 1$.

16. Consider the c.d.f. of T with gamma (r, λ) distribution, $F(r, \lambda, t) = P(T \leq t)$. Section 4.2 gives a formula for $F(r, \lambda, t)$ for integer r, but for r not an integer there is no simple formula for $F(r, \lambda, t)$.

a) Show that for fixed r and t, $F(r, \lambda, t)$ is an increasing function of λ. [*Hint*: Rescale to the gamma $(r, 1)$ distribution.]

b) Show that for fixed λ and t, $F(r, \lambda, t)$ is a decreasing function of r. [*Hint*: Use sums of independent gamma variables.]

17. Take a unit cube in three dimensions. Cut the cube by a plane perpendicular to the line from its corners $(0, 0, 0)$ and $(1, 1, 1)$, that cuts this line at the point $(t/3, t/3, t/3)$.

a) What is the cross-sectional area of this slice through the cube?

b) Check your answer by describing geometrically the shape of the cross section in the case when $t \leq 1$ and $t = 3/2$.

18. Let f_n be the density function and F_n the c.d.f. of the sum S_n of n independent uniform $(0, 1)$ random variables.

a) Show that $f_n(x) = F_{n-1}(x) - F_{n-1}(x - 1)$.

b) Show that on each of the n intervals $(i - 1, i)$ for $i = 1$ to n, f_n is equal to a polynomial of degree $n - 1$, and F_n is equal to a polynomial of degree n.

c) Find $f_n(x)$ and $F_n(x)$ for $0 \leq x \leq 1$.

d) Find $f_n(x)$ and $F_n(x)$ for $n - 1 \leq x \leq n$.

Find also: e) $P(0 \leq S_4 \leq 1)$; f) $P(1 \leq S_4 \leq 2)$; g) $P(1.5 \leq S_4 \leq 2)$.

19. Let X and Y be independent variables with gamma (r, λ) and gamma (s, λ) distribution, respectively. Show that $X/(X+Y)$ has beta (r, s) distribution, independently of $X+Y$.

Continuous Joint Distributions: Summary

Differential Formula for Joint Density

$$P(X \in dx, Y \in dy) = f(x,y)dx\,dy$$

The density $f(x,y)$ is the probability per unit area for values near (x,y). See pages 348 and 349 of Section 5.2 for properties of joint densities, and comparison with joint distributions in the discrete case.

Central Limit Theorem

Let $X_1, X_2, \ldots$ be a sequence of independent, identically distributed random variables, each with mean μ and variance σ^2. Let $S_n = X_1 + \cdots + X_n$. Provided $\sigma^2 < \infty$, the limit distribution, as $n \to \infty$, of the standardized sum $Z_n = [S_n - n\mu]/(\sqrt{n}\sigma)$ is the standard normal distribution.

Formula for Density of $X + Y$

If (X, Y) has density $f(x,y)$ in the plane, then $X + Y$ has density on the line

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f(x, z-x)dx.$$

Convolution Formula

If X and Y are independent, then

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx.$$

Exact distribution of various functions of particular variables. See distribution summaries.

The Rayleigh Distribution

If X and Y are independent standard normal variables, then $R = \sqrt{X^2 + Y^2}$ has the Rayleigh distribution, with density

$$f_R(r) = re^{-\frac{1}{2}r^2}, \qquad r > 0,$$

and distribution function

$$F_R(r) = 1 - e^{-\frac{1}{2}r^2}, \qquad r > 0.$$

The variable R represents the distance from the origin of the random point (X, Y).

Review Exercises

1. For X and Y independent and uniform $(0,1)$, find $P(Y \geq 1/2 \,|\, Y \geq X^2)$.

2. For X and Y independent and both uniform $(-1,1)$, find

   a) $P(|X+Y| \leq 1)$; b) $E(|X+Y|)$.

3. **Coin in a can.** A coin of diameter 1 inch is tossed in the air and caught in an empty soup can of bottom radius 3 inches. The coin lies flat on the bottom.

   a) What is the chance that the coin covers the center point of the bottom of the can?

   Suppose that instead of the soup can, the coin is dropped into a box whose bottom is a square with sides of length 5 inches.

   b) What is the chance that the coin covers the center point of the bottom of the box?

   c) Consider one of the main diagonals of the bottom of the box. What is the probability that part of the coin crosses that diagonal line?

   State any assumptions you make.

4. Let X and Y be independent with uniform $(-1,1)$ distribution. Find

   a) $P(X^2+Y^2) \leq r^2$; b) the c.d.f. of $R^2 = X^2+Y^2$; c) the density of R^2.

5. A point is chosen uniformly at random from a unit square. Let D be the distance of the point from the midpoint of one side of the square. Find a) $P(D \geq \frac{1}{2})$; b) $E(D^2)$.

6. For a particular kind of call, the phone company charges \$1 for the first minute or any portion thereof, and one cent per second for time after the first minute. Calculate the approximate value of the long-run average charge per call assuming the distribution of call duration is:

   a) exponential with mean 1 minute;

   b) exponential with mean 2 minutes;

   c) gamma with shape parameter 2 and mean 1 minute.

7. Suppose that $X_1, X_2, \ldots, X_{100}$ are independent random variables, with normal $(\mu, 1)$ distribution, representing 100 measurements whose average $\bar{X} = (X_1 + \cdots + X_{100})/100$ should be close to the number μ. Calculate the probability that $|\bar{X} - \mu| \geq 0.25$.

8. Suppose that $X_1, \ldots, X_{100}$ are independent random variables with common distribution with mean μ and variance 1, but not necessarily normally distributed. Repeat Exercise 7 with these assumptions. Explain why the answer will be approximately the same.

9. Let X be the number of heads in two fair coin tosses. Suppose U has uniform distribution on $(0,1)$, independently of X.

   a) Find the density of $X+U$ and sketch its graph.

   b) Find an alternate distribution for U such that for any integer-valued random variable X independent of U, the graph of the density of $X+U$ is simply the usual histogram of the distribution of X.

10. Let X, Y be independent exponential random variables with parameters λ and μ.

a) Find the density function for $Z = \min(X, Y)$.

b) Calculate $P(X \geq Y)$.

c) Calculate $P(\frac{1}{2} < X/Y < 2)$, in the case $\lambda = \mu$. [*Hint*: Use the result of b).]

11. Let U and V be two independent uniform $(0, 1)$ random variables. Let $X = U/V$.

a) For $0 < x < 1$, calculate $P(X > x)$.

b) Find the c.d.f. F of the random variable X. Sketch the graph of F.

c) Find the density function f of X. Sketch the graph of f.

12. A marksman fires at the center of a target; he hits a random point (X, Y) (measured relative to the center of the target) such that X and Y are independent normal (0, a^2) random variables. A second marksman fires, and hits at (X', Y') where X' and Y' are independent with normal (0, b^2) distributions. What is the chance that the second marksman hits closer to the center of the target than the first marksman?

13. Suppose (X, Y) is uniformly distributed according to relative arc length on the circumference of the circle $\{(x, y) : x^2 + y^2 = 1\}$. Find the c.d.f. of

a) X; b) Y; c) $X + Y$.

14. Suppose U_1, U_2, U_3 are independent and uniform $(0, 1)$. Find: a) $P(U_1 < U_2 < U_3)$;

b) $E(U_1U_2U_3)$; c) $Var(U_1U_2U_3)$; d) $P(U_1U_2 > U_3)$; e) $P(\max(U_1, U_2) > U_3)$.

15. Repeat Exercise 14 for Z_i instead of U_i, where the Z_i are independent normal $(0, 1)$ random variables. Find also:

f) $P(Z_1^2 + Z_2^2 > 1)$ g) $P(Z_1 + Z_2 + Z_3 < 2)$;

h) $P(Z_1/Z_2 < 1)$; i) $P(3Z_1 - 2Z_2 < 4Z_3 + 1)$.

16. A point is picked randomly in space. Its three coordinates X, Y and Z are independent standard normal variables. Let $R = \sqrt{X^2 + Y^2 + Z^2}$ be the distance of the point from the origin. Find

a) the density of R^2; b) the density of R; c) $E(R)$; d) $Var(R)$.

17. Let $X_1, X_2, \ldots,$ be independent normally distributed random variables having mean 0 and variance 1. Use the normal approximation to find:

a) $P(X_1^2 + X_2^2 + \cdots + X_{100}^2 \geq 80)$;

b) a number c such that $P(100 - c \leq X_1^2 + \cdots + X_{100}^2 \leq 100 + c) \approx 0.95$.

18. For X and Y independent normal $(0, 1)$ variables, show that for $r > 0$

$$P(aX + bY \leq r\sqrt{a^2 + b^2} \text{ for all } a, b \geq 0) = \Phi(r) - \frac{1}{4}e^{-\frac{1}{2}r^2}$$

19. Independent Poisson processes. Suppose particles of d different kinds, labeled $k = 1, 2, \ldots, d$, arrive at a counter according to independent Poisson processes at rates λ_k. Let W_k be arrival time of the first particle of kind k. Let K_1 be the kind of the first particle to arrive, K_2 the kind of the second particle to arrive, and so on. So the K_n are discrete random variables with values in the set $\{1, \ldots, d\}$.

a) Express the event $(K_1 = k)$ in terms of the random variables $W_1, \ldots, W_d$.

b) Use this expression to find $p_k = P(K_1 = k), 1 \le k \le d$ in terms of $\lambda_1, \ldots, \lambda_d$.

c) Explain informally why $K_1, K_2, \ldots$ are independent with identical distribution.

d) Assuming the result of c), derive the formula for p_k in another way after filling in the blanks in the following statements e), f) and g): After a very long time T,

e) the number of arrivals of type k should be about ________.

f) the number of all arrivals of all types should be about ________.

g) the fraction of all arrivals that are of type k should be about ________.

20. Minimum of independent exponential variables. Let T_1 and T_2 be two independent exponential variables, with rates λ_1 and λ_2. Think of T_i as the lifetime of component i, $i = 1, 2$. Let $T_{\min}$ represent the lifetime of a system which fails whenever the first of the two components fails, so $T_{\min} = \min(T_1, T_2)$. Let $X_{\min}$ designate which component failed first, so $X_{\min}$ has value 1 if $T_1 < T_2$ and value 2 if $T_2 < T_1$. Show:

a) that the distribution of $T_{\min}$ is exponential $(\lambda_1 + \lambda_2)$;

b) that the distribution of $X_{\min}$ is given by the formula $P(X_{\min} = i) = \dfrac{\lambda_i}{\lambda_1 + \lambda_2}$ for $i = 1, 2$;

c) that the random variables $T_{\min}$ and $X_{\min}$ are independent;

d) how these results generalize simply to describe the minimum of n independent exponential random variables with rates $\lambda_1, \ldots, \lambda_n$.

21. Closest point. Consider a Poisson random scatter of points in a plane with mean intensity λ per unit area. Let R be the distance from 0 to the closest point of the scatter.

a) Find formulae for the c.d.f. and density of R, and sketch their graphs.

b) Show that $\sqrt{2\lambda\pi}R$ has the Rayleigh distribution described in Section 5.3

c) Use b) to find formulae for the mean and SD of R from results of Section 5.3.

d) Find the mode and the median of the distribution of R.

22. In Maxwell's model of a gas, molecules of mass m are assumed to have velocity components, V_x, V_y, V_z that are independent, with a joint distribution that is invariant under rotation of the three-dimensional coordinate system. Maxwell showed that V_x, V_y, V_z must have normal $(0, \sigma^2)$ distribution for some σ. Taking this result for granted:

a) find a formula for the density of the kinetic energy

$$K = \frac{1}{2}mV_x^2 + \frac{1}{2}mV_y^2 + \frac{1}{2}mV_z^2$$

b) find the mean and mode of the energy distribution.

23. Let Y be the minimum of three independent random variables with uniform distribution on $(0, 1)$, and let Z be their maximum. Find:

a) $P(Z \le \frac{2}{3} | Y \ge \frac{1}{3})$; b) $P(Z \le \frac{2}{3} | Y \le \frac{1}{3})$.

24. A coin of diameter d is tossed at random on a grid of squares of side s. Making appropriate assumptions, to be stated clearly, calculate:

a) the probability that the coin lands inside some square (i.e., not touching any line);

b) the probability that the coin lands heads inside some square.

Suppose now that the coin is tossed four times. Let X be the number of times it lands inside a square, Y the number of heads. Assume $d = s/2$. Calculate:

c) $P(X = Y)$; d) $P(X < Y)$; e) $P(X > Y)$.

25. Joint distribution of order statistics. Let $V_1 < V_2 < \cdots < V_n$ be the order statistics of n independent uniform $(0, 1)$ variables. (Refer to Section 4.6.) Let $1 \le k < m \le n$.

a) Find the joint density of V_k and V_m.

Now show that each of the following variables has a beta distribution, and identify the parameters: b) $V_m - V_k$; c) V_k/V_m;

26. Averages of order statistics. Let $V_1, \ldots, V_n$ be the order statistics of n independent uniform $(0, 1)$ variables. Let

$A_{\text{all}} = (V_1 + \cdots + V_n)/n$

$A_{1n} = (V_1 + V_n)/2$

$A_{\text{mid}} = V_{(n+1)/2}$ the middle value, where you can assume n is odd.

a) Show that for sufficiently large n, each of these three variables is most likely very close to $1/2$.

b) For all large enough values of n, one of these variables can be expected to be very much closer to $1/2$ than either of the two others. Which one, and why?

c) Confirm your answer to b) for $n = 100$ by finding for each of the A's a good approximation to the probability that it is between 0.49 and 0.51.

27. A box contains n balls numbered $1, \ldots, n$. Balls are drawn at random until the first draw that produces a ball obtained on some previous draw. Let D_n be the random number of draws required. So the possible values of D_n are $2, \ldots, n + 1$.

a) Check that for $0 < x < \infty$,

$$\lim_{n\to\infty} P(D_n/\sqrt{n} > x) = e^{-x^2/2}$$

That is to say, the limit distribution of $D_n/\sqrt{n}$ is the Rayleigh distribution.

b) Assuming a switch in the order of the limit and integration can be justified (it can, but do not worry about that), deduce that

$$\lim_{n\to\infty} E(D_n/\sqrt{n}) = \sqrt{\pi/2}$$

c) There seems to be no simpler expression for $E(D_n)$ than a sum of n or $n + 1$ terms. But the terms can be arranged in some interesting ways. Show by writing $E(D_n)$ as the sum of the tail probabilities $P(D_n > k)$ in reverse order that

$$E(D_n) = P(X_n \le n)\, n!\, n^{-n} e^n$$

where X_n is a Poisson random variable with mean n.

d) Deduce the limit of $P(X_n \le n)$ as $n \to \infty$ from the central limit theorem, then combine b) and c) to give a derivation of Stirling's formula

$$n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$$

e) Derive the following formula, which is surprisingly simple in view of c)

$$E(D_n^2 - D_n) = 2n$$

f) Transform the identity e) as in the calculation c) to derive the formula

$$E(|X_n - n|) = \frac{2n^{n+1}}{e^n n!}$$

and give yet another derivation of Stirling's formula, much as in d) above, this time using the central limit theorem instead of a).

28. Volumes in higher dimensions. Use the derivation of the chi-square distribution to derive part a), then use a) for the remaining parts:

a) Find the constant c_n such that the $(n-1)$-dimensional volume of the "surface" of a sphere of radius r in n-dimensional space is $c_n r^{n-1}$.

b) Find d_n so the n-dimensional volume inside a sphere of radius r is $d_n r^n$.

c) An n-dimensional sphere of radius r is packed inside an n-dimensional cube with sides of length $2r$. What proportion p_n of the volume of the cube is inside the sphere?

d) Use Stirling's formula $\Gamma(s) \sim \sqrt{2\pi} s^{s-1/2} e^{-s}$ as $s \to \infty$ to find a simple approximation for p_n for large n. What is the limit of p_n as $n \to \infty$?

e) Interpret p_n probabilistically in terms of n independent uniform $(-1, 1)$ variables.

29. A needle is tossed at random on a grid of equally spaced parallel lines. Assume the needle is so much longer than the spacing between the lines that the possibility of the needle not crossing any line can be neglected. Let X be the distance between the center of the needle and the closest point at which the needle crosses one of the lines. Find:

a) the distribution function of X;

b) the density function of X.

30. Random walk inside squares. Draw a square centered at $(0, 0)$ with sides of length 2 parallel to the axes, so the corners are at $(\pm 1, \pm 1)$. Let (X_1, Y_1) be picked uniformly at random from the area inside this square. Given (X_1, Y_1), draw a square centered at (X_1, Y_1), with sides of length 2 parallel to the axes, so the corners are at $(X_1 \pm 1, X_2 \pm 1)$. Let (X_2, Y_2), be picked uniformly at random from the area inside this square, and so on: Given $(X_1, Y_1), \ldots, (X_n, Y_n)$ let (X_{n+1}, Y_{n+1}) be picked uniformly at random from the area inside the square with corners at $(X_n \pm 1, Y_n \pm 1)$. For $n = 300$, use a normal approximation to find the following probabilities:

a) $P(|X_n| > 10)$; b) $P(|Y_n| > 10)$.

c) The probability that (X_n, Y_n) lies outside the square with corners at $(\pm 10, \pm 10)$.

d) The probability that (X_n, Y_n) lies outside the circle of radius 10 centered at $(0, 0)$.

31. Random walk inside circles. Fix $r > 0$. Draw a circle centered at $(0, 0)$ with radius r. Let (X_1, Y_1) be picked uniformly at random from the area inside this circle. Given (X_1, Y_1), draw a circle with radius r centered at (X_1, Y_1). Let (X_2, Y_2), be picked uniformly at random from the area inside this circle, and so on. Given $(X_1, Y_1), \ldots, (X_n, Y_n)$, let (X_{n+1}, Y_{n+1}) be picked uniformly at random from the area inside the circle around (X_n, Y_n) with radius r.

a) Find r so that for large n the distribution of X_n in this problem is nearly the same as in Exercise 30 for a square of side 2 instead of the circle of radius r. [*Hint*: Find $E[X_1^2]$ by considering $E[Y_1^2]$ as well.]

b) Are X_n and Y_n independent?

c) The point (X_n, Y_n) is projected onto the line rotated an angle θ from the X-axis at $X_n \cos\theta + Y_n \sin\theta$ measured from the origin along this line. Use the normal approximation for sums of independent random variables to show that with r as in part a), for every $\theta \in [0, 2\pi]$ and for large n, the distribution of $X_n \cos\theta + Y_n \sin\theta$ is nearly the same for both the circle of radius r and the square of side 2.

d) It is known that a joint distribution of (X, Y) in the plane is determined by the distributions of all the projections $X \cos\theta + Y \sin\theta$ as θ ranges over $[0, 2\pi]$. In particular if $X \cos\theta + Y \sin\theta$ has standard normal distribution for every θ then X and Y are independent standard normal variables. An approximate version of this result is also true: if $X \cos\theta + Y \sin\theta$ has approximately the standard normal distribution for every θ, then X and Y are approximately independent standard normal variables. Apply this result and part c) to approximate the probability that for r as in part a), and $n = 300$, the point (X_n, Y_n) defined using circles of radius r lies outside the circle of radius 10 centered at the origin.

32. Random walk on circles. Repeat Exercise 31 for the motion defined by picking points at random according to the uniform distribution on the *perimeter* of the circle of radius r, so each new point is at distance r from the previous one, in a random direction.

33. Mixture of discrete and continuous. Repeat Exercise 31 for the motion defined by repeatedly picking points at random according to the uniform distribution (proportional to length) on the perimeter of a square centered at the current point with sides of length $2r$. Note that the distribution of X_n in this case is neither discrete nor continuous but a mixture of the two kinds. The second moment of X_1 is defined by adding the discrete and continuous parts. It can be shown that the usual method of calculating the second moment of X_n is still valid, and that the normal approximation is still correct in the limit of large n. Following parts a) to d) as in Exercise 31,

e) Calculate and plot the graph of the distribution function of X_1.

f) Calculate and plot the graph of the distribution function of X_2.

g) What is the total probability in the discrete part of the distribution of X_n?

34. Ratios of sums of squares.

a) Use the result of Exercise 5.4.19 to show that if X, Y and Z are independent normal $(0, 1)$ random variables, then $X^2/(X^2 + Y^2 + Z^2)$ has beta $(1/2, 1)$ distribution, independent of $X^2 + Y^2 + Z^2$.

b) Suppose that (U, V, W) has uniform distribution on the surface of the unit sphere in three dimensions. Deduce from a) that $U^2/(U^2 + V^2 + W^2)$ has beta $(1/2, 1)$ distribution.

c) What is the distribution of $U^2/(U^2 + V^2 + W^2)$ if (U, V, W) has uniform distribution over the volume inside the unit sphere in three dimensions?

d) Suppose that $U_1, U_2, \ldots, U_n$ are independent uniform $(-1, 1)$ variables. For $1 \leq k \leq n$, let $S_k = U_1^2 + \cdots + U_k^2$. Find the the conditional distribution of S_k/S_n given that $S_n \leq 1$.

6
Dependence

This chapter treats features of a joint distribution which give insight into the nature of dependence between random variables. Sections 6.1 and 6.2 concern conditional distributions and expectations in the discrete case. Then parallel formulae for the density case are developed in Section 6.3. Covariance and correlation are introduced in Section 6.4. All these ideas are combined in Section 6.5 in a study of the bivariate normal distribution.

6.1 Conditional Distributions: Discrete Case

This section translates into the language of random variables the conditioning ideas of Section 1.4. The dependence between two variables X and Y can be understood in terms of the marginal distribution of X and the conditional distribution of Y given $X = x$, which may be a different distribution for each possible value x of X. Given this information, the distribution of Y is found by the rule of average conditional probabilities, and the conditional distribution of X given $Y = y$ is found by Bayes' rule.

Example 1. Number of successes in a random number of trials.

Suppose a fair die is rolled. Then as many fair coins are tossed as there are spots showing on the die.

Problem 1. Find the distribution of the number of heads showing among the coins.

Solution. Let Y denote the number of heads showing among the coins. The problem is to calculate the probabilities

$$P(Y = y) = P(y \text{ heads}) \qquad (y = 0, 1, 2, \ldots, 6)$$

Let X represent the number showing on the die. If $X = x$, that is to say the die rolls x, then x coins are tossed, so the chance of y heads given the die rolls x is given by the binomial formula for the probability of y successes in x trials with probability $1/2$ of success on each trial:

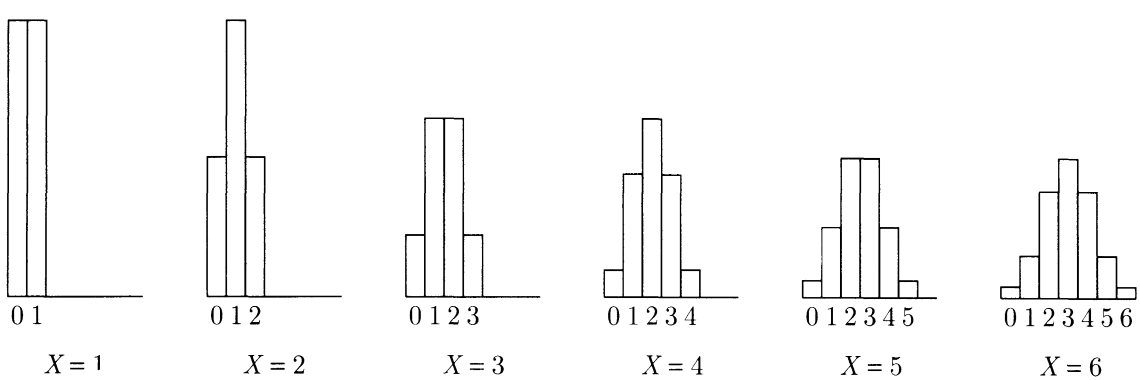
$$P(y \text{ heads} \,|\, \text{die rolls } x) = P(y \text{ heads in } x \text{ fair coin tosses}) = \binom{x}{y} 2^{-x}$$

where $\binom{x}{y} = 0$ if $x < y$. In random variable notation,

$$P(Y = y \,|\, X = x) = \binom{x}{y} 2^{-x}$$

This formula states that the *conditional distribution of Y given $X = x$* is the binomial distribution with parameters $n = x$ and $p = 1/2$.

FIGURE 1. Conditional distribution of Y given $X = x$ for $x = 1, 2, \ldots, 6$ in Example 1.



The assumption that the die is fair specifies the unconditional distribution of X:

$$P(X = x) = P(\text{die rolls } x) = 1/6 \qquad (x = 1, 2, \ldots, 6)$$

These ingredients are combined by the rule of average conditional probabilities to give $P(Y = y)$, the unconditional probability of getting y heads:

$$\begin{aligned}
P(Y = y) = P(y \text{ heads}) &= \sum_{x=1}^{6} P(\text{die rolls } x \text{ and } y \text{ heads}) \\
&= \sum_{x=1}^{6} P(y \text{ heads} \,|\, \text{die rolls } x) P(\text{die rolls } x) \\
&= \sum_{x=1}^{6} P(Y = y \,|\, X = x) P(X = x) \\
&= \frac{1}{6} \sum_{x=1}^{6} \binom{x}{y} 2^{-x} \qquad\qquad (0 \le y \le 6)
\end{aligned}$$

where $\binom{x}{y} = 0$ if $x < y$. For example,

$$P(Y = 0) = \frac{1}{6}\left[\frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^6}\right] = \frac{1}{6} \times \frac{63}{64} = \frac{63}{384}$$

$$P(Y = 4) = \frac{1}{6}\left[\binom{4}{4}\frac{1}{2^4} + \binom{5}{4}\frac{1}{2^5} + \binom{6}{4}\frac{1}{2^6}\right] = \frac{29}{384}$$

and so on. Continuing in this way we obtain $P(Y = y)$ for each $y = 0, 1, 2, \ldots, 6$, as shown in Table 1.

TABLE 1. Probability $P(Y = y)$ of getting y heads.

y	0	1	2	3	4	5	6
$P(Y = y)$	$\frac{63}{384}$	$\frac{120}{384}$	$\frac{99}{384}$	$\frac{64}{384}$	$\frac{29}{384}$	$\frac{8}{384}$	$\frac{1}{384}$

Example 1 introduces the important idea of conditional distributions.

Conditional Distribution of *Y* Given *X* = *x*

For each possible value x of X, as y varies over all possible values of y, the probabilities $P(Y = y \,|\, X = x)$ form a probability distribution, depending on x, called the *conditional distribution of* Y *given* $X = x$.

The given value x of X can be thought of a as *parameter* in the distribution of Y given $X = x$. In Example 1, the distribution of Y given $X = x$ is the binomial distribution with parameters $n = x$ and $p = 1/2$.

According to the rule of average conditional probabilities, the unconditional distribution of Y, found in Example 1, is the *average* or *mixture* of these conditional

Example 1 introduces the important idea of conditional distributions.

Conditional Distribution of Y Given $X = x$

For each possible value x of X, as y varies over all possible values of y, the probabilities $P(Y = y \mid X = x)$ form a probability distribution, depending on x, called the *conditional distribution of Y given $X = x$*.

The given value x of X can be thought of as a *parameter* in the distribution of Y given $X = x$. In Example 1, the distribution of Y given $X = x$ is the binomial distribution with parameters $n = x$ and $p = 1/2$.

According to the rule of average conditional probabilities, the unconditional distribution of Y, found in Example 1, is the *average* or *mixture* of these conditional distributions, with equal weights $1/6$ defined by the uniform distribution of X. This distribution of Y may be called the *overall*, *marginal*, or *unconditional* distribution of Y, to distinguish it from the conditional distributions used to calculate it. The key step in the calculation of Example 1 was the following:

Rule of Average Conditional Probabilities

$$P(Y = y) = \sum_x P(Y = y \mid X = x) P(X = x)$$

This is just a basic rule of probability expressed in random variable notation. The rule holds for every pair of discrete random variables X and Y defined in the same probabilistic setting. The method of finding the distribution of a random variable Y by using this formula is called *conditioning on the value of X*. Note that in the sum for $P(Y = y)$ the term

$$P(Y = y \mid X = x) P(X = x) = P(X = x, Y = y)$$

is the generic entry in the joint probability table for X and Y. See Table 2 for example. You can use the above formula to calculate the distribution of a random variable Y if you can find a random variable X such that you either know or can easily calculate:

(i) the distribution of X;

(ii) the conditional probabilities $P(Y = y \mid X = x)$ for all possible values x of X.

If Y is determined by some two-stage or multistage process, the distribution of Y can often be calculated this way by letting X be the result of the first stage.

Example 1. **(Continued.)**

As in the previous example, let Y be the number of heads in X fair coin tosses, where X is uniformly distributed on $\{1, \ldots, 6\}$.

Problem 1. Find the conditional distribution of X given $Y = y$ for $y = 0, 1, \ldots 6$.

Solution. The problem now is to find $P(X = x \,|\, Y = y)$ as x varies, for each possible value y of Y. These conditional probabilities are calculated using Bayes' rule, as in Section 1.5. All that is new here is the random variable notation and terminology. As a start, the division rule for conditional probabilities gives

$$P(X = x \,|\, Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

where the joint probabilities

$$P(X = x, Y = y) = P(Y = y \,|\, X = x)P(X = x)$$

are the individual terms in the sum used previously to calculate $P(Y = y)$. Substituting the values of $P(X = x)$ and $P(Y = y \mid X = x)$, the joint probabilities $P(X = x, Y = y)$ are displayed in Table 2.

TABLE 2. Joint distribution table for (X, Y).

		Possible values x for X						Marginal
		1	2	3	4	5	6	distn. of Y
	0	$\frac{1}{6}\frac{1}{2}$	$\frac{1}{6}\frac{1}{4}$	$\frac{1}{6}\frac{1}{8}$	$\frac{1}{6}\frac{1}{16}$	$\frac{1}{6}\frac{1}{32}$	$\frac{1}{6}\frac{1}{64}$	$\frac{63}{384}$
	1	$\frac{1}{6}\frac{1}{2}$	$\frac{1}{6}\frac{2}{4}$	$\frac{1}{6}\frac{3}{8}$	$\frac{1}{6}\frac{4}{16}$	$\frac{1}{6}\frac{5}{32}$	$\frac{1}{6}\frac{6}{64}$	$\frac{120}{384}$
Possible	2	0	$\frac{1}{6}\frac{1}{4}$	$\frac{1}{6}\frac{3}{8}$	$\frac{1}{6}\frac{6}{16}$	$\frac{1}{6}\frac{10}{32}$	$\frac{1}{6}\frac{15}{64}$	$\frac{99}{384}$
values y	3	0	0	$\frac{1}{6}\frac{1}{8}$	$\frac{1}{6}\frac{4}{16}$	$\frac{1}{6}\frac{10}{32}$	$\frac{1}{6}\frac{20}{64}$	$\frac{64}{384}$
for Y	4	0	0	0	$\frac{1}{6}\frac{1}{16}$	$\frac{1}{6}\frac{5}{32}$	$\frac{1}{6}\frac{15}{64}$	$\frac{29}{384}$
	5	0	0	0	0	$\frac{1}{6}\frac{1}{32}$	$\frac{1}{6}\frac{6}{64}$	$\frac{8}{384}$
	6	0	0	0	0	0	$\frac{1}{6}\frac{1}{64}$	$\frac{1}{384}$
	Marginal distn. of X	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

In column x of the table you see numbers proportional to the binomial $(x, 1/2)$ probabilities forming the conditional distribution of Y given $X = x$. The constant of proportionality is $1/6$, which is the marginal probability of $(X = x)$. Similarly, in row y of the table you see numbers proportional to the conditional distribution of X given $Y = y$. The conditional probabilities themselves are obtained by dividing the numbers in the row y by the constant factor $P(Y = y)$, their sum, which appears in the margin. For example, the conditional distribution of X given $Y = 2$ is displayed in Table 3.

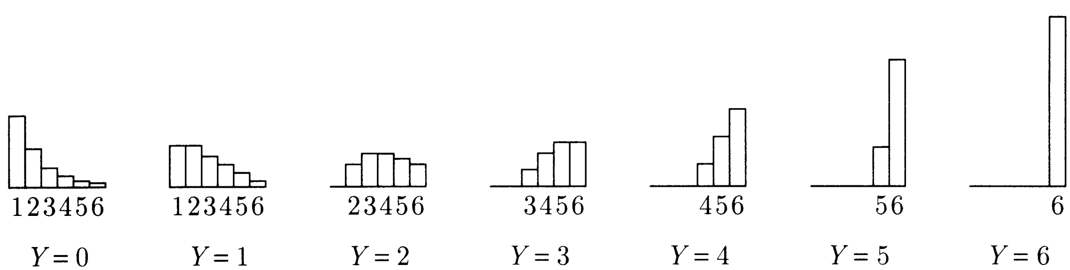
TABLE 3. Conditional distribution of X given $Y = 2$.

x	1	2	3	4	5	6
$P(X = x \mid Y = 2)$	0	$\frac{16}{99}$	$\frac{24}{99}$	$\frac{24}{99}$	$\frac{20}{99}$	$\frac{15}{99}$

So, given two heads, the number of coins tossed is equally likely to be either 3 or 4, and these are the most likely values.

Similar tables of the conditional distributions are easily made for other values y of Y. Here is a graphical display of all seven of these conditional distributions using histograms.

FIGURE 3. Conditional distribution of X given $Y = y$.



Exercises 6.1

1. Suppose I toss three coins. Some of them land heads and some land tails. Those that land tails I toss again. Let X be the number of heads showing after the first tossing, Y the total number showing after the second tossing, including the X heads appearing on the first tossing. So X and Y are random variables such that $0 \le X \le Y \le 3$ no matter how the coins land. Write out distribution tables and sketch histograms for each of the following distributions:

a) the distribution of X;

b) the conditional distribution of Y given $X = x$ for $x = 0, 1, 2, 3$;

c) the joint distribution of X and Y (no histogram in this case);

d) the distribution of Y;

e) the conditional distribution of X given $Y = y$ for $y = 0, 1, 2, 3$.

f) What is the best guess of the value of X given $Y = y$ for $y = 0, 1, 2, 3$? That is, for each y, choose x depending on y to maximize $P(X = x|Y = y)$.

g) Suppose the random experiment generating X and Y is repeated independently over and over again. Each time you observe the value of Y, and then guess the value of X using the rule found in f). Over the long run, what proportion of times will you guess correctly?

2. In a particular town 10% of the families have no children, 20% have one child, 40% have two children, 20% have three children, and 10% have four. Let T represent the total number of children, and G the number of girls, in a family chosen at random from this town. Assuming that children are equally likely to be boys or girls, find the distribution of G. Display your answer in a table and sketch the histogram.

3. Suppose the names of all the children in the town of Exercise 2 are put into a hat, and a name is picked out at random. So now a child is picked at random instead of a family being picked at random. Let U be the total number of children in the family of the child chosen at random.

a) Find the distribution of U. Why is this distribution different from the distribution of T in Exercise 2?

b) What is the probability that the child picked at random comes from a family consisting of two girls and a boy?

c) Is this the same as the probability that a family picked at random consists of two girls and a boy? Calculate and explain.

4. Let $A_1, \ldots, A_{20}$ be independent events each with probability $1/2$. Let X be the number of events among the first 10 which occur and let Y be the number of events among the last 10 which occur. Find the conditional probability that $X = 5$, given that $X + Y = 12$.

5. Let X_1 and X_2 be independent Poisson random variables with parameters λ_1 and λ_2.

a) Show that for every $n \geq 1$, the conditional distribution of X_1, given $X_1 + X_2 = n$, is binomial, and find the parameters of this binomial distribution.

b) The number of eggs laid by a certain kind of insect follows a Poisson distribution quite closely. It is known that two such insects have laid their eggs in a particular area. If the total number of eggs in the area is 150, what is the chance that the first insect laid at least 90 eggs? (State your assumptions, and give approximate decimal answer.)

6. Conditioning independent Poisson variables on their sum. Let N_i be independent Poisson variables with parameters λ_i. Think of the N_i as the number of points of a Poissson scatter in disjoint parts of the plane with areas λ_i, where the mean intensity is one point per unit area.

a) What is the conditional joint distribution of $(N_1, \ldots, N_m)$ given $N_1 + \ldots + N_m = n$? [*Hint*: See Exercise 5 for a special case.]

b) Suppose now that N has Poisson(λ) distribution, and given $N = n$ the conditional joint distribution of some m-tuple of random variables $(N_1, \ldots, N_m)$ is exactly what you found in part a). What can you conclude about the unconditional distribution of $(N_1, \ldots, N_m)$?

7. Poissonization of the binomial distribution. Let N have Poisson (λ) distribution. Let X be a random variable with the following property: for every n, the conditional distribution of X given $(N = n)$ is binomial (n, p).

a) Show that the unconditional distribution of X is Poisson, and find its parameter.

It is known that X-rays produce chromosome breakages in cells. The number of such breakages usually follows a Poisson distribution quite closely, where the parameter depends on the time of exposure, etc. For a particular dosage and time of exposure, the number of breakages follows the Poisson (0.4) distribution. Assume that each breakage heals with probability 0.2, independently of the others.

b) Find the chance that after such an X-ray, there are 4 healed breakages.

8. Independence in Poissonization of the binomial distribution. Suppose you roll a random number of dice. If the number of dice follows the Poisson (λ) distribution, show that the number of sixes is independent of the number of nonsixes. [*Hint*: Let N be the number of dice, X the number of sixes, and Y the number of nonsixes. Exercise 7 gives you the marginal distributions of X and Y. To show that the joint distribution of X and Y is the product of the marginals, show

$$P(X = x, Y = y) = P(N = x + y, X = x, Y = y)$$

and then use the multiplication rule.]

9. Conditional independence. Random variables X and Y are called *conditionally independent given* Z if given the value of Z, X, and Y are independent. That is,

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z)$$

for all possible values x, y, and z. Prove that X and Y are conditionally independent given Z if and only if the conditional distribution of Y given $X = x$ and $Z = z$ is a distribution which depends only on z:

$$P(Y = y \mid X = x, Z = z) = P(Y = y \mid Z = z)$$

for all possible values x, y, and z. Give a further equivalent condition in terms of the conditional distribution of X given $Y = y$ and $Z = z$.

10. Conditional independence (continued). Suppose as in Example 5 of Section 3.1 that two sequences of n draws with replacement are made from a box containing an unknown number of red tickets among a total of 10 tickets. Regard the number of red tickets in the box as the value of a random variable R, with probability distribution $P(R = r) = \pi_r$, $r = 0, 1, \ldots, 10$. Let X_1 be the number of red tickets in the first n draws, and X_2 the number in the second n draws. Assuming that X_1 and X_2 are conditionally independent and binomially distributed given $R = r$, find expressions for the following:

a) $P(R = r, X_1 = x_1, X_2 = x_2)$; b) $P(R = r \mid X_1 = x_1)$;

c) $P(X_2 = x_2 \mid R = r, X_1 = x_1)$; d) $P(X_2 = x_2 \mid X_1 = x_1)$.

e) Calculate numerical values for the conditional probabilities in d) assuming that $\pi_r = 1/11$ for $r = 0, 1, \ldots, 10$ and $n = 1$. Are X_1 and X_2 independent?

6.2 Conditional Expectation: Discrete Case

Conditional expectations are simply expectations relative to conditional distributions.

Conditional Expectation Given an Event

The *conditional expectation of a random variable Y given an event A*, denoted by $E(Y \mid A)$, is the expectation of Y under the conditional probability distribution given A:

$$E(Y \mid A) = \sum_{\text{all } y} y P(Y = y \mid A)$$

This is just the definition of $E(Y)$, with probabilities replaced by conditional probabilities given A. Intuitively, $E(Y \mid A)$ is the expected value of Y, given the information that event A has occurred.

Example 1. Conditioning on at most 2 heads on 4 coin tosses.

Let Y be the number of heads in four tosses of a fair coin. Calculate the conditional expectation of Y given 2 or less heads. What is the long-run interpretation of this quantity?

Solution. Here the conditioning event is $A = (Y \leq 2)$. Since Y has the binomial $(4, \frac{1}{2})$ distribution

$$P(Y = y) = \binom{4}{y} \Big/ 2^4 \qquad (y = 0 \text{ to } 4)$$

$$P(Y \leq 2) = (1 + 4 + 6)/16 = 11/16$$

Hence

$$P(Y = y \mid Y \leq 2) = \binom{4}{y} \Big/ 11 \qquad (y = 0, 1, 2)$$

and

$$E(Y \mid Y \leq 2) = \sum_{y=0}^{2} y \binom{4}{y} \Big/ 11 = (1 \cdot 4 + 2 \cdot 6)/11 = 16/11$$

The long-run interpretation is that if you repeatedly toss four fair coins, the long-run average number of heads, averaging only over the trials that produce 0, 1, or 2 heads, will be $16/11$.

Properties of Conditional Expectation

For a fixed conditioning event A, conditional expectation has familiar properties of expectation like linearity. For instance, there is the addition rule

$$E(X + Y \mid A) = E(X \mid A) + E(Y \mid A)$$

and so on. For a fixed random variable Y, as A varies, there is a useful generalization of the rule of average conditional probabilities, a rule of *average conditional expectations*: If $A_1, \ldots, A_n$ is a partition of the whole outcome space, then

$$E(Y) = \sum_{i=1}^{n} E(Y \mid A_i)P(A_i)$$

In the special case when Y is an indicator random variable, say $Y = I_B$, the indicator of event B, this reduces to the rule of average conditional probabilities

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i)P(A_i)$$

The general case can be derived from this special case by linear operations. It is most convenient for applications to express the general rule as follows, for the partition generated by values of a discrete random variable X:

Rule of Average Conditional Expectations

For any random variable Y with finite expectation and any discrete random variable X,

$$E(Y) = \sum_{\text{all } x} E(Y \mid X = x)P(X = x)$$

This formula is also called the *formula for $E(Y)$ by conditioning on X*. This formula gives a useful method of calculating expectations, as shown by the examples below. The next box introduces a useful short notation:

Definition of $E(Y \mid X)$

The *conditional expectation of Y given X*, denoted $E(Y \mid X)$, is the function of X whose value is $E(Y \mid X = x)$ if $X = x$.

Here $E(Y \mid X)$ is actually a random variable, since by definition it is a particular function of X, and a function of a random variable defines another random variable. It can be shown that $E(Y \mid X)$ is the best predictor of Y based on X, in the sense of mean-square error. That is to say, $E(Y \mid X)$ is the function $g(X)$ that minimizes the mean square prediction error $E[(Y - g(X))^2]$. See Exercise 17. Because $E(Y \mid X)$ is a random variable, it makes sense to consider its expectation. The result is stated in the next box.

