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, $\mu_{1/2}$:

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

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:

$$\mathrm{CDF}_X(x) = P(X \quad x)$$

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, , 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 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 . 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

$$\mathrm{PDF}_{expo}(x) = \quad e^{-\ x}$$

The PDF of a normal distribution is

$$\text{PDF}_{normal}(x) = \frac{1}{\overline{\quad 2}} \exp \quad -\frac{1}{2} \quad \underline{x-\mu} \quad ^2$$

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:

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

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

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

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 falls between 1 and 20? Express your answer as a function of . 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 ^{2} = 59.4 cm for men, and μ = 163 cm and ^{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

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

   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.

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

   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:

$$P(Z \quad z) = \int_{-} P(Z \quad z \mid X = x)\, \text{PDF}_X(x)\, dx$$

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

$$P(Z \quad z) = \quad _{-} \quad \mathrm{CDF}_Y(z - x)\ \mathrm{PDF}_X(x)\ dx$$

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

$$\mathrm{CDF}_Z(z) = \quad _{-} \quad \mathrm{CDF}_Y(z - x)\ \mathrm{PDF}_X(x)\ dx$$

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

$$\mathrm{PDF}_Z(z) = \quad _{-} \quad \mathrm{PDF}_Y(z - x)\ \mathrm{PDF}_X(x)\ dx$$

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 .

$$\mathrm{PDF}_Z = \mathrm{PDF}_Y \quad \mathrm{PDF}_X$$

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 . The distribution of $Z = X + Y$ is:

$$\mathrm{PDF}_Z(z) = \quad _{-} \quad \mathrm{PDF}_X(x)\ \mathrm{PDF}_Y(z - x)\ dx = \quad _{-} \quad e^{-\ x} \quad e^{-\ (z-x)} dx$$

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:

$$\mathrm{PDF}_Z(z) = \quad _{0}^{z} \quad e^{-\ x} \quad e^{-\ (z-x)}\ dx$$

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

$$\mathrm{PDF}_Z(z) = \quad ^2 e^{-\ z} \quad _{0}^{z} dx = \quad ^2 z\, e^{-\ z}$$

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 , and Y has an Erlang distribution with parameters k and , 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 , you can write

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

where the symbol 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 $X \quad \mathcal{N}(\mu, \ ^2)$,

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

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

$$Z \quad \mathcal{N}(\mu_X + \mu_Y, \ {}^2_X + \ {}^2_Y)$$

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

Exercise 6.10 If $X \quad \mathcal{N}(\mu_X, \ {}_X{}^2)$ and $Y \quad \mathcal{N}(\mu_Y, \ {}_Y{}^2)$, 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 , the distribution of the sum is approximately $\mathcal{N}(n\mu, n \ {}^2)$.

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 ^{2}, what is the distribution of the sample mean:

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

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.

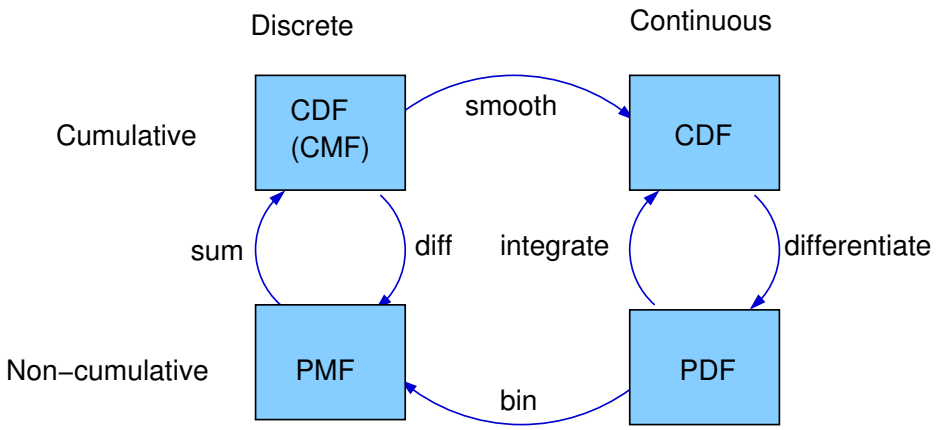

Figure 6.1: A framework that relates representations of distribution functions.

