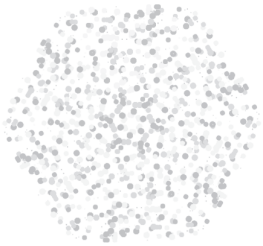
7



DISTRIBUTIONS FOR DESCRIPTION

This chapter covers some methods of describing a data set, via a number of strategies of increasing complexity. The first approach, in Section 7.1, consists of simply looking at summary statistics for a series of observations about a single variable, like its mean and variance. It imposes no structure on the data of any sort. The next level of structure is to assume that the data is drawn from a distribution; instead of finding the mean or variance, we would instead use the data to estimate the parameters that describe the distribution. The simple statistics and distributions in this chapter are already sufficient to describe rather complex models of the real world, because we can chain together multiple distributions to form a larger model. Chapter 8 will take a slightly different approach to modeling a multidimensional data set, by projecting it onto a subspace of few dimensions.

7.1 MOMENTS

The first step in analyzing a data set is always to get a quick lay of the land: where do the variables generally lie? How far do they wander? As variable A goes up, does variable B follow?

※ ESTIMATOR VOCABULARY A *statistic* is the output of a function that takes in data, typically of the form $f : \mathbb{R}^n \to \mathbb{R}$. That is, a statistic takes in data and summarizes it to a single dimension. Common statistics include the mean of $\mathbf{x}$, its variance, $\max(\mathbf{x})$, the covariance of $\mathbf{x}$ and $\mathbf{y}$, or the regression parameter β_2 from a regression of $\mathbf{X}$ on $\mathbf{y}$ (which could be written in the form $\beta_2(\mathbf{X}, \mathbf{y})$).

Thus, many of the means of describing a data set, such as writing down its mean, could be described as the generation of statistics. One goal of writing down a statistic is dimension reduction: simply summarizing the data via a few human-comprehensible summary statistics, such as the data set's mean and variance.

Another goal, which is more often the case, is to use the statistics of the data, $\mathbf{X}$, to estimate the same statistic of the population. Let $\mathcal{P}$ signify the population. When the US Census Bureau said in an August 2006 press release[1] that 46.6 million people in the United States have no health insurance, they meant that the count of people in the Current Population Survey that did not have health insurance (a sample statistic) indicated that the count of people in the United States without health insurance (a population statistic) was 46.6 million. Is the estimate of the statistic based on the sample data, $\hat{\beta} \equiv f(\mathbf{X})$, a valid estimate of the population statistic, $\beta \equiv f(\mathcal{P})$? There are several ways to make this a precise question. For example, as $\mathbf{X}$ grows larger, does $\hat{\beta} \to \beta$? Do there exist estimates of β that have smaller variance than $\hat{\beta}$? After discussing some desirable qualities in an estimator, we will begin with some simple statistics.

※ *Evaluating an estimator* From a given population, one could take many possible samples, say $\mathbf{X}_1$, $\mathbf{X}_2$, ..., which means that there could be many possible calculated statistics, $\hat{\beta}_1 = f(\mathbf{X}_1)$, $\hat{\beta}_2 = f(\mathbf{X}_2)$, ...

There are many means of describing whether the collection of statistics $\hat{\beta}_i$ (for $i = 1, i = 2, \dots$) is a precise and accurate estimate of the true value of β. You will see in the sections to follow that intuitive means of estimating a population statistic sometimes work on all of these scales at once, and sometimes fail.

- Unbiasedness: The expected value of $\hat{\beta}$ (discussed in great detail below) equals the true population value: $E(\hat{\beta}_i) = \beta$.
- Variance: The variance is the expected value of the squared distance to the expected value: $E\left((\hat{\beta}_i - E(\hat{\beta}))^2\right)$. The variance is also discussed in detail below.
- Mean squared error: MSE of $\hat{\beta} \equiv E(\hat{\beta} - \beta)^2$. Below we will see that the MSE equals the variance plus the square of bias. So if $\hat{\beta}$ is an unbiased estimator of β, meaning that $E(\hat{\beta}) = \beta$, then the MSE is equivalent to the variance, but as the bias increases, the difference between MSE and variance grows.
- Efficiency: An estimator $\hat{\beta}$ is efficient if, for any other estimator $\tilde{\beta}$, $MSE(\hat{\beta}) \leq MSE(\tilde{\beta})$. If $\hat{\beta}$ and $\tilde{\beta}$ are unbiased estimators of the same β, then this reduces to $\mathrm{var}(\hat{\beta}) \leq \mathrm{var}(\tilde{\beta})$, so some authors describe an efficient estimator as the unbiased

[1] US Census Bureau, "Income Climbs, Poverty Stabilizes, Uninsured Rate Increases," Press release #CB06-136, 29 August 2006, `http://www.census.gov/Press-Release/www/releases/archives/income_wealth/007419.html`.

estimator with minimum variance among all unbiased estimators.
We test this using inequality 10.1.7 (page 333), that the variance must be greater than or equal to the Cramér–Rao lower bound. If $\mathrm{var}(\hat{\beta})$ equals the CRLB, we know we have a minimum variance.

- BLUE: $\hat{\beta}$ is the Best Linear Unbiased Estimator if $\mathrm{var}(\hat{\beta}) \leq \mathrm{var}(\tilde{\beta})$ for all linear unbiased estimators $\tilde{\beta}$, and $\hat{\beta}$ is itself a linear function and unbiased.
- Aymptotic unbiasedness: Define $\hat{\beta}_n = f(x_1, \ldots, x_n)$. For example, $\hat{\mu}_1 = x_1$, $\hat{\mu}_2 = (x_1 + x_2)/2$, $\hat{\mu}_3 = (x_1 + x_2 + x_3)/3$, ... Then $\lim_{n\to\infty} E(\hat{\beta}_n) = \beta$. Clearly, unbiasedness implies asymptotic unbiasedness.
- Consistency: $\mathrm{plim}(\hat{\beta}_n) = \beta$, i.e., for a fixed small ϵ, $\lim_{n\to\infty} P(|\hat{\beta}_n - \beta| > \epsilon) = 0$. Equivalently, $\lim_{n\to\infty} P((\hat{\beta}_n - \beta)^2 > \epsilon^2) = 0$.
  One can verify consistency using *Chebychev's inequality*; see, e.g., Casella & Berger (1990, p 184).
  In a sense, consistency is the asymptotic analog to the low MSE condition. If MSE goes to zero, then consistency follows (but not necessarily vice versa).
  However, a biased estimator or a high-variance estimator may have a few things going for it, but an inconsistent estimator is just a waste. You could get yourself two near-infinite samples and find that $\hat{\boldsymbol{\beta}}$ is different for each of them—and then what are you supposed to pick?
- Asymptotic efficiency: $\mathrm{var}(\hat{\beta}) \to$ the Cramér–Rao lower bound. This makes sense only if $\hat{\beta}$'s asymptotic distribution has a finite mean and variance and $\hat{\beta}$ is consistent.

EXPECTED VALUE Say that any given value of x has probability $p(x)$. Then if $f(x)$ is an arbitrary function,

$$E\left(f(x)\right) = \int_{\forall x} f(x)p(x)dx.$$

The $p(x)dx$ part of the integral is what statisticians call a *measure* and everyone else calls a weighting. If $p(x)$ is constant for all x, then every value of x gets equal weighting, as does every value of $f(x)$. If $p(x_1)$ is twice as large as $p(x_2)$, meaning that we are twice as likely to observe x_1, then $f(x_1)$ gets double weighting in the integral.

If we have a vector of data points, $\mathbf{x}$, consisting of n elements x_i, then we take each single observation to be equally likely: $p(x_i) = \frac{1}{n}$, $\forall i$. The expected value for a sample then becomes the familiar calculation

$$E(\mathbf{x}) = \frac{\sum_i x_i}{n},$$

and (given no further information about the population) is the Best Unbiased Estimator of the true mean μ.[2]

[2]The term *expected value* implies that the mean is what we humans actually expect will occur. But if I have

VARIANCE AND ITS DISSECTIONS The *variance* for discrete data is the familiar formula of the mean of the squared distance to the average. Let $\overline{\mathbf{x}}$ indicate the mean of the data vector $\mathbf{x}$; then the best unbiased estimate of the variance of the sample is

$$\mathrm{var}(\mathbf{x}) = \frac{1}{n}\sum_i (x_i - \overline{\mathbf{x}})^2 . \tag{7.1.1}$$

Degrees of freedom

Rather than calculating the variance of a sample, say that we seek the variance of a population, and have only a sample from which to estimate the variance. The best unbiased estimate of the variance of the population is

$$\widehat{\mathrm{var}}(\mathbf{x}) = \frac{1}{n-1}\sum_i (x_i - \overline{\mathbf{x}})^2 . \tag{7.1.2}$$

We can think of the sum being divided by $n-1$ instead of n (as in Equation 7.1.1) because there are only $n-1$ random elements in the sum: given the mean $\overline{\mathbf{x}}$ and $n-1$ elements, the nth element is deterministically solved. That is, there are only $n-1$ *degrees of freedom*. An online appendix to this book provides a more rigorous proof that Equation 7.1.2 is an unbiased estimator of the population variance.

As $n \to \infty$, $1/n \approx 1/(n-1)$, so both the estimate of variance based on $1/n$ and on $1/(n-1)$ are asymptotically unbiased estimators of the population variance.

The number of degrees of freedom (*df*) will appear in other contexts throughout the book. The *df* indicates the number of dimensions in which the data could possibly vary. With no additional information, this is just the number of independently drawn variables, but there may be more restrictions on the data. Imagine three variables, which would normally have three dimensions, with the added restriction that $x_1 + 2x_2 = x_3$. Then this defines a plane (which happens to be orthogonal to the vector $(1, 2, -1)$ and goes through the origin). That is, by adding the restriction, the data points have been reduced to a two-dimensional surface. For the sample variance, the restriction is that the mean of the sample is $\hat{\mu}$.

The square root of the variance is called the *standard deviation*. It is useful because the Normal distribution is usually described in terms of the standard deviation (σ) rather than the variance (σ^2). Outside of the context of the Normal, the variance is far more common.

The variance is useful in its own right as a familiar measure of dispersion. But it can also be decomposed various ways, depending on the situation, to provide still more information, such as how much of the variance is due to bias, or how much variation is explained by a linear regression. Since information is extracted from the decomposition of variance time and time again throughout classical statistics, it is worth going over these various dissections.

Recall from basic algebra that the form $(x+y)^2$ expands to $x^2 + y^2 + 2xy$. In some special cases the unsightly $2xy$ term can be eliminated or merged with another term, leaving the pleasing result that $(x + y)^2 = x^2 + y^2$.

a one in a million chance of winning a two million dollar lottery, there are no states of the world where I am exactly two dollars wealthier. Further, research pioneered by Kahneman and Tversky (e.g., Kahneman *et al.* (1982)) found that humans tend to focus on other features of a probability distribution. They will consider events with small probability to either have zero probability or a more manageable value (e.g., reading $p = 1\mathrm{e}{-7}$ as $p = 1\mathrm{e}{-3}$). Or, they may assume the most likely state of the world occurs with certainty. Bear in mind that human readers of your papers may be interested in many definitions of expectation beyond the mean.

Throughout the discussion below $\bar{\mathbf{x}} = E[\mathbf{x}]$; that is $\bar{\mathbf{x}}$ is constant for a given data set. The expectation of a constant is the constant itself, so $E[\bar{\mathbf{x}}]$ is simply $\bar{\mathbf{x}}$; and $E[\mathbf{y}^2\bar{\mathbf{x}}]$ would expand to $\frac{1}{n}\sum_{i=1}^{n}\left[y_i^2\bar{\mathbf{x}}\right] = \bar{\mathbf{x}}\cdot\frac{1}{n}\sum_{i=1}^{n} y_i^2 = \bar{\mathbf{x}}\cdot E[\mathbf{y}^2]$.

The first breakdown of variance is the equation as above:

$$\begin{aligned}\mathrm{var}(\mathbf{x}) &= E\left[(\mathbf{x}-\bar{\mathbf{x}})^2\right] \\ &= E\left[\mathbf{x}^2 - 2\mathbf{x}\bar{\mathbf{x}} + \bar{\mathbf{x}}^2\right] \\ &= E[\mathbf{x}^2] - E[2\mathbf{x}\bar{\mathbf{x}}] + E[\bar{\mathbf{x}}^2] \\ &= E[\mathbf{x}^2] - 2E[\mathbf{x}]^2 + E[\mathbf{x}]^2 \\ &= E[\mathbf{x}^2] - E[\mathbf{x}]^2.\end{aligned}$$

Read this as: $\mathrm{var}(\mathbf{x})$ is the expectation of the squared values minus the square of the expected value. This form simplifies many transformations and basic results of the type that frequently appear in probability homework questions.

$\mathbb{Q}_{7.1}$ Write a function to display $\mathrm{var}(\mathbf{x})$, $E[\mathbf{x}^2]$, $E[\mathbf{x}]^2$, and $E[\mathbf{x}^2] - E[\mathbf{x}]^2$ for any input data, then use it to verify that the first and last expressions are equal for a few columns of data selected from any source on hand.

Mean squared error The next breakdown appears with the mean squared error. Say that we have a biased estimate of the mean, $\tilde{\mathbf{x}}$; if you had the true mean $\bar{\mathbf{x}}$, then you could define the bias as $(\tilde{\mathbf{x}} - \bar{\mathbf{x}})$. It turns out that the MSE is a simple function of the true variance and the bias. The value can be derived by inserting $-\bar{\mathbf{x}} + \bar{\mathbf{x}} = 0$ and expanding the square:

$$\begin{aligned}MSE \equiv E\left[(\mathbf{x}-\tilde{\mathbf{x}})^2\right] \\ &= E\left[\left((\mathbf{x}-\bar{\mathbf{x}}) + (\bar{\mathbf{x}}-\tilde{\mathbf{x}})\right)^2\right] \\ &= E\left[(\mathbf{x}-\bar{\mathbf{x}})\right]^2 + 2E\left[(\mathbf{x}-\bar{\mathbf{x}})(\bar{\mathbf{x}}-\tilde{\mathbf{x}})\right] + E\left[(\bar{\mathbf{x}}-\tilde{\mathbf{x}})^2\right] \\ &= \mathrm{var}(\mathbf{x}) - 2\cdot\mathrm{bias}(\tilde{\mathbf{x}})E\left(\mathbf{x}-\bar{\mathbf{x}}\right) + \mathrm{bias}(\tilde{\mathbf{x}})^2 \\ &= \mathrm{var}(\mathbf{x}) + \mathrm{bias}(\tilde{\mathbf{x}})^2\end{aligned}$$

In this case the middle term drops out because $E(\mathbf{x}-\bar{\mathbf{x}}) = 0$, and the MSE breaks down to simply being the variance of $\mathbf{x}$ plus the bias squared.