> **Expectation is the Expectation of the Conditional Expectation**
>
> $$E(Y) = E\,[E(Y \mid X)]$$

This is a condensed form of the rule of average conditional expectations, obtained by application to $g(x) = E(Y \mid X = x)$ of the formula

$$E[g(X)] = \sum_{\text{all } x} g(x)P(X = x)$$

Examples

Example 2. Tossing a random number of coins.

As in Example 1 of the previous section, let Y be the number of heads in X tosses of a fair coin, where X is generated by a fair die roll.

Problem 1. Find the conditional expectation of Y given $X = x$.

Solution. Since the conditional distribution of Y given $X = x$ is binomial with parameters $n = x$ and $p = 1/2$, the conditional expectation of Y given $X = x$ is the mean of the binomial(n, p) distribution, that is np, for $n = x$ and $p = 1/2$:

$$E(Y \mid X = x) = x/2 \qquad (x = 1, 2, \ldots, 6)$$

Problem 2. Find $E(Y)$.

Solution. Since from the previous solution $E(Y \mid X) = X/2$, and $E(X) = 3.5$

$$E(Y) = E[E(Y \mid X)] = E(X/2) = E(X)/2 = (3.5)/2 = 1.75$$

Discussion. Of course, the expectation of Y can also be calculated from the distribution of Y, shown in Table 1 of Section 6.1. But the method of conditioning on X gives the result more quickly. Also, the method of computing $E(Y)$ by conditioning on a suitable random variable X can be applied in problems where it is difficult to obtain a formula for the distribution of Y.

Problem 3. Find $E(X \mid Y = 2)$

Solution. There is no simple formula for $E(X \mid Y = y)$ as a function of y in this problem. But these conditional expectations can be calculated one by one from the various conditional distributions of X given $Y = y$ for $y = 0$ to 6. Using the conditional distribution of X given $Y = 2$ displayed in Table 3 of Section 6.1 gives

$$E(X \mid Y = 2) = (2\times16 + 3\times24 + 4\times24 + 5\times20 + 6\times15)/99 = 390/99 \approx 3.94$$

Example 3. Number of girls in a family.

Suppose the number of children in a family is a random variable X with mean μ, and given $X = n$ for $n \geq 1$, each of the n children in the family is a girl with probability p and a boy with probability $1 - p$.

Problem. What is the expected number of girls in a family?

Solution. Intuitively, the answer should be $p\mu$. To show this is correct, let G be the random number of girls in a family. Given $X = n$, G is the sum of n indicators of events with probability p, so

$$E(G \mid X = n) = np$$

Note that this is correct even for $n = 0$. By conditioning on X,

$$E(G) = \sum_n E(G \mid X = n)P(X = n) = p \sum_n nP(X = n) = p\mu$$

Remark. In short notation,

$$E(G \mid X) = pX$$

$$E(G) = E\left[E(G \mid X)\right] = E(pX) = pE(X)$$

Example 4. Success counts in overlapping series of trials.

Let S_n be the number of successes in n independent trials with probability p of success on each trial.

Problem. Calculate $E(S_m \mid S_n = k)$ for $m \leq n$.

Solution. Since $S_m = X_1 + \cdots + X_m$ where X_j is the indicator of success on the jth trial

$$E(S_m \,|\, S_n = k) = \sum_{j=1}^{m} E(X_j \,|\, S_n = k) \quad \text{where}$$

$$\begin{aligned}
E(X_j \,|\, S_n = k) &= P(j\text{th trial is a success} \,|\, S_n = k) \\
&= \frac{P(j\text{th trial is a success},\ S_n = k)}{P(S_n = k)} \\
&= \frac{P(j\text{th trial success},\ k-1 \text{ of other } n-1 \text{ trials are successes})}{P(S_n = k)} \\
&= \frac{p\binom{n-1}{k-1}p^{k-1}(1-p)^{n-k}}{\binom{n}{k}p^k(1-p)^{n-k}} \qquad \text{using independence and the binomial distribution} \\
&= \frac{k}{n} \quad \text{so}
\end{aligned}$$

$$E(S_m \,|\, S_n = k) = \frac{mk}{n}$$

Discussion. In short notation, the conclusion is that for $1 \le m \le n$

$$E(S_m \,|\, S_n) = \frac{m}{n} S_n$$

This is a rather intuitive formula. It says that given S_n successes in n trials, the number of successes to be expected in m of the trials is proportional to m. The formula can be derived in other ways. By symmetry, $E(X_j \,|\, S_n)$ must be the same for all j, and equal to $E(X_1 \,|\, S_n)$. Since

$$S_n = E(S_n \,|\, S_n) = \sum_{j=1}^{n} E(X_j \,|\, S_n) = nE(X_1 \,|\, S_n)$$

it follows that $E(X_1 \,|\, S_n) = S_n/n$ and hence

$$E(S_m \,|\, S_n) = \sum_{j=1}^{m} E(X_j \,|\, S_n) = mE(X_1 \,|\, S_n) = \frac{m}{n} S_n$$

This argument shows that formula

$$E(S_m \,|\, S_n) = \frac{m}{n} S_n \qquad (1 \le m \le n)$$

holds whenever S_n is a sum of n independent and identically distributed variables $X_1, \ldots, X_n$. In fact all that is required is that the variables $X_1, \ldots, X_n$ are *exchangeable*, as defined in Section 3.6. This is an example where a conditional expectation can be calculated using symmetry and linearity, even though there is no nice formula for the conditional distribution.

Treating a conditioned variable as a constant. When computing conditional probabilities or expectations given $X = x$, the random variable X may be treated as if it were the constant x. Intuitively, this is quite obvious: on the restricted outcome space $(X = x)$, the random variable X has only one value, namely, x. To illustrate, if g is a function of two random variables X and Y, the conditional distribution of $g(X, Y)$ given $X = x$, is the same as the conditional distribution of $g(x, Y)$ given $X = x$. And if g has numerical values

$$E\left[g(X, Y) \,|\, X = x\right] = E\left[g(x, Y) \,|\, X = x\right]$$

For instance

$$E[XY \,|\, X = x] = E[xY \,|\, X = x] = xE[Y \,|\, X = x]$$

which reads in short notation

$$E[XY \,|\, X] = XE[Y \,|\, X]$$

Another example is

$$E[aX + bY \,|\, X = x] = E[ax + bY \,|\, X = x] = ax + bE[Y \,|\, X = x]$$

which reads in short notation

$$E[aX + bY \,|\, X] = aX + bE[Y \,|\, X]$$

Example 5. Conditional expectation of a sum given one of the terms.
Suppose X and Y are independent.

Problem. Find $E(X + Y \,|\, X = x)$.

Solution.

$$\begin{aligned} E(X + Y \,|\, X = x) &= E(X \,|\, X = x) + E(Y \,|\, X = x) \\ &= x + E(Y) \end{aligned}$$

Here $E(X \,|\, X = x) = x$ because X may be treated as the constant x given $X = x$. And $E(Y \,|\, X = x)$ is the mean of the conditional distribution of Y given $X = x$, and by independence this is just the unconditional distribution of Y with mean $E(Y)$.

Exercises 6.2

1. Let X_1 and X_2 be the numbers on two independent fair-die rolls. Let X be the minimum and Y the maximum of X_1 and X_2. Calculate: a) $E(Y|X = x)$; b) $E(X|Y = y)$.

2. Repeat Exercise 1 above, with X_1 and X_2 independent and uniformly distributed on $\{1, 2, \dots, n\}$.

3. Repeat Exercise 1 with X_1 and X_2 two draws without replacement from $\{1, 2, \dots, n\}$.

4. An item is selected randomly from a collection labeled $1, 2, ..., n$. Denote its label by X. Now select an integer Y uniformly at random from $\{1, \dots, X\}$. Find:

a) $E(Y)$; b) $E(Y^2)$; c) $SD(Y)$; d) $P(X + Y = 2)$.

5. Suppose an event A is independent of a pair of random variables X_1 and X_2, whose c.d.f's are F_1 and F_2. Define a random variable X by:

$$X = \begin{cases} X_1 & \text{if } A \text{ occurs} \\ X_2 & \text{if } A \text{ does not occur} \end{cases}$$

Find and justify formulae for:

a) the c.d.f. $F(x)$ of X, in terms of $F_1(x)$, $F_2(x)$, and $p = P(A)$;

b) $E(X)$ in terms of $E(X_1)$, $E(X_2)$, and p.

c) $Var(X)$ in terms of $E(X_1)$, $E(X_2)$, $Var(X_1)$, $Var(X_2)$ and p.

6. Suppose that N is a Poisson random variable with parameter μ. Suppose that given $N = n$, random variables $X_1, X_2, ..., X_n$ are independent with uniform $(0, 1)$ distribution. So there are a random number of X's.

a) Given $N = n$, what is the probability that all the X's are less than t?

b) What is the (unconditional) probability that all the X's are less than t?

c) Let $S_N = X_1 + \cdots + X_N$ denote the sum of the random number of X's. (If $N = 0$ then $S_N = 0$.) Find $P(S_N = 0)$. Explain.

d) Find $E(S_N)$.

7. Suppose that N is a counting random variable, with values $\{0, 1, \dots, n\}$, and that given $(N = k)$, for $k \geq 1$, there are defined random variables $X_1, \dots, X_k$ such that

$$E(X_j | N = k) = \mu \qquad (1 \leq j \leq k)$$

Define a random variable S_N by

$$S_N = \begin{cases} X_1 + X_2 + \cdots + X_k & \text{if } (N = k),\ 1 \leq k \leq n \\ 0 & \text{if } (N = 0) \end{cases}$$

Show that $E(S_N) = \mu E(N)$.

8. Suppose that each individual in a population produces a random number of children, and the distribution of the number of children has mean μ. Starting with one individual, show, using the result of Exercise 7, that the expected number of descendants of that individual in the nth generation is μ^n.

9. Let T_i be the place at which the ith good element appears in a random ordering of $N - k$ bad elements and k good ones. Use the results of Exercise 3.6.13 to calculate:

a) $E(T_1 | T_2 = j)$; b) $E(T_2 | T_1 = j)$;

c) $E(T_h|T_i = j)$ first for $h < i$, then for $h > i$.

10. What is the expected number of black balls among $n \leq b+w+d$ balls drawn at random from a box containing b black balls, w white balls, and d balls drawn at random from another box of b_0 black balls and w_0 white balls? Assume all draws are made without replacement.

11. A deck of cards is cut into two halves of 26 cards each. As it turns out, the top half contains 3 aces and the bottom half just one ace. The top half is shuffled, then cut into two halves of 13 cards each. One of these packs of 13 cards is shuffled into the bottom half of 26 cards, and from this pack of 39 cards, 5 cards are dealt. What is the expected number of aces among these 5 cards?

12. **Conditional expectations in Polya's urn scheme.** An urn contains 1 black and 2 white balls. One ball is drawn at random and its color noted. The ball is replaced in the urn, together with an additional ball of its color. There are now four balls in the urn. Again, one ball is drawn at random from the urn, then replaced along with an additional ball of its color. The process continues in this way.

a) Let B_n be the number of black balls in the urn just before the nth ball is drawn. (Thus B_1 is 1.) For $n \geq 1$, find $E(B_{n+1}|B_n)$.

b) For $n \geq 1$, find $E(B_n)$. [*Hint*: $E(B_1) = 1$; now use part a) and induction on n.]

c) For $n \geq 1$, what is the expected proportion of black balls in the urn just before the nth ball is drawn?

13. **Conditioning on the number of successes in Bernoulli trials.** Let $S_n = X_1 + \cdots + X_n$ be the number of successes in of n independent Bernoulli(p) trials $X_1, X_2, \ldots, X_n$.

a) For $1 \leq m \leq n$, show that the conditional distribution of S_m, the number of successes in the first m trials, given $S_n = k$, is identical to the distribution of the number of good elements in a random sample of size m without replacement from a population of k good and $n - k$ bad elements.

b) Use the result of a) to rederive the result of Example 4 that $E(S_m \,|\, S_n = k) = mk/n$.

c) Find $Var(S_m \,|\, S_n = k)$.

14. **Sufficiency of the number of successes in Bernoulli trials.** Let $S_n = X_1 + \cdots + X_n$ be the number of successes in n independent Bernoulli (p) trials $X_1, X_2, \ldots, X_n$. As a continuation of Exercise 13, show that conditionally given $S_n = k$, the sequence of zeros and ones $X_1, \ldots, X_n$ is distributed like an exhaustive sample without replacement from a population of k ones and $n - k$ zeros. [Note that this conditional distribution does not depend on p. In the language of statistics, when p is an unknown parameter S_n is called a *sufficient statistic* for p. If you want to estimate an unknown p given observed values of $X_1, \ldots, X_n$, and are committed to the assumption of Bernoulli (p) trials, it makes no sense to use any aspect of the data besides S_n in the estimation problem, because given $S_n = k$, the parameter p does not affect the distribution of the data at all. One natural estimate of p given the data is S_n/n, the observed proportion of successes. But other functions of S_n may be considered. See Exercise 6.3.15.]

15. Let Π be a random proportion between 0 and 1, for example, the proportion of black balls in an urn picked at random from some population of urns. Let S be the number of successes in n Bernoulli trials, which given $\Pi = p$ are independent with probability p, for example, the number of black balls in n draws at random, with replacement from the urn picked at random.

a) Find a formula for $E(S)$ in terms of n and $E(\Pi)$.

b) Find a formula for $Var(S)$ in terms of n, $E(\Pi)$, and $Var(\Pi)$.

c) For given n and $E(\Pi) = p$, say, which distribution of Π makes $Var(S)$ as large as possible? Which as small as possible? Prove your answers using your answer to b).

16. Expectation of a product by conditioning. Let X and Y be random variables, and let h be a function of X. Show that

$$E[h(X)Y] = E[h(X)E(Y|X)]$$

[*Hint*: Look at $E(h(X)Y|X = x)$.] *Remark:* This identity, for indicator functions $h(x)$, is used in more advanced treatments of probability to define conditional expectations given a continuous random variable X.

17. Prediction by functions. Suppose you want to predict the value of a random variable Y. Instead of just trying to predict the value of Y by a constant, as was done in Section 3.2, suppose that some additional information pertinent to the prediction of Y is available. For instance, you might know the value of some other random variable X, whose joint distribution with Y is assumed known. The problem here is to predict the value of Y by a function of X, call it $g(X)$. Once the value x of X is known, the value $g(x)$ of $g(X)$ can be calculated and used to predict the unknown value of Y.
One measure of the goodness of the predictor $g(X)$ is its *mean square error* (MSE)

$$MSE(g(X)) = E[(Y - g(X))^2]$$

It is a measure of, on average, how far off the prediction is. Show that $g(X) = E(Y|X)$ minimizes the MSE. [*Hint*: Condition on the value of X

$$E[(Y - g(X))^2] = \sum_x E[(Y - g(X))^2|X = x]P(X = x)$$

and minimize each term in the sum separately.]

18. Conditional variance. Define $Var(Y|X)$, the *conditional variance of Y given X*, to be the random variable whose value, if $(X = x)$, is the variance of the conditional distribution of Y given $X = x$. So $Var(Y|X)$ is a function of X, namely $h(X)$, where $h(x) = E(Y^2|X = x) - [E(Y|X = x)]^2$. Show that

$$Var(Y) = E[Var(Y|X)] + Var[E(Y|X)]$$

In words, the variance is the expectation of the conditional variance plus the variance of the conditional expectation.

6.3 Conditioning: Density Case

This section treats conditional probabilities given the value of a random variable X with a continuous distribution. In the discrete case, the conditional probability of an event A, given that X has value x, is defined by

$$P(A\,|\,X = x) = \frac{P(A, X = x)}{P(X = x)}$$

whenever $P(X = x) > 0$. In the continuous case $P(X = x) = 0$ for every x, so the above formula gives the undefined expression $0/0$. This must be replaced, as in the usual calculus definition of a derivative dy/dx, by the following:

Infinitesimal Conditioning Formula

$$P(A\,|\,X = x) = \frac{P(A, X \in dx)}{P(X \in dx)}$$

Intuitively, $P(A \mid X = x)$ should be understood as $P(A \mid X \in dx)$, the chance of A given that X falls in a very small interval near x. It is assumed here that in the limit of small intervals this chance does not depend on what interval is chosen near x. So, like a derivative dy/dx, $P(A\,|\,X \in dx)$ is a function of x, hence the notation $P(A\,|\,X = x)$. In terms of limits,

$$P(A\,|\,X = x) = \lim_{\Delta x \to 0} P(A\,|\,X \in \Delta x) = \lim_{\Delta x \to 0} \frac{P(A, X \in \Delta x)}{P(X \in \Delta x)}$$

where Δx stands for an interval of length Δx containing the point x. It is assumed here that the limit exists, except perhaps for a finite number of exceptional points x such as endpoints of an interval defining the range of X, or places where the density of X has a discontinuity. See the book *Probability and Measure* by P. Billingsley for a rigorous treatment of conditioning on a continuously distributed variable.

Most often, the event A of interest is determined by some random variable Y, for instance, $A = (Y > 3)$. If (X, Y) has a joint density $f(x, y)$, then $P(A\,|\,X = x)$ can be found by integration of the conditional density of Y given $X = x$, defined as follows:

Conditional Density of Y given $X = x$

For random variables X and Y with joint density $f(x,y)$, for each x such that the marginal density $f_X(x) > 0$, the *conditional density of* Y *given* $X = x$ is the probability density function with dummy variable y defined by

$$f_Y(y \,|\, X = x) = f(x,y)/f_X(x)$$

Intuitively, the formula for $f_Y(y \,|\, X = x)$ is justified by the following calculation of the chance of $(Y \in dy)$ given $X = x$:

$$\begin{aligned} P(Y \in dy \,|\, X = x) &= P(Y \in dy \,|\, X \in dx) \\ &= \frac{P(X \in dx, Y \in dy)}{P(X \in dx)} \\ &= \frac{f(x,y)\,dx\,dy}{f_X(x)\,dx} \\ &= f_Y(y \,|\, X = x)dy \end{aligned}$$

The formula $\int f(x,y)dy = f_X(x)$, the marginal density of X, implies that

$$\int f_Y(y \,|\, X = x)dy = 1$$

So for each fixed x with $f_X(x) > 0$, the formula for $f_Y(y \,|\, X = x)$ gives a probability density in y. This conditional density given x defines a probability distribution parameterized by x, called the *conditional distribution of* Y *given* $X = x$. In examples, this will often be a familiar distribution, for example, a uniform or a normal distribution, with parameters depending on x.

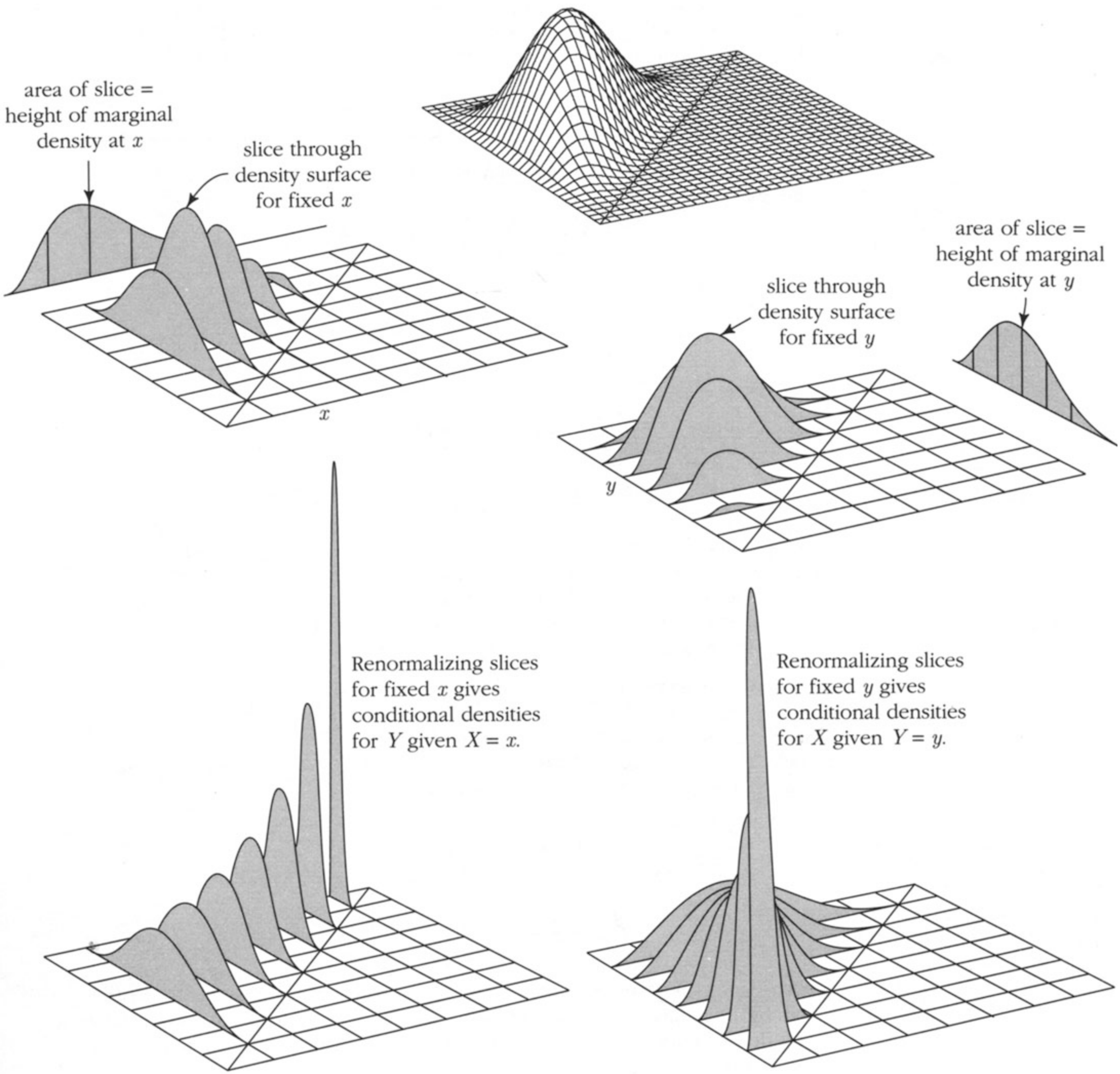
The conditional density of Y given $X = x$ can be understood geometrically by taking a vertical slice through the joint density surface at x, and renormalizing the resulting function of y by its total integral, which is $f_X(x)$. Conditional probabilities given $X = x$ of events determined by X and Y can be calculated by integrating with respect to this conditional density. For example

$$P(Y > b \,|\, X = x) = \int_b^\infty f_Y(y \,|\, X = x)dy$$

$$P(Y > 2X \,|\, X = x) = \int_{2x}^\infty f_Y(y \,|\, X = x)dy$$

Such expressions are obtained formally from their discrete analogs by replacing a sum by an integral, and replacing the probability of an individual point by the value of a density times an infinitesimal length. See the display at the end of this section for details of this analogy.

FIGURE 1. Joint, marginal, and conditional densities.



Key to Figure 1

Top: Joint density surface. This is a perspective projection of the surface

$$z = f(x, y)$$

defined by a particular joint density function $f(x, y)$.

Middle left: Slices for some values of X and the marginal density of X. Here are seven slices, or cross sections through the density surface for given values X ranging from $1/8$ to $7/8$. (The last two are so low that they are invisible.) The probability that X falls in a short interval of length Δ near x is the volume of such a slice of thickness Δ, which for small enough Δ is essentially Δ times the area of the slice at x. This area equals

$$\int f(x, y)dy = f_X(x)$$

the height of the *marginal density of X at x*, graphed at back. This marginal density shows how probability is distributed between slices according to the distribution of X. The heights of the vertical segments shown in the graph of the marginal density are proportional to the areas of corresponding slices.

Middle right: Slices for some values Y and the marginal density of Y. Here are perpendicular slices through the density surface for given values of Y. The area of the slice at y equals

$$\int f(x, y)\, dx = f_Y(y),$$

the height of the *marginal density of Y at y*, shown at right.

Bottom left: Conditional density of Y for some given values of X. Rescaling each section of the diagram above by its total area, the marginal density of X at x, gives the *conditional density of Y given $X = x$*, shown here using the same vertical scale as for the marginal densities in the middle diagrams. Given $X = x$, Y is distributed with density proportional to the section of the density surface $f(x, y)$ through x. Dividing by the total area of the section through x gives the conditional density of Y given $X = x$. Note how the shape of the two invisible sections in the middle left diagram can now be seen, due to the normalization of each section by its total area. The marginal density of Y (see middle right) is the average of all the conditional densities of Y given $X = x$ weighted according to the marginal distribution of X (middle left).

Bottom right: Conditional density of X for some given values of Y. These are interpreted just as above, with the roles of X and Y switched.

Example 1. Uniform on a triangle.

Problem. Suppose that a point (X, Y) is chosen uniformly at random from the triangle $\{(x,y) : x \geq 0, y \geq 0, x + y \leq 2\}$. Find $P(Y > 1 \mid X = x)$.

To illustrate the basic concepts, three slightly different solutions will be presented.

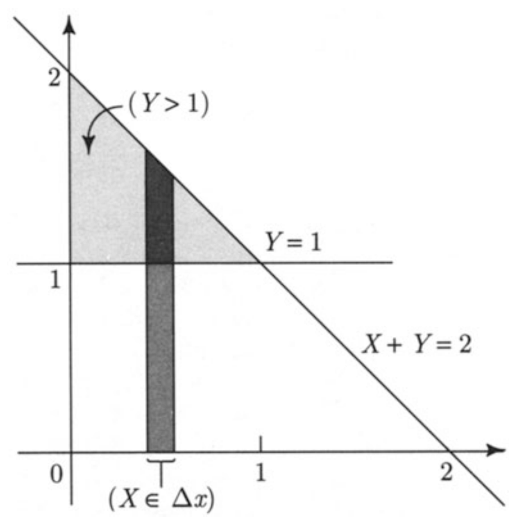
Solution 1. *Informal approach.* Intuitively, it seems obvious that given $X = x$, the random point (X, Y) should be regarded as uniformly distributed on the vertical line segment $\{(x,y) : y \geq 0, x + y \leq 2\}$ with length $2 - x$. This is the conditional distribution of (X, Y) given $X = x$. If x is between 0 and 1, the portion of this segment above $y = 1$ has length $(2 - x) - 1 = 1 - x$. Otherwise, no portion of the segment is above $y = 1$. So the answer is

$$P(Y > 1 \mid X = x) = \begin{cases} (1 - x)/(2 - x) & 0 \leq x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Solution 2. *Definition of conditional probability.* To see that Solution 1 agrees with the formal definition

$$P(Y > 1 \mid X = x) = \lim_{\Delta x \to 0} P(Y > 1 \mid X \in \Delta x)$$

look at the following diagram which shows the events $(Y > 1)$ and $(X \in \Delta x) = (x \leq X \leq x + \Delta x)$:



Since the triangle has area 2, the probability of an event is half its area. So, for $0 \leq x < 1, x + \Delta x \leq 1$, there are the exact formulae

$$P(X \in \Delta x) = \frac{1}{2}\Delta x(2 - x - \frac{1}{2}\Delta x)$$

$$P(Y > 1, X \in \Delta x) = \frac{1}{2}\Delta x(1 - x - \frac{1}{2}\Delta x)$$

Therefore, for $0 \le x < 1$,

$$\begin{aligned} P(Y > 1 \mid X \in \Delta x) &= \frac{P(Y > 1, X \in \Delta x)}{P(X \in \Delta x)} \\ &= \frac{1 - x - \frac{1}{2}\Delta x}{2 - x - \frac{1}{2}\Delta x} \\ &\to \frac{1-x}{2-x} \quad \text{as} \quad \Delta x \to 0 \end{aligned}$$

This verifies the formula of Solution 1 for $0 \le x < 1$. The formula for $x \ge 1$ is obvious because the event $(Y > 1, X \in \Delta x)$ is empty if $x \ge 1$.

Solution 3. *Calculation with densities.* Let us recalculate $P(Y > 1 \mid X = x)$ using the conditional density $f_Y(y \mid X = x)$. The uniform distribution on the triangle makes the joint density

$$f(x, y) = \begin{cases} 1/2 & x \ge 0, y \ge 0, x + y \le 2 \\ 0 & \text{otherwise} \end{cases}$$

So for $0 \le x \le 2$,

$$f_X(x) = \int_0^\infty f(x, y)dy = \int_0^{2-x} \frac{1}{2}dy = \frac{1}{2}(2 - x)$$

and

$$f_Y(y \mid X = x) = \begin{cases} \dfrac{f(x,y)}{f_X(x)} = \dfrac{1}{2-x} & 0 \le y \le 2 - x \\ 0 & \text{otherwise} \end{cases}$$

That is, given $X = x$ for $0 \le x \le 2$, Y has uniform $(0, 2 - x)$ distribution, as is to be expected intuitively. So

$$P(Y > 1 \mid X = x) = \begin{cases} \displaystyle\int_1^{2-x} \frac{dy}{2-x} = \frac{1-x}{2-x} & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

as before.

Discussion. The point of the first solution is that conditional distributions are often intuitively obvious, and once identified they can be used to find conditional probabilities very quickly. The second solution shows how this kind of calculation is justified by the formal definition. This method is not recommended for routine calculations. The third solution is essentially a more detailed version of the first. While rather pedantic in the present problem, this kind of calculation is essential in more difficult problems where you cannot guess the answer by intuitive reasoning.

Rules for conditional densities. These are analogs of corresponding rules in the discrete case. Note that every concept defined by the distribution of a real-valued random variable Y, in particular, the notions of density function, distribution function, expectation, variance, moments, and so on, can be considered for conditional distributions, just as well as for unconditional ones. There is just an extra parameter, x, the given value of X.

When the density of X is known, and a conditional density for Y given $X = x$ is specified for each x in the range of X, the joint density of X and Y is calculated by the following rearrangement of the formula $f_Y(y \mid X = x) = f(x, y)/f_X(x)$.

Multiplication Rule for Densities

$$f(x, y) = f_X(x) f_Y(y \mid X = x)$$

Example 2. **Gamma and uniform.**

Suppose X has gamma $(2, \lambda)$ distribution, and that given $X = x$, Y has uniform $(0, x)$ distribution.

Problem 1. Find the joint density of X and Y.

Solution. By the definition of the gamma distribution

$$f_X(x) = \begin{cases} \lambda^2 x e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

and from the uniform $(0, x)$ distribution of Y given $X = x$

$$f_Y(y \mid X = x) = \begin{cases} 1/x & 0 < y < x \\ 0 & \text{otherwise} \end{cases}$$

So by the multiplication rule for densities

$$f(x, y) = f_X(x) f_Y(y \mid X = x) = \begin{cases} \lambda^2 e^{-\lambda x} & 0 < y < x \\ 0 & \text{otherwise} \end{cases}$$

Problem 2. Find the marginal density of Y.

Solution. Integrating out x in the joint density gives the marginal density of Y: for $y > 0$

$$f_Y(y) = \int_0^\infty f(x, y)\, dx = \int_y^\infty \lambda^2 e^{-\lambda x}\, dx = \lambda e^{-\lambda y}$$

The density is of course 0 for $y \leq 0$. That is to say, Y has exponential (λ) distribution.

Problem 3. Show that X and Y have the same joint distribution as T_2 and T_1, where T_1 is the first arrival time and T_2 is the second arrival time in a Poisson arrival process with rate λ.

Solution. That X has the same distribution as T_2, and that Y has the same distribution as T_1, follows from the above calculation and the result of Section 4.2 that the ith arrival time in a Poisson process with rate λ has gamma(i, λ) distribution. That the *joint* distribution of X and Y is the same as the joint distribution of T_2 and T_1 requires a little more calculation, because a joint distribution is not determined by its marginals. The simplest way to verify this is to observe that for $0 < y < x$

$$P(T_1 \in dy, T_2 \in dx)$$

is the probability of no arrivals in the time interval $[0, y]$ of length y, one arrival in time dy, no arrivals in the time interval $[y + dy, x]$ of length $x - y - dy \approx x - y$, and finally one arrival in dx. By independence and Poisson distribution of counts in disjoint intervals, and neglecting a term of order $(dy)^2$, this event has probability

$$e^{-\lambda y}\,\lambda\,dy\,e^{-\lambda(x-y)}\,\lambda\,dx = \lambda^2 e^{-\lambda x} dy\,dx$$

Dividing the last expression by $dy\,dx$ shows that the joint density of (T_2, T_1) at (x, y) with $0 < y < x$ is identical to the joint density found in Problem 1. Since obviously $P(T_1 < T_2) = 1$, the joint density of (T_2, T_1) can be taken to be zero except if $0 < y < x$. Thus (T_2, T_1) has the same joint density function as (X, Y), hence the same joint distribution.

Problem 4. For T_1 and T_2 the first two arrival times in a Poisson process with rate λ, find the conditional distribution of T_1 given $T_2 = x$.

Solution. Since according to the solution of the previous problem, T_2 and T_1 have the same joint density as X and Y, found in Problem 1, the conditional distribution of T_1 given $T_2 = x$ is identical to the conditional distribution of Y given $X = x$, which was given at the start, that is to say, uniform on $(0, x)$.

Averaging Conditional Probabilities

For a random variable X with density f_X, the rule of average conditional probabilities becomes the following:

Integral Conditioning Formula

$$P(A) = \int P(A \mid X = x) f_X(x)\,dx$$

The integral breaks up the probability of A according to the values of X:

$$P(A \mid X = x) f_X(x)\, dx = P(A \mid X \in dx) P(X \in dx) = P(A, X \in dx)$$

Just as in the discrete case, $P(A \mid X = x)$ is often specified in advance by the formulation of a problem. Then $P(A)$ can be calculated by the integral conditioning formula, assuming also that the distribution of X is known. Bayes' rule then gives the conditional density of X given that A has occurred:

$$P(X \in dx \mid A) = \frac{P(X \in dx) P(A \mid X = x)}{P(A)} = \frac{f_X(x) P(A \mid X = x)}{P(A)}\, dx$$

The following example shows how the integral conditioning formula arises naturally by taking limits of discrete problems. In this example, as is often the case, the limits defined by integrals are much easier to work with than the discrete sums. The example makes precise the idea of independent trials with probabilty p of success in a setting where it makes clear sense to think of p as picked at random from some distribution before the trials are performed. In the first problem p is picked from a discrete uniform distribution on $N + 1$ evenly spaced points in $[0, 1]$. Passing to the limit as $N \to \infty$ leads to p that is uniformly distributed on $[0, 1]$. Bayesian statisticians view this as a model for independent trials with unknown probability of success.

Example 3. Discrete uniform–binomial.

Suppose there are $N + 1$ boxes labeled by $b = 0, 1, 2, \ldots, N$. Box b contains b black and $N - b$ white balls. A box is picked uniformly at random, and then n balls are drawn at random with replacement from whatever box is picked (the same box for each of the n draws). Let S_n denote the total number of black balls that appear among the n balls drawn.

Problem 1. Find the distribution of S_n.

Solution. Let Π denote the proportion of black balls in the box picked. Let G_N denote the grid of $N + 1$ possible values p of Π:

$$G_N = \{0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}, 1\}$$

For each $p \in G_N$ the binomial formula for n independent trials with probability p of success on each trial gives

$$P(S_n = k \mid \Pi = p) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Averaging with respect to the uniform distribution of Π over the $N + 1$ values in G_N, and substituting $p = b/N$, gives the unconditional distribution of S_n:

$$P(S_n = k) = \sum_{p \in G_N} \binom{n}{k} p^k (1-p)^{n-k} \frac{1}{N+1} \tag{1}$$
$$= \binom{n}{k} \frac{1}{(N+1)N^n} \sum_{b=0}^{N} b^k (N-b)^{n-k}$$

It is hard to simplify this expression further. But the expression is easily evaluated for small values of n and N. To illustrate, for $N = n = 2$ the result is shown in the next table. The limiting behavior for large N is the subject of the next problem.

Distribution of S_2 for $N = 2$

k	0	1	2
$P(S_2 = k)$	$\frac{5}{12}$	$\frac{2}{12}$	$\frac{5}{12}$

Problem 2. For a fixed value of n, find the limiting distribution of S_n, the number of black balls that appear in n draws, as the number of boxes N tends to ∞.

Solution. Expression (1) for $P(S_n = k)$ is $\binom{n}{k}$ times a discrete approximation to the beta integral

$$B(k+1, n-k+1) = \int_0^1 p^k (1-p)^{n-k} dp$$

The approximation in (1) is obtained by taking the average value of the function $p^k(1-p)^{n-k}$ at $N+1$ evenly spaced points p, beween 0 and 1. In the limit as $N \to \infty$, the discrete average converges to the continuous integral. Using the expression for the beta integral in terms of the gamma function, and $\Gamma(m+1) = m!$ for integers m, gives

$$B(k+1, n-k+1) = \frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(k+1+n-k+1)} = \binom{n}{k}^{-1} \frac{1}{n+1} \tag{2}$$

The conclusion is that as $N \to \infty$

$$P(S_n = k) \to \binom{n}{k}\binom{n}{k}^{-1} \frac{1}{n+1} = \frac{1}{n+1}$$

for every $0 \le k \le n$. That is, the limiting distribution of S_n as $N \to \infty$ is uniform on $\{0, 1, \ldots, n\}$.

Example 4. Continuous uniform–binomial.

Suppose that Π is picked uniformly at random from $(0,1)$. Given that $\Pi = p$, let S_n be the number of successes in n independent trials with probability p of success on each trial.

Problem 1. Find the distribution of S_n.

Solution. By the limiting result obtained in the previous example as $N \to \infty$, the answer must be uniform on $\{0, 1, \ldots n\}$. This can be derived directly in the continuous model using the integral conditioning formula. Since the density of Π is $f_\Pi(p) = 1$ for $0 < p < 1$, and 0 otherwise,

$$\begin{aligned} P(S_n = k) &= \int P(S_n = k \mid \Pi = p) f_\Pi(p) dp \qquad (3) \\ &= \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} dp \\ &= \frac{1}{n+1} \end{aligned}$$

by evaluation of the beta integral as in the previous problem.

Discussion. Note the close parallel between the expression (3) for $P(S_n = k)$ obtained by the integral conditioning formula for Π with uniform distribution on $(0,1)$, and the corresponding expression (1) for $P(S_n = k)$ in the previous example for Π with uniform distribution on the set of $N+1$ values in G_N. All that happens is that the sum is replaced by an integral, and $1/(N+1)$, which is both the probability of each point in G_N and the difference between adjacent points in G_N, is replaced by the calculus differential dp representing the probability that the uniform variable falls in an infinitesimal length dp near p.

Problem 2. Find the conditional distribution of Π given that $S_n = k$.

Solution. Using Bayes' rule, for $0 < p < 1$,

$$\begin{aligned} P(\Pi \in dp \mid S_n = k) &= \frac{P(\Pi \in dp) P(S_n = k \mid \Pi = p)}{P(S_n = k)} \\ &= (n+1) \binom{n}{k} p^k (1-p)^{n-k} dp \end{aligned}$$

This is the density at p of the beta distribution with parameters $k+1$ and $n-k+1$, times dp. Conclusion: the conditional distribution of Π given $S_n = k$ is beta$(k+1,\ n-k+1)$.

Problem 3. In the above setup, given that n trials have produced k successes, what is the probability that the next trial is a success?

Solution. Given $\Pi = p$ and $S_n = k$, the next trial is a success with probability p, by the assumption of independent trials with constant probability p of success given $\Pi = p$. Given just $S_n = k$, the value of Π is unknown. Rather, Π is a random variable with beta$(k+1,\ n-k+1)$ distribution. By the integral conditioning formula, the required probability is the conditional expectation of Π given $S_n = k$, which is $(k+1)/(n+2)$, by the formula $a/(a+b)$ for the mean of the beta (a, b) distribution. In detail:

$$\begin{aligned} & P(\text{next trial a success} \mid S_n = k) \\ &= \int_0^1 P(\text{next trial a success} \mid S_n = k, \Pi = p) f_\Pi(p \mid S_n = k) dp \\ &= \int_0^1 p f_\Pi(p \mid S_n = k) dp = E(\Pi \mid S_n = k) = \frac{k+1}{n+2} \end{aligned}$$

Discussion. In particular, for $k = n$, given n successes in a row, the chance of one more success is $(n+1)/(n+2)$. This formula, for the probability of one more success given a run of n successes in independent trials with unknown success probability assumed uniformly distributed on $(0, 1)$, is known as *Laplace's law of succession.* Laplace illustrated his formula by calculating the probability that the sun will rise tomorrow, given that it has risen daily for 5000 years, or $n = 1{,}826{,}213$ days. But this kind of application is of doubtful value. Both the assumption of independent trials with unknown p and the uniform prior distribution of p make little sense in this context.

Example 5. Simulation of uniform–binomial.

Suppose you have available a random number generator which you are willing to believe generates independent uniform $(0, 1)$ variables $U_0, U_1, \ldots$.

Problem 1. How could you simulate a pair of values from the joint distribution of Π and S_n considered above, with Π uniform on $(0, 1)$, and S_n binomial(n, p) given $\Pi = p$?

Solution. Set

$$\Pi = U_0, \text{ and } S_n = \sum_{i=1}^{n} I(U_i < U_0)$$

where $I(U_i < U_0)$ is an indicator variable that is 1 if $(U_i < U_0)$ and 0 otherwise. If $\Pi = p$, then $S_n = \sum_{i=1}^n I(U_i < p)$ is the sum of n independent indicator variables, each of which is 1 with probability p and 0 with probability $1 - p$, exactly as required.

Problem 2. Use this construction to calculate $P(S_n = k)$ without integration.

Solution. By construction of S_n from $U_0, U_1, \ldots, U_n$

$(S_n = 0)$ if and only if U_0 is the smallest of the $U_0, U_1, \ldots, U_n$

$(S_n = 1)$ if and only if U_0 is the second smallest of the $U_0, U_1, \ldots, U_n$

... ...

$(S_n = n)$ if and only if U_0 is the largest of the $U_0, U_1, \ldots, U_n$

Since all $(n+1)!$ possible orderings of the $U_0, U_1, \ldots, U_n$ are equally likely, each of these events has the same probability $1/(n+1)$.

Remark. This calculation is closely related to the distribution of order statistics treated in Section 4.6. For $j = 1, \ldots, n+1$, let $U_{(j)}$ denote the jth smallest of the $n+1$ variables $U_0, \ldots, U_n$. Then the event $S_n = j - 1$, that there are exactly $j - 1$ values U_i less than U_0, is identical to the event $U_{(j)} = U_0$, that the jth smallest of the U_i equals U_0. The solution of Problem 2 in Example 4 now translates into the following: the conditional distribution of U_0, or of $U_{(j)}$, given that $U_{(j)} = U_0$, is beta $(j, n - j + 2)$. By symmetry, the same is true for U_k instead of U_0 for any $1 \leq k \leq n$. Consequently, the distribution of $U_{(j)}$, the jth smallest of $n+1$ independent uniform $(0, 1)$ variables, is beta $(j, n - j + 2)$, independently of K, where K is the random index k such that $U_k = U_{(j)}$. This agrees with the result of Section 4.6, with the present $n + 1$ and j instead of n and k in that section.