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.

Chapter 7

Hypothesis testing

Exploring the data from the NSFG, we saw several "apparent effects," including a number of differences between first babies and others. So far we have taken these effects at face value; in this chapter, finally, we put them to the test.

The fundamental question we want to address is whether these effects are real. For example, if we see a difference in the mean pregnancy length for first babies and others, we want to know whether that difference is real, or whether it occurred by chance.

That question turns out to be hard to address directly, so we will proceed in two steps. First we will test whether the effect is **significant**, then we will try to interpret the result as an answer to the original question.

In the context of statistics, "significant" has a technical definition that is different from its use in common language. As defined earlier, an apparent effect is statistically significant if it is unlikely to have occurred by chance.

To make this more precise, we have to answer three questions:

1. What do we mean by "chance"?

2. What do we mean by "unlikely"?

3. What do we mean by "effect"?

All three of these questions are harder than they look. Nevertheless, there is a general structure that people use to test statistical significance:

Null hypothesis: The **null hypothesis** is a model of the system based on the assumption that the apparent effect was actually due to chance.

p-value: The **p-value** is the probability of the apparent effect under the null hypothesis.

Interpretation: Based on the p-value, we conclude that the effect is either statistically significant, or not.

This process is called **hypothesis testing**. The underlying logic is similar to a proof by contradiction. To prove a mathematical statement, A, you assume temporarily that A is false. If that assumption leads to a contradiction, you conclude that A must actually be true.

Similarly, to test a hypothesis like, "This effect is real," we assume, temporarily, that is is not. That's the null hypothesis. Based on that assumption, we compute the probability of the apparent effect. That's the p-value. If the p-value is low enough, we conclude that the null hypothesis is unlikely to be true.

7.1 Testing a difference in means

One of the easiest hypotheses to test is an apparent difference in mean between two groups. In the NSFG data, we saw that the mean pregnancy length for first babies is slightly longer, and the mean weight at birth is slightly smaller. Now we will see if those effects are significant.

For these examples, the null hypothesis is that the distributions for the two groups are the same, and that the apparent difference is due to chance.

To compute p-values, we find the pooled distribution for all live births (first babies and others), generate random samples that are the same size as the observed samples, and compute the difference in means under the null hypothesis.

If we generate a large number of samples, we can count how often the difference in means (due to chance) is as big or bigger than the difference we actually observed. This fraction is the p-value.

For pregnancy length, we observed $n = 4413$ first babies and $m = 4735$ others, and the difference in mean was $= 0.078$ weeks. To approximate the p-value of this effect, I pooled the distributions, generated samples with sizes n and m and computed the difference in mean.

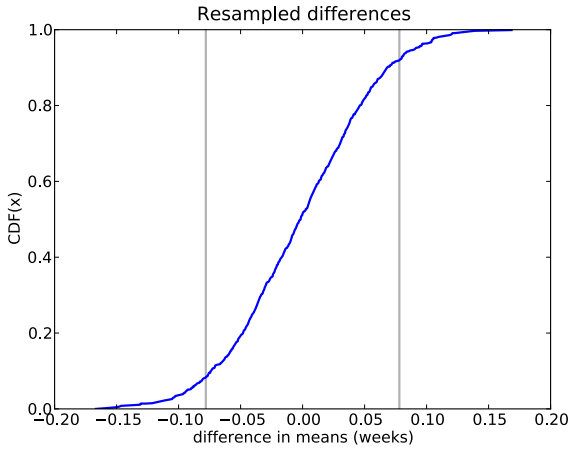

Figure 7.1: CDF of difference in mean for resampled data.

This is another example of resampling, because we are drawing a random sample from a dataset that is, itself, a sample of the general population. I computed differences for 1000 sample pairs; Figure 7.1 shows their distribution.

The mean difference is near 0, as you would expect with samples from the same distribution. The vertical lines show the cutoffs where $x = -$ or $x =$.

Of 1000 sample pairs, there were 166 where the difference in mean (positive or negative) was as big or bigger than , so the p-value is approximately 0.166. In other words, we expect to see an effect as big as about 17% of the time, even if the actual distribution for the two groups is the same.

