**Purpose:** These concise notes contain the definitions and results for Purdue University's course IE 230, "Probability and Statistics for Engineers, I".

The purpose of these notes is to provide a complete, clear, and concise compendium. The purpose of the lectures, textbook, homework assignments, and office hours is to help understand the meanings and implications of these notes via discussion and examples.

Essentially everything here is in Chapters 2–7 of the textbook, often in the highlighted blue boxes. Topic order roughly follows the textbook.

**Textbook:** D.C. Montgomery and G.C. Runger, *Applied Statistics and Probability for Engineers*, John Wiley & Sons, New York, 2007 (fourth edition).

**Table of Contents**

**Set-Theory Review**

A *set* is a collection of items; each such item is called a *member* of the set.

If a set $A$ has members $x$, $y$, and $z$, we can write $A = \{x, y, z\}$ and, for example, $x \in A$.

If a set has members defined by a condition, we write $A = \{x \mid x \text{ satisfies the condition}\}$. The vertical line is read "such that".

The largest set is the *universe*, the set containing all relevant items.

The smallest set is the *empty set* (or, sometimes, the *null set*), the set containing no items; it is denoted by $\varnothing$ or, occasionally, by $\{\}$.

If all members of a set $A$ are contained in a set $B$, then $A$ is a *subset* of $B$, written $A \subset B$.

If two sets $A$ and $B$ contain the same members, then they are equal, written $A = B$.

The *union* of two sets $A$ and $B$ is the set of items contained in at least one of the sets; that is, $A \cup B = \{x \mid x \in A \text{ or } x \in B\}$.

The *intersection* of two sets $A$ and $B$ is the set of items contained in both sets; that is, $A \cap B = \{x \mid x \in A \text{ and } x \in B\}$. This intersection is also written $AB$.

The *complement* of a set $A$ is the set of all items not contained in the set; that is, $A' = \{x \mid x \notin A\}$.

Sets $E_1$, $E_2$,..., $E_n$ *partition* the set $A$ if each member of $A$ lies in exactly one of the $n$ partitioning sets. (Equivalently, if their union is $A$ and their pairwise intersections are empty. That is, if $A = \cup_{i=1}^{n} E_i$ and $E_i \cap E_j = \varnothing$ for every pair $E_i$ and $E_j$.)

*Distributive Laws.* For any sets $A$, $B$, and $C$, $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ and $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$.

*DeMorgan's Laws.* For any sets $A$ and $B$, $(A \cup B)' = A' \cap B'$ and $(A \cap B)' = A' \cup B'$.

The *cardinal number* of a set $A$, denoted by $\#(A)$,is the number of members in $A$.

— A set $A$ is *finite* if $\#(A)$ is finite; otherwise $A$ is *infinite*.

— An infinite set $A$ is *countably infinite* if its members can be counted (to *count* a set means to assign a unique integer to each member);

— otherwise $A$ is *uncountably infinite* (e.g., the set of real numbers).

The *open interval* $(a, b)$ is the set $\{x \mid a < x < b\}$, all real numbers between, but not including, $a$ and $b$. Square brackets are used to include $a$ and/or $b$. Except for $[a, a]$, the *closed interval* containing only the one member $a$, non-empty intervals are uncountably infinite.

The *real-number line* is the open interval $(-\infty, \infty)$.

A *function* assigns a single *value* to each *argument*. The set of arguments is called the *domain* and the set of values is called the *range*. For example, $f(x)=x^2$ has the real-number line as its domain and $[0, \infty)$ as its range. For example, let the domain be a set of students, let the function be the weight of each student; then the range is the set of student weights.

**Probability Basics** (M&R Chapter 2)

All of probability and statistics depends upon the concept of a *random experiment*, defined to be a procedure that can result in a different *outcome* each time it is performed. Each *replication* of the experiment has exactly one outcome.

Experiments are sometimes classified into two types:

*Enumerative:* sampling from a well-defined finite set. (e.g., persons in the room).

*Analytical:* other sampling (e.g., next year's students).

A set containing all possible outcomes is called a *sample space*. An experiment can have many sample spaces; we try to choose the simplest one that is sufficient to answer the question at hand. Denote the chosen sample space by $S$.

A sample space is *discrete* if it has a finite or countably infinite number of members.

A set, say $E$, is an *event* if it is a subset of $S$; that is, if $E \subset S$. For a given replication of the experiment, $E$ *occurs* if it contains the outcome; otherwise it does not occur.

In practice, an event $E$ can be a verbal statement that is true if the event occurs and false if the event does not occur.

The complement of the event $E$ is the event $E'$.

—$E'$ occurs if (and only if) the outcome does not lie in $E$.

—The statement for $E'$ is the negation of the statement for $E$.

—For each replication, exactly one of $E$ and $E'$ occurs.

The intersection of events $E_1$ and $E_2$ is the event $E_1 \cap E_2$.

—$E_1 \cap E_2$ occurs if (and only if) the outcome lies in both $E_1$ and $E_2$.

—The statements for $E_1$ and $E_2$ must both be true for $E_1 \cap E_2$ to occur.

—The complement of $E_1 \cap E_2$ is $(E_1 \cap E_2)' = E'_1 \cup E'_2$.

That is, $(E_1 \cap E_2)'$ occurs if either $E_1$ or $E_2$ or both do not occur.

The union of events $E_1$ and $E_2$ is the event $E_1 \cup E_2$.

—$E_1 \cup E_2$ occurs if (and only if) the outcome lies in $E_1$, or in $E_2$, or in both.

—Either or both of the statements for $E_1$ and $E_2$ must be true for $E_1 \cup E_2$ to occur.

—The complement of $E_1 \cup E_2$ is $(E_1 \cup E_2)' = E'_1 \cap E'_2$.

That is, $(E_1 \cup E_2)'$ occurs if both $E_1$ and $E_2$ do not occur.

Every subset of $S$ is an event, so an experiment always has multiple events.

—The largest event is $S$, which always occurs.

— The smallest event is the empty set, $\varnothing$, which never occurs.

Definition. Two events, say $E_1$ and $E_2$, are *mutually exclusive* if they cannot both occur in the same replication of the experiment; that is, if $E_1 \cap E_2 = \varnothing$. More generally, $n$ events, say $E_1, E_2, \ldots, E_n$, are *mutually exclusive* if only one can occur in the same replication; that is, if $E_i \cap E_j = \varnothing$ for every pair of events.

The correspondence between set theory and probability theory is that the universe is the sample space, the items are outcomes, and sets are events.

The *probability* of an event $E$, denoted by $\mathrm{P}(E)$, is a numerical measure of how likely the event $E$ is to occur when the experiment is performed.

There are two commonly used interpretations of probability.

*Relative frequency*: If the experiment were repeated infinitely often, $\mathrm{P}(E)$ is the fraction of the replications in which $E$ occurs. (This interpretation is sometimes called "objective".)

*Subjective*: $\mathrm{P}(E)$ is a measure of belief about the likelihood that $E$ will occur in a particular replication.

Alternative statements of probability.

"The odds of $E$ are 2 in 3" is equivalent to "$\mathrm{P}(E) = 2/3$".

"The odds of $E$ are 2 to 3" is equivalent to "$\mathrm{P}(E) = 2/5$".

"$E$ has a 70% chance" is equivalent to "$\mathrm{P}(E) = 0.7$".

"$E$ has a 50–50 chance" is equivalent to "$\mathrm{P}(E) = 0.5$".

A baseball player batting 282 has hit successfully with relative frequency 0.282.

All results of probability follow from three *axioms*.

*Axiom* 1. $\mathrm{P}(S) = 1$. (That is, the probability of the "sure" event is one.)

*Axiom* 2. For every event $E$, $0 \le \mathrm{P}(E)$. (That is, probabilities are non-negative.)

*Axiom* 3. For all mutually exclusive events $E_1$ and $E_2$,

$$\mathrm{P}(E_1 \cup E_2) = \mathrm{P}(E_1) + \mathrm{P}(E_2) .$$

(That is, if two events cannot occur simultaneously, then the probability that one or the other occurs is the sum of their probabilities.)

Five useful probability results that are easily proven using the three axioms:

Result. (Complement) For every event $E$, $\mathrm{P}(E') = 1 - \mathrm{P}(E)$. (In particular, $\mathrm{P}(\varnothing) = 1 - \mathrm{P}(S) = 0$; that is, the "impossible" event has probability zero.)

