PROBABILITY AND STATISTICAL INFERENCE

Probability vs. Statistics – Standard Viewpoint:

"Probability" postulates a probability model and uses this to predict the behavior of observed data.

"Statistics" uses observed data to infer the probability model (= distribution) from which the data was generated.

1. Probability Distributions and Random Variables.

1.1. The components (Ω, S, P) of a probability model (\equiv random experiment):

 $\Omega := sample \ space = set of all possible outcomes of the random experiment.$

 Ω either *discrete* (finite or countable) or *continuous* (\approx open subset of \mathbb{R}^n).

Example 1.1. Toss a coin *n* times: Ω = all sequences $HHTH \dots TH$ of length *n* (*H* = Heads, *T* = Tails). Thus $|\Omega| = 2^n$ (finite), so Ω is discrete.

Example 1.2. Toss a coin repeatedly until Heads appears and record the number of tosses needed: $\Omega = \{1, 2, ...\}$ (countable) so Ω is again discrete.

Example 1.3. Spin a pointer and record the angle where the pointer comes to rest: $\Omega = [0, 2\pi) \subset \mathbf{R}^1$, an entire interval, so Ω is continuous.

Example 1.4. Toss a dart at a circular board of radius d and record the impact point: $\Omega = \{(x, y) \mid x^2 + y^2 \leq d^2\} \subset \mathbb{R}^2$, a solid disk; Ω is continuous.

Example 1.5. Toss a coin *infinitely* many times: $\Omega = \text{all infinite sequences}$ $HHTH \ldots$ Here $\Omega \stackrel{1-1}{\longleftrightarrow} [0,1] \subset \mathbf{R}^1$ [why?], so Ω is continuous.

Example 1.6. (Brownian motion) Observe the path of a particle suspended in a liquid or gaseous medium: Ω is the set of all continuous paths (functions), so Ω is continuous but *not* finite-dimensional.

Note: Examples 1.5 and 1.6 are examples of *discrete-time* and *continuous-time stochastic processes*, respectively.

 $\mathcal{S} := class of (measurable) events A \subseteq \Omega$.

We assume that S is a σ -field ($\equiv \sigma$ -algebra), that is:

- (a) $\emptyset, \Omega \in \mathcal{S};$
- (b) $A \in \mathcal{S} \Rightarrow A^c \in \mathcal{S};$
- (c) $\{A_i\} \subseteq \mathcal{S} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{S}.$

In the discrete case, $S = 2^{\Omega}$ = all possible subsets of Ω is a σ -field.

In the continuous case, usually $S = \mathcal{B}^n(\Omega)$, the Borel σ -field generated by all open subsets of Ω . (Write \mathcal{B}^n for $\mathcal{B}^n(\mathbf{R}^n)$.)

(For stochastic processes, must define \mathcal{B}^{∞} carefully.)

$$P := a \text{ probability measure: } P(A) = \text{probability that } A \text{ occurs. Require:}$$
(a) $0 \le P(A) \le 1$;
(b) $P(\emptyset) = 0, P(\Omega) = 1$.
(c) $\{A_i\} \subseteq S, \text{ disjoint, } \Rightarrow P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$. (count'le additivity)
(c) $\Rightarrow P(A^c) = 1 - P(A)$.

In the **discrete case** where $\Omega = \{\omega_1, \omega_2, \ldots\}$, *P* is completely determined by the elementary probabilities $p_k \equiv P(\{\omega_k\}), k = 1, 2, \ldots$ This is because countable additivity implies that

(1.1)
$$P(A) = \sum_{\omega \in A} P(\{\omega\}) \quad \forall A \in \mathcal{S} \equiv 2^{\Omega}.$$

Conversely, given any set of numbers p_1, p_2, \ldots that satisfy

(a) $p_k \ge 0$, (b) $\sum_{k=1}^{\infty} p_k = 1$,

we can *define* a probability measure P on 2^{Ω} via (1.1) [verify countable additivity.] Here, $\{p_1, p_2, \ldots\}$ is called a *probability mass function (pmf)*.

Example 1.7. The following $\{p_k\}$ are pmfs [verify (a) and (b)]:

(1.2)
$$\binom{n}{k} p^k (1-p)^{n-k}, \ k = 0, 1, \dots, n \quad (0$$

(1.3)
$$(1-p)^{k-1}p, \ k = 1, 2, \dots$$
 $(0 [Geometric(p)];$

(1.4) $e^{-\lambda}\lambda^k/k!, \quad k = 0, 1, \dots \quad (\lambda > 0)$ [Poisson(λ)].

The binomial distribution occurs in Example 1.1; the geometric distribution occurs in Example 1.2; the Poisson distribution arises as the limit of the binomial distributions Bin(n, p) when $n \to \infty$ and $p \to 0$ such that $np \to \lambda$; [Discuss details].

In the **continuous case** where $(\Omega, S) = (\mathbf{R}^n, \mathcal{B}^n)$, let $f(\cdot)$ be a (Borelmeasurable) function on \mathbf{R}^n that satisfies

(a)
$$f(x_1, \ldots, x_n) \ge 0 \quad \forall (x_1, \ldots, x_n) \in \mathbf{R}^n$$
,
(b) $\int \ldots \int_{\mathbf{R}^n} f(x_1, \ldots, x_n) \, dx_1 \ldots dx_n = 1$.

Then $f(\cdot)$ defines a probability measure P on $(\mathbf{R}^n, \mathcal{B}^n)$ by

(1.5)
$$P(B) = \int \dots \int_{B} f(x_1, \dots, x_n) \, dx_1 \dots dx_n \quad \forall B \in \mathcal{B}^n.$$

The function $f(\cdot)$ is called a *probability density function* (pdf) on \mathbb{R}^n . Note that in the continuous case, unlike the discrete case, it follows from (1.5) that singleton events $\{x\}$ have probability 0.

Example 1.8. The following f(x) are pdfs on \mathbb{R}^1 or \mathbb{R}^n [verify all]:

$$(1.6) \ \lambda e^{-\lambda x} I_{(0,\infty)}(x) \quad (\lambda > 0) \qquad [\text{Exponential}(\lambda)];$$

$$(1.7) \ \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2} \quad (\sigma > 0) \qquad [\text{Normal}(\mu, \sigma^2) \equiv N(\mu, \sigma^2)];$$

$$(1.8) \ \frac{1}{\pi\sigma} \frac{1}{1+(x-\mu)^2/\sigma^2} \quad (\sigma > 0) \qquad [\text{Cauchy}(\mu, \sigma^2) \equiv C(\mu, \sigma^2)];$$

$$(1.9) \ \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} I_{(0,\infty)}(x) \quad (\alpha, \lambda > 0) \qquad [\text{Gamma}(\alpha, \lambda)];$$

$$(1.10)\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}I_{(0,1)}(x) \quad (\alpha,\beta>0) \quad [\text{Beta}(\alpha,\beta)];$$

$$(1.11)\frac{e^{x}}{(1+e^{x})^{2}} = \frac{e^{-x}}{(1+e^{-x})^{2}} \qquad [\text{standard Logistic}].$$

$$(1.12)\frac{1}{b-a}I_{(a,b)}(x) \quad ((a,b) \in \mathbf{R}^{1}) \qquad [\text{Uniform}(a,b)].$$

$$(1.13)\frac{1}{\text{volume}(C)}I_{C}(x) \quad (x=(x_{1},\ldots,x_{n}), C \in \mathbf{R}^{n}) \quad [\text{Uniform}(C)].$$

Here, I_A is the indicator function of the set A: $I_A(x) = 1(0)$ if $x \in (\notin)A$. For the Uniform(C) pdf in (1.13), it follows from (1.5) that for any $A \subseteq C$,

(1.14)
$$P(A) = \frac{\text{volume}(A)}{\text{volume}(C)}.$$

The exponential distribution appears as the distribution of waiting times between events in a *Poisson process* – cf. §3.5. According to the *Central Limit Theorem* (cf. §3.5), the normal \equiv Gaussian distribution occurs as the limiting distribution of sample averages (suitably standarized).

1.2. Random variables, pmfs, cdfs, and pdfs.

Often it is convenient to represent a feature of the outcome of a random experiment by a random variable (rv), usually denoted by a capital letter X, Y, Z, etc. Thus in Example 1.1, $X \equiv$ the total number of Heads in the n trials and $Y \equiv$ the length of the longest run of Tails in the same n trials are both random variables. This shows already that two or more random variables may arise from the same random experiment. Additional random variables may be constructed by arithmetic operations, e.g., $Z \equiv X + Y$ and $W \equiv XY^3$ are also random variables arising in Example 1.1.

Formally, a random variable defined from a probability model (Ω, S, P) is simply a (measurable) function defined on Ω . Each random variable Xdetermines its own *induced* probability model $(\Omega_X, \mathcal{S}_X, P_X)$, where Ω_X is the *range* of possible values of X, \mathcal{S}_X the appropriate σ -field of measurable events in Ω , and P_X the probability distribution induced on $(\Omega_X, \mathcal{S}_X)$ from P by X: for any $B \in \mathcal{S}_X$,

(1.15)
$$P_X(B) \equiv P[X \in B] := P[X^{-1}(B)] \equiv P[\{\omega \in \Omega \mid X(\omega) \in B\}].$$

If in Example 1.5 we define X := the number of trials needed to obtain the first Head, then the probability model for X is exactly that of Example 1.2. This shows that the same random experiment can give rise to different probability models.

A multivariate rv (X_1, \ldots, X_n) is called a random vector (rvtr). It is important to realize that the individual rvs X_1, \ldots, X_n are related in that they must arise from the same random experiment. Thus they may (or may not) be correlated. [Example: (X, Y) = (height,weight); other examples?] One goal of statistical analysis is to study the relationship among correlated rvs for purposes of prediction.

The random variable (or random vector) X is called *discrete* if its range Ω_X is discrete, and *continuous* if Ω_X is continuous. As in (1.1), the probability distribution P_X of a discrete random variable is completely determined by its *probability mass function (pmf)*

(1.16)
$$f_X(x) := P[X = x], \qquad x \in \Omega_X.$$

For a univariate continuous random variable X with pdf f_X on \mathbb{R}^1 , it is convenient to define the cumulative distribution function (cdf) as follows:

(1.17)
$$F_X(x) := P[X \le x] = \int_{-\infty}^x f_X(t) dt, \quad x \in \mathbf{R}^1,$$

The pdf f_X can be recovered from the cdf F_X as follows:

(1.18)
$$f_X(x) = \frac{d}{dx} F_X(x), \qquad x \in \mathbf{R}^1.$$

Clearly F_X directly determines the probabilities of all intervals in \mathbb{R}^1 :

(1.19)
$$P_X[(a,b]] \equiv \Pr[X \in (a,b]] = F_X(b) - F_X(a).$$

In fact, F_X completely determines¹ the probability distribution P_X on \mathbb{R}^1 . *Note:* The cdf F_X is also defined for *univariate discrete* random variables by (1.17). Now F_X determines the pmf p_X not by (1.18) but by

(1.20)
$$f_X(x) \equiv P[X = x] = F_X(x) - F_X(x-), \quad x \in \mathbf{R}^1 \quad [verify].$$

F:

Basic properties of a cdf F on \mathbb{R}^1 :

(i)
$$F(-\infty) = 0 \le F(x) \le 1 = F(+\infty)$$
.

(ii) $F(\cdot)$ is non-decreasing and right-continuous: F(x) = F(x+).

For a continuous multivariate rvtr (X_1, \ldots, X_n) the joint cdf is

(1.21)
$$F_{X_1,...,X_n}(x_1,...,x_n) := P[X_1 \le x_1,...,X_n \le x_n],$$

from which the *joint pdf* f is recovered as follows:

(1.22)
$$f_{X_1,\dots,X_n}(x_1,\dots,x_n):\frac{\partial^n}{\partial x_1\cdots\partial x_n}F_{X_1,\dots,X_n}(x_1,\dots,x_n).$$

Exercise 1.1. Extend (1.19) to show that for n = 2, the cdf F directly determines the probabilities of all rectangles in \mathbb{R}^2 .

¹ Since any Borel set $B \subset R^1$ can be approximated by finite disjoint unions of intervals.

For a discrete multivariate rvtr (X_1, \ldots, X_n) , the joint cdf F_{X_1, \ldots, X_n} is again defined by (1.21). The joint pmf is given by

(1.23)
$$f_{X_1,\dots,X_n}(x_1,\dots,x_n) := P[X_1 = x_1,\dots,X_n = x_n],$$

from which all joint probabilities can be determined as in (1.1).

The marginal pmf or pdf of any X_i can be recovered from the joint pmf or pdf by summing or integrating over the other variables. The marginal cdf can also be recovered from the joint cdf. In the bivariate case (n = 2), for example, if the rvtr (X, Y) has joint pmf $f_{X,Y}$ or joint pdf $f_{X,Y}$, and joint cdf $F_{X,Y}$, then, respectively,

(1.24) $f_X(x) = \sum_y f_{X,Y}(x,y);$ [verify via countable additivity] (1.25) $f_X(x) = \int f_{X,Y}(x,y) \, dy.$ [verify via (1.18) and (1.17)] (1.26) $F_X(x) = F_{X,Y}(x,\infty).$ [verify via (1.21)]

The joint distribution contains information about X and Y beyond their marginal distributions, i.e., information about the nature of any dependence between them. Thus, the joint distribution determines all marginal distributions but not conversely (except under independence – cf. (1.32), (1.33).)

1.3. Conditional probability.

Let (Ω, \mathcal{S}, P) be a probability model. Let $B \in \mathcal{S}$ be an event such that P(B) > 0. If we are told that B has occurred but given no other information, then the original probability model is reduced to the *conditional* probability model $(\Omega, \mathcal{S}, P[\cdot | B])$, where for any $A \in \mathcal{S}$,

(1.27)
$$P[A \mid B] = \frac{P(A \cap B)}{P(B)}.$$

Then $P[\cdot | B]$ is also a probability measure [verify] and P[B | B] = 1, i.e., $P[\cdot | B]$ assigns probability 1 to B. Thus Ω is reduced to B and, by (1.27), events within B retain the same relative probabilities.

Example 1.9. Consider the Uniform(C) probability model ($\mathbb{R}^n, \mathcal{B}^n, P_C$) determined by (1.13). If $B \subset C$ and volume(B) > 0, then the conditional distribution $P_C[\cdot|B] = P_B$, the Uniform(B) distribution [verify via (1.14)].

Example 1.10. Let X be a random variable whose distribution on $[0, \infty)$ is determined by the exponential pdf in (1.6). Then for x, y > 0,

$$P[X > x + y \mid X > y] = \frac{P[X > x + y]}{P[X > y]}$$
$$= \frac{\int_{x+y}^{\infty} \lambda e^{-\lambda t} dt}{\int_{y}^{\infty} \lambda e^{-\lambda t} dt} = \frac{e^{-\lambda(x+y)}}{e^{-\lambda y}} = e^{-\lambda x}.$$

Because $e^{-\lambda x} = P[X > x]$, this can be interpreted as follows: the exponential distribution is *memory-free*; i.e., given that we have waited at least y time units, the probability of having to wait an additional x time units is the same as the unconditional probability of waiting at least x units from the start.

Exercise 1.2. Show that the exponential distribution is the *only* distribution on $(0, \infty)$ with this memory-free property. That is, show that if X is a continuous rv on $(0, \infty)$ such that $P[X > x + y \mid X > y] = P[X > x]$ for every x, y > 0, then $f_X(x) = \lambda e^{-\lambda x} I_{(0,\infty)}(x)$ for some $\lambda > 0$.

1.4. Conditional pmfs and pdfs.

Let (X, Y) be a *discrete* bivariate rvtr with joint pmf $f_{X,Y}$. For any $x \in \Omega_X$ such that P[X = x] > 0, the *conditional pmf* of Y given X = x is defined by

(1.28)
$$f_{Y|X}(y|x) \equiv P[Y = y \mid X = x] = \frac{f_{X,Y}(x,y)}{f_X(x)},$$

where the second equality follows from (1.27). As in (1.1), the conditional pmf completely determines the conditional distribution of Y given X = x:

(1.29)
$$P[Y \in B \mid X = x] = \sum_{y \in B} f_{Y|X}(y|x) \quad \forall B. \qquad [verify]$$

"Slicing": discrete case:

continuous case:

Next let (X, Y) be a *continuous* bivariate rvtr with joint pdf $f_{X,Y}$. By analogy with (1.28), for any $x \in \Omega_X$ such that the marginal pdf $f_X(x) > 0$, the *conditional pdf* of Y given X = x is defined by

(1.30)
$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

As in (1.29), the conditional pdf (1.30) completely determines the conditional distribution of Y given X = x:

(1.31)
$$P[Y \in B \mid X = x] = \int_B f_{Y|X}(y|x) \, dy \qquad \forall B$$

Note that $P[Y \in B \mid X = x]$ cannot be interpreted as a conditional probability for events via (1.27), since P[X = x] = 0 for every x in the continuous case. Instead, (1.31) will be given a more accurate interpretation in §4.

1.5. Independence.

Two events A and B are *independent* under the probability model (Ω, \mathcal{S}, P) , denoted as $A \perp B$ [P] or simply $A \perp B$, if any of the following three equivalent [verify!] conditions hold:

- $(1.32) P[A \cap B] = P[A]P[B];$
- (1.33) $P[A \mid B] = P[A];$

$$(1.34) P[B \mid A] = P[B].$$

Intuitively, $A \perp B$ means that information about the occurrence (or non-occurrence!) of either event does not change the probability of occurrence or non-occurrence for the other.

Exercise 1.3. Show that $A \perp\!\!\!\perp B \Leftrightarrow A \perp\!\!\!\perp B^c \Leftrightarrow A^c \perp\!\!\!\perp B \Leftrightarrow A^c \perp\!\!\!\perp B^c$.

	B	B^c	Venn:	
A	$A \cap B$	$A\cap B^c$		
A^c	$A^c\cap B$	$A^c\cap B^c$		Ū

Two rvs X and Y are independent under the model (Ω, \mathcal{S}, P) , denoted as $X \perp \!\!\!\perp Y [P]$ or simply $X \perp \!\!\!\perp Y$, if $\{X \in A\}$ and $\{Y \in B\}$ are independent for each pair of measurable events $A \in \Omega_X$ and $B \in \Omega_Y$. It is straightforward to show that for a *jointly discrete or jointly continuous bivariate* rvtr $(X, Y), X \perp \!\!\!\perp Y$ iff any of the following four equivalent conditions hold [verify]:

- (1.35) $f_{X,Y}(x,y) = f_X(x)f_Y(y) \quad \forall (x,y) \in \Omega_{X,Y};$
- (1.36) $f_{Y|X}(y|x) = f_Y(y) \qquad \forall (x,y) \in \Omega_{X,Y};$
- (1.37) $f_{X|Y}(x|y) = f_X(x) \qquad \forall (x,y) \in \Omega_{X,Y};$
- (1.38) $F_{X,Y}(x,y) = F_X(x)F_Y(y) \quad \forall (x,y) \in \Omega_{X,Y}.$

Intuitively, it follows from (1.36) and (1.37) that independence of rvs means that information about the values of one of the rvs does not change the probability distribution of the other rv. It is important to note that this requires that the *joint range of* (X, Y) *is the Cartesian product of the marginal ranges:*

(1.39)
$$\Omega_{X,Y} = \Omega_X \times \Omega_Y.$$

Example 1.11. Let U, V be independent Uniform(0,1) rvs and set $X = \min(U, V), Y = \max(U, V)$. Then the range of (X, Y) is given by (1.40) $\Omega_{X,Y} = \{(x, y) \mid 0 \le x \le y \le 1\}$:

Because $\Omega_{X,Y}$ is not a Cartesian product set, X and Y cannot be independent. [In fact, they are *positively correlated* – why?]

Exercise 1.4. Condition (1.39) is necessary for mutual independence. Show by counterexample that it is not sufficient.

Example 1.12. Let $(X, Y) \sim \text{Uniform}(D)$, where $D \equiv \{x^2 + y^2 \leq 1\}$ denotes the unit disk in \mathbb{R}^2 . (Recall Example 1.4.) By (1.13), the joint pdf of X, Y is

(1.41)
$$f_{X,Y}(x,y) = \frac{1}{\pi} I_D(x,y):$$

In particular, the range of (X, Y) is D. However, the marginal ranges of X and Y are both [-1, 1], so (1.39) fails, hence X and Y are not independent. More precisely, it follows from (1.25) that the marginal pdf of X is

More precisely, it follows from (1.25) that the marginal pdf of X is

(1.42)
$$f_X(x) = \frac{1}{\pi} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} dy = \frac{2}{\pi} \sqrt{1-x^2} I_{(-1,1)}(x)$$

and similarly $f_Y(y) = \frac{2}{\pi}\sqrt{1-y^2}I_{(-1,1)}(y)$. Thus by (1.41), (1.35) fails, hence X and Y are not independent. [But they are uncorrelated: no *linear* trend – verify.]

The dependence of X and Y can also be seen from the conditional pdf $f_{Y|X}$ (recall (1.36)). From (1.30) and (1.41),

(1.43)
$$f_{Y|X}(y|x) = \frac{1}{2\sqrt{1-x^2}} I_{(-\sqrt{1-x^2},\sqrt{1-x^2})}(y) \neq f_Y(y),$$

so (1.36) fails and X and Y are not independent. Note that (1.43) is equivalent to the statement that the conditional distribution of Y|X is uniform on the interval $\left(-\sqrt{1-x^2},\sqrt{1-x^2}\right)$, i.e.,

(1.44)
$$Y|X = x \sim \text{Uniform}(-\sqrt{1-x^2}, \sqrt{1-x^2}),$$

which is already obvious from the following figure:

If, however, we represent the rvtr (X, Y) in polar coordinates (R, Θ) , then $R \perp \Box \Theta$. This is readily verified: clearly $\Omega_{R,\Theta} = \Omega_R \times \Omega_{\Theta}$ [verify], while by (1.41) (uniformity),

(1.45)

$$F_{R,\Theta}(r,\theta) \equiv P[0 \le R \le r, \ 0 \le \Theta \le \theta]$$

$$= \frac{\pi r^2 \cdot [\theta/(2\pi)]}{\pi}$$

$$= r^2 \cdot [\theta/(2\pi)]$$

$$= P[0 \le R \le r] P[0 \le \Theta \le \theta]$$

$$\equiv F_R(r) \cdot F_{\Theta}(\theta)$$

so (1.38) holds. It follows too that

(1.46a) $f_R(r) = 2rI_{(0,1)}(r);$

(1.46b)
$$f_{\Theta}(\theta) = \frac{1}{2\pi} I_{[0,2\pi)}(\theta);$$

the latter states that $\Theta \sim \text{Uniform}[0, 2\pi)$.

 \Box

Mutual independence. Events A_1, \ldots, A_n $(n \ge 3)$ are mutually independent iff

(1.47)
$$P(A_1^{\epsilon_1} \cap \dots \cap A_n^{\epsilon_n}) = P(A_1^{\epsilon_1}) \cdots P(A_n^{\epsilon_n}) \ \forall (\epsilon_1, \dots, \epsilon_n) \in \{0, 1\}^n,$$

where $A^1 := A$ and $A^0 := A^c$. A finite family² of rvs X_1, \ldots, X_n are *mutually independent* iff $\{X_1 \in B_1\}, \ldots, \{X_n \in B_n\}$ are mutually independent for every choice of measurable events B_1, \ldots, B_n . An infinite family X_1, X_2, \ldots of rvs are *mutually independent* iff every finite subfamily is mutually independent. Intuitively, mutual independence of rvs means that information about the values of some of the rvs does not change the (joint) probability distribution of the other rvs.

Exercise 1.5. (i) For $n \geq 3$ events A_1, \ldots, A_n , show that mutual independence implies pairwise independence. (ii) Show by counterexample that pairwise independence does not imply mutual independence. (iii) Show that

(1.48)
$$P(A \cap B \cap C) = P(A)P(B)P(C)$$

is not by itself sufficient for mutual independence of A, B, C.

Example 1.13. In Example 1.1, suppose that the *n* trials are mutually independent and that p: P(H) and $q \equiv (1-p) \equiv P(T)$ do not vary from trial to trial. Let X denote the total number of Heads in the *n* trials. Then by independence, $X \sim \text{Binomial}(n, p)$, i.e., the pmf p_X of X is given by (1.2). [Verify!].

Example 1.14. In Example 1.2, suppose that the entire infinite sequence of trials are mutually independent and that p := P(H) does not vary from trial to trial. Let X denote the number of trials needed to obtain the first Head. Then by independence, $X \sim \text{Geometric}(p)$, i.e., the pmf p_X is given by (1.3). [Verify!]

Mutual independence of rvs X_1, \ldots, X_n can be expressed in terms of their joint pmf (discrete case), joint pdf (continuous case), or joint cdf (both

² More precisely, (X_1, \ldots, X_n) must be a rvtr, i.e., X_1, \ldots, X_n arise from the same random experiment.

cases): X_1, \ldots, X_n are mutually independent iff either

(1.49)
$$f_{X_1,\dots,X_n}(x_1,\dots,x_n) = f_{X_1}(x_1)\cdots f_{X_n}(x_n);$$

(1.50)
$$F_{X_1,\dots,X_n}(x_1,\dots,x_n) = F_{X_1}(x_1)\cdots F_{X_n}(x_n).$$

Again, these conditions implicitly require that the joint range of X_1, \ldots, X_n is the Cartesian product of the marginal ranges:

(1.51)
$$\Omega_{X_1,\dots,X_n} = \Omega_{X_1} \times \dots \times \Omega_{X_n}.$$

Example 1.15. Continuing Example 1.13, let X_1, \ldots, X_n be indicator variables (\equiv *Bernoulli variables*) that denote the outcomes of trials $1, \ldots, n$: $X_i = 1$ or 0 according to whether Heads or Tails occurs on the *i*th trial. Here X_1, \ldots, X_n are *mutually independent and identically distributed (i.i.d.)* rvs, and X can be represented as their sum: $X = X_1 + \cdots + X_n$. Therefore we expect that X and X_1 are *not* independent. In fact, the joint range

$$\Omega_{X,X_1} = \{ (x, x_1) \mid x = 0, 1, \dots, n, \ x_1 = 0, 1, \ x \ge x_1 \}.$$

The final inequality implies that this is not a Cartesian product of an x-set and an x_1 -set [verify], hence by (1.39), X and X_1 cannot be independent.

1.6. Composite events and total probability.

Equation (1.27) can be rewritten in the following useful form(s):

(1.52)
$$P(A \cap B) = P[A \mid B] P(B) \ \left(= P[B \mid A] P(A) \right).$$

By (1.28) and (1.30), similar formulas hold for joint pmfs and pdfs:

(1.53)
$$f(x,y) = f(x|y)f(y) \ \left(= f(y|x)f(x) \right).$$

Now suppose that the sample space Ω is partitioned into a finite or countable set of disjoint events: $\Omega = \bigcup_{i=1}^{\infty} B_i$. Then by the countable additivity of Pand (1.52), we have the *law of total probability*:

(1.54)
$$P(A) = P\left[A \cap \left(\bigcup_{i=1}^{\infty} B_i\right)\right] = P\left[\bigcup_{i=1}^{\infty} \left(A \cap B_i\right)\right]$$
$$= \sum_{i=1}^{\infty} P[A \cap B_i] = \sum_{i=1}^{\infty} P[A \mid B_i] P(B_i).$$

Example 1.16. Let $X \sim \text{Poisson}(\lambda)$, where $\lambda > 0$ is unknown and is to be estimated. (For example, λ might be the decay rate of a radioactive process and X the number of emitted particles recorded during a unit time interval.) We shall later see that the *expected value* of X is given by $E(X) = \lambda$, so if X were observed then we would estimate λ by $\hat{\lambda} \equiv X$.

Suppose, however, that we do not observe X but instead only observe the value of Y, where

(1.55)
$$Y|X=x \sim \text{Binomial}(n=x, p).$$

(This would occur if each particle emitted has probability p of being observed, independently of the other emitted particles.) If p is known then we may still obtain a reasonable estimate of λ based on Y, namely $\tilde{\lambda} = \frac{1}{p}Y$.

To obtain the distribution of Y, apply (1.54) as follows (set q = (1-p)):

$$P[Y = y] = P[Y = y, X = y, y + 1, ...] \qquad [since X \ge Y]$$
$$= \sum_{x=y}^{\infty} P[Y = y \mid X = x] \cdot P[X = x] \qquad [by (1.54)]$$
$$= \sum_{x=y}^{\infty} {x \choose y} p^y q^{x-y} \cdot e^{-\lambda} \frac{\lambda^x}{x!} \qquad [by (1.2), (1.4)]$$
$$= \frac{e^{-\lambda} (p\lambda)^y}{y!} \sum_{k=0}^{\infty} \frac{(q\lambda)^k}{k!} \qquad [let \ k = x - y]$$
$$(1.56) \qquad = \frac{e^{-p\lambda} (p\lambda)^y}{y!}.$$

This implies that $Y \sim \text{Poisson}(p\lambda)$, so

$$\mathbf{E}(\tilde{\lambda}) \equiv \mathbf{E}(\frac{1}{p}Y) = \frac{1}{p}\mathbf{E}(Y) = \frac{1}{p}(p\lambda) = \lambda,$$

which justifies $\tilde{\lambda}$ as an estimate of λ based on Y.

 \Box

1.7. Bayes formula.

If P(A) > 0 and P(B) > 0, then (1.27) yields Bayes formula for events:

(1.57)
$$P[A \mid B] = \frac{P[A \cap B]}{P(B)} = \frac{P[B \mid A]P(A)}{P(B)}$$

Similarly, (1.28) and (1.30) yield Bayes formula for joint pmfs and pdfs:

(1.58)
$$f(x|y) = \frac{f(y|x)f(x)}{f(y)} \quad \text{if } f(x), f(y) > 0.$$

[See §4 for extensions to the *mixed* cases where X is discrete and Y is continuous, or vice versa.]

Example 1.17. In Example 1.16, what is the conditional distribution of X given that Y = y? By (1.58), the conditional pmf of $X \mid Y = y$ is

$$f(x|y) = \frac{\binom{x}{y}p^{y}q^{x-y} \cdot e^{-\lambda}\lambda^{x}/x!}{e^{-p\lambda}(p\lambda)^{y}/y!}$$
$$= \frac{e^{-q\lambda}(q\lambda)^{x-y}}{(x-y)!}, \qquad x = y, \ y+1, \dots$$

Thus, if we set Z = X - Y, then

(1.59)
$$P[Z = z \mid Y = y] = \frac{e^{-q\lambda}(q\lambda)^z}{z!}, \quad z = 0, 1, \dots,$$

so $Z | Y = y \sim \text{Poisson}(q\lambda)$. Because this conditional distribution does not depend on y, it follows from (1.36) that $X - Y \perp Y$. (In the radioactivity scenario, this states that the number of uncounted particles is independent of the number of counted particles.)

Note: this also shows that if $U \sim \text{Poisson}(\mu)$ and $V \sim \text{Poisson}(\nu)$ with Uand V independent, then $U + V \sim \text{Poisson}(\mu + \nu)$. [Why?] \Box **Exercise 1.6.** (i) Let X and Y be independent Bernoulli rvs with

$$P[X = 1] = p,$$
 $P[X = 0] = 1 - p;$
 $P[Y = 1] = r,$ $P[Y = 0] = 1 - r.$

Let Z = X + Y, a discrete rv with range $\{0, 1, 2\}$. Do there exist p, r such that Z is uniformly distributed on its range, i.e., such that $P[Z = k] = \frac{1}{3}$ for k = 0, 1, 2? (Prove or disprove.)

(ii)* (unfair dice.) Let X and Y be independent discrete rvs, each having range $\{1, 2, 3, 4, 5, 6\}$, with pmfs

$$p_X(k) = p_k, \quad p_Y(k) = r_k, \quad k = 1, \dots, 6.$$

Let Z = X + Y, a discrete rv with range $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. First note that if X and Y are the outcomes of tossing two fair dice, i.e. $p_X(k) = p_Y(k) = \frac{1}{6}$ for $k = 1, \ldots, 6$, then the pmf of Z is given by

$$\frac{1}{36}, \frac{2}{36}, \frac{3}{36}, \frac{4}{36}, \frac{5}{36}, \frac{6}{36}, \frac{5}{36}, \frac{4}{36}, \frac{3}{36}, \frac{2}{36}, \frac{1}{36}, \frac{1}$$

which is *not* uniform over its range. Do there exist *unfair* dice such that Z is uniformly distributed over its range, i.e., such that $P[Z = k] = \frac{1}{11}$ for $k = 2, 3, \ldots, 12$? (Prove or disprove.)

1.8. Conditional independence.

Consider three events A, B, C with P(C) > 0. We say that A and B are *conditionally independent given* C, written $A \perp B \mid C$, if any of the following three equivalent conditions hold (recall (1.32) - (1.34)):

(1.60) $P[A \cap B \mid C] = P[A \mid C]P[B \mid C];$

(1.61) $P[A \mid B, C] = P[A \mid C];$

(1.62) $P[B \mid A, C] = P[B \mid C].$

As with ordinary independence, $A \perp\!\!\!\perp B \mid C \Leftrightarrow A \perp\!\!\!\perp B^c \mid C \Leftrightarrow A^c \perp\!\!\!\perp B \mid C$ $\Leftrightarrow A^c \perp\!\!\!\perp B^c \mid C$ (see Exercise 1.3). However, $A \perp\!\!\!\perp B \mid C \not\Leftrightarrow A \perp\!\!\!\perp B \mid C^c$ [Examples are easy – verify]. The rvs X and Y are conditionally independent given Z, denoted as $X \perp\!\!\!\perp Y \mid Z$, if

$$(1.63) \qquad \qquad \{X \in A\} \perp \{Y \in B\} \mid \{Z \in C\}$$

for each triple of (measurable) events A, B, C. It is straightforward to show that for a *jointly discrete or jointly continuous trivariate* rvtr (X, Y, Z), $X \perp \!\!\!\perp Y \mid Z$ iff any of the following three equivalent conditions hold [verify]:

(1.64)
$$f(x, y \mid z) = f(x \mid z)f(y \mid z);$$

(1.65)
$$f(y \mid x, z) = f(y \mid z);$$

(1.66)
$$f(x \mid y, z) = f(x \mid z);$$

(1.67) f(x, y, z)f(z) = f(x, z)f(y, z);

Exercise 1.7. Conditional independence \Leftrightarrow independence.

- (i) Construct (X, Y, Z) such that $X \perp \!\!\!\perp Y \mid Z$ but $X \not\perp Y$.
- (ii) Construct (X, Y, Z) such that $X \perp \!\!\!\perp Y$ but $X \not\perp \!\!\!\perp Y \mid Z$.

 \Box

Graphical Markov model representation of $X \perp\!\!\!\perp Y \mid Z$:

(1.68)
$$X < --- Z - --> Y.$$

2. Transforming Continuous Distributions.

2.1. One function of one random variable.

Let X be a continuous rv with pdf f_X on the range $\Omega_X = (a, b)$ $(-\infty \leq a < b \leq \infty)$. Define the new rv Y = g(X), where g is a strictly increasing and continuous function on (a, b). Then the pdf f_Y is determined as follows:

Theorem 2.1.

(2.1)
$$f_Y(y) = \begin{cases} \frac{f_X(g^{-1}(y))}{g'(g^{-1}(y))}, & g(a) < y < g(b); \\ 0, & otherwise. \end{cases}$$

Proof.

Example 2.1. Consider $g(x) = x^2$. In order that this g be strictly increasing we must have $0 \le a$. Then g'(x) = 2x and $g^{-1}(y) = \sqrt{y}$, so from (2.1) with $Y = X^2$,

(2.3)
$$f_Y(y) = \frac{1}{2\sqrt{y}} f_X(\sqrt{y}), \qquad a^2 < y < b^2.$$

In particular, if $X \sim \text{Uniform}(0,1)$ then $Y \equiv X^2$ has pdf

(2.4)
$$f_Y(y) = \frac{1}{2\sqrt{y}}, \quad 0 < y < 1.$$
 [decreasing]

Example 2.2a. If X has cdf F then $Y \equiv F(X) \sim \text{Uniform}(0,1)$. [Verify]

Example 2.2b. How to generate a rv Y with a pre-specified pdf f:

Solution: Let F be the cdf corresponding to f. Use a computer to generate $X \sim \text{Uniform}(0,1)$ and set $Y = F^{-1}(X)$. Then Y has cdf F. [Verify] \square

Note: If g is strictly decreasing then (2.1) remains true with $g'(g^{-1}(y))$ replaced by $|g'(g^{-1}(y))|$ [Verify].

Now suppose that g is not monotone. Then (2.2) remains valid, but the region $\{g(X) \leq y\}$ must be specifically determined before proceeding.

Example 2.3. Again let $Y = X^2$, but now suppose that the range of X is $(-\infty, \infty)$. Then for y > 0,

$$f_Y(y) = \frac{d}{dy} P[Y \le y]$$

$$= \frac{d}{dy} P[X^2 \le y]$$

$$= \frac{d}{dy} P[-\sqrt{y} \le X \le \sqrt{y}]$$

$$= \frac{d}{dy} \Big[F_X(\sqrt{y}) - F_X(-\sqrt{y}) \Big]$$

$$= \frac{1}{2\sqrt{y}} \Big[f_X(\sqrt{y}) + f_X(-\sqrt{y}) \Big].$$

If in addition the distribution of X is symmetric about 0, i.e., $f_X(x) = f_X(-x)$, then (2.5) reduces to

(2.6)
$$f_Y(y) = \frac{1}{\sqrt{y}} f_X(\sqrt{y}).$$

Note that this is similar to (2.3) (where the range of X was restricted to $(0,\infty)$) but without the factor $\frac{1}{2}$. To understand this, consider

$$X_1 \sim \text{Uniform}(0,1), \qquad X_2 \sim \text{Uniform}(-1,1).$$

Then

$$f_{X_1}(x) = I_{(0,1)}(x), \qquad f_{X_2}(x) = \frac{1}{2}I_{(-1,1)}(x),$$

but $Y_i = X_i^2$ has pdf $\frac{1}{2\sqrt{y}}I_{(0,1)}$ for $i = 1, 2$. [Verify – recall (2.4)]. \square

Example 2.4. Let $X \sim N(0,1)$, i.e., $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} I_{(-\infty,\infty)}(x)$ and $Y = X^2$. Then by (2.6),

(2.7)
$$f_Y(y) = \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2} I_{(0,\infty)}(y),$$

the Gamma $(\frac{1}{2}, \frac{1}{2})$ pdf, which is also called the *chi-square pdf with one degree* of freedom, denoted as χ_1^2 (see Remark 6.3). Note that (2.7) shows that [verify!]

(2.8)
$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

Exercise 2.1. Let $\Theta \sim \text{Uniform}(0, 2\pi)$, so we may think of Θ as a *random angle*. Define $X = \cos \Theta$. Find the pdf f_X .

Hint: Always begin by specifying the range of X, which is [-1, 1] here. On this range, what shape do you expect f_X to have, among the following three possibilities? (Compare this f_X to that in Example 1.12, p. 11.)

Exercise 2.1 suggests the following problem. A bivariate pdf $f_{X,Y}$ on \mathbb{R}^2 is called *radial* if it has the form

(2.9)
$$f_{X,Y}(x,y) = g(x^2 + y^2)$$

for some (non-negative) function g on $(0, \infty)$. Note that the condition

$$\iint_{\mathbf{R}^2} f(x,y) dx dy = 1$$

requires that

(2.10)
$$\int_{0}^{\infty} rg(r^{2})dr = \frac{1}{2\pi} \qquad [\text{why?}]$$

Exercise 2.2*. Does there exist a radial pdf $f_{X,Y}$ on the unit disk in \mathbb{R}^2 such that the marginal distribution of X is Uniform(-1,1)? More precisely, does there exist g on (0,1) that satisfies (2.10) and

(2.11)
$$f_X(x) \equiv \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} g(x^2 + y^2) dy = \frac{1}{2} I_{(-1,1)}(x)?$$

Note: if such a radial pdf $f_{X,Y}$ on the unit disk exists, it could be called a bivariate uniform distribution, since both X and Y (by symmetry) have the Uniform(-1, 1) distribution. Of course, there are simpler bivariate distributions with these uniform marginal distributions but which are not radial on the unit disk. [Can you think of two?]

2.2. One function of two or more random variables.

Let (X, Y) be a *continuous* bivariate rvtr with pdf $f_{X,Y}$ on a subset of \mathbb{R}^2 . Define a new rv

$$U = g(X, Y), \text{ e.g.}, U = X + Y, X - Y, \frac{X}{Y}, \frac{1 + \exp(X + Y)}{1 + \exp(X - Y)}.$$

Then the pdf f_U can be determined via two methods:

Method One: Apply

(2.12)
$$f_U(u) = \frac{d}{du} F_U(u) = \frac{d}{du} P[U \le u] = \frac{d}{du} P[g(X, Y) \le u],$$

then determine the region $\{(X, Y) \mid g(X, Y) \leq u\}.$

Example 2.5. Let (X, Y) be uniformly distributed on the unit square. [Note that this is equivalent to assuming that X and Y are independent Uniform(0, 1) rvs – why?] To find the pdf f_U of U = X + Y, begin by noting that the range of U is the interval [0, 2]. Then (see Figure)

$$P[X + Y \le u] = \begin{cases} \frac{1}{2}u^2, & 0 < u < 1;\\ 1 - \frac{1}{2}(2 - u)^2, & 1 < u < 2; \end{cases}$$

 \mathbf{SO}

(2.13)
$$f_U(u) = \begin{cases} u, & 0 < u < 1; \\ 2 - u, & 1 < u < 2. \end{cases}$$

Next let $U = \max(X, Y)$. The range of U is (0, 1). For 0 < u < 1,

$$P[\max(X,Y) \le u] = P[X \le u, Y \le u]$$
$$= P[X \le u]P[Y \le u]$$
$$= u^{2},$$

 \mathbf{SO}

(2.14)
$$f_U(u) = 2uI_{(0,1)}(u).$$

Finally, let $V = \min(X, Y)$. Again the range of V is [0, 1], and for 0 < v < 1,

$$P[\min(X, Y) \le v] = 1 - P[\min(X, Y) > v]$$

= 1 - P[X > v, Y > v]
= 1 - P[X > v]P[Y > v]
= 1 - (1 - v)^2,

 \mathbf{SO}

(2.15)
$$f_V(v) = 2(1-v)I_{(0,1)}(v).$$

Exercise 2.3*. Let X, Y, Z be independent, identically distributed (i.i.d.) Uniform(0, 1) rvs. Find the pdf of $U \equiv X + Y + Z$. [What is range(U)?]