Within group/among group variance The next breakdown of variance, common in ANOVA estimations (where ANOVA is short for analysis of variance), arises when the data is divided into a set of groups. Then the total variance over the entire data set could be expressed as *among group variance* and *within group variance*. Above, $\mathbf{x}$ consisted of a homogeneous sequence of elements x_i, $i = \{1, \ldots, n\}$, but now break it down into subgroups x_{ij}, where j indicates the group and i indicates the elements within the group. There is thus a mean $\bar{\mathbf{x}}_j$ for each group j, which is the simple mean for the n_j elements in that group. The unsubscripted $\bar{\mathbf{x}}$ continues to represent the mean of the entire sample. With that notation in hand, a similar breakdown to those given above can be applied to the groups:

$$\begin{aligned}
\operatorname{var}(\mathbf{x}) &= E\left[(\mathbf{x} - \bar{\mathbf{x}})^2\right] \\
&= \frac{1}{n}\sum_j \left[\sum_{i=1}^{n_j}(x_{ij} - \bar{\mathbf{x}}_j + \bar{\mathbf{x}}_j - \bar{\mathbf{x}})^2\right] \\
&= \frac{1}{n}\sum_j \left[\sum_{i=1}^{n_j}(x_{ij} - \bar{\mathbf{x}}_j)^2 + \sum_{i=1}^{n_j}(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^2 + 2\sum_{i=1}^{n_j}(x_{ij} - \bar{\mathbf{x}}_j)(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})\right] \qquad (7.1.3) \\
&= \frac{1}{n}\sum_j \left[\sum_{i=1}^{n_j}(x_{ij} - \bar{\mathbf{x}}_j)^2 + \sum_{i=1}^{n_j}(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^2\right] \qquad (7.1.4) \\
&= \frac{1}{n}\sum_j \left[n_j \operatorname{var}(\mathbf{x}_j)\right] + \frac{1}{n}\sum_j \left[n_j(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^2\right]
\end{aligned}$$

The transition from Equation 7.1.3 to 7.1.4 works because $(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})$ is constant for a given j, and $\sum_{i=1}^{n_j}(x_{ij} - \bar{\mathbf{x}}_j) = \bar{\mathbf{x}}_j - \bar{\mathbf{x}}_j = 0$. Once again, the unsightly middle term cancels out, and we are left with an easily interpretable final equation. In this case, the first element is the weighted mean of within-group variances, and the second is the weighted among-group variance, where each group is taken to be one unit at $\bar{\mathbf{x}}_j$, and then the variance is taken over this set of group means.

The `data-metro.db` set gives average weekday passenger boardings at every station in the Washington Metro subway system, from the founding of the system in November 1977 to 2007.[3] The system has five lines (Blue, Green, Orange, Red, Yellow), and Listing 7.1 breaks down the variance of ridership on the Washington Metro into within-line and among-line variances.

- Line 20 is the query to join the `riders` and `lines` tables. The parens mean that it

[3] As a Washington-relevant detail, all post-'77 measurements were made in May, outside the peak tourist season.

```
1  #include <apop.h>
2
3  void variance_breakdown(char *table, char *data, char *grouping){
4      apop_data* aggregates = apop_query_to_mixed_data("mmw",
5                  "select var_pop(%s) var, avg(%s) avg, count(*) from %s group by %s",
6                  data, data, table, grouping);
7      APOP_COL_T(aggregates, "var", vars);
8      APOP_COL_T(aggregates, "avg", means);
9      double total= apop_query_to_float("select var_pop(%s) from %s", data, table);
10     double mean_of_vars = apop_vector_weighted_mean(vars, aggregates->weights);
11     double var_of_means = apop_vector_weighted_var(means, aggregates->weights);
12     printf("total variance: %g\n", total);
13     printf("within group variance: %g\n", mean_of_vars);
14     printf("among group variance: %g\n", var_of_means);
15     printf("sum within+among: %g\n", mean_of_vars + var_of_means);
16 }
17
18 int main(){
19     apop_db_open("data-metro.db");
20     char joinedtab[] = "(select riders/100 as riders, line from riders, lines \
21                 where lines.station =riders.station)";
22     variance_breakdown(joinedtab, "riders", "line");
23 }
```

Listing 7.1 Decomposing variance between among-group and within-group. Online source: `amongwithin.c`.

can comfortably be inserted into a `from` clause, as in the query on line four.[4]

- The query on line 9 pulls the total variance—the total sum of squares—and the query on lines 4–6 gets the within-group variances and means.
- Lines 10 and 11 take the weighted mean of the variances and the weighted variance of the means.
- Lines 14–17 print the data to screen, showing that these two sub-calculations do indeed add up to the total variance.

Rewrite the program to use the `data-wb.db` data set (including the `classes` table) to break down the variance in GDP per capita into within-class and among-class variance.

Within-group and among-group variance is interesting by itself. To give one example, Glaeser *et al.* (1996, Equation 8) break down the variance in crime into

[4]Using a subquery like this may force the SQL interpreter to re-generate the subtable for every query, which is clearly inefficient. Therefore, when using functions like `apop_anova` in the wild, first run a `create table ...` query to join the data, perhaps index the table, and then send that table to the `apop_anova` function.

within-city and among-city variances, and find that among-city variance is orders of magnitude larger than within-city variance.

Returning to the Metro data, we could group data by year, and look for within- and among-group variation in that form, or we could group data by line and ask about within- and among-group variation there.

```
#include <apop.h>

int main(){
    apop_db_open("data-metro.db");
    char joinedtab[] = "(select year, riders, line \
                        from riders, lines \
                        where riders.station = lines.station)";
    apop_data_show(apop_anova(joinedtab, "riders", "line", "year"));
}
```

Listing 7.2 Produce a two-way ANOVA table breaking variance in per-station passenger boardings into by-year effects, by-line effects, an interaction term, and the residual. Online source: `metroanova.c`.

Listing 7.2 produces an *ANOVA table*, which is a spreadsheet-like table giving the within-group and among-group variances. The form of the table dates back to the mid-1900s—ANOVA is basically the most complex thing that one can do without a matrix-inverting computer, and the tabular form facilitates doing the calculation with paper, pencil, and a desk calculator. But it still conveys meaning even for those of us who have entirely forgotten how to use a pencil.

The first three rows of the output present the between-group sum of squares. That is, if we were to aggregate all the data points for a given group into the mean, how much variation would remain? With *grouping1* and *grouping2*, there are three ways to group the data: `group by` *grouping1* [(Green line), (Red line), ...], `group by` *grouping2* [(1977), (1978), ..., (2007)], and the *interaction*: `group by` *grouping1*, *grouping2* [(Green Line, 1977), (Red Line, 1977), ...(Green Line, 2007), (Red Line, 2007)]. Using algebra much like that done above, we can break down the total sum of squares into (weighted sum of squares, *grouping1*) + (weighted sum of squares, *grouping2*) + (weighted sum of squares, *grouping1*, *grouping2*) + (weighted sum of squares, residual).

We can compare the weighted grouped sums to the residual sum, which is listed as the F statistic in the ANOVA table. As will be discussed in the chapter on testing (see page 309), an F statistic over about two is taken to indicate that the grouping explains more variation than would be explained via a comparable random grouping of the data. The output of this example shows that the grouping by year is very significant, as is the more refined interaction grouping by line and year, but the

grouping by line is not significant, meaning that later studies may be justified in not focusing on how the subway stations are broken down into lines.

$\mathbb{Q}_{7.3}$ Most stations are on multiple lines, so a station like Metro Center is included in the Blue, Orange, and Red groups. In fact, the Yellow line has only two stations that it doesn't share with other lines. [You can easily find an online map of the Washington Metro to verify this.] This probably causes us to underestimate the importance of the per-line grouping. How would you design a grouping that puts all stations in only one group? It may help in implementation to produce an intermediate table that presents your desired grouping. Does the ANOVA using your new grouping table show more significance to the line grouping?

By changing the second group in the code listing from `"year"` to `NULL`, we would get a one-way ANOVA, which breaks down the total sum of squares into just (weighted sum of squares, *grouping1*) + (weighted sum of squares, residual). The residual sum of squares is therefore larger, the *df* of the residual is also larger, and in this case the overall change in the F statistic is not great.

※ *Regression variance* Next, consider the OLS model, which will be detailed in Section 8.2.1. In this case, we will break down the observed value to the estimated value plus the error: $\mathbf{y} = \mathbf{y}_{\text{est}} + \boldsymbol{\epsilon}$.

$$\begin{aligned}\text{var}(\mathbf{y}) &= E\left[(\mathbf{y}_{\text{est}} + \boldsymbol{\epsilon} - \bar{\mathbf{y}})^2\right] \\ &= E\left[(\mathbf{y}_{\text{est}} - \bar{\mathbf{y}})^2\right] + E[\boldsymbol{\epsilon}^2] + 2E\left[(\mathbf{y}_{\text{est}} - \bar{\mathbf{y}})\boldsymbol{\epsilon}\right]\end{aligned}$$

It will be shown below that $\mathbf{y}_{\text{est}}\boldsymbol{\epsilon} = \boldsymbol{\beta}_{\text{OLS}}\mathbf{X}\boldsymbol{\epsilon} = 0$ (because $\mathbf{X}\boldsymbol{\epsilon} = 0$), and $\bar{\epsilon} = 0$, so $E[\bar{\mathbf{y}}\boldsymbol{\epsilon}] = \bar{\mathbf{y}}E[\boldsymbol{\epsilon}] = 0$. Thus, the $2E[\dots]$ term is once again zero, and we are left with

$$\text{var}(\mathbf{y}) = E\left[(\mathbf{y}_{\text{est}} - \bar{\mathbf{y}})^2\right] + E[\boldsymbol{\epsilon}^2] \tag{7.1.5}$$

Make the following definitions:

$$\begin{aligned} SST &\equiv \text{total sum of squares} \\ &= E\left[(\mathbf{y} - \bar{\mathbf{y}})^2\right] \\ &= \text{var}(\mathbf{y}) \\ SSR &\equiv \text{Regression sum of squares} \\ &= E\left[(\mathbf{y}_{\text{est}} - \bar{\mathbf{y}})^2\right] \\ SSE &\equiv \text{Sum of squared errors} \\ &= E[\boldsymbol{\epsilon}^2] \end{aligned}$$

Then the expansion of $\text{var}(\mathbf{y})$ in Equation 7.1.5 could be written as

$$SST = SSR + SSE.$$

This is a popular breakdown of the variance, because it is relatively easy to calculate and has a reasonable interpretation: total variance is variance explained by the regression plus the unexplained, error variance. As such, these elements will appear on page 311 with regard to the F test, and are used for the common *coefficient of determination*, which is an indicator of how well a regression fits the data. It is defined as:

$$\begin{aligned} R^2 &\equiv \frac{SSR}{SST} \\ &= 1 - \frac{SSE}{SST}. \end{aligned}$$

You will notice that the terminology about the sum of squared components and the use of the F test matches that used in the ANOVA breakdowns, which is not just a coincidence: in both cases, there is a portion of the data's variation explained by the model (grouping or regression), and a portion that is unexplained by the model (residual). In both cases, we can use this breakdown to gauge whether the model explains more variation than would be explained by a random grouping. The exact details of the F test will be delayed until the chapter on hypothesis testing.

COVARIANCE The population *covariance* is $\sigma^2_{\mathbf{xy}} = \frac{1}{n}\sum_i (x_i - \overline{\mathbf{x}})(y_i - \overline{\mathbf{y}})$, which is equivalent to $E[\mathbf{xy}] - E[\mathbf{x}]E[\mathbf{y}]$. [ℚ: Re-apply the first variance expansion above to prove this.] The variance is a special case where $\mathbf{x} = \mathbf{y}$.

As with the variance, the unbiased estimate of the sample covariance is $s^2_{xy} = \sigma^2_{xy} \cdot \frac{n}{n-1}$, i.e., $\sum_i (x - \overline{\mathbf{x}})(y - \overline{\mathbf{y}})$ divided by $n - 1$ instead of n.

Given a vector of variables $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n$, we typically want the covariance of every combination. This can neatly be expressed as a matrix

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 & \dots & \sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_2^2 & \dots & \sigma_{2n}^2 \\ \vdots & & \ddots & \\ \sigma_{n1}^2 & \sigma_{n2}^2 & \dots & \sigma_n^2 \end{bmatrix},$$

where the diagonal elements are the variances (i.e., the covariance of $\mathbf{x}_i$ with itself for all i), and the off-diagonal terms are symmetric in the sense that $\sigma_{ij}^2 = \sigma_{ji}^2$ for all i and j.

Correlation and Cauchy–Schwarz: The *correlation coefficient* (sometimes called the *Pearson correlation coefficient*) is

$$\rho_{\mathbf{xy}} \equiv \frac{\sigma_{\mathbf{xy}}}{\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}}.$$

By itself, the statistic is useful for looking at the relations among columns of data, and can be summarized into a matrix like the covariance matrix above. The correlation matrix is also symmetric, and has ones all along the diagonal, since any variable is perfectly correlated with itself.