Result. (Dominance) If $E_1 \subset E_2$, then $\mathrm{P}(E_1) \le \mathrm{P}(E_2)$. (That is, if two events differ only in that one contains more outcomes than the other, the larger event cannot be less likely.)

Result. (Axiom 3 extended to $n$ events) If events $E_1, E_2, \ldots, E_n$ are mutually exclusive, then

$$\mathrm{P}\left( \cup_{i=1}^{n} E_i \right) = \sum_{i=1}^{n} \mathrm{P}(E_i) .$$

(That is, if only one of the $n$ events can occur, then the probability that one of them does occur is the sum of their probabilities.)

Result. (Equally likely events) If equally likely events $E_1, E_2, \ldots, E_n$ partition the sample space, then $\mathrm{P}(E_i) = 1/n$ for $i = 1, 2, \ldots, n$.

Result. (Always true) For any two events $E_1$ and $E_2$,

$$\mathrm{P}(E_1 \cup E_2) = \mathrm{P}(E_1) + \mathrm{P}(E_2) - \mathrm{P}(E_1 \cap E_2) .$$

More generally, for any three events $E_1$, $E_2$, and $E_3$,

$$\begin{aligned} \mathrm{P}(E_1 \cup E_2 \cup E_3) = {} & \mathrm{P}(E_1) + \mathrm{P}(E_2) + \mathrm{P}(E_3) \\ & - \mathrm{P}(E_1 \cap E_2) - \mathrm{P}(E_1 \cap E_3) - \mathrm{P}(E_2 \cap E_3) \\ & + \mathrm{P}(E_1 \cap E_2 \cap E_3) . \end{aligned}$$

Yet more generally, for $n$ events continue to alternate signs.

Definition. The *conditional probability* of an event $E_1$, given that an event $B$ has occurred, is

$$\mathrm{P}(E_1 \mid B) \equiv \frac{\mathrm{P}(E_1 \cap B)}{\mathrm{P}(B)}.$$

—If $\mathrm{P}(B) = 0$, then $\mathrm{P}(E_1 \mid B)$ is undefined.

—The *given* event $B$ is assumed to have occurred; that is, the outcome is in $B$.

—The given event $B$ becomes the sample space.

—$\mathrm{P}(E_1) \equiv \mathrm{P}(E_1 \mid S)$ is the *unconditional* or *marginal* probability of $E$.

*Multiplication Rule*. For any nonempty events $B$ and $E_1$,

$$\mathrm{P}(B \cap E_1) = \mathrm{P}(B)\,\mathrm{P}(E_1 \mid B) = \mathrm{P}(E_1)\,\mathrm{P}(B \mid E_1).$$

*Baby Bayes's Rule*. For any events $E_1$ and $B$, if $\mathrm{P}(B) > 0$, then

$$\mathrm{P}(E_1 \mid B) = \frac{\mathrm{P}(B \mid E_1)\,\mathrm{P}(E_1)}{\mathrm{P}(B)}.$$

The full *Bayes's Rule* is obtained by expanding $\mathrm{P}(B)$ using Total Probability.

*Total Probability*. (Applying this rule is sometimes called *conditioning*.)

For any events $B$ and $E_1$,

$$\mathrm{P}(B) = \mathrm{P}(B \cap E_1) + \mathrm{P}(B \cap E'_1) = \mathrm{P}(B \mid E_1)\,\mathrm{P}(E_1) + \mathrm{P}(B \mid E'_1)\,\mathrm{P}(E'_1).$$

More generally, if events $E_1, E_2, \ldots, E_n$ partition $S$, then for any event $B$

$$\mathrm{P}(B) = \sum_{i=1}^{n} \mathrm{P}(B \cap E_i) = \sum_{i=1}^{n} \mathrm{P}(B \mid E_i)\,\mathrm{P}(E_i).$$

Result. (Independent events) The following four statements are equivalent; that is, either all are false or all are true. Often, Statement (2) is chosen to define independence.

(1) Events $A$ and $B$ are independent.

(2) $\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)$

(3) $\mathrm{P}(A \mid B) = \mathrm{P}(A)$

(4) $\mathrm{P}(B \mid A) = \mathrm{P}(B)$

Result. (Independence of complements) The following four statements are equivalent.

(1) Events $A$ and $B$ are independent.

(2) Events $A'$ and $B$ are independent.

(3) Events $A$ and $B'$ are independent.

(4) Events $A'$ and $B'$ are independent.

*Extended Multiplication Rule*. For any nonempty events $A_1, A_2, \ldots, A_n$,

$$\mathrm{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathrm{P}(A_1)\,\mathrm{P}(A_2 \mid A_1)\,\mathrm{P}(A_3 \mid A_1 \cap A_2) \cdots \mathrm{P}(A_n \mid A_1 \cap \cdots \cap A_{n-1}).$$

Definition. The $n$ events $A_1, A_2, \ldots, A_n$ are *(jointly) independent* if and only if

$$\mathrm{P}(A_{i1} \cap A_{i2} \cap \cdots \cap A_{ik}) = \mathrm{P}(A_{i1})\,\mathrm{P}(A_{i2}) \cdots \mathrm{P}(A_{ik}),$$

for every subset $A_{i1}, A_{i2}, \ldots, A_{ik}$ of the $n$ events, for $k = 2, 3, \ldots, n$.

A weaker form of independence is *pairwise independence*, which requires only that every pair of events be independent ($k = 2$ in the definition).

**Discrete Random Variables and Probability Distributions** (M&R Chapter 3)

Definition. A *random variable* is a function that assigns a real number to each outcome in the sample space of an experiment.

—Traditionally, we denote random variables by upper-case letters toward the end of the English alphabet; e.g., $X$.

—In practice, a random variable can be a verbal statement; e.g., the experiment is to select a random student and $X$ = "the student's grade point."

—Events can be constructed from random variables; e.g., "$X > 3.0$" is an event (defined by the set of students whose grade point is greater than 3.0).

Definition. The *probability distribution* of a random variable $X$ is a description (in whatever form) of the likelihoods associated with the values of $X$.

Definition. The *cumulative distribution function*, often abbreviated *cdf*, of a random variable $X$ is $F(x) = \mathrm{P}(X \le x)$ for every real number $-\infty < x < \infty$.

Result. For every random variable $X$, if $a \le b$, then $\mathrm{P}(a < X \le b) = F(b) - F(a)$.

Result. $F(-\infty) = 0$ and $F(\infty) = 1$. More generally, if $x \le y$, then $F(x) \le F(y)$.

(That is, every cdf is nondecreasing. This is a special case of $\mathrm{P}(A) \le \mathrm{P}(B)$ if $A \subset B$.)

Definition. A random variable is *discrete* if its range is finite or countably infinite. A random variable is *continuous* if its range is uncountably infinite.

—Often, discrete random variables arise from *counting*.

—Often, continuous random variables arise from *measuring*.

Definition. For a discrete random variable $X$, the *probability mass function*, often abbreviated *pmf*, is $f(x) = \mathrm{P}(X = x)$ for every real number $-\infty < x < \infty$.

(Notice that the cdf is denoted by an upper-case $F$, whereas the pmf is denoted by a lower-case $f$. Later, when dealing with more than one random variable, we will use the more-explicit notation $F_X$ and $f_X$.)

Result. For a discrete random variable $X$ having possible values $x_1, x_2, \ldots, x_n$, the cdf is

$$F(x) = \sum_{\text{all } x_i \le x} f(x_i) \quad \text{for every real number} \quad -\infty < x < \infty .$$

Definition. For a discrete random variable $X$ having possible values $x_1, x_2, \ldots, x_n$, the *mean* or *expected value* is the constant $\mathrm{E}(X) = \sum_{i=1}^{n} x_i \, f(x_i)$.

—Traditionally, the mean is also denoted by $\mu$ or, more explicitly, by $\mu_X$.

—Because $\sum_{i=1}^{n} f(x_i) = 1$, the mean is the *first moment*, or *center of gravity*.

Definition. The *variance* of a random variable $X$ is the constant $\mathrm{V}(X) = \mathrm{E}[(X - \mu)^2]$.