 \Box

Example 2.6. Let X, Y be i.i.d. Exponential(1) rvs and set U = X + Y. Then for $0 < u < \infty$,

$$P[X + Y \le u] = \iint_{x+y \le u} e^{-x-y} dx dy$$

= $\int_0^u e^{-x} \Big[\int_0^{u-x} e^{-y} dy \Big] dx$
= $\int_0^u e^{-x} \Big[1 - e^{-(u-x)} \Big] dx$
= $\int_0^u e^{-x} dx - e^{-u} \int_0^u dx$
= $1 - e^{-u} - u e^{-u}$,

so by (2.12),

(2.16)
$$f_U(u) = \frac{d}{du} [1 - e^{-u} - ue^{-u}] = ue^{-u}.$$

Next let $V = \min(X, Y)$. Then for $0 < v < \infty$,
 $P[\min(X, Y) \le v] = 1 - P[\min(X, Y) > v]$
 $= 1 - P[X > v, Y > v]$
 $= 1 - \left[\int_v^\infty e^{-x} dx\right] \left[\int_v^\infty e^{-y} dy\right]$
 $= 1 - e^{-2v},$

 \mathbf{SO}

(2.17)
$$f_V(v) = 2e^{-2v},$$

that is, $V \equiv \min(X, Y) \sim \text{Exponential}(2)$.

More generally: If X_1, \ldots, X_n are i.i.d. Exponential(λ) rvs, then [verify!]

(2.18)
$$\min(X_1, \ldots, X_n) \sim \operatorname{Exponential}(n\lambda).$$

However: if $T = \max(X, Y)$, then T is not an exponential rv [verify!]:

(2.19)
$$f_T(t) = 2(e^{-t} - e^{-2t}).$$

Now let $Z = |X - Y| \equiv \max(X, Y) - \min(X, Y)$. The range of Z is $(0, \infty)$. For $0 < z < \infty$,

$$\begin{split} P[|X - Y| &\leq z] \\ = 1 - P[Y \geq X + z] - P[Y \leq X - z] \\ = 1 - 2P[Y \geq X + z] \text{ [by symmetry]} \\ = 1 - 2\int_0^\infty e^{-x} \Big[\int_{x+z}^\infty e^{-y} dy\Big] dx \\ = 1 - 2\int_0^\infty e^{-x} e^{-(x+z)} dx \\ = 1 - 2e^{-z}\int_0^\infty e^{-2x} dx \\ = 1 - e^{-z}, \end{split}$$

 \mathbf{SO}

(2.20)
$$f_Z(z) = e^{-z}.$$

That is, $Z \equiv \max(X, Y) - \min(X, Y) \sim \text{Exponential}(1)$, the same as X and Y themselves.

Note: This is another "memory-free" property of the exponential distribution. It is stronger in that it involves a random starting time, namely $\min(X, Y)$.

Finally, let $W = \frac{X}{X+Y}$. The range of W is (0,1). For 0 < w < 1,

$$P\left[\frac{X}{X+Y} \le w\right]$$
$$=P[X \le w(X+Y)]$$
$$=P\left[Y \ge \left(\frac{1-w}{w}\right)X\right]$$
$$=\int_{0}^{\infty} \left[\int_{\left(\frac{1-w}{w}\right)x}^{\infty} e^{-y}\right]e^{-x}dx$$
$$=\int_{0}^{\infty} \left[e^{-\left(\frac{1-w}{w}\right)x}e^{-x}dx\right]$$
$$=\int_{0}^{\infty} e^{-\frac{x}{w}}dx$$
$$=w,$$

 \mathbf{SO}

(2.21)
$$f_W(w) = I_{(0,1)}(w),$$

that is, $W \equiv \frac{X}{X+Y} \sim \text{Uniform}(0,1).$

Note: In Example 6.3 we shall show that $\frac{X}{X+Y} \perp (X+Y)$. Then (2.21) can be viewed as a "backward" memory-free property of the exponential distribution: given X + Y, the location of X is uniformly distributed over the interval (0, X + Y).

 \Box

Method Two: Introduce a second rv V = h(X, Y), where h is chosen cleverly so that it is relatively easy to find the joint pdf $f_{U,V}$ via the "Jacobian method", then marginalize to find f_U . (This method appears in §6.2.)

3. Expected Value of a RV: Mean, Variance, Covariance; Moment Generating Function; Normal & Poisson Approximations.

The **expected value** (expectation, mean) of a rv X is defined by

(3.1)
$$\mathbf{E}X = \sum_{x} x f_X(x),$$

[discrete case]

(3.2)
$$EX = \int x f_X(x) dx,$$
 [continuous case]

provided that the sum or integral is absolutely convergent. If not convergent, then the expectation does not exist.

The Law of Large Numbers states that if EX exists, then for i.i.d. copies X_1, X_2, \ldots , of X the sample averages $\bar{X}_n \equiv \frac{1}{n} \sum_{i=1}^n X_i$ converge to EX as $n \to \infty$. If the sum in (3.1) or integral in (3.2) equals $+\infty$ $(-\infty)$, then $\bar{X}_n \to +\infty$ $(-\infty)$. If the sum or integral has no value (i.e. has the form $\infty - \infty$), then \bar{X}_n will oscillate indefinitely.

If the probability distribution of X is thought of as a (discrete or continuous) mass distribution on \mathbb{R}^1 , then EX is just the *center of gravity* of the mass. With this interpretation, we can often use *symmetry* to find the expected value without actually calculating the sum or integral; however, absolute convergence still must be verified! [Eg. Cauchy distribution]

Example 3.1: [verify, including convergence; for Var see (3.9) and (3.10)]

$$\begin{aligned} X \sim \operatorname{Binomial}(n, p) \Rightarrow \mathrm{E}X &= np, \ \operatorname{Var}X = np(1-p); \qquad \text{[sum]} \\ X \sim \operatorname{Geometric}(p) \Rightarrow \mathrm{E}X = 1/p, \ \operatorname{Var}X = (1-p)^2/p; \qquad \text{[sum]} \\ X \sim \operatorname{Poisson}(\lambda) \Rightarrow \mathrm{E}X = \lambda, \ \operatorname{Var}X = \lambda; \qquad \text{[sum]} \\ X \sim \operatorname{Exponential}(\lambda) \Rightarrow \mathrm{E}X = 1/\lambda, \ \operatorname{Var}X = 1/\lambda^2; \qquad \text{[integrate]} \\ X \sim \operatorname{Normal} N(0, 1) \Rightarrow \mathrm{E}X = 0, \ \operatorname{Var}X = 1; \qquad \text{[symmetry, integrate]} \\ X \sim \operatorname{Cauchy} C(0, 1) \Rightarrow \mathrm{E}X \text{ and } \operatorname{Var}X \text{ do not exist}; \\ X \sim \operatorname{Gamma}(\alpha, \lambda) \Rightarrow \mathrm{E}X = \alpha/\lambda, \ \operatorname{Var}X = \alpha/\lambda^2; \qquad \text{[integrate]} \\ X \sim \operatorname{Beta}(\alpha, \beta) \Rightarrow \mathrm{E}X = \alpha/(\alpha + \beta); \qquad \text{[integrate]} \\ X \sim \operatorname{std.} \ \operatorname{Logistic} \Rightarrow \mathrm{E}X = 0 \qquad \qquad \text{[symmetry]} \\ X \sim \operatorname{Uniform}(a, b) \Rightarrow \mathrm{E}X = \frac{a+b}{2}, \ \operatorname{Var}X = \frac{(b-a)^2}{12}; \quad \text{[symm., integ.} \end{aligned}$$

The expected value E[g(X)] of a function of a rv X is defined similarly:

(3.3)
$$E[g(X)] = \sum_{x} g(x) f_X(x), \qquad [discrete case]$$

(3.4)
$$\operatorname{E}[g(X)] = \int g(x) f_X(x) \, dx,$$
 [continuous case]

In particular, the *r*th moment of X (if it exists) is defined as $E(X^r), r \in \mathbb{R}^1$.

Expectations of functions of random vectors are defined similarly. For example in the bivariate case,

(3.5)
$$\mathbb{E}[g(X,Y)] = \sum_{x} \sum_{y} g(x,y) f_{X,Y}(x,y), \qquad \text{[discrete case]}$$

(3.6)
$$\operatorname{E}[g(X,Y)] = \int \int g(x,y) f_{X,Y}(x,y) \, dx \, dy$$
, [continuous case]

Linearity: It follows from (3.5) and (3.6) that expectation is *linear*:

(3.7)
$$E[ag(X,Y) + bh(X,Y)] = aE[g(X,Y)] + bE[h(X,Y)].$$
 [verify]

Order-preserving: $X \ge 0 \Rightarrow EX \ge 0$ (and EX = 0 iff $X \equiv 0$).

 $X \ge Y \Rightarrow EX \ge EY$ (and EX = EY iff $X \equiv Y$). [Pf: EX - EY = E(X - Y)]

Linearity ($\equiv additivity$) simplifies many calculations:

Binomial mean: We can find the expected value of $X \sim \text{Binomial}(n, p)$ easily as follows: Because X is the total number of successes in n independent Bernoulli trials, i.e., trials with exactly two outcomes (H,T, or S,F, etc.), we can represent X as

$$(3.8) X = X_1 + \dots + X_n,$$

where $X_i = 1$ (or 0) if S (or F) occurs on the *i*th trial. (Recall Example 1.15.) Thus by linearity,

$$\mathbf{E}X = \mathbf{E}(X_1 + \dots + X_n) = \mathbf{E}X_1 + \dots + \mathbf{E}X_n = p + \dots + p = np.$$

Variance. The variance of X is defined to be

(3.9)
$$\operatorname{Var} X = \operatorname{E}[(X - \operatorname{E} X)^2],$$

the average of the square of the deviation of X about its mean. The standard deviation of X is

$$\operatorname{sd}(X) = \sqrt{\operatorname{Var} X}.$$

Properties.

(a) Var $X \ge 0$; equality holds iff X is degenerate (constant).

(b) location - scale : $\operatorname{Var}(aX + b) = a^2 \operatorname{Var}X;$ $\operatorname{sd}(aX + b) = |a| \cdot \operatorname{sd}(X).$

(c)
$$\operatorname{Var} X \equiv \operatorname{E}[(X - \operatorname{E} X)^2] = \operatorname{E}[X^2 - 2X \operatorname{E} X + (\operatorname{E} X)^2]$$

= $\operatorname{E}(X^2) - 2(\operatorname{E} X)(\operatorname{E} X) + (\operatorname{E} X)^2$
= $\operatorname{E}(X^2) - (\operatorname{E} X)^2.$

The standard deviation is a measure of the *spread* \equiv *dispersion* of the distribution of X about its mean value. An alternative measure of spread is E[|X - EX|]. Another measure of spread is the difference between the 75th and 25th *percentiles* of the distribution of X.

Covariance: The *covariance* between X and Y indicates the nature of the *linear* dependence (if any) between X and Y:

(3.11)
$$\operatorname{Cov}(X,Y) = \operatorname{E}[(X - \operatorname{E} X)(Y - \operatorname{E} Y)].$$
 [interpret; also see §4]

Properties of covariance:

(a)
$$\operatorname{Cov}(X, Y) = \operatorname{Cov}(Y, X).$$

(b) $\operatorname{Cov}(X, Y) = \operatorname{E}[XY - XEY - YEX + (EX)(EY)]$
 $= \operatorname{E}(XY) - 2(EX)(EY) + (EX)(EY)$
(3.12) $= \operatorname{E}(XY) - (EX)(EY).$

(c) $\operatorname{Cov}(X, X) = \operatorname{Var} X.$

(d) If X or Y is a degenerate rv (a constant), then Cov(X, Y) = 0.

(e) Bilinearity:
$$\operatorname{Cov}(aX, bY + cZ) = ab \operatorname{Cov}(X, Y) + ac \operatorname{Cov}(X, Z).$$

 $\operatorname{Cov}(aX + bY, cZ) = ac \operatorname{Cov}(X, Z) + bc \operatorname{Cov}(Y, Z).$

(f) Variance of a sum or difference:

(3.13)
$$\operatorname{Var}(X \pm Y) = \operatorname{Var}X + \operatorname{Var}Y \pm 2\operatorname{Cov}(X, Y).$$
 [verify]

(g) Product rule. If X and Y are independent it follows from (1.35), (3.5) and (3.6) that

(3.14)
$$\mathbf{E}[g(X)h(Y)] = \mathbf{E}[g(X)] \cdot \mathbf{E}[h(Y)].$$
 [verify]

Thus, by (3.12) and (3.13),

$$(3.15) X \perp Y \Rightarrow Cov(X, Y) = 0,$$

(3.16)
$$X \perp \!\!\!\perp Y \Rightarrow \operatorname{Var}(X \pm Y) = \operatorname{Var}X + \operatorname{Var}Y.$$

Exercise 3.1. Show by counterexample that the converse of (3.15) is not true. [Example 1.12 provides one counterexample: Suppose that (X, Y) is uniformly distributed over the unit disk D. Then by the symmetry of $D, (X, Y) \sim (X, -Y)$. Thus Cov(X, Y) = Cov(X, -Y) = -Cov(X, Y), so Cov(X, Y) = 0. But we have already seen that $X \not \perp Y$.]

Binomial variance: We can find the variance of $X \sim \text{Binomial}(n, p)$ easily as follows (recall Example 1.12):

(3.17)

$$\operatorname{Var} X = \operatorname{Var}(X_1 + \dots + X_n)$$
 [by (3.8)]
 $= \operatorname{Var} X_1 + \dots + \operatorname{Var} X_n$ [by (3.16)]
 $= p(1-p) + \dots + p(1-p)$ [by (3.10)]
 $= np(1-p).$

Variance of a sample average \equiv sample mean: Let X_1, \ldots, X_n be i.i.d. rvs, each with mean μ and variance $\sigma^2 < \infty$ and set $\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$. Then by (3.16),

(3.18)
$$E(\bar{X}_n) = \mu$$
, $Var(\bar{X}_n) = \frac{Var(X_1 + \dots + X_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$

3.1. The Weak Law of Large Numbers (WLLN).

Let X_1, \ldots, X_n be i.i.d. rvs, each with mean μ and variance $\sigma^2 < \infty$. Then \bar{X}_n converges to μ in probability $(X_n \xrightarrow{p} \mu)$, that is, for each $\epsilon > 0$,

$$P[|\bar{X}_n - \mu| \le \epsilon] \to 1 \text{ as } n \to \infty.$$

Proof. By Chebyshev's Inequality (below) and (3.18),

$$P[|\bar{X}_n - \mu| \ge \epsilon] \le \frac{\operatorname{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0 \quad \text{as } n \to \infty.$$

Lemma 3.1. Chebyshev's Inequality. Let $EY = \nu$, $VarY = \tau^2$. Then

(3.19)
$$P[|Y - \nu| \ge \epsilon] \le \frac{\tau^2}{\epsilon^2}.$$

Proof. Let $X = Y - \nu$, so E(X) = 0. Assume that X is continuous with pdf f. (The discrete case is similar, with sums replacing integrals.) Then

$$\begin{split} \tau^2 &\equiv \mathcal{E}(X^2) = \int_{|x| \ge \epsilon} x^2 f(x) dx + \int_{|x| < \epsilon} x^2 f(x) dxy \\ &\geq \int_{|x| \ge \epsilon} \epsilon^2 f(x) dx = \epsilon^2 P[|X| \ge \epsilon]. \end{split}$$

Example 3.2. Sampling without replacement - the hypergeometric distribution.

Suppose an urn contains r red balls and w white balls. Draw n balls at random from the urn and let X denote the number of red balls obtained. If the balls are sampled with replacement, then clearly $X \sim \text{Binomial}(n, p)$, where p = r/(r+w), so EX = np, VarX = np(1-p).

Suppose, however, that the balls are sampled *without* replacement. Note that we now require that $n \leq r + w$. The probability distribution of X is described as follows: its range is $\max(0, n - w) \leq x \leq \min(r, n)$ [why?], and its pmf is given by

(3.20)
$$P[X = x] = \frac{\binom{r}{x}\binom{w}{n-x}}{\binom{r+w}{n}}, \quad \max(0, n-w) \le x \le \min(r, n).$$

[Verify the range and verify the pmf. This probability distribution is called *hypergeometric* because these ratios of binomial coefficients occurs as the coefficients in the expansion of hypergeometric functions such as Bessel functions. The name is unfortunate because it has no relationship to the probability model that gives rise to (3.20).]

To determine EX and VarX, rather than combining (3.1) and (3.20), it is easier again to use the representation

$$X = X_1 + \dots + X_n,$$

where, as in (3.8), $X_i = 1$ (or 0) if a red (or white) ball is obtained on the *i*th trial. Unlike (3.13), however, clearly X_1, \ldots, X_n are *not mutually independent*. [Why?] Nonetheless, the joint distribution of (X_1, \ldots, X_n) is *exchangeable* \equiv *symmetric* \equiv *permutation-invariant*, that is

$$(X_1,\ldots,X_n)\sim(X_{i_1},\ldots,X_{i_n})$$

for every permutation (i_1, \ldots, i_n) of $(1, \ldots, n)$. This is intuitively evident but we will not prove it. As supporting evidence, however, note that

$$P[X_{2} = 1] = P[X_{2} = 1 | X_{1} = 1] P[X_{1} = 1] + P[X_{2} = 1 | X_{1} = 0] P[X_{1} = 0]$$

$$= \frac{r-1}{r+w-1} \cdot \frac{r}{r+w} + \frac{r}{r+w-1} \cdot \frac{w}{r+w}$$

(3.21)
$$= \frac{r}{r+w} \equiv P[X_{1} = 1],$$

so $X_1 \sim X_2$. Note too that since $X_1 X_2$ has range $\{0, 1\}$,

$$E(X_1X_2) = P[X_1X_2 = 1]$$

= $P[X_1 = 1, X_2 = 1]$
= $P[X_2 = 1|X_1 = 1]P[X_1 = 1]$
= $\frac{r-1}{r+w-1}\frac{r}{r+w}$,

(3.22)

$$Cov(X_1, X_2) = E(X_1 X_2) - (EX_1)(EX_2) \\
= \frac{r-1}{r+w-1} \frac{r}{r+w} - \left[\frac{r}{r+w}\right]^2 \\
= \frac{r}{r+w} \left[\frac{r-1}{r+w-1} - \frac{r}{r+w}\right] \\
= \frac{-rw}{(r+w)^2(r+w-1)}.$$

[Thus X_1 and X_2 are negatively correlated, which is intuitively clear. [Why?] By (2.21) and exchanges bility $\mathbf{E} \mathbf{Y} = r$ for i = 1 and $\mathbf{x} = 0$

By (3.21) and exchangeability, $EX_i = \frac{r}{r+w} \equiv p$ for i = 1, ..., n, so

(3.23)
$$\mathbf{E}X = \mathbf{E}X_1 + \cdots \mathbf{E}X_n = n\left(\frac{r}{r+w}\right) \equiv np,$$

the same as for sampling with replacement.

Key question: do we expect VarX also to be the same as for sampling with replacement, namely, np(1-p)? Larger? Smaller? Answer: By (3.22) and exchangeability,

$$\operatorname{Var} X = \sum_{i=1}^{n} \operatorname{Var} X_{i} + 2 \sum_{1 \le i < j \le n} \operatorname{Cov}(X_{i}, X_{j})$$

$$= np(1-p) + n(n-1)\operatorname{Cov}(X_{1}, X_{2})$$

$$= n \frac{r}{r+w} \frac{w}{r+w} + n(n-1) \left[\frac{-rw}{(r+w)^{2}(r+w-1)} \right]$$

$$= \frac{nrw}{(r+w)^{2}} \left[1 - \frac{n-1}{r+w-1} \right]$$

$$(3.24) \qquad = np(1-p) \left[1 - \frac{n-1}{N-1} \right],$$

where $N \equiv r + w$ is the total number of balls in the urn. [Discuss $\frac{n-1}{N-1}$.]

By comparing (3.24) to (3.17), we see that sampling without replacement from a finite population reduces the variability of the outcome. This is to be expected from the representation $X = X_1 + \cdots + X_n$ and the fact that each pair (X_i, X_j) is negatively correlated (by (3.22) and exchangeability).

3.2. Correlated and conditionally correlated events.

The events A and B are *positively correlated* if any of the following three equivalent conditions hold:

$$(3.25) P[A \cap B] > P[A]P[B];$$

$$(3.26) P[A \mid B] > P[A];$$

(3.27) $P[B \mid A] > P[B].$

Note that (3.25) is equivalent to $Cov(I_A, I_B) > 0$. Negative correlation is definited similarly with > replaced by <.

It is easy to see that A and B are positively correlated iff [verify:]

(3.28) either
$$P[A \mid B] > P[A \mid B^c]$$

(3.29) or $P[B \mid A] > P[B \mid A^c].$

The events A and B are *conditionally positively correlated given* C if any of the following three equivalent conditions hold:

[verify:]

- (3.30) $P[A \cap B \mid C] > P[A \mid C]P[B \mid C];$
- (3.31) $P[A \mid B, C] > P[A \mid C];$
- (3.32) $P[B \mid A, C] > P[B \mid C].$

Then A and B are positively correlated given C iff

- (3.34) or $P[B \mid A, C] > P[B \mid A^c, C].$

Example 3.3: Simpson's paradox.

(3.35)
$$\left\{ \begin{array}{l} P[A \mid B, \ C] > P[A \mid B^c, \ C] \\ P[A \mid B, \ C^c] > P[A \mid B^c, \ C^c] \end{array} \right\} \not\Rightarrow P[A \mid B] > P[A \mid B^c] !$$

To see this, consider the famous Berkeley Graduate School Admissions data. To simplify, assume there are only two graduate depts, Physics and English. Let

 $A = \{ \text{Applicant is Accepted to Berkeley Grad School} \}$ $B = \{ \text{Applicant is Female} \} \equiv F$ $B^c = \{ \text{Applicant is Male} \} \equiv M$ $C = \{ \text{Applied to Physics Department} \} \equiv Ph$ $C^c = \{ \text{Applied to English Department} \} \equiv En.$

Physics: 100 F Applicants, 60 Accepted: $P[A \mid F, Ph] = 0.6$.

Physics: 100 M Applicants, 50 Accepted: $P[A \mid M, Ph] = 0.5$.

English: 1000 F Applicants, 250 Accepted: $P[A \mid F, En] = 0.25$.

English: 100 M Applicants, 20 Accepted: $P[A \mid M, En] = 0.2$.

Total: 1100 F Applicants, 310 Accepted: $P[A \mid F] = 0.28$. Total: 200 M Applicants, 70 Accepted: $P[A \mid M] = 0.35$.

Therefore: P[A | F] < P[A | M]. Is this evidence of discrimination?

No:
$$\begin{aligned} &P[A \mid F, \ Ph] > P[A \mid M, \ Ph], \\ &P[A \mid F, \ En] > PA \mid M, \ En], \end{aligned}$$

so F's are more likely to be accepted into each dept Ph and En separately! *Explanation:* Most F's applied to English, where the overall acceptance rate is low:

$$P[A | F] = P[A | F, Ph]P[Ph | F] + P[A | F, En]P[En | F],$$

$$P[A | M] = P[A | M, Ph]P[Ph | M] + P[A | M, En]P[En | M].$$

$$\Box$$

Exercise 3.2. Show that the implication (3.35) does hold if $B \perp \!\!\!\perp C$. \Box

3.3. Moment generating functions: they uniquely determine moments, distributions, and convergence of distributions.

The moment generating function (mgf) of (the distribution of) the rv X is:

(3.36)
$$m_X(t) = \mathbf{E}(e^{tX}), \qquad -\infty < t < \infty.$$

Clearly $m_X(0) = 1$ and $0 < m_X(t) \le \infty$, with ∞ possible. If $m_X(t) < \infty$ for $|t| < \delta$ then the Taylor series expansion of e^{tX} yields

(3.37)
$$m_X(t) = \mathbf{E}\Big[\sum_{k=0}^{\infty} \frac{(tX)^k}{k!}\Big] = \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathbf{E}(X^k), \quad |t| < \delta,$$

a power series whose coefficients are the kth moments of X. In this case the moments of X are recovered from the mgf m_X by differentiation at t = 0:

(3.38)
$$E(X^k) = m_X^{(k)}(0), \quad k = 1, 2, \dots$$
 [verify]

Location-scale:

(3.39)
$$m_{aX+b}(t) = \mathrm{E}(e^{t(aX+b)}) = e^{bt}\mathrm{E}(e^{atX}) = e^{bt}m_X(at).$$

Multiplicativity: X, Y independent \Rightarrow

(3.40)
$$m_{X+Y}(t) = \mathrm{E}(e^{t(X+Y)}) = \mathrm{E}(e^{tX})\mathrm{E}(e^{tY}) = m_X(t)m_Y(t).$$

In particular, if X_1, \ldots, X_n are i.i.d. then

(3.41)
$$m_{X_1 + \dots + X_n}(t) = [m_{X_1}(t)]^n.$$

Example 3.4.

Bernoulli(p): Let $X = \begin{cases} 1, & \text{with probability } p; \\ 0, & \text{with probability 1-p.} \end{cases}$ Then

(3.42)
$$m_X(t) = pe^t + (1-p).$$

Binomial (n, p): We can represent $X = X_1 + \cdots + X_n$, where X_1, \ldots, X_n are i.i.d. Bernoulli(p) 0-1 rvs. Then by (3.41) and (3.42),

(3.43)
$$m_X(t) = \left[pe^t + (1-p)\right]^n.$$

 $Poisson(\lambda)$:

(

(3.44)
$$m_X(t) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{tk} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}.$$

Standard univariate normal N(0, 1):

(3.45)
$$m_X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-\frac{x^2}{2}} dx$$
$$= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-t)^2}{2}} dx$$
$$= e^{\frac{t^2}{2}}.$$

General univariate normal $N(\mu, \sigma^2)$: We can represent $X = \sigma Z + \mu$ where $Z \sim N(0, 1)$. Then by (3.39) (location-scale) and (3.45),

(3.46)
$$m_X(t) = e^{\mu t} m_Z(\sigma t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}.$$

 $Gamma(\alpha, \lambda)$ (includes Exponential(λ)):

$$m_X(t) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \int_0^{\infty} e^{tx} \cdot x^{\alpha-1} e^{-\lambda x} dx$$
$$= \frac{\lambda^{\alpha}}{(\lambda - t)^{\alpha}} \cdot \frac{(\lambda - t)^{\alpha}}{\Gamma(\alpha)} \int_0^{\infty} x^{\alpha-1} e^{-(\lambda - t)x} dx$$
$$= \frac{\lambda^{\alpha}}{(\lambda - t)^{\alpha}}, \quad -\infty < t < \lambda.$$

Uniqueness: The moment generating function $m_X(t)$ (if finite for some interval $|t| < \delta$) uniquely determines the distribution of X.

That is, suppose that $m_X(t) = m_Y(t) < \infty$ for $|t| < \delta$. Then $X \stackrel{\text{distn}}{=} Y$, i.e., $P[X \in A] = P[Y \in A]$ for all events A. [See §3.3.1]

Application 3.3.1: The sum of independent Poisson rvs is Poisson. Suppose that X_1, \ldots, X_n are independent, $X_i \sim \text{Poisson}(\lambda_i)$. Then by (3.40), (3.44),

$$m_{X_1+\dots+X_n}(t) = e^{\lambda_1(e^t-1)} \cdots e^{\lambda_n(e^t-1)} = e^{(\lambda_1+\dots+\lambda_n)(e^t-1)},$$

 \Box

so $X_1 + \cdot + X_n \sim \text{Poisson}(\lambda_1 + \cdot \cdot + \lambda_n)$.

37

Application 3.3.2: The sum of independent normal rvs is normal. Suppose X_1, \ldots, X_n are independent, $X_i \sim N(\mu_i, \sigma_i^2)$. Then by (3.40) and (3.46),

$$m_{X_1+\dots+X_n}(t) = e^{\mu_1 t + \frac{\sigma_1^2 t^2}{2}} \cdots e^{\mu_n t + \frac{\sigma_n^2 t^2}{2}} = e^{(\mu_1 + \dots + \mu_n)t + \frac{(\sigma_1^2 + \dots + \sigma_n^2)t^2}{2}},$$

so $X_1 + \cdot + X_n \sim N(\mu_1 + \cdots + \mu_n, \sigma_1^2 + \cdots + \sigma_n^2).$

Application 3.3.3: The sum of independent Gamma rvs with the same scale parameter is Gamma. Suppose that X_1, \ldots, X_n are independent rvs with $X_i \sim G(\alpha_i, \lambda)$. Then by (3.40) and (3.47),

$$m_{X_1+\dots+X_n}(t) = \frac{\lambda^{\alpha_1}}{(\lambda-t)^{\alpha_1}} \cdots \frac{\lambda^{\alpha_n}}{(\lambda-t)^{\alpha_n}} = \frac{\lambda^{\alpha_1+\dots+\alpha_n}}{(\lambda-t)^{\alpha_1+\dots+\alpha_n}}, \ -\infty < t < \lambda,$$

so $X_1 + \cdot + X_n \sim \text{Gamma}(\alpha_1 + \cdots + \alpha_n, \lambda).$ \Box

Convergence in distribution: A sequence of rvs $\{X_n\}$ converges in distribution to X, denoted as $X_n \xrightarrow{d} X$, if $P[X_n \in A] \to P[X \in A]$ for every³ event A. Then if $m_X(t) < \infty$ for $|t| < \delta$, we have

(3.48)
$$X_n \xrightarrow{d} X \iff m_{X_n}(t) \to m_X(t) \quad \forall |t| < \delta.$$
 [See §3.4.1]

Application 3.3.4: The normal approximation to the binomial distribution $(\equiv the Central Limit Theorem for Bernoulli rvs).$

Let $S_n \sim \text{Binomial}(n, p)$, that is, S_n represents the total number of successes in n independent trials with P[Success] = p, 0 . (This is called asequence of*Bernoulli trials.* $) Since <math>E(S_n) = np$ and $Var(S_n) = np(1-p)$, the standardized version of S_n is

$$Y_n \equiv \frac{S_n - np}{\sqrt{np(1-p)}},$$

so $E(Y_n) = 0$, $Var(Y_n) = 1$. We apply (3.48) to show that $Y_n \xrightarrow{d} N(0, 1)$, or equivalently, that

(3.49)
$$X_n \equiv \sqrt{p(1-p)} Y_n \xrightarrow{d} N(0, p(1-p)):$$

³ Actually A must be restricted to be such that $P[X \in \partial A] = 0$; see §10.2.

$$\begin{split} m_{X_n}(t) &= \mathbf{E} \left[e^{\frac{t(S_n - np)}{\sqrt{n}}} \right] \\ &= e^{-tp\sqrt{n}} \mathbf{E} \left[e^{\frac{tS_n}{\sqrt{n}}} \right] \\ &= e^{-tp\sqrt{n}} \left[pe^{\frac{t}{\sqrt{n}}} + (1 - p) \right]^n \qquad \text{[by (3.43)]} \\ &= \left[pe^{\frac{t(1 - p)}{\sqrt{n}}} + (1 - p)e^{-\frac{tp}{\sqrt{n}}} \right]^n \\ &= \left[p \left(1 + \frac{t(1 - p)}{\sqrt{n}} + \frac{t^2(1 - p)^2}{2n} + O\left(\frac{1}{n^{3/2}}\right) \right) \\ &+ (1 - p) \left(1 - \frac{tp}{\sqrt{n}} + \frac{t^2p^2}{2n} + O\left(\frac{1}{n^{3/2}}\right) \right) \right]^n \\ &= \left[1 + \frac{t^2p(1 - p)}{2n} + O\left(\frac{1}{n^{3/2}}\right) \right]^n \\ &\to e^{\frac{t^2p(1 - p)}{2}} \qquad \text{[by CB Lemma 2.3.14.]} \end{split}$$

Since $e^{\frac{t^2p(1-p)}{2}}$ is the mgf of N(0, p(1-p)), (3.49) follows from (3.48). \Box

Application 3.3.5: The Poisson approximation to the binomial distribution for "rare" events.

Let $X_n \sim \text{Binomial}(n, p_n)$, where $p_n = \frac{\lambda}{n}$ for some $\lambda \in (0, \infty)$. Thus $E(X_n) \equiv np_n = \lambda$ remains fixed while $P[\text{Success}] \equiv p_n \to 0$, so "Success" becomes a rare event as $n \to \infty$. From (3.43),

$$m_{X_n}(t) = \left[\left(\frac{\lambda}{n}\right) e^t + \left(1 - \frac{\lambda}{n}\right) \right]^n$$
$$= \left[1 + \left(\frac{\lambda}{n}\right) \left(e^t - 1\right) \right]^n$$
$$\to e^{\lambda(e^t - 1)},$$

so by (3.44) and (3.48), $X_n \xrightarrow{d} \text{Poisson}(\lambda)$ as $n \to \infty$.

Note: This also can be proved directly: for k = 0, 1, ...,

$$P[X_n = k] = {\binom{n}{k}} {\left(\frac{\lambda}{n}\right)^k} {\left(1 - \frac{\lambda}{n}\right)^{n-k}} = \frac{n(n-1)\cdots(n-k+1)}{n^k} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^{n-k} \to \frac{\lambda^k}{k!} e^{-\lambda} \text{ as } n \to \infty. \quad \Box$$

3.3.1. Proofs of the uniqueness and convergence properties of mgfs for discrete distributions with finite support.

Consider rvs X, Y, and $\{X_n\}$ with a common, finite support $\{1, 2, \ldots, s\}$. The probability distributions of X, Y, and X_n are given by the vectors

$$\mathbf{p}_X \equiv \begin{pmatrix} p_1 \\ \vdots \\ p_s \end{pmatrix}, \quad \mathbf{p}_Y \equiv \begin{pmatrix} q_1 \\ \vdots \\ q_s \end{pmatrix} \quad \mathbf{p}_{X_n} \equiv \begin{pmatrix} p_{n,1} \\ \vdots \\ p_{n,s} \end{pmatrix},$$

respectively, where $p_j = P[X = j]$, $q_j = P[Y = j]$, $p_{n,j} = P[X_n = j]$, $j = 1, \ldots, s$. Choose any s distinct points $t_1 < \cdots < t_s$ and let

$$\mathbf{m}_X = \begin{pmatrix} m_X(t_1) \\ \vdots \\ m_X(t_s) \end{pmatrix}, \quad \mathbf{m}_Y = \begin{pmatrix} m_Y(t_1) \\ \vdots \\ m_Y(t_s) \end{pmatrix}, \quad \mathbf{m}_{X_n} = \begin{pmatrix} m_{X_n}(t_1) \\ \vdots \\ m_{X_n}(t_s) \end{pmatrix}.$$

Since

$$m_X(t_i) = \sum_{j=1}^s e^{t_i j} p_j \equiv (e^{t_i}, e^{2t_i} \dots, e^{st_i}) \mathbf{p}_X,$$

etc., we can write [verify!]

(3.50)
$$\mathbf{m}_X = A \mathbf{p}_X, \quad \mathbf{m}_Y = A \mathbf{p}_Y, \quad \mathbf{m}_{X_n} = A \mathbf{p}_{X_n},$$

where A is the $s \times s$ matrix given by

$$A = \begin{pmatrix} e^{t_1} & e^{2t_1} & \dots & e^{st_1} \\ \vdots & \vdots & & \vdots \\ e^{t_s} & e^{2t_s} & \dots & e^{st_s} \end{pmatrix}.$$

 \Box

Exercise 3.3*. Show that A is a nonsingular matrix, so A^{-1} exists. *Hint:* show that $det(A) \neq 0$.

Thus

(3.51)
$$\mathbf{p}_X = A^{-1}\mathbf{m}_X \text{ and } \mathbf{p}_Y = A^{-1}\mathbf{m}_Y,$$

so if $m_X(t_i) = m_Y(t_i) \ \forall i = 1, \dots, s$ then $\mathbf{m}_X = \mathbf{m}_Y$, hence $\mathbf{p}_X = \mathbf{p}_Y$ by (3.51), which establishes the uniqueness property of mgfs in this special case. Also, if $m_{X_n}(t_i) \to m_X(t_i) \forall$ then $\mathbf{m}_{X_n} \to \mathbf{m}_X$, hence $\mathbf{p}_{X_n} \to \mathbf{p}_X$ by (3.51), which established the convergence property of mgfs in this case.

Remark 3.1. (3.50) simply shows that the mgf is a *nonsingular linear transform* of the probability distribution. In engineering, the mgf would be called the *Laplace transform* of the probability distribution. An alternative transformation is the Fourier transform defined by $\phi_X(t) = E(e^{itX})$, where $i = \sqrt{-1}$, which we call the *characteristic function* of (the distribution of) X. ϕ_X is complex-valued, but has the advantage that it is always finite. In fact, since $|e^{iu}| = 1$ for all real $u, |\phi_X(t)| \leq 1$ for all real t. Ū

3.4. Multivariate moment generating functions.

Let
$$X \equiv \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$$
 and $t \equiv \begin{pmatrix} t_1 \\ \vdots \\ t_p \end{pmatrix}$. The moment generating function (mgf) of (the distribution of) the rvtr X is

ヮ (*י*ן/

(3.52)
$$m_X(t) = \mathbb{E}\left(e^{t'X}\right) \equiv \mathbb{E}\left(e^{t_1X_1 + \dots + t_pX_p}\right)$$

Again, $m_X(0) = 1$ and $0 < m_X(t) \le \infty$, with ∞ possible. Note that if X_1, \ldots, X_p are independent, then

(3.53)

$$m_X(t) \equiv \mathbf{E} \left(e^{t_1 X_1 + \dots + t_p X_p} \right)$$

$$= \mathbf{E} \left(e^{t_1 X_1} \right) \cdots \mathbf{E} \left(e^{t_p X_p} \right)$$

$$\equiv m_{X_1}(t_1) \cdots m_{X_p}(t_p).$$

All properties of the mgf, including uniqueness, convergence in distribution, the location-scale formula, and multiplicativity, extend to the multivariate case. For example:

If $m_X(t) < \infty$ for $||t|| < \delta$ then the multiple Taylor series expansion of $e^{t'X}$ vields

(3.54)
$$m_X(t) = \sum_{k=0}^{\infty} \sum_{k_1 + \dots + k_p = k} \frac{t_1^{k_1} \cdots t_p^{k_p} \mathcal{E}(X_1^{k_1} \cdots X_p^{k_p})}{k_1! \cdots k_p!}, \quad ||t|| < \delta,$$

 \mathbf{SO}

(3.55)
$$E(X_1^{k_1}\cdots X_p^{k_p}) = m_X^{(k_1,\dots,k_p)}(0), \quad k_1,\dots,k_p \ge 0.$$
 [verify]

Multivariate location-scale: For any fixed $q \times p$ matrix A and $q \times 1$ vector b,

(3.56)
$$m_{AX+b}(t) = \mathbb{E}\left(e^{t'(AX+b)}\right) = e^{t'b}\mathbb{E}\left(e^{t'AX}\right) = e^{t'b}m_X(A't).$$

Example 3.5. The multivariate normal distribution $N_p(\mu, \Sigma)$.

First suppose that Z_1, \ldots, Z_q are i.i.d. standard normal N(0, 1) rvs. Then by (3.45) and (3.53) the rvtr $Z \equiv (Z_1, \ldots, Z_q)'$ has mgf

(3.57)
$$m_Z(t) = e^{t_1^2/2} \cdots e^{t_q^2/2} = e^{t't/2}.$$

Now let $X = AZ + \mu$, with $A : p \times q$ and $\mu : p \times 1$. Then by (3.56) and (3.57),

(3.58)

$$m_X(t) = e^{t'\mu} m_Z(A't)$$

$$= e^{t'\mu} e^{(A't)'(A't)/2}$$

$$= e^{t'\mu} e^{t'(AA')t/2}$$

$$\equiv e^{t'\mu+t'\Sigma t/2},$$

where $\Sigma = AA'$. We shall see in §8.3 that $\Sigma = \text{Cov}(X)$, the covariance matrix of X. Thus the distribution of $X \equiv AZ + \mu$ depends on (μ, A) only through (μ, Σ) , so we denote this distribution by $N_p(\mu, \Sigma)$, the pdimensional multivariate normal distribution (MVND) with mean vector μ and covariance matrix Σ . We shall derive its pdf in §8.3. However, we can use the representation $X = AZ + \mu$ to derive its basic linearity property:

Linearity of $N_p(\mu, \Sigma)$: If $X \sim N_p(\mu, \Sigma)$ then for $C: r \times p$ and $d: r \times 1$,

(3.59)

$$Y \equiv CX + d = (CA)Z + (C\mu + d)$$

$$\sim N_r (C\mu + d, (CA)(CA)')$$

$$= N_r (C\mu + d, C\Sigma C').$$

In particular, if r = 1 then for $c : p \times 1$ and $d : 1 \times 1$,

(3.60)
$$c'X + d \sim N_1(c'\mu + d, \ c'\Sigma c).$$

3.5. The Central Limit Theorem (CLT) \equiv normal approximation.

The normal approximation to the binomial distribution (see Application 3.3.4) can be viewed as an approximation to the distribution of the sum of i.i.d. Bernoulli (0-1) rvs. This extends to any sum of i.i.d. rvs with finite second moments. We will state the result here and defer the proof to (???).