The *Cauchy–Schwarz inequality*, $0 \leq \rho^2 \leq 1$, puts bounds on the correlation coefficient. That is, ρ is in the range $[-1, 1]$, where $\rho = -1$ indicates that one variable always moves in the opposite direction of the other, and $\rho = 1$ indicates that they move in perfect sync.[5]

The matrix of correlations is another popular favorite for getting basic descriptive information about a data set; produce it via `apop_data_correlation`. The correlation matrix will be the basis of the Cramér–Rao lower bound on page 333.

MORE MOMENTS Given a continuous probability distribution from which the data was taken, you could write out the expectation in the variance equation as an integral,

$$\begin{aligned} \text{var}(f(x)) &= E\left(\left(f(x) - \overline{f(x)}\right)^2\right) \\ &= \int_{\forall x} \left(f(x) - \overline{f(x)}\right)^2 p(x)dx. \end{aligned}$$

[5]It would be a digression to prove the Cauchy–Schwarz inequality here, but see Hölder's inequality in any probability text, such as Feller (1966, volume II, p 155).

Similarly for higher powers as well:

$$\text{skew}\left(f(x)\right) \equiv \int_{\forall x} \left(f(x) - \overline{f(x)}\right)^3 p(x)dx$$

$$\text{kurtosis}\left(f(x)\right) \equiv \int_{\forall x} \left(f(x) - \overline{f(x)}\right)^4 p(x)dx.$$

These three integrals are *central moments* of $f(x)$. They are central because we subtracted the mean from the function before taking the second, third, or fourth power.[6]

Transformed moments

Let $\mathcal{S}$ and $\mathcal{K}$ be the third and fourth central moments as given here. Some use a *standardized moment* for kurtosis, which may equal $\mathcal{K}_1' = \mathcal{K}/(\sigma^2)^2$, $\mathcal{K}_2' = \mathcal{K}/(\sigma^2)^2 - 3$, or whatever else the author felt would be convenient. Similarly, some call $\mathcal{S}' = \mathcal{S}/(\sigma^2)^{3/2}$ the skew. These adjustments are intended to ease comparisons to the standard Normal and to acommodate differences in scale.

The only way to know what a given source means when it refers to skew and kurtosis is to look up the definitions. The GSL uses $\mathcal{K}_2'$ (because engineers are probably comparing their data to a standard Normal); Apophenia uses $\mathcal{K}$ (because the corrections can add complication in situations outside the Normal distribution, and is easy to make when needed).

What information can we get from the higher moments? Section 9.1 will discuss the powerful Central Limit Theorem, which says that if a variable represents the mean of a set of independent and identical draws, then it will have an $\mathcal{N}(\mu, \sigma)$ distribution, where μ and σ are unknowns that can be estimated from the data. These two parameters completely define the distribution: the skew of a Normal is always zero, and the kurtosis is always $3\sigma^4$. If the kurtosis is larger, then this often means that the assumption of independent draws is false—the observations are interconnected. One often sees this among social networks, stock markets, or other systems where independent agents observe and imitate each other.

Positive skew indicates that a distribution is upward leaning, and a negative skew indicates a downward lean. Kurtosis is typically put in plain English as *fat tails*: how much density is in the tails of the distribution? For example, the kurtosis of an $\mathcal{N}(0, 1)$ is three, while the kurtosis of a Student's t distribution with n degrees of freedom is greater than three, and decreases as $n \to \infty$, converging to three (see page 365 for a full analysis). An un-normalized kurtosis $> 3\sigma^4$ is known as *leptokurtic* and $< 3\sigma^4$ is known as *platykurtic*; see Figure 7.3 for a mnemonic.

The caveats about unbiased estimates of the sample versus population variance (see box, page 222) also hold for skew and kurtosis: calculating the mean as done in the definitions above leads to a biased estimate of the population skew or kurtosis, but there are simple corrections that can produce an unbiased estimate. An online

[6]The central first moment is always zero; the non-central second, third, ..., moments are difficult to interpret and basically ignored. Since there is no ambiguity, some authors refer to the useful moments as *the nth moment*, $n \in \{1, 2, 3, 4\}$, and leave it as understood when the moment is central or non-central.

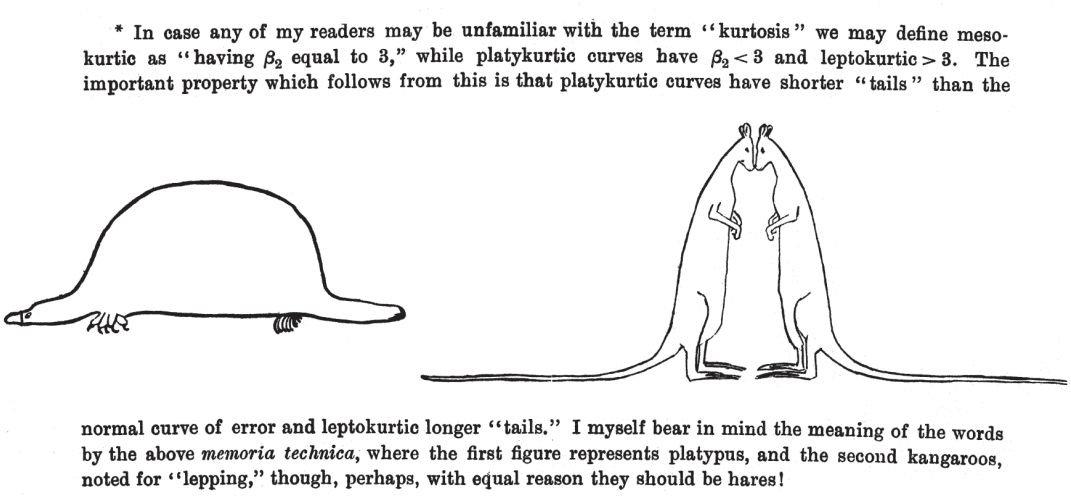


Figure 7.3 Leptokurtic, mesokurtic and platykurtic, illustration by Gosset in *Biometrika* (Student, 1927, p 160). In the notation of the time, $\beta_2 \equiv$ kurtosis/(variance squared).

appendix to this book offers a few more facts about central moments, and derives the correction factors.

But in all cases, the population vs sample detail is relevant only for small n. Efron & Tibshirani (1993, p 43) state that estimating variance via n is "just as good" as estimating it via $n-1$, so there is highly-esteemed precedent for ignoring this detail. For the higher moments, the sample and population estimates converge even more quickly.

Coding it Given a vector, Apophenia provides functions to calculate most of the above, e.g.:

```
apop_data *set = gather_your_data_here();
apop_data *corr = apop_data_correlation(set);
APOP_COL(set, 0, v1);
APOP_COL(set, 1, v2);
double mean1 = apop_vector_mean(v1);
double var1 = apop_vector_var(v1);
double skew1 = apop_vector_skew(v1);
double kurt1 = apop_vector_kurtosis(v1);
double cov = apop_vector_cov(v1, v2);
double cor = apop_vector_correlation(v1, v2);
apop_data_show(apop_data_summarize(set));
```

The last item in the code, `apop_matrix_summarize`, produces a table of some summary statistics for every column of the data set.

Your data may be aggregated so that one line of data represents multiple observations. For example, sampling efficiency can be improved by sampling subpopulations differently depending upon their expected variance (Särndal *et al.*, 1992). For this and other reasons, data from statistical agencies often includes weightings.

This is not the place to go into details about statistically sound means of weighting data, but if you have a separate vector with weights, you can use `apop_vector_weighted_mean`, `apop_vector_weighted_var`, et cetera, to use those weights. Or, if your `apop_data` set's `weights` vector is filled, `apop_data_summarize` will make use of it.

$Q_{7.4}$

- Write a query that pulls the number of males per 100 females and the population density from the Census data (`data-census.db`). The query will
  - join together the `geography` and `demos` tables by county number, and
  - exclude states and the national total, and
  - return a two-column table.
- Write a function `void summarize_paired_data(char *q)` that takes in a query that produces a two-column table, and outputs some of the above summary information about both columns, including the mean, variance, and correlation coefficients.
- Write a `main()` that sends your query to the above function, and run the program. Is population density positively or negatively correlated to males per female?
- Write another query that pulls the ratio of (median income for full-time workers, female)/(median income for full-time workers, male) and the population density.
- Add a line to `main()` to send that query to your summarizing function as well. How is the new pair of variables correlated?

Quantiles The mean and variance can be misleading for skewed data. The first option for describing a data set whose distribution is likely ab-Normal is to plot a histogram of the data, as per page 172.

A numeric option is to print the quartiles, quintiles, or deciles of the data. For quartiles, sort the data, and then display the values of the data points 0%, 25%, 50%, 75%, and 100% of the way through the set. The 0% value is the minimum of

the data set, the 100% value is the maximum, and the 50% value is probably the median (see below). For quintiles, print the values of data points 0%, 20%, 40%, 60%, 80%, and 100% of the way through the set, and for deciles, print the values every ten percent.

Sorting your data is simple. If you have an `apop_data` set and a `gsl_vector`, then

```
apop_data_sort(my_data, 2, 'd');
gsl_vector_sort(my_vector);
```

would sort `my_data` in place so that column 2 is in 'd'escending order, and sort the vector in place to ascending order, so `gsl_vector_get(my_vector, 0)` is the minimum of the data, `gsl_vector_get(my_vector, my_vector->size)` is the maximum, and `gsl_vector_get(my_vector, my_vector->size/2)` is about the median.

Alternatively, the function `apop_vector_percentiles` takes in a `gsl_vector` and returns the percentiles—the value of the data point 0%, 1%, ..., 99%, 100% of the way through the data. It takes in two arguments: the data vector, and a character describing the rounding method—`'u'` for rounding up, `'d'` for rounding down, and `'a'` for averaging. Since the number of elements in a data set is rarely divisible by a hundred and one, the position of most percentile points likely falls between two data points. For example, if the data set has 107 points, then the tenth data point is 9.47% through the data set, and the eleventh data point is 10.38% through the set, so which is the tenth percentile? If you specify `'u'`, then it is the eleventh data point; if you specify `'d'` then it is the tenth data point, and if you specify `'a'`, then it is the simple average of the two.

The standard definition of the median is that it is the middle value of the data point if the data set has an odd number of elements, and it is the average of the two data points closest to the middle if the data set has an even number of elements. Thus, here is a function to find the median of a data set. It finds the percentiles using the averaging rule for interpolation, marks down the 50th percentile, then cleans up and returns that value.

```
double find_median(gsl_vector *v){
    double *pctiles = apop_vector_percentiles(v, 'a');
    double out = pctiles[50];
    free(pctiles);
    return out;
}
```

$\mathbb{Q}_{7.5}$

Write a function with header `double show_quantiles(gsl_vector *v, char rounding_method, int divisions)` that passes `v` and `rounding_method` to `apop_vector_percentiles`, and then prints a table of selected quantiles to the screen. For example, if `divisions==4`, print quartiles, if `divisions==5`, print quintiles, if `divisions==10`, print deciles, et cetera.

On page 88 you tabulated GDP per capita for the countries of the world. Use your function to print the deciles for country incomes from `data-wb`.

$\mathbb{Q}_{7.6}$

The *trimean* is $\frac{1}{4}$ the sum of the first quartile, third quartile, and two times the median (Tukey, 1977, p 46). It uses more information about the distribution than the median alone, but is still robust to extreme values (unlike the mean).

Write a function that takes in a vector of data points, applies `apop_vector_percentiles` internally, and returns the trimean. How does the trimean of GDP per capita compare to the mean and median, and why?

See also page 319, which compares percentiles of the data to percentiles of an assumed distribution to test whether the data were drawn from the distribution.

$\sum$

- The most basic means of describing data is via its moments. The basic moments should be produced and skimmed for any data set; in simple cases, there is no need to go further.
- The variance can often be decomposed into smaller parts, thus revealing more information about how a data set's variation arose.
- The mean and variance are well known, but there is also information in higher moments—the skew and kurtosis.
- It is also important to know how variables interrelate, which can be summarized using the correlation matrix.
- There is a one-line function to produce each of these pieces of information. Notably, `apop_data_summarize` produces a summary of each column of a data set.
- You can get a more detailed numerical description of a data set's distribution using quartiles or quintiles; to do so, use `apop_vector_percentiles`.

7.2 SAMPLE DISTRIBUTIONS

Here are some distributions that an observed variable may take on. They are not just here so you can memorize them before the next statistics test. Each has a story attached, which is directly useful for modeling. For example, if you think that a variable is the outcome of n independent, binary events, then the variable should be modeled as a Binomial distribution, and once you estimate the parameter to the distribution, you will have a full model of that variable, that you can even test if so inclined. Table 7.4 presents a table of what story each distribution is telling. After the catalog of models, I will give a few examples of such modeling.

The distribution	The story
Bernoulli	A single success/failure draw, fixed p.
Binomial	What are the odds of getting x successes from n Bernoulli draws with fixed p?
Hypergeometric	What are the odds of getting x successes from n Bernoulli draws, where p is initially fixed, but drawing is without replacement?
Normal/Gaussian	Binomial as $n \to \infty$; if $\mu_j \equiv \sum_{i=1}^{n} x_{ij}/n$, then $\mu_j \sim$ Normal.
Lognormal	If $\mu_j \equiv \prod_{i=1}^{n} x_{ij}$, then $\mu_j \sim$ Lognormal.
Multinomial	n draws from m possibilities with probabilities p_1, $\ldots$, p_m, $\sum_{i=1}^{m} p_i = 1$.
Multivariate Normal	Multinomial as $n \to \infty$.
Negative binomial	How many Bernoulli draws until n successes?
Poisson	Given λ events per period, how many events in t periods?
Gamma	The 'Negative Poisson': given a Poisson setup, how long until n events?
Exponential	A proportion λ of the remaining stock leaves each period; how much is left at time t?
Beta	A versatile way to describe the odds that p takes on any value $\in (0, 1)$.
Uniform	No information but the upper and lower bounds.

Table 7.4 Every probability distribution tells a story.

Common distributions of statistical parameters (as opposed to natural populations) are discussed in Section 9.2.