—Traditionally, the variance is also denoted by $\sigma^2$ or, more explicitly, by $\sigma_X^2$.

—The variance is the *second moment about the mean*, or *moment of inertia*.

Result. For a discrete random variable $X$ having possible values $x_1, x_2, \ldots, x_n$,

$$\mathrm{V}(X) = \sum_{i=1}^{n} (x_i - \mu)^2 \, f(x_i) = \sum_{i=1}^{n} x_i^2 \, f(x_i) - \mu^2 = \mathrm{E}(X^2) - \mu^2 .$$

Definition. The *standard deviation* of $X$ is the constant $\sigma \equiv \sigma_X \equiv +\sqrt{\mathrm{V}(X)}$.

Definition. A random variable $X$ has a *discrete uniform* distribution if each of the $n$ numbers in its range, say $x_1, x_2, ..., x_n$, has equal probability.

Result. Suppose that $X$ has a discrete uniform distribution on the $n > 1$ equally spaced numbers from $a$ to $b$. Then the mean and variance of $X$ are

$$\mathrm{E}(X) = \frac{a+b}{2} \quad \text{and} \quad \mathrm{V}(X) = \left[\frac{(b-a)^2}{12}\right]\left[\frac{n+1}{n-1}\right].$$

Definition. A sequence of *Bernoulli trials* has three properties:

(i) Each trial has exactly two outcomes (often called "success" and "failure").

(ii) Each trial has $\mathrm{P}(\text{success}) = p$, which is a constant.

(iii) Each trial is independent of every other trial.

Definition. A *binomial experiment* is an experiment composed of $n$ Bernoulli trials.

Definition. An ordering of $r$ elements from a set of $n$ elements is called a *permutation*.

Definition. A selection (without regard to order) of $r$ elements from a set of $n$ elements is called a *combination*.

Result. A set of $n$ elements has $n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$ permutations.
(An empty set has $0! = 1$ permutation.)

Result. The number of permutations of $r$ elements from a set of $n$ elements is

$$P_r^n = \frac{n!}{(n-r)!} \quad \text{for } r = 0, 1, ..., n.$$

Result. The number of combinations of $r$ elements from a set of $n$ elements is

$$C_r^n = \binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{P_r^n}{r!} \quad \text{for } r = 0, 1, ..., n.$$

Definition (not precise). Suppose that counts occur at random throughout a real-number interval (typically of time or space). The random experiment is called a *(homogeneous) Poisson process* if the interval can be partitioned into equal-length non-overlapping subintervals of small enough length that

(i) the probability of more than one count in a subinterval is zero,

(ii) the probability of one count in a subinterval is the same for all subintervals and proportional to the length of the subintervals, and

(iii) the count in each subinterval is independent of other subintervals.

(An example: Customer arrivals. Condition (i) says that customers arrive separately; Condition (ii) says that the arrival rate is constant (that is, homogeneous) and that the probability of a customer arriving during a two-second subinterval is twice the probability of a customer arriving during a one-second subinterval; Condition (iii) says that the number of customers in one subinterval is not useful in predicting the number of customers in another subinterval.)

Result: *Poisson Approximation to the Binomial Distribution.* The Poisson process is (asymptotically) equivalent to "many", $n$, Bernoulli trials, each with a "small" probability of success, $p$. Therefore, the Poisson distribution with mean $np$ is a good approximation to the binomial distribution when $n$ is large and $p$ is small.

Discrete Distributions: Summary Table

| random variable | distribution name | range | probability mass function | expected value | variance |
|---|---|---|---|---|---|
| $X$ | general | $x_1, x_2, \ldots, x_n$ | $\mathrm{P}(X = x)$<br>$= f(x)$<br>$= f_X(x)$ | $\sum_{i=1}^{n} x_i f(x_i)$<br>$= \mu = \mu_X$<br>$= \mathrm{E}(X)$ | $\sum_{i=1}^{n} (x_i - \mu)^2 f(x_i)$<br>$= \sigma^2 = \sigma_X^2$<br>$= \mathrm{V}(X)$<br>$= \mathrm{E}(X^2) - \mu^2$ |
| $X$ | discrete uniform | $x_1, x_2, \ldots, x_n$ | $1/n$ | $\sum_{i=1}^{n} x_i / n$ | $[\sum_{i=1}^{n} x_i^2 / n] - \mu^2$ |
| $X$ | equal-space uniform | $x = a, a+c, \ldots, b$ where $n = (b-a+c)/c$ | $1/n$ | $\frac{a+b}{2}$ | $\frac{c^2(n^2-1)}{12}$ |
| "# successes in 1 Bernoulli trial" | indicator variable | $x = 0, 1$ | $p^x (1-p)^{1-x}$ | $p$<br>where $p = \mathrm{P}(\text{"success"})$ | $p(1-p)$ |
| "# successes in $n$ Bernoulli trials" | binomial | $x = 0, 1, \ldots, n$ | $C_x^n p^x (1-p)^{n-x}$ | $np$<br>where $p = \mathrm{P}(\text{"success"})$ | $np(1-p)$ |
| "# successes in a sample of size $n$ from a population of size $N$ containing $K$ successes" | hyper-geometric (sampling without replacement) | $x = (n-(N-K))^+, \ldots, \min\{K, n\}$ and integer | $C_x^K C_{n-x}^{N-K} / C_n^N$ | $np$<br>where $p = K/N$ | $np(1-p)\frac{(N-n)}{(N-1)}$ |
| "# Bernoulli trials until 1st success" | geometric | $x = 1, 2, \ldots$ | $p(1-p)^{x-1}$ | $1/p$<br>where $p = \mathrm{P}(\text{"success"})$ | $(1-p)/p^2$ |
| "# Bernoulli trials until $r$th success" | negative binomial | $x = r, r+1, \ldots$ | $C_{r-1}^{x-1} p^r (1-p)^{x-r}$ | $r/p$<br>where $p = \mathrm{P}(\text{"success"})$ | $r(1-p)/p^2$ |
| "# of counts in time $t$ from a Poisson process with rate $\lambda$" | Poisson | $x = 0, 1, \ldots$ | $e^{-\mu} \mu^x / x!$ | $\mu$<br>where $\mu = \lambda t$ | $\mu$ |

Result. For $x = 1, 2, \ldots$, the geometric cdf is $F_X(x) = 1 - (1-p)^x$.

Result. The geometric distribution is the only discrete memoryless distribution. That is, $\mathrm{P}(X > x + c \mid X > x) = \mathrm{P}(X > c)$.

Result. The binomial distribution with $p = K/N$ is a good approximation to the hypergeometric distribution when $n$ is small compared to $N$.

**Continuous Random Variables and Probability Distributions** (M&R Chapter 4)

Comment. Concepts of, and notations for, the continuous case are analogous to the discrete case, with integrals replacing sums.

Definition. For a continuous random variable $X$, the *probability density function*, often abbreviated *pdf*, is a function satisfying

(i) $f(x) \geq 0$ for every real number $x$,

(ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$, and

(iii) $\mathrm{P}(a \leq X \leq b) = \int_{a}^{b} f(x)dx$ for all real numbers $a \leq b$.

(The pdf is analogous to the pmf. The pmf is a probability, however, while the pdf must be integrated to obtain a probability.)

Result. If $X$ is continuous, then $\mathrm{P}(X = x) = 0$ for every real number $x$.

(Therefore, e.g., $\mathrm{P}(a < X) = \mathrm{P}(a \leq X)$, which is different from the discrete case.)

Result. For a continuous random variable $X$, the cdf is $F(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f(u)\,du$ .

Definition. The *mean* of a continuous random variable $X$ is the constant

$$\mu \equiv \mu_X \equiv \mathrm{E}(X) = \int_{-\infty}^{\infty} x\, f(x)\,dx \ .$$

Definition. The *variance* of a random variable $X$ is the constant $\mathrm{V}(X) = \mathrm{E}[(X - \mu)^2]$. (Unchanged from the discrete case.)

Result. For a continuous random variable $X$

$$\sigma^2 \equiv \sigma_X^2 \equiv \mathrm{V}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\,dx = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2 = \mathrm{E}(X^2) - \mu^2 \ .$$

Definition. The *standard deviation* of $X$ is the constant $\sigma \equiv \sigma_X \equiv +\sqrt{\mathrm{V}(X)}$.
(Unchanged from the discrete case.)