So the apparent effect is not very likely, but is it unlikely enough? I'll address that in the next section.

Exercise 7.1 In the NSFG dataset, the difference in mean weight for first births is 2.0 ounces. Compute the p-value of this difference.

Hint: for this kind of resampling it is important to sample with replacement, so you should use `random.choice` rather than `random.sample` (see Section 3.8).

You can start with the code I used to generate the results in this section, which you can download from `http://thinkstats.com/hypothesis.py`.

7.2 Choosing a threshold

In hypothesis testing we have to worry about two kinds of errors.

- A Type I error, also called a **false positive**, is when we accept a hypothesis that is actually false; that is, we consider an effect significant when it was actually due to chance.
- A Type II error, also called a **false negative**, is when we reject a hypothesis that is actually true; that is, we attribute an effect to chance when it was actually real.

The most common approach to hypothesis testing is to choose a threshold[1], , for the p-value and to accept as significant any effect with a p-value less than . A common choice for is 5%. By this criterion, the apparent difference in pregnancy length for first babies is not significant, but the difference in weight is.

For this kind of hypothesis testing, we can compute the probability of a false positive explicitly: it turns out to be .

To see why, think about the definition of false positive—the chance of accepting a hypothesis that is false—and the definition of a p-value—the chance of generating the measured effect if the hypothesis is false.

Putting these together, we can ask: if the hypothesis is false, what is the chance of generating a measured effect that will be considered significant with threshold ? The answer is .

We can decrease the chance of a false positive by decreasing the threshold. For example, if the threshold is 1%, there is only a 1% chance of a false positive.

But there is a price to pay: decreasing the threshold raises the standard of evidence, which increases the chance of rejecting a valid hypothesis.

In general there is a tradeoff between Type I and Type II errors. The only way to decrease both at the same time is to increase the sample size (or, in some cases, decrease measurement error).

Exercise 7.2 To investigate the effect of sample size on p-value, see what happens if you discard half of the data from the NSFG. Hint: use `random.sample`. What if you discard three-quarters of the data, and so on?

[1] Also known as a "Significance criterion."

What is the smallest sample size where the difference in mean birth weight is still significant with = 5%? How much larger does the sample size have to be with = 1%?

You can start with the code I used to generate the results in this section, which you can download from `http://thinkstats.com/hypothesis.py`.

7.3 Defining the effect

When something unusual happens, people often say something like, "Wow! What were the chances of *that*?" This question makes sense because we have an intuitive sense that some things are more likely than others. But this intuition doesn't always hold up to scrutiny.

For example, suppose I toss a coin 10 times, and after each toss I write down H for heads and T for tails. If the result was a sequence like THHTHTTTHH, you wouldn't be too surprised. But if the result was HHHHHHHHHH, you would say something like, "Wow! What were the chances of *that*?"

But in this example, the probability of the two sequences is the same: one in 1024. And the same is true for any other sequence. So when we ask, "What were the chances of *that*," we have to be careful about what we mean by "that."

For the NSFG data, I defined the effect as "a difference in mean (positive or negative) as big or bigger than ." By making this choice, I decided to evaluate the magnitude of the difference, ignoring the sign.

A test like that is called **two-sided**, because we consider both sides (positive and negative) in the distribution from Figure 7.1. By using a two-sided test we are testing the hypothesis that there is a significant difference between the distributions, without specifying the sign of the difference.

The alternative is to use a **one-sided** test, which asks whether the mean for first babies is significantly *higher* than the mean for others. Because the hypothesis is more specific, the p-value is lower—in this case it is roughly half.

7.4 Interpreting the result

At the beginning of this chapter I said that the question we want to address is whether an apparent effect is real. We started by defining the null hypothesis, denoted H_0, which is the hypothesis that the effect is not real. Then we

defined the p-value, which is $P(E \mid H_0)$, where E is an effect as big as or bigger than the apparent effect. Then we computed p-values and compared them to a threshold, .

That's a useful step, but it doesn't answer the original question, which is whether the effect is real. There are several ways to interpret the result of a hypothesis test:

Classical: In classical hypothesis testing, if a p-value is less than , you can say that the effect is statistically significant, but you can't conclude that it's real. This formulation is careful to avoid leaping to conclusions, but it is deeply unsatisfying.