Bernoulli and Poisson events The core of the system is an event, which sometimes happens and sometimes does not. Some people have a disease, some do not; some days it rains, some days it does not. Events add up to more-or-less continuous quantities: some cities see a 22% chance of rain on a given day, and some see a 23.2% chance; some populations have high rates of disease and some do not.

From there, there are variants: instead of asking how many successes we will see in n trials (the Binomial and Poisson distributions), we can ask how many trials we can expect to make to get n successes (the Negative binomial and Gamma distributions). We can look at sampling without replacement (Hypergeometric). We can look at the case where the number of trials goes to infinity (Normal) or aggregate trials and take their product (Lognormal). In short, a surprisingly wide range of situations can be described from the simple concept of binary events aggregated in various ways.

The snowflake problem For a controlled study (typical of physical sciences), the claim of a fixed probability (p or λ) is often plausible, but for most social science experiments, where each observation is an individual or a very different geographic region, the assumption that all observations are identical is often unacceptable—every snowflake is unique.

It may seem like the fixed-p and fixed-λ models below are too simple to be applicable to many situations—and frankly, they often are. However, they can be used as building blocks to produce more descriptive models. Section 8.2.1 presents a set of linear regression models, which handle the snowflake problem well, but throw out the probability framework presented in this chapter. But we can solve the snowflake problem and still make use of simple probability models by embedding a linear model inside a distribution: let p_i differ among each element i, and estimate p_i using a linear combination of element i's characteristics. Page 288 covers examples of the many possibilities provided by models-within-models.

Statistics and their estimators

- The catalog in this section includes the three most typical items one would want from a distribution: a random number generator (*RNG*) that would produce data with the given distribution, a *probability density function* (*PDF*), and a *cumulative density function* (*CDF*).[7]

[7]This is for the case when the independent variable is continuous. When it is discrete, the PDF is named a *probability mass function* (*PMF*) and the CDF a *cumulative mass function* (*CMF*). For every result or statement about PDFs, there is an analogous result about PMFs; for every result or statement about CDFs, there is an analogous result about CMFs.

- The catalog also includes the expected value and variance of these distributions, which are distinct from the means and variances to this point in a key manner: given a data set, the mean is a statistic—a function of data of the form $f(x)$. Meanwhile, given a model, the mean and variance of a draw are functions of the parameters, e.g., given a Binomial distribution with parameters n and p, $E(x|n,p) = np$. Of course, we rarely know all the parameters, so we are left with estimating them from data, but our estimate of p, $\hat{p}$ is once again a function of the data set on hand. We will return to this back-and-forth between estimates from data and estimates from the model after the catalog of models.
- The full details of RNG use will be discussed in Chapter 11, but the RNG functions are included in the catalog here for easy reference; each requires a pointer to a `gsl_rng`, which will be named `r`.

THE BERNOULLI FAMILY The first set of distributions are built around a narrative of drawing from a pool of events with a fixed probability. The most commonly-used example is flipping a coin, which is a single event that comes out *heads* with probability $p = 1/2$. But other examples abound: draw one recipient from a sales pitch out of a list of such people and check whether he purchased or did not purchase, or pull a single citizen from a population and see if she was the victim of a crime. For one event, a draw can take values of only zero or one; this is known as a *Bernoulli draw*.

Bernoulli The Bernoulli distribution represents the result of one Bernoulli draw, meaning that $P(x = 1|p) = p$, $P(x = 0|p) = 1-p$, and $P(x = \text{anything else}\,|p) = 0$. Notice that $E(x|p) = p$, even though x can be only zero or one.

$$\begin{aligned} P(x,p) &= p^x(1-p)^{(1-x)}, x \in \{0,1\} \\ &= \texttt{gsl\_ran\_bernoulli\_pdf(x,p)} \\ E(x|p) &= p \\ \text{var}(x|p) &= p(1-p) \\ RNG &: \texttt{gsl\_ran\_bernoulli(r,p)} \end{aligned}$$

Binomial Now take n Bernoulli draws, so we can observe between 0 and n events. The output is now a count: how many people dunned by a telemarketer agree to purchase the product, or how many crime victims there are among a fixed population. The probability of observing exactly k events is $p(k) \sim \text{Binomial}(n,p)$.

Counting x successes out of n trials is less than trivial. For example, there are six ways by which two successes could occur over four trials: $(0,0,1,1)$, $(0,1,0,1)$,

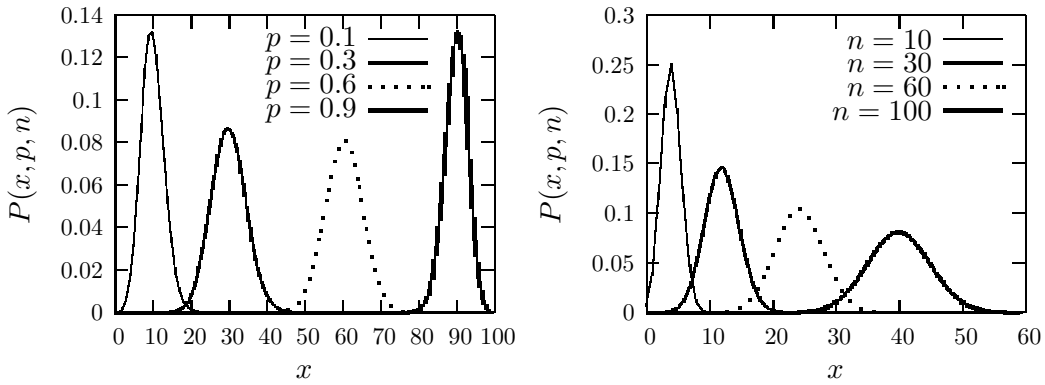


Figure 7.5 Left: The Binomial distribution with $n = 100$ and various values of p. Right: the Binomial distribution with $p = 0.4$ and various values of n. As n grows, the distribution approaches an $\mathcal{N}(np, \sqrt{np(1-p)})$.

$(0, 1, 1, 0)$, $(1, 1, 0, 0)$, $(1, 0, 0, 1)$, or $(1, 0, 1, 0)$, and the model underlying the Binomial model treats them all equally.

The form of the distribution therefore borrows a counting technique from combinatorics. The notation $\binom{n}{x}$ indicates n *choose* x, the number of unordered sets of x elements that can be pulled from n objects. The equation is

$$\binom{n}{x} = \frac{n!}{x!(n-x)!},$$

and the function is `gsl_sf_choose(n,x)` (in the GSL's Special Functions section). For example, we could get exactly thirty successful trials out of a hundred in $\binom{100}{30}$ ways ($\approx$ 2.94e25).

Combinatorics also dictates the shape of the curve. There is only one way each to list four zeros or four ones—$(0, 0, 0, 0)$ and $(1, 1, 1, 1)$—and there are four ways to list one one—$(0, 0, 0, 1)$, $(0, 0, 1, 0)$, ... —and symmetrically for one zero. In order, the counts for zero through four ones are 1, 4, 6, 4, and 1. This simple counting scheme already produces something of a bell curve. Returning to coin-flipping, if $p = 1/2$ and the coin is flipped 100 times ($n = 100$), $p(50 \text{ heads})$ is relatively high, while $p(0 \text{ heads})$ or $p(100 \text{ heads})$ is almost nil.

In fact, as $n \to \infty$, the probability distribution approaches the familiar Normal distribution with mean np and variance $np(1-p)$, as in Figure 7.5. Assuming that every telemarketer's probability of selling is equal, we expect that a plot of many months' telemarketer results will look like a Normal distribution, with many telemarketers successfully selling to np victims, and others doing exceptionally well or poorly. The assumption that every telemarketer is equally effective can even be tested, by checking for digression from the Normal distribution.

$$
\begin{aligned}
P(x,p,n) &= \binom{n}{x} p^x (1-p)^{(n-x)} \\
&= \texttt{gsl\_ran\_binomial\_pdf}(\texttt{x},\texttt{p},\texttt{n}) \\
E(x|n,p) &= np \qquad (7.2.1) \\
\mathrm{var}(x|n,p) &= np(1-p) \qquad (7.2.2) \\
RNG &: \texttt{gsl\_ran\_binomial}(\texttt{r},\texttt{p},\texttt{n})
\end{aligned}
$$

- If $X \sim$ Bernoulli(p), then for the sum of n independent draws, $\sum_{i=1}^{n} X_i \sim$ Binomial(n,p).
- As $n \to \infty$, Binomial$(n,p) \to$ Poisson(np) or $\mathcal{N}(np, \sqrt{np(1-p)})$.

Since n is known and $E(x)$ and $\mathrm{var}(x)$ can be calculated from the data, Equations 7.2.1 and 7.2.2 are an oversolved system of two variables for one unknown, p. Thus, you can test for *excess variance*, which indicates that there are interactions that falsify that the observations were *iid* (independent and identically distributed) Bernoulli events.

A variant: one p from n draws The statistic of interest often differs from that calculated in this catalog, but it is easy to transform the information here. Say that we multiply the elements of a set x by k. Then the mean goes from being $\mu_x \equiv \sum_i x_i/n$ to being $\mu_{kx} \equiv \sum_i (kx_i)/n = k\sum_i x_i/n = k\mu_x$. The variance goes from being $\sigma_x^2 \equiv \sum_i (x_i - \mu_x)^2/n$ to $\sigma_{kx}^2 \equiv \sum_i (kx_i - k\mu_x)^2/n = k^2 \sum_i (x_i - \mu_x)^2/n = k^2\sigma_k^2$.

For example, we are often interested in estimating $\hat{p}$ from data with a Binomial-type story. Since $E(x) = np$ under a Binomial model, one could estimate $\hat{p}$ as $E(x/n)$. As for the variance, let k in the last paragraph be $1/n$; then the variance of x/n is the original variance ($\mathrm{var}(x) = n\hat{p}(1-\hat{p})$) times $1/n^2$, which gives $\mathrm{var}(\hat{p}) = \hat{p}(1-\hat{p})/n$.

Hypergeometric Say that we have a pool of N elements, initially consisting of s successes and $f \equiv N - s$ failures. So $N = s + f$, and the Bernoulli probability for the entire pool is $p = s/N$. What are the odds that we get x successes from n draws *without replacement*? In this case, the probability of a success changes as the draws are made. The counting is more difficult, resulting in a somewhat more involved equation.

$$
\begin{aligned}
P(x,s,f,n) &= \frac{\binom{s}{x}\binom{f}{n-x}}{\binom{N}{n}} \\
&= \texttt{gsl\_ran\_hypergeometric\_pdf}(\texttt{x},\texttt{s},\texttt{f},\texttt{n}) \\
E(x|n,s,f) &= \frac{ns}{N} \\
\mathrm{var}(x|n,s,f) &= \frac{n(\frac{s}{N})(1-\frac{s}{N})(N-n)}{(N-1)} \\
RNG &: \texttt{gsl\_ran\_hypergeometric}(\texttt{r},\texttt{s},\texttt{f},\texttt{n})
\end{aligned}
$$

- As $N \to \infty$, drawing with and without replacement become equivalent, so the Hypergeometric distribution approaches the Binomial.

Multinomial The Binomial distribution was based on having a series of events that could take on only two states: success/failure, sick/well, heads/tails, et cetera. But what if there are several possible events, like left/right/center, or Africa/Eurasia/Australia/Americas? The Multinomial distribution extends the Binomial distribution for such cases.

The Binomial case could be expressed with one parameter, p, which indicated success with probability p and failure with probability $1-p$. The Multinomial case requires k variables, $p_1, \ldots, p_k$, such that $\sum_{i=1}^{k} p_i = 1$.

$$
\begin{aligned}
P(\mathbf{x},\mathbf{p},n) &= \frac{n!}{x_1!\cdots x_k!} p_1^{x_1}\cdots p_k^{x_k} \\
&= \texttt{gsl\_ran\_multinomial\_pdf}(\texttt{k},\texttt{p},\texttt{n}) \\
E\left(\left[\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \end{array}\right] \middle| n,\mathbf{p}\right) &= n\left[\begin{array}{c} p_1 \\ p_2 \\ \vdots \\ p_n \end{array}\right] \\
\mathrm{var}(\mathbf{x}|n,\mathbf{p}) &= n\left[\begin{array}{cccc} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_k \\ -p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_k \\ \vdots & & \ddots & \vdots \\ -p_kp_1 & -p_kp_2 & \cdots & p_k(1-p_k) \end{array}\right] \\
RNG &: \texttt{gsl\_ran\_multinomial}(\texttt{r},\texttt{draws},\texttt{k},\texttt{p},\texttt{out})
\end{aligned}
$$

- There are two changes from the norm for the GSL's functions. First, `p` and `n` are arrays of `doubles` of size `k`. If $\sum_{\texttt{i}=1}^{\texttt{k}} \texttt{p[i]} \neq 1$, then the system normalizes the

probabilities to make this the case. Also, most RNGs draw one number at a time, but this one draws K elements at a time, which will be put into the bins of the `out` array.

- You can verify that when $k = 2$, this is the Binomial distribution.

Normal You know and love the bell curve, aka the Gaussian distribution. It is pictured for a few values of σ^2 in Figure 7.6.

As Section 9.1 will explain in detail, any set of means generated via iid draws will have a Normal distribution. That is,

- Draw K items from the population (which can have any nondegenerate distribution), $x_1, x_2, \ldots, x_k$. The Normal approximation works best when K is large.
- Write down the mean of those items, $\bar{x}_i$.
- Repeat the first two steps n times, producing a set $\mathbf{x} = \{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n\}$.

Then $\mathbf{x}$ has a Normal distribution.