Theorem 3.2. Let $\{Y_n\}$ be a sequence of i.i.d. rvs with finite mean μ and variance σ^2 . Set $S_n = Y_1 + \cdots + Y_n$ and $\overline{Y}_n = \frac{S_n}{n}$. Their standardized distributions converge to the standard normal N(0, 1) distribution: for any a < b,

(3.61)
$$P\left[a \le \frac{S_n - n\mu}{\sqrt{n\sigma}} \le b\right] \to P[a \le N(0,1) \le b] \equiv \Phi(b) - \Phi(a),$$

$$(3.62)P\left[a \le \frac{\sqrt{n}\left(Y_n - \mu\right)}{\sigma} \le b\right] \to P\left[a \le N(0, 1) \le b\right] \equiv \Phi(b) - \Phi(a),$$

where Φ is the cdf of N(0,1). Thus, if n is "large", for any c < d we have

(3.63)
$$P[c \le S_n \le d] = P\left[\frac{c - n\mu}{\sqrt{n\sigma}} \le \frac{S_n - n\mu}{\sqrt{n\sigma}} \le \frac{d - n\mu}{\sqrt{n\sigma}}\right] \approx \Phi\left[\frac{d - n\mu}{\sqrt{n\sigma}}\right] - \Phi\left[\frac{c - n\mu}{\sqrt{n\sigma}}\right].$$

Continuity correction: Suppose that $\{Y_n\}$ are integer-valued, hence so is S_n . Then if c, d are integers, the accuracy of (3.63) can be improved significantly as follows:

(3.64)
$$P[c \le S_n \le d] = P[c - 0.5 \le S_n \le d + 0.5]$$
$$\approx \Phi\left[\frac{d + 0.5 - n\mu}{\sqrt{n\sigma}}\right] - \Phi\left[\frac{c - 0.5 - n\mu}{\sqrt{n\sigma}}\right].$$

3.6. The Poisson process.

We shall construct the *Poisson process* (PP) as a limit of *Bernoulli processes*. First, we restate the Poisson approximation to the binomial distribution (recall Application 3.3.5):

Lemma 3.2. Let $Y^{(n)} \sim \text{Binomial}(n, p_n)$. Assume that $n \to \infty$ and $p_n \to 0$ s.t. $E(Y^{(n)}) \equiv np_n = \lambda > 0$. Then $Y^{(n)} \stackrel{d}{\to} \text{Poisson}(\lambda)$ as $n \to \infty$.

[Note that the range of $Y^{(n)}$ is $\{0, 1, \ldots, n\}$, which converges to the Poisson range $\{0, 1, \ldots\}$ as $n \to \infty$.]

This result says that if a very large number (n) of elves toss identical coins independently, each with a very small success probability p_n , so that the expected number of successes $np_n = \lambda$, then the total number of successes approximately follows a Poisson distribution. Suppose now that these *n* elves are spread uniformly over the unit interval [0, 1), and that *n* more elves with identical coins are spread uniformly over the interval [1, 2), and *n* more spread over [2, 3), and so on:

For $0 < t < \infty$, let $Y_t^{(n)}$ denote the total number of successes occurring in the interval [0, t]. Considered as a function of t,

(3.65)
$$\{Y_t^{(n)} \mid 0 \le t < \infty\}$$

is a stochastic process \equiv random function. Because it is constructed from Bernoulli (0-1) rvs, it is called a *Bernoulli process*. A typical realization \equiv sample path looks like:

Thus the random function $\{Y_t^{(n)} \mid 0 \le t < \infty\}$ is a nonnegative nondecreasing step function with jump size 1. Here, each jump occurs at a rational

point $t = \frac{r}{n}$ corresponding to the location point of a lucky elf who achieves a success. Furthermore, because all the elves are probabilistically independent, for any $0 < t_0 < t_1 < t_2 \cdots$, the successive increments $Y_{t_1}^{(n)} - Y_{t_0}^{(n)}$, $Y_{t_2}^{(n)} - Y_{t_1}^{(n)}$, ..., are mutually independent. Thus the Bernoulli process has independent increments. Furthermore, for each increment

(3.66)
$$Y_{t_i}^{(n)} - Y_{t_{i-1}}^{(n)} \sim \text{Binomial}((t_i - t_{i-1})n, p_n).$$

In particular,

(3.67)
$$E[Y_{t_i}^{(n)} - Y_{t_{i-1}}^{(n)}] = (t_i - t_{i-1})np_n \equiv (t_i - t_{i-1})\lambda,$$

which does not depend on n. In fact, here it depends only on λ and the length of the interval (t_{i-1}, t_i) .

Now let $n \to \infty$ and let

$$(3.68) \qquad \qquad \{Y_t \mid 0 \le t < \infty\}$$

be the limiting stochastic process. Since independence is preserved under limits in distribution [cf. §10], for any $0 < t_0 < t_1 < t_2 \cdots$ the increments $Y_{t_1} - Y_{t_0}, Y_{t_2} - Y_{t_1}$, ...of the limiting process remain mutually independent, and by Lemma 3.2 each increment

(3.69)
$$Y_{t_i} - Y_{t_{i-1}} \sim \operatorname{Poisson}((t_i - t_{i-1})\lambda).$$

The process (3.68) is called a *Poisson process* (PP) with *intensity* λ . Its sample paths are again nonnegative nondecreasing step functions, again with jump size 1. Such a process is called a *point process* because the sample path is completely determined by the locations of the jump points T_1, T_2, \ldots , and these jump points can occur anywhere on $(0, \infty)$ (not just at rational points).

Note that (3.67) continues to hold for $\{Y_t\}$:

(3.70)
$$E[Y(t_i) - Y(t_{i-1})] = (t_i - t_{i-1})\lambda,$$

which again depends only on λ and the length of the interval (t_{i-1}, t_i) .

[Such a process is called *homogeneous*. Non-homogeneous Poisson processes also can be defined, with λ replaced by an *intensity function* $\lambda(t) \geq 0$. A non-homogeneous PP retains all the above properties of a homogeneous PP except that in (3.69) and (3.70), $(t_i - t_{i-1})\lambda$ is replaced by $\int_{t_{i-1}}^{t_i} \lambda(t) dt$. A non-homogeneous PP can be thought of as the limit of non-homogeneous Bernoulli processes, where the success probabilities vary (continuously) along the sequence of elves.]

There is a duality between a (homogeneous) PP $\{Y_t \mid 0 \leq t < \infty\}$ and sums of i.i.d. exponential random variables. This is (partially) seen as follows. Let T_1, T_2, \ldots be the times of the jumps of the PP.⁴

Proposition 3.1. T_1 , $T_2 - T_1$, $T_3 - T_2$,... are *i.i.d.* Exponential (λ) rvs. In particular, $E(T_i - T_{i-1}) = \frac{1}{\lambda}$, which reflects the intuitive fact that the expected waiting time to the next success is inversely proportional to the intensity rate λ .

Partial Proof. $P[T_1 > t] = P[\text{no successes occur in } (0, t]] = P[Y_t = 0] = e^{-\lambda t}$, since $Y_t \sim \text{Poisson}(\lambda t)$. (The proof continues with Exercise 6.4.) \Box

Note: $T_k \sim \text{Gamma}(k, \lambda)$, because the sum of i.i.d. exponential rvs has a Gamma distribution. Since $\{T_k > t\} = \{Y_t \le k - 1\}$, this implies a relation between the cdfs of the Gamma and Poisson distributions: see CB Example 3.3.1 and CB Exercise 3.19; also Ross Exercise 26, Ch. 4.]

In view of Proposition 3.1 and the fact that a PP is completely determined by the location of the jumps, a PP can be constructed from a sequence of i.i.d. Exponential (λ) rvs V_1, V_2, \ldots : just define $T_1 = V_1$,

⁴ There must be infinitely many jumps in $(0, \infty)$. This follows from the Borel-Cantelli Theorem, which says that if $\{A_n\}$ is a sequence of *independent* events, then $P[\text{infinitely many } A_n \text{ occur}]$ is 0 or 1 according to whether $\sum P(A_n)$ is $< \infty$ or $= \infty$. Now let A_n be the event that at least one success occurs in the interval (n-1, n].

 $T_2 = V_1 + V_2$, $T_3 = V_1 + V_2 + X_3$, etc. Then T_1, T_2, T_3, \ldots determine the jump points of the PP, from which the entire sample path can be constructed.

PPs arise in many applications, for example as a model for radioactive decay over time. Here, an "elf" is an individual atom – each atom has a tiny probability p of decaying (a "success") in unit time, but there are a large number n of atoms. PPs also serve as models for the number of traffic accidents over time (or location) on a busy freeway.

PPs can be extended in several ways: from homogeneous to nonhomogeneous as mentioned above, and/or to processes indexed by more than one parameter, e.g., t_1, \ldots, t_k , so the process is written as

(3.71)
$$\{Y_{t_1,\ldots,t_k} \mid t_1 \ge 0,\ldots,t_k \ge 0\}.$$

This can be thought of as a limit of Bernoulli processes where the elves are distributed in a k-dimensional space rather than along a 1-dimensional line.

The basic properties continue to hold, i.e., the number of successes occurring in non-overlapping regions are independent Poisson variates determined by an intensity function $\lambda(t_1, \ldots, t_k) \geq 0$, as follows: for any region R, the number of counts occurring in R has a Poisson distribution with parameter $\int_R \lambda(t_1, \ldots, t_k) dt_1, \ldots, dt_k$. Examples of 2-dimensional random processes that may be PPs include the spatial distribution of weeds in a field, of ore deposits in a region, or of erroneous pixels in a picture transmitted from a Mars orbiter. These are called "spatial" processes because the parameters (t_1, \ldots, t_k) determine a *location*, not because the process originates from outer space. They are only "possibly" PP because the independence property may not hold if spatial correlation is present.

The waiting-time paradox for a homogeneous Poisson Process.

Does it seem that your waiting time for a bus is usually longer than you had expected? This can be explained by the memory-free property of the exponential distribution of the waiting times.

We will model the bus arrival times as the jump times $T_1 < T_2 < \cdots$ of a homogeneous PP $\{Y_t\}$ on $[0, \infty)$ with intensity λ . Thus the interarrival times $V_i \equiv T_i - T_{i-1}$ $(i \ge 1, T_0 \equiv 0)$ are i.i.d Exponential (λ) rvs and

(3.72)
$$\operatorname{E}(V_i) = \frac{1}{\lambda}, \qquad i \ge 1.$$

Now suppose that you arrive at the bus stop at a fixed time $t^* > 0$. Let the index $j \ge 1$ be such that $T_{j-1} < t^* < T_j$ $(j \ge 1)$, so V_j is the length of the interval that contains your arrival time. We expect from (3.72) that

(3.73)
$$\operatorname{E}(V_j) = \frac{1}{\lambda}.$$

Paradoxically, however [see figure],

(3.74)
$$E(V_j) = E(T_j - t^*) + E(t^* - T_{j-1}) > E(T_j - t^*) = \frac{1}{\lambda},$$

since $T_j - t^* \sim \text{Expo}(\lambda)$ by the memory-free property of the exponential distribution: if the next bus has not arrived by time t^* then the additional waiting time to the next bus still has the $\text{Expo}(\lambda)$ distribution. Thus you appear always to be unlucky to arrive at the bus stop during a longer-than-average interarrival time!

This paradox is resolved by noting that the index j is random, not fixed: it is the random index such that V_j includes t^* . The fact that this interval includes a prespecified point t^* tends to make V_j larger than average: a larger interval is more likely to include t^* than a shorter one! Thus it is not so suprising that $E(V_j) > \frac{1}{\lambda}$.

Exercise 3.4. A simpler example of this phenomenon can be seen as follows. Let U_1 and U_2 be two random points chosen independently and uniformly on the (circumference of the) unit circle and let L_1 and L_2 be the lengths of the two arcs thus obtained:

Thus, $L_1 + L_2 = 2\pi$ and $L_1 \stackrel{\text{distn}}{=} L_2$ by symmetry, so

(3.76)
$$E(L_1) = E(L_2) = \pi.$$

(i) Find the distributions of L_1 and of L_2 .

(ii) Let L^* denote the length of the arc that contains the point $u^* \equiv (1,0)$ and let L^{**} be the length of the other arc [see figure]. Find the distributions of L^* and L^{**} . Find $E(L^*)$ and $E(L^{**})$ and show that $E(L^*) > E(L^{**})$. *Hint:* There is a simplifying geometric trick.

Remark 3.2. In (3.74), it is tempting to apply the memory-free property in reverse to assert that also $t^* - T_{j-1} \sim \text{Expo}(\lambda)$. This is actually true whenever $j \geq 2$, but not when j = 1: $t^* - T_0 \equiv t^* \not\sim \text{Expo}(\lambda)$. However this may be achieved by assuming that the bus arrival times $\ldots, T_{-2}, T_{-1}, T_0, T_1, T_2, \ldots$ follow a "doubly-infinite" homogeneous PP on the entire real line $(-\infty, \infty)$. Just as the PP on $(0, \infty)$ can be thought of in terms of many coin-tossing elves spread homogeneously over $(0, \infty)$, this PP can be thought of in terms of many coin-tossing elves spread homogeneously over $(0, \infty)$. The PP properties remain the same, in particular, the interarrival times $T_i - T_{i-1}$ are i.i.d. Exponential (λ) rvs. In this case it *is* true that $t^* - T_{j-1} \sim \text{Expo}(\lambda)$, hence we have the *exact* result that

(3.75)
$$\operatorname{E}(V_j) = \frac{2}{\lambda}.$$

(In fact, $V_j \sim \text{Expo}(\lambda) + \text{Expo}(\lambda) \stackrel{\text{distn}}{=} \text{Gamma}(2, \lambda).$) \square

4. Conditional Expectation and Conditional Distribution.

Let (X, Y) be a bivariate rvtr with joint pmf (discrete case) or joint pdf (continuous case) f(x, y). The conditional expectation of Y given X = x is defined by

(4.1)
$$\operatorname{E}[Y \mid X = x] = \sum_{y} yf(y|x),$$
 [discrete case]

(4.2)
$$E[Y | X = x] = \int yf(y|x) \, dy,$$
 [continuous case]

provided that the sum or integral exists, where f(y|x) is given by (1.28) or (1.30). More generally, for any (measurable) function g(x, y),

(4.3)
$$\operatorname{E}[g(X,Y) \mid X = x] = \sum_{y} g(x,y)f(y|x),$$
 [discrete case]

(4.4)
$$\operatorname{E}[g(X,Y) \mid X=x] = \int g(x,y)f(y|x) \, dy$$
, [continuous case]

Note that (1.29) and (1.31) are special cases of (4.3) and (4.4), respectively, with $g(x, y) = I_B(y)$. For simplicity, we often omit "=x" when writing a conditional expectation or conditional probability.

Because $f(\cdot|x)$ is a bona fide pmf or pdf, conditional expectation enjoys all the properties of ordinary expectation, in particular, *linearity*:

(4.5)
$$E[ag(X,Y) + bh(X,Y) \mid X] = aE[g(X,Y) \mid X] + bE[h(X,Y) \mid X].$$

The key *iteration formula* follows from (4.3), (1.28), (3.5) or (4.4), (1.30), (3.6):

(4.6)
$$\mathbf{E}[g(X,Y)] = \mathbf{E}\big(\mathbf{E}[g(X,Y) \mid X]\big). \qquad \text{[verify]}$$

As a special case (set $g(x, y) = I_C(x, y)$), for any (measurable) event C,

(4.7)
$$P[(X,Y) \in C] = E(P[(X,Y) \in C \mid X]).$$

We now discuss the extension of these results to the two "mixed" cases.

(i) X is discrete and Y is continuous.

First, (4.6) and(4.7) continue to hold: (4.7) follows immediately from the law of total probability (1.54) [verify], then (4.6) follows since any (measurable) g can be approximated as $g(x, y) \approx \sum c_i I_{C_i}(x, y)$. Thus, although we cannot calculate $P[(X, Y) \in C]$ or E[g(X, Y)] directly since we do not have a joint pmf or joint pdf, we can obtain them by the iteration formulas in (4.6) and (4.7). For this we can apply a formula analogous to (4.4), which requires us to determine f(y|x) as follows.

For any event B and any x s.t. P[X = x] > 0,

(4.8)
$$P[Y \in B \mid X = x] \equiv \frac{P[Y \in B, X = x]}{P[X = x]}$$

is well defined by the usual conditional probability formula (1.27). Thus we can define

(4.9)
$$f(y|x) = \frac{d}{dy}F(y|x) \equiv \frac{d}{dy}P[Y \le y \mid X = x].$$

Clearly $f(y|x) \ge 0$ and $\int f(y|x)dy = 1$ for each x s.t. P[X = x] > 0. Thus for each such x, $f(\cdot|x)$ determines a bona fide pdf. Furthermore (cf. (1.31))

(4.10)
$$P[Y \in B \mid X = x] = \int_B f(y|x) \, dy \qquad \forall B.$$

so this f(y|x) does in fact represent the conditional distribution of Y given X = x. [Using (4.9), verify (4.10) for $B = (-\infty, y]$, then extend to all B.] Now (4.10) can be extended by the approximation $g(y) \approx \sum b_i I_{B_i}(y)$ to give

(4.11)
$$\operatorname{E}[g(Y) \mid X = x] = \int g(y)f(y|x)dy$$

and similarly to give (4.4) (since E[g(X, Y) | X = x] = E[g(x, Y) | X = x].)

Note: In applications, f(y|x) is not found via (4.9) but instead is either specified directly or else is found using f(x|y) and Bayes formula (4.14) – see Example 4.3.

(ii) X is continuous and Y is either discrete or continuous.

For any (measurable) event C and any x such that f(x) > 0, we define the conditional probability

$$(4.12) P[(X,Y) \in C | X = x] = \lim_{\delta \downarrow 0} P[(X,Y) \in C | x \leq X \leq x + \delta]$$
$$= \lim_{\delta \downarrow 0} \frac{P[(X,Y) \in C, x \leq X \leq x + \delta]}{P[x \leq X \leq x + \delta]}$$
$$= \frac{1}{f(x)} \lim_{\delta \downarrow 0} \frac{P[(X,Y) \in C, x \leq X \leq x + \delta]}{\delta}$$
$$(4.13) \qquad \equiv \frac{1}{f(x)} \frac{d}{dx} P[(X,Y) \in C, X \leq x].$$

Then the iteration formulas (4.6) and (4.7) continue to hold. For (4.7),

$$E(P[(X,Y) \in C \mid X]) = \int_{-\infty}^{\infty} \left(\frac{d}{dx} P[(X,Y) \in C, X \le x]\right) dx \quad [by (4.13)]$$
$$= P[(X,Y) \in C].$$

Again (4.6) follows by the approximation $g(x, y) \approx \sum c_i I_{C_i}(x, y)$.

In particular, if X is continuous and Y is discrete, then by (4.13), f(y|x) is given by

$$f(y|x) \equiv P[Y = y \mid X = x]$$

$$= \frac{1}{f(x)} \frac{d}{dx} P[Y = y, X \le x]$$

$$= \frac{1}{f(x)} \frac{d}{dx} P[X \le x \mid Y = y] \cdot P[Y = y]$$

$$= \frac{f(x|y)f(y)}{f(x)}.$$
[by (4.9)]

This is Bayes formula for pdfs in the mixed case, and extends (1.58).

Remark 4.1. By (4.14),

(4.)

(4.15)
$$f(y|x)f(x) = f(x|y)f(y),$$

even in the mixed cases where a joint pmf or pdf f(x, y) does not exist. In such cases, the joint distribution is specified either by specifying f(y|x) and f(x), or f(x|y) and f(y) – see Example 4.3.

Remark 4.2. If (X, Y) is jointly continuous then we now have two definitions of f(y|x): the "slicing" definition (1.30): $f(y|x) = \frac{f(x,y)}{f(x)}$, and the following definition (4.16) obtained from (4.13):

$$f(y|x) \equiv \frac{d}{dy} F(y|x)$$

$$\equiv \frac{d}{dy} P[Y \le y \mid X = x]$$

(4.16)

$$= \frac{d}{dy} \Big[\frac{1}{f(x)} \frac{d}{dx} P[Y \le y, X \le x] \Big] \qquad \text{[by (4.13)]}$$

$$= \frac{1}{f(x)} \frac{\partial^2}{\partial x \partial y} F(x, y)$$

$$\equiv \frac{f(x, y)}{f(x)}.$$

Thus the two definitions coincide in this case.

Remark 4.3. To illustrate the iteration formula (4.6), we obtain the following useful result:

 \Box

$$Cov(g(X), h(Y)) = E(g(X)h(Y)) - (Eg(X))(Eh(Y)) = E(E[g(X)h(Y) | X]) - (Eg(X))[E(E[h(Y) | X])] = E(g(X)E[h(Y) | X]) - (Eg(X))[E(E[h(Y) | X])] = Cov(g(X), E[h(Y)|X]).$$

Here we have used the *Product Formula*:

(4.18)
$$\operatorname{E}[g(X) h(Y) \mid X] = g(X) \operatorname{E}[h(Y) \mid X]. \quad \Box$$

Example 4.1. (Example 1.12 revisited.) Let $(X, Y) \sim \text{Uniform}(D)$, where D is the unit disk in \mathbb{R}^2 . In (1.44) we saw that

$$Y|X \sim \text{Uniform}\left(-\sqrt{1-X^2},\sqrt{1-X^2}\right),$$

which immediately implies $E[Y|X] \equiv 0$. Thus the iteration formula (4.6) and the covariance formula (4.17) yield, respectively,

$$E(Y) = E(E[Y|X]) = E(0) = 0,$$

$$Cov(X, Y) = Cov(X, E[Y|X]) = Cov(X, 0) = 0,$$

which are also clear from considering the joint distribution of (X, Y). \Box

Example 4.2. Let $(X, Y) \sim \text{Uniform}(T)$, where T is the triangle below, so

(4.19)
$$f(x,y) = \begin{cases} 2, & 0 < x < 1, \ 0 < y < x; \\ 0, & \text{otherwise.} \end{cases}$$

Thus

(4.20)
$$f(x) = 2xI_{(0,1)}(x),$$

hence

$$f(y|x) = \begin{cases} \frac{1}{x}, & 0 < y < x; \\ 0, & \text{otherwise.} \end{cases}$$

That is,

(4.21)
$$Y|X \sim \text{Uniform}(0, X)$$

[verify from the figure by "slicing"], so

(4.22)
$$E[Y \mid X] = \frac{X}{2}$$

From (4.20) we have $E(X) = \frac{2}{3}$ [verify], so the iteration formula gives

(4.23)
$$E(Y) = E(E[Y|X]) = E(\frac{X}{2}) = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}$$

Also from (4.17), (4.22), and the bilinearity of covariance,

$$\operatorname{Cov}(X, Y) = \frac{1}{2}\operatorname{Cov}(X, X) \equiv \frac{1}{2}\operatorname{Var}(X).$$

But [verify from (4.20)]

(4.24)
$$\operatorname{Var}(X) = \operatorname{E}(X^2) - (\operatorname{E}(X))^2 = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18},$$

so $Cov(X, Y) = \frac{1}{36} > 0$. Thus X and Y are positively correlated.

 \Box

Example 4.3. (A "mixed" joint distribution: Binomial-Uniform). Suppose that the joint distribution of (X, Y) is specified by the conditional distribution of X|Y and the marginal distribution of $Y(\equiv p^{*})$ as follows:

(4.25)
$$\begin{aligned} X|Y \sim \text{Binomial}(n,Y), & (\text{discrete}) \\ Y \sim & \text{Uniform}(0,1), & (\text{continuous}) \end{aligned}$$

 \mathbf{SO}

(4.26)
$$f(x|y) = \binom{n}{x} y^x (1-y)^{n-x}, \quad x = 0, \dots, n,$$
$$f(y) = 1, \qquad 0 < y < 1.$$

Here, X is discrete and Y is continuous, and their joint range is

(4.25)
$$\Omega_{X,Y} = \Omega_X \times \Omega_Y = \{0, 1, \dots, n\} \times (0, 1).$$

However, (4.25) shows that $X \not\perp Y$, since the conditional distribution of X|Y varies with Y. In particular,

$$(4.28) E[X | Y] = nY.$$

Suppose that only X is observed and we wish to estimate Y by E[Y|X]. For this we first need to find f(y|x) via Bayes formula (4.14). First,

$$f(x) \equiv P[X = x] = E(P[X = x | Y])$$

$$= E\left[\binom{n}{x}Y^{x}(1 - Y)^{n - x}\right]$$

$$= \binom{n}{x}\int_{0}^{1}y^{x}(1 - y)^{n - x}dy$$

$$= \binom{n}{x}\int_{0}^{1}y^{(x+1)-1}(1 - y)^{(n - x + 1)-1}dy$$

$$= \binom{n}{x}\frac{\Gamma(x + 1)\Gamma(n - x + 1)}{\Gamma(n + 2)} \quad [\text{see } (1.10)]$$

$$= \frac{n!}{x!(n - x)!} \cdot \frac{x!(n - x)!}{(n + 1)!}$$

$$= \frac{1}{n + 1}, \qquad x = 0, 1, \dots, n.$$

This shows that, marginally, X has the discrete uniform distribution over the integers $0, \ldots, n$. Then from (4.14),

$$f(y|x) = \frac{\binom{n}{x}y^x(1-y)^{n-x} \cdot 1}{\frac{1}{n+1}}$$

= $\frac{(n+1)!}{x!(n-x)!}y^x(1-y)^{n-x}$
$$\equiv \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)}y^x(1-y)^{n-x}, \quad 0 < y < 1.$$

Thus, the conditional ($\equiv posterior$) distribution of Y given X is

(4.31)
$$Y|X \sim \text{Beta}(X+1, n-X+1),$$

(

so the posterior \equiv Bayes estimator of Y|X is given by

$$E[Y \mid X] = \int_{0}^{1} y \cdot \left[\frac{\Gamma(n+2)}{\Gamma(X+1)\Gamma(n-X+1)} y^{X} (1-y)^{n-X} \right]$$

$$= \frac{\Gamma(n+2)}{\Gamma(X+1)\Gamma(n-X+1)} \int_{0}^{1} y^{(X+2)-1} (1-y)^{(n-X+1)-1}$$

$$= \frac{\Gamma(n+2)}{\Gamma(X+1)\Gamma(n-X+1)} \cdot \frac{\Gamma(X+2)\Gamma(n-X+1)}{\Gamma(n+3)}$$

$$= \frac{(n+1)!(X+1)!}{X!(n+2)!}$$

(4.32)
$$= \frac{X+1}{n+2}.$$

Remark 4.4. If we observe X = n successes (so no failures), then the Bayes estimator is $\frac{n+1}{n+2}$, not 1. In general, the Bayes estimator can be written as

(4.33)
$$\frac{X+1}{n+2} = \frac{n}{n+2} \left(\frac{X}{n}\right) + \frac{2}{n+2} \left(\frac{1}{2}\right),$$

which is a convex combination of the usual estimate $\frac{X}{n}$ and the *a priori* estimate $\frac{1}{2} \equiv E(Y)$. Thus the Bayes estimator adjusts the usual estimate to reflect the *a priori* assumption that $Y \sim \text{Uniform}(0, 1)$. Note, however, that the weight $\frac{n}{n+2}$ assigned to $\frac{X}{n}$ increases to 1 as the sample size $n \to \infty$, i.e., the prior information becomes less influential as $n \to \infty$. (See §16.) \Box

Example 4.4: Borel's Paradox. (This example shows the need for the limit operation (4.12) in the definition of the conditional probability $P[(X, Y) \in C | X = x]$ when X is continuous:)

As in Examples 1.12 and 4.1, let (X, Y) be uniformly distributed over the unit disk D in \mathbb{R}^2 and consider the conditional distribution of |Y| given X = 0. The "slicing" formula (1.30) applied to $f(x, y) \equiv I_D(x, y)$ gives

$$Y \mid \{X = 0\} \sim \text{Uniform}(-1, 1),$$

so [verify]

(4.34)
$$|Y| | \{X = 0\} \sim \text{Uniform}(0, 1).$$

However, if we represent (X, Y) in terms of polar coordinates (R, Θ) as in Example 1.12, then the event $\{X = 0\}$ is equivalent to $\{\Theta = \frac{\pi}{2} \text{ or } \frac{3\pi}{2}\}$, while under this event, |Y| = R. However, we know that $R \perp \Theta$ and $f(r) = 2rI_{(0,1)}(r)$ (recall (1.45) and (1.46a)), hence the "slicing" formula (1.30) applied to $f(r, \theta) \equiv f(r)f(\theta)$ shows that

(4.35)
$$R \mid \left\{ \Theta = \frac{\pi}{2} \text{ or } \frac{3\pi}{2} \right\} \sim f(r) \neq \text{Uniform}(0,1).$$

Since the left sides of (4.34) and (4.35) appear identical, this yields Borel's Paradox. The paradox is resolved by noting that, according to (4.12), conditioning on X is not equivalent to conditioning on Θ :

Conditioning on $\{X = 0\}$ Conditioning on $\left\{\Theta = \frac{\pi}{2} \text{ or } \frac{3\pi}{2}\right\}$ \square

5. Correlation, Prediction, and Regression.

5.1. Correlation. The covariance Cov(X, Y) indicates the nature (positive or negative) of the linear relationship (if any) between X and Y, but does not indicate the strength, or exactness, of this relationship. The *Pearson correlation coefficient*

(5.1)
$$\operatorname{Cor}(X, Y) = \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var} X} \sqrt{\operatorname{Var} Y}} \equiv \rho_{X,Y},$$

the standardized version of Cov(X, Y), does serve this purpose.

Properties:

(a) $\operatorname{Cor}(X, Y) = \operatorname{Cor}(Y, X).$

- (b) location-scale: $\operatorname{Cor}(aX + b, cY + d) = \operatorname{sgn}(ac) \cdot \operatorname{Cor}(X, Y).$
- (c) $-1 \le \rho_{X,Y} \le 1$. Equality holds $(\rho_{X,Y} = \pm 1)$ iff Y = aX + b for some a, b, i.e., iff X and Y are perfectly linearly related.

Proof. Let U = X - EX, V = Y - EY, and

$$g(t) = \mathrm{E}[(tU+V)^2] = t^2 \mathrm{E}(U^2) + 2t \mathrm{E}(UV) + \mathrm{E}(V^2).$$

Since this quadratic function is always ≥ 0 , its discriminant is ≤ 0 , i.e.

(5.2)
$$\left[\mathrm{E}(UV)\right]^2 \le \mathrm{E}(U^2) \cdot \mathrm{E}(V^2),$$

 \mathbf{SO}

(5.3)
$$\left[\operatorname{Cov}(X, Y)\right]^2 \le \operatorname{Var} X \cdot \operatorname{Var} Y,$$

which is equivalent to $\rho_{X,Y}^2 \leq 1$. [(5.2) is the Cauchy-Schwartz Inequality.]

Next, equality holds in (5.2) iff the discriminant of g is 0, so $g(t_0) = 0$ for some t_0 . But $g(t_0) = \mathbb{E}[(t_0U + V)^2]$, hence $t_0U + V \equiv 0$, so V must be exactly a linear function of U, i.e., Y is exactly a linear function of X.

Property (c) suggests that the closer $\rho_{X,Y}^2$ is to 1, the closer the relationship between X and Y is to exactly linear. (Also see (5.22).)

5.2. Mean-square error prediction (general regression).

For any rv Y s.t. $E(Y^2) < \infty$ and any $-\infty < c < \infty$,

$$E[(Y-c)^{2}] = E([(Y-EY) + (EY-c)]^{2})$$

= $E[(Y-EY)^{2}] + (EY-c)^{2} + 2(EY-c)E(Y-EY)$
(5.4) = $Var Y + (EY-c)^{2}$.

Thus c = EY is the best predictor of Y w.r.to mean-square error (MSE) in the absence of any other information, and the minimum MSE is

(5.5)
$$\min_{-\infty < c < \infty} \mathbf{E} \left[(Y - c)^2 \right] = \operatorname{Var} Y.$$

Now consider a bivariate rvtr (X, Y) with $E(Y^2) < \infty$. How can we best use the information in X to obtain a better predictor g(X) of Y than EY? That is, what function g(X) minimizes the MSE

(5.6)
$$E[(Y - g(X))^2] = E(E[(Y - g(X))^2 | X])?$$

But this follows immediately if we hold X fixed and in (5.4) and (5.5), replace the marginal distribution of Y by the conditional distribution of Y given X and replace c by g(X):

(5.7)
$$E[(Y - g(X))^2 \mid X] = Var[Y|X] + (E[Y|X] - g(X))^2;$$

(5.8)
$$\min_{-\infty < g(X) < \infty} E[(Y - g(X))^2 \mid X] = Var[Y|X].$$

Thus g(X) = E[Y|X] is the best predictor of Y based on X, and from (5.6) the minimum (unconditional) MSE is

(5.9)
$$\min_{g(X)} \operatorname{E}\left[(Y - g(X))^2\right] = \operatorname{E}\left(\operatorname{Var}[Y|X]\right).$$

The best predictor E[Y|X] is often called the regression function [explain] of Y on X. The prediction error Y - E[Y|X] is called the residual. The basic decomposition formula for the prediction of Y by X is:

(5.10)
$$Y = \mathbf{E}[Y|X] + (Y - \mathbf{E}[Y|X]) \equiv \text{best predictor} + \text{residual}.$$

Note that: E(best predictor) = EY and E(residual) = 0. [verify]

Proposition 5.1. (Variance decomposition). The variance of Y can be decomposed as follows:

(5.11)
$$\operatorname{Var} Y = \operatorname{Var} \left(\operatorname{E}[Y|X] \right) + \operatorname{E} \left(\operatorname{Var} [Y|X] \right)$$
$$= \operatorname{Var} (\operatorname{best predictor}) + \operatorname{Var} (\operatorname{residual})$$

Equivalently, the best predictor and residual are uncorrelated:

(5.12)
$$\operatorname{Cov}(\operatorname{E}[Y|X], Y - \operatorname{E}[Y|X]) = 0.$$

Proof. Set g(X) = EY in (5.7) to obtain

$$\mathbf{E}\left[(Y - \mathbf{E}Y)^2 \mid X\right] = \mathbf{Var}[Y|X] + \left(\mathbf{E}[Y|X] - \mathbf{E}Y\right)^2.$$

Take expectations w.r.to X to obtain (5.11) (since E(E[Y|X]) = EY). Now

(5.13)

$$\operatorname{Var}(Y - \operatorname{E}[Y|X]) = \operatorname{E}[(Y - \operatorname{E}[Y|X])^{2}] \quad [\operatorname{verify}] = \operatorname{E}\left(\operatorname{E}[(Y - \operatorname{E}[Y|X])^{2}] \mid X\right)$$
$$= \operatorname{E}(\operatorname{Var}[Y|X]), \quad [\operatorname{verify}]$$

so (5.12) follows from (3.13), (5.11), and (5.13).

Exercise 5.1. Prove (5.12) directly from (3.11) or (3.12) by conditioning on X and using the iteration formula.

 \Box

Remark 5.1. Consider a rvtr (X_1, \ldots, X_k, Y) with $E(Y^2) < \infty$. Then the best predictor $g(X_1, \ldots, X_k)$ of $Y \mid (X_1, \ldots, X_k)$ is $E[Y \mid X_1, \ldots, X_k]$. All of the results above remain valid with X replaced by X_1, \ldots, X_k . For example, (5.9) becomes

(5.14)
$$\min_{g(X_1,...,X_k)} \mathbb{E}\left[(Y - g(X_1,...,X_k))^2 \right] = \mathbb{E}\left(\operatorname{Var}[Y|X_1,...,X_k] \right),$$

which implies that

(5.15)
$$E(\operatorname{Var}[Y|X_1,\ldots,X_k]) \le \cdots \le E(\operatorname{Var}[Y|X_1]) \le \operatorname{Var} Y.$$

5.3. Linear prediction (\equiv linear regression).

In practice, the best predictor \equiv regression function E[Y|X] is unavailable, since to find it would require knowing the entire joint distribution of (X, Y). As a first step, we might ask to find the *linear* prediction function g(X) = a + bX that minimizes the MSE

(5.16)
$$E([Y - (a + bX)]^2).$$

First hold b fixed. From (5.4) with Y replaced by Y - bX, the MSE is minimized when

(5.17)
$$a = \hat{a}(b) = E(Y - bX) = E(Y) - bE(X),$$

so
$$\min_{a} \mathbb{E}\left[(Y - (a + bX)]^2\right] = \mathbb{E}\left([Y - \mathbb{E}Y] - b[X - \mathbb{E}X]\right)^2$$
$$= \operatorname{Var} Y - 2b \operatorname{Cov}(X, Y) + b^2 \operatorname{Var} X.$$

This is a quadratic in b and is minimized when

(5.18)
$$\hat{b} = \frac{\operatorname{Cov}(X, Y)}{\operatorname{Var} X},$$

so
$$\min_{a,b} E([Y - (a + bX)]^2) = Var Y - 2 \frac{[Cov(X, Y)]^2}{Var X} + \frac{[Cov(X, Y)]^2}{Var X}$$

(5.19) $= Var Y - \frac{[Cov(X, Y)]^2}{Var X}.$

Thus from (5.17) and (5.18), the best linear predictor (BLP) of Y|X is

$$\hat{a}(\hat{b}) + \hat{b}X = \left[\operatorname{E} Y - \left(\frac{\operatorname{Cov}(X, Y)}{\operatorname{Var} X} \right) \operatorname{E} X \right] + \left[\frac{\operatorname{Cov}(X, Y)}{\operatorname{Var} X} \right] X$$

$$(5.20) \qquad \qquad = \operatorname{E} Y + \left(\frac{\operatorname{Cov}(X, Y)}{\operatorname{Var} X} \right) [X - \operatorname{E} X].$$

(5.21)
$$\equiv \mu_Y + \rho_{X,Y} \left(\frac{\sigma_Y}{\sigma_X}\right) (X - \mu_X),$$

where $\mu_X = E X$, $\mu_Y = E Y$, $\sigma_X = sd(X)$, $\sigma_Y = sd(Y)$. From (5.19) the MSE of the BLP $\hat{a} + \hat{b}X$ is

(5.22)
$$E([Y - (\hat{a} + \hat{b}X)]^2) = \sigma_Y^2 - \frac{\rho_{X,Y}^2 \sigma_X^2 \sigma_Y^2}{\sigma_X^2} = (1 - \rho_{X,Y}^2) \sigma_Y^2,$$

The error of linear prediction $Y - (\hat{a} + \hat{b}X)$ is again called the *residual*. [See scatter plots.] The *basic decomposition formula* for linear prediction is:

(5.23)
$$Y = (\hat{a} + \hat{b}X) + [Y - (\hat{a} + \hat{b}X)]$$

$$\equiv \text{best linear predictor} + \text{residual.}$$

Again: E(best linear predictor) = EY and E(residual) = 0. [verify]

Proposition 5.2. (Linear variance decomposition). The variance of Y can be decomposed as follows:

(5.24)

$$\operatorname{Var} Y \equiv \rho_{X,Y}^2 \sigma_Y^2 + (1 - \rho_{X,Y}^2) \sigma_Y^2$$

$$\equiv \rho_{X,Y}^2 \left(\frac{\sigma_Y^2}{\sigma_X^2}\right) \sigma_X^2 + (1 - \rho_{X,Y}^2) \sigma_Y^2$$

$$= \operatorname{Var}(\operatorname{best\ linear\ predictor}) + \operatorname{Var}(\operatorname{residual}).$$

Equivalently, the best linear predictor and residual are uncorrelated:

(5.25)
$$\operatorname{Cov}\left(\hat{a} + \hat{b}X, Y - (\hat{a} + \hat{b}X)\right) = 0$$

Proof. From (5.21) and (5.22) [verify],

(5.26)
$$\operatorname{Var}(\mathrm{BLP}) = \rho_{X,Y}^2 \left(\frac{\sigma_Y^2}{\sigma_X^2}\right) \sigma_X^2,$$

(5.27)
$$\operatorname{Var}(\operatorname{residual}) = (1 - \rho_{X,Y}^2)\sigma_Y^2,$$

so (5.24) holds. Then (5.25) follows from (3.13) and (5.24).

Exercise 5.2. Prove (5.25) directly using the bilinearity of covariance.

Remark 5.2. Note that (5.22) gives another proof that $-1 \leq \rho_{X,Y} \leq 1$, or equivalently, $\rho_{X,Y}^2 \leq 1$. Also, it follows from (5.24) that $\rho_{X,Y}^2$, not $|\rho_{X,Y}|$, expresses the strength of the linear relationship between X and Y: if $\rho_{X,Y}^2 = 1$ then there is an exact linear relationship, while if $\rho_{X,Y}^2 = 0$ there is no overall linear relationship and the BLP reduces to the constant EY. Note that it is possible that $\rho_{X,Y}^2 < 1$ even if there is an exact nonlinear relationship between X and Y; for example if $Y = e^X$ exactly. (Here, however, there is a perfect linear relationship between $\log Y$ and X.) \square FPP scatter plot 1

FPP scatter plot 2

Remark 5.3. Suppose we know that E[Y|X] is a linear function, i.e., E[Y|X] = c + dX. (This holds in multinomial and multivariate normal distributions – see §7.5 and §8.3). Then necessarily $E[Y|X] = \hat{a} + \hat{b}X$, i.e., the best general predictor must coincide with the best linear predictor [why?]. In this case, all the results in §5.2 reduce to those in §5.3.

Exercise 5.3. Prove or disprove: E[Y|X] is linear $\Rightarrow E[X|Y]$ is linear.