Definition. A random variable $X$ has a *continuous uniform* (or *rectangular*) distribution if its pdf forms a rectangle with base $[a, b]$.

Result. If $X$ has a continuous uniform distribution on $[a, b]$, then

(i) the mean of $X$ is $\mathrm{E}(X) = (a + b)/2$, and

(ii) the variance of $X$ is $\mathrm{V}(X) = (b - a)^2/12$.

(Notice the analogy to the discrete uniform distribution.)

Definition. A random variable $X$ has a *triangular* distribution if its pdf forms a triangle with base $[a, b]$ and mode at $m$, where $a \leq m \leq b$.

Result. If $X$ has a triangular distribution on $[a, b]$, with mode at $m$, then

(i) The mean of $X$ is $\mathrm{E}(X) = (a + m + b)/3$.

(ii) The variance of $X$ is $\mathrm{V}(X) = [(b - a)^2 - (m - a)(b - m)]/18$.

The *normal* (or *Gaussian* ) distribution (see table) with mean $\mu$ and standard deviation $\sigma$
— is often an adequate model for the sum of many random variables (as a result of the "Central Limit Theorem"),
— has a symmetric bell-shaped pdf centered at $\mu$, points of inflection at $\mu - \sigma$ and $\mu + \sigma$, and range $(-\infty, \infty)$,
— is the only famous distribution for which the general notation $\mu$ and $\sigma$ are used directly as the distribution's mean and standard deviation.
— $X \sim N(\mu, \sigma^2)$ is read "$X$ is normally distributed with mean $\mu$ and variance $\sigma^2$".

Definition. The *standard normal* distribution is the special case of $\mu = 0$ and $\sigma = 1$. This random variable is usually denoted by $Z$, its cdf by $\Phi$, and its pdf by $\phi$.

Result. To convert between general and standardized normal distributions, use...

— If $X \sim N(\mu, \sigma^2)$, then $(X - \mu)/\sigma \sim N(0, 1)$.

— If $Z \sim N(0, 1)$, then $(\mu + \sigma Z) \sim N(\mu, \sigma^2)$.

Probability calculations for the normal distribution are not closed form. Use numerical methods, approximations, or tabled values.

— For $X \sim N(\mu, \sigma^2)$, relate $p$ and $x_p$, where $p = \mathrm{P}(X \le x_p) \equiv F_X(x_p)$.
° MSExcel: Given $p$, use "norminv"; given $z_p$, use "normdist".
° Sketch the normal density and visually approximate the relevant area. As an aid, remember that

(i) $\mathrm{P}(\mu - \sigma < X < \mu + \sigma) \approx 0.68$
(ii) $\mathrm{P}(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.95$
(iii) $\mathrm{P}(\mu - 3\sigma < X < \mu + 3\sigma) \approx 0.997.$

° Probabilities can be converted to standard normal probabilities using

$$F_X(x_p) \equiv \mathrm{P}(X \le x_p) = \mathrm{P}\left[\frac{X - \mu}{\sigma} \le \frac{x_p - \mu}{\sigma}\right] = \mathrm{P}(Z \le z_p) \equiv \Phi(z_p),$$

where $z_p = \dfrac{x_p - \mu}{\sigma}$ is the $z$-value obtained by *standardizing* $X = x_p$.

— For $Z \sim N(0, 1)$, relate $p$ and $z_p$, where $p = \mathrm{P}(Z \le z_p) \equiv \Phi(z_p)$.
° MSExcel: Given $p \in [0, 1]$, use "normsinv"; given $z_p$, use "normsdist".
° Table: $p = \Phi(z_p)$ for $z_p = a, a + b, \ldots, c - b, c$.
(Table III of M&R uses $a = -4$, $b = 0.01$, and $c = 4$.)
° One-line Approximations:

$$z_p \approx [p^{0.135} - (1-p)^{0.135}] / 0.1975$$
$$p \approx 1. / [1. + \exp(-z_p \times (1.5966 + (z_p^2 / 14.)))]$$

Definition. The *order statistics* from a sample $x_1, x_2, \ldots, x_n$ are the sorted values $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$. That is, $x_{(1)}$ is the minimum observation and $x_{(n)}$ is the maximum. (This definition is general, not dependent upon normality.)

A graphical test of normality. Given a sample of size $n$, plot each order statistic $x_{(j)}$ against its *standardized normal score* $z_p$, the $p = (j - 0.5)/n$ quantile. If the resulting curve is approximately a straight line, then the sample data are consistent with the observations having been drawn from a normal distribution.

Result. *Normal approximation to the binomial distribution.* The normal $(n\,p, n\,p\,(1-p))$ distribution is a good approximation to the binomial $(n, p)$ distribution when $\min\{\mu_X, n-\mu_X\} = n\,\min\{p, (1-p)\}$ is "large". (A common value for "large" is 5; the approximation is asymptotically exact. See also "continuity correction".)

Result. *Normal approximation to the Poisson distribution.* The normal $(\lambda, \lambda)$ distribution is a good approximation to the Poisson $(\lambda)$ distribution when $\mu_X = \lambda$ is "large". (A common value for "large" is 5; the approximation is asymptotically exact. See also "continuity correction".)

Definition. *Continuity correction.* Rewriting an event to reduce the error from approximating a discrete distribution with a continuous distribution.

Let $X$ be a discrete random variable. For simplicity, assume that the possible values of $X$ are integers. Then, for integer constants $a$ and $b$,

$$\mathrm{P}(a \le X \le b) = \mathrm{P}(a - 0.5 \le X \le b + 0.5).$$

The continuity-correction approximation of $\mathrm{P}(a \le X \le b)$ is the continuous-distribution value of $\mathrm{P}(a - 0.5 \le X \le b + 0.5)$.

(The correction is crucial when $a = b$, but is less important as $b - a$ increases.)

Result. From any time $t$, let $X$ denote the time until the next count of a Poisson Process having rate $\lambda$. Then $X$ has an exponential distribution with mean $1/\lambda$. (Here $t$ can be any time, including zero or the time of a particular count.) (See Table.)

Result. For any time $t$, let $X$ denote the time until the $r$th count of a Poisson Process having rate $\lambda$. Then $X$ is the sum of $r$ independent exponential random variables and $X$ has an Erlang distribution with parameters $r$ and $\lambda$. (See Table.)

Analogy. The exponential distribution is analogous to the geometric distribution; the Erlang distribution is analogous to the negative binomial distribution.

Definition. The *gamma* distribution is the generalization of the Erlang distribution in which the parameter $r > 0$ is not necessarily an integer. (See Table.)

Result. If $Y$ is an exponential random variable with mean 1, then $X = \delta\, Y^{1/\beta}$ is *Weibull* with parameters $\delta > 0$ and $\beta > 0$. (See Table.)

Definitions. Three types of distribution parameters:

— A *location* parameter is additive: $Y = a + X$. The distribution of $Y$ is identical to that of $X$ except that its location is shifted $a$ units to the right.

— A *scale* parameter is multiplicative: $Y = b\,X$. The distribution of $Y$ is identical to that of $X$, except that each unit of $Y$ is $b$ units of $X$; the location of zero is unchanged.

— A *shape* parameter is nonlinear: $Y = g(X; c)$, where the function $g$ is nonlinear in $c$. The distributions of $Y$ and $X$ have different shapes.

Result. *Chebyshev's Inequality*. The probability that a random variable $X$ differs from its mean by at least $c$ standard deviations is no more than $1/c^2$. Notationally,

$$\mathrm{P}(|X - \mu| \ge c\,\sigma) \le 1/c^2$$

for every constant $c > 0$. Equivalently, the inequality can be written as

$$\mathrm{P}(\mu - c\,\sigma \le X \le \mu + c\,\sigma) \ge 1 - 1/c^2.$$

(For example, every distribution has at least $8/9$ of the probability within three standard deviations of the mean. Chebyshev's inequality holds for all distributions, but seldom is a good approximation for a particular distribution.)