Independence

In the continuous case, just as in the discrete case, it can be shown that each of the following conditions is equivalent to independence of random variables X and Y:

- the conditional distribution of Y given $X = x$ does not depend on x;
- the conditional distribution of X given $Y = y$ does not depend on y.

By integration with respect to the distribution of X, the common conditional distribution of Y given $X = x$ then equals the unconditional distribution of Y. That is to say, for all subsets B in the range of Y

$$P(Y \in B \,|\, X = x) = P(Y \in B)$$

Similarly for all subsets A in the range of X

$$P(X \in A \,|\, Y = y) = P(X \in A)$$

These are variations of the basic definition of independence of X and Y, which is

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

for all subsets A and B in the ranges of X and Y respectively. When X and Y have densities, X and Y are independent if and only if $f_Y(y \,|\, X = x) = f_Y(y)$ for all x and y, and again if and only if $f_X(x \,|\, Y = y) = f_X(x)$ for all x and y. So the general multiplication rule for densities reduces in this case to the formula

$$f(x, y) = f_X(x) f_Y(y)$$

for independent variables X and Y. This formula was applied in Section 5.2.

Conditional Expectations

The conditional expectation of Y given $X = x$, denoted $E(Y \mid X = x)$, is defined as the expectation of Y relative to the conditional distribution of Y given $X = x$. More generally, for a function g, assuming that Y has a conditional density $f_Y(y \mid X = x)$,

$$E[g(Y) \mid X = x] = \int g(y) f_Y(y \mid X = x) dy$$

Taking $g(y) = y$ gives $E(Y \mid X = x)$. And integrating the conditional expectation with respect to the distribution of X gives the unconditional expectation

$$E[g(Y)] = \int E[g(Y) \mid X = x] f_X(x)\, dx$$

These formulae are extensions to general functions g of the basic conditional probability formulae, which are the special cases when g is an indicator. As a general rule, all the basic properties of conditional expectations, considered in the discrete case in Section 6.2, remain valid in the density case.

Example 6. Uniform distribution on a triangle.

Problem. Suppose, as in Example 1, that (X, Y) is chosen uniformly at random from the triangle $\{(x, y) : x \geq 0, y \geq 0, x + y \leq 2\}$. Find $E(Y \mid X)$ and $E(X \mid Y)$.

Solution. As argued before, given $X = x$, for $0 < x < 2$, Y has uniform distribution on $(0, 2 - x)$. Since the mean of this conditional distribution is $(2 - x)/2$,

$$E(Y \mid X = x) = (2 - x)/2$$

In short notation

$$E(Y \mid X) = (2 - X)/2$$

Similarly, because joint density of X and Y is symmetric in x and y,

$$E(X \mid Y) = (2 - Y)/2$$

Conditioning Formulae: Discrete Case

Multiplication rule: The joint probability is the product of the marginal and the conditional

$$P(X = x, Y = y) = P(X = x)P(Y = y \mid X = x)$$

Division rule: The conditional probability of $Y = y$ given $X = x$ is

$$P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

Bayes' rule:

$$P(X = x \mid Y = y) = \frac{P(Y = y \mid X = x)P(X = x)}{P(Y = y)}$$

Conditional distribution of Y given $X = x$: Sum the conditional probabilities

$$P(Y \in B \mid X = x) = \sum_{y \in B} P(Y = y \mid X = x)$$

Conditional expectation of $g(Y)$ given $X = x$: Sum g against the conditional probabilities

$$E\,(g(Y) \mid X = x) = \sum_{\text{all } y} g(y)P(Y = y \mid X = x)$$

Average conditional probability:

$$P(B) = \sum_{\text{all } x} P(B \mid X = x)P(X = x)$$

$$P(Y = y) = \sum_{\text{all } x} P(Y = y \mid X = x)P(X = x)$$

Average conditional expectation:

$$E(Y) = \sum_{\text{all } x} E(Y \mid X = x)P(X = x)$$

Conditioning Formulae: Density Case

Multiplication rule: The joint density is the product of the marginal and the conditional

$$f(x, y) = f_X(x) f_Y(y \mid X = x)$$

Division rule: The conditional density of Y at y given $X = x$ is

$$f_Y(y \mid X = x) = \frac{f(x, y)}{f_X(x)}$$

Bayes' rule:

$$f_X(x \mid Y = y) = \frac{f_Y(y \mid X = x) f_X(x)}{f_Y(y)}$$

Conditional distribution of Y given $X = x$: Integrate the conditional density

$$P(Y \in B \mid X = x) = \int_B f_Y(y \mid X = x) dy$$

Conditional expectation of $g(Y)$ given $X = x$: Integrate g against the conditional density:

$$E(g(Y) \mid X = x) = \int g(y) f_Y(y \mid X = x) dy$$

Average conditional probability:

$$P(B) = \int P(B \mid X = x) f_X(x)\, dx$$

$$f_Y(y) = \int f_Y(y \mid X = x) f_X(x)\, dx$$

Average conditional expectation:

$$E(Y) = \int E(Y \mid X = x) f_X(x)\, dx$$

Exercises 6.3

1. Suppose X has uniform $(0, 1)$ distribution and $P(A|X = x) = x^2$. What is $P(A)$?

2. Let X and Y have the following joint density:

$$f(x, y) = \begin{cases} 2x + 2y - 4xy & \text{for } 0 \le x \le 1 \text{ and } 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

a) Find the marginal densities of X and Y.

b) Find $f_Y(y \,|\, X = \frac{1}{4})$. c) Find $E(Y \,|\, X = \frac{1}{4})$.

3. Let (X, Y) be as in Example 1. Find a formula for $P(Y \le y|X = x)$.

4. Suppose X, Y are random variables with joint density

$$f(x, y) = \begin{cases} \lambda^3 x e^{-\lambda y} & \text{for } 0 < x < y \\ 0 & \text{otherwise} \end{cases}$$

a) Find the density of Y. What is $E(Y)$? b) Compute $E(X|Y = 1)$.

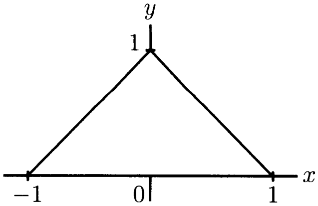
5. Suppose (X, Y) has uniform distribution on the triangle shown in the diagram. For x between -1 and 1, find:

a) $P(Y \ge \frac{1}{2}|X = x)$;

b) $P(Y < \frac{1}{2}|X = x)$;

c) $E(Y|X = x)$;

d) $Var(Y|X = x)$.



6. Suppose X, Y are random variables with joint density

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sqrt{x(y - x)}}\, e^{-y/2} \qquad (0 < x < y)$$

a) Find the distribution of Y. [*Hint*: For integration use the substitution $x = y\,s$.]

b) Compute $E(X|Y = 1)$.

7. Suppose that Y and Z are random variables with the following joint density:

$$f(y, z) = \begin{cases} k(z - y) & \text{for } 0 \le y \le z \le 1 \\ 0 & \text{otherwise} \end{cases}$$

for some constant k. Find:

a) the marginal distribution of Y; b) $P\left(Z < \frac{2}{3} \,|\, Y = \frac{1}{2}\right)$.

8. The random variable X has a uniform distribution on $(0, 1)$. Given that $X = x$, the random variable Y is binomial with parameters $n = 5$ and $p = x$.

a) Find $E(Y)$ and $E(Y^2)$. b) Find $P(Y = y \text{ and } x < X < x + dx)$.

c) Find the density of X given $Y = y$. Do you recognize it? If yes, as what?

9. Let A and B be events and let Y be a random variable uniformly distributed on $(0,1)$. Suppose that, conditional on $Y = p$, A and B are independent, each with probability p. Find:

a) the conditional probability of A given that B occurs;

b) the conditional density of Y given that A occurs and B does not.

10. Conditioning a Poisson process on the number of arrivals in a fixed time. Let T_1 and T_5 be the time of the first and fifth arrivals in a Poisson process with rate λ, as in Section 4.2.

a) Find the conditional density of T_1 given that there are 10 arrivals in the time interval $(0,1)$.

b) Find the conditional density of T_5 given that there are 10 arrivals in the time interval $(0,1)$.

c) Recognize the answers to a) and b) as named densities, and find the parameters.

11. Suppose X has uniform distribution on $(-1,1)$ and, given $X = x$, Y is uniformly distributed on $(-\sqrt{1-x^2}, \sqrt{1-x^2})$. Is (X,Y) then uniformly distributed over the unit disk $\{(x,y) : x^2 + y^2 < 1\}$? Explain carefully.

12. Suppose there are ten atoms, each of which decays by emission of an α-particle after an exponentially distributed lifetime with rate 1, independently of the others. Let T_1 be the time of the first α-particle emission, T_2 the time of the second. Find:

a) the distribution of T_1;

b) the conditional distribution of T_2 given T_1;

c) the distribution of T_2.

13. Let X and Y be independent random variables, X with uniform distribution on $(0,3)$, Y with Poisson (λ) distribution. Find:

a) a formula in terms of λ for $P(X < Y)$;

b) the conditional density of X given $X < Y$, and sketch its graph in the cases $\lambda = 1, 2, 3$;

c) $E(X|X < Y)$.

14. Bayesian sufficiency. Let $S_n = X_1 + \cdots + X_n$ be the number of successes in a sequence of n independent Bernoulli (p) trials $X_1, X_2, \ldots, X_n$ with unknown success probability p. Regard p as the value of a random variable Π whose prior distribution has some density $f(p)$ on $(0,1)$. Show that the conditional (posterior) distribution of Π given $X_1 = x_1, \ldots, X_n = x_n$, for any particular sequence of zeros and ones $x_1, \ldots, x_n$ with $x_1 + \cdots + x_n = k$, depends only the observed number of successes k in the n trials, and not on the order in which the k successes and $n-k$ failures appear. Deduce that this conditional distribution is identical to the posterior distribution of Π given $S_n = k$. [This is another expression of the fact that S_n is a sufficient statistic for p. See Exercise 6.2.14.]

15. Beta-binomial. As in Exercise 14 let $S_n = X_1 + \cdots + X_n$ be the number of successes in a sequence of n independent Bernoulli (p) trials $X_1, X_2, \ldots, X_n$, with unknown success probability p, regarded as the value of a random variable Π.

a) Suppose the prior distribution of Π is beta (r, s) for some $r > 0$ and $s > 0$. Show that the posterior distribution of Π given $S_n = k$ is beta $(r + k, s + n - k)$. [*Hint for quick solution*: It is enough to show that the posterior density is *proportional* to the beta $(r + k, s + n - k)$ density. See Chapter 4 Review Exercise 8.]

b) Using the fact that the total integral of the beta $(r+k, s+n-k)$ density is 1, find a formula for the unconditional probability $P(S_n = k)$.

c) Check your result in part b) agrees with the distribution of S_n found in Example 4 in the case $r = s = 1$.

d) For general r and s find the posterior mean $E(\Pi \mid S_n = k)$ and the posterior variance $Var(\Pi \mid S_n = k)$.

e) Suppose n is very large and the observed proportion of successes $\hat{p} = k/n$ is not very close to either 0 or 1. Show that no matter what r and s, provided n is large enough, $E(\Pi \mid S_n = k) \approx \hat{p}$ and $Var(\Pi \mid S_n = k) \approx \hat{p}(1 - \hat{p})/n$.

[It can be shown that the posterior distribution of Π given $S_n = k$ is approximately normal under the assumptions in e). So
for large enough n, the conditional distribution of the unknown value of p, given the observed proportion of successes $\hat{p}$ in n trials, is approximately normal with mean $\hat{p}$ and standard deviation $\sqrt{\hat{p}(1-\hat{p})}/\sqrt{n}$,
regardless of the prior parameters r and s. The same conclusion holds for any strictly positive and continuous prior density $f(p)$ instead of a beta prior. In the long run, any reasonable prior opinion is overwhelmed by the data. The italicized assertion should be compared to the following paraphrase of the normal approximation to the binomial distribution:
for large enough n, the distribution of proportion of successes $\hat{p}$ in n trials, given the probability p of success on each trial, is approximately normal with mean p and standard deviation $\sqrt{p(1-p)}/\sqrt{n}$.
While the assertions are very similar, and both true, it is not a trivial matter to pass from one to the other. There is a big conceptual difference between, on the one hand, the distribution of $\hat{p}$ for a fixed and known value of p, which has a clear frequency interpretation in terms of repeated blocks of n trials with the same p, and on the other hand, the posterior distribution of p given $\hat{p}$, which while intuitive from a subjective standpoint, is almost impossible to interpret in terms of long-run frequencies. Long-run frequency of what? The problem is that for large n, in any model of repeated blocks of n trials, the exact value of $\hat{p}$ observed in the first block will typically not be observed even once again until after a very large number of blocks have been examined. The number of blocks required to find the first repeat is of order $\sqrt{n}$ if the same p is used in each block, and order n if p is randomized for each block using the prior distribution: this is because the probability of the most likely values of $\hat{p}$ is of order $1/\sqrt{n}$ in the first case, by the normal approximation to the binomial, and order $1/n$ in the second case, as typified when the prior is uniform on $(0, 1)$ and the distribution of $\hat{p}$ is uniform on the $n + 1$ possible multiples of $1/n$. Either way, it is hard to make a convincing frequency interpretation of the conditional distribution of p given an exact observed value of $\hat{p}$.]

16. Negative binomial distribution for number of accidents. Consider a large population of individuals subject to accidents at various rates. Suppose the empirical distribution of accident rates over the whole population is well approximated by the gamma (r, α) distribution for some $r > 0$ and $\alpha > 0$. Suppose that given an individual has accident rate λ per day, the number of accidents that individual has in t days has Poisson (λt) distribution. Let Λ be the accident rate and N be the number of accidents in t days for an individual picked at random from this population. So Λ has gamma (r, α) distribution, and given $\Lambda = \lambda$, N has Poisson (λt) distribution.

a) Show by integration that

$$P(N = k) = \frac{\Gamma(r+k)}{\Gamma(r)k!} p^r q^k \quad (k = 0, 1, 2 \ldots) \text{ where } p = \alpha/(t+\alpha), \quad q = t/(t+\alpha)$$

b) Evaluate $\Gamma(r + k)/\Gamma(r)$ as a product of k factors. Deduce that if r is a positive integer, the distribution of N is the same as the distribution of the number of failures before the rth success in Bernoulli (p) trials, as found in Section 3.4.

[In general, the distribution of N defined in a) is called the *negative binomial*(r, p) distribution, now defined for arbitrary $r > 0$ and $0 < p < 1$. The terminology is explained by the following relation between this distribution and the binomial expansion for the negative power $-r$.]

c) Show, either by conditioning on Λ, or from a) and b), that N has generating function

$$E(z^N) = p^r(1 - zq)^{-r} \qquad (|z| < 1)$$

d) Find $E(N)$ and $E(N^2)$ in terms of r and p by conditioning on Λ. Deduce a formula for $Var(N)$. Check for integer r that your results agree with those obtained in Section 3.4.

e) Derive $E(N)$ and $Var(N)$ another way by differentiating the generating function. (Refer to Exercise 3.4.22.)

f) Show that for each integer $k \geq 0$, the conditional density of Λ given $N = k$ is a gamma density, and find its parameters.

17. Sums of independent negative binomial variables. Consider, as in Exercise 16, a large population of individuals subject to accidents at various rates. Suppose now that an individual picked at random from the population is subject to one kind of accident at rate Λ_1 per day, and another kind of accident at rate Λ_2 per day, where Λ_1 and Λ_2 are independent gamma variables with parameters (r_1, α) and (r_2, α) for some $\alpha > 0$. Assume that given $\Lambda_1 = \lambda_1$ and $\Lambda_2 = \lambda_2$ the two types of accidents occur according to independent Poisson processes with rates λ_1 and λ_2. Let N_1 and N_2 be the numbers of accidents of these two kinds the individual has in t days.

a) Describe the joint distribution of N_1 and N_2.

b) What is the distribution of $N_1 + N_2$? [*Hint*: No calculation required. Use results about sums of independent random variables with gamma or Poisson distributions.] Check your conclusion is consistent with the mean and variance formulae of Exercise 16.

c) Suppose $X_i, 1 \leq i \leq k$ are k independent random variables, and that X_i has negative binomial (r_i, p) distribution for some $r_i > 0$, $0 < p < 1$. What is the distribution of $X_1 + \cdots + X_n$? Explain carefully how your conclusion follows from parts a) and b).

d) Derive the result of c) another way using generating functions [see Chapter 3 Review Exercise 34].

6.4 Covariance and Correlation

Covariance is a quantity which appears in calculation of the variance of a sum of possibly dependent random variables. This quantity is useful in variance calculations, but like variance is hard to interpret intuitively. Correlation is a standardized covariance which is easier to interpret. It provides a measure of the degree of linear dependence between two variables. In Section 3.3, the formula

$$Var(X+Y) = Var(X) + Var(Y) \qquad \text{if } X \text{ and } Y \text{ are independent}$$

was derived from the more general formula

$$Var(X+Y) = Var(X) + Var(Y) + 2E\left[(X-\mu_X)(Y-\mu_Y)\right]$$

where $\mu_X = E(X)$ and $\mu_Y = E(Y)$. For independent random variables, the last term vanishes. In general, for two random variables X and Y with finite second moments, there is the following:

Definition of Covariance

The *covariance of X and Y*, denoted $Cov(X,Y)$, is the number

$$Cov(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right]$$

where $\quad \mu_X = E(X), \quad \mu_Y = E(Y)$

Alternative Formula

$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

Variance of a Sum

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

Proof of alternative formula for covariance. Expand

$$(X-\mu_X)(Y-\mu_Y) = XY - \mu_X Y - X\mu_Y + \mu_X\mu_Y$$

and take expectations.□

Variance. Notice that $Cov(X,X) = Var(X)$, so these formulae for covariance are extensions of old formulae for variance.

Independence. If X and Y are independent then $Cov(X,Y) = 0$.

Warning. $Cov(X,Y) = 0$ *does not* imply X and Y are independent. See Exercises.

Indicators

Let $X = I_A$ be the indicator of event A, and $Y = I_B$ the indicator of another event B. These could be events in any outcome space, where there is given a probability distribution P. In this case

$$XY = I_A I_B = I_{AB}$$

is the indicator of the intersection of the events A and B. Thus

$$E(I_A) = P(A); \qquad E(I_B) = P(B); \qquad E(I_A I_B) = P(AB)$$

$$Cov(I_A, I_B) = P(AB) - P(A)P(B)$$

This covariance is

positive	iff	$P(AB) > P(A)P(B)$, when A and B are called *positively dependent*;
zero	iff	$P(AB) = P(A)P(B)$, when A and B are *independent;*
negative	iff	$P(AB) < P(A)P(B)$, when A and B are called *negatively dependent.*

In the case of positive dependence, learning that B has occurred increases the chance of A:

$$P(A|B) > P(A) \quad \text{and vice versa} \quad P(B|A) > P(B)$$

For negative dependence, learning that B has occurred decreases the chance of A:

$$P(A|B) < P(A) \quad \text{and vice versa} \quad P(B|A) < P(B)$$

These formulations of positive and negative dependence are easily seen to be equivalent to those in the box, by using the formula for $P(A|B)$, and rearranging inequalities. The most extreme case of positive dependence is if A is a subset of B, with $0 < P(A) \leq P(B) < 1$. Then, given that A occurs, B is certain to occur. In this case, given that B occurs, A is more likely to occur than before

$$P(A|B) = P(AB)/P(B) = P(A)/P(B) > P(A)$$

The most extreme case of negative dependence is if A and B are mutually exclusive events B with $P(A) > 0$ and $P(B) > 0$. Then, given that A occurs, B cannot occur, and vice versa.

Example 1. **Draws with and without replacement.**

Consider two draws at random from a box of b black balls and w white balls, where $b > 0$, $w > 0$. Let Black_i and White_i denote the events of getting a black or a white ball on the ith draw, $i = 1, 2$. Then you can check that the dependence between pairs of these events from different draws is affected by whether the sampling is done with or without replacement, as shown in the following table.

Dependence Between Events on Different Draws

Pairs of events	Sampling with replacement	Sampling without replacement
$\text{Black}_1, \text{Black}_2$	independent	− dependent
$\text{Black}_1, \text{White}_2$	independent	+ dependent
$\text{White}_1, \text{White}_2$	independent	− dependent
$\text{White}_1, \text{Black}_2$	independent	+ dependent

The Sign of the Covariance

As a general rule, the sign of $Cov(X, Y)$ is *positive* if above-average values of X tend to be associated with above-average values of Y, and below-average values of X with below-average values of Y. The random variable $(X - \mu_X)(Y - \mu_Y)$ is then most likely positive, with a positive expectation.

The sign of $Cov(X, Y)$ is *negative* if above-average values of X tend to be associated with below-average values of Y, and vice versa. Then $(X - \mu_X)(Y - \mu_Y)$ is most likely negative, with a negative expectation.

$Cov(X, Y)$ is *zero* only in special cases when there is no such association between the variables X and Y. Then $(X - \mu_X)(Y - \mu_Y)$ has positive values balanced by negative values, and expected value zero.

While the sign of the covariance can be interpreted as above, its magnitude is hard to interpret. It is easier to interpret the *correlation of X and Y*, denoted here by $Corr(X, Y)$, which is defined as follows:

Definition of Correlation

$$Corr(X, Y) = \frac{Cov(X, Y)}{SD(X)SD(Y)}$$

Assume now that neither X nor Y is a constant, so $SD(X)SD(Y) > 0$. The sign of $Cov(X, Y)$ is then the same as the sign of $Corr(X, Y)$.

Conditions for *X* and *Y* to be Uncorrelated

The following three conditions are equivalent:

$$Corr(X, Y) = 0$$

$$Cov(X, Y) = 0$$

$$E(XY) = E(X)E(Y)$$

in which case X and Y are called *uncorrelated.* Independent variables are uncorrelated, but uncorrelated variables are not necessarily independent.

Let X^* and Y^* now denote X and Y rescaled to standard units. So

$$X^* = (X - \mu_X)/SD(X) \quad \text{and} \quad Y^* = (Y - \mu_Y)/SD(Y)$$

Then

$$E(X^*) = E(Y^*) = 0 \quad \text{and} \quad SD(X^*) = SD(Y^*) = 1$$

by the scaling properties of E and SD. And you can check that

$$Corr(X, Y) = Cov(X^*, Y^*) = E(X^*Y^*)$$

So correlation is a kind of standardized covariance that is unaffected by changes of origin or units of measurement. See Exercises.

Correlations are between -1 and +1

$$-1 \leq Corr(X, Y) \leq 1$$

no matter what the joint distribution of X and Y.

Proof. Since $E(X^{*2}) = E(Y^{*2}) = 1$

$$0 \leq E(X^* - Y^*)^2 = 1 + 1 - 2E(X^*Y^*)$$
$$0 \leq E(X^* + Y^*)^2 = 1 + 1 + 2E(X^*Y^*)$$

Thus $-1 \leq E(X^*Y^*) \leq 1$, and $Corr(X, Y) = E(X^*Y^*)$ by the preceding discussion. □

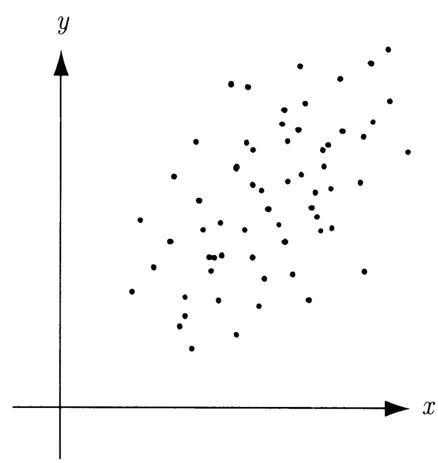
Correlations of ± 1. The proof that correlations are between ± 1 shows $Corr(X, Y) = +1$ if and only if $E(X^* - Y^*)^2 = 0$, that is, if and only if $X^* = Y^*$ with probability one. This means there are constants a and b with $a > 0$ such that

$$Y = aX + b$$

with probability 1. That is to say, a correlation of $+1$ indicates a deterministic linear relationship between X and Y with positive slope. Similarly, a correlation of -1 indicates a deterministic linear relationship between X and Y with negative slope. Correlations between -1 and $+1$ indicate intermediate degrees of linear association between the two variables.

Example 2. Empirical correlations.

Like expectation and variance, covariance and correlation are generalizations to random variables of corresponding notions for empirical variables. Suppose $(x_1, y_1), \ldots, (x_n, y_n)$ is a list of n pairs of numbers, and (X, Y) is one of these pairs picked uniformly at random. Then the joint distribution of (X, Y) puts probability $1/n$ at each of the pairs, as suggested by the scatter diagram:



$$E(X) = \bar{x} \quad \text{and} \quad SD(X) = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

and similarly for Y instead of X. Also

$$E(XY) = \frac{1}{n}\sum_{k=1}^{n} x_k y_k \qquad \text{so}$$

$$Cov(X, Y) = E(XY) - E(X)E(Y) \qquad \text{and} \qquad Corr(X, Y) = \frac{Cov(X, Y)}{SD(X)SD(Y)}$$

can be computed from the list of number pairs. If the list of number pairs is a list of empirical measurements, or a sample of some kind, these may be called empirical or sample quantities. These quantities are all defined in terms of averages, which may be expected to converge to theoretical expectations as the sample size n increases, under conditions of random sampling. For example, the empirical correlation of n observed values of independent random variables $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, all with the same joint distribution, will most probably be close to the theoretical correlation of X_1 and Y_1, provided n is sufficiently large. Thus a correlation in a theoretical model is often estimated by an empirically observed correlation based on a random sample. In particular, the empirical correlation of two variables over a large population can be estimated this way by the procedure of random sampling.

Example 3. Correlation and distribution of the sum.

This example shows in a simple case how the distribution of the sum of random variables X and Y is affected by their correlation. Suppose a gambler can bet on the value of a number U chosen uniformly at random from the numbers $1, 2, \ldots, 8$. The gambler can choose any set A of four numbers, such as $A = \{1, 2, 3, 4\}$, and place an even-money bet of \$1 on A. So the gambler wins \$1 if $U \in A$, and loses \$1 if $U \in A^c$. Let \$$X$ denote the gambler's net gain from this contract. Then, X has value $+1$ if $U \in A$, -1 if $U \in A^c$. In terms of indicators,

$$X = 2I_A - 1$$

Clearly $E(X) = 0$. The bet is fair no matter what set A the gambler chooses, because $P(A) = P(A^c) = 1/2$ for every set of four numbers A.

Suppose now that in addition to placing a bet on A, the gambler is also free to place at the same time a similar bet on a second set of four numbers B, for example $B = \{1, 3, 5, 7\}$. Let

$$Y = 2I_B - 1$$

denote the net gain to the gambler from this second bet. Then the gambler's overall gain from the placement of the two bets is the sum

$$S = X + Y$$

Notice that the distribution of X and the distribution of Y are the same, uniform on $\{-1, 1\}$, regardless of the gambler's choice of sets A and B. But the distribution of S is affected by the degree of dependence between X and Y, which is governed in turn by the amount of overlap between A and B. Clearly, $E(S)$ is zero no matter what the choice of A and B. But $SD(S)$ is affected by the gambler's choice of A and B. This standard deviation gives an indication of the likely size of the fluctuation in the gambler's fortune due to the combined bet.

Problem. Find how the standard deviation of S is determined by the choice of A and B.

Solution. Use the addition rule for variance

$$\begin{aligned} Var(S) &= Var(X) + Var(Y) + 2Cov(X,Y) \\ &= 2 + 2Corr(X,Y) \end{aligned}$$

because $SD(X) = SD(Y) = 1$, so $Corr(X,Y) = Cov(X,Y)$ in this case. Because $X = 2I_A - 1$, $Y = 2I_B - 1$, and the correlation coefficient is unchanged by linear transformations,

$$Corr(X,Y) = Corr(I_A, I_B) = \frac{Cov(I_A, I_B)}{SD(I_A)SD(I_B)} = \left(P(AB) - \frac{1}{4}\right) \Big/ \frac{1}{4} = 4P(AB) - 1$$

This used $P(A) = P(B) = 1/2$, which makes $SD(I_A) = SD(I_B) = 1/2$. Using the earlier expression for $Var(S)$ this gives

$$SD(S) = \sqrt{8P(AB)} = \sqrt{\#(AB)}$$

where $\#(AB)$ is the number of points in the intersection of A and B, so $P(AB) = \#(AB)/8$.

Discussion. The formula shows that the larger the overlap between A and B, the larger will be the likely size of the fluctuation in the gambler's fortune as a result of betting on both A and B. This is intuitively clear if you think about the following special cases:

Case $\#(AB) = 0$, $Corr(X,Y) = -1$, $SD(S) = 0$. This means $B = A^c$. Then $Y = -X$, because whatever is gained on one bet is lost on the other. So $S = X + Y = 0$. This is a strategy of extreme hedging, with zero result.

Case $\#(AB) = 1$, $Corr(X,Y) = -1/2$, $SD(S) = 1$. Intuitively, this is still hedging. The two bets tend to cancel each other.

Case $\#(AB) = 2$, $Corr(X,Y) = 0$, $SD(S) = \sqrt{2}$. In this case A and B are independent. Therefore, so too are the indicator random variables I_A and I_B, and the random variables $X = 2I_A - 1, Y = 2I_B - 1$ representing the net gains from the two bets. So the net effect of betting on both A and B in one game is the same as the effect of betting on A in one game, then betting on A again in a second game, independent of the first. The distribution of S in this case is the familiar binomial $(2, 1/2)$ distribution, but centered at 0 and rescaled by a factor of 2, because

$$S = X + Y = 2(I_A + I_B) - 2$$

where $I_A + I_B$ is the number of successes in two independent trials with probability $1/2$ of success on each trial, with binomial $(2, 1/2)$ distribution. The appearance of $\sqrt{2}$ as the standard deviation in this case illustrates the square root law for the standard deviation of the sum of $n = 2$ independent variables.

Case $\#(AB) = 3$, $Corr(X, Y) = 1/2$, $SD(S) = \sqrt{3}$. This is a bolder strategy.

Case $\#(AB) = 4$, $Corr(X, Y) = 1$, $SD(S) = 2$. Now $A = B$. All the gambler's eggs are in one basket. This is the boldest strategy for the gambler, effectively doubling the stake on A from \$1 to \$2.

Example 4. Red and black.

Let N_R be the number of reds that appear, N_B the number of blacks, in n spins of a roulette wheel that has proportion r of its numbers red, proportion b black, and the rest of its numbers green. (So $r + b < 1$. For a Nevada roulette wheel, as described at the end of Section 1.1, $r = b = 18/38$.)

Problem. Find $Corr(N_R, N_B)$.

Solution. Notice first, without calculation, that the answer ought to be negative for the usual case with $r+b \approx 1$. If $r+b = 1$ (no green numbers on the wheel) then $N_B = n - N_R$ which makes $Corr(N_R, N_B) = -1$. For $r + b \approx 1$ this relation is still approximately correct, so you should expect a correlation close to -1. Since N_R is a binomial (n, r) random variable,

$$E(N_R) = nr \text{ and } SD(N_R) = \sqrt{nr(1-r)}$$

and similarly for N_B, with b instead of r. Since

$$Cov(N_R, N_B) = E(N_R N_B) - E(N_R)E(N_B)$$

to calculate

$$Corr(N_R, N_B) = \frac{Cov(N_R, N_B)}{SD(N_R)SD(N_B)}$$

the only missing ingredient is $E(N_R N_B)$. You might try to calculate this from the joint distribution of N_R and N_B, but you will find this a frightful task. It is difficult to calculate even the variance of N_R directly from its binomial distribution, and the covariance with N_B is worse. The way around this difficulty is to use the connection between $Cov(N_R, N_B)$ and the variance of $N_R + N_B$

$$Var(N_R + N_B) = Var(N_R) + Var(N_B) + 2\,Cov(N_R, N_B)$$

The point is that $N_R + N_B$ is just the number of spins which are either red or black, which is a binomial $(n, r + b)$ random variable, with variance $n(r + b)(1 - r - b)$. Rearrange the equation and substitute all the variances to get

$$Cov(N_R, N_B) = \frac{1}{2}n\left[(r + b)(1 - r - b) - r(1 - r) - b(1 - b)\right] = -nrb,$$

hence,

$$Corr(N_R, N_B) = \frac{-nrb}{\sqrt{nr(1-r)}\sqrt{nb(1-b)}} = -\sqrt{\frac{rb}{(1-r)(1-b)}}$$

Discussion. In particular, for a Nevada roulette wheel,

$$r/(1-r) = b/(1-b) = 18/20 = 0.9 \qquad \text{so}$$

$$Corr(N_R, N_B) = -0.9$$

Note the interesting fact that the correlation does not depend at all on the number of spins n, only on the proportions of red and black. Also, the correlation is always negative, no matter what the proportions r and b.

Example 5. Correlations in the multinomial distribution.

Suppose the joint distribution of $(N_1, \ldots, N_m)$ is multinomial with parameters n and $(p_1, \ldots, p_m)$.

Problem. Find $Corr(N_i, N_j)$.

Solution. Call results in category i red, results in category j black, and results in all other categories green. Then the joint distribution of N_i and N_j is is the same as the joint distribution of N_R and N_B in the previous problem, for $r = p_i$, $b = p_j$. Since the correlation between two variables is determined by their joint distribution (by definition of correlation and the change of variable principle) this choice of r and b makes $Corr(N_i, N_j) = Corr(N_R, N_B)$. That is to say, from the solution of the previous problem,

$$Corr(N_i, N_j) = -\sqrt{\frac{p_i p_j}{(1-p_i)(1-p_j)}}$$

Correlation and Conditioning

An important connection between the ideas of correlation and conditioning is brought out by the following example.

Example 6. Sharkey's Casino.

At Sharkey's Casino the roulette wheels spin an average of one thousand times a day. Every day, Sharkey records the total numbers of red and black spins for the day on a computer. One day he notices that over the years he has been keeping data, the correlation between the number of reds and number of blacks has come out around

$+0.8$, rather than around -0.9 as predicted by the above calculation. Sharkey is very concerned that his roulette wheels are not obeying the laws of chance, and that someone might take advantage of it.

Problem. Should Sharkey get new roulette wheels?

Solution. Despite the fact that no matter what the number of spins n, the correlation between numbers of reds and blacks is -0.9, this does not imply that the same is true for a random number of spins, say N, the number of spins in a day picked at random at Sharkey's. While the expected value of N may be estimated as 1000 based on the long-run average of 1000 spins a day, it is reasonable to expect some spread in the distribution of N due to fluctuations in the number of customers and the rate of play. Since to a first approximation $N_B \approx \frac{18}{38}N$, $N_R \approx \frac{18}{38}N$, both N_B and N_R are positively correlated with N. If there is enough spread in the distribution of N, this will make for a positive correlation between N_B and N_R. So Sharkey need not be concerned, provided his data give a standard deviation of N consistent with a correlation of $+0.8$ between N_R and N_B.

To find the precise relation between $SD(N)$ and $Corr(N_R, N_B)$, for N_R and N_B, now numbers of reds and blacks in a random number N of spins, use the formula

$$Cov(N_R, N_B) = E(N_R N_B) - E(N_R)E(N_B)$$

where each expectation can be computed by conditioning on N. First, if N is treated as a constant, then by previous calculations,

$$E(N_R) = Nr \qquad E(N_B) = Nb$$

$$E(N_R N_B) = E(N_R)E(N_B) + Cov(N_R, N_B) = N^2 rb - Nrb$$

For random N, these are *conditional* expectations given N. But since expectations are expectations of conditional expectations, this gives

$$\begin{aligned}
E(N_R) &= E(N)r, \qquad E(N_B) = E(N)b \\
E(N_R N_B) &= E(N^2)rb - E(N)rb, \qquad \text{hence} \\
Cov(N_R, N_B) &= E(N_R N_B) - E(N_R)E(N_B) \\
&= rb\left[E(N^2) - E(N) - [E(N)]^2\right] \\
&= rb\left[Var(N) - E(N)\right]
\end{aligned}$$

In particular, $Cov(N_R, N_B)$ will be positive provided $Var(N) > E(N)$. Thus for $E(N) = 1000$, if $SD(N) > \sqrt{1000} \approx 32$, there will be a positive correlation between N_R and N_B. The same method of calculation gives

$$Var(N_B) = b^2 Var(N) + b(1-b)E(N)$$

For $b = r$ this gives

$$\begin{aligned} Corr(N_R, N_B) &= \frac{b^2[Var(N) - E(N)]}{b^2 Var(N) + b(1-b)E(N)} \\ &= \frac{9\,Var(N) - 9000}{9\,Var(N) + 10,000} \qquad \text{for} \quad b = \frac{18}{38}, \quad E(N) = 1,000. \end{aligned}$$

If $Var(N) = 0$ this simplifies to -0.9 as before. But as $Var(N)$ increases the correlation increases, and approaches 1 for large values of $Var(N)$. Set $Corr(N_R, N_B) = \rho$ and solve for $SD(N) = \sqrt{Var(N)}$ to get

$$\begin{aligned} SD(N) &= \sqrt{\frac{9000 + 10,000\rho}{9(1-\rho)}} \\ &= \sqrt{\frac{17,000}{9 \times 0.2}} \qquad \text{for } \rho = 0.8 \\ &\approx 100 \end{aligned}$$

So a correlation of 0.8 between N_R and N_B is consistent with a standard deviation of about 100 for the number of spins per day. Provided that is the case, Sharkey need not be concerned.

Discussion. The example makes the important point that two variables, like N_R and N_B, may be positively correlated due to association with some third variable, like N, even if there is zero or negative correlation between the two variables for a fixed value of N. Here is another example. For children of a fixed age, the correlation between height and reading ability would most likely come out around zero. But if you looked at children of ages from 5 to 10, there would be a high positive correlation between height and reading ability, because both variables are closely associated with age. For data variables, looking at distributions or relationships between some variables for a fixed value of another variable, N say, is called *controlling for N*. In a probability model the corresponding thing is *conditioning on N*. Whether or not you condition or control on one variable typically has major effects on relationships between other variables.

The calculations in the example show in general that for two mutually exclusive outcomes in independent trials, like red and black at roulette, the counts of results of the two kinds that occur in any fixed number of trials will be negatively correlated. If the number of trials N is random, the two counts will be positively or negatively correlated according to whether $Var(N) > E(N)$ or $Var(N) < E(N)$. In the case where $Var(N) = E(N)$, the two counts will be uncorrelated. In particular this is the case if N has a Poisson distribution. Then the two counts are actually independent. See Exercise 6.1.8.

Variance of a Sum of *n* Variables

The general formula involving covariance for the variance of a sum of two random variables has the following extension to n variables. The formula shows that the simple addition rule for the variance of a sum of independent random variables works just as well for uncorrelated ones, but in general there are $\binom{n}{2}$ covariance terms to be considered as well.

> **Variance of a Sum of *n* Variables**
>
> $$Var\left(\sum_k X_k\right) = \sum_k Var(X_k) + 2\sum_{j<k} Cov(X_j X_k)$$
>
> where $\sum_k$ denotes a sum of n terms from $k = 1$ to n, and $\sum_{j<k}$ denotes a sum of $\binom{n}{2}$ terms indexed by j and k with $1 \le j < k \le n$.

Proof: The variance of the sum is by definition the expectation of

$$\begin{aligned}\left[\sum_k X_k - E(\sum_k X_k)\right]^2 &= \left[\sum_k X_k - \sum_k \mu_k\right]^2 \qquad \text{where } \mu_k = E(X_k)\\ &= \left[\sum_k (X_k - \mu_k)\right]^2 \\ &= \sum_k (X_k - \mu_k)^2 + 2\sum_{j<k}(X_j - \mu_j)(X_k - \mu_k)\end{aligned}$$

by the algebraic identity

$$\left(\sum a_k\right)^2 = \sum_k a_k^2 + 2\sum_{j<k} a_j a_k$$

applied to $a_k = X_k - \mu_k$. Now use the linearity of expectation and the definition of $Cov(X_j, X_k)$. In the sum over all $j < k$, there are exactly $\binom{n}{2}$ terms, one for each way of choosing two indices $j < k$ from the set $\{1, 2, \ldots, n\}$. □

Example 7. Variance of sample averages.

Let $x(1), x(2), \ldots, x(N)$ be a list of N numbers. Think of $x(k)$ as representing the height of the kth individual in a population of size N. Let

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x(k) \quad \text{and} \quad \sigma^2 = \frac{1}{n}\sum_{k=1}^{n}[x(k) - \bar{x}]^2$$

So $\bar{x}$ is the *population mean*, and σ^2 is the *population variance*. Let $X_1, X_2, \ldots, X_n$ be the heights obtained in a random sample of size n from this population. More formally, for $i = 1, 2, \ldots, n$, the ith height in the sample is $X_i = x(K_i)$, where $K_1, K_2, \ldots, K_n$ is a random sample of size n from the index set $\{1, 2, \ldots, N\}$. This random sample might be taken either with replacement or without replacement. Either way, each random index K_i has uniform distribution over $\{1, 2, \ldots, N\}$, by symmetry. So each X_i is distributed according to the distribution of the list of heights in the total population, with

$$E(X_i) = \bar{x} \quad \text{and} \quad SD(X_i) = \sigma \qquad (i = 1, 2, \ldots, n)$$

Let

$$\bar{X}_n = (X_1 + X_2 + \ldots + X_n)/n$$

be the *sample average*. This is the average height of individuals in the sample of size n. Note that this is a random variable: repeating the sampling procedure will typically produce a different sample average. Whereas $\bar{x}$, the population average, is a constant. Since $E(X_i) = \bar{x}$ for $i = 1, 2, \ldots, n$, the rules of expectation imply that also

$$E(\bar{X}_n) = \bar{x}$$

still no matter whether the sampling is done with or without replacement. In the case with replacement, the random variables X_i are independent, all with standard deviation σ, so

$$SD(\bar{X}_n) = \sigma/\sqrt{n} \qquad \text{(with replacement)}$$

by the square root law of Section 3.3. So the average height in a random sample of size n is most likely only a few multiples of $\sigma/\sqrt{n}$ away from the population average $\bar{x}$. If σ can be bounded or estimated, this gives an indication of the quality of the sample average $\bar{X}_n$ as an estimator of the unknown population average $\bar{x}$.

Intuitively, for sampling without replacement, $\bar{X}_n$ should provide a better estimate of $\bar{x}$ than for sampling with replacement. In this case, the random variables $X_1, \ldots, X_n$ turn out to be negatively correlated, which affects the formula for $SD(\bar{X}_n)$. The problem is how to correct for the dependence.