Example 5.1. (Another Bayesian example.) Let X = height of father, Y = height of son. Suppose that the joint distribution of (X, Y) is specified by the conditional distribution of Y|X and the marginal distribution of X as follows:

(5.28)
$$\begin{aligned} Y|X \sim \operatorname{Normal}(X, \tau^2), \\ X \sim \operatorname{Normal}(\mu, \sigma^2), \end{aligned}$$

 \mathbf{SO}

(5.29)
$$f(y|x) = \frac{1}{\sqrt{2\pi\tau}} e^{-\frac{(y-x)^2}{2\tau^2}},$$
$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Suppose that we observe a son's height Y and want to estimate (\equiv predict) his father's height X using E[X|Y]. For this we find f(x|y) via Bayes' formula (4.14):

$$f(x|y) = \frac{f(y|x)f(x)}{f(y)}$$
$$= \frac{\text{const}}{f(y)} \cdot e^{-h(x)},$$

where

(5.30)
$$h(x) \equiv \frac{(y-x)^2}{2\tau^2} + \frac{(x-\mu)^2}{2\sigma^2}$$

is a quadratic in x with leading term

(5.31)
$$\frac{x^2}{2} \left(\frac{1}{\tau^2} + \frac{1}{\sigma^2} \right) \equiv \frac{x^2}{2} \left(\frac{1}{\gamma^2} \right).$$

Since h(x) is minimized when

$$h'(x) \equiv \frac{x-y}{\tau^2} + \frac{x-\mu}{\sigma^2} = 0,$$

i.e., when

(5.32)
$$x = \frac{\frac{y}{\tau^2} + \frac{\mu}{\sigma^2}}{\frac{1}{\tau^2} + \frac{1}{\sigma^2}} \equiv c(y),$$

it follows that

$$h(x) = \frac{(x - c(y))^2}{2\gamma^2} + d(y),$$

where d(y) does not involve x. Thus f(x|y) must have the following form:

$$f(x|y) = \frac{\operatorname{const} \cdot e^{-d(y)}}{f(y)} \cdot e^{-\frac{(x-c(y))^2}{2\gamma^2}}.$$

However, since f(x|y) must be a pdf in x for each fixed y, we conclude that

$$f(x|y) = \frac{1}{\sqrt{2\pi\gamma}} e^{-\frac{(x-c(y))^2}{2\gamma^2}}.$$

(Note that it was not necessary to find f(y)!) Thus the conditional (\equiv posterior) distribution of X given Y is

(5.33)
$$X|Y \sim N(c(Y), \gamma^2) \\ \equiv N\left(\frac{\frac{Y}{\tau^2} + \frac{\mu}{\sigma^2}}{\frac{1}{\tau^2} + \frac{1}{\sigma^2}}, \frac{1}{\frac{1}{\tau^2} + \frac{1}{\sigma^2}}\right),$$

so the best estimator \equiv predictor of X|Y is

(5.34)
$$E[X|Y] = \frac{\frac{Y}{\tau^2} + \frac{\mu}{\sigma^2}}{\frac{1}{\tau^2} + \frac{1}{\sigma^2}},$$

a linear function of Y. In fact, this is again a convex combination of the "unbiased" estimate Y (recall (5.30) and the *a priori* estimate $\mu \equiv E(X)$.

Note that the weights assigned to Y and μ are inversely proportional to τ^2 and σ^2 , respectively [interpret!]. Also, since E[X|Y] is a linear function of Y, (5.34) is also the best linear predictor of X|Y (recall Remark 5.3).

How good is E[X|Y]? From (5.13) (with X and Y interchanged), its MSE is given by

$$E[(X - E[X|Y])^{2}] = Var(X - E[X|Y])$$

= $E(Var[X|Y])$
(5.35) = $\frac{1}{\frac{1}{\tau^{2}} + \frac{1}{\sigma^{2}}}$
 $\approx \begin{cases} 0, & \text{if } \sigma^{2} \approx 0 \text{ (i.e., if prior info very good);} \\ \tau^{2}, & \text{if } \sigma^{2} \approx \infty \text{ (i.e., if prior info not good).} \end{cases}$

The result (5.35) can also be derived in terms of $\rho_{X,Y}$. First,

$$\operatorname{Var} X = \sigma^2; \qquad \qquad [by \ (5.28)]$$

$$\operatorname{Var} Y = \operatorname{Var} \left(\operatorname{E}[Y|X] \right) + \operatorname{E} \left(\operatorname{Var} \left[Y|X \right] \right) \qquad \text{[by (5.11)]}$$

$$= \operatorname{Var} X + \operatorname{E}(\tau^2) \qquad [by (5.28)]$$
$$= \sigma^2 + \tau^2:$$

$$Cov(X, Y) = Cov(X, E[Y|X])$$

$$= Cov(X, X)$$

$$= \sigma^{2}.$$
[by (4.17)]

Thus

(5.36)
$$1 - \rho_{X,Y}^2 = 1 - \frac{\sigma^2}{\sigma^2 + \tau^2} = \frac{\tau^2}{\sigma^2 + \tau^2},$$

so by (5.22) (with X and Y interchanged), the MSE of E[X|Y] is given by

(5.37)
$$(1 - \rho_{X,Y}^2) \, \sigma^2 = \frac{\tau^2 \sigma^2}{\sigma^2 + \tau^2},$$

which agrees with (5.35).

 \Box

Remark 5.4. Consider a rvtr (X_1, \ldots, X_k, Y) with $E(Y^2) < \infty$. All the above results remain valid but matrix notation is required – see §8.2. For example, the best linear predictor $g(\mathbf{X}_k) \equiv a + b^t \mathbf{X}_k$ of Y based on $\mathbf{X}_k \equiv (X_1, \ldots, X_k)^t$ is given by (compare to (5.20) and (8.76))

(5.38)
$$\hat{a} + \hat{b}^t \mathbf{X}_k = \mathbf{E} Y + \operatorname{Cov}(Y, \mathbf{X}_k) [\operatorname{Cov}(\mathbf{X}_k)]^{-1} (\mathbf{X}_k - \mathbf{E}(\mathbf{X}_k)),$$

while (5.24)-(5.27) remain valid if $\rho_{X,Y}^2$ is replaced by the *multiple correla*tion coefficient

(5.39)
$$\rho_{\mathbf{X}_k,Y}^2 \equiv \operatorname{Cov}(Y,\mathbf{X}_k)[\operatorname{Cov}(\mathbf{X}_k)]^{-1}\operatorname{Cov}(\mathbf{X}_k,Y)(\operatorname{Var} Y)^{-1}.$$

In particular, the extension of (5.22) implies that

(5.40)
$$1 \ge \rho_{\mathbf{X}_k,Y}^2 \ge \dots \ge \rho_{\mathbf{X}_1,Y}^2 \ge 0.$$

Remark 5.5. In practice, the population quantities μ_X , μ_Y , σ_X^2 , σ_Y^2 , and $\rho_{X,Y}$ that appear in the BLP $\hat{a} + \hat{b}X$ (5.21) are unknown, so must be estimated from a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. The usual estimators are:

$$\hat{\mu}_X = \bar{X}_n, \qquad \hat{\mu}_Y = \bar{Y}_n,$$
(5.41)
$$\hat{\sigma}_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2, \quad \hat{\sigma}_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2,$$

$$\hat{\rho}_{X,Y} = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2}}.$$

When (X, Y) has a bivariate normal distribution these estimators are the MLEs (CB Exercise 7.18), hence asymptotically optimal by Theorem 14.20.

Remark 5.6. If a linear predictor is not appropriate, i.e., if it is *not* the case that $Y \approx a + bX$ (+ error), it may be the case that a transformation will convert the non-linear relation into a linear one. For example,

(5.42) $Y \approx ae^{bX} \Rightarrow Y' \equiv \log Y \approx (\log a) + bX$

(5.43)
$$Y \approx aX^b \Rightarrow Y' \equiv \log Y \approx (\log a) + b(\log X),$$

(5.44) $Y \approx a + bX^2 \Rightarrow Y' \equiv Y \approx a + b(X^2).$

(Of course, one must assume that the error in Y', not Y, is additive.) \Box Note: The examples (5.42) - (5.44) emphasize that the "linearity" in a "linear model" comes from the (approximate) linear dependence of the response Y on the unknown parameters a, b, log a, etc., not on X! (See §8.5.)

5.4. Covariance and Regression.

Proposition 5.3. If E[Y|X] is a strictly increasing function of X, then Cov(X, Y) > 0.

Proof. This is an immediate consequence of (4.17) and the following lemma.

Lemma 5.1. (Chebyshev's Other Inequality) If X is a non-degenerate rv and g(X) and h(X) are both strictly increasing in X, then

(5.45)
$$\operatorname{Cov}(g(X), h(X)) > 0.$$

Proof. Let Y be another rv with the same distribution as X and independent of X. Then

$$[g(X) - g(Y)][h(X) - h(Y)] \ge 0,$$

with strict inequality whenever $X \neq Y$, which occurs with positive probability since X and Y are independent and non-degenerate [verify]. Thus

$$0 < E([g(X) - g(Y)][h(X) - h(Y)])$$

= E[g(X)h(X)] - E[g(X)h(Y)] - E[g(Y)h(X)] + E[g(Y)h(Y)]
= 2(E[g(X)h(X)] - E[g(X)]E[h(X)]),

since X and Y are i.i.d., which yields (5.45).

 \Box

6. Transforming Continuous Multivariate Distributions.

6.1. Two functions of two random variables.

Let (X, Y) be a continuous random vector with joint pdf $f_{X,Y}(x, y)$ and consider a transformation $(X, Y) \to (U, V)$, where

(6.1)
$$U = u(X, Y), \quad V = v(X, Y).$$

The pdf $f_{U,V}(u,v)$ is obtained by differentiating the joint cdf $F_{U,V}(u,v)$:

(6.2)

$$f_{U,V}(u,v) = \frac{\partial^2}{\partial u \partial v} F_{U,V}(u,v)$$

$$= \frac{\partial^2}{\partial u \partial v} P[U \le u, V \le v]$$

$$= \frac{\partial^2}{\partial u \partial v} \iint_{R(u,v)} f_{X,Y}(x,y) dx dy,$$

where R(u, v) is the region $\{(x, y) | u(x, y) \le u, v(x, y) \le v\}$. If the double integral can be evaluated⁵ explicitly, then the derivatives can be taken to obtain $f_{U,V}(u, v)$. Three examples of this method⁶ are now presented.

Example 6.1. Let (X, Y) be uniformly distributed on the unit square; set

(6.3)
$$U = \max(X, Y), \quad V = \min(X, Y).$$

In Example 2.5 we found the marginal pdfs of U and V separately. Here we find the joint pdf of (U, V) by using (6.2).

First specify the range of (U, V): this is just the triangle

$$T \equiv \{ (u, v) \mid 0 < u < 1, \ 0 < v < u \}$$

from Example 4.2. Then for $(u, v) \in T$,

⁵ If the mapping $(X, Y) \to (U, V)$ given by (6.1) is differentiable and 1-1, then the integral in (6.2) need not be evaluated – instead, $f_{U,V}(u, v)$ can be obtained by simply multiplying $f_{X,Y}(x, y)$ by the Jacobian – see §6.2.

⁶ Another example was given earlier - see (1.45).

$$F_{U,V}(u,v) = P[U \le u] - P[U \le u, V > v]$$

= $P[X \le u, Y \le u] - P[v < X \le u, v < Y \le u]$
= $u^2 - (u - v)^2$, [by independence - verify]
(6.4) $f_{U,V}(u,v) = 2I_T(u,v)$,

which is the same as (4.19). Thus (U, V) is uniformly distributed on T. Note:

$$E(UV) = E(XY) = (EX)(EY) = \frac{1}{4},$$

$$E(U+V) = E(X+Y) = 1,$$

$$E(V) = \int_0^1 [1 - F_V(v)] dv = \int_0^1 (1-v)^2 dv = \frac{1}{3},$$

so $E(U) = \frac{2}{3}$, hence (recall Example 4.2)

$$Cov(U, V) = \frac{1}{4} - (EU)(EV) = \frac{1}{4} - \left(\frac{2}{3}\right)\left(\frac{1}{3}\right) = \frac{1}{36}.$$
 \Box

Example 6.2. Let X, Y be i.i.d. Exponential(1) rvs and set

(6.5)
$$V = \min(X, Y), \qquad Z = |X - Y| \equiv \max(X, Y) - \min(X, Y).$$

In Example 2.6 we found the marginal pdfs of V and Z separately. Here we find the joint pdf of (V, Z) via (6.2).

The range of (V, Z) is $(0, \infty) \times (0, \infty)$. For $0 < v < \infty, 0 < z < \infty$,

$$P[V \le v, Z \le z] = P[Z \le z] - P[V > v, Z \le z],$$

$$P[V > v, Z \le z] = P[X > v, Y > v, |X - Y| \le z]$$

= $P[(X, Y) \text{ in shaded region}]$
(6.6)
= $2P[(X, Y) \text{ in half of region}]$
= $2P[X > v, X \le Y \le X + z]$

$$= 2 \int_{v}^{\infty} e^{-x} \left(\int_{x}^{x+z} e^{-y} dy \right) dx$$

= $2 \int_{v}^{\infty} e^{-x} \left(e^{-x} - e^{-x-z} \right) dx$
= $2 (1 - e^{-z}) \int_{v}^{\infty} e^{-2x} dx$
= $e^{-2v} (1 - e^{-z}),$

where (6.6) follows by symmetry: $(X, Y) \sim (Y, X)$. Therefore

(6.7)

$$f_{V,Z}(v,z) = -\frac{\partial^2}{\partial v \partial z} \left[e^{-2v} \left(1 - e^{-z} \right) \right]$$

$$= \left(2e^{-2v} \right) \left(e^{-z} \right)$$

$$= f_V(v) f_Z(z), \quad \text{[recall (2.17), (2.20)]}$$

so V and Z are *independent*, with

 $V \sim \text{Exponential}(2), \qquad Z \sim \text{Exponential}(1).$

Interpretation: Suppose that X, Y represent the lifetimes of two lightbulbs. Thus, $V \equiv \min(X, Y)$ is the time to the *first* burnout (either X or Y). Once the first burnount occurs, the time to the second burnout has the *original* exponential distribution Expo(1), not Expo(2). This is another memoryfree property of the exponential distribution. It is stronger in that it shows that the process renews itself at the random time V. (The first memory-free property concerned any fixed renewal time t.)

Exercise 6.1.** (Converse of Example 6.2: a second characterization of the exponential distribution (compare to Exercise 1.2).) Let X, Y be nonnegative i.i.d. rvs with common pdf f on $(0, \infty)$. Show that if $V \equiv \min(X, Y)$ and $Z \equiv |X - Y|$ are independent, then X and Y must be exponential rvs, i.e., $f(x) = \lambda e^{-\lambda x}$ for some $\lambda > 0$.

Hint: follow the approach of Example 6.2. For simplicity, you may assume that f is strictly positive and continuous.

Example 6.3. Again let X, Y be i.i.d. Exponential(1) rvs and set

(6.8)
$$U = X + Y, \qquad W = \frac{X}{X + Y}.$$

In Example 2.6 we found the marginal pdfs of U and W separately. Here we find the joint pdf of (U, W) via (6.2).

The range of (U, W) is $(0, \infty) \times (0, 1)$. For $0 < u < \infty, 0 < w < 1$,

$$P[U \le u, W \le w] = P[X + Y \le u, X \le w(X + Y)]$$

$$= E\left(P\left[Y \le u - X, Y \ge \left(\frac{1 - w}{w}\right)X \mid X\right]\right)$$

$$= \int_{0}^{uw} e^{-x} \left(\int_{\left(\frac{1 - w}{w}\right)x}^{u - x} e^{-y} dy\right) dx$$

$$= \int_{0}^{uw} \left(e^{-\frac{x}{w}} - e^{-u}\right) dx$$

$$= \left[1 - e^{-u} - ue^{-u}\right] \cdot w.$$

Therefore

(6.10)

$$f_{U,W}(u,w) = \frac{\partial^2}{\partial u \partial w} [1 - e^{-u} - u e^{-u}] \cdot w$$

$$= (u e^{-u}) \cdot 1$$

$$= f_U(u) f_W(w), \qquad [\text{recall } (2.16), (2.21)]$$

so U and W are *independent*, with

$$U \sim \text{Gamma}(2,1), \qquad W \sim \text{Uniform}(0,1).$$

Interpretation: As noted in Example 2.6, (6.9) and (6.10) can be viewed as a "backward" memory-free property of the exponential distribution: given X+Y, the location of X is uniformly distributed over the interval (0, X+Y), i.e., over the "past".

6.2. The Jacobian method.

Let A, B be open sets in \mathbb{R}^n and

(6.11)
$$T: A \to B$$
$$x \equiv (x_1, \dots, x_n) \mapsto y \equiv (y_1, \dots, y_n)$$

a smooth bijective (1-1 and onto) mapping (\equiv diffeomorphism). The Jacobian matrix of this mapping is given by

(6.12)
$$J_T(x) \equiv \left(\frac{\partial y}{\partial x}\right) := \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial y_1}{\partial x_n} & \cdots & \frac{\partial y_n}{\partial x_n} \end{pmatrix};$$

the *Jacobian* of the mapping is given by

$$|J_T(x)| \equiv \left|\frac{\partial y}{\partial x}\right| := |\det\left(\frac{\partial y}{\partial x}\right)| \ge 0.$$

Theorem 6.1. (Substitution formula for multiple integrals.) Let A, B be open sets in \mathbb{R}^n and $T: A \to B$ a diffeomorphism such that $|J_T(x)| > 0$ for "a.e." x. Let f be a real-valued integrable function on A. Then

(6.13)
$$\int_{A} f(x)dx = \int_{B=T(A)} f(T^{-1}(y)) \left| \frac{\partial x}{\partial y} \right| dy. \quad \text{[explain]}$$

[See (6.16) for the relation between $\left|\frac{\partial x}{\partial y}\right|$ and $\left|\frac{\partial y}{\partial x}\right|$.]

Corollary 6.2. Let X be a random vector (rvtr) with pdf $f_X(x)$ (wrto Lebesgue measure) on \mathbb{R}^n . Suppose that $A := \{x \mid f_X(x) > 0\}$ is open and that $T : A \to B$ is a diffeomorphism with $J_T(x) > 0$ a.e. Then the pdf of Y := T(X) is given by

(6.14)
$$f_Y(y) = f_X(T^{-1}(y)) \cdot \left|\frac{\partial x}{\partial y}\right| \cdot I_B(y).$$

Proof. For any (measurable) set $C \subseteq \mathbf{R}^n$,

$$P[Y \in C] = P[X \in T^{-1}(C)]$$
$$= \int_{T^{-1}(C)} f_X(x) dx$$
$$= \int_C f(T^{-1}(y)) \left| \frac{\partial x}{\partial y} \right| dy,$$

which confirms (6.14) [Why?]

The calculation of Jacobians can be simplified by application of the following rules.

Chain Rule: Suppose that $x \mapsto y$ and $y \mapsto z$ are diffeomorphisms. Then $x \mapsto z$ is a diffeomorphism and

(6.15)
$$\left|\frac{\partial z}{\partial x}\right| = \left|\frac{\partial z}{\partial y}\right|_{y=y(x)} \cdot \left|\frac{\partial y}{\partial x}\right|.$$

[This follows from the chain rule for partial derivatives:

$$\frac{\partial z_i(y_1(x_1,\ldots,x_n),\ldots,y_n(x_1,\ldots,x_n))}{\partial x_j} = \sum_k \frac{\partial z_i}{\partial y_k} \frac{\partial y_k}{\partial x_j} = \left[\left(\frac{\partial z}{\partial y} \right) \left(\frac{\partial y}{\partial x} \right) \right]_{ij}.$$

Therefore $\left(\frac{\partial z}{\partial x}\right) = \left(\frac{\partial z}{\partial y}\right) \left(\frac{\partial y}{\partial x}\right)$; now take determinants.]

Inverse Rule: Suppose that $x \mapsto y$ is a diffeomorphism. Then

$$\frac{\partial x}{\partial y}\Big|_{y=y(x)} = \Big|\frac{\partial y}{\partial x}\Big|^{-1}$$

[Set z = x in (6.15).] Reversing the roles of x and y we obtain

(6.16)
$$\left|\frac{\partial y}{\partial x}\right|_{x=x(y)} = \left|\frac{\partial x}{\partial y}\right|^{-1}.$$

Combination Rule: Suppose that $x \mapsto u$ and $y \mapsto v$ are (unrelated) diffeomorphisms. Then

(6.17)
$$\left|\frac{\partial(u,v)}{\partial(x,y)}\right| = \left|\frac{\partial u}{\partial x}\right| \cdot \left|\frac{\partial v}{\partial y}\right|.$$

 \Box

The Jacobian matrix is given by

$$\left(\frac{\partial(u,v)}{\partial(x,y)}\right) = \left(\begin{array}{cc}\frac{\partial u}{\partial x} & 0\\ 0 & \frac{\partial v}{\partial y}\end{array}\right).$$

Extended Combination Rule: Suppose that $(u, v) \mapsto (x, y)$ is a diffeomorphism of the form u = u(x), v = v(x, y). Then (6.17) remains valid.

[The Jacobian matrix is given by

$$\left(\frac{\partial(u,v)}{\partial(x,y)}\right) = \left(\begin{array}{cc}\frac{\partial u}{\partial x} & \frac{\partial v}{\partial x}\\ 0 & \frac{\partial v}{\partial y}\end{array}\right).$$

Jacobians of linear mappings. Let $A: p \times p$ and $B: n \times n$ be nonsingular matrices and c a nonzero scalar (A, B, L, M, U, V, c fixed.) Then:

(a) vectors: y = cx, $x, y: p \times 1$: $\left|\frac{\partial y}{\partial x}\right| = |c|^p$. [combination rule] (b) matrices: Y = cX, $X, Y: p \times n$: $\left|\frac{\partial Y}{\partial X}\right| = |c|^{pn}$. [comb. rule] (c) symmetric matrices: Y = cX, $X, Y: p \times p$, symmetric: $\left|\frac{\partial Y}{\partial X}\right| = |c|^{\frac{p(p+1)}{2}}$. [comb. rule] (d) vectors: y = Ax, $x, y: p \times 1$, $A: p \times p$: $\left|\frac{\partial y}{\partial x}\right| = |\det A|$. [verify] (e) matrices: Y = AX, $X, Y: p \times n$: $\left|\frac{\partial Y}{\partial X}\right| = |\det A|^n$. [comb. rule] $Y = XB, X, Y: p \times n$: $\left|\frac{\partial Y}{\partial X}\right| = |\det B|^p$. [comb. rule] $Y = AXB, X, Y: p \times n$: $\left|\frac{\partial Y}{\partial X}\right| = |\det A|^n |\det B|^p$. [chain rule]

Example 6.4. The Gamma (α, λ) distribution with shape parameter $\alpha > 0$ and intensity parameter $\lambda > 0$ has pdf

(6.18)
$$g(x;\alpha,\lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} I_{(0,\infty)}(x).$$

Let $X \sim \text{Gamma}(\alpha, \lambda)$ and $Y \sim \text{Gamma}(\beta, \lambda)$ be independent Gamma random variables with the same scale parameter and define

(6.19)
$$U = X + Y, \qquad W = \frac{X}{X + Y}.$$

Find the joint pdf of (U, W), find the marginal pdfs of U and W, and show that U and W are independent. (This extends Example 6.3.)

Solution. The ranges of (X, Y) and (U, W) are

$$A := \{ (x, y) \mid x > 0, y > 0 \} \equiv (0, \infty) \times (0, \infty),$$

$$B := \{ (u, w) \mid u > 0, 0 < w < 1 \} \equiv (0, \infty) \times (0, 1),$$

respectively. Notice that both A and B are Cartesian product sets. The transformation

(6.20)
$$T: A \to B$$
$$(x, y) \mapsto (x + y, \ x/(x + y))$$

is bijective, with inverse given by

(6.21)
$$T^{-1}: B \to A$$
$$(u, w) \mapsto (uw, u(1-w)).$$

Thus T^{-1} is continuously differentiable and bijective, and its Jacobian is given by [verify!]

(6.22)
$$\left| \frac{\partial(x,y)}{\partial(u,w)} \right| = \left| \begin{array}{cc} \frac{\partial(uw)}{\partial u} & \frac{\partial(uw)}{\partial w} \\ \frac{\partial(u(1-w))}{\partial u} & \frac{\partial(u(1-w))}{\partial w} \end{array} \right| = \left| \begin{array}{cc} w & u \\ 1-w & -u \end{array} \right| = u.$$

Because

(6.23)
$$f_{X,Y}(x,y) = \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} e^{-\lambda x} y^{\beta-1} e^{-\lambda y} I_A(x,y),$$

it follows from (6.14) that

$$f_{U,W}(u,w) = \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} (uw)^{\alpha-1} e^{-\lambda uw} (u(1-w))^{\beta-1} e^{-\lambda u(1-w)} \cdot u \cdot I_B(u,w)$$

$$(6.24) = \frac{\lambda^{\alpha+\beta} u^{\alpha+\beta-1} e^{-\lambda u}}{\Gamma(\alpha+\beta)} I_{(0,\infty)}(u) \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} w^{\alpha-1} (1-w)^{\beta-1} I_{(0,1)}(w)$$

$$\equiv f_U(u) \cdot f_W(w).$$

Thus $U \perp W$, $U \sim \text{Gamma}(\alpha + \beta, \lambda)$, $W \sim \text{Beta}(\alpha, \beta)$.

Remark 6.1. (The converse of Example 6.4 – a characterization of the Gamma distribution.) The Gamma family is the *only* family of distributions on $(0, \infty)$ with the property that X + Y and $\frac{X}{X+Y}$ are independent. [References: Hogg (1951 Ann. Math. Statist.); Lukacs (1955 Ann. Math. Statist.); G. Marsaglia (Festschrift for I. Olkin); Kagan, Linnik, Rao (book).

Remark 6.2. Let U and W be independent with $U \sim \text{Gamma}(\alpha + \beta, \lambda)$ and $W \sim \text{Beta}(\alpha, \beta)$. It follows from Example 6.4 that $UW \sim \text{Gamma}(\alpha, \lambda)$, $U(1-W) \sim \text{Gamma}(\beta, \lambda)$, and $UW \perp U(1-W)$ [verify].

Exercise 6.2. Let X, Y, Z be i.i.d. Exponential(1) rvs and let

$$U = \frac{X}{X+Y}, \qquad V = \frac{X+Y}{X+Y+Z}, \qquad W = X+Y+Z.$$

Show that $U \perp \!\!\!\perp V \perp \!\!\!\perp W$ and find the distributions of U, V, and W.

Example 6.5. Let X, Y be i.i.d. random variables each having an Exponential (λ) distribution on $(0, \infty)$ with pdf

(6.25)
$$f_{\lambda}(x) = \lambda e^{-\lambda x} I_{(0,\infty)}(x)$$

w.r.to Lebesgue measure. Find the joint pdf of (V, Z), where

$$V = \min(X, Y), \qquad Z = X - Y.$$

(Be sure to specify the range of (V, Z).) Find the marginal pdfs of V and Z, and show they are independent.

Solution. The ranges of (X, Y) and (V, Z) are

(6.26)
$$A := \{(x, y) \mid x > 0, y > 0\} \equiv (0, \infty) \times (0, \infty), B := \{(v, z) \mid v > 0\} \equiv (0, \infty) \times (-\infty, \infty),$$

respectively. Notice that both A and B are Cartesian product sets. The transformation

(6.27)
$$T: A \to B$$
$$(x, y) \mapsto (\min(x, y), \ x - y) \equiv (v, z)$$

is bijective, with inverse given by

(6.28)
$$\begin{array}{c} T^{-1} : B \to A \\ (v, z) \mapsto (v + z^+, \ v + z^-) \equiv (x, y), \end{array}$$

where $z^+ = \max(z, 0), z^- = -\min(z, 0)$. Then T^{-1} is continuously differentiable on the open set $B^* := B \setminus N_1$, where $N_1 := \{(v, z) \mid z = 0\}$ is a (Lebesgue-) null set. The Jacobian of T^{-1} on B^* is given by [verify!]

(6.29)
$$\left| \frac{\partial(x,y)}{\partial(v,z)} \right| = \begin{cases} + \begin{vmatrix} 1 & 1 \\ 1 & 0 \end{vmatrix} = 1, & \text{if } z > 0; \\ + \begin{vmatrix} 1 & 1 \\ 0 & -1 \end{vmatrix} = 1, & \text{if } z < 0. \end{cases}$$

Because

(6.30)
$$f_{X,Y}(x,y) = \lambda^2 e^{-\lambda(x+y)} I_A(x,y),$$

it follows from (6.14) and (6.28) that

(6.31)

$$f_{V,Z}(v,z) = \lambda^2 e^{-\lambda(2v+|z|)} I_B(v,z)$$

$$= 2\lambda e^{-2\lambda v} I_{(0,\infty)}(v) \cdot \frac{\lambda}{2} e^{-\lambda|z|} I_{(-\infty,\infty)}(z)$$

$$\equiv f_V(v) \cdot f_Z(z).$$

Thus, V and Z are independent, V has an Exponential (2λ) distribution on $(0,\infty)$, and Z has a "double exponential distribution" on $(-\infty,\infty)$. \Box

Exercise 6.3.** (Converse of Example 6.5: another characterization of the exponential distribution (compare to Exercise 6.1**).) Let X, Y be i.i.d. positive random variables, each having a continuous and positive pdf f on $(0, \infty)$. Define V, Z as in Example 6.5 and assume that V and Z are independent. Show that X and Y each must have an Exponential (λ) distribution, i.e., $f = f_{\lambda}$ for some $\lambda > 0$, as in (6.25)

Hint: Express the joint pdf $f_{V,Z}$ and the marginal pdfs f_V , f_Z in terms of f. By independence, $f_{V,Z} = f_V f_Z$. Deduce that

(6.32)
$$f(v+|z|) = [1 - F(v)]f(|z|) \text{ for } v > 0, -\infty < u < \infty,$$

where $F(v) = \int_0^v f(x) dx$. (To be perfectly rigorous, you have to beware of null sets, i.e., exceptional sets of measure 0.)

For extra credit, prove this result without the simplifying assumption that f is positive (MDP STAT 383). For double super extra credit, prove this result without assuming that f is either positive or continuous.

Exercise 6.4. (Continuation of Proposition 3.1.) Let T_1, T_2, \ldots be the jump times in a homogeneous Poisson process with intensity parameter λ .

- (i) Show that T_1 and $T_2 T_1$ are independent Exponential(λ) rvs.
- (ii) For any $n \ge 3$, show that $T_1, T_2 T_1, \ldots, T_n T_{n-1}$ are i.i.d. Exponential (λ) rvs.

Example 6.6. (Polar coordinates in \mathbb{R}^2). Let (X, Y) be a continuous bivariate rvtr with joint pdf $f_{X,Y}(x,y)$ on \mathbb{R}^2 . Find the joint pdf $f_{R,\Theta}(r,\theta)$ of (R,Θ) , where $(X,Y) \to (R,\Theta)$ is the 1-1 transformation [verify] whose inverse is given by

(6.33)
$$X = R\cos\Theta, \qquad Y = R\sin\Theta.$$

Solution. The range of (R, Θ) is $(0, \infty) \times [0, 2\pi)$, the Cartesian product of the ranges of R and Θ . The Jacobian of (6.33) is

(6.34)
$$\left|\frac{\partial(x,y)}{\partial(r,\theta)}\right| = \left|\begin{array}{cc}\cos\theta & \sin\theta\\-r\sin\theta & r\cos\theta\end{array}\right| = r,$$

so from (6.14),

(6.35)
$$f_{R,\Theta}(r,\theta) = f_{X,Y}(r\cos\theta, r\sin\theta) \cdot r \; .$$

In the case where $f_{X,Y}(x,y)$ is radial, i.e.,

$$f_{X,Y}(x,y) = g(x^2 + y^2)$$

(recall (2.9), (2.10)), (6.35) becomes

(6.36)
$$f_{R,\Theta}(r,\theta) = rg(r^2)$$
$$= 2\pi rg(r^2) \cdot \frac{1}{2\pi}.$$

This shows that:

(i) R and Θ are independent; (ii) $f_R(r) = 2\pi r g(r^2);$ (iii) $\Theta \sim \text{Uniform}[0, 2\pi).$

A special case appeared in Example 1.12, where (X, Y) was uniformly distributed over the unit disk D in \mathbb{R}^2 , i.e., (cf. (1.45))

$$f_{X,Y}(x,y) = \frac{1}{\pi} I_D(x,y) = \frac{1}{\pi} I_{(0,1)}(x^2 + y^2) \equiv g(x^2 + y^2).$$

Thus R and Θ are independent, $\Theta \sim \text{Uniform}[0, 2\pi)$, and (cf. (1.46a))

$$f_R(r) = 2\pi r g(r^2) = 2r I_{(0,1)}(r).$$

Another special case occurs when X, Y are i.i.d. N(0, 1) rvs. Here

$$f_{X,Y}(x,y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}} \equiv g(x^2+y^2),$$

so again R and Θ are independent, $\Theta \sim \text{Uniform}[0, 2\pi)$, and

$$f_R(r) = 2\pi r g(r^2) = re^{-\frac{r^2}{2}}$$

Finally, set $S = R^2$ (= $X^2 + Y^2$). Then $\frac{ds}{dr} = 2r$, so

$$f_S(s) = f_R(r(s)) \cdot \frac{dr}{ds} = re^{-\frac{r^2}{2}} \cdot \frac{1}{2r} = \frac{1}{2}e^{-\frac{s}{2}}$$

hence $S \sim \text{Exponential}\left(\lambda = \frac{1}{2}\right) \equiv \text{Gamma}\left(\frac{2}{2}, \frac{1}{2}\right) \equiv \chi_2^2$ (see Remark 6.3). $\overline{\sqcup}$

These results extend to polar coordinates in \mathbb{R}^n . Suppose that $X \equiv (X_1, \ldots, X_n)$ is a continuous rvtr with joint pdf $f(x_1, \ldots, x_n)$. Then X can be represented by polar coordinates $R, \Theta_1, \ldots, \Theta_{n-1}$ in several different ways, depending on how the angles Θ_i are defined. However, in each case $R = \sqrt{X_1^2 + \cdots + X_n^2}$ and the Jacobian has the form

(6.37)
$$\left|\frac{\partial(x_1,\ldots,x_n)}{\partial(r,\theta_1,\ldots,\theta_{n-1})}\right| = r^{n-1} \cdot h(\theta_1,\ldots,\theta_{n-1})$$

for some function h. Thus, if $f(x_1, \ldots, x_n) = g(x_1^2 + \cdots + x_n^2)$, i.e., if f is radial, then by (6.14) the joint pdf of $(R, \Theta_1, \ldots, \Theta_{n-1})$ again factors:

(6.38)
$$f_{R,\Theta_1,\dots,\Theta_{n-1}}(r,\theta_1,\dots,\theta_{n-1}) = r^{n-1}g(r^2) \cdot h(\theta_1,\dots,\theta_{n-1}).$$

Thus R is independent of $(\Theta_1, \ldots, \Theta_{n-1})$ and has pdf of the form

(6.39)
$$f_R(r) = c_n \cdot r^{n-1} g(r^2).$$

For example, if X_1, \ldots, X_n are i.i.d. N(0, 1) then

(6.40)
$$f_R(r) = c_n \cdot r^{n-1} e^{-\frac{r^2}{2}}.$$

Again set $S = R^2$, so [verify]

(6.41)
$$f_S(s) = c'_n \cdot s^{\frac{n}{2}-1} e^{-\frac{s}{2}},$$

from which we recognize that

(6.42)
$$S \equiv R^2 \equiv X_1^2 + \dots + X_n^2 \sim \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right) \ (=\chi_n^2).$$

Remark 6.3. The chi-square distribution with n degrees of freedom, denoted by χ_n^2 , is defined as in (6.42) to be the distribution of $X_1^2 + \cdots + X_n^2$ where X_1, \ldots, X_n are i.i.d. standard normal N(0, 1) rvs. The pdf of χ_1^2 was shown directly in (2.7) of Example 2.4 to be that of the Gamma $(\frac{1}{2}, \frac{1}{2})$ distribution, from which (6.42) also follows from Application 3.3.3 in §3.3 on moment generating functions, or from Example 6.4 in the present section.

Remark 6.4. The transformations in Examples 6.3 - 6.6 are 1-1, while those in Examples 6.1 and 6.2 are 2-1 [verify].

7. The Multinomial Distribution.

The multinomial experiment. Consider a random experiment with k possible outcomes (or "categories", or "cells"). Let $p_i = P(C_i)$. Suppose the experiment is repeated independently n times, i.e., n independent trials are carried out, resulting in X_i observations in cell C_i :

Cells:
$$C_1, C_2, \ldots, C_k,$$

Probabilities $p_1, p_2, \ldots, p_k, \sum_{i=1}^k p_i = 1,$
Counts: $X_1, X_2, \ldots, X_k, \sum_{i=1}^k X_i = n.$

Definition. 7.1. The distribution of (X_1, \ldots, X_k) is called the *multinomial* distribution for k cells, n trials, and cell probabilities p_1, \ldots, p_k . We write

(7.1)
$$(X_1,\ldots,X_k) \sim M_k(n; p_1,\ldots,p_k).$$

The multinomial distribution is a discrete multivariate distribution. Note that the marginal distribution of each X_i is $Binomial(n; p_i)$. In fact, when k = 2, M_2 essentially reduces to the binomial distribution:

(7.2)
$$X \sim \text{Binomial}(n; p) \iff (X, n - X) \sim M_2(n; p, 1 - p)$$

The multinomial pmf and mgf. For $x_1 \ge 0, \ldots, x_k \ge 0$, $\sum_{i=1}^k x_i = n$ $(x_i$'s integers),

(7.3)
$$P[(X_1, \dots, X_k) = (x_1, \dots, x_k)] = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

[Draw picture of range; discuss complete vs. incomplete multinomial dist'n.] Here the *multinomial coefficient*

$$\frac{n!}{x_1!\cdots x_k!} = \binom{n}{x_1,\ldots,x_k}$$

is the number of ways that the labels " C_1, \ldots, C_k " can be assigned to $1, \ldots n$ such that label C_i occurs x_i times, $i = 1, \ldots, k$.

[First, the *n* distinguishable labels $C_1^1, \ldots, C_1^{x_1}, \ldots, C_k^1, \ldots, C_k^{x_k}$ can be assigned to $1, \ldots, n$ in n! ways. But there are $x_i!$ permutations of $C_i^1, \ldots, C_i^{x_i}$.]

The fact that the probabilities in (7.3) sum to 1 follows either from their interpretation as probabilities or from the *multinomial expansion*

(7.4)
$$(p_1 + \dots + p_k)^n = \sum_{x_i \ge 0, \sum x_i = n} \binom{n}{x_1, \dots, x_k} p_1^{x_1} \cdots p_k^{x_k}.$$

If we replace p_i by $p_i e^{t_i}$ in (7.4) we obtain the multinomial mgf (7.5):

$$m_{X_{1},...,X_{k}}(t_{1},...,t_{k}) \equiv \mathbb{E}\left(e^{t_{1}X_{1}+\cdots+t_{k}X_{k}}\right)$$

$$(7.5) = \left(p_{1}e^{t_{1}}+\cdots+p_{k}e^{t_{k}}\right)^{n} \quad [p_{i} \mapsto p_{i}e^{t_{i}} \text{ in } (7.4)]$$

$$(alternatively:) = e^{t_{k}n}\left(p_{1}e^{t_{1}-t_{k}}+\cdots+p_{k-1}e^{t_{k-1}-t_{k}}+p_{k}\right)^{n}$$

$$(7.6) = e^{t_{k}n}m_{X_{1},...,X_{k-1}(t_{1}-t_{k},...,t_{k-1}-t_{k}}),$$

since $X_k = n - X_1 - \cdots - X_{k-1}$. Expression (7.6) shows that the distribution of (X_1, \ldots, X_k) is determined by that of (X_1, \ldots, X_{k-1}) , which must be the case since $X_k = n - X_1 - \cdots - X_{k-1}$.