Continuous Distributions: Summary Table

| random variable | distribution name | range | cumulative distrib. func. | probability density func. | expected value | variance |
|---|---|---|---|---|---|---|
| $X$ | general | $(-\infty, \infty)$ | $P(X \le x)$<br>$= F(x)$<br>$= F_X(x)$ | $\left.\frac{dF(y)}{dy}\right\|_{y=x}$<br>$= f(x)$<br>$= f_X(x)$ | $\int_{-\infty}^{\infty} x f(x)\,dx$<br>$= \mu = \mu_X$<br>$= E(X)$ | $\int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx$<br>$= \sigma^2 = \sigma_X^2$<br>$= V(X)$<br>$= E(X^2) - \mu^2$ |
| $X$ | continuous uniform | $[a, b]$ | $\frac{x-a}{b-a}$ | $\frac{1}{b-a}$ | $\frac{a+b}{2}$ | $\frac{(b-a)^2}{12}$ |
| $X$ | triangular | $[a, b]$ | $(x-a)f(x)/2$ if $x \le m$, else $1-(b-x)f(x)/2$ | $\frac{2(x-d)}{(b-a)(m-d)}$ ($d = a$ if $x \le m$, else $d = b$) | $\frac{a+m+b}{3}$ | $\frac{(b-a)^2-(m-a)(b-m)}{18}$ |
| sum of random variables | normal (or Gaussian) | $(-\infty, \infty)$ | Table III | $\frac{e^{\frac{-1}{2}\left[\frac{x-\mu}{\sigma}\right]^2}}{\sqrt{2\pi}\,\sigma}$ | $\mu$ | $\sigma^2$ |
| time to Poisson count 1 | exponential | $[0, \infty)$ | $1 - e^{-\lambda x}$ | $\lambda e^{-\lambda x}$ | $1/\lambda$ | $1/\lambda^2$ |
| time to Poisson count $r$ | Erlang | $[0, \infty)$ | $\sum_{k=r}^{\infty} \frac{e^{-\lambda x}(\lambda x)^k}{k!}$ | $\frac{\lambda^r x^{r-1} e^{-\lambda x}}{(r-1)!}$ | $r/\lambda$ | $r/\lambda^2$ |
| lifetime | gamma | $[0, \infty)$ | numerical | $\frac{\lambda^r x^{r-1} e^{-\lambda x}}{\Gamma(r)}$ | $r/\lambda$ | $r/\lambda^2$ |
| lifetime | Weibull | $[0, \infty)$ | $1 - e^{-(x/\delta)^\beta}$ | $\frac{\beta x^{\beta-1} e^{-(x/\delta)^\beta}}{\delta^\beta}$ | $\delta\Gamma(1+\frac{1}{\beta})$ | $\delta^2\Gamma(1+\frac{2}{\beta})-\mu^2$ |

Definition. For any $r > 0$, the *gamma function* is $\Gamma(r) = \int_0^\infty x^{r-1} e^{-x}\,dx$.

Result. $\Gamma(r) = (r-1)\Gamma(r-1)$. In particular, if $r$ is a positive integer, then $\Gamma(r) = (r-1)!$.

Result. The exponential distribution is the only continuous memoryless distribution. That is, $P(X > x + c \mid X > x) = P(X > c)$.

Definition. A *lifetime* distribution is continuous with range $[0, \infty)$.

Modeling lifetimes. Some useful lifetime distributions are the exponential, Erlang, gamma, and Weibull.

**Joint Probability Distributions** (M&R Chapter 5)

Comment. The topic is now two (or more) random variables defined on the same sample space. Concepts and notation are analogous to those from Chapters 2–4. In particular, independence of random variables is analogous to independence of events, lower-case letters denote constants and upper-case letters denote random variables, except that lower-case $f$ denotes pmf's and pdf's, and upper-case $F$ denotes cdf's. Subscripts become important.

Definition. A *random vector* is a vector whose components are (scalar) random variables defined on a common sample space.

Definition. The *joint probability distribution* of a random vector $(X_1, X_2, \ldots, X_n)$ is a description (in whatever form) of the likelihoods associated with the possible values of $(X_1, X_2, \ldots, X_n)$. (When we consider $n = 2$, the two random variables often will be denoted by $X$ and $Y$. Definitions are given here only for $n = 2$; they extend to general values of $n$ by analogy, as done in Section 5.2.)

Definition. The *joint cumulative distribution function (cdf)* of the random vector $(X, Y)$ is, for all real numbers $x$ and $y$,

$$F_{XY}(x, y) = \mathrm{P}(X \le x, Y \le y),$$

where the comma denotes intersection and is read as "and".

Result. For all real numbers $a$, $b$, $c$, and $d$,

$$\mathrm{P}(a < X \le b, c < Y \le d) = F_{XY}(b, d) - F_{XY}(b, c) - F_{XY}(a, d) + F_{XY}(a, c)$$

Definition. The *joint probability mass function (pmf)* of the discrete random vector $(X, Y)$ is, for all real numbers $x$ and $y$,

$$f_{XY}(x, y) = \mathrm{P}(X = x, Y = y).$$

Result. Every joint pmf satisfies

(1) $f_{XY}(x, y) \ge 0$, for all real numbers $x$ and $y$, and

(2) $\sum_{\text{all } x} \sum_{\text{all } y} f_{XY}(x, y) = 1.$

Definition. The *marginal distribution* of $X$ is the distribution of $X$ alone, unconditional on $Y$ (just as discussed in Chapters 3 and 4.)

Result. The *marginal cdf* of $X$ is, for every real number $x$,

$$F_X(x) = F_{XY}(x, \infty).$$

Result. If $(X, Y)$ is a discrete random vector, then the *marginal pmf* of $X$ is

$$f_X(x) = \mathrm{P}(X = x) = \sum_{\text{all } y} f_{XY}(x, y) \text{ for every real number } x,$$

the marginal mean of $X$ is

$$\mathrm{E}(X) = \mu_X = \sum_{\text{all } x} x\, f_X(x) = \sum_{\text{all } x} \sum_{\text{all } y} x\, f_{XY}(x, y)$$

and the marginal variance of $X$ is

$$\mathrm{V}(X) = \sigma_X^2 = \sum_{\text{all } x} (x - \mu_X)^2 f_X(x) = \sum_{\text{all } x} \sum_{\text{all } y} (x - \mu_X)^2 f_{XY}(x, y).$$

Analogous results hold for the marginal cdf, pmf, mean and variance of $Y$.

Definition. If $(X, Y)$ is a discrete random vector with joint pmf $f_{XY}$, then the *conditional pmf* of $Y$ given $X = x$ is

$$f_{Y \mid X=x}(y) = \frac{f_{XY}(x,y)}{f_X(x)} \quad \text{if } f_X(x) > 0.$$

The conditional pmf of $X$ given $Y = y$ is defined analogously.

Result. For every real number $x$ at which $f_{Y \mid X=x}$ is defined,

(1) $f_{Y \mid X=x}(y) \geq 0$ for every real number $y$,

(2) $\sum_{\text{all } y} f_{Y \mid X=x}(y) = 1$, and

(3) $\mathrm{P}(Y = y \mid X = x) = f_{Y \mid X=x}(y)$ for every real number $y$.

Result. The conditional mean and conditional variance of $Y$ given $X = x$ are

$$\mathrm{E}(Y \mid X = x) = \mu_{Y \mid X=x} = \sum_{\text{all } y} y\, f_{Y \mid X=x}(y)$$

and

$$\mathrm{V}(Y \mid X = x) = \sigma^2_{Y \mid X=x} = \sum_{\text{all } y} (y - \mu_{Y \mid X=x})^2 f_{Y \mid X=x}(y) = \mathrm{E}(Y^2 \mid X = x) - \mu^2_{Y \mid X=x}.$$

Result. If $(X, Y)$ is a discrete random vector, then

$$f_{XY}(x,y) = f_{X \mid Y=y}(x)\, f_Y(y) = f_{Y \mid X=x}(y)\, f_X(x) \quad \text{for all } x \text{ and } y.$$

Definition. If $(X, Y)$ is a discrete random vector, then $X$ and $Y$ are *independent* if and only if $f_{XY}(x,y) = f_X(x)\, f_Y(y)$ for all $x$ and $y$.

Result. If $(X, Y)$ is a discrete random vector, then the following four statements are equivalent; that is, either none are true or all are true.

(1) $X$ and $Y$ are independent.

(2) $f_{Y \mid X=x}(y) = f_Y(y)$ for all $x$ and $y$ with $f_X(x) > 0$.