Problem. Calculate $SD(\bar{X}_n)$ for sampling without replacement.

Solution. Let $S_n = X_1 + \cdots + X_n$, so $\bar{X}_n = S_n/n$. Then

$$\begin{aligned} Var(S_n) &= \sum_j Var(X_j) + 2 \sum_{j<k} Cov(X_j, X_k) \\ &= n\sigma^2 + n(n-1)\, Cov(X_1, X_2), \end{aligned}$$

because $Cov(X_j, X_k) = Cov(X_1, X_2)$ by the symmetry of sampling without replacement discussed in Section 3.6: (X_j, X_k) is for every $j < k$ a simple random sample of size 2, with the same distribution as (X_1, X_2). This formula for $Var(S_n)$ holds for every sample size n with $1 \le n \le N$. But for $n = N$

$$S_N = x_1 + x_2 + \cdots + x_N$$

is constant, because in a complete sample of the population each element appears exactly once, so the sum defining S_N is just the sum on the right done in a random order. Thus $Var(S_N) = 0$. Comparison with the previous formula for $Var(S_n)$, in the case where $n = N$, shows

$$Cov(X_1, X_2) = -\sigma^2/(N-1)$$

hence

$$Var(S_n) = n\sigma^2 \left[1 - \frac{n-1}{N-1}\right]$$

and

$$SD(\bar{X}_n) = SD(S_n)/n = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Discussion. This shows that the standard deviation for the average in sampling without replacement is the corresponding standard deviation for sampling with replacement, reduced by the *correction factor* $\sqrt{\frac{N-n}{N-1}}$. The same is true for the sum as well as the average, by scaling.

The same correction factor appears in the formula for the variance of the hypergeometric distribution, calculated in Section 3.6. Though covariances are not used in that calculation, it is still a special case of the current example, with $x_j = 0$ or 1 for every j.

It is remarkable that the same correction factor works no matter what the distribution of the empirical variable x. The correction factor takes care of the slight negative correlation between terms, which also does not depend on the distribution of x:

$$Corr(X_j, X_k) = \frac{Cov(X_j, X_k)}{SD(X_j)SD(X_k)} = -1/(N-1)$$

The correlation is negative because observation of a large value of X_j removes a large value from the population, and tends to make large values of X_k less likely. Similarly, small values of X_j tend to make small values of X_k less likely. This means there is a greater tendency for the deviations $X_j - E(X_j)$ to cancel each other out for sampling without replacement than for sampling with replacement, when these deviations are independent. This reduces the likely size of the deviation for the sum

$$S_n - E(S_n) = \sum_{j=1}^{n} (X_j - E(X_j))$$

Ultimately, for $n = N$, the deviation of S_N is zero, which was the key to calculating the correction factor.

Bilinearity of Covariance

The following formulae for covariances of linear combinations of variables are easily derived from the definition. These formulae can often be used to simplify covariance calculations.

$$Cov(X, Y + Z) = Cov(X, Y) + Cov(X, Z)$$

$$Cov(W + X, Y) = Cov(W, Y) + Cov(X, Y)$$

For constants a and b

$$Cov(aX, Y) = a\,Cov(X, Y) \quad \text{and} \quad Cov(X, bY) = b\,Cov(X, Y)$$

and so on for linear combinations of several variables. For example

$$Cov(aW+bX,\ cY+dZ) = a\,c\,Cov(W, Y) + a\,d\,Cov(W, Z) + b\,c\,Cov(X, Y) + b\,d\,Cov(X, Z)$$

To summarize:

Covariance is Bilinear

$$Cov\left(\sum_i a_i X_i\,,\ \sum_j b_j Y_j\right) = \sum_i \sum_j a_i\, b_j\, Cov(X_i, Y_j)$$

Here the a_i and b_j are arbitrary constants. If there are n terms in the sum over i and m terms in the sum over j there are nm terms in the double sum on the right side. Taking $n = m$, $a_i = b_i = 1$ and $X_i = Y_i$ for $1 \le i \le n$, this formula reduces to the formula for the variance of $\sum_i X_i$.

Exercises 6.4

1. Suppose A, B are two events such that $P(A) = 0.3$, $P(B) = 0.4$, and $P(A \cup B) = 0.5$.

a) Find $P(A|B)$. b) Are A and B independent, positively or negatively dependent?

c) Find $P(A^c B)$. d) Let $X = I_A$, $Y = I_B$. Find $Corr(X, Y)$.

2. Use the formula $P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$ to prove:

a) if $P(A|B) = P(A|B^c)$ then A and B are independent;

b) if $P(A|B) > P(A|B^c)$ then A and B are positively dependent;

c) if $P(A|B) < P(A|B^c)$ then A and B are negatively dependent.

Now prove the converses of a), b), and c).

3. Suppose that the failures of two components are positively dependent. If the first component fails, does that make it more or less likely that the second component works? What if the first component works?

4. Let (X, Y) have uniform distribution on the four points $(-1, 0), (0, 1), (0, -1), (1, 0)$. Show that X and Y are uncorrelated but not independent.

5. Let X have uniform distribution on $\{-1, 0, 1\}$ and let $Y = X^2$. Are X and Y uncorrelated? Are X and Y independent? Explain carefully.

6. Let X_1 and X_2 be the numbers on two independent fair die rolls, $X = X_1 - X_2$ and $Y = X_1 + X_2$. Show that X and Y are uncorrelated, but not independent.

7. Let X_2 and X_3 be indicators of independent events with probabilities $1/2$ and $1/3$, respectively.

a) Display the joint distribution table of $X_2 + X_3$ and $X_2 - X_3$.

b) Calculate $E(X_2 - X_3)^3$.

c) Are X_2 and X_3 uncorrelated? Prove your answer.

8. You have N boxes labeled Box1, Box2, $\ldots$, BoxN, and you have k balls. You drop the balls at random into the boxes, independently of each other. For each ball the probability that it will land in a particular box is the same for all boxes, namely $1/N$. Let X_1 be the number of balls in Box1 and X_N be the number of balls in BoxN. Calculate $Corr(X_1, X_N)$.

9. Suppose n cards numbered $1, 2, \ldots, n$ are shuffled and k of the cards are dealt. Let S_k be the sum of the numbers on the k cards dealt. Find formulae in terms of n and k for:

a) the mean of S_k; b) the variance of S_k.

10. **Overlapping counts.** A fair coin is tossed 300 times. Let H_{100} be the number of heads in the first 100 tosses, and H_{300} the total number of heads in the 300 tosses. Find $Corr(H_{100}, H_{300})$.

11. Let T_1 and T_3 be the times of the first and third arrivals in a Poisson process with rate λ. Find $Corr(T_1, T_3)$.

12. Suppose α, β, γ denote the proportions of Democrats (D), Republicans (R) and Others (O) in a large population of voters. (So $0 \le \alpha, \beta, \gamma \le 1$ and $\alpha+\beta+\gamma = 1$.) An individual is selected at random from the population. Write $X = 1, Y = 0, Z = 0$ if that individual is D, write $X = 0, Y = 1, Z = 0$ if the individual is R and write $X = 0, Y = 0, Z = 1$ if the individual is O. Find:

a) $E(X)$, $E(Y)$; b) $Var(X)$, $Var(Y)$; c) $Cov(X, Y)$.

Suppose next that n individuals are selected independently and randomly with replacement from the population. The total number of D's may be written, $D_n = X_1 + ... + X_n$. Similarly let $R_n = Y_1 + ... + Y_n$. and let $O_n = Z_1 + ... + Z_n$. Let $D_n - R_n$ denote the excess of D's over R's selected. Find d) $E(D_n - R_n)$; e) $Var(D_n - R_n)$.

13. Let A and B be two possible results of a trial, not necessarily mutually exclusive. Let N_A be the number of times A occurs in n independent trials, N_B the number of times B occurs in the same n trials. True or false and explain: If N_A and N_B are uncorrelated, then they are independent.

14. Show that for any two random variables X and Y

$$|SD(X) - SD(Y)| \le SD(X + Y) \le |SD(X) + SD(Y)|$$

15. **Covariance is bilinear.** Show from the definition of covariance that:

a) $Cov(X, Y + Z) = Cov(X, Y) + Cov(X, Z)$

b) $Cov(W + X, Y) = Cov(W, Y) + Cov(X, Y)$

c) $Cov(\sum_i X_i, \sum_j Y_j) = \sum_i \sum_j Cov(X_i, Y_j)$

d) Use c) to rederive the formula for $Cov(N_R, N_B)$ in Example 6.

16. **Invariance of the correlation coefficient under linear transformations.** Show that for arbitrary random variables X and Y, and constants a, b, c, d with $a \ne 0$, $c \ne 0$,

$$Corr(aX + b, cY + d) = \begin{cases} Corr(X, Y) & \text{if } a \text{ and } c \text{ have the same sign} \\ -Corr(X, Y) & \text{if } a \text{ and } c \text{ have opposite signs.} \end{cases}$$

Thus the correlation coefficients are affected only by the sign of a linear change of variable. They are therefore unaffected by shifts of origin or changes of units.

17. Show that for indicator random variables I_A and I_B of events A and B

$$Corr(I_A, I_B) = Corr(I_{A^c}, I_{B^c}) = -Corr(I_A, I_{B^c}) = -Corr(I_{A^c}, I_B)$$

Deduce that if A and B are positively dependent, then so are A^c and B^c, but A and B^c are negatively dependent, as are A^c and B.

18. Random variables $X_1, \ldots, X_n$ are *exchangeable* if their joint distribution is the same, no matter what order they are presented (see Section 3.6). Show that if $X_1, \ldots, X_n$ are exchangeable, then

$$Var(\sum_{k=1}^{n} X_k) = n\,Var(X_1) + n(n-1)\,Cov(X_1, X_2)$$

19. A box contains 5 nickels, 10 dimes, and 25 quarters. Suppose 20 draws are made at random without replacement from this box. Let X be the total sum obtained in these 20 draws. Calculate: a) $E(X)$; b) $SD(X)$;

c) $P(X \leq \$3)$ using the normal approximation.

d) Can you imagine why these calculations might give results inconsistent with long-run repetitions of the sampling experiment? For each of a) and c), say whether your reasoning would suggest higher or lower long-run averages.

20. Correlation and conditioning. A random variable X assumes values x_1 and x_2 with probabilities p_1 and p_2, where $p_1 + p_2 = 1$. Given $X = x_i$, random variable Y has mean equal to μ_i and SD equal to σ_i. Find formulae in terms of x_i, p_i, μ_i, and σ_i, $i = 1, 2$, for the following quantities:

a) $E(X)$; b) $E(Y)$; c) $SD(X)$; d) $SD(Y)$; e) $Cov(X, Y)$; f) $Corr(X, Y)$.

Indicate how these formulae could be generalized to the case of X with n possible values $x_1, \ldots, x_n$.

21. A box contains 5 red balls and 8 blue ones. A random sample of size 3 is drawn *without* replacement. Let X be the number of red balls and let Y be the number of blue balls selected. Compute: a) $E(X)$; b) $E(Y)$; c) $Var(X)$; d) $Cov(X, Y)$.

22. Suppose there were m married couples, but that d of these $2m$ people have died. Regard the d deaths as striking the $2m$ people at random. Let X be the number of surviving couples. Find:

a) $E(X)$; b) $Var(X)$.

23. Linear prediction and the correlation coefficient. For random variables X and Y, the *linear prediction problem* for predicting Y based on knowledge of X is the problem of finding a linear function of X, $\beta X + \gamma$, which minimizes the *mean square* of the prediction error

$$MSE = E[Y - (\beta X + \gamma)]^2$$

(Compare with Exercise 6.2.17 where the predictor of Y could be an arbitrary function of X.) This exercise derives the basic formulae for the best linear predictor according to this criterion.

a) Expand out the MSE using algebra, and regard it as a quadratic function of γ and β with coefficients involving the numbers $E(X)$, $E(Y)$, $E(XY)$, etc.

b) Differentiate this function with respect to γ to show that for fixed β, the unique γ which minimizes the MSE is $\hat{\gamma}(\beta) = E(Y) - \beta E(X)$. What is the resulting minimal MSE called when $\beta = 0$?

c) Consider now the MSE as a function of β, with $\gamma = \hat{\gamma}(\beta)$ the best γ for the given β. Differentiate this function with respect to β, and show that it is minimized at $\hat{\beta} = Cov(X, Y)/Var(X)$ where it is assumed that $Var(X) > 0$.

d) Deduce that the unique pair (β, γ) which minimizes the MSE is $(\hat{\beta}, \hat{\gamma}(\hat{\beta}))$.

e) Let $\hat{Y} = \hat{\beta} X + \hat{\gamma}$ now denote this best linear predictor. Show that

$$E(\hat{Y}) = E(Y); \quad Var(\hat{Y}) = \hat{\beta}^2 Var(X); \quad E[\hat{Y}(Y - \hat{Y})] = 0$$

f) Deduce that the variance of Y can be decomposed into the sum of the variance of the best predictor $\hat{Y}$ and the minimum MSE according to the formula

$$Var(Y) = Var(\hat{Y}) + E[(Y - \hat{Y})^2]$$

with $Var(\hat{Y}) = \rho^2 Var(Y)$ and $E[(Y - \hat{Y})^2] = (1 - \rho^2) Var(Y)$ where $\rho = Corr(X, Y)$.

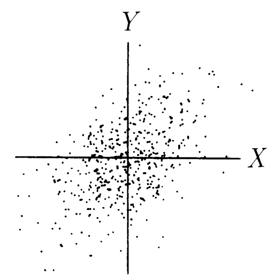
g) It is customary to express the slope $\hat{\beta}$ of the best linear predictor $\hat{Y} = \hat{\beta}X + \hat{\gamma}$ in terms of ρ. Show that $\hat{\beta} = \rho SD(Y)/SD(X)$ and that the intercept $\hat{\gamma}$ is then uniquely determined by the requirement that the line $y = \hat{\beta}x + \hat{\gamma}$ passes through the point $(E(X), E(Y))$.

h) Let $Y^* = (Y - E(Y))/SD(Y)$, $X^* = (X - E(X))/SD(X)$. Show that the best linear predictor of Y^* based on X^* is just ρX^*. So the correlation coefficient ρ is simply the slope of the best linear predictor when the variables are expressed in standard units.

6.5 Bivariate Normal

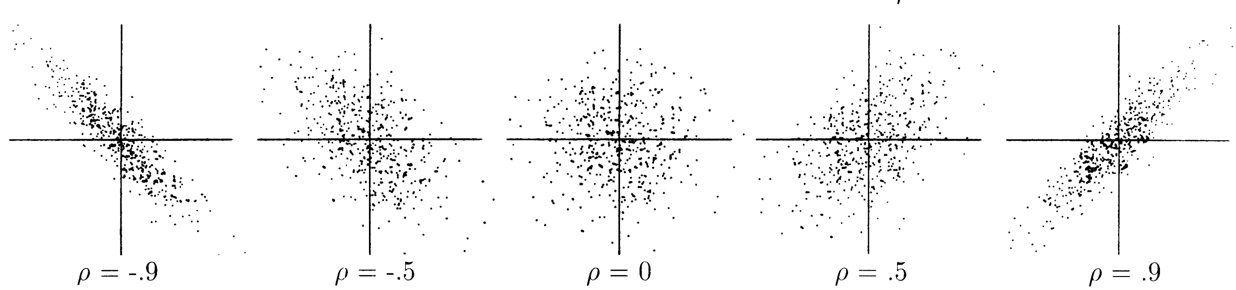
The radially symmetric bivariate normal distribution corresponding to independent normal variables was considered in Section 5.3. This section uses the tools of previous sections to analyze correlated normal variables by making a linear transformation to the simpler case of independent variables.

FIGURE 1. Bivariate normal scatter.
The diagram shows points picked at random from a bivariate distribution, in which the coordinates X and Y each have the same normal distribution, but are not independent. The two variables are *positively correlated*, which makes the cloud elliptical, sloping upward to the right and downwards to the left.



Clouds of data like this are very common in statistical analysis. They were first examined by the British scientist Francis Galton (1822–1911), who studied relations between variables like a father's height and his son's height. To display visually how two variables are related, a *scatter diagram* like Figure 1 may be used. In such a diagram, data pairs are represented by plotting a point at the coordinates of each pair. The hereditary connection between a father's height and his son's height makes the variables positively correlated—taller fathers tend to have taller sons, taller sons tend to have taller fathers. But the relation is not a rigid one, since the son's height is not a deterministic function of his father's height. The dependence between the two variables is more interesting and subtle. When variables are measured in their standard units, this dependence shows up in a scatter diagram as a tendency to form an elliptical cloud along a diagonal. The cloud has a major axis along the line $Y = X$ at 45° to the axes in the case of positive correlation, and a major axis along the perpendicular line $Y = -X$ in the case of negative correlation.

FIGURE 2. Bivariate normal scatters for various correlations ρ.



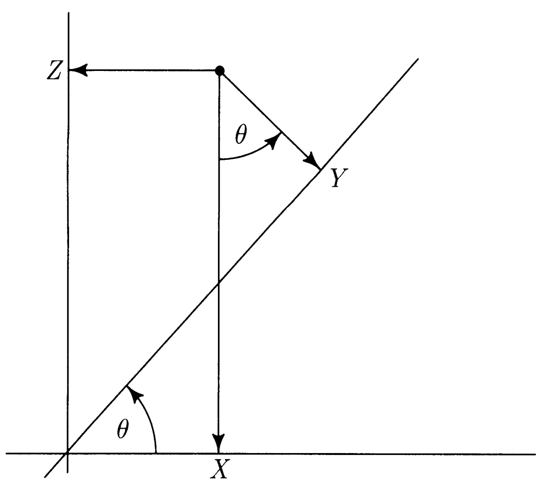
The object now is to describe this kind of dependence between variables by representing correlated normal variables as linear functions of independent ones. This is a powerful technique which is the basis for much statistical analysis of two or more variables. A basic ingredient is the correlation coefficient, denoted here by ρ, often also by r:

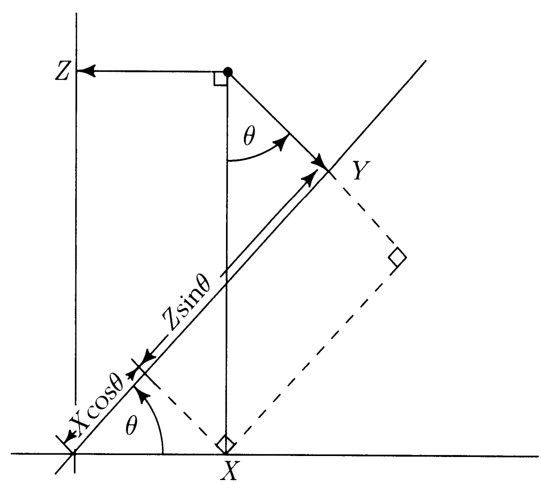
$$\rho = Corr(X,Y) = E(X^*Y^*)$$

where X^* is X in standard units, and Y^* is Y in standard units. This correlation ρ is a theoretical quantity, defined by expected values or integrals with respect to a bivariate distribution. In practice, such correlations are usually estimated by the corresponding empirical correlation obtained from data, with the empirical distribution of a data list $(x_1, y_1), \ldots, (x_n, y_n)$ instead of the theoretical distribution, and averages instead of expectations.

Constructing Correlated Normal Variables

To get a pair of correlated standard normal variables X and Y, start with a pair of independent standard normal variables, say X and Z. Let Y be the projection of (X,Z) onto an axis at an angle θ to the X-axis, as in the left-hand diagram:





By the geometry of the right-hand diagram

$$Y = X\cos\theta + Z\sin\theta$$

By rotational symmetry of the joint distribution of X and Z, the distribution of Y is standard normal. Thus

$$\begin{aligned} E(X) &= E(Y) = E(Z) = 0 \\ SD(X) &= SD(Y) = SD(Z) = 1 \\ \rho(X,Y) = E(XY) &= E[X(X\cos\theta + Z\sin\theta)] \\ &= E(X^2)\cos\theta + E(XZ)\sin\theta \\ &= \cos\theta \end{aligned}$$

since $E(X^2) = 1$, and $E(XZ) = E(X)E(Z) = 0$ by independence of X and Z. To summarize, X and Y are standard normal variables with correlation $\rho = \cos\theta$. Note the special cases

$$\begin{array}{lll} \theta = 0 & \text{when } \rho = 1 & Y = X \\ \theta = \pi/2 & \text{when } \rho = 0 & Y = Z \text{ is independent of } X \\ \theta = \pi & \text{when } \rho = -1 & Y = -X \end{array}$$

For each ρ between -1 and 1, there is an angle $\theta = \arccos\rho$, which makes X and Y have correlation ρ. Then $\cos\theta = \rho$, $\sin\theta = \sqrt{1-\rho^2}$, and

$$Y = \rho X + \sqrt{1-\rho^2}Z$$

where X and Z are independent normal (0, 1). The joint distribution of X and Y so defined is the *standard bivariate normal distribution with correlation ρ*.

Standard Bivariate Normal Distribution

X and Y have standard bivariate normal distribution with correlation ρ if and only if

$$Y = \rho X + \sqrt{1-\rho^2}Z$$

where X and Z are independent standard normal variables.

Marginals. Both X and Y have standard normal distribution.

Conditionals. Given $X = x$, Y has normal $(\rho x, 1-\rho^2)$ distribution. Given $Y = y$, X has normal $(\rho y, 1-\rho^2)$ distribution.

Joint density. The joint density of X and Y is

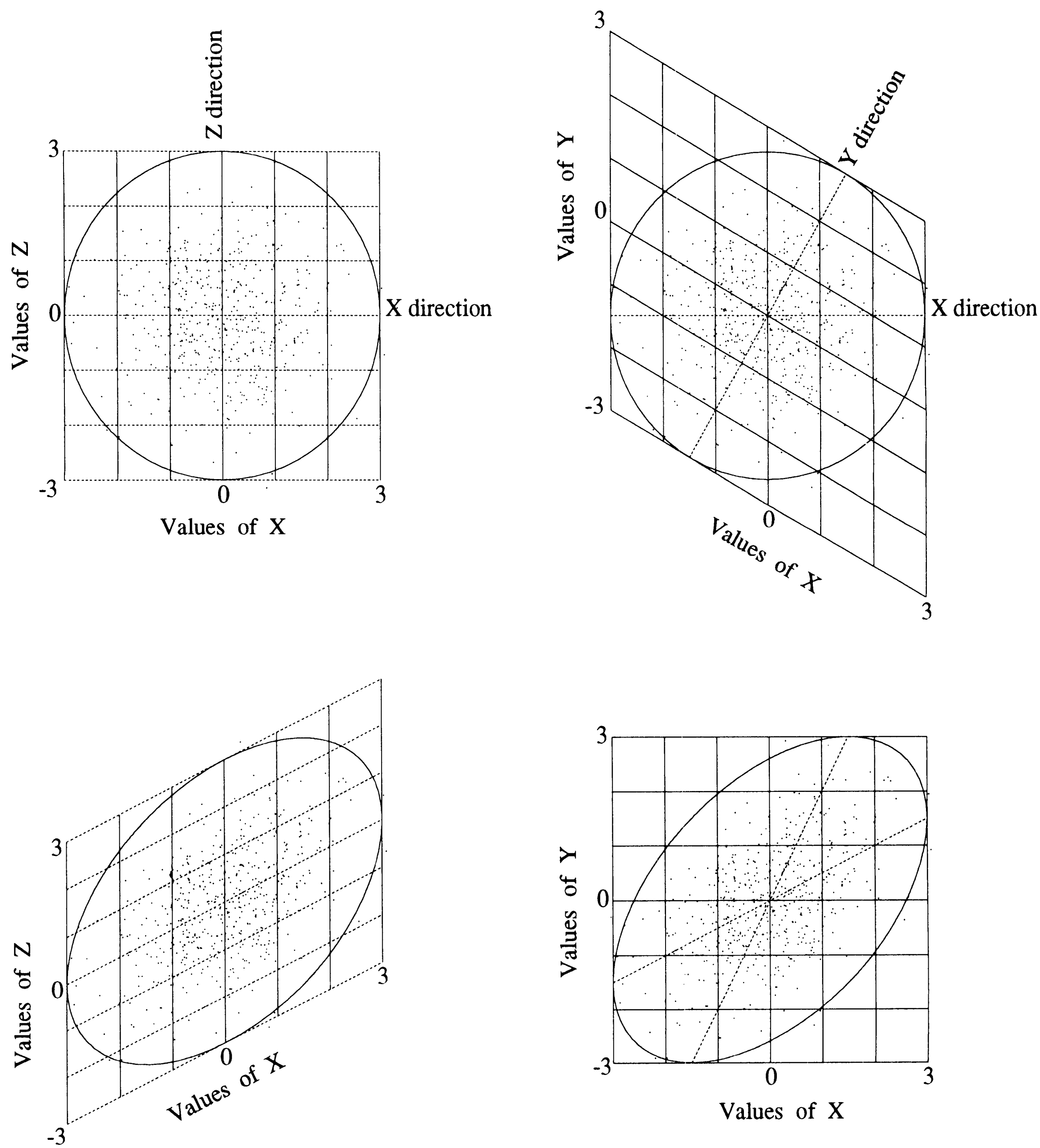
$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right\}$$

Independence. For X and Y with standard bivariate normal distribution, X and Y are independent if and only if $\rho = 0$.

The next two pages display the geomerty of linear transformation from (X, Z) to (X, Y). Following these pages is a discussion of the results presented in the above box.

FIGURE 3. **Geometry of the bivariate normal distribution.** Properties of the standard bivariate normal distribution with correlation ρ may be understood in terms of the simplest case $\rho = 0$ by the geometry of the linear transformation $(X, Z) \mapsto (X, Y)$, displayed here for $\theta = 60°$, so

$$\rho = \cos\theta = \frac{1}{2}, \qquad \sqrt{1-\rho^2} = \sin\theta = \frac{\sqrt{3}}{2} \quad \text{and} \quad Y = \frac{1}{2}X + \frac{\sqrt{3}}{2}Z.$$



Key to Figure 3.

Top left panel. This shows a computer-generated scatter of 500 points picked at random according to the joint distribution of X and Z, plotted in the usual way with rectangular X and Z coordinates. This is a roughly circular cloud, due to the rotational symmetry of the distribution of two independent standard normals. The circle is the contour of constant density for (X, Z), of radius 3 standard units, containing 98.9% of the probability. The vertical lines represent the events $X = 0, \pm 1, \pm 2, \pm 3$. The dashed horizontal lines represent $Z = 0, \pm 1, \pm 2, \pm 3$.

Top right panel. This is the same scatter in the (X, Z) plane, but with the diagonal lines $Y = 0, \pm 1, \pm 2, \pm 3$. The Y direction is the dotted line at angle $\theta = 60°$ to the horizontal X direction. The diagonals $Y = \mathit{constant}$ are at angle θ to the vertical lines $X = \mathit{constant}$.

Bottom right panel. This is the image of the top right panel after shearing and shrinking to represent X and Y by new rectangular axes. Each point in the top scatter is transformed into one in the bottom scatter. Thus the cloud becomes a random scatter of 500 points picked at random according to the bivariate normal distribution of X and Y, with correlation $\rho = \cos\theta$. Think of the lines in the top right panel as a lattice of rigid rods attached by pins. Keep the vertical axis $X = 0$ fixed, and shear the lattice so the diagonals become horizontal. This makes a lattice of squares of side $1/\sin\theta$. Now shrink everything by a factor of $\sin\theta$ to get the bottom-right panel.

The shearing which turns the diamonds into squares turns the circle into an ellipse, with major axis on the 45-degree line through the new origin. This is an ellipse of constant density for (X, Y). The images of the dotted lines in the old X and Y directions are the dotted lines $Y = \rho X$ and $X = \rho Y$. These are the *regression lines* discussed further in the next paragraph.

Bottom left panel. This is the image of the top left panel by the same transformation from (X, Z) to (X, Y). The ellipse and the cloud of points are the same as in the bottom right panel. But now the lines representing $X = 0, \pm 1, \pm 2, \pm 3$ are shown, along with those representing $Z = 0, \pm 1, \pm 2, \pm 3$. The line $Z = 0$ plays a particularly important role. This is the *regression line*. The equation of this line $Z = 0$ in the (X, Y) plane is

$$Y = \rho X$$

where ρ is the correlation. Geometrically, this is the line of midpoints of vertical sections of the ellipse. Statistically, it is the best predictor of Y based on X.

The properties of the standard bivariate normal distribution stated in the box on page 451 all follow from the basic representation

$$Y = \rho X + \sqrt{1-\rho^2}\, Z \tag{1}$$

in terms of independent standard normal X and Z.

Conditionals. The formula for the distribution of Y given $X = x$ is immediate from (1). Conditioning on X does not affect the distribution of Z. And given $X = x$ you can treat X in (1) as the constant x, so Y is then just a linear transformation of the standard normal variable Z with coefficients involving ρ and x. This gives the conditional distribution of Y given $X = x$. The distribution of X given $Y = y$ follows by symmetry, or from (1′) below.

Symmetry. The standard bivariate normal distribution of (X, Y) is symmetric with respect to switching X and Y. This can be seen from the formula for the joint density, which is a symmetric function of x and y, or from the geometric description of X and Y. This symmetry is obscured in formula (1) however. You should check as an exercise that (1) has a dual

$$X = \rho Y + \sqrt{1-\rho^2}\, Z' \tag{1'}$$

where Z' is a linear combination of X and Z that is independent of Y.

Joint density. The derivation of this is an exercise: Write out the formulae for the marginal and conditional densities, multiply, and simplify. There is no point remembering this formula. Rather, take the following:

Advice. Do not attempt to compute bivariate normal probabilities or expectations by integrating against the joint density. It is always simpler to rewrite the problem in terms of independent variables X and Z, using (1). This technique is used in all the examples below.

Bivariate Normal Distribution

Random variables U and V have *bivariate normal distribution* with parameters μ_U, μ_V, σ_U^2, σ_V^2, and ρ if and only if the standardized variables

$$X = (U - \mu_U)/\sigma_U \qquad Y = (V - \mu_V)/\sigma_V$$

have standard bivariate normal distribution with correlation ρ. Then

$$\rho = Corr(X, Y) = Corr(U, V)$$

and U and V are independent if and only if $\rho = 0$.

Examples

The point of the following examples is to show how any problem involving random variables U and V with a bivariate normal distribution can be solved by a simple three-step procedure:

- **Step 1.** Express U and V in terms of the standardized variables X and Y.
- **Step 2.** Write $Y = \rho X + \sqrt{1-\rho^2}Z$ to reduce the problem to one involving two independent standard normal variables X and Z.
- **Step 3.** Solve the reduced problem involving X and Z by exploiting independence or rotational symmetry.

Example 1. Fathers and sons.

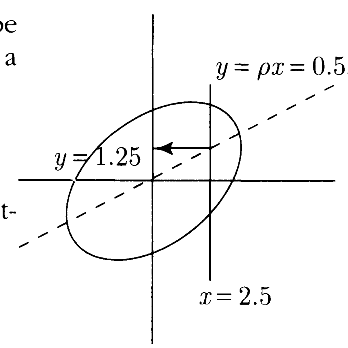
Galton's student Karl Pearson carried out a study on the resemblances between parents and children. He measured the heights of 1078 fathers and sons, and found that the sons averaged one inch taller than the fathers:

Fathers:	mean height: 5′9″	SD: 2″
Sons:	mean height: 5′10″	SD: 2″
	correlation: 0.5	

Problem 1. Predict the height of the son of a father who is 6′2″ tall.

Solution. Assume that the data are approximately bivariate normal in distribution. Then the parameters can be estimated by the corresponding empirical measurements.

Let X be the father's height in standard units, and Y be the son's height in standard units. The assumption of a bivariate normal distribution makes



$$Y = \rho X + \sqrt{1-\rho^2}Z$$

where Z is standard normal independent of X. The natural prediction for Y given $X = x$ is

$$E(Y|X=x) = \rho x$$

Here the given value of X is

$$\begin{aligned} x &= 6'2'' \text{ converted to standard units} \\ &= (6'2'' - 5'9'')/2'' = 2.5 \text{ standard units} \end{aligned}$$

So the predicted value of Y is

$$E(Y|X=x) = 0.5 \times 2.5 = 1.25 \text{ standard units,}$$

That is,

$$\begin{aligned}\text{predicted son's height} &= 5'10'' + 2''Y \\ &= 5'10'' + 2'' \times 1.25 = 6'\,0.5''\end{aligned}$$

Discussion. Though the father is exceptionally tall (height $6'2''$), the son is not predicted to be $6'2''$, but only $6'0.5''$ tall. Galton called this phenomenon *regression to the mean.*

Problem 2. What is the chance that your prediction is off by more than 1 inch?

Solution. Since 1 inch is 0.5 times the SD of sons' heights, and we are given $X = 2.5$, the problem in standard units is to find

$$P(|Y - \rho X| > 0.5 \,|\, X = 2.5).$$

But since $Y - \rho X = \sqrt{1-\rho^2}Z$ is independent of X with normal $(0, 1-\rho^2)$ distribution, where

$$\sqrt{1-\rho^2} = \sqrt{0.75} \approx 0.87,$$

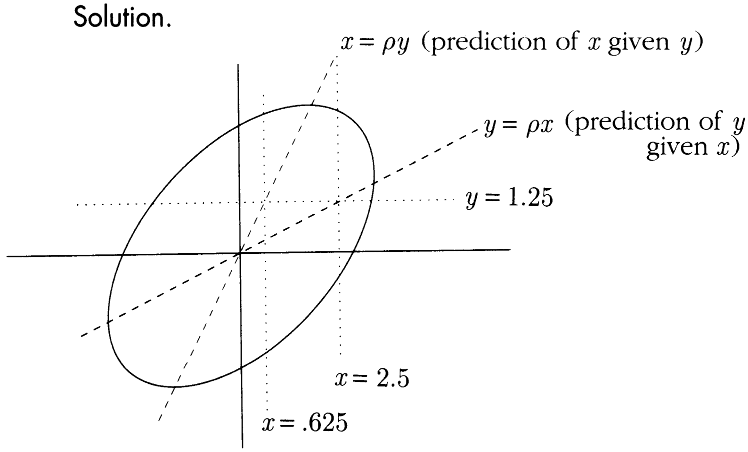
this is the same as

$$\begin{aligned}P(|Y - \rho X| > 0.5) &= P(\sqrt{1-\rho^2}|Z| > 0.5) \\ &= P(|Z| > 0.5/\sqrt{1-\rho^2}) \\ &= P(|Z| > 0.5/0.87) \\ &= 2[1 - \Phi(0.5/0.87)] \approx 2[1 - \Phi(0.57)] \approx 0.57\end{aligned}$$

So with about 57% chance, the prediction will be off by more than an inch.

Problem 3. Estimate the height of a father whose son is $6'0.5''$ tall.

Solution.



From above, $6'0.5''$ is the mean height of sons of $6'2''$ fathers. So you might guess that $6'2''$ was the mean height of fathers of $6'0.5''$ sons. But this is wrong, because a given father's height corresponds to a vertical slice through the scatter, whereas a given son's height corresponds to a horizontal slice, which is something quite different. See diagrams. The roles of X and Y must simply be switched in the calculation of Problem 1. The son's height of $6'0.5''$ is 1.25 in standard units. So

$$\begin{aligned}\text{estimated father's height} &= 0.5 \times 1.25 \quad \text{in standard units} \\ &= 0.625 \text{ in standard units} \\ &= 5'9'' + 0.625 \times 2'' = 5'10.25''\end{aligned}$$

Example 2. The probability that both variables are above average.

Problem 1. For the data in Example 1, what fraction of father–son pairs have both father and son of above average height?

Solution. Expressed in terms of the standardized variables X and Y, the problem is to find $P(X \geq 0, Y \geq 0)$. In principle, the answer can be computed as a double integral

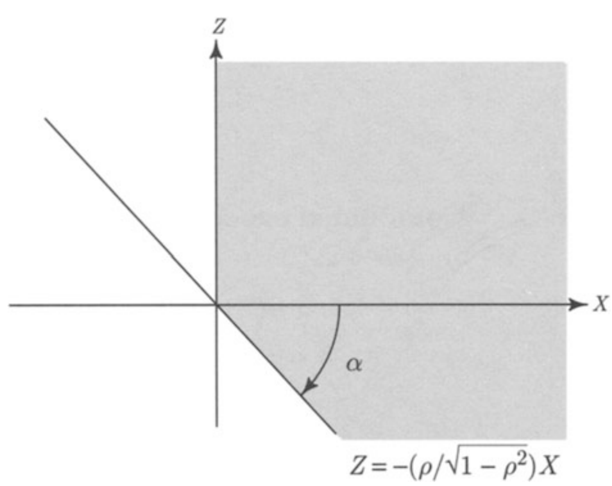
$$\iint_{\text{positive quadrant}} f(x,y)\,dx\,dy$$

where $f(x,y)$ is the standard bivariate normal density with $\rho = 0.5$. But, as usual, it is easier to first express X and Y in terms of independent standard normal variables X and Z:

$$Y = \rho X + \sqrt{1-\rho^2}Z$$

Now the problem is to find

$$\begin{aligned} P(X \geq 0, Y \geq 0) &= P(X \geq 0, \rho X + \sqrt{1-\rho^2}Z \geq 0) \\ &= P\left(X \geq 0, Z \geq \frac{-\rho}{\sqrt{1-\rho^2}}X\right) \end{aligned}$$



The diagram shows the (X,Z) plane, with the line $Z = -\rho/\sqrt{1-\rho^2}X$. The shaded region corresponds to the event above. The slope of the line is $-\rho/\sqrt{1-\rho^2}$.

So for α as in the diagram, considered a negative angle,

$$\tan\alpha = \frac{-\rho}{\sqrt{1-\rho^2}} = \frac{-0.5}{\sqrt{0.75}} = -1/\sqrt{3}$$

So $\alpha = -30°$. Thus the angle at the corner of the shaded region is $-\alpha + 90° = 120°$. By rotational symmetry, the chance that (X, Z) lies in the shaded region is the ratio of angles $120°/360° = 1/3$. So

$$P(X \geq 0, Y \geq 0) = 1/3$$

In other words, about one-third of the father–son pairs had both father and son above average height.

Problem 2. Suppose you have data on two variables with a bivariate normal distribution, and 3/8 of the data is above average in both variables. Estimate ρ.

Solution. Transform to standard units and use the same linear change of variable as in the solution of the previous problem. Now

$$\frac{3}{8} = \frac{135°}{360°}$$

so the angle of the corner at the origin is $135°$. Thus α in the diagram is $-45°$, and by the previous solution

$$\frac{-\rho}{\sqrt{1-\rho^2}} = \tan\alpha = \tan(-45°) = -1$$

So $\rho = 1/\sqrt{2}$.

Example 3. Conditional expectation of Y given X in an interval.

Suppose (X, Y) has standard bivariate normal density with correlation ρ.

Problem. For $a < b$, find $E(Y \mid a < X < b)$.

Solution. Given that X has a particular value $x \in (a, b)$, the expected value of Y is

$$E(Y \mid X = x) = \rho x.$$

Given just $(a < X < b)$ the precise value of X is unknown. But by the rule of average conditional expectations, $E(Y \mid a < X < b)$ can be found by integration of the conditional expectation $E(Y \mid X = x) = \rho x$ with respect to the conditional density of X given $a < X < b$. This gives

$$E(Y \mid a < X < b) = \int_a^b \rho x f_X(x \mid a < X < b)\, dx$$

where for $a < x < b$

$$\begin{aligned} f_X(x \,|\, a < X < b)\, dx &= P(X \in dx \,|\, a < X < b) \\ &= \frac{P(X \in dx, a < X < b)}{P(a < X < b)} \\ &= \frac{P(X \in dx)}{P(a < X < b)} \\ &= \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}}{\Phi(b) - \Phi(a)}\, dx \end{aligned}$$

Substituting this expression gives

$$E(Y \,|\, a < X < b) = \int_a^b \rho x \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}}{\Phi(b) - \Phi(a)}\, dx = \frac{\frac{\rho}{\sqrt{2\pi}} \left[e^{-\frac{1}{2}a^2} - e^{-\frac{1}{2}b^2} \right]}{\Phi(b) - \Phi(a)}$$

Example 4. Midterm and final.

Midterm and final scores in a large class have an approximately bivariate normal distribution, with parameters

midterm scores:	mean: 65	SD: 18
final scores:	mean: 60	SD: 20
	correlation: 0.75	

Problem. Estimate the average final score of students who were above average on the midterm.

Solution. Let X and Y denote the midterm and final scores in standard units. The event "midterm score above average" is the same as the event $X > 0$. Take $a = 0$ and $b = \infty$ in the previous example to get

$$E(Y | X > 0) = \frac{\rho}{\sqrt{2\pi}} \left[\frac{1 - 0}{0.5} \right] = \frac{0.75 \times 2}{\sqrt{2\pi}} \approx 0.6$$

So the average final score of those who scored above average on the midterm is 0.6, in standard units. Thus the required score is

$$60 + 20 \times 0.6 = 72$$

Linear Combinations of Several Independent normal variables

The standard bivariate normal distribution was defined as the joint distribution of a particular pair of linear combinations of independent standard normal variables X and Z, namely, X and $\rho X + \sqrt{1-\rho^2}Z$. While this representation seems at first artificial, the examples show how it is the basis of all calculations involving the more general bivariate normal distribution, which is obtained by allowing arbitrary means and variances, but insisting that the two standardized variables are standard bivariate normal.

The rotational symmetry of the joint distribution of two independent standard normal variables Z_1 and Z_2 implies that the joint distribution of any two linear combinations of Z_1 and Z_2, say

$$V = a_1 Z_1 + a_2 Z_2 \qquad \text{and} \qquad W = a_1 Z_1 + a_2 Z_2$$

is bivariate normal. By reducing to this case by scaling, the same conclusion is obtained for any two independent normal variables Z_1 and Z_2 (not necessarily standard). It can be shown that this extends to linear combinations of any number of independent normal variables Z_i:

Two Linear Combinations of Independent Normal Variables

Let

$$V = \sum_i a_i Z_i \qquad \text{and} \qquad W = \sum_i b_i Z_i$$

be two linear combinations of independent normal (μ_i, σ_i^2) variables Z_i. Then the joint distribution of V and W is bivariate normal.