Practical: In practice, people are not so formal. In most science journals, researchers report p-values without apology, and readers interpret them as evidence that the apparent effect is real. The lower the p-value, the higher their confidence in this conclusion.

Bayesian: What we really want to know is $P(H_A \mid E)$, where H_A is the hypothesis that the effect is real. By Bayes's theorem

$$P(H_A \mid E) = \frac{P(E \mid H_A)\, P(H_A)}{P(E)}$$

where $P(H_A)$ is the prior probability of H_A before we saw the effect, $P(E \mid H_A)$ is the probability of seeing E, assuming that the effect is real, and $P(E)$ is the probability of seeing E under any hypothesis. Since the effect is either real or it's not,

$$P(E) = P(E \mid H_A)\, P(H_A) + P(E \mid H_0)\, P(H_0)$$

As an example, I'll compute $P(H_A \mid \mathrm{E})$ for pregnancy lengths in the NSFG. We have already computed $P(E \mid H_0) = 0.166$, so all we have to do is compute $P(E \mid H_A)$ and choose a value for the prior.

To compute $P(E \mid H_A)$, we assume that the effect is real—that is, that the difference in mean duration, , is actually what we observed, 0.078. (This way of formulating H_A is a little bit bogus. I will explain and fix the problem in the next section.)

By generating 1000 sample pairs, one from each distribution, I estimated $P(\mathrm{E} \mid H_A) = 0.494$. With the prior $P(H_A) = 0.5$, the posterior probability of H_A is 0.748.

So if the prior probability of H_A is 50%, the updated probability, taking into account the evidence from this dataset, is almost 75%. It makes sense that

the posterior is higher, since the data provide some support for the hypothesis. But it might seem surprising that the difference is so large, especially since we found that the difference in means was not statistically significant.

In fact, the method I used in this section is not quite right, and it tends to overstate the impact of the evidence. In the next section we will correct this tendency.

Exercise 7.3 Using the data from the NSFG, what is the posterior probability that the distribution of birth weights is different for first babies and others?

You can start with the code I used to generate the results in this section, which you can download from `http://thinkstats.com/hypothesis.py`.

7.5 Cross-validation

In the previous example, we used the dataset to formulate the hypothesis H_A, and then we used the same dataset to test it. That's not a good idea; it is too easy to generate misleading results.

The problem is that even when the null hypothesis is true, there is likely to be some difference, , between any two groups, just by chance. If we use the observed value of to formulate the hypothesis, $P(H_A \mid E)$ is likely to be high even when H_A is false.

We can address this problem with **cross-validation**, which uses one dataset to compute and a *different* dataset to evaluate H_A. The first dataset is called the **training set**; the second is called the **testing set**.

In a study like the NSFG, which studies a different cohort in each cycle, we can use one cycle for training and another for testing. Or we can partition the data into subsets (at random), then use one for training and one for testing.

I implemented the second approach, dividing the Cycle 6 data roughly in half. I ran the test several times with different random partitions. The average posterior probability was $P(H_A \mid E) = 0.621$. As expected, the impact of the evidence is smaller, partly because of the smaller sample size in the test set, and also because we are no longer using the same data for training and testing.

7.6 Reporting Bayesian probabilities

In the previous section we chose the prior probability $P(H_A) = 0.5$. If we have a set of hypotheses and no reason to think one is more likely than another, it is common to assign each the same probability.

Some people object to Bayesian probabilities because they depend on prior probabilities, and people might not agree on the right priors. For people who expect scientific results to be objective and universal, this property is deeply unsettling.

One response to this objection is that, in practice, strong evidence tends to swamp the effect of the prior, so people who start with different priors will converge toward the same posterior probability.

Another option is to report just the **likelihood ratio**, $P(\mathrm{E} \mid H_A) \ / \ P(E \mid H_0)$, rather than the posterior probability. That way readers can plug in whatever prior they like and compute their own posteriors (no pun intended). The likelihood ratio is sometimes called a Bayes factor (see `http://wikipedia.org/wiki/Bayes_factor`).