(3) $f_{X \mid Y=y}(x) = f_X(x)$ for all $x$ and $y$ with $f_Y(y) > 0$.

(4) $\mathrm{P}(X \in A, Y \in B) = \mathrm{P}(X \in A)\,\mathrm{P}(Y \in B)$ for all subsets $A$ and $B$ of $(-\infty, \infty)$.

Definition. A *multinomial experiment* is composed of $n$ trials satisfying

(1) each trial has exactly one of $k$ outcomes,

(2) the probability of outcome $i$ is $p_i$ for $i = 1, 2, ..., k$
(and therefore $p_1 + p_2 + \cdots + p_k = 1$), and

(3) the trials are independent.

Definition. In a multinomial experiment, let $X_i$ denote the number of trials that result in outcome $i$ for $i = 1, 2, ..., k$. (Then $X_1 + X_2 + \cdots + X_k = n$.) The random vector $(X_1, X_2, \ldots, X_k)$ has a *multinomial distribution* with joint pmf

$$\mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{x_1!\, x_2!\, \cdots\, x_k!}\, p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

when each $x_i$ is a nonnegative integer and $x_1 + x_2 + \cdots + x_k = n$; zero elsewhere.

Result. If the random vector $(X_1, X_2, \ldots, X_k)$ has a multinomial distribution, then the marginal distribution of $X_i$ is binomial with parameters $n$ and $p_i$ for $i = 1, 2, ..., k$. (And therefore $\mathrm{E}(X_i) = np_i$ and $\mathrm{V}(X_i) = np_i(1-p_i)$.)

Comment. The topic is now two (or more) continuous random variables defined on the same sample space. Concepts and notation are analogous to those from Chapters 2–4 and to Sections 5.1–5.2. All previous cdf results hold for both discrete and continuous random variables; they are not repeated here.

Definition. The *joint probability density function (pdf)* of the continuous random vector $(X, Y)$ is, for all real numbers $x$ and $y$, denoted by $f_{XY}(x, y)$ and satisfies

$$\mathrm{P}((X, Y) \in R) = \iint_R f_{XY}(x, y)\, dx\, dy\ ,$$

for every region $R$ in two-dimensional space.

Result. Every joint pdf satisfies

(1) $f_{XY}(x, y) \geq 0$ for all real numbers $x$ and $y$, and

(2) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) = 1.$

Result. If $(X, Y)$ is a continuous random vector, then the *marginal pdf* of $X$ is

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy \quad \text{for every real number } x\ .$$

Analogously, the *marginal pdf* of $Y$ is

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx \quad \text{for every real number } y\ .$$

Result. The marginal mean and variance of a continuous random variable $X$ having marginal pdf $f_X$ are

$$\mathrm{E}(X) = \mu_X = \int_{-\infty}^{\infty} x\, f_X(x)\, dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, f_{XY}(x, y)\, dy\, dx$$

and

$$\mathrm{V}(X) = \sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)\, dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)^2 f_{XY}(x, y)\, dy\, dx\ .$$

Analogous results hold for the marginal mean and variance of $Y$.

Definition. If $h(X, Y)$ is a scalar function of the continuous random vector $(X, Y)$, then the expected value of $h(X, Y)$ is

$$\mathrm{E}[h(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y)\, f_{XY}(x, y)\, dy\, dx\ .$$

If $X$ and/or $Y$ is discrete, replace the corresponding integral with a summation.

Definition. If $(X, Y)$ is a continuous random vector with joint pdf $f_{XY}$, then the *conditional pdf* of $Y$ given $X = x$ is

$$f_{Y \mid X=x}(y) = \frac{f_{XY}(x,y)}{f_X(x)} \quad \text{if } f_X(x) > 0.$$

The conditional pdf of $X$ given $Y = y$ is defined analogously.

Result. For every real number $x$ at which $f_{Y|x}$ is defined,

(1) $f_{Y \mid X=x}(y) \geq 0$ for every real number $y$,

(2) $\int_{-\infty}^{\infty} f_{Y \mid X=x}(y)\, dy = 1$, and

(3) $\mathrm{P}(Y \in B \mid X = x) = \int_B f_{Y \mid X=x}(y)\, dy$ for every subset $B$ of $(-\infty, \infty)$.

Result. The conditional mean and conditional variance of $Y$ given $X = x$ are

$$\mathrm{E}(Y \mid X = x) = \mu_{Y \mid X=x} = \int_{-\infty}^{\infty} y\, f_{Y \mid X=x}(y)\, dy$$

and

$$\mathrm{V}(Y \mid X = x) = \sigma^2_{Y \mid X=x} = \int_{-\infty}^{\infty} (y - \mu_{Y \mid X=x})^2 f_{Y \mid X=x}(y)\, dy = \mathrm{E}(Y^2 \mid X = x) - \mu^2_{Y \mid X=x}.$$

Result. If $(X, Y)$ is a continuous random vector, then

$$f_{XY}(x,y) = f_{X \mid Y=y}(x)\, f_Y(y) = f_{Y \mid X=x}(y)\, f_X(x) \quad \text{for all } x \text{ and } y.$$

Definition. If $(X, Y)$ is a continuous random vector, then $X$ and $Y$ are *independent* if and only if $f_{XY}(x,y) = f_X(x)\, f_Y(y)$ for all $x$ and $y$.

Result. If $(X, Y)$ is a continuous random vector, then the following four statements are equivalent; that is, either none are true or all are true.

(1) $X$ and $Y$ are independent.

(2) $f_{Y \mid X=x}(y) = f_Y(y)$ for all $x$ and $y$ with $f_X(x) > 0$.

(3) $f_{X \mid Y=y}(x) = f_X(x)$ for all $x$ and $y$ with $f_Y(y) > 0$.

(4) $\mathrm{P}(X \in A, Y \in B) = \mathrm{P}(X \in A)\, \mathrm{P}(Y \in B)$ for all subsets $A$ and $B$ of $(-\infty, \infty)$.

Definition. If $(X, Y)$ is a random vector, then the *covariance* of $X$ and $Y$, denoted by $\text{cov}(X, Y)$ or by $\sigma_{XY}$, is

$$\text{cov}(X, Y) = \sigma_{XY} = \text{E}[(X - \mu_X)(Y - \mu_Y)] = \text{E}(XY) - \mu_X \mu_Y$$

Definition. If $(X, Y)$ is a random vector, then the *correlation* of $X$ and $Y$, denoted by $\text{corr}(X, Y)$ or by $\rho_{XY}$, is

$$\text{corr}(X, Y) = \rho_{XY} = \frac{\text{cov}(X, Y)}{\sqrt{\text{V}(X)\text{V}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Result. If $(X, Y)$ is a random vector, then $-1 \le \rho_{XY} \le 1$.

Result. If all observations of $(X, Y)$ lie in a straight line, then $|\rho_{XY}| = 1$.

Result. If $X$ and $Y$ are independent, then $\sigma_{XY} = \rho_{XY} = 0$.

Comment. Covariance and correlation are measures of *linear* dependence. Zero correlation does not imply independence; consider $Y = |X|$ for $X$ uniformly distributed on $[-1, 1]$. Nevertheless, informally the phrase "$X$ and $Y$ are correlated" sometimes is used to mean that "$X$ and $Y$ are dependent."

Definition. Suppose that the continuous random vector $(X, Y)$ has means $(\mu_X, \mu_Y)$, positive variances $(\sigma_X^2, \sigma_Y^2)$, and correlation $\rho$ with $|\rho|<1$. (Notice that zero variances and $|\rho|=1$ are excluded.)

Then $(X, Y)$ has the *bivariate normal distribution* if its pdf at any point $(x, y)$ is

$$f_{XY}(x, y) = \frac{\exp\left[\dfrac{z^2(x) - 2\rho z(x) z(y) + z^2(y)}{-2(1-\rho^2)}\right]}{2\pi \sigma_X \sigma_Y \sqrt{1-\rho^2}},$$

where $z(x) = (x - \mu_X)/\sigma_X$ and $z(y) = (y - \mu_Y)/\sigma_Y$ (i.e., the $z$-values of $X = x$ and $Y = y$).