Granted this, the parameters of the bivariate normal distribution of V and W are easily computed:

$$\mu_V = \sum_i a_i \mu_i \quad \text{and} \quad \mu_W = \sum_i b_i \mu_i$$

$$\sigma_V^2 = \sum_i a_i^2 \sigma_i^2 \quad \text{and} \quad \sigma_W^2 = \sum_i b_i^2 \sigma_i^2$$

$$Cov(V, W) = \sum_i a_i b_i \sigma_i^2$$

$$\rho = Cov(V, W)/\sigma_V \sigma_W$$

Thus the bivariate normal distribution adequately describes the dependence between any two linear combinations of independent normal variables. In particular, this discussion implies the following result:

Independence of Linear Combinations

Two linear combinations $V = \sum_i a_i Z_i$ and $W = \sum_i b_i Z_i$ of independent normal(μ_i, σ_i^2) variables Z_i are independent if and only if they are uncorrelated, that is, if and only if $\sum_i a_i b_i {\sigma_i}^2 = 0$.

Just as the bivariate normal distribution is the joint distribution of two linear combinations of independent normal variables, the *multivariate normal distribution* is the joint distribution of several linear combinations of independent normal variables. It can be shown that several linear combinations of independent normal variables are mutually independent if and only if the covariance between every pair of them is zero. This is a special and important property of normally distributed random variables. It makes covariance and correlation perfectly suited to the analysis of linear combinations of such variables. Keep in mind however, that in general uncorrelated random variables are not necessarily independent.

Exercises 6.5

1. Here is a summary of Pre-SAT and SAT scores of a large group of students.

PSAT scores:	average: 1200	SD: 100
SAT scores:	average: 1300	SD: 90
correlation: 0.6		

Assume the data are approximately bivariate normal in distribution.

a) Of the students who scored 1000 on the PSAT, about what percentage scored above average on the SAT?

b) Of the students who scored below average on the PSAT, about what percentage scored above average on the SAT?

c) About what percentage of students got at least 50 points more on the SAT than on the PSAT?

2. Data from a large population indicate that the heights of mothers and daughters in this population follow the bivariate normal distribution with correlation 0.5. Both variables have mean 5 feet 4 inches, and standard deviation 2 inches. Among the daughters of above average height, what percent were shorter than their mothers?

3. Heights and weights of a large group of people follow a bivariate normal distribution, with correlation 0.75. Of the people in the 90th percentile of weights, about what percentage are above the 90th percentile of heights?

4. Suppose X and Y are standard normal variables. Find an expression for $P(X+2Y \leq 3)$ in terms of the standard normal distribution function Φ,

   a) in case X and Y are independent;

   b) in case X and Y have bivariate normal distribution with correlation $1/2$.

5. Let X and Y have bivariate normal distribution with parameters μ_X, μ_Y, ${\sigma_X}^2$, ${\sigma_Y}^2$, and ρ. Let $P(X > \mu_X, Y > \mu_Y) = q$. Find:

   a) a formula for q in terms of ρ; b) a formula for ρ in terms of q.

6. Let X and Y be independent standard normal variables.

   a) For a constant k, find $P(X > kY)$.

   b) If $U = \sqrt{3}X + Y$, and $V = X - \sqrt{3}Y$, find $P(U > kV)$.

   c) Find $P(U^2 + V^2 < 1)$.

   d) Find the conditional distribution of X given $V = v$.

7. Let X and Y have bivariate normal distribution with parameters μ_X, μ_Y, ${\sigma_X}^2$, ${\sigma_Y}^2$, and ρ.

   a) Show that X and Y are independent if and only if they are uncorrelated.

   b) Find $E(Y|X = x)$. c) Find $Var(Y|X = x)$.

   d) Show that for constants a, b, and c, $aX + bY + c$ has a normal distribution. Find its mean and variance in terms of the parameters of X and Y.

   e) Show that if $\mu_X = \mu_Y = 0$, then $X \cos\theta + Y \sin\theta$ and $-X \sin\theta + Y \cos\theta$ are independent normal variables, where

   $$\theta = \frac{1}{2} \cot^{-1} \left[\frac{{\sigma_X}^2 - {\sigma_Y}^2}{2\rho\sigma_X\sigma_Y} \right]$$

   Explain the geometric significance of θ in terms of the axes of an ellipse of constant density for (X, Y).

8. Let X_1 and X_2 be two independent standard normal random variables. Define two new random variables as follows: $Y_1 = X_1 + X_2$ and $Y_2 = \alpha X_1 + 2X_2$. You are not given the constant α but it is known that $Cov\{Y_1, Y_2\} = 0$. Find

   a) the density of Y_2; b) $Cov\{X_2, Y_2\}$.

9. Suppose that W has normal (μ, σ^2) distribution. Given that $W = w$, suppose that Z has normal $(a\,w + b, \tau^2)$ distribution.

   a) Show the joint distribution of W and Z is bivariate normal, and find its parameters.

   b) What is the distribution of Z?

   c) What is the conditional distribution of W given $Z = z$?

10. Show that if V and W have a bivariate normal distribution then

a) every linear combination $aV + bW$ has a normal distribution;

b) every pair of linear combinations $(aV + bW,\ cV + dW)$ has a bivariate normal distribution.

c) Find the parameters of the distributions obtained in a) and b) in terms of the parameters of the joint distribution of V and W.

11. Show that for standard bivariate normal variables X and Y with correlation ρ,

$$E(\max(X, Y)) = \sqrt{\frac{1-\rho}{\pi}}$$

12. Suppose that the magnitude of a signal received from a satellite is

$$S = a + bV + W$$

where V is a voltage which the satellite is measuring, a and b are constants, and W is a noise term. Suppose V and W are independent and normally distributed with means 0 and variances ${\sigma_V}^2$ and ${\sigma_W}^2$.

a) Find $Corr(S, V)$.

b) Given that $S = s$, what is the distribution of V?

c) What is the best estimate of V given $S = s$?

d) If this estimate is used repeatedly for different values of S coming from a sequence of independent values of V and W with the given normal distributions, what is the long-run average absolute value of the error of estimation?

13. Find a formula in terms of ρ for the ratio of the lengths of the axes of an ellipse of constant density in the standard bivariate normal distribution with correlation ρ. (Let the ratio be the length of the axis at $+45°$ over the length of the axis at $-45°$.)

Check your answer by measurement with a ruler in Figure 3 in the case where $\rho = 1/2$. [*Hint*: Let $\rho = \cos\theta$ and reason from Figure 3 that an ellipse of constant density is the image in the (X, Y) plane of the unit circle in the (X, Z) plane. Now consider the images of the points $(\cos\theta/2, \sin\theta/2)$ and $(\cos(\theta/2 + \pi/2), \sin(\theta/2 + \pi/2))$ in the (X, Y) plane which end up on the $\pm 45°$ lines in the (X, Z) plane, and use trigonometric identities.]

Dependence: Summary

Conditional Distributions: Let X be a discrete random variable. The conditional probability of an event A given $X = x$ is

$$P(A|X = x) = \frac{P(A, X = x)}{P(X = x)}$$

by the division rule of Section 1.4.

For continuously distributed X, there is instead the *infinitesimal conditioning formula*

$$P(A|X = x) = \frac{P(A, X \in dx)}{P(X \in dx)}$$

Understand $P(A|X = x)$ as the chance of A given that X falls in a very small interval near x.

If X and Y are discrete random variables, the conditional probability of $Y = y$ given $X = x$ is

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

If X and Y are continuous random variables with joint density $f_{X,Y}$, the conditional density of Y at y given $X = x$ is $f_Y(y|X = x)$ where

$$f_Y(y|X = x)dy = P(Y \in dy|X \in dx) = \frac{f_{X,Y}(x, y)dx\, dy}{f_X(x)dx} = \frac{f_{X,Y}(x, y)}{f_X(x)}dy$$

Once you have conditioned on $X = x$, you can treat the random variable X as the constant x. Conditional distributions given $X = x$ behave exactly like ordinary distributions, with the constant x as a parameter.

Conditional expectation: The *conditional expectation of Y given $X = x$*, denoted $E(Y|X = x)$, is defined as the expectation of Y relative to the conditional distribution of Y given $X = x$.

The *conditional expectation of Y given X*, denoted $E(Y|X)$, is a random variable, whose value is $E(Y|X = x)$ if $(X = x)$. Thus the random variable $E(Y|X)$ is a function of the random variable X, namely, $f(X)$, where $f(x) = E(Y|X = x)$ for every x.

Expectation is the expectation of conditional expectation: $E(Y) = E[E(Y|X)]$.

See boxes on pages 424 and 425 for important properties of conditional distributions and expectations, and a comparison of the discrete and continuous cases.

Independence: Random variables X and Y are independent if and only if for all subsets B in the range of Y, and all x

$$P(Y \in B|X = x) = P(Y \in B)$$

That is, the conditional distribution of Y given $X = x$ does not depend on x.

Equivalently, X and Y are independent if the conditional distribution of X given $Y = y$ does not depend on y.

Covariance and correlation: $Cov(X,Y) = E\Big[[X-E(X)][Y-E(Y)]\Big] = E(XY)-E(X)E(Y)$

$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$

$Corr(X,Y) = \dfrac{Cov(X,Y)}{SD(X)SD(Y)} \in [-1,1]$

X and Y independent $\Longrightarrow Corr(X,Y) = 0$ *but not conversely.*

X and Y uncorrelated $\iff Corr(X,Y) = 0$
$\iff Cov(X,Y) = 0 \iff E(XY) = E(X)E(Y)$.

Bivariate normal: X and Y have standard bivariate normal distribution with correlation ρ if and only if

$$Y = \rho X + \sqrt{1-\rho^2}Z,$$

where X and Z are independent standard normal variables.

Marginals. Both X and Y have standard normal distribution.

Conditionals.
Given $X = x$, Y has normal $(\rho x, 1-\rho^2)$ distribution.
Given $Y = y$, X has normal $(\rho y, 1-\rho^2)$ distribution.

X and Y have bivariate normal distribution with parameters $\mu_X, \mu_Y, {\sigma_X}^2, {\sigma_Y}^2$, and ρ if and only if the standardized variables $X^* = (X - \mu_X)/\sigma_X$ and $Y^* = (Y - \mu_Y)/\sigma_Y$ have standard bivariate normal distribution with correlation ρ. Conditional distributions in this case are derived from the standardized case by a linear change of variable. All probabilities and expectations for bivariate normal variables are found by a linear change of variable to independent standard normal variables.

Independence. X and Y with bivariate normal distribution are independent if and only if they are uncorrelated.

Review Exercises

1. Let X and Y be independent random variables. Suppose X has Poisson distribution with parameter λ_1, and Y has Poisson distribution with parameter λ_2.

a) Given that $X + Y = 100$, what are the possible values of X?

b) For each possible value k, find $P(X = k \mid X + Y = 100)$.

c) Take $\lambda_1 = 1$ and $\lambda_2 = 99$. Given $X + Y = 100$, estimate the chance that X is 4 or 5 or 6.

2. Let N denote the number of children in a randomly picked family. Suppose N has geometric distribution:

$$P(N = n) = (1/3)(2/3)^{n-1} \qquad (n = 1, 2, 3, \ldots)$$

and suppose each child is equally likely to be male or female. Let X be the number of male children and Y the number of female children, in a randomly picked family:

a) Find the joint distribution of (X, Y).

b) Given $Y = 0$, what is the most likely value of X?

c) What is the conditional expectation of X given $Y = 0$?

3. A list of $2n$ numbers has mean μ and variance σ^2. Suppose that n numbers are picked at random from the list. Let A_n be the average of these n numbers, B_n the average of the other numbers. Find: a) $E(A_n - B_n)$; b) $SD(A_n - B_n)$.

4. Suppose X and Y have joint density function

$$f(x, y) = \begin{cases} c/x^3 & x > y > 1 \\ 0 & \text{otherwise} \end{cases}$$

where c is a constant.

a) Find c. b) Find the marginal density of X.

c) What is the conditional distribution of Y given $X = x$?

5. Suppose X and Y are random variables with joint density in the plane $f(x, y) = ce^{-(x^2+xy+y^2)}$ where c is a constant. a) Find c. b) Find $Corr(X, Y)$.

6. Let X and Y be independent exponential random variables each with mean 1. Find
a) the joint density of $X + Y$ and $X - Y$;
b) $Corr(X + Y, X - Y)$.

7. Suppose that a point (X, Y) is chosen according to the uniform distribution on the triangle with vertices $(0, 0), (0, 1), (1, 0)$. Calculate:

a) the mean and variance of X;

b) the conditional mean and variance of X given that $Y = 1/3$;

c) the mean and variance of $\max(X, Y)$;

d) the mean and variance of $\min(X, Y)$.

8. Let Y have exponential distribution with mean 0.5. Let X be such that, conditional on $Y = y$, X has exponential distribution with mean y. Find:

(a) the joint density of (X, Y); b) $E(X)$; c) $Corr(X, Y)$.

9. Let X, Y, and Z be independent uniform $(0, 1)$ variables. Find $P[(X/Y) > (Y/Z)]$.

10. Let T_A, T_B, and T_C be the failure times of components A, B, and C. Assume these are independent exponential random variables with rates α, β, and γ, respectively.

a) What is the distribution of the time until the first failure?

b) What is the probability that the first component to fail is component C?

c) Given that the first component to fail is component C, what is the distribution of the time between the first and second failures?

d) Write a formula for the (unconditional) c.d.f. of the time between the first and second failures.

11. Insurance claims arrive at an insurance company according to a Poisson process with rate λ. The amount of each claim has exponential distribution with rate μ, independently of times and amounts of all other claims. Let X_t denote the accumulated total of claims between time 0 and time t. Find simple formulae for

a) $E(X_t)$; $E(X_t^2)$; c) $SD(X_t)$; d) $Corr(X_s, X_t)$ for $s \le t$.

12. An elevator has an occupancy warning of no more than 26 people and of total weight no more than 4000 pounds. For the population of users, suppose weights are approximately normal with mean 150 pounds and standard deviation 30 pounds.

a) What is the probability that the total weight of a random sample of 26 people from the population exceeds 4000 pounds?

b) Suppose next that the people are carrying things and that the weight of these for an individual of weight X pounds, is approximately normal with mean $0.05X$ pounds and standard deviation 2 pounds. What is the probability that the total weight in the elevator now exceeds 4000 pounds?

c) The dimensions of the floor of the elevator are 54 inches by 92 inches. Suppose the amount of floor space needed by users is normally distributed with mean μ square inches and standard deviation 0.1μ. Find μ such that the probability 20 people can be accommodated is 0.99.

13. a) Let X and Y be two random variables with finite and nonzero variances. Show that $X - Y$ and $X + Y$ are uncorrelated if and only if $Var(X) = Var(Y)$.

b) Let X and Y have standard bivariate normal distribution with correlation 0.6. Find $P(X - Y < 1, X + Y > 2)$.

14. Heights. A population consists of 50% men and 50% women. The empirical distribution of heights over the population yields the following statistics:

	Average	Standard deviation
Men's heights	67 inches	3 inches
Women's heights	63 inches	3 inches

a) What is the average height over the whole population?

b) What is the standard deviation of heights over the whole population?

c) Suppose that men's heights are approximately normally distributed, and that women's heights are as well. Calculate the approximate proportion of individuals in the whole population with heights between 63 and 67 inches.

d) Repeat c), assuming instead that heights are normally distributed over the whole population. Explain why the answers are slightly different.

e) Suppose that a man and a woman are picked at random from this population. Making assumptions as in c), what is the probability that the man is taller than the woman? [*Hint:* No integration required!]

15. Sums of normals in the positive quadrant. Let X and Y be two independent standard normal variables.

a) Calculate $P(X \geq 0, Y \geq 0, X + Y \leq 1)$.

b) Find the conditional density of $X + Y$ given $X \geq 0$ and $Y \geq 0$, and sketch its graph.

c) Find, approximately, the median and the mode of this distribution.

16. Rainfall. Suppose that the distribution of annual rainfall in a particular place, measured in inches, is approximately gamma with shape parameter $r = 3$. If the mean annual rainfall is 20 inches, find approximations to the following:

a) the probability of more than 35 inches of rain in any particular year;

b) the probability that in ten consecutive years, it never rains more than 35 inches, assuming different years are independent;

c) still assuming independence of different years, the probability that the record rainfall over the last 20 years is exceeded in at least one of the next ten years, assuming the record rainfall over the last 20 years, R_{20} say, is known;

d) same as c), but assuming the value of R_{20} is unknown.

17. Symmetry under rotations.

a) Suppose the joint distribution of X and Y is symmetric under rotations. Are X and Y necessarily independent? Are they necessarily uncorrelated? Explain by arguments or examples.

b) Suppose (X, Y) is a point picked at random from the unit circle $X^2 + Y^2 = 1$. Calculate $E(X^2)$, $E(Y^2)$, and $E(XY)$.

c) Suppose U is uniformly distributed on $(0, 1)$, $X = \cos 2\pi U, Y = \sin 2\pi U$. Are X and Y uncorrelated? Are X and Y independent? Explain carefully the connection betweeen b) and c).

18. Maxima and minima of normal variables.
Calculate the expected values of $\max(X, Y)$ and $\min(X, Y)$:

a) if X and Y are independent standard normal variables;

b) if X and Y are independent normal (μ, σ^2);

c) if X and Y are standard bivariate normal with correlation ρ.

19. Suppose you sample with replacement n times from a population of n elements.

a) What fraction of the n elements should you expect to see in the sample?

b) For example, what fraction of all $\binom{52}{5}$ poker hands should you expect to see in $\binom{52}{5}$ independent deals?

c) Compute the variance of the fraction in a), and show that it is less than $1/4n$.

d) Evaluate for the example in b), and estimate the chance that your prediction in b) is off by more than 1%.

20. Craps. Find the expectated total number of times Y the pair of dice must be rolled in a craps game (see Exercise 3.4.8) by conditioning on the result of the first roll.

21. I toss a coin which lands heads with probability p. Let W_H be the number of tosses till I get a head, W_{HH} the number of tosses till I get two heads in a row, and W_{HHH} the number of tosses till I get three heads in a row. Find:

a) $E(W_H)$; b) $E(W_{HH})$ [*Hint*: condition on whether the first toss was heads or tails]; c) $E(W_{HHH})$ [*Hint*: condition on W_T].

d) Generalize to find the expected number of tosses to obtain m heads in a row.

22. Long runs of heads. In the play *Rosencrantz and Guildenstern are dead* by Tom Stoppard, the results of 101 apparently fair coin tosses are recorded: 100 heads in a row, followed by a tail. Suppose a fair coin is tossed independently once every second. About how many years do you expect it would take before 100 heads in a row came up? How long for it to be 99% sure that such a run will have appeared?

23. Suppose an insect lays a Poisson (λ) number of eggs. Suppose each egg hatches with probability p and dies with probability q, independently of each other egg. Show that the number of eggs that hatch and the number of eggs that die are independent Poisson random variables, and find their parameters.

24. I roll a random number of dice. If the number of dice rolled has the Poisson (12) distribution, find (and justify your answers)

a) the expectation of the total number of spots showing;

b) the standard deviation of the total number of spots showing.

25. Suppose the number of accidents in an interval of time has Poisson (λ) distribution. Suppose that in each accident there are k persons injured with probability p_k, independently of all other accidents. Let N_k be the number of accidents in which k persons are injured.

a) What is the joint distribution of N_1 and N_2?

b) Let M be the total number of persons injured. Find formulae for $E(M)$ and $SD(M)$ in terms of $p_1, p_2, \ldots$ and λ.

26. Distinguishing points in a Poisson scatter. In practical situations, if two points in a scatter are closer than some distance δ, it may not be possible to distinguish them. Suppose that this is the case, and that there is a Poisson scatter over the unit square, with intensity λ. Show that the probability of the event D, that all points in the scatter can be distinguished, is at least $1 - \frac{\pi}{2}\lambda^2\delta^2$.

[*Hint.* Show that $P(D|N=2) \geq 1 - \pi\delta^2$ and $P(D|N=3) \geq (1-\pi\delta^2)(1-2\pi\delta^2)$ and so on. Use the inequality

$$(1-\alpha)(1-\beta) \geq 1 - (\alpha + \beta) \qquad (\alpha > 0,\ \beta > 0)$$

repeatedly, to obtain

$$P(D|N) \geq 1 - \frac{1}{2}N(N-1)\pi\delta^2]$$

27. Inhomogeneous Poisson scatter. Let Q be a probability distribution over a set S, $\lambda > 0$. Consider a random scatter of points over the set S, where a Poisson (λ) number N of points are distributed independently at random according to Q. More formally, for B a subset of S, let $N(B) = 0$ if $N = 0$, and

$$N(B) = \sum_{i=1}^{n} I(X_i \in B) \qquad \text{if} \quad (N = n), \quad n = 1, 2, \ldots$$

where $X_1, \ldots, X_n$ are conditionally independent with common distribution Q given $(N = n)$, and N has Poisson (λ) distribution. Prove that
for disjoint $B_1, \ldots, B_j$, *the* $N(B_1), \ldots, N(B_j)$ *are mutually independent Poisson random variables with parameters* $\lambda(B_1), \ldots, \lambda(B_j)$ *where* $\lambda(B) = \lambda Q(B)$.
[*Hint*: Start by considering the case of B_1 and B_2 with $B_1 + B_2 = S$, and calculate $P(N(B_1) = n_1, N(B_2) = n_2)$ by conditioning on $N = n_1 + n_2$. Argue that, in general, it suffices to consider a partition $B_1, \ldots, B_j$ of S, and proceed similarly. The multinomial coefficients $n!/(n_1!n_2!\cdots n_j!)$ should appear.]
Note. Such a collection of random variables $N(B)$ is called a *Poisson process with intensity measure* $\lambda(B)$ *on* S. For S the unit square and $\lambda(B) = \lambda \times Area(B)$ this is a construction of the Poisson scatter over the unit square considered in Section 3.5. Such a scatter is called *homogeneous.* If $Q(B)$ is not the uniform distribution, the scatter is called *inhomogeneous.* Note that if $Q\{s\} > 0$ for a point $s \in S$, there may be more than one "hit" counted at s. In particular, if Q is a discrete measure with probabilities $q_1, \ldots, q_n$ at points $s_1, \ldots, s_n$, then $N(s_1), \ldots, N(s_n)$ are independent Poisson random variables with parameters $\lambda q_1, \ldots, \lambda q_n$.
Illustration. Suppose you roll a Poisson (λ) number N of dice. Then the number of times each of the six faces appears is an independent Poisson ($\lambda/6$) random variable. And the number of odd faces and the number of even faces are two independent Poisson ($\lambda/2$) random variables. But if you throw a fixed number n of dice these numbers are dependent, because they must add up to n.

28. You and I both toss a fair coin N times. You get X heads and I get Y heads.

a) If $P(X = Y)$ is approximately 10%, then approximately how large must N be?

b) The normal approximation says $P(\left|X - \frac{1}{2}N\right| \leq \frac{1}{2}\sqrt{N}) \approx 68\%$.
Given $X = Y$, is the conditional probability that $\left|X - \frac{1}{2}N\right| \leq \frac{1}{2}\sqrt{N}$ still about 68%, somewhat larger than 68%, or somewhat smaller than 68%? Explain which, *without* doing detailed calculations.

29. Variance of discrete order statistics. Let T_i be the place at which the ith good element appears in a random ordering of k good and $N - k$ bad elements. From Exercise 3.6.13, the mean of T_i is $E(T_i) = i(N+1)/(k+1)$. Calculate $SD(T_i)$ by the following steps.

a) Let $\alpha(k,N) = E(T_1(T_1-1))$, $1 \le k \le N$. Show by conditioning on whether the first element is good or bad that

$$\alpha(k,N) = (N-k)\left[\frac{2}{k+1} + \frac{\alpha(k,N-1)}{N}\right]$$

b) Deduce that

$$\alpha(k,N) = \frac{2(N+1)(N-k)}{(k+1)(k+2)}$$

c) Deduce that

$$Var(T_1) = \frac{(N+1)(N-k)k}{(k+1)^2(k+2)}$$

d) Check the case $k=1$ by calculating $Var(T_1)$ directly from the distribution of T_1.

e) Let $W_i = T_{i+1} - T_i$, $\quad i = 1, \ldots, k+1$, $\quad$ where $T_0 = 0$ and $T_{k+1} = N+1$. Use the exchangeability of $W_1, \ldots, W_{k+1}$ to show that for each $i = 1, \ldots, k+1$

$$Var(T_i) = i\,Var(T_1) + i(i-1)\,Cov(W_1, W_2)$$

Deduce that

$$Cov(W_1, W_2) = -Var(T_1)/k$$

and hence that

$$Var(T_i) = \frac{i(k+1-i)(N+1)(N-k)}{(k+1)^2(k+2)}$$

f) Give an intuitive explanation of why $SD(T_i) = SD(T_{k+1-i})$.

g) Suppose that $T_1, \ldots, T_4$ are the places at which the aces appear in a well-shuffled deck of 52 cards. Find numerical values of $E(T_i)$ and $SD(T_i)$ for $i = 1, \ldots, 4$.

30. Let $V_1, \ldots, V_n$ be the order statistics of n independent uniform $(0,1)$ variables. Let
$A_{\text{all}} = (V_1 + \cdots + V_n)/n$, $\quad$ average of all the order statistics,
$A_{\text{ext}} = (V_1 + V_n)/2$, $\quad$ average of the extremes,
$A_{\text{mid}} = V_{(n+1)/2}$, $\quad$ the middle value, where you can assume n is odd.

a) Show that for large n, each of the A's is most likely very close to $1/2$.

b) For large n, one of the A's can be expected to be very much closer to $1/2$ than the two others. Which one, and why?

c) For $n = 101$ find for each of the A's a good approximation to the probability that it is between .49 and .51.

31. **From discrete to continuous spacings.** Let $U_{(1)} < U_{(2)} < \ldots < U_{(n)}$ be the order statistics of n independent uniform $(0,1)$ variables $U_1, \ldots, U_n$. Let $V_1 = U_{(1)}$, $V_i = U_{(i)} - U_{(i-1)}$ for $1 \le i \le n$, and let $V_{n+1} = 1 - U_{(n)}$. Imagine the unit interval is cut into subintervals at each of the n random points U_i for $1 \le i \le n$. Then $V_1, V_2, \ldots, V_{n+1}$ are the lengths of the $n+1$ subintervals so obtained, in order from left to right. This model for cutting an interval at random is of interest in genetics. The V_i could represent the relative lengths of strands obtained by random cutting of a long molecule such as DNA. For a positive integer $N > n$ let $U'_1, \ldots, U'_{N-n}$ denote $N-n$ more uniform $(0,1)$ variables, independent of each other and of the cut points $U_1, \ldots, U_n$. For $1 \le i \le n+1$ let N_i denote the number of U'_i that fall in the interval $(U_{(i-1)}, U_{(i)})$ of length V_i (where $U_{(0)} = 0$ and $U_{(n+1)} = 1$ to make the definition work for N_1 and N_{n+1}).

a) Show that the joint distribution of $N_1, \ldots, N_{n+1}$ is identical to the joint distribution of the discrete spacings $W_1, \ldots, W_{n+1}$ derived from a random ordering of n aces and $N-n$ nonaces as in Exercise 3.6.13. That is to say, $(N_1, \ldots, N_{n+1})$ has uniform distribution over the set of all $(n+1)-$tuples of non-negative integers $(n_1, \ldots, n_{n+1})$ with $n_1 + \cdots + n_{n+1} = N-n$. In particular, $N_1, \ldots, N_{n+1}$ are exchangeable.

b) Conditionally given the continuous spacings $(V_1, \ldots, V_{n+1})$, the sequence $(N_1, \ldots, N_{n+1})$ is distributed like the number of results in each of $n+1$ categories in a sequence of $N-n$ independent trials with probability V_i of a result in category i on each trial. Explain why this is so. Deduce that for large N, N_i/N is almost equal to V_i for each i with overwhelmingly high probability. It follows that in the limit as $N \to \infty$ for fixed n, as discussed at the end of Exercise 3.6.12, the joint distribution of the normalized discrete spacings $(N_1/N, N_2/N, \ldots, N_{n+1}/N)$ converges to the joint distribution of the continuous spacings $V_1, V_2, \ldots, V_{n+1}$.

(Keep in mind that the distribution of N_i depends on N, so N_i/N does not just tend to zero: the sum over i of the N_i/N is identically equal to 1.) Since the N_i/N are exchangeable for every N, it follows that the V_i are exchangeable, something that is not obvious in the continuous model.

32. Joint distribution of continuous spacings. Continuing with the same notation as in Exercise 31,

a) Show that for $v_i \geq 0$ with $v_1 + \cdots + v_{n+1} = v \leq 1$

$$\lim_{N\to\infty} P(N_i/N \geq v_i \text{ for every } 1 \leq i \leq n+1) = (1-v)^n$$

by explicit evaluation of the limit, using Exercise 3.6.15 and the fact that $(N)_k \sim N^k$ as $N \to \infty$ for every $k = 1, 2, \ldots$. This yields the corresponding probabilty for the continuous model: for $v_i \geq 0$ with $v_1 + \cdots + v_{n+1} = v \leq 1$

$$P(V_i \geq v_i \text{ for every } 1 \leq i \leq n+1) = (1-v)^n$$

b) Show that the V_i have identical distribution with

$$P(V_i \geq v) = (1-v)^n \qquad (0 \leq v \leq 1)$$

c) Deduce that V_i has beta $(1, n)$ distribution.

33. Maximum and minimum spacings. Continuing with the notation of the preceding exercises, let $V_{\min} = \min_i V_i$ where the min is over $1 \leq i \leq n+1$.

a) Show that $V_{\min}$ has the same distribution as $V_1/(n+1)$. Deduce the mean and variance of $V_{\min}$ from the mean and variance of the beta $(1, n)$ distribution.

b) Let $V_{\max} = \max_i V_i$. Parallel to the discrete formula of Exercise 3.6.16, show that for $0 \leq v \leq 1$

$$P(V_{\max} \geq v) = \sum_{i=1}^{n+1} (-1)^{i-1} \binom{n+1}{i} (1-iv)_+^n$$

where $(1-iv)_+^n$ equals $(1-iv)^n$ if $iv \leq 1$, and equals 0 otherwise.

c) Deduce by integration of this tail probability from 0 to 1 that

$$E(V_{\max}) = \sum_{i=1}^{n+1} (-1)^{i-1} \binom{n+1}{i} \frac{1}{i(n+1)}$$

It is intuitively clear, and can be verified analytically, that as the number of cuts $n \to \infty$, $V_{\max} \to 0$, which forces the distribution of $V_{\max}$ to pile up around zero. But the rate of convergence is rather slow.

d) Find the numerical values of $E(V_{\min})$ and $E(V_{\max})$ for $n = 1, \ldots, 10$.

34. Dirichlet distribution. A sequence of random variables $(Q_1, \ldots, Q_m)$ has *Dirichlet distribution* with parameters $(r_1, \ldots, r_m)$ if $Q_i \geq 0$, $Q_1 + \cdots + Q_m = 1$, and

$$\frac{P(Q_i \in dq_i, 1 \leq i \leq m-1)}{dq_1 dq_2 \cdots dq_{m-1}} = \frac{\Gamma(r_1 + \cdots + r_m)}{\Gamma(r_1) \cdots \Gamma(r_m)} \prod_{i=1}^{m} q_i^{r_i - 1} \qquad (q_i \geq 0,\ q_1 + \cdots + q_m = 1)$$

For $m = 2$, (Q_1, Q_2) has Dirichlet distribution with parameters r and s if and only if $Q_2 = 1 - Q_1$ for Q_1 with beta (r, s) distribution. So the Dirichlet distribution is a multivariate extension of the beta distribution. There is a multivariate version of the result of Exercise 5.4.19: If $Y_i, 1 \leq i \leq m$ are independent with gamma (r_i, λ) distributions, $\sum = \sum_i Y_i$ and $Q_i = Y_i / \sum$, then $(Q_1, \ldots, Q_m)$ has Dirichlet distribution with parameters $(r_1, \ldots, r_m)$, independently of $\sum$, which has gamma (r, λ) distribution for $r = r_1 + \cdots + r_m$. Assuming this result, deduce the following properties of this Dirichlet distribution of $(Q_1, \ldots, Q_m)$:

a) The marginal distribution of Q_i is beta $(r_i, r - r_i)$.

b) For $i \neq j$ the distribution of $Q_i + Q_j$ is beta $(r_i + r_j, r - r_i - r_j)$. Similarly for any finite sum of at most $m - 1$ different Q_i.

c) The joint distribution of the continuous spacings $V_1, \cdots, V_{n+1}$ derived from n independent uniform$(0, 1)$ random variables as in Exercises 31 and 32 is Dirichlet with parameters $r_i = 1$ for $1 \leq i \leq m = n + 1$.

35. Dirichlet–multinomial. Suppose that $X_1, X_2, \ldots$ is a sequence of independent trials with m possible values $\{1, \ldots, m\}$, with probability q_i for value i on each trial. The parameters $(q_1, \ldots, q_m)$ are unknown, and regarded as the values of random variables $(Q_1, \ldots, Q_m)$. Suppose the prior distribution of $(Q_1, \ldots, Q_m)$ is Dirichlet with parameters $(r_1, \ldots, r_m)$, as in Exercise 34. After n trials, let N_i be the number of results i, that is the number of times that $X_j = i$ for $1 \leq j \leq n$. So the conditional distribution of $(N_1, \ldots, N_m)$ given $(Q_1, \ldots, Q_m)$ is multinomial with parameters n and $(Q_1, \ldots, Q_m)$.

a) Show that the posterior distribution of $(Q_1, \ldots, Q_m)$ given the results $(N_1, \ldots, N_m)$ of n trials is Dirichlet with parameters $(r_1 + N_1, \ldots, r_m + N_m)$.

b) Find a formula for the unconditional probability $P(N_i = n_i \text{ for } 1 \leq i \leq m)$ for any sequence of m non-negative integers n_i with $n_1 + \cdots + n_m = n$. [*Hint*: Use the fact that the total integral of the Dirichlet joint density with parameters $(r_1 + n_1, \ldots, r_m + n_m)$ is 1].

c) Deduce in particular that if $r_i = 1$ for $1 \leq i \leq m$ then the unconditional distribution of $(N_1, \ldots, N_m)$ is uniform over its range of possible values.

d) Explain the result of part c) without integration by reference to Exercise 31.

Distribution Summaries

Discrete

name and range	$P(k) = P(X = k)$ for $k \in$ range	mean	variance
uniform on $\{a, a+1, \ldots, b\}$	$\dfrac{1}{b-a+1}$	$\dfrac{a+b}{2}$	$\dfrac{(b-a+1)^2-1}{12}$
Bernoulli (p) on $\{0, 1\}$	$P(1) = p$; $P(0) = 1-p$	p	$p(1-p)$
binomial (n, p) on $\{0, 1, \ldots, n\}$	$\binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$
Poisson (μ) on $\{0, 1, 2, \ldots\}$	$\dfrac{e^{-\mu}\mu^k}{k!}$	μ	μ
hypergeometric (n, N, G) on $\{0, \ldots, n\}$	$\dfrac{\binom{G}{k}\binom{N-G}{n-k}}{\binom{N}{n}}$	$\dfrac{nG}{N}$	$n\left(\dfrac{G}{N}\right)\left(\dfrac{N-G}{N}\right)\left(\dfrac{N-n}{N-1}\right)$
geometric (p) on $\{1, 2, 3 \ldots\}$	$(1-p)^{k-1}p$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$
geometric (p) on $\{0, 1, 2 \ldots\}$	$(1-p)^k p$	$\dfrac{1-p}{p}$	$\dfrac{1-p}{p^2}$
negative binomial (r, p) on $\{0, 1, 2, \ldots\}$	$\binom{k+r-1}{r-1} p^r (1-p)^k$	$\dfrac{r(1-p)}{p}$	$\dfrac{r(1-p)}{p^2}$

Continuous

† undefined.

name	range	density $f(x)$ for $x \in$ range	c.d.f. $F(x)$ for $x \in$ range	Mean	Variance
uniform (a, b)	(a, b)	$\frac{1}{b-a}$	$\frac{x-a}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
normal $(0, 1)$	$(-\infty, \infty)$	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$	$\Phi(x)$	0	1
normal (μ, σ^2)	$(-\infty, \infty)$	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}$	$\Phi\left(\frac{x-\mu}{\sigma}\right)$	μ	σ^2
exponential (λ) = gamma $(1, \lambda)$	$(0, \infty)$	$\lambda e^{-\lambda x}$	$1 - e^{-\lambda x}$	$1/\lambda$	$1/\lambda^2$
gamma (r, λ)	$(0, \infty)$	$\Gamma(r)^{-1}\lambda^r x^{r-1} e^{-\lambda x}$	$1 - e^{-\lambda x}\sum_{k=0}^{r-1}\frac{(\lambda x)^k}{k!}$ for integer r	r/λ	r/λ^2
chi-square (n) =gamma $\left(\frac{n}{2}, \frac{1}{2}\right)$	$(0, \infty)$	$\Gamma(\frac{n}{2})^{-1}(\frac{1}{2})^{\frac{n}{2}} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}$	as above for $\lambda = \frac{1}{2}$, $r = \frac{n}{2}$ if n is even	n	$2n$
Rayleigh	$(0, \infty)$	$xe^{-\frac{1}{2}x^2}$	$1 - e^{-\frac{1}{2}x^2}$	$\sqrt{\frac{\pi}{2}}$	$\frac{4-\pi}{2}$
beta (r, s)	$(0, 1)$	$\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} x^{r-1}(1-x)^{s-1}$	see Exercise 4.6.5 for integer r and s	$\frac{r}{r+s}$	$\frac{rs}{(r+s)^2(r+s+1)}$
Cauchy	$(-\infty, \infty)$	$\frac{1}{\pi(1+x^2)}$	$\frac{1}{2} + \frac{1}{\pi}\arctan(x)$	†	†
arcsine =beta $(1/2, 1/2)$	$(0, 1)$	$\frac{1}{\pi\sqrt{x(1-x)}}$	$\frac{2}{\pi}\arcsin(\sqrt{x})$	$\frac{1}{2}$	$\frac{1}{8}$

Beta

Parameters: $r > 0$ and $s > 0$

Range: $x \in [0, 1]$

Density function:

$$P(X_{r,s} \in dx)/dx = \frac{1}{B(r,s)} x^{r-1}(1-x)^{s-1} \quad (0 \le x \le 1)$$

where

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

is the *beta function*, and $\Gamma(r)$ is the *gamma function* (see gamma distributions).

Cumulative distribution function: (Exercises in Section 4.6.) No simple general formula for r or s not an integer. See tables of the incomplete beta function. For integers r and s

$$P(X_{r,s} \le x) = \sum_{i=r}^{r+s-1} \binom{r+s-1}{i} x^i (1-x)^{r+s-1-i} \qquad (0 \le x \le 1)$$

Mean and standard deviation: (4.6)

$$E(X_{r,s}) = \frac{r}{r+s} \qquad SD(X_{r,s}) = \frac{\sqrt{rs}}{(r+s)\sqrt{(r+s+1)}}$$

Special cases:

- $r = s = 1$: The uniform $[0, 1]$ distribution.
- $r = s = 1/2$: The arcsine distribution.

Sources and applications:

- Order statistics of uniform variables (4.6).
- Ratios of gamma variables (5.4).
- Bayesian inference for unknown probabilities.

Normal approximation:

- Good for large r and s.

Binomial

Parameters:

$n =$ number of trials $(n = 1, 2, \ldots)$
$p =$ probability of success on each trial $(0 \le p \le 1)$

Range: $k \in \{0, 1, \ldots, n\}$

Probability function: (2.1)

$$P(k) = P(S = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad (k = 0, 1, \ldots, n)$$

$$\text{where} \quad S = \begin{pmatrix} \text{number of successes in } n \text{ independent trials with} \\ \text{probability } p \text{ of success on each trial} \end{pmatrix}$$
$$= X_1 + \cdots + X_n \quad \text{where } X_i = \text{indicator of success on trial } i.$$

Mean and standard deviation: (3.2, 3.3)

$$E(S) = \mu = np \qquad SD(S) = \sigma = \sqrt{np(1-p)}$$

Mode: (2.1) $\quad \text{int}(np + p)$

Consecutive odds ratios: (2.1)

$$\frac{P(k)}{P(k-1)} = \frac{(n-k+1)}{k} \cdot \frac{p}{1-p} \qquad \text{(decreasing)}$$

Special case: (1.3)

Binomial $(1, p) \equiv$ Bernoulli (p), distribution of the indicator of an event A with probability $P(A) = p$.

Normal approximation: (2.2, 2.3)

If $\sigma = \sqrt{np(1-p)}$ is sufficiently large

$$P(k) \approx \frac{1}{\sigma} \phi \left(\frac{k - \mu}{\sigma} \right)$$

where $\phi(z)$ is the standard normal density function

$$P(a \text{ to } b) \approx \Phi \left(\frac{b + \frac{1}{2} - \mu}{\sigma} \right) - \Phi \left(\frac{a - \frac{1}{2} - \mu}{\sigma} \right)$$

where Φ is the standard normal cumulative distribution function.

Poisson approximation: (2.4)

If p is close to zero

$$P(k) \approx e^{-\mu} \mu^k / k! \quad \text{where} \quad \mu = np$$

Exponential

Parameter: $\lambda > 0$, the *rate* of an exponential random variable T.

Range: $t \in [0, \infty)$

Density function: (4.2)

$$P(T \in dt)/dt = \lambda e^{-\lambda t} \quad (t \geq 0)$$

Cumulative distribution function: (4.2)

$$P(T \leq t) = 1 - e^{-\lambda t} \quad (t \geq 0)$$

Often T is interpreted as a lifetime.

Survival function:

$$P(T > t) = e^{-\lambda t} \quad (t \geq 0)$$

Mean and Standard Deviation: (4.2)

$$E(T) = 1/\lambda \qquad SD(T) = 1/\lambda$$

Interpretation of λ:

$$\lambda = P(T \in dt \,|\, T > t)/dt$$

is the constant *hazard rate* or chance per unit time of death given survival to time t. See Section 4.3 for a discussion of non-constant hazard rates.

Characterizations:

- Only distribution with constant hazard rate.
- Only distribution with the *memoryless property*

$$P(T > t + s \,|\, T > t) = P(T > s)$$

for all $s, t > 0$

Sources:

- Time until the next arrival in a Poisson process with rate λ.
- Approximation to geometric (p) distribution for small p.
- Approximation to beta $(1, s)$ distribution for large s.
- Spacings and shortest spacings of uniform order statistics.

Gamma

Parameters: $r > 0$ (shape) $\lambda > 0$ (rate or inverse scale)

Range: $t \in [0, \infty)$

Density function: (4.2, 5.4) $P(T_{r,\lambda} \in dt)/dt = \Gamma(r)^{-1}\lambda^r t^{r-1}e^{-\lambda t} \quad (t \geq 0)$
where $\Gamma(r) = \int_0^\infty t^{r-1}e^{-t}dt$ is the *gamma function. Note:* $\Gamma(r) = (r-1)!$ for integer r.