Additional trials. Let

$$(X_1,\ldots,X_k) \sim M_k(m; p_1,\ldots,p_k),$$

$$(Y_1,\ldots,Y_k) \sim M_k(n; p_1,\ldots,p_k),$$

denote the cell counts based on m and n independent multinomial trials with the same k and same p_i 's. Then obviously

(7.7)
$$(X_1 + Y_1, \dots, X_k + Y_k) \sim M_k(m+n; p_1, \dots, p_k).$$

Combining cells. Suppose $(X_1, \ldots, X_k) \sim M_k(n; p_1, \ldots, p_k)$ and define new "combined cells" D_1, D_2, \ldots, D_r as follows:

(7.8)
$$\overbrace{C_{1},\ldots,C_{k_{1}}}^{D_{1}}, \overbrace{C_{k_{1}+1},\ldots,C_{k_{2}}}^{D_{2}}, \ldots, \overbrace{C_{k_{r-1}+1},\ldots,C_{k_{r}}}^{D_{r}},$$

where $1 \leq r < k$ and $1 \leq k_1 < k_2 < \ldots < k_{r-1} < k_r \equiv k$. Define the *combined cell counts* to be

$$Y_1 = X_1 + \dots + X_{k_1}, Y_2 = X_{k_1+1} + \dots + X_{k_2}, \dots, Y_r = X_{k_{r-1}+1} + \dots + X_{k_r}$$

and the *combined cell probabilities* to be

 $q_1 = p_1 + \dots + p_{k_1}, \ q_2 = p_{k_1+1} + \dots + p_{k_2}, \dots, \ q_r = p_{k_{r-1}+1} + \dots + p_{k_r}.$ Then obviously

(7.9)
$$(Y_1, \ldots, Y_r) \sim M_r(n; q_1, \ldots, q_r).$$
 [same n]

In particular, for any l with $1 \le l < k$, (7.10) $(X_1, \ldots, X_l, n - (X_1 + \cdots + X_l)) \sim M_{l+1}(n; p_1, \ldots, p_l, 1 - (p_1 + \cdots + p_l)).$

Conditional distributions. First consider k = 4:

 $(X_1, X_2, X_3, X_4) \sim M_4(n; p_1, p_2, p_3, p_4).$

In (7.8) let $r = 2, k_1 = 2, k_2 = 4$, so

$$Y_1 = X_1 + X_2,$$
 $Y_2 = X_3 + X_4,$
 $q_1 = p_1 + p_2,$ $q_2 = p_3 + p_4,$

hence by (7.9),

$$(Y_1, Y_2) \sim M_2(n; q_1, q_2).$$

Therefore the conditional distribution of (X_1, X_2, X_3, X_4) given (Y_1, Y_2) is as follows: for integers x_1, x_2, x_3, x_4 and y_1, y_2 such that $x_1 + x_2 = y_1$, $x_3 + x_4 = y_2$, and $y_1 + y_2 = n$,

$$P[(X_1, X_2, X_3, X_4) = (x_1, x_2, x_3, x_4) \mid (Y_1, Y_2) = (y_1, y_2)]$$

$$= \frac{P[(X_1, X_2, X_3, X_4) = (x_1, x_2, x_3, x_4)]}{P[(Y_1, Y_2) = (y_1, y_2)]}$$
$$= \frac{\frac{n!}{x_1! x_2! x_3! x_4!} p_1^{x_1} p_2^{x_2} p_3^{x_3} p_4^{x_4}}{\frac{n!}{y_1! y_2!} q_1^{y_1} q_2^{y_2}} \qquad \text{[by (7.3) and (7.9)]}$$
$$= \frac{y_1!}{y_1! y_2!} (p_1)^{x_1} (p_2)^{x_2} = y_2! = (p_3)^{x_3} (p_4)^{x_4}$$

(7.11)
$$= \frac{y_1!}{x_1!x_2!} \left(\frac{p_1}{q_1}\right)^{x_1} \left(\frac{p_2}{q_1}\right)^{x_2} \cdot \frac{y_2!}{x_3!x_4!} \left(\frac{p_3}{q_2}\right)^{x_3} \left(\frac{p_4}{q_2}\right)^{x_4}.$$

This shows that (X_1, X_2) and (X_3, X_4) are conditionally independent given $X_1 + X_2$ and $X_3 + X_4$ ($\equiv n - (X_1 + X_2)$), and that [discuss]

(7.12)
$$(X_1, X_2) \mid X_1 + X_2 \sim M_2 \left(X_1 + X_2; \frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2} \right) \\ (X_3, X_4) \mid X_3 + X_4 \sim M_2 \left(X_3 + X_4; \frac{p_3}{p_3 + p_4}, \frac{p_4}{p_3 + p_4} \right).$$

In particular,

(7.13)
$$X_1 \mid X_1 + X_2 \sim \text{Binomial}\Big(X_1 + X_2, \frac{p_1}{p_1 + p_2}\Big),$$

 \mathbf{SO}

(7.14)
$$E[X_1 | X_1 + X_2] = (X_1 + X_2) \left(\frac{p_1}{p_1 + p_2}\right).$$

Verify:

$$np_1 = \mathcal{E}(X_1) = \mathcal{E}(\mathcal{E}[X_1 \mid X_1 + X_2]) = \mathcal{E}\left[(X_1 + X_2)\left(\frac{p_1}{p_1 + p_2}\right)\right] = np_1.$$

It also follows from (7.12) that

(7.15)
$$(X_1, X_2) \mid X_3 + X_4 \sim M_2 \Big(n - (X_3 + X_4); \frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2} \Big),$$

which shows the *negative* (linear!) relation between (X_1, X_2) and (X_3, X_4) .

Now consider the general case, as in (7.8). Then a similar argument shows that

$$(X_1, \ldots, X_{k_1}), (X_{k_1+1}, \ldots, X_{k_2}), \ldots, (X_{k_{r-1}+1}, \ldots, X_{k_r})$$

are conditionally independent given

$$X_1 + \dots + X_{k_1}, \ X_{k_1+1} + \dots + X_{k_2}, \ \dots, \ X_{k_{r-1}+1} + \dots + X_{k_r},$$

with

(7.16)
$$(X_{k_{i-1}+1}, \dots, X_{k_i}) \mid X_{k_{i-1}+1} + \dots + X_{k_i}$$

$$\sim M_{k_i-k_{i-1}} \left(X_{k_{i-1}+1} + \dots + X_{k_i}; \frac{p_{k_{i-1}+1}}{p_{k_{i-1}+1} + \dots + p_{k_i}}, \dots, \frac{p_{k_i}}{p_{k_{i-1}+1} + \dots + p_{k_i}} \right)$$

for $i = 1, \dots, r$, where $k_0 \equiv 0$.

In particular, for $2 \le l < k$,

(7.17)
$$(X_1, \dots, X_l) \mid X_{l+1}, \dots, X_k$$
$$\sim M_l \Big(n - (X_{l+1} + \dots + X_k); \frac{p_1}{p_1 + \dots + p_l}, \dots, \frac{p_l}{p_1 + \dots + p_l} \Big).$$

Because $X_1 + \cdots + X_k = n$, we know there is a negative linear relation between any X_i and X_j . Let's verify this explicitly. For $i \neq j$,

(7.18)
$$X_i \mid X_j \sim \operatorname{Binomial}\left(n - X_j; \frac{p_i}{1 - p_j}\right),$$

(7.19)
$$E[X_i \mid X_j] = (n - X_j) \left(\frac{p_i}{1 - p_j}\right),$$
$$Cov(X_i, X_j) = Cov(E[X_i \mid X_j], X_j)$$
$$= Cov\left((n - X_j) \left(\frac{p_i}{1 - p_j}\right), X_j\right)$$
$$= -\left(\frac{p_i}{1 - p_j}\right) Var(X_j)$$
$$= -\left(\frac{p_i}{1 - p_j}\right) np_j(1 - p_j)$$
$$= -np_i p_j \le 0.$$

An alternative derivation is based on the variance formula (3.13):

$$2 \operatorname{Cov}(X_i, X_j) = \operatorname{Var}(X_i + X_j) - \operatorname{Var}(X_i) - \operatorname{Var}(X_j)$$

= $n(p_i + p_j)(1 - p_i - p_j) - np_i(1 - p_i) - np_j(1 - p_j)$
= $-n(p_i + p_j)^2 + np_i^2 + np_j^2$
= $-2np_ip_j$,

which yields (7.20). (The second line follows from the fact that $X_i + X_j \sim \text{Binomial}(n; p_i + p_j)$.)

Remark 7.1. For $k \ge 3$ the range of (X_i, X_j) shown here, indicates the negative relation between X_i and X_j :

Exercise 7.1. (Representation of a multinomial rvtr in terms of independent Poisson rvs.) Let Y_1, \ldots, Y_k be independent Poisson rvs with $Y_i \sim \text{Poisson}(\lambda_i), i = 1, \ldots, k$. Show that

$$(Y_1,\ldots,Y_k)|\{Y_1+\cdots+Y_k=n\}\sim M_k\Big(n;\,\frac{\lambda_1}{\lambda_1+\cdots+\lambda_k},\ldots,\frac{\lambda_k}{\lambda_1+\cdots+\lambda_k}\Big).$$

Exponential family. From (7.3), the multinomial pmf can be written as an exponential family (see Example 11.11). When $x_1 + \cdots + x_k = n$,

$$(7.21)
\frac{n!}{x_1!\cdots x_k!} p_1^{x_1}\cdots p_k^{x_k}
= \frac{n!}{x_1!\cdots x_k!} e^{x_1\log p_1+\cdots+x_{k-1}\log p_{k-1}+x_k\log p_k}
= \frac{n!}{x_1!\cdots x_k!} e^{n\log(1-p_1-\cdots-p_{k-1})} e^{x_1\log\frac{p_1}{1-p_1-\cdots-p_{k-1}}+\cdots+x_{k-1}\log\frac{p_{k-1}}{1-p_1-\cdots-p_{k-1}}}.$$

This is a (k-1)-parameter exponential family with natural parameters

(7.22)
$$\theta_i = \log \frac{p_i}{1 - p_1 - \dots - p_{k-1}}, \quad i = 1, \dots, k-1.$$

(Note that the Binomial(n, p) family is a one-parameter exponential family (see Example 11.10) with natural parameter $\log \frac{p}{1-p}$.)

Maximum likelihood estimates (MLE). The MLE of (p_1, \ldots, p_k) is $(\hat{p}_1, \ldots, \hat{p}_k) = \left(\frac{X_1}{n}, \ldots, \frac{X_k}{n}\right)$. [See Example 14.24.]

Representation of a multinomial rvtr as a sum of i.i.d. Bernoulli (0-1) rvtrs. First recall that a binomial rv $X \sim Bin(n, p)$ can be represented as the sum of n i.i.d. Bernoulli rvs:

(7.23)
$$X = U_1 + \dots + U_n,$$

where
$$U_j = \begin{cases} 1, & \text{if Success on trial } j;\\ 0, & \text{if Failure on trial } j. \end{cases}$$

Now extend this to multinomial trials. Consider a multinomial experiment as in §7.1 with n i.i.d. trials and k possible outcomes (cells) C_1, \ldots, C_k . Again let p_i be the probability of cell C_i and X_i be the total number of outcomes in cell C_i . Form the column vectors

(7.24)
$$\mathbf{X} \equiv \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}, \quad \mathbf{p} \equiv \begin{pmatrix} p_1 \\ \vdots \\ p_k \end{pmatrix},$$

so we can write $\mathbf{X} \sim M_k(n; \mathbf{p})$. Then the multinomial rvtr \mathbf{X} can be represented as

(7.25)
$$\mathbf{X} = \mathbf{U}_{1} + \dots + \mathbf{U}_{n},$$
where
$$\mathbf{U}_{j} = \begin{pmatrix} U_{1j} \\ \vdots \\ U_{kj} \end{pmatrix} : k \times 1$$
with
$$U_{ij} = \begin{cases} 1, & \text{if cell } C_{i} \text{ occurs on trial } j; \\ 0, & \text{if cell } C_{i} \text{ does not occur on trial } j. \end{cases}$$

Note that each \mathbf{U}_j is a *Bernoulli rvtr:* it has exactly one 1 and k - 1 0's. Clearly (7.25) generalizes (7.23).

The representations (7.23) and (7.25) are very convenient for finding moments and applying the Central Limit Theorem to obtain normal approximations to the binomial and multinomial distributions. For example, since $E(U_j) = p$ and $V(U_j) = p(1-p)$ in (7.23), it follows that in the binomial case, E(X) = np and Var(X) = np(1-p), and that as $n \to \infty$,

(7.26)
$$\frac{X - np}{\sqrt{np(1-p)}} \stackrel{d}{\to} N(0,1) \quad \text{if } 0$$

(7.27) or equivalently,
$$\frac{X - np}{\sqrt{n}} \stackrel{d}{\to} N(0, p(1-p))$$

For the multinomial, the mean vector and covariance matrix of \mathbf{U}_j are

$$\mathbf{E}(\mathbf{U}_j) = \begin{pmatrix} \mathbf{E}(U_{1j}) \\ \vdots \\ \mathbf{E}(U_{kj}) \end{pmatrix} = \begin{pmatrix} p_1 \\ \vdots \\ p_k \end{pmatrix} = \mathbf{p},$$

$$\operatorname{Cov}(\mathbf{U}_{j}) = \begin{pmatrix} \operatorname{Var}(U_{1j}) & \operatorname{Cov}(U_{1j}, U_{2j}) & \cdots & \operatorname{Cov}(U_{1j}, U_{kj}) \\ \operatorname{Cov}(U_{2j}, U_{1j}) & \operatorname{Var}(U_{2j}) & \cdots & \operatorname{Cov}(U_{2j}, U_{kj}) \\ \vdots & \vdots & & \vdots \\ \operatorname{Cov}(U_{kj}, U_{1j}) & \operatorname{Cov}(U_{kj}, U_{2j}) & \cdots & \operatorname{Var}(U_{kj}) \end{pmatrix} \\ = \begin{pmatrix} p_{1}(1 - p_{1}) & -p_{1}p_{2} & \cdots & -p_{1}p_{k} \\ -p_{2}p_{1} & p_{2}(1 - p_{2}) & \cdots & -p_{2}p_{k} \\ \vdots & \vdots & & \vdots \\ -p_{k}p_{1} & -p_{k}p_{2} & \cdots & p_{k}(1 - p_{k}) \end{pmatrix} \quad [\text{verify}] \\ \equiv D_{\mathbf{p}} - \mathbf{pp}', \end{cases}$$

where $D_{\mathbf{p}} = \text{diag}(p_1, \ldots, p_k)$. (Note that $D_{\mathbf{p}} - \mathbf{pp'}$ is a *singular* matrix of rank k - 1, so it has no inverse.) Thus by (7.25) and the independence of $\mathbf{U}_1, \ldots, \mathbf{U}_k$,

(7.28)
$$\operatorname{E}(\mathbf{X}) = n\mathbf{p}, \quad \operatorname{Cov}(\mathbf{X}) = n(D_{\mathbf{p}} - \mathbf{pp}').$$

Therefore, it follows from the multivariate Central Limit Theorem that

(7.29)
$$\frac{\mathbf{X} - n\mathbf{p}}{\sqrt{n}} \stackrel{d}{\to} N_k \left(0, D_{\mathbf{p}} - \mathbf{p}\mathbf{p}' \right).$$

Now suppose that $p_1 > 0, \ldots, p_k > 0$. Then $D_{\mathbf{p}}$ is nonsingular, so by the continuity of convergence in distribution (§10.2),

(7.30)
$$D_{\mathbf{p}}^{-\frac{1}{2}}\left(\frac{\mathbf{X}-n\mathbf{p}}{\sqrt{n}}\right) \xrightarrow{d} N_k\left(0, I_k - \mathbf{u}\mathbf{u}'\right),$$

where $D_{\mathbf{p}}^{-\frac{1}{2}} = \operatorname{diag}\left(p_{1}^{-\frac{1}{2}}, \ldots, p_{k}^{-\frac{1}{2}}\right)$ and $\mathbf{u} \equiv D_{\mathbf{p}}^{-\frac{1}{2}}\mathbf{p}$ is a unit vector, i.e., $\mathbf{u}'\mathbf{u} = 1$. Again by the continuity of convergence in distribution,

(7.31)
$$\left\| D_{\mathbf{p}}^{-\frac{1}{2}} \left(\frac{\mathbf{X} - n\mathbf{p}}{\sqrt{n}} \right) \right\|^2 \stackrel{d}{\to} \left\| N_k \left(0, I_k - \mathbf{u}\mathbf{u}' \right) \right\|^2 \sim \chi_{k-1}^2,$$

since $I_k - \mathbf{u}\mathbf{u}'$ is a projection matrix of rank k - 1 (apply Fact 8.5).

But

(7.32)
$$\left\| D_{\mathbf{p}}^{-\frac{1}{2}} \left(\frac{\mathbf{X} - n\mathbf{p}}{\sqrt{n}} \right) \right\|^{2} = \sum_{i=1}^{k} \frac{(X_{i} - np_{i})^{2}}{np_{i}}$$
$$\equiv \sum_{i=1}^{k} \frac{(\text{Observed}_{i} - \text{Expected}_{i})^{2}}{\text{Expected}_{i}} \equiv \chi^{2},$$

which is (Karl) Pearson's classical *chi-square goodness-of-fit statistic* for testing the simple null hypothesis **p**. Thus we have derived Pearson's classic result that $\chi^2 \xrightarrow{d} \chi^2_{k-1}$.

(However, Pearson got the degrees of freedom wrong! He first asserted that $\chi^2 \xrightarrow{d} \chi_k^2$, but was corrected by Fisher, which Pearson did not entirely appreciate!)

Remark 7.2. Note that (7.29) is an extension of (7.27), not (7.26), which has no extension to the multinomial case since $D_{\mathbf{p}} - \mathbf{pp}'$ is singular, hence has no inverse. However, if we reduce \mathbf{X} to $\tilde{\mathbf{X}} \equiv (X_1, \ldots, X_{k-1})'$, then

(7.33)
$$\operatorname{Cov}(\tilde{\mathbf{X}}) \equiv n(\tilde{D}_{\mathbf{p}} - \tilde{\mathbf{p}}\tilde{\mathbf{p}}')$$

is nonsingular provided that $p_1 > 0, \ldots, p_k > 0$ [verify], where $\tilde{D}_{\mathbf{p}} = \text{diag}(p_1, \ldots, p_{k-1})$ and $\tilde{\mathbf{p}} = (p_1, \ldots, p_{k-1})'$. Then (7.26) can be extended:

(7.34)
$$\left(\tilde{D}_{\mathbf{p}} - \tilde{\mathbf{p}}\tilde{\mathbf{p}}'\right)^{-\frac{1}{2}} (\tilde{\mathbf{X}} - n\tilde{\mathbf{p}}) \xrightarrow{d} N_{k-1}(0, I_{k-1}),$$

from which (7.31) can also be obtained (but with more algebra).

8. Linear Models and the Multivariate Normal Distribution.

8.1. Review of vectors and matrices. (The results are stated for vectors and matrices with real entries but also hold for complex entries.) An $m \times n$ matrix $A \equiv \{a_{ij}\}$ is an array of mn numbers:

 $A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mm} \end{pmatrix}.$

This matrix represents the linear mapping (\equiv linear transformation)

where $x \in \mathbf{R}^n$ is written as an $n \times 1$ column vector and

$$Ax \equiv \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \equiv \begin{pmatrix} \sum_{j=1}^n a_{1j} x_j \\ \vdots \\ \sum_{j=1}^n a_{mj} x_j \end{pmatrix} \in \mathbf{R}^m.$$

The mapping (8.1) clearly satisfies the *linearity property:*

$$A(ax+by) = aAx+bBy.$$

Matrix addition: If $A \equiv \{a_{ij}\}$ and $B \equiv \{b_{ij}\}$ are $m \times n$ matrices, then

$$(A+B)_{ij} = a_{ij} + b_{ij}.$$

Matrix multiplication: If A is $m \times n$ and B is $n \times p$, then the matrix product AB is the $m \times p$ matrix AB whose *ij*-th element is

(8.2)
$$(AB)_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}.$$

Then AB is the matrix of the composition $\mathbf{R}^p \xrightarrow{B} \mathbf{R}^n \xrightarrow{A} \mathbf{R}^m$ of the two linear mappings determined by A and B [verify]:

$$(AB)x = A(Bx) \qquad \forall x \in \mathbf{R}^p.$$

Rank of a matrix: The row (column) rank of a matrix $A : m \times n$ is the dimension of the linear space spanned by its rows (columns). The rank of A is the dimension r of the largest nonzero minor (= $r \times r$ subdeterminant) of A. Then [verify]

$$\operatorname{row} \operatorname{rank}(A) \leq \min(m, n),$$

$$\operatorname{column} \operatorname{rank}(A) \leq \min(m, n),$$

$$\operatorname{rank}(A) \leq \min(m, n),$$

$$\operatorname{row} \operatorname{rank}(A) = m - \dim \left([\operatorname{row} \operatorname{space}(A)]^{\perp} \right),$$

$$\operatorname{column} \operatorname{rank}(A) = n - \dim \left([\operatorname{column} \operatorname{space}(A)]^{\perp} \right),$$

$$\operatorname{row} \operatorname{rank}(A) = \operatorname{column} \operatorname{rank}(A)$$

$$= \operatorname{rank}(A) = \operatorname{rank}(A')$$

$$= \operatorname{rank}(AA') = \operatorname{rank}(A'A).$$

Furthermore, for $A: m \times n$ and $B: n \times p$,

$$\operatorname{rank}(AB) \le \min(\operatorname{rank}(A), \operatorname{rank}(B)).$$

Inverse matrix: If $A : n \times n$ is a square matrix, its *inverse* A^{-1} (if it exists) is the unique matrix that satisfies

$$AA^{-1} = A^{-1}A = I,$$

where I is the $n \times n$ identity matrix⁷ diag $(1, \ldots, 1)$. If A^{-1} exists then A is called *nonsingular* (or *regular*). The following are equivalent:

- (a) A is nonsingular.
- (b) The *n* columns of *A* are linearly independent (i.e., column rank(A) = n). Equivalently, $Ax \neq 0$ for every nonzero $x \in \mathbf{R}^n$.
- (c) The *n* rows of *A* are linearly independent (i.e., row rank(*A*) = *n*). Equivalently, $x'A \neq 0$ for every nonzero $x \in \mathbf{R}^n$.
- (d) The determinant $|A| \neq 0$ (i.e., rank(A) = n). [Define det geometrically.]

⁷ It is called the "identity" matrix since $Ix = x \ \forall x \in \mathbb{R}^n$.

Note that if A is nonsingular then A^{-1} is nonsingular and $(A^{-1})^{-1} = A$.

If $A: m \times m$ and $C: n \times n$ are nonsingular and B is $m \times n$, then [verify]

$$\operatorname{rank}(AB) = \operatorname{rank}(B) = \operatorname{rank}(BC).$$

If $A: n \times n$ and $B: n \times n$ are nonsingular then so is AB, and [verify]

$$(8.3) (AB)^{-1} = B^{-1}A^{-1}.$$

If $A \equiv \operatorname{diag}(d_1, \dots, d_n)$ with all $d_i \neq 0$ then $A^{-1} = \operatorname{diag}(d_1^{-1}, \dots, d_n^{-1})$.

Transpose matrix: If $A \equiv \{a_{ij}\}$ is $m \times n$, its *transpose* is the $n \times m$ matrix A' (also denoted by A^t) whose *ij*-th element is a_{ji} . That is, the m row vectors (n column vectors) of A are the m column vectors (n row vectors) of A'. Note that [verify]

(8.4)
$$(A+B)' = A' + B';$$

(8.5)
$$(AB)' = B'A' \qquad (A:m \times n, B:n \times p);$$

(8.6)
$$(A^{-1})' = (A')^{-1}$$
 $(A: n \times n, \text{ nonsingular}).$

Trace: For a square matrix $A \equiv \{a_{ij}\} : n \times n$, the *trace* of A is

(8.7)
$$\operatorname{tr}(A) = \sum_{i=1}^{n} a_{ii},$$

the sum of the diagonal entries of A. Then

$$\begin{array}{ll} (8.8) & \operatorname{tr}(aA+bB) = a\operatorname{tr}(A) + b\operatorname{tr}(B); \\ (8.9) & \operatorname{tr}(AB) = \operatorname{tr}(BA); & (Note:A:m\times n,\ B:n\times m) \\ (8.10) & \operatorname{tr}(A') = \operatorname{tr}(A). & (A:n\times n) \end{array}$$

Proof of (8.9):

$$\operatorname{tr}(AB) = \sum_{i=1}^{m} (AB)_{ii} = \sum_{i=1}^{m} \left(\sum_{k=1}^{n} a_{ik} b_{ki} \right)$$
$$= \sum_{k=1}^{n} \left(\sum_{i=1}^{m} b_{ki} a_{ik} \right) = \sum_{k=1}^{n} (BA)_{kk} = \operatorname{tr}(BA).$$

Determinant: For a square matrix $A \equiv \{a_{ij}\} : n \times n$, its *determinant* is

$$|A| = \sum_{\pi} \epsilon(\pi) \prod_{i=1}^{n} a_{i\pi(i)}$$
$$= \pm \text{Volume}(A([0,1]^n)) \qquad \text{[discuss]},$$

where π ranges over all n! permutations of $1, \ldots, n$ and $\epsilon(\pi) = \pm 1$ according to whether π is an even or odd permutation. Then

(8.11)
$$|AB| = |A| \cdot |B|$$
 $(A, B : n \times n);$

$$(8.12) |A^{-1}| = |A|^{-1}$$

$$(8.13) |A'| = |A|$$

|A'| = |A|; $|A| = \prod_{i=1}^{n} a_{ii} \text{ if } A \text{ is triangular (or diagonal).}$ (8.14)

Orthogonal matrix. An $n \times n$ matrix U is orthogonal if

$$(8.15) UU' = I.$$

This is equivalent to the fact that the n row vectors of U form an orthonormal basis for \mathbb{R}^n . Note that (8.15) implies that $U' = U^{-1}$, hence also U'U = I, which is equivalent to the fact that the n column vectors of U also form an orthonormal basis for \mathbf{R}^n .

Note that U preserves angles and lengths, i.e., preserves the usual inner product and norm in \mathbf{R}^n : for $x, y \in \mathbf{R}^n$,

$$(Ux, Uy) \equiv (Ux)'(Uy) = x'U'Uy = x'y \equiv (x, y),$$

 \mathbf{SO}

$$||Ux||^2 \equiv (Ux, Ux) = (x, x) \equiv ||x||^2$$

In fact, any orthogonal transformation is a product of rotations and reflections. Also, from (8.13) and (8.15), $|U|^2 = 1$, so $|U| = \pm 1$.

Complex numbers and matrices. For any complex number $c \equiv a + ib \in \mathbf{C}$, let $\bar{c} \equiv a - ib$ denote the *complex conjugate* of c. Note that $\bar{\bar{c}} = c$ and

$$c\bar{c} = a^2 + b^2 \equiv |c|^2,$$
$$\overline{cd} = \bar{c}\overline{d}.$$

For any complex matrix $C \equiv \{c_{ij}\}$, let $\overline{C} = \{\overline{c}_{ij}\}$ and define $C^* = \overline{C'}$. Note that

$$(8.16) (CD)^* = D^*C^*.$$

The characteristic roots \equiv eigenvalues of the $n \times n$ matrix A are the n roots l_1, \ldots, l_n of the polynomial equation

$$(8.17) |A - l I| = 0.$$

These roots may be real or complex; the complex roots occur in conjugate pairs. Note that the eigenvalues of a triangular or diagonal matrix are just its diagonal elements.

By (b) (for matrices with possibly complex entries), for each eigenvalue l there exists some nonzero (possibly complex) vector $u \in \mathbb{C}^n$ s.t.

(A - lI)u = 0,

equivalently,

$$(8.18) Au = lu.$$

The vector u is called a *characteristic vector* \equiv *eigenvector* for the eigenvalue l. Since any nonzero multiple cu is also an eigenvector for l, we will usually normalize u to be a unit vector, i.e., $||u||^2 \equiv u^*u = 1$.

For example, if A is a diagonal matrix, say

$$A = \operatorname{diag}(d_1, \dots, d_n) \equiv \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{pmatrix},$$

then its eigenvalues are just d_1, \ldots, d_n , with corresponding eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$, where

(8.19)
$$\mathbf{u}_i \equiv (0, \dots, 0, \underbrace{1}_i, 0, \dots, 0)'$$

is the *i*-th unit coordinate vector.

Note, however, that in general, eigenvalues need not be distinct and eigenvectors need not be unique. For example, if A is the identity matrix I, then its eigenvalues are $1, \ldots, 1$ and every unit vector $u \in \mathbf{R}^n$ is an eigenvector for the eigenvalue 1: $Iu = 1 \cdot u$.

However, eigenvectors u, v associated with two *distinct* eigenvalues l, m cannot be proportional: if u = cv then

$$lu = Au = cAv = cmv = mu,$$

which contradicts the assumption that $l \neq m$.

Symmetric matrix. An $n \times n$ matrix $S \equiv \{s_{ij}\}$ is symmetric if S = S', i.e., if $s_{ij} = s_{ji} \forall i, j$.

Fact 8.1. Let S be a real symmetric $n \times n$ matrix.

(a) Each eigenvalue l of S is real and has a real eigenvector $u \in \mathbf{R}^n$.

(b) If $l \neq m$ are distinct eigenvalues of S with corresponding real eigenvectors u and v, then $u \perp v$, i.e., u'v = 0. Thus if all the eigenvalues of S are distinct, each eigenvalue l has exactly one real eigenvector u.

Proof. (a) Let *l* be an eigenvalue of *S* with eigenvector $u \neq 0$. Then

$$Su = lu \Rightarrow u^*Su = lu^*u = l.$$

But S is real and symmetric, so $S^* = S$, hence

$$\overline{u^*Su} = (u^*Su)^* = u^*S^*(u^*)^* = u^*Su.$$

Thus u^*Su is real, hence l is real. Since S - lI is real, the existence of a real eigenvector u for l now follows from (b), p.93.

(b) We have Su = lu and Sv = mv, hence

$$l(u'v) = (lu)'v = (Su)'v = u'Sv = u'(mv) = m(u'v),$$

so u'v = 0 since $l \neq m$.

Fact 8.2. (Spectral decomposition of a real symmetric matrix.) Let S be a real symmetric $n \times n$ matrix with eigenvalues l_1, \ldots, l_n (necessarily real). Then there exists a real orthogonal matrix U such that

$$(8.20) S = U D_l U',$$

where $D_l = \text{diag}(l_1, \ldots, l_n)$. Since $SU = UD_l$, the *i*-th column vector u_i of U is a real eigenvector for l_i .

Proof. For simplicity we suppose that l_1, \ldots, l_n are distinct. Let u_1, \ldots, u_n be the corresponding unique real unit eigenvectors (apply Fact 8.1b). Since u_1, \ldots, u_n is an orthonormal basis for \mathbb{R}^n , the matrix

$$(8.21) U \equiv (u_1 \quad \cdots \quad u_n) \quad : n \times n$$

satisfies U'U = I, i.e., U is an orthogonal matrix. Since each u_i is an eigenvector for l_i , $SU = UD_l$ [verify], which is equivalent to (8.20).

(The case where the eigenvalues are not distinct can be established by a "perturbation" argument. Perturb S slightly so that its eigenvalues become distinct [non-trivial] and apply the first case. Now use a limiting argument based on the compactness of the set of all $n \times n$ orthogonal matrices.) \Box

Fact 8.3. If S is a real symmetric matrix with eigenvalues l_1, \ldots, l_n ,

(8.22)
$$\operatorname{tr}(S) = \sum_{i=1}^{n} l_i ;$$

(8.23)
$$|S| = \prod_{i=1}^{n} l_i$$
.

Proof. This is immediate from the spectral decomposition (8.20) of S. \Box

 \Box

Positive definite matrix. An $n \times n$ matrix S is positive semi-definite (psd) if it is symmetric and its quadratic form is nonnegative:

(8.24)
$$x'Sx \ge 0 \qquad \forall x \in \mathbf{R}^n;$$

S is positive definite (pd) if it is symmetric and its quadratic form is positive:

(8.25)
$$x'Sx > 0 \quad \forall \text{ nonzero } x \in \mathbf{R}^n.$$

- The identity matrix is pd: $x'Ix = ||x||^2 > 0$ if $x \neq 0$.
- A diagonal matrix diag (d_1, \ldots, d_n) is psd (pd) iff each $d_i \ge 0$ (> 0).
- If $S: n \times n$ is psd, then ASA' is psd for any $A: m \times n$.
- If $S: n \times n$ is pd, then ASA' is pd for any $A: m \times n$ of full rank $m \leq n$.
- AA' is psd for any $A: m \times n$.
- AA' is pd for any $A: m \times n$ of full rank $m \leq n$.

Note: This shows that the proper way to "square" a matrix A is to form AA' (or A'A), not A^2 (which need not even be symmetric).

• S pd \Rightarrow S has full rank \Rightarrow S^{-1} exists \Rightarrow $S^{-1} \equiv (S^{-1})S(S^{-1})'$ is pd.

Fact 8.4. (a) A symmetric $n \times n$ matrix S with eigenvalues l_1, \ldots, l_n is psd (pd) iff each $l_i \ge 0$ (> 0). In particular, $|S| \ge 0$ (> 0) if S is psd (pd), so a pd matrix is nonsingular.

(b) Suppose S is pd with distinct eigenvalues $l_1 > \cdots > l_n > 0$ and corresponding unique real unit eigenvectors u_1, \ldots, u_n . Then the set

(8.26)
$$\mathcal{E} \equiv \{ x \in \mathbf{R}^n \mid x'S^{-1}x = 1 \}$$

is the ellipsoid with principle axes $\sqrt{l_1}u_1, \ldots, \sqrt{l_n}u_n$.

Proof. (a) Apply the above results and the spectral decomposition (8.20).

(b) From (8.20), $S = UD_lU'$ with $U = (u_1 \cdots u_n)$, so $S^{-1} = UD_l^{-1}U'$ and,

$$\mathcal{E} = \{ x \in \mathbf{R}^n \mid (U'x)' D_l^{-1}(U'x) = 1 \}$$

= $U\{ y \in \mathbf{R}^n \mid y' D_l^{-1}y = 1 \}$ $(y = Ux)$
= $U\{ y \equiv (y_1, \dots, y_n)' \mid \frac{y_1^2}{l_1} + \dots + \frac{y_n^2}{l_n} = 1 \}$
= $U\mathcal{E}_0.$

But \mathcal{E}_0 is the ellipsoid with principle axes $\sqrt{l_1}\mathbf{u}_1, \ldots, \sqrt{l_n}\mathbf{u}_n$ (recall (8.19)) and $U\mathbf{u}_i = u_i$, so \mathcal{E} is the ellipsoid with principle axes $\sqrt{l_1}u_1, \ldots, \sqrt{l_n}u_n$.

Square root of a pd matrix. Let S be an $n \times n$ pd matrix. Any $n \times n$ matrix A such that AA' = S is called a square root of S, denoted by $S^{\frac{1}{2}}$. From the spectral decomposition $S = UD_lU'$, one version of $S^{\frac{1}{2}}$ is

(8.27)
$$S^{\frac{1}{2}} = U \operatorname{diag}(l_1^{\frac{1}{2}}, \dots, l_n^{\frac{1}{2}}) U' \equiv U D_l^{\frac{1}{2}} U'.$$

Note that

(8.28)
$$|S^{\frac{1}{2}}| = |S|^{\frac{1}{2}}.$$

Partitioned pd matrix. Partition the pd matrix $S: n \times n$ as

(8.29)
$$S = \frac{n_1}{n_2} \begin{pmatrix} n_1 & n_2 \\ S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix},$$

where
$$n_1 + n_2 = n$$
. Then both S_{11} and S_{22} are symmetric pd [why?],
 $S_{12} = S'_{21}$, and [verify!]
(8.30)
 $\begin{pmatrix} I_{n_1} & -S_{12}S_{22}^{-1} \\ 0 & I_{n_2} \end{pmatrix} \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} I_{n_1} & 0 \\ -S_{22}^{-1}S_{21} & I_{n_2} \end{pmatrix} = \begin{pmatrix} S_{11\cdot 2} & 0 \\ 0 & S_{22} \end{pmatrix},$

where

(8.31)
$$S_{11\cdot 2} \equiv S_{11} - S_{12}S_{22}^{-1}S_{21}$$

is necessarily pd [why?] This in turn implies the two fundamental identities

$$(8.32) \quad \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \begin{pmatrix} I_{n_1} & S_{12}S_{22}^{-1} \\ 0 & I_{n_2} \end{pmatrix} \begin{pmatrix} S_{11\cdot 2} & 0 \\ 0 & S_{22} \end{pmatrix} \begin{pmatrix} I_{n_1} & 0 \\ S_{22}^{-1}S_{21} & I_{n_2} \end{pmatrix},$$

$$(8.33) \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}^{-1} = \begin{pmatrix} I_{n_1} & 0 \\ -S_{22}^{-1}S_{21} & I_{n_2} \end{pmatrix} \begin{pmatrix} S_{11\cdot 2}^{-1} & 0 \\ 0 & S_{22}^{-1} \end{pmatrix} \begin{pmatrix} I_{n_1} & -S_{12}S_{22}^{-1} \\ 0 & I_{n_2} \end{pmatrix},$$

The following three consequences of (8.32) and (8.33) are immediate:

(8.34) S is $pd \iff S_{11\cdot 2}$ and S_{22} are $pd \iff S_{22\cdot 1}$ and S_{11} are pd;

(8.35)
$$|S| = |S_{11\cdot 2}| \cdot |S_{22}| = |S_{22\cdot 1}| \cdot |S_1|;$$

for $x \equiv \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbf{R}^n$, the quadratic form $x'S^{-1}x$ can be decomposed as

(8.36)
$$x'S^{-1}x = (x_1 - S_{12}S_{22}^{-1}x_2)'S_{11\cdot 2}^{-1}(x_1 - S_{12}S_{22}^{-1}x_2) + x_2'S_{22}^{-1}x_2.$$

Projection matrix. An $n \times n$ matrix P is a projection matrix if it is symmetric and *idempotent*: $P^2 = P$.

Fact 8.5. *P* is a projection matrix iff it has the form

(8.37)
$$P = U \begin{pmatrix} I_m & 0\\ 0 & 0 \end{pmatrix} U'$$

for some orthogonal matrix $U : n \times n$ and some $m \leq n$. In this case, $\operatorname{rank}(P) = m = \operatorname{tr}(P)$.

Proof. Since P is symmetric, $P = UD_lU'$ by its spectral decomposition. But the idempotence of P implies that each $l_i = 0$ or 1. (A permutation of the rows and columns, which is also an orthogonal transformation, may be necessary to obtain the form (8.37).)

Interpretation of (8.37): Partition U as

(8.38) m n-m $U=n (U_1 U_2),$ so (8.37) becomes

(8.39)
$$P = U_1 U_1'$$

But U is orthogonal so U'U = I, hence

(8.40)
$$\begin{pmatrix} I_m & 0\\ 0 & I_{n-m} \end{pmatrix} = U'U = \begin{pmatrix} U'_1U_1 & U'_1U_2\\ U'_2U_1 & U'_2U_2 \end{pmatrix}.$$

Thus from (8.39) and (8.40),

$$PU_1 = (U_1U'_1) U_1 = U_1,$$

$$PU_2 = (U_1U'_1) U_2 = 0.$$

This shows that P represents the linear transformation that projects \mathbb{R}^n orthogonally onto the column space of U_1 , which has dimension $m = \operatorname{tr}(P)$.

Furthermore, I - P is also symmetric and idempotent [verify] with $\operatorname{rank}(I - P) = n - m$. In fact,

$$I - P = UU' - P = (U_1U_1' + U_2U_2') - U_1U_1' = U_2U_2',$$

so I - P represents the linear transformation that projects \mathbb{R}^n orthogonally onto the column space of U_2 ; the dimension of this space is $n-m = \operatorname{tr}(I-P)$.

Note that the column spaces of U_1 and U_2 are perpendicular, since $U'_1U_2 = 0$. Equivalently, P(I - P) = (I - P)P = 0, i.e., applying P and I - P successively sends any $x \in \mathbf{R}^n$ to 0.

8.2. Random vectors and covariance matrices. Let $X \equiv \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ be

a rvtr in \mathbb{R}^n . The *expected value* of X is the vector

$$\mathbf{E}(X) \equiv \begin{pmatrix} \mathbf{E}(X_1) \\ \vdots \\ \mathbf{E}(X_n) \end{pmatrix}.$$

which is the center of gravity of the probability distribution of X in \mathbb{R}^n . Note that expectation is linear: for rvtrs X, Y and constant matrices A, B,

(8.41)
$$\operatorname{E}(AX + BY) = A \operatorname{E}(X) + B \operatorname{E}(Y).$$

Similarly, if
$$Z \equiv \begin{pmatrix} Z_{11} & \cdots & Z_{1n} \\ \vdots & & \vdots \\ Z_{m1} & \cdots & Z_{mn} \end{pmatrix}$$
 is a random matrix in $\mathbf{R}^{m \times n}$, $\mathbf{E}(Z)$ is also defined component-wise:

$$\mathbf{E}(Z) = \begin{pmatrix} \mathbf{E}(Z_{11}) & \cdots & \mathbf{E}(Z_{1n}) \\ \vdots & & \vdots \\ \mathbf{E}(Z_{m1}) & \cdots & \mathbf{E}(Z_{mn}) \end{pmatrix}.$$

Then for constant matrices $A: k \times m$ and $B: n \times p$,

(8.42)
$$E(AZB) = A E(Z) B.$$

The covariance matrix of X (\equiv the variance-covariance matrix), is

$$\operatorname{Cov}(X) = \operatorname{E} \left[(X - \operatorname{E} X)(X - \operatorname{E} X)' \right]$$
$$= \begin{pmatrix} \operatorname{Var}(X_1) & \operatorname{Cov}(X_1, X_2) & \cdots & \operatorname{Cov}(X_1, X_n) \\ \operatorname{Cov}(X_2, X_1) & \operatorname{Var}(X_2) & \cdots & \operatorname{Cov}(X_2, X_n) \\ \vdots & \vdots & & \vdots \\ \operatorname{Cov}(X_n, X_1) & \operatorname{Cov}(X_n, X_2) & \cdots & \operatorname{Var}(X_n) \end{pmatrix}.$$

The following formulas are essential: for $X : n \times 1$, $A : m \times n$, $a : n \times 1$,

(8.43)
$$\operatorname{Cov}(X) = \operatorname{E}(XX') - (\operatorname{E}X)(\operatorname{E}X)';$$

(8.44)
$$\operatorname{Cov}(AX + b) = A\operatorname{Cov}(X)A';$$

(8.45)
$$\operatorname{Var}(a'X+b) = a'\operatorname{Cov}(X)a.$$

Fact 8.6. Let $X \equiv (X_1, \ldots, X_n)'$ be a rotr in \mathbb{R}^n .

(a) $\operatorname{Cov}(X)$ is psd.

(b) $\operatorname{Cov}(X)$ is pd unless \exists a nonzero $a \equiv (a_1, \ldots, a_n)' \in \mathbf{R}^n$ s.t. the linear combination

$$a'X \equiv a_1X_1 + \dots + a_nX_n = \text{const.}$$

In this case the support of X is contained in a hyperplane of dimension $\leq n-1$.

Proof. (a) This follows directly from (8.24) and (8.45).

(b) If Cov(X) is not pd, then \exists a nonzero $a \in \mathbb{R}^n$ s.t.

$$0 = a' \operatorname{Cov}(X) a = \operatorname{Var}(a'X).$$

But this implies that a'X = const.