Alternatively, the Binomial distribution already produced something of a bell curve with $n = 4$ above; as $n \to \infty$, the Binomial distribution approaches a Normal distribution.

$$
\begin{aligned}
P(x, \mu, \sigma) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{(x-\mu)}{\sigma}\right]^2\right) \\
&= \texttt{gsl\_ran\_gaussian\_pdf}(\texttt{x}, \texttt{sigma}) + \texttt{mu} \\
E(x|\mu, \sigma) &= \mu \\
\mathrm{var}(x|\mu, \sigma) &= \sigma^2 \\
\int_{-\infty}^{x} \mathcal{N}(y|\mu, \sigma)dy &= \texttt{gsl\_cdf\_gaussian\_P}(\texttt{x} - \texttt{mu}, \texttt{sigma}) \\
\int_{x}^{\infty} \mathcal{N}(y|\mu, \sigma)dy &= \texttt{gsl\_cdf\_gaussian\_Q}(\texttt{x} - \texttt{mu}, \texttt{sigma}) \\
RNG &: \texttt{gsl\_ran\_gaussian}(\texttt{r}, \texttt{sigma}) + \texttt{mu}
\end{aligned}
$$

- If $X \sim \mathcal{N}(\mu_1, \sigma_1)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2)$ then $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.
- Because the Normal is symmetric, $X - Y \sim \mathcal{N}(\mu_1 - \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.
- Section 9.1 (p 297) discusses the Central Limit Theorem in greater detail.

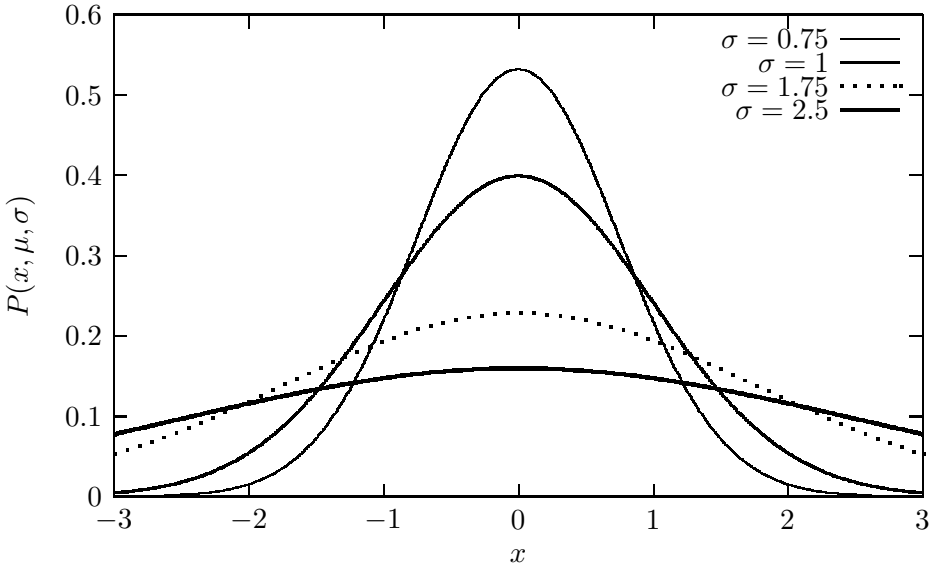


Figure 7.6 The Normal distribution, with $\mu = 0$.

Multivariate Normal Just as the Normal distribution is the extension of the Binomial, the Multivariate Normal is the extension of the Multinomial. Say that we have a data set $\mathbf{X}$ that includes a thousand observations and seven variables (so $\mathbf{X}$ is a 1000×7 matrix). Let its mean be μ (a vector of length seven) and the covariance among the variables be $\mathbf{\Sigma}$ (a seven by seven matrix). Then the Multivariate Normal distribution that you could fit to this data is

$$P(\mathbf{X}, \boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^n \det(\mathbf{\Sigma})}}$$
$$E(\mathbf{X}|\boldsymbol{\mu}, \mathbf{\Sigma}) = \boldsymbol{\mu}$$
$$\mathrm{var}(\mathbf{X}|\boldsymbol{\mu}, \mathbf{\Sigma}) = \mathbf{\Sigma}$$

- When $\mathbf{X}$ has only one column and $\mathbf{\Sigma} = [\sigma^2]$, this reduces to the univariate Normal distribution.

Lognormal The Normal distribution is apropos when the items in a sample are the mean of a set of draws from a population, $\overline{x}_i = (s_1 + s_2 + \cdots + s_k)/k$. But what if a data point is the *product* of a series of iid samples, $\tilde{x}_i = s_1 \cdot s_2 \cdot \cdots \cdot s_k$? Then the log of $\tilde{x}_i$ is $\ln(\tilde{x}_i) = \ln(s_1) + \ln(s_2) + \cdots + \ln(s_k)$, so the log is a sum

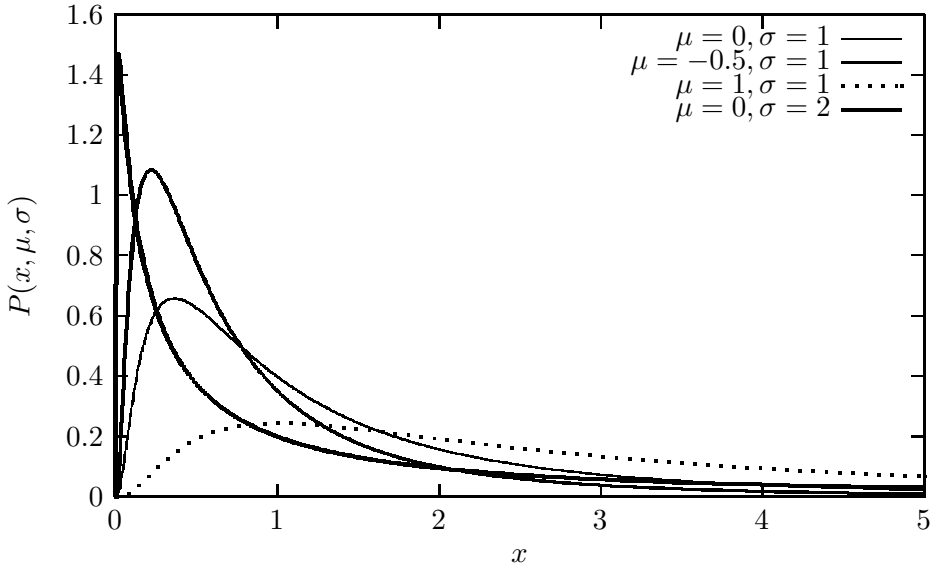


Figure 7.7 The Lognormal distribution.

of independent elements (i.e., n times a mean). Very broadly, when a point in the data set is produced by summing iid draws, it will be Normally distributed; when a point in the data set is produced by taking the product of iid draws, its log will be Normally distributed—i.e., it will have a lognormal distribution. The next section will present an example. Figure 7.7 shows some Lognormal distributions.

A notational warning: in the typical means of expressing the lognormal distribution, μ and σ refer to the mean of the Normal distribution that you would get if you replaced every element x in your data set with e^x, thus producing a standard Normal distribution. Be careful not to confuse this with the mean and variance of the data you actually have.

$$
\begin{aligned}
p(x,\mu,\sigma) &= \frac{\exp\left(-(\ln x - \mu)^2/(2\sigma^2)\right)}{x\sigma\sqrt{2\pi}} \\
&= \texttt{ran\_lognormal\_pdf}(\texttt{x}, \texttt{mu}, \texttt{sigma}) \\
E(x|\mu,\sigma) &= e^{\left(\mu + \frac{\sigma^2}{2}\right)} \\
\mathrm{var}(x|\mu,\sigma) &= (e^{\sigma^2} - 1)e^{(2\mu+\sigma^2)} \\
RNG &: \texttt{ran\_lognormal}(\texttt{rng}, \texttt{mu}, \texttt{sigma})
\end{aligned}
$$

Negative binomial Say that we have a sequence of Bernoulli draws. How many failures will we see before we see n successes? If p percent of cars are illegally parked, and a meter reader hopes to write n parking tickets, the Negative binomial tells her the odds that she will be able to stop with $n + x$ cars.

The form is based on the Gamma function,

$$\begin{aligned}\Gamma(z) &= \int_0^\infty x^{z-1}e^{-x}dx \\ &= \texttt{gsl\_sf\_gamma(z)}.\end{aligned}$$

You can easily verify that $\Gamma(z+1) = z\Gamma(z)$. Also, $\Gamma(1) = 1, \Gamma(2) = 1, \Gamma(3) = 2$, $\Gamma(4) = 6$, and generally, $\Gamma(z) = (z-1)!$ for positive integers. Thus, if n and x are integers, formulas based on the Gamma function reduce to more familiar factorial-based counting formulas.

$$\begin{aligned}P(x,n,p) &= \frac{\Gamma(n+x)}{\Gamma(x+1)\Gamma(n)}p^n(1-p)^x \\ &= \texttt{gsl\_ran\_negative\_binomial\_pdf(x,p,n)} \\ E(x|n,p) &= \frac{n(1-p)}{p} \\ \mathrm{var}(x|n,p) &= \frac{n(1-p)}{p^2} \\ RNG &: \texttt{gsl\_ran\_negative\_binomial(rng,p,n)}\end{aligned}$$

RATES A Poisson process is very much like a Bernoulli draw, but the unit of measurement is continuous—typically a measure of time or space. It makes sense to have half of an hour, but not half of a coin flip, so the stories above based on Bernoulli draws are modified slightly to allow for a rate of λ events per hour to be applied to half an hour or a week.

Baltimore, Maryland, sees about 110 days of precipitation per year, somewhat consistently spaced among the months. But for how many days will it rain or snow in a single week? The Poisson distribution answers this question. We can also do a count of weeks: how often does it rain once in a week, twice in a week, et cetera? The Exponential distribution answers this question. Turning it around, if we want a week with three rainfalls, how long would we have to wait? The Gamma distribution answers this question.

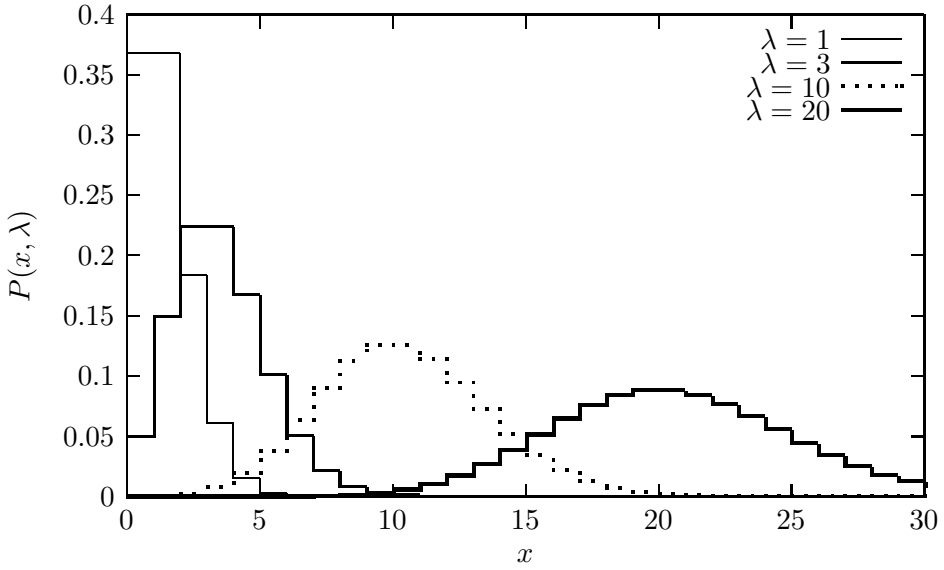


Figure 7.8 The Poisson distribution.

Poisson Say that independent events (rainy day, landmine, bad data) occur at the mean rate of λ events per span (of time, space, et cetera). What is the probability that there will be x events in a single span?

We are assuming that events occur at a sufficiently even rate that the same rate applies to different time periods: if the rate per day is λ_1, then the rate per week is $7\lambda_1$, and the rate per hour is $\lambda_1/24$. See Figure 7.8.

$$
\begin{aligned}
P(x,\lambda) &= \frac{e^{-\lambda}\lambda^x}{x!} \\
&= \texttt{gsl\_ran\_poisson\_pdf}(\texttt{x}, \texttt{lambda}) \\
E(x|\lambda) &= \lambda \\
\mathrm{var}(x|\lambda) &= \lambda \\
RNG &: \texttt{gsl\_ran\_poisson}(\texttt{r}, \texttt{lambda})
\end{aligned}
$$

- As $n \to \infty$, Binomial(n,p) →Poisson(np).
- If $X \sim \mathrm{Poisson}(\lambda_1)$, $Y \sim \mathrm{Poisson}(\lambda_2)$, and X and Y are independent, then $(X+Y) \sim \mathrm{Poisson}(\lambda_1+\lambda_2)$.
- As $\lambda \to \infty$, Poisson$(\lambda) \to \mathcal{N}(\lambda, \sqrt{\lambda})$.

$\mathbb{Q}_{7.7}$

- Calculate the Binomial-distributed probability of three rainfalls in seven days, given the probability of rain in one day of $p = (110/365)$.
- Calculate the Poisson-distributed probability of three rainfalls in seven days, given a one-day $\lambda = (110/365)$.

Gamma distribution The Gamma *distribution* is so-named because it relies heavily on the Gamma *function*, first introduced on page 244. Along with the Beta distribution below, this naming scheme is one of the great notational tragedies of mathematics.

A better name in the statistical context would be 'Negative Poisson,' because it relates to the Poisson distribution in the same way the Negative binomial relates to the Binomial. If the timing of events follows a Poisson distribution, meaning that events come by at the rate of λ per period, then this distribution tells us how long we would have to wait until the nth event occurs.