Cumulative distribution function: (4.2)
No formula for non-integer r. See tables of the incomplete gamma function. For integer r

$$P(T_{r,\lambda} \leq t) = P(N_{t,\lambda} \geq r) = 1 - \sum_{k=0}^{r-1} e^{-\lambda t}\frac{(\lambda t)^k}{k!}$$

where $N_{t,\lambda}$ denotes the number of points up to time t in a Poisson process with rate λ, and has Poisson (λt) distribution.

Mean and standard deviation: (4.2) $E(T_{r,\lambda}) = r/\lambda \quad SD(T_{r,\lambda}) = \sqrt{r}/\lambda$

Special cases:

- gamma $(1, \lambda)$ is exponential (λ).
- gamma $(n/2, 1/2)$ is chi-square (n), the distribution of the sum of the squares of n independent standard normals.

Sources:

- Sum of r independent exponential (λ) variables.
- Time until the rth arrival in a Poisson process with rate λ.
- Bayesian inference for unknown Poisson rates.
- Approximation to negative binomial (r, p) for small p.
- Approximation to beta (r, s) for large s.

Transformations: (Notation: $X \sim F$ means X is a random variable with distribution F.)

Scaling: $T \sim$ gamma $(r, \lambda) \iff \lambda T \sim$ gamma $(r, 1)$

Sums: For independent $T_i \sim$ gamma (r_i, λ) $\quad \sum_i T_i \sim$ gamma $(\sum_i r_i, \lambda)$

Ratios: For independent $T_{r,\lambda}$ and $T_{s,\lambda}$

$$\frac{T_{r,\lambda}}{T_{r,\lambda} + T_{s,\lambda}} \sim \text{beta}(r, s) \text{ independent of the sum } T_{r,\lambda} + T_{s,\lambda} \sim \text{gamma } (r + s, \lambda)$$

Higher moments: For $s > 0$ $\quad E[(T_{r,\lambda})^s] = \dfrac{\Gamma(r+s)}{\Gamma(r)\lambda^s}$

Normal approximation: If r is sufficiently large, the distribution of the standardized gamma variable $Z_{r,\lambda} = [T_{r,\lambda} - E(T_{r,\lambda})]/SD(T_{r,\lambda})$ is approximately standard normal.

Geometric and Negative Binomial

Geometric

Parameter: $p =$ success probability.

Range: $n \in \{1, 2, \ldots\}$

Definition: Distribution of the waiting time T to first success in independent trials with probability p of success on each trial.

Probability function: (1.6, 3.4)

$$P(n) = P(T = n) = (1-p)^{n-1}p \quad (n = 1, 2, \ldots)$$

Let $F = T - 1$ denote the number of failures before the first success. The distribution of F is the geometric distribution on $\{0, 1, 2, \ldots\}$.

Tail probabilities:

$$P(T > n) = P(\text{first } n \text{ trials are failures}) = (1-p)^n$$

Mean and Standard Deviation: (3.4)

$$E(T) = 1/p \qquad SD(T) = \sqrt{(1-p)}/p$$

Negative Binomial

Parameters: $p =$ success probability, $r =$ number of successes.

Range: $n \in \{0, 1, 2, \ldots\}$

Definition: Distribution of the number of failures F_r before the rth success in Bernoulli trials with probability p of success on each trial.

Probability function: (3.4)

$$P(F_r = n) = P(T_r = n + r) = \binom{n+r-1}{r-1} p^r (1-p)^n \quad (n = 0, 1, \ldots)$$

where T_r is the waiting time to the rth success. The distribution of $T_r = F_r + r$ is the negative binomial distribution on $\{r, r+1, \ldots\}$.

Mean and standard deviation: (3.4)

$$E(F_r) = r(1-p)/p \qquad SD(F_r) = \sqrt{r(1-p)}/p$$

Sum of geometrics: The sum of r independent geometric (p) random variables on $\{0, 1, 2, \ldots\}$ has negative binomial (r, p) distribution.

Hypergeometric

Parameters: n = sample size
N = total population size
G = number of good elements in population

Range: $g \in \{0, 1, \ldots, n\}$

Definition: The *hypergeometric* (n, N, G) *distribution* is the distribution of the number S of good elements in a random sample of size n without replacement from a population of size N with G good elements and $B = N - G$ bad ones.

Probability function: (2.5)

$$P(g) = P(S = g) = \binom{n}{g} \frac{(G)_g (B)_b}{(N)_n} = \frac{\binom{G}{g}\binom{B}{b}}{\binom{N}{n}}$$

is the chance of getting g good elements and b bad elements in the random sample of size n. Here $b = n - g$. The random variable is

$$S = \text{number of good elements in sample} = X_1 + \cdots + X_n$$

where X_i = indicator of the event that the ith element sampled is good. These indicators are dependent due to sampling without replacement. But each indicator has the same Bernoulli(p) distribution, where

$$p = G/N = P(X_i = 1) = P(i\text{th element is good}) \quad \text{for each } i = 1, \ldots, n$$

Compare with the binomial (n, p) distribution of S for sampling with replacement, when the indicators are independent.

Mean and standard deviation: (3.6, 6.4)

$$E(S) = \mu = np \qquad SD(S) = \sigma = \sqrt{np(1-p)} \cdot \sqrt{\frac{N-n}{N-1}}$$

Note: Mean is the same as for sampling with replacement. But the SD is decreased by the *correction factor* of $\sqrt{(N-n)/(N-1)}$.

Normal approximation: As for binomial if σ is large enough, for σ as above with correction factor.

Poisson approximation: As for binomial if $p = G/N$ sufficiently small but both G and N are large.

Binomial approximation: Ignores the distinction between sampling with and without replacement. Works well if $n \ll N$ and both G and B are large.

Conditioned binomial: Let S_n be the number of successes in n independent trials which are part of a larger sequence of N independent trials. Then no matter what the probability of success p, provided it is the same on all trials, the conditional distribution of S_n given $S_N = G$ is hypergeometric (n, N, G).

Normal

Standard Normal

Range: $z \in (-\infty, \infty)$

Standard normal density function: (2.2, 4.1)

$$P(Z \in dz)/dz = \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad (-\infty < z < \infty)$$

Standard normal cumulative distribution function:

$$P(Z \leq z) = \Phi(z) = \int_{-\infty}^{z} \phi(x)dx$$

No simpler formula—use a normal table (Appendix 5).

$$\Phi(-z) = 1 - \Phi(z) \qquad \text{(by symmetry and rule of complements)}$$

Mean: 0 **Standard deviation:** 1

Other moments:

$$E(Z^m) = 0, \quad m \text{ odd} \qquad \text{(by symmetry)}$$
$$E(Z^{2m}) = \frac{(2m)!}{m!2^m}, \quad m = 0, 1, 2, \ldots$$
$$E(|Z|) = \sqrt{\frac{2}{\pi}}$$
$$E(e^{tZ}) = e^{t^2/2}$$

Sources:

- Approximation to standardized sums of independent random variables (2.2, 3.3, 4.1, 5.4).
- Standardized normal (μ, σ^2) (4.1).
- Approximation to binomial, Poisson, negative binomial, gamma, beta. See summaries of these distributions for conditions under which the approximation is good.

Transformations: (5.3, 5.4) (*Notation:* $X \sim F$ means X is a random variable with distribution F.) Let $Z_1, Z_2, \ldots$ be independent standard normal.

Linear: $(Z_1 + Z_2)/\sqrt{2} \sim$ standard normal.

$$\sum_i \alpha_i Z_i \sim \text{standard normal} \quad \text{iff} \quad \sum_i \alpha_i^2 = 1 \qquad \text{(rotational symmetry)}$$

Quadratic: $Z^2 \sim$ gamma $(1/2, 1/2)$

$$Z_1^2 + Z_2^2 + \cdots + Z_n^2 \sim \text{gamma } (n/2, 1/2) \equiv \text{chi-square } (n)$$

Ratios: $Z_1/Z_2 \sim$ Cauchy $(0, 1)$

Normal (μ, σ^2)

$$X = \mu + \sigma Z \quad \text{where} \quad Z \sim \text{normal } (0, 1)$$

Note: All formulae follow from this linear change of variable.

Mean: μ **Standard deviation:** σ

Density function: (4.1)

$$P(X \in dx)/dx = \frac{1}{\sigma}\phi\left((x - \mu)/\sigma\right) \quad (-\infty < x < \infty)$$

Cumulative distribution function:

$$P(X \le x) = \Phi\left(\frac{x - \mu}{\sigma}\right) \quad (-\infty < x < \infty)$$

Sources:

- Approximation to distribution of heights, weights, etc., over human and biological populations.
- Measurement errors.
- Random fluctuations.

Sums: (5.4)
If X_i are independent normal (μ_i, σ_i^2), then $\sum_i X_i$ is normal $\left(\sum_i \mu_i, \sum_i \sigma_i^2\right)$.

Bivariate normal (6.5)

X and Y have standard bivariate normal distribution with correlation ρ if and only if

$$Y = \rho X + \sqrt{1 - \rho^2}\, Z$$

where X and Z are independent standard normal variables.

Poisson

Parameter: $\mu =$ mean number.

Range: $k \in \{0, 1, 2, \ldots\}$

Probability function: (2.4, 3.5)

$$P(k) = P(N_\mu = k) = e^{-\mu}\mu^k/k! \quad (k = 0, 1, \ldots)$$

where N_μ is the number of arrivals in a given time period in a Poisson arrival process, or the number of points in a given area in a Poisson random scatter, when the expected number is μ.

Mean and standard deviation: (3.5)

$$E(N_\mu) = \mu \qquad SD(N_\mu) = \sqrt{\mu}$$

Sources:

- Poisson process.
- Approximation to binomial as $p \to 0$ with $\mu = np$ (2.4).
- From independent exponential variables $W_1, W_2, \ldots$ with rate 1. Let

$$N_\mu = \{\text{first } n \text{ such that } W_1 + \cdots + W_n > \mu\} - 1$$

Transformations: (3.5)

Sums: Let $N_{\mu_1}, N_{\mu_2}, \ldots, N_{\mu_n}$ be independent Poisson. Their sum has Poisson distribution with parameter $\sum_{i=1}^n \mu_i$.

Poissonization of the binomial: (3.5)

If $S_N =$ number of successes and $F_N =$ number of failures in N trials, where N has Poisson (μ) distribution, and given N the trials are independent with probability p of success on each trial, then S_N and F_N are independent with

$$S_N \sim \text{Poisson } (\mu p) \qquad F_N \sim \text{Poisson } (\mu q) \qquad p + q = 1$$

Notation: $X \sim F$ means X is a random variable having distribution F. Similarly for multinomial trials. A consequence is the following:

Binomial distribution of Poisson terms given their sum:

If N_α and N_β are independent Poisson variables with mean α and β, then the conditional distribution of N_α given $N_\alpha + N_\beta = n$ is binomial $\left(n, \dfrac{\alpha}{\alpha+\beta}\right)$

Normal approximation: Good for large μ.

Uniform

Uniform distribution on a finite set: (1.1)

This is the distribution of a point picked at random from a finite set Ω, so that all points are equally likely to be picked. For $A \subset \Omega$,

$$P(A) = \frac{\#(A)}{\#(\Omega)}$$

Special cases:

- Bernoulli $(1/2)$ distribution: uniform distribution on $\{0, 1\}$.
- The number on a fair-die roll: uniform distribution on $\{1, 2, 3, 4, 5, 6\}$.
- Uniform distribution on $\{1, 2, \ldots, n\}$: Let X have uniform distribution on the integers 1 to n. Then

$$E(X) = \frac{n+1}{2} \qquad SD(X) = \sqrt{\frac{n^2 - 1}{12}}$$

Uniform distribution on an interval: (4.1)

Parameters: $a < b$, endpoints of the interval.

Range: $x \in (a, b)$ (or (a, b), or $(a, b]$ or $[a, b)$—the endpoints don't matter.)

Density function:

$$P(X \in dx)/dx = 1/(b - a) \quad a \leq x \leq b$$

The probability of any subinterval of (a, b) is proportional to the length of the subinterval.

Cumulative distribution function:

$$P(X \leq x) = \begin{cases} 0 & \text{if } x < a \\ (x-a)/(b-a) & \text{if } a \leq x \leq b \\ 1 & \text{if } x > b \end{cases}$$

Mean and standard deviation:

$$E(X) = (a + b)/2 \qquad SD(X) = \sqrt{\frac{(b-a)^2}{12}}$$

Transformations:

Linear: (4.4)

If X has uniform (a, b) distribution, then for constants $c > 0$ and d, the random variable $Y = cX + d$ has uniform distribution on $(ca + d, cb + d)$.

Cumulative distribution functions: (4.5)

Let X be a random variable with continuous c.d.f. F. Then the random variable $F(X)$ has uniform $(0, 1)$ distribution.

Inverse cumulative distribution functions:

Let U have uniform distribution on $(0, 1)$, and let F be any cumulative distribution function. Then $F^{-1}(U)$ is a random variable whose c.d.f. is F.

This allows random numbers with any given distribution to be generated from uniform $(0, 1)$ random numbers.

Order statistics: (4.6)

The kth order statistic of n independent uniform $(0, 1)$ random variables has beta distribution with parameters k and $n - k + 1$.

Sums: (5.4)

The density of the sum of n independent uniform $(0, 1)$ random variables is defined by polynomials of degree $n - 1$ on each of the intervals $[0, 1)$, $[1, 2)$, $\ldots$, $(n - 1, n)$.

For $n \geq 5$, the distribution of the standardized sum is very well approximated by the standard normal distribution.

Uniform distribution on a region in the plane (5.1)

A random point (X, Y) has uniform distribution on a region D of the plane, where D has finite area, if:

(i) (X, Y) is certain to lie in D

(ii) the chance that (X, Y) falls into a subregion C of D is proportional to the area of C:

$$P[(X, Y) \in C] = \text{area}(C)/\text{area}(D) \quad C \subset D$$

If X and Y are independent random variables, each uniformly distributed on an interval, then (X, Y) is uniformly distributed on the rectangle (range of X) $\times$ (range of Y).

Suppose (X, Y) is uniformly distributed on a region D in the plane. Then given $X = x$, Y has uniform distribution on the values of Y which are possible when $X = x$.

Uniform distributions on regions in three or higher dimensions have similar properties, with volumes replacing lengths and areas.

Examinations

Midterm Examination 1 (1 hour)

1. Ten dice are rolled. Five dice are red and five are green. Write down numerical expressions for:

a) The probability of the event that exactly four of the ten dice are sixes.

b) The probability of the event that exactly two of the red dice are sixes and exactly three of the green dice show even numbers.

c) The probability that there are the same number of sixes among the red dice as among the green dice.

d) The probability that there are strictly more sixes among the red dice than among the green dice.

2. Five cards are dealt from a standard deck of 52. Write down numerical expressions for

a) The probability that the third card is an ace.

b) The probability that the third card is an ace given that the last two cards are not aces.

c) The probability that all cards are of the same suit.

d) The probability of two or more aces.

3. A student takes a multiple choice examination where each question has 5 possible answers. He works a question correctly if he knows the answer, otherwise he guesses at random. Suppose he knows the answer to 70% of the questions.

a) What is the probability that on a question chosen at random the student gets the correct answer?

b) Given that the student gets the correct answer to this question chosen at random, what is the probability that he actually knew the answer?

Suppose there are 20 questions on the examination. Let N be the number of questions that the student gets correct.

c) Find $E(N)$. d) Find $SD(N)$.

4. Let A, B, and C be events which are mutually independent, with probabilities a, b, and c. Let N be the random number of events which occur.

a) What is the event $(N = 2)$ in terms of A, B and C?

b) What is the probability of this event in terms of a, b, and c?

c) What is $E(N)$ in terms of a, b, and c?

d) What is $SD(N)$ in terms of a, b, and c?

5. Let X_2 and X_3 be indicators of independent events with probabilities $\frac{1}{2}$ and $\frac{1}{3}$, respectively.

a) Display the joint distribution table of $X_2 + X_3$ and $X_2 - X_3$.

b) Calculate $E(X_2 - X_3)$.

c) Calculate $SD(X_2 - X_3)$.

Midterm Examination 2 (1 hour)

1. **Coin spinning.** I have two coins. One shows heads with probability $1/10$ when spun. The other shows heads with probability $1/2$. Suppose you pick one of my two coins at random and spin it twice. Find:

   a) $P(\text{heads on first spin})$;

   b) $P(\text{heads on second spin})$;

   c) $P(\text{heads on both spins})$;

   d) the probability that the coin is the $1/2$ coin given heads on both spins.

2. **True or false.** A student answers a set of 100 true/false questions by answering 36 questions correctly, and guessing the other 64 at random.

   a) If the pass mark is 70 questions correct, what is the student's chance of passing? Give your answer as a decimal correct to two places.

   b) Another student also knows 36 correct answers and guesses the rest at random. What is the chance that just one of these two students passes?

3. **Rare white balls**. A box contains 998 black and 2 white balls. Let $X =$ the number of whites in 500 random draws *with* replacement from this box. Calculate:

   a) $P(X = 1)/P(X = 2)$;

   b) $P(X = 1 \text{ given } X = 1 \text{ or } 2)$;

   c) repeat b) assuming draws without replacement.

4. **Reliability**. A system consists of four components which work independently with probabilities 0.9, 0.8, 0.7, and 0.6. Let $X =$ the number of components that work. Find:

   a) $E(X)$;

   b) $SD(X)$;

   c) $P(X > 0)$;

   d) $P(X = 2)$.

Final Examination 1 (3 hours)

1. A random variable N is uniformly distributed on $\{1, 2, \ldots, 10\}$. Let X be the indicator of the event $(N \leq 5)$ and Y the indicator of the event (N is even).

   a) Find $E(X)$ and $E(Y)$.

   b) Are X and Y independent?

   c) Find $Cov(X, Y)$.

   d) Find $E\big[(X+Y)^2\big]$.

2. A box contains 5 tickets. An unknown number of them are red, the rest are green. Suppose that to start off with you think there are equally likely to be 0, 1, 2, 3, 4, or 5 red tickets in the box.

   a) Three tickets are drawn from the box with replacement between draws. The tickets drawn are red, green, and red. Given this information, what is the chance that there are actually 3 red tickets in the box?

   b) What would your answer to (a) be if you knew the draws were made without replacement?

3. In the World Series, two teams play a series of games, and the first team to win four games wins the series. Suppose that each game ends in either a win or a loss for your team, and that for each game that is played the chance of a win for your team is p, independently of what happens in other games. What is the probability that your team wins the series?

4. Let X, Y, and Z be three independent normal $(0, 1)$ random variables. Calculate:

   a) $P(|X| \leq 1, |Y| \leq 2, |Z| \leq 3)$;

   b) $E\left[(X+Y+Z)^2\right]$;

   c) $P(X+Y \leq 2Z)$.

5. Suppose that T is a random variable such that $P(T > t) = e^{-t}, \quad t \geq 0$.

   a) Find a formula for the probability density function f_X of the random variable $X = 1/T$.

   b) What is the value of $E(X)$?

6. A fair coin is tossed 100 times. The probability of getting *exactly* 50 heads is close to one of the following numbers.

   $$0.001, \quad 0.01, \quad 0.1, \quad 0.5, \quad 0.9, \quad 0.99, \quad 0.999$$

   a) Circle which number you think is closest and explain your choice.

   b) How many times do you have to toss the coin to make the probability of getting exactly as many heads as tails very close to one tenth of this probability of getting 50 heads in 100 tosses?

7. A pair of dice is rolled n times, where n is chosen so that the chance of getting at least one double six in the n rolls is very close to $1/2$.

a) The number of rolls n must be very close to one of the following numbers:

$$6, \quad 12, \quad 18, \quad 20, \quad 25, \quad 30, \quad 36, \quad 50, \quad 72, \quad 100.$$

Circle which number you think n must be close to, and explain your choice.

b) What, approximately, is the chance that you actually get two or more double sixes in this many rolls? Give your answer as a decimal.

8. Let U_1 and U_2 be two independent uniform $[0, 1]$ random variables. Let

$$X = \min(U_1, U_2)$$

$$Y = \max(U_1, U_2)$$

where $\min(u_1, u_2)$ is the smaller and $\max(u_1, u_2)$ the larger of two numbers u_1 and u_2. Find:

a) the probability density function f_X of X;

b) the joint density function $f_{X,Y}$ of (X, Y);

c) $P(X \leq 1/2 | Y \geq 1/2)$.

9. Suppose that on average one person in a hundred has a particular genetic defect, which can be detected only by a laboratory test.

a) Suppose fifty people chosen at random are tested. What is the probability that at least one of them will have the defect? [Answer as a decimal.]

b) About how many people have to be tested in order for the probability to be at least 99% that at least one person has the defect?

c) If this number of people are tested, what is the expected number of individuals with the defect?

10. Let $U_1, U_2, \ldots, U_n$ be independent uniform $[0, 1]$ random variables. If n is large the geometric mean G_n of $U_1, U_2, \ldots, U_n$, defined by $G_n = (U_1 U_2 \ldots U_n)^{1/n}$, is most likely to be very close to a certain number g. Explain why, and find g.
[*Hint*: Use logarithms.]

Final Examination 2 (3 hours)

1. Suppose one morning you pick two eggs for lunch at random from a dozen eggs in your refrigerator, thinking that they are all hard-boiled. You then learn that in fact four of the eggs have not been hard-boiled.

a) What is the probability that your two lunch eggs are both hard-boiled?

b) Given that you crack one of your lunch eggs and find it is hard-boiled, what is the probability that the second egg is hard boiled?

2. A hat contains n coins, f of which are fair, and b of which are biased to land with heads with probability $2/3$, with $f+b=n$. A coin is drawn at random from the hat and tossed once. It lands heads. What is the probability that it is a biased coin?

3. A die has one spot painted on one face, two spots painted on each of two faces, and three spots painted on each of three faces. The die is rolled twice.

a) Calculate the distribution of the sum S_2 of the numbers on the two rolls. Display your answer in a table.

b) Calculate the numerical value of $E(S_2)$ in two different ways to check your answer to a).

c) Calculate the standard deviation of S_2.

4. Suppose the average family income in a particular area is $\$10,000$.

a) Find an upper bound for the fraction of families in the area with incomes over $\$50,000$.

b) Find a smaller upper bound than in a), given that the standard deviation is $8000.

c) Do you think the normal approximation would give a good estimate for the fraction in question?

5. A random variable X has probability density function of the form

$$f_X(x) = \begin{cases} cx^2, & 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

a) Find the constant c.

b) Find $P(X \le a)$ for $0 \le a \le 1$.

c) Calculate $E(X)$.

d) Calculate $SD(X)$.

6. Telephone calls arrive at an exchange at an average rate of one every second. Find the probabilities of the following events, explaining briefly your assumptions.

a) No calls arriving in a given five-second period.

b) Between four and six calls arriving in the five-second period.

c) Between 90 and 110 calls arriving in a 100-second period. (Give answer as a decimal.)

7. Let T be the number of times you have to roll a die before each face has appeared at least once. Let N be the number of different faces appearing in the first six rolls. Calculate:

a) $E(T)$;

b) $E(N)$;

c) $E(T|N = 3)$.

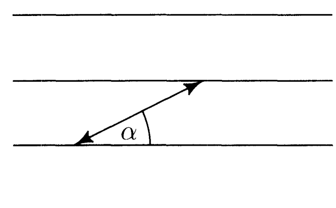
8. Let X, Y, and Z be independent standard normal random variables. Find the probability density functions of each of the following random variables:

a) X^2;

b) $X^2 + Y^2$;

c) $X + Y + Z$.

9. A floor is ruled with equally spaced parallel lines. A needle is such that if its two ends are placed on adjacent lines the angle between the needle and the lines is α, where $0 \leq \alpha \leq \pi/2$. Calculate the probability that the needle crosses at least one of the lines when tossed at random on the floor:



a) for $\alpha = 45°$;

b) for a general α between 0 and $\pi/2$.

10. A fair coin is tossed $2n$ times. Let p_{2n} be the probability of getting the same number of heads as tails.

a) Find constants a and b such that

$$p_{2n} \sim \frac{1}{an^b} \qquad \text{as } n \to \infty.$$

b) Show that $p_{2n} \to 0 \quad$ as $n \to \infty$.

c) Why does this not contradict the law of large numbers?

Final Examination 3 (3 hours)

1. Suppose you try 5 times to hit the bull's eye. The first time you have a 0.2 chance of a hit, and each time you try your chance of hitting increases by 0.1. Let H be the number of hits in the five attempts. Assuming your attempts are independent, calculate the following quantities. Answers should be decimals.

a) $E(H)$; b) $Var(H)$.

2. Suppose that in a network of 3 computers, at a given time the event that the kth computer is down has overall probability p_k, $k = 1, 2, 3$. Calculate the probability that at this time there is at least one computer up:

a) assuming the computers are up or down independently of each other;

b) assuming that there is probability p of power failure, in which case all the computers are down, but given that there is no power failure the computers are up or down independently of each other.

3. A fair six-sided die has:

the 1 spot face opposite the 6 spot face;
the 2 spot face opposite the 5 spot face;
the 3 spot face opposite the 4 spot face.

Suppose the die is rolled once. Let X be the number of spots showing on top, Y the number of spots showing on one of its side faces, say the leftmost face from a particular point of view.

a) Display the joint probability distribution of X and Y in a suitable table.

b) Are X and Y independent?

c) Find $Cov(X, Y)$.

d) Find $Var(X + Y)$.

4. Suppose there are 50 married couples. After some years, 20 of these 100 people have died. Regard the 20 deaths as striking the 100 people at random. Find numerical expressions for:

a) the probability that a particular couple has survived;

b) the expected number of couples surviving;

c) the probability that two particular couples have survived;

d) the variance of the number of couples surviving.

5. Suppose X and Y are independent random variables, each uniformly distributed on $[0, 1]$. Calculate:

a) $P(X^2 + Y^2 \le 1)$; b) $P(Y^2 > 3X^2)$; c) $P(X^2 + Y^2 \le 1$ given $Y^2 > 3X^2)$.

6. Suppose a particle has velocity V which is normally distributed with mean 0 and variance σ^2. Let $X = mV^2/2$ where $m > 0$ is a positive constant. Find formulae in terms of m and σ for;

a) $E(X)$; b) the probability density function of X; c) $Var(X)$.

7. A particle counter records two types of particles, Types 1 and 2. Type 1 particles arrive at an average rate of 1 per minute, Type 2's at an average rate of 2 per minute. Assume these are two independent Poisson processes. Give numerical expressions for the following probabilities:

a) Three Type 1 particles and four of Type 2's arrive in a two-minute period;

b) the total number of particles of either type in a two-minute period is 5;

c) the fourth particle arrives in the first 5 minutes;

d) the first particle to arrive is of Type 1;

e) the second particle of Type 1 turns up before the third of Type 2.

8. Consider the average $\bar{X}_n = (X_1 + X_2 + \cdots + X_n)/n$ of n independent random variables, each uniformly distributed on $[0, 1]$. Find n so that $P(\bar{X}_n < 0.51)$ is approximately 90%.

9. Two statisticians are watching a sequence of independent Bernoulli trials with probability p of success on each trial. The first statistician estimates p by the proportion of successes in the first 100 trials. The second statistician estimates p by the proportion of successes in the next 300 trials. Consider the probability that the second estimate is closer to p than the first.

a) Explain why this probability hardly depends at all on p, provided p is fairly close to $1/2$.

b) Assuming p is fairly close to $1/2$, this probability is very close to one of the following numbers:

$$0, \quad 1/10, \quad 1/5, \quad 1/4, \quad 1/3, \quad 1/2, \quad 2/3, \quad 3/4, \quad 4/5, \quad 9/10, \quad 1.$$

Which one, and why?

10. Suppose 10 dice are shaken together and rolled. Any that turn up six are set aside. The remaining dice are shaken and rolled again. Any of these that turn up six are set aside. And so on, until all the dice show six. Let N be the number of times the dice are shaken and rolled. To illustrate, if after the first roll of 10 dice, 7 non-sixes remain, and after the second roll of these 7 dice 2 non-sixes remain, and after the third roll of these 2 dice no non-sixes remain, then $N = 3$.

a) Describe the distribution of N.

[*Hint:* Consider the number of times each die is rolled.]

b) Let T be the total number of individual die rolls. To illustrate, $T = 10 + 7 + 2 = 19$ for the outcome described above. Describe the distribution of T.

c) Let L be the number of dice shaken on the last roll. To illustrate, $L = 2$ for the outcome described above. Describe the distribution of L.

Midterm Examination 1—Solutions

1. a) $\binom{10}{4}(1/6)^4(5/6)^6$

b) $\binom{5}{2}(1/6)^2(5/6)^3\binom{5}{3}(1/2)^5$

c) $\binom{5}{0}^2(5/6)^{10}+\binom{5}{1}^2(1/6)^2(5/6)^8+\binom{5}{2}^2(1/6)^4(5/6)^6+\cdots+\binom{5}{5}^2(1/6)^{10}$

d) $\frac{1}{2}$(1 − answer to c))

2. a) $1/13$ b) $4/50$

c) $\dfrac{4\times\binom{13}{5}}{\binom{52}{5}}$ d) $1-\dfrac{\binom{4}{0}\binom{48}{5}+\binom{4}{1}\binom{48}{4}}{\binom{52}{5}}$

3. a) $0.7+(0.3)(0.2)=0.76$ b) $\dfrac{0.7}{0.76}$

c) 20×0.76 d) $\sqrt{6\times\frac{1}{5}\times\frac{4}{5}}$

4. a) $ABC^c\cup AB^cC\cup A^cBC$

b) $ab(1-c)+a(1-b)c+(1-a)bc$

c) $a+b+c$

d) $\sqrt{a(1-a)+b(1-b)+c(1-c)}$

5. a)

X_2-X_3	X_2+X_3: 0	1	2
-1	0	1/6	0
0	1/3	0	1/6
1	0	1/3	0

b) $\dfrac{1}{6}$

c) $\sqrt{\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{3}\cdot\frac{2}{3}}$

Midterm Examination 2—Solutions

1. a) $P(H_1) = P(1/10 \text{ coin})P(H_1 \mid 1/10 \text{ coin}) + P(1/2 \text{ coin})P(H_1 \mid 1/2 \text{ coin})$
$= \frac{1}{2} \times \frac{1}{10} + \frac{1}{2} \times \frac{1}{2} = \frac{3}{10}$

b) $P(H_2) = P(H_1) = \frac{3}{10}$

c) $P(H_1H_2) = P(1/10 \text{ coin})P(H_1H_2 \mid 1/10 \text{ coin}) + P(1/2 \text{ coin})P(H_1H_2 \mid 1/2 \text{ coin})$
$= \frac{1}{2} \times \left(\frac{1}{10}\right)^2 + \frac{1}{2} \times \left(\frac{1}{2}\right)^2 = \frac{13}{100}.$

d) $P(1/2 \text{ coin} \mid H_1H_2) = \dfrac{P(1/2 \text{ coin})P(H_1H_2 \mid 1/2 \text{ coin})}{P(H_1H_2)} = \dfrac{\frac{1}{2} \times \left(\frac{1}{2}\right)^2}{\frac{13}{100}} = \dfrac{25}{26}$

2. a) Let X be the number of correct answers in 64 questions; then X has binomial ($n = 64$, $p = 1/2$) distribution, so $EX = 32$, $SD(X) = 4$.

$$P(\text{passing}) = P(X \geq 34) = P(X \geq 33.5) = P\left(\frac{X - 32}{4} \geq \frac{33.5 - 32}{4}\right)$$
$$\approx 1 - \Phi(0.375) = 1 - 0.646 = 0.354$$

b) This is just $2pq$ for $p = 0.354$, $q = 0.646$, i.e.,

$$2pq = 2 \times 0.354 \times 0.646 = 0.457$$

3. a) X has binomial ($n = 500, p = 2/1000$) distribution, so

$$\frac{P(X=1)}{P(X=2)} = \frac{2}{n-2+1} \cdot \frac{1-p}{p} = \frac{2}{499} \cdot \frac{\frac{998}{1000}}{\frac{2}{1000}} = 2$$

b) $P(1 \mid 1 \text{ or } 2) = \dfrac{P(X=1)}{P(X=1) + P(X=2)} = \dfrac{P(X=1)/P(X=2)}{P(X=1)/P(X=2) + 1} = \dfrac{2}{3}$

c) Now $\dfrac{P(X=1)}{P(X=2)} = \dfrac{\binom{2}{1}\binom{998}{499}}{\binom{2}{2}\binom{998}{498}} = 2 \times \dfrac{498!500!}{499!499!} = 2 \times \dfrac{500}{499} = \dfrac{1000}{499}$

Continue as before: $P(X = 1 \mid 1 \text{ or } 2) = \dfrac{(1000/499)}{(1000/499) + 1} = \dfrac{1000}{1499}$

4. a) $X = X_1 + X_2 + X_3 + X_4$, where X_i is the indicator that the ith component works.
So $E(X) = \sum_{i=1}^{4} P(X_i = 1) = 0.9 + 0.8 + 0.7 + 0.6 = 3.0$

b) $Var(X) = \sum_{i=1}^{4} Var(X_i) = 0.9 \times 0.1 + 0.8 \times 0.2 + 0.7 \times 0.3 + 0.6 \times 0.4 = 0.7$
$SD(X) = 0.8367$

c) $1 - P(X = 0) = 1 - 0.1 \times 0.2 \times 0.3 \times 0.4 = 0.9976$

d) $0.9 \times 0.8 \times 0.3 \times 0.4 + 0.9 \times 0.2 \times 0.7 \times 0.4 + 0.9 \times 0.2 \times 0.3 \times 0.6 + 0.1 \times 0.8 \times 0.7 \times 0.4 + 0.1 \times 0.8 \times 0.3 \times 0.6 + 0.1 \times 0.2 \times 0.7 \times 0.6 = 0.2144$

Final Examination 1—Solutions

1. a) $E(X) = E(Y) = 1/2$.

b) No.

c) $-1/20$.

d) 1.4.

2. a) 0.36 b) 0.4

3.

$$\binom{7}{4}p^4q^3 + \binom{7}{5}p^5q^2 + \binom{7}{6}p^6q + p^7 = p^4 + 4p^4q + \binom{5}{2}p^4q^2 + \binom{6}{3}p^4q^3,$$

where $q = 1 - p$.

4. a) 0.65 b) 3 c) 1/2

5. a) $f_X(x) = \dfrac{e^{-1/x}}{x^2}, \quad x > 0, \quad 0$ otherwise.

b) $E(X) = \infty$.

6. a) 0.1, by normal approximation or Stirling's formula.

b) 10^4 tosses.

7. a) 25.

b) 0.152 (Poisson approximation)

8. a) $f_X(x) = 2 - 2x, \quad 0 \le x \le 1, \qquad 0$ otherwise.

b) $f_{X,Y}(x, y) = 2, \quad 0 \le x \le y \le 1, \qquad 0$ otherwise.

c) 2/3

9. a) $1 - e^{-1/2}$.

b) $100 \log 100$.

c) $\log 100$.

10. e^{-1}, by law of large numbers.

Final Examination 2—Solutions

1. a) $\dfrac{\binom{8}{2}\binom{4}{0}}{\binom{12}{2}} = \dfrac{56}{132} = \dfrac{14}{33} = 0.424$ (sampling without replacement) b) $7/11$

2. $\dfrac{4b}{4b+3f} = \dfrac{\frac{2}{3}b}{\frac{2}{3}b + \frac{1}{2}f}$ (Bayes' rule)

3. a)

2	3	4	5	6
1/36	4/36	10/36	12/36	9/36

b) $\frac{14}{3}$ c) $\frac{\sqrt{10}}{3}$

4. a) $\frac{10{,}000}{50{,}000} = \frac{1}{5}$ (Markov's inequality)

b) $\left(\frac{8{,}000}{40{,}000}\right)^2 = \frac{1}{25}$ (Chebychev's inequality)

c) No, because income ≥ 0.

5. a) 3 b) a^3 c) $\frac{3}{4}$ d) $\sqrt{\frac{3}{80}} \approx 0.194$

6. a) e^{-5} (Poisson process) b) $e^{-5}\left(\dfrac{5^4}{4!} + \dfrac{5^5}{5!} + \dfrac{5^6}{6!}\right)$ c) 0.68 (normal approximation)

7. a) $1 + \frac{6}{5} + \frac{6}{4} + \cdots + 6 \approx 14.7$

b) $6\left(1 - (\frac{5}{6})^6\right)$

c) $6 + \frac{6}{3} + \frac{6}{2} + 6 = 17$

8. a) gamma $(1/2, 1/2)$ b) gamma $(1, 1/2)$ c) normal $(0, 3)$

9. a) $\dfrac{2}{\pi}\left(\sqrt{2} - 1\right) + \dfrac{1}{2}$ b) $\dfrac{2}{\pi \sin\alpha}(1 - \cos\alpha) + 1 - \dfrac{2\alpha}{\pi}$

10. a) $a = \sqrt{\pi}$, $b = 1/2$

b) Follows from $\frac{1}{\sqrt{n}} \to 0$ as $n \to \infty$

c) The law of large numbers says that the proportion of heads is very likely to be very close to $1/2$, not that it is very likely to be *exactly* $1/2$.

Final Examination 3—Solutions

1. a) $0.2 + 0.3 + 0.4 + 0.5 + 0.6 = 2.0$ b) $0.2 \times 0.8 + 0.3 \times 0.7 + 0.4 \times 0.6 + 0.5 \times 0.5 + 0.6 \times 0.4 = 1.1$

2. a) $1 - P(\text{all down}) = 1 - p_1 p_2 p_3$

b) Answer is $(1-p)(1-r_1 r_2 r_3)$, where r_i is the conditional probability that computer i is down given no power failure. But $p_i = p + (1-p)r_i$, so $r_i = (p_i - p)/(1-p)$, and the answer is

$$(1-p)\left[1 - \frac{(p_1-p)(p_2-p)(p_3-p)}{(1-p)^3}\right]$$

3. a)

Values of X (columns), Values of Y (rows)

	1	2	3	4	5	6
1	0	1/24	1/24	1/24	1/24	0
2	1/24	0	1/24	1/24	0	1/24
3	1/24	1/24	0	0	1/24	1/24
4	1/24	1/24	0	0	1/24	1/24
5	1/24	0	1/24	1/24	0	1/24
6	0	1/24	1/24	1/24	1/24	0

b) No.

c) $Cov(X, Y) = 0$.

d) $Var(X+Y) = Var(X) + Var(Y) = \frac{35}{12} + \frac{35}{12} = \frac{35}{6} = 5.833$

4. a) $\frac{80}{100} \times \frac{79}{99} = 0.638$

b) Let $I_i =$ indicator that couple i has survived. Then

$$E(\#\text{ couples surviving}) = E\sum_{i=1}^{50} I_i = \sum_{i=1}^{50} E(I_i) = 50 \times 0.638 = 31.9$$

c) $\frac{80}{100} \times \frac{79}{99} \times \frac{78}{98} \times \frac{77}{97} = 0.4033 = E(I_1 I_2)$

d) $Var(S_{50}) = \sum_{i=1}^{50} Var(I_i) + 2\sum_{j<k} Cov(I_j, I_k)$; here

$$Var(I_i) = E(I_1^2) - [E(I_1)]^2 = 0.638 - 0.638^2 = 0.23085,$$

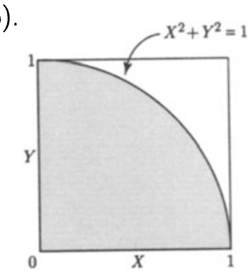
$$Cov(I_j, I_k) = Cov(I_1, I_2) = E(I_1 I_2) - E(I_1)E(I_2) = 0.4033 - 0.638^2 = -0.004196,$$

so $Var(S_{50}) = 50 Var(I_1) + 50 \cdot 49 Cov(I_1, I_2) = 50 \cdot (0.23085) + 2450(-0.004196)$.