For rvtrs $X: m \times 1$ and $Y: n \times 1$, define

 $\operatorname{Cov}(X,Y) = \operatorname{E}\left[(X - \operatorname{E} X)(Y - \operatorname{E} Y)'\right]$

$$= \begin{pmatrix} \operatorname{Cov}(X_1, Y_1) & \operatorname{Cov}(X_1, Y_2) & \cdots & \operatorname{Cov}(X_1, Y_n) \\ \operatorname{Cov}(X_2, Y_1) & \operatorname{Cov}(X_2, Y_2) & \cdots & \operatorname{Cov}(X_2, Y_n) \\ \vdots & \vdots & & \vdots \\ \operatorname{Cov}(X_m, Y_1) & \operatorname{Cov}(X_m, Y_2) & \cdots & \operatorname{Cov}(X_m, Y_n) \end{pmatrix} : m \times n.$$

Clearly Cov(X, Y) = [Cov(Y, X)]'. Thus, if m = n then [verify]

(8.46)
$$\operatorname{Cov}(X \pm Y) = \operatorname{Cov}(X) + \operatorname{Cov}(Y) \pm \operatorname{Cov}(X, Y) \pm \operatorname{Cov}(Y, X).$$

and [verify]

(8.47)
$$\begin{array}{l} X \perp Y \Rightarrow \operatorname{Cov}(X,Y) = 0 \\ \Rightarrow \operatorname{Cov}(X \pm Y) = \operatorname{Cov}(X) + \operatorname{Cov}(Y). \end{array}$$

Variance of sample average (sample mean) of rvtrs: Let X_1, \ldots, X_n be i.i.d. rvtrs in \mathbb{R}^p , each with mean vector μ and covariance matrix Σ . Set

$$\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n).$$

Then $E(\bar{X}_n) = \mu$ and, by (8.47),

(8.48)
$$\operatorname{Cov}(\bar{X}_n) = \frac{1}{n^2} \operatorname{Cov}(X_1 + \dots + X_n) = \frac{1}{n} \Sigma.$$

Exercise 8.1. Verify the Weak Law of Large Numbers (WLLN) for rvtrs: \bar{X}_n converges to μ in probability $(X_n \xrightarrow{p} \mu)$, that is, for each $\epsilon > 0$,

$$P[||X_n - \mu|| \le \epsilon] \to 1 \quad as \ n \to \infty.$$

Example 8.1a. (Extension of (3.18) to identically distributed but correlated rvs.) Let X_1, \ldots, X_n be rvs with common mean μ and variance σ^2 . Suppose they are *equicorrelated*, i.e., $\operatorname{Cor}(X_i, X_j) = \rho \ \forall i \neq j$. Let

(8.49)
$$\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n), \qquad s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2,$$

the sample mean and sample variance, respectively. Then

(8.50)
$$E(\bar{X}_n) = \mu \qquad (\text{so } \bar{X}_n \text{ is } unbiased \text{ for } \mu); .$$

$$Var(\bar{X}_n) = \frac{1}{n^2} Var(X_1 + \dots + X_n)$$

$$= \frac{1}{n^2} [n\sigma^2 + n(n-1)\rho\sigma^2] \qquad [\text{why?}]$$

$$(8.51) \qquad \qquad = \frac{\sigma^2}{n} [1 + (n-1)\rho].$$

When X_1, \ldots, X_n are uncorrelated $(\rho = 0)$, in particular when they are independent, then (8.51) reduces to $\frac{\sigma^2}{n}$, which $\rightarrow 0$ as $n \rightarrow \infty$. When $\rho \neq 0$, however, $\operatorname{Var}(\bar{X}_n) \rightarrow \sigma^2 \rho \neq 0$, so the WLLN fails for equicorrelated *i.d. rvs.* Also, (8.51) imposes the constraint

(8.52)
$$-\frac{1}{n-1} \le \rho \le 1.$$

Next, using (8.51),

$$E(s_n^2) = \left(\frac{1}{n-1}\right) E\left(\sum_{i=1}^n X_i^2 - n(\bar{X}_n)^2\right) \\ = \left(\frac{1}{n-1}\right) \left[n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n}[1 + (n-1)\rho] + \mu^2\right)\right] \\ (8.53) = (1-\rho)\sigma^2.$$

Thus s_n^2 is unbiased for σ_n^2 if $\rho = 0$ but not otherwise!

Example 8.1b. We now re-derive (8.51) and (8.53) via covariance matrices, using properties (8.44) and (8.45). Set $X = (X_1, \ldots, X_n)'$, so

$$\operatorname{Cov}(X) = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{pmatrix}$$

(8.55)
$$\equiv \sigma^2 [(1-\rho)I_n + \rho \mathbf{e}_n \mathbf{e}'_n].$$

Then $\overline{X}_n = \frac{1}{n} \mathbf{e}'_n X$, so by (8.45),

$$\operatorname{Var}(\bar{X}_n) = \frac{\sigma^2}{n^2} \mathbf{e}'_n [(1-\rho)I_n + \rho \, \mathbf{e}_n \mathbf{e}'_n] \mathbf{e}_n$$
$$= \frac{\sigma^2}{n^2} [(1-\rho)n + \rho n^2] \qquad [\text{since } \mathbf{e}'_n \mathbf{e}_n = n]$$
$$= \frac{\sigma^2}{n} [1 + (n-1)\rho],$$

 \Box

which agrees with (8.51). Next, to find $E(s_n^2)$, write

(8.56)

$$\sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = \sum_{i=1}^{n} X_i^2 - n(\bar{X}_n)^2$$
$$= X'X - \frac{1}{n} (\mathbf{e}'_n X)^2$$
$$= X'X - \frac{1}{n} (X'\mathbf{e}_n) (\mathbf{e}'_n X)$$
$$\equiv X' (I_n - \mathbf{e}_n \mathbf{e}_n') X$$
$$\equiv X'QX,$$

where $\vec{\mathbf{e}}_n \equiv \left(\frac{\mathbf{e}_n}{\sqrt{n}}\right)$ is a unit vector, $P \equiv \vec{\mathbf{e}}_n \vec{\mathbf{e}}_n'$ is the projection matrix of rank $1 \equiv \operatorname{tr}(\vec{\mathbf{e}}_n \vec{\mathbf{e}}_n')$ that projects \mathbf{R}^n orthogonally onto the 1-dimensional subspace spanned by \mathbf{e}_n , and $Q \equiv I_n - \vec{\mathbf{e}}_n \vec{\mathbf{e}}_n'$ is the projection matrix of rank $n-1 \equiv \operatorname{tr} Q$ that projects \mathbf{R}^n orthogonally onto the (n-1)-dimensional subspace \mathbf{e}_n^{\perp} (see figure). Now complete the following exercise:

Exercise 8.2. Prove Fact 8.7 below, and use it to show that

(8.57)
$$E(X'QX) = (n-1)(1-\rho)\sigma^2,$$

which is equivalent to (8.53).

Fact 8.7. Let $X : n \times 1$ be a rvtr with $E(X) = \theta$ and $Cov(X) = \Sigma$. Then for any $n \times n$ symmetric matrix A,

(8.58)
$$E(X'AX) = tr(A\Sigma) + \theta'A\theta.$$

(This generalizes the relation $E(X^2) = Var(X) + (EX)^2$.)

Exercise 8.3. Show that $\operatorname{Cov}(X) \equiv \sigma^2[(1-\rho)I_n + \rho \mathbf{e}_n \mathbf{e}'_n]$ in (8.55) has one eigenvalue $= \sigma^2[1+(n-1)\rho]$ with eigenvector \mathbf{e}_n , and n-1 eigenvalues $= \sigma^2(1-\rho)$.

Example 8.1c. Eqn. (8.53) also can be obtained from the properties of the projection matrix Q. First note that [verify]

Define

(8.60)
$$Y \equiv \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = QX : n \times 1,$$

 \mathbf{SO}

(8.61)
$$\operatorname{E}(Y) = Q \operatorname{E}(X) = \mu Q \mathbf{e}_n = 0,$$

(8.62)
$$\operatorname{Cov}(Y) = \sigma^2 Q[(1-\rho)I_n + \rho \mathbf{e}_n \mathbf{e}'_n]Q$$
$$= \sigma^2 (1-\rho)Q.$$

Thus, since Q is idempotent $(Q^2 = Q)$,

$$E(X'QX) = E(Y'Y) = E[tr(Y'Y)]$$

= $E[tr(YY')] = tr[E(YY')] = tr[Cov(Y)]$
= $\sigma^2(1-\rho)tr(Q) = \sigma^2(1-\rho)(n-1),$

which again is equivalent to (8.53).

8.3. The multivariate normal distribution. [This section will revisit some results from §4, 5, 6.] As in Example 3.5, first let Z_1, \ldots, Z_n be i.i.d. standard normal N(0, 1) rvs. Then the rvtr $Z \equiv (Z_1, \ldots, Z_n)'$ has pdf

(8.63)
$$f_Z(z) = \prod_{i=1}^n \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}z_i^2} \\ = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}z'z},$$

where $z = (z_1, \ldots, z_n)'$. This is an extension of the univariate standard normal pdf to \mathbf{R}^n with E Z = 0 and $\operatorname{Cov} Z = I_n$, so we write $Z \sim N_n(0, I_n)$.

Now let $X = AZ + \mu$, with $A : n \times n$ nonsingular! and $\mu : n \times 1$. This is a linear (actually, affine) transformation with inverse given by

$$Z = A^{-1}(X - \mu)$$

and Jacobian $|A|^+$, so by the Jacobian method for transformations,

(8.64)

$$f_X(x) = \frac{1}{(2\pi)^{n/2} |A|^+} e^{-\frac{1}{2}(x-\mu)'(A^{-1})'A^{-1}(x-\mu)}$$

$$= \frac{1}{(2\pi)^{n/2} |AA'|^{1/2}} e^{-\frac{1}{2}(x-\mu)'(AA')^{-1}(x-\mu)}$$

$$= \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)},$$

which depends on (μ, A) only through [verify!]

$$(\mu, \Sigma) \equiv (\mu, AA') = (E X, Cov X).$$

Thus we write

$$X \sim N_n(\mu, \Sigma),$$

the (nonsingular, *n*-dimensional) multivariate normal distribution (MVND) with mean vector μ and covariance matrix Σ . (Note that $\Sigma \equiv AA'$ is pd.)

We can use the representation $X = AZ + \mu$ to derive the basic linearity property of the nonsingular MVND: **8.3.1. Linearity of** $N_n(\mu, \Sigma)$: If $X \sim N_n(\mu, \Sigma)$ then for $C : n \times n$ and $d : n \times 1$ with C nonsingular,

(8.65)

$$Y \equiv CX + d = (CA)Z + (C\mu + d)$$

$$\sim N_n (C\mu + d, (CA)(CA)')$$

$$= N_n (C\mu + d, C\Sigma C').$$

Remark 8.1. It is important to remember that the *general* (possibly singular) MVND was already defined in Example 3.5 via its moment generating function, where it was shown (recall (3.59)) that the linearity property (8.65) holds for any $C : m \times n$. Thus the following results are valid for general MVNDs.

8.3.2. Marginal distributions are normal. Let

(8.66)
$$n_1 \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_{n_1+n_2} \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right).$$

Then by the linearity property (8.65) (actually (3.59)) with $C = (I_{n_1} \quad 0)$,

(8.67)
$$X_1 = (I_{n_1} \quad 0) \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_{n_1} (\mu_1, \Sigma_{11}),$$

and similarly

(8.68)
$$X_2 = \begin{pmatrix} 0 & I_{n_2} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_{n_2} (\mu_2, \Sigma_{22}).$$

8.3.3. Linear combinations are normal. If X_1 , X_2 satisfy (8.66) then for $A_1 : m \times n_1$ and $A_2 : m \times n_2$, (8.65) (actually (3.59)) implies that

$$A_1 X_1 + A_2 X_2 = (A_1 \quad A_2) \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

(8.69) ~ $N_m (A_1 \mu_1 + A_2 \mu_2, A_1 \Sigma_{11} A'_1 + A_1 \Sigma_{12} A'_2 + A_2 \Sigma_{21} A'_1 + A_2 \Sigma_{22} A'_2).$

8.3.4. Independence \iff Uncorrelation. If X_1, X_2 satisfy (8.66) then

(8.70)
$$X_1 \perp \perp X_2 \iff \operatorname{Cov}(X_1, X_2) \equiv \Sigma_{12} = 0.$$

Proof. \Rightarrow is obvious. Next, suppose that $\Sigma_{12} = 0$. Then \Leftarrow is established for a general MVND via its mgf (recall (3.58)):

$$m_{X_1,X_2}(t_1,t_2) = e^{\binom{t_1}{t_2}'\binom{\mu_1}{\mu_2}} e^{\binom{t_1}{t_2}'\binom{\Sigma_{11}}{0} \cdot \binom{0}{\Sigma_{22}}\binom{t_1}{t_2}/2$$
$$= e^{t'_1\mu_1}e^{t'_1\Sigma_{11}t_1/2} \cdot e^{t'_2\mu_2}e^{t'_2\Sigma_{22}t_2/2}$$
$$= m_{X_1}(t_1) \cdot m_{X_2}(t_2).$$

Since this mgf factors, $X_1 \perp \perp X_2$ [by (8.67), (8.68), and the uniqueness property of mgfs - verify!]

Exercise 8.4. Prove (8.70) for a *nonsingular* MVND using the pdf (8.64).

Exercise 8.5. Extend (8.70) as follows. If $X \sim N_n(\mu, \Sigma)$ then for any two matrices $A: l \times n$ and $B: m \times n$,

(8.71) $AX \perp BX \iff A\Sigma B' = 0.$ Hint: Consider $\begin{pmatrix} A \\ B \end{pmatrix} X$ and apply (8.65).

8.3.5. Conditional distributions are normal. If X_1 , X_2 satisfy (8.66) and Σ_{22} is pd (in particular if Σ itself is pd), then

(8.72)
$$X_1 \mid X_2 \sim N_{n_1} \left(\mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (X_2 - \mu_2), \Sigma_{11 \cdot 2} \right).$$

Proof. Again apply linearity ((8.65), actually (3.59)) and the identity (8.32):

$$\begin{pmatrix} X_1 - \Sigma_{12} \Sigma_{22}^{-1} X_2 \\ X_2 \end{pmatrix}$$

$$= \begin{pmatrix} I_{n_1} & -\Sigma_{12} \Sigma_{22}^{-1} \\ 0 & I_{n_2} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

$$\sim N_{n_1+n_2} \begin{pmatrix} I_{n_1} & -\Sigma_{12} \Sigma_{22}^{-1} \\ 0 & I_{n_2} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},$$

$$\begin{pmatrix} I_{n_1} & -\Sigma_{12} \Sigma_{22}^{-1} \\ 0 & I_{n_2} \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} I_{n_1} & -\Sigma_{12} \Sigma_{22}^{-1} \\ 0 & I_{n_2} \end{pmatrix}' \end{pmatrix}$$

$$= N_{n_1+n_2} \begin{pmatrix} \mu_1 - \Sigma_{12} \Sigma_{22}^{-1} \mu_2 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11 \cdot 2} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \end{pmatrix}.$$

By (8.67) and (8.70), this implies that

(8.73)
$$X_1 - \Sigma_{12} \Sigma_{22}^{-1} X_2 \sim N_{n_1} \left(\mu_1 - \Sigma_{11} \Sigma_{22}^{-1} \mu_2, \Sigma_{11\cdot 2} \right)$$

and

(8.74)
$$X_1 - \Sigma_{12} \Sigma_{22}^{-1} X_2 \perp \perp X_2.$$

Thus (8.73) holds conditionally:

(8.75)
$$X_1 - \Sigma_{12} \Sigma_{22}^{-1} X_2 \mid X_2 \sim N_{n_1} \left(\mu_1 - \Sigma_{11} \Sigma_{22}^{-1} \mu_2, \Sigma_{11 \cdot 2} \right),$$

which implies (8.72) [verify].

Exercise 8.6. Prove (8.70) for a nonsingular MVND using the pdf (8.64). That is, apply the formula $f(x_1|x_2) = \frac{f(x_1,x_2)}{f(x_2)}$.

 \Box

Hint: Apply (8.35) and (8.36) for Σ .

8.3.6. Regression is linear, covariance is constant (\equiv homoscedastic). It follows immediately from (8.72) that

(8.76)
$$E(X_1 \mid X_2) = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (X_2 - \mu_2),$$

(8.77) $\operatorname{Cov}(X_1 \mid X_2) = \Sigma_{11 \cdot 2}.$

(Compare (8.76) with (5.38) in Remark 5.4.)

8.3.7. The bivariate case. Take $n_1 = n_2 = 1$ and consider

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2 \begin{bmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \end{bmatrix}$$

$$(8.78) \equiv N_2 \begin{bmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix} \end{bmatrix}.$$

Note that

(8.79)
$$\Sigma \equiv \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix},$$
so

(8.80)
$$|\Sigma| = \sigma_1^2 \sigma_2^2 (1 - \rho^2),$$

(8.81) $\Sigma^{-1} = \frac{1}{1 - \rho^2} \begin{pmatrix} \sigma_1^{-1} & 0\\ 0 & \sigma_2^{-1} \end{pmatrix} \begin{pmatrix} 1 & -\rho\\ -\rho & 1 \end{pmatrix} \begin{pmatrix} \sigma_1^{-1} & 0\\ 0 & \sigma_2^{-1} \end{pmatrix}$

Thus from (8.64) the pdf of (X_1, X_2) is

(8.82)
$$\frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2}}e^{-\frac{1}{2(1-\rho^2)}\left[(\frac{x_1-\mu_1}{\sigma_1})^2+(\frac{x_2-\mu_2}{\sigma_2})^2-2\rho(\frac{x_1-\mu_1}{\sigma_1})(\frac{x_2-\mu_2}{\sigma_2})\right]},$$

while (8.76) and (8.77) become

(8.83)
$$E(X_1 \mid X_2) = \mu_1 + \rho\left(\frac{\sigma_1}{\sigma_2}\right)(X_2 - \mu_2),$$

(8.84) $\operatorname{Var}(X_1 \mid X_2) = (1 - \rho^2) \sigma_1^2.$

Note that (8.83) agrees with the best linear predictor \hat{X}_1 given by (5.21), and (8.84) agrees with (5.27), the variance of the linear residual $X_1 - \hat{X}_1$.

Remark 8.2. In the special case where $\sigma_1^2 = \sigma_2^2 \equiv \sigma^2$ then

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]$$

$$\Rightarrow \begin{pmatrix} X_1 + X_2 \\ X_1 - X_2 \end{pmatrix} \equiv \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

$$\sim N_2 \left[\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \right]$$

$$\sim N_2 \left[\begin{pmatrix} \mu_1 + \mu_2 \\ \mu_1 - \mu_2 \end{pmatrix}, 2\sigma^2 \begin{pmatrix} 1 + \rho & 0 \\ 0 & 1 - \rho \end{pmatrix} \right].$$

In particular, this implies that $(X_1 + X_2) \perp (X_1 - X_2)$ in this case. (This extends CB Exercise 4.27, where it was assumed that $\rho = 0$.)

Exercise 8.7. (cf. Example 5.1.) Define the joint distribution of X, Y via the hierarchy

$$Y \mid X \sim N_1(\beta X, \tau^2),$$
$$X \sim N_1(0, \sigma^2).$$

Show that the joint distribution of X, Y is $N_2(\mu, \Sigma)$ and find μ and Σ . Find the conditional distribution of X|Y and the marginal distribution of Y. **Exercise 8.8.** Clearly X_1, \ldots, X_n i.i.d. $\Rightarrow X_1, \ldots, X_n$ are exchangeable.

(i) Find a trivariate normal example showing that \Leftarrow need not hold.

(ii) Clearly X_1, \ldots, X_n exchangeable $\Rightarrow X_1, \ldots, X_n$ are identically distributed. Find a trivariate normal example showing that X_1, \ldots, X_n identically distributed $\Rightarrow X_1, \ldots, X_n$ exchangeable. Show, however, that \Rightarrow holds for a bivariate normal distribution.

(ii) Find a non-normal bivariate example showing that X_1, X_2 identically distributed $\neq X_1, X_2$ exchangeable.

8.4. The MVND and the chi-square distribution.

In Remark 6.3, the *chi-square distribution* χ_n^2 with *n* degrees of freedom was defined to be the distribution of

$$Z_1^2 + \dots + Z_n^2 \equiv Z'Z \equiv ||Z||^2,$$

where $Z \equiv (Z_1, \ldots, Z_n)' \sim N_n(0, I_n)$. (That is, Z_1, \ldots, Z_n are i.i.d. standard N(0, 1) rvs.) Recall that

(8.85)
$$\chi_n^2 \sim \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right),$$

$$(8.86) E(\chi_n^2) = n$$

(8.87)
$$\operatorname{Var}(\chi_n^2) = 2n$$

Now consider $X \sim N_n(\mu, \Sigma)$ with Σ pd. Then

(8.88)
$$Z \equiv \Sigma^{-1/2} (X - \mu) \sim N_n(0, I_n),$$

(8.89)
$$(X-\mu)'\Sigma^{-1}(X-\mu) = Z'Z \sim \chi_n^2.$$

Suppose, however, that we omit Σ^{-1} in (8.89) and seek the distribution of $(X - \mu)'(X - \mu)$. Then this will *not* have a chi-square distribution in general. Instead, by the spectral decomposition $\Sigma = UD_{\lambda}U'$, (8.88) yields

(8.90)
$$(X - \mu)'(X - \mu) = Z'\Sigma Z = (U'Z)'D_{\lambda}(U'Z)$$
$$\equiv V'D_{\lambda}V = \lambda_1 V_1^2 + \cdots + \lambda_n V_n^2$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of Σ and $V \equiv U'Z \sim N_n(0, I_n)$. Thus the distribution of $(X - \mu)'(X - \mu)$ is a *positive linear combination of in*dependent χ_1^2 rvs, which is not (proportional to) a χ_n^2 rv. [Check via mgfs!]

8.4.1. Quadratic forms and projection matrices. Let $X \sim N_n(\xi, \sigma^2 I_n)$ and let P be an $n \times n$ projection matrix with rank $(P) = \operatorname{tr}(P) \equiv m$. Then the quadratic form determined by $X - \xi$ and P satisfies

(8.91) $(X - \xi)' P(X - \xi) \sim \sigma^2 \chi_m^2.$

Proof. By Fact 8.5, there exists an orthogonal matrix $U: n \times n$ s.t.

$$P = U \begin{pmatrix} I_m & 0\\ 0 & 0 \end{pmatrix} U'.$$

Then $Y \equiv U'(X - \xi) \sim N_n(0, \sigma^2 I_n)$, so with $Y = (Y_1, \dots, Y_n)'$,

$$(X - \xi)' P(X - \xi) = Y' \begin{pmatrix} I_m & 0\\ 0 & 0 \end{pmatrix} Y = Y_1^2 + \cdots Y_m^2 \sim \sigma^2 \chi_m^2.$$

8.4.2. Joint distribution of \bar{X}_n and s_n^2 . Let X_1, \ldots, X_n be a random (i.i.d) sample from the univariate normal distribution $N_1(\mu, \sigma^2)$ and let

$$\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n), \qquad s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

(Recall (3.18), also Example 8.1a,b,c.) Then:

(8.92) \bar{X}_n and s_n^2 are independent;

(8.93)
$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right);$$

(8.94) $s_n^2 \sim \frac{\sigma^2}{n-1} \chi_{n-1}^2.$

Proof. As in Example 8.1b, let $\vec{\mathbf{e}}_n = \frac{\mathbf{e}_n}{\sqrt{n}}$, $P = \vec{\mathbf{e}}_n \vec{\mathbf{e}_n'}$, $Q = I_n - \vec{\mathbf{e}}_n \vec{\mathbf{e}_n'}$, and

(8.95)
$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \sim N_n(\mu \mathbf{e}_n, \, \sigma^2 I_n).$$

Then $P(\sigma^2 I_n)Q' = \sigma^2 PQ = 0$, so $PX \perp QX$ by (8.71). But [recall figure]

$$\begin{pmatrix} \bar{X}_n \\ \vdots \\ \bar{X}_n \end{pmatrix} = \bar{X}_n \mathbf{e}_n = PX \sim N_n(\mu P \mathbf{e}_n, \, \sigma^2 P) = N_n\left(\mu \mathbf{e}_n, \, \frac{\sigma^2}{n} \mathbf{e}_n \mathbf{e}'_n\right),$$
$$\begin{pmatrix} X_1 - \bar{X}_n \\ \vdots \\ X_n - \bar{X}_n \end{pmatrix} = X - \bar{X}_n \mathbf{e}_n = QX,$$

so this implies (8.92) and (8.93) Finally,

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 = (QX)'(QX) = (X - \mu \mathbf{e}_n)'Q(X - \mu \mathbf{e}_n) \quad \text{[verify]}$$

 \Box

and rank(Q) = tr(Q) = n - 1, so (8.94) follows from (8.91).

Geometrical interpretation of (8.95): The i.i.d. normal model (8.95) is the simplest example of the *univariate linear model*. If we let L denote the 1-dimensional linear subspace spanned by \mathbf{e}_n , then (8.95) can be expressed as follows:

(8.96)
$$X \sim N_n(\xi, \sigma^2 I_n)$$
 with $\xi \in L$.

If we let $P \equiv P_L \equiv \vec{\mathbf{e}}_n \vec{\mathbf{e}}_n'$ denote projection onto L, then $Q \equiv Q_L = I_n - P_L$ is the projection onto the "residual" subspace L^{\perp} [recall Figure], and Pythagoras gives us the following *Analysis of Variance (ANOVA)*:

(8.97)

$$I_{n} = P_{L} + Q_{L},$$

$$X = P_{L}X + Q_{L}X,$$

$$\|X\|^{2} = \|P_{L}X\|^{2} + \|Q_{L}X\|^{2},$$

$$X_{1}^{2} + \dots + X_{n}^{2} = n(\bar{X}_{n})^{2} + \sum_{i=1}^{n} (X_{i} - \bar{X})^{2},$$
Total Sum of Squares = SS(L) + SS(L^{\perp}).

 $[SS(L^{\perp})$ is often called the "residual sum of squares".] Then we have seen that under the linear model (8.96),

(8.98)
$$\operatorname{SS}(L) \perp \operatorname{SS}(L^{\perp}) \text{ and } \operatorname{SS}(L^{\perp}) \sim \sigma^2 \chi^2_{\dim(L^{\perp})}.$$

Furthermore we will see in $\S8.4.3$ below that

(8.99)
$$||P_L X||^2 \equiv SS(L) \sim \sigma^2 \chi^2_{\dim(L)} \left(\frac{||P_L \xi||^2}{\sigma^2}\right),$$

a noncentral chi-square distribution with noncentrality parameter (§8.4.3). $\frac{\|P_L\xi\|^2}{\sigma^2}$. Note that for the model (8.95), $\xi = \mu \mathbf{e}_n$, so

$$P_L \xi = \mu P_L \mathbf{e}_n = \mu \mathbf{e}_n,$$
$$\frac{\|P_L\|^2}{\sigma^2} = \frac{n\mu^2}{\sigma^2} \ge 0,$$

so the noncentrality parameter = 0 iff $\mu = 0$. Thus the "null hypothesis" $\mu = 0$ can be tested by means of the *F*-statistic \equiv *F*-ratio

(8.100)
$$F \equiv \frac{\mathrm{SS}(L)}{\mathrm{SS}(L^{\perp})} \sim \frac{\chi^2_{\mathrm{dim}(L)} \left(\frac{\|P_L \xi\|^2}{\sigma^2}\right)}{\chi^2_{\mathrm{dim}(L^{\perp})}} = \frac{\chi^2_1 \left(\frac{n\mu^2}{\sigma^2}\right)}{\chi^2_{n-1}}$$

The null hypothesis $\mu = 0$ is rejected in favor of the alternative hypothesis $\mu \neq 0$ for sufficiently large values of F. [See §8.4.3 for the noncentral chisquare distribution, also Remark 8.4 and Exercise 18.27.]

Exercise 8.9. Extend (8.92) - (8.94) to the case where X_1, \ldots, X_n are equicorrelated, as in Examples 8.1a,b,c. That is, if as in Example 8.1b, $X \sim N_n(\mu \mathbf{e}_n, \sigma^2[(1-\rho)I_n + \rho \mathbf{e}_n \mathbf{e}'_n])$, show that

(8.101) $\bar{X}_n \text{ and } s_n^2 \text{ are independent};$

(8.102)
$$\bar{X}_n \sim N_1\left(\mu, \frac{\sigma^2}{n}[1+(n-1)\rho]\right);$$

(8.103)
$$s_n^2 \sim \frac{\sigma^2(1-\rho)}{n-1} \chi_{n-1}^2.$$

Hint: Follow the matrix formulation in Example 8.1b.

8.4.3. The noncentral chi-square distribution. Extend the results of §8.4 as follows: First let $Z \equiv (Z_1, \ldots, Z_n)' \sim N_n(\mu, I_n)$, where $\mu \equiv (\mu_1, \ldots, \mu_n)' \in \mathbf{R}^n$. The distribution of

$$Z_1^2 + \dots + Z_n^2 \equiv Z'Z \equiv ||Z||^2$$

is called the noncentral chi-square distribution with n degrees of freedom and noncentrality parameter $\|\mu\|^2$, denoted by $\chi_n^2(\|\mu\|^2)$. Note that as in §8.4, Z_1, \ldots, Z_n are independent, each with variance = 1, but now $E(Z_i) = \mu_i$.

To show that the distribution of $||Z||^2$ depends on μ only through its (squared) length $||\mu||^2$, choose⁸ an orthogonal (rotation) matrix $U : n \times n$ such that $U\mu = (||\mu||, 0, ..., 0)'$, i.e., U rotates μ into $(||\mu||, 0, ..., 0)'$, and set

$$Y = UZ \sim N_n(U\mu, UU') = N_n((\|\mu\|, 0, \dots, 0)', I_n)$$

Then the desired result follows since

$$||Z||^{2} = ||Y||^{2} \equiv Y_{1}^{2} + Y_{2}^{2} + \dots + Y_{n}^{2}$$

$$\sim [N_{1}(||\mu||, 1)]^{2} + [N_{1}(0, 1)]^{2} + \dots + [N_{1}(0, 1)]^{2}$$

$$\equiv \chi_{1}^{2}(||\mu||^{2}) + \chi_{1}^{2} + \dots + \chi_{1}^{2}$$

(8.104)

$$\equiv \chi_{1}^{2}(||\mu||^{2}) + \chi_{n-1}^{2},$$

where the chi-square variates in each line are mutually independent.

To find the pdf of $V \equiv Y_1^2 \sim \chi_1^2(\delta) \sim [N_1(\sqrt{\delta}, 1)]^2$, where $\delta = \|\mu\|^2$, recall the method of Example 2.3:

$$f_V(v) = \frac{d}{dv} P[-\sqrt{v} \le Y_1 \le \sqrt{v}]$$
$$= \frac{d}{dv} \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{v}}^{\sqrt{v}} e^{-\frac{1}{2}(t-\sqrt{\delta})^2} dt$$
$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{\delta}{2}} \frac{d}{dv} \int_{-\sqrt{v}}^{\sqrt{v}} e^{t\sqrt{\delta}} e^{-\frac{t^2}{2}} dt$$

⁸ Let the first row of U be $\vec{\mu} \equiv \frac{\mu}{\|\mu\|}$ and let the remaining n-1 rows be any orthonormal basis for L^{\perp} .

$$(8.105) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\delta}{2}} \frac{d}{dv} \int_{-\sqrt{v}}^{\sqrt{v}} \left[\sum_{k=0}^{\infty} \frac{t^k \delta^{\frac{k}{2}}}{k!} \right] e^{-\frac{t^2}{2}} dt$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{\delta}{2}} \sum_{k=0}^{\infty} \frac{\delta^{\frac{k}{2}}}{k!} \frac{d}{dv} \int_{-\sqrt{v}}^{\sqrt{v}} t^k e^{-\frac{t^2}{2}} dt$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{\delta}{2}} \sum_{k=0}^{\infty} \frac{\delta^k}{(2k)!} \frac{d}{dv} \int_{-\sqrt{v}}^{\sqrt{v}} t^{2k} e^{-\frac{t^2}{2}} dt \qquad [why?]$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{\delta}{2}} \sum_{k=0}^{\infty} \frac{\delta^k}{(2k)!} v^{k-\frac{1}{2}} e^{-\frac{v}{2}} \qquad [verify]$$

$$= \underbrace{e^{-\frac{\delta}{2}} \sum_{k=0}^{\infty} \frac{(\frac{\delta}{2})^k}{k!}}_{\text{Poisson}(\frac{\delta}{2})} \underbrace{\left[\frac{v^{\frac{1+2k}{2}-1}e^{-\frac{v}{2}}}{2^{\frac{1+2k}{2}}\Gamma(\frac{1+2k}{2})}\right]}_{\text{pdf of } \chi^2_{1+2k}} \cdot c_k,$$

where

$$c_k = \frac{2^k k! 2^{\frac{1+2k}{2}} \Gamma\left(\frac{1+2k}{2}\right)}{(2k)! \sqrt{2\pi}} = 1$$

by the Legendre "duplication formula" for the Gamma function! Thus we have represented the pdf of a $\chi_1^2(\delta)$ rv as a mixture (weighted average) of central chi-square pdfs with Poisson weights. This can be written as follows:

(8.106) $\chi_1^2(\delta) \mid K \sim \chi_{1+2K}^2$ where $K \sim \text{Poisson}(\delta/2)$.

(Compare this to the result of Example 2.4.) Thus by (8.104) this implies that $Z'Z \equiv ||Z||^2 \sim \chi_n^2(\delta)$ satisfies

(8.107) $\chi_n^2(\delta) \mid K \sim \chi_{n+2K}^2$ where $K \sim \text{Poisson}(\delta/2)$.

That is, the pdf of the noncentral chi-square $rv V \equiv \chi_n^2(\delta)$ is a $Poisson(\delta/2)$ mixture of the pdfs of central chi-square rvs with n + 2k d.f., k = 0, 1, ...

The representation (8.107) can be used to obtain the mean and variance of $\chi_n^2(\delta)$:

$$E[\chi_n^2(\delta)] = E\{E[\chi_{n+2K}^2 \mid K]\} = E(n+2K) = n+2(\delta/2) = n+\delta; Var[\chi_n^2(\delta)] = E[Var(\chi_{n+2K}^2 \mid K)] + Var[E(\chi_{n+2K}^2 \mid K)] = E[2(n+2K)] + Var(n+2K) = [2n+4(\delta/2)] + 4(\delta/2) = 2n+4\delta.$$

Exercise 8.10*. Show that the noncentral chi-square distribution $\chi_n^2(\delta)$ is stochastically increasing in both n and δ .

Next, consider $X \sim N_n(\mu, \Sigma)$ with a general pd Σ . Then

(8.110)
$$X'\Sigma^{-1}X = (\Sigma^{-\frac{1}{2}}X)'(\Sigma^{-\frac{1}{2}}X) \sim \chi_n^2(\mu'\Sigma^{-1}\mu),$$

since $Z \equiv \Sigma^{-\frac{1}{2}} X \sim N_n(\Sigma^{-\frac{1}{2}} \mu, I_n)$ and $\|\Sigma^{-\frac{1}{2}} \mu\|^2 = \mu' \Sigma^{-1} \mu$. Thus, by Exercise 8.10, the distribution of $X' \Sigma^{-1} X$ in (8.110) is stochastically increasing in $\mu' \Sigma^{-1} \mu$.

Finally, let $Y \sim N_n(\xi, \sigma^2 I_n)$ and let P be a projection matrix with $\operatorname{rank}(P) = m$. Then $P = U_1 U_1'$ where $U_1' U_1 = I_m$ (cf. (8.38) - (8.40)), so

$$||PY||^{2} = ||U_{1}U_{1}'Y||^{2} = (U_{1}U_{1}'Y)'(U_{1}U_{1}'Y) = Y'U_{1}U_{1}'Y = ||U_{1}'Y||^{2}.$$

But

$$U'_1 Y \sim N_m(U'_1\xi, \, \sigma^2 U'_1 U_1) = N_m(U'_1\xi, \, \sigma^2 I_m),$$

so by (8.110) with $X = U'_1 Y$, $\mu = U'_1 \xi$, and $\Sigma = \sigma^2 I_m$,

$$\frac{\|PY\|^2}{\sigma^2} = \frac{(U_1'Y)'(U_1'Y)}{\sigma^2} \sim \chi_m^2 \left(\frac{\xi'U_1U_1'\xi}{\sigma^2}\right) = \chi_m^2 \left(\frac{\|P\xi\|^2}{\sigma^2}\right),$$

 \mathbf{SO}

(8.111)
$$||PY||^2 \sim \sigma^2 \chi_m^2 \left(\frac{||P\xi||^2}{\sigma^2}\right).$$

Remark 8.4. By taking $P = P_L$ for an *m*-dimensional linear subspace *L* of \mathbb{R}^n , this confirms (8.98). Furthermore, under the general univariate linear model (8.96) it is assumed that $\xi \in L$, so $P_L \xi = \xi$ and $||P_L \xi||^2 = ||\xi||^2$. In view of Exercise 8.10, this shows why the *F*-ratio

(8.112)
$$F \equiv \frac{\mathrm{SS}(L)}{\mathrm{SS}(L^{\perp})} \equiv \frac{\|P_L Y\|^2}{\|Q_L Y\|^2} \sim \frac{\chi^2_{\mathrm{dim}(L)}\left(\frac{\|\xi\|^2}{\sigma^2}\right)}{\chi^2_{\mathrm{dim}(L^{\perp})}}$$

in (8.100) will tend to take larger values when $\xi \neq 0$ ($\xi \in L$) than when $\xi = 0$, hence why this *F*-ratio is a reasonable statistic for testing $\xi = 0$ vs. $\xi \neq 0$ ($\xi \in L$).

8.5. Further examples of univariate linear models.

As indicated by (8.96), the *univariate normal linear model* has the following form: observe

(8.113)
$$X \sim N_n(\xi, \sigma^2 I_n) \quad \text{with} \quad \xi \in L,$$

where L is a d-dimensional linear subspace of \mathbb{R}^n (0 < d < n). The components X_1, \ldots, X_n of X are independent⁹ normal rvs with common unknown variance σ^2 . Thus the model (8.113) imposes the linear constraint that $\xi \equiv \mathbb{E}(X) \in L$. Our goal is to estimate ξ and σ^2 subject to this constraint.

Let P_L denote projection onto L, so $Q_L \equiv I_n - P_L$ is the projection onto the "residual" subspace L^{\perp} (recall the figure in Example 8.1b). Then it can be shown (compare to §8.4.2) that

(8.114) $\hat{\xi} \equiv P_L X$ is the best linear unbiased estimator (BLUE) and maximum likelihood estimator (MLE) of ξ (see §14.1);

(8.115)
$$\hat{\sigma}^2 \equiv \frac{\|Q_L X\|^2}{n}$$
 is the MLE of σ^2 ; $\tilde{\sigma}^2 \equiv \frac{\|Q_L X\|^2}{n-d}$ is unbiased for σ^2 ;

(8.116)
$$(\tilde{\xi}, \tilde{\sigma}^2)$$
 is a complete sufficient statistic for (ξ, σ^2) (see §11, §12);

(8.117)
$$\xi$$
 and $\tilde{\sigma}^2$ are independent,

⁹ More generally, we may assume that $X \sim N_n(\xi, \sigma^2 \Sigma_0)$ where Σ_0 is a known p.d. matrix. This can be reduced to the form (8.113) by the transformation $Y = \Sigma_0^{-1/2} X$.

(8.118) $\hat{\xi} \sim N_n(\xi, \sigma^2 P_L);$ (8.119) $\tilde{\sigma}^2 \sim \frac{\sigma^2}{n-d} \chi^2_{n-d}.$

This leaves only one task: the algebraic problem of finding the matrices P_L and Q_L and thereby calculating $\hat{\xi}$ and $\tilde{\sigma}^2$. Typically the linear subspace L is specified as the subspace spanned by d linearly independent $n \times 1$ vectors $z_1, \ldots, z_d \in \mathbf{R}^n$. That is,

(8.120)
$$L = \{\beta_1 z_1 + \dots + \beta_d z_d \mid \beta_1, \dots \beta_d \in \mathbf{R}^1\} \\\equiv \{Z\beta \mid \beta : d \times 1 \in \mathbf{R}^d\},$$

where

$$Z = (z_1 \quad \cdots \quad z_d), \qquad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_d \end{pmatrix}$$

The matrix $Z : n \times d$, called the *design matrix*, is a matrix of full rank d, so Z'Z is nonsingular. The linear model (8.113) can now be written as

(8.121)
$$X \sim N_n(Z\beta, \sigma^2 I_n) \text{ with } \beta \in \mathbf{R}^d,$$

and our goal becomes that of estimating β and σ^2 .

