Chapter 6

Operations on distributions

6.1 Skewness

Skewness is a statistic that measures the asymmetry of a distribution. Given a sequence of values, x_i, the sample skewness is:

$$g_1 = m_3 / m_2^{3/2}$$

$$m_2 = \frac{1}{n} \sum_i (x_i - \mu)^2$$

$$m_3 = \frac{1}{n} \sum_i (x_i - \mu)^3$$

You might recognize m_2 as the mean squared deviation (also known as variance); m_3 is the mean cubed deviation.

Negative skewness indicates that a distribution "skews left;" that is, it extends farther to the left than the right. Positive skewness indicates that a distribution skews right.

In practice, computing the skewness of a sample is usually not a good idea. If there are any outliers, they have a disproportionate effect on g_1.

Another way to evaluate the asymmetry of a distribution is to look at the relationship between the mean and median. Extreme values have more effect on the mean than the median, so in a distribution that skews left, the mean is less than the median.

Pearson's median skewness coefficient is an alternative measure of skewness that explicitly captures the relationship between the mean, μ, and the

median, $\mu_{1/2}$:

$$g_p = 3(\mu - \mu_{1/2})/\sigma$$

This statistic is **robust**, which means that it is less vulnerable to the effect of outliers.

Exercise 6.1 Write a function named `Skewness` that computes g_1 for a sample.

Compute the skewness for the distributions of pregnancy length and birth weight. Are the results consistent with the shape of the distributions?

Write a function named `PearsonSkewness` that computes g_p for these distributions. How does g_p compare to g_1?

Exercise 6.2 The "Lake Wobegon effect" is an amusing nickname[1] for **illusory superiority**, which is the tendency for people to overestimate their abilities relative to others. For example, in some surveys, more than 80% of respondents believe that they are better than the average driver (see `http://wikipedia.org/wiki/Illusory_superiority`).

If we interpret "average" to mean median, then this result is logically impossible, but if "average" is the mean, this result is possible, although unlikely.

What percentage of the population has more than the average number of legs?

Exercise 6.3 The Internal Revenue Service of the United States (IRS) provides data about income taxes, and other statistics, at `http://irs.gov/taxstats`. If you did Exercise 4.13, you have already worked with this data; otherwise, follow the instructions there to extract the distribution of incomes from this dataset.

What fraction of the population reports a taxable income below the mean?

Compute the median, mean, skewness and Pearson's skewness of the income data. Because the data has been binned, you will have to make some approximations.

The Gini coefficient is a measure of income inequality. Read about it at `http://wikipedia.org/wiki/Gini_coefficient` and write a function called `Gini` that computes it for the income distribution.

[1] If you don't get it, see `http://wikipedia.org/wiki/Lake_Wobegon`.

Hint: use the PMF to compute the relative mean difference (see `http://wikipedia.org/wiki/Mean_difference`).

You can download a solution to this exercise from `http://thinkstats.com/gini.py`.

6.2 Random Variables

A **random variable** represents a process that generates a random number. Random variables are usually written with a capital letter, like X. When you see a random variable, you should think "a value selected from a distribution."

For example, the formal definition of the cumulative distribution function is:

$$\text{CDF}_X(x) = P(X \leq x)$$

I have avoided this notation until now because it is so awful, but here's what it means: The CDF of the random variable X, evaluated for a particular value x, is defined as the probability that a value generated by the random process X is less than or equal to x.

As a computer scientist, I find it helpful to think of a random variable as an object that provides a method, which I will call `generate`, that uses a random process to generate values.

For example, here is a definition for a class that represents random variables:

```
class RandomVariable(object):
    """Parent class for all random variables."""
```

And here is a random variable with an exponential distribution:

```
class Exponential(RandomVariable):
    def __init__(self, lam):
        self.lam = lam

    def generate(self):
        return random.expovariate(self.lam)
```

The init method takes the parameter, λ, and stores it as an attribute. The `generate` method returns a random value from the exponential distribution with that parameter.

Each time you invoke `generate`, you get a different value. The value you get is called a **random variate**, which is why many function names in the `random` module include the word "variate."

If I were just generating exponential variates, I would not bother to define a new class; I would use `random.expovariate`. But for other distributions it might be useful to use RandomVariable objects. For example, the Erlang distribution is a continuous distribution with parameters λ and k (see `http://wikipedia.org/wiki/Erlang_distribution`).