The form of the Gamma distribution, shown for some parameter values in Figure 7.9, is typically expressed in terms of a shape parameter $\theta \equiv 1/\lambda$, where λ is the Poisson parameter. Here is the summary for the function in terms of both parameters:

$$
\begin{aligned}
P(x, n, \theta) &= \frac{1}{\Gamma(n)\theta^n} x^{n-1} e^{-x/\theta}, x \in [0, \infty) \\
&= \texttt{gsl\_ran\_gamma\_pdf}(\texttt{x}, \texttt{n}, \texttt{theta}) \\
P(x, n, \lambda) &= \frac{1}{\Gamma(n)(\frac{1}{\lambda})^n} x^{n-1} e^{-\lambda x}, x \in [0, \infty) \quad (7.2.3) \\
E(x|n, \theta \text{ or } \lambda) &= n\theta = n/\lambda \\
\text{var}(x|n, \theta \text{ or } \lambda) &= k\theta^2 = k/\lambda^2 \\
\int_{-\infty}^{x} \text{G}(y|n, \theta) dy &= \texttt{gsl\_cdf\_gamma\_P}(\texttt{x}, \texttt{theta}) \\
\int_{x}^{\infty} \text{G}(y|n, \theta) dy &= \texttt{gsl\_cdf\_gamma\_Q}(\texttt{n}, \texttt{theta}) \\
RNG &: \texttt{gsl\_ran\_gamma}(\texttt{r}, \texttt{n}, \texttt{theta})
\end{aligned}
$$

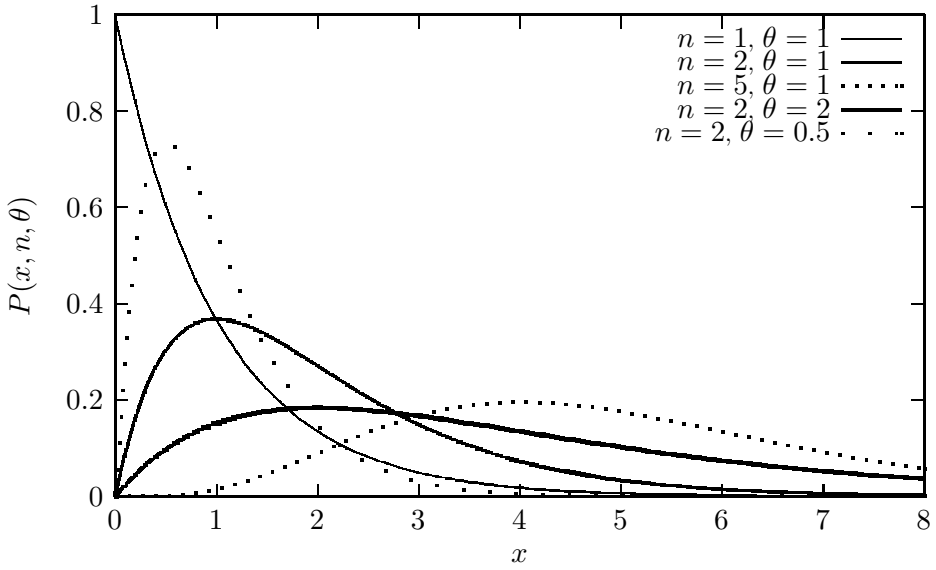


Figure 7.9 The Gamma distribution.

- With $n = df/2$ and $\theta = 2$, the Gamma distribution becomes a χ^2_{df} distribution (introduced on page 301).
- With $n = 1$, the Gamma distribution becomes an Exponential(λ) distribution.

Exponential distribution The Gamma distribution found the time until n events occur, but consider the time until the first event occurs. $\Gamma(1) \equiv 1$, $1^\lambda = 1$ for all positive λ, and $x^0 = 1$ for all positive x, so at $n = 1$, Equation 7.2.3 defining the PDF of the Gamma distribution reduces to simply $e^{-\lambda x}$.

If we had a population of items, $\int_0^t e^{-\lambda x} dx$ percent would have had a first event between time zero and time t. If the event causes the item to leave the population, then one minus this percent are still in the population at time t. The form e^x is very easy to integrate, and doing so gives that the percent left at time $t = e^{-\lambda t}/\lambda$.

So we now have a story of a population where members leave via a Poisson process. Common examples include the stock of unemployed workers as some find a job every period, radioactive particles emanating from a block, or drug dosage remaining in a person's blood stream. Figure 7.10 shows a few examples of the Exponential distribution.

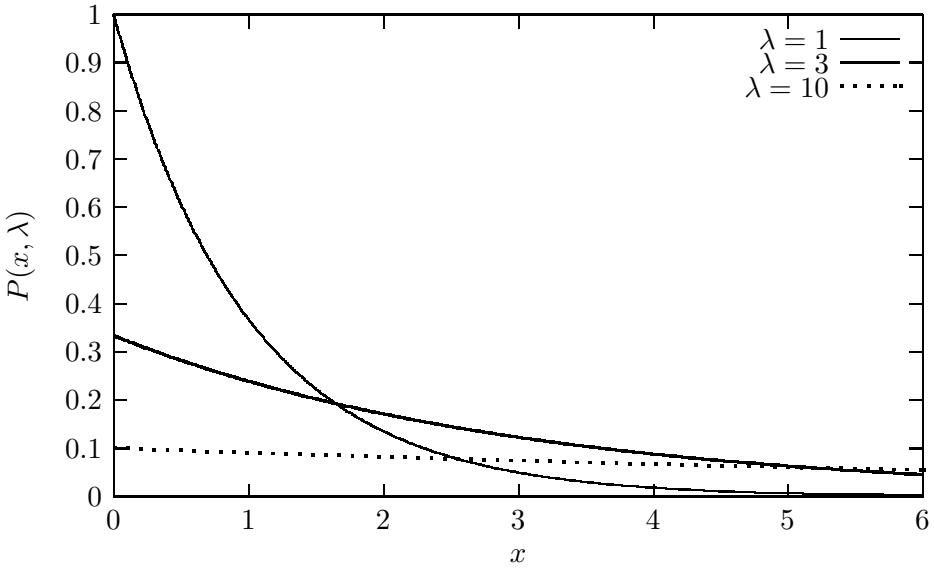


Figure 7.10 The Exponential distribution.

Since the exponent is $-\lambda$, this is sometimes called the *Negative exponential distribution*.

$$
\begin{aligned}
P(x,\lambda) &= \frac{1}{\lambda} e^{\frac{-x}{\lambda}} \\
&= \texttt{gsl\_ran\_exponential\_pdf}(\texttt{x}, \texttt{lambda}) \\
E(x|\lambda) &= \lambda \\
\mathrm{var}(x|\lambda) &= \lambda^2 \\
\int_{-\infty}^{x} \mathrm{Exp}(\lambda) dy &= \texttt{gsl\_cdf\_exponential\_P}(\texttt{x}, \texttt{lambda}) \\
\int_{x}^{\infty} \mathrm{Exp}(\lambda) dy &= \texttt{gsl\_cdf\_exponential\_Q}(\texttt{x}, \texttt{lambda}) \\
RNG &: \texttt{gsl\_ran\_exponential}(\texttt{r}, \texttt{lambda})
\end{aligned}
$$

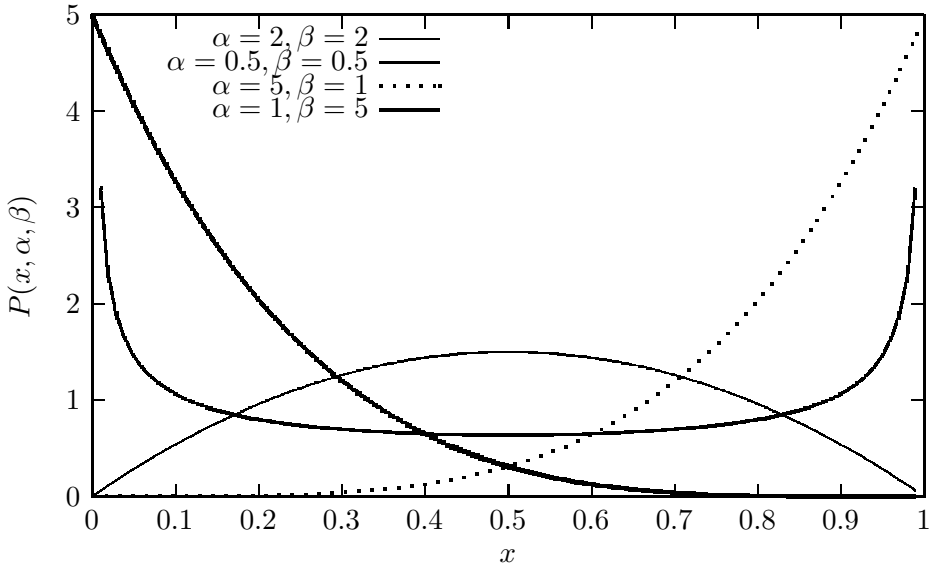


Figure 7.11 The Beta distribution.

DESCRIPTION Here are a few more distributions that are frequently used in modeling to describe the shape of a random variable.

Beta distribution Just as the Gamma distribution is named for the Gamma function, the Beta distribution is named after the Beta function—whose parameters are typically notated as α and β. This book will spell out $\text{Beta}(\cdot)$ for the Beta function and use $\mathcal{B}(\cdot, \cdot)$ for the Beta distribution.

The Beta function can be described via the following forms:

$$\begin{aligned}\text{Beta}(\alpha, \beta) &= \int_0^1 x^{(\alpha-1)}(1-x)^{(\beta-1)}dx \\ &= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \\ &= \texttt{gsl\_sf\_beta}(\texttt{alpha}, \texttt{beta}).\end{aligned}$$

The Beta distribution is a flexible way to describe data inside the range $[0, 1]$. Figure 7.11 shows how different parameterizations could lead to a left-leaning, right-leaning, concave, or convex curve; see page 358 for more.

$$
\begin{aligned}
P(x, \alpha, \beta) &= \text{Beta}(\alpha, \beta) x^{\alpha-1}(1-x)^{\beta-1} \\
&= \texttt{gsl\_ran\_beta\_pdf}(\texttt{x}, \texttt{alpha}, \texttt{beta}) \\
E(x|\alpha, \beta) &= \frac{\alpha}{\alpha+\beta} \\
\text{var}(x|\alpha, \beta) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\
\int_{-\infty}^{x} \mathcal{B}(y|\alpha, \beta)dy &= \texttt{gsl\_cdf\_beta\_P}(\texttt{x}, \texttt{alpha}, \texttt{beta}) \\
\int_{x}^{\infty} \mathcal{B}(y|\alpha, \beta)dy &= \texttt{gsl\_cdf\_beta\_Q}(\texttt{x}, \texttt{alpha}, \texttt{beta}) \\
RNG &: \texttt{gsl\_ran\_beta}(\texttt{r}, \texttt{alpha}, \texttt{beta})
\end{aligned}
$$

- If $\alpha < 1$ and $\beta < 1$, then the distribution is bimodal, with peaks at zero and one.
- If $\alpha > 1$ and $\beta > 1$, then the distribution is unimodal.
- As α rises, the distribution leans toward one; as β rises, the distribution leans toward zero; if $\alpha = \beta$, then the distribution is symmetric.
- If $\alpha = \beta = 1$, then this is the $\text{Uniform}[0, 1]$ distribution.

※ *The Beta distribution and order statistics* The first *order statistic* of a set of numbers $\mathbf{x}$ is the smallest number in the set; the second is the next-to-smallest, up to the largest order statistic, which is $\max(\mathbf{x})$.

Assume that the $\alpha + \beta - 1$ elements of $\mathbf{x}$ are drawn from a $\text{Uniform}[0, 1]$ distribution. Then the αth order statistic has a $\mathcal{B}(\alpha, \beta)$ distribution.

$\mathbb{Q}_{7.8}$

- Write a function that takes in a `gsl_rng` and two integers `a` and `b`, produces a list of `a+b-1` random numbers in $[0, 1]$, sorts them, and returns the $a\text{th}$ order statistic.
- Write a function to call that function 10,000 times and plot the PDF of the returned data (using `apop_plot_histogram`). It helps to precede the plot output to Gnuplot with `set xrange [0:1]` to keep the range consistent.
- Write a `main` that produces an animation of the PDFs of the first through 100th order statistic for a set of 100 numbers.
- Replace the call to the draw-and-sort function with a draw from the $\mathcal{B}(\texttt{a}, \texttt{b})$ distribution, and re-animate the results.

Uniform distribution What discussion of distributions would be complete without mention of the Uniform? It represents a belief that any value within $[\alpha, \beta]$ is equally possible.

$$P(x,\alpha,\beta) = \begin{cases} \frac{1}{\beta-\alpha} & x \in [\alpha,\beta] \\ 0 & x < \alpha, x > \beta \end{cases}$$
$$= \texttt{gsl\_ran\_flat\_pdf}(\texttt{x}, \texttt{alpha}, \texttt{beta})$$
$$E(x|\alpha,\beta) = \frac{\beta-\alpha}{2}$$
$$\text{var}(x|\alpha,\beta) = \frac{(\beta-\alpha)^2}{12}$$
$$\int_{-\infty}^{x} \mathcal{U}(y|\alpha,\beta)dy = \begin{cases} 0 & x < \alpha \\ \frac{x-\alpha}{\beta-\alpha} & x \in [\alpha,\beta] \\ 1 & x > \beta \end{cases}$$
$$= \texttt{gsl\_cdf\_flat\_P}(\texttt{x}, \texttt{alpha}, \texttt{beta})$$
$$\int_{x}^{\infty} \mathcal{U}(y|\alpha,\beta)dy = \texttt{gsl\_cdf\_flat\_Q}(\texttt{x}, \texttt{alpha}, \texttt{beta})$$
$$RNG, \text{general} : \texttt{gsl\_ran\_flat}(\texttt{r}, \texttt{alpha}, \texttt{beta})$$
$$RNG, \alpha = 0, \beta = 1 : \texttt{gsl\_rng\_uniform}(\texttt{r})$$

- Probability theorists through the ages have developed models that indicate that if a process follows certain guidelines, the data will have a predictable form.
- A single draw from a binary event with fixed probability has a Bernoulli distribution; from this, a wealth of other distributions can be derived.
- An event which occurs with frequency λ per period (or λ per volume, et cetera) is known as a Poisson process; a wealth of distributions can be derived for such a process.
- If $\overline{x}$ is the mean of a set of independent, identically distributed draws from *any* nondegenerate distribution, then the distribution of $\overline{x}$ approaches a Normal distribution. This is the Central Limit Theorem.
- The Beta distribution is useful for modeling a variety of variables that are restricted to $[0, 1]$. It can be unimodal, bimodal, lean in either direction, or can simply match the Uniform distribution.