Exercise 7.4 If your prior probability for a hypothesis, H_A, is 0.3 and new evidence becomes available that yields a likelihood ratio of 3 relative to the null hypothesis, H_0, what is your posterior probability for H_A?

Exercise 7.5 This exercise is adapted from MacKay, *Information Theory, Inference, and Learning Algorithms*:

> Two people have left traces of their own blood at the scene of a crime. A suspect, Oliver, is tested and found to have type O blood. The blood groups of the two traces are found to be of type O (a common type in the local population, having frequency 60%) and of type AB (a rare type, with frequency 1%). Do these data (the blood types found at the scene) give evidence in favor of the proposition that Oliver was one of the two people whose blood was found at the scene?

Hint: Compute the likelihood ratio for this evidence; if it is greater than 1, then the evidence is in favor of the proposition. For a solution and discussion, see page 55 of MacKay's book.

7.7 Chi-square test

In Section 7.2 we concluded that the apparent difference in mean pregnancy length for first babies and others was not significant. But in Section 2.10, when we computed relative risk, we saw that first babies are more likely to be early, less likely to be on time, and more likely to be late.

So maybe the distributions have the same mean and different variance. We could test the significance of the difference in variance, but variances are less robust than means, and hypothesis tests for variance often behave badly.

An alternative is to test a hypothesis that more directly reflects the effect as it appears; that is, the hypothesis that first babies are more likely to be early, less likely to be on time, and more likely to be late.

We proceed in five easy steps:

1. We define a set of categories, called **cells**, that each baby might fall into. In this example, there are six cells because there are two groups (first babies and others) and three bins (early, on time or late).

   I'll use the definitions from Section 2.10: a baby is early if it is born during Week 37 or earlier, on time if it is born during Week 38, 39 or 40, and late if it is born during Week 41 or later.

2. We compute the number of babies we expect in each cell. Under the null hypothesis, we assume that the distributions are the same for the two groups, so we can compute the pooled probabilities: P(early), P(ontime) and P(late).

   For first babies, we have n = 4413 samples, so under the null hypothesis we expect n P(early) first babies to be early, n P(ontime) to be on time, etc. Likewise, we have m = 4735 other babies, so we expect m P(early) other babies to be early, etc.

3. For each cell we compute the deviation; that is, the difference between the observed value, O_i, and the expected value, E_i.

4. We compute some measure of the total deviation; this quantity is called the **test statistic**. The most common choice is the chi-square statistic:

$$^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

5. We can use a Monte Carlo simulation to compute the p-value, which is the probability of seeing a chi-square statistic as high as the observed value under the null hypothesis.

When the chi-square statistic is used, this process is called a **chi-square test**. One feature of the chi-square test is that the distribution of the test statistic can be computed analytically.

Using the data from the NSFG I computed $^2 = 91.64$, which would occur by chance about one time in 10,000. I conclude that this result is statistically significant, with one caution: again we used the same dataset for exploration and testing. It would be a good idea to confirm this result with another dataset.

You can download the code I used in this section from `http://thinkstats.com/chi.py`.

Exercise 7.6 Suppose you run a casino and you suspect that a customer has replaced a die provided by the casino with a "crooked die;" that is, one that has been tampered with to make one of the faces more likely to come up than the others. You apprehend the alleged cheater and confiscate the die, but now you have to prove that it is crooked.

You roll the die 60 times and get the following results:

Value	1	2	3	4	5	6
Frequency	8	9	19	6	8	10

What is the chi-squared statistic for these values? What is the probability of seeing a chi-squared value as large by chance?

7.8 Efficient resampling

Anyone reading this book who has prior training in statistics probably laughed when they saw Figure 7.1, because I used a lot of computer power to simulate something I could have figured out analytically.

Obviously mathematical analysis is not the focus of this book. I am willing to use computers to do things the "dumb" way, because I think it is easier for beginners to understand simulations, and easier to demonstrate that they are correct. So as long as the simulations don't take too long to run, I don't feel guilty for skipping the analysis.

However, there are times when a little analysis can save a lot of computing, and Figure 7.1 is one of those times.

Remember that we were testing the observed difference in the mean between pregnancy lengths for $n = 4413$ first babies and $m = 4735$ others. We formed the pooled distribution for all babies, drew samples with sizes n and m, and computed the difference in sample means.