For this we establish the following relation between P_L and Z:

(8.122)
$$P_L = Z(Z'Z)^{-1}Z'.$$

To see this, simply note that $Z(Z'Z)^{-1}Z'$ is symmetric and idempotent, $Z(Z'Z)^{-1}Z'(\mathbf{R}^n) \subseteq Z(\mathbf{R}^d) = L$, and

$$\operatorname{rank}[Z(Z'Z)^{-1}Z'] = \operatorname{tr}[Z(Z'Z)^{-1}Z'] = d,$$

so $Z(Z'Z)^{-1}Z'(\mathbf{R}^n) = L$ [why?], which establishes (8.122). Now by (8.114) and (8.122),

$$Z\hat{\beta} \equiv \hat{\xi} = Z(Z'Z)^{-1}Z'X,$$
$$Z'Z\hat{\beta} = Z'X,$$

 \mathbf{SO}

(8.123) $\hat{\beta} = (Z'Z)^{-1}Z'X.$

Finally,

$$Q_L X = (I_n - P_L) X = [I_n - Z(Z'Z)^{-1}Z']X,$$

 \mathbf{SO}

(8.124)
$$\tilde{\sigma}^2 = \frac{1}{n-d} \|Q_L X\|^2 = \frac{1}{n-d} X' [I_n - Z(Z'Z)^{-1}Z'] X.$$

An alternative expression for $\tilde{\sigma}^2$ is

(8.125)
$$\tilde{\sigma}^2 = \frac{1}{n-d} \|X - P_L X\|^2 = \frac{1}{n-d} \|X - \hat{\xi}\|^2.$$

It follows from (8.121) and (8.123) that

(8.126) $\hat{\beta} \sim N_d[\beta, \sigma^2 (Z'Z)^{-1}],$

 \mathbf{SO}

(8.127)
$$(\hat{\beta} - \beta)'(Z'Z)(\hat{\beta} - \beta) \sim \sigma^2 \chi_d^2.$$

Thus by (8.117), (8.119), and (8.127),

(8.128)
$$\frac{(\hat{\beta}-\beta)'(Z'Z)(\hat{\beta}-\beta)}{d\,\tilde{\sigma}^2} \sim F_{d,n-d},$$

from which a $(1 - \alpha)$ -confidence ellipsoid for β easily can be obtained:

(8.129)
$$(1-\alpha) = P\left[(\hat{\beta}-\beta)'(Z'Z)(\hat{\beta}-\beta) \le d\,\tilde{\sigma}^2 F_{d,n-d;\,\alpha}\right].$$

Example 8.2. (The one-sample model.) As in §8.4.2, let X_1, \ldots, X_n be a random (i.i.d) sample from the univariate normal distribution $N_1(\mu, \sigma^2)$, so that

$$X \equiv (X_1, \dots, X_n)' \sim N_n(\mu \mathbf{e}_n, \, \sigma^2 I_n)$$

satisfies (8.121) with $Z = \mathbf{e}_n, d = 1, \beta = \mu$. Then from (8.123) and (8.124),

$$\hat{\mu} = (\mathbf{e}'_{n}\mathbf{e}_{n})^{-1}\mathbf{e}'_{n}X = \frac{1}{n}\sum_{i=1}^{n}X_{i} = \bar{X}_{n},$$
$$\tilde{\sigma}^{2} = \frac{1}{n-1}X'[I_{n} - \mathbf{e}_{n}(\mathbf{e}'_{n}\mathbf{e}_{n})^{-1}\mathbf{e}'_{n}]X = \frac{1}{n-1}\sum_{i=1}^{n}(X_{i} - \bar{X})^{2} = s_{n}^{2},$$

the sample mean and sample variance as before. The $(1 - \alpha)$ -confidence ellipsoid (8.129) becomes the usual Student-*t* confidence interval

(8.130)
$$\hat{\mu} \pm \frac{s_n}{\sqrt{n}} t_{n-1;\,\alpha/2}.$$

Example 8.3. (Simple linear regression.) Let X_1, \ldots, X_n be independent rvs that depend linearly on known *regressor variables* t_1, \ldots, t_n , that is,

(8.131)
$$X_i = a + bt_i + \epsilon_i, \quad i = 1, ..., n,$$

where a and b are unknown parameters and $\epsilon_1, \ldots, \epsilon_n$ are i.i.d unobservable random errors with $\epsilon_i \sim N_1(0, \sigma^2)$ (σ^2 unknown). We can write (8.131) in the vector form (8.121) as follows:

$$\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

or equivalently,

$$(8.132) X = Z\beta + \epsilon$$

Since $\epsilon \sim N_n(0, \sigma^2 I_n)$, (8.126) is a special case of the univariate normal linear model (8.121). Here the design matrix Z is of full rank d = 2 iff (t_1, \ldots, t_n) and $(1, \ldots, 1)$ are linearly dependent, i.e., iff at least two t_i 's are different. (That is, we can't fit a straight line through $(t_1, X_1), \ldots, (t_n, X_n)$ if all the t_i 's are the same.) In this case

(8.133)
$$Z'Z = \begin{pmatrix} n & \sum t_i \\ \sum t_i & \sum t_i^2 \end{pmatrix},$$

so (8.123) and (8.125) become [verify]

$$\hat{\beta} \equiv \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} n & \sum t_i \\ \sum t_i & \sum t_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum X_i \\ \sum t_i X_i \end{pmatrix} = \begin{pmatrix} \bar{X}_n - \hat{b}\bar{t}_n \\ \frac{\sum (t_i - \bar{t}_n)(X_i - \bar{X}_n)}{\sum (t_i - \bar{t}_n)^2} \end{pmatrix},$$
$$\tilde{\sigma}^2 = \frac{1}{n-2} \sum [X_i - (\hat{a} + \hat{b}t_i)]^2 = \frac{1}{n-2} \sum [(X_i - \bar{X}_n) - \hat{b}(t_i - \bar{t}_n)]^2.$$

(The expressions for \hat{a} and \hat{b} should be compared to (5.17) and (5.18), where (X, Y) corresponds to (t, X).)

Note that \hat{b} can be expressed as

(8.134)
$$\hat{b} = \frac{\sum (t_i - \bar{t}_n)^2 \left(\frac{X_i - \bar{X}_n}{t_i - \bar{t}_n}\right)}{\sum (t_i - \bar{t}_n)^2},$$

a weighted average of the slopes $\frac{X_i - \bar{X}_n}{t_i - \bar{t}_n}$ determined by the individual data points (t_i, X_i) , where the weights are proportional to the squared distances $(t_i - \bar{t}_n)^2$. Furthermore, it follows from (8.126) and(8.133) that [verify]

(8.135)
$$\hat{b} \sim N_1 \left(b, \frac{\sigma^2}{\sum (t_i - \bar{t}_n)^2} \right),$$

so a $(1 - \alpha)$ -confidence interval for the slope b is given by [verify]

(8.136)
$$\hat{b} \pm \frac{\tilde{\sigma}}{\sqrt{\sum(t_i - \bar{t}_n)^2}} t_{n-2;\,\alpha/2}$$

Note that this confidence interval can be made narrower by increasing the dispersion of t_1, \ldots, t_n . In fact, the most accurate (narrowest) confidence interval for the slope b is obtained by placing half the t_i 's at each extreme of their range. (Of course, this precludes the possibility of detecting departures from the assumed linear regression model.)

Exercise 8.11. (Quadratic regression.) Replace the simple linear regression model (8.131) by the quadratic regression model

(8.137)
$$X_i = a + bt_i + ct_i^2 + \epsilon_i, \qquad i = 1, \dots, n,$$

where a, b, c are unknown parameters and the ϵ_i 's are as in Example 8.3. Then (8.137) can be written in the vector form (8.121) with

(8.138)
$$Z \equiv \begin{pmatrix} 1 & t_1 & t_1^2 \\ \vdots & \vdots & \vdots \\ 1 & t_n & t_n^2 \end{pmatrix}, \qquad \beta = \begin{pmatrix} a \\ b \\ c \end{pmatrix}.$$

Assume that the design matrix Z has full rank d = 3. (Note that although the regression is not linear in the (known) t_i 's, this qualifies as a linear model because $E(X) \equiv Z\beta$ is linear in the unknown parameters a, b, c.) Find the MLEs \hat{b} and \hat{c} , find $\tilde{\sigma}^2$, and find (individual) confidence intervals for b and c. Express your answers in terms of the t_i 's and the X_i 's.

Exercise 8.12 (The two-sample model.) Let X_1, \ldots, X_m and Y_1, \ldots, Y_n be independent random (i.i.d) samples from the univariate normal distributions $N_1(\mu, \sigma^2)$ and $N_1(\nu, \sigma^2)$, respectively, where $m \ge 1$ and $n \ge 1$. Express this as a univariate normal linear model (8.121) – what are Z, d, and β , and what if any additional condition(s) on m and n are needed for the existence of $\tilde{\sigma}^2$? Find the MLEs $\hat{\mu}$ and $\hat{\nu}$, find $\tilde{\sigma}^2$, and find a confidence interval for $\mu - \nu$. Express your answers in terms of the X_i 's and Y_j 's.

9. Order Statistics from a Continuous Univariate Distribution.

Let X_1, \ldots, X_n be an i.i.d. random sample from a *continuous* distribution with pdf $f_X(x)$ and cdf $F_X(x)$ on $(-\infty, \infty)$. The *order statistics* $Y_1 < Y_2 < \cdots < Y_n$ are the ordered values of X_1, \ldots, X_n . Often the notation $X_{(1)} < \cdots < X_{(n)}$ is used instead. Thus, for example,

$$Y_1 = X_{(1)} = \min(X_1, \dots, X_n),$$

 $Y_n = X_{(n)} = \max(X_1, \dots, X_n).$

The mapping

$$(X_1, \ldots, X_n) \to (Y_1 < Y_2 < \cdots < Y_n)$$

is not 1-1 but rather is (n!) - 1: Each of the n! permutations of (X_1, \ldots, X_n) is mapped onto the same value of $(Y_1 < Y_2 < \cdots < Y_n)$:

We now present approximate (but valid) derivations of the pdfs of:

- (i) a single order statistic Y_i ;
- (ii) a pair of order statistics Y_i, Y_j ;
- (iii) the entire set of order statistics Y_1, \ldots, Y_n .

(These derivations are based on the multinomial distribution).

(i) Approximately, $f_{Y_i}(y)dy \approx P[y < Y_i < y + dy]$:

But the event $\{y < Y_i < y + dy\}$ is approximately the same as the event

$$\{(i-1) \ X' s \in (-\infty, y), \ 1 \ X \in (y, \ y+dy), \ (n-i) \ X' s \in (y+dy, \ \infty).\}$$

The probability of this event is approximated by the trinomial probability

$$\frac{n!}{(i-1)!\,1!\,(n-i)!}\,[F_X(y)]^{i-1}\,[f_X(y)dy]^1\,[1-F_X(y)]^{n-i}\,.$$

Thus, cancelling the dy's we find that

(9.1)
$$f_{Y_i}(y) = \frac{n!}{(i-1)!(n-i)!} \left[F_X(y)\right]^{i-1} \left[1 - F_X(y)\right]^{n-i} f_X(y).$$

(ii) Similarly, for i < j and y < z,

$$f_{Y_{i}, Y_{j}}(y, z)dydz \approx P[y < Y_{i} < y + dy, z < Y_{j} < z + dz] \approx P[(i - 1) X's \in (-\infty, y), 1X \in (y, y + dy), (j - i - 1) X's \in (y + dy, z), 1X \in (z, z + dz), (n - j) X's \in (z + dz, \infty)] \qquad [see figure] \approx \frac{n!}{(i - 1)! 1! (j - i - 1)! 1! (n - j)!} [F_{X}(y)]^{i - 1} [f_{X}(y)dy]^{1} \cdot [F_{X}(z) - F_{X}(y)]^{j - i - 1} [f_{X}(z)dz]^{1} [1 - F_{X}(z)]^{n - j}.$$

Thus, cancelling dydz we obtain

$$f_{Y_i,Y_j}(y,z) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} [F_X(y)]^{i-1} [F_X(z) - F_X(y)]^{j-i-1}$$

$$(9.2) \qquad \cdot [1 - F_X(z)]^{n-j} f_X(y) f_X(z).$$

(iii) Finally, for $y_1 < \cdots < y_n$,

$$f_{Y_{1},...,Y_{n}}(y_{1},...,y_{n})dy_{1}\cdots dy_{n}$$

$$\approx P[y_{1} < Y_{1} < y_{1} + dy_{1},...,y_{n} < Y_{n} < y_{n} + dy_{n}]$$

$$\approx \frac{n!}{1!\cdots 1!} [f_{X}(y_{1})dy_{1}]^{1}\cdots [f_{X}(y_{n})dy_{n}]^{1}.$$

Thus, cancelling $dy_1 \cdots dy_n$ we obtain

(9.3)
$$f_{Y_1,...,Y_n}(y_1,...,y_n) = n! f_X(y_1) \cdots f_X(y_n).$$

The factor n! occurs because $(X_1, \ldots, X_n) \rightarrow (Y_1 < Y_2 < \cdots < Y_n)$ is an (n!) - 1 mapping.

Exercise 9.1. Extend (9.3) to the case of exchangeable rvs: Let $X \equiv (X_1, \ldots, X_n)$ have pdf $f_X(x_1, \ldots, x_n)$ with f_X symmetric \equiv permutation-invariant, that is,

(9.4)
$$f_X(x_1,\ldots,x_n) = f_X(x_{\pi(1)},\ldots,x_{\pi(n)})$$
 for all permutations π .

Show that the order statistics $Y_1 < \cdots < Y_n$ have joint pdf given by

(9.5)
$$f_{Y_1,...,Y_n}(y_1,...,y_n) = n! f_X(y_1,...,y_n).$$

Example 9.1. For the rest of this section assume that X_1, \ldots, X_n are i.i.d Uniform(0, 1) rvs with order statistics $Y_1 < \cdots < Y_n$. Here

(9.6)
$$f_X(x) = 1$$
 and $F_X(x) = x$ for $0 < x < 1$,

so (9.1) becomes

(9.7)
$$f_{Y_i}(y) = \frac{n!}{(i-1)!(n-i)!} y^{i-1} (1-y)^{n-i}, \quad 0 < y < 1.$$

Thus we see that the *i*-th order statistic has a Beta distribution:

(9.8)
$$Y_i \equiv X_{(i)} \sim \text{Beta}(i, n-i+1),$$

(9.9)
$$E(Y_i) = E(X_{(i)}) = \frac{i}{n+1},$$

(9.10)
$$\operatorname{Var}(Y_i) = \operatorname{Var}(X_{(i)}) = \frac{i(n-i+1)}{(n+1)^2(n+2)}.$$

Note that if n is odd, $Var(X_{(i)})$ is maximized when $i = \frac{n+1}{2}$, i.e., when $X_{(i)}$ is the sample median. In this case,

(9.11)
$$\operatorname{Var}(X_{\left(\frac{n+1}{2}\right)}) = \frac{1}{4(n+2)} = O\left(\frac{1}{n}\right).$$

On the other hand, the variance is minimized by the extreme order statistics:

(9.12)
$$\operatorname{Var}(X_{(1)}) = \operatorname{Var}(X_{(n)}) = \frac{n}{(n+1)^2(n+2)} = O\left(\frac{1}{n^2}\right),$$

a smaller order of magnitude. The asymptotic distributions of the sample mean and sample extremes are also different: the median is asymptotically normal (§10.6), the extremes are asymptotically exponential (Exercise 9.4.).

Relation between the Beta and Binomial distributions: It follows from (9.7) that for any 0 < y < 1,

$$\frac{n!}{(i-1)!(n-i)!} \int_{y}^{1} u^{i-1} (1-u)^{n-i} du = P[Y_i > y]$$
$$= P[(i-1) \text{ or fewer } X's < y] = \sum_{k=0}^{i-1} {n \choose k} y^k (1-y)^{n-k}.$$

Now let j = i - 1 and t = 1 - u to obtain the following relation between the Beta and Binomial cdfs:

(9.13)
$$\frac{n!}{(n-j-1)!j!} \int_0^{1-y} t^{n-j-1} (1-t)^j du = \sum_{k=0}^j \binom{n}{k} y^k (1-y)^{n-k}.$$

Joint distribution of the sample spacings from Uniform(0,1) rvs. Define the sample spacings $W_1, \ldots, W_n, W_{n+1}$ by

$$W_{1} = Y_{1}$$

$$W_{2} = Y_{2} - Y_{1}$$

$$\vdots$$

$$W_{n} = Y_{n} - Y_{n-1}$$

$$W_{n+1} = 1 - Y_{n}.$$

Note that $0 < W_i < 1$ and $W_1 + \cdots + W_n + W_{n+1} = 1$. Furthermore,

(9.15)

$$Y_1 = W_1$$

$$Y_2 = W_1 + W_2$$

$$\vdots$$

$$Y_n = W_1 + \dots + W_n,$$

From (9.3) and (9.6) the joint pdf of Y_1, \ldots, Y_n is

(9.16)
$$f_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = n!, \quad 0 < y_1 < \dots < y_n < 1,$$

and the Jacobian of the mapping (9.15) is $\left|\frac{\partial Y}{\partial W}\right| = 1$ [verify], so¹⁰

(9.17) $f_{W_1,\ldots,W_n}(w_1,\ldots,w_n) = n!, \quad 0 < w_i < 1, \ 0 < w_1 + \cdots + w_n < 1.$

Clearly both $f_{W_1,\ldots,W_n}(w_1,\ldots,w_n)$ and its range are invariant under all permutations of w_1,\ldots,w_n , so W_1,\ldots,W_n are exchangeable:

(9.18)
$$(W_1, \dots, W_n) \stackrel{\text{distn}}{=} (W_{\pi(1)}, \dots, W_{\pi(n)})$$

for all permutations $(\pi(1), \ldots, \pi(n))$ of $(1, \ldots, n)$. Thus, from (9.8)-(9.10),

(9.19) $W_i \stackrel{\text{distn}}{=} W_1 = Y_1 \sim \text{Beta}(1, n), \qquad 1 \le i \le n,$

(9.20)
$$E(W_i) = \frac{1}{n+1}, \qquad 1 \le i \le n.$$

(9.21)
$$\operatorname{Var}(W_i) = \frac{n}{(n+1)^2(n+2)}, \qquad 1 \le i \le n.$$

¹⁰ Note that (9.17) is a special case of the *Dirichlet distribution* – cf. CB Exercise 4.40 with a = b = c = 1.

Also, for $1 \leq i < j \leq n$, $(W_i, W_j) \stackrel{\text{distn}}{=} (W_1, W_2)$, so

(9.22)

$$Cov(W_i, W_j) = Cov(W_1, W_2)$$

$$= \frac{1}{2} [Var(W_1 + W_2) - Var(W_1) - Var(W_2)]$$

$$= \frac{1}{2} [Var(Y_2) - 2Var(Y_1)]$$

$$= \frac{1}{2} \left[\frac{2(n-1)}{(n+1)^2(n+2)} - \frac{2n}{(n+1)^2(n+2)} \right]$$

$$= -\frac{1}{(n+1)^2(n+2)} < 0.$$

Exercise 9.2. Extend the above results to $(W_1, \ldots, W_n, W_{n+1})$. That is: (i)* Show that $W_1, \ldots, W_n, W_{n+1}$ are exchangeable. *Note:* since $W_1 + \cdots + W_n + W_{n+1} = 1, W_1, \ldots, W_n, W_{n+1}$ do *not* have a joint pdf on \mathbb{R}^{n+1} .

(ii) Show that (9.19) - (9.21) remain valid for i = n + 1. Use (i) to give a short proof of (9.22) for all $1 \le i < j = n + 1$.

Exercise 9.3. Find $Cov(Y_i, Y_j) \equiv Cov(X_{(i)}, X_{(j)})$ for $1 \le i < j \le n$.

Exercise 9.4. Show that $nW_i \xrightarrow{d} \text{Exponential}(\lambda = 1)$ as $n \to \infty$.

Hint: use (9.19), Example 6.4, and Slutsky's Theorem 10.6 (as applied in Example 10.3).

10. Asymptotic (Large Sample) Distribution Theory.

10.1. (Nonparametric) Estimation of a cdf and its functionals.

Let X_1, \ldots, X_n be a random (i.i.d.) sample from an unknown probability distribution P with cdf F on $(-\infty, \infty)$. This distribution may be discrete, continuous, or a mixture. As it stands this is a *nonparametric* statistical model¹¹ because no parametric form is assumed for F. Our goal is to estimate F based on the data X_1, \ldots, X_n .

Definition 10.1. The empirical (\equiv sample) distribution P_n is the (discrete!) probability distribution that assigns probability $\frac{1}{n}$ to each observation X_i . Equivalently, P_n assigns probability $\frac{1}{n}$ to each order statistic $X_{(i)}$, so P_n depends on the data only through the order statistics. The empirical $cdf F_n$ is the cdf associated with P_n , i.e.,

(10.1)
$$F_n(x) = \frac{1}{n} \{ \text{number of } X'_i s \le x \}$$
$$= \frac{1}{n} \sum_{i=1}^n I_{(-\infty,x]}(X_i).$$

(10.2)
$$= \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(X_{(i)}):$$

(Note that F_n is a random function.) Eqn. (10.2) shows that F_n depends only on the order statistics, while (10.1) shows that for each fixed x,

(10.3)
$$nF_n(x) \sim \text{Binomial}(n, p \equiv F(x)).$$

¹¹ Some parametric models: $\{N(\mu, \sigma^2)\}$, $\{\text{Exponential}(\lambda)\}$, $\{\text{Poisson}(\lambda)\}$, etc.

Thus by the LLN and the CLT, for each fixed $x, \hat{p}_n \equiv F_n(x)$ is a consistent, asymptotically normal estimator of $p \equiv F(x)$: as $n \to \infty$,

(10.4)
$$F_n(x) \to F(x),$$

(10.5)
$$\sqrt{n}[F_n(x) - F(x)] \xrightarrow{d} N_1[0, F(x)(1 - F(x))].$$

In fact, (10.4) and (10.5) can be greatly strengthened by considering the asymptotic behavior of the *entire random function* $F_n(\cdot)$ on $(-\infty, \infty)$:

The Glivenko-Cantelli Theorem: If F is continuous, then (10.4) holds uniformly in x:

(10.6)
$$\sup_{-\infty < x < \infty} |F_n(x) - F(x)| \to 0 \quad \text{as } n \to \infty.$$

The Brownian Bridge: First consider the case where U_1, \ldots, U_n are i.i.d. Uniform(0,1) rvs with cdf G(u) = u for $0 \le u \le 1$ and empirical cdf G_n . Here $G_n(\cdot) - G(\cdot)$ is a random function of $u \in [0,1]$ that is "tied down" to 0 at each endpoint. Then

(10.7)
$$\sqrt{n} [G_n - G] \xrightarrow{d} B_0$$
 as random functions on $[0, 1]$ as $n \to \infty$,

where B_0 denotes the *Brownian Bridge* stochastic process. This is a random function on [0, 1] that can be thought of as the conditional distribution of a standard Brownian motion $B(\cdot)$ starting at 0, given that B(1) = 0.

Now let X_1, \ldots, X_n be i.i.d. ~ any (continuous) F on $(-\infty, \infty)$ with empirical cdf F_n . By (10.7) and the fact that $X_i \sim F^{-1}(U_i)$, as $n \to \infty$,

(10.8)
$$\sqrt{n} \left[F_n(F^{-1}) - F(F^{-1}) \right] \sim \sqrt{n} \left[G_n - G \right] \xrightarrow{d} B_0 \text{ on } [0, 1].$$

This gives the asymptotic distribution of the Kolmogorov-Smirnov goodnessof-fit statistic for testing a specified null hypothesis F: with $x = F^{-1}(u)$,

(10.9)
$$\sup_{-\infty < x < \infty} \sqrt{n} |F_n(x) - F(x)| \xrightarrow{d} \sup_{0 \le u \le 1} |B_0(u)| \quad \text{as } n \to \infty.$$

probability distribution:

 $P \text{ (range} = (-\infty, \infty)) \quad \leftrightarrow P_n \text{ (range} = \{X_{(1)}, \dots, X_{(n)}\})$

cumulative distribution function (cdf):

 $F \quad \leftrightarrow \quad F_n$

probability of a set A:

$$P(A) = \int_A dP \qquad \leftrightarrow \qquad P_n(A) = \frac{\sum I_A(X_{(i)})}{n} \equiv \int_A dP_n$$

probability of $A \equiv (\infty, x]$:

 $F(x) = \int_{-\infty}^{x} dF \qquad \leftrightarrow \quad F_n(x) = \frac{\sum I_{(-\infty,x]}(X_{(i)})}{n} \equiv \int_{-\infty}^{x} dF_n$

mean:

$$E_F(X) = \int x dF \qquad \leftrightarrow \qquad \bar{X}_n = \frac{\sum_{i=1}^n X_{(i)}}{n} \equiv \int x dF_n$$

variance:

$$\operatorname{Var}_{F}(X) = \int (x - \operatorname{E}_{F}(X))^{2} dF \quad \leftrightarrow \tilde{s}_{n}^{2} = \frac{\sum_{i=1}^{n} (X_{(i)} - \bar{X}_{n})^{2}}{n} \equiv \int (x - \bar{X}_{n})^{2} dF_{n}$$

p-th quantile (0 < p < 1):

Basic method of nonparametric estimation: To estimate a "functional" $\phi(F)$ of F, use $\phi(F_n)$.^{12,13} (Equivalently, to estimate a "functional" $\phi(P)$ of P, use $\phi(P_n)$.) Simple examples are given on the preceding page. Since F_n is a consistent and asymptotically normal (CAN) estimator of F(both for fixed x and as a random function), we would like to conclude that $\phi(F_n)$ is a CAN estimator of $\phi(F)$. For this we need to show:

- The functional ϕ of interest is continuous, in fact differentiable, in $F_n(!)$;
- Consistency and asymptotic normality are preserved by such a ϕ .

These results require the general theory of "weak convergence" (\equiv convergence in distribution) of stochastic processes. We will not prove them in general, but only for the above examples. We will need the following:

- (i) Definition and properties: convergence in probability $(X_n \xrightarrow{p} X)$ in \mathbf{R}^k , convergence in distribution $(X_n \xrightarrow{d} X)$ in \mathbf{R}^k ,
- (ii) Perturbation \equiv Slutsky-type results ("c" represents a constant in \mathbf{R}^k):

(10.10) $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{p} 0 \Rightarrow X_n + Y_n \xrightarrow{d} X$; (10.11) $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{p} c \Rightarrow h(X_n, Y_n) \xrightarrow{d} h(X, c)$ if *h* is continuous. (iii) Central Limit Theorem (CLT) (for \mathbf{R}^1 , see CB Theorems 5.5.14-15):

Let X_1, X_2, \ldots i.i.d. rvtrs in \mathbf{R}^k , $\mathbf{E}(X_i) = \mu$, $\operatorname{Cov}(X_i) = \Sigma$. Then

(10.12)
$$\sqrt{n} \left(\bar{X}_n - \mu \right) \xrightarrow{d} N_k(0, \Sigma).$$

(iv) Propagation of error $\equiv \delta$ -method \equiv Taylor approximation: Let $\{Y_n\}$ be a sequence of rvtrs in \mathbf{R}^k such that

(10.13)
$$\sqrt{n}(Y_n - \mu) \xrightarrow{d} N_k(0, \Sigma).$$

^{12,} Sometimes we prefer to adjust $\phi(F_n)$ slightly to obtain an unbiased estimator. For example, the unbiased sample variance is $s_n^2 = \frac{n}{n-1}\tilde{s}_n^2$. This remains a c.a.d. estimator.

¹³ If the statistical model is parametric, a nonparametric estimator need not be efficient. In the Poisson(λ) model, for example, the optimal estimator of $E(X) \equiv \lambda$ is \bar{X}_n (see Example 12.13), so the nonparametric estimator s_n^2 of $Var(X) \equiv \lambda$ is inefficient.

If $g(y_1, \ldots, y_k)$ is differentiable at $\mu \equiv (\mu_1, \ldots, \mu_k)^t$, then

(10.14)
$$\sqrt{n} \left[g(Y_n) - g(\mu)\right] \xrightarrow{d} N_1 \left(0, \left(\frac{\partial g}{\partial y_1}, \dots, \frac{\partial g}{\partial y_k}\right) \Sigma \begin{pmatrix} \frac{\partial g}{\partial y_1} \\ \vdots \\ \frac{\partial g}{\partial y_k} \end{pmatrix}\right),$$

where the partial derivatives are evaluated at $y = \mu$, provided that at least one $\frac{\partial g}{\partial y_i} \neq 0$. In the univariate case (k = 1) this can be written as

(10.15)
$$\sqrt{n} \left[g(Y_n) - g(\mu)\right] \xrightarrow{d} N_1\left(0, \left[g'(\mu)\right]^2 \sigma^2\right) \quad (\text{if } g'(\mu) \neq 0).$$

10.2. Modes of convergence.

Let X, X_1, X_2, \ldots be a sequence of rvs or rvtrs with range \mathcal{X} , all occurring in the *same* random experiment, so we can meaningfully write $X_n - X$, $X_1 + \cdots + X_n$, etc. (These rvs/rvtrs are *not* necessarily independent.) We treat the case that $\mathcal{X} = \mathbf{R}^k$ (although most of these concepts and results are valid when \mathcal{X} is any complete separable metric space).

Definition 10.2. X_n converges to X in probability $(X_n \xrightarrow{p} X)$ if $\forall \epsilon > 0$,

(10.16) $P[\|X_n - X\| \ge \epsilon] \to 0 \quad \text{as } n \to \infty.$

 X_n converges to X almost surely $(X_n \stackrel{a.s.}{\rightarrow} X)$ if

(10.17)
$$P[\lim_{n \to \infty} X_n = X] = 1,$$

or equivalently, if $\forall \epsilon > 0$,

(10.18) $P[||X_{n+k} - X|| \ge \epsilon \text{ for at least one } k] \to 0 \quad \text{as } n \to \infty.$

Note that convergence in prob. is weaker than a.s. convergence; in fact the former only involves the joint distributions of all pairs (X_n, X) , while the latter involves the joint distribution of the entire infinite sequence X, X_1, X_2, \ldots Furthermore, if $X \equiv c$, then $X_n \xrightarrow{p} c$ only involves the marginal distribution of each X_i . (We will see that $X_n \xrightarrow{p} c \iff X_n \xrightarrow{d} c$.)

Example 10.1. The Weak (Strong) Law of Large Numbers states that if X_1, X_2, \ldots is a sequence of *i.i.d.* rvs in \mathbb{R}^1 with finite mean $\mathbb{E}(X_i) \equiv \mu$, then

(10.19)
$$\bar{X}_n \xrightarrow{p} \mu \quad (\bar{X}_n \xrightarrow{a.s.} \mu).$$

Under the added assumption of finite second moments, the WLLN is proved easily by means of Chebyshev's inequality (cf. §3.1). The proof of the SLLN is nontrivial.

Next, if we also assume that $Var(X_i) \equiv \sigma^2$ is finite, then the sample variance s_n^2 converges to σ^2 :

(10.20)
$$s_n^2 = \frac{n}{n-1} \left(\frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2 \right) \xrightarrow{p} (\sigma^2 + \mu^2) - \mu^2 = \sigma^2$$

by (10.19) and the WLLN applied to X_1^2, X_2^2, \ldots , and also $s_n^2 \xrightarrow{a.s.} \sigma^2$ by the SLLN. (These results also extend to a sequence of rvtrs in \mathbf{R}^k .) \Box

Next let P, P_1, P_2, \ldots be a sequence of probability measures on a common probability space (Ω, S) (cf. §1.1). We say that $A \in S$ is a *P*-continuity set if $P(\partial A) = 0$, where ∂A is the boundary of A:

Definition 10.3. P_n converges weakly to $P(P_n \xrightarrow{w} P)$ if

(10.21) $P_n(A) \to P(A)$ as $n \to \infty \quad \forall P - \text{continuity sets } A$.

Now let X, X_1, X_2, \ldots be a sequence of rvs or rvtrs, *not* necessarily all occurring in the same random experiment, but still with common range \mathcal{X} . Let P_X (resp. P_{X_n}) denote the probability distribution of X (resp. X_n).

Definition 10.4. X_n converges to X in distribution $(X_n \xrightarrow{d} X)$ if $P_{X_n} \xrightarrow{w} P_X$, that is, if

(10.22) $P[X_n \in A] \to P[X \in A]$ as $n \to \infty \forall P_X$ -continuity sets A.

The need for restriction to *P*-continuity sets is easily seen: Suppose that X_n is the rv degenerate at $\frac{1}{n}$. Then we want that $X_n \xrightarrow{d} X \equiv 0$, but

$$P[X_n \in (-\infty, 0]] = 0 \not\to 1 = P[X \in (-\infty, 0]].$$

Here the set $A \equiv (-\infty, 0]$ is not a P_X -continuity set, since $\partial A = \{0\}$ so $P(\partial A) = 1 \neq 0$. Thus it does hold that $X_n \stackrel{d}{\to} X$.

Note that convergence in distribution is a property of the sequence of distributions $P_X, P_{X_1}, P_{X_2}, \ldots$, not of the values of X, X_1, X_2, \ldots themselves. Thus, for example, if $X_n \xrightarrow{d} X$ we cannot conclude anything about the limiting behavior of $X_n - X$ (unless $X \equiv c$).

Theorem 10.1. (Basic characterization of convergence in distribution.) the following are equivalent:

(a) $X_n \xrightarrow{d} X$. (b) $F_{X_n}(x) \to F_X(x)$ for every continuity point x of F_X . (c) $E[g(X_n)] \to E[g(X)]$ for every bounded continuous function g. Note re (b): F_X is continuous at x iff $(-\infty, x]$ is a P_X -continuity set.

This theorem, whose proof is omitted, is "basic" because condition (c) is much easier to work with than (a) or (b). Here is an example:

Corollary 10.1. If $X_n \xrightarrow{d} X$ and h is continuous on \mathcal{X} , then $h(X_n) \xrightarrow{d} h(X)$. **Proof.** By (c) it suffices to show that $E[g(h(X_n)] \to E[g(h(X)]]$ for any bounded continuous g. But this holds since $X_n \xrightarrow{d} X$ and $g(h(\cdot))$ is bounded and continuous.

Example 10.2. If $X_n \xrightarrow{d} X \sim N_1(0,1)$, then $X_n^2 \xrightarrow{d} X^2 \sim \chi_1^2$.

Remark 10.1. The conclusion of Corollary 10.1 remains valid if h is not continuous but $P[X \in D_h] = 0$, where D_h is the set of discontinuity points of h [proof omitted]. For example, if $h(x) = \frac{1}{x}$ on $(-\infty, \infty)$, then $D_h = \{0\}$. Thus

(10.23)
$$X_n \xrightarrow{d} X$$
 and $P[X=0] = 0 \implies \frac{1}{X_n} \xrightarrow{d} \frac{1}{X}$

Remark 10.2. Condition (c) need not hold if $X_n \xrightarrow{d} X$ and g is continuous but unbounded. For example, suppose that g(x) = x on $(-\infty, \infty)$ and take

$$X_n = \begin{cases} n, & \text{with probability } \frac{1}{n}; \\ 0, & \text{with probability } 1 - \frac{1}{n}. \end{cases}$$

Then $X_n \xrightarrow{d} X \equiv 0$ [verify] but $1 = E(X_n) \not\rightarrow E(X) = 0$. More generally, $X_n \xrightarrow{d} X$ need not imply that the moments of X_n converge to those of X.

Theorem 10.2. (a) $X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{d} X$.

(b) If
$$X = c$$
 then $X_n \xrightarrow{p} c \iff X_n \xrightarrow{d} c$.

Proof. (a) The first \Rightarrow is immediate by comparing (10.18) to (10.16).

We prove the second \Rightarrow for the case of 1-dimensional rvs, i.e., $\mathcal{X} = \mathbf{R}^1$, by verifying condition (b) of Theorem 10.1. Suppose that $X_n \xrightarrow{p} X$ and let xbe a continuity point of F_X . We must show that $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$. For any x' < x < x'',

$$F_X(x') = P[X \le x', X_n \le x] + P[X \le x', X_n > x] \\ \le P[X_n \le x] + P[X \le x', X_n > x].$$

But the last probability $\rightarrow 0$ since $X_n \xrightarrow{p} X$ and x' < x, hence

$$F_X(x') \le \liminf_{n \to \infty} P[X_n \le x] \equiv \liminf_{n \to \infty} F_{X_n}(x)$$

Similarly $F_X(x'') \ge \limsup_{n \to \infty} F_{X_n}(x)$, so

$$F_X(x') \le \liminf_{n \to \infty} F_{X_n}(x) \le \limsup_{n \to \infty} F_{X_n}(x) \le F_X(x'').$$

Now let $x' \uparrow x$ and $x'' \downarrow x$. Since F_X is continuous at x, we conclude that $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$, as required.

(b) (This proof is valid for $\mathcal{X} = \mathbf{R}^k$.) First suppose that $X_n \xrightarrow{p} X \equiv c$. Note that $A \subset \mathbf{R}^k$ is a P_X -continuity set iff $c \notin \partial A$, i.e., iff either $c \in \text{interior}(A)$ or $c \in \text{interior}(\mathbf{R}^k \setminus A)$. In the first case, for all sufficiently small ϵ we have

$$1 \ge P[X_n \in A] \ge P[\|X_n - X\| \le \epsilon] \to 1 = P[X \in A],$$

hence $P[X_n \in A] \to P[X \in A]$. Similarly this limit holds in the second case, hence $X_n \stackrel{d}{\to} X$.

Now suppose $X_n \xrightarrow{d} X \equiv c$. To show that $X_n \xrightarrow{p} c$, consider the set $A_{\epsilon} \equiv \{x \mid ||x - c|| > \epsilon\}$. Clearly A_{ϵ} is a P_X -continuity set $\forall \epsilon > 0$, hence

$$P[||X_n - c|| > \epsilon] \to P[||X - c|| > \epsilon] = 0.$$

Remark 10.3. In Theorem 10.2(a), neither \Leftarrow holds in general. (See CB Example 5.5.8 (as modified by me) for the first counterexample.)

Theorem 10.3. Let h be a continuous function on \mathcal{X} .

(a) If $X_n \xrightarrow{a.s.} X$ then $h(X_n) \xrightarrow{a.s.} h(X)$. (b) If $X_n \xrightarrow{p} X$ then $h(X_n) \xrightarrow{p} h(X)$.

Proof. (a) follows easily from the definition (10.17) of a.s. convergence.

(b) Fix $\epsilon > 0$ and select a sufficiently large $\rho(\epsilon)$ such that the ball

$$B_{\rho(\epsilon)} \equiv \{ x \in \mathbf{R}^k \mid ||x|| \le \rho(\epsilon) \} \subset \mathbf{R}^k$$

satisfies $P[X \in B_{\rho(\epsilon)}] \ge 1 - \epsilon$. Increase $\rho(\epsilon)$ slightly if necessary to insure that $B_{\rho(\epsilon)}$ is a P_X -continuity set [why possible?]. Since $X_n \xrightarrow{d} X$ by Theorem 10.2a, $\exists n(\epsilon)$ s.t.

$$n \ge n(\epsilon) \Rightarrow P[X_n \in B_{\rho(\epsilon)}] \ge 1 - 2\epsilon$$

$$\Rightarrow P[X_n \in B_{\rho(\epsilon)} \text{ and } X \in B_{\rho(\epsilon)}]$$

$$\ge P[X_n \in B_{\rho(\epsilon)}] + P[X \in B_{\rho(\epsilon)}] - 1$$

$$\ge 1 - 3\epsilon.$$

Furthermore, h is uniformly continuous on $B_{\rho(\epsilon)}$, that is, $\exists \delta(\epsilon) > 0$ s.t.

$$x, y \in B_{\rho(\epsilon)}, \ \|x - y\| < \delta(\epsilon) \Rightarrow |h(x) - h(y)| < \epsilon.$$

Also, since $X_n \xrightarrow{p} X$, $\exists n'(\epsilon)$ s.t.

$$n \ge n'(\epsilon) \implies P[||X_n - X|| \ge \delta(\epsilon)] \le \epsilon.$$

Thus $h(X_n) \xrightarrow{p} h(X)$, since for $n \ge \max\{n(\epsilon), n'(\epsilon)\},\$

$$P[|h(X_n) - h(X)| \ge \epsilon] = P[|h(X_n) - h(X)| \ge \epsilon, X_n \in B_{\rho(\epsilon)}, X \in B_{\rho(\epsilon)}] + P[|h(X_n) - h(X)| \ge \epsilon, \{X_n \in B_{\rho(\epsilon)}, X \in B_{\rho(\epsilon)}\}^c] \le P[||X_n - X|| \ge \delta(\epsilon), X_n \in B_{\rho(\epsilon)}, X \in B_{\rho(\epsilon)}] + 1 - P[X_n \in B_{\rho(\epsilon)} \text{ and } X \in B_{\rho(\epsilon)}] \le 4\epsilon.$$

Exercise 10.1. It can be shown that if $X_n \xrightarrow{p} X$, there is a subsequence $\{n'\} \subseteq \{n\}$ s.t. $X_{n'} \xrightarrow{a.s.} X$. Use this to give another proof of (b).

Theorem 10.4 (Slutsky). If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} 0$, then $X_n + Y_n \xrightarrow{d} X$. **Exercise 10.2.** Prove Theorem 10.4 for rvs in \mathbb{R}^1 . *Hint:* Similar to the proof of the second \Rightarrow in Theorem 10.2a.

Theorem 10.5. Suppose that $X_n \xrightarrow{d} X$ in \mathbf{R}^k and $Y_n \xrightarrow{p} c$ in \mathbf{R}^l . Then $(X_n, Y_n) \xrightarrow{d} (X, c)$ in \mathbf{R}^{k+l} .

Proof. Write $(X_n, Y_n) = (X_n, c) + (0, Y_n - c)$. By Theorem 10.1c, $(X_n, c) \stackrel{d}{\rightarrow} (X, c)$, and clearly $(0, Y_n - c) \stackrel{d}{\rightarrow} (0, 0)$, so the result follows from Theorem 10.4.

Theorem 10.6 (Slutsky). Suppose that $X_n \xrightarrow{d} X$ in \mathbb{R}^k and $Y_n \xrightarrow{p} c$ in \mathbb{R}^l . If h(x, y) is continuous then $h(X_n, Y_n) \xrightarrow{d} h(X, c)$.

Proof. Apply Theorems 10.5 and Corollary 10.1.

Remark 10.4. If h is not continuous but $P[(X, c) \in D_h] = 0$, then $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ still imply $h(X_n, Y_n) \xrightarrow{d} h(X, c)$ (use Remark 10.1).

