8

Random Variables

8.1 Objectives

This chapter describes the means by which we label and treat known and unknown values. Basically there are two types of observable data, and the abstract terminology for yet-to-be observed values should also reflect this distinction. We first talk here about the levels of measurement for observed values where the primary distinction is discrete versus continuous. We will then see that the probability functions used to describe the distribution of such variables preserves this distinction. Many of the topics here lead to the use of statistical analysis in the social sciences.

8.2 Levels of Measurement

It is important to classify data by the precision of measurement. Usually in the social sciences this is an inflexible condition because many times we must take data "as is" from some collecting source. The key distinction is between **discrete data**, which take on a set of categorical values, and **continuous data**, which take on values over the real number line (or some bounded subset of it). The difference can be subtle. While discreteness requires *countability*, it

can be infinitely countable, such as the set of positive integers. In contrast, a continuous random variable takes on uncountably infinite values, even if only in some range of the real number line, like $[0:1]$, because any interval of the real line, finitely bounded or otherwise, contains an infinite number of rational and irrational numbers.

To see why this is an uncountably infinite set, consider any two points on the real number line. It is always possible to find a third point between them. Now consider finding a point that lies between the first point and this new point; another easy task. Clearly we can continue this process infinitely and can therefore never fully count the number of values between any two points on the real number line.

It is customary to divide levels of measurement into four types, the first two of which are discrete and the second two of which are either continuous or discrete. Stevens (1946, 1951) introduced the following (now standard) language to describe the four differing measurement precisions for observed data.

Nominal. Nominal data are purely categorical in that there is no logical way to order a set of events. The classic example is religions of the world: Without specifying some additional criteria (such as age, origin, or number of adherents) there is no nonnormative way to rank them. A synonym for nominal is **polychotomous**, and sometimes just "categorical" is used as well, but this latter term can be confusing in this context because there are two types of categorical data types. In addition, dichotomous (yes/no, on/off, etc.) data are also considered nominal, because with two outcomes ordering does not change any interpretive value. Examples of nominal data include

- male/female
- regions of the U.S.
- football jersey numbers
- war/peace
- political parties
- telephone numbers.

Ordinal. Ordinal data are categorical (discrete) like nominal data, but with the key distinction that they can be ranked (i.e., ordered). While we could

treat ordinal data as if they were just nominal, both are discrete, we would be ignoring important qualitative information about how to relate categories. Examples include

- seniority in Congress (it is naive to treat years in office more literally);
- lower/middle/upper socio-economic class;
- Likert scales (agree/no opinion/disagree, and other variants);
- Guttman scale (survey respondents are presented with increasingly hard-to-agree-with statements until they disagree);
- levels of democratization.

Often ordinal data are the result, not of directly measured data, but artificial indices created by researchers to measure some latent characteristic. For instance, sociologists are sometimes concerned with measuring tolerance within societies. This may be tolerance of different races, cultures, languages, sexual orientations, or professions. Unfortunately it is not possible to measure such underlying attitudes directly either by observation or a single survey query. So it is common to ask a multitude of questions and combine the resulting information into an index: multi-item measures from accumulating scores to create a composite variable. Political scientists do this to a slightly lesser extent when they are concerned with levels of freedom, volatility, political sophistication, ideology, and other multifaceted phenomenon.

Interval. The key distinction between interval data and ordinal data is that interval data have equal spacing between ordered values. That is, the difference between 1 and 2 is exactly the difference between 1001 and 1002. In this way the ordering of interval data has a higher level of measurement, allowing more precise comparisons of values. Consider alternatively the idea of measuring partisanship in the U.S. electorate from a survey. It may or may not be the case that the difference between **somewhat conservative** and **conservative** is the same as the distance between **conservative** and **extremely conservative**. Therefore it would incorrect, in general, to treat this as interval data.

Interval data can be discrete or continuous, but if they are measured on the real number line, they are obviously continuous. Examples of interval measured data include

- temperature measured in Fahrenheit or Celsius;
- a "feeling thermometer" from 0 to 100 that measures how survey respondents feel about political figures;
- size of legislature (it does not exist when $n = 0$);
- time in years (0 AD is a construct).

Ratio. Ratio measurement is exactly like interval measurement except that the point at zero is "meaningful." There is nothing in the definition of interval measure that asserts that zero (if even included in the set of possible values) is really an indicator of a true lack of effect. For example, Fahrenheit and Celsius both have totally arbitrary zero points. Zero Fahrenheit is attributed to the coldest level that the Dutch scientist Daniel Fahrenheit could make a water and salt solution in his lab (he set 100 degrees as his own measured body temperature at that same time). Zero Celsius was established a bit more scientifically by the Swedish astronomer Anders Celsius as the point where water alone freezes (and, as is generally known, 100 degrees Celsius is the point where water boils). While the zero point in both cases has some physical basis, the choice of water and salt is completely arbitrary. Suppose we were to meet developed creatures from Jupiter. It is likely that their similarly constructed scales would be based on ammonia, given the dominant chemical content of the Jovian atmosphere. So what does this limitation to interval measure mean for these two scales? It means that **ratios** are meaningless: 80 degrees is not twice as hot as 40 degrees (either scale) because the reference point of true zero does not exist. Is there a measure of temperature that is ratio? Fortunately, yes; zero degrees Kelvin is the point at which all molecular motion in matter stops. It is an absolute zero because there can be no lower temperature.

Ratio measurement is useful specifically because it does allow direct ***ratio***

comparison. That is, 10 Kelvin is twice as "warm" as 5 Kelvin. More relevant examples include

- appropriations
- crime rates
- unemployment
- votes
- war deaths
- group membership.

Ratio measurement, like interval measurement, can be either discrete or continuous. There is also a subtle distinction between interval and ratio measurement that sometimes gets ignored. Previously the example of the size of a legislature was given as interval rather than ratio. Although zero has some meaning in this context, a legislature with zero members does not exist as a legislature and this then voids the utility of the zero point so that it has no practical meaning.

Notice that the scale of measurement here is ascending in precision (and actually in desirability as well). This direction is easy to remember with the mnemonic NOIR, as the French word for the color black. Any level of measurement, except nominal, can always be reduced to a lower one simply by ignoring some information. This makes sense at times when the defining characteristic is suspicious or perhaps measured poorly.

8.3 Distribution Functions

Distribution functions are central in statistical and probabilistic analyses. They provide a description of how we believe that some data-generating process is operating. Since all models are simplifications of reality, probability statements are really just rough simplifications of the way things actually work. Nobody believes that events like wars, marriages, or suicides occur for underlying *mathematical* reasons. Yet, we can learn a lot about natural, social, and political phenomenon by fitting a parsimonious description based on probability statements.

What do we mean by the word **random** here? We will shortly review a formal and rigorous definition, but it helps to first think intuitively about the meaning. Everyone is accustomed to the idea that some events are more likely

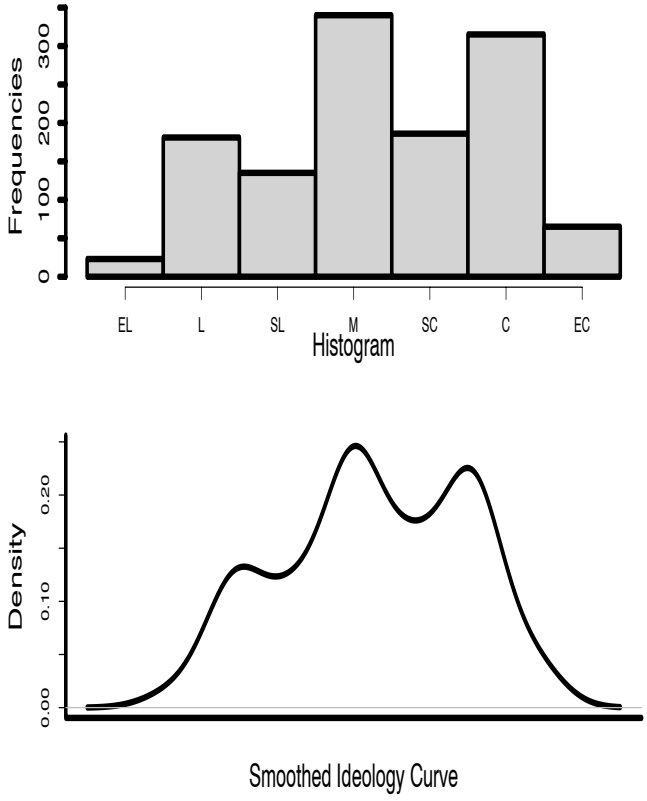
to occur than others. We are more likely to eat lunch today than to win the lottery; it is more likely to rain on a given Glasgow winter day than to be sunny; the stock market is more likely to rise on good economic news than to fall. The key idea here is the expression of *relative difference* between the likelihood of events. Probability formalizes this notion by standardizing such comparisons to exist between zero and one inclusive, where zero means the event will absolutely not occur and one means that the event will certainly occur.† Every other assignment of probability represents some measure of uncertainty and is between these two extreme values, where higher values imply a greater likelihood of occurrence. So probability is no more than a conventional standard for comparisons that we all readily make.

★ **Example 8.1: Measuring Ideology in the American Electorate**. As a simple example, consider a question from the 2002 American National Election Study that asks respondents to identify their ideology on a seven-point scale that covers extremely liberal, liberal, slightly liberal, moderate, slightly conservative, conservative, and extremely conservative. A total of 1245 in the survey placed themselves on this scale (or a similar one that was merged), and we will assume for the moment that it can be treated as an interval measure. Figure 8.1 shows a histogram of the ideology placements in the first panel. This histogram clearly demonstrates the multimodality of the ideology placements with three modes at liberal, moderate, and conservative.

The second panel of Figure 8.1 is a "smoothed" version of the histogram, called a *density plot*. The y-axis is now rescaled because the area under this curve is normalized to integrate to one. The point of this density plot is to estimate an underlying probability structure that supposedly governs the placement of ideology. The key point is that we do not really believe that a mathematical law of some sort determines political ideology, but hopefully

† There is actually a subtlety here. Impossible events have probability zero and exceedingly, exceedingly unlikely events also have probability zero for various reasons. The same logic exists for probability one events. For our purposes these distinctions are not important, however.

Fig. 8.1. Self-Identified Ideology, ANES 2002



by constructing this density plot we have captured an accurate probabilistic description of the underlying structure that determines the observed phenomenon. So a probability function can be taken as a description of the long-run relative frequencies.

There is actually an old simmering controversy behind the interpretation of these probability functions. One group, who are called "frequentists," believe that probability statements constitute a long-run likelihood of occurrence for specific events. Specifically, they believe that these are objective, permanent

statements about the likelihood of certain events relative to the likelihood of other events. The other group, who are usually termed "Bayesians" or "subjectivists," believe that all probability statements are inherently subjective in the sense that they constitute a certain "degree of belief" on the part of the person making the probability statement. More literally, this last interpretation constitutes the odds with which one would be willing to place a bet with his or her own money. There are strong arguments for both perspectives, but to a great degree this discussion is more philosophical than practical.

8.3.1 Randomness and Variables

Randomness does not actually mean what many (nonconcerned) people think that it means. Colloquially "random" is synonymous with equally likely outcomes, and that is how we explicitly treated randomness in part of the last chapter. So it may be common to describe the experiment of rolling a single fair die as random because each of the six sides are equally likely. But think about how restrictive this definition would be if that was the only type of uncertainty allowed: All countries are equally likely to go to war, all eligible citizens in a given country are equally likely to vote in the next election, every surveyed household is equally likely to be below the poverty level.

What randomness really means is that the outcome of some experiment (broadly defined) is not *deterministic*: guaranteed to produce a certain outcome. So, as soon as the probability for some described event slips below one or above zero, it is a random occurrence. Thus if the probability of getting a jackpot on some slot machine is 0.001 for a given pull of the handle, then it is still a random event.

Random variables describe unoccurred events abstractly for pedagogical purposes. That is, it is often convenient to describe the results of an experiment *before it has actually occurred*. In this way we may state that the outcome of a coin flip is designated as X. So for a fair coin we can now say that the probability that X is going to be a heads on the next flip is 0.5.

Formally, a random variable, often denoted with a capital Latin letter such as X or Y, is a function that maps the sample space on which it is "created" to a subset of the real number line (including possibly the whole real number line itself). So we now have a new sample space that corresponds not to the physical experiment performed but to the possible outcomes of the random variable itself. For example, suppose our experiment is to flip a coin 10 times ($n = 10$). The random variable X is defined to be the number of heads in these 10 tosses. Therefore, the sample space of a single iteration of the experiment is $\{H, T\}$, and the sample space of X is $\{0, 1, 2, \ldots, 10\}$.