One way to generate values from an Erlang distribution is to add k values from an exponential distribution with the same λ. Here's an implementation:

```
class Erlang(RandomVariable):
    def __init__(self, lam, k):
        self.lam = lam
        self.k = k
        self.expo = Exponential(lam)

    def generate(self):
        total = 0
        for i in range(self.k):
            total += self.expo.generate()
        return total
```

The init method creates an Exponential object with the given parameter; then `generate` uses it. In general, the init method can take any set of parameters and the `generate` function can implement any random process.

Exercise 6.4 Write a definition for a class that represents a random variable with a Gumbel distribution (see `http://wikipedia.org/wiki/Gumbel_distribution`).

6.3 PDFs

The derivative of a CDF is called a **probability density function**, or PDF. For example, the PDF of an exponential distribution is

$$\text{PDF}_{expo}(x) = \lambda e^{-\lambda x}$$

The PDF of a normal distribution is

$$\text{PDF}_{normal}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

Evaluating a PDF for a particular value of x is usually not useful. The result is not a probability; it is a probability *density*.

In physics, density is mass per unit of volume; in order to get a mass, you have to multiply by volume or, if the density is not constant, you have to integrate over volume.

Similarly, probability density measures probability per unit of x. In order to get a probability mass[2], you have to integrate over x. For example, if x is a random variable whose PDF is PDF_X, we can compute the probability that a value from X falls between -0.5 and 0.5:

$$P(-0.5 \le X < 0.5) = \int_{-0.5}^{0.5} \text{PDF}_X(x)dx$$

Or, since the CDF is the integral of the PDF, we can write

$$P(-0.5 \le X < 0.5) = \text{CDF}_X(0.5) - \text{CDF}_X(-0.5)$$

For some distributions we can evaluate the CDF explicitly, so we would use the second option. Otherwise we usually have to integrate the PDF numerically.

Exercise 6.5 What is the probability that a value chosen from an exponential distribution with parameter λ falls between 1 and 20? Express your answer as a function of λ. Keep this result handy; we will use it in Section 8.8.

Exercise 6.6 In the BRFSS (see Section 4.5), the distribution of heights is roughly normal with parameters μ = 178 cm and σ^2 = 59.4 cm for men, and μ = 163 cm and σ^2 = 52.8 cm for women.

In order to join Blue Man Group, you have to be male between 5'10" and 6'1" (see `http://bluemancasting.com`). What percentage of the U.S. male population is in this range? Hint: see Section 4.3.

[2]To take the analogy one step farther, the mean of a distribution is its center of mass, and the variance is its moment of inertia.

6.4 Convolution

Suppose we have two random variables, X and Y, with distributions CDF_X and CDF_Y. What is the distribution of the sum $Z = X + Y$?

One option is to write a RandomVariable object that generates the sum:

```
class Sum(RandomVariable):
    def __init__(X, Y):
        self.X = X
        self.Y = Y

    def generate():
        return X.generate() + Y.generate()
```

Given any RandomVariables, X and Y, we can create a Sum object that represents Z. Then we can use a sample from Z to approximate CDF_Z.

This approach is simple and versatile, but not very efficient; we have to generate a large sample to estimate CDF_Z accurately, and even then it is not exact.

If CDF_X and CDF_Y are expressed as functions, sometimes we can find CDF_Z exactly. Here's how:

1. To start, assume that the particular value of X is x. Then $\text{CDF}_Z(z)$ is

$$P(Z \leq z \mid X = x) = P(Y \leq z - x)$$

   Let's read that back. The left side is "the probability that the sum is less than z, given that the first term is x." Well, if the first term is x and the sum has to be less than z, then the second term has to be less than $z - x$.

2. To get the probability that Y is less than $z - x$, we evaluate CDF_Y.

$$P(Y \leq z - x) = \text{CDF}_Y(z - x)$$

   This follows from the definition of the CDF.

3. Good so far? Let's go on. Since we don't actually know the value of x, we have to consider all values it could have and integrate over them:

$$P(Z \leq z) = \int_{-\infty}^{\infty} P(Z \leq z \mid X = x)\ \text{PDF}_X(x)\ dx$$

The integrand is "the probability that Z is less than or equal to z, given that $X = x$, times the probability that $X = x$."

Substituting from the previous steps we get

$$P(Z \leq z) = \int_{-\infty}^{\infty} \text{CDF}_Y(z - x)\ \text{PDF}_X(x)\ dx$$

The left side is the definition of CDF_Z, so we conclude:

$$\text{CDF}_Z(z) = \int_{-\infty}^{\infty} \text{CDF}_Y(z - x)\ \text{PDF}_X(x)\ dx$$

4. To get PDF_Z, take the derivative of both sides with respect to z. The result is

$$\text{PDF}_Z(z) = \int_{-\infty}^{\infty} \text{PDF}_Y(z - x)\ \text{PDF}_X(x)\ dx$$

If you have studied signals and systems, you might recognize that integral. It is the **convolution** of PDF_Y and PDF_X, denoted with the operator $\star$.