7.3 USING THE SAMPLE DISTRIBUTIONS

Here are some examples of how you could use the distributions described above to practical benefit.

LOOKING UP FIGURES If I have fifty draws from a Bernoulli event with probability .25, what is the likelihood that I will have more than twenty successes?

Statistics textbooks used to include an appendix listing tables of common distributions, but those tables are effectively obsolete, and more modern textbooks refer the reader to the appropriate function in a stats package. For those who long for the days of grand tables, the code supplement includes `normaltable.c`, code for producing a neatly formatted table of CDFs for a set of Normal distributions (the p-value often reported with hypothesis tests is one minus the listed value).

The code is not printed here because it is entirely boring, but the tables it produces provide another nice way to get a feel for the distributions.

Alternatively, Apophenia's command-line program `apop_lookup` will look up a quick number for you.

GENERATING DATA FROM A DISTRIBUTION Each distribution neatly summarizes an oft-modeled story, and so each can be used as a capsule simulation of a process, either by itself or as a building block for a larger simulation.

Listing 7.12 gives a quick initial example. It is based on work originated by Gibrat (1931) and extended by many others, including Axtell (2006), regarding *Zipf's law*, that the distribution of the sizes of cities, firms, or other such agglomerations tends toward an Exponential-type form. In the model here, this comes about because agents' growth rates are assumed to be the mean of a set of iid random shocks, and so are Normally distributed.

- First, the program produces a set of agents with one characteristic: size, stored in a `gsl_vector`. The `initialize` function draws agent sizes from a Uniform$[0, 100]$ distribution. To do this, it requires a `gsl_rng`, which `main` allocates using `apop_rng_alloc` and passes to `initialize`. See Chapter 11 for more on using random number generators.
- Each period, the firms grow by a Normally distributed rate (via the `grow` function). That is, the `grow` function randomly draws g from a `gsl_rng_gaussian`, and then reassigns the firm size to $size \leftarrow size * \exp(g)$. The most likely growth rate is therefore $\exp(0) = 1$. When $g < 0$, $\exp(g) < 1$; and when $g > 0$, $\exp(g) > 1$.

```
#include <apop.h>

int agentct = 5000;
int periods = 50;
int binct = 30;
double pauselength = 0.6;
gsl_rng *r;

void initialize(double *setme){
    *setme = gsl_rng_uniform(r)*100;
}

void grow(double *val){
    *val *= exp(gsl_ran_gaussian(r,0.1));
}

double estimate(gsl_vector *agentlist){
    return apop_vector_mean(agentlist);
}

int main(){
  gsl_vector *agentlist = gsl_vector_alloc(agentct);
    r = apop_rng_alloc(39);
    apop_vector_apply(agentlist, initialize);
    for (int i=0; i< periods; i++){
        apop_plot_histogram(agentlist, binct, NULL);
        printf("pause %g\n", pauselength);
        apop_vector_apply(agentlist, grow);
    }
    fprintf(stderr, "the mean: %g\n", estimate(agentlist));
}
```

Listing 7.12 A model of Normally distributed growth. Online source: `normalgrowth.c`.

Also, $\exp(g) * \exp(-g) = 1$, and by the symmetry of the Normal distribution, g and $-g$ have equal likelihood, so it is easy for an agent to find good luck in period one countered by comparable bad luck in period two, leaving it near where it had started.

- The output is a set of Gnuplot commands, so use `./normalgrowth | gnuplot`. With a `pause` between each histogram, the output becomes an animation, showing a quick transition from a Uniform distribution to a steep Lognormal distribution, where most agents are fast approaching zero size, but a handful have size approaching 1,000.[8]

[8]Here, the x-axis is the firm size, and the y-axis is the number of firms. Typically, Zipf-type distributions are displayed somewhat differently: the x-axis is the *rank* of the firm, 1st, 2nd, 3rd, et cetera, and the y-axis is the size of the so-ranked firm. Converting to this form is left as an exercise to the reader. (*Hint*: use `gsl_vector_sort`.)

- The last step is a model estimation, to which we will return in a few pages. Its output is printed to `stderr`, aka the screen, so that the pipe to Gnuplot is not disturbed.

SIMULATION Fein *et al.* (1988) found that their depressive patients responded well to a combination of Lithium and a monoamine oxidase inhibitor (MAOI). But both types of drug require careful monitoring: Lithium overdoses are common and potentially damaging, while the combination of MAOIs and chocolate can be fatal.

```
#include <apop.h>

double find_lambda(double half_life){
  double lambda = -half_life/log(1/2.);
  return gsl_cdf_exponential_Q(1, lambda);
}

int main(){
  double li = 0, maoi = 0;
  int days = 10;
  gsl_matrix *d = gsl_matrix_alloc(days*24,4);
  double hourly_decay1 = find_lambda(20.); //hours; lithium carbonate
  double hourly_decay2 = find_lambda(11.); //hours; phenelzine
    for (size_t i=0; i < days*24; i ++){
      li *= hourly_decay1;
      maoi *= hourly_decay2;
      if (i % 24 == 0)
          li+= 600;
      if ((i+12) % 24 == 0)
          maoi+= 45;
      APOP_MATRIX_ROW(d, i, onehour);
      apop_vector_fill(onehour, i/24., li/10., maoi, maoi/li*100.);
    }
    printf("plot 'maoi.out' using 1:2 with lines title 'Li/10', \
                  'maoi.out' using 1:3 with lines title 'MAOI', \
                  'maoi.out' using 1:4 with lines title 'MAOI/Li, pct\n");
    remove("maoi.out");
    apop_matrix_print(d, "maoi.out");
}
```

Listing 7.13 A simulation of the blood stream of a person taking two drugs. Online source: `maoi.c`.

Listing 7.13 simulates a patient's blood stream as she follows a regime of Lithium carbonate (average half life: about 20 hours, with high variance) and an MAOI named phenelzine (average half life: 11 hours). As per the story on page 247, when the drug leaves the blood stream via a Poisson process, the amount of a drug remaining in the blood is described by an Exponential distribution.

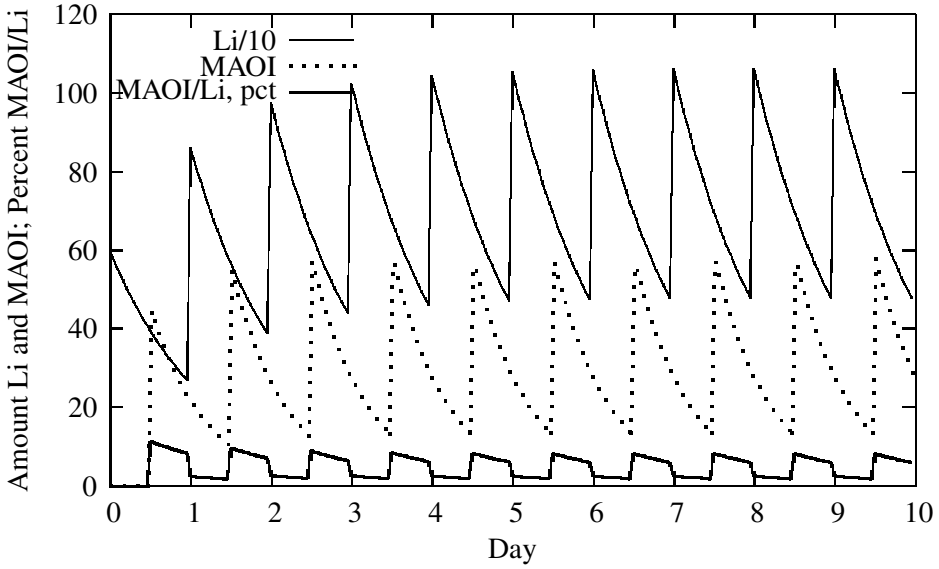


Figure 7.14 The typical sawtooth pattern of decay and renewal.

- The first step is to convert from the half life commonly used by pharmacists to the λ parameter in the exponential distribution. The `find_lambda` function does this.
- Given λ, `gsl_cdf_exponential_Q(1, lambda)` answers the question of what percentage of a given initial level is remaining after one hour.
- The main simulation is a simple hourly `for` loop, that decrements the amount of drug in the blood stream by the amount calculated above, then checks whether it is time for our patient to take one of her meds, and then records the various levels on her chart.
- The ungainly `printf` statement at the end plots the result. Gnuplot does not save data, so it needs to reread the data file three times to plot the three columns.

Figure 7.14 shows the density in blood as our subject takes 600 mg of Lithium at midnight every day, and 45 mg of MAOI at noon every day. For convenience in scaling of the plot, the amount of Lithium in the blood stream is divided by ten. In the later days, after the subject's system has reached its dynamic equilibrium, she takes in 600 mg of Li per day, and loses about 600 mg per day; similarly for the 45 mg of MAOI. The ratio of MAOI/Li jumps constantly over the range of 187% to 824%.

$\mathbb{Q}_{7.9}$ Derive or verify that the `find_lambda` function correctly converts between half life and the Exponential distribution's λ parameter.

$\mathbb{Q}_{7.10}$ Let there be two stocks: employed and unemployed. Let the half-life of employment (i.e., transition to unemployment) be 365 days, and the half-life of unemployment (i.e., finding a job) be 3 weeks (21 days).
Modify the Lithium/MAOI program to model the situation. For each period, calculate the loss from both the employment and the unemployment stocks, and then transfer the appropriate number of people to the other stock. What is the equilibrium unemployment rate?

FITTING EXISTING DATA The common goal throughout the book is to estimate the parameters of the model with data, so given a data set, how can we find the parameters for the various distributions above?

You can see above that almost every parameter can be solved—sometimes over-solved—using the mean and variance. For example, Equations 7.2.1 and 7.2.2 (describing the parameters of a Binomial distribution) are a system of two equations in two unknowns:

$$\mu = np$$
$$\sigma^2 = np(1-p)$$

It is easy to calculate estimates of μ and σ from data, $\hat{\mu}$ and $\hat{\sigma}$, and we could plug those estimates into the above system of equations to find the parameters of the distribution. You can verify that for these two equations we would have

$$\hat{n} = \frac{\hat{\mu}^2}{\hat{\mu} - \hat{\sigma}^2}$$
$$\hat{p} = 1 - \frac{\hat{\sigma}^2}{\hat{\mu}}.$$

This is *method of moments* estimation (see, e.g., Greene (1990, pp 117*ff*)). To summarize the method, we write down the parameter estimates as functions of the mean, variance, skew, and kurtosis, then we find estimates of those parameters from the data, and use those parameter estimates to solve for the parameter estimates of the model itself.

But problems easily crop up. For example, we can just count observations to find the value of n for our data set, so given n, $\hat{\mu}$, and $\hat{\sigma}^2$, our system of equations is now two equations with only one unknown (p). The Poisson distribution had a similar but simpler story, because its single parameter equals two different moments:

$$\mu = \lambda$$
$$\sigma^2 = \lambda.$$

So if our data set shows $\hat{\mu} = 1.2$ and $\hat{\sigma}^2 = 1.4$, which do we use for $\hat{\lambda}$? Apophenia doesn't fret much about this issue and just uses $\hat{\mu}$, because this is also the maximum

likelihood estimator (MLE) of λ (where MLE will be discussed fully in Chapter 10).

For the Uniform, the method of moments doesn't work either: the expression $(\beta - \alpha)$ is oversolved with the two equations, but there is no way to solve for α or β alone. However, a few moments' thought will show that the most likely value for (α, β) given data $\mathbf{x}$ is simply $(\min(\mathbf{x}), \max(\mathbf{x}))$.

Most of the above distributions have an `apop_model` associated (`apop_normal`, `apop_gamma`, `apop_uniform`, et cetera), and if you have a data set on hand, you can quickly estimate the above parameters:

```
apop_data *d = your_data;
apop_model *norm = apop_estimate(d, apop_normal);
apop_model *beta = apop_estimate(d, apop_beta);
apop_model_show(norm);
apop_model_show(beta);
apop_model_show(apop_estimate(d, apop_gamma));
```

$\mathbb{Q}_{7.11}$ Listing 7.12 produces a data set that should be Zipf distributed. Add an estimation in the `estimate` function to see how well it fits.
Better still, run a tournament. First, declare an array of several models, say the Lognormal, Zipf, Exponential, and Gamma. Write a `for` loop to estimate each model with the data, and fill an array of confidence levels based on log-likelihood tests. [Is such a tournament valid? See the notes on the multiple testing problem on 316.]

The method of moments provides something of a preview of the working of the various model-based estimations in the remainder of the book. It took in data, and produced an estimate of the model parameters, or an estimate of a statistic using the estimate of the model parameters that were produced using data.

As the reader may have noticed, all these interactions between data, model parameters, and statistics create many opportunities for confusion. Here are some notes to bear in mind:

- The expected value, variance, and other such measures of a *data set*, when no model is imposed, is a function of the data. [E.g., $E(\mathbf{x})$.]
- The expected value, variance, and other such measures of a *model* are functions of the ideal parameters, not any one data set. [E.g., $E(\mathbf{x}|\beta)$ is only a function of β.]
- Our *estimate* of model parameters given a data set is a function of the given data set (and perhaps any known parameters, if there are any on hand). For example, the Normal parameter μ is a part of the model specification, but the estimate of μ,

which we write as $\hat{\mu}$, is a function of the data. Any variable with a hat, like $\hat{p}$, could be notated as a function of the data, $\hat{p}(\mathbf{x})$.

- We will often have a statistic like $E(\mathbf{x})$ that is a function of the data—in fact, we define a *statistic* to be *a function of data*. But models often have data-free analogues to these statistics. Given a probability distribution $P(x, \beta)$, the expected value $E(f(x)|\beta) = \int_{\forall x} f(x)P(x, \beta)dx$, meaning that we integrate over all x, and so $E(f(x)|\beta)$ is a function of only β. The model in which β lives is almost always taken as understood by context, and many authors take the parameters as understood by context as well, leaving the expected value to be written as $E(f(x))$, even though this expression is a function of β, not x.