Result. If $(X, Y)$ is bivariate normal with parameters $\mu_X$, $\mu_Y$, $\sigma_X$, $\sigma_Y$ and $\rho$, then

(a) the marginal distributions of $X$ and $Y$ are normal,

(b) if $\rho = 0$, then $X$ and $Y$ are independent, and

(c) if $X = x$, then the conditional distribution of $Y$ is normal with mean $\mu_{Y|x} = \mu_Y + \rho \sigma_Y z(x)$ and variance $\sigma_{Y|x}^2 = (1 - \rho^2)\sigma_Y^2$ and

if $Y = y$, then the conditional distribution of $X$ is normal with mean $\mu_{X|y} = \mu_X + \rho \sigma_X z(y)$ and variance $\sigma_{X|y}^2 = (1 - \rho^2)\sigma_X^2$.

Result. If $Z_1$ and $Z_2$ are two independent standard-normal random variables and if

$$X = \mu_X + \sigma_X Z_1$$

$$Y = \mu_Y + \sigma_Y (\rho Z_1 + \sqrt{1-\rho^2} Z_2),$$

then $(X, Y)$ is bivariate normal with parameters $\mu_X$, $\mu_Y$, $\sigma_X$, $\sigma_Y$, and $\rho$. (This result can be used to generate bivariate normal observations in a Monte Carlo simulation.)

Definition. Given random variables $X_1, X_2, \ldots, X_n$ and any constants $c_1, c_2, \ldots, c_n$, then

$$Y = c_1 X_1 + c_2 X_2 + \cdots + c_n X_n$$

is a *linear combination* of $X_1, X_2, \ldots, X_n$.

Result. The linear combination $Y = c_1 X_1 + c_2 X_2 + \cdots + c_n X_n$ has mean and variance

$$\mathrm{E}(Y) = \sum_{i=1}^{n} \mathrm{E}(c_i X_i) = \sum_{i=1}^{n} c_i \mathrm{E}(X_i)$$

and

$$\mathrm{V}(Y) = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{cov}(c_i X_i, c_j X_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \mathrm{cov}(X_i, X_j)$$

$$= \sum_{i=1}^{n} \mathrm{c}_i^2 V(X_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c_i c_j \mathrm{cov}(X_i, X_j) .$$

Recall: $\mathrm{cov}(X_i, X_j) = \mathrm{cov}(X_j, X_i)$ and $\mathrm{cov}(X_i, X_i) = \mathrm{V}(X_i)$.

Corollary. If $X_1, X_2, \ldots, X_n$ are mutually independent, then the linear combination $Y = c_1 X_1 + c_2 X_2 + \cdots + c_n X_n$ has variance $\mathrm{V}(Y) = \sum_{i=1}^{n} c_i^2 \mathrm{V}(X_i)$.

Definition. The *sample mean* of $X_1, X_2, \ldots, X_n$ is the linear combination

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} .$$

Result. If $X_1, X_2, \ldots, X_n$ each have mean $\mu$, then the mean of the sample mean is

$$\mathrm{E}(\bar{X}) = \mu .$$

Result. If $X_1, X_2, \ldots, X_n$ have common mean $\mu$, common variance $\sigma^2$, and are independent, then the variance of the sample mean is

$$\mathrm{V}(\bar{X}) = \sigma^2 / n .$$

Result. *Reproductive Property of the Normal Distribution*. If $X_1, X_2, \ldots, X_n$ are independent normally distributed random variables, then the linear combination $Y = c_1 X_1 + c_2 X_2 + \cdots + c_n X_n$ is normally distributed.

Corollary. If $X_1, X_2, \ldots, X_n$ are mutually independent normal random variables with common mean $\mu$ and common variance $\sigma^2$, then the sample mean $\bar{X}$ is normally distributed with mean $\mu$ and variance $\sigma^2 / n$.

Result. *Central Limit Theorem*. If $X_1, X_2, \ldots, X_n$ is a random sample of size $n$ taken from a population with mean $\mu$ and variance $\sigma^2$, then as $n \to \infty$

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

goes in the limit to the standard normal distribution.

(That is, when $n$ is "large" $\bar{X}$ can be assumed to be normally distributed with mean $\mu$ and variance $\sigma^2 / n$.)

**Descriptive Statistics** (M&R Chapter 6)

A *sample* of *data*, denoted here by $x_1, x_2, \ldots, x_n$, can be summarized (that is, described) numerically or graphically.

A numerical summary value is called a *statistic*. Commonly used statistics include

— sample average (the center of gravity, a measure of location):

$$\bar{x} = \sum_{i=1}^{n} x_i \; / \; n$$

— sample variance (the moment of inertia, a measure of dispersion):

$$s^2 = \frac{\sum_{i=1}^{n} \left( x_i - \bar{x} \right)^2}{n-1} = \frac{\left( \sum_{i=1}^{n} x_i^2 \right) - n(\bar{x})^2}{n-1}$$

— sample standard deviation (an alternative measure of dispersion):

$$s = +\sqrt{s^2}$$

— sample range (an alternative measure of dispersion. Often $4s \le r \le 6s$.):

$$r = \max\{x_1, x_2, \ldots, x_n\} - \min\{x_1, x_2, \ldots, x_n\}$$

— $100k$ sample percentiles for $k = 1, 2, ..., 99$ (alternative measures of location):

$$p_i = \text{the data value greater than approximately } 100k \text{ \% of the data}$$

— first, second, and third sample quartiles (alternative measures of location):

$$q_1 = p_{25}, \; q_2 = p_{50}, \; q_3 = p_{75}$$

— sample median (an alternative measure of location):

$$\tilde{m} = q_2 = p_{50}$$

— sample inter-quartile range (an alternative measure of dispersion):

$$IQR = q_3 - q_1 = p_{75} - p_{25}$$

— sample mode (most-common data value, an alternative measure of the location).

Result. For constants $a$ and $b$, consider the *coded data* $y_i = a + b\, x_i$ for $i = 1, 2, ..., n$. Then
(a) *location* measures are multiplied by $b$ and then increased by $a$, and
(b) *dispersion* measures (except for sample variance) are multiplied by $|b|$.

Graphical summaries include dot plots, histograms, cumulative distribution plots, stem-and-leaf diagrams, and box plots. Data values plotted against time is a time-series plot; appending a stem-and-leaf plot to a time-series plot yields a digidot plot.

— *Frequency* is the number of data values satisfying a specified condition, such as lying in a particular interval. *Relative frequency* is the frequency divided by $n$, the fraction of the data values satisfying the condition.

— A *histogram* is a bar graph showing data-value frequency for several adjacent equal-length intervals. (The intervals are sometimes called *cells* or *bins*).

— A *cumulative distribution plot* is analogous to a histogram, but each bar shows *cumulative frequency*, the number of data values in the bin or to the left of the bin. (Alternatively, each bar can show *cumulative relative frequency*.)

**Parameter Estimation** (M&R Chapter 7)

Comment. We now begin *inferential statistics*, the study of data drawn from a system to infer conclusions about the system. (Chapter 6 discussed *descriptive statistics*.)

Definition. A *population* is the set of all possible observations of the relevant system. (For example, the students in this course.)

Definition. A constant $\theta$ is a *population parameter* if it is a characteristic of the population. (For example, class average gpa.)

Definition. A *sample* is a subset selected from the population. (For example, the students in the front row.)

Definition. The random vector $(X_1, X_2, \ldots, X_n)$ is *independent and identically distributed*, often abbreviated *iid*, if

(a) the $X_i$s are mutually independent and

(b) every $X_i$ has the same probability distribution, say with cdf $F_X$.

Definition. The random vector $(X_1, X_2, \ldots, X_n)$ is a *random sample* (of size $n$) if it is iid.

(For now, every random sample is iid. Later, the definition will be generalized in order to discuss more-sophisticated sampling ideas. At that time this simplest kind of sample will become an "iid random sample" or a "simple random sample".)

Definition. A *statistic* $\hat{\Theta}$ is (a random variable that is) a function of a random sample, that is, there is some function $h$ so that $\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$. (For example, the average gpa of the students in the sample.)

Definition. A statistic $\hat{\Theta}$ is a *point estimator* of the population parameter $\theta$ if its purpose is to guess the value of the population parameter $\theta$.