5. a) $P(X^2 + Y^2 \le 1)$

$$= \frac{\text{area of shaded region}}{\text{area of square}}$$

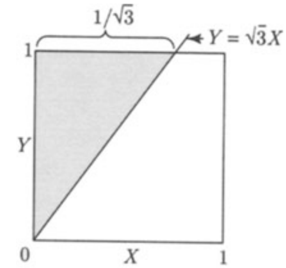
$$= \frac{\frac{1}{4}\pi(1)^2}{1^2} = \frac{\pi}{4} = 0.785$$



b) $P(Y^2 > 3X^2)$

$$= \frac{\text{area of shaded region}}{\text{area of square}}$$

$$= \frac{\frac{1}{2}(1)\left(\frac{1}{\sqrt{3}}\right)}{1} = \frac{1}{2\sqrt{3}} = 0.2887$$

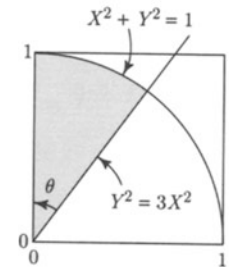


c) $\tan\theta = \frac{1}{\sqrt{3}}$, so $\theta = 30°$, and

$$P(X^2 + Y^2 \leq 1 \text{ given } Y^2 > 3X^2\ )$$

$$= \frac{\text{area of shaded region}}{\text{area of shaded region in (b)}}$$

$$= \frac{\frac{1}{4}\pi\left(\frac{1}{3}\right)}{\frac{1}{2\sqrt{3}}} = \frac{\sqrt{3}\pi}{6} = \frac{\pi}{2\sqrt{3}} = 0.91$$



6. a) V^2 has gamma $\left(\frac{1}{2}, \frac{1}{2\sigma^2}\right)$ distribution, so

$$E(V^2) = \frac{\frac{1}{2}}{\frac{1}{2\sigma^2}} = \sigma^2 \text{ and } Var(V^2) = \frac{\frac{1}{2}}{\left(\frac{1}{2\sigma^2}\right)^2} = 2\sigma^4.$$

So $E(X) = E\left(\frac{1}{2}mV^2\right) = \frac{1}{2}m\sigma^2$.

b) $X = \frac{1}{2}mV^2$; $\quad V = \sqrt{\frac{2X}{m}}$;

$$\frac{dV}{dX} = \sqrt{\frac{2}{m}}\left(\frac{1}{2\sqrt{X}}\right) = \frac{1}{\sqrt{2mX}}$$

$$f_V(v) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{v^2}{2\sigma^2}}$$

$$\Rightarrow \quad f_X(x) = 2f_V(v)\left|\frac{dv}{dx}\right| = 2\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-\left(\frac{2x}{m}\right)}{2\sigma^2}}\left(\frac{1}{\sqrt{2mx}}\right) = \frac{1}{\sqrt{\pi m x}\sigma}e^{-x/m\sigma^2}$$

c) By part a), $Var(X) = Var\left(\frac{1}{2}mV^2\right) = \frac{m^2}{4}2\sigma^4 = m^2\sigma^4/2$

7. a) $P(3 \text{ of Type } 1)P(4 \text{ of Type } 2) = e^{-2}\frac{2^3}{3!}\cdot e^{-4}\frac{4^4}{4!}$

b) $e^{-6}6^5/5!$, since the total number of particles of either type in a two-minute period has Poisson $[2(1+2)]$ distribution.

c) $P(T_4 \leq 5) = P(N_5 \geq 4) = 1 - P(N_5 < 4) = 1 - e^{-15}\left(1 + 15 + \frac{15^2}{2!} + \frac{15^3}{3!}\right)$.

d) 1/3

e) $P(\text{3rd of Type 2 after 2nd of Type 1})$
$= P(\text{3rd of Type 2 at or after 5th of either type})$
$= P(\text{0 or 1 or 2 Type 2's in first 4 particles})$
$= (1/3)^4 + 4(1/3)^3\cdot(2/3) + 6(1/3)^2\cdot(2/3)^2 = \frac{1+8+24}{81} = \frac{11}{27}$

8. $\bar{X}_n$ has approximately normal $(0.5, \frac{1}{12n})$ distribution. Want n so that

$$0.9 = P(\bar{X}_n < 0.51) \approx \Phi\left(\frac{0.01}{\sqrt{1/12n}}\right)$$

$$\iff \frac{0.01}{\sqrt{1/12n}} \approx 1.29$$

$$\iff \sqrt{12n} \approx \frac{1.29}{0.01} \iff n \approx \left(\frac{1.29}{0.01}\right)^2 \cdot \frac{1}{12} \iff n \approx 1387$$

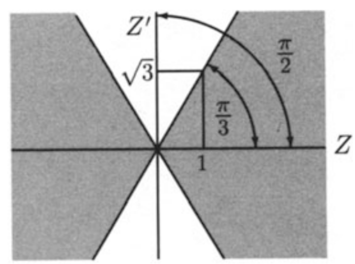
9. a) By the normal approximation

$$P\left(|\bar{X}_{100} - p| > |\bar{Y}_{300} - p|\right) = P\left(\frac{|\bar{X}_{100} - p|}{\frac{\sqrt{pq}}{10}} > \frac{|\bar{Y}_{300} - p|}{\frac{\sqrt{pq}}{\sqrt{3}\cdot 10}}\frac{1}{\sqrt{3}}\right) \approx P\left(|Z| > \frac{|Z'|}{\sqrt{3}}\right)$$

where Z and Z' are independent normal $(0, 1)$.

b) 2/3. Reason: by circular symmetry of (Z, Z'), the desired probability is

$$\frac{\arctan(\sqrt{3})}{\pi/2} = \frac{\pi/3}{\pi/2} = \frac{2}{3}$$



10. a) Let W_i be the number of times die i is rolled. Then $N = \max_i W_i$, and the W_i are independent with geometric $(1/6)$ distribution, so

$$P(N \le n) = P(W_i \le n \text{ for } i = 1 \text{ to } 10) = P(W \le n)^{10} = \left[1 - \left(\frac{5}{6}\right)^n\right]^{10}$$

$$P(N = n) = P(N \le n) - P(N \le n - 1).$$

b) $T = W_1 + \cdots + W_{10}$, so $T - 10$ has negative binomial $(10, 1/6)$ distribution

$$P(T = t) = \binom{t-1}{9}\left(\frac{1}{6}\right)^{10}\left(\frac{5}{6}\right)^{t-10}$$

c) Use $P(L = l) = \sum_{k=1}^{\infty} P(N = k, L = l)$ with

$$P(N = k, L = l) = P(10 - l \text{ dice have fallen 6 by roll } k - 1, l \text{ dice fall 6 on roll } k)$$

$$= \binom{10}{l}[P(W \le k - 1)]^{10-l}[P(W = k)]^l$$

$$= \binom{10}{l}\left[1 - \left(\frac{5}{6}\right)^{k-1}\right]^{10-l}\left[\frac{1}{6}\left(\frac{5}{6}\right)^{k-1}\right]^l$$

Appendices

Appendix 1
Counting

Basic Rules

Let $\#(B)$ denote the number of elements in a finite set B. There are three basic rules to help evaluate $\#(B)$, the *correspondence rule*, the *addition rule*, and the *multiplication rule*. The first of these is the basis of counting on your fingers:

The Correspondence Rule

If the elements of B can be put in one-to-one correspondence with the elements of another set C, then $\#(B) = \#(C)$.

The trick to using this rule is to find a one-to-one correspondence between a set you are trying to count, and some other set you already know how to count. See examples below.

The Addition Rule

If B can be split into disjoint sets $B_1, B_2, \ldots, B_n$, then

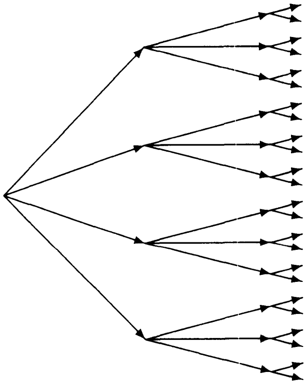
$$\#(B) = \#(B_1) + \#(B_2) + \cdots + \#(B_n).$$

The multiplication rule is generally less applicable, but nonetheless extremely useful. The number of elements of a set B is the number of different ways in which an element B may be chosen. In many problems it is possible to regard the choice of an element in B as being made by stages. For example, if B is a set of sequences, the choice of a sequence in B can be made by choosing the first element of the sequence, then the second element, and so on. The number of elements in B is then equal to the number of ways of making the successive choices. If there are k choices to be made, and at each stage $j \leq k$ there are n_j possible choices available, where n_j does not depend on what choices were made previously, then $\#(B)$ is equal to the product $n_1 n_2 \cdots n_k$, by the following rule:

The Multiplication Rule

Suppose that k successive choices are to be made, with exactly n_j choices available at each stage $j \leq k$, no matter what choices have been made at previous stages. Then the total number of successive choices which can be made is $n_1 n_2 \cdots n_k$.

The same rule can be expressed in other words. For example, *choices* can be replaced by *decisions* or *selections.* Notice that exactly which choices are available at stage j may depend on what choices have been made earlier, provided that the *number* of these choices available, that is n_j, does not. The multiplication rule can be proved by mathematical induction, using the addition rule. A good way to visualize the setup for the multiplication rule is to think in terms of a *decision tree*, where at each stage the decision is which branch of the tree to follow. Here is a decision tree representing $k = 3$ successive choices, with $n_1 = 4$ choices available at stage 1, $n_2 = 3$ choices at stage 2, and $n_3 = 2$ choices at stage 3. The number of possible successive choices is the total number of paths through the tree. In accordance with the multiplication rule, there are $4 \times 3 \times 2 = 24$ paths.



Sequences, Orderings, and Combinations

Let S be a finite set. A *sequence of length k of elements of S* is an ordered k-tuple $(s_1, s_2, \ldots, s_k)$ with $s_j \in S$ for each j. If S has n elements, the first element of the sequence can be chosen in n ways. However the first is chosen, the second can be chosen in n ways, making $n \times n = n^2$ ways to choose the first two elements of the sequence. Whichever one of those n^2 choices is made, there are n ways to choose a third element, so $n^2 \times n = n^3$ ways to choose the first three elements. This is the multiplication rule in action. Continuing in this way gives the following:

Formula for Number of Sequences

The number of sequences of length k from a set of n elements is n^k.

Example 1.

Let S be the alphabet, that is, the set of 26 letters {a, b, ..., z}. Call a sequence of five letters, such as "aargh", a *five-letter word*, no matter whether it is meaningful or not. Define a *k-letter word* similarly. There are $n = 26$ letters in the alphabet S. Hence there are

26 one-letter words,

$26^2 = 26 \times 26$ two-letter words,

$26^3 = 26 \times 26 \times 26$ three-letter words, and so on. In general, there are

$26^k = \underbrace{26 \times 26 \times \cdots \times 26}_{k \text{ factors}}$ k-letter words.

An *ordering* or *permutation of k elements of S* is a sequence of length k of elements of S with no duplications. That is to say, an arrangement of k distinct elements of S. If S has n elements, the first element in an ordering can be chosen in n ways. However this choice is made, the second element can be any one of the $n - 1$ remaining elements. So there are $n(n-1)$ ways to choose the first two elements in an ordering. Whichever one of these choices is made, there are $n - 2$ remaining elements from which to choose a third element for the permutation, so $n(n-1)(n-2)$ ways to choose the first three elements in an ordering. Continue in this way, choosing one element of the ordering at a time from among the remaining possibilities, and use the multiplication rule to obtain the following:

Formula for Number of Orderings

The number of orderings of k out of n elements is

$$n(n-1)(n-2)\cdots(n-k+1).$$

The product of k decreasing factors $n(n-1)(n-2)\cdots(n-k+1)$ is denoted $(n)_k$, a symbol which may be read "n order k". It is the number of ways of ordering k out of n elements. Alternative notations for $(n)_k$, found in some other texts, are ${}_nP_k$ and P_n^k. Compare with the larger number n^k, the number of sequences of length k, without the restriction that there be no repetitions.

Example 2.

Let S be the alphabet as in Example 1. A permutation of length k of the 26 letters of the alphabet is a word of length k with no repetitions of letters. For example, for $k = 5$, "aorgh" is such a permutation, but "aargh" and "gargh" are not. There are

26 one-letter permutations,

$(26)_2 = 26 \times 25$ two-letter permutations,

$(26)_3 = 26 \times 25 \times 24$ three-letter permutations, and so on.

In general, there are

$$(26)_k = \underbrace{26 \times 25 \times 24 \times \cdots \times (26-k+1)}_{k \text{ factors}} \quad k\text{-letter permutations.}$$

Example 3.

In the birthday problem (Section 1.6), the probability that a group of k people all have different birthdays, assuming all possible sequences of k birthdays are equally likely, is

$$\frac{(365)_k}{365^k} = \frac{365 \times 364 \times \cdots \times (365-k+1)}{365 \times 365 \times \cdots \times 365}$$

because the denominator is the number of all possible sequences of birthdays of length k, while the numerator is the number of possible sequences with no duplication, that is, the number of possible permutations of k birthdays.

Factorials

The notation $n!$ is used for

$$(n)_n = n(n-1)\cdots 2 \cdot 1$$

and by convention $0! = 1$. The symbol $n!$ is read "n factorial". By the formula for the number of permutations in the special case $k = n$,

the number of ways of ordering a set of n elements is $n!$.

Put another way,

$n!$ is the number of different ways to arrange n objects in a row

and from above

$$(n)_k = \frac{n!}{(n-k)!}$$

is the number of different ways of arranging k of these n objects in a row.

The above expression for $(n)_k$ is correct because the factor $(n-k)!$ in the denominator cancels the last $n-k$ factors in the numerator, leaving just

$$(n)_k = n \times (n-1) \times \cdots \times (n-k+1)$$

The formula works even when $k = n$, because of the convention that $0! = 1$.

A permutation is a particular kind of sequence, namely one with no repetitions. But *combination* is just another name for "subset". A *combination of k elements from a set of n elements* is a subset consisting of k of the n elements. A combination may also be called an *unordered sample*. The number of combinations of k elements from a set of n elements is denoted $\binom{n}{k}$, a symbol which is read "n choose k". This is the number of ways of choosing k out of n elements. An ordering of k of a set of n elements can be made by the following two-stage procedure:

(i) choose a combination of k elements;

(ii) order the combination.

The number of ways of making the first choice is $\binom{n}{k}$. And no matter what combination is chosen, the number of ways of ordering it is $(k)_k = k!$. Thus by the multiplication rule,

$$(n)_k = \binom{n}{k} k!$$

Dividing both sides by $k!$ yields the following basic formula:

Formula for Number of Combinations (Subsets)

The number of ways of choosing k out of n elements is

$$\binom{n}{k} = \frac{(n)_k}{k!} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1}$$

By the convention $0! = 1$, $\binom{n}{0} = 1$, every set has just one subset with no elements, the empty set. To make the second formula in terms of $(n)_k$ work in this case, make the convention that $(n)_0 = 1$.

As well as being the number of subsets of size k of a set of n elements, one-to-one correspondences show that $\binom{n}{k}$ is:

- the number of different ways to choose k places out of n places in a row;
- the number of different ways to arrange k symbols p and $n-k$ symbols q in a row.

The numbers $\binom{n}{k}$ are also called *binomial coefficients*, as they appear in the *binomial theorem*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

Number of Subsets of a Set of *n* Elements

The number of subsets of a set of n elements is 2^n.

Note that the subsets include both the empty set and the whole set. A subset may be chosen by deciding for each of the n elements whether that element should belong to the subset, or not. There are n successive choices to be made, with two possible choices at each stage. The product rule applies once more, to show that there are 2^n subsets in all. Since each subset may be classified according to its size, the number of subsets may also be expressed using the addition rule as

$$\sum_{k=0}^{n} \binom{n}{k}$$

The equality of this expression with 2^n is the binomial theorem for $x = y = 1$.

Exercises: Appendix 1

(i) Prove that

$$\binom{n}{k} = \binom{n}{n-k}$$

(a) by using the formula for $\binom{n}{k}$;

(b) by exhibiting a one-to-one correspondence between subsets of size k and subsets of size $n - k$.

(ii) Prove that

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

(a) by using the formula for $\binom{n}{k}$;

(b) by breaking subsets of size k into two mutually exclusive classes, one class comprising all those subsets which contain a given element, and the other all those which don't.

(iii) Use (i) and (ii) to generate the next two rows in the following table (called Pascal's triangle), where $\binom{n}{k}$ appears in the kth column of the nth row.

		k: 0	1	2	3	4	5
	0	1					
	1	1	1				
	2	1	2	1			
n	3	1	3	3	1		
	4	?	?	?	?	?	
	5	?	?	?	?	?	?

(iv) Check that the formula

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n$$

holds for rows $n = 0$ to 5 in Pascal's triangle. (If it doesn't work for $n = 4$ or 5, go back and redo (iii)!)

(v) Prove the formula of (iv) using (ii).

(a) using (ii);

(b) by proving that both sides of the formula represent the ***number of subsets of a set of n elements.*** For the left side use the ***addition rule*** for counting after partitioning the collection of all subsets according to size. And for the right side use the ***product rule*** for counting after identifying a subset $A \subset \{1, 2, \dots, n\}$ with the sequence of zeros and ones which is the *indicator* of A.

(vi) Prove that

$$\binom{2n}{n} = \sum_{k=0}^{n} \binom{n}{k}\binom{n}{n-k} = \sum_{k=0}^{n} \binom{n}{k}^2$$

(vii) Find a formula for the number of sequences of 0's and 1's of length n such that the sum of the 0's and 1's in the sequence is k. [*Hint*: choose the places for the 1's.]

(viii) a) Prove that for $k_0 + k_1 + k_2 = n$, the number of sequences of 0's, 1's, and 2's of length n which contain exactly k_0 0's, k_1 1's and k_2 2's is $\dfrac{n!}{k_0!k_1!k_2!}$

(b) Generalize your formula to find the number of sequences of the numbers 0, 1, 2, ..., m of length n in which the number j appears k_j times. These numbers are called *multinomial coefficients.*

(ix) Prove the binomial theorem by counting the number of terms of the form $x^k y^{n-k}$ in the expansion of $(x + y)^n$.

(x) How many different eleven-letter words (not necessarily pronounceable or meaningful!) can be made from the letters in the word MISSISSIPPI?

(xi) How many different 5-card poker hands can be dealt from a regular 52-card deck?

(xii) How many of these hands contain no aces?

(xiii) How many contain a aces, for $a = 0$ to 4?

(xiv) How many contain all cards of the same suit?

Appendix 2
Sums

The symbol $\sum_{i=1}^{n} a_i$ stands for the sum of the terms a_i from $i = 1$ to n, also denoted

$$a_1 + a_2 + \cdots + a_n$$

Note that the symbol i is an *index* or *dummy variable*. It can be replaced by any other symbol without changing the value of the sum. So

$$\sum_{i=1}^{n} a_i = \sum_{j=1}^{n} a_j$$

Sums are often made over other index sets than the first n integers. For example,

$$\sum_{i=3}^{5} a_i = a_3 + a_4 + a_5$$

If the range of i is clear from the context, a sum may be written simply $\sum_i a_i$.

General Properties of Sums

All sums are assumed to be over the same range of i:

Constants: If $x_i = c$ for every i, then

$$\sum_i x_i = (\text{number of terms}) \times c$$

Indicators: If $x_i = 0$ or 1 for every i, then

$$\sum_i x_i = (\text{number of } i \text{ such that } x_i = 1)$$

Constant factors:

$$\sum_i c x_i = c \sum_i x_i$$

Addition:

$$\sum_i (x_i + y_i) = \sum_i x_i + \sum_i y_i$$

Inequalities: If $x_i \le y_i$ for every i, then

$$\sum x_i \le \sum y_i$$

Particular Sums

$$1 + 2 + \cdots + n = \sum_{i=1}^{n} i = n(n+1)/2$$

Provided $R \neq 1$, $$1 + R + R^2 + \cdots + R^n = \sum_{i=0}^{n} R^i = \frac{1 - R^{n+1}}{1 - R}$$

Appendix 3
Calculus

Infinite Series

Let $a_1, a_2, \ldots$ be a sequence of numbers. The infinite sum

$$\sum_{i=1}^{\infty} a_i = a_1 + a_2 + \cdots$$

is called an *infinite series.*

The finite sum

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + \cdots + a_n$$

is called the nth *partial sum* of the a's.

Convergence of Infinite Series

The series $\sum_{i=1}^{\infty} a_i$ *converges* if the sequence of partial sums converges to a finite limit, that is, if $\lim_{n\to\infty} \sum_{i=1}^{n} a_i$ exists and is finite.

The series $\sum_{i=1}^{\infty} a_i$ *diverges* if the sequence of partial sums does not converge to a finite limit, that is, if $\lim_{n\to\infty} \sum_{i=1}^{n} a_i$ either does not exist or is infinite.

If $a_1, a_2, \ldots$ are all positive, then the sequence of partial sums is increasing, and thus has a limit, though the limit may be $+\infty$. So the series $\sum_{i=1}^{\infty} a_i$ either converges, or diverges to $+\infty$.

Some Common Infinite Series

$$\sum_{n=1}^{\infty} \frac{1}{n} = +\infty$$

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$$

Geometric series: If $|r| < 1$, $\displaystyle\sum_{i=i_0}^{\infty} r^i = \frac{r^{i_0}}{1-r} = \frac{\text{first term}}{1 - \text{common ratio}}$

Exponential series: $\displaystyle\sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x$ (see Appendix 4)

Derivatives

The function $f(x)$ is said to be *differentiable at* x_0 if

$$\lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

exists. In that case, the ***derivative of $f(x)$ at x_0*** is defined as the limit

$$f'(x_0) = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

If $f(x)$ is differentiable at every x_0 in its domain, then $f(x)$ is called ***differentiable.***

Interpretations of the derivative

The derivative $f'(x_0)$ may be interpreted as

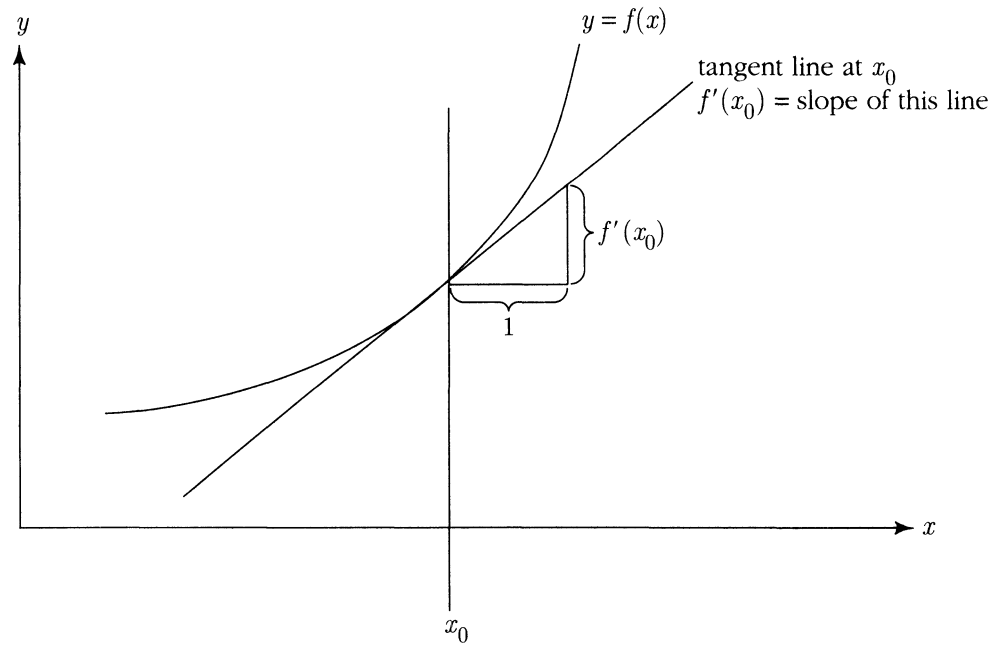
the ***rate of change*** of $f(x)$ at x_0,

or

the ***slope*** of the graph of $f(x)$ at x_0.

If $y = f(x)$, the derivative function $f'(x)$ is often written as

$$\frac{dy}{dx} \quad \text{or} \quad \frac{d}{dx} f(x)$$



Properties of Derivatives

Constants: If $f(x) = c$ for all x, then $f'(x) = 0$ for all x.

Constant factors: $(cf)'(x) = c(f'(x))$

Addition: $(f+g)'(x) = f'(x) + g'(x)$

Multiplication: $(fg)'(x) = f'(x)g(x) + f(x)g'(x)$

Chain rule: $\dfrac{d}{dx} f(g(x)) = f'(g(x))g'(x)$

Some Common Derivatives

$$\frac{d}{dx}x^n = nx^{n-1}, \qquad n = 1, 2, \ldots$$

$$\frac{d}{dx}\log(x) = \frac{1}{x}, \qquad x > 0$$

$$\frac{d}{dx}e^{\beta x} = \beta e^{\beta x}$$

$$\frac{d}{d\theta}\sin(\theta) = \cos(\theta)$$

$$\frac{d}{d\theta}\cos(\theta) = -\sin(\theta)$$

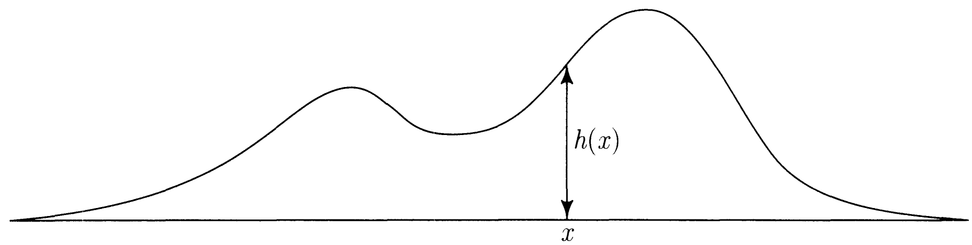
$$\frac{d}{dx}\arcsin(x) = \frac{1}{\sqrt{1-x^2}}, \qquad |x| < 1$$

$$\frac{d}{dx}\arccos(x) = \frac{-1}{\sqrt{1-x^2}}, \qquad |x| < 1$$

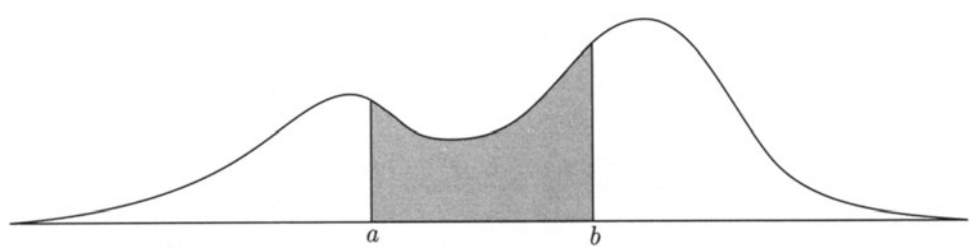
$$\frac{d}{dx}\arctan(x) = \frac{1}{1+x^2}$$

Integrals

Consider a non-negative function $h(x)$ defined for x on the line $(-\infty, \infty)$. For example, $h(x)$ might be the following curve.

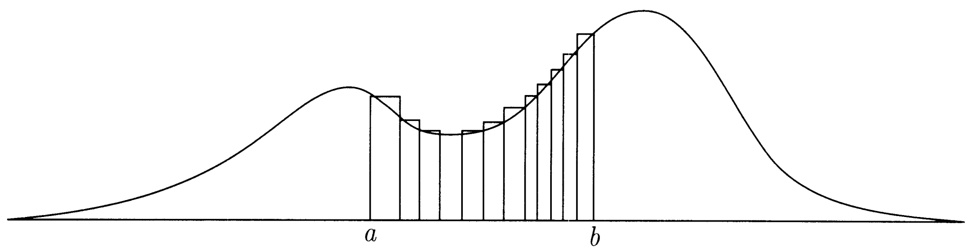


The area under the graph of $h(x)$ between two points a and b on the line is by definition the integral of $h(x)$ from a to b:



$$\text{Area } (a \text{ to } b) = \int_a^b h(x)dx.$$

This area integral is a limit of areas obtained by approximating $h(x)$ with step functions which take a finite number of different values on a finite number of disjoint intervals, as in the following diagram:



It is shown by calculus that such area integrals exist, and that they can be evaluated by finding a function H which is an *anti-derivative* or *indefinite integral* of h,

$$H'(x) = h(x), \qquad \frac{d}{dx}H(x) = h(x), \qquad \text{or} \qquad \int h(x)dx = H(x),$$

to express the same relation with three different standard notations. Such an indefinite integral H is unique apart from the addition of arbitrary constants, and

$$\text{Area } (a \text{ to } b) = \int_a^b h(x)dx = H(b) - H(a) \stackrel{def}{=} H(x)\Big|_a^b$$

The total area under the graph of h

$$\text{Area } (-\infty \text{ to } \infty) = \int_{-\infty}^{\infty} h(x)dx$$

is defined as the limit of Area (a to b) as $a \to -\infty$ and $b \to \infty$.

Fundamental Theorem of Calculus

$$\int_a^b F'(x)dx = F(b) - F(a) \stackrel{def}{=} F(x)\Big|_a^b$$

Some Indefinite Integrals

$$\int x^n dx = \frac{x^{n+1}}{n+1} \qquad (n \neq -1, \quad x > 0)$$

$$\int \frac{1}{x} dx = \log(x) \qquad (x > 0)$$

$$\int e^{\beta x} dx = \frac{1}{\beta} e^{\beta x}$$

$$\int \log(x) dx = x \log(x) - x \qquad (x > 0)$$

$$\int \sin(\theta) d\theta = -\cos(\theta)$$

$$\int \cos(\theta) d\theta = \sin(\theta)$$

$$\int \frac{1}{\sqrt{1-x^2}} dx = \arcsin(x) \qquad (|x| < 1)$$

$$\int \frac{1}{1+x^2} dx = \arctan(x)$$

Some Definite Integrals

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\, dx = \sqrt{2\pi}$$

$$\int_0^{\infty} x^n e^{-x}\, dx = n! \qquad (n \textit{ integer}, n \geq 0)$$

$$\int_0^1 x^m (1-x)^n\, dx = \frac{m!\, n!}{(m+n+1)!}$$

The first four properties of integrals in the box below should be compared to the corresponding properties of sums, listed in Appendix 2.

Properties of Integrals

Assume $a < b$.

Constants: If $f(x) = c$ for all x, then

$$\int_a^b f(x)dx = \int_a^b cdx = (b-a)c = (\text{length of interval}) \times c$$

Constant factors:

$$\int_a^b cf(x)dx = c\int_a^b f(x)dx$$

Addition:

$$\int_a^b (f(x)+g(x))dx = \int_a^b f(x)dx + \int_a^b g(x)dx$$

Inequalities: If $f(x) \leq g(x)$ for all x, then

$$\int_a^b f(x)dx \leq \int_a^b g(x)dx$$

Splitting the range of integration: If $a < b < c$,

$$\int_a^c f(x)dx = \int_a^b f(x)dx + \int_b^c f(x)dx$$

Integration by parts:

$$\int_a^b f(x)g'(x)dx = [f(x)g(x)]_a^b - \int_a^b f'(x)g(x)dx$$

Appendix 4
Exponents and Logarithms

Suppose that $b > 0$. For each positive integer x, a number b^x, called *b to the power x*, *b to the exponent x*, or just *b to the x*, is defined by $b^1 = b$, $b^2 = b.b$, and so on. So b^x is the product of x factors of b. This implies the first two rules stated in the box for positive integer exponents x and y. The definition of b^x is extended to $x = 0$, negative integers x, and rational numbers x, by requiring these two rules to hold for all these values of x as well. This implies the rest of the laws stated for rational x and y. The definition of b^x is further extended to all real x by assuming b^x is a continuous function of x.

Laws of Exponents

For $b, c > 0$, and all real numbers x and y:

$$\text{(i)}\quad b^{x+y} = b^x b^y$$

$$\text{(ii)}\quad b^{xy} = (b^x)^y$$

$$\text{(iii)}\quad b^0 = 1$$

$$\text{(iv)}\quad b^{-x} = 1/b^x$$

$$\text{(v)}\quad b^{x-y} = b^x/b^y$$

$$\text{(vi)}\quad (bc)^x = b^x c^x$$

To illustrate, for a positive integer n, $b^{1/n}$ is the positive nth root of b, also denoted $\sqrt[n]{b}$. This comes from rule (ii) for $x = 1/n$ and $y = n$. For positive rational $x = m/n$, (ii) gives

$$b^x = (b^{1/n})^m = (b^m)^{1/n}$$

Negative exponents are defined by rule (iv). The idea of multiplying together x factors of b does not make sense if x is not a positive integer. But the extended definition of exponents is very useful for algebraic manipulations with powers.

For $y > 0$ and $b > 0$ the equation $y = b^x$ is solved by a unique number $x = \log_b(y)$, called the logarithm of x to base b. In other words, the function $y \mapsto \log_b(y)$ is the *inverse function* of $x \mapsto b^x$. The laws of exponents imply the following:

Laws of Logarithms

For $b > 0$, $x > 0$, $y > 0$,

$$\log_b(xy) = \log_b(x) + \log_b(y)$$

$$\log_b(x^y) = y\log_b(x) \qquad \text{(true also for } y \leq 0\text{)}$$

$$\log_b(1) = 0$$

$$\log_b(1/x) = -\log_b(x)$$

$$\log_b(y/x) = \log_b(y) - \log_b(x)$$

$$\log_a(x) = \log_a(b)\log_b(x) \qquad \text{(change of base)}$$

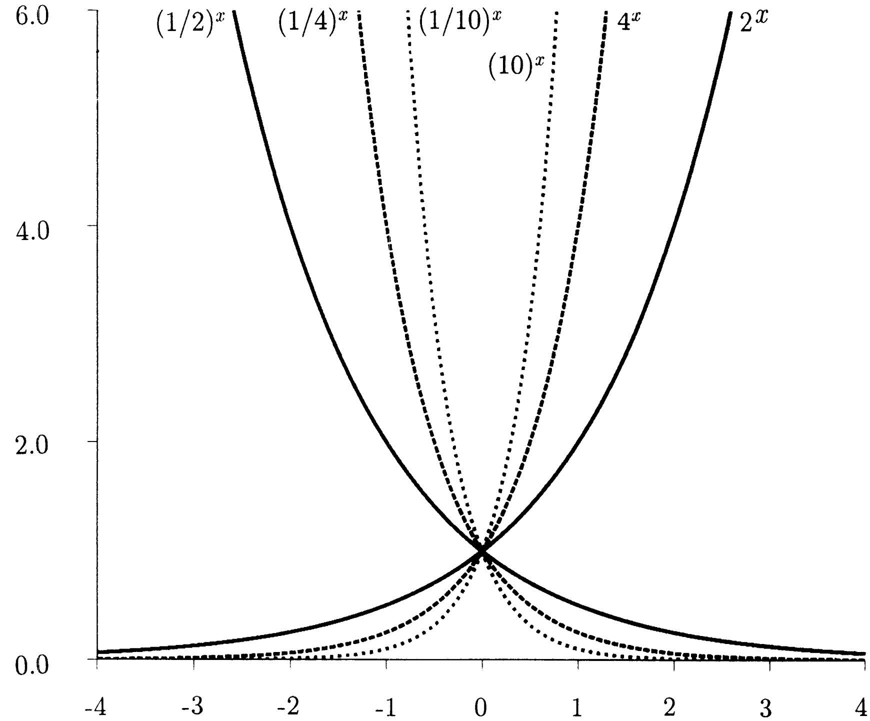
As the graphs suggest, b^x is a differentiable function of x for every $b > 0$. This involves the constant

$$e = 2.71828\ldots$$

defined precisely by any of the formulae in the next box. While the function $x \rightarrow b^x$ may be called an exponential function for any $b > 0$, *the* exponential function is

$$\exp(x) = e^x$$

FIGURE 1. Graphs of $y = b^x$



Exponentials with Base *e*

Derivative:

$$\frac{d}{dx}(e^x) = e^x.$$

Tangent approximation near zero:

$$e^x = 1 + x + \epsilon(x) \qquad \text{where } \epsilon(x)/x \to 0 \text{ as } x \to 0$$

Convex inequality:

$$e^x \geq 1 + x \qquad \text{for all } x$$

Series:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

Product limit:

$$e^x = \lim_{n\to\infty} (1 + \frac{x}{n})^n$$

The inverse of the exponential function is the *logarithm to base e*, or *natural logarithm*,

$$\log(x) = \log_e(x), \quad x > 0.$$

Logarithms with Base *e*

Derivative:

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

Tangent approximation near one:

$$\log(1+z) = z - \delta(z) \qquad \text{where } \delta(z)/z \to 0 \text{ as } z \to 0$$

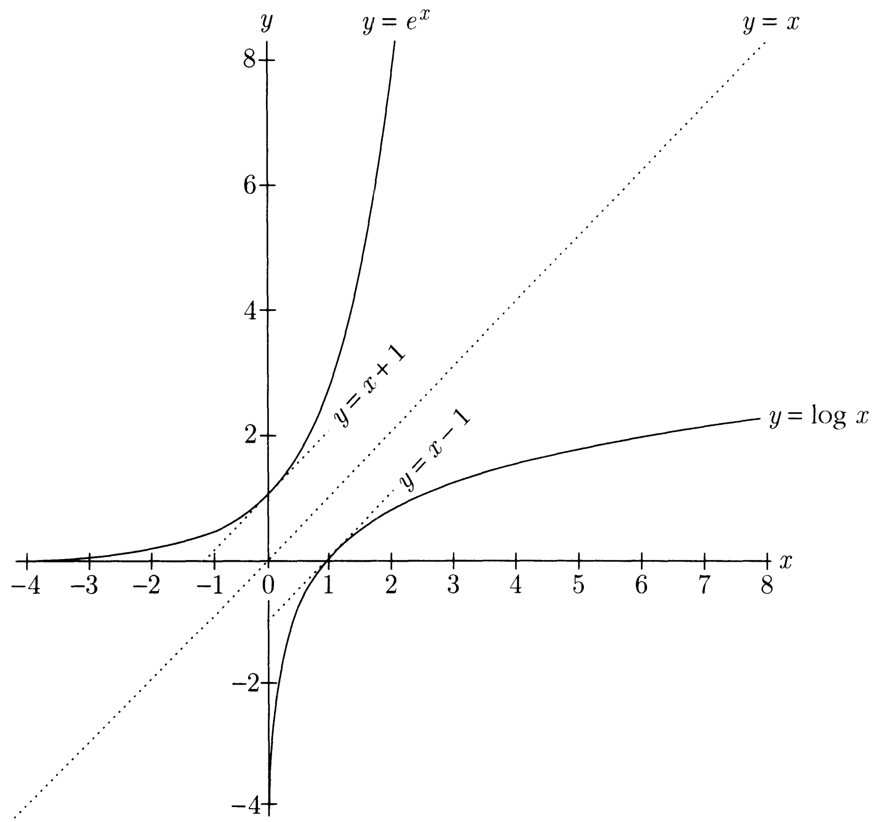
Concave inequality:

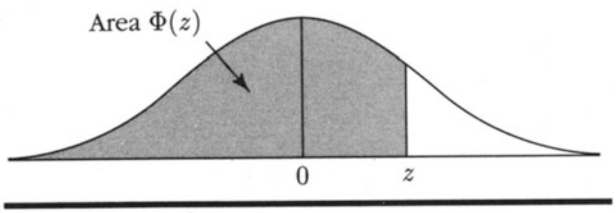
$$\log(1+z) \le z \qquad \text{for all } z$$

Series:

$$\log(1+z) = z - \frac{z^2}{2} + \frac{z^3}{3} - \frac{z^4}{4} + \cdots \qquad \text{for } -1 < z \le 1$$

FIGURE 2. **Graphs of e^x and log x.** The graph of log x, the inverse function of e^x, is obtained by reflection of the graph of e^x about the $45°$ line $y = x$. Just as the slope of $y = e^x$ is $e^0 = 1$ as this curve passes through the point $(0, 1)$, the slope of the curve $y = \log x$ is also 1 as it passes through the point $(1, 0)$. So
the $45°$ line $y = x + 1$ is tangent below the curve $y = e^x$ at $x = 0$,
the $45°$ line $y = x - 1$ is tangent above the curve $y = \log x$ at $x = 1$.
This gives the tangent approximations and inequalities for exp and log.





Appendix 5

Normal Table

Table shows values of $\Phi(z)$ for z from 0 to 3.59 by steps of .01. Example: to find $\Phi(1.23)$, look in row 1.2 and column .03 to find $\Phi(1.2 + .03) = \Phi(1.23) = .8907$. Use $\Phi(z) = 1 - \Phi(-z)$ for negative z.

	.0	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7703	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998

Brief Solutions to Odd-Numbered Exercises

1.1.1. a) 2/3 b) 66.67% c) 0.6667 d) 4/7 e) 57.14% f) 0.5714

1.1.3. a) $1/n^2$ b) $(n-1)/n^2$ c) $(1-1/n)/2$ d) $1/n(n-1)$, $1/n$, $1/2$

1.1.5. a) 2652 b) 1/13 c) 1/13 d) 1/221 e) 33/221

1.1.7. a) 1/9 b) 1/4 c) 5/36 d) $x^2/36, (2x-1)/36$ e) 1

1.1.9. 1/11, 1/6

1.1.11. Use the definition of fair odds, substitute in the formula for the house percentage.

1.2.1. The opinion of the judge.

1.2.3. a) $\Sigma > 1$ b) Yes. In this situation, you can get back more than you bet.

1.3.1. 4/7 of the cake

1.3.3. $\Omega = \{1, 2, \ldots, 500\}$ a) $\{17, 93, 202\}$
b) $\{17, 93, 202, 4, 101, 102, 398\}^c$ c) $\{16, 18, 92, 94, 201, 203\}$

1.3.5. a) first coin lands heads b) second coin lands tails c) first coin lands heads d) at least two heads e) exactly two tails f) first two coins land the same way

1.3.7. a) $P(1) = P(6) = p/2$ and $P(2) = P(3) = P(4) = P(5) = (1-p)/4$
b) $(3-p)/4$

1.3.9. a) 0.9 b) 1 c) 0.1

1.3.13. Use inclusion–exclusion for two sets.

1.3.15. *Hint*: Let $A_i = B_i^c$.

1.4.1. a) can't be decided; the rest are true.

1.4.3. 75%

1.4.5. c) 17/35

1.4.7. a) 0.3 b) 0.6

1.4.9. $p_1 = 0.1$, $p_2 = 0.4$, and $p_3 = 0.5$

1.4.11. a) $\frac{1+p}{4}$ b) $\frac{1-p}{4}$ c) $\frac{1-p}{2}$ d) $\frac{1+p}{2}$

1.5.1. a) 7/24 b) 8/17

1.5.3. a) 40/41 b) 1/41

1.5.5. a) 0.0575 b) 0.002 c) 0.9405 d) $16/115 \approx 0.139$ e) yes

1.5.7. a) 5/12 b) no c) You would be right 7/16 of the time.
d) Respond by always guessing box 1. Your probability of correct guessing is 1/2.

1.6.1. 5

1.6.3. a) 0.7692 b) 0.2308

1.6.5. a) $1 - (364/365)^{n-1}$ b) at least 254

1.6.7. a) $p_3p_1q_2 + p_3q_1p_2 + p_3p_1p_2$
b) $p_4 + P(\text{flows along top}) - p_4 \cdot P(\text{flows along top})$ where P(flows along top) was calculated in a)

1.rev.1. 6/11

1.rev.3. False

1.rev.5. The chance of passing when you use the first order is $zh(2-z)$. With the second order, it's $hz(2-h)$.

1.rev.7. a) $\frac{20}{50} \cdot \frac{19}{49} \cdot \frac{18}{48} \cdot \frac{17}{47} = .021$ b) 1 – answer to a) c) $4 \cdot \frac{30}{50} \cdot \frac{20}{49} \cdot \frac{19}{48} \cdot \frac{18}{47}$
d) $4 \cdot \frac{30}{50} \cdot \frac{29}{49} \cdot \frac{28}{48} \cdot \frac{20}{47} + \frac{30}{50} \cdot \frac{29}{49} \cdot \frac{28}{48} \cdot \frac{27}{47}$

1.rev.9. a) 1/60 b) 3/5 c) 13/30

1.rev.11. $\frac{9f}{9f+8b}$

1.rev.13. *Hint*: Write $P(A|B) = \sum_{i=1}^{n} P(AB_i|B)$.

1.rev.15. 2/3

1.rev.17. False.

2.1.1. a) $\binom{7}{4}$; b) $\binom{7}{4}(1/6)^4(5/6)^3$

2.1.3. a) 0.1608 b) 0.1962 c) 0.9645 d) 0.3125 e) 0.5

2.1.5. a) $\frac{\binom{19}{11}}{\binom{20}{12}}$ b) $\frac{\binom{18}{10}}{\binom{20}{12}}$ c) $1 - \left\{ \frac{\binom{15}{12}}{\binom{20}{12}} + 5 \times \frac{\binom{15}{11}}{\binom{20}{12}} \right\}$

2.1.7. 0.1005

2.1.9. a) 8, with probability 0.1387 b) 0.1128 c) 0.1133

2.1.11. a) 11 b) 0.2186

2.1.13. a) no b) 0.5 c) 0.75 d) 0.5

2.1.15. b) Note that $np = int\ (np) + [np - int\ (np)]$.

2.2.1. a) 0.7062 b) 0.1509 c) 0.0398 d) 0.0242

2.2.3. a) the first one b) $0.1841 > 0.0256$

2.2.5. 0.3974

2.2.7. a) City B has better accuracy. b) Both have same accuracy. c) City B has better accuracy.