$\text{PDF}_Z = \text{PDF}_Y \star \text{PDF}_X$

So the distribution of the sum is the convolution of the distributions. See `http://wiktionary.org/wiki/booyah`!

As an example, suppose X and Y are random variables with an exponential distribution with parameter λ. The distribution of $Z = X + Y$ is:

$$\text{PDF}_Z(z) = \int_{-\infty}^{\infty} \text{PDF}_X(x)\ \text{PDF}_Y(z - x)\ dx = \int_{-\infty}^{\infty} \lambda e^{-\lambda x}\ \lambda e^{-\lambda(z-x)} dx$$

Now we have to remember that PDF_{expo} is 0 for all negative values, but we can handle that by adjusting the limits of integration:

$$\text{PDF}_Z(z) = \int_{0}^{z} \lambda e^{-\lambda x}\ \lambda e^{-\lambda(z-x)}\ dx$$

Now we can combine terms and move constants outside the integral:

$$\text{PDF}_Z(z) = \lambda^2 e^{-\lambda z} \int_{0}^{z} dx = \lambda^2 z\ e^{-\lambda z}$$

This, it turns out, is the PDF of an Erlang distribution with parameter $k = 2$ (see `http://wikipedia.org/wiki/Erlang_distribution`). So the convolution of two exponential distributions (with the same parameter) is an Erlang distribution.

Exercise 6.7 If X has an exponential distribution with parameter λ, and Y has an Erlang distribution with parameters k and λ, what is the distribution of the sum $Z = X + Y$?

Exercise 6.8 Suppose I draw two values from a distribution; what is the distribution of the larger value? Express your answer in terms of the PDF or CDF of the distribution.

As the number of values increases, the distribution of the maximum converges on one of the extreme value distributions; see `http://wikipedia.org/wiki/Gumbel_distribution`.

Exercise 6.9 If you are given Pmf objects, you can compute the distribution of the sum by enumerating all pairs of values:

```
for x in pmf_x.Values():
    for y in pmf_y.Values():
        z = x + y
```

Write a function that takes PMF_X and PMF_Y and returns a new Pmf that represents the distribution of the sum $Z = X + Y$.

Write a similar function that computes the PMF of $Z = \max(X, Y)$.

6.5 Why normal?

I said earlier that normal distributions are amenable to analysis, but I didn't say why. One reason is that they are closed under linear transformation and convolution. To explain what that means, it will help to introduce some notation.

If the distribution of a random variable, X, is normal with parameters μ and σ, you can write

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

where the symbol $\sim$ means "is distributed" and the script letter $\mathcal{N}$ stands for "normal."

A linear transformation of X is something like $X' = aX + b$, where a and b are real numbers. A family of distributions is closed under linear transformation if X' is in the same family as X. The normal distribution has this property; if $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$X' \sim \mathcal{N}(a\mu + b, a^2\,\sigma^2)$$

Normal distributions are also closed under convolution. If $Z = X + Y$and $X \sim \mathcal{N}(\mu_X, {\sigma_X}^2)$ and $Y \sim \mathcal{N}(\mu_Y, {\sigma_Y}^2)$ then

$$Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

The other distributions we have looked at do not have these properties.

Exercise 6.10 If $X \sim \mathcal{N}(\mu_X, {\sigma_X}^2)$ and $Y \sim \mathcal{N}(\mu_Y, {\sigma_Y}^2)$, what is the distribution of $Z = aX + bY$?

Exercise 6.11 Let's see what happens when we add values from other distributions. Choose a pair of distributions (any two of exponential, normal, lognormal, and Pareto) and choose parameters that make their mean and variance similar.

Generate random numbers from these distributions and compute the distribution of their sums. Use the tests from Chapter 4 to see if the sum can be modeled by a continuous distribution.

6.6 Central limit theorem

So far we have seen:

- If we add values drawn from normal distributions, the distribution of the sum is normal.
- If we add values drawn from other distributions, the sum does not generally have one of the continuous distributions we have seen.

But it turns out that if we add up a large number of values from almost any distribution, the distribution of the sum converges to normal.

More specifically, if the distribution of the values has mean and standard deviation μ and σ, the distribution of the sum is approximately $\mathcal{N}(n\mu, n\sigma^2)$.

This is called the **Central Limit Theorem**. It is one of the most useful tools for statistical analysis, but it comes with caveats:

- The values have to be drawn independently.
- The values have to come from the same distribution (although this requirement can be relaxed).