Instead, we could directly compute the distribution of the difference in sample means. To get started, let's think about what a sample mean is: we draw n samples from a distribution, add them up, and divide by n. If the distribution has mean μ and variance ^{2}, then by the Central Limit Theorem, we know that the sum of the samples is $\mathcal{N}(n\mu, n\ ^2)$.

To figure out the distribution of the sample means, we have to invoke one of the properties of the normal distribution: if X is $\mathcal{N}(\mu,\ ^2)$,

$$aX + b \quad \mathcal{N}(a\mu + b, a^2\ ^2)$$

When we divide by n, $a = 1/n$and $b = 0$, so

$$X/n \quad \mathcal{N}(\mu/n,\ ^2/\ n^2)$$

So the distribution of the sample mean is $\mathcal{N}(\mu,\ ^2/n)$.

To get the distribution of the difference between two sample means, we invoke another property of the normal distribution: if X_1 is $\mathcal{N}(\mu_1,\ _1{}^2)$ and X_2 is $\mathcal{N}(\mu_2,\ _2{}^2)$,

$$aX_1 + bX_2 \quad \mathcal{N}(a\mu_1 + b\mu_2, a^2\ _1^2 + b^2\ _2^2)$$

So as a special case:

$$X_1 - X_2 \quad \mathcal{N}(\mu_1 - \mu_2,\ _1^2 +\ _2^2)$$

Putting it all together, we conclude that the sample in Figure 7.1 is drawn from $\mathcal{N}(0, f\ ^2)$, where $f = 1/n + 1/m$. Plugging in $n = 4413$ and $m = 4735$, we expect the difference of sample means to be $\mathcal{N}(0, 0.0032)$.

We can use `erf.NormalCdf` to compute the p-value of the observed difference in the means:

```
delta = 0.078
sigma = math.sqrt(0.0032)
left = erf.NormalCdf(-delta, 0.0, sigma)
right = 1 - erf.NormalCdf(delta, 0.0, sigma)
```

The sum of the left and right tails is the p-value, 0.168, which is pretty close to what we estimated by resampling, 0.166. You can download the code I used in this section from `http://thinkstats.com/hypothesis_analytic.py`

7.9 Power

When the result of a hypothesis test is negative (that is, the effect is not statistically significant), can we conclude that the effect is not real? That depends on the power of the test.

Statistical **power** is the probability that the test will be positive if the null hypothesis is false. In general, the power of a test depends on the sample size, the magnitude of the effect, and the threshold .

Exercise 7.7 What is the power of the test in Section 7.2, using = 0.05 and assuming that the actual difference between the means is 0.078 weeks?

You can estimate power by generating random samples from distributions with the given difference in the mean, testing the observed difference in the mean, and counting the number of positive tests.

What is the power of the test with = 0.10?

One way to report the power of a test, along with a negative result, is to say something like, "If the apparent effect were as large as x, this test would reject the null hypothesis with probability p."

7.10 Glossary

significant: An effect is statistically significant if it is unlikely to occur by chance.

null hypothesis: A model of a system based on the assumption that an apparent effect is due to chance.

p-value: The probability that an effect could occur by chance.

hypothesis testing: The process of determining whether an apparent effect is statistically significant.

false positive: The conclusion that an effect is real when it is not.

false negative: The conclusion that an effect is due to chance when it is not.

two-sided test: A test that asks, "What is the chance of an effect as big as the observed effect, positive or negative?"

one-sided test: A test that asks, "What is the chance of an effect as big as the observed effect, and with the same sign?"

cross-validation: A process of hypothesis testing that uses one dataset for exploratory data analysis and another dataset for testing.

training set: A dataset used to formulate a hypothesis for testing.

testing set: A dataset used for testing.

test statistic: A statistic used to measure the deviation of an apparent effect from what is expected by chance.

chi-square test: A test that uses the chi-square statistic as the test statistic.

likelihood ratio: The ratio of $P(E \mid A)$ to $P(E \mid B)$ for two hypotheses A and B, which is a way to report results from a Bayesian analysis without depending on priors.

cell: In a chi-square test, the categories the observations are divided into.

power: The probability that a test will reject the null hypothesis if it is false.