Example 10.3. If $X_n \xrightarrow{d} N_1(0, 1)$ and $Y_n \to c \neq 0$ then by Remark 10.4,

$$h(X_n, Y_n) \equiv \frac{X_n}{Y_n} \xrightarrow{d} \frac{N_1(0, 1)}{c} = N_1\left(0, \frac{1}{c^2}\right).$$

In particular, if X_1, \ldots, X_n are i.i.d. rvs with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$,

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N_1(0, 1) \quad \text{and} \quad \frac{s_n}{\sigma} \xrightarrow{p} 1$$

by the CLT and (10.20), so

(10.24)
$$t_n \equiv \frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n} = \frac{\sqrt{n}(\bar{X}_n - \mu)/\sigma}{s_n/\sigma} \xrightarrow{d} N_1(0, 1).$$

[This is a "robustness" property of the Student *t*-statistic, since it shows that the large-sample distribution of t_n does not depend on the actual distribution being sampled. (As long as it has finite variance.)]

Example 10.4. (Asymptotic distribution of the sample variance.) Let X_1, \ldots, X_n be an i.i.d. sample from a univariate distribution with *finite* fourth moment. Set $\mu = E(X_i), \sigma^2 = Var(X_i)$, and

(10.25)
$$\lambda_4 = \operatorname{Var}\left[\left(\frac{X_i - \mu}{\sigma}\right)^2\right] = \operatorname{E}\left[\left(\frac{X_i - \mu}{\sigma}\right)^4\right] - 1 \equiv \kappa_4 - 1.$$

Then the sample variance

$$s_n^2 \equiv \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n-1} \left[\sum_{i=1}^n X_i^2 - n\bar{X}_n^2 \right]$$

is an unbiased, consistent estimator for σ^2 (cf. Examples 8.1 and 10.1). We wish to find the limiting distribution of s_n (suitably normalized) as $n \to \infty$:

Let $V_i = X_i - \mu$. We have the following normal approximation for s_n^2 :

$$\sqrt{n}(s_n^2 - \sigma^2) = \sqrt{n} \left[\frac{n}{n-1} \left(\frac{1}{n} \sum_{i=1}^n V_i^2 - \bar{V}_n^2 \right) - \sigma^2 \right] \\
= \frac{n}{n-1} \left[\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n V_i^2 - \sigma^2 \right) \right] + \frac{\sqrt{n}\sigma^2}{n-1} - \frac{n}{\sqrt{n}(n-1)} \left(\sqrt{n}\bar{V}_n \right)^2 \\
(10.26) \qquad \stackrel{d}{\to} N_1(0, \lambda_4 \sigma^4)$$

by Slutsky's Theorem, the CLT, and the fact that $\sqrt{n}\overline{V}_n \xrightarrow{d} N_1(0, \sigma^2)$. [Since λ_4 depends on the form of the distribution being sampled, this shows that s_n^2 , unlike t_n , is not robust to departures from normality.] In the special case that $X_i \sim N_1(\mu, \sigma^2), \lambda_4 = \operatorname{Var}(\chi_1^2) = 2$ [verify], so

(10.27)
$$\sqrt{n}(s_n^2 - \sigma^2) \xrightarrow{d} N_1(0, 2\sigma^4).$$
 \square

10.3. Propagation of error $\equiv \delta$ -method \equiv Taylor approximation.

Theorem 10.7. (a) Let $\{Y_n\}$ be a sequence of rvs in \mathbb{R}^1 such that

(10.28)
$$\sqrt{n}(Y_n - \mu) \xrightarrow{d} N_1(0, \sigma^2), \quad \sigma^2 > 0.$$

If g(y) is differentiable at μ and $g'(\mu) \neq 0$ then

(10.29)
$$\sqrt{n} \left[g(Y_n) - g(\mu) \right] \xrightarrow{d} N_1 \left(0, \left[g'(\mu) \right]^2 \sigma^2 \right).$$

(b) Let $\{Y_n\}$ be a sequence of rvtrs in \mathbf{R}^k such that

(10.30)
$$\sqrt{n}(Y_n - \mu) \xrightarrow{d} N_k(0, \Sigma), \quad \Sigma \text{ pd.}$$

If $g(y_1, \ldots, y_k)$ is differentiable at $\mu \equiv (\mu_1, \ldots, \mu_k)^t$, then

(10.31)
$$\sqrt{n} \left[g(Y_n) - g(\mu)\right] \xrightarrow{d} N_1 \left(0, \left(\frac{\partial g}{\partial y_1}, \dots, \frac{\partial g}{\partial y_k}\right) \Sigma \left(\begin{array}{c} \frac{\partial g}{\partial y_1} \\ \vdots \\ \frac{\partial g}{\partial y_k} \end{array}\right)\right),$$

where the partial derivatives are evaluated at $y = \mu$, provided that at least one $\frac{\partial g}{\partial y_i} \neq 0$.

Proof. (a) Since g is differentiable at μ , its first-order Taylor expansion is

(10.32)
$$g(y) = g(\mu) + (y - \mu)g'(\mu) + O(|y - \mu|^2).$$

Thus by (10.28) and Slutsky's theorems,

$$\sqrt{n} [g(Y_n) - g(\mu)] = \sqrt{n} (Y_n - \mu) g'(\mu) + O(\sqrt{n} |Y_n - \mu|^2)$$

$$\stackrel{d}{\to} N_1 (0, [g'(\mu)]^2 \sigma^2),$$

since $n |Y_n - \mu|^2 \xrightarrow{d} |N_1(0, \sigma^2)|^2 < \infty \Rightarrow O(\sqrt{n} |Y_n - \mu|^2) = O_p\left(\frac{1}{\sqrt{n}}\right) \xrightarrow{p} 0.$

(b) The multivariate first-order Taylor approximation of g at $y = \mu$ is

$$g(y) = g(\mu) + \sum_{i=1}^{k} (y_i - \mu_i) \frac{\partial g}{\partial y_i} \Big|_{\mu} + O(||y - \mu||^2)$$

$$\equiv g(\mu) + \left(\frac{\partial g}{\partial y_1}, \dots, \frac{\partial g}{\partial y_k}\right) (y - \mu) + O(||y - \mu||^2)$$

Thus by (10.30) and Slutsky's theorems,

$$\sqrt{n} \left[g(Y_n) - g(\mu) \right] = \left(\frac{\partial g}{\partial y_1}, \dots, \frac{\partial g}{\partial y_k} \right) \sqrt{n} (Y_n - \mu) + O(\sqrt{n} \|Y_n - \mu\|^2)$$
$$\stackrel{d}{\to} N_1 \left(0, \left(\frac{\partial g}{\partial y_1}, \dots, \frac{\partial g}{\partial y_k} \right) \Sigma \begin{pmatrix} \frac{\partial g}{\partial y_1} \\ \vdots \\ \frac{\partial g}{\partial y_k} \end{pmatrix} \right),$$

since $n ||Y_n - \mu||^2 \xrightarrow{d} ||N_k(0, \Sigma)||^2 < \infty \Rightarrow O(\sqrt{n} ||Y_n - \mu||^2) \xrightarrow{p} 0.$ \Box

Remark 10.5. Often $Y_n = \overline{X}_n$, a sample mean of i.i.d. rvs or rvtrs, but Y_n also may be a sample median (see §10.6), a maximum likelihood estimator (see §14.3), etc.

Example 10.5. (a) Let $g(y) = \frac{1}{y}$. Then g is differentiable at $y = \mu$ with $g'(\mu) = -\frac{1}{\mu^2}$ (provided that $\mu \neq 0$), so (10.26) becomes

(10.33)
$$\sqrt{n}\left(\frac{1}{Y_n} - \frac{1}{\mu}\right) \xrightarrow{d} N_1\left(0, \frac{\sigma^2}{\mu^4}\right).$$

(b) Similarly, if $g(y) = \log y$ and $\mu > 0$, then $g'(\mu) = \frac{1}{\mu}$, hence

(10.34)
$$\sqrt{n} \left(\log Y_n - \log \mu\right) \xrightarrow{d} N_1\left(0, \frac{\sigma^2}{\mu^2}\right).$$

Exercise 10.3. Assume that $\sqrt{n}(Y_n - \mu) \xrightarrow{d} N_1(0, \sigma^2)$.

(i) Find the asymptotic distribution of $\sqrt{n} (Y_n^2 - \mu^2)$ when $\mu \neq 0$.

(ii) When $\mu = 0$, show that this asymptotic distribution is degenerate (constant). Find the (non-degenerate!) asymptotic distribution of $n(Y_n^2 - 0)$. Note: if $\mu = 0$ the first-order (linear) term in the Taylor expansion of $g(y) \equiv y^2$ at y = 0 vanishes, so the second-order (quadratic) term determines the limiting distribution – see CB Theorem 5.5.26. Of course, here no expansion is needed [why?].

Example 10.6. For a bivariate example, take g(x, y) = xy and suppose

(10.35)
$$\sqrt{n} \left[\begin{pmatrix} X_n \\ Y_n \end{pmatrix} - \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \right] \xrightarrow{d} N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} \right]$$

Then $\frac{\partial g}{\partial x} = \mu_y$ and $\frac{\partial g}{\partial y} = \mu_x$ at (μ_x, μ_y) , so if $(\mu_x, \mu_y) \neq (0, 0)$, (10.34) yields

(10.36)
$$\sqrt{n}(X_nY_n - \mu_x\mu_y) \xrightarrow{d} N_1 \left[0, \ (\mu_y, \ \mu_x) \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix} \right]$$
$$= N_1(0, \ \mu_y^2 \sigma_x^2 + \mu_x^2 \sigma_y^2 + 2\mu_y \mu_x \sigma_{xy})$$

In particular, if X_n and Y_n are asymptotically independent, i.e., if $\sigma_{xy} = 0$,

(10.37)
$$\sqrt{n}(X_nY_n - \mu_x\mu_y) \xrightarrow{d} N_1(0, \,\mu_y^2\sigma_x^2 + \mu_x^2\sigma_y^2).$$

[Note the interchange of the subscripts x and y – can you explain this?]

Exercise 10.4. In Example 10.6, suppose that $(\mu_x, \mu_y) = (0, 0)$. Show that $\sqrt{n}(X_nY_n - 0) \xrightarrow{d} 0$ but that nX_nY_n has a non-degenerate limiting distribution. Express this limiting distribution in terms of chi-square variates and find its mean and variance.

Exercise 10.5. (i) Repeat Example 10.6 with $g(x, y) = \frac{x}{y}$. (Take $\mu_y \neq 0$.)

(ii) Let $F_{m,n}$ denote a rv having the *F*-distribution with *m* and *n* degrees of freedom. (See CB Definition 5.36 and especially page 624.) First suppose that m = n. Show that as $n \to \infty$,

(10.38)
$$\sqrt{n} (F_{n,n} - 1) \xrightarrow{d} N_1(0, 4).$$

(iii) Now let $m \to \infty$ and $n \to \infty$ s.t. $\frac{m}{n} \to \gamma$ $(0 < \gamma < \infty)$. Show that

(10.39)
$$\sqrt{m} \left(F_{m,n} - 1 \right) \xrightarrow{d} N_1(0, 2 + 2\gamma),$$

(10.40)
$$\sqrt{n} \left(F_{m,n}-1\right) \xrightarrow{d} N_1\left(0, 2+\frac{2}{\gamma}\right).$$

10.4. Variance-stabilizing transformations.

Let $\{Y_n\}$ be a consistent, asymptotically normal (CAN) sequence of estimators of a real-valued statistical parameter θ , e.g., $\theta = p$ in the Binomial(n, p)model, $\theta = \lambda$ in the Poisson (λ) model, $\theta = \lambda$ or $\frac{1}{\lambda}$ in the Exponential (λ) model, $\theta = \mu$ or σ or $\frac{\mu}{\sigma}$ in the normal model $N(\mu, \sigma^2)$. Assume that the asymptotic variance of Y_n depends only on θ :

(10.41)
$$\sqrt{n} (Y_n - \theta) \xrightarrow{d} N_1 (0, \sigma^2(\theta)).$$

Define $z_{\beta} = \Phi^{-1}(1-\beta)$, the $(1-\beta)$ -quantile of $N_1(0, 1)$. Then (10.41) gives

(10.42)
$$\begin{aligned} 1 - \alpha &\approx P\left[-z_{\frac{\alpha}{2}} \leq \frac{\sqrt{n}(Y_n - \theta)}{\sigma(\theta)} \leq z_{\frac{\alpha}{2}}\right] \\ &= P\left[Y_n - \frac{\sigma(\theta)}{\sqrt{n}} z_{\frac{\alpha}{2}} \leq \theta \leq Y_n + \frac{\sigma(\theta)}{\sqrt{n}} z_{\frac{\alpha}{2}}\right]. \end{aligned}$$

hence $Y_n \pm \frac{\sigma(\theta)}{\sqrt{n}} z_{\frac{\alpha}{2}}$ is an approximate $(1 - \alpha)$ -confidence interval for θ . However, this confidence interval has two drawbacks:

- (i) If $\sigma(\theta)$ is not constant but varies with θ , it must be estimated,¹⁴ which may be difficult and introduces additional variability in the confidence limits, so the actual confidence probability will be *less* the nominal 1α .
- (ii) The accuracy of the normal approximation (10.41) may vary with θ . That is, how large *n* must be to insure the accuracy of (10.41) may depend on the unknown parameter θ , hence is not subject to control.

¹⁴ Sometimes the inequalities in (10.42) may be solved directly for θ . This occurs. e.g., in the Binomial(n, p) model with $\theta = p$, $\sigma^2(p) = p(1-p)$ - see Example 10.8.

These two drawbacks can be resolved by a variance-stabilizing transformation $g(Y_n)$, found as follows: For any g that is differentiable at $y = \theta$, it follows from (10.41) and the propagation of error formula that

(10.43)
$$\sqrt{n}[g(Y_n) - g(\theta)] \xrightarrow{d} N_1\left(0, [g'(\theta)]^2 \sigma^2(\theta)\right).$$

Therefore, if we can find a differentiable function g such that

(10.44)
$$[g'(\theta)]^2 \sigma^2(\theta) = 1 \text{ (not depending on } \theta),$$

then (10.40) becomes

(10.45)
$$\sqrt{n}[g(Y_n) - g(\theta)] \xrightarrow{d} N_1(0, 1).$$

This averts the difficulty (i), since it implies that $g(Y_n) \pm \sqrt{\frac{1}{n}} z_{\frac{\alpha}{2}}$ is an approximate $(1 - \alpha)$ -confidence interval for $g(\theta)$ that does not involve the unknown θ . If in addition $g(\theta)$ is monotone in θ , this interval can be converted to a confidence interval for θ . Furthermore, it may also alleviate difficulty (ii) since the normal approximation (10.45) is usually accurate over a wide range of θ -values uniformly in n. (See Example 10.8.)

To find g that satisfies (10.44), simply note that (10.44) yields

(10.46)
$$g'(\theta) = \frac{1}{\sigma(\theta)},$$

$$g(\theta) = \int \frac{d\theta}{\sigma(\theta)} \quad \text{(an indefinite integral)}.$$

If we can solve this for g, then $g(Y_n)$ will satisfy (10.45).

Example 10.7. Suppose that X_1, \ldots, X_n is an i.i.d. sample from the Exponential $(\lambda = \frac{1}{\theta})$ distribution. Then $E(X_i) = \theta$, $Var(X_i) = \theta^2$, hence \overline{X}_n is a CAN estimator of θ :

(10.47)
$$\sqrt{n} (\bar{X}_n - \theta) \xrightarrow{d} N_1(0, \theta^2).$$

Here $\sigma^2(\theta) = \theta^2$, so the variance-stabilizing function $g(\theta)$ in (10.46) becomes

$$g(\theta) = \int \frac{d\theta}{\theta} = \log \theta.$$

We conclude that

(10.48)
$$\sqrt{n} \left[\log \bar{X}_n - \log \theta \right] \xrightarrow{d} N_1(0, 1),$$

which yields confidence intervals for $\log \theta$ and thence for θ .

Note: If $X \sim \text{Expo}(\frac{1}{\theta})$ then θ is a scale parameter, i.e., $X \sim \theta Y$ where $Y \sim \text{Expo}(1)$. Thus

$$\log X \sim \log Y + \log \theta,$$

which easily shows why $\log \bar{X}_n$ stabilizes the variance for $0 < \theta < \infty$.

Example 10.8. Suppose we want a confidence interval for $p \in (0, 1)$ based on $X_n \sim \text{Binomial}(n, p)$. Then $\hat{p}_n \equiv \frac{X_n}{n}$ is a CAN estimator of p:

(10.49)
$$\sqrt{n} (\hat{p}_n - p) \xrightarrow{d} N_1 (0, p(1-p))$$

Here $\sigma(p) = \sqrt{p(1-p)}$, so the function g(p) in (10.46) is given by

$$g(p) = \int \frac{dp}{\sqrt{p(1-p)}} = 2 \arcsin(\sqrt{p})$$
 [verify!]

Thus $\arcsin(\sqrt{\hat{p}_n})$ stabilizes the variance of \hat{p}_n (see the following figures):

(10.50)
$$\sqrt{n} \left[\arcsin(\sqrt{\hat{p}_n}) - \arcsin(\sqrt{p}) \right] \xrightarrow{d} N_1 \left(0, \frac{1}{4} \right),$$

which yields confidence intervals for $\arcsin(\sqrt{p})$ and thence for p.

 \Box

 $\operatorname{arcsin}(\sqrt{p})$:

This non-linear transformation "stretches" the interval (0, 1) more near p = 0 and p = 1 than near p = 1/2.

Asymptotic distn. of \hat{p}_n :

 $\hat{p}_n \rightarrow$

 $\operatorname{Var}(\hat{p}_n) = \frac{p(1-p)}{n}$ depends on p. It is very small for $p \approx 0, 1$, where the distribution of \hat{p}_n is relatively *skewed* due to its truncation at the endpoint 0 or 1 and the normal approximation is not very good unless n is very large.

Asymptotic distn. of $\arcsin(\sqrt{\hat{p}_n})$:

 $\arcsin(\sqrt{\hat{p}_n}) \rightarrow$

 $\operatorname{Var}(\operatorname{arcsin}(\sqrt{\hat{p}_n})) \approx \frac{1}{4n}$ does not depend on p. The distribution of $\operatorname{arcsin}(\sqrt{\hat{p}_n})$ is not very skewed for $p \approx 0, 1$ and the normal approximation is fairly good uniformly in p for moderately large n.

Remark 10.6. A variance-stabilizing transformation can be used to make comparisons between two or more parameters based on independent samples. For example, if $m\hat{p}_1 \sim \text{Binomial}(m, p_1)$ and $n\hat{p}_2 \sim \text{Binomial}(n, p_2)$ with $m \to \infty$ and $n \to \infty$ s.t. $\frac{m}{n} \to \gamma$ ($0 < \gamma < \infty$), then [verify]

$$\sqrt{n} \left[\left(\arcsin(\sqrt{\hat{p}_1}) - \arcsin(\sqrt{\hat{p}_2}) \right) - \left(\arcsin(\sqrt{p_1}) - \arcsin(\sqrt{p_2}) \right) \right]$$

$$(10.51) \qquad \qquad \stackrel{d}{\to} N_1 \left(0, \frac{1}{4} + \frac{1}{4\gamma} \right).$$

Thus $\left(\operatorname{arcsin}(\sqrt{\hat{p}_1}) - \operatorname{arcsin}(\sqrt{\hat{p}_2})\right) \pm \sqrt{\frac{1+\gamma}{4n\gamma}} z_{\frac{\alpha}{2}}$ is an approximate $(1-\alpha)$ confidence interval for $\left(\operatorname{arcsin}(\sqrt{p_1}) - \operatorname{arcsin}(\sqrt{p_2})\right)$, which can in turn be
used to test the hypothesis $p_1 = p_2$ vs. $p_1 \neq p_2$.

Exercise 10.6. (i) Assume $X_{\lambda} \sim \text{Poisson}(\lambda)$. Find a variance-stabilizing transformation for X_{λ} as $\lambda \to \infty$. That is, find a function h and constant c > 0 s.t.

(10.52)
$$[h(X_{\lambda}) - h(\lambda)] \xrightarrow{d} N_1(0, c).$$

Use this to obtain an approximate $(1 - \alpha)$ -confidence interval for λ .

Hint: Write $\lambda = n\theta$ with $n \to \infty$ and θ the fixed unknown parameter.

(ii) Let $X_{\lambda} \sim \text{Poisson}(\lambda)$ and $X_{\mu} \sim \text{Poisson}(\mu)$, where $X_{\lambda} \perp \perp X_{\mu}$ and λ and μ are both large. Based on (i), describe an approximate procedure for testing $\lambda = \mu$ vs. $\lambda \neq \mu$.

10.5. Asymptotic distribution of a sample covariance matrix.

Let $W_1 \equiv \begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \dots, W_n \equiv \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ be an i.i.d. sample from a *bivariate* distribution with *finite fourth moments*. Let

$$\mathbf{E}\begin{pmatrix} X_i\\Y_i \end{pmatrix} = \begin{pmatrix} \mu\\\nu \end{pmatrix} \equiv \xi,$$

(10.53)

$$\operatorname{Cov}\begin{pmatrix}X_i\\Y_i\end{pmatrix} = \begin{pmatrix}\sigma^2 & \rho \sigma \tau\\ \rho \sigma \tau & \tau^2\end{pmatrix} \equiv \begin{pmatrix}\sigma & 0\\ 0 & \tau\end{pmatrix}\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\begin{pmatrix}\sigma & 0\\ 0 & \tau\end{pmatrix} \equiv \Sigma.$$

The sample covariance matrix is defined to be

(10.54)
$$S_n = \frac{1}{n-1} \sum_{i=1}^n (W_i - \bar{W}_n) (W_i - \bar{W}_n)'$$

(10.55) $= \frac{1}{n-1} \begin{pmatrix} \sum (X_i - \bar{X}_n)^2 & \sum (X_i - \bar{X}_n) (Y_i - \bar{Y}_n) \\ \sum (X_i - \bar{X}_n) (Y_i - \bar{Y}_n) & \sum (Y_i - \bar{Y}_n)^2 \end{pmatrix}.$

First, S_n is an unbiased, consistent estimator of Σ . To see this, set

$$V_i = \begin{pmatrix} X_i - \mu \\ Y_i - \nu \end{pmatrix} \equiv W_i - \xi,$$

so $E(V_i) = 0$, $Cov(V_i) = \Sigma$, $Cov(\overline{V}_n) = \frac{1}{n}\Sigma$. Then from (10.54),

(10.56)

$$E(S_n) = \frac{1}{n-1} E\left[\sum (V_i - \bar{V}_n)(V_i - \bar{V}_n)'\right]$$

$$= \frac{1}{n-1} E\left[\sum V_i V_i' - n\bar{V}_n \bar{V}_n'\right]$$

$$= \frac{1}{n-1} \left[n \operatorname{Cov}(V_i) - n \operatorname{Cov}(\bar{V}_n)\right]$$

$$= \frac{1}{n-1} \left[n \Sigma - n \cdot \frac{1}{n} \Sigma\right]$$

$$= \Sigma,$$

so S_n is unbiased. Also, S_n is consistent: from (10.56) and the WLLN,

(10.57)
$$S_n = \frac{n}{n-1} \left(\frac{1}{n} \sum V_i V_i' - \bar{V}_n \bar{V}_n' \right) \xrightarrow{p} \Sigma.$$

We wish to determine the limiting distribution of $\sqrt{n}(S_n - \Sigma)$ as $n \to \infty$ (recall Example 10.4 re s_n^2), from which we can determine, for example, the limiting distribution of the sample correlation coefficient. Since S_n are 2×2 symmetric matrices, equivalently we can consider the limiting distribution of the 3×1 rvtr $\sqrt{n}(\tilde{S}_n - \tilde{\Sigma})$, where for any 2×2 symmetric matrix,

$$\left(\begin{array}{cc} a & b \\ b & c \end{array}\right) = \left(\begin{array}{c} a \\ b \\ c \end{array}\right).$$

Note that this mapping $A \to \widetilde{A}$ is a linear operation. Thus from (10.57), \tilde{S}_n and $\tilde{\Sigma}$ are the 3-dimensional vectors

If we denote the 3-dimensional rvtr $\widetilde{V_iV_i'}$ by

(

(10.59)
$$\widetilde{Z}_{i} = \widetilde{V_{i}V_{i}'} = \begin{pmatrix} (X_{i} - \mu)^{2} \\ (X_{i} - \mu)(Y_{i} - \nu) \\ (Y_{i} - \nu)^{2} \end{pmatrix},$$

then $\tilde{Z}_1, \ldots, \tilde{Z}_n$ are i.i.d., $E(\tilde{Z}_i) = \tilde{\Sigma}$, and

$$\operatorname{Cov}(\tilde{Z}_i) = \begin{pmatrix} \kappa_{40}\sigma^4 & \kappa_{31}\sigma^3\tau & \kappa_{22}\sigma^2\tau^2 \\ \kappa_{31}\sigma^3\tau & \kappa_{22}\sigma^2\tau^2 & \kappa_{13}\sigma\tau^3 \\ \kappa_{22}\sigma^2\tau^2 & \kappa_{13}\sigma\tau^3 & \kappa_{04}\tau^4 \end{pmatrix} - \tilde{\Sigma}\,\tilde{\Sigma}'$$
$$= D(\sigma,\tau)[K - B(\rho)B(\rho)']D(\sigma,\tau)$$

(10.61)
$$\equiv D(\sigma,\tau)[K - R(\rho)R(\rho)']D(\sigma,\tau)$$
$$\equiv \Lambda,$$

where (recall (10.25))

(10.62)
$$\kappa_{jk} = \mathbf{E}\left[\left(\frac{X_i - \mu}{\sigma}\right)^j \left(\frac{Y_i - \nu}{\tau}\right)^k\right],$$

$$D(\sigma,\tau) = \begin{pmatrix} \sigma^2 & 0 & 0\\ 0 & \sigma\tau & 0\\ 0 & 0 & \tau^2 \end{pmatrix}, \quad K = \begin{pmatrix} \kappa_{40} & \kappa_{31} & \kappa_{22}\\ \kappa_{31} & \kappa_{22} & \kappa_{13}\\ \kappa_{22} & \kappa_{13} & \kappa_{04} \end{pmatrix}, \quad R(\rho) = \begin{pmatrix} 1\\ \rho\\ 1 \end{pmatrix}.$$

(The factorization (10.61) shows how the scale parameters σ and τ , the standardized fourth moments $\kappa_{j,k}$, and the correlation ρ contribute to the covariance matrix of \tilde{Z}_i , hence to that of S_n .) Thus the CLT yields

(10.63)
$$\sqrt{n}\left(\overline{\tilde{Z}}_n - \tilde{\Sigma}\right) \xrightarrow{d} N_3(0, \Lambda).$$

so if we can show that

(10.64)
$$\tilde{\Delta} \equiv \sqrt{n} \left(\overline{\tilde{Z}}_n - \tilde{\Sigma} \right) - \sqrt{n} \left(\tilde{S}_n - \tilde{\Sigma} \right) \xrightarrow{p} 0,$$

then it will follow from (10.63) and Slutsky's Theorem that

(10.65)
$$\sqrt{n}\left(\tilde{S}_n - \tilde{\Sigma}\right) \xrightarrow{d} N_3(0, \Lambda).$$

By linearity, however,

$$\tilde{\Delta} = \sqrt{n} \left(\overline{\tilde{Z}}_n - \tilde{S}_n \right) = \sqrt{n} \left(\overline{\tilde{Z}_n - S_n} \right)$$

and $\overline{Z}_n = \frac{1}{n} \sum V_i V_i'$, so from (10.57) we obtain (10.64):

$$\begin{split} \sqrt{n} \left(\bar{Z}_n - S_n \right) &= \sqrt{n} \left[\frac{1}{n} \sum V_i V_i' - \frac{n}{n-1} \left(\frac{1}{n} \sum V_i V_i' - \bar{V}_n \bar{V}_n' \right) \right] \\ &= \frac{\sqrt{n}}{n-1} \left(n \bar{V}_n \bar{V}_n' - \frac{1}{n} \sum V_i V_i' \right) \\ &\stackrel{p}{\to} 0 \qquad [\text{verify}]. \end{split}$$

Example 10.8 (Consistency and asymptotic distribution of the sample correlation coefficient and Fisher's z-transform.) The sample correlation r_n is a consistent estimator of the population correlation ρ : since

(10.66)
$$r_n \equiv \frac{\frac{1}{n-1}\sum(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\frac{1}{n-1}\sum(X_i - \bar{X}_n)^2}\sqrt{\frac{1}{n-1}\sum(Y_i - \bar{Y}_n)^2}} = g(\tilde{S}_n),$$

where

(10.67)
$$g(x, y, z) \equiv \frac{y}{\sqrt{x}\sqrt{z}}$$

is continuous for x, y > 0, and since $\tilde{S}_n \xrightarrow{p} \tilde{\Sigma}$ by (10.57), we have that

(10.68)
$$r_n \xrightarrow{p} \rho$$
 as $n \to \infty$.

Exercise 10.7. (i) Apply (10.65) to find the asymptotic distribution of $\sqrt{n} (r_n - \rho)$. Express the asymptotic variance in terms of ρ and the moments κ_{jk} in (10.62). (Since r_n is invariant under location and scale changes, its distribution does not depend on μ , ν , σ , or τ .)

(ii) Specialize the result in (i) to the *bivariate normal* case, i.e., assume that

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \mu \\ \nu \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho \sigma \tau \\ \rho \sigma \tau & \tau^2 \end{pmatrix} \equiv \Sigma \right]$$

[Evaluate the κ_{jk} in (10.62).]

(iii) In (ii), find a variance-stabilizing transformation for r_n . That is, find a function $g(r_n)$ such that $\sqrt{n} [g(r_n) - g(\rho)] \xrightarrow{d} N_1(0, c)$ where c does not depend on ρ (specify c).

10.6. Asymptotic distribution of sample quantiles.

Let X_1, \ldots, X_n be an i.i.d. sample from a *continuous* distribution on \mathbb{R}^1 with unknown cdf F. We shall show that for (0 , the*p* $-th sample quantile <math>X_{([np]+1)} \equiv F_n^{-1}(p)$ is a CAN estimator of the *p*-th population quantile $\equiv F^{-1}(p)$. That is, we shall show:

(10.69)
$$X_{([np]+1)} \xrightarrow{p} F^{-1}(p)$$

provided that F^{-1} exists and is continuous at p;

(10.70)
$$\sqrt{n} \left(X_{([np]+1)} - F^{-1}(p) \right) \xrightarrow{d} N_1 \left(0, \frac{p(1-p)}{\left[f(F^{-1}(p)) \right]^2} \right)$$

provided that F^{-1} exists and is differentiable at p. (This requires that $f(F^{-1}(p)) > 0$.)

To derive (10.69) and (10.70), first consider the case where U_1, \ldots, U_n are i.i.d. Uniform(0,1) rvs with cdf $G(u) = u, 0 \le u \le 1$ (recall §10.1). From (10.4) and (10.5) we know that the empirical cdf $G_n(u) \sim \frac{1}{n}$ Binomial(n, u) is a CAN estimator of G(u):

(10.71)
$$G_n(u) \xrightarrow{p} G(u),$$

(10.72)
$$\sqrt{n} \left[G_n(u) - G(u) \right] \xrightarrow{d} N_1(0, u(1-u)).$$

Let $0 < U_{(1)} < \cdots < U_{(n)} < 1$ be the order statistics based on U_1, \ldots, U_n .

Proposition 10.1. $U_{([np]+1)} \xrightarrow{p} p \ (\equiv G^{-1}(p)).$ **Proof.** Fix $\epsilon > 0$. Then

$$\left\{ U_{([np]+1)} \le p - \epsilon \right\} = \left\{ [np] + 1 \text{ or more } U'_i s \le p - \epsilon \right\}$$
$$= \left\{ G_n(p - \epsilon) \ge \frac{[np] + 1}{n} \right\}.$$

But $G_n(p-\epsilon) \xrightarrow{p} p-\epsilon$ and $\frac{[np]+1}{n} \to p$ (since $|np-([np]+1)| \le 1$), so $P\left[U_{([np]+1)} \le p-\epsilon\right] \to 0$ as $n \to \infty$.

Similarly,

$$P\left[U_{([np]+1)} \ge p + \epsilon\right] \to 0 \quad \text{as } n \to \infty.$$

Proof of (10.69): Since $(X_1, ..., X_n) \stackrel{\text{distn}}{=} (F^{-1}(U_1), ..., F^{-1}(U_n)),$ (10.73) $(X_{(1)}, ..., X_{(n)}) \stackrel{\text{distn}}{=} (F^{-1}(U_{(1)}), ..., F^{-1}(U_{(n)}))$ because *F* is increasing, so

(10.74)
$$X_{([np]+1)} \stackrel{\text{distn}}{=} F^{-1}(U_{([np]+1)}).$$

Since F^{-1} is assumed continuous at $p, X_{([np]+1)} \xrightarrow{p} F^{-1}(p)$ by Prop. 10.1.

Proposition 10.2.

(10.75)
$$\sqrt{n} \left(U_{([np]+1)} - p \right) \xrightarrow{d} N_1 \left(0, \, p(1-p) \right).$$

Heuristic proof: By (10.71), $G_n \approx G$. Thus

$$\sqrt{n} \left(U_{([np]+1)} - p \right) = \sqrt{n} \left[G_n^{-1}(p) - p \right]$$

$$\approx \sqrt{n} \left[G_n^{-1}(p) - G_n^{-1}(G_n(p)) \right]$$

$$\approx \sqrt{n} \left[p - G_n(p) \right] \qquad [\text{since } G_n^{-1} \approx G^{-1} = \text{ident.}]$$

$$\stackrel{d}{\to} N_1 \left(0, p(1-p) \right) \qquad [\text{by } (10.72)].$$

Rigorous proof: First, for any $t \in \mathbf{R}^1$,

$$\left\{\sqrt{n} \left(U_{([np]+1)} - p\right) \le t\right\} = \left\{U_{([np]+1)} \le p + \frac{t}{\sqrt{n}}\right\}$$
$$= \left\{[np] + 1 \text{ or more } U'_i s \le p + \frac{t}{\sqrt{n}}\right\}$$
$$= \left\{G_n\left(p + \frac{t}{\sqrt{n}}\right) \ge \frac{[np] + 1}{n}\right\}$$
$$= \left\{A_n + B_n + C_n \ge -t\right\}$$

where ((10.76) is verified below)

(10.76)
$$A_n \equiv \sqrt{n} \left[G_n \left(p + \frac{t}{\sqrt{n}} \right) - G_n(p) - \frac{t}{\sqrt{n}} \right] \xrightarrow{p} 0,$$

(10.77)
$$B_n \equiv \sqrt{n} \left[G_n(p) - p \right] \xrightarrow{d} N_1(0, p(1-p)), \quad \text{[by (10.72)]}$$

(10.78)
$$C_n \equiv \sqrt{n} \left[p - \left(\frac{[np] + 1}{n} \right) \right] = \frac{|np - ([np] + 1)|}{\sqrt{n}} \le \frac{1}{\sqrt{n}} \to 0.$$

Thus the asserted result follows by Slutsky and symmetry:

$$P\left[\sqrt{n} \left(U_{([np]+1)} - p\right) \le t\right] \to P[N_1(0, p(1-p)) \ge -t]$$

= $P[N_1(0, p(1-p)) \le t].$

To verify (10.76): if t > 0 then for $\sqrt{n} > t$,

$$A_n \stackrel{\text{distn}}{=} \sqrt{n} \left[\frac{\text{Binomial}(n, \frac{t}{\sqrt{n}})}{n} - \frac{t}{\sqrt{n}} \right],$$

hence

$$E(A_n) = 0,$$
 $Var(A_n) = \frac{t}{\sqrt{n}} \left(1 - \frac{t}{\sqrt{n}}\right) \to 0$ as $n \to \infty,$

so (10.76) follows from Chebyshev's inequality. The proof is similar if t < 0.

Proof of (10.70): By (10.74), (10.76), and propagation of error,

$$\sqrt{n} \left(X_{([np]+1)} - F^{-1}(p) \right) \stackrel{\text{distn}}{=} \sqrt{n} \left[F^{-1}(U_{([np]+1)}) - F^{-1}(p) \right]$$
$$\stackrel{d}{\to} N_1 \left(0, \left[(F^{-1})'(p) \right]^2 p(1-p) \right)$$
$$= N_1 \left(0, \frac{p(1-p)}{\left[f(F^{-1}(p)) \right]^2} \right) \quad [\text{since } F' = f].$$

Example 10.9. Asymptotic distribution of the sample median. When $p = \frac{1}{2}$, $F^{-1}(1/2) \equiv m$ is the population median, $X_{([n/2]+1)} \equiv \tilde{X}_n$ is the sample median, and (10.70) becomes

(10.79)
$$\sqrt{n} \left(\tilde{X}_n - m\right) \xrightarrow{d} N_1 \left(0, \frac{1}{4\left[f(m)\right]^2}\right).$$

This shows that the precision \equiv accuracy of the sample median as an estimator of the population median is directly proportional to f(m). This makes sense because the larger f(m) is, the more the observations X_i will accrue in the vicinity of m (see following figure):

sample median has high precision sample median has low precision

Remark 10.7. (a) If $f(F^{-1}(p)) = \infty$ or = 0 then (10.70) is inapplicable. In the first case $X_{([np]+1)}$ may converge to $F^{-1}(p)$ at a rate faster than $\frac{1}{\sqrt{n}}$, while in the second case it may converge at a slower rate or may not converge at all:

(b) The asymptotic normality (10.70) of the *p*-th sample quantile $X_{([np]+1)}$ is valid when p is fixed with $0 . It is not valid when <math>p \equiv p(n) \to 0$ or 1, e.g., the extreme order statistics $X_{(1)} \equiv X_{\min}$ and $X_{(n)} \equiv X_{\max}$ are not asymptotically normal. For example, in the Uniform(0,1) case of Example 9.1, the variance of these extreme order statistics are $O\left(\frac{1}{n^2}\right)$ rather than $O\left(\frac{1}{n}\right)$, thus to obtain a nontrivial limiting distribution they must be multiplied by an "inflation factor" n rather than \sqrt{n} – see Exercise 9.4.

10.7. Asymptotic efficiency of sample mean vs. sample median as estimators of the center of a symmetric distribution.

Suppose that X_1, \ldots, X_n is an i.i.d. sample from a distribution with pdf $f_{\theta}(x) \equiv f(x-\theta)$ on \mathbb{R}^1 , where θ is an unknown location parameter. Suppose that f is symmetric about 0, i.e., $f(x) = f(-x) \forall x$, so f_{θ} is symmetric about θ . Thus θ serves as both the population mean (provided it exists) and the population median. Thus it is natural to compare the sample mean \bar{X}_n and the sample median \tilde{X}_n as estimators¹⁵ of θ .

Suppose that $\tau^2 \equiv \operatorname{Var}(X_i) \equiv \int x^2 f(x) dx < \infty$ and f(0) > 0. Then from the CLT and (10.79),

(10.80)
$$\sqrt{n} \left(\bar{X}_n - \theta \right) \xrightarrow{d} N_1(0, \tau^2),$$

(10.81)
$$\sqrt{n} \left(\tilde{X}_n - \theta \right) \xrightarrow{d} N_1 \left(0, \frac{1}{4 \left[f(0) \right]^2} \right)$$
 [verify!]

¹⁵ Neither \bar{X}_n nor \tilde{X}_n need be the (asymptotically) optimal estimator of θ . The maximum likelihood estimator usually is asymptotically optimal (Theorem 14.9), and it need not be equivalent to \bar{X}_n or \tilde{X}_n ; e.g., if f is a Cauchy density (Example 14.5).

Thus the asymptotic efficiency of \tilde{X}_n relative to \bar{X}_n (as measured by the ratio of their asymptotic variances) is $4[f(0)]^2\tau^2$.

Example 10.10. If $X_i \sim N_1(\theta, 1)$ (so $\tau^2 = 1$), then $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, so

(10.82)
$$4[f(0)]^2 \tau^2 = \frac{2}{\pi} \approx 0.637,$$

hence \tilde{X}_n is (asymptotically) less efficient than \bar{X}_n : in the normal case the precision of the sample median based on n observations is about the same as that of the sample mean based on only .637n observations.

Actually, this should *not* be interpreted as a strong argument in favor of \bar{X}_n . The efficiency of the sample median \tilde{X}_n relative to \bar{X}_n is about 64%, which, while a significant loss, is not catastrophic. If, however, our assumption of a normal model is wrong, then the performance of the sample mean \bar{X}_n may itself be catastrophic. For example, if $f(x) \equiv \frac{1}{\pi(1+x^2)}$ is the standard Cauchy density then the asymptotic variance of $\sqrt{n}(\tilde{X}_n - \theta)$ is

(10.83)
$$\frac{1}{4[f(0)]^2} = \frac{\pi^2}{4} \approx 2.47,$$

but $\tau^2 = \infty$ so the asymptotic variance of \bar{X}_n is infinite. In fact, \bar{X}_n is not even a consistent estimator of θ . Because of this, we say that the sample median is a *robust* estimator of the location parameter θ (for heavy-tailed departures from normality), whereas the sample mean is not robust.

Note: in the Cauchy case the sample median, while robust, is not optimal: the MLE is better (see Exercise 14.36(i)). \Box

Exercise 10.8. Let $f(x) \equiv \frac{1}{2}e^{-|x|}$ be the standard double exponential density on \mathbf{R}^1 , so $f_{\theta}(x) = \frac{1}{2}e^{-|x-\theta|}$.

(i) Find the asymptotic efficiency of the sample median \tilde{X}_n relative to the sample mean \bar{X}_n as estimators of θ .

(ii) Find the MLE $\hat{\theta}$ for θ , i.e., the value of θ that maximizes the joint pdf

(10.84)
$$f_{\theta}(x_1, \dots, x_n) \equiv \frac{1}{2^n} \prod_{i=1}^n e^{-|x_i - \theta|}.$$