2.2.9. a) 0.0495 b) Increase c) 0.1093

2.2.11. a) 0.4562 b) 0.2929 c) 0.2929 d) Increase e) Could be due to chance

2.2.13. Sample 9604 people

2.3.1. *Hint*: $P(k) = R(k) \cdot R(k-1) \cdots R(1)P(0)$.

2.3.3. a) Use odds ratios. b) Condition. e) Use the inequality $1 - x \le e^{-x}$.

2.4.1. a) Approximately Poisson(1) b) Approximately Poisson(2) c) Approximately Poisson(0.3284)

2.4.3. a) 0.999674. b) 0.997060.

2.4.5. 0.5945, 0.3092, 0.0804.

2.4.7. a) 2 b) 0.2659 c) 0.2475 d) 0.2565 e) $m = 250$; Normal approx: 0.0266 f) $m = 2$; Poisson approx: 0.2565

2.4.9. Use Poisson approximation: 0.9828

2.5.1. a) $\frac{\binom{20}{4}\binom{30}{6}}{\binom{50}{10}}$ b) $\binom{10}{4}(2/5)^4(3/5)^6$

2.5.3. a) $\frac{\binom{4}{4}\binom{48}{9}}{\binom{52}{13}}$ b) $\frac{\binom{3}{3}\binom{48}{9}}{\binom{51}{12}}$ c) $\frac{\binom{4}{4}\binom{48}{9}}{\binom{52}{13}-\binom{48}{13}}$ d) 0

2.5.5. $n \geq 537$ will do.

2.5.7. a) 0.1456 b) 0.3716 c) 0.0929

2.5.9. a) 0.282409 b) 0.459491

2.5.11. $\max\{0, n - N + G\}$ to $\min\{n, G\}$

2.5.13. 0.0028

2.rev.1. a) $\binom{10}{4}(1/6)^4(5/6)^6$ b) $\binom{10}{4}(1/5)^4(4/5)^6$ c) $\frac{10!}{4!3!2!}/6^{10}$ d) $\frac{\binom{7}{4}}{\binom{10}{4}}$

2.rev.3. a) $1/6$ b) $1/4$

2.rev.5. $\frac{\binom{97}{57}}{\binom{100}{60}}$

2.rev.7. $k \approx 1025$

2.rev.9. a) 0.8 b) guess 3 c) 0.4375

2.rev.11. 0.0102

2.rev.13. 0.99; the chance that any particular packet needs to be replaced is about 0.0144.

2.rev.15. a) $\binom{20}{5}(0.4)^5(0.6)^15$ b) $\frac{20!}{2!4!6!8!}(0.1)^2(0.2)^4(0.3)^6(0.4)^8$ c) $\binom{24}{2}(0.1)^3(0.9)^{22}$

2.rev.17. a) $\frac{\binom{6}{1}}{6^4}$ b) $\binom{6}{1} \times \binom{5}{1} \times \frac{\binom{4}{1}}{6^4}$ c) $\binom{6}{2} \times \frac{\binom{4}{2}}{6^4}$

2.rev.19. a) $(2/3)^4$ b) $\binom{4}{1}(2/3)^4(1/3) + (2/3)^4$

2.rev.21. $\sum_{x=0}^{(n-1)/2} \binom{n}{x} q^x p^{n-x}$.

2.rev.23. $\frac{(.4\times 1/2)+(.2\times 6/8)+(.1\times 14/16)}{(.2\times 1/2)+(.4\times 3/4)+(.2\times 7/8)+(.1\times 15/16)}$

2.rev.25. a) p^3, $3p^3q$, $6p^3q^2$ b) $p^3 + 3p^3q + 6p^3q^2$ c) $\frac{1}{1+3q+6q^2}$ d) 0.375 e) no.

2.rev.27. 0.3971

2.rev.29. 0.0579

2.rev.31. *Hint*: $np \geq npq$

2.rev.33. *Hint*: Think about conditional probabilities.

2.rev.35. a) $\sum_{k=20}^{35} \binom{1000}{k}(1/38)^k(37/38)^{1000-k}$ b) Use normal: 0.876

3.1.1. a) $P(X=0)=1/8$, $P(X=1)=3/8$, $P(X=2)=3/8$, $P(X=3)=1/8$
b) $P(|X-1|=0)=3/8$, $P(|X-1|=1)=4/8$, $P(|X-1|=2)=1/8$

3.1.3. a) All integers from 2 to 12 inclusive. b) Partial answer: $P(S=2)=1/36$, $P(S=3)=2/36$, $P(S=4)=3/36$, $P(S=5)=4/36$

3.1.5. Partial answer: $P(X_1X_2=1)=1/36$, $P(X_1X_2=2)=2/36$, $P(X_1X_2=3)=2/36$, $P(X_1X_2=4)=3/36$, $P(X_1X_2=5)=2/36$, $P(X_1X_2=6)=4/36$

3.1.7. a) $(ABC^c)\cup(AB^cC)\cup(A^cBC)$ b) $ab(1-c)+a(1-b)c+(1-a)bc$

3.1.9. Partial answer: $P(X=2)=5/35$, $P(X=3)=10/35$

3.1.11. a) binomial $(n+m,p)$ e) $\binom{2n}{n}$

3.1.13. a) $\frac{2n-2}{2n}\cdot\frac{2n-4}{2n}\cdot\cdots\cdot\frac{2n-2(k-1)}{2n}$ b) $\sqrt{n\log 4}$

3.1.15. a) $1/n$ b) $(n-1)/2n$ c) $(n-1)/2n$ d) $(2k-1)/n^2$ e) $[2(n+1-k)-1]/n^2$
f) $(k-1)/n^2$ for $k=2$ to $n+1$; $(2n-k+1)/n^2$ for $k=n+2$ to $2n$

3.1.17. a) $P(Z=k)=(k/21)\binom{20}{k}(1/2)^{20}+(1/21)\sum_{i=0}^{k}\binom{20}{i}(1/2)^{20}$

3.1.19. a) Partial answer: $P(S=7)=p_1r_6+p_2r_5+p_3r_4+p_4r_3+p_5r_2+p_6r_1$
d) yes

3.1.21. yes

3.1.23. $P(X\le T)\le P(Y\le T)$

3.2.1. 41.5

3.2.3. The expected number of sixes is $1/2$, the expected number of odds is $3/2$.

3.2.5. Expect to lose about 8 cents per game.

3.2.7. $\sum_{i=1}^{n} p_i$

3.2.9. $p-2pr+r$

3.2.11. Simple upper bound: 0.3 Actual probability: 0.271

3.2.13. a) 35 b) 8.458 c) 5.43 d) 10/3 e) 0.9690 f) 5.0310

3.2.15. Show that $E[L(Y,b)]=(\lambda+\pi)\sum_{y\le b}(b-y)p(y)-\pi b$.

3.2.17. a) $\frac{\binom{10}{9-3}}{\binom{13}{9}}$ b) $\frac{\binom{10}{9-3}}{\binom{13}{9}}-\frac{\binom{10}{8-3}}{\binom{13}{8}}$ c) 10.5

3.2.19. a) $\frac{5!}{1!1!1!2!}[2(1/7)^3(2/7)(3/7)+(1/7)^2(2/7)^2(3/7)+(1/7)^2(2/7)(3/7)^2]$
b) $2[1-(6/7)^5]+1-(5/7)^5+1-(4/7)^5$

3.3.1. a) $E(X)=30.42$, $SD(X)=0.86$ b) $E(X)=30.44$, $SD(X)=0.86$

3.3.3. a) 5 b) 26 c) 1 d) 26

3.3.5. *Hint*: Use the computational formula for the variance.

3.3.7. a) no b) $E(X)=\sum_{i=1}^3 n_i p_i$, $Var(X)=\sum_{i=1}^3 n_i p_i q_i$

3.3.9. a) $r(1-p_1)+(n-r)p_2$ b) $r(1-p_1)p_1+(n-r)p_2(1-p_2)$

3.3.11. $E(Y)=a+b\left(\frac{n-1}{2}\right)$, $Var(Y)=b^2(n^2-1)/12$

3.3.13. a) 111112 b) 55556 c) 1300

3.3.15. b) $10\sqrt{8}$

3.3.17. a) 0.05 b) 0.03 c) 0.92

3.3.19. Approximately $1-\Phi(1.66)=1-0.9515=0.0485$

3.3.21. a) 0.0876 b) 0.0489

3.3.23. Approximately $\Phi(-0.77)\approx 0.22$

3.3.27. For b), reduce to a). For c): Half the list are zeros, the rest are nines.

3.3.29. a) Guess 4. b) ($n=1$) 1/10; ($n=2$) 19/100; ($n=33$) 0.6826; ($n=66$) 0.8414; ($n=132$) 0.9544. c) $n\geq 220$ will do.

3.3.31. a) $9/2, \sqrt{33}/2$ d) $2\Phi(2b/\sqrt{33})-1$

3.4.1. a) $\binom{9}{5}p^5(1-p)^4$ b) $(1-p)^6\cdot p$ c) $\binom{11}{4}p^4(1-p)^7\cdot p$
d) $\sum_{k=0}^5 \binom{8}{k}p^k(1-p)^{8-k}\cdot\binom{5}{k}p^k(1-p)^{5-k}$

3.4.3. 12

3.4.5. Let $q_i=1-p_i$. a) q_2^n b) $(q_1q_2q_3)^n$ c) $(q_1q_2q_3)^{n-1}-(q_1q_2q_3)^n$
d) $p_2/(1-q_1q_2q_3)$

3.4.7. a) $\frac{1}{1+q}$ b) $\frac{q}{1+q}$ c) $\frac{p}{1-q^3}, \frac{q-q^3}{1-q^3}$ d) $p=\frac{3-\sqrt{5}}{2}$ e) 2/3

3.4.9. Expect to lose \$4 per game.

3.4.11. a) $\frac{p_Aq_B}{1-q_Aq_B}$ b) $\frac{q_Ap_B}{1-q_Aq_B}$ c) $\frac{p_Ap_B}{1-q_Aq_B}$ d) $P(N=k)=(q_Aq_B)^{k-1}(1-q_Aq_B)$ for $k=1,2,3,\ldots$

3.4.13. a) $P(\text{Black wins})=\frac{p}{1-qp}$ b) $(3-\sqrt{5})/2$ c) no d) 13

3.4.15. a) Use Exercise 3.4.6 b) *Hint*: Look at the tail probabilities $P(F\geq k)$

3.4.17. $\frac{2(1-p)p^k}{(2-p)^{k+1}}$ $(k \geq 0)$

3.4.19. *Hints*: a) Negative binomial b) Symmetry c) Use the result of b)

3.4.21. a) μ b) μ

3.4.23. a) $p/(1-qz)$ b) $f_1 = \frac{q}{p}$, $f_2 = 2(\frac{q}{p})^2$, $f_3 = 6(\frac{q}{p})^3$

3.5.1. 0.1428

3.5.3. a) 0.222 b) About 44

3.5.5. 0.39

3.5.7. a) Poisson(3), Poisson(2), Poisson(5) b) 0.3679

3.5.9. a) 0.0996 b) 0.8008 c) 0.3951

3.5.11. a) $e^{-2}2^4/4!$ b) 6 c) $e^{-3}3^4/4!$

3.5.13. a) $2.69 \times 10^{19}x^3$, $5.19 \times 10^9 x^{3/2}$ b) 7.19×10^{-6} cm

3.5.15. a) 198.01, 1.97 b) 1.79 c) 0.59

3.5.17. a) 0.0067 b) 0.0037

3.5.19. b) μ, μ^2, μ^3 c) μ, $\mu^2 + \mu$, $\mu^3 + 3\mu^2 + \mu$

3.5.21. c) 0.58304 d) 0.5628 e) 0.58306

3.6.1. a) $1/13$ b) $4/50$ c) $4 \times \frac{\binom{13}{5}}{\binom{52}{5}}$ d) $1 - \frac{\binom{4}{0}\binom{48}{5}}{\binom{52}{5}} - \frac{\binom{4}{1}\binom{48}{4}}{\binom{52}{5}}$

3.6.3. a) $8/47$ b) $(12 \times 11 \times 10 \times 9 \times 8)/(51 \times 50 \times 49 \times 48 \times 47)$ c) $1/4$
d) $1/13$ e) $1/13$ f) $1/4$

3.6.5. a) $b\left(\frac{b-1}{b}\right)^n$ b) $b\left(\frac{b-1}{b}\right)^n + b(b-1)\left(\frac{b-2}{b}\right)^n - b^2\left(\frac{b-1}{b}\right)^{2n}$

3.6.7. a) $n \cdot \frac{26}{52}$ b) $\left(\frac{52-n}{52-1}\right) \cdot n \cdot \frac{26}{52} \cdot \frac{26}{52}$

3.6.9. a) $\frac{N+1}{G+1}$ b) $\sqrt{\frac{BG(N+1)}{(G+1)^2(G+2)}}$

3.6.11. a) $P(x_1, \ldots, x_n) = 1/\binom{n}{g}$ if $x_1 + \cdots + x_n = g$ and 0 otherwise
b) no c) yes

3.6.13. a) Uniform on all ordered $(n+1)$-tuples of non-negative integers with sum $N-n$ c) $(N-n)_w n/(N)_{w+1}$ d) $E(W_i) = (N-n)/(n+1)$,
$E(T_i) = i(N+1)/(n+1)$, 9.6, 10.6, 21.2, 31.8, 42.4 e) $\frac{\binom{n}{1}\binom{N-n}{t}}{\binom{N}{t+1}} \frac{(n-1)}{(N-t-1)}$
f) $P(D_n = d) = P(W_1 + W_{n+1} = N - 2 - d)$. Now use e).
$E(D_n) = \frac{(n-1)(N+1)}{(n+1)} - 1$

3.6.15. c) $t_1 = 0 = t_{n+1}$; $t_2 = t_3 = \cdots = t_n = 1$; so $t = n - 1$.

3.rev.1. a) $1-(5/6)^{10}$ b) $10/6$ c) 35 d) $\frac{\binom{5}{2}\binom{5}{2}}{\binom{10}{4}}$ e) $\frac{1}{2}\left(1-\sum_{k=0}^{5}\left[\binom{5}{k}(1/6)^k(5/6)^{5-k}\right]^2\right)$

3.rev.3. a) $(2x-1)/36$ b) $2/5$ for $y = 1, 2$ and $1/5$ for $y = 3$.
c) $2/36$ for $1 \le y < x \le 6$ and $1/36$ for $y = x$ d) 7

3.rev.5. a) -18.4 cents b) 2.111 c) 12.667

3.rev.7. a) 0.1875 b) 0.5 c) 0.219

3.rev.9. a) $5/12$ b) $7/12$ c) 441 d) Approximately 796

3.rev.11. $P(X < 2)$ is largest, $P(X > 2)$ is smallest.

3.rev.15. a) Binomial$(100, 1/38)$ b) Poisson$(100/38)$ c) Negative binomial $(3, 1/38)$ shifted to $\{3, 4, ...\}$ d) 3×38

3.rev.17. a) $N/6$ b) 0.3604

3.rev.19. a) $e^{-p\mu}$ b) 0.6065

3.rev.21. Negative binomial distribution on $\{\ 0, 1, ...\ \}$ with parameters $r = 3$ and p

3.rev.23. a) $\frac{2^k (n)_k}{(2n)_k}$ b) $H/\sqrt{2n}$ tends to the Rayleigh distribution (See section 6.3).
c) $\sqrt{\pi n}$ d) 17 or so.

3.rev.25. a) Partial answer: $P(Y_1 + Y_2 = 0) = 9/36$, $P(Y_1 + Y_2 = 1) = 12/36$,
$P(Y_1 + Y_2 = 2) = 10/36$. b) $10/3$

3.rev.27. 343.047

3.rev.29. c) uniform on $\{0, 1, \ldots, n\}$ d) no, yes e) $\frac{b}{b+w}$ f) $\frac{b+d}{b+w+d}$

3.rev.33. b) $\frac{1}{2^{n-1}}$, $\frac{1}{2^{n-2}\left(1+\frac{1}{n}\right)}$ d) $\frac{1}{2^{n-3}\left(1+\frac{3}{n}\right)}$

3.rev.37. a) $\binom{n}{k}\frac{(G)_k}{(N)_k}$

3.rev.41. a) 2350 b) 70 c) 9400 d) 8700 e) 730

4.1.1. a) 0.000399 b) 0.000242

4.1.3. a) 6 b) $1/2$ c) $7/27$ d) $13/54$ e) $1/2$, $1/20$

4.1.5. b) $7/12$ c) $1/2$ d) no

4.1.7. 0.096

4.1.9. 0.0418

4.1.11. a) 0.2325 b) 0.6102 c) 0.84

4.1.13. a) $1/16$ b) $n \geq 134$

4.1.15. a) $(0, 1/2)$ b) $\text{erf}(x) = 2\Phi(\sqrt{2}x) - 1$ c) $\Phi(z) = (\text{erf}(z/\sqrt{2}) + 1)/2$

4.2.1. a) $1/32$ b) 3.32 years c) 10 years d) 0.3679

4.2.3. a) 0.6321 b) 0.3935 c) 0.8647 d) 0.99995

4.2.5. a) 0.86 b) 0.73 c) 4 seconds

4.2.7. $-\frac{1}{\lambda}\log(1-p)$

4.2.9. c) $E(T^n) = \Gamma(n+1)$.

4.2.11. b) $e^{-\lambda} - e^{-2\lambda}$

4.2.13. a) 5% per day
b) $(d = 10)$ 6065, 49; $(d = 20)$ 3679, 48; $(d = 30)$ 2231, 42.

4.2.15. a) 80 days b) 40 days c) 0.6472

4.2.17. a) $E(T_{\text{total}}) = 80$ days, $SD(T_{\text{total}}) = 20\sqrt{2}$ days, $P(T_{\text{total}} \geq 60) \approx 0.744$.
b) four spares will do.

4.3.1. a) $1 - G(b)$ b) $G(a) - G(b)$

4.3.3. a) $\left(\frac{b}{b+t}\right)^a$ if $t > 0$. b) $\left(\frac{a}{b+t}\right)\left(\frac{b}{b+t}\right)^a = \frac{ab^a}{(b+t)^{a+1}}$ if $t > 0$.

4.3.5. b) Mean: $\lambda^{-1/\alpha}\Gamma\left(\frac{1}{\alpha}+1\right)$. Variance: $\lambda^{-2/\alpha}\left\{\Gamma\left(\frac{2}{\alpha}+1\right) - \left[\Gamma\left(\frac{1}{\alpha}+1\right)\right]^2\right\}$

4.3.7. b) 9.265 c) About $1 - \Phi(2.456) = 0.007$

4.4.1. Exponential (λ/c)

4.4.3. $f_Y(y) = \frac{1}{2\sqrt{y}}$ if $0 < y < 1$.

4.4.5. If $0 < y < 1$ then $f_Y(y) = \frac{1}{3\sqrt{y}}$; if $1 < y < 4$ then $f_Y(y) = \frac{1}{6\sqrt{y}}$.

4.4.7. Apply Exercise 4.4.6

4.4.9. One to one change of variable formula

4.5.3. a) Y has the same distribution as X.
b) If $0 < r < 1$ then $F_R(r) = r^2$ and $f_R(r) = 2r$.

4.5.5. If $x \leq 0$, then $F_X(x) = \frac{1}{2}e^x$; if $x \geq 0$, then $F_X(x) = 1 - \frac{1}{2}e^{-x}$.

4.5.7. a) $f_Y(y) = 2\lambda y e^{-\lambda y^2}$ $(y > 0)$ b) 0.51 c) Let $Y = \sqrt{-\log(1-U)/\lambda}$

4.6.1. a) 0.0881 b) 0.0056 c) 0.0399

4.6.3. a) $(y-x)^n$ b) $(1-x)^n-(y-x)^n$ c) $y^n-(y-x)^n$
d) $1-(1-x)^n-y^n+(y-x)^n$ e) $\binom{n}{k}x^k(1-y)^{n-k}$
f) $\binom{n}{k}x^k(1-y)^{n-k}+\binom{n}{k+1}x^{k+1}(1-y)^{n-k-1}+\frac{n!}{k!1!(n-k-1)!}x^k(y-x)(1-y)^{n-k-1}$

4.6.5. a) $P(X_{(k)} \le x) = \sum_{i=k}^n \binom{n}{i}[F(x)]^i[1-F(x)]^{n-i}$

4.rev.1. a) $ne^{-\lambda t}$ b) $ne^{-\lambda t}(1-e^{-\lambda t})$

4.rev.3. Density: $3x^2$ if $0 < x < 1$. Expectation: $3/4$

4.rev.5. a) 0.4 b) If $0 < t \le 30$, then $P(T > t) = \frac{100-2t}{100}$; If $30 < t \le 70$, then $P(T > t) = \frac{70-t}{100}$. c) $f_T(t) = 2/100$ if $0 < t \le 30$, $= 1/100$ if $30 < t \le 70$.
d) mean 29, SD 19.8 e) Locate the station at the midpoint of the road.

4.rev.7. a) $\alpha = \frac{\beta}{2}$ b) $E(X) = 0$, $Var(X) = 2/\beta^2$ c) $e^{-\beta y}$ if $y > 0$
d) $1-(1/2)e^{-\beta x}$ if $x > 0$; $(1/2)e^{\beta x}$ if $x < 0$

4.rev.11. a) $1/2$ b) $\frac{7!}{2^8}$ c) $\frac{100^5}{30}$

4.rev.13. a) $\frac{e^{-\lambda_{\text{loc}}}(\lambda_{\text{loc}})^5}{5!}\frac{e^{-\lambda_{\text{dis}}}(\lambda_{\text{dis}})^3}{3!}$ b) $e^{-3(\lambda_{\text{loc}}+\lambda_{\text{dis}})}\frac{[3(\lambda_{\text{loc}}+\lambda_{\text{dis}})]^{50}}{50!}$ c) $\left(\frac{\lambda_{\text{loc}}}{\lambda_{\text{dis}}+\lambda_{\text{loc}}}\right)^{10}$

4.rev.15. 0.2518

4.rev.19. a) $(20-2)\log_2 10$ b) $20\log_2 10 - \log_e 2$

4.rev.21. a) $f_Y(y) = 2ye^{-y^2}$ $(y > 0)$ b) exponential (1) c) 1

4.rev.23. a) 5, 4 b) $f_M(m) = 0.5e^{-0.5(m-3)}$ $(m > 3)$ c) 0.3679

4.rev.25. a) uniform $(0, 1/2)$ b) uniform $(0, 1)$ c) $1/4$, $1/48$

4.rev.27. a) Use the fact that all the $n!$ orderings of $U_1, \ldots, U_n$ are equally likely.

4.rev.29. a) When $c < \sqrt{\frac{2}{\pi}}$
b) Expected net gain is maximized at b satisfying $e^{-b^2/2} = \sqrt{\frac{\pi}{2}}c$.

5.1.1. a) $7/12$ b) $5/36$

5.1.3. $7/12$

5.1.5. a) 0.1 b) 0.81

5.1.7. a) $(1-x)^2$ b) If $0 < x < 1$ then $P(M \le x) = 1-(1-x)^2$ and $f_M(x) = 2(1-x)$

5.1.9. $1/4$

5.2.1. a) If $0 < |y| < x < 1$ then $f_{X,Y}(x,y) = 1$ b) If $0 < x < 1$ then $f_X(x) = 2x$; if $0 < y < 1$ then $f_Y(y) = 1-|y|$ c) no d) $E(X) = 2/3$, $E(Y) = 0$

5.2.3. a) $3/4$ b) $\frac{3}{4}\left(\frac{a^3}{3}+a^2\right)$ c) $\frac{3}{4}\left(\frac{b}{3}+b^2\right)$

5.2.5. $\frac{\mu}{3\lambda+\mu}$

5.2.7. $1/8$

5.2.9. a) $2\lambda^2 e^{-\lambda(x+y)}$ $(0 < x < y)$, no; b) $2\lambda^2 e^{-2\lambda x - \lambda z}$ $(x > 0, z > 0)$, yes; c) X is exponential (2λ) and Z is exponential (λ).

5.2.11. a) $3/2$ b) $1/2$ c) $4/3$ d) ∞

5.2.13. The distributions are all the same, with density $2(1-x)$ for $0 < x < 1$.

5.2.15. a) $F(b,d) - F(a,d) - F(b,c) + F(a,c)$ b) $F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v)dudv$
c) $f(x,y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}F(x,y)$ d) $F(x,y) = F_X(x)F_Y(y)$
e) $F(x,y) = y^n - (y-x)^n$ for $0 < x < y < 1$;
$f(x,y) = n(n-1)(y-x)^{n-2}$ for $0 < x < y < 1$

5.2.17. a) $f(x,r) = \frac{2}{\pi}\frac{r}{\sqrt{r^2-x^2}}$ for $0 \le r \le 1$ and $-r \le x \le r$
b) $f(x,r) = \frac{3}{2}r$ for $0 \le r \le 1$ and $-r \le x \le r$

5.2.19. a) $f_{\text{Lon}}(x) = 1/360$ if $-180 < x < 180$
b) $f_{\text{Lat}}(y) = \frac{\pi}{360}\cos\left(\frac{\pi}{180}y\right)$ if $-90 < y < 90$
c) $f(x,y) = \frac{1}{360} \cdot \frac{\pi}{360}\cos\left(\frac{\pi}{180}y\right)$ if $-180 < x < 180$ and $-90 < y < 90$
d) yes

5.2.21. a) 0.3825, 0.765 b) 1/3 c) 0.577 d) 0.5197 ± 0.0048

5.3.1. a) 0.1175 b) 0.1178 c) $\sqrt{\frac{2}{\pi}}$ d) 0.762 e) 0.58 f) 0.3521 g) 0.29

5.3.3. a) $1 - \Phi(0.5)$ b) $1/2$ c) 5 d) $\sqrt{14}$

5.3.5. About 2.1

5.3.7. a) 97.72% b) 88.49% c) 0.9795

5.3.9. a) 0.1307 b) 0.0062 c) The answer to b) will be approximately the same.

5.3.11. a) normal with mean 0 and variance $t\sigma^2$ b) $\frac{R_t}{\sigma\sqrt{t}}$ has Rayleigh distribution so R_t has expectation $\sigma\sqrt{\frac{t\pi}{2}}$ and SD $\sigma\sqrt{t\left(\frac{4-\pi}{2}\right)}$ c) 0.1353

5.3.13. c) Try $h(u) = \sqrt{-2\log(1-u)}$ and $k(v) = 2\pi v$

5.3.15. *Hints*: a) Example 4.4.5 b) induction c) linear change of variable

5.3.17. a) Skew-normal approximations: 0.1377, 0.5940, 0.9196, 0.9998, 1.0000
Compare to the exact values: 0.0902, 0.5940, 0.9389, 0.9970, 1.0000
b) 0.441, 0.499. Skew-normal is better.

5.4.1. a) $3/4$ b) $f_{X_1+X_2}(z) = z/2$ if $0 \le z \le 1$; $= 1/2$ if $1 \le z \le 2$; $= (3-z)/2$ if $2 \le z \le 3$ c) $F_{X_1+X_2}(z) = z^2/4$ if $0 \le z \le 1$; $= (2z-1)/4$ if $1 \le z \le 2$; $= 1 - (3-z)^2/4$ if $2 \le z \le 3$.

5.4.3. a) If $\alpha \neq \beta$, $f_{X+Y}(z) = \frac{\alpha\beta}{\alpha-\beta}(e^{-\beta z} - e^{-\alpha z})$ b) $\frac{1}{\alpha} + \frac{1}{\beta}$ c) $\frac{\sqrt{\alpha^2+\beta^2}}{\alpha\beta}$

5.4.5. a) Uniform over $(10, 70)$ b) 0.483

5.4.7. a) $f_{XY}(z) = \int_{-\infty}^{\infty} \frac{1}{|x|} f_{X,Y}(z, \frac{z}{x})dx$. b) $f_{X-Y}(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, x-z)dx$. c) $f_{X+2Y}(z) = \int_{-\infty}^{\infty} \frac{1}{2} f_{X,Y}(x, \frac{z-x}{2})dx$.

5.4.9. $f_X(x) = -\log(x) \qquad (0 < x < 1)$

5.4.11. uniform $(0, 1)$

5.4.13. $f_Z(z) = \frac{1}{2}\lambda e^{-\lambda|z|}$

5.4.15. a) reduce to the case $\lambda = 1$ by scaling. b) $P(Z \leq z) = 2z/(1+z)$ c) $f_Z(z) = 2/(1+z)^2$ for $0 < z < 1$.

5.4.17. a) $\frac{\sqrt{3}}{2}t^2$ if $0 < t < 1$; $\sqrt{3}(-t^2 + 3t - \frac{3}{2})$ if $1 < t < 2$; $\frac{\sqrt{3}}{2}(3-t)^2$ if $2 < t < 3$ b) If $t \leq 1$ then the cross section is an equilateral triangle having side length $t\sqrt{2}$; if $t = 3/2$ then the cross section is a regular hexagon having side length $1/\sqrt{2}$.

5.rev.1. $1 - \frac{\sqrt{2}}{4}$

5.rev.3. a) 0.04 b) 0.039 c) 0.29

5.rev.5. a) $1 - \pi/8$ b) $5/12$

5.rev.7. 0.0124

5.rev.9. a) $f_{X+U}(x) = 1/4$ if $0 < x < 1$; $= 1/2$ if $1 < x < 2$; $= 1/4$ if $2 < x < 3$. b) Uniform$(-1/2,\ 1/2)$

5.rev.11. a)$P(X > x) = 1 - (1/2)x$ for $0 < x < 1$. b)$F_X(x) = (1/2)x$ for $0 < x < 1$ and $F_X(x) = 1 - \frac{1}{2x}$ for $x > 1$. c)$f_X(x) = 1/2$ for $0 < x < 1$ and $f_X(x) = \frac{1}{2x^2}$ for $x > 1$

5.rev.13. a) $F_X(x) = 1 - \frac{1}{\pi}\arccos(x)$ for $|x| \leq 1$ b) Y has the same distribution function as X. c) $F_{X+Y}(z) = 1 - \frac{1}{\pi}\arccos\frac{z}{\sqrt{2}}$ for $|z| \leq \sqrt{2}$

5.rev.15. a) $1/6$ b) 0 c) 1 d) $1/2$ e) $2/3$ f) $e^{-\frac{1}{2}}$ g) 0.8759 h) $3/4$ i) 0.5737

5.rev.17. a) 0.92 b) About 27.7

5.rev.19. a) $(K_1 = k) = (W_k < \min_{i \neq k} W_i)$; b) $p_k = \lambda_k/(\lambda_1 + \cdots \lambda_d)$; c) use the memoryless property of the exponential waiting times; d) the answer to g) must be p_k by the law of large numbers; e) $\lambda_k T$; f) $(\lambda_1 + \cdots \lambda_d)T$; g) $\lambda_k/(\lambda_1 + \cdots \lambda_d)$.

5.rev.21. a)$F_R(r) = 1 - e^{-\lambda\pi r^2}$ and $f_R(r) = 2\lambda\pi r e^{-\lambda\pi r^2}$ for $r > 0$ c) $E(R) = \frac{1}{2\sqrt{\lambda}}$; $SD(R) = \frac{1}{2\sqrt{\lambda}}\sqrt{\frac{4-\pi}{\pi}}$ d) mode: $\frac{1}{\sqrt{2\lambda\pi}}$; median: $\sqrt{\frac{\log 2}{\lambda\pi}}$

5.rev.23. a) 1/8 b) 7/19

5.rev.25. a) $\frac{n!x^{k-1}(y-x)^{m-k-1}(1-y)^{n-m-1}}{(k-1)!(m-k-1)!(n-m-1)!}$ $(0 < x < y < 1)$
b) beta $(m-k, n-m+k+1)$ c) beta $(k, m-k+1)$

5.rev.29. Let the spacing between the parallel lines be $2a$. a) If $0 < x \leq a$ then $P(X \leq x) = \frac{2}{\pi a}x$; if $x \geq a$ then $P(X \leq x) = \frac{2}{\pi a}\left[a \arccos \frac{a}{x} + x - \sqrt{x^2 - a^2}\right]$
b) If $0 < x < a$ then $f_X(x) = \frac{2}{\pi a}$; if $x \geq a$ then $f_X(x) = \frac{2}{\pi a}\left[1 - \sqrt{\frac{x^2-a^2}{x}}\right]$

5.rev.31. a) $r = \sqrt{4/3}$ b) no, X_1 and Y_1 are not independent d) $e^{-1/2}$

5.rev.33. a)$r = \sqrt{1/2}$ b) no, X_1 and Y_1 are not independent d) $e^{-1/2}$ g) 2^{-n}

6.1.1. a) binomial $(3, 1/2)$ b) binomial $(3-x, 1/2)$ distribution shifted to $\{x, x+1, \ldots 3\}$ c) Partial answer: $P(X=0, Y=1) = 3/64$, $P(X=1, Y=1) = 3/32$, $P(X=2, Y=1) = 0$, $P(X=3, Y=1) = 0$. d) $P(Y=y) = 1/64, 9/64, 27/64, 27/64$ for $y = 0, 1, 2, 3$. e) Partial answer: The conditional distribution of X given $Y=0$ is given by $P(X=0|Y=0) = 1/3$, $P(X=1|Y=0) = 2/3$. f) For $y = 0, 1, 2, 3$, guess $x = 0, 1, 1$ (or 2), 2 respectively. g) 31/64

6.1.3. a) $P(U=u) = 0, 0.1, 0.4, 0.3, 0.2$ for $u = 0, 1, 2, 3, 4$ b) 0.1125 c) 0.075

6.1.5. b) By the normal approximation to the binomial, $1 - \Phi\left(\frac{89.5-75}{6.124}\right) = 0.0089$

6.1.7. a) Write $P(X=k) = \sum_{n=k}^{\infty} P(X=k, N=n)$. b) 0.0000016

6.1.9. Further equivalent condition: $P(X=x|Y=y, Z=z) = P(X=x|Z=z)$.

6.2.1. a) $E(Y|X=x) = 41/11, 38/9, 33/7, 26/5, 17/3, 6/1$ for $x = 1, 2, 3, 4, 5, 6$
b) $E(X|Y=y) = 1, 4/3, 9/5, 16/7, 25/9, 36/11$ for $y = 1, 2, 3, 4, 5, 6$

6.2.3. a) $E(Y|X=x) = \frac{n+x+1}{2}$ for $x = 1$ to $n-1$ b) $E(X|Y=y) = \frac{y}{2}$ for $y = 2$ to n.

6.2.5. a) $F_1(x)p + F_2(x)(1-p)$ b) $E(X_1)p + E(X_2)(1-p)$
c) $Var(X_1)p + Var(X_2)(1-p) + p(1-p)(E(X_1) - E(X_2))^2$

6.2.7. Condition on the value of N.

6.2.9. a) $j/2$ b) $j + \frac{N-j+1}{k}$ c) $h\frac{j}{i}$ if $h < i$; $j + (h-i)\left(\frac{N-j+1}{k-i+1}\right)$ if $h > i$.

6.2.11. 25/78

6.2.13. c) $\frac{(n-m)}{(n-1)} m \frac{k}{n} \frac{(n-k)}{n}$

6.2.15. a) $E(S) = nE(\Pi)$ b) $Var(S) = nE(\Pi)(1 - E(\Pi)) + n(n-1)Var(\Pi)$ c) As large as possible: Π with values only 0 and 1. As small as possible: Π constant.

6.3.1. 1/3

6.3.3. $P(Y \leq y|X = x) = y/(2-x)$ for $0 < x < 2$ and $0 < y < 2 - x$

6.3.5. a) If $|x| < 1/2$ then $P(Y \geq 1/2|X = x) = \frac{1/2-|x|}{1-|x|}$. b) One minus the answer in a). c) If $|x| < 1$ then $E(Y|X = x) = (1-|x|)/2$ d) If $|x| < 1$ then $Var(Y|X = x) = (1-|x|)^2/12$

6.3.7. a) $f_Y(y) = 3(1-y)^2$ for $0 < y < 1$ b) 1/9

6.3.9. a) 2/3 b) $P(Y \in dp|AB^c) = 6p(1-p)dp$ for $0 < p < 1$

6.3.11. no

6.3.13. a) $1 - \frac{1}{3}e^{-\lambda}(3 + 2\lambda + \lambda^2/2)$ b) Partial answer: $P(X \in dx|X < Y) = \frac{1-e^{-\lambda}}{3-e^{-\lambda}(3+2\lambda+\frac{\lambda^2}{2})}dx$ for $0 \leq x < 1$ c) $\frac{9-e^{-\lambda}(9+8\lambda+\frac{5}{2}\lambda^2)}{6-2e^{-\lambda}(3+2\lambda+\frac{1}{2}\lambda^2)}$

6.3.15. b) $\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\binom{n}{k}\frac{\Gamma(r+k)\Gamma(s+n-k)}{\Gamma(r+s+n)}$
d) mean: $\frac{r+k}{r+s+n}$, variance: $\frac{(r+k)(s+n+k)}{(r+s+n)^2(r+s+n+1)}$

6.3.17. a)Independent negative binomial (r_i, p) $(i = 1, 2)$ b) negative binomial $(r_1 + r_2, p)$ c) negative binomial $(\sum_i r_i, p)$

6.4.1. a) 0.5 b) positively dependent c) 0.2 d) 0.356

6.4.3. less likely; more likely.

6.4.5. Uncorrelated, not independent.

6.4.7. a) Partial answer: $P(X_2 + X_3 = 0, X_2 - X_3 = 0) = 1/3$,
$P(X_2 + X_3 = 1, X_2 - X_3 = 0) = 0$, $P(X_2 + X_3 = 2, X_2 - X_3 = 0) = 1/6$
b) 1/6
c) uncorrelated

6.4.9. a) $k(n+1)/2$ b) $\frac{k(n^2-1)(n-k)}{12(n-1)}$

6.4.11. $\sqrt{1/3}$

6.4.13. True: note that $E(N_A N_B) = nP(AB) + n(n-1)P(A)P(B)$.

6.4.15. d) Write $N_R = \sum_{i=1}^{n} X_i$ and $N_B = \sum_{j=1}^{n} Y_j$, where X_i = 1 if the ith spin is red, = 0 otherwise; and Y_j = 1 if the jth spin is black, = 0 otherwise.

6.4.17. Apply Exercise 6.4.16.

6.4.19. a) 375 b) 26.25 c) 0.0021 d) higher; lower.

5.2.9. a) $2\lambda^2 e^{-\lambda(x+y)}$ $(0 < x < y)$, no; b) $2\lambda^2 e^{-2\lambda x - \lambda z}$ $(x > 0, z > 0)$, yes; c) X is exponential (2λ) and Z is exponential (λ).

5.2.11. a) 3/2 b) 1/2 c) 4/3 d) ∞

5.2.13. The distributions are all the same, with density $2(1-x)$ for $0 < x < 1$.

5.2.15. a) $F(b,d) - F(a,d) - F(b,c) + F(a,c)$ b) $F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)dudv$
c) $f(x,y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}F(x,y)$ d) $F(x,y) = F_X(x)F_Y(y)$
e) $F(x,y) = y^n - (y-x)^n$ for $0 < x < y < 1$;
$f(x,y) = n(n-1)(y-x)^{n-2}$ for $0 < x < y < 1$

5.2.17. a) $f(x,r) = \frac{2}{\pi}\frac{r}{\sqrt{r^2 - x^2}}$ for $0 \le r \le 1$ and $-r \le x \le r$
b) $f(x,r) = \frac{3}{2}r$ for $0 \le r \le 1$ and $-r \le x \le r$

5.2.19. a) $f_{\text{Lon}}(x) = 1/360$ if $-180 < x < 180$
b) $f_{\text{Lat}}(y) = \frac{\pi}{360}\cos\left(\frac{\pi}{180}y\right)$ if $-90 < y < 90$
c) $f(x,y) = \frac{1}{360}\cdot\frac{\pi}{360}\cos\left(\frac{\pi}{180}y\right)$ if $-180 < x < 180$ and $-90 < y < 90$
d) yes

5.2.21. a) 0.3825, 0.765 b) 1/3 c) 0.577 d) 0.5197 ± 0.0048

5.3.1. a) 0.1175 b) 0.1178 c) $\sqrt{\frac{2}{\pi}}$ d) 0.762 e) 0.58 f) 0.3521 g) 0.29

5.3.3. a) $1 - \Phi(0.5)$ b) 1/2 c) 5 d) $\sqrt{14}$

5.3.5. About 2.1

5.3.7. a) 97.72% b) 88.49% c) 0.9795

5.3.9. a) 0.1307 b) 0.0062 c) The answer to b) will be approximately the same.

5.3.11. a) normal with mean 0 and variance $t\sigma^2$ b) $\frac{R_t}{\sigma\sqrt{t}}$ has Rayleigh distribution so R_t has expectation $\sigma\sqrt{\frac{t\pi}{2}}$ and SD $\sigma\sqrt{t\left(\frac{4-\pi}{2}\right)}$ c) 0.1353

5.3.13. c) Try $h(u) = \sqrt{-2\log(1-u)}$ and $k(v) = 2\pi v$

5.3.15. *Hints:* a) Example 4.4.5 b) induction c) linear change of variable

5.3.17. a) Skew-normal approximations: 0.1377, 0.5940, 0.9196, 0.9998, 1.0000
Compare to the exact values: 0.0902, 0.5940, 0.9389, 0.9970, 1.0000
b) 0.441, 0.499. Skew-normal is better.

5.4.1. a) 3/4 b) $f_{X_1+X_2}(z) = z/2$ if $0 \le z \le 1$; $= 1/2$ if $1 \le z \le 2$; $= (3-z)/2$ if $2 \le z \le 3$ c) $F_{X_1+X_2}(z) = z^2/4$ if $0 \le z \le 1$; $= (2z-1)/4$ if $1 \le z \le 2$; $= 1 - (3-z)^2/4$ if $2 \le z \le 3$.

5.4.3. a) If $\alpha \ne \beta$, $f_{X+Y}(z) = \frac{\alpha\beta}{\alpha-\beta}(e^{-\beta z} - e^{-\alpha z})$ b) $\frac{1}{\alpha} + \frac{1}{\beta}$ c) $\frac{\sqrt{\alpha^2+\beta^2}}{\alpha\beta}$

Index

Springer Texts in Statistics *(continued from page ii)*

Santner and Duffy: The Statistical Analysis of Discrete Data
Saville and Wood: Statistical Methods: The Geometric Approach
Sen and Srivastava: Regression Analysis: Theory, Methods, and Applications
Whittle: Probability via Expectation, Third Edition
Zacks: Introduction to Reliability Analysis: Probability Models and Statistical Methods