- The values have to be drawn from a distribution with finite mean and variance, so most Pareto distributions are out.

- The number of values you need before you see convergence depends on the skewness of the distribution. Sums from an exponential distribution converge for small sample sizes. Sums from a lognormal distribution do not.

The Central Limit Theorem explains, at least in part, the prevalence of normal distributions in the natural world. Most characteristics of animals and other life forms are affected by a large number of genetic and environmental factors whose effect is additive. The characteristics we measure are the sum of a large number of small effects, so their distribution tends to be normal.

Exercise 6.12 If I draw a sample, x_1 .. x_n, independently from a distribution with finite mean μ and variance σ^2, what is the distribution of the sample mean:

$$\bar{x} = \frac{1}{n} \sum x_i$$

As n increases, what happens to the variance of the sample mean? Hint: review Section 6.5.

Exercise 6.13 Choose a distribution (one of exponential, lognormal or Pareto) and choose values for the parameter(s). Generate samples with sizes 2, 4, 8, etc., and compute the distribution of their sums. Use a normal probability plot to see if the distribution is approximately normal. How many terms do you have to add to see convergence?

Exercise 6.14 Instead of the distribution of sums, compute the distribution of products; what happens as the number of terms increases? Hint: look at the distribution of the log of the products.

6.7 The distribution framework

At this point we have seen PMFs, CDFs and PDFs; let's take a minute to review. Figure 6.1 shows how these functions relate to each other.

We started with PMFs, which represent the probabilities for a discrete set of values. To get from a PMF to a CDF, we computed a cumulative sum. To be more consistent, a discrete CDF should be called a cumulative mass function (CMF), but as far as I can tell no one uses that term.

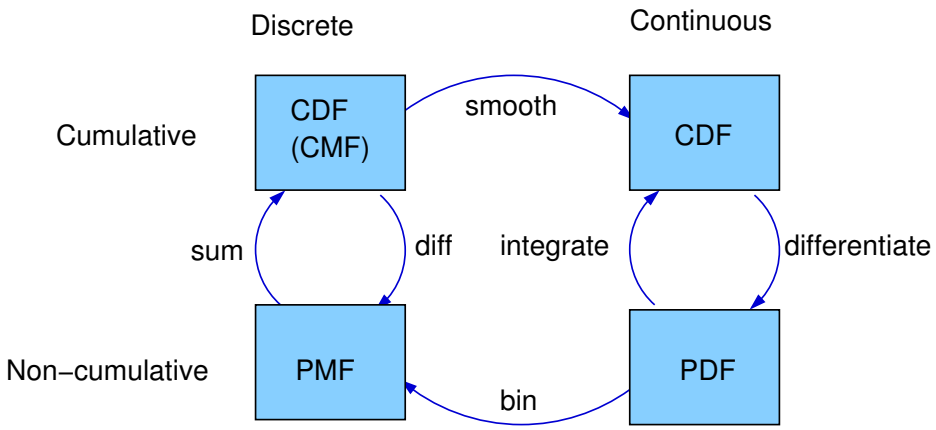

Figure 6.1: A framework that relates representations of distribution functions.

To get from a CDF to a PMF, you can compute differences in cumulative probabilities.

Similarly, a PDF is the derivative of a continuous CDF; or, equivalently, a CDF is the integral of a PDF. But remember that a PDF maps from values to probability densities; to get a probability, you have to integrate.

To get from a discrete to a continuous distribution, you can perform various kinds of smoothing. One form of smoothing is to assume that the data come from an analytic continuous distribution (like exponential or normal) and to estimate the parameters of that distribution. And that's what Chapter 8 is about.

If you divide a PDF into a set of bins, you can generate a PMF that is at least an approximation of the PDF. We use this technique in Chapter 8 to do Bayesian estimation.

Exercise 6.15 Write a function called `MakePmfFromCdf` that takes a Cdf object and returns the corresponding Pmf object.

You can find a solution to this exercise in `thinkstats.com/Pmf.py`.

6.8 Glossary

skewness: A characteristic of a distribution; intuitively, it is a measure of how asymmetric the distribution is.

robust: A statistic is robust if it is relatively immune to the effect of outliers.

illusory superiority: The tendency of people to imagine that they are better than average.

random variable: An object that represents a random process.

random variate: A value generated by a random process.

PDF: Probability density function, the derivative of a continuous CDF.

convolution: An operation that computes the distribution of the sum of values from two distributions.

Central Limit Theorem: "The supreme law of Unreason," according to Sir Francis Galton, an early statistician.