Random variables provide the connection between events and probabilities because they give the abstraction necessary for talking about uncertain and unobserved events. Sometimes this is as easy as mapping a probability function to a set of discrete outcomes in the sample space of the random variable. To continue the example, we can calculate (more details below) the probability that X takes on each possible value in the sample space determined by 10 flips of a fair coin:

X	0	1	2	3	4
$p(X)$	0.0010	0.0098	0.0439	0.1172	0.2051
X	5	6	7	8	9	10
$p(X)$	0.2461	0.2051	0.1172	0.0439	0.0098	0.0010

Here each possible event for the random variable X, from 0 heads to 10 heads, is paired with a specific probability value. These probability values necessarily sum to unity because one of the 11 values *must* occur.

8.3.2 Probability Mass Functions

When the state space is discrete, we can assign probability values to each single event, even if the state space is countably infinite (discrete with an infinite number of distinct outcomes). So, for example, in the case of flipping a possibly unfair coin, we can assign a probability to heads, $p(H)$, and therefore a

complementary probability to tails, $p(T)$.

The essence of a **probability mass function** is that it assigns probabilities to each unique event, such that the Kolmogorov Axioms still apply. It is common to abbreviate the expression "probability mass function" with "PMF" as a shorthand. We denote such PMF statements

$$f(x) = p(X = x),$$

meaning that the PMF $f(x)$ is a function which assigns a probability for the random variable X equaling the specific numerical outcome x. This notation often confuses people on introduction because of the capital and lower case notation for the same letter. It is important to remember that X is a random variable that can take on multiple discrete values, whereas x denotes a hypothetical single value. Customarily, the more interesting versions of this statement insert actual values for x. So, for instance, the coin-flipping statements above are more accurately given as:

$$p(X = H) = 1 - p(X = T).$$

Notice that in this setup the coin need not be "fair" in the sense that the probability expression accommodates weighted versions such as $p(X = H) = 0.7$ and $p(X = T) = 0.3$.

8.3.3 Bernoulli Trials

The coin-flipping example above is actually much more general than it first appears. Suppose we are studying various political or social phenomenon such as whether a coup occurs, whether someone votes, cabinet dissolution or continuation, whether a new person joins some social group, if a bill passes or fails, and so on. These can all be modeled as **Bernoulli outcomes** whereby the occurrence of the event is assigned the value "1," denoting success, and the nonoccurrence of the event is assigned the value "0," denoting failure. Success and failure can be an odd choice of words when we are talking about coups or

wars or other undesirable events, but this vocabulary is inherited from statistics and is quite well entrenched.

The basic premise behind the **Bernoulli PMF** is that the value one occurs with probability p and the value zero occurs with probability $1 - p$. Thus these outcomes form a partition of the sample space and are complementary. If x denotes the occurrence of the event of interest, then

$$p(x) = p \qquad \text{and} \qquad p(x^{\complement}) = 1 - p.$$

So it is natural to want to estimate p given some observations. There are many ways to do this, but the most direct is to take an average of the events (this process actually has substantial theoretical support, besides being quite simple). So if we flip a coin 10 times and get 7 heads, then a reasonable estimate of p is 0.7.

8.3.4 Binomial Experiments

The **binomial PMF** is an extension to the Bernoulli PMF whereby we simultaneously analyze multiple Bernoulli trials. This is historically called an experiment because it was originally applied to controlled tests. The random variable is no longer binary but instead is the sum of the observed binary Bernoulli events and is thus a count: $Y = \sum_{i=1}^{n} X_i$. A complication to this Bernoulli extension is figuring out how to keep track of all of the possible sets of results leading to a particular sum.

To make things easy to start with, suppose we are studying three senior legislators who may or may not be retiring at the end of the current term. We believe that there is an underlying shared probability p governing their independent decisions and denote the event of retiring with R. We thus have a number of events E dictated by the three individual values and their ordering, which produce a sum bounded by zero (no retirements) and three (all retirements). These events are given in the first column of Table 8.1 with their respective sums in the second column.

Table 8.1. Binomial Outcomes and Probabilities

E	Y	$p(Y = y)$	
$\{R^{\complement}, R^{\complement}, R^{\complement}\}$	0	$(1-p)(1-p)(1-p)$	$p^0(1-p)^{3-0}$
$\{R, R^{\complement}, R^{\complement}\}$	1	$(p)(1-p)(1-p)$	$p^1(1-p)^{3-1}$
$\{R^{\complement}, R, R^{\complement}\}$	1	$(1-p)(p)(1-p)$	$p^1(1-p)^{3-1}$
$\{R^{\complement}, R^{\complement}, R\}$	1	$(1-p)(1-p)(p)$	$p^1(1-p)^{3-1}$
$\{R, R, R^{\complement}\}$	2	$(p)(p)(1-p)$	$p^2(1-p)^{3-2}$
$\{R, R^{\complement}, R\}$	2	$(p)(1-p)(p)$	$p^2(1-p)^{3-2}$
$\{R^{\complement}, R, R\}$	2	$(1-p)(p)(p)$	$p^2(1-p)^{3-2}$
$\{R, R, R\}$	3	$(p)(p)(p)$	$p^3(1-p)^{3-3}$

The third column of Table 8.1 gives the probabilities for each of these events, which is simply rewritten in the fourth column to show the structure relating Y and the number of trials, 3. Since the retirement decisions are assumed independent, we can simply multiply the underlying individual probabilities according the definition given on page 317 in Chapter 7 to get the joint probability of occurrence. If we lump these together by the term Y, it is easy to see that there is one way to get zero retirements with probability $(1-p)^3$, three ways to get one retirement with probability $p(1-p)^2$, three ways to get to two retirements with probability $p^2(1-p)$, and one way to get three retirements with probability p^3. Recalling that this is really choosing by unordered selection without replacement (page 288), we can note that the ways to get each of these events is given by the expression $\binom{3}{y}$.

There is also a clear pattern to binomial distribution probabilities. The outcome that receives the highest probability is the one that corresponds to $n \times p$ (more on this calculation below), and probabilities slope down in both directions from this modal point. A particularly elegant picture results from experiments with so-called "fair" probabilities. Suppose we flip such a fair coin 10 times. What is the full probability outcome for the number of heads? We can obviously make these calculations in exactly the same way that was done in the example above. If such probabilities were then graphed with a barplot, the result would look like Figure 8.2.

This is really useful because we can now state the binomial PMF for this particular "experiment" for the sum of retirement events:

$$p(Y = y) = \binom{3}{y} p^y (1-p)^{3-y},$$

which has the more general form for n Bernoulli trials:

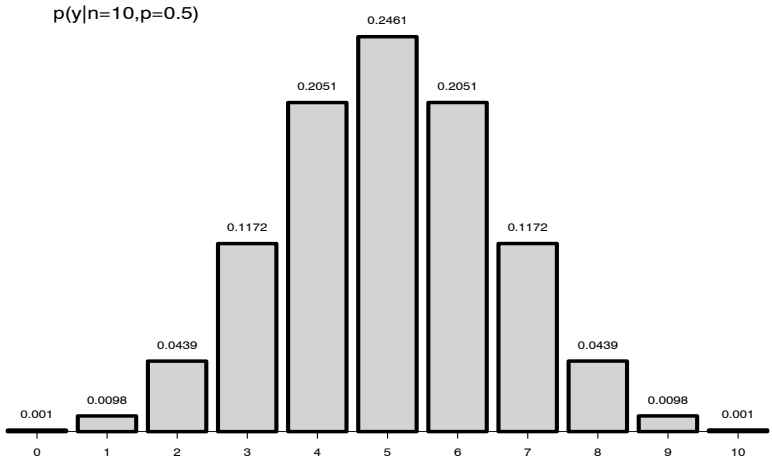
$$p(Y = y|n, p) = \binom{n}{y} p^y (1-p)^{n-y}, \quad n \geq y,\ n, y \in \mathcal{I}^+,\ p \in [0{:}1].$$

We can denote this general form or any specific case with the shorthand $\mathcal{B}(n, p)$. So in this way we define a general form for the binomial PMF and a mechanism for specifying specific cases, that is, $\mathcal{B}(10, 5)$, $\mathcal{B}(100, 75)$, and so on.

★ **Example 8.2: Binomial Analysis of Bill Passage**. Suppose we know that a given legislature has a 0.7 probability of passing routine bills (perhaps from historical analysis). If 10 routine bills are introduced in a given week, what is the probability that:

(i) Exactly 5 bills pass? We can simply plug three values into the binomial

Fig. 8.2. Example Binomial Probabilities



PMF for this question:

$$p(Y = 5|n = 10, p = 0.7) = \binom{10}{5}(0.7)^5(1 - 0.7)^{10-5}$$

$$= (252)(0.16807)(0.00243) = 0.10292.$$

(ii) Less than three bills pass? The most direct method is to add up the three probabilities associated with zero, one, and two occurrences:

$$p(Y < 3|10, 0.7) = p(Y = 0|10, 0.7) + p(Y = 1|10, 0.7) + p(Y = 2|10, 0.7)$$

$$= \binom{10}{0}(0.7)^0(1 - 0.7)^{10-0} + \binom{10}{1}(0.7)^1(1 - 0.7)^{10-1} + \binom{10}{2}(0.7)^2(1 - 0.7)^{10-2}$$

$$= 0.0000059049 + 0.000137781 + 0.001446701 = 0.00159.$$

(iii) Nine or less bills pass? The obvious, but time-consuming way to answer this question is the way the last answer was produced, by summing up all (nine here) applicable individual binomial probabilities. However, recall that because this binomial PMF is a probability function, the sum of the probability of all possible events must be one. So this

suggests the following trick:

$$p(Y \leq 9|10, 0.7) = \sum_{i=1}^{9} p(Y = i|10, 0.7)$$

$$= \sum_{i=1}^{10} p(Y = i|10, 0.7) - p(Y = 10|10, 0.7)$$

$$= 1 - p(Y = 10|10, 0.7)$$

$$= 1 - \binom{10}{10}(0.7)^{10}(1 - 0.7)^{10-10}$$

$$= 1 - 0.02825 = 0.97175.$$

8.3.5 Poisson Counts

Suppose that instead of counting the number of successes out of a fixed number of trials, we were concerned with the number of events (which can still be considered successes, if one likes) without an upper bound. That is, we might consider the number of wars on a continent, the number of alliances between protest groups, or the number of cases heard by a court. While there may be some practical upper limit imposed by the number of hours in day, these sorts of events are usually counted as if there is no upper bound because the number of attempts is unknown a priori. Another way of thinking of such count data is in terms of durations: the length of time waiting for some prescribed event. If the probability of the event is proportional to the length of the wait, then the length of wait can be modeled with the **Poisson PMF**. This discrete distributional form is given by

$$p(y|\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \quad y \in \mathcal{I}^{+},\ \lambda \in \mathfrak{R}^{+}.$$

The assumption of proportionality is usually quite reasonable because over longer periods of time the event has more "opportunities" to occur. Here the single PMF parameter λ is called the intensity parameter and gives the expected

number of events. This parametric form is very useful but contains one limiting feature: λ is also assumed to be the dispersion (variance, defined on page 366) of the number of events.

★ **Example 8.3: Poisson Counts of Supreme Court Decisions**. Recent Supreme Courts have handed down roughly 8 unanimous decisions per term. If we assume that $\lambda = 8$ for the next Court, then what is the probability of observing:

(i) Exactly 6 decisions? Plugging these values into the Poisson PMF gives

$$p(Y = 6|\lambda = 8) = \frac{e^{-8}8^6}{6!} = 0.12214.$$

(ii) Less than three decisions? Here we can use a sum of three events:

$$p(Y < 3|\lambda = 8) = \sum_{i=0}^{2} \frac{e^{-8}8^{y_i}}{y_i!}$$

$$= 0.00034 + 0.00268 + 0.01073 = 0.01375.$$

(iii) Greater than 2 decisions? The easiest way to get this probability is with the following "trick" using the quantity from above:

$$p(Y > 2|\lambda = 8) = 1 - p(Y < 3|\lambda = 8)$$

$$= 1 - 0.01375 = 0.98625.$$

The Poisson distribution is quite commonly applied to events in international systems because of the discrete nature of many studied events. The two examples that follow are typical of simple applications. To directly apply the Poisson distribution two assumptions are required:

- Events in different time periods are independent.
- For small time periods, the probability of an event is proportional to the length of time passed in the period so far, and not dependent on the number of previous events in this period.

These are actually not as restrictive as it might appear. The first condition says that rates of occurrence in one time period are not allowed to influence subsequent rates in another. So if we are measuring conflicts, the outset of a widespread war will certainly influence the number of actual battles in the next period, and this thus obviates the continued use of the same Poisson parameterization as was used prior to the war. The second condition means that time matters in the sense that, for some bounded slice of time, as the waiting time increases, the probability of the event increases. This is intuitive; if we are counting arrivals at a traffic light, then it is reasonable to expect more arrivals as the recording period is extended.

★ **Example 8.4: Modeling Nineteenth-Century European Alliances**. McGowan and Rood (1975) looked at the frequency of alliance formation from 1814 to 1914 in Europe between the "Great Powers:" Austria-Hungary, France, Great Britain, Prussia-Germany, and Russia. They found 55 alliances during this period that targeted behavior within Europe between these powers and argued that the observed pattern of occurrence follows the Poisson distribution. The mean number of alliances per year total is 0.545, which they used as their empirical estimate of the Poisson parameter, $\lambda = 0.545$. If we use this value in the Poisson PMF, we can compare observed events against predicted events:

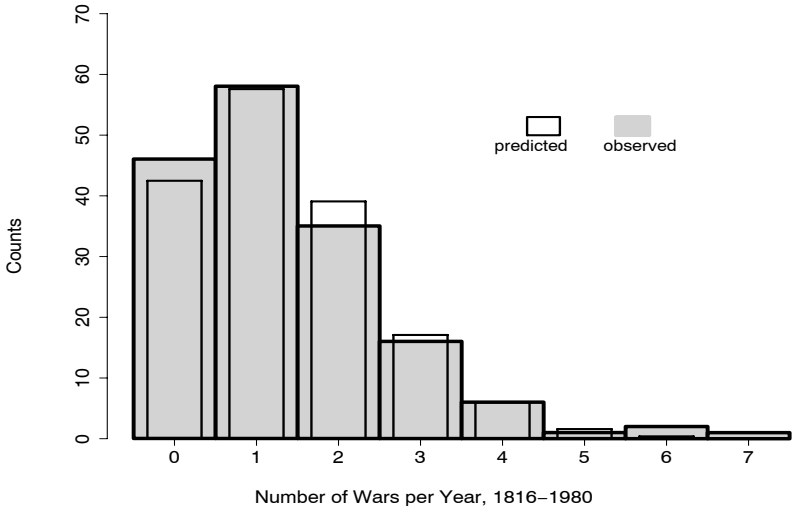
Alliances/Year	$y = 0$	$y = 1$	$y = 2$	$y \geq 3$
Observed	61	31	6	3
Predicted	58.6	31.9	8.7	1.8

This seems to fit the data reasonably well in terms of prediction. It is important to recall that $\lambda = 0.545$ is the intensity parameter for *five* countries to enter into alliances, so assuming that each country is equally likely, the intensity parameter for an individual country is $\lambda_i = 0.545/5 = 0.109$.

★ **Example 8.5: Poisson Process Model of Wars**. Houweling and Kuné (1984) looked at wars as discrete events in a paper appropriately titled "Do

Outbreaks of War Follow a Poisson-Process?" They compared 224 events of international and civil wars from 1816 to 1980 to that predicted by estimating the Poisson intensity parameter with the empirical mean: $\lambda = 1.35758$. Evidence from Figure 8.3 indicates that the Poisson assumption fits the data quite nicely (although the authors quibbled about the level of statistical significance).

Fig. 8.3. Poisson Probabilities of War



Interestingly, the authors found that the Poisson assumption fits less well when the wars were disaggregated by region. The events in the Western Hemisphere continue to fit, while those in Europe, the Middle East, and Asia deviate from the expected pattern. They attribute this latter effect to not meeting the second condition above.

8.3.6 The Cumulative Distribution Function: Discrete Version

If X is a discrete random variable, then we can define the sum of the probability mass to the left of some point $X = x$: the mass associated with values less

than X. Thus the function

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

defines the **cumulative distribution function** (CDF) for the random variable X. A couple of points about notation are worth mentioning here. First, note that the function uses a capital "F" rather than the lower case notation given for the PMF. Sometimes the CDF notation is given with a random variable subscript, $F_X(x)$, to remind us that this function corresponds to the random variable X.

If the values that X can take on are indexed by order: $x_1 < x_2 < \cdots < x_n$, then the CDF can be calculated with a sum for the chosen point x_j:

$$F(x_j) = \sum_{i=1}^{j} p(x_i).$$

That is, $F(x_j)$ is the sum of the probability mass for events less than or equal to x_j. Using this definition of the random variable, it follows that

$$F(x < x_1) = 0 \qquad \text{and} \qquad F(x \geq x_n) = 1.$$

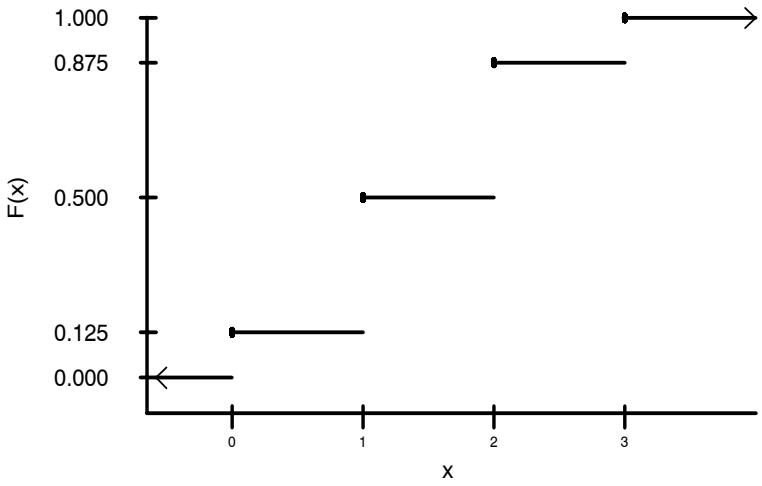
Therefore, CDF values are bounded by $[0{:}1]$ under all circumstances, even if the random variable is not indexed in this convenient fashion. In fact, we can now state technically the three defining properties of a CDF:

- **[Cumulative Distribution Function Definition.]** $F(x)$ is a CDF for the random variable X iff it has the following properties:
  - **bounds:** $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$,
  - **nondecreasing:** $F(x_i) \leq F(x_j)$ for $x_i < x_j$,
  - **right-continuous:** $\lim_{x \downarrow x_i} F(x) = x_i$ for all x_i defined by $f(x)$.

The idea of a **right-continuous function** is best understood with an illustration. Suppose we have a binomial experiment with $n = 3$ trials and $p = 0.5$. Therefore the sample space is $S = \{0, 1, 2, 3\}$, and the probabilities associated with each event are $[0.125, 0.375, 0.375, 0.125]$. The graph of $F(x)$ is given in Figure 8.4, where the discontinuities reflect the discrete nature of a

binomial random variable. The solid circles on the left-hand side of each interval emphasize that this value at the integer belongs to that CDF level, and the lack of such a circle on the right-hand side denotes otherwise. The function is right-continuous because for each value of x_i ($i = 0, 1, 2, 3$) the limit of the function reaches x_i moving from the right. The arrows pointing left and right at 0 and 1, respectively, are just a reminder that the CDF is defined towards negative and positive infinity at these values. Note also that while the values are cumulative, the jumps between each level correspond to the PMF values $f(x_i),\ i = 0, 1, 2, 3$.

Fig. 8.4. Binomial CDF Probabilities, $n = 3, p = 0.5$



It is important to know that a CDF fully defines a probability function, as does a PMF. Since we can readily switch between the two by noting the step sizes (CDF→PMF) or by sequentially summing (PMF→CDF), then the one we use is completely a matter of convenience.

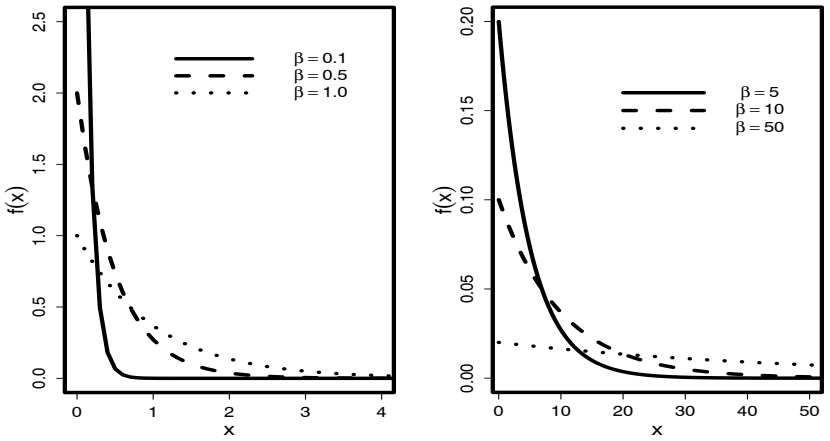
8.3.7 Probability Density Functions

So far the random variables have only taken on discrete values. Clearly it would be a very limiting restriction if random variables that are defined over some interval of the real number line (or even the entire real number line) were excluded. Unfortunately, the interpretation of probability functions for continuous random variables is a bit more complicated.

As an example, consider a spinner sitting flat on a table. We can measure the direction of the spinner relative to some reference point in radians, which vary from 0 to 2π (Chapter 2). How many outcomes are possible? The answer is infinity because the spinner can theoretically take on any value on the real number line in $[0:2\pi]$. In reality, the number of outcomes is limited to our measuring instrument, which is by definition discrete. Nonetheless, it is important to treat continuous random variables in an appropriate manner.

For continuous random variables we replace the probability mass function with the **probability density function** (PDF). Like the PMF, the PDF assigns probabilities to events in the sample space, but because there is an infinite number of alternatives, we cannot say $p(X = x)$ and so just use $f(x)$ to denote the function value at x. The problem lies in questions such as, if we survey a large population, what is the probability that the average income were $\$65{,}123.97$? Such an event is sufficiently rare that its probability is essentially zero. It goes to zero as a measurement moves toward being truly continuous (money in dollars and cents is still discrete, although granular enough to be treated as continuous in most circumstances). This seems ultimately frustrating, but the solution lies in the ability to replace probabilities of specific events with probabilities of ranges of events. So instead with our survey example we may ask questions such as, what is the probability that the average income amongst respondents is greater than $\$65{,}000$?

Fig. 8.5. Exponential PDF Forms



8.3.8 Exponential and Gamma PDFs

The **exponential PDF** is a very general and useful functional form that is often used to model durations (how long "things last"). It is given by

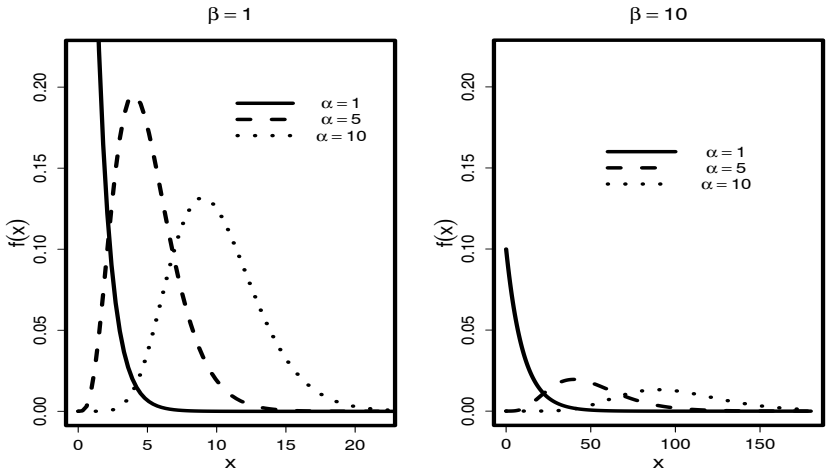
$$f(x|\beta) = \frac{1}{\beta}\exp[-x/\beta], \qquad 0 \leq x < \infty, \quad 0 < \beta,$$

where, similar to the Poisson PMF, the function parameter (β here) is the mean or expected duration. One reason for the extensive use of this PDF is that it can be used to model a wide range of forms. Figure 8.5 gives six different parameterizations in two frames. Note the broad range of spread of the distribution evidenced by the different axes in the two frames. For this reason β is called a **scale parameter**: It affects the scale (extent) of the main density region.

Although we have praised the exponential distribution for being flexible, it is still a special case of the even more flexible **gamma PDF**. The gamma distribution adds a **shape parameter** that changes the "peakedness" of the distribution: how sharply the density falls from a modal value. The gamma PDF is given by

$$f(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}\exp[-x/\beta], \quad 0 \leq x < \infty, \quad 0 < \alpha, \beta,$$

Fig. 8.6. GAMMA PDF FORMS



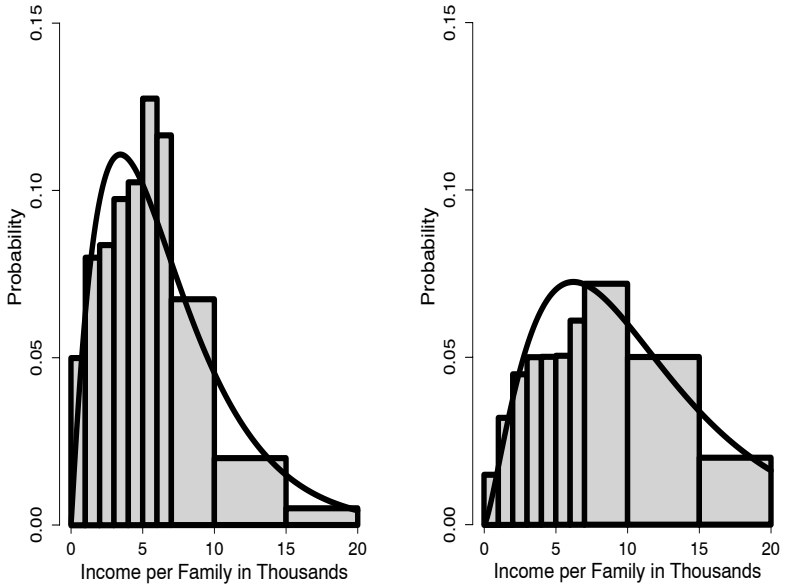
where α is the new shape parameter, and the mean is now $\alpha\beta$. Note the use of the gamma function (hence the name of this PDF). Figure 8.6 shows different forms based on varying the α and β parameters where the y-axis is fixed across the two frames to show a contrast in effects.

An important special case of the gamma PDF is the χ^2 distribution, which is used in many statistical tests, including the analysis of tables. The χ^2 distribution is a gamma where $\alpha = \frac{df}{2}$ and $\beta = 2$, and df is a positive integer value called the **degrees of freedom**.

★ **Example 8.6: Characterizing Income Distributions**. The gamma distribution is particularly well suited to describing data that have a mode near zero and a long right (positive) skew. It turns out that income data fit this description quite closely. Pareto (1897) first noticed that income in societies, no matter what kind of society, follows this pattern, and this effect is sometimes called **Pareto's Law**. Subsequent studies showed that the gamma distribution could be easily tailored to describe a range of income distributions.

Salem and Mount (1974) looked at family income in the United States from 1960 to 1969 using survey data from the Current Population Report

Fig. 8.7. Fitting Gamma Distributions to Income



Series (CPS) published by the Census Bureau and fit gamma distributions to categories. Figure 8.7 shows histograms for 1960 and 1969 where the gamma distributions are fit according to

$$f_{1960}(income) = \mathcal{G}(2.06, 3.2418) \quad \text{and}$$
$$f_{1969}(income) = \mathcal{G}(2.43, 4.3454)$$

(note: Salem and Mount's table contains a typo for β_{1969}, and this is clearly the correct value given here, as evidenced from their graph and the associated fit).

The unequal size categories are used by the authors to ensure equal numbers of sample values in each bin. It is clear from these fits that the gamma distribution can approximately represent the types of empirical forms that income data takes.

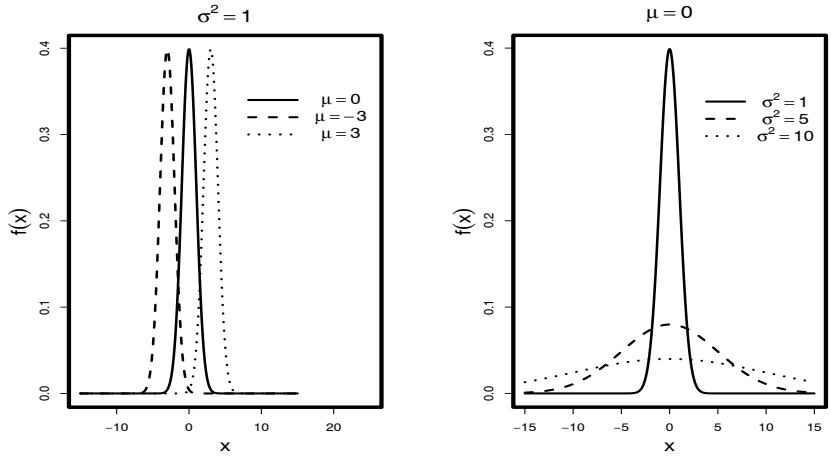
8.3.9 Normal PDF

By far the most famous probability distribution is the **normal PDF**, sometimes also called the **Gaussian PDF** in honor of its "discoverer," the German mathematician Carl Friedrich Gauss. In fact, until replacement with the Euro currency on January 1, 2002, the German 10 Mark note showed a plot of the normal distribution and gave the mathematical form

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right], \qquad -\infty < x, \mu < \infty, 0 < \sigma^2,$$

where μ is the mean parameter and σ^2 is the dispersion (variance) parameter. These two terms completely define the shape of the particular normal form where μ moves the modal position along the x-axis, and σ^2 makes the shape more spread out as it increases. Consequently, the normal distribution is a member of the **location-scale family** of distributions because μ moves only the location (and not anything else) and σ^2 changes only the scale (and not the location of the center or modal point). Figure 8.8 shows the effect of varying these two parameters individually in two panels.

Fig. 8.8. Normal PDF Forms



The reference figure in both panels of Figure 8.8 is a normal distribution with

$\mu = 0$ and $\sigma^2 = 1$. This is called a **standard normal** and is of great practical as well as theoretical significance. The PDF for the standard normal simplifies to

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}x^2\right], \qquad -\infty < x < \infty.$$

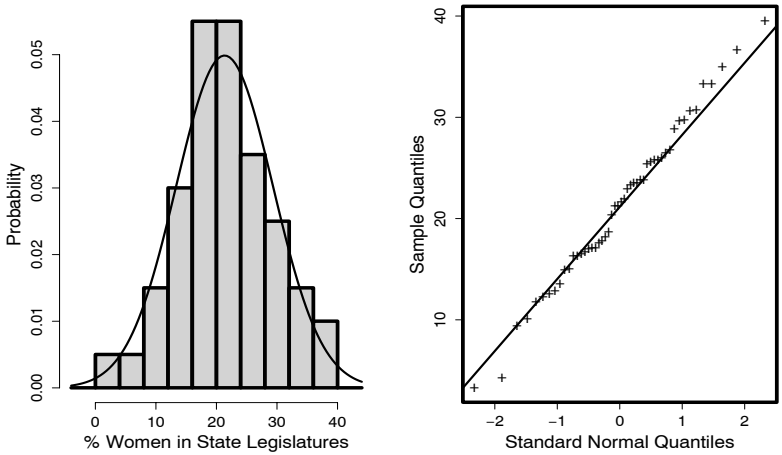
The primary reason that this is an important form is that, due to the location-scale characteristic, any other normal distribution can be transformed to a standard normal and then back again to its original form. As a quick example, suppose $x \sim \mathcal{N}(\mu, \sigma^2)$; then $y = (x - \mu)/\sigma^2 \sim \mathcal{N}(0, 1)$. We can then return to x by substituting $x = y\sigma^2 + \mu$. Practically, what this means is that textbooks need only include one normal table (the standard normal) for calculating tail values (i.e., integrals extending from some point out to infinity), because all other normal forms can be transformed to the standard normal in this way.

One additional note relates to the normal distribution. There are quite a few other common distributions that produce unimodal symmetric forms that appear similar to the normal. Some of these, however, have quite different mathematical properties and thus should not be confused with the normal. For this reason it is not only lazy terminology, but it is also very confusing to refer to a distribution as "bell-shaped."

★ **Example 8.7: Levels of Women Serving in U.S. State Legislatures**. Much has been made in American politics about the role of women in high level government positions (particularly due to "the year of the woman" in 1992). The first panel of Figure 8.9 shows a histogram of the percent of women in legislatures for the 50 states with a normal distribution ($\mu = 21, \sigma = 8$) superimposed (source: Center for American Women and Politics).

The obvious question is whether the data can be considered normally distributed. The normal curve appears to match well the distribution given in the histogram. The problem with relying on this analysis is that the shape of a histogram is greatly affected by the number of bins selected. Consequently,

Fig. 8.9. Fitting the Normal to Legislative Participation



the second panel of Figure 8.9 is a "qqplot" that plots the data against standard normal quantiles (a set of ordered values from the standard normal PDF of length equal to the evaluated vector). The closer the data points are to the line, the closer they are to being normally distributed. We can see here that the fit is quite close with just a little bit of deviation in the tails. Asserting that these data are actually normal is useful in that it allows us to describe typical or atypical cases more precisely, and perhaps to make predictive claims about future legislatures.

8.3.10 The Cumulative Distribution Function: Continuous Version

If X is a continuous random variable, then we can also define the sum of the probability mass to the left of some point $X = x$: the density associated with all values less than X. Thus the function

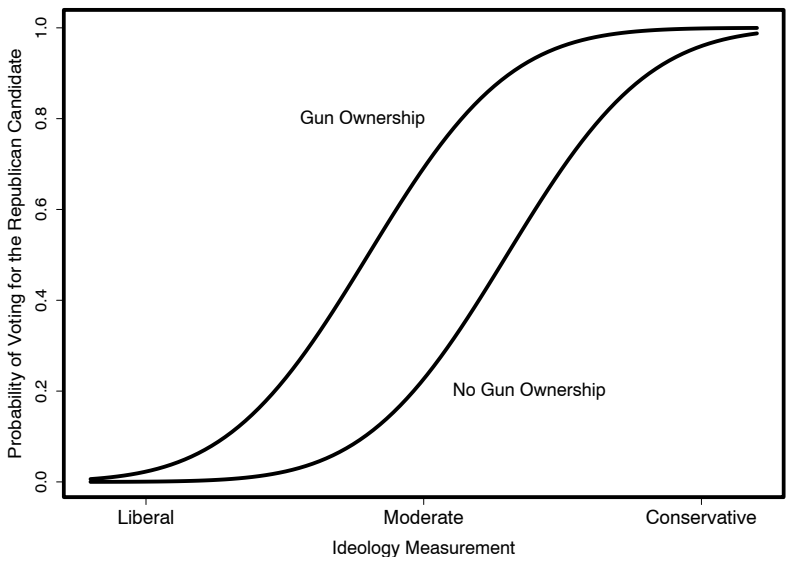
$$F(x) = p(X \leq x) = \int_{-\infty}^{x} f(x)dx$$

defines the cumulative distribution function (CDF) for the continuous random variable X. Even though this CDF is given with an integral rather than a sum,

it retains the three key defining properties, see page 308. The difference is that instead of being a step function (as shown in Figure 8.4), it is a smooth curve monotonically nondecreasing from zero to one.

★ **Example 8.8: The Standard Normal CDF: Probit Analysis**. The CDF of the standard normal is often abbreviated $\Phi(X)$ for $\mathcal{N}(X \leq x|\mu = 0, \sigma^2 = 1)$ (the associated PDF notation is $\phi(X)$). One application that occurs in empirical models is the idea that while people may make dichotomous choices (vote/not vote, purchase/not purchase, etc.), the underlying mechanism of decision is really a smooth, continuous preference or utility function that describes more subtle thinking. If one event (usually the positive/action choice) is labeled as "1" and the opposite event as "0," and if there is some interval measured variable X that affects the choice, then $\Phi(X) = p(X = 1)$ is called the **probit model**. In the basic formulation higher levels of X are assumed to push the subject toward the "1" decision, and lower levels of X are assumed to push the subject toward the "0" decision (although the opposite effect can easily be modeled as well).

Fig. 8.10. Probit Models for Partisan Vote Choice



To give a concrete example, consider the dichotomous choice outcome of voting for a Republican congressional candidate against an interval measured explanatory variable for political ideology. One certainly would not be surprised to observe that more conservative individuals tend to vote Republican and more liberal individuals tend not to vote Republican. We also obtain a second variable indicating whether the respondent owns a gun. A simple probit model is specified for these data with no directly indicated interaction term:

$$p(Y_i = 1) = \Phi(IDEOLOGY_i + GUN_i).$$

Here $IDEOLOGY_i$ is the political ideology value for individual i, GUN_i is a dichotomous variable equaling one for gun ownership and zero otherwise (it is common to weight these two values in such models, but we can skip it here without losing the general point). This model is depicted in Figure 8.10 where gun owners and nongun owners are separated. Figure 8.10 shows that gun ownership shifts the curve affecting the probability of voting for the Republican candidate by making it more likely at more liberal levels of ideology. Also, for very liberal and very conservative respondents, gun ownership does not really affect the probability of voting for the Republican. Yet for respondents without a strong ideological orientation, gun ownership matters considerably: a difference of about 50% at the center.

8.3.11 The Uniform Distributions

There is an interesting distributional form that accommodates both discrete and continuous assumptions. The **uniform distribution** is a perfectly flat form that can be specified in either manner:

k-Category Discrete Case (PMF):

$$p(Y = y|k) = \begin{cases} \frac{1}{k}, & \text{for } y = 1, 2, \ldots, k \\ 0, & \text{otherwise;} \end{cases}$$

Continuous Case (PDF):

$$f(y|a, b) = \begin{cases} \frac{1}{b-a}, & \text{for } a \leq y \leq b \\ 0, & \text{otherwise.} \end{cases}$$

The discrete case specifies k outcomes (hence the conditioning on k in $p(Y = y|k)$) that can be given any range desired (obviously greater ranges make $\frac{1}{k}$ smaller for fixed k), and the continuous case just gives the bounds (a and b), which are often zero and one. So the point is that each outcome has equal individual probability (PMF) or equal density (PDF). This distribution is sometimes used to reflect great uncertainty about outcomes (although it is definitely saying something specific about the probability of events). The continuous case with $a = 0$ and $b = 1$ is particularly useful in modeling probabilities.

★ **Example 8.9: Entropy and the Uniform Distribution**. Suppose we wanted to identify a particular voter by serial information on this person's characteristics. We are allowed to ask a consecutive set of yes/no questions (i.e., like the common guessing game). As we get answers to our series of questions we gradually converge (hopefully, depending on our skill) on the desired voter. Our first question is, does the voter reside in California? Since about 13% of voters in the United States reside in California, a yes answer gives us different information than a no answer. Restated, a yes answer reduces our uncertainty more than a no answer because a yes answer eliminates 87% of the choices whereas a no answer eliminates 13%. If P_i is the probability of the ith event (residing in California), then the improvement in information as defined by Shannon (1948) is defined as

$$I_{P_i} = \log_2 \left[\frac{1}{P_i} \right] = -\log_2 P_i.$$

The probability is placed in the denominator here because the smaller the probability, the greater the investigative information supplied by a yes answer. The log function is required to obtain some desired properties (discussed below) and is justified by various limit theorems. The logarithm is base-2 because there are only two possible answers to our question (yes and no), making the units of information bits. In this example,

$$H_i = -\log_2(0.13) = 2.943416$$

bits, whereas if we had asked, does the voter live in the state of Arkansas? then an affirmative reply would have increased our information by

$$H_i = -\log_2(0.02) = 5.643856$$

bits, or about twice as much. However, there is a much smaller probability that we would have gotten an affirmative reply had the question been asked about Arkansas. What Slater (1939) found, and Shannon (1948) later refined, was the idea that the "value" of the question was the information returned by a positive response times the probability of a positive response. So if the value of the ith binary-response question is

$$H_i = f_i \log_2\left[\frac{1}{f_i}\right] = -f_i \log_2 f_i,$$

then the value of a series of n of these questions is

$$\sum_{i=1}^{n} H_i = k \sum_{i=1}^{n} f_i \log_2\left[\frac{1}{f_i}\right] = -k \sum_{i=1}^{n} f_i \log_2 f_i,$$

where f_i is the frequency distribution of the ith yes answer and k is an arbitrary scaling factor that determines choice of units. This is called the **Shannon entropy** or **information entropy** form. The arbitrary scaling factor here makes the choice of base in the logarithm unimportant because we can change this base by manipulating the constant. For instance, if this form were expressed in terms of the natural log, but $\log_2$ was more appropriate for the application (such as above), then setting $k = \frac{1}{\ln 2}$ converts the entropy form to base 2.

We can see that the total improvement in information is the additive value of the series of individual information improvements. So in our simple example we might ask a series of questions narrowing down on the individual of interest. Is the voter in California? Is the voter registered as a Democrat? Does the voter reside in an urban area? Is the voter female? The total information supplied by this vector of yes/no responses is the total information improvement in units of bits because the response space is binary. Its important to remember that the information obtained is defined only with regard to a well-defined question having finite, enumerated responses

The uniform prior distribution as applied provides the greatest entropy because no single event is more likely to occur than any others:

$$H = -\sum \frac{1}{n} \ln\left(\frac{1}{n}\right) = \ln(n),$$

and entropy here increases logarithmically with the number of these equally likely alternatives. Thus the uniform distribution of events is said to provide the minimum information possible with which to decode the message. This application of the uniform distribution does not imply that this is a "no information" assumption because equally likely outcomes are certainly a type of information. A great deal of controversy and discussion has focused around the erroneous treatment of the uniform distribution as a zero-based information source. Conversely, if there is certainty about the result, then a degenerate distribution describes the m_i, and the message does not change our information level:

$$H = -\sum_{i=1}^{n-1}(0) - \log(1) = 0.$$

8.4 Measures of Central Tendency: Mean, Median, and Mode

The first and most useful step in summarizing observed data values is determining its **central tendency**: a measure of where the "middle" of the data resides

on some scale. Interestingly, there is more than one definition of what constitutes the center of the distribution of the data, the so-called average. The most obvious and common choice for the average is the mean. For n data points $x_1, x_2, \ldots, x_n$, the mean is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

where the bar notation is universal for denoting a mean average. The mean average is commonly called just the "average," although this is a poor convention in the social sciences because we use other averages as well.

The median average has a different characteristic; it is the point such that as many cases are greater as are less: For n data points $x_1, x_2, \ldots, x_n$, the median is X_i such that $i = \lceil n/2 \rceil$ (even n) or $i = \frac{n+1}{2}$ (odd n). This definition suits an odd size to the dataset better than an even size, but in the latter case we just split the difference and define a median point that is halfway between the two central values. More formally, the median is defined as

$$M_x = X_i \colon \int_{-\infty}^{x_i} f_x(X)dx = \int_{x_i}^{\infty} f_x(X) = \frac{1}{2}.$$

Here $f_x(X)$ denotes the **empirical distribution** of the data, that is, the distribution observed rather than that obtained from some underlying mathematical function generating it (see Chapter 7).

The mode average has a totally different flavor. The mode is the most frequently observed value. Since all observed data are countable, and therefore discrete, this definition is workable for data that are actually continuously measured. This occurs because even truly continuous data generation processes, which should be treated as such, are measured or observed with finite instruments. The mode is formally given by the following:

$$m_x = X_i \colon n(X_i) > n(X_j)\ \forall j \neq i,$$

where the notation "$n()$" means "number of" values equal to the X stipulated in the cardinality sense (page 294).

★ **Example 8.10: Employment by Race in Federal Agencies**. Table 8.2 gives the percent of employment across major U.S. government agencies by four racial groups. Scanning down the columns it is clear that there is a great deal of similarity across agencies, yet there exist some interesting dissimilarities as well.

Table 8.2. PERCENT EMPLOYMENT BY RACE, 1998

Agency	Black	Hispanic	Asian	White
Agriculture	10.6	5.6	2.4	81.4
Commerce	18.3	3.4	5.2	73.1
DOD	14.2	6.2	5.4	74.3
Army	15.3	5.9	3.7	75.0
Navy	13.4	4.3	9.8	72.6
Air Force	10.6	9.5	3.1	76.8
Education	36.3	4.7	3.3	55.7
Energy	11.5	5.2	3.8	79.5
EOP	24.2	2.4	4.2	69.3
HHS	16.7	2.9	5.1	75.4
HUD	34.0	6.7	3.2	56.1
Interior	5.5	4.3	1.6	88.6
Justice	16.2	12.2	2.8	68.9
Labor	24.3	6.6	2.9	66.5
State	14.9	4.2	3.7	77.1
Transportation	11.2	4.7	2.9	81.2
Treasury	21.7	8.4	3.3	66.4
VA	22.0	6.0	6.7	65.4
GSA	28.4	5.0	3.4	63.2
NASA	10.5	4.6	4.9	80.1
EEOC	48.2	10.6	2.7	38.5

Source: Office of Personnel Management

The mean values by racial group are $\bar{X}_{\text{Black}} = 19.43$, $\bar{X}_{\text{Hispanic}} = 5.88$, $\bar{X}_{\text{Asian}} = 4.00$, and $\bar{X}_{\text{White}} = 70.72$. The median values differ somewhat: $M_{\text{Black}} = 16.2$, $M_{\text{Hispanic}} = 5.2$, $M_{\text{Asian}} = 3.4$, and $M_{\text{White}} = 73.1$. Cases where the mean and median differ noticeably are where the data are skewed (asymmetric) with the longer "tail" in the direction of the mean. For example,

the white group is negatively skewed (also called left-skewed) because the mean is noticeably less than the median.

These data do not have a modal value in unrounded form, but we can look at modality through a **stem and leaf plot**, which groups data values by leading digits and looks like a histogram that is turned on its side. Unlike a histogram, though, the bar "heights" contain information in the form of the lower digit values. For these data we have the following four stem and leaf plots (with rounding):

Black:

The decimal point is 1 digit to the right of the |

0|6
1|111123455678
2|2244
2|8
3|4
3|6
4|
4|8

Hispanic:

The decimal point is at the |

2|49
3|4
4|233677
5|0269
6|0267
7|
8|4
9|5
10|6
11|
12|2

Asian:

The decimal point is at the |

1|6
2|47899
3|12334778
4|29
5|124
6|7
7|
8|
9|8

White:

The decimal point is 1 digit to the right of the |

3|9
4|
5|66
6|356799
7|3345577
8|00119

Due to the level of measurement and the relatively small number of agency cases (21), we do not have an exact modal value. Nonetheless, the stem and

leaf plot shows that values tend to clump along a single modal region in each case. For instance, if we were to round the Asian values to integers (although this would lose information), then the mode would clearly be 2%.

One way to consider the utility of the three different averages is to evaluate their resistance to large outliers. The **breakdown bound** is the proportion of data values that can become unbounded (go to plus or minus infinity) before the statistic of interest becomes unbounded itself. The mean has a breakdown bound of 0 because even one value of infinity will take the sum to infinity. The median is much more resistant because almost half the values on either side can become unbounded before the median itself goes to infinity. In fact, it is customary to give the median a breakdown bound of 0.5 because as the data size increases, the breakdown bound approaches this value. The mode is much more difficult to analyze in this manner as it depends on the relative placement of values. It is possible for a high proportion of the values to become unbounded provided a higher proportion of the data is concentrated at some other point. If these points are more spread out, however, the infinity point may become the mode and thus the breakdown bound lowers. Due to this uncertainty, the mode cannot be given a definitive breakdown bound value.

8.5 Measures of Dispersion: Variance, Standard Deviation, and MAD

The second most important and common data summary technique is calculating a measure of spread, how dispersed are the data around a central position? Often a measure of centrality and a measure of spread are all that are necessary to give researchers a very good, general view of the behavior of the data. This is particularly true if we know something else, such as that the data are unimodal and symmetric. Even when we do not have such complementary information, it is tremendously useful to know how widely spread the data points are around some center.

The most useful and common measure of dispersion is the **variance**. For n data points $x_1, x_2, \ldots, x_n$, the variance is given by

$$\text{Var}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

The preceding fraction, $\frac{1}{n-1}$, is slightly surprising given the more intuitive $\frac{1}{n}$ for the mean. It turns out that without the -1 component the statistic is **biased**: not quite right on average for the true underlying population quantity. A second closely related quantity is the **standard deviation**, which is simply the square root of the variance:

$$SD(X) = \sqrt{\text{Var}(X)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Since the variance and the standard deviation give the same information, the choice of which one to use is often a matter of taste. There are times, however, when particular theoretical discussions need to use one form over the other.

A very different measure of dispersion is the **median absolute deviation** (MAD). This is given by the form

$$MAD(X) = median(|x_i - median(x)|),$$

for $i = 1, 2, \ldots, n$. That is, the MAD is the median of the absolute deviations from the data median. Why is this useful? Recall our discussion of resistance. The variance (and therefore the standard deviation) is very sensitive to large outliers, more so even than the mean due to the squaring. Conversely, the MAD obviously uses medians, which, as noted, are far more resistant to atypical values. Unfortunately, there are some differences in the way the MAD is specified. Sometimes a mean is used instead of the innermost median, for instance, and sometimes there is a constant multiplier to give asymptotic properties. This is irritating because it means that authors need to say which version they are providing.

★ **Example 8.11: Employment by Race in Federal Agencies, Continued**.

Returning to the values in Table 8.2, we can calculate the three described

measures of dispersion for the racial groups. It is important to remember that none of these three measures is necessarily the "correct" view of dispersion in the absolute sense, but rather that they give different views of it.

Table 8.3. Measures of Dispersion, Race in Agencies

	Black	Hispanic	Asian	White
variance	107.20	6.18	3.16	122.63
standard deviation	10.35	2.49	1.78	11.07
MAD	5.60	1.00	0.60	6.60

8.6 Correlation and Covariance

A key question in evaluating data is to what extent two variables vary together. We expect income and education to vary in the same direction: Higher levels of one are associated with higher levels of the other. That is, if we look at a particular case with a high level of education, we expect to see a high income. Note the use of the word "expect" here, meaning that we are allowing for cases to occur in opposition to our notion without necessarily totally disregarding the general idea. Of course, if a great many cases did not reflect the theory, we would be inclined to dispense with it.

Covariance is a measure of variance with two paired variables. Positive values mean that there is positive varying effect between the two, and negative values mean that there is negative varying effect: High levels of one variable are associated with low levels of another. For two variables of the same length, $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$, the covariance is given by

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

This is very useful because it gives us a way of determining if one variable tends to vary positively or negatively with another. For instance, we would expect income and education levels to vary positively together, and income and prison time to vary negatively together. Furthermore, if there is no relationship

between two variables, then it seems reasonable to expect a covariance near zero. But there is one problem with the covariance: We do not have a particular scale for saying what is large and what is small for a given dataset.

What happens if we calculate the covariance of some variable with itself? Let's take $\text{Cov}(X, Y)$ and substitute in $Y = X$:

$$\begin{aligned}\text{Cov}(X, X) &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \text{Var}(X).\end{aligned}$$

This means that the covariance is a direct generalization of the variance where we have two variables instead of one to consider. Therefore, while we do not generally know the context of the magnitude of the covariance, we can compare it to the magnitude of the variance for X as a reference. So one solution to the covariance scale problem is to measure the covariance in terms of units of the variance of X:

$$\text{Cov}^*(X, Y) = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

In this way units of the covariance are expressed in units of the variance that we can readily interpret. This seems unfair to the Y variable as there may not be anything special about X that deserves this treatment. Now instead let us measure the covariance in units of the standard deviation of X and Y:

$$\text{Cov}^{**}(X, Y) = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

The reason we use the standard deviation of X and Y in the denominator is it scales this statistic conveniently to be bounded by $[-1 : 1]$. That is, if we re-performed our trick of substituting $Y = X$ (or equivalently $X = Y$), then the statistic would be equal to one. In substantive terms, a value of one means that Y covaries *exactly* as X covaries. On the other hand, if we substituted

$Y = -X$ (or conversely $X = -Y$), then the statistic would be equal to negative one, meaning that Y covaries in exactly the opposite manner as X. Since these are the limits of the ratio, any value inbetween represents lesser degrees of absolute scaled covariance. This statistic is important enough to have a name: It is the **correlation coefficient** between X and Y (more formally, it is ***Pearson's Product Moment Correlation Coefficient***), and is usually denoted $\text{cor}(X, Y)$ or r_{XY}.

★ **Example 8.12: An Ethnoarchaeological Study in the South American Tropical Lowlands**. Siegel (1990) looked at the relationship between the size of buildings and the number of occupants in a South Amerindian tropical-forest community located in the upper Essequibo region of Guyana. In such communities the household is the key structural focus in social, economic, and behavioral terms. The substantive point is that the overall settlement area of the community is a poor indicator of ethnographic context in terms of explaining observed relationships, but other ethnoarchaeological measures are quite useful in this regard. Siegel points out that ethnographic research tends not to provide accurate and useful quantitative data on settlement and building dimensions. Furthermore, understanding present-day spatial relationships in such societies has the potential to add to our understanding in historical archaeological studies.

The main tool employed by Siegel was a correlational analysis between floor area of structures and occupational usage. There are four types of structures: residences, multipurpose work structures, storage areas, and community buildings. In these tribal societies it is common for extended family units to share household space including kitchen and storage areas but to reserve a component of this space for the nuclear family. Thus there is a distinction between *households* that are encompassing structures, and the individual *residences* within. Table 8.4 gives correlation coefficients between the size of the floor area for three definitions of space and the family unit for the village of Shefariymo where **Total** is the sum of **Multipurpose** space, **Residence**

space, and **Storage** space.

Table 8.4. CORRELATION COEFFICIENT OF FAMILY UNIT VS. SPACE

Structure Type	Family Unit	Cases	Correlation
Multipurpose	Nuclear	16	0.137
Residence	Nuclear	24	0.662
Total	Nuclear	24	0.714
Multipurpose	Extended	12	0.411
Residence	Extended	11	0.982
Total	Extended	11	0.962

What we see from this analysis is that there exists a positive but weak relationship between the size of the nuclear family and the size of the multipurpose space (0.137). Conversely, there are relatively strong relationships between the size of these same nuclear families and the size of their residences (0.662) and the total family space (0.714). A similar pattern emerges when looking at the size of the extended families occupying these structures except that the relationships are now noticeably stronger. Not surprisingly, the size of the extended family is almost perfectly correlated with residence size and total size.

8.7 Expected Value

Expected value is essentially a probability-weighted average over possible events. Thus, with a fair coin, there are 5 ***expected*** heads in 10 flips. This does not mean that 5 heads will necessarily occur, but that we would be inclined to bet on 5 rather than any other number. Interestingly, with real-life interval measured data you *never* get the expected value in a given experiment because the probability of any one point on the real number line is zero.

The discrete form of the expected value of some random variable X is

$$E[X] = \sum_{i=1}^{k} X_i p(X_i),$$

for k possible events in the discrete sample space $\{X_1, X_2, \ldots, X_k\}$. The continuous form is exactly the same except that instead of summing we need to integrate:

$$E[X] = \int_{-\infty}^{\infty} Xp(X)dX,$$

where the integral limits are adjusted if the range is restricted for this random variable, and these are often left off the integral form if these bounds are obvious. Intuitively, this is easier to understand initially for the discrete case. Suppose someone offered you a game that consisted of rolling a single die with the following payoffs:

die face	1	2	3	4	5	6
X, in dollars	0	1	1	1	2	2

Would you be inclined to play this game if it costs \$2? The expected value of a play is calculated as

$$E[X] = \sum_{i=1}^{6} X_i p(X_i) = \frac{1}{6}(0) + \frac{1}{6}(1) + \frac{1}{6}(1) + \frac{1}{6}(1) + \frac{1}{6}(2) + \frac{1}{6}(2)$$

$$= \$1.67 \text{ (rounded)}.$$

Therefore it would not make sense to pay \$2 to play this game! This is exactly how all casinos around the world function: They calculate the expected value of a play for each of the games and make sure that it favors them. This is not to say that any one person cannot beat the casino, but *on average* the casino always comes out ahead.

So far we have just looked at the expected value of X, but it is a simple matter to evaluate the expected value of some *function* of X. The process inserts the function of X into the calculation above instead of X itself. Discrete and continuous forms are given by

$$E[f(X)] = \sum_{i=1}^{k} f(X_i)p(X_i), \qquad E[f(X)] = \int_{-\infty}^{\infty} f(X)p(X)dX.$$

For instance, if we have the function $f(X) = X^2$, then, given a continuous random variable, $E[X^2] = \int X^2 p(X)dX$.

The calculation of expected value for vectors and matrices is only a little bit more complicated because we have to keep track of the dimension. A $k \times 1$ vector $\mathbf{X}$ of discrete random variables has the expected value $E[\mathbf{X}] = \sum \mathbf{X} p(\mathbf{X})$. For the expected value of a function of the continuous random vector it is common to use the *Riemen-Stieltjes integral* form:

$$E[f(\mathbf{X})] = \int f(\mathbf{X}) dF(\mathbf{X}),$$

where $F(\mathbf{X})$ denotes the joint distribution of the random variable vector $\mathbf{X}$.

In much statistical work expected value calculations are "conditional" in the sense that the average for the variable of interest is taken conditional on another. For instance, the discrete form for the expected value of Y given a specific level of X is

$$E[Y|X] = \sum_{i=1}^{k} Y_i p(Y_i|X).$$

Sometimes expectations are given subscripts when there are more than one random variables in the expression and it is not obvious to which one the expectation holds:

$$\text{Var}_x[E_y[Y|X]] = \text{Var}_x \left[\sum_{i=1}^{k} Y_i p(Y_i|X) \right].$$

8.8 Some Handy Properties and Rules

Since expectation is a summed quantity, many of these rules are obvious, but some are worth thinking about. Let X, Y, and Z be random variables defined in $\Re$ (the real number line), whose expectations are finite.

Finite Expectation Properties for X, Y, Z

- $E[a + bX] = a + bE[X]$, for constants a, b.
- $E[X + Y] = E[X] + E[Y]$
- $E[X + Y|Z] = E[X|Z] + E[Y|Z]$
- $E[Y|Y] = Y$
- $E[E[Y|X]] = E[Y]$, "double expectation"
- If $X \geq Y$, then $E[X] \geq E[Y]$ with probability one
- $|E[X|Y]| \leq E[|X||Y]$ with probability one
- If X and Y are independent, then $E[XY] = E[X]E[Y]$.
  (also holds under the weaker assumption of uncorrelatedness)

The fifth rule, labeled as double expectation, is also called the **law of iterated expectation**, and can be given more generally for functions of Y: $E[E[f(Y)|X]] = E[f(Y)]$. There is also a related set of properties for the variance function:

Finite Variance Properties for X and Y

- Without assuming independence (or uncorrelatedness):
  $\text{Cov}[X, Y] = E[XY] - E[X]E[Y]$
- $\text{Var}[X] = E[X^2] - (E[X])^2$
- $\text{Var}[X|Y] = E[X^2|Y] - (E[X|Y])^2$
- $\text{Cov}[X, Y] = \text{Cov}[X, E[Y|X]]$
- $\text{Var}[Y] = \text{Var}_x[E_y[Y|X]] + E_x[\text{Var}_y[Y|X]]$, "decomposition"

★ **Example 8.13: Craps in a Casino**. In this example we review the rules and winning probabilities at the game craps as a way to illustrate expected

value. The key principle guiding casino management is that every game has negative expected value to the customer. However, craps has many bets that are very nearly "fair" in that the probability of winning is just below 0.5. This tends to attract the more "sophisticated" gamblers, but of course craps still makes money for the house.

The basic process of a craps game is that one person (the shooter) rolls two dice and people bet on the outcome. The most common bet is a "pass," meaning that the player has an unconditional win if the result is 7 or 11 and an unconditional loss if the result is 2, 3, or 12 (called craps). If the result, however, is 4, 5, 6, 8, 9, or 10, then the outcome is called a "point" and the shooter repeats until either the outcome is repeated (a win) or a 7 appears (a loss).

The probabilities associated with each of the 11 possible sums on any given role are

Result	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

So the probability of winning on the first roll is

$$p(7) + p(11) = \frac{6}{36} + \frac{2}{36} = 0.2222222\ldots,$$

and the probability of losing on the first roll is

$$p(2) + p(3) + p(12) = \frac{1}{36} + \frac{2}{36} + \frac{1}{36} = 0.1111111\ldots.$$

Whereas the probability of winning with the point is slightly more compli-

cated:

$$
\begin{aligned}
&p(\text{point}) \times p(\text{repeating point before 7}) \\
&\quad = \frac{24}{36} p(\text{repeating point before 7}) \\
&\quad = \frac{24}{36}\Big[p(4\ or\ 10|\text{point})p(4\ or\ 10 \text{ before } 7) \\
&\qquad + p(5\ or\ 9|\text{point})p(5\ or\ 9 \text{ before } 7) \\
&\qquad + p(6\ or\ 8|\text{point})p(6\ or\ 8 \text{ before } 7)\Big] \\
&\quad = \frac{24}{36}\left[\frac{1}{4}\frac{3/36}{(3+6)/36} + \frac{1}{3}\frac{4/36}{(4+6)/36} + \frac{5}{12}\frac{5/36}{(5+6)/36}\right] \\
&\quad = 0.270707.
\end{aligned}
$$

This means that the probability of winning including the pass is $0.270707 + 0.2222222 = 0.4929292$. So the expected value of a \$5 bet is the expected winnings minus the cost to play: $10 \times 0.4929292 + 0 \times 0.5070708 - 5 = -0.070708$, meaning about negative seven cents. A player can also play "don't pass bar 12," which is the opposite bet except that 12 is considered a tie (the gamblers bet is returned). The probability of winning on this bet is

$$1 - p(\text{pass}) - \frac{1}{2}p(12) = 1 - 0.4929292 - \frac{1}{2}\left(\frac{1}{36}\right) = 0.4931818,$$

which has for a \$5 bet the expected value $10 \times 0.4931818 - 5 = -0.068182$, which is slightly better than a pass. Two variants are the "come" bet where the player starts a pass bet in the middle of the shooter's sequence, and the "don't come" bet where the player starts a don't pass bar 12 in the middle of the shooter's sequence. Predictably these odds are identical to the pass and don't pass bar 12 bets, respectively. Another common bet is the "field," which bets that a 2, 3, 4, 8, 10, 11, or 12 occurs, with the probability of winning: $p(2) + p(3) + p(4) + p(9) + p(10) + p(11) + p(12) = 0.4444444$, but a 2 or 12 pays double, thereby increasing the total expected value: $1.5p(2) + 1p(3) + 1p(4) + 1p(9) + 1p(10) + 1p(11) + 1.5p(12) = 0.4722222$.

The probability of winning with a "big six" or "big eight" bet (6 or 8 comes up before 7) is $\frac{5/36}{(5+6)/36} = 0.4545454$. It is also possible to bet on a specific value for the next roll, and the house sets differing payoffs according to

Probability	**Payoff Odds**	**Expected Value per 0.50**
$p(2) = \frac{1}{36}$	$29 - 1$	$\frac{(29+1)(1)}{2(36)} = 0.4166667$
$p(3) = \frac{2}{36}$	$14 - 1$	$\frac{(14+1)(2)}{2(36)} = 0.4166667$
$p(7) = \frac{6}{36}$	$4 - 1$	$\frac{(4+1)(6)}{2(36)} = 0.4166667$
$p(11) = \frac{2}{36}$	$14 - 1$	$\frac{(14+1)(2)}{2(36)} = 0.4166667$
$p(12) = \frac{1}{36}$	$29 - 1$	$\frac{(29+1)(1)}{2(36)} = 0.4166667$

Obviously these are very poor bets. Another interesting but ill-advised bet is the "hard way," where the player bets that specified doubles occur before the same number in nondoubled form (i.e., the "easy way") or a 7. The associated probabilities are

$$p[(2,2) \text{ before } 7 \textit{ or } (3,1),(1,3)] = \left(\frac{1}{9}\right)\frac{7+1}{2} = 0.4444444$$

$$p[(3,3) \text{ before } 7 \textit{ or } (4,2),(2,4),(5,1),(1,5)] = \left(\frac{1}{11}\right)\frac{9+1}{2} = 0.4545455$$

$$p[(4,4) \text{ before } 7 \textit{ or } (5,3),(3,5),(6,2),(2,6)] = \left(\frac{1}{11}\right)\frac{9+1}{2} = 0.4545455$$

$$p[(10,10) \text{ before } 7 \textit{ or } (6,4),(4,6)] = \left(\frac{1}{9}\right)\frac{7+1}{2} = 0.4444444$$

Note that the payouts used above may differ by casino/area/country/etc. Another bet in this category is "any craps," which means betting on the occurrence of 2, 3, or 12. The payoff is $7/1$, so

$$p(2,3,12) = \left(\frac{4}{36}\right)\frac{7+1}{2} = 0.4444444,$$

and the expected value of a \$5 bet is $10 \times 0.4444444 - 5 = -0.55556$.

The really interesting bet is called "odds," which is sometime billed falsely as giving even money to the player. This bet is allowed only during an

ongoing pass, don't pass, come, or don't come bet, and the actual bet takes place *during* a point. Sometimes an equal or smaller bet compared to the original bet only is allowed, but some casinos will let you double here. The actual bet is that the point value re-occurs before a 7, and one can also bet the "contrary" bet that it won't (evidence of fairness for the second component of the bet). What does this mean in terms of probabilities and payoffs? The "old" game continues with the same probability of winning that benefits the house (0.4929292), but now a new game begins with new odds and the same rules as the point part of a pass play. A new betting structure starts with the payoffs

$$4 \text{ or } 10 \text{ before } 7 : 2{-}1, \quad 5 \text{ or } 9 \text{ before } 7 : 1.5{-}1, \quad 6 \text{ or } 8 \text{ before } 7 : 1.2{-}1.$$

The probabilities of each point value occurring before 7 (recall that you have *one* of these) are

$$p(4\,\text{before}\,7) = \frac{3}{3+6} = \left(\frac{1}{3}\right) \qquad p(10\,\text{before}\,7) = \frac{3}{3+6} = \left(\frac{1}{3}\right)$$

$$p(5\,\text{before}\,7) = \frac{4}{4+6} = \left(\frac{2}{5}\right) \qquad p(9\,\text{before}\,7) = \frac{4}{4+6} = \left(\frac{2}{5}\right)$$

$$p(6\,\text{before}\,7) = \frac{5}{5+6} = \left(\frac{5}{11}\right) \qquad p(8\,\text{before}\,7) = \frac{5}{5+6} = \left(\frac{5}{11}\right).$$

So what are the odds on this new game (betting \$1)? They are

$$4, 10: \quad \left(\frac{1}{3}\right)\left(\frac{2+1}{2}\right) = \frac{1}{2}$$

$$5, 9: \quad \left(\frac{2}{5}\right)\left(\frac{1.5+1}{2}\right) = \frac{1}{2}$$

$$6, 8: \quad \left(\frac{5}{11}\right)\left(\frac{1.2+1}{2}\right) = \frac{1}{2}.$$

The house still comes out ahead because you cannot play the even-money game independently, so the total probability is the average of 0.4929292 and 0.5, which is still below 0.5 (weighted by the relative bets). Also, the

second half of pass bet when you are on the points has the probabilities $p(4, 5, 6, 8, 9, 10 \text{ before } 7) = 0.40606\ldots$ and $p(7 \text{ first}) = 0.59393\ldots$. But apparently most craps players aren't sophisticated enough to use this strategy anyway.

8.9 Inequalities Based on Expected Values

There are a number of "famous" inequalities related to expected values that are sometimes very useful. In all cases X and Y are random variables with expected values that are assumed to exist and are finite. The positive constants k and ℓ are also assumed finite. These assumptions are actually important and the compelling book by Romano and Siegel, *Counterexamples in Probability and Statistics* (1986), gives cases where things can go awry otherwise. A classic reference on just inequalities is *Inequalities* by Hardy, Littlewood, and Polya (1988).

- **Chebychev's Inequality**. If $f(X)$ is a positive and nondecreasing function on $[0, \infty]$, then for all (positive) values of k

$$p(f(X) > k) \leq E[f(X)/k].$$

  A more common and useful form of Chebychev's inequality involves μ and σ, the mean and standard deviation of X (see Section 8.5). For k greater than or equal to 1:

$$p(|X - \mu| \geq k\sigma) \leq 1/k^2.$$

  To relate these two forms, recall that μ is the expected value of X.

- **Markov Inequality**. Similar to Chebychev's Inequality:

$$P[|X| \geq k] \leq E[|X|^{\ell}]/k^{\ell}.$$

- **Jensen's Inequality**. If $f(X)$ is a concave function (open toward the x-axis, like the natural log function), then

$$E[f(X)] \leq f(E[X]).$$

Conversely, if $f(X)$ is a convex function (open away from the x-axis, like the absolute value function), then

$$E[f(X)] \geq f(E[X]).$$

- **Minkowski's Inequality**. For $k > 1$,

$$(E[|X+Y|^k])^{1/k} \leq (E[|X|^k])^{1/k} + (E[|Y|^k])^{1/k}.$$

- **Hölder's Inequality**. For $\frac{1}{k} + \frac{1}{\ell} = 1$,

$$|E[XY]| \leq E[|XY|] \leq (E[|X|]^k)^{1/k} (E[|Y|^\ell])^{1/\ell}.$$

- **Schwarz Inequality**. An interesting special case of Hölder's Inequality where $k = \ell = \frac{1}{2}$:

$$E[|XY|] \leq \sqrt{E[X^2]E[Y^2]}$$

- **Liapounov's Inequality**. For $1 < k < \ell$,

$$(E[|X|]^k)^{1/k} \leq (E[|X|]^\ell)^{1/\ell}.$$

- **Cramer-Rao Inequality**. Given a PDF or PMF conditional on a parameter vector, $f(\mathbf{X}|\boldsymbol{\theta})$, define the *information matrix* as

$$I(\boldsymbol{\theta}) = E\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log(f(\mathbf{X}|\boldsymbol{\theta}))' \frac{\partial}{\partial \boldsymbol{\theta}} \log(f(\mathbf{X}|\boldsymbol{\theta}))\right],$$

and define the vector quantity

$$\alpha = \frac{\partial}{\partial \boldsymbol{\theta}} E[f(\mathbf{X}|\boldsymbol{\theta})].$$

Then

$$\text{Var}(f(\mathbf{X}|\boldsymbol{\theta})) \geq \alpha I(\boldsymbol{\theta}) \alpha.$$

- **Berge Inequality**. Suppose $E[X] = E[Y] = 0$, $\text{Var}[X] = \text{Var}[Y] = 1$, and $\sigma^2 = \text{Cov}[X, Y]$. In other words, these are *standardized* random variables. Then for (positive) k

$$P[\max(|X|, |Y|) \geq k] \leq (1 + \sqrt{1 - \text{Cov}(X,Y)^2})/k^2.$$

Note also that these inequalities apply for conditional expectations as well. For instance, the statement of Liapounov's Inequality conditional on Y is $(E[|X|]^k|Y)^{1/k} \leq (E[|X|]^\ell|Y)^{1/\ell}$.

8.10 Moments of a Distribution

Most (but not all) distributions have a series of **moments** that define important characteristics of the distribution. In fact, we have already seen the first moment, which is the mean or expected value of the distribution. The general formula for the kth moment is based on the expected value:

$$m_k = E[X^k] = \int_X x^k dF(x)$$

for the random variable X with distribution $f(X)$ where the integration takes place over the appropriate support of X. It can also be expressed as

$$m_k = \int_X e^{kx} dF(x),$$

which is more useful in some circumstances. The kth **central moment** is (often called just the "kth moment")

$$m'_k = E[(X - m_1)^k] = \int_X (x - m_1)^k dF(x).$$

So the central moment is defined by a deviation from the mean. The most obvious and important central moment is the variance: $\sigma^2 = E[(X - \bar{X})^2]$. We can use this second central moment to calculate the variance of a PDF or

PMF. This calculation for the exponential is

$$\begin{aligned}
\text{Var}[X] &= E[(X - E[X])^2] \\
&= \int_0^\infty (X - E[X])^2 f(x|\beta) dx \\
&= \int_0^\infty (X - \beta)^2 \frac{1}{\beta} \exp[-x/\beta] dx \\
&= \int_0^\infty (X^2 - 2X\beta + \beta^2) \frac{1}{\beta} \exp[-x/\beta] dx \\
&= \int_0^\infty X^2 \frac{1}{\beta} \exp[-x/\beta] dx \\
&\qquad + \int_0^\infty 2X \exp[-x/\beta] dx + \int_0^\infty \beta \exp[-x/\beta] dx \\
&= (0 - 2\beta^2) + (2\beta^2) + (\beta^2) \\
&= \beta^2,
\end{aligned}$$

where we use integration by parts and L'Hospital's Rule to do the individual integrations.

An important theory says that a distribution function is "determined" by its moments of all orders (i.e., all of them), and some distributions have an infinite number of moments defined. The normal distribution actually has an infinite number of moments. Conversely, the Cauchy PDF has no finite moments at all, even though it is "bell shaped" and looks like the normal (another reason not to use that expression). The **Cauchy distribution**] has the PDF

$$f(x|\beta) = \frac{1}{\beta} \frac{1}{1 + (x - \beta)^2}, \qquad -\infty < x, \beta < \infty.$$

Calculating the first moment with integration by parts gives

$$E[X] = \int_{-\infty}^{\infty} \frac{1}{\beta} \frac{x}{1+(x-\beta)^2} dx$$

$$= \left[x\text{arctan}(x-\beta) - (x-\beta)\text{arctan}(x-\beta) + \frac{1}{2}\log(1+x^2) \right] \Bigg|_{-\infty}^{\infty}$$

$$= \beta\text{arctan}(x-\beta) \Big|_{-\infty}^{\infty} + \frac{1}{2}\log(1+x^2) \Big|_{-\infty}^{\infty}.$$

While the first term above is finite because $\text{arctan}(\pm\infty) = \pm\frac{1}{2}\pi$, the second term is clearly infinite. It is straightforward to show that higher moments are also infinite and therefore undefined (an exercise), and clearly every central moment is also infinite as they all include m_1.

8.11 New Terminology

Berge Inequality, 379
Bernoulli PMF, 340
Bernoulli outcome, 339
beta function, 387
biased, 366:
binomial PMF, 340
breakdown bound, 365
Cauchy PDF, 381
central moment, 380
central tendency, 361
Chebychev's Inequality, 378
continuous data, 330
correlation coefficient, 369
Covariance, 367
Cramer-Rao Inequality, 379
cumulative distribution function, 348
degrees of freedom, 352
discrete, 330
empirical distribution, 362
exponential PDF, 351
gamma PDF, 351
Hölder's Inequality, 379
information entropy, 360
Interval, 332
Jensen's Inequality, 378
law of iterated expectation, 373
Liapounov's Inequality, 379
location-scale family, 354
Markov Inequality, 378
median absolute deviance (MAD), 366
Minkowski's Inequality, 379
moments, 380
Nominal, 331
normal (Gaussian) PDF, 354
Ordinal, 331
Pareto's Law, 352
Poisson PMF, 344
polychotomous, 331
probability density function, 350
probability mass functions (PMF), 339
probit model, 357
random, 334
Ratio, 333
right-continuous function, 348
scale parameter, 351
Schwarz Inequality, 379
Shannon entropy, 360
shape parameter, 351
standard deviation, 366
standard normal, 355
stem and leaf plot, 364
uniform distribution, 358
variance, 366

Exercises

8.1 Indicate the level of measurement and which measure(s) of central tendency can be used for the following:

(a) college education: none, AA, BA/BS, JD, MD/DVM/DDO, Ph.D.;

(b) letter grades;

(c) income given as 0–10K, 10–20K, 30–50K, 50–80K, 100K+;

(d) distance of commute from home to work;

(e) marital status: single, married, widowed, divorced;

(f) working status: employed, unemployed, retired, student;

(g) governmental level: local, state, federal, international;

(h) party: Democrat, Republican, Green, Bull Moose.

8.2 The following data are exam grade percentages: 37, 39, 28, 73, 50, 59, 41, 57, 46, 41, 62, 28, 26, 66, 53, 54, 37, 46, 25.

(a) What is the level of measurement for these data?

(b) Suppose we change the data to create a new dataset in the following way: values from 25 to 45 are assigned "Low," values from 45 to 60 are assigned "Medium," and values from 60 to 75 are assigned "High." Now what is the level of measurement for these data?

(c) Now suppose we take the construction from (b) and assign "Low" and "High" to "Atypical" and assign "Medium" to "Typical." What is the level of measurement of this new dataset?

(d) Calculate the mean and standard deviation of each of the three datasets created above.

8.3 Morrison (1977) gave the following data for Supreme Court vacancies from 1837 to 1932:

Number of Vacancies/Year	0	1	2	3	4+
Number of Years for Event	59	27	9	1	0

Fit a distribution to these data, estimating any necessary parameters. Using this model, construct a table of expected versus observed frequencies by year.

8.4 The National Taxpayers Union Foundation (NTUF), an interest group that advocates reduced government spending, scores House members on the budgetary impact of their roll call votes. A "spending" vote is one in favor of a bill or amendment that increases federal outlays and a "saving" vote is one that specifically decreases federal spending (i.e., program cuts). The fiscal impact of each House member's vote is cross-indexed and calculated as the total increase to the budget or the total decrease to the budget. The NTUF supplies these values along with a ranking of each member's "fiscal responsibility," calculated by adding all positive and negative fiscal costs of each bill voted on by each member and then ranking members by total cost. What is the level of measurement of the NTUF fiscal responsibility scale? Since House members' values are compared in NTUF public statements, is there a different level of measurement being implied?

8.5 Suppose you had a Poisson process with intensity parameter $\lambda = 5$. What is the probability of getting exactly 7 events? What is the probability of getting exactly 3 events? These values are the same distance from the expected value of the Poisson distribution, so why are they different?

8.6 Given the following PMF:

$$f(x) = \begin{cases} \frac{3!}{x!(3-x)!}\left(\frac{1}{2}\right)^3 & x = 0, 1, 2, 3 \\ 0 & \text{otherwise,} \end{cases}$$

(a) prove that this *is* in fact a PMF;

(b) find the expected value;

(c) find the variance;

(d) Derive the CDF.

8.7 Let X be the event that a single die is rolled and the resulting number is even. Let Y be the event describing the actual number that results from the roll (1–6). Prove the independence or nonindependence of these two events.

8.8 Suppose X_1 and X_2 are independent identically distributed (iid) random variables according to the PMF Bernoulli, $p = \frac{1}{2}$. Are $Y_1 = X_1 + X_2$ and $Y_2 = |X_1 - X_2|$ correlated? Are they independent?

8.9 Suppose we have a PMF with the following characteristics: $p(X = -2) = \frac{1}{5}$, $p(X = -1) = \frac{1}{6}$, $p(X = 0) = \frac{1}{5}$, $p(X = 1) = \frac{1}{15}$, and $p(X = 2) = \frac{11}{30}$. Define the random variable $Y = X^2$. Derive the PMF of Y and prove that it is a PMF. Calculate the expected value and variance of Y.

8.10 Charles Manski (1989) worried about missing data when the outcome variable of some study had missing values rather than when the variables assumed to be causing the outcome variable had missing values, which is the more standard concern. Missing values can cause series problems in making probabilistic statements from observed data. His first concern was notated this way: "Suppose each member of a population is characterized by a triple (y, x, z) where y is a real number, z is a binary indicator, and x is a real number vector." The problem is that, in collecting these data, (z, x) are always observed, but y is observed only when $z = 1$. The quantity of interest is $E(y|x)$. Use the Theorem of Total Probability to express this conditional probability when the data only provide $E(y|x, z = 1)$ and we cannot assume mean independence: $E(y|x) = E(y|x, z = 1) = E(y|x, z = 0)$.

8.11 Twenty developing countries each have a probability of military coup of 0.01 in any given year. We study these countries over a 10-year period.

(a) How many coups do you expect in total?

(b) What is the probability of four coups?

(c) What is the probability that there will be no coups during this period?

8.12 Show that the full parameter normal PDF $f(X|\mu, \sigma^2)$ reduces to the standard normal PDF when $\mu = 0$ and $\sigma^2 = 1$.

8.13 Use the exponential PDF to answer the following questions.

(a) Prove that the exponential form *is* a PDF.

(b) Derive the CDF.

(c) Prove that the exponential distribution is a special case of the gamma distribution.

8.14 Use the normal PDF to answer the following questions.

(a) If a normal distribution has $\mu = 25$ and $\sigma = 25$, what is the 91st percentile of the distribution?

(b) What is the 6th percentile of the distribution of part (a)?

(c) The width of a line etched on an integrated circuit chip is normally distributed with mean $3.000\mu m$ and standard deviation 0.150. What width value separates the widest 10% of all such lines from the other 90%?

8.15 A function that can be used instead of the probit function is the logit function: $\Lambda(X) = \frac{\exp(X)}{1+\exp(X)}$. Plot both the logit function and the probit function in the same graph and compare. What differences do you observe?

8.16 The **beta function** is defined for nonnegative values a and b as:

$$B(a, b) = \int_0^1 x^{a-1}(1 - x)^{b-1}dx.$$

This form is used in some statistical problems and elsewhere. The relationship between the beta and gamma functions is given by

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}.$$

Prove this using the properties of PDFs.

8.17 Prove that $E[Y|Y] = Y$.

8.18 Suppose that the performance of test-takers is normally distributed around a mean, μ. If we observe that 99% of the students are within 0.194175 of the mean, what is the value of σ?

8.19 Calculate the entropy of the distribution $\mathcal{B}(n = 5, p = 0.1)$ and the distribution $\mathcal{B}(n = 3, p = 0.5)$. Which one is greater? Why?

8.20 We know that the reaction time of subjects to a specific visual stimuli is distributed gamma with $\alpha = 7$ and $\beta = 3$, measured in seconds.

(a) What is the probability that the reaction time is greater that 12 seconds?

(b) What is the probability that the reaction time will be between 15 and 21 seconds?

(c) What is the 95th percentile value of this distribution?

8.21 Show that the second moment of the Cauchy distribution is infinite and therefore undefined.

8.22 The following data are temperature measurements in Fahrenheit. Use these data answer the following questions.

38.16	52.68	53.47	50.18	49.13

(a) Is the median bigger or smaller than the mean?

(b) Calculate the mean and standard deviation.

(c) What is the level of measurement for these data?

(d) Suppose we transformed the data in the following way: Values from 0 to 40 are assigned "Cold," values from 41 to 70 are assigned "Medium," and values above 71 are assigned "Hot." Now what is the level of measurement for these data?

(e) Suppose we continue to transform the data in the following way: "Cold" and "Hot" are combined into "Uncomfortable," and "Medium" is renamed "Comfortable." What is the level of measurement for these data?

8.23 The following is a stem and leaf plot for 20 different observations (stem = tens digit). Use these data to answer the questions.

0	7	8	9	9	9	
1	0	1	5	7	7	9
2	0	1	1	8		
3	2	4				
4						
5						
6						
7	5					
8	3	9				

(a) Is the median bigger or smaller than the mean?

(b) Calculate the 10% trimmed mean.

(c) Make a frequency distribution with relative *and* relative cumulative frequencies.

(d) Calculate the standard deviation.

(e) Identify the IQR.

8.24 Nine students currently taking introductory statistics are randomly selected, and both the first midterm score (x) and the second midterm score (y) are determined. Three of the students have the class at 8 A.M., another three have it at noon, and the remaining three have a night class.

8 A.M.	(70,60)	(72,83)	(94,85)
Noon	(80,72)	(60,74)	(55,58)
Night	(45,63)	(50,40)	(35,54)

(a) Calculate the sample correlation coefficient for the nine (x, y) pairs.

(b) Let $\bar{x}_1$ = the average score on the first midterm for the 8 A.M. students and $\bar{y}_1$ = the average score on the second midterm for these students. Let $\bar{x}_2$ and $\bar{y}_2$ be these averages for the noon students, and $\bar{x}_3$ and $\bar{y}_3$ be these averages for the evening students. Calculate r for these three $(\bar{x}, \bar{y})$ pairs.

(c) Construct a scatterplot of the nine (x, y) pairs and construct another one of the three $(\bar{x}, \bar{y})$ pairs. Does this suggest that a correlation coefficient based on averages (an "ecological" correlation) might be misleading? Explain.

8.25 The *Los Angeles Times* (Oct. 30, 1983) reported that a typical customer of the 7-Eleven convenience stores spends \$3.24. Suppose that the average amount spent by customers of 7-Eleven stores is the reported value of \$3.24 and that the standard deviation for the amount of sale is \$8.88.

(a) What is the level of measurement for these data?

(b) Based on the given mean and standard deviation, do you think that the distribution of the variable *amount of sale* could have been symmetric in shape? Why or why not?

(c) What can be said about the proportion of all customers that spend more than \$20 on a purchase at 7-Eleven?

8.26 The following are Grunfeld's General Electric data (see Maddala 1977) with the following variables: GEI = "Gross investment GE," GEC = "Lagged Capital Stock GE," and GEF = "Lagged Value of GE shares."

Year	GEI	GEC	GEF
1935	33.1	1170.6	97.8
1936	45.0	2015.8	104.4
1937	77.2	2803.3	118.0
1938	44.6	2039.7	156.2
1939	48.1	2256.2	172.6
1940	74.4	2132.2	186.6
1941	113.0	1834.1	220.9
1942	91.9	1588.0	287.8
1943	61.3	1749.4	319.9
1944	56.8	1687.2	321.3
1945	93.6	2007.7	319.6
1946	159.9	2208.3	346.0
1947	147.2	1656.7	456.4
1948	146.3	1604.4	543.4
1949	98.3	1431.8	618.3
1950	93.5	1610.5	647.4
1951	135.2	1819.4	671.3
1952	157.3	2079.7	726.1
1953	179.5	2371.6	800.3
1954	189.6	2759.9	888.9

(a) Calculate the variance and the MAD of each of the three variables.

(b) Calculate the correlation coefficients. Truncate the variables such that there are no values to the right of the decimal point and recalculate the correlation coefficients. Do you see a difference? Why or why not?