Definition. A *point estimate* $\hat{\theta}$ is a single observation of $\hat{\Theta}$. (Notice that $\hat{\Theta} = \hat{\theta}$ is an event in the same sense that $X = x$ is an event.)

Definition. A sequence of point estimators $\hat{\Theta}_n$ is *consistent* if $\lim_{n\to\infty} P(|\hat{\Theta}_n - \theta| < \varepsilon) = 1$ for every positive constant $\varepsilon$. (Usually $n$ is sample size. A consistent estimator, then, is guaranteed to be arbitrarily close to $\theta$ for large sample sizes.)

Definition. The *bias* of the point estimator $\hat{\Theta}$ is $\text{bias}(\hat{\theta}, \theta) = E(\hat{\Theta}) - \theta$.

Definition. The point estimator $\hat{\Theta}$ is an *unbiased* estimator of $\theta$ if $E(\hat{\Theta}) = \theta$.

Definition. The *standard error* of a point estimator $\hat{\Theta}$ is its standard deviation, $\sigma_{\hat{\Theta}} = \sqrt{V(\hat{\Theta})}$. (For example, $\sigma_{\bar{X}} = \sigma_X / \sqrt{n}$.)

Definition. The *mean squared error (MSE)* of a point estimator $\hat{\Theta}$ of the parameter $\theta$ is

$$\text{MSE}(\hat{\Theta}, \theta) \equiv E[(\hat{\Theta} - \theta)^2].$$

Result. $\text{MSE}(\hat{\Theta}, \theta) = [\text{bias}(\hat{\Theta}, \theta)]^2 + V(\hat{\Theta})$.

Definition. The *root mean squared error (RMSE)* is the square root of the MSE.

Comments.

— The concept of variance is generalized by MSE, which is useful when the point estimator is biased. RMSE is analogous to standard deviation.

— For most commonly used point estimators, squared bias goes to zero faster than variance as $n$ increases, so asymptotically $\text{MSE}(\hat{\Theta}, \theta) / V(\hat{\Theta}) = 1$.

— Some biased point estimators are good in the sense of having a small MSE.

Point Estimators (from iid data): Summary Table

| Distribution Parameter | Point Estimator | ...Sampling Distribution... Mean | Variance | Standard-Error Estimator |
|---|---|---|---|---|
| $\theta$ | $\hat{\Theta}$ | $E(\hat{\Theta})$ | $V(\hat{\Theta}) \equiv [ste(\hat{\Theta})]^2$ | $\widehat{se}(\hat{\Theta})$ |
| $p \equiv P(A)$ | $\hat{p} \equiv$ "# of successes" $/ n$ | $p$ | $p\,(1-p)/n$ | $\sqrt{\hat{p}\,(1-\hat{p})/(n-1)}$ |
| $\mu \equiv E(X)$ | $\bar{X} \equiv \sum_{i=1}^{n} X_i / n$ | $\mu$ | $\sigma^2 / n$ | $S / \sqrt{n}$ |
| $\mu_1 - \mu_2$ | $\bar{X}_1 - \bar{X}_2$ | $\mu_1 - \mu_2$ | $\sigma_1^2 / n_1 + \sigma_2^2 / n_2$ | $\sqrt{S_1^2 / n_1 + S_2^2 / n_2}$ |
| $\sigma^2 \equiv V(X)$ | $S^2 \equiv \sum_{i=1}^{n} (X_i - \bar{X})^2 / (n-1)$ | $\sigma^2$ | $\frac{\sigma^4}{n}\left(\alpha_4 - \frac{n-3}{n-1}\right)$ | difficult |

Definition. The unbiased estimator of $\theta$ having the smallest variance is called the *minimum-variance unbiased estimator (MVUE)*.

(More precisely, for the iid random vector $(X_1, X_2, \ldots, X_n)$ drawn from a particular distribution having parameter $\theta$, consider all functions $h$ for which $E[h(X_1, X_2, \ldots, X_n)] = \theta$. The MVUE of $\theta$ is the point estimator defined by the function $h$ that minimizes $V[h(X_1, X_2, \ldots, X_n)]$).

Result. If $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with mean $\mu$ and variance $\sigma^2$, then the sample mean $\bar{X}$ is the MVUE for $\mu$.

(We already knew that $\bar{X}$ is unbiased for $\mu$, and that $V(\bar{X}) = \sigma^2 / n$. The new point is that the functional form $h$ of the MVUE is that of a sample average.)

Definition. The distribution of a point estimator $\hat{\Theta}$ is its *sampling distribution*.

— The quality of a point estimator is determined by its sampling distribution.

— The sampling distribution (and therefore the bias and standard error of $\hat{\Theta}$) depends upon the sample size $n$, the distribution $F_X$, and the function $h$.

— For all point estimators in a first course, and for almost all point estimators in general, as $n$ becomes large

(i) the bias goes to zero (at a rate inversely proportional to $n$),

(ii) the standard error goes to zero (at a rate inversely proportional to $\sqrt{n}$), and

(iii) the sampling distribution becomes normal.

Definition. The *estimated standard error*, $\hat{\sigma}_{\hat{\Theta}}$, is a point estimate of the standard error.

— A common notational convenience, illustrated here, is to denote an estimator by placing a "hat" over a the quantity being estimated.

— The reason to estimate a standard error $\sigma_{\hat{\Theta}}$ is to evaluate the quality of $\hat{\Theta}$.
The reason to estimate $\theta$ is to make an inference about the system of interest.

— Many point estimators are created by estimating unknown constants.
(For example, $\hat{\sigma}_{\bar{X}} = S / \sqrt{n}$, where $S$ is the sample standard deviation.)

— The *bootstrap* is an alternative method of estimating $\sigma_{\hat{\theta}}$, especially when the function $h$ is complicated. The method used is Monte Carlo sampling.
(We do not provide bootstrapping details in these notes.)

Distribution fitting. Choosing values of the distribution parameters to obtain the desired properties. Two classical methods are Method of Moments and Maximum Likelihood Estimation.

Method of Moments (MOM). Fitting a $k$-parameter distribution to a real-world context by matching the values of the first $k$ distribution moments to the corresponding $k$ sample moments.

Definition. The *likelihood* $L$ of a sample $x_1, x_2, \ldots, x_n$ is $f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n)$, the $n$-dimensional joint pmf (if discrete) or joint pdf (if continuous) evaluated at the observed values $x_1, x_2, \ldots, x_n$.

(In the discrete case the likelihood is simply the probability of the sample; in the continuous case it is the density of the sample. The word *likelihood* is commonly used to encompass both cases while still being descriptive.)

Definition. The *likelihood function* $L(\theta)$ of an observed sample $x_1, x_2, \ldots, x_n$ is $f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n; \theta)$, where $\theta$ is a distribution parameter.

Result. Assume that $x_1, x_2, \ldots, x_n$ is a random sample from pmf or pdf $f(x; \theta)$. Then the sample's likelihood function is

$$L(\theta) = f(x_1; \theta)\, f(x_2; \theta) \cdots f(x_n; \theta),$$

(The observed sample is known, so $L$ is a function of only the unknown parameter $\theta$. The analyst must assume that the observations $x_i$ are from a family of distributions parameterized by $\theta$; for example, the normal family with $\theta = (\mu, \sigma^2)$.)

Definition. The *maximum likelihood estimator (MLE)* of $\theta$ is the (feasible) value of theta that maximizes $L(\theta)$.

Result. The value of $\theta$ that maximizes $L$ also maximizes any continuous monotonic function of $L$.

(In particular, $L$ and $\ln L$ are both maximized by the same value of $\theta$, a useful result because often $\ln L$ is simpler, especially if maximization is accomplished by setting the first derivative to zero.)

Result. Except in unusual situations, MLEs have these large-sample properties:

(1) approximately unbiased,

(2) nearly minimum variance,

(3) approximately normally distributed, and

(4) consistent.

Result. *MLE Invariance*. Let $\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_k$ be the mle's of the parameters $\theta_1, \theta_2, \ldots, \theta_k$. The mle of any function $h(\theta_1, \theta_2, \ldots, \theta_k)$ is the same function $h(\Theta_1, \Theta_2, \ldots, \Theta_k)$ of the estimators $\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_k$.

(For example, if the mle of $\sigma^2$ is $\hat{\sigma}^2$, then the mle of $\sigma$ is $\hat{\sigma}$.)