BAYESIAN UPDATING The definition of a conditional probability is based on the statement $P(A \cap B) = P(A|B)P(B)$; in English, the likelihood of A and B occurring at the same time equals the likelihood of A occurring given that B did, times the likelihood that B occurs. The same could be said reversing A and B: $P(A \cap B) = P(B|A)P(A)$. Equating the two complementary forms and shunting over $P(B)$ gives us the common form of Bayes's rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Now to apply it. Say that we have a prior belief regarding a parameter, such as that the distribution of the mean of a data set is $\sim \mathcal{N}(0, 1)$; let this be $Pri(\beta)$. We gather a data set $\mathbf{X}$, and can express the likelihood that we would have gathered this data set given any haphazard value of β, $P(\mathbf{X}|\boldsymbol{\beta})$. Let $\mathbb{B}$ be the entire range of values that β could take on. We can then use Bayes's rule to produce a *posterior distribution*:

$$Post(\beta|\mathbf{X}) = \frac{P(\mathbf{X}|\beta)Pri(\beta)}{P(\mathbf{X})}$$

So on the right-hand side, we had a prior belief about β's value expressed as a distribution, and a likelihood function $P(\mathbf{X}|\beta)$ expressing the odds of observing the data we observed given any one parameter. On the left-hand side, we have a new distribution for β, which takes into account the fact that we have observed the data $\mathbf{X}$. In short, this equation used the data to update our beliefs about the distribution of β from $Pri(\beta)$ to $Post(\beta)$.

The numerator is relatively clear, and requires only local information, but we can write $P(\mathbf{X})$ in its full form—

$$Post(\beta|\mathbf{X}) = \frac{P(\mathbf{X}|\beta)Pri(\beta)}{\int_{\forall B \in \mathbb{B}} P(\mathbf{X}|B)Pri(B)dB}$$

—to reveal that the denominator is actually global information, because calculating it requires covering the entire range that β could take on. Local information is easy and global information is hard (see pages 325 *ff*), so Bayesian updating is often described via a form that just ignores the global part:

$$Post(\beta|\mathbf{X}) \propto P(\mathbf{X}|\beta)Pri(\beta).$$

That is, the posterior equals the amount on the right-hand side times a fixed amount (the denominator above) that does not depend on any given value of β. This is already enough to compare ratios like $Post(\beta_1|\mathbf{X})/Post(\beta_2|\mathbf{X})$, and given the right conditions, such a ratio is already enough for running likelihood ratio tests (as discussed in Chapter 10).

Computationally, there are two possibilities for moving forward given the problem of determining the global scale of the distribution. First, there are a number of *conjugate distribution* pairs that can be shown to produce an output model that matches the prior in form but has updated parameters. In this case, the `apop_update` function simply returns the given model and its new parameters; see the example below.

Chapter 11 will present a computationally-intensive method of producing a posterior distribution when the analytic route is closed (i.e., *Monte Carlo Maximum Likelihood*). But for now we can take `apop_update` as a black box that takes in two models and outputs an updated conjugate form where possible, and an empirical distribution otherwise. We could then make draws from the output distribution, plot it, use it as the prior to a new updating procedure when a new data set comes in, et cetera.

An example: Beta ♡ Binomial For now, assume that the likelihood that someone has a tattoo is constant for all individuals, regardless of age, gender, ...(we will drop this clearly false assumption in the section on multilevel modeling, page 288). We would like to know the value of that overall likelihood. That is, the statistic of interest is $p \equiv$ (count of people who have tattoos)/(total number of people in the sample).

Because we have weak knowledge of p, we should describe our beliefs about its value using a distribution: p has small odds of being near zero or one, a reasonable chance of being about 10%, and so on. The Beta distribution is a good way to describe the distribution, because it is positive only for inputs between zero and one. Let $\mathcal{B}$ indicate the Beta distribution; then $\mathcal{B}(1,1)$ is a Uniform(0, 1) distribution, which is a reasonably safe way to express a neutral prior belief about p. Alternatively, setting $Pri(p)$ to be $\mathcal{B}(2,2)$ will put more weight around $p = 1/2$ and less at the extremes, and raising the second parameter a little more will bring the mode of our beliefs about p below $1/2$ [See Figure 7.11].

Given p, the distribution of the expected count of tattooed individuals is Binomial. For each individual, there is a p chance of having a tattoo—a simple Bernoulli draw. The overall study makes $n = 500$ such draws, and thus fits the model underlying the Binomial distribution perfectly. But we do not yet know p, so this paragraph had to begin by taking p as given. That is, the Binomial distribution describes $P(\text{data}|p)$.

It so happens that the Beta and Binomial distributions are conjugate. This means that, given that $Pri(p)$ is a Beta distribution and $P(\text{data}|p)$ is a Binomial distribution, the posterior $Post(p|\text{data})$ is a Beta distribution, just like the prior. Tables of other such conjugate pairs are readily available online.

However, the parameters are updated to accommodate the new information. Let x be the number of tattoos observed out of n subjects, and the prior distribution be $\mathcal{B}(\alpha, \beta)$. Then the posterior is a $\mathcal{B}(\alpha + x, \beta + n - x)$ distribution. The discussion of the prior offered possibilities like $\alpha = \beta = 1$ or $\alpha = \beta = 2$. But the survey has 500 subjects; the count of tattooed individuals alone dwarfs $\alpha = 2$. Therefore, we can approximate the posterior as simply $\mathcal{B}(x, n - x)$.

The catalog above listed the expected value of a Beta distribution as $\frac{\alpha}{\alpha+\beta}$. With $\alpha = x$ and $\beta = n - x$, this reduces simply to x/n. That is, the expected posterior value of p is the percentage of people in our sample who have tattoos ($\hat{p}$). Bayesian updating gave us a result exactly as we would expect.

The variance of a Beta distribution is

$$\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

Again, with N around 500, the 1 in the denominator basically disappears. Filling in $\alpha = x$ and $\beta = n - x$, we get

$$\frac{\hat{p}(1 - \hat{p})}{n}.$$

Again, this is what we would get from the Binomial distribution.

We call a $\mathcal{B}(\alpha, \beta)$ distribution with small α and β a *weak prior*, by which we mean that a moderately-sized data set entirely dwarfs the beliefs we expressed in the prior. So what is the point of the updating process? First, we could use a stronger prior, like $\alpha = 200, \beta = 300$, which would still have some effect on the posterior distribution even after updating with the data set.

Second, the system provides a consistent mechanism for combining multiple data sets. The posterior distribution that you have after accounting for a data set will have a form appropriate for use as a prior distribution to be updated by the next data set. Thus, Bayesian updating provides a natural mechanism for running metastudies.

$\mathbb{Q}_{7.12}$ Verify that `apop_update` using a Beta prior, a Binomial likelihood function, and the tattoo data does indeed produce the estimated mean and variance as the simple x/n estimate. Gather the data from the column `tattoos.'ego has tattoos'`, which is coded 1=yes, 2=no, and calculate $\hat{\mu}$ and $\widehat{\sigma}^2$ using the formulæ in the above few paragraphs.
What results do you get when you assume a stronger prior, like $\mathcal{B}(200, 800)$ or $\mathcal{B}(800, 200)$?

7.4 NON-PARAMETRIC DESCRIPTION

Say that we have a data set and would like to know the distribution from which the data was drawn. To this point, we assumed the form of the distribution (Normal, Binomial, Poisson, et cetera) and then had only to estimate the parameters of the distribution from data. But without assuming a simple parametric form, how else could we describe the distribution from which the data was drawn?

The simplest answer would be a plain old histogram of the drawn data. This is often sufficient. But especially for small data sets, the histogram has dissatisfactory features. If we make four draws, and three have value 20 and one has value 22, does this mean that 21 has probability zero, or we just didn't have the luck of drawing a 21 this time?

Thus, a great deal of nonparametric modeling consists of finding ways to smooth a histogram based on the claim that the actual distribution is not as lumpy as the data.

The histogram The histogram is the most assumption-free way to describe the likelihood distribution from which the data was drawn. Simply lay down a row of bins, pile each data point into the appropriate bin, normalize the bins to sum to one if so desired, and plot. Because the most common use of a histogram (after just plotting it) is using it to make random draws, the full discussion of histogram production will appear in the chapter on random draws, on page 361.

The key free variable in a histogram is the *bandwidth*—the range over the x-axis that goes into each data point. If the bandwidth is too small, then the histogram will have many slices and generally be as spiky as the data itself. A too-large bandwidth oversmooths—at an infinite bandwidth, the histogram would have only one bar, which is not particularly informative. Formally, there is a bias-variance trade-off between the two extremes, but most of us just try a few bandwidths until we get something that looks nice. See Givens & Hoeting (2005, ch 11) for an extensive discussion of the question in the context of data smoothing.

Moving average The simplest means of smoothing data is a moving average, replacing each data point with the mean of the adjacent b data points (where b is the bandwidth). You could use this for histograms or for any other series. For example, `movingavg.c` in the online code supplement plots the temperature deviances as shown in `data-climate.db`, and a moving average that replaces each data point with the mean deviation over a two-year window, based on the `apop_vector_moving_average` function.

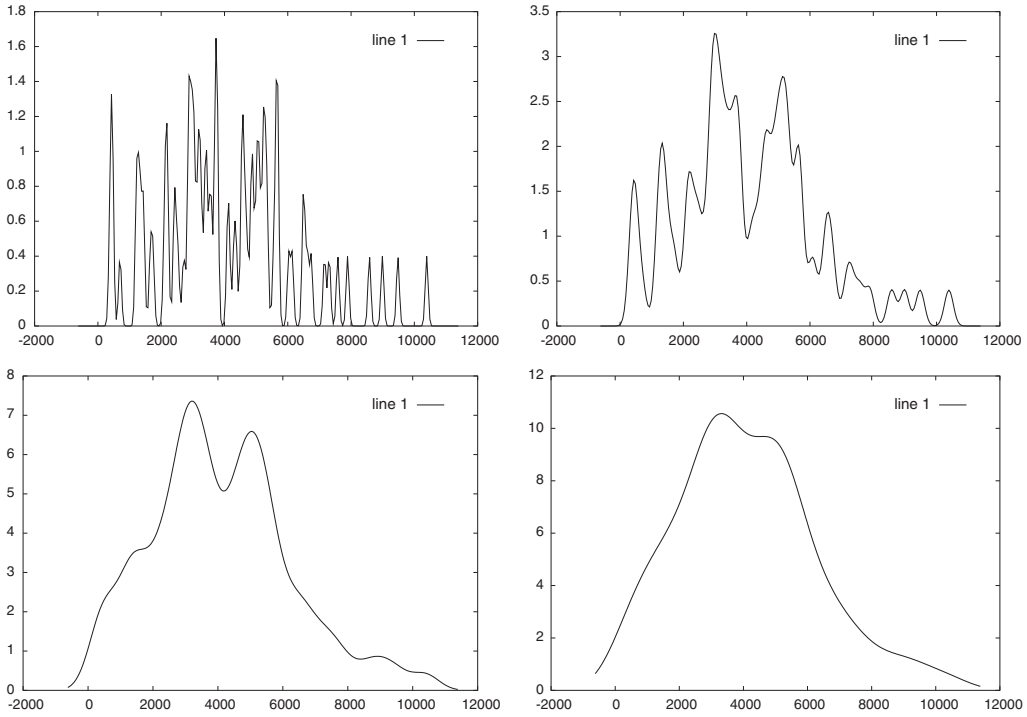


Figure 7.15 A series of density plots. As h rises, the kernel density smooths out and has fewer peaks.

Kernel smoothing The *kernel density estimate* is based on this function:

$$\hat{f}(t, X, h) = \frac{\sum_{i=1}^{n} \mathcal{N}((t - X_i)/h)}{n \cdot h},$$

where $X_1, X_2, \ldots X_n \in \mathbb{R}$ are the n data points observed, $\mathcal{N}(y)$ is a Normal$(0, 1)$ density function evaluated at y, and $h \in \mathbb{R}^+$ is the bandwidth. Thus, the overall curve is the sum of a set of subcurves, each centered over a different data point. Figure 7.15 shows the effect of raising h on the shape of a set of fitted curves.[9] When h is very small, the Normal distributions around each data point are sharp spikes, so there is a mode at every data point. As h grows, the spikes spread out and merge, until the sum of subdistributions produces a single bell curve. See page

[9]The data is the male viewership for 86 TV specials, from Chwe (2001).

376 for more on how these plots were generated; see also Silverman (1985).

As usual, there is a simple form for code to produce a default kernel density from a data set, and a more extensive form that allows more control. Try Listing 7.16, which plots the histogram of precipitation figures and the kernel-density smoothed version based on a $\mathcal{N}(0, 0.1)$ kernel. Also try $\sigma = 0.001$, 0.01, and 0.2 to see the progression from the data's spikes to a smooth bell curve.

```
#include <apop.h>

int main(){
    apop_db_open("data-climate.db");
    apop_data *data = apop_query_to_data("select pcp from precip");
    apop_model *h = apop_estimate(data, apop_histogram);
    apop_histogram_normalize(h);
    remove("out.h"); remove("out.k");
    apop_histogram_print(h, "out.h");
    apop_model *kernel = apop_model_set_parameters(apop_normal, 0., 0.1);
    apop_model *k = apop_model_copy(apop_kernel_density);
    Apop_settings_add_group(k, apop_kernel_density, NULL, h, kernel, NULL);
    apop_histogram_print(k, "out.k");
    printf("plot 'out.h' with lines title 'data', 'out.k' with lines title 'smoothed'\n");
}
```

Listing 7.16 A histogram before and after smoothing via kernel densities. Run via `smoothing | gnuplot`. Online source: `smoothing.c`.

Q 7.13

Plot the `data-tv` set using:

- a histogram, using 40 and 100 bins,
- a smoothed version of the 40-bin histogram, via a moving average of bandwidth four,
- the 40-bin histogram smoothed via a Normal(x, 100.0) kernel density,
- the 40-bin histogram smoothed via a Uniform($x - 500.0$, $x + 500.0$